RISK-SENSITIVE PREFERENCE GAMES: LEARNING FROM STRATEGIC INTERACTION & UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

A growing line of work reframes preference-based fine-tuning of large language models game-theoretically: Nash Learning from Human Feedback (NLHF) recasts the problem as a zero-sum game over policies. However, optimization is over expected pairwise payoffs, thereby conflating policies with similar win rates but different tail behavior. As such, these methods are agnostic to where in the data distribution they succeed or fail: strong average performance can mask systematic failure across prompts, annotators, or safety-critical strata. We introduce *risk-sensitive preference games*, in which players optimize convex risk measures of their preference loss, exploiting structure in preference uncertainty. While risk-sensitivity generally breaks the zero-sum structure, we show that translation invariance of many risk metrics ensures that we retain monotonicity, yielding fast convergence of sample-efficient self-play methods. Furthermore, we establish algorithmic stability and offline sample complexity bounds that scale with risk, requiring simultaneous control of structural bias from nonlinear risk transformations, statistical bias in risk estimation, and concentration tailored to the risk-sensitive setting. To address statistical bias, we introduce a hierarchical game formulation and a two-timescale extragradient algorithm with bias correction that converges to the Stackelberg equilibrium and is especially effective in low-sample regimes. Empirically, risk-adjusted policies are robust across data strata, stable across risk choices, and match or exceed risk-neutral performance thereby achieving robustness without a performance tax.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed at scale, and with that scale comes a growing record of harm: chatbots coaching teenagers toward suicide (Hill, 2025a); ordinary adults with no prior history of mental illness driven into weeks-long delusional spirals—over simulation theory, AI sentience (Hill, 2025b), fabricated mathematical frameworks that a Toronto recruiter spent three weeks publicly broadcasting to cryptography experts and national security agencies before the spell broke (Hill & Freedman, 2025); models that learn to identify and manipulate the small fraction of users most vulnerable to manipulation (Williams et al., 2025), systems that confidently reinforce delusions while satisfying every aggregate benchmark.

These are not isolated failures to be patched one at a time. They are the *canary in the mine* for a deeper problem in how alignment research is conceived. The dominant paradigm—RLHF and its variants—collapses rich, heterogeneous human preference data into a single scalar reward and optimizes its expectation—see, e.g., (Ouyang et al., 2022; Ziegler et al., 2020; Bai et al., 2022; Munos et al., 2024; Christiano et al., 2017b; Stiennon et al., 2020; Ziegler et al., 2019). The structure of human disagreement, the shape of the response distribution, the existence of a long tail of users whose needs differ from the median: all of this information is in the data we collect, and all of it is averaged away before the model ever sees it. Alignment, framed this way, can only be aligned with a fictitious average human; *it has no language for the tails where harm actually occurs, and no language for the diversity of legitimate responses that real users require*.

We argue that alignment research must be reframed. The objective is not to chase down failure modes after the fact, nor to align to a representative user, but to make the tail behavior and the diversity of the response distribution quantities the learning algorithm explicitly reasons about, rather

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

than artifacts it discards. This motivates our central contribution: *extending RLHF with risk-sensitive objectives, so that the equilibrium concept itself reasons about the distribution of outcomes rather than collapsing it to a mean*.

Building on RLHF methods that optimize human preference feedback via scalar rewards (Christiano et al., 2017a; Ziegler et al., 2020; Rafailov et al., 2023), recent work has extended preference optimization to a game-theoretic setting. NLHF (Munos et al., 2024) recasts alignment as a two-player constant-sum game in which the preference oracle defines the payoff. With a KL penalty, the resulting solution concept is a *quantal response equilibrium* (QRE) (McKelvey & Palfrey, 1995), characterized by a fixed-point equation targeted by Nash-MD (Munos et al., 2024) and follow-up algorithms including IPO-MD (Calandriello et al., 2024b), MPO (Wang et al., 2025), and EGPO (Zhou et al., 2025). NLHF recovers the *strategic interaction* structure that RLHF discards: it sees a preference distribution, not a point estimate. But it stops there. Methods in this line optimize the *expected* pairwise payoff and remain blind to the second piece of structure that preference data carries: *uncertainty*. A response strongly preferred by part of the population and dispreferred by another is, in expectation, indistinguishable from a broadly acceptable one; a policy that wins on average while occasionally producing catastrophic outputs is conflated with one that never does.

This work closes that gap by introducing *risk-sensitive preference games* (RSPG). Rather than optimizing the expected pairwise payoff, RSPG players reason about the full distribution of preference outcomes—variance, tails, and worst-case responses—yielding equilibria sensitive to exactly the features that scalar-reward RLHF and expectation-based NLHF discard.

Contributions. Beyond introducing RSPGs, we develop algorithms for learning equilibria. RSPGs replace expected payoffs with *convex risk measures* (Föllmer & Schied, 2002; 2016). This changes what alignment optimizes: the choice of risk measure encodes the structure of the harms of interest, making the corresponding features of the outcome distribution explicit optimization targets. Our contributions overcome two obstacles introduced by this change. The first is structural: nonlinear risk generally breaks the constant-sum property that makes computationally efficient single-policy self-play sufficient for equilibrium computation in NLHF. The second is statistical: most risk measures do not admit unbiased sample estimators, so stochastic gradient methods inherit a persistent bias floor that does not vanish with the step size—especially damaging in the low-sample regime of LLM fine-tuning, where data is expensive and bias compounds across iterates. We resolve both.

- **Resolving the structural obstacle.** *Translation invariance*, a property of convex risk measures, preserves constant-sum structure of the risk-adjusted payoff-operator (Lemma 1), even as it fails at the certainty-equivalent level. Symmetric self-play thus retains the computational benefits of maintaining one policy rather than two, while yielding a robust solution (Theorem 1). We characterize $\mu_{\mathcal{R}}$ strong monotonicity of RSPGs (Theorem 2), and therefore existence of a unique equilibrium; prove $\mathcal{O}((1-\eta\mu_{\mathcal{R}})^t)$ last-iterate convergence of deterministic extragradient, and linear contraction to a neighborhood of size $\mathcal{O}(1/(m\mu_{\mathcal{R}}))$ for stochastic-biased extragradient (Proposition 2, and Theorem 8 and 9 (Appendix H) for the merely- and strongly-monotone settings.
- **Resolving the statistical obstacle.** A Stackelberg game with a fast follower (Theorem 3, Appendix I) reduces the bias of plug-in risk estimators to the variance floor, especially valuable in the low-sample m regime. For smooth risks, operator averaging attains $\mathcal{O}(T^{-1}+m^{-4})$ last-iterate convergence, matching the optimal rate of *unbiased* stochastic methods in the strongly monotone regime (Corollary 10). Algorithmic stability (Theorem 15, Appendix J) and an offline sample-complexity guarantee of $\widetilde{\mathcal{O}}(1/n)$ (Theorem 4) match the risk-neutral setting up to risk-dependent constants, so risk introduces no additional sample-complexity bottleneck, and is explicitly tunable.
- **Empirical Pareto-dominance.** Risk-adjusted policies are robust across strata and match or exceed risk-neutral baselines (Section 4): safety and combined win-rates improve consistently across opponents, metrics, and held-out harm strata, with *no statistically significant* drop in win-rate. Two-timescale bias tracking reduces finite-sample risk-estimation bias and improves severe-prompt tail performance without a detectable loss in preference win-rate (Appendix L.8.2). Risk sensitivity is tunable, with quantified consequences for learning, convergence, and generalization.

2 Risk-Sensitive Preference Game Preliminaries

Let $\mathcal{X}, \mathcal{Y}$ denote a finite set of prompts and responses, respectively. A policy is a probability distribution $\pi \in \Delta(\mathcal{Y})$. Fix a reference policy $\pi_{\text{ref}} \in \Delta^\circ(\mathcal{Y})$, where $\Delta^\circ(\mathcal{Y})$ denotes the relative interior,

and a reference distribution $\mu \in \Delta(\mathcal{Y})$ over opponents. To simplify notation we often treat the $|\mathcal{X}| = 1$ case; all results extend to $|\mathcal{X}| > 1$ by prompt conditioning.

The NLHF framework (Munos et al., 2024) formulates preference-based fine-tuning as a two-player constant-sum matrix game. It assumes access to preference model $\mathcal{P} : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ mapping pairs of completions to preference probabilities, which are collected into matrix $P = [\mathcal{P}(y \succ y')]_{(y,y')}$. Critically, since $\mathcal{P}$ outputs probabilities, $P + P^\top = \mathbf{1}\mathbf{1}^\top$. In places where clear from context and notationally convenient, we conflate P and $\mathcal{P}$. If both players choose policies $\pi_1, \pi_2 \in \Delta(\mathcal{Y})$, then player 1's payoff against player 2 is $V(\pi_1, \pi_2) = \mathbb{E}_{y_1 \sim \pi_1, y_2 \sim \pi_2}[\mathcal{P}(y_1 \succ y_2)] = \pi_1^\top P \, \pi_2$ and

$$V_\beta(\pi_1, \pi_2) = \pi_1^\top P \, \pi_2 - \beta \, \mathrm{KL}(\pi_1 \,\|\, \pi_{\mathrm{ref}}) + \beta \, \mathrm{KL}(\pi_2 \,\|\, \pi_{\mathrm{ref}}). \tag{1}$$

The KL-regularized game admits a unique QRE $\pi^\star_\beta$, characterized by the fixed-point condition over logits θ, where $\pi \propto \exp(\theta)$:

$$\theta^\star_\beta = \theta_{\mathrm{ref}} + \mathcal{P}\pi^\star_\beta/\beta. \tag{2}$$

Note that the equilibrium logits equal the reference logits plus a term that is *linear* in the opponent's policy through the operator $\mathcal{P}/\beta$. This is the fixed point every NLHF algorithm is chasing. The form of $\theta^\star_\beta$ relies on *both* the constant-sum structure $P + P^\top = \mathbf{1}\mathbf{1}^\top$ and the KL-regularization in (1). We provide an overview of related game-theoretic algorithms in Appendix C.

2.1 Risk-Sensitive Preference Games

We now introduce risk into the NLHF framework. Replacing the expected pairwise payoff with a risk functional yields a risk-adjusted preference operator and shifts the solution concept to the risk-sensitive quantal response equilibrium (RQRE) (Mazumdar et al., 2025; Gonzales et al., 2026). This formulation lets the player express sensitivity to the variability of preference outcomes rather than only their mean, addressing the tail-behavior limitations of expectation-based alignment discussed in Sec. 1. To formalize this, we draw on the class of convex risk measures studied extensively in mathematical finance Föllmer & Schied (2002), the definition of which is stated here for payoffs.

Definition 1 (Risk adjusted payoff functional)**.** A risk functional $\mathcal{R} : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ is a risk-adjusted payoff functional if it satisfies the following properties:

a. Concavity: For all $X, Y \in L^\infty$ and $\lambda \in [0, 1]$, $\mathcal{R}(\lambda X + (1-\lambda)Y) \geq \lambda\mathcal{R}(X) + (1-\lambda)\mathcal{R}(Y)$.

b. Monotonicity: For all $X, Y \in L^\infty$ such that $X \leq Y$, $\mathcal{R}(X) \leq \mathcal{R}(Y)$.

c. Translation invariance: For all $X \in L^\infty$ and $c \in \mathbb{R}$, $\mathcal{R}(X + c) = \mathcal{R}(X) + c$.

Without loss of generality, we normalize so that $\mathcal{R}(0) = 0$. To connect this convention to the standard convex-risk convention on losses, if L is a loss we define $\varrho(L) := -\mathcal{R}(-L)$. Then larger $\varrho(L)$ is worse, ϱ is convex, and $\varrho(L + c) = \varrho(L) + c$.

The risk-adjusted preference operator & RQRE. Define the *risk-adjusted preference operator*

$$(\mathcal{P}_\mathcal{R} \, \mu)_y \;:=\; \mathcal{R}[P_{y,y''} \mid y'' \sim \mu], \tag{3}$$

which replaces the expected win-rate of y against $y'' \sim \mu$ with the risk-adjusted analog. Letting $\mathcal{P}^i_\mathcal{R}$ denote player i's risk-adjusted preference operator, each player i optimizes a risk-adjusted value

$$V_{\beta,\mathcal{R}}(\pi_1, \pi_2) = \pi_i^\top \mathcal{P}^i_\mathcal{R} \pi_{-i} - \beta \, \mathrm{KL}(\pi_i \| \pi_{\mathrm{ref}}) + \beta \, \mathrm{KL}(\pi_{-i} \| \pi_{\mathrm{ref}}). \tag{4}$$

A *risk-adjusted equilibrium* $(\theta^\star_1, \theta^\star_2)$ satisfies

$$\theta^\star_1 = \theta_{\mathrm{ref}} + \mathcal{P}^1_\mathcal{R} \, \pi_{\theta^\star_2}/\beta, \qquad \theta^\star_2 = \theta_{\mathrm{ref}} + \mathcal{P}^2_\mathcal{R} \, \pi_{\theta^\star_1}/\beta. \tag{5}$$

Taking $\mathcal{R} \equiv \mathbb{E}$ recovers the NLHF game (1) and its QRE fixed point as a special case of (5).

Choice of Loss. The risk framework we develop is not specific to any one preference loss. As shown in Appendix E, the framework applies to the standard family of preference-tuning objectives i.e., IPO, DPO, GRPO, KTO, and others (Calandriello et al., 2024b; Rafailov et al., 2023; Ethayarajh et al., 2024; Shao et al., 2024), by replacing expectation with a risk measure. Here, we focus on IPO Azar et al. (2024) for two reasons. It is the natural choice for game-theoretic analysis: its gradient is a preconditioned residual of the QRE fixed-point equation, and the opponent distribution μ enters distinctly from the sampling distribution ρ, allowing us to directly add risk over the

opponent while reaping the rewards of preconditioning. Second, existing game-theoretic algorithms (Nash-MD (Munos et al., 2024), IPO-MD (Calandriello et al., 2024b), MPO (Wang et al., 2025), EGPO (Zhou et al., 2025)) are formulated in the IPO framework, thereby enabling direct comparison theoretically and empirically.

Risk-Adjusted IPO Objective. Write $\boldsymbol{\mathcal{R}} = (\mathcal{R}^{\mathcal{W}}, \mathcal{R}^{\mu})$ for payoff-side risk functionals acting inside the preference operator, and $\boldsymbol{\varrho} = (\varrho^{\mathcal{X}}, \varrho^{\rho})$ for loss-side risk measures acting on the squared training residual, with $\varrho(L) = -\mathcal{R}(-L)$ when derived from the same primitive. Here $\mathcal{X}$ is the prompt space, ρ is the response sampling distribution, $\mathcal{W}$ broadly captures stochasticity in groups (e.g., annotators, safety, harm, etc.), and μ is the opponent distribution. Define

$$\ell_\theta^{\boldsymbol{\mathcal{R}}}(x, y, y'; \mu) := \left(\log \frac{\pi_\theta(y|x)\pi_{\text{ref}}(y'|x)}{\pi_\theta(y'|x)\pi_{\text{ref}}(y|x)} - \frac{1}{\beta}\left((\mathcal{P}_{\mathcal{R}^\mu, \mathcal{W}}\mu)_{y|x} - (\mathcal{P}_{\mathcal{R}^\mu, \mathcal{W}}\mu)_{y'|x}\right)\right)^2,$$

where the inner risk-adjusted preference operator is

$$(\mathcal{P}_{\mathcal{R}^\mu, \mathcal{W}}\mu)_y^x = \mathcal{R}_{w|x}^{\mathcal{W}}\left[\mathcal{R}_{y''\sim\mu}^{\mu}[\mathbb{P}(y \succ y'' \mid x, w)]\right].$$

Setting $\varrho_o \equiv \varrho_{x\sim\rho_{\mathcal{X}}}^{\mathcal{X}} \circ \varrho_{(y,y')\sim\rho|x}^{\rho}$, $\mathcal{R}_o \equiv \mathcal{R}_{x\sim\rho_{\mathcal{X}}}^{\mathcal{X}} \circ \mathcal{R}_{(y,y')\sim\rho|x}^{\rho}$, the risk-adjusted IPO objective is

$$\mathcal{L}_{\text{IPO}}^{\boldsymbol{\varrho},\boldsymbol{\mathcal{R}}}(\theta; \rho, \mu) = \varrho_o\left[\ell_\theta^{\boldsymbol{\mathcal{R}}}(x, y, y'; \mu)\right] = -\mathcal{R}_o\left[-\ell_\theta^{\boldsymbol{\mathcal{R}}}(x, y, y'; \mu)\right], \tag{6}$$

Setting components of $\boldsymbol{\mathcal{R}}$ or $\boldsymbol{\varrho}$ to $\mathbb{E}$ recovers the corresponding risk-neutral aggregation, and setting all four to $\mathbb{E}$ recovers the standard, risk-neutral IPO loss (Munos et al., 2024; Calandriello et al., 2024b). The two inner payoff risks $\mathcal{R}^{\mathcal{W}}, \mathcal{R}^{\mu}$ enter the preference operator $\mathcal{P}$ thereby modifying the equilibrium (5). The two outer loss risks $\varrho^{\mathcal{X}}, \varrho^{\rho}$ leave the per-prompt fixed point unchanged and act on the training dynamics. We focus on $\mathcal{R}^{\mu}$, which captures strategic uncertainty. The remaining three risk choices and composition rules are formalized in Appendix D.

To simplify presentation, we instantiate (6) with risk only over the opponent, and write $\mathcal{R} := \mathcal{R}^{\mu}$ and $\mathcal{P}_{\mathcal{R}} := \mathcal{P}_{\mathcal{R}^\mu}$ for the rest of the main body; the appendix treats the more general case. In this specialization, the outer training objective remains risk-neutral and the payoff-side risk enters only through $\mathcal{P}_{\mathcal{R}}$. The resulting *opponent-risk IPO loss* is

$$\mathcal{L}_{\text{IPO}}^{\mathcal{R}}(\theta; \rho, \mu) = \mathbb{E}_{(y,y')\sim\rho}\left[\left(\log \frac{\pi_\theta(y)\pi_{\text{ref}}(y')}{(\pi_\theta(y')\pi_{\text{ref}}(y)} - \frac{1}{\beta}\left((\mathcal{P}_{\mathcal{R}}\,\mu)_y - (\mathcal{P}_{\mathcal{R}}\,\mu)_{y'}\right)\right)^2\right]. \tag{7}$$

A direct calculation gives the gradient

$$\nabla_\theta \mathcal{L}_{\text{IPO}}^{\mathcal{R}}(\theta; \rho, \mu) = 2\,\Sigma(\rho)(\theta - \theta_{\text{ref}} - \mathcal{P}_{\mathcal{R}}\mu/\beta),$$

where $\Sigma(\rho) := \mathbb{E}_{(y,y')\sim\rho}[(\mathbf{1}_y - \mathbf{1}_{y'})(\mathbf{1}_y - \mathbf{1}_{y'})^\top]$ is the response-pair preconditioner. The operator $\mathcal{P}_{\mathcal{R}}$ is generally nonlinear in μ, and the following examines its structural effect on the game.

2.2 Translation invariance preserves the constant-sum structure

The risk-neutral analysis of NLHF relies on the constant-sum identity $P + P^\top = \mathbf{1}\mathbf{1}^\top$ to collapse (5) into $\theta_\beta^\star = \theta_{\text{ref}} + \mathcal{P}\pi_\beta^\star/\beta$. Nonlinear risk breaks this structure at the certainty equivalence level.

Example 1 (Risk-adjusted payoff asymmetry)**.** Let $X \in 0, 1$ with $\mathbb{P}(X = 1) = \mathbb{P}(X = 0) = \frac{1}{2}$, so $X + (-X) = 0$. Nonlinear risk transformations break this structure: for any lower-tail mass $q \in (0, 1/2]$, $\text{CVaR}_q(X) = 0$, and $\text{CVaR}_q(-X) = -1$, so $\text{CVaR}_q(X) + \text{CVaR}_q(-X) \neq 0$.

Despite this shortcoming, when $\mathcal{R}$ is translation-invariant, the identity $\mathcal{R}(1 - X) = 1 + \mathcal{R}(-X)$ ensures that constant-sum structure is preserved at the level of risk-adjusted payoff operators.

Lemma 1 (Operator-level constant-sum)**.** Let $P \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ satisfy $P + P^\top = \mathbf{1}\mathbf{1}^\top$. Let $\mathcal{R}$ be translation invariant, i.e., $\mathcal{R}(X + c) = \mathcal{R}(X) + c$ for all constants $c \in \mathbb{R}$. For $\mu \in \Delta(\mathcal{Y})$, define the risk-adjusted payoff operators $(\mathcal{P}_{\mathcal{R}}^1\mu)_a := \mathcal{R}(\mathcal{P}_{a,Y'})$, and $(\mathcal{P}_{\mathcal{R}}^2\mu)_a := \mathcal{R}(-\mathcal{P}_{Y',a})$, with $Y' \sim \mu$. Then, for every $\mu \in \Delta(\mathcal{Y})$, we have $\mathcal{P}_{\mathcal{R}}^1\mu = \mathbf{1} + \mathcal{P}_{\mathcal{R}}^2\mu$.

Preservation of constant sum structure at the operator level is sufficient to prove the equilibrium set is invariant under player exchange, and more importantly, that self-play fixed points lift to symmetric Nash equilibrium of the RSPG. To see this, consider the *single-player risk-adjusted operator*

$$F_{\mathcal{R}}(\theta) := F_{\mathcal{R}}(\theta; \pi_\theta) = \beta(\theta - \theta_{\text{ref}}) - \mathcal{P}_{\mathcal{R}}\,\pi_\theta,$$

which is the residual of player 1's fixed point equation if the opponent is set to be player 1's policy.

Theorem 1 (Self-play fixed points lift to symmetric Nash equilibria)**.** Under the assumptions of Lemma 1, suppose that $\theta^\star$ solves the self-play fixed-point equation $\theta^\star = \theta_{\text{ref}} + \mathcal{P}^1_{\mathcal{R}}\pi_{\theta^\star}/\beta$. Then $\theta^\star$ induces a Nash equilibrium of the original two-player RSPG: if $\pi^\star := \pi_{\theta^\star}$, $\theta^\star_1 := \theta^\star$ and $\theta^\star_2 := \theta^\star - \frac{1}{\beta}\mathbf{1}$, then $\pi_{\theta^\star_1} = \pi_{\theta^\star_2} = \pi^\star$ and $(\theta^\star_1, \theta^\star_2)$ satisfies the coupled RSPG fixed-point equations (5).

Proof. The player-1 condition follows immediately from the self-play fixed point: $\theta^\star_1 = \theta^\star = \theta_{\text{ref}} + \mathcal{P}^1_{\mathcal{R}}\pi_{\theta^\star}/\beta = \theta_{\text{ref}} + \mathcal{P}^1_{\mathcal{R}}\pi_{\theta^\star_2}/\beta$, where the last equality uses softmax gauge invariance, $\pi_{\theta^\star_2} = \pi_{\theta^\star - \beta^{-1}\mathbf{1}} = \pi_{\theta^\star}$. For player 2, Lemma 1 yields $\mathcal{P}^1_{\mathcal{R}}\pi^\star = \mathbf{1} + \mathcal{P}^2_{\mathcal{R}}\pi^\star$. Therefore

$$\theta^\star_2 = \theta^\star - \mathbf{1}/\beta = \theta_{\text{ref}} + \mathcal{P}^1_{\mathcal{R}}\pi^\star/\beta - \mathbf{1}/\beta = \theta_{\text{ref}} + \mathcal{P}^2_{\mathcal{R}}\pi^\star/\beta = \theta_{\text{ref}} + \mathcal{P}^2_{\mathcal{R}}\pi_{\theta^\star_1}/\beta.$$

Thus the fixed-point equations hold, so $(\theta^\star_1, \theta^\star_2)$ is a Nash equilibrium of the two-player RSPG. □

Proposition 1 (Equilibrium-set symmetry)**.** Under the hypotheses of Lemma 1, the RSPG equilibrium set is invariant under player exchange, modulo the softmax gauge. Consequently, if the RSPG equilibrium is unique in policy space, then it is symmetric.

Monotonicity of the operator $F_{\mathcal{R}}$ determines not only uniqueness but also the rate of convergence for many algorithms. In the symmetric self-play setting, monotonicity is determined by displacements within the tangent space $T^{\mathsf{sym}} = \{(\xi, \xi) : \xi \in \mathbf{1}^\perp\}$ of the diagonal constraint set, where it reduces to a single-player condition:

$$\begin{aligned}\langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2),\, \pi_{\theta_1} - \pi_{\theta_2}\rangle &= \beta\langle \theta_1 - \theta_2,\, \pi_{\theta_1} - \pi_{\theta_2}\rangle - \langle \mathcal{P}_{\mathcal{R}}\pi_{\theta_1} - \mathcal{P}_{\mathcal{R}}\pi_{\theta_2},\, \pi_{\theta_1} - \pi_{\theta_2}\rangle \\ &= \beta\big(\mathrm{KL}(\pi_{\theta_1}\|\pi_{\theta_2}) + \mathrm{KL}(\pi_{\theta_2}\|\pi_{\theta_1})\big) - (\pi_{\theta_1} - \pi_{\theta_2})^\top \mathcal{J}_{\mathcal{R}}(\bar\pi)(\pi_{\theta_1} - \pi_{\theta_2}),\end{aligned}$$

where $\mathcal{J}_{\mathcal{R}}(\pi) := \frac{1}{2}(J_{\mathcal{R}} + J_{\mathcal{R}}^\top)(\pi) - \frac{1}{2}\mathbf{1}\mathbf{1}^\top$ is the distortion matrix, $\bar\pi$ is a policy on the line segment between π_{θ_1} and π_{θ_2}, and $J_{\mathcal{R}}(\pi) := \nabla_\mu P_{\mathcal{R}}(\mu)|_{\mu=\pi}$ is the Jacobian. The first term is bounded below by strong convexity of the log partition function; the second is controlled by the worst-case risk-distortion eigenvalue restricted to T^{sym}. Define the worst-case risk-distortion eigenvalue on the symmetric tangent space $\bar\lambda_{\mathcal{R}} := \sup_{\bar\pi\in\Delta(\mathcal{Y})} \sup\left\{\frac{1}{2}\xi^\top\big(J_{\mathcal{R}}(\bar\pi) + J_{\mathcal{R}}(\bar\pi)^\top\big)\xi \;\middle|\; \xi \in \mathbf{1}^\perp,\ \|\xi\|_2 = 1\right\}$.

Theorem 2 (Informal)**.** If $\bar\lambda_{\mathcal{R}} < \beta$, then the self-play operator $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone with $\mu_{\mathcal{R}} = \beta - \bar\lambda_{\mathcal{R}} > 0$. Hence projected extragradient on the self-play fixed-point equation converges last-iterate linearly to the unique self-play fixed point. By Theorem 1, the limiting policy induces a Nash equilibrium of the original two-player RSPG.

The formal statement for the symmetric self-play result is provided in Appendix F.4, where Lemma 5 shows that this single-player form of $\bar\lambda_{\mathcal{R}}$ is what the joint-game definition reduces to under symmetric self-play. The more general version for the full joint pseudogradient is in the preceding Appendix F.3, and the full stochastic bounds are given in the sequel.

3 Learning in Risk-Sensitive Preference Games

We now prove efficient convergence of stochastic self-play algorithms applied to the RSPG. The biggest challenge is that estimators of risk operators are generally *not unbiased*. We introduce a novel Stackelberg game framework to systematically reduce that bias at no-extra cost. Finally, we show generalization guarantees, not afforded in the risk-neutral regime.

Risk Estimators. The results require understanding the statistical structure of risk. Recall $F_{\mathcal{R}}(\theta) = \beta(\theta - \theta_{\text{ref}}) - \mathcal{P}_{\mathcal{R}}\pi_\theta$, with $\mathcal{P}_{\mathcal{R}}(\pi)_y := \mathcal{R}_{Y''\sim\pi}[\mathcal{P}(y \succ Y'')]$. Let $\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi)_y$ denote the sample-based estimator of $\mathcal{P}_{\mathcal{R}}(\pi)_y$ using $Y''_1, \ldots, Y''_m \overset{\text{iid}}{\sim} \pi$ (the same samples shared across components y). With $b_m(\theta) := \mathbb{E}\big[\widehat{F}_{\mathcal{R},m}(\theta)\big] - F_{\mathcal{R}}(\theta)$ and $\zeta_m(\theta) := \widehat{F}_{\mathcal{R},m}(\theta) - \mathbb{E}\big[\widehat{F}_{\mathcal{R},m}(\theta)\big]$,

$$\widehat{F}_{\mathcal{R},m}(\theta) := \beta(\theta - \theta_{\text{ref}}) - \widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta) = F_{\mathcal{R}}(\theta) + b_m(\theta) + \zeta_m(\theta).$$

Here $b_m(\theta)$ is deterministic given θ and $\mathbb{E}[\zeta_m(\theta)|\theta] = 0$. By construction $b_m(\theta) = \mathcal{P}_{\mathcal{R}}(\pi_\theta)) - \mathbb{E}[\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta)]$ and $\zeta_m(\theta) = \mathbb{E}[\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta)] - \widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta)$. The convergence theorems in the sequel (and Appendix H) rely on uniform bias and variance bounds.

Assumption 1 (bias and variance)**.** There exist constants $B_m, V_m > 0$ decreasing in m such that $\|b_m(\theta)\|_2 \le B_m$ and $\mathbb{E}\,\|\zeta_m(\theta)\|_2^2 \le V_m$ for all $\theta \in \theta_{\text{ref}} + \mathcal{W}$.

In Appendix G we establish that smooth plug-in risk estimators $\mathcal{R}[Z] = h(\mathbb{E}_\mu(g(Z)))$ for smooth h and bounded statistic g (e.g., entropic risk) admit bias and variance (at most) $B_m, V_m = \mathcal{O}(1/m)$.

3.1 Iteration Complexity of Biased Extra-Gradient Preference Learning

Stochastic extragradient on a biased operator is well-studied Juditsky et al. (2011): under strong monotonicity it contracts geometrically to a neighborhood of the optimum whose radius is set by the bias and variance of the oracle. Specializing to the risk-adjusted $F_{\mathcal{R}}$ with the plug-in estimator $\widehat{F}_m$, in each round, the stochastic extragradient method Korpelevich (1976) takes two steps:

$$\theta_{\tau-1/2} \leftarrow \mathrm{proj}_{\mathcal{K}}(\theta_{\tau-1} - \eta\,\Sigma(\rho)\,\widehat{F}^{\tau}_{m,1}), \quad \theta_{\tau} \leftarrow \mathrm{proj}_{\mathcal{K}}(\theta_{\tau-1} - \eta\,\Sigma(\rho)\,\widehat{F}^{\tau}_{m,2}),$$

where $\widehat{F}^{\tau}_{m,1}(\theta_{\tau-1})$ and $F^{\tau}_{m,2}(\theta_{\tau-1/2})$ are estimated via a batches of size m at $\theta_{\tau-1}$ and $\theta_{\tau-1/2}$, respectively. Under Assumption 1, the $\Sigma(\rho)$-norm bias and variance satisfy

$$\widetilde{B}_m := \sup_{\theta} \|\,\mathbb{E}[\hat{F}_m(\theta)] - F_{\mathcal{R}}(\theta)\|_{\Sigma(\rho)} \le \sqrt{\sigma_{\mathrm{m}}} B_m,\ \widetilde{V}_m := \mathbb{E}\,\|\hat{F}_m(\theta) - \mathbb{E}[\hat{F}_m(\theta)]\|^2_{\Sigma(\rho)} \le \sigma_{\mathrm{m}} V_m,$$

where $\sigma_{\mathrm{m}} := \|\Sigma(\rho)\|_{\mathrm{op}}$ is the largest eigenvalue of the preconditioner. Analysis yields three error terms: (i) deterministic linear decay $\mathcal{O}((1-\eta\mu_{\mathcal{R}})^T)$ from the contraction; (ii) a persistent bias floor $\Theta(1/(\mu_{\mathcal{R}} m))$, the cost of plug-in risk estimation, which does not vanish in T or η; (iii) a standard variance floor $\Theta(\eta/(\mu_{\mathcal{R}} m))$ that shrinks with the step size.

Proposition 2 (Biased stochastic EG under strong monotonicity)**.** Consider the strongly monotone regime, $\mu_{\mathcal{R}} \ge \beta - \bar{\lambda}_{\mathcal{R}} > 0$. Suppose $\mu \in \Delta(\mathcal{Y}) \mapsto \mathcal{P}_{\mathcal{R}}(\mu)$ is $L_{\mathcal{R}}$ Lipschitz continuous, and Assumption 1 holds. Let $\Omega := \sup_{\theta,\theta' \in \mathcal{K}} \|\theta - \theta'\|_{\Sigma(\rho)^+}$ be the $\Sigma(\rho)^+$-diameter of $\mathcal{K}$. Run stochastic extra-gradient with step size $\eta \le \min\{\frac{1}{4\mu_{\mathcal{R}}}, \frac{1}{\sqrt{6}\,\ell_{\mathcal{R}}}\}$. For any $T \ge 1$ and $m \ge 1$, the estimate holds:

$$\mathbb{E}\,\|\theta_T - \theta^\star\|^2_{\Sigma(\rho)^+} \le (1-\eta\mu_{\mathcal{R}})^T \|\theta_0 - \theta^\star\|^2_{\Sigma(\rho)^+} + 4\,\Omega\,\widetilde{B}_m/\mu_{\mathcal{R}} + 6\,\eta\,(\widetilde{B}^2_m + \widetilde{V}_m)/\mu_{\mathcal{R}}.$$

This proposition is proved in Appendix H. Specifically, we show biased stochastic extragradient achieves $\mathcal{O}(1/\sqrt{t}) + \mathcal{O}(\widetilde{B}_m)$ averaged-iterate convergence in the merely-monotone regime (Theorem 8) and last-iterate contraction at rate $\mathcal{O}((1-\gamma\widetilde{\mu})^t)$ to a neighborhood of size $\mathcal{O}(\widetilde{B}_m/\widetilde{\mu})$ in the strongly monotone regime (Theorem 9), where $\widetilde{\mu} = \mu_{\mathcal{R}}\sigma_{\min}$ and $\widetilde{B}_m = \mathcal{O}(1/m)$ is the oracle bias. The bias term is the binding constraint in the practically relevant regime: samples are expensive, m is constrained by per-iteration compute, and no amount of further iteration drives this floor down.

Why extragradient? Risk distortion can weaken the monotonicity induced by KL regularization. We thus analyze convergence for extragradient across both the monotone and strongly monotone regimes. Its extrapolation step exploits Lipschitz continuity to control the rotational component of the game dynamics. Our analysis gives averaged-iterate gap bounds for biased stochastic oracles under monotonicity, and geometric last-iterate contraction to an oracle-dependent neighborhood under strong monotonicity (Appendix H). We highlight the strongly monotone specialization as it separates optimization error, persistent estimation bias, and sampling variance, directly motivating the bias correction developed next. Analogous guarantees for the mirror descent alternative are given in Appendix H.10.

3.2 Stackelberg Game for Fast Two Time-scale Bias Tracking

We now show the bias floor can be removed essentially for free, by augmenting extragradient with a fast-timescale bias tracker. For smooth plug-in risk estimators, the persistent $\mathcal{O}(1/m)$ floor collapses to a residual of order $\mathcal{O}(1/m^2)$—the same order as the variance floor of an unbiased oracle.

Bias tracking as a Stackelberg game. Let $b_m(\theta) := \mathbb{E}[\widehat{F}_m(\theta)] - F_{\mathcal{R}}(\theta)$ be the oracle's population-level bias and $\widehat{b}_m$ a sample-based estimator. Introduce a tracker $\xi \in \mathbb{R}^d$ with target $\xi^\star(\theta) = b_m(\theta)$. The composite system is a Stackelberg game: the *leader* is risk-adjusted self-play on θ, using the *de-biased* gradient $\widehat{F}_m - \xi$; the *follower* is ξ, best-responding to the leader's current θ. The Stackelberg equilibrium is $(\theta^\star_{\mathcal{R}}, b_m(\theta^\star_{\mathcal{R}}))$: the unique risk-adjusted Nash paired with the perfect bias estimate at that Nash. For the equilibrium analysis to translate into convergence the follower must sit on its best-response curve along the trajectory, which requires a faster timescale: $\eta_\tau \ll \gamma_\tau$. The bias estimator $\hat{b}_m$ uses the same batch already drawn for $\widehat{F}_m$, so tracking costs *no additional samples*. The two-timescale extragradient method with bias tracking updates via

$$(\theta_\tau, \theta_{\tau-1/2}) \leftarrow \texttt{ExtraGrad}(\eta_\tau) \quad \text{and} \quad \xi_\tau \leftarrow (1-\gamma_\tau)\xi_{\tau-1} + \gamma_\tau\,\hat{b}_m(\theta_{\tau-1/2}), \qquad \texttt{(TT-EG)}$$

for batch size m. For smooth plug-in risks of the form $h(\mathbb{E}_\mu[g(Z)])$, both delta-method and leave-one-out jackknife bias corrections have residual bias $\mathcal{O}(m^{-2})$ and variance $\mathcal{O}(m^{-3})$ under the regularity conditions of Appendix I.1.

Theorem 3 (Two-Timescale Convergence)**.** Let $F_\mathcal{R}$ be $\widetilde{\mu}$-strongly monotone and Lipschitz continuous. Assume $\|\mathbb{E}[\widehat{F}_m(\theta)] - F_\mathcal{R}(\theta)\| \le \widetilde{B}_m$, $\mathbb{E}\,\|\widehat{F}_m(\theta) - \mathbb{E}\,\widehat{F}_m(\theta)\|^2 \le \widetilde{V}_m$, and that $\widehat{b}_m$ has residual bias $\|\mathbb{E}\,\widehat{b}_m(\theta) - b_m(\theta)\| \le R_m$, bounded variance $\mathbb{E}\,\|\widehat{b}_m(\theta) - \mathbb{E}\,\widehat{b}_m(\theta)\|^2 \le V_m^b$, and that b_m is Lipschitz continuous in θ. Then (TT−EG) satisfies:

i. **Non-asymptotic.** For appropriate η, γ, the iterates contract geometrically to a steady-state floor of order $\mathcal{O}(\eta/\gamma + R_m + \sqrt{\gamma V_m^b} + \eta\widetilde{V}_m)$. Tuning $\eta = \mathcal{O}(\gamma/m^2)$ collapses this to $\mathcal{O}(1/m^{3/2})$, thereby matching the variance scale.

ii. **Asymptotic.** Letting $\eta_t \to 0$, $\gamma_t \to 0$, $\eta_t/\gamma_t \to 0$, and $\sum_t \eta_t = \infty$, the error estimate $\limsup_{T\to\infty} \mathbb{E}\,\|\theta_T - \theta^\star_\mathcal{R}\|^2 = \mathcal{O}(R_m)$ holds.

The merely-monotone case (Theorem 12) gives the analogous averaged-iterate gap bound with the same $\mathcal{O}(1/m^{3/2})$ floor. Precise statements, the Stackelberg derivation, and horizon-tuned bounds are in Appendix I. In the strongly monotone case, (TT−EG) achieves $\mathcal{O}(T^{-2/3} + R_m)$ expected squared error (Corollary 7); operator averaging improves this to $\mathcal{O}(T^{-1} + R_m^2)$ (Corollary 10), matching the optimal unbiased iteration rate (Kannan & Shanbhag, 2019). For smooth plug-in risks, growing preference batches remove the remaining floor (Corollary 11). Mirror descent analogues of (TT−EG) appear in Appendix I.4. More concretely, when the un-tracked algorithm has a bias term $\widetilde{B}_m^2 \asymp 1/m^2$ and a slow-variance term $\eta\widetilde{V}_m \asymp \eta/m$, the bias dominates when $m \lesssim 1/\eta$, where the implicit constants depend on the risk measure. Below this threshold, the un-tracked algorithm wastes a full power of $1/m$ on a floor that (TT−EG) removes; above it, both algorithms are variance-limited. Practical training sits in the bias-dominated regime: m is constrained by memory and rollout cost (cf. Appendix L.8).

3.3 Equilibrium Sensitivity & Statistical Guarantees

The learning dynamics seek an equilibrium of a RSPG whose preferences are estimated from data. Convergence therefore leaves a statistical question: how accurately does this equilibrium represent the underlying preference population? Reducing optimization error brings the policy closer to the empirical equilibrium; what remains is how estimation error in the preference model propagates to the population target, which is governed by the sensitivity of the equilibrium to perturbations in $\mathcal{P}_\mathcal{R}$.

For possibly nonlinear operators, write $\|\mathcal{P}_\mathcal{R} - \mathcal{P}'_\mathcal{R}\|_{\mathrm{op}} := \sup_{\pi\in\Delta(\mathcal{Y})} \|\mathcal{P}_\mathcal{R}(\pi) - \mathcal{P}'_\mathcal{R}(\pi)\|_2$.

Proposition 3. Let $\pi^\star$ and $\pi'^\star$ be the RQRE induced by $\mathcal{P}_\mathcal{R}$ and $\mathcal{P}'_\mathcal{R}$, respectively, and suppose we are in the strongly monotone regime, with respective moduli $\mu_\mathcal{R}, \mu'_\mathcal{R} > 0$. Then $\mathrm{KL}(\pi^\star\|\pi'^\star) \le \|\mathcal{P}_\mathcal{R} - \mathcal{P}'_\mathcal{R}\|_{\mathrm{op}}^2/(2\min\{\mu_\mathcal{R}, \mu'_\mathcal{R}\}^2)$.

Specializing Proposition 3 to neighboring datasets gives algorithmic stability of the risk-adjusted equilibrium map. In particular, if the preference estimator is ζ-uniformly stable, then $S \mapsto \pi^\star_S$ is $\beta_n = \mathcal{O}(|\mathcal{Y}|^{1/2}\zeta/\mu_\mathcal{R})$-uniformly stable. Standard stability arguments yield an expected generalization gap of order β_n, while the moment-based concentration of Bousquet et al. (2020) gives the high-probability rate $\widetilde{\mathcal{O}}(\beta_n) + \mathcal{O}(\sqrt{\log(1/\delta)/n})$. Full statements and proofs are given in Appendix J.

Optimization yields convergence to an empirical equilibrium; equilibrium sensitivity controls its deviation from the population equilibrium. Together, these give a fast $\widetilde{\mathcal{O}}(1/n)$ KL rate, matching the n-dependence recently obtained for risk-neutral KL-regularized games (Zhang et al., 2026). Achieving this rate under nonlinear risk requires overcoming the loss of constant-sum structure and the bias induced by applying risk directly to observed comparison labels. We address these challenges by first estimating the pairwise preference probabilities and then applying the risk functional; monotonicity and translation invariance yield uniform operator control, which equilibrium sensitivity converts from $\widetilde{\mathcal{O}}(n^{-1/2})$ estimation error to the $\widetilde{\mathcal{O}}(1/n)$ KL rate.

Theorem 4 (Offline sample complexity)**.** Consider the strongly monotone regime with $\mu_\mathcal{R} = \beta - \bar{\lambda}_\mathcal{R} > 0$, and let $\mathcal{R}$ be a *risk-adjusted payoff functional* (Definition 1). Let $\widehat{\pi}_n$ be the empirical risk-

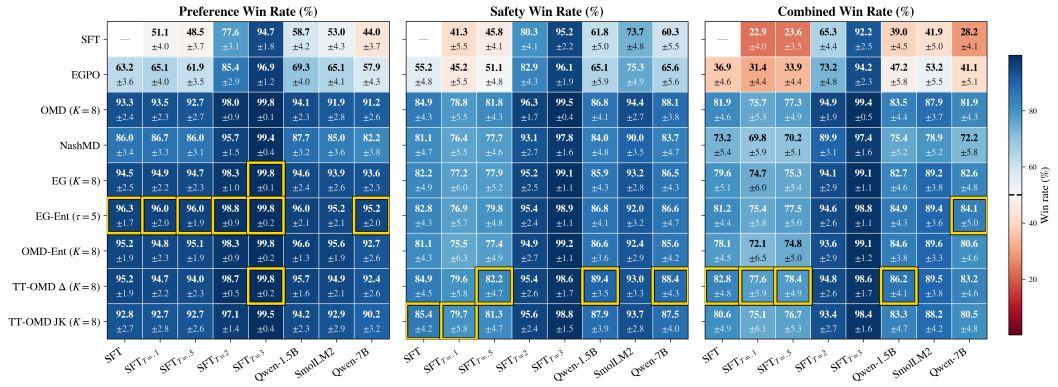


Figure 1: **Cross-play win-rates on a random prompts.** Opponents: off-the-shelf LLMs and SFT decoded at softmax temperature ($T = 1$ base; $T \leq 0.5$ near-greedy, $T \geq 2$ near-uniform). Each cell reports the row policy's win-rate against the column opponent, with 95% bootstrap CIs over prompts. TT-OMD Δ and JK use delta-method and jackknife bias correction, respectively.

adjusted equilibrium computed from n offline preference samples. With probability at least $1 - \delta$, the estimate holds: $\mathrm{KL}(\pi^\star_{\mathcal{R}} \| \widehat{\pi}_n) \lesssim \log(|\mathcal{Y}|/\delta)/(\mu^2_{\mathcal{R}} n)$.

The proof, given in Appendix K, combines uniform concentration of the pairwise preference table, a deterministic nonexpansiveness bound for the risk-adjusted operator, and strong-metric stability of the equilibrium under operator perturbations. The resulting rate depends explicitly on pairwise sampling coverage and the population monotonicity modulus. For entropic risk, the bound is $\widetilde{\mathcal{O}}(1/(\mu^2_{\mathcal{R}} n))$ up to the pairwise sampling-coverage factor, so risk affects statistical conditioning through the monotonicity modulus $\mu_{\mathcal{R}}$ (cf. Appendix K.7).

Together, these results connect solving the empirical game to learning from the underlying preference population. Algorithmic stability guarantees (Appendix J) ensure that a small empirical IPO loss translates into a small population IPO loss up to the generalization error, while Theorem 4 ensures that the empirical equilibrium approximates the population risk-adjusted equilibrium.

4 Experiments

We empirically evaluate risk-sensitive preference games[1] on safety alignment of LLMs using the PKU-SafeRLHF preference dataset (Ji et al., 2024), which provides paired responses with both helpfulness preferences and per-response harmfulness annotations across three severity levels. The central question is whether risk-adjusted training delivers the robustness predicted by our theory—i.e., improved tail behavior and stable performance across data strata—and whether, surprisingly, this robustness comes at no cost to average-case preference win-rate: *robustness without a performance tax*. We probe *tail behavior* through evaluation on held-out severity strata, unseen preference models trained on the same data, and cross-play; significant degradation on any of these would be the signature of the *reward hacking* we wish to avoid. Full experimental details are in Appendix L.

Setup. All policies are LoRA fine-tuned from a common SFT base (Touvron et al., 2023). We compare risk-neutral NLHF baselines (EGPO (Zhou et al., 2025), OIPO (Calandriello et al., 2024b), and Nash-MD (Munos et al., 2024)) against the risk-adjusted IPO (7) instantiated with entropic risk at level τ. The two-timescale variants TT-OMD Δ and JK add delta-method and jackknife bias tracking to entropic OMD $\tau = 10$. In Appendix L, we also explore other losses, including group DRO, that incorporate risk from structured structured uncertainty such as *a priori* known data categories. We train risk-adjusted models with extragradient (EG) and online mirror descent (OMD). Our theoretical guarantees for EG carry over to OMD (with weaker guarantees); we include OMD as an easier-to-implement variant still informative of the impact of risk. Preferences are scored by an LLM judge and harmfulness by the PKU cost model. We report win-rates (mean±std) for preference, safety, and their combination for cross-play amongst method pairs.

[1] Code: anonymous.4open.science/r/risk-sensitive-preference-games-436C.

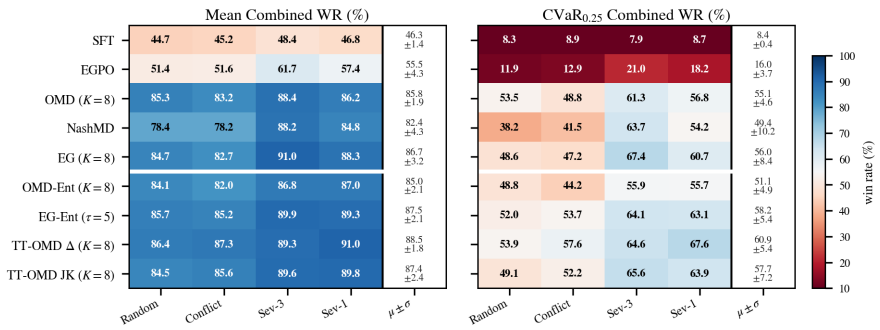


Figure 2: **Mean/lower-tail $\mathrm{CVaR}_{0.25}$ combined win-rate across held-out strata**. Random: uniform test prompts; Conflict: preference and safety disagree; Sev-3/1: severity high/low unsafe prompts. The right column gives cross-strata mean±std-dev. Both two-timescale corrections improve on uncorrected OMD-ENT in every stratum on both metrics; TT-OMD Δ has the highest averages across strata.

Average-case: robustness without a win-rate performance tax. Figure 1 reports cross-play win-rates on the Random stratum. Risk-adjusted methods retain high preference win-rates: EG-ENT ($\tau = 5$) reaches 96.3% against SFT, compared with 94.5% for risk-neutral EG. Across metrics, risk-adjusted and two-timescale methods remain comparable to OMD and EG within the displayed uncertainty; gains over EGPO in most comparisons are supported by nonoverlapping 95% CIs. TT-OMD Δ also exceeds NASH-MD in combined win-rate with nonoverlapping CIs against Qwen-1.5B, SmolLM2, and Qwen-7B. At $K = 8$, TT-OMD Δ has higher combined win-rate point estimates than uncorrected OMD-ENT against six of eight opponents (highlighted in gold), including 82.8% vs 78.1% against SFT and 77.6% vs 72.1% against SFT at temperature 0.1, while retaining comparable preference win-rates. TT-OMD JK attains the highest safety win-rates against these two opponents (85.4%, 79.7%), although the best-performing correction varies.

Cross-strata robustness. The harder test is whether a single policy remains strong across held-out strata. Figure 2 reports mean and lower-tail $\mathrm{CVaR}_{0.25}$ combined win-rates on Random, Conflict, Sev-3, and Sev-1 prompts. Both TT corrections improve on uncorrected OMD-ENT in every stratum on both metrics. TT-OMD Δ achieves the highest mean and CVaR scores on three of four strata, with averages across strata of 88.5% and 60.9%, compared with 85.0% and 51.1% for OMD-ENT. TT-OMD JK reaches 87.4% and 57.7%, respectively. The largest CVaR improvements over OMD-ENT occur on Conflict for the delta correction (44.2% to 57.6%) and on Sev-3 for TT-OMD JK (55.9% to 65.6%). Additional results across strata and metrics are in Appendix L.

Two-timescale bias correction in LLM fine-tuning. To examine the correction directly, we sweep opponent sample counts $K = 2$–256 using 4096-response reference pools. The uncorrected estimator is bias-dominated up to $K \approx 150$ (Fig. 27): squared bias exceeds batch-averaged estimator variance (batch size 64). Small rollout budgets such as $K = 2$–8, motivated by the cost of LLM sampling, lie squarely in this regime (cf. Tab. 1). The largest biases occur on rare-loss conditionals (Fig. 29). In controlled training runs, JK tracking removes 56–76% of measured bias at $K = 4$ and 50–70% at $K = 8$ at checkpoints 1500, 3000, and 4500 (Fig. 33). The corrected $K = 8$ policy also has lower estimated self-play entropic risk than its uncorrected control at every evaluated checkpoint from step 2000 onward, with nonoverlapping 95% CIs at steps 3500 and 4000 (cf. Fig. 36). Appendix L.8 reports the bias-tracking, risk, and performance curves.

5 CONCLUSION

This work develops RSPGs as a way to align over preference distributions, while retaining the structural, computational, and statistical guarantees that make NLHF tractable. Our results show that this shift can be made without sacrificing mean performance and that the consequences of risk for equilibrium structure, learning, and generalization can be quantified explicitly. Several promising future directions are proposed in Appendix B.

AI USE STATEMENT

For ease of review, we list the required disclosures and associate a number to them: (1) generate synthetic data sets, (2) help develop theoretical models or conceptual frameworks, (3) formulate mathematical claims, (4) provide critical ingredients for proving mathematical claims, (5) assist in the writing of proofs, (6) propose or refine hypotheses, (7) design or provide feedback on research methodology or experiments, (8) implement methods, (9) assist with translation, (10) clean and reformat dataset, (11) support qualitative and thematic data analysis, (12) interpret results.

In this work, we used generative AI tools for (1) (specifically, the synthetic risk preference game used for the "toy" two time scale analysis plots was generated with the help of a generative AI model) and (8) (specifically, we used a generative AI model to help implement and edit code for empirical experiments). We have not used generative AI tools for (2–8) and (10–12), and (9) is not applicable to this work. Additionally, we used generative AI tools for checking for grammatical errors and typos. We have reviewed all AI-assisted work. For example, any grammatical errors or typos where reviewed before making edits proposed by the AI assistant. Additionally, LLM-generated code was verified and tested for correctness by Ratliff, Horwitz, and Gonzales. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies risk-sensitive preference learning and bias correction to improve robustness in language-model alignment. Our experiments use existing preference datasets and automated evaluators, including prompts concerning potentially harmful behavior. These datasets and evaluators may encode social and cultural biases, and improvements in their scores do not establish safety or fairness across populations and deployment settings. Safety is assessed using a separate automated cost model; our empirical conclusions are therefore limited to the evaluated models, prompts, and evaluators.

The choice of preference model, risk functional, and risk parameter determines which outcomes receive greater emphasis. Risk-sensitive optimization can amplify biases or misspecifications in these components, and the methods could be misused with objectives that reward harmful behavior. Our theoretical guarantees concern the stated mathematical objectives and assumptions; they do not certify real-world safety. Deployment should include evaluation of both safety and usefulness across relevant user groups, attention to underrepresented failure modes, and appropriate human oversight. Reuse or redistribution of the underlying datasets and models must respect their licenses, access restrictions, and privacy requirements.

REPRODUCIBILITY STATEMENT

We have made all code required to reproduce the experiments and main results available in an anonymized Git repository. The repository includes the implementations, experiment configurations, and instructions needed to reproduce the reported results.

REFERENCES

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Francis Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson,

Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626. PMLR, 2020.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, et al. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*, 2024a.

Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024b.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4302–4310, Red Hook, NY, USA, 2017a. Curran Associates Inc. ISBN 9781510860964.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017b.

Thinh T. Doan. Nonlinear two-time-scale stochastic approximation: Convergence and finite-time performance. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pp. 47–47. PMLR, 07 – 08 June 2021.

Thinh T Doan. Fast Nonlinear Two-Time-Scale Stochastic Approximation: Achieving $\mathcal{O}(1/k)$ Finite-Sample Complexity. *IEEE Transactions on Automatic Control*, 71(1):230–242, 2025.

Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009a.

Asen L Dontchev and R Tyrrell Rockafellar. Robinson's implicit function theorem and its extensions. *Mathematical Programming*, 117(1):129–147, 2009b.

Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Improved rates for derivative free gradient play in strongly monotone games. In *IEEE Conference on Decision and Control (CDC)*, pp. 3403–3408. IEEE, 2022.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Hans Föllmer and Alexander Schied. Convex measures of risk and trading constraints. *Finance and stochastics*, 6(4):429–447, 2002.

Hans Föllmer and Alexander Schied. *Stochastic Finance: An Introduction in Discrete Time*. De Gruyter Graduate. De Gruyter, Berlin, Boston, 4 edition, 2016.

Jake Gonzales, Max Horwitz, Eric Mazumdar, and Lillian J. Ratliff. Strategically robust multi-agent reinforcement learning with linear function approximation. *Reinforcement Learining Conference (arxiv:2603.09208)*, 2026.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Matthias Heger. Consideration of risk in reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 105–111. 1994.

Daniel Hernández-Hernández and Steven I. Marcus. Risk sensitive control of markov processes in countable state space. *Systems Control Letters*, 29(3):147–155, 1996. ISSN 0167-6911.

Kashmir Hill. A teen was suicidal. ChatGPT was the friend he confided in. The New York Times, August 2025a. URL https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html. Published August 26, 2025.

Kashmir Hill. They asked an A.I. chatbot questions. the answers sent them spiraling. The New York Times, June 2025b. URL https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html. Published June 13, 2025.

Kashmir Hill and Dylan Freedman. Chatbots can go into a delusional spiral. here's how it happens. The New York Times, August 2025. URL https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html. Published August 8, 2025.

Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*, 2024.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Aswin Kannan and Uday V Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820, 2019.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

Jonathan Lacotte, Mohammad Ghavamzadeh, Yinlam Chow, and Marco Pavone. Risk-sensitive generative adversarial imitation learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2154–2163. PMLR, 2019.

Nicolas Lanzetti, Sylvain Fricker, Saverio Bolognani, Florian Dörfler, and Dario Paccagnan. Strategically robust game theory via optimal transport. *arxiv preprint arxiv:2507.15325*, 2025.

Eric Mazumdar, Kishan Panaganti, and Laixi Shi. Tractable multi-agent reinforcement learning through behavioral economics. In *The Thirteenth International Conference on Learning Representations*, 2025.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian J Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202):1–56, 2023.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Lillian J. Ratliff and Eric Mazumdar. Inverse risk-sensitive reinforcement learning. *IEEE Transactions on Automatic Control*, 65(3):1256–1263, 2020. doi: 10.1109/TAC.2019.2926674.

Stephen M Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980.

R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

J Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica: Journal of the Econometric Society*, pp. 520–534, 1965.

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Math. Program.*, 125(2):235–261, October 2010. ISSN 0025-5610.

Mahdi Sabbaghi, George Pappas, Adel Javanmard, and Hamed Hassani. Robust policy optimization to prevent catastrophic forgetting. *arxiv preprint arxiv:2602.08813*, 2026.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arxiv preprint arxiv:1707.06347*, 2017.

Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. Multi-turn reinforcement learning with preference human feedback. In *Advances in Neural Information Processing Systems*, volume 37, pp. 118953–118993, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arxiv preprint arxiv:2402.03300*, 2024.

Yun Shen, Michael J. Tobia, Tobias Sommer, and Klaus Obermayer. Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328, 07 2014.

Oliver Slumbers, David Henry Mguni, Stefano B. Blumberg, Stephen McAleer, Yaodong Yang, and Jun Wang. A game-theoretic framework for managing risk in multi-agent systems. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Hamidou Tembine, Quanyan Zhu, and Tamer Başar. Risk-sensitive mean-field games. *IEEE Transactions on Automatic Control*, 59(4):835–850, 2014. doi: 10.1109/TAC.2013.2289711.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arxiv preprint arxiv:2307.09288*, 2023.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. doi: 10.1007/BF01448847.

Mingzhi Wang, Chengdong Ma, Qizhi Chen, Linjian Meng, Yang Han, Jiancong Xiao, Zhaowei Zhang, Jing Huo, Weijie J Su, and Yaodong Yang. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zifan Wang, Yi Shen, Michael M. Zavlanos, and Karl H. Johansson. Learning of nash equilibria in risk-averse games. In *2024 American Control Conference (ACC)*, pp. 3270–3275, 2024. doi: 10.23919/ACC60939.2024.10644891.

P. Whittle. Risk-sensitive linear/quadratic/gaussian control. *Advances in Applied Probability*, 13(4): 764–777, 1981.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing llms for user feedback. *ICLR (arXiv:2411.02306)*, 2025.

Ali Yekkehkhany, Timothy Murray, and Rakesh Nagi. Risk-averse equilibrium for games. *arxiv preprint arXiv:2002.08414*, 2020.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. *arXiv preprint arXiv:2407.00617*, 2024.

Yuheng Zhang, Claire Chen, and Nan Jiang. Beyond pessimism: Offline learning in kl-regularized games. *arXiv preprint arXiv:2604.06738*, 2026.

Runlong Zhou, Maryam Fazel, and Simon S Du. Extragradient preference optimization (EGPO): Beyond last-iterate convergence for Nash learning from human feedback. *COLM (arXiv preprint arXiv:2503.08942)*, 2025.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arxiv preprint arXiv:1909.08593*, 2020.

APPENDIX CONTENTS

A Extended Related Work **17**

B Future Directions & Discussion **18**

C Overview of Game-Theoretic Algorithms **20**

C.1 Reinforcement Learning from Human Feedback 20

C.2 Nash Learning from Human Feedback 21

C.3 The IPO loss as a unifying lens 21

C.4 Algorithms in the (ρ, μ) framework 22

D Sources of Uncertainty in Preference Learning Data **23**

D.1 Risk over the prompt distribution 24

D.2 Risk over within-prompt heterogeneity 24

D.3 Risk over the response-pair sampler 24

D.4 Composition of $\mathcal{R}^{\mu}$ and $\mathcal{R}^{\mathcal{W}}$ 25

E Risk on Other Preference Losses **25**

E.1 IPO with entropic risk 25

E.2 Reverse-KL risk with group sampling 26

E.3 GRPO with risk-adjusted rewards 27

E.4 DPO with entropic risk 28

E.5 KTO with entropic risk 28

F Monotonicity of the Risk-Adjusted Operator **29**

F.1 Translation Invariance Preserves Constant Sum Structure 29

F.2 Symmetric Self-play is Enough 30

F.3 The two-player risk-adjusted game 31

F.4 Symmetric Game via Self-Play 35

F.5 Approaches for Bounding Δ 37

G Statistical Bias Analysis **39**

G.1 Decomposition and abstract assumptions 39

G.2 Scalar-to-vector reduction 40

G.3 The bias lemma 40

G.4 Examples 41

G.5 Aggregate bounds: (O1)-(O2) for the operator 43

H Stochastic Convergence Analysis with Bias **44**

H.1 Preliminaries 44

H.2 Properties of $G = \Sigma F_{\mathcal{R}}$ in the Σ^+ metric . . . 45
H.3 Distance-generating function and projection . . . 46
H.4 Stochastic oracle, decomposition, and filtration . . . 46
H.5 Filtration and noise quantities . . . 47
H.6 Three core technical lemmas . . . 47
H.7 Deterministic-style Bound . . . 49
H.8 Monotone setting: bounding the expected gap . . . 51
H.8.1 Corollary: optimal step size . . . 52
H.9 Strongly monotone setting . . . 53
H.10 Mirror Descent: corollary from extra-gradient analysis . . . 56
H.10.1 Monotone case . . . 57
H.10.2 Strongly monotone case . . . 60

I Stackelberg Game and Two-Timescale Convergence **63**
I.1 Bias estimator and assumptions . . . 63
I.1.1 The Two-timescale extragradient algorithm with time-varying steps . . . 65
I.1.2 Convergence theorem . . . 65
I.1.3 Constant step-size regime . . . 72
I.1.4 Finite-horizon tuned rate . . . 75
I.1.5 An operator-averaging variant with a T^{-1} rate . . . 77
I.1.6 Summary: comparison of methods and step-size regimes . . . 80
I.2 Numerical illustration: a Bradley-Terry preference game . . . 81
I.3 Stackelberg interpretation: a follower tracking joint Nash play . . . 84
I.3.1 Monotone-only case: averaged-iterate convergence . . . 86
I.4 Two-Timescale Mirror Descent: Corollary from extra-gradient analysis . . . 89
I.4.1 Strongly monotone case, decaying steps . . . 90
I.4.2 Strongly monotone case, constant steps . . . 90
I.4.3 Monotone case, constant steps . . . 91

J Stability and Generalization **92**
J.1 Structural stability of the solution mapping . . . 92
J.2 From stability to generalization under the risk-adjusted IPO loss . . . 95
J.3 Algorithmic Stability Generalization Bounds in Expectation . . . 97
J.4 Algorithmic Stability High Probability Generalization Bounds . . . 98

K Offline sample complexity **100**
K.1 Technical Novelties & Hurdles . . . 100
K.2 Preliminaries and main result . . . 101
K.3 Uniform estimation of the risk-adjusted operator . . . 103

K.4 Existence and stability of the equilibrium . . . 104
K.5 Assembling the offline guarantee . . . 104
K.6 Entropic risk and plug-in bias . . . 105
K.7 Explicit risk-parameter dependence for entropic risk . . . 106
K.8 Coherent payoff risks and CVaR . . . 107
K.9 Interpretation and scope . . . 109

L Additional Experimental Results and Details **110**
L.1 Experimental Setup and Implementation Details . . . 110
L.2 Win-Rate Distributions Across Strata . . . 110
L.3 Tail Robustness Across Opponents . . . 112
L.4 Response Consistency & Variance . . . 113
L.5 Cross-Play Win Rates Across Strata . . . 114
L.6 Robustness Across Harm Categories . . . 117
L.7 Training Dynamics . . . 117
L.8 Two-Timescale RSPG for LLM Fine-Tuning . . . 118
L.8.1 Understanding the Bias-Variance Tradeoff for Fine-Tunning LLMs . . . 119
L.8.2 Two-timescale bias correction at small K: a controlled study . . . 124
L.9 Additional Two-Timescale Crossplay Results for $K = 8$ . . . 133
L.10 Risk-Sensitive IPO with Uncertainty over Safety Categories . . . 134
L.10.1 What "risk over group uncertainty" actually means here . . . 135
L.10.2 Risk over the Safety Category $\xi = e$ . . . 135
L.10.3 Sign convention: robust vs. optimistic . . . 136
L.10.4 Sample-Level Implementation . . . 136

A Extended Related Work

RLHF. Reinforcement learning from human feedback (RLHF) has emerged as a standard approach for aligning language models with human preferences (Christiano et al., 2017a; Ziegler et al., 2020). The canonical pipeline first learns a scalar reward model from pairwise comparisons and then optimizes a policy against it via PPO (Schulman et al., 2017). A fundamental limitation of this approach is reward hacking, where the policy is optimized against a learned proxy rather than true human preferences and can exploit gaps in the reward model, producing outputs that score highly without being genuinely preferred. Rafailov et al. (2023) proposed DPO to sidestep reward modeling entirely by reparameterizing the optimal policy directly in terms of preference pairs, though it inherits the Bradley-Terry assumption and is inherently offline, making it susceptible to distribution shift. Shao et al. (2024) introduced GRPO, which retains the PPO framework but discards the value network in favor of group-relative advantage estimation. A separate line of work drops the Bradley-Terry assumption entirely, observing that a scalar reward cannot represent non-transitive preferences (Swamy et al., 2024; Wang et al., 2025). Closest to our setting, Sabbaghi et al. (2026) extends GRPO with an entropic risk objective over a KL-bounded neighborhood of policies, with the goal of reducing brittleness to downstream fine-tuning. However, this approach remains single-player and does not engage with the game-theoretic structure of preference optimization, which is the focus of our work.

Game-theoretic preference learning. Motivated by the limitations of reward-based approaches, a growing body of work recasts alignment as a two-player game over policies using pairwise preferences directly as the payoff. Azar et al. (2024) introduced the Ψ-PO framework and the IPO loss as an offline approach to directly optimize pairwise preference probabilities against a fixed policy, without reward modeling or the Bradley-Terry assumption. Building on this IPO loss, Munos et al. (2024) and Calandriello et al. (2024a) were among the first to cast preference optimization as a two-player constant-sum game. Munos et al. (2024) recast this problem as a two-player constant-sum game, introducing Nash Mirror Descent (Nash-MD) as an online self play algorithm for finding the Nash equilibrium of a preference model, and Calandriello et al. (2024b) showed that online IPO approximates the Nash equilibrium of the preference game via self-play, proposing IPO-MD with explicit convergence guarantees to the quantal response equilibrium. Several subsequent algorithms target the same fixed point, including MPO (Wang et al., 2025), MTPO (Shani et al., 2024), and INPO Zhang et al. (2024). Zhou et al. (2025) further generalized the IPO loss framework and showed that extragradient updates outperform mirror descent variants, thereby outperforming several prior works. However, all of these methods optimize the expected pairwise payoff, making them blind to the distribution of preference outcomes and leaving no principled way to express or control robustness to tail behavior, annotator disagreement, or safety-critical features. This work addresses this gap by introducing risk-sensitive preference games, in which players optimize convex risk measures of their preference loss rather than its expectation.

Risk Sensitive Games. Risk sensitivity has a long history across a range of fields. In mathematical finance, Artzner et al. (1999) introduced coherent risk measures and Föllmer & Schied (2002) formalized the broader class of convex risk measures which have found widespread application. Risk-sensitive objectives are equally established in sequential decision-making, starting with risk-sensitive MDPs (Howard & Matheson, 1972; Hernández-Hernández & Marcus, 1996), and extended to optimal control Whittle (1981); Hernández-Hernández & Marcus (1996), dynamic programming (Ruszczyński, 2010), and reinforcement learning (Heger, 1994; Shen et al., 2014; Lacotte et al., 2019; Ratliff & Mazumdar, 2020), where it consistently improves robustness to environmental noise and modeling error potentially at a cost to expected performance. Risk in multi-agent and game-theoretic settings is less developed: Tembine et al. (2014) study risk in mean-field games, while Wang et al. (2024); Yekkehkhany et al. (2020); Slumbers et al. (2023); Lanzetti et al. (2025) each incorporate risk into finite-player games under varying structural assumptions and with varying degrees of tractability. The Risk Sensitive Quantal Response Equilibrium (RQRE) of Mazumdar et al. (2025) unifies risk aversion and bounded rationality in a framework that is both unique and computationally tractable, subject to conditions on risk aversion and bounded rationality that are independent of the underlying game. Gonzales et al. (2026) extend this to general-sum Markov games in the linear function approximation setting via an optimistic value iteration algorithm, establishing finite-sample regret bounds and proving the RQRE policy map is Lipschitz stable under payoff perturbations, thereby yielding robustness and convergence guarantees that Nash-based approaches cannot provide. We bring risk sensitivity into preference learning, replacing expected pairwise payoffs with convex risk measures so that the equilibrium concept itself reasons about the diversity in the distribution of preference outcomes rather than collapsing to the mean.

B Future Directions & Discussion

The framework introduced in this work opens several directions, both theoretical and applied. We sketch the most promising below.

Beyond the monotone regime. Our convergence and stability theory covers the monotone regime $\overline{\lambda}_{\mathcal{R}} \leq \beta/2$, with linear last-iterate rates in the strongly monotone interior $\overline{\lambda}_{\mathcal{R}} < \beta/2$ and slower convergence at the boundary. Beyond this regime, when $\overline{\lambda}_{\mathcal{R}} > \beta/2$, the joint pseudogradient is non-monotone on the symmetric tangent space and last-iterate cycling becomes generically unavoidable: this is a fundamental barrier inherent to non-monotone games rather than an artifact of our analysis. This is also the regime in which the most distinctive behaviors of risk-sensitive equilibria emerge—where the risk-adjusted operator genuinely departs from its risk-neutral counterpart rather than merely deforming it—and where the question of what constitutes a meaningful solution concept becomes non-trivial. Productive directions here lie not in chasing last-iterate convergence (which is impossible in general) but in characterizing average-iterate dynamics, identifying structural sub-

classes where convergence can be recovered, or designing algorithms that explicitly target the cycle structure.

The two timescale method provides a clean bias correction framework with a nice game-theoretic interpretation. However, the guarantee depends on the quality of the bias estimator: clean for delta and jack-knife method estimators ($R_m = \mathcal{O}(1/m^2)$) and oracle CVaR-RU ($R_m = 0$), yet constructing high-quality bias estimators for general convex risks (e.g., non-smooth distortions) is not automatic.

Tightening the offline rate. Our offline sample complexity bound (Theorem 16, Corollary 18) achieves a fast $\widetilde{\mathcal{O}}(1/n)$ rate via a direct argument combining bias control, Bernstein-type concentration, and VI-based stability. Three quantitative gaps remain. First, the dependence on the regularization strength is $1/\beta^2$, whereas the risk-neutral analysis of Zhang et al. (2026) achieves $1/\beta$ by exploiting skew-symmetry of the preference operator; this structure is generically lost under risk, but a refined concentration argument exploiting the residual constant-sum structure of $P_{\mathcal{R}}$ may close the gap. Second, the polynomial dependence on $|\mathcal{Y}|$ is $|\mathcal{Y}|^3$ in our analysis and could be reduced to $|\mathcal{Y}|^2$ via importance-weighted sampling that equalizes pair counts, or potentially to $|\mathcal{Y}|$ via a row-wise concentration argument that bypasses the ℓ_2-vs-ℓ_∞ conversion in our stability lemma. Third, the threshold for empirical strong monotonicity scales as $e^{8\lambda}$ in our worst-case bound; a direct concentration argument on $\bar{\lambda}_{\mathcal{R}}(\widehat{P})$ (rather than going through the operator's ℓ_∞ deviation) should yield a substantially tighter threshold. None of these are barriers to the qualitative result, but each would sharpen the constants.

Additionally, our online generalization bound is $\widetilde{\mathcal{O}}(n^{-1/2})$ in the regime relevant for deep preference models (concentration-only stability $\zeta = \mathcal{O}(n^{-1/2})$), while our offline result attains the fast $\widetilde{\mathcal{O}}(1/n)$ rate. Closing this gap likely requires either stronger preference-model stability or a different analytical route.

Generality of the Stackelberg correction. The bias floor that motivates our two-timescale Stackelberg formulation is not specific to preference games. Any equilibrium problem in which the operator depends nonlinearly on the data distribution—i.e., distributionally robust optimization, mean-field games with risk-sensitive agents, multi-agent reinforcement learning under coherent risk constraints—faces an analogous obstacle, and the same fast-follower architecture should apply with minimal modification. Formalizing the conditions under which such corrections recover unbiased gradients and characterizing their efficiency relative to direct importance-sampling alternatives is a natural extension.

Risk specification at the population level. Our framework accommodates any convex risk measure satisfying translation invariance, applied at any of four structurally distinct levels of the IPO loss: the prompt distribution ($\mathcal{R}^{\mathcal{X}}$), the response sampling distribution ($\mathcal{R}^{\rho}$), the group/annotator/severity distribution ($\mathcal{R}^{\mathcal{W}}$), and the opponent distribution ($\mathcal{R}^{\mu}$). This decomposition already captures sub-population heterogeneity, annotator disagreement, and context-dependent risk aversion as design choices. Two extensions go beyond what the current framework handles.

The first concerns *biased preference sources*. Practical preference pipelines increasingly rely on AI judges, automated annotators, or distilled reward models alongside human raters, each of which introduces its own systematic biases and failure modes — biases that compound when judges themselves are trained on human data and inherit the same coverage gaps. Modeling such sources as additional components of the $\mathcal{R}^{\mathcal{W}}$ stratification is straightforward in principle, but quantifying *how* judge bias propagates through the risk-adjusted equilibrium, and whether risk-sensitivity at the policy level can correct for systematic miscalibration at the judge level, is an open question with significant practical stakes.

The second is *(multiplayer) performative prediction* (Perdomo et al., 2020; Narang et al., 2023). Our analysis treats the latent distribution Ω over which the risk operates—annotators, sub-populations, contexts—as exogenous to the players' actions. In deployment this is rarely true: the population of users a model interacts with, the prompts it encounters, and the annotator pool that evaluates it all shift in response to the policy itself. Performative effects of this kind break the i.i.d. assumption underlying our sample complexity analysis and turn the risk-adjusted equilibrium into a fixed point of a coupled system: Ω is now a function of the joint policy (π_1, π_2), not an independent draw.

Extending the risk-sensitive framework to this setting—e.g., where the risk measure itself depends on the actions over which we are optimizing—is a substantively different problem, and one that we view as the most important applied direction beyond this work as users respond to these models and the models are typically adapting in some way on the fly.

Empirical scope. Our empirical results establish that risk adjustment provides distributional control on held-out severity strata at the scale of our experiments. Three directions extend this scope. First, scaling: how do the constants in our generalization bounds behave at modern frontier scale, and does the variance floor become the binding constraint? Second, deployment-time guarantees: turning risk-adjusted equilibria into auditable safety properties on specified subpopulations, rather than only on the strata used at training time. Third, alignment to populations rather than representative users: a normative as well as a technical project, requiring practitioners to specify which tails matter and to what degree.

Our experiments are on safety alignment on PKU-SafeRLHF with LoRA fine-tuning; the qualitative findings should generalize to other alignment domains and scales, but quantitative effects and optimal risk parameters are likely domain-dependent.

Choosing the Risk Measure. Finally, the framework gives tools for optimizing against a chosen risk measure but is silent on *which* risk measure to choose. Selecting the right risk measure for a given alignment objective is a modeling question we view as the natural next step rather than a gap in the present work.

C Overview of Game-Theoretic Algorithms

In this section we provide an overview of reinforcement learning from human feedback and relevant game-theoretic algorithms for preference learning from human feedback.

C.1 Reinforcement Learning from Human Feedback

Reinforcement learning from human feedback (RLHF) (Christiano et al., 2017a; Ziegler et al., 2020) fits a scalar reward model from pairwise comparison data and then optimizes a policy against it. Given a dataset $\{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$ where y_w is preferred to y_l conditional on prompt x, pairwise preferences are modeled using a Bradley–Terry model (Bradley & Terry, 1952):

$$\mathbb{P}(y \succ y' \mid x) = \sigma\left(r_\phi(x, y) - r_\phi(x, y')\right), \qquad \sigma(t) = \frac{1}{1 + e^{-t}},$$

and the reward model $r_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is learned by minimizing the cross-entropy loss

$$\mathcal{L}_r(\phi) = -\frac{1}{N} \sum_{i=1}^{N} \log \sigma\left(r_\phi(x^{(i)}, y_w^{(i)}) - r_\phi(x^{(i)}, y_l^{(i)})\right).$$

The policy is then optimized against this reward subject to a KL penalty toward a reference policy π_{ref},

$$\pi_\phi^\star = \arg\max_\pi \mathop{\mathbb{E}}_{x \sim \rho_\mathcal{X}} \left[\mathop{\mathbb{E}}_{y \sim \pi(\cdot|x)} [r_\phi(x, y)] - \beta \, \mathrm{KL}(\pi(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x)) \right],$$

typically using PPO (Schulman et al., 2017) or a variant such as GRPO (Shao et al., 2024). For $\beta > 0$ and a full-support reference policy, optimization over unrestricted tabular policies separates across prompts. For each x with $\rho_\mathcal{X}(x) > 0$, one optimal logit representative is

$$\theta_\phi^\star(\cdot \mid x) = \theta_{\text{ref}}(\cdot \mid x) + \frac{r_\phi(x, \cdot)}{\beta}.$$

Adding a prompt-dependent constant to all response logits leaves the optimal policy unchanged. DPO (Rafailov et al., 2023) exploits this closed form to optimize the policy directly from preference pairs without explicit reward modeling. The Bradley–Terry model induces a transitive ordering through its scalar rewards and cannot represent cyclic population preferences. This restriction of the preference model motivates the preference-game formulation of NLHF.

C.2 Nash Learning from Human Feedback

Nash learning from human feedback (Munos et al., 2024) replaces the scalar reward model of RLHF with a preference model $\mathcal{P} : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ that takes both responses as input

$$\mathcal{P}(y \succ y' \mid x) = \mathbb{P}[\text{annotator prefers } y \text{ to } y' \mid x] + \frac{1}{2}\mathbb{P}[\text{annotator is indifferent between } y \text{ and } y' \mid x].$$

Here probabilities are over a randomly chosen annotator, and ties receive half credit. Consequently, $\mathcal{P}(y \succ y' \mid x) + \mathcal{P}(y' \succ y \mid x) = 1$, including $\mathcal{P}(y \succ y \mid x) = 1/2$. Following the setup in Section 2, we suppress the prompt x and assume $|\mathcal{X}| = 1$ for notational simplicity; all definitions extend by conditioning on the prompt. Collecting the preference probabilities into a matrix $P = [\mathcal{P}(y \succ y')]_{(y,y') \in \mathcal{Y} \times \mathcal{Y}}$, the complementarity of the preference probabilities gives the constant-sum identity

$$P + P^\top = \mathbf{1}\mathbf{1}^\top. \tag{8}$$

For two policies, $\pi_1, \pi_2 \in \Delta(\mathcal{Y})$, the preference of π_1 over π_2 is

$$V(\pi_1, \pi_2) = \mathcal{P}(\pi_1 \succ \pi_2) = \mathop{\mathbb{E}}_{y \sim \pi_1,\, y' \sim \pi_2} [\mathcal{P}(y \succ y')] = \pi_1^\top P \, \pi_2,$$

which together with (8) defines a two-player constant-sum matrix game. The NLHF objective is to maximize the worst-case expected preference score:

$$\pi^\star \in \arg\max_{\pi \in \Delta(\mathcal{Y})} \min_{\pi' \in \Delta(\mathcal{Y})} V(\pi, \pi').$$

The constant-sum identity and the minimax theorem (von Neumann, 1928) imply that the game value is $1/2$. Thus $V(\pi^\star, \pi') \geq 1/2$ for every opponent π', and $(\pi^\star, \pi^\star)$ is a Nash equilibrium.

Regularized game and the QRE. We regularize toward a full-support reference policy $\pi_{\text{ref}} \in \Delta^\circ(\mathcal{Y})$ with coefficient $\beta > 0$. The KL-regularized value function and corresponding objective are

$$V_\beta(\pi_1, \pi_2) = \pi_1^\top P \, \pi_2 - \beta \, \mathrm{KL}(\pi_1 \| \pi_{\text{ref}}) + \beta \, \mathrm{KL}(\pi_2 \| \pi_{\text{ref}}),$$
$$\pi_\beta^\star = \arg\max_{\pi_1} \min_{\pi_2} V_\beta(\pi_1, \pi_2).$$

The KL term makes V_β strictly concave in π_1 and strictly convex in π_2, so the regularized game has a unique equilibrium (Munos et al., 2024). Under a softmax parameterization $\pi_\theta \propto \exp(\theta)$, the first-order optimality conditions determine each player's logits up to an additive constant. Choosing a representative for each player gives the coupled equations

$$\theta_1^\star = \theta_{\text{ref}} + \frac{P \, \pi_{\theta_2^\star}}{\beta}, \qquad \theta_2^\star = \theta_{\text{ref}} - \frac{P^\top \pi_{\theta_1^\star}}{\beta}. \tag{9}$$

The constant-sum identity gives $V_\beta(\pi_1, \pi_2) = 1 - V_\beta(\pi_2, \pi_1)$. Thus exchanging the players preserves equilibria, and uniqueness implies $\pi_1^\star = \pi_2^\star = \pi_\beta^\star$. Moreover, $-P^\top \pi = P\pi - \mathbf{1}$, so the two representatives in (9) differ only by a constant vector and induce the same policy. A representative of this common policy satisfies

$$\theta_\beta^\star = \theta_{\text{ref}} + \frac{P \, \pi_{\theta_\beta^\star}}{\beta}. \tag{10}$$

The induced policy $\pi_\beta^\star = \pi_{\theta_\beta^\star}$ is the unique *quantal response equilibrium* (QRE) of the regularized preference game (McKelvey & Palfrey, 1995). Every game-theoretic NLHF algorithm we discuss below targets (10) at its fixed point.

C.3 The IPO loss as a unifying lens

Solving the fixed-point equation (10) requires a tractable loss whose gradient drives the iterates toward the QRE. The generalized IPO loss (Azar et al., 2024; Calandriello et al., 2024b; Zhou et al., 2025) provides exactly this. Let $\rho \in \Delta(\mathcal{Y} \times \mathcal{Y})$ be a sampling distribution over response pairs and $\mu \in \Delta(\mathcal{Y})$ be an opponent distribution. The *generalized IPO loss* is

$$\mathcal{L}_{\text{IPO}}(\theta; \rho, \mu) = \mathop{\mathbb{E}}_{(y,y') \sim \rho} \left[\left(\log \frac{\pi_\theta(y) \pi_{\text{ref}}(y')}{\pi_\theta(y') \pi_{\text{ref}}(y)} - \frac{1}{\beta} \mathop{\mathbb{E}}_{y'' \sim \mu} [\mathcal{P}(y \succ y'') - \mathcal{P}(y' \succ y'')] \right)^2 \right]. \tag{11}$$

Define the response-pair preconditioner

$$\Sigma(\rho) := \mathop{\mathbb{E}}_{(y,y')\sim\rho}\left[(\mathbf{1}_y - \mathbf{1}_{y'})(\mathbf{1}_y - \mathbf{1}_{y'})^\top\right]. \tag{12}$$

When ρ and μ are constructed from the current policy, we stop gradients through their construction. Differentiating with respect to θ while holding ρ and μ fixed yields

$$\nabla_\theta \mathcal{L}_{\mathrm{IPO}}(\theta;\rho,\mu) = 2\,\Sigma(\rho)\left(\theta - \theta_{\mathrm{ref}} - \frac{P\mu}{\beta}\right). \tag{13}$$

Three structural observations follow:

1. The bracketed expression in (13) is the residual of the regularized best-response equation against the fixed opponent μ. When $\mu = \pi_\theta$ is evaluated with stop-gradient, this bracket becomes the residual of the self-play QRE equation (10).
2. The terms ρ and μ play distinct structural roles: ρ enters the gradient only through $\Sigma(\rho)$, which acts as a preconditioner, while μ enters the best-response target through $P\mu$. Thus ρ changes the geometry of the update without changing the underlying best-response or self-play QRE equation.
3. Lastly, recovering the target policy from a zero IPO gradient requires sufficient sampling coverage. If $\ker\Sigma(\rho) = \mathrm{span}\{\mathbf{1}\}$, then the gradient vanishes exactly when $\theta - \theta_{\mathrm{ref}} - P\mu/\beta$ is a constant vector. Thus stationarity identifies the regularized best-response policy against μ, and in self-play identifies the QRE. This condition holds precisely when the undirected graph on $\mathcal{Y}$, with edges between distinct y, y' whenever $\rho(y,y') + \rho(y',y) > 0$, is connected. Without this coverage, the preconditioner can annihilate nonconstant residuals, introducing additional stationary policies even though the target equation is unchanged.

This decomposition organizes the game-theoretic NLHF literature. Every algorithm we discuss below is determined by three choices: a sampling distribution ρ, an opponent μ, and a step rule (a single forward step, a magnetic-adjusted forward step, or a two-step extragradient update) applied to the gradient (13). The expression (10) characterizes the equilibrium for a fixed reference policy and regularization coefficient. Methods that update the reference policy can instead target a sequence of such equilibria. For example, MPO uses this mechanism to approach a Nash equilibrium of the unregularized game (Wang et al., 2025).

C.4 Algorithms in the (ρ, μ) framework

Throughout this subsection $\bar\pi := \mathrm{Uniform}(\mathcal{Y}) \otimes \mathrm{Uniform}(\mathcal{Y})$ denotes uniform sampling over response pairs, $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator, and $\eta > 0$ is a step size. All algorithms perform parameter updates on θ using the IPO gradient (13) evaluated at specific choices of (ρ, μ).

Online IPO and tabular OMD. Online IPO (Calandriello et al., 2024b) samples both responses from the current policy. In our notation, its expected gradient corresponds to $\rho = \mathrm{sg}[\pi^{(t)} \otimes \pi^{(t)}]$ and $\mu = \mathrm{sg}[\pi^{(t)}]$.

For comparison, uniform pair sampling gives a tabular representation of OMD. Writing $n = |\mathcal{Y}|$, we have $\Sigma(\bar\pi) = \frac{2}{n}(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$. A Euclidean gradient step on $\mathcal{L}_{\mathrm{IPO}}(\theta;\bar\pi,\mathrm{sg}[\pi^{(t)}])$ with step size $\alpha = n\beta\eta/4$ therefore yields, up to an additive constant in the logits, the recursion

$$\theta^{(t+1)} = (1-\eta\beta)\,\theta^{(t)} + \eta\beta\left(\theta_{\mathrm{ref}} + \frac{P\,\pi^{(t)}}{\beta}\right).$$

This is a relaxed Picard iteration for (10), with relaxation parameter $\eta\beta$ and the current policy serving as the opponent. Standard no-regret analyses of OMD control the exploitability of averaged policies under suitable step-size schedules (Munos et al., 2024). These averaged-policy guarantees do not by themselves establish last-iterate convergence, whose behavior depends on the game, regularization strength, and step size.

Nash-MD. The tabular Nash-MD update of Munos et al. (2024) admits an IPO representation with $\rho = \bar\pi$, $\mu = \mathrm{sg}[\pi^{(t)}_{\mathrm{mix}}]$, and gradient step size $\alpha = n\beta\eta/4$. This representation produces the

same policy update; the logit representatives may differ by an additive constant. Here $\pi^{(t)}_{\mathrm{mix}}$ is the geometric mixture

$$\pi^{(t)}_{\mathrm{mix}}(y) \propto \pi^{(t)}(y)^{1-\eta\beta}\,\pi_{\mathrm{ref}}(y)^{\eta\beta}.$$

The geometric mixture moves the opponent toward π_{ref}. For exact tabular updates with full-support initialization, the schedule $\eta = \eta_t = 2/(\beta(t+2))$ at iteration $t \geq 0$ yields

$$\mathrm{KL}(\pi^\star_\beta \| \pi^{(T)}) \leq \frac{8}{\beta^2(T+1)}, \qquad T \geq 1.$$

This gives last-iterate convergence to the regularized equilibrium at rate $\mathcal{O}(1/T)$ in KL divergence. Convergence of the iterates themselves has classical precedents in variational inequality methods, including linear convergence under suitable error-bound conditions (Tseng, 1995). The specific bound above for the Nash-MD update is given in Munos et al. (2024, Theorem 1).

Nash-MD-PG. Munos et al. (2024) also proposes a policy-gradient counterpart of Nash-MD. Each outer iteration approximately solves an inner regularized best-response problem against a geometric-mixture opponent whose mixing weight is tuned independently of the KL regularization strength via a single gradient step, which introduces an error whose effect propagates across iterations. Extending the $\mathcal{O}(1/T)$ last-iterate KL bound stated above for exact tabular Nash-MD to this policy-gradient implementation requires controlling that error.

MPO. Wang et al. (2025) replace the geometric-mixture magnet of Nash-MD with a slowly-moving anchor: $\mu = \mathrm{sg}[\pi^{(\tau(t))}]$ for a lagged index $\tau(t) < t$, with the anchor refreshed on a slower timescale. The same anchor also serves as the KL regularization reference, and the update includes an additional KL penalty relative to the current iterate. For a fixed anchor, they show last-iterate linear convergence to the QRE at rate $\mathcal{O}((1+\eta\beta)^{-t})$ with step size constraint $\eta \leq \mathcal{O}(\beta)$. MPO updates the anchor to target a Nash equilibrium of the original unregularized game; the fixed-anchor rate does not by itself establish a linear rate for the full moving-anchor procedure.

EGPO. Zhou et al. (2025) introduce a two-step extragradient update keeping $\rho = \bar{\pi}$ throughout but alternating μ. The extrapolation step uses $\mu = \mathrm{sg}[\pi^{(t)}]$ to produce an intermediate iterate, and the correction step uses $\mu = \pi^{(t+1/2)}$ as the opponent:

$$\theta^{(t+1/2)} = (1-\eta\beta)\,\theta^{(t)} + \eta\beta\left(\theta_{\mathrm{ref}} + \frac{P\,\pi^{(t)}}{\beta}\right),$$

$$\theta^{(t+1)} = (1-\eta\beta)\,\theta^{(t)} + \eta\beta\left(\theta_{\mathrm{ref}} + \frac{P\,\pi^{(t+1/2)}}{\beta}\right).$$

The correction step queries preferences against the predicted opponent rather than the current one, which removes the rotational component responsible for the cycling of OMD. The authors establish last-iterate linear convergence in KL divergence at rate $(1-\eta\beta)^T$ for any step size $\eta \leq 1/(\beta+3)$, and, under their conditionally mean-zero, coordinate-independent sub-Gaussian noise assumptions with variance proxy σ^2, establish the same geometric decay for $\mathbb{E}[\mathrm{KL}(\pi^\star_\beta \| \pi^{(T)})]$, up to an additive $\mathcal{O}(\sigma^2 \log|\mathcal{Y}|/\beta)$ term.

D Sources of Uncertainty in Preference Learning Data

The main paper develops payoff-side risk over the opponent distribution μ. Three other places where risk can enter are the prompt distribution, an auxiliary variable indexing within-prompt heterogeneity, and the response-pair sampler. We treat each in turn and then discuss composition of risks. Following Definition 1, $\mathcal{R}$ denotes a payoff-side risk functional. When risk is applied to a minimized training loss, we write the associated loss-side functional as $\varrho(L) = -\mathcal{R}(-L)$.

Setup. Restoring the prompt $x \in \mathcal{X}$ explicitly, the preference oracle returns $\mathcal{P}(y \succ y' \mid x)$ and may depend on an auxiliary variable $w \in \mathcal{W}$ giving $\mathcal{P}(y \succ y' \mid x, w)$. The per-prompt risk-adjusted IPO loss with risk over μ is

$$\ell^{\mathcal{R}}_{\mathrm{IPO}}(\theta; x, \rho, \mu) = \mathop{\mathbb{E}}_{(y,y')\sim\rho|x}\left[\left(\log\frac{\pi_\theta(y \mid x)\pi_{\mathrm{ref}}(y' \mid x)}{\pi_\theta(y' \mid x)\pi_{\mathrm{ref}}(y \mid x)} - \frac{1}{\beta}\big((P_{\mathcal{R}}\,\mu)_y - (P_{\mathcal{R}}\,\mu)_{y'}\big)\right)^2\right], \quad (14)$$

and the aggregate training loss takes the outer expectation over prompts,

$$\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\theta;\rho,\mu) = \mathop{\mathbb{E}}_{x\sim\rho_{\mathcal{X}}}\left[\ell^{\mathcal{R}}_{\mathrm{IPO}}(\theta;x,\rho,\mu)\right]. \tag{15}$$

The remaining cases correspond to placing risk on x, on w, or on the response-pair sampler ρ.

D.1 Risk over the prompt distribution

Replace the outer expectation $\mathbb{E}_{x\sim\rho_{\mathcal{X}}}$ in (15) with a loss-side risk functional $\varrho^{\mathcal{X}}$ over prompts,

$$\mathcal{L}^{\mathcal{R},\varrho^{\mathcal{X}}}_{\mathrm{IPO}}(\theta;\rho,\mu) = \varrho^{\mathcal{X}}\left[\ell^{\mathcal{R}}_{\mathrm{IPO}}(\theta;x,\rho,\mu)\right].$$

This is distributionally robust optimization over prompts. Natural choices for $\varrho^{\mathcal{X}}$ are loss-side CVaR, which focuses on high-loss difficult prompts, and KL-ball DRO, which optimizes over the worst reweighting of $\rho_{\mathcal{X}}$ within a KL ball. Since this outer loss risk does not enter the preference operator, it leaves the per-prompt RQRE equations unchanged. It need not, however, preserve the stationary points or minimizers of the aggregate training objective.

With ρ and μ held fixed during differentiation, an active subgradient of $\varrho^{\mathcal{X}}$ with respect to the vector of per-prompt losses supplies nonnegative prompt weights $q_\theta(x)$ summing to one. The corresponding update direction is

$$g_\theta = \sum_{x\in\mathrm{supp}(\rho_{\mathcal{X}})} q_\theta(x)\,\nabla_\theta \ell^{\mathcal{R}}_{\mathrm{IPO}}(\theta;x,\rho,\mu).$$

For independent tabular parameter blocks across prompts, positive weights on every prompt in this support, and connected response-pair sampling graphs, $g_\theta = 0$ forces the per-prompt best-response equations, modulo additive logit constants. Under self-play, $\mu(\cdot \mid x) = \pi_\theta(\cdot \mid x)$, these are the RQRE equations. Without these conditions, shared parameters can couple the prompt gradients and zero weights can leave prompts unconstrained. Thus prompt risk changes which games are prioritized and can change the fitted policy. Any policy that simultaneously makes every per-prompt residual zero remains a global minimizer of the frozen-opponent loss for any normalized $\varrho^{\mathcal{X}}$, by monotonicity and nonnegativity of the squared losses.

D.2 Risk over within-prompt heterogeneity

When the preference oracle depends on an auxiliary variable w indexing groups, annotators, or latent labels, the risk-neutral preference operator integrates over $w \mid x$,

$$\mathbb{E}_{w|x}\left[\mathcal{P}(y \succ y' \mid x, w)\right].$$

Using payoff-side risk functionals $\mathcal{R}^{\mu}$ and $\mathcal{R}^{\mathcal{W}}$ as in Definition 1, we first aggregate over the opponent within each group and then over groups, defining the doubly risk-adjusted preference operator

$$(P_{\mathcal{R}^{\mathcal{W}},\mathcal{R}^{\mu}}\,\mu)^x_y := \mathcal{R}^{\mathcal{W}}\left[\mathcal{R}^{\mu}\left[\mathcal{P}(y \succ y' \mid x, w) \,:\, y' \sim \mu\right] \,:\, w \mid x\right],$$

which sits in the same position as $P_{\mathcal{R}}$ inside the per-prompt loss (14). The per-prompt QRE fixed point becomes

$$\theta^\star_{\beta,\mathcal{R}^{\mathcal{W}},\mathcal{R}^{\mu}}(\cdot \mid x) = \theta_{\mathrm{ref}}(\cdot \mid x) + \frac{P_{\mathcal{R}^{\mathcal{W}},\mathcal{R}^{\mu}}(\cdot,\cdot \mid x)\,\pi^\star(\cdot \mid x)}{\beta}.$$

Structurally, risk over w acts on the inner preference operator in the same way as risk over μ—both replace an expectation in the preference computation with a risk functional and produce the same kind of fixed-point modification. The difference is what each models. Risk over μ targets robustness against an adversarial opponent in the game, whereas risk over w targets robustness of each response's opponent-risk-adjusted payoff across heterogeneous evaluators. Here, CVaR over annotator identity means lower-tail CVaR of these payoffs, and a worst-case over safety labels means their minimum over labels. This gives action-wise group robustness, closely related to group-DRO.

D.3 Risk over the response-pair sampler

Risk over the response pair sampler ρ replaces the outer expectation $\mathbb{E}_{(y,y')\sim\rho}$ in the per-prompt loss (14) with a loss-side risk functional ϱ^{pair}. For tabular logits, hold ρ and μ fixed during differentiation and suppress the prompt conditioning. A subgradient of ϱ^{pair} with respect to the vector of

pairwise squared losses supplies nonnegative pair weights $q_\theta(y, y')$ summing to one. The resulting update direction is

$$g_\theta = 2\,\Sigma(q_\theta)\left(\theta - \theta_{\text{ref}} - \frac{P_{\mathcal{R}}\,\mu}{\beta}\right),$$

where $\Sigma(q_\theta)$ is the response-pair preconditioner evaluated under these effective weights. Thus pair risk changes the preconditioner, which may now depend on θ, but leaves the fixed-point bracket unchanged. If the graph of pairs with positive effective weight is connected, stationarity under stop-gradient self-play still characterizes the same RQRE, modulo additive logit constants. Without this coverage condition, the preconditioner can have additional null directions, so stationarity alone need not imply the RQRE equations.

D.4 Composition of $\mathcal{R}^\mu$ and $\mathcal{R}^{\mathcal{W}}$

Risk over μ and risk over w both modify the inner preference operator and the two can be composed. Risk functionals do not commute in general, so the order of composition is a modeling choice. The two natural orderings are

$$\text{(A)}\quad (P_{\mathcal{R}^{\mathcal{W}}\circ\mathcal{R}^\mu}\,\mu)^x_y = \mathcal{R}^{\mathcal{W}}_w\left[\mathcal{R}^\mu_{y''}\left[\mathcal{P}(y \succ y'' \mid x, w)\right]\right],$$

$$\text{(B)}\quad (P_{\mathcal{R}^{\mu}\circ\mathcal{R}^{\mathcal{W}}}\,\mu)^x_y = \mathcal{R}^\mu_{y''}\left[\mathcal{R}^{\mathcal{W}}_w\left[\mathcal{P}(y \succ y'' \mid x, w)\right]\right].$$

Ordering (A) corresponds to applying the opponent-risk first and group-risk on the outside, meaning the policy is robust against an adversarial opponent within each group, and then robust across groups. Ordering (B) reverses this, the policy is robust across groups for each opponent draw, and then robust against an adversarial opponent. The two need not coincide when only one of $\mathcal{R}^\mu$ and $\mathcal{R}^{\mathcal{W}}$ is an expectation, and the remaining risk functional still affects the preference operator. When both reduce to expectations, the two orderings coincide and recover the risk-neutral preference operator. The gradient identity

$$\nabla_\theta \mathcal{L}_{\text{IPO}}^{\mathcal{R}^\mu,\mathcal{R}^{\mathcal{W}}}(\theta;\rho,\mu) = 2\,\Sigma(\rho)\left(\theta - \theta_{\text{ref}} - \frac{P_{\mathcal{R}^{\mathcal{W}}\circ\mathcal{R}^\mu}\,\mu}{\beta}\right)$$

holds with either ordering substituted into the operator. The bracket remains the residual of the corresponding fixed-point equation. The strong-monotonicity analysis of the main paper applies to the composed operator provided its regularity assumptions and risk-distortion bound are verified for that operator; strong monotonicity does not follow from composition alone.

E Risk on Other Preference Losses

The risk framework can be applied either to uncertain payoffs or to the losses used to fit feedback data. We consider IPO (Calandriello et al., 2024b), GRPO (Shao et al., 2024), DPO (Rafailov et al., 2023), and KTO (Ethayarajh et al., 2024). In the IPO and GRPO constructions, payoff-side risk changes the action values or rewards. For DPO and KTO, we replace the expected training loss with a loss-side risk measure, keeping each method's per-example loss unchanged.

Throughout, we write $h_\pi(y, y') := \log \frac{\pi(y)\pi_{\text{ref}}(y')}{\pi(y')\pi_{\text{ref}}(y)}$ for the log-ratio term, $X_y := P(y \succ Y')$ for the random variable representing the win-rate of y against an opponent $Y' \sim \mu$, and $R_\tau(y) := \mathcal{R}_\tau(X_y)$ for the risk-adjusted action value. The squared-residual form

$$\widehat{L}_\tau(\pi) = \mathop{\mathbb{E}}_{(y,y')\sim\mathcal{D}}\left[\left(h_\pi(y, y') - \beta^{-1}\big(\widehat{R}_{\tau,m}(y) - \widehat{R}_{\tau,m}(y')\big)\right)^2\right] \tag{16}$$

is the generalized IPO surrogate used for the payoff-target constructions below; what changes is the choice of $\mathcal{R}_\tau$, μ, and the resulting estimator $\widehat{R}_{\tau,m}$. The GRPO, DPO, and KTO extensions retain their respective objectives, as specified in their subsections.

E.1 IPO with entropic risk

Take $\mu \in \Delta(\mathcal{Y})$ to be the data-generating distribution over individual responses, distinct from the response-pair distribution $\mathcal{D}$, and $\mathcal{R}_\tau$ to be the entropic payoff risk with $\tau > 0$:

$$\mathcal{R}^{\text{ent}}_\tau(X_y) = -\frac{1}{\tau}\log \mathop{\mathbb{E}}_{Y'\sim\mu}\left[e^{-\tau P(y\succ Y')}\right],$$

giving the risk-adjusted action value

$$R_\tau^{\mathrm{IPO}}(y) = \mathcal{R}_\tau^{\mathrm{ent}}(X_y).$$

Given $Y'_1, \dots, Y'_m \overset{\text{iid}}{\sim} \mu$, the plug-in estimator is

$$\widehat{R}_{\tau,m}^{\mathrm{IPO}}(y) = -\frac{1}{\tau} \log \left(\frac{1}{m} \sum_{j=1}^{m} e^{-\tau P(y \succ Y'_j)} \right), \tag{17}$$

and substituting into (16) yields the risk-adjusted IPO loss. As $\tau \to 0$, a Taylor expansion gives $\widehat{R}_{\tau,m}^{\mathrm{IPO}}(y) \to \frac{1}{m} \sum_j P(y \succ Y'_j)$, the empirical mean win-rate. Thus the limit recovers the risk-neutral generalized IPO loss (11) with its opponent expectations replaced by empirical averages.

E.2 Reverse-KL risk with group sampling

GRPO (Shao et al., 2024) draws a group of responses from π_{old} and, for outcome supervision, centers and scales their rewards to obtain advantages for a clipped policy-gradient objective. Motivated by this group sampling, we take $\mu = \pi_{\text{old}}$ as an opponent distribution and construct reverse-KL risk-adjusted payoff targets for (16). Assume that μ has full support on the finite response space and $\tau > 0$. The payoff-side counterpart of the reverse-KL loss risk is

$$\mathcal{R}_\tau^{\mathrm{RKL}}(X_y) = - \sup_{p \in \Delta(\mathcal{Y})} \left[\mathbb{E}_p[-X_y] - \frac{1}{\tau} \mathrm{KL}(\mu \,\|\, p) \right].$$

Equivalently, the inner adversary minimizes $\mathbb{E}_p[X_y] + \tau^{-1} \mathrm{KL}(\mu \,\|\, p)$. Its KKT optimality condition gives

$$p^\star(y'') = \frac{\mu(y'')}{\tau(P(y \succ y'') - \mu_0)},$$

where μ_0 is determined by requiring these weights to sum to one:

$$\sum_{y'' \in \mathcal{Y}} \frac{\mu(y'')}{\tau(P(y \succ y'') - \mu_0)} = 1, \qquad \mu_0 < \min_{y'' \in \mathcal{Y}} P(y \succ y'').$$

The inequality ensures that every denominator is positive. On this interval, the left-hand side increases continuously and strictly from 0 to $+\infty$, so the normalization equation has exactly one solution. Substituting and applying the Fenchel identity for $\varphi(t) = -\log t$ produces

$$R_\tau^{\mathrm{RKL}}(y) = \mu_0 + \frac{1}{\tau} + \frac{1}{\tau} \mathop{\mathbb{E}}_{y'' \sim \mu} \left[\log\big(\tau(P(y \succ y'') - \mu_0)\big) \right]. \tag{18}$$

The likelihood ratios $p^\star(y'')/\mu(y'')$ are largest for opponents against which y has the lowest win probability, emphasizing the hardest matchups. The plug-in estimator $\widehat{R}_{\tau,m}^{\mathrm{RKL}}$ uses the empirical distribution of a group $Y''_1, \dots, Y''_m \overset{\text{iid}}{\sim} \mu$, with the adversary restricted to the sampled responses. It replaces the expectation in (18) by the empirical mean and uses $\widehat{\mu}_0 < \min_j P(y \succ Y''_j)$ solving

$$\frac{1}{m} \sum_{j=1}^{m} \frac{1}{\tau(P(y \succ Y''_j) - \widehat{\mu}_0)} = 1.$$

As $\tau \to 0$, we have that $p^\star \to \mu$ and $R_\tau^{\mathrm{RKL}}(y) \to \mathbb{E}_{Y'' \sim \mu}[P(y \succ Y'')]$; as $\tau \to \infty$, the value tends to $\min_{y''} P(y \succ y'')$. The empirical construction has the corresponding limits over the sampled group. Substitution into (16) therefore recovers the generalized IPO surrogate with empirical group win-rates in the risk-neutral limit. These win-rates are payoff targets, not the centered and scaled reward advantages of standard GRPO; the reverse-KL penalty here acts on the opponent distribution rather than regularizing the learned policy against a reference policy.

E.3 GRPO WITH RISK-ADJUSTED REWARDS

We extend outcome-supervised GRPO (Shao et al., 2024) by applying payoff-side risk to the reward of each response before computing group-relative advantages. This retains GRPO's group normalization, token-level clipping, and policy-reference KL penalty.

For a prompt x, let $u(x, y, Z)$ be a bounded payoff for response y, where $Z \sim \nu(\cdot \mid x)$ represents uncertainty in its evaluation. Define

$$r_\tau(x, y) = \mathcal{R}_\tau\big(u(x, y, Z)\big).$$

In the preference game, take $Z = Y' \sim \mu(\cdot \mid x)$ and $u(x, y, Y') = P(y \succ Y' \mid x)$, so this reward is the risk-adjusted action value used above. More generally, Z can index annotators or uncertain reward evaluations. The evaluation law ν, including the opponent policy when $\nu = \mu$, is held fixed during each policy update.

Sample $G \geq 2$ responses $Y_1, \ldots, Y_G$ independently from $\pi_{\text{old}}(\cdot \mid x)$ and an independent evaluation batch $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} \nu(\cdot \mid x)$. The same evaluation batch may be shared across responses. Write $u_{ij} = u(x, Y_i, Z_j)$. For example, the reverse-KL construction in Section E.2 gives the empirical reward

$$\widehat{r}_{\tau,i} = \min_{w \in \Delta_m} \left\{ \sum_{j=1}^{m} w_j u_{ij} + \frac{1}{\tau m} \sum_{j=1}^{m} \log \frac{1}{m w_j} \right\}, \tag{19}$$

where Δ_m is the probability simplex on the m evaluation samples. Each response has its own minimizing weights. The entropic estimator in (17), with $P(y \succ Y'_j)$ replaced by u_{ij}, is another choice. Both choices emphasize low-payoff evaluation outcomes; the KL penalty in (19) is included in the reward value.

Compute the group mean, standard deviation, and advantages from these risk-adjusted rewards:

$$\overline{r}_\tau = \frac{1}{G} \sum_{i=1}^{G} \widehat{r}_{\tau,i}, \quad s_\tau^2 = \frac{1}{G} \sum_{i=1}^{G} (\widehat{r}_{\tau,i} - \overline{r}_\tau)^2, \quad \widehat{A}_{\tau,i} \quad = \frac{\widehat{r}_{\tau,i} - \overline{r}_\tau}{s_\tau + \delta}, \delta > 0.$$

Here δ is a fixed numerical stabilizer, used also in the risk-neutral baseline. Risk is applied across evaluation outcomes for each response; group normalization then compares the resulting rewards across responses.

Let T_i be the length of Y_i and $H_{it} = (x, Y_{i,<t})$ its history before token t. Define the policy ratio and sampled reference-policy penalty by

$$\omega_{it}(\theta) = \frac{\pi_\theta(Y_{i,t} \mid H_{it})}{\pi_{\text{old}}(Y_{i,t} \mid H_{it})}, \quad v_{it}(\theta) = \frac{\pi_{\text{ref}}(Y_{i,t} \mid H_{it})}{\pi_\theta(Y_{i,t} \mid H_{it})}, \quad k_{it}(\theta) = v_{it}(\theta) - \log v_{it}(\theta) - 1.$$

For a clipping parameter $\epsilon \in (0, 1)$, write

$$C_\epsilon(\omega, A) = \min\{\omega A, \text{clip}(\omega, 1 - \epsilon, 1 + \epsilon)A\}.$$

The empirical risk-sensitive GRPO objective to maximize is

$$\widehat{J}_\tau(\theta) = \frac{1}{G} \sum_{i=1}^{G} \frac{1}{T_i} \sum_{t=1}^{T_i} \left[C_\epsilon\big(\omega_{it}(\theta), \text{sg}[\widehat{A}_{\tau,i}]\big) - \beta k_{it}(\theta) \right], \qquad \beta > 0,$$

averaged over prompts and sampled groups. The sampled responses, evaluation batch, and π_{old} are fixed during optimization; gradients pass through the policy ratios and reference-policy penalty. The outcome advantage $\widehat{A}_{\tau,i}$ is used for every token of response i, as in outcome-supervised GRPO.

For both the reverse-KL and entropic choices, $\widehat{r}_{\tau,i} \to m^{-1} \sum_j u_{ij}$ as $\tau \downarrow 0$. With the same samples and stabilizer, the advantages and clipped objective therefore converge to those of GRPO using these empirical mean rewards. Taking $\delta = 0$ gives the original, unstabilized normalization whenever the group standard deviation is positive; the corresponding limit holds when the risk-neutral standard deviation is positive. If $u(x, y, Z) = r(x, y)$ is deterministic, normalization and cash additivity of $\mathcal{R}_\tau$ give $r_\tau(x, y) = r(x, y)$ for every τ, recovering ordinary GRPO with that scalar reward. Thus nontrivial risk sensitivity here comes from uncertainty in each response's payoff, including the opponent uncertainty of the preference game.

E.4 DPO WITH ENTROPIC RISK

Fix the reference policy π_{ref}. With σ denoting the logistic function and prompt conditioning restored in h_π, DPO (Rafailov et al., 2023) uses the fitted Bradley–Terry preference probability

$$P_\theta(y \succ y' \mid x) = \sigma\big(\beta h_{\pi_\theta}(y, y' \mid x)\big), \qquad \beta > 0.$$

For the frozen-reference construction, draw prompts $x \sim \rho_{\mathcal{X}}$ from the fixed empirical prompt distribution and response pairs $Y_1, Y_2 \overset{\text{iid}}{\sim} \pi_{\mathrm{ref}}(\cdot \mid x)$. The preference label $B \in \{0, 1\}$ has the conditional distribution

$$B \mid x, Y_1, Y_2 \sim \mathrm{Bernoulli}\big(P(Y_1 \succ Y_2 \mid x)\big).$$

Here $B = 1$ means that Y_1 is preferred to Y_2, and $B = 0$ means the reverse. The per-comparison negative log-likelihood is

$$\ell_\theta^{\mathrm{DPO}}(x, y_1, y_2, b) = -b \log P_\theta(y_1 \succ y_2 \mid x) - (1 - b) \log\big(1 - P_\theta(y_1 \succ y_2 \mid x)\big).$$

The expected DPO objective is

$$\mathcal{L}^{\mathrm{DPO}}(\theta) = \mathop{\mathbb{E}}_{x \sim \rho_{\mathcal{X}}} \mathop{\mathbb{E}}_{Y_1, Y_2 \overset{\text{iid}}{\sim} \pi_{\mathrm{ref}}(\cdot|x)} \left[\mathop{\mathbb{E}}_{B|x, Y_1, Y_2} \left[\ell_\theta^{\mathrm{DPO}}(x, Y_1, Y_2, B) \right] \right]. \tag{20}$$

We replace the joint expectation over prompts, reference-policy draws, and labels with loss-side risk applied to the negative log-likelihood, keeping π_{ref} fixed. Using the convention $\varrho(L) = -\mathcal{R}(-L)$, entropic risk gives, for $\tau > 0$,

$$\mathcal{L}_\tau^{\mathrm{DPO}}(\theta) = \varrho_\tau\big(\ell_\theta^{\mathrm{DPO}}(x, Y_1, Y_2, B)\big) = \frac{1}{\tau} \log \mathop{\mathbb{E}}_{x \sim \rho_{\mathcal{X}}} \mathop{\mathbb{E}}_{Y_1, Y_2 \overset{\text{iid}}{\sim} \pi_{\mathrm{ref}}(\cdot|x)} \left[\mathop{\mathbb{E}}_{B|x, Y_1, Y_2} \left[e^{\tau \ell_\theta^{\mathrm{DPO}}(x, Y_1, Y_2, B)} \right] \right].$$

The expectations are estimated by averaging over the sampled, labeled comparisons. As $\tau \downarrow 0$, this recovers (20), the standard expected DPO loss under the same frozen-reference sampling scheme.

E.5 KTO WITH ENTROPIC RISK

Fix the reference policy π_{ref} and a sample of N labeled feedback examples $Z_i = (x_i, y_i, s_i)$, where $s_i \in \{D, U\}$ indicates desirable or undesirable feedback. For an example $Z = (x, y, s)$, following Equation (8) of Ethayarajh et al. (2024), define the implicit reward and reference point by

$$r_\theta(x, y) = \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}, \qquad z_\theta(x) = \mathrm{KL}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\big).$$

KTO's sigmoid value function is

$$v_\theta(Z) = \begin{cases} \lambda_D \, \sigma\big(\beta(r_\theta(x, y) - z_\theta(x))\big), & s = D, \\ \lambda_U \, \sigma\big(\beta(z_\theta(x) - r_\theta(x, y))\big), & s = U, \end{cases}$$

where $\beta > 0$ controls saturation and $\lambda_D, \lambda_U > 0$ control the relative weighting of desirable and undesirable examples. As in KTO, the reference point is treated with stop-gradient during policy updates. Writing $\lambda_s = \lambda_D$ for $s = D$ and $\lambda_s = \lambda_U$ for $s = U$, define the loss of example i by $\ell_i = \lambda_{s_i} - v_\theta(Z_i)$. The usual empirical objective is

$$\mathcal{L}^{\mathrm{KTO}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_i.$$

We replace this average with loss-side risk while retaining KTO's value function and reference-point treatment. With the convention $\varrho(L) = -\mathcal{R}(-L)$, applying entropic risk to the N equally weighted example losses gives, for $\tau > 0$,

$$\mathcal{L}_\tau^{\mathrm{KTO}}(\theta) = \frac{1}{\tau} \log \left(\frac{1}{N} \sum_{i=1}^{N} e^{\tau \ell_i} \right).$$

The corresponding reweighting of the feedback examples is given by

$$\mathcal{L}_\tau^{\mathrm{KTO}}(\theta) = \max_{p \in \Delta_N} \left\{ \sum_{i=1}^{N} p_i \ell_i - \frac{1}{\tau} \sum_{i=1}^{N} p_i \log(N p_i) \right\},$$

where Δ_N is the probability simplex on the feedback examples. Its KKT condition gives

$$p_i^\star = \frac{e^{\tau \ell_i}}{\sum_{j=1}^N e^{\tau \ell_j}} \propto \exp\big(\tau\big[\lambda_{s_i} - v_\theta(Z_i)\big]\big)\,.$$

Substituting these weights into the maximization above gives the displayed entropic objective. Thus entropic risk places greater relative weight on high-loss feedback examples, with the sigmoid retained inside v_θ. As $\tau \downarrow 0$, $p_i^\star \to 1/N$ and $\mathcal{L}_\tau^{\mathrm{KTO}} \to \mathcal{L}^{\mathrm{KTO}}$, recovering the standard empirical KTO objective.

F Monotonicity of the Risk-Adjusted Operator

In the body, we demonstrated the conditions under which it was sufficient to consider a single player fixed point iteration (Theorem 1), as many of our practical implementations followed that framework. These conditions need not hold when the players use different risk functionals or the preference matrix does not satisfy $P + P^\top = \mathbf{1}\mathbf{1}^\top$. In those settings, the operator identity $P_{\mathcal{R}}^{(1)}\mu = \mathbf{1} + P_{\mathcal{R}}^{(2)}\mu$ need not hold, so the same argument does not justify a reduction to symmetric self-play.

F.1 Translation Invariance Preserves Constant Sum Structure

The risk-neutral analysis of NLHF relies on the constant-sum identity $P + P^\top = \mathbf{1}\mathbf{1}^\top$ to collapse the coupled fixed-point equations (5) into the single expression (2) satisfied by both players at equilibrium. Under nonlinear risk transformations, this structure no longer holds at the level of certainty equivalents: for a general payoff-side risk functional $\mathcal{R}$, one has $\mathcal{R}(X) \neq -\mathcal{R}(-X)$, so zero-sum structure is broken at the utility level. Indeed, as observed in Example 1, at the certainty equivalent level (aka the utility level) risk transformations are nonlinear and therefore break the zero-sum structure of the preference learning problem (or any zero sum matrix game for that matter).

We first show that despite this apparent shortcoming, when both players use the same translation-equivariant payoff risk functional, the property $\mathcal{R}(1 - X) = 1 + \mathcal{R}(-X)$ ensures that their payoff operators differ by the constant vector $\mathbf{1}$. This operator identity does not assert that their risk-adjusted utilities sum to a constant. We formally restate and prove Lemma 1 from the main.

Lemma 1 (Operator-level constant-sum)**.** Let $P \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ satisfy $P + P^\top = \mathbf{1}\mathbf{1}^\top$. Let $\mathcal{R}$ be translation invariant, i.e., $\mathcal{R}(X + c) = \mathcal{R}(X) + c$ for all constants $c \in \mathbb{R}$. For $\mu \in \Delta(\mathcal{Y})$, define the risk-adjusted payoff operators $(\mathcal{P}_{\mathcal{R}}^1\mu)_a := \mathcal{R}(\mathcal{P}_{a,Y'})$, and $(\mathcal{P}_{\mathcal{R}}^2\mu)_a := \mathcal{R}(-\mathcal{P}_{Y',a})$, with $Y' \sim \mu$. Then, for every $\mu \in \Delta(\mathcal{Y})$, we have $\mathcal{P}_{\mathcal{R}}^1\mu = \mathbf{1} + \mathcal{P}_{\mathcal{R}}^2\mu$.

Proof. Fix $a \in \mathcal{Y}$ and let $Y' \sim \mu$. Since $P + P^\top = \mathbf{1}\mathbf{1}^\top$, we have

$$P_{a,Y'} = 1 - P_{Y',a}.$$

Therefore, by translation equivariance of $\mathcal{R}$, we have that

$$(P_{\mathcal{R}}^{(1)}\mu)_a = \mathcal{R}(P_{a,Y'}) = \mathcal{R}(1 - P_{Y',a}) = 1 + \mathcal{R}(-P_{Y',a}) = 1 + (P_{\mathcal{R}}^{(2)}\mu)_a.$$

Since this holds for every $a \in \mathcal{Y}$, the vector identity follows. □

Gauge invariance. The softmax map $\mathrm{softmax}(\theta) := \frac{e^\theta}{\mathbf{1}^\top e^\theta}$ satisfies $\mathrm{softmax}(\theta + c\mathbf{1}) = \mathrm{softmax}(\theta)$ for every scalar c. Adding a constant to all entries of a logit vector therefore leaves the policy unchanged; this is the one-dimensional gauge of the parameterization.

Translation-equivariant payoff risk functionals. The payoff-side risk functionals used in the paper are translation equivariant, in the sense of Definition 1. In the remainder of this section we verify this property for the two examples we use most often, entropic risk and lower-tail payoff CVaR.

Lemma 2 (Entropic payoff risk obeys translation equivariance)**.** Recall that the entropic payoff risk of a random variable Z with risk parameter τ is given by $\mathrm{Ent}_\tau(Z) = -\frac{1}{\tau}\log\left(\mathbb{E}[\exp(-\tau Z)]\right)$. For any constant c we have $\mathrm{Ent}_\tau(Z + c) = \mathrm{Ent}_\tau(Z) + c$.

Proof. The result is immediate from the following simple algebraic manipulations:

$$\begin{aligned}\mathrm{Ent}_\tau(Z+c) &= -\frac{1}{\tau}\log\left(\mathbb{E}[\exp(-\tau Z-\tau c)]\right)\\ &= -\frac{1}{\tau}\log\big(\exp(-\tau c)\,\mathbb{E}[\exp(-\tau Z)]\big)\\ &= c-\frac{1}{\tau}\log\mathbb{E}[\exp(-\tau Z)]\\ &= \mathrm{Ent}_\tau(Z)+c.\end{aligned}$$

□

We now prove the same for payoff-side CVaR. The mechanism is different: entropic risk-aversion factors a constant shift through the exponential, whereas CVaR uses translation equivariance directly. Yet, the conclusion is identical.

For a real-valued random variable Z with distribution function F_Z, define the lower q-quantile $\mathrm{Var}_q(Z) := \inf\{z : F_Z(z) \geq q\}$ and the lower-tail conditional expectation

$$\mathrm{CVaR}_q(Z) \;:=\; \frac{1}{q}\,\mathbb{E}\big[Z\cdot\mathbf{1}\{Z\leq \mathrm{Var}_q(Z)\}\big] \;+\; \text{atom-correction} \;=\; \frac{1}{q}\int_0^q \mathrm{Var}_u(Z)\,\mathrm{d}u, \tag{21}$$

the second equality holding for general distributions (Rockafellar & Uryasev, 2002). Risk-averse use of CVaR takes the *lower* tail, with $q \in (0,1)$ the tail probability mass; $q \to 1$ recovers $\mathbb{E}[Z]$, $q \to 0$ recovers $\mathrm{essinf}(Z)$. The Rockafellar–Uryasev variational formula

$$\mathrm{CVaR}_q(Z) \;=\; \sup_\eta\Big\{\eta - \tfrac{1}{q}\,\mathbb{E}\big[(\eta-Z)_+\big]\Big\} \tag{22}$$

makes the translation property transparent: the supremum over η shifts by c when Z is replaced by $Z+c$.

Lemma 3 (Translation equivariance of payoff CVaR)**.** For any real-valued random variable Z on a probability space and any deterministic constant $c \in \mathbb{R}$,

$$\mathrm{CVaR}_q(Z+c) \;=\; \mathrm{CVaR}_q(Z)+c.$$

Proof. By (22), the change of variable $\eta \mapsto \eta + c$ gives

$$\begin{aligned}\mathrm{CVaR}_q(Z+c) &= \sup_\eta\big\{\eta - \tfrac{1}{q}\,\mathbb{E}[(\eta-(Z+c))_+]\big\}\\ &= \sup_{\eta'}\big\{(\eta'+c) - \tfrac{1}{q}\,\mathbb{E}[(\eta'-Z)_+]\big\}\\ &= c+\mathrm{CVaR}_q(Z).\end{aligned}$$

Equivalently, quantiles shift by c—i.e., $\mathrm{Var}_q(Z+c) = \mathrm{Var}_q(Z)+c$—and (21) preserves this shift through the conditional expectation. □

F.2 Symmetric Self-play is Enough

For the ease of reference, we recall Proposition 1 from the main body before proving it.

Proposition 1 (Equilibrium-set symmetry)**.** Under the hypotheses of Lemma 1, the RSPG equilibrium set is invariant under player exchange, modulo the softmax gauge. Consequently, if the RSPG equilibrium is unique in policy space, then it is symmetric.

Proof. Let $(\pi_{\theta_1^\star}, \pi_{\theta_2^\star})$ be any Nash equilibrium, with logit representatives satisfying the coupled fixed-point equations (5). Fix an action $a \in \mathcal{Y}$ and draw $Y' \sim \pi_{\theta_1^\star}$. By the definitions in Lemma 1,

$$(\mathcal{P}^1_{\mathcal{R}}\,\pi_{\theta_1^\star})_a = \mathcal{R}(P_{a,Y'}), \qquad (\mathcal{P}^2_{\mathcal{R}}\,\pi_{\theta_1^\star})_a = \mathcal{R}(-P_{Y',a}).$$

In particular, player 2's operator applies $\mathcal{R}$ to its negative payoff $-P_{Y',a}$. The constant-sum identity $P + P^\top = \mathbf{1}\mathbf{1}^\top$ gives, for each realization of Y',

$$P_{a,Y'} = 1 - P_{Y',a}, \qquad \Longleftrightarrow \qquad -P_{Y',a} = P_{a,Y'} - 1.$$

Applying the shared payoff risk functional to this equality yields

$$(\mathcal{P}^2_{\mathcal{R}}\,\pi_{\theta_1^\star})_a = \mathcal{R}(-P_{Y',a}) = \mathcal{R}(P_{a,Y'} - 1) = \mathcal{R}(P_{a,Y'}) - 1 = (\mathcal{P}^1_{\mathcal{R}}\,\pi_{\theta_1^\star})_a - 1.$$

The third equality uses translation equivariance, $\mathcal{R}(X + c) = \mathcal{R}(X) + c$, with $X = P_{a,Y'}$ and $c = -1$. Since this holds for every a, player 2's fixed-point equation becomes

$$\theta_2^\star = \theta_{\text{ref}} + \frac{\mathcal{P}^2_{\mathcal{R}}\,\pi_{\theta_1^\star}}{\beta} = \theta_{\text{ref}} + \frac{\mathcal{P}^1_{\mathcal{R}}\,\pi_{\theta_1^\star}}{\beta} - \beta^{-1}\mathbf{1}.$$

Thus define $\widetilde{\theta}_2 := \theta_2^\star + \beta^{-1}\mathbf{1}$ to absorb this constant shift, and note $\pi_{\theta_2^\star} = \pi_{\widetilde{\theta}_2}$ by gauge invariance. Together with player 1's fixed-point equation, this yields

$$\theta_1^\star = \theta_{\text{ref}} + \frac{\mathcal{P}^1_{\mathcal{R}}\,\pi_{\widetilde{\theta}_2}}{\beta}, \qquad \widetilde{\theta}_2 = \theta_{\text{ref}} + \frac{\mathcal{P}^1_{\mathcal{R}}\,\pi_{\theta_1^\star}}{\beta},$$

which are symmetric in $\theta_1^\star$ and $\widetilde{\theta}_2$. Exchanging them and undoing the shift in the second component shows that $(\widetilde{\theta}_2, \theta_1^\star - \beta^{-1}\mathbf{1})$ also satisfies (5). By gauge invariance, this pair induces the swapped policies $(\pi_{\theta_2^\star}, \pi_{\theta_1^\star})$. Thus the equilibrium set is invariant under player exchange. If the equilibrium is unique in policy space, the swapped pair must coincide with the original pair, giving $\pi_{\theta_1^\star} = \pi_{\theta_2^\star}$. □

F.3 The two-player risk-adjusted game

In the specific case of a constant sum preference matrix and shared translation-equivariant payoff risk functionals, we have just shown how symmetric self-play fixed points yield Nash equilibria. However, there are many settings under which it may be desirable to break one or more of those conditions. For example one might want to solve a preference game with *asymmetric* risk profiles, considering both a risk-averse and a risk-seeking player. In these settings, symmetry is no longer guaranteed, and it need not suffice to consider symmetric self play. The goal of this section is to study sufficient conditions for monotonicity of the joint game operator. The argument depends on the regularity and distortion bounds stated below and also applies to risk-seeking evaluations satisfying those bounds, even though they fall outside Definition 1. We specialize these results to the symmetric self play operator in Section F.4.

Recall that each player $i \in \{1, 2\}$ chooses logits $\theta_i \in \mathbb{R}^{|\mathcal{Y}|}$ inducing a softmax policy $\pi_{\theta_i} \in \Delta(\mathcal{Y})$. The risk-adjusted utility for player i is

$$V_\beta^{\mathcal{R}_i}(\theta_1, \theta_2) \;=\; \pi_{\theta_i}^\top P^{(i)}_{\mathcal{R}_i}\,\pi_{\theta_{-i}} \;-\; \beta\,\mathrm{KL}(\pi_{\theta_i} \| \pi_{\text{ref}}),$$

where $P^{(1)}_{\mathcal{R}_1} = P_{\mathcal{R}_1}$ is the risk-adjusted preference operator from (3), and player 2 applies its payoff risk functional to its own payoff $-\mathcal{P}(y \succ y')$:

$$\big(P^{(2)}_{\mathcal{R}_2}\,\mu\big)_{y'} \;=\; \mathcal{R}_2\big[-\mathcal{P}(y \succ y') \;:\; y \sim \mu\big].$$

The joint Nash fixed-point equation. Player i's first-order condition (set $\nabla_{\theta_i} V_\beta^{\mathcal{R}_i} = 0$ and use softmax stationarity) gives, after choosing logit representatives to absorb the per-player additive constants,

$$\theta_i^\star \;=\; \theta_{\text{ref}} \;+\; \frac{P^{(i)}_{\mathcal{R}_i}\,\pi_{\theta^\star_{-i}}}{\beta}, \qquad i = 1, 2. \tag{23}$$

This is a *coupled* pair of fixed-point equations. The natural object for studying learning dynamics targeting these is the policy-space *pseudogradient* (Rosen, 1965), represented up to per-player constant vectors in logit coordinates by

$$\mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}) \;:=\; \begin{pmatrix} \beta(\theta_1 - \theta_{\text{ref}}) - P^{(1)}_{\mathcal{R}_1}\,\pi_{\theta_2} \\ \beta(\theta_2 - \theta_{\text{ref}}) - P^{(2)}_{\mathcal{R}_2}\,\pi_{\theta_1} \end{pmatrix}, \qquad \boldsymbol{\theta} := (\theta_1, \theta_2), \tag{24}$$

whose zeros coincide with the Nash equilibria (23). Operationally, with the sampling distribution ρ_i and opponent policy held fixed during differentiation, player i's IPO gradient is

$$g_i(\boldsymbol{\theta}) = \frac{2}{\beta}\,\Sigma(\rho_i)\,\mathbf{F}_{\mathcal{R},i}(\boldsymbol{\theta}),$$

where $\mathbf{F}_{\mathcal{R},i}$ is the ith block of (24) and $\Sigma(\rho_i)$ is defined in (12). Thus IPO-based updates use a block-diagonal preconditioner on $\mathbf{F}_{\mathcal{R}}$. Last-iterate linear convergence requires conditions on this preconditioner, the update geometry, and the step size in addition to strong monotonicity.

We will need the following lemma in order to study (24).

Lemma 4 (Softmax monotonicity)**.** For any $\theta, \theta' \in \mathbb{R}^{|\mathcal{Y}|}$ with induced softmax policies π, π',

$$\langle \theta - \theta', \pi - \pi' \rangle = \mathrm{KL}(\pi \| \pi') + \mathrm{KL}(\pi' \| \pi).$$

Proof. Write $Z(\theta) = \sum_y e^{\theta_y}$. Since

$$\log \frac{\pi_y}{\pi'_y} = \theta_y - \theta'_y - \log Z(\theta) + \log Z(\theta'),$$

and $\sum_y (\pi_y - \pi'_y) = 0$, we obtain

$$\mathrm{KL}(\pi \| \pi') + \mathrm{KL}(\pi' \| \pi) = \sum_y (\pi_y - \pi'_y) \log \frac{\pi_y}{\pi'_y} = \langle \theta - \theta', \pi - \pi' \rangle.$$

□

We now seek $\mu_{\mathcal{R}} > 0$ such that

$$\langle \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}) - \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}'),\, \boldsymbol{\pi} - \boldsymbol{\pi}' \rangle \;\geq\; \mu_{\mathcal{R}} \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2^2, \tag{25}$$

where $\boldsymbol{\pi} = (\pi_{\theta_1}, \pi_{\theta_2})$, the inner product and norm are Euclidean on the product policy space. For joint policies $\boldsymbol{\pi} = (\pi_{\theta_1}, \pi_{\theta_2})$ and $\boldsymbol{\pi}' = (\pi_{\theta'_1}, \pi_{\theta'_2})$, define the *product KL divergence*

$$D(\boldsymbol{\pi}, \boldsymbol{\pi}') := \sum_{i=1}^{2} \mathrm{KL}(\pi_{\theta_i} \| \pi_{\theta'_i}).$$

Thus D is the sum of the two players' KL divergences, with no factor of $1/2$. The reverse divergence $D(\boldsymbol{\pi}', \boldsymbol{\pi})$ reverses the arguments in each KL term.

Write $\Delta\theta_i := \theta_i - \theta'_i$ and $\Delta\pi_i := \pi_{\theta_i} - \pi'_{\theta_i}$ for $i = 1, 2$. Expanding (25) using (24), we have

$$\langle \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}) - \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}'), \boldsymbol{\pi} - \boldsymbol{\pi}' \rangle = \beta \sum_{i=1}^{2} \underbrace{\langle \Delta\theta_i, \Delta\pi_i \rangle}_{(\mathrm{A}_i)\text{: own-player softmax}} - \sum_{i=1}^{2} \underbrace{\langle P^{(i)}_{\mathcal{R}_i} \pi_{\theta_{-i}} - P^{(i)}_{\mathcal{R}_i} \pi'_{\theta_{-i}},\, \Delta\pi_i \rangle}_{(\mathrm{B}_i)\text{: cross-player coupling}}.$$

The own-player terms (A_i) are handled by softmax monotonicity (Lemma 4) applied to each player separately:

$$\sum_i (\mathrm{A}_i) \;\geq\; \sum_{i=1}^{2} \big[\mathrm{KL}(\pi_{\theta_i} \| \pi'_{\theta_i}) + \mathrm{KL}(\pi'_{\theta_i} \| \pi_{\theta_i}) \big]. \tag{26}$$

The cross-player terms (B_i) are the new objects, and they are now genuinely *bilinear across players* rather than quadratic in a single difference vector.

For the following Jacobian argument, assume that each risk-adjusted preference operator is continuously differentiable on the relative interior of the simplex. To study (B_i), define the Jacobians

$$J^{(i)}_{\mathcal{R}_i}(\pi) \;:=\; \nabla_\mu P^{(i)}_{\mathcal{R}_i}(\mu)\Big|_{\mu=\pi} \;\in\; \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}.$$

Only their action on zero-sum directions is needed. Set $\pi_i(t) := \pi_{\theta'_i} + t\Delta\pi_i$ for $t \in [0, 1]$. The fundamental theorem of calculus gives

$$P^{(i)}_{\mathcal{R}_i} \pi_{\theta_{-i}} - P^{(i)}_{\mathcal{R}_i} \pi_{\theta'_{-i}} = \int_0^1 J^{(i)}_{\mathcal{R}_i}(\pi_{-i}(t))\, \Delta\pi_{-i}\, dt.$$

In general, this vector identity cannot be replaced by a single Jacobian evaluated at an intermediate point. Thus $(\mathrm{B}_i) = \int_0^1 \Delta\pi_i^\top J^{(i)}_{\mathcal{R}_i}(\pi_{-i}(t))\, \Delta\pi_{-i}\, dt$. Stacking $\Delta\boldsymbol{\pi} := (\Delta\pi_1, \Delta\pi_2) \in \mathbb{R}^{2|\mathcal{Y}|}$, the

cross-player contribution is

$$\sum_{i=1}^{2}(\mathrm{B}_i) = \int_0^1 \Delta\boldsymbol{\pi}^\top \mathcal{J}_{\mathcal{R}}(\pi_1(t),\pi_2(t))\,\Delta\boldsymbol{\pi}\,dt,$$
$$\mathcal{J}_{\mathcal{R}}(\pi_1,\pi_2) := \begin{pmatrix} 0 & J^{(1)}_{\mathcal{R}_1}(\pi_2) \\ J^{(2)}_{\mathcal{R}_2}(\pi_1) & 0 \end{pmatrix}. \tag{27}$$

Only the symmetric part of $\mathcal{J}_{\mathcal{R}}$ contributes to the quadratic form:

$$\mathcal{J}^{\mathsf{sym}}_{\mathcal{R}}(\pi_1,\pi_2) \;=\; \tfrac{1}{2}\begin{pmatrix} 0 & J^{(1)}_{\mathcal{R}_1}(\pi_2) + J^{(2)}_{\mathcal{R}_2}(\pi_1)^\top \\ J^{(2)}_{\mathcal{R}_2}(\pi_1) + J^{(1)}_{\mathcal{R}_1}(\pi_2)^\top & 0 \end{pmatrix}.$$

Risk-neutral baseline. When $\mathcal{R}_1 = \mathcal{R}_2 = \mathbb{E}$, $J^{(1)}_{\mathbb{E}} = P$ and $J^{(2)}_{\mathbb{E}} = -P^\top$ identically, so the off-diagonal block of $\mathcal{J}^{\mathsf{sym}}_{\mathbb{E}}$ is

$$\tfrac{1}{2}\big(J^{(1)}_{\mathbb{E}} + (J^{(2)}_{\mathbb{E}})^\top\big) \;=\; \tfrac{1}{2}\big(P + (-P^\top)^\top\big) \;=\; \tfrac{1}{2}(P-P) \;=\; 0.$$

Hence $\mathcal{J}^{\mathsf{sym}}_{\mathbb{E}} \equiv 0$, on the full space and a fortiori on the joint zero-sum subspace. This is the constant-sum miracle: the cross-player coupling makes *no* contribution to the monotonicity quadratic form, leaving only the softmax-monotonicity bound and yielding the β-strong-monotonicity bound in policy space. This cancellation uses the risk-neutral payoff matrices P and $-P^\top$ and holds even without the additional preference symmetry $P + P^\top = \mathbf{1}\mathbf{1}^\top$.

Joint risk-distortion matrix. Since the risk-neutral subtraction is zero, define

$$\boldsymbol{\mathcal{J}}_{\mathcal{R}}(\pi_1,\pi_2) \;:=\; \mathcal{J}^{\mathsf{sym}}_{\mathcal{R}}(\pi_1,\pi_2). \tag{28}$$

The joint constraint set and its tangent space. The pairing in (25) is in policy space. Let $C \subseteq \Delta(\mathcal{Y}) \times \Delta(\mathcal{Y})$ be a nonempty closed convex policy constraint set, and write $\Pi := C \cap (\Delta^\circ(\mathcal{Y}) \times \Delta^\circ(\mathcal{Y}))$, assumed nonempty. Let $T_C := \mathrm{span}(C - C)$ be its direction space, which contains all policy displacements. Convexity ensures that the policy segments in the preceding integrals remain in Π. Choosing logit representatives $\theta_i = \theta_{\mathrm{ref}} + \log(\pi_i/\pi_{\mathrm{ref}})$ gives

$$\mathbf{F}_{\mathcal{R},i}(\boldsymbol{\theta}) = \beta \log \frac{\pi_i}{\pi_{\mathrm{ref}}} - P^{(i)}_{\mathcal{R}_i}\pi_{-i}, \qquad i = 1,2.$$

Other logit representatives add only constant vectors, which vanish when paired with policy displacements. The relevant variational inequality is

$$\big\langle \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}^\star), \boldsymbol{\pi} - \boldsymbol{\pi}^\star \big\rangle \geq 0 \qquad \text{for every } \boldsymbol{\pi} \in C.$$

For a product set $C = C_1 \times C_2$, this is the Nash condition for the game with the corresponding policy constraints. For a coupled set, we refer to its solutions as solutions of this joint variational inequality. The worst-case risk-distortion eigenvalue is

$$\bar{\lambda}_{\mathcal{R}} \;:=\; \sup_{(\bar\pi_1,\bar\pi_2)\in\Pi} \; \sup_{v\in T_C,\,\|v\|_2=1} v^\top \boldsymbol{\mathcal{J}}_{\mathcal{R}}(\bar\pi_1,\bar\pi_2)\,v. \tag{29}$$

The inner supremum is the largest Rayleigh quotient restricted to T_C. Symmetrization does not introduce an additional factor of two: $v^\top \mathcal{J}^{\mathsf{sym}}_{\mathcal{R}} v = v^\top \mathcal{J}_{\mathcal{R}} v$. Thus (29) bounds the distortion quadratic form by $\bar{\lambda}_{\mathcal{R}}\|v\|_2^2$ on T_C. In the unconstrained game, $C = \Delta(\mathcal{Y}) \times \Delta(\mathcal{Y})$ and $T_C = \mathbf{1}^\perp \times \mathbf{1}^\perp$. On this full product space, $\bar{\lambda}_{\mathcal{R}} \geq 0$ because the distortion matrix has zero diagonal blocks. For symmetric policy pairs, $T_C = T^{\mathsf{sym}} = \{(\xi,\xi) : \xi \in \mathbf{1}^\perp\}$, as treated in Lemma 5.

Combining (26) with (27) and (28), we have the signed bound

$$\begin{aligned}\big\langle \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}) - \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}'),\, \boldsymbol{\pi} - \boldsymbol{\pi}' \big\rangle &\geq \beta\big[D(\boldsymbol{\pi},\boldsymbol{\pi}') + D(\boldsymbol{\pi}',\boldsymbol{\pi})\big] - \int_0^1 \Delta\boldsymbol{\pi}^\top \boldsymbol{\mathcal{J}}_{\mathcal{R}}(\pi_1(t),\pi_2(t))\,\Delta\boldsymbol{\pi}\,dt \\ &\geq \beta\big[D(\boldsymbol{\pi},\boldsymbol{\pi}') + D(\boldsymbol{\pi}',\boldsymbol{\pi})\big] - \bar{\lambda}_{\mathcal{R}}\|\Delta\boldsymbol{\pi}\|_2^2 \\ &\geq (\beta - \bar{\lambda}_{\mathcal{R}})\|\Delta\boldsymbol{\pi}\|_2^2, \end{aligned} \tag{30}$$

where $\bar{\lambda}_{\mathcal{R}}$ is defined in (29). By that definition, the integral is at most $\bar{\lambda}_{\mathcal{R}}\|\Delta\boldsymbol{\pi}\|_2^2$. For the last inequality, apply Pinsker's inequality separately to each of the two KL terms:

$$D(\boldsymbol{\pi}, \boldsymbol{\pi}') \geq \tfrac{1}{2}\sum_{i=1}^{2}\|\pi_{\theta_i} - \pi_{\theta_i'}\|_2^2 = \tfrac{1}{2}\|\Delta\boldsymbol{\pi}\|_2^2, \quad \text{and} \quad D(\boldsymbol{\pi}', \boldsymbol{\pi}) \geq \tfrac{1}{2}\|\Delta\boldsymbol{\pi}\|_2^2.$$

Adding the two inequalities gives

$$D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi}) \geq \|\Delta\boldsymbol{\pi}\|_2^2. \tag{31}$$

Only the regularization term is bounded this way, so the argument holds for either sign of $\bar{\lambda}_{\mathcal{R}}$.

Theorem 5 (Strong monotonicity of the joint pseudogradient)**.** Let C be a closed convex policy constraint set with direction space T_C, and let $\bar{\lambda}_{\mathcal{R}}$ be defined by (29) on T_C. If $\bar{\lambda}_{\mathcal{R}} < \beta$, then the pseudogradient $\mathbf{F}_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone with respect to the Euclidean norm on the positive policies in C, in the sense of (25), with modulus

$$\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}} > 0.$$

KL consequence. We derive the KL bound directly from the penultimate line of (30), which retains both the regularization and distortion terms. This is a separate consequence of that line, not a conversion of the final Euclidean bound. The sign of $\bar{\lambda}_{\mathcal{R}}$ determines how Pinsker's inequality can be used.

Case 1: $0 \leq \bar{\lambda}_{\mathcal{R}} < \beta$. Multiplying (31) by $-\bar{\lambda}_{\mathcal{R}} \leq 0$ gives

$$-\bar{\lambda}_{\mathcal{R}}\|\Delta\boldsymbol{\pi}\|_2^2 \geq -\bar{\lambda}_{\mathcal{R}}\big[D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi})\big].$$

Substituting into the entire regularization-minus-distortion expression, we obtain

$$\begin{aligned}\big\langle \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}) - \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}'), \boldsymbol{\pi} - \boldsymbol{\pi}' \big\rangle &\geq \beta\big[D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi})\big] - \bar{\lambda}_{\mathcal{R}}\|\Delta\boldsymbol{\pi}\|_2^2 \\ &\geq \beta\big[D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi})\big] - \bar{\lambda}_{\mathcal{R}}\big[D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi})\big] \\ &= (\beta - \bar{\lambda}_{\mathcal{R}})\big[D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi})\big] \\ &\geq \mu_{\mathcal{R}} D(\boldsymbol{\pi}, \boldsymbol{\pi}'). \end{aligned} \tag{32}$$

The last step uses $\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}} > 0$ and drops the nonnegative reverse-KL term. The distortion subtraction has already been included in the coefficient $\beta - \bar{\lambda}_{\mathcal{R}}$.

Case 2: $\bar{\lambda}_{\mathcal{R}} < 0$. Here $-\bar{\lambda}_{\mathcal{R}}\|\Delta\boldsymbol{\pi}\|_2^2 \geq 0$. We retain this sign and do not replace the squared norm by KL:

$$\begin{aligned}\big\langle \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}) - \mathbf{F}_{\mathcal{R}}(\boldsymbol{\theta}'), \boldsymbol{\pi} - \boldsymbol{\pi}' \big\rangle &\geq \beta\big[D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi})\big] - \bar{\lambda}_{\mathcal{R}}\|\Delta\boldsymbol{\pi}\|_2^2 \\ &\geq \beta\big[D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi})\big] \geq \beta D(\boldsymbol{\pi}, \boldsymbol{\pi}'). \end{aligned}$$

Thus a positive Euclidean modulus $\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}}$ also gives a positive product-KL bound: its coefficient is $\beta - \bar{\lambda}_{\mathcal{R}}$ in Case 1 and β in Case 2. In Case 2 the favorable distortion term still improves the Euclidean modulus; the KL coefficient β is a lower bound, not a claim of optimality.

For comparison, when $\bar{\lambda}_{\mathcal{R}} \geq 0$, using only the forward-direction bound $\|\Delta\boldsymbol{\pi}\|_2^2 \leq 2D(\boldsymbol{\pi}, \boldsymbol{\pi}')$ would give

$$\beta\big[D(\boldsymbol{\pi}, \boldsymbol{\pi}') + D(\boldsymbol{\pi}', \boldsymbol{\pi})\big] - \bar{\lambda}_{\mathcal{R}}\|\Delta\boldsymbol{\pi}\|_2^2 \geq (\beta - 2\bar{\lambda}_{\mathcal{R}})D(\boldsymbol{\pi}, \boldsymbol{\pi}') + \beta D(\boldsymbol{\pi}', \boldsymbol{\pi}).$$

Dropping the reverse divergence gives the weaker coefficient $\beta - 2\bar{\lambda}_{\mathcal{R}}$. Keeping both KL terms as in (32) improves this to $\beta - \bar{\lambda}_{\mathcal{R}}$ without changing the normalization of the operator, the distortion matrix, or the product KL divergence.

Consequently, any Nash equilibrium satisfying (23) in C is unique. Under the standard regularity and step-size conditions for the corresponding VI geometry, deterministic extragradient and Euclidean mirror descent applied to the policy-space VI converge last-iterate linearly in squared Euclidean distance at rate $\mathcal{O}\big((1 - \eta\mu_{\mathcal{R}})^T\big)$, with the usual constant rescaling of the step size (Tseng, 1995; Nemirovski, 2004).

For nonsmooth risks, the argument requires an absolutely continuous restriction to each policy segment, the integral identity used in (27), and a uniform bound on the derivatives appearing in that identity.

Three regimes. The sufficient bound above distinguishes three regimes:

- *Risk-aligned with regularization* ($\bar{\lambda}_{\mathcal{R}} \leq 0$): the cross-player contribution cannot reduce the monotonicity pairing, and the modulus is strengthened relative to the risk-neutral baseline: $\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}} \geq \beta$, with strict improvement when $\bar{\lambda}_{\mathcal{R}} < 0$. Negative values can occur on restricted direction spaces such as symmetric self-play; on the full product direction space, this regime has $\bar{\lambda}_{\mathcal{R}} = 0$.
- *Moderate risk* ($0 < \bar{\lambda}_{\mathcal{R}} < \beta$): strong monotonicity survives with reduced modulus $0 < \mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}} < \beta$.
- *Beyond the sufficient threshold* ($\bar{\lambda}_{\mathcal{R}} \geq \beta$): this bound supplies no positive modulus. At equality it still proves monotonicity, but not strong monotonicity; above the threshold it is inconclusive. Neither case alone establishes nonuniqueness or cycling.

Discussion. A consequence of the joint formulation is that the two players need not share a risk functional: the monotonicity criterion applies whenever the stated regularity and distortion bounds hold. Setting $\mathcal{R}_1 \neq \mathcal{R}_2$ corresponds to modeling scenarios in which the players evaluate uncertainty differently, for example one might wish to run an optimistic player against a pessimistic player. We also note that Theorem 1 relies on the constant-sum identity. For preference operators P that are not constant-sum (i.e. where $P + P^\top \neq \mathbf{1}\mathbf{1}^\top$), $P^{(1)}_{\mathcal{R}}\,\mu - P^{(2)}_{\mathcal{R}}\,\mu$ is not in general a multiple of $\mathbf{1}$. Thus the gauge equivalence argument does not apply in general, and an equilibrium need not lie on the diagonal. The joint formulation covers such asymmetric equilibria; symmetry or convergence requires the corresponding conditions to be checked.

F.4 Symmetric Game via Self-Play

The above results establish the monotonicity of the full joint pseudogradient (24) over an arbitrary constraint set of joint policies. However, in the main paper we study the symmetric self-play setting and the associated symmetric self-play operator. We now specialize the monotonicity analysis to this case. Throughout we adopt the notation and standards of Dontchev & Rockafellar (2009a); Rockafellar & Wets (1998).

Let us first formalize the symmetric self-play constraint space: symmetric self-play corresponds to the constraint set

$$C^{\mathsf{sym}} \;:=\; \big\{\, v = (\theta, \theta) \;:\; \theta \in \mathbb{R}^{|\mathcal{Y}|}/\mathbf{1} \,\big\} \;\subseteq\; \mathbb{R}^{|\mathcal{Y}|}/\mathbf{1} \times \mathbb{R}^{|\mathcal{Y}|}/\mathbf{1},$$

i.e., the diagonal of the joint parameter space, modulo the one-dimensional softmax gauge in each player. In policy space, the corresponding constraint set is the diagonal of the product simplex,

$$\Delta^{\mathsf{sym}} \;:=\; \big\{\, (\pi, \pi) \;:\; \pi \in \Delta(\mathcal{Y}) \,\big\} \;\subseteq\; \Delta(\mathcal{Y}) \times \Delta(\mathcal{Y}),$$

and the softmax map carries C^{sym} into Δ^{sym}.

Since C^{sym} is a linear subspace, its tangent cone is constant in v and coincides with C^{sym} itself:

$$T_{C^{\mathsf{sym}}}(v) \;=\; C^{\mathsf{sym}} \;=\; \big\{\, (\xi, \xi) \;:\; \xi \in \mathbf{1}^\perp \,\big\} \qquad \forall v \in C^{\mathsf{sym}},$$

where the identification with $\mathbf{1}^\perp$ in each component arises from the gauge quotient: a representative $\xi \in \mathbb{R}^{|\mathcal{Y}|}$ of an element of $\mathbb{R}^{|\mathcal{Y}|}/\mathbf{1}$ may be taken to satisfy $\sum_y \xi(y) = 0$—i.e., $\xi \in \mathbf{1}^\perp$. We will use the shorthand $T^{\mathsf{sym}} := T_{C^{\mathsf{sym}}}(v)$ throughout, suppressing the v-dependence since C^{sym} is linear.

In this notation, the symmetric-self-play variational inequality is the generalized equation

$$0 \;\in\; \boldsymbol{F}_{\mathcal{R}}(v) + N_{C^{\mathsf{sym}}}(v),$$

and the relevant strong-monotonicity condition is on $\boldsymbol{F}_{\mathcal{R}}$ restricted to displacements in T^{sym}.

Reduction to a single-player monotonicity condition. For two diagonal points $v_1 = (\theta_1, \theta_1)$ and $v_2 = (\theta_2, \theta_2)$ in C^{sym}, the displacement $v_1 - v_2 = (\theta_1 - \theta_2, \theta_1 - \theta_2)$ lies in T^{sym}. Direct computation gives

$$\big\langle \boldsymbol{F}_{\mathcal{R}}(v_1) - \boldsymbol{F}_{\mathcal{R}}(v_2),\, v_1 - v_2 \big\rangle \;=\; 2\,\big\langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2),\, \theta_1 - \theta_2 \big\rangle.$$

For this parameter-space analysis, we use $J^{(i)}_{\mathcal{R}_i}(\pi)$ to denote the Jacobian with respect to the opponent's logits, rather than the policy derivatives used in the preceding subsection. For $\pi = \pi_\theta$, define

$$J^{(i)}_{\mathcal{R}_i}(\pi_\theta) \;:=\; \frac{\partial}{\partial\theta}\big(P^{(i)}_{\mathcal{R}_i}\pi_\theta\big) \;=\; \nabla_\mu P^{(i)}_{\mathcal{R}_i}(\mu)\Big|_{\mu=\pi_\theta}\big[\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top\big],$$

where the bracketed matrix is the softmax Jacobian. In the setting of Theorem 1, we have $\mathcal{R}_1 = \mathcal{R}_2$ and $P + P^\top = \mathbf{1}\mathbf{1}^\top$, so Lemma 1 gives $P^{(1)}_{\mathcal{R}_1}\pi_\theta = P^{(2)}_{\mathcal{R}_2}\pi_\theta + \mathbf{1}$ for every θ. Differentiating this identity with respect to θ gives

$$J^{(1)}_{\mathcal{R}_1}(\pi_\theta) = \frac{\partial}{\partial\theta}\big(P^{(2)}_{\mathcal{R}_2}\pi_\theta + \mathbf{1}\big) = \frac{\partial}{\partial\theta}\big(P^{(2)}_{\mathcal{R}_2}\pi_\theta\big) = J^{(2)}_{\mathcal{R}_2}(\pi_\theta).$$

We therefore write $J^{(1)}_{\mathcal{R}_1}(\pi) = J^{(2)}_{\mathcal{R}_2}(\pi) = J_{\mathcal{R}}(\pi)$ for brevity, with $\pi = \pi_\theta$.

The risk-distortion eigenvalue on T^{sym}. In this subsection, $\boldsymbol{\mathcal{J}}_{\mathcal{R}}$ denotes the symmetric part of the joint payoff Jacobian in parameter space, formed from the logit Jacobians just defined:

$$\boldsymbol{\mathcal{J}}_{\mathcal{R}}(\pi_1,\pi_2) \;:=\; \tfrac{1}{2}\begin{pmatrix} 0 & J^{(1)}_{\mathcal{R}_1}(\pi_2) + J^{(2)}_{\mathcal{R}_2}(\pi_1)^\top \\ J^{(2)}_{\mathcal{R}_2}(\pi_1) + J^{(1)}_{\mathcal{R}_1}(\pi_2)^\top & 0 \end{pmatrix}. \tag{33}$$

We define the corresponding parameter-space risk-distortion eigenvalue on symmetric displacements by[2]

$$\bar\lambda_{\mathcal{R}} \;:=\; \sup_{(\bar\pi_1,\bar\pi_2)\in\Delta^{\mathsf{sym}}}\;\sup_{v\in T^{\mathsf{sym}},\,\|v\|_2=1} v^\top \boldsymbol{\mathcal{J}}_{\mathcal{R}}(\bar\pi_1,\bar\pi_2)\,v. \tag{34}$$

Here $\bar\lambda_{\mathcal{R}}$ refers to the parameter-space quantity and need not equal the policy-space eigenvalue in (29). Parameterizing T^{sym} by $v = (\xi,\xi)/\sqrt{2}$, so that $\|v\|_2 = \|\xi\|_2$, gives the following single-player form.

Lemma 5 (Reduction of the risk-distortion eigenvalue to a single-player form)**.** In the symmetric self-play setting of Theorem 1, with constraint sets C^{sym} and Δ^{sym} defined above and tangent space $T^{\mathsf{sym}} = \{(\xi,\xi) : \xi \in \mathbf{1}^\perp\}$, the parameter-space risk-distortion eigenvalue (34) admits the single-player form

$$\bar\lambda_{\mathcal{R}} \;=\; \sup_{\bar\pi\in\Delta(\mathcal{Y})}\;\sup_{\xi\in\mathbf{1}^\perp,\,\|\xi\|_2=1} \tfrac{1}{2}\,\xi^\top\big(J_{\mathcal{R}}(\bar\pi) + J_{\mathcal{R}}(\bar\pi)^\top\big)\,\xi.$$

Proof. We simplify each of the two suprema in (34) in turn.

Step 1: Collapsing the outer supremum. By definition of Δ^{sym}, every element is of the form $(\bar\pi,\bar\pi)$ for a single $\bar\pi \in \Delta(\mathcal{Y})$. Therefore

$$\sup_{(\bar\pi_1,\bar\pi_2)\in\Delta^{\mathsf{sym}}}(\,\cdot\,) \;=\; \sup_{\bar\pi\in\Delta(\mathcal{Y})}(\,\cdot\,) \quad \text{evaluated at } (\bar\pi_1,\bar\pi_2) = (\bar\pi,\bar\pi).$$

Moreover, by Theorem 1 and Lemma 1, when evaluated on the diagonal the two players' Jacobians agree: $J^{(1)}_{\mathcal{R}_1}(\bar\pi) = J^{(2)}_{\mathcal{R}_2}(\bar\pi) = J_{\mathcal{R}}(\bar\pi)$. The parameter-space matrix (33) therefore takes the block-antidiagonal form

$$\boldsymbol{\mathcal{J}}_{\mathcal{R}}(\bar\pi,\bar\pi) \;=\; \tfrac{1}{2}\begin{pmatrix} 0 & J_{\mathcal{R}}(\bar\pi) + J_{\mathcal{R}}(\bar\pi)^\top \\ J_{\mathcal{R}}(\bar\pi) + J_{\mathcal{R}}(\bar\pi)^\top & 0 \end{pmatrix}. \tag{35}$$

Step 2: Reducing the inner supremum. Every $v \in T^{\mathsf{sym}}$ is of the form $v = (\xi,\xi)$ with $\xi \in \mathbf{1}^\perp$. Reparameterize as $v = (\xi,\xi)/\sqrt{2}$ so that $\|v\|_2 = \|\xi\|_2$, making the unit-norm constraint $\|v\|_2 = 1$ equivalent to $\|\xi\|_2 = 1$.

Substituting into the quadratic form using (35), denote $A := J_{\mathcal{R}}(\bar\pi) + J_{\mathcal{R}}(\bar\pi)^\top$ for brevity. Then

$$\begin{aligned} v^\top\boldsymbol{\mathcal{J}}_{\mathcal{R}}(\bar\pi,\bar\pi)\,v \;&=\; \frac{1}{2}\cdot\frac{1}{2}\begin{pmatrix}\xi^\top & \xi^\top\end{pmatrix}\begin{pmatrix}0 & A\\ A & 0\end{pmatrix}\begin{pmatrix}\xi\\ \xi\end{pmatrix} \;=\; \frac{1}{4}\big(\xi^\top A\xi + \xi^\top A\xi\big) \\ &=\; \tfrac{1}{2}\,\xi^\top A\xi \;=\; \tfrac{1}{2}\,\xi^\top\big(J_{\mathcal{R}}(\bar\pi) + J_{\mathcal{R}}(\bar\pi)^\top\big)\xi. \end{aligned}$$

[2] Whenever the Jacobian extends continuously to the simplex boundary, the same supremum is obtained over the relative interior, by density. This equality is independent of the parameterization; softmax simply parameterizes the relative interior.

Now we combine these two steps to obtain the claimed expression. Indeed, substituting both simplifications into (34),

$$\bar{\lambda}_{\mathcal{R}} \;=\; \sup_{\bar{\pi}\in\Delta(\mathcal{Y})} \; \sup_{\xi\in\mathbf{1}^{\perp},\,\|\xi\|_2=1} \; \tfrac{1}{2}\,\xi^{\top}\big(J_{\mathcal{R}}(\bar{\pi}) + J_{\mathcal{R}}(\bar{\pi})^{\top}\big)\,\xi,$$

which is the claimed form. □

This is the parameter-space risk-distortion eigenvalue for symmetric self-play.

Recall the convergence result from the main paper—namely, Theorem 2. Let us restate it a little more formally.

Theorem 6 (Formal version of Theorem 2)**.** Define the worst-case risk-distortion eigenvalue on the symmetric tangent space $\bar{\lambda}_{\mathcal{R}} := \sup_{\bar{\pi}\in\Delta(\mathcal{Y})} \sup\{\tfrac{1}{2}\,\xi^{\top}\big(J_{\mathcal{R}}(\bar{\pi}) + J_{\mathcal{R}}(\bar{\pi})^{\top}\big)\,\xi|\ \xi\in\mathbf{1}^{\perp},\ \|\xi\|_2 = 1\}$. Suppose $\bar{\lambda}_{\mathcal{R}} < \beta$. Then $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone in the Euclidean norm on the gauge-fixed parameter space $\mathbf{1}^{\perp}$, with $\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}} > 0$, and deterministic extragradient converges with optimization error rate $\mathcal{O}\big((1-\eta\mu_{\mathcal{R}})^{T}\big)$.

Proof of the monotonicity bound. Let $\theta,\theta'\in\mathbf{1}^{\perp}$, set $h := \theta-\theta'$, and write $\theta(t) := \theta' + th$ for $t\in[0,1]$. By Lemma 5, $h^{\top}J_{\mathcal{R}}(\pi_{\theta(t)})h \le \bar{\lambda}_{\mathcal{R}}\|h\|_2^2$. Since $\nabla_\theta F_{\mathcal{R}}(\theta) = \beta I - J_{\mathcal{R}}(\pi_\theta)$, the fundamental theorem of calculus gives

$$\big\langle F_{\mathcal{R}}(\theta) - F_{\mathcal{R}}(\theta'), h\big\rangle = \int_0^1 h^{\top}\big[\beta I - J_{\mathcal{R}}(\pi_{\theta(t)})\big]h\,dt \ge (\beta-\bar{\lambda}_{\mathcal{R}})\|h\|_2^2 = \mu_{\mathcal{R}}\|\theta-\theta'\|_2^2.$$

□

F.5 Approaches for Bounding Δ

This subsection works in policy space. To distinguish the two coordinate systems, write

$$J^{\pi}_{\mathcal{R}}(\pi) := \nabla_\mu P_{\mathcal{R}}(\mu)\big|_{\mu=\pi}, \quad \text{and} \quad J^{\theta}_{\mathcal{R}}(\pi) = J^{\pi}_{\mathcal{R}}(\pi)\big[\mathrm{diag}(\pi) - \pi\pi^{\top}\big].$$

Here $J^{\theta}_{\mathcal{R}}$ is the parameter-space Jacobian denoted by $J_{\mathcal{R}}$ in Section F.4. We write $\bar{\lambda}^{\theta}_{\mathcal{R}}$ for its eigenvalue bound (34), and define the policy-space counterpart by

$$\bar{\lambda}^{\pi}_{\mathcal{R}} := \sup_{\bar{\pi}\in\Delta(\mathcal{Y})} \; \sup_{\substack{\xi\in\mathbf{1}^{\perp}\\ \|\xi\|_2=1}} \; \tfrac{1}{2}\,\xi^{\top}\big(J^{\pi}_{\mathcal{R}}(\bar{\pi}) + J^{\pi}_{\mathcal{R}}(\bar{\pi})^{\top}\big)\xi.$$

Throughout the remainder of this subsection, $J_{\mathcal{R}} = J^{\pi}_{\mathcal{R}}$ and $\mathcal{J}_{\mathcal{R}}$ denotes the corresponding policy-space distortion matrix. The monotonicity bounds below concern pairings with policy displacements $\pi-\pi'$. A bound on $\bar{\lambda}^{\pi}_{\mathcal{R}}$ alone does not give a bound on $\bar{\lambda}^{\theta}_{\mathcal{R}}$: applying these calculations to the parameter-space theorem 6 requires a bound on the symmetric part of $J^{\theta}_{\mathcal{R}}$ above. We present two approaches to bounding the policy-space distortion below.

Route A: dual representation. For a payoff-side risk functional with a divergence-penalized dual representation, write

$$\mathcal{R}[Z] = \inf_{q\in\mathcal{Q}}\big\{\mathbb{E}_q[Z] + D(q\|p)\big\},$$

where p is the reference law, $\mathcal{Q}$ is an ambiguity set, and D is a nonnegative divergence penalty. The coherent ambiguity-set representation is the special case $D\equiv 0$ on $\mathcal{Q}$. From this perspective, the risk transforms the preference operator as follows:

$$\begin{aligned}(P_{\mathcal{R}}\,\mu)_y &= \inf_{q\in\mathcal{Q}(\mu)}\left\{\int \mathcal{P}(y\succ y'')\,dq(y'') + D(q\|\mu)\right\}\\ &= \inf_{q\in\mathcal{Q}(\mu)}\big\{(P\,q)_y + D(q\|\mu)\big\}.\end{aligned}$$

Choose a minimizing distribution separately for each payoff row:

$$q^{\star}_y(\mu)\in\arg\min_{q\in\mathcal{Q}(\mu)}\big\{(Pq)_y + D(q\|\mu)\big\}.$$

Writing $d_y(\mu) := D(q_y^\star(\mu)\|\mu)$, the value is

$$(P_{\mathcal{R}}\mu)_y = (Pq_y^\star(\mu))_y + d_y(\mu).$$

Both the optimizer and its penalty can depend on y, so the penalty vector cannot generally be discarded as a multiple of $\mathbf{1}$. Where the selected maps $q_y^\star$ and d_y are differentiable, direct differentiation gives the rowwise policy Jacobian

$$[J_{\mathcal{R}}(\mu)]_{y,:} = P_{y,:}\nabla_\mu q_y^\star(\mu) + \nabla_\mu d_y(\mu)^\top.$$

The derivative of d_y includes its dependence on both $q_y^\star(\mu)$ and μ.

Sufficient condition. The common-map factorization applies if there is a differentiable minimizer $q^\star(\mu)$ shared by all rows. In that case set $S(\mu) := \nabla_\mu q^\star(\mu)$ and $d(\mu) := D(q^\star(\mu)\|\mu)$. Then

$$J_{\mathcal{R}}(\mu) = PS(\mu) + \mathbf{1}\nabla_\mu d(\mu)^\top.$$

Using $P = \frac{1}{2}\mathbf{1}\mathbf{1}^\top + A$ with $A = -A^\top$, the terms with a left factor $\mathbf{1}$ vanish in zero-sum quadratic forms. Thus, for $\xi \in \mathbf{1}^\perp$,

$$\xi^\top \mathcal{J}_{\mathcal{R}}(\mu)\xi = \tfrac{1}{2}\xi^\top\big(AS(\mu) - S(\mu)^\top A\big)\xi. \tag{36}$$

If S is symmetric, the matrix $AS - SA$ is symmetric as well: $(AS - SA)^\top = AS - SA$. Its quadratic form need not vanish. A sufficient condition for zero curvature is the additional commutation identity $AS = SA$.

Proposition 4 (Zero curvature for a commuting symmetric dual Jacobian)**.** Suppose the coherent dual representation admits a common continuously differentiable minimizer $q^\star(\mu)$ for all payoff rows. If $S(\mu) = \nabla_\mu q^\star(\mu)$ is symmetric positive semidefinite and $AS(\mu) = S(\mu)A$ throughout the policy domain, then the distortion vanishes on the zero-sum subspace and the policy-space operator retains the Euclidean strong-monotonicity bound $\mu_{\mathcal{R}} = \beta$.

Indeed, (36) vanishes, leaving the regularization term $\beta[\mathrm{KL}(\pi_1\|\pi_2)+\mathrm{KL}(\pi_2\|\pi_1)]$ in the monotonicity pairing. The common-minimizer and commutation conditions must be verified for the particular risk; symmetry of an ambiguity set or a quadratic divergence alone does not establish them.

When this fails. Entropic payoff risk and CVaR generally do not admit a common minimizing distribution across payoff rows. For entropic risk, the minimizing weights are

$$q_y^\star(\mu)_{y'} = \frac{\mu(y')e^{-\tau P_{y,y'}}}{\sum_z \mu(z)e^{-\tau P_{y,z}}}.$$

The normalization couples the weights, so the Jacobian with respect to μ generally has off-diagonal entries and is not symmetric. Both the weights and the penalty $d_y(\mu) = \tau^{-1}\,\mathrm{KL}(q_y^\star(\mu)\|\mu)$ depend on the payoff row. This penalty must be retained when differentiating the entropic payoff value; entropic risk is not a penalty-free coherent risk.

For CVaR, the minimizing distribution emphasizes the lowest-payoff opponents in each row, with partial weight at the quantile threshold when needed. The lower-tail set depends on the row, and its membership can change with μ. Within a region with fixed tail membership and threshold, the minimizing weights can be chosen affine in μ; their Jacobian is generally nonsymmetric, and threshold changes can produce kinks. Thus the common-map zero-curvature argument does not apply in general.

These examples motivate working with the rowwise Jacobians or bounding the changes in the minimizing distributions directly. Route B below provides the latter bound for coherent risks such as CVaR, once its Lipschitz assumption is verified for the chosen minimizing selections. For entropic risk, the corresponding analysis must also account for the divergence penalty, as in Route A.

Route B: Direct bound via variational representation. For a coherent payoff risk, write $\mathcal{R}[Z] = \inf_{q\in\mathcal{Q}(\mu)} \mathbb{E}_q[Z]$ and select a minimizer $q_y^\star(\mu) \in \mathcal{Q}(\mu)$ for each row y. Then $(P_{\mathcal{R}}\mu)_y = (Pq_y^\star(\mu))_y$. View $q^\star(\mu)$ as the matrix with rows $q_y^\star(\mu)^\top$ and set $\delta\pi := \pi_1 - \pi_2$. The single-player coupling term is

$$(\mathrm{B}) := \big\langle P_{\mathcal{R}}\pi_1 - P_{\mathcal{R}}\pi_2, \delta\pi\big\rangle = \sum_y \delta\pi_y \left(q_y^\star(\pi_1) - q_y^\star(\pi_2)\right)^\top P^\top e_y.$$

Cauchy–Schwarz and $\|P^\top e_y\|_2 \le \|P\|_2$ give

$$|(\mathrm{B})| \le \|P\|_2 \, \|\delta\pi\|_2 \, \|q^\star(\pi_1) - q^\star(\pi_2)\|_F.$$

The spectral norm $\|P\|_2$ need not be at most one merely because the entries of P lie in $[0,1]$. If $q^\star$ is L-Lipschitz in Frobenius norm with respect to the policy ℓ_2 norm, then

$$|(\mathrm{B})| \le L\|P\|_2 \, \|\delta\pi\|_2^2.$$

In the differentiable setting, this also bounds the policy-space eigenvalue: $\bar{\lambda}^{\pi}_{\mathcal{R}} \le L\|P\|_2$. For $\pi_j = \pi_{\theta_j}$, the softmax identity therefore yields

$$\begin{aligned}\big\langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2), \delta\pi \big\rangle &= \beta\big[\mathrm{KL}(\pi_1\|\pi_2) + \mathrm{KL}(\pi_2\|\pi_1)\big] - (\mathrm{B}) \\ &\ge \beta\big[\mathrm{KL}(\pi_1\|\pi_2) + \mathrm{KL}(\pi_2\|\pi_1)\big] - L\|P\|_2 \, \|\delta\pi\|_2^2.\end{aligned}$$

Applying Pinsker's inequality separately to the two KL terms gives $\|\delta\pi\|_2^2 \le \mathrm{KL}(\pi_1\|\pi_2) + \mathrm{KL}(\pi_2\|\pi_1)$. Keeping both terms, as in (32), gives the following bound.

Proposition 5 (Monotonicity via Lipschitz dual map)**.** Suppose the coherent dual representation admits minimizing selections $q^\star : \Delta(\mathcal{Y}) \to \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ that are L-Lipschitz in Frobenius norm with respect to the policy ℓ_2 norm. If $L\|P\|_2 < \beta$, then the self-play operator is strongly monotone in policy space with Euclidean modulus $\mu_{\mathcal{R}} = \beta - L\|P\|_2 > 0$. More precisely,

$$\big\langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2), \pi_1 - \pi_2 \big\rangle \ge \mu_{\mathcal{R}}\big[\mathrm{KL}(\pi_1\|\pi_2) + \mathrm{KL}(\pi_2\|\pi_1)\big] \ge \mu_{\mathcal{R}}\|\pi_1 - \pi_2\|_2^2.$$

On a closed convex policy constraint set where the policy-space operator is Lipschitz, projected extragradient with a sufficiently small fixed step size converges last-iterate linearly to the VI solution, with squared Euclidean error $\mathcal{O}((1-\eta\mu_{\mathcal{R}})^T)$ (Tseng, 1995).

G Statistical Bias Analysis

Before stating the convergence theorems, we characterize the bias and variance of the empirical risk-adjusted preference operator $\widehat{F}_{\mathcal{R},m}$ that drives the algorithms. This section establishes that the abstract bias and variance assumptions—which will be used in the convergence theorems—hold with constants explicitly derived from the structure of the risk functional and the boundedness of the underlying preferences.

G.1 Decomposition and abstract assumptions

Recall the self play operator

$$F_{\mathcal{R}}(\theta) = \beta(\theta - \theta_{\mathrm{ref}}) - \mathcal{P}_{\mathcal{R}}(\pi_\theta),$$

with $P_{\mathcal{R}}(\pi)_y := \mathcal{R}_{Y''\sim\pi}[P(y \succ Y'')]$. Let $(\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi))_y$ denote the sample-based estimator of $(\mathcal{P}_{\mathcal{R}}(\pi))_y$ using m samples $Y_1'', \dots, Y_m'' \overset{\mathrm{iid}}{\sim} \pi$ (the same samples shared across components y). The empirical operator is

$$\widehat{F}_{\mathcal{R},m}(\theta) := \beta(\theta - \theta_{\mathrm{ref}}) - \widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta). \tag{37}$$

Expectations below are over the sampling batch with θ held fixed. For an adaptive query, the same statements hold conditionally on the past when a fresh i.i.d. batch is drawn from π_θ. Define

$$b_m(\theta) := \mathbb{E}\big[\widehat{F}_{\mathcal{R},m}(\theta)\big] - F_{\mathcal{R}}(\theta), \qquad \zeta_m(\theta) := \widehat{F}_{\mathcal{R},m}(\theta) - \mathbb{E}\big[\widehat{F}_{\mathcal{R},m}(\theta)\big], \tag{38}$$

so $b_m(\theta)$ is deterministic given θ and $\mathbb{E}[\zeta_m(\theta) \mid \theta] = 0$. By construction, we therefore have that

$$b_m(\theta) = \mathcal{P}_{\mathcal{R}}(\pi_\theta) - \mathbb{E}[\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta)]$$

and

$$\zeta_m(\theta) = \mathbb{E}[\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta)] - \widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta).$$

Abstract bias and variance assumptions. The convergence theorems in Section H rely on the following uniform bias and variance bounds.

(O1) **Uniform bias bound.** There exists $B_m \geq 0$ (non-increasing in m) such that for all $\theta \in \theta_{\mathrm{ref}} + \mathcal{W}$,

$$\|b_m(\theta)\|_2 \leq B_m.$$

(O2) **Uniform variance bound.** There exists $V_m \geq 0$ (non-increasing in m) such that for all $\theta \in \theta_{\mathrm{ref}} + \mathcal{W}$,

$$\mathbb{E}\,\|\zeta_m(\theta)\|_2^2 \leq V_m.$$

The remainder of this section establishes (O1)-(O2) with explicit constants $B_m, V_m = O(1/m)$ for entropic risk and $B_m = 0$, $V_m = O(1/m)$ for the oracle-threshold CVaR estimator obtained from the Rockafellar-Uryasev (CVaR-RU) representation. The zero-bias CVaR claim is for the estimator that evaluates the RU objective at the population optimizer $\nu^\star$; estimating $\nu^\star$ from the same finite batch is a different oracle and is not covered by the $B_m = 0$ statement below.

G.2 Scalar-to-vector reduction

Each component $(\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta))_y$ is a *scalar* functional of a scalar sample mean. For entropic risk at parameter $\lambda > 0$, we have that

$$(\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta))_y = \widehat{\mathcal{R}}^{\mathrm{ent}}_{\lambda,m}(\mathcal{P}(y \succ \cdot); \pi_\theta) := -\lambda^{-1} \log\Big(\tfrac{1}{m} \sum_{i=1}^{m} e^{-\lambda \mathcal{P}(y \succ Y_i'')}\Big),$$

which has the form $h(\widehat{q}_m^{(y)})$ where $h(q) = -\lambda^{-1} \log q$ and $\widehat{q}_m^{(y)} := m^{-1} \sum_{i=1}^m e^{-\lambda \mathcal{P}(y \succ Y_i'')}$ is the sample mean of the scalar quantity $e^{-\lambda \mathcal{P}(y \succ Y'')}$ for $Y'' \sim \pi_\theta$.

Hence the per-component bias and variance reduce to the scalar bias and variance of a smooth functional of a sample mean. The vector-norm bounds (O1)-(O2) aggregate per-component bounds:

$$\begin{aligned} \|b_m(\theta)\|_2^2 &= \sum_y b_m(\theta)_y^2 \leq |\mathcal{Y}| \max_y b_m(\theta)_y^2, \\ \mathbb{E}\,\|\zeta_m(\theta)\|_2^2 &= \sum_y \mathbb{E}[\zeta_m(\theta)_y^2] = \sum_y \mathrm{Var}(\zeta_m(\theta)_y) \leq |\mathcal{Y}| \max_y \mathrm{Var}(\zeta_m(\theta)_y). \end{aligned} \tag{39}$$

Note that (39) ignores covariances between components. For $\mathbb{E}\,\|\zeta_m\|_2^2$ specifically, only diagonal variances enter (since $\mathbb{E}\,\|\zeta\|_2^2 = \sum_y \mathbb{E}[\zeta_y^2]$ and $\mathbb{E}\,\zeta_y = 0$). The covariance matrix $\mathrm{Cov}(\zeta_m(\theta))$ can have nonzero off-diagonal entries because all components share the same samples Y_i'', but these off-diagonals do not contribute to the ℓ_2 norm.[3]

G.3 The bias lemma

Lemma 6 (Bias and variance of a smooth functional of a sample mean)**.** Let $X \in [a,b]$ be a $[a,b]$-valued random variable with $\mathbb{E}\,X = q$, $\mathrm{Var}(X) = \sigma_X^2 \leq (b-a)^2/4$. Let $\widehat{q}_m := m^{-1} \sum_{i=1}^m X_i$ be the sample mean of m i.i.d. copies of X. Let h be three times continuously differentiable on an open interval containing $[a,b]$, and let $C_h''' := \sup_{u \in [a,b]} |h'''(u)|$. In particular, h is defined at every possible sample mean and its third derivative is bounded on $[a,b]$.

i. **Bias (second-order delta method).** The bias is given by

$$\mathbb{E}[h(\widehat{q}_m)] - h(q) = \frac{h''(q)\,\sigma_X^2}{2m} + O(m^{-3/2}), \tag{40}$$

and in particular

$$|\,\mathbb{E}[h(\widehat{q}_m)] - h(q)| \leq \frac{C_h'''\,(b-a)^2}{8\,m}. \tag{41}$$

[3] The convergence analysis uses the preconditioned noise $\Sigma\zeta_m$ in the Σ^+ metric. Its variance is $\mathbb{E}\,\|\Sigma\zeta_m\|_{\Sigma^+}^2 = \mathrm{tr}(\Sigma \mathrm{Cov}(\zeta_m)) \leq \sigma_{\max}\,\mathbb{E}\,\|\zeta_m\|_2^2$, using $\Sigma\Sigma^+\Sigma = \Sigma$ and $\sigma_{\max} = \|\Sigma\|_2$; see Section H.1. The trace can include off-diagonal covariances, while the spectral upper bound needs only the component variances.

ii. **Variance of plug-in.** The variance of the plug-in estimator is

$$\mathrm{Var}(h(\widehat{q}_m)) = h'(q)^2 \, \frac{\sigma_X^2}{m} + O(m^{-2}). \tag{42}$$

iii. **Linear functionals are unbiased.** If h is affine, then $\mathbb{E}[h(\widehat{q}_m)] = h(q)$ exactly and $\mathrm{Var}(h(\widehat{q}_m)) = h'(q)^2\sigma_X^2/m$ exactly, with no remainder.

Proof. We prove each of the three parts of the lemma separately.

(iii) Linear case. If $h(q) = aq + b$, then $h(\widehat{q}_m) - h(q) = a(\widehat{q}_m - q)$. Taking the expectation, we have

$$\mathbb{E}[h(\widehat{q}_m)] - h(q) = a\,\mathbb{E}[\widehat{q}_m - q] = 0.$$

For the variance, we have that $\mathrm{Var}(h(\widehat{q}_m)) = a^2\mathrm{Var}(\widehat{q}_m) = a^2\sigma_X^2/m$ exactly.

(i) Bias. Write $\delta_m := \widehat{q}_m - q$. Apply Taylor's theorem with Lagrange remainder to h at q to get that

$$h(\widehat{q}_m) = h(q) + h'(q)(\widehat{q}_m - q) + \tfrac{1}{2}h''(q)(\widehat{q}_m - q)^2 + R_m, \tag{43}$$

where $|R_m| \le \frac{1}{6}\|h'''\|_{\infty,[a,b]}|\delta_m|^3$ for every sample realization. Independence and centering give

$$\mathbb{E}\,\delta_m^2 = \frac{\sigma_X^2}{m}, \quad \mathbb{E}\,\delta_m^3 = \frac{\mathbb{E}[(X-q)^3]}{m^2}, \quad \mathbb{E}\,\delta_m^4 = \frac{m\,\mathbb{E}[(X-q)^4] + 3m(m-1)\sigma_X^4}{m^4} = O\left(\frac{1}{m^2}\right).$$

Therefore, applying Cauchy–Schwarz yields the bound $\mathbb{E}\,|\delta_m|^3 \le \sqrt{\mathbb{E}\,\delta_m^2\ \mathbb{E}\,\delta_m^4} = O(m^{-3/2})$.

We now take the expectation of (43). The first-order term is $\mathbb{E}[h'(q)(\widehat{q}_m - q)] = h'(q)\,\mathbb{E}[\widehat{q}_m - q] = 0$. The second-order term is $\mathbb{E}[\frac{1}{2}h''(q)(\widehat{q}_m - q)^2] = \frac{1}{2}h''(q)\mathrm{Var}(\widehat{q}_m) = h''(q)\sigma_X^2/(2m)$. The remainder satisfies $\mathbb{E}\,|R_m| = O(m^{-3/2})$ by the moment bound above.

Combining these bounds, we have that $\mathbb{E}[h(\widehat{q}_m)] - h(q) = h''(q)\sigma_X^2/(2m) + O(m^{-3/2})$, which is (40). For the finite-sample bound, the first-order Taylor remainder satisfies

$$|h(\widehat{q}_m) - h(q) - h'(q)\delta_m| \le \tfrac{1}{2}C_h''\delta_m^2.$$

Taking expectations and using $\mathbb{E}\,\delta_m = 0$ gives $|\,\mathbb{E}[h(\widehat{q}_m)] - h(q)| \le C_h''\sigma_X^2/(2m)$. Popoviciu's inequality $\sigma_X^2 \le (b-a)^2/4$ yields (41) for every $m \ge 1$.

(ii) Variance. By the same Taylor expansion applied to $h(\widehat{q}_m) - \mathbb{E}[h(\widehat{q}_m)]$, we have that

$$h(\widehat{q}_m) - \mathbb{E}[h(\widehat{q}_m)] = h'(q)(\widehat{q}_m - q) + \widetilde{R}_m,$$

where $\widetilde{R}_m = \frac{1}{2}h''(q)(\delta_m^2 - \mathbb{E}\,\delta_m^2) + R_m - \mathbb{E}\,R_m$ is centered and satisfies $\mathbb{E}\,|\widetilde{R}_m|^2 = O(m^{-2})$ by the fourth-moment bound and boundedness of δ_m. Taking the variance gives

$$\mathrm{Var}(h(\widehat{q}_m)) = h'(q)^2\mathrm{Var}(\widehat{q}_m) + \mathbb{E}[\widetilde{R}_m^2] + 2\,\mathbb{E}[h'(q)(\widehat{q}_m - q)\widetilde{R}_m].$$

The first term is $h'(q)^2\sigma_X^2/m$. The second is $O(m^{-2})$. For the cross term, use the signed third moment rather than an absolute third-moment bound:

$$2\,\mathbb{E}[h'(q)\delta_m\widetilde{R}_m] = h'(q)h''(q)\,\mathbb{E}\,\delta_m^3 + 2h'(q)\,\mathbb{E}[\delta_m R_m] = O(m^{-2}),$$

because $\mathbb{E}\,\delta_m^3 = O(m^{-2})$ and $|\,\mathbb{E}[\delta_m R_m]| \le \frac{1}{6}\|h'''\|_{\infty,[a,b]}\,\mathbb{E}\,\delta_m^4 = O(m^{-2})$. Combining these bounds gives us $\mathrm{Var}(h(\widehat{q}_m)) = h'(q)^2\sigma_X^2/m + O(m^{-2})$, which is (42). □

G.4 Examples

Let us now present examples to help concretize ideas.

Entropic payoff risk. The per-component bias and variance of the entropic payoff risk are easily characterized via the following proposition.

Proposition 6 (Per-component bias and variance: entropic payoff risk)**.** Let $\mathcal{R}_\lambda^{\mathrm{ent}}(X) = -\lambda^{-1}\log\mathbb{E}[e^{-\lambda X}]$ be the entropic payoff risk at parameter $\lambda > 0$ for $X \in [0,1]$, with plug-in estimator $\widehat{\mathcal{R}}_{\lambda,m}^{\mathrm{ent}} := -\lambda^{-1}\log(m^{-1}\sum_{i=1}^m e^{-\lambda X_i})$. Here the X_i are i.i.d. copies of X. For every $m \geq 1$, the estimates hold:

$$\left|\mathbb{E}[\widehat{\mathcal{R}}_{\lambda,m}^{\mathrm{ent}}] - \mathcal{R}_\lambda^{\mathrm{ent}}(X)\right| \leq \frac{(1-e^{-\lambda})^2}{8\lambda\, e^{-2\lambda}\, m}, \tag{44}$$

$$\mathrm{Var}(\widehat{\mathcal{R}}_{\lambda,m}^{\mathrm{ent}}) \leq \frac{(1-e^{-\lambda})^2}{4\lambda^2\, e^{-2\lambda}\, m}. \tag{45}$$

Proof. Set $Y := e^{-\lambda X}$. Since $X \in [0,1]$, we have that $Y \in [e^{-\lambda}, 1]$, so $(b-a)$ for Y is $(1-e^{-\lambda})$. Set $h(q) := -\lambda^{-1}\log q$, so $\widehat{\mathcal{R}}_{\lambda,m}^{\mathrm{ent}} = h(\widehat{q}_m)$ where $\widehat{q}_m = m^{-1}\sum_i Y_i$ and $\mathbb{E}\,\widehat{q}_m = q := \mathbb{E}\,Y \in [e^{-\lambda}, 1]$.

Computing derivatives, we have that $h'(q) = -(\lambda q)^{-1}$, $h''(q) = (\lambda q^2)^{-1}$. On the range $q \in [e^{-\lambda}, 1]$, we have that

$$|h'(q)| \leq (\lambda e^{-\lambda})^{-1} = e^{\lambda}/\lambda \quad \text{and} \quad |h''(q)| \leq (\lambda e^{-2\lambda})^{-1} = e^{2\lambda}/\lambda.$$

Apply Lemma 6(i) with $C_h'' = e^{2\lambda}/\lambda$ and $(b-a) = 1-e^{-\lambda}$ to get that

$$\left|\mathbb{E}[\widehat{\mathcal{R}}_{\lambda,m}^{\mathrm{ent}}] - \mathcal{R}_\lambda^{\mathrm{ent}}\right| \leq \frac{e^{2\lambda}}{\lambda}\cdot\frac{(1-e^{-\lambda})^2}{8m} = \frac{(1-e^{-\lambda})^2 e^{2\lambda}}{8\lambda m},$$

which equals the right-hand side of (44).

For the variance, use the derivative bound on the entire sample-mean range to obtain a finite-sample estimate:

$$\mathrm{Var}(h(\widehat{q}_m)) \leq \mathbb{E}\big[(h(\widehat{q}_m) - h(q))^2\big] \leq \frac{e^{2\lambda}}{\lambda^2}\,\mathbb{E}[(\widehat{q}_m - q)^2] = \frac{e^{2\lambda}}{\lambda^2}\frac{\sigma_Y^2}{m} \leq \frac{e^{2\lambda}}{\lambda^2}\frac{(1-e^{-\lambda})^2}{4m},$$

which equals (45). □

Remark 1 (Constants for entropic)**.** For convenience, define

$$C_b^{\mathrm{ent}}(\lambda) := \frac{(1-e^{-\lambda})^2 e^{2\lambda}}{8\lambda}, \qquad C_\zeta^{\mathrm{ent}}(\lambda) := \frac{(1-e^{-\lambda})^2 e^{2\lambda}}{4\lambda^2},$$

so per-component bias is less than or equal to C_b^{ent}/m and per-component variance is less than or equal to C_ζ^{ent}/m for every $m \geq 1$. Both constants blow up as λ grows; this reflects the increasing "concentration on the worst case" of the entropic functional, which is harder to estimate.

CVaR via Rockafellar-Uryasev. The per-component bias and variance of payoff-side CVaR are easily characterized for the oracle-threshold estimator obtained from the well-known Rockafellar-Uryasev transformation Rockafellar & Uryasev (2002).

Proposition 7 (Per-component bias and variance: oracle-threshold CVaR-RU)**.** Let $X \in [0,1]$ and let $X_1, \ldots, X_m$ be i.i.d. copies of X. The lower-tail payoff CVaR at tail mass $q \in (0,1)$ admits the variational representation

$$\mathrm{CVaR}_q(X) = \sup_{\nu\in\mathbb{R}}\Big\{\nu - q^{-1}\,\mathbb{E}[(\nu - X)_+]\Big\}, \tag{46}$$

attained at $\nu^\star = \mathrm{VaR}_q(X)$. Define the empirical estimator at fixed ν as

$$\widehat{C}_m(\nu) := \nu - q^{-1}m^{-1}\sum_{i=1}^m (\nu - X_i)_+.$$

The oracle-threshold estimator is $\widehat{C}_m^\star := \widehat{C}_m(\nu^\star)$.

i. For each fixed ν, $\widehat{C}_m(\nu)$ is unbiased for its population value $\nu - q^{-1}\,\mathbb{E}[(\nu - X)_+]$. At $\nu = \nu^\star$, $\mathbb{E}[\widehat{C}_m^\star] = \mathrm{CVaR}_q(X)$.

ii. The per-component variance is $\mathrm{Var}(\widehat{C}^{\star}_m) = q^{-2}\mathrm{Var}((\nu^\star - X)_+)/m \le 1/(4q^2m)$, since $X \in [0,1]$ implies $(\nu^\star - X)_+ \in [0,1]$.

Proof. For fixed ν, set $Z_i := (\nu - X_i)_+$. Then $\widehat{C}_m(\nu) = h(m^{-1}\sum_i Z_i)$ with the affine function $h(z) := \nu - q^{-1}z$. By Lemma 6(iii), it is unbiased for the population value. At $\nu = \nu^\star$, the population value equals $\mathrm{CVaR}_q(X)$ by (46).

For the variance, $(\nu^\star - X)_+ \in [0,1]$ with variance $\le 1/4$ by Popoviciu. Sample-mean variance is $q^{-2}\mathrm{Var}((\nu^\star - X)_+)/m \le 1/(4q^2m)$. □

Remark 2 (Constants for oracle-threshold CVaR-RU)**.** The constants for the oracle-threshold CVaR-RU estimator are given by

$$C_b^{\mathrm{CVaR-RU}} = 0, \qquad C_\zeta^{\mathrm{CVaR-RU}}(q) = \frac{1}{4q^2}.$$

The bias is exactly zero because $\nu^\star$ is fixed at its population value before the sample average is taken. If instead one plugs in an empirical optimizer $\widehat{\nu}_m$ computed from the same samples, the threshold and empirical average are correlated; that estimator generally has its own finite-sample quantile error and is not the oracle analyzed here. The variance bound grows as $q \to 0$ (estimating the extreme lower tail requires more samples).

G.5 Aggregate bounds: (O1)-(O2) for the operator

Proposition 8 (Operator-level bias and variance)**.** Let $\widehat{F}_{\mathcal{R},m}$ be the empirical operator (37) for the risk-adjusted preference operator with samples $Y_1'', \dots, Y_m'' \sim \pi_\theta$.

(i) **Entropic payoff risk at parameter λ.** Assumption (O1) holds with

$$B_m^{\mathrm{ent}} := \frac{\sqrt{|\mathcal{Y}|}\, C_b^{\mathrm{ent}}(\lambda)}{m}. \tag{47}$$

Assumption (O2) holds with

$$V_m^{\mathrm{ent}} := \frac{|\mathcal{Y}|\, C_\zeta^{\mathrm{ent}}(\lambda)}{m}. \tag{48}$$

(ii) **Payoff CVaR at lower-tail mass q via oracle-threshold Rockafellar-Uryasev.** If each component uses the population threshold $\nu_y^\star(\theta) = \mathrm{VaR}_q(P(y \succ Y''))$ under $Y'' \sim \pi_\theta$, then Assumption (O1) holds with $B_m^{\mathrm{CVaR-RU}} = 0$. Assumption (O2) holds with

$$V_m^{\mathrm{CVaR-RU}} := \frac{|\mathcal{Y}|}{4q^2\, m}. \tag{49}$$

Both sets of bounds are uniform in θ (and hence in π_θ): the constants depend only on the risk parameter and the boundedness of $P(y \succ Y'') \in [0,1]$, not on π_θ.

Proof. By the decomposition (37)-(38), we have that $\|b_m(\theta)\|_2^2 = \sum_y \left|\mathbb{E}[\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta)_y] - \mathcal{P}_{\mathcal{R}}(\pi_\theta)_y\right|^2$ and $\mathbb{E}\,\|\zeta_m(\theta)\|_2^2 = \sum_y \mathrm{Var}(\widehat{\mathcal{P}}_{\mathcal{R},m}(\pi_\theta)_y)$. The deterministic regularization term contributes neither bias nor variance. For entropic, apply Proposition 6 per-component to conclude that each per-component bias is $\le C_b^{\mathrm{ent}}/m$ and each per-component variance is $\le C_\zeta^{\mathrm{ent}}/m$. Summing over $|\mathcal{Y}|$ components (worst case) and taking square root for the bias gives (47). Summing variances gives (48).

For oracle-threshold CVaR-RU, apply Proposition 7: per-component bias is 0 at $\nu_y^\star(\theta)$, so $\|b_m\|_2 = 0$ giving $B_m^{\mathrm{CVaR-RU}} = 0$. Per-component variance is less than or equal to $1/(4q^2m)$, and summing gives (49).

Uniformity in π_θ yields the fact that constants $C_b^{\mathrm{ent}}, C_\zeta^{\mathrm{ent}}$ depend only on λ and on the range $X \in [0,1]$ (which is independent of π_θ), not on π_θ itself. Same for CVaR. □

Looseness of the bounds. The bounds in Proposition 8 are loose in two ways:

1. **Worst-case per-component aggregation.** We took the per-component variance bound and multiplied by $|\mathcal{Y}|$ (and the bias bound by $\sqrt{|\mathcal{Y}|}$), ignoring the possibility that not every component achieves the maximum. In practice, for many distributions π_θ, the per-component variances are much smaller than the worst case.
2. **Off-diagonal covariance ignored.** The bound on $\mathbb{E}\,\|\zeta_m\|_2^2$ uses only diagonal variances, which is exact for the ℓ_2 norm. However, for the preconditioned noise in the Σ^+ norm used in the convergence proofs, we apply the loose bound $\mathbb{E}\,\|\Sigma\zeta_m\|_{\Sigma^+}^2 \le \sigma_{\max}\,\mathbb{E}\,\|\zeta_m\|_2^2 \le \sigma_{\max} V_m$ (see Section H.1), which incurs an additional factor of $\sigma_{\max}$. A tighter bound via direct computation of $\mathbb{E}\,\|\Sigma\zeta_m\|_{\Sigma^+}^2 = \mathrm{tr}(\Sigma\mathrm{Cov}(\zeta_m))$ would exploit the full covariance structure of the empirical risk estimator and the spectrum of Σ. The improvement depends on the covariance's alignment with the eigenspaces of Σ. We do not pursue this refinement here.

The constants given provide explicit $1/m$ bounds for the bias and variance inputs to the convergence analysis.

H Stochastic Convergence Analysis with Bias

We prove convergence of stochastic projected Mirror-Prox (extragradient) on the IPO gradient flow with a biased stochastic oracle, in the Σ^+-weighted Euclidean geometry on $\mathcal{W}$. The proof structure follows Juditsky et al. (2011) with several core modifications that arise from handling the risk case. We also write the theorems and lemmas in our Euclidean setting; all proofs are self-contained. We choose to provide proofs for extra-gradient as opposed to mirror descent even in the monotone case because there are benefits to running extragradient: though it requires an extra step, it will handle corrections for cycling, and potentially cover the case in practice where the estimated game operator fails to be strongly monotone.

H.1 Preliminaries

Throughout this section, fix the response-pair preconditioner $\Sigma = \Sigma(\rho)$ and write $\mathcal{W} := \mathrm{range}(\Sigma)$. Let $\sigma_{\min}$ and $\sigma_{\max}$ denote its smallest and largest positive eigenvalues, respectively.

Let $F_{\mathcal{R}} : \theta_{\mathrm{ref}} + \mathcal{W} \to \mathbb{R}^d$ be the risk-adjusted game operator on the affine slice $\theta_{\mathrm{ref}} + \mathcal{W}$. As is typical, we require some regularity assumptions. Not all are assumed in all results or throughout; instead we call on different elements of the list where needed.

Assumption 2. The following regularity properties hold:

a. **(P-mono)** *Monotonicity.* For all $\theta_1, \theta_2 \in \theta_{\mathrm{ref}} + \mathcal{W}$, the lower bound holds:
$$\langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2), \theta_1 - \theta_2\rangle \ge 0.$$

b. **(P-strong)** *Strong monotonicity (when assumed).* The map $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone on $\theta_{\mathrm{ref}} + \mathcal{W}$:
$$\langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2), \theta_1 - \theta_2\rangle \ge \mu_{\mathcal{R}}\|\theta_1 - \theta_2\|_2^2,$$
with $\mu_{\mathcal{R}} > 0$. When this holds, **(P-mono)** is implied.

c. **(P-lip)** *Lipschitz Continuity.* The map $F_{\mathcal{R}}$ is $\ell_{\mathcal{R}}$-Lipschitz continuous on $\theta_{\mathrm{ref}} + \mathcal{W}$:
$$\|F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2)\|_2 \le \ell_{\mathcal{R}}\|\theta_1 - \theta_2\|_2.$$

In the symmetric self-play setting of Theorem 6, **(P-strong)** holds with $\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}} > 0$ when $\bar{\lambda}_{\mathcal{R}} < \beta$, where $\bar{\lambda}_{\mathcal{R}}$ is the parameter-space risk-distortion eigenvalue defined in (34).

With the opponent policy held fixed during differentiation, the IPO gradient is $\nabla\mathcal{L}_{\mathrm{IPO}}(\theta) = (2/\beta)\Sigma F_{\mathcal{R}}(\theta)$. Let $G(\theta) := \Sigma F_{\mathcal{R}}(\theta)$ be the operator we analyze, with the factor $2/\beta$ absorbed into the step size.

On $\mathcal{W}$, define $\langle u, v\rangle_{\Sigma^+} := u^\top\Sigma^+ v$ and $\|u\|_{\Sigma^+}^2 := u^\top\Sigma^+ u$, where Σ^+ is the Moore-Penrose pseudoinverse, positive definite on $\mathcal{W}$ with eigenvalues $1/\sigma_{\max}, \ldots, 1/\sigma_{\min}$. Norm equivalence on $\mathcal{W}$ implies
$$\sigma_{\min}\|v\|_{\Sigma^+}^2 \le \|v\|_2^2 \le \sigma_{\max}\|v\|_{\Sigma^+}^2. \tag{50}$$

Risk-adjusted Psuedo-gradient Operator. Let $F_{\mathcal{R}} : \theta_{\text{ref}} + \mathcal{W} \to \mathbb{R}^d$ be the risk-adjusted game operator, assumed monotone and $\ell_{\mathcal{R}}$-Lipschitz on $\theta_{\text{ref}} + \mathcal{W}$: for all $\theta_1, \theta_2 \in \theta_{\text{ref}} + \mathcal{W}$,

$$\langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2), \theta_1 - \theta_2 \rangle \geq 0, \quad \text{and} \quad \|F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2)\|_2 \leq \ell_{\mathcal{R}} \|\theta_1 - \theta_2\|_2.$$

Lemma 7 (Lipschitz continuity of $F_{\mathcal{R}}$)**.** The map $F_{\mathcal{R}}$ is Lipschitz continuous on $\mathbb{R}^{|\mathcal{Y}|}$ with constant $\ell_{\mathcal{R}} \leq \beta + L_{\mathcal{R}}$, where $L_{\mathcal{R}}$ is the Lipschitz constant of $\mu \mapsto P_{\mathcal{R}}(\mu)$ as a map between $\Delta(\mathcal{Y})$ and $\mathbb{R}^{|\mathcal{Y}|}$.

Proof. We derive directly from the definition of $F_{\mathcal{R}}$, the sequence of bounds

$$\begin{aligned} \|F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2)\|_2 &\leq \beta \|\theta_1 - \theta_2\|_2 + \|P_{\mathcal{R}}(\pi_1) - P_{\mathcal{R}}(\pi_2)\|_2 \\ &\leq \beta \|\theta_1 - \theta_2\|_2 + L_{\mathcal{R}} \|\pi_1 - \pi_2\|_2 \\ &\leq (\beta + L_{\mathcal{R}}) \|\theta_1 - \theta_2\|_2, \end{aligned}$$

using that $\theta \mapsto \pi_\theta$ (softmax) is 1-Lipschitz from ℓ_2 to ℓ_2. □

H.2 Properties of $G = \Sigma F_{\mathcal{R}}$ in the Σ^+ metric

The next two lemmas establish that G inherits monotonicity and Lipschitz properties from $F_{\mathcal{R}}$ when measured in the Σ^+ inner product.

Lemma 8 (Preconditioning preserves monotonicity in Σ^+ metric)**.** For any $\theta_1, \theta_2 \in \theta_{\text{ref}} + \mathcal{W}$, the following two relationships hold: (i) $\langle G(\theta_1) - G(\theta_2), \theta_1 - \theta_2 \rangle_{\Sigma^+} = \langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2), \theta_1 - \theta_2 \rangle_2$, and (ii) $\|G(\theta_1) - G(\theta_2)\|_{\Sigma^+}^2 \leq \sigma_{\max} \|F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2)\|_2^2$. Consequently, we have that

a. Under **(P-mono)**, the map G is monotone in the Σ^+ inner product.

b. Under **(P-strong)**, the map G is $\mu_{\mathcal{R}} \sigma_{\min}$-strongly monotone in the Σ^+ inner product: $\langle G(\theta_1) - G(\theta_2), \theta_1 - \theta_2 \rangle_{\Sigma^+} \geq \mu_{\mathcal{R}} \sigma_{\min} \|\theta_1 - \theta_2\|_{\Sigma^+}^2$.

Proof. Let $\Delta := \theta_1 - \theta_2 \in \mathcal{W}$ and $\xi := F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2)$. Note $G(\theta_i) = \Sigma F_{\mathcal{R}}(\theta_i)$, so $G(\theta_1) - G(\theta_2) = \Sigma \xi$.

Proof for claim (i). We directly compute

$$\langle G(\theta_1) - G(\theta_2), \Delta \rangle_{\Sigma^+} = \langle \Sigma \xi, \Delta \rangle_{\Sigma^+} = \Delta^\top \Sigma^+ \Sigma \xi = \Delta^\top \Pi_{\mathcal{W}} \xi,$$

where $\Pi_{\mathcal{W}} = \Sigma^+ \Sigma$ is the orthogonal projector onto $\mathcal{W}$ (in the standard inner product), using the Moore-Penrose property $\Sigma^+ \Sigma = \Pi_{\text{range}(\Sigma)} = \Pi_{\mathcal{W}}$. Since $\Delta \in \mathcal{W}$, $\Delta^\top \Pi_{\mathcal{W}} \xi = \Delta^\top \xi = \langle \xi, \Delta \rangle_2$.

Proof for claim (ii). Analogously, we compute

$$\|G(\theta_1) - G(\theta_2)\|_{\Sigma^+}^2 = (\Sigma \xi)^\top \Sigma^+ (\Sigma \xi) = \xi^\top \Sigma \Sigma^+ \Sigma \xi = \xi^\top \Sigma \xi,$$

using $\Sigma \Sigma^+ \Sigma = \Sigma$ (i.e., Moore-Penrose). Then $\xi^\top \Sigma \xi \leq \sigma_{\max} \|\xi\|_2^2$ by the eigenvalue bound on Σ.

Proof for claim a. By (i), under **(P-mono)**, we deduce that $\langle G(\theta_1) - G(\theta_2), \Delta \rangle_{\Sigma^+} = \langle \xi, \Delta \rangle_2 \geq 0$.

Proof for claim b. By (i), under **(P-strong)**, we further deduce that $\langle G(\theta_1) - G(\theta_2), \Delta \rangle_{\Sigma^+} = \langle \xi, \Delta \rangle_2 \geq \mu_{\mathcal{R}} \|\Delta\|_2^2 \geq \mu_{\mathcal{R}} \sigma_{\min} \|\Delta\|_{\Sigma^+}^2$, the last inequality from (50). □

Lemma 9 (Lipschitz constant of G in Σ^+ metric)**.** Under (P-lip), G is L_G-Lipschitz in the Σ^+ inner product on $\mathcal{W}$ with $L_G := \sigma_{\max} \ell_{\mathcal{R}}$. That is, for all $\theta_1, \theta_2 \in \theta_{\text{ref}} + \mathcal{W}$, the upper bound holds: $\|G(\theta_1) - G(\theta_2)\|_{\Sigma^+} \leq L_G \|\theta_1 - \theta_2\|_{\Sigma^+}$.

Proof. By Lemma 8 part (ii), we have that

$$\|G(\theta_1) - G(\theta_2)\|_{\Sigma^+}^2 \leq \sigma_{\max} \|F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2)\|_2^2.$$

By **(P-lip)**, we have that

$$\|F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2)\|_2^2 \leq \ell_{\mathcal{R}}^2 \|\theta_1 - \theta_2\|_2^2.$$

By (50), we also have that

$$\|\theta_1 - \theta_2\|_2^2 \leq \sigma_{\max} \|\theta_1 - \theta_2\|_{\Sigma^+}^2.$$

Combining these three bounds, we deduce that

$$\|G(\theta_1) - G(\theta_2)\|_{\Sigma^+}^2 \leq \sigma_{\max}^2 \ell_{\mathcal{R}}^2 \|\theta_1 - \theta_2\|_{\Sigma^+}^2.$$

Taking the square root gives us the conclusion—indeed, we see that $\|G(\theta_1) - G(\theta_2)\|_{\Sigma^+} \leq \sigma_{\max} \ell_{\mathcal{R}} \|\theta_1 - \theta_2\|_{\Sigma^+} = L_G \|\theta_1 - \theta_2\|_{\Sigma^+}$. □

H.3 Distance-generating function and projection

Set $D := \{\theta \in \theta_{\mathrm{ref}} + \mathcal{W} : \|\theta - \theta_{\mathrm{ref}}\|_{\Sigma^+} \le R\}$ for $R > 0$ to be determined by the boundedness lemma. Following Juditsky et al. (2011), we use the distance generating function $\omega(\theta) = \frac{1}{2}\|\theta - \theta_{\mathrm{ref}}\|_{\Sigma^+}^2$ on the slice $\theta_{\mathrm{ref}} + \mathcal{W}$. This is strongly convex with modulus one with respect to $\|\cdot\|_{\Sigma^+}$. The associated prox-function (Bregman divergence), using the Euclidean gradient $\nabla\omega(z) = \Sigma^+(z - \theta_{\mathrm{ref}})$, is

$$V(z,u) := \omega(u) - \omega(z) - \langle \nabla\omega(z), u - z\rangle_2 = \tfrac{1}{2}\|z-u\|_{\Sigma^+}^2.$$

The prox-mapping is

$$P_z(\xi) := \arg\min_{u\in D}\{V(z,u) + \langle \xi, u\rangle_{\Sigma^+}\} = \Pi_D(z - \xi),$$

where Π_D is the Euclidean projection onto D in the Σ^+ metric.

The prox-center of the projection set D is $z_c = \theta_{\mathrm{ref}}$. We have that

$$\Theta(z_c) := \max_{u\in D} V(z_c, u) = \tfrac{1}{2}R^2, \quad \text{and} \quad \Omega := \sqrt{2\Theta(z_c)} = R. \tag{51}$$

Thus $D \subseteq \{\theta : \|\theta - z_c\|_{\Sigma^+} \le \Omega\}$, i.e., the Σ^+-radius of D from z_c is exactly $\Omega = R$. The Σ^+-diameter of D is at most $2\Omega = 2R$.

Lemma 10 (Boundedness of equilibrium in Σ^+)**.** The risk-adjusted equilibrium $\theta^\star \in \theta_{\mathrm{ref}} + \mathcal{W}$ satisfies

$$\|\theta^\star - \theta_{\mathrm{ref}}\|_{\Sigma^+} \le \sqrt{|\mathcal{Y}|}/(\beta\sqrt{\sigma_{\min}}).$$

Proof. The equilibrium condition $F_{\mathcal{R}}(\theta^\star) \in \ker\Sigma$ on the slice $\theta_{\mathrm{ref}} + \mathcal{W}$ means $\beta(\theta^\star - \theta_{\mathrm{ref}}) - P_{\mathcal{R}}(\pi^\star) \in \ker\Sigma$, so the $\mathcal{W}$ component of $\theta^\star - \theta_{\mathrm{ref}}$ equals the $\mathcal{W}$ component of $\beta^{-1}P_{\mathcal{R}}(\pi^\star)$—i.e.,

$$\theta^\star - \theta_{\mathrm{ref}} = \beta^{-1}\Pi_{\mathcal{W}}P_{\mathcal{R}}(\pi^\star).$$

Components of $P_{\mathcal{R}}(\pi)$ lie in $[0,1]$ since $P(y \succ Y'') \in [0,1]$ and the risk functional preserves this range. Hence $\|P_{\mathcal{R}}(\pi^\star)\|_2 \le \sqrt{|\mathcal{Y}|}$ and $\|\Pi_{\mathcal{W}}P_{\mathcal{R}}(\pi^\star)\|_2 \le \sqrt{|\mathcal{Y}|}$ (projection non-expansive). By the lower bound in (50), for $v \in \mathcal{W}$, we have that $\|v\|_{\Sigma^+}^2 \le \|v\|_2^2/\sigma_{\min}$, so $\|\theta^\star - \theta_{\mathrm{ref}}\|_{\Sigma^+} \le \beta^{-1}\sqrt{|\mathcal{Y}|/\sigma_{\min}}$ as claimed. □

Set the projection radius to be

$$R := \frac{\sqrt{|\mathcal{Y}|}}{\beta\sqrt{\sigma_{\min}}} + R_0, \qquad R_0 := \|\theta^{(0)} - \theta_{\mathrm{ref}}\|_{\Sigma^+},$$

so that both $\theta^\star \in D$ (by Lemma 10) and $\theta^{(0)} \in D$. Hence $\Omega = R$ from (51).

H.4 Stochastic oracle, decomposition, and filtration

At a fixed input $\theta \in D$ with sample budget m, the oracle returns $\widehat{F}_m(\theta) \in \mathbb{R}^d$ satisfying

$$\big\|\mathbb{E}[\widehat{F}_m(\theta)] - F_{\mathcal{R}}(\theta)\big\|_2 \le B_m \tag{O1}$$

$$\mathbb{E}\big\|\widehat{F}_m(\theta) - \mathbb{E}[\widehat{F}_m(\theta)]\big\|_2^2 \le V_m \tag{O2}$$

with $B_m, V_m \ge 0$ non-increasing in m (typically $\mathcal{O}(1/m)$; explicit constants from Section G). Decompose $\widehat{F}_m(\theta) = F_{\mathcal{R}}(\theta) + b_m(\theta) + \zeta_m(\theta)$ with $b_m(\theta) := \mathbb{E}[\widehat{F}_m(\theta)] - F_{\mathcal{R}}(\theta)$ deterministic given θ and ζ_m mean-zero given θ.

The preconditioned oracle $\widehat{G}_m := \Sigma\widehat{F}_m$ satisfies (in the Σ^+ norm) the following:

$$\big\|\mathbb{E}[\widehat{G}_m(\theta)] - G(\theta)\big\|_{\Sigma^+} \le \widetilde{B}_m, \quad \widetilde{B}_m := \sqrt{\sigma_{\max}}B_m, \tag{52}$$

$$\mathbb{E}\big\|\widehat{G}_m(\theta) - \mathbb{E}[\widehat{G}_m(\theta)]\big\|_{\Sigma^+}^2 \le \widetilde{V}_m, \quad \widetilde{V}_m := \sigma_{\max}V_m. \tag{53}$$

To see these constructions observe that $\|\Sigma b_m\|_{\Sigma^+}^2 = b_m^\top\Sigma\Sigma^+\Sigma b_m = b_m^\top\Sigma b_m \le \sigma_{\max}\|b_m\|_2^2$ using $\Sigma\Sigma^+\Sigma = \Sigma$. Same for ζ_m.

For adaptive queries, each oracle call uses fresh sampling, as in stochastic Mirror-Prox (Juditsky et al., 2011): conditionally on the history before that call, draw m i.i.d. samples from the queried policy. Thus (O1)–(O2) also hold conditionally on that history. Write $\zeta_{2\tau-1}$ and $\zeta_{2\tau}$ for the centered errors ζ_m from the extrapolation and correction calls, respectively. Each error has conditional mean zero and conditional second moment at most V_m; its distribution can depend on previous iterates.

Algorithm. Stochastic projected Mirror-Prox proceeds as follows. First, set $r_0 := \theta^{(0)}$. Then, for $\tau = 1, \dots, t$, update

$$w_\tau := P_{r_{\tau-1}}\big(\gamma_\tau \widehat{G}_m(r_{\tau-1})\big) = \Pi_D\big(r_{\tau-1} - \gamma_\tau \widehat{G}_m(r_{\tau-1})\big), \tag{54}$$

$$r_\tau := P_{r_{\tau-1}}\big(\gamma_\tau \widehat{G}_m(w_\tau)\big) = \Pi_D\big(r_{\tau-1} - \gamma_\tau \widehat{G}_m(w_\tau)\big), \tag{55}$$

with a fresh batch at each of the two oracle calls. The corresponding centered errors in $\widehat{F}_m$ are $\zeta_{2\tau-1}$ at $r_{\tau-1}$ and $\zeta_{2\tau}$ at w_τ. The output is

$$\widehat{z}_t := \left(\sum_{\tau=1}^{t} \gamma_\tau\right)^{-1} \sum_{\tau=1}^{t} \gamma_\tau w_\tau.$$

The variational-inequality (VI) error of $z \in D$ is

$$\mathrm{Gap}_{\mathrm{vi}}(z) := \sup_{u \in D} \langle F_{\mathcal{R}}(u), z - u \rangle.$$

Note that $\mathrm{Gap}_{\mathrm{vi}}$ is defined using $F_{\mathcal{R}}$—not G—and the standard Euclidean inner product—not Σ^+. This is because $F_{\mathcal{R}}$ is the underlying VI operator; the preconditioner Σ enters only through the algorithm's geometry.

H.5 Filtration and noise quantities

Let $\mathcal{G}_0 := \sigma(r_0)$ and, for $\tau \geq 1$, define

$$\mathcal{F}_\tau := \sigma(r_0, \zeta_1, \dots, \zeta_{2\tau-1}), \qquad \mathcal{G}_\tau := \sigma(r_0, \zeta_1, \dots, \zeta_{2\tau}),$$

nested as $\mathcal{F}_1 \subset \mathcal{G}_1 \subset \mathcal{F}_2 \subset \mathcal{G}_2 \subset \cdots$. By construction, $r_{\tau-1}$ is $\mathcal{G}_{\tau-1}$-measurable, w_τ is $\mathcal{F}_\tau$-measurable (uses $\zeta_{2\tau-1}$), and r_τ is $\mathcal{G}_\tau$-measurable (uses $\zeta_{2\tau}$). Define the per-step noise and bias quantities (in the Σ^+ norm):

$$\begin{aligned} \Delta_\tau &:= G(w_\tau) - \widehat{G}_m^{(\tau,2)}, \\ \widehat{G}_m^{(\tau,2)} &:= \Sigma \widehat{F}_m^{(\tau,2)}(w_\tau) \text{ which uses } \zeta_{2\tau} \\ \varepsilon_z &:= \|\widehat{G}_m(z) - G(z)\|_{\Sigma^+} \text{ for the relevant call at } z. \end{aligned}$$

Specifically, $\varepsilon_{r_{\tau-1}}$ uses $\zeta_{2\tau-1}$ (call at $r_{\tau-1}$ in the extrapolation step), and ε_{w_τ} uses $\zeta_{2\tau}$ (call at w_τ in the correction step). By (52) and (53) and the bias-variance decomposition, we have that

$$\|\,\mathbb{E}[\Delta_\tau \mid \mathcal{F}_\tau]\|_{\Sigma^+} \leq \widetilde{B}_m, \tag{56}$$

$$\mathbb{E}[\|\Delta_\tau\|_{\Sigma^+}^2 \mid \mathcal{F}_\tau] \leq \widetilde{B}_m^2 + \widetilde{V}_m, \tag{57}$$

$$\mathbb{E}[\varepsilon_{r_{\tau-1}}^2 \mid \mathcal{G}_{\tau-1}] \leq \widetilde{B}_m^2 + \widetilde{V}_m, \tag{58}$$

$$\mathbb{E}[\varepsilon_{w_\tau}^2 \mid \mathcal{F}_\tau] \leq \widetilde{B}_m^2 + \widetilde{V}_m. \tag{59}$$

To see that (56) and (57) hold, observe that $\Delta_\tau = G(w_\tau) - \widehat{G}_m^{(\tau,2)} = -\Sigma b_m(w_\tau) - \Sigma \zeta_m^{(\tau,2)}$. Conditional on $\mathcal{F}_\tau$, w_τ is fixed, $b_m(w_\tau)$ is deterministic, and $\mathbb{E}[\zeta_m^{(\tau,2)} \mid \mathcal{F}_\tau] = 0$ by fresh sampling. So $\mathbb{E}[\Delta_\tau \mid \mathcal{F}_\tau] = -\Sigma b_m(w_\tau)$ with $\|\Sigma b_m\|_{\Sigma^+} \leq \widetilde{B}_m$. For the second moment, we have $\|\Delta_\tau\|_{\Sigma^+}^2 = \|\Sigma b_m\|_{\Sigma^+}^2 + 2\langle \Sigma b_m, \Sigma \zeta_m^{(\tau,2)} \rangle_{\Sigma^+} + \|\Sigma \zeta_m^{(\tau,2)}\|_{\Sigma^+}^2$. Conditional on $\mathcal{F}_\tau$, the cross term has mean zero, giving $\mathbb{E}[\|\Delta_\tau\|_{\Sigma^+}^2 \mid \mathcal{F}_\tau] \leq \widetilde{B}_m^2 + \widetilde{V}_m$.

H.6 Three core technical lemmas

There are three core technical results required for the typical proof for biased stochastic mirror prox which we translate to our setting in the Σ^+ Euclidean setting.

The first technical lemma is a descent lemma for the prox-mapping; it is Juditsky et al. (2011) Lemma 3 converted to our setting.

Lemma 11 (Prox-mapping descent)**.** For any $z \in D$ and $\xi \in \mathbb{R}^d$, let $w := P_z(\xi) = \Pi_D(z - \xi)$. Then for all $u \in D$, the following upper bounds hold:

$$V(w,u) \le V(z,u) + \langle \xi, u - w\rangle_{\Sigma^+} - V(z,w), \tag{60}$$

$$V(w,u) \le V(z,u) + \langle \xi, u - z\rangle_{\Sigma^+} + \tfrac{1}{2}\|\xi\|_{\Sigma^+}^2. \tag{61}$$

Proof. By optimality of $w = \arg\min_{v\in D}\{\frac{1}{2}\|v - (z-\xi)\|_{\Sigma^+}^2\}$, we have that, for all $u \in D$,

$$\langle w - (z - \xi), u - w\rangle_{\Sigma^+} \ge 0, \tag{62}$$

i.e., $\langle w - z, u - w\rangle_{\Sigma^+} + \langle \xi, u - w\rangle_{\Sigma^+} \ge 0$. Compute

$$\begin{aligned}
V(w,u) - V(z,u) &= \tfrac{1}{2}\|w-u\|_{\Sigma^+}^2 - \tfrac{1}{2}\|z-u\|_{\Sigma^+}^2 \\
&= \tfrac{1}{2}\|(w-z)+(z-u)\|_{\Sigma^+}^2 - \tfrac{1}{2}\|z-u\|_{\Sigma^+}^2 \\
&= \tfrac{1}{2}\|w-z\|_{\Sigma^+}^2 + \langle w-z, z-u\rangle_{\Sigma^+} \\
&= V(z,w) + \langle w-z, z-u\rangle_{\Sigma^+} \\
&= V(z,w) - \langle w-z, u-z\rangle_{\Sigma^+} \\
&= V(z,w) - \langle w-z, u-w\rangle_{\Sigma^+} - \langle w-z, w-z\rangle_{\Sigma^+} \\
&= V(z,w) - \langle w-z, u-w\rangle_{\Sigma^+} - 2V(z,w) \\
&= -V(z,w) - \langle w-z, u-w\rangle_{\Sigma^+}.
\end{aligned}$$

By (62), we have $\langle w-z, u-w\rangle_{\Sigma^+} \ge -\langle \xi, u-w\rangle_{\Sigma^+}$, so that $-\langle w-z, u-w\rangle_{\Sigma^+} \le \langle \xi, u-w\rangle_{\Sigma^+}$. Combining the decomposition of V with this bound we get

$$V(w,u) - V(z,u) \le -V(z,w) + \langle \xi, u-w\rangle_{\Sigma^+},$$

which is (60). For (61), write $\langle \xi, u-w\rangle_{\Sigma^+} = \langle \xi, u-z\rangle_{\Sigma^+} + \langle \xi, z-w\rangle_{\Sigma^+}$. Apply Young's inequality $\langle \xi, z-w\rangle_{\Sigma^+} \le \frac{1}{2}\|\xi\|_{\Sigma^+}^2 + \frac{1}{2}\|z-w\|_{\Sigma^+}^2 = \frac{1}{2}\|\xi\|_{\Sigma^+}^2 + V(z,w)$. Substitute this into (60) to get that

$$V(w,u) \le V(z,u) + \langle \xi, u-z\rangle_{\Sigma^+} + \tfrac{1}{2}\|\xi\|_{\Sigma^+}^2 + V(z,w) - V(z,w).$$

This gives (61). □

The next technical lemma is the classical two-step identity (e.g., see Juditsky et al. (2011) Lemma 4), again adapted to our setting.

Lemma 12 (Two-step descent identity)**.** For $z \in D$, $\zeta, \eta \in \mathbb{R}^d$, let $w := P_z(\zeta) = \Pi_D(z-\zeta)$ and $r_+ := P_z(\eta) = \Pi_D(z-\eta)$. Then for all $u \in D$, the following upper bound holds:

$$\|w - r_+\|_{\Sigma^+} \le \|\zeta - \eta\|_{\Sigma^+}, \tag{63}$$

$$V(r_+,u) - V(z,u) \le \langle \eta, u-w\rangle_{\Sigma^+} + \tfrac{1}{2}\|\zeta-\eta\|_{\Sigma^+}^2 - \tfrac{1}{2}\|w-z\|_{\Sigma^+}^2. \tag{64}$$

Proof. To see (63), the optimality conditions for w and r_+, evaluated at r_+ and w, respectively, give

$$\langle w - z + \zeta, r_+ - w\rangle_{\Sigma^+} \ge 0,$$
$$\langle r_+ - z + \eta, w - r_+\rangle_{\Sigma^+} \ge 0.$$

Adding these two conditions, we get that

$$\langle (w-z+\zeta) - (r_+ - z + \eta), r_+ - w\rangle_{\Sigma^+} \ge 0 \Leftrightarrow \langle (w - r_+) + (\zeta - \eta), r_+ - w\rangle_{\Sigma^+} \ge 0.$$

That is, $-\|w-r_+\|_{\Sigma^+}^2 + \langle \zeta - \eta, r_+ - w\rangle_{\Sigma^+} \ge 0$, so that

$$\|w-r_+\|_{\Sigma^+}^2 \le \langle \zeta - \eta, r_+ - w\rangle_{\Sigma^+} \le \|\zeta-\eta\|_{\Sigma^+}\|w-r_+\|_{\Sigma^+}.$$

Dividing by $\|w-r_+\|_{\Sigma^+}$ (where we assume its nonzero, otherwise the argument is trivial), we get that $\|w-r_+\|_{\Sigma^+} \le \|\zeta-\eta\|_{\Sigma^+}$, proving (63).

Now for (64), apply (60) of Lemma 11 to $w = P_z(\zeta)$ at $u = r_+$ to get that

$$V(w,r_+) \le V(z,r_+) + \langle \zeta, r_+ - w\rangle_{\Sigma^+} - V(z,w).$$

Rearrange the above inequality to get

$$V(z,r_+) \ge V(w,r_+) + V(z,w) + \langle \zeta, w - r_+\rangle_{\Sigma^+}. \tag{65}$$

Apply (60) to $r_+ = P_z(\eta)$ at any $u \in D$ to get that

$$\begin{aligned} V(r_+, u) &\le V(z, u) + \langle \eta, u - r_+ \rangle_{\Sigma^+} - V(z, r_+) \\ &= V(z, u) + \langle \eta, u - w \rangle_{\Sigma^+} + \langle \eta, w - r_+ \rangle_{\Sigma^+} - V(z, r_+) \\ &\le V(z, u) + \langle \eta, u - w \rangle_{\Sigma^+} + \langle \eta, w - r_+ \rangle_{\Sigma^+} - V(w, r_+) - V(z, w) - \langle \zeta, w - r_+ \rangle_{\Sigma^+} \\ &= V(z, u) + \langle \eta, u - w \rangle_{\Sigma^+} + \langle \eta - \zeta, w - r_+ \rangle_{\Sigma^+} - V(w, r_+) - V(z, w), \end{aligned}$$

where in the last inequality we used (65). Apply Young's

$$\langle \eta - \zeta, w - r_+ \rangle_{\Sigma^+} \le \tfrac{1}{2}\|\eta - \zeta\|_{\Sigma^+}^2 + \tfrac{1}{2}\|w - r_+\|_{\Sigma^+}^2 = \tfrac{1}{2}\|\eta - \zeta\|_{\Sigma^+}^2 + V(w, r_+).$$

Now substitute in to get that

$$V(r_+, u) \le V(z, u) + \langle \eta, u - w \rangle_{\Sigma^+} + \tfrac{1}{2}\|\zeta - \eta\|_{\Sigma^+}^2 - V(z, w),$$

which is (64) (using $V(z, w) = \frac{1}{2}\|w - z\|_{\Sigma^+}^2$). □

Finally we prove a corollary on the ghost iterates, namely, an adaptation of (Juditsky et al., 2011, Corollary 2).

Lemma 13 (Ghost iterate corollary)**.** Let $\xi_1, \xi_2, \dots$ be a sequence of vectors in $\mathbb{R}^d$. Define the sequence $\{y_\tau\}_{\tau \ge 0}$ by

$$y_\tau := P_{y_{\tau-1}}(\xi_\tau) = \Pi_D(y_{\tau-1} - \xi_\tau), \qquad y_0 := r_0. \tag{66}$$

Then for every $u \in D$, the upper bound holds:

$$\sum_{\tau=1}^{t} \langle \xi_\tau, y_{\tau-1} - u \rangle_{\Sigma^+} \le V(y_0, u) + \frac{1}{2}\sum_{\tau=1}^{t} \|\xi_\tau\|_{\Sigma^+}^2. \tag{67}$$

Proof. Apply (61) of Lemma 11 with $z = y_{\tau-1}$, $\xi = \xi_\tau$, $w = y_\tau$, to get that

$$V(y_\tau, u) \le V(y_{\tau-1}, u) + \langle \xi_\tau, u - y_{\tau-1} \rangle_{\Sigma^+} + \frac{1}{2}\|\xi_\tau\|_{\Sigma^+}^2.$$

Rearranging we have that $\langle \xi_\tau, y_{\tau-1} - u \rangle_{\Sigma^+} \le V(y_{\tau-1}, u) - V(y_\tau, u) + \frac{1}{2}\|\xi_\tau\|_{\Sigma^+}^2$. Summing from $\tau = 1$ to t (i.e., telescoping), we get that

$$\sum_{\tau} \langle \xi_\tau, y_{\tau-1} - u \rangle_{\Sigma^+} \le V(y_0, u) - V(y_t, u) + \tfrac{1}{2}\sum_{\tau} \|\xi_\tau\|_{\Sigma^+}^2.$$

Drop the non-positive term $-V(y_t, u)$ to get (67). □

H.7 Deterministic-style Bound

We first prove a deterministic-style bound that holds path-by-path (a la Juditsky et al. (2011), Theorem 2). This bound holds uniformly in $u \in D$ and is the foundation for all subsequent expected-error bounds. Recall that

$$\mathrm{Gap}_{\mathrm{vi}}(z) := \sup_{u \in D} \langle F_{\mathcal{R}}(u), z - u \rangle.$$

For the initial point $r_0 \in D$, write $\Theta(r_0) := \max_{u \in D} V(r_0, u)$.

Theorem 7 (Deterministic-style bound)**.** Assume **(P-mono)** and **(P-lip)**. Run the stochastic projected Mirror-Prox algorithm (54)-(55) from any $r_0 \in D$ with constant step size $\gamma_\tau \equiv \gamma > 0$ satisfying $\gamma \le \frac{1}{\sqrt{3}\, L_G}$ where $L_G = \sigma_{\max}\ell_{\mathcal{R}}$. Then for any sample path, the error is bounded as follows:

$$\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t) \le \frac{1}{t\gamma}\,\Gamma(t), \tag{68}$$

where

$$\Gamma(t) := 2\Theta(r_0) + \sum_{\tau=1}^{t}\Big[\frac{3\gamma^2}{2}\varepsilon_{r_{\tau-1}}^2 + 2\gamma^2\varepsilon_{w_\tau}^2\Big] + \sum_{\tau=1}^{t}\langle \gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+},$$

with $\Delta_\tau, \varepsilon_{r_{\tau-1}}, \varepsilon_{w_\tau}$ defined in Section H.5 and $\{y_\tau\}$ the ghost-iterate sequence (66) driven by $\xi_\tau = \gamma\Delta_\tau$.

Proof. The proof proceeds in steps.

Step 1: descent identity for one iteration. Apply Lemma 12 with

$$z = r_{\tau-1},\ \zeta = \gamma\widehat{G}_m^{(\tau,1)} = \gamma\Sigma\widehat{F}_m^{(\tau,1)}(r_{\tau-1}),\ \eta = \gamma\widehat{G}_m^{(\tau,2)} = \gamma\Sigma\widehat{F}_m^{(\tau,2)}(w_\tau),$$

so $w = w_\tau$ and $r_+ = r_\tau$. By (64), for all $u \in D$, we have that

$$V(r_\tau, u) - V(r_{\tau-1}, u) \le \langle\gamma\widehat{G}_m^{(\tau,2)}, u - w_\tau\rangle_{\Sigma^+} + \frac{1}{2}\|\gamma\widehat{G}_m^{(\tau,1)} - \gamma\widehat{G}_m^{(\tau,2)}\|_{\Sigma^+}^2 - \frac{1}{2}\|w_\tau - r_{\tau-1}\|_{\Sigma^+}^2. \quad (69)$$

Step 2: bound the difference of oracle calls. By the triangle inequality, we have

$$\begin{aligned}\|\widehat{G}_m^{(\tau,1)} - \widehat{G}_m^{(\tau,2)}\|_{\Sigma^+} &\le \|\widehat{G}_m^{(\tau,1)} - G(r_{\tau-1})\|_{\Sigma^+} + \|G(r_{\tau-1}) - G(w_\tau)\|_{\Sigma^+} + \|G(w_\tau) - \widehat{G}_m^{(\tau,2)}\|_{\Sigma^+} \\ &= \varepsilon_{r_{\tau-1}} + \|G(r_{\tau-1}) - G(w_\tau)\|_{\Sigma^+} + \varepsilon_{w_\tau} \\ &\le \varepsilon_{r_{\tau-1}} + L_G\|r_{\tau-1} - w_\tau\|_{\Sigma^+} + \varepsilon_{w_\tau}\end{aligned}$$

using Lipschitz of G in Σ^+ metric (Lemma 9). Using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we get that

$$\|\widehat{G}_m^{(\tau,1)} - \widehat{G}_m^{(\tau,2)}\|_{\Sigma^+}^2 \le 3\varepsilon_{r_{\tau-1}}^2 + 3L_G^2\|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2 + 3\varepsilon_{w_\tau}^2.$$

Substituting into (69), we have that

$$\begin{aligned}V(r_\tau, u) - V(r_{\tau-1}, u) \le \langle\gamma\widehat{G}_m^{(\tau,2)}, u - w_\tau\rangle_{\Sigma^+} + \frac{3\gamma^2}{2}\big(\varepsilon_{r_{\tau-1}}^2 + L_G^2\|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2 + \varepsilon_{w_\tau}^2\big) \\ - \frac{1}{2}\|w_\tau - r_{\tau-1}\|_{\Sigma^+}^2.\end{aligned}$$

The $\|w_\tau - r_{\tau-1}\|^2$ terms combine: coefficient $-\frac{1}{2} + \frac{3\gamma^2 L_G^2}{2}$. Since $\gamma \le 1/(\sqrt{3}\,L_G)$ by design, we have that $3\gamma^2 L_G^2 \le 1$, so this coefficient is non-positive and dropping the term, we have that

$$V(r_\tau, u) - V(r_{\tau-1}, u) \le \langle\gamma\widehat{G}_m^{(\tau,2)}, u - w_\tau\rangle_{\Sigma^+} + \frac{3\gamma^2}{2}\big(\varepsilon_{r_{\tau-1}}^2 + \varepsilon_{w_\tau}^2\big). \quad (70)$$

Step 3: relate the inner product to $F_\mathcal{R}$. Recall $\widehat{G}_m^{(\tau,2)} = \Sigma\widehat{F}_m^{(\tau,2)}(w_\tau)$, so that

$$\langle\widehat{G}_m^{(\tau,2)}, u - w_\tau\rangle_{\Sigma^+} = (u - w_\tau)^\top\Sigma^+\Sigma\widehat{F}_m^{(\tau,2)} = (u - w_\tau)^\top\widehat{F}_m^{(\tau,2)},$$

using $\Sigma^+\Sigma = \Pi_\mathcal{W}$ and $u - w_\tau \in \mathcal{W}$ (both in $\theta_{\text{ref}} + \mathcal{W}$). Decompose $\widehat{F}_m^{(\tau,2)} = F_\mathcal{R}(w_\tau) - \Delta_\tau^F$ where $\Delta_\tau^F := F_\mathcal{R}(w_\tau) - \widehat{F}_m^{(\tau,2)}$ (the unpreconditioned discrepancy). Then

$$\langle\widehat{G}_m^{(\tau,2)}, u - w_\tau\rangle_{\Sigma^+} = \langle F_\mathcal{R}(w_\tau), u - w_\tau\rangle - \langle\Delta_\tau^F, u - w_\tau\rangle.$$

Note that Δ_τ in Section H.5 was defined as the preconditioned discrepancy $G(w_\tau) - \widehat{G}_m^{(\tau,2)} = \Sigma\Delta_\tau^F$. The relation between the two: for $u, w_\tau \in \theta_{\text{ref}} + \mathcal{W}$, $\langle\Delta_\tau, u - w_\tau\rangle_{\Sigma^+} = (u - w_\tau)^\top\Sigma^+\Sigma\Delta_\tau^F = (u - w_\tau)^\top\Delta_\tau^F = \langle\Delta_\tau^F, u - w_\tau\rangle$. Hence we have that

$$\langle\widehat{G}_m^{(\tau,2)}, u - w_\tau\rangle_{\Sigma^+} = \langle F_\mathcal{R}(w_\tau), u - w_\tau\rangle - \langle\Delta_\tau, u - w_\tau\rangle_{\Sigma^+}.$$

By substituting into (70), we get that

$$V(r_\tau, u) - V(r_{\tau-1}, u) \le \gamma\langle F_\mathcal{R}(w_\tau), u - w_\tau\rangle - \langle\gamma\Delta_\tau, u - w_\tau\rangle_{\Sigma^+} + \frac{3\gamma^2}{2}(\varepsilon_{r_{\tau-1}}^2 + \varepsilon_{w_\tau}^2). \quad (71)$$

Step 4: telescoping. Summing (71) over $\tau = 1, \ldots, t$ yields

$$V(r_t, u) - V(r_0, u) \le \sum_{\tau=1}^t \gamma\langle F_\mathcal{R}(w_\tau), u - w_\tau\rangle - \sum_{\tau=1}^t\langle\gamma\Delta_\tau, u - w_\tau\rangle_{\Sigma^+} + \sum_{\tau=1}^t \frac{3\gamma^2}{2}(\varepsilon_{r_{\tau-1}}^2 + \varepsilon_{w_\tau}^2).$$

By rearranging, with $V(r_t, u) \ge 0$ and $V(r_0, u) \le \Theta(r_0)$ by definition, we have that

$$\sum_\tau \gamma\langle F_\mathcal{R}(w_\tau), w_\tau - u\rangle \le \Theta(r_0) + \sum_\tau\langle\gamma\Delta_\tau, w_\tau - u\rangle_{\Sigma^+} + \sum_\tau \frac{3\gamma^2}{2}(\varepsilon_{r_{\tau-1}}^2 + \varepsilon_{w_\tau}^2). \quad (72)$$

Step 5: ghost-iterate decomposition (sup over u). The quantity $\sum_\tau \langle \gamma\Delta_\tau, w_\tau - u\rangle_{\Sigma^+}$ on the right of (72) depends on u; to take sup over u uniformly, decompose:

$$\sum_\tau \langle \gamma\Delta_\tau, w_\tau - u\rangle_{\Sigma^+} = \sum_\tau \langle \gamma\Delta_\tau, y_{\tau-1} - u\rangle_{\Sigma^+} + \sum_\tau \langle \gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+}. \tag{73}$$

For the first sum, applying Lemma 13 with $\xi_\tau = \gamma\Delta_\tau$ gives

$$\sum_\tau \langle \gamma\Delta_\tau, y_{\tau-1} - u\rangle_{\Sigma^+} \le V(y_0, u) + \frac{1}{2}\sum_\tau \|\gamma\Delta_\tau\|^2_{\Sigma^+} \le \Theta(r_0) + \frac{\gamma^2}{2}\sum_\tau \|\Delta_\tau\|^2_{\Sigma^+},$$

using $y_0 = r_0$ and $V(r_0, u) \le \Theta(r_0)$ for all $u \in D$. This bounds the first sum in (73). Substitute back into (72) via (73), we get that

$$\begin{aligned}\sum_\tau \gamma\langle F_{\mathcal{R}}(w_\tau), w_\tau - u\rangle \le \Theta(r_0) + \Theta(r_0) + \frac{\gamma^2}{2}\sum_\tau \|\Delta_\tau\|^2_{\Sigma^+} + \sum_\tau \langle \gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+} \\ + \sum_\tau \frac{3\gamma^2}{2}(\varepsilon^2_{r_{\tau-1}} + \varepsilon^2_{w_\tau}).\end{aligned} \tag{74}$$

Step 6: bound $\|\Delta_\tau\|^2_{\Sigma^+}$. By definition $\Delta_\tau = G(w_\tau) - \widehat{G}^{(\tau,2)}_m = -\Sigma b_m(w_\tau) - \Sigma\zeta^{(\tau,2)}_m$ so that

$$\|\Delta_\tau\|_{\Sigma^+} = \|\Sigma\zeta^{(\tau,2)}_m + \Sigma b_m(w_\tau)\|_{\Sigma^+}.$$

By the definition of $\varepsilon_{w_\tau} := \|\widehat{G}^{(\tau,2)}_m(w_\tau) - G(w_\tau)\|_{\Sigma^+}$, we have that $\|\Delta_\tau\|^2_{\Sigma^+} = \varepsilon^2_{w_\tau}$. Substituting into (74) yields

$$\sum_\tau \gamma\langle F_{\mathcal{R}}(w_\tau), w_\tau - u\rangle \le 2\Theta(r_0) + \frac{\gamma^2}{2}\sum_\tau \varepsilon^2_{w_\tau} + \sum_\tau \langle \gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+} + \sum_\tau \frac{3\gamma^2}{2}(\varepsilon^2_{r_{\tau-1}} + \varepsilon^2_{w_\tau}). \tag{75}$$

Combining the $\varepsilon^2_{w_\tau}$ coefficients, we get that $\frac{\gamma^2}{2} + \frac{3\gamma^2}{2} = 2\gamma^2$. Combining all of this into $\Gamma(t)$, we have

$$\Gamma(t) = 2\Theta(r_0) + \sum_\tau \frac{3\gamma^2}{2}\varepsilon^2_{r_{\tau-1}} + 2\gamma^2\sum_\tau \varepsilon^2_{w_\tau} + \sum_\tau \langle \gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+}. \tag{76}$$

Step 7: convert to $\mathrm{Gap}_{\mathrm{vi}}$. By monotonicity of $F_{\mathcal{R}}$, for all $u, w \in D$, we have $\langle F_{\mathcal{R}}(u), w - u\rangle \le \langle F_{\mathcal{R}}(w), w - u\rangle$. Hence

$$\langle F_{\mathcal{R}}(u), w_\tau - u\rangle \le \langle F_{\mathcal{R}}(w_\tau), w_\tau - u\rangle,$$

and summing then averaging by $\sum \gamma_\tau = t\gamma$, we get that

$$\langle F_{\mathcal{R}}(u), \widehat{z}_t - u\rangle = (t\gamma)^{-1}\sum_\tau \gamma\langle F_{\mathcal{R}}(u), w_\tau - u\rangle \le (t\gamma)^{-1}\sum_\tau \gamma\langle F_{\mathcal{R}}(w_\tau), w_\tau - u\rangle.$$

By (75), the right side is $\le (t\gamma)^{-1}\Gamma(t)$ for *every fixed* $u \in D$. Take supremum over u to get

$$\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t) = \sup_u \langle F_{\mathcal{R}}(u), \widehat{z}_t - u\rangle \le (t\gamma)^{-1}\Gamma(t),$$

which is (68). □

H.8 Monotone setting: bounding the expected gap

Building on Juditsky et al. (2011), we prove the convergence of stochastic mirror prox (extragradient) in the montone setting with a biased oracle for learning in risk-sensitive preference games. This requires handling the terms in both the bias and variance from estimating the payoff-side risk functional.

Theorem 8 (Stochastic Mirror-Prox, monotone case, biased oracle)**.** Under the assumptions of Theorem 7, let $r_0 \in D$ be arbitrary. Suppose the oracle satisfies (O1) and (O2), with the squared bias and variance in the Σ^+ metric bounded by $\widetilde{B}^2_m := \sigma_{\max} B^2_m$ and $\widetilde{V}_m := \sigma_{\max} V_m$. Then the estimate holds:

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \frac{2\Theta(r_0)}{t\gamma} + \frac{7\gamma}{2}\big(\widetilde{B}^2_m + \widetilde{V}_m\big) + 2\Omega\widetilde{B}_m. \tag{77}$$

Proof. Take expectations of $\Gamma(t)/(t\gamma)$ from (76), we have that

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \frac{1}{t\gamma}\,\mathbb{E}[\Gamma(t)] = \frac{2\Theta(r_0)}{t\gamma} + \frac{1}{t\gamma}\sum_{\tau}\Big[\tfrac{3\gamma^2}{2}\,\mathbb{E}\,\varepsilon^2_{r_{\tau-1}} + 2\gamma^2\,\mathbb{E}\,\varepsilon^2_{w_\tau} + \mathbb{E}\langle\gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+}\Big].$$

Bound 1: $\mathbb{E}\,\varepsilon^2_{r_{\tau-1}}$ and $\mathbb{E}\,\varepsilon^2_{w_\tau}$. By (58)–(59), we have that

$$\mathbb{E}\,\varepsilon^2_{r_{\tau-1}} \le \widetilde{B}_m^2 + \widetilde{V}_m, \quad \text{and} \quad \mathbb{E}\,\varepsilon^2_{w_\tau} \le \widetilde{B}_m^2 + \widetilde{V}_m.$$

Therefore,

$$\frac{1}{t\gamma}\sum_{\tau=1}^{t}\left[\frac{3\gamma^2}{2}\,\mathbb{E}\,\varepsilon^2_{r_{\tau-1}} + 2\gamma^2\,\mathbb{E}\,\varepsilon^2_{w_\tau}\right] \le \frac{1}{t\gamma}\sum_{\tau=1}^{t}\frac{7\gamma^2}{2}\left(\widetilde{B}_m^2 + \widetilde{V}_m\right) = \frac{7\gamma}{2}\left(\widetilde{B}_m^2 + \widetilde{V}_m\right).$$

Bound 2: $\mathbb{E}\langle\gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+}$. Recall w_τ is $\mathcal{F}_\tau$-measurable (uses the first oracle call of step τ); $y_{\tau-1}$ is $\mathcal{G}_{\tau-1}$-measurable (depends on $\Delta_1,\dots,\Delta_{\tau-1}$ through the ghost recursion); Δ_τ is $\mathcal{G}_\tau$-measurable but need not be $\mathcal{F}_\tau$-measurable (uses the second oracle call of step τ). Thus, by the tower property of expectations, we have that

$$\mathbb{E}\langle\gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+} = \gamma\,\mathbb{E}\left[\langle\mathbb{E}[\Delta_\tau \mid \mathcal{F}_\tau], w_\tau - y_{\tau-1}\rangle_{\Sigma^+}\right].$$

Here $y_{\tau-1}$ is $\mathcal{G}_{\tau-1}$-measurable and hence $\mathcal{F}_\tau$-measurable; w_τ is $\mathcal{F}_\tau$-measurable.

By (56), the upper bound holds: $\|\,\mathbb{E}[\Delta_\tau \mid \mathcal{F}_\tau]\|_{\Sigma^+} \le \widetilde{B}_m$. Hence, by Cauchy–Schwarz and the fact that $w_\tau, y_{\tau-1} \in D$, we have

$$\mathbb{E}\langle\gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+} \le \gamma\,\mathbb{E}\left[\|\mathbb{E}[\Delta_\tau \mid \mathcal{F}_\tau]\|_{\Sigma^+}\,\|w_\tau - y_{\tau-1}\|_{\Sigma^+}\right] \le 2\gamma\Omega\widetilde{B}_m.$$

Therefore, we deduce that

$$\frac{1}{t\gamma}\sum_{\tau=1}^{t}\mathbb{E}\langle\gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+} \le \frac{1}{t\gamma}\sum_{\tau=1}^{t}2\gamma\Omega\widetilde{B}_m = 2\Omega\widetilde{B}_m.$$

Combining the bounds. Substituting the two bounds above into

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \frac{2\Theta(r_0)}{t\gamma} + \frac{1}{t\gamma}\sum_{\tau=1}^{t}\left[\frac{3\gamma^2}{2}\,\mathbb{E}\,\varepsilon^2_{r_{\tau-1}} + 2\gamma^2\,\mathbb{E}\,\varepsilon^2_{w_\tau} + \mathbb{E}\langle\gamma\Delta_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+}\right]$$

yields

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \frac{2\Theta(r_0)}{t\gamma} + \frac{7\gamma}{2}\left(\widetilde{B}_m^2 + \widetilde{V}_m\right) + 2\Omega\widetilde{B}_m,$$

which is (77). This concludes the proof. □

H.8.1 Corollary: optimal step size

Theorem 8 bounds the expected gap as a function of the step size γ, leaving open the question of how to choose γ to minimize the bound. The optimal choice depends on whether the dominant source of error is the operator's Lipschitz continuity (favoring smaller γ) or the noise from finite-sample estimation of the risk-adjusted operator (favoring a step size that trades off iteration count against per-step variance). The following corollary makes this trade-off explicit.

Corollary 1 (Optimal step size, monotone case)**.** Choosing

$$\gamma = \min\left\{\frac{1}{\sqrt{3}L_G},\ \sqrt{\frac{4\Theta(r_0)}{7t(\widetilde{B}_m^2 + \widetilde{V}_m)}}\right\} \tag{78}$$

in Theorem 8, with $\gamma = 1/(\sqrt{3}L_G)$ when $\widetilde{B}_m^2 + \widetilde{V}_m = 0$, gives

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \max\left\{\frac{7L_G\Theta(r_0)}{t},\ 7\sqrt{\frac{2\Theta(r_0)(\widetilde{B}_m^2 + \widetilde{V}_m)}{3t}}\right\} + 2\Omega\widetilde{B}_m. \tag{79}$$

The leading rate is $\mathcal{O}(t^{-1/2})$ in the variance-dominated regime and $\mathcal{O}(t^{-1})$ in the Lipschitz-dominated regime; the bias contributes the term $2\Omega\widetilde{B}_m$, independent of t.

Proof. Write $S := \widetilde{B}_m^2 + \widetilde{V}_m$. The chosen step size minimizes $f(\gamma) := 2\Theta(r_0)/(t\gamma) + (7/2)\gamma S$ subject to $0 < \gamma \le 1/(\sqrt{3}L_G)$. If the cap is selected, then $S \le 12L_G^2\Theta(r_0)/(7t)$, and

$$f(\gamma) = \frac{2\sqrt{3}L_G\Theta(r_0)}{t} + \frac{7S}{2\sqrt{3}L_G} \le \frac{4\sqrt{3}L_G\Theta(r_0)}{t} \le \frac{7L_G\Theta(r_0)}{t}.$$

Otherwise, $f(\gamma) = \sqrt{28\Theta(r_0)S/t} \le 7\sqrt{2\Theta(r_0)S/(3t)}$. Adding $2\Omega\widetilde{B}_m$ proves (79). □

Substituting concrete entropic constants. For fixed entropic risk parameter λ, using $\widetilde{B}_m^2 = \sigma_{\max}B_m^2 = \mathcal{O}(\sigma_{\max}|\mathcal{Y}|/m^2)$ and $\widetilde{V}_m = \sigma_{\max}V_m = \mathcal{O}(\sigma_{\max}|\mathcal{Y}|/m)$ from Proposition 8, the optimal-step-size bound becomes

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] = O\Big(\frac{L_G\Theta(r_0)}{t}\Big) + O\Big(\sqrt{\frac{\Theta(r_0)\sigma_{\max}|\mathcal{Y}|}{mt}}\Big) + O\Big(\frac{\Omega\sqrt{\sigma_{\max}|\mathcal{Y}|}}{m}\Big).$$

The bias contribution scales as $1/m$ (last term); the variance contribution scales as $1/\sqrt{mt}$ (middle term); the deterministic Lipschitz contribution scales as $1/t$ (first term). For fixed t, increasing m makes the sampling terms vanish, leaving the deterministic $1/t$ term.

Initialization at the reference policy. Finally, setting $r_0 = z_c = \theta_{\mathrm{ref}}$ gives $2\Theta(r_0) = \Omega^2$, as in Juditsky et al. (2011, Section 3.4). The bound (77) then becomes

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \frac{\Omega^2}{t\gamma} + \frac{7\gamma}{2}(\widetilde{B}_m^2 + \widetilde{V}_m) + 2\Omega\widetilde{B}_m,$$

and the step size (78) reduces to

$$\gamma = \min\left\{\frac{1}{\sqrt{3}L_G},\ \Omega\sqrt{\frac{2}{7t(\widetilde{B}_m^2 + \widetilde{V}_m)}}\right\}.$$

The optimized and entropic bounds specialize by the same substitution $\Theta(r_0) = \Omega^2/2$.

H.9 Strongly monotone setting

We now prove last-iterate linear contraction in the strongly monotone case. The proof uses the same descent identity as in Juditsky et al. (2011) (Lemma 12), specialized to $u = \theta^\star$ and combined with strong monotonicity to extract a contraction factor.

Effective strong-monotonicity modulus in Σ^+. By b., $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone in the standard inner product. Translating to the Σ^+ inner product on $\mathcal{W}$, we have that

$$\begin{aligned}\langle G(\theta_1) - G(\theta_2), \theta_1 - \theta_2\rangle_{\Sigma^+} = \langle F_{\mathcal{R}}(\theta_1) - F_{\mathcal{R}}(\theta_2), \theta_1 - \theta_2\rangle &\ge \mu_{\mathcal{R}}\|\theta_1 - \theta_2\|_2^2 \\ &\ge \mu_{\mathcal{R}}\sigma_{\min}\|\theta_1 - \theta_2\|_{\Sigma^+}^2,\end{aligned}$$

where the first equality uses Lemma 8(i) and the second inequality uses (50). Define the effective strong-monotonicity modulus

$$\widetilde{\mu} := \mu_{\mathcal{R}}\sigma_{\min}.$$

Then G is $\widetilde{\mu}$-strongly monotone in the Σ^+ inner product on $\mathcal{W}$.

Theorem 9 (Stochastic Mirror-Prox, strongly monotone, biased oracle)**.** Suppose $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone (b.) and $\ell_{\mathcal{R}}$-Lipschitz (c.) on $\theta_{\mathrm{ref}} + \mathcal{W}$, the oracle satisfies (O1)–(O2) of Section G, and the algorithm (54)-(55) is run with constant step size $\gamma_\tau \equiv \gamma$ satisfying

$$0 < \gamma \le \min\left\{1/(4\widetilde{\mu}),\ 1/(\sqrt{6}\,L_G)\right\}, \qquad \text{with} \quad L_G = \sigma_{\max}\ell_{\mathcal{R}}. \tag{80}$$

Writing $\theta_\tau := r_\tau$, for all $T \ge 1$ the estimate holds:

$$\mathbb{E}\,\|\theta_T - \theta^\star\|_{\Sigma^+}^2 \le (1 - \gamma\widetilde{\mu})^T\|\theta_0 - \theta^\star\|_{\Sigma^+}^2 + \frac{4\Omega\widetilde{B}_m}{\widetilde{\mu}} + \frac{6\gamma(\widetilde{B}_m^2 + \widetilde{V}_m)}{\widetilde{\mu}}. \tag{81}$$

Proof. We work step by step from the two-step descent identity (Lemma 12), specialized to $u = \theta^\star$ and followed by strong monotonicity.

Step 1: descent identity at $u = \theta^\star$. First apply (64) of Lemma 12 with $z = r_{\tau-1}$, $\zeta = \gamma\widehat{G}_m^{(\tau,1)}$, $\eta = \gamma\widehat{G}_m^{(\tau,2)}$ (so $w = w_\tau$, $r_+ = r_\tau$), and $u = \theta^\star$. This gives

$$V(r_\tau, \theta^\star) - V(r_{\tau-1}, \theta^\star) \le \gamma\langle \widehat{G}_m^{(\tau,2)}, \theta^\star - w_\tau\rangle_{\Sigma^+} + \frac{\gamma^2}{2}\|\widehat{G}_m^{(\tau,1)} - \widehat{G}_m^{(\tau,2)}\|_{\Sigma^+}^2 - \tfrac{1}{2}\|w_\tau - r_{\tau-1}\|_{\Sigma^+}^2. \tag{82}$$

Step 2: decompose the inner product term. Decompose $\widehat{G}_m^{(\tau,2)} = G(w_\tau) + \Sigma b_m(w_\tau) + \Sigma\zeta_m^{(\tau,2)}$ (unpreconditioned bias b_m, noise $\zeta_m^{(\tau,2)}$), so that

$$\langle \widehat{G}_m^{(\tau,2)}, \theta^\star - w_\tau\rangle_{\Sigma^+} = \langle G(w_\tau), \theta^\star - w_\tau\rangle_{\Sigma^+} + \langle \Sigma b_m(w_\tau), \theta^\star - w_\tau\rangle_{\Sigma^+} + \langle \Sigma\zeta_m^{(\tau,2)}, \theta^\star - w_\tau\rangle_{\Sigma^+}.$$

Using Lemma 8 part (i), we have that $\langle \Sigma v, w\rangle_{\Sigma^+} = \langle v, w\rangle$ for $w \in \mathcal{W}$ with v arbitrary; we use $w = \theta^\star - w_\tau \in \mathcal{W}$. Apply this three times to get that

$$\langle \widehat{G}_m^{(\tau,2)}, \theta^\star - w_\tau\rangle_{\Sigma^+} = \langle F_{\mathcal{R}}(w_\tau), \theta^\star - w_\tau\rangle + \langle b_m(w_\tau), \theta^\star - w_\tau\rangle + \langle \zeta_m^{(\tau,2)}, \theta^\star - w_\tau\rangle. \tag{83}$$

Step 3: apply strong monotonicity. The risk-adjusted equilibrium condition gives $F_{\mathcal{R}}(\theta^\star) \in \ker\Sigma$. Since $\theta^\star - w_\tau \in \mathcal{W}$ and $\ker\Sigma = \mathcal{W}^\perp$, we have that

$$\langle F_{\mathcal{R}}(\theta^\star), \theta^\star - w_\tau\rangle = 0.$$

Hence $\langle F_{\mathcal{R}}(w_\tau), \theta^\star - w_\tau\rangle = \langle F_{\mathcal{R}}(w_\tau) - F_{\mathcal{R}}(\theta^\star), \theta^\star - w_\tau\rangle = -\langle F_{\mathcal{R}}(w_\tau) - F_{\mathcal{R}}(\theta^\star), w_\tau - \theta^\star\rangle$. By b., the latter inner product is $\ge \mu_{\mathcal{R}}\|w_\tau - \theta^\star\|_2^2 \ge \widetilde{\mu}\|w_\tau - \theta^\star\|_{\Sigma^+}^2$ (using (50) and $\widetilde{\mu} = \mu_{\mathcal{R}}\sigma_{\min}$). Thus we deduce that

$$\langle F_{\mathcal{R}}(w_\tau), \theta^\star - w_\tau\rangle \le -\widetilde{\mu}\|w_\tau - \theta^\star\|_{\Sigma^+}^2. \tag{84}$$

Substitute (84) into (83) leads to

$$\langle \widehat{G}_m^{(\tau,2)}, \theta^\star - w_\tau\rangle_{\Sigma^+} \le -\widetilde{\mu}\|w_\tau - \theta^\star\|_{\Sigma^+}^2 - \langle b_m(w_\tau), w_\tau - \theta^\star\rangle - \langle \zeta_m^{(\tau,2)}, w_\tau - \theta^\star\rangle. \tag{85}$$

Step 4: bound the Lipschitz / variance squared term. By the triangle inequality and Lipschitz continuity of G in Σ^+ metric (Lemma 9), we have that

$$\begin{aligned}\|\widehat{G}_m^{(\tau,1)} - \widehat{G}_m^{(\tau,2)}\|_{\Sigma^+} &\le \|\widehat{G}_m^{(\tau,1)} - G(r_{\tau-1})\|_{\Sigma^+} + \|G(r_{\tau-1}) - G(w_\tau)\|_{\Sigma^+} + \|G(w_\tau) - \widehat{G}_m^{(\tau,2)}\|_{\Sigma^+} \\ &\le \varepsilon_{r_{\tau-1}} + L_G\|r_{\tau-1} - w_\tau\|_{\Sigma^+} + \varepsilon_{w_\tau}.\end{aligned}$$

Using the property $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we deduce that

$$\|\widehat{G}_m^{(\tau,1)} - \widehat{G}_m^{(\tau,2)}\|_{\Sigma^+}^2 \le 3\varepsilon_{r_{\tau-1}}^2 + 3L_G^2\|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2 + 3\varepsilon_{w_\tau}^2. \tag{86}$$

Step 5: combine. Substitute (85) and (86) into (82) to get that

$$\begin{aligned}V(r_\tau, \theta^\star) - V(r_{\tau-1}, \theta^\star) \le{}& -\gamma\widetilde{\mu}\|w_\tau - \theta^\star\|_{\Sigma^+}^2 - \gamma\langle b_m(w_\tau), w_\tau - \theta^\star\rangle - \gamma\langle \zeta_m^{(\tau,2)}, w_\tau - \theta^\star\rangle \\ &+ \frac{3\gamma^2}{2}\big(\varepsilon_{r_{\tau-1}}^2 + L_G^2\|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2 + \varepsilon_{w_\tau}^2\big) - \frac{1}{2}\|w_\tau - r_{\tau-1}\|_{\Sigma^+}^2.\end{aligned} \tag{87}$$

Step 6: bridge $\|w_\tau - \theta^\star\|^2$ to $V(r_{\tau-1}, \theta^\star)$. Apply the elementary inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ with $a = r_{\tau-1} - w_\tau$, $b = w_\tau - \theta^\star$ (both in $\mathcal{W}$) to get that

$$\|r_{\tau-1} - \theta^\star\|_{\Sigma^+}^2 \le 2\|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2 + 2\|w_\tau - \theta^\star\|_{\Sigma^+}^2,$$

which rearranges to

$$\|w_\tau - \theta^\star\|_{\Sigma^+}^2 \ge \tfrac{1}{2}\|r_{\tau-1} - \theta^\star\|_{\Sigma^+}^2 - \|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2,$$

i.e., using $V(\cdot,\cdot) = \frac{1}{2}\|\cdot - \cdot\|_{\Sigma^+}^2$, we have

$$\|w_\tau - \theta^\star\|_{\Sigma^+}^2 \ge V(r_{\tau-1}, \theta^\star) - \|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2. \tag{88}$$

Hence

$$-\gamma\widetilde{\mu}\|w_\tau - \theta^\star\|_{\Sigma^+}^2 \le -\gamma\widetilde{\mu}V(r_{\tau-1}, \theta^\star) + \gamma\widetilde{\mu}\|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2. \tag{89}$$

Substituting (89) into (87), we get that

$$\begin{aligned}V(r_\tau, \theta^\star) \le{}& V(r_{\tau-1}, \theta^\star) - \gamma\widetilde{\mu}V(r_{\tau-1}, \theta^\star) + \gamma\widetilde{\mu}\|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2 \\ &- \gamma\langle b_m(w_\tau), w_\tau - \theta^\star\rangle - \gamma\langle \zeta_m^{(\tau,2)}, w_\tau - \theta^\star\rangle \\ &+ \frac{3\gamma^2}{2}\big(\varepsilon_{r_{\tau-1}}^2 + L_G^2\|r_{\tau-1} - w_\tau\|_{\Sigma^+}^2 + \varepsilon_{w_\tau}^2\big) - \frac{1}{2}\|w_\tau - r_{\tau-1}\|_{\Sigma^+}^2.\end{aligned}$$

Step 7: drop the $\|r_{\tau-1}-w_\tau\|^2$ term under the step-size constraint. The coefficient on $\|r_{\tau-1}-w_\tau\|^2_{\Sigma^+}$ in (H.9) is

$$C_w := \gamma\widetilde{\mu} + \frac{3\gamma^2 L_G^2}{2} - \frac{1}{2}.$$

Under (80), the bound $\gamma \le 1/(4\widetilde{\mu})$ gives $\gamma\widetilde{\mu} \le 1/4$ and $\gamma \le 1/(\sqrt{6}\,L_G)$ gives $\gamma^2 L_G^2 \le 1/6$, so that $\frac{3\gamma^2 L_G^2}{2} \le 1/4$. Hence $C_w \le 1/4 + 1/4 - 1/2 = 0$. We drop the $\|r_{\tau-1}-w_\tau\|^2_{\Sigma^+}$ term, to get that

$$V(r_\tau,\theta^\star) \le (1-\gamma\widetilde{\mu})V(r_{\tau-1},\theta^\star) - \gamma\langle b_m(w_\tau)+\zeta_m^{(\tau,2)}, w_\tau-\theta^\star\rangle + \frac{3\gamma^2}{2}(\varepsilon^2_{r_{\tau-1}}+\varepsilon^2_{w_\tau}). \tag{90}$$

Step 8: take expectation. Take $\mathbb{E}[\cdot \mid \mathcal{F}_\tau]$ of (90). Recall the filtration structure (Section H.5): $r_{\tau-1}$ is $\mathcal{G}_{\tau-1}$-measurable, w_τ is $\mathcal{F}_\tau$-measurable. $\varepsilon_{r_{\tau-1}}$ is $\mathcal{F}_\tau$-measurable. Fresh sampling gives $\mathbb{E}[\zeta_m^{(\tau,2)} \mid \mathcal{F}_\tau] = 0$. $\varepsilon^2_{w_\tau}$ depends on w_τ (via the bias) and on $\zeta_m^{(\tau,2)}$ (via the noise); given $\mathcal{F}_\tau$, w_τ is fixed.

Term 1: bias cross. The bias $b_m(w_\tau)$ is deterministic given w_τ, hence given $\mathcal{F}_\tau$. Thus $\mathbb{E}[\langle b_m(w_\tau), w_\tau-\theta^\star\rangle \mid \mathcal{F}_\tau] = \langle b_m(w_\tau), w_\tau-\theta^\star\rangle$ (deterministic). Bound the bias as follows:

$$\begin{aligned}|\langle b_m(w_\tau), w_\tau-\theta^\star\rangle| &\le \|b_m(w_\tau)\|_2\|w_\tau-\theta^\star\|_2\\ &\le B_m\cdot\sqrt{\sigma_{\max}}\|w_\tau-\theta^\star\|_{\Sigma^+}\\ &\le \widetilde{B}_m\cdot 2\Omega,\end{aligned}$$

using (O1) ($\|b_m\|_2 \le B_m$), (50) ($\|v\|_2 \le \sqrt{\sigma_{\max}}\|v\|_{\Sigma^+}$), $\widetilde{B}_m = \sqrt{\sigma_{\max}}B_m$, and the Σ^+-diameter of D is 2Ω.

Term 2: noise cross. The term $\zeta_m^{(\tau,2)}$ has conditional mean zero given $\mathcal{F}_\tau$. Thus $\mathbb{E}[\langle\zeta_m^{(\tau,2)}, w_\tau-\theta^\star\rangle \mid \mathcal{F}_\tau] = \langle\mathbb{E}[\zeta_m^{(\tau,2)} \mid \mathcal{F}_\tau], w_\tau-\theta^\star\rangle = \langle 0, w_\tau-\theta^\star\rangle = 0$.

Term 3: noise-squared $\varepsilon^2_{r_{\tau-1}}$. This is $\mathcal{F}_\tau$-measurable (uses the first oracle call of step τ, included in $\mathcal{F}_\tau$). Take unconditional expectation later.

Term 4: noise-squared $\varepsilon^2_{w_\tau}$. Recall $\varepsilon_{w_\tau} = \|G(w_\tau) - \widehat{G}_m^{(\tau,2)}\|_{\Sigma^+}$. Decompose the operator as follows: $G(w_\tau) - \widehat{G}_m^{(\tau,2)} = -\Sigma b_m(w_\tau) - \Sigma\zeta_m^{(\tau,2)}$. Hence

$$\varepsilon^2_{w_\tau} = \|\Sigma b_m(w_\tau)\|^2_{\Sigma^+} + 2\langle\Sigma b_m(w_\tau), \Sigma\zeta_m^{(\tau,2)}\rangle_{\Sigma^+} + \|\Sigma\zeta_m^{(\tau,2)}\|^2_{\Sigma^+}.$$

Conditional on $\mathcal{F}_\tau$, w_τ is fixed and $\mathbb{E}[\zeta_m^{(\tau,2)} \mid \mathcal{F}_\tau] = 0$, so the cross term has zero conditional expectation—that is,

$$\mathbb{E}[\varepsilon^2_{w_\tau} \mid \mathcal{F}_\tau] = \|\Sigma b_m(w_\tau)\|^2_{\Sigma^+} + \mathbb{E}[\|\Sigma\zeta_m^{(\tau,2)}\|^2_{\Sigma^+} \mid \mathcal{F}_\tau] \le \widetilde{B}_m^2 + \widetilde{V}_m,$$

using (52)-(53).

Combine. Finally, we combine these bounds to get that

$$\mathbb{E}[V(r_\tau,\theta^\star) \mid \mathcal{F}_\tau] \le (1-\gamma\widetilde{\mu})V(r_{\tau-1},\theta^\star) + 2\gamma\Omega\widetilde{B}_m + 0 + \tfrac{3\gamma^2}{2}\varepsilon^2_{r_{\tau-1}} + \tfrac{3\gamma^2}{2}(\widetilde{B}_m^2+\widetilde{V}_m).$$

Now take unconditional expectation. In particular, $\mathbb{E}\,\varepsilon^2_{r_{\tau-1}} \le \widetilde{B}_m^2 + \widetilde{V}_m$ by the same decomposition argument applied to $\varepsilon_{r_{\tau-1}}$ (the first oracle call)—indeed, we have

$$\mathbb{E}\,V(r_\tau,\theta^\star) \le (1-\gamma\widetilde{\mu})\,\mathbb{E}\,V(r_{\tau-1},\theta^\star) + 2\gamma\Omega\widetilde{B}_m + 3\gamma^2(\widetilde{B}_m^2+\widetilde{V}_m). \tag{91}$$

Step 9: telescope. Define $A := 2\gamma\Omega\widetilde{B}_m + 3\gamma^2(\widetilde{B}_m^2+\widetilde{V}_m)$. Then (91) reads $\Lambda_\tau \le (1-\gamma\widetilde{\mu})\Lambda_{\tau-1} + A$ where $\Lambda_\tau := \mathbb{E}\,V(r_\tau,\theta^\star)$. Iterating we have that

$$\begin{aligned}\Lambda_T &\le (1-\gamma\widetilde{\mu})^T\Lambda_0 + A\sum_{j=0}^{T-1}(1-\gamma\widetilde{\mu})^j\\ &\le (1-\gamma\widetilde{\mu})^T\Lambda_0 + \frac{A}{\gamma\widetilde{\mu}}\\ &= (1-\gamma\widetilde{\mu})^T V(r_0,\theta^\star) + \frac{2\Omega\widetilde{B}_m}{\widetilde{\mu}} + \frac{3\gamma(\widetilde{B}_m^2+\widetilde{V}_m)}{\widetilde{\mu}}.\end{aligned}$$

Step 10: convert to $\|\cdot\|_{\Sigma^+}^2$. Since $V(z,u) = \frac{1}{2}\|z-u\|_{\Sigma^+}^2$, we have that

$$\mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 = 2\,\mathbb{E}\,V(r_T,\theta^\star) \le (1-\gamma\widetilde{\mu})^T\|r_0-\theta^\star\|_{\Sigma^+}^2 + \frac{4\Omega\widetilde{B}_m}{\widetilde{\mu}} + \frac{6\gamma(\widetilde{B}_m^2+\widetilde{V}_m)}{\widetilde{\mu}},$$

which is (81). □

Understanding the bound. The bound (81) has three core terms which can be interpreted or understood as follows.

- **Linear contraction.** $(1-\gamma\widetilde{\mu})^T\|r_0-\theta^\star\|_{\Sigma^+}^2$ contracts geometrically with rate $\gamma\widetilde{\mu} = \gamma\mu_{\mathcal{R}}\sigma_{\min}$. A sufficient number of iterations to halve the initial-distance contribution is $T = \lceil \log 2/(\gamma\widetilde{\mu})\rceil = \mathcal{O}(1/(\gamma\mu_{\mathcal{R}}\sigma_{\min}))$.
- **Bias floor.** $4\Omega\widetilde{B}_m/\widetilde{\mu} = 4\Omega B_m\sqrt{\sigma_{\max}}/(\mu_{\mathcal{R}}\sigma_{\min})$. Linear in B_m (which scales as $1/m$ for entropic, 0 for oracle-threshold CVaR-RU). The bias term in this upper bound is independent of step size γ and iteration count; increasing m reduces it for entropic risk.
- **Variance floor.** $6\gamma(\widetilde{B}_m^2+\widetilde{V}_m)/\widetilde{\mu} = \mathcal{O}(\gamma\sigma_{\max}V_m/(\mu_{\mathcal{R}}\sigma_{\min}))$ (the $\widetilde{B}_m^2$ contribution is dominated by $\widetilde{V}_m$ for $B_m^2 \le V_m$, which holds whenever variance V_m is order $1/m$ and bias is also $1/m$, giving $B_m^2 \sim 1/m^2 \ll V_m \sim 1/m$). The variance floor scales as γ/m; can be reduced by taking smaller γ or larger m.

Corollary 2 (Optimal step size, strongly monotone case)**.** With sample budget m fixed and target accuracy $\delta > 0$, choose

$$\gamma = \min\left\{\frac{1}{4\widetilde{\mu}},\ \frac{1}{\sqrt{6}\,L_G},\ \frac{\widetilde{\mu}\delta}{12(\widetilde{B}_m^2+\widetilde{V}_m)}\right\}, \tag{92}$$

omitting the last constraint when $\widetilde{B}_m^2+\widetilde{V}_m = 0$. If $2\|r_0-\theta^\star\|_{\Sigma^+}^2 > \delta$, run for

$$T = \left\lceil \frac{1}{\gamma\widetilde{\mu}}\log\frac{2\|r_0-\theta^\star\|_{\Sigma^+}^2}{\delta}\right\rceil$$

iterations; otherwise take $T=1$. Then $\mathbb{E}\,\|r_T-\theta^\star\|_{\Sigma^+}^2 \le \delta + 4\Omega\widetilde{B}_m/\widetilde{\mu}$. To make this upper bound at most a target $\delta' > 0$, take $\delta = \delta'/2$ and choose m large enough that $\widetilde{B}_m \le \widetilde{\mu}\delta'/(8\Omega)$. For entropic risk, it suffices to take $m \ge 8\Omega\sqrt{\sigma_{\max}|\mathcal{Y}|}\,C_b^{\mathrm{ent}}(\lambda)/(\widetilde{\mu}\delta')$, using Proposition 8.

Proof. With γ as in (92), the variance floor is $6\gamma(\widetilde{B}_m^2+\widetilde{V}_m)/\widetilde{\mu} \le \delta/2$. The contraction term is at most $\delta/2$ for the chosen T, using $(1-\gamma\widetilde{\mu})^T \le e^{-\gamma\widetilde{\mu}T}$. Thus the upper bound is $\delta/2 + \delta/2 + 4\Omega\widetilde{B}_m/\widetilde{\mu}$. □

H.10 Mirror Descent: corollary from extra-gradient analysis

We now state and prove the analogue of Theorems 8 and 9 for stochastic projected Mirror Descent (mirror descent), obtained by specializing the extragradient analysis to the case where the extrapolation step is omitted. The proofs reuse the technical lemmas of Section H (prox-mapping descent, ghost iterate) directly.

Algorithm. Stochastic projected Mirror Descent proceeds as follows. First, set $r_0 := \theta^{(0)}$. Then, for $\tau = 1,\dots,t$, update

$$r_\tau := P_{r_{\tau-1}}\big(\gamma_\tau\widehat{G}_m(r_{\tau-1})\big) = \Pi_D\big(r_{\tau-1} - \gamma_\tau\widehat{G}_m(r_{\tau-1})\big), \tag{93}$$

with each iteration using a fresh batch. Write ζ_τ for the corresponding centered oracle error in $\widehat{F}_m$. The output is

$$\widehat{z}_t := \Big(\sum_{\tau=1}^t\gamma_\tau\Big)^{-1}\sum_{\tau=1}^t\gamma_\tau r_{\tau-1}.$$

Filtration and noise quantities. Let

$$\mathcal{F}_\tau := \sigma(r_0, \zeta_1, \dots, \zeta_\tau),$$

so that $r_{\tau-1}$ is $\mathcal{F}_{\tau-1}$-measurable. Define the per-step discrepancy

$$\Delta_\tau := G(r_{\tau-1}) - \widehat{G}_m^{(\tau)} = -\Sigma b_m(r_{\tau-1}) - \Sigma\zeta_m^{(\tau)}, \tag{94}$$

where $\widehat{G}_m^{(\tau)} := \Sigma\widehat{F}_m^{(\tau)}(r_{\tau-1})$ uses ζ_τ. By the same bias-variance decomposition as in Section H.5, we have

$$\|\,\mathbb{E}[\Delta_\tau \mid \mathcal{F}_{\tau-1}]\|_{\Sigma^+} \le \widetilde{B}_m, \tag{95}$$

$$\mathbb{E}[\|\Delta_\tau\|_{\Sigma^+}^2 \mid \mathcal{F}_{\tau-1}] \le \widetilde{B}_m^2 + \widetilde{V}_m. \tag{96}$$

Indeed, conditional on $\mathcal{F}_{\tau-1}$, $r_{\tau-1}$ is fixed, $b_m(r_{\tau-1})$ is deterministic, and $\mathbb{E}[\zeta_m^{(\tau)} \mid \mathcal{F}_{\tau-1}] = 0$ by fresh sampling. So $\mathbb{E}[\Delta_\tau \mid \mathcal{F}_{\tau-1}] = -\Sigma b_m(r_{\tau-1})$ with $\|\Sigma b_m\|_{\Sigma^+} \le \widetilde{B}_m$ by (52). For the second moment, $\|\Delta_\tau\|_{\Sigma^+}^2 = \|\Sigma b_m\|_{\Sigma^+}^2 + 2\langle\Sigma b_m, \Sigma\zeta_m^{(\tau)}\rangle_{\Sigma^+} + \|\Sigma\zeta_m^{(\tau)}\|_{\Sigma^+}^2$, and conditional on $\mathcal{F}_{\tau-1}$ the cross term has zero mean, giving (96).

H.10.1 Monotone case

Corollary 3 (Stochastic Mirror Descent, monotone case, biased oracle)**.** Suppose $F_\mathcal{R}$ satisfies **(P-mono)** and **(P-lip)**, and the oracle satisfies (O1) and (O2). Run the mirror descent algorithm (93) with constant step size $\gamma_\tau \equiv \gamma > 0$ satisfying

$$\gamma \le 1/(2L_G), \qquad L_G = \sigma_{\max}\ell_\mathcal{R}.$$

Then

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \frac{2\Theta(r_0)}{t\gamma} + \frac{3\gamma}{2}\big(\widetilde{B}_m^2 + \widetilde{V}_m\big) + 4L_G^2\gamma\Omega^2 + 2\Omega\widetilde{B}_m. \tag{97}$$

Proof. The proof proceeds in steps mirroring those of Theorem 7 and Theorem 8, simplified by the fact that there is only one oracle call per iteration.

Step 1: descent identity for one iteration. Apply (61) of Lemma 11 with

$$z = r_{\tau-1}, \qquad \xi = \gamma\widehat{G}_m^{(\tau)}, \qquad w = r_\tau,$$

to get, for all $u \in D$,

$$V(r_\tau, u) \le V(r_{\tau-1}, u) + \langle\gamma\widehat{G}_m^{(\tau)}, u - r_{\tau-1}\rangle_{\Sigma^+} + \frac{1}{2}\|\gamma\widehat{G}_m^{(\tau)}\|_{\Sigma^+}^2.$$

Rearranging,

$$V(r_\tau, u) - V(r_{\tau-1}, u) \le \gamma\langle\widehat{G}_m^{(\tau)}, u - r_{\tau-1}\rangle_{\Sigma^+} + \frac{\gamma^2}{2}\|\widehat{G}_m^{(\tau)}\|_{\Sigma^+}^2. \tag{98}$$

Step 2: relate the inner product to $F_\mathcal{R}$. Recall $\widehat{G}_m^{(\tau)} = \Sigma\widehat{F}_m^{(\tau)}(r_{\tau-1})$. Decompose $\widehat{G}_m^{(\tau)} = G(r_{\tau-1}) - \Delta_\tau$ by (94). Following Step 3 of the proof of Theorem 7: since $u - r_{\tau-1} \in \mathcal{W}$ and $\Sigma^+\Sigma = \Pi_\mathcal{W}$ acts as the identity on $\mathcal{W}$,

$$\begin{aligned}\langle G(r_{\tau-1}), u - r_{\tau-1}\rangle_{\Sigma^+} &= (u - r_{\tau-1})^\top\Sigma^+\Sigma F_\mathcal{R}(r_{\tau-1}) \\ &= (u - r_{\tau-1})^\top F_\mathcal{R}(r_{\tau-1}) = \langle F_\mathcal{R}(r_{\tau-1}), u - r_{\tau-1}\rangle.\end{aligned}$$

Hence

$$\langle\widehat{G}_m^{(\tau)}, u - r_{\tau-1}\rangle_{\Sigma^+} = \langle F_\mathcal{R}(r_{\tau-1}), u - r_{\tau-1}\rangle - \langle\Delta_\tau, u - r_{\tau-1}\rangle_{\Sigma^+}. \tag{99}$$

Step 3: bound the squared gradient term. By the triangle inequality,

$$\|\widehat{G}_m^{(\tau)}\|_{\Sigma^+} \le \|G(r_{\tau-1})\|_{\Sigma^+} + \|\Delta_\tau\|_{\Sigma^+}.$$

Using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\|\widehat{G}_m^{(\tau)}\|_{\Sigma^+}^2 \le 2\|G(r_{\tau-1})\|_{\Sigma^+}^2 + 2\|\Delta_\tau\|_{\Sigma^+}^2. \tag{100}$$

For the first term, recall the equilibrium condition (Step 3 of the proof of Theorem 9): $F_{\mathcal{R}}(\theta^\star) \in \ker\Sigma$, so $G(\theta^\star) = \Sigma F_{\mathcal{R}}(\theta^\star) = 0$. By Lemma 9,

$$\|G(r_{\tau-1})\|_{\Sigma^+} = \|G(r_{\tau-1}) - G(\theta^\star)\|_{\Sigma^+} \le L_G\|r_{\tau-1} - \theta^\star\|_{\Sigma^+} \le 2L_G\Omega,$$

where the last inequality uses $r_{\tau-1}, \theta^\star \in D$ and the Σ^+-diameter of D is 2Ω. Hence $\|G(r_{\tau-1})\|_{\Sigma^+}^2 \le 4L_G^2\Omega^2$. Substituting into (100), we have the bound

$$\|\widehat{G}_m^{(\tau)}\|_{\Sigma^+}^2 \le 8L_G^2\Omega^2 + 2\|\Delta_\tau\|_{\Sigma^+}^2. \tag{101}$$

Step 4: substitute and telescope. Substitute (99) and (101) into (98) to get that

$$V(r_\tau, u) - V(r_{\tau-1}, u) \le \gamma\langle F_{\mathcal{R}}(r_{\tau-1}), u - r_{\tau-1}\rangle - \langle\gamma\Delta_\tau, u - r_{\tau-1}\rangle_{\Sigma^+} + 4\gamma^2L_G^2\Omega^2 + \gamma^2\|\Delta_\tau\|_{\Sigma^+}^2.$$

Summing over $\tau = 1, \dots, t$, we have that

$$V(r_t, u) - V(r_0, u) \le \sum_\tau \gamma\langle F_{\mathcal{R}}(r_{\tau-1}), u - r_{\tau-1}\rangle - \sum_\tau\langle\gamma\Delta_\tau, u - r_{\tau-1}\rangle_{\Sigma^+} + 4t\gamma^2L_G^2\Omega^2 + \gamma^2\sum_\tau\|\Delta_\tau\|_{\Sigma^+}^2.$$

Using $V(r_t, u) \ge 0$ and $V(r_0, u) \le \Theta(r_0)$, rearrange to obtain

$$\sum_\tau \gamma\langle F_{\mathcal{R}}(r_{\tau-1}), r_{\tau-1} - u\rangle \le \Theta(r_0) + \sum_\tau\langle\gamma\Delta_\tau, r_{\tau-1} - u\rangle_{\Sigma^+} + 4t\gamma^2L_G^2\Omega^2 + \gamma^2\sum_\tau\|\Delta_\tau\|_{\Sigma^+}^2. \tag{102}$$

Step 5: ghost-iterate decomposition. The quantity $\sum_\tau\langle\gamma\Delta_\tau, r_{\tau-1} - u\rangle_{\Sigma^+}$ on the right of (102) depends on u; to take sup over u uniformly, define the ghost iterate sequence $\{y_\tau\}_{\tau\ge0}$ by

$$y_\tau := \Pi_D(y_{\tau-1} - \gamma\Delta_\tau), \qquad y_0 := r_0, \tag{103}$$

and decompose

$$\sum_\tau\langle\gamma\Delta_\tau, r_{\tau-1} - u\rangle_{\Sigma^+} = \sum_\tau\langle\gamma\Delta_\tau, y_{\tau-1} - u\rangle_{\Sigma^+} + \sum_\tau\langle\gamma\Delta_\tau, r_{\tau-1} - y_{\tau-1}\rangle_{\Sigma^+}. \tag{104}$$

For the first sum on the right, applying Lemma 13 with $\xi_\tau = \gamma\Delta_\tau$ gives

$$\sum_\tau\langle\gamma\Delta_\tau, y_{\tau-1} - u\rangle_{\Sigma^+} \le V(y_0, u) + \tfrac{1}{2}\sum_\tau\|\gamma\Delta_\tau\|_{\Sigma^+}^2 \le \Theta(r_0) + \tfrac{\gamma^2}{2}\sum_\tau\|\Delta_\tau\|_{\Sigma^+}^2,$$

using $y_0 = r_0$ and $V(r_0, u) \le \Theta(r_0)$. Substituting into (102) via (104),

$$\sum_\tau \gamma\langle F_{\mathcal{R}}(r_{\tau-1}), r_{\tau-1} - u\rangle \le 2\Theta(r_0) + \tfrac{\gamma^2}{2}\sum_\tau\|\Delta_\tau\|_{\Sigma^+}^2 + \sum_\tau\langle\gamma\Delta_\tau, r_{\tau-1} - y_{\tau-1}\rangle_{\Sigma^+} + 4t\gamma^2L_G^2\Omega^2 + \gamma^2\sum_\tau\|\Delta_\tau\|_{\Sigma^+}^2.$$

Combine the $\|\Delta_\tau\|_{\Sigma^+}^2$ coefficients: $\frac{\gamma^2}{2} + \gamma^2 = \frac{3\gamma^2}{2}$. Hence

$$\sum_\tau \gamma\langle F_{\mathcal{R}}(r_{\tau-1}), r_{\tau-1} - u\rangle \le 2\Theta(r_0) + \tfrac{3\gamma^2}{2}\sum_\tau\|\Delta_\tau\|_{\Sigma^+}^2 + \sum_\tau\langle\gamma\Delta_\tau, r_{\tau-1} - y_{\tau-1}\rangle_{\Sigma^+} + 4t\gamma^2L_G^2\Omega^2. \tag{105}$$

Step 6: convert to $\mathrm{Gap}_{\mathrm{vi}}$. By **(P-mono)**, for all $u, w \in D$, $\langle F_{\mathcal{R}}(u), w-u\rangle \le \langle F_{\mathcal{R}}(w), w-u\rangle$. Applied with $w = r_{\tau-1}$: $\langle F_{\mathcal{R}}(u), r_{\tau-1}-u\rangle \le \langle F_{\mathcal{R}}(r_{\tau-1}), r_{\tau-1}-u\rangle$. Summing, multiplying by γ, and dividing by $\sum_\tau \gamma_\tau = t\gamma$:

$$\langle F_{\mathcal{R}}(u), \widehat{z}_t - u\rangle \le (t\gamma)^{-1}\sum_\tau \gamma\langle F_{\mathcal{R}}(r_{\tau-1}), r_{\tau-1}-u\rangle.$$

By (105), the right side is bounded above (for every fixed u) by

$$\frac{2\Theta(r_0)}{t\gamma} + \frac{3\gamma}{2t}\sum_\tau \|\Delta_\tau\|_{\Sigma^+}^2 + \frac{1}{t\gamma}\sum_\tau \langle \gamma\Delta_\tau, r_{\tau-1}-y_{\tau-1}\rangle_{\Sigma^+} + 4\gamma L_G^2\Omega^2.$$

The first three terms above contain u-free upper bounds (the third was obtained by sup-over-u via the ghost iterate); taking supremum over $u \in D$,

$$\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t) \le \frac{2\Theta(r_0)}{t\gamma} + \frac{3\gamma}{2t}\sum_\tau \|\Delta_\tau\|_{\Sigma^+}^2 + \frac{1}{t\gamma}\sum_\tau \langle \gamma\Delta_\tau, r_{\tau-1}-y_{\tau-1}\rangle_{\Sigma^+} + 4\gamma L_G^2\Omega^2. \quad (106)$$

Step 7: take expectation. Take expectations of (106).

Bound 1: $\mathbb{E}\,\|\Delta_\tau\|_{\Sigma^+}^2$. By (96) and the tower property, $\mathbb{E}\,\|\Delta_\tau\|_{\Sigma^+}^2 \le \widetilde{B}_m^2 + \widetilde{V}_m$. Summing: $\sum_\tau \mathbb{E}\,\|\Delta_\tau\|_{\Sigma^+}^2 \le t(\widetilde{B}_m^2 + \widetilde{V}_m)$. Multiplying by $3\gamma/(2t)$: $\le \frac{3\gamma}{2}(\widetilde{B}_m^2 + \widetilde{V}_m)$.

Bound 2: $\mathbb{E}\langle \gamma\Delta_\tau, r_{\tau-1}-y_{\tau-1}\rangle_{\Sigma^+}$. Both $r_{\tau-1}$ (by definition) and $y_{\tau-1}$ (which depends only on $\Delta_1, \dots, \Delta_{\tau-1}$ via the ghost recursion (103), and these are $\mathcal{F}_{\tau-1}$-measurable) are $\mathcal{F}_{\tau-1}$-measurable. By the tower property,

$$\mathbb{E}\langle \gamma\Delta_\tau, r_{\tau-1}-y_{\tau-1}\rangle_{\Sigma^+} = \gamma\,\mathbb{E}\langle \mathbb{E}[\Delta_\tau \mid \mathcal{F}_{\tau-1}], r_{\tau-1}-y_{\tau-1}\rangle_{\Sigma^+}.$$

By (95), $\|\,\mathbb{E}[\Delta_\tau \mid \mathcal{F}_{\tau-1}]\|_{\Sigma^+} \le \widetilde{B}_m$. Cauchy–Schwarz in Σ^+ then gives

$$|\langle \mathbb{E}[\Delta_\tau \mid \mathcal{F}_{\tau-1}], r_{\tau-1}-y_{\tau-1}\rangle_{\Sigma^+}| \le \widetilde{B}_m\|r_{\tau-1}-y_{\tau-1}\|_{\Sigma^+} \le 2\widetilde{B}_m\Omega,$$

since $y_{\tau-1}, r_{\tau-1} \in D$ and the Σ^+-diameter of D is 2Ω. Hence $|\,\mathbb{E}\langle \gamma\Delta_\tau, r_{\tau-1}-y_{\tau-1}\rangle_{\Sigma^+}| \le 2\gamma\widetilde{B}_m\Omega$. Summing over τ and dividing by $t\gamma$: $\le 2\widetilde{B}_m\Omega$.

Step 8: combine. Combining the bounds,

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \frac{2\Theta(r_0)}{t\gamma} + \tfrac{3\gamma}{2}(\widetilde{B}_m^2 + \widetilde{V}_m) + 2\widetilde{B}_m\Omega + 4\gamma L_G^2\Omega^2.$$

This is (97). □

Corollary 4 (Optimal step size, mirror descent monotone case)**.** Choosing

$$\gamma = \min\left\{\frac{1}{2L_G},\ \sqrt{\frac{2\Theta(r_0)}{t\left[\frac{3}{2}(\widetilde{B}_m^2 + \widetilde{V}_m) + 4L_G^2\Omega^2\right]}}\right\} \quad (107)$$

in Corollary 3 gives

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \max\left\{\frac{8L_G\Theta(r_0)}{t},\ 2\sqrt{\frac{2\Theta(r_0)\left[\frac{3}{2}(\widetilde{B}_m^2 + \widetilde{V}_m) + 4L_G^2\Omega^2\right]}{t}}\right\} + 2\Omega\widetilde{B}_m. \quad (108)$$

Proof. Write $Q := \frac{3}{2}(\widetilde{B}_m^2 + \widetilde{V}_m) + 4L_G^2\Omega^2$. The terms depending on γ in (97) are $f(\gamma) := 2\Theta(r_0)/(t\gamma) + \gamma Q$. Their unconstrained minimizer is $\sqrt{2\Theta(r_0)/(tQ)}$. If this is at most $1/(2L_G)$, then $f(\gamma) = 2\sqrt{2\Theta(r_0)Q/t}$. Otherwise, $Q \le 8L_G^2\Theta(r_0)/t$, so the capped step gives

$$f(\gamma) = \frac{4L_G\Theta(r_0)}{t} + \frac{Q}{2L_G} \le \frac{8L_G\Theta(r_0)}{t}.$$

Adding $2\Omega\widetilde{B}_m$ gives (108). □

Initialization at the reference policy. For $r_0 = z_c = \theta_{\mathrm{ref}}$, we have $2\Theta(r_0) = \Omega^2$. Thus (97) becomes

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \frac{\Omega^2}{t\gamma} + \frac{3\gamma}{2}(\widetilde{B}_m^2 + \widetilde{V}_m) + 4\gamma L_G^2\Omega^2 + 2\Omega\widetilde{B}_m.$$

The step size (107) specializes by replacing $2\Theta(r_0)$ with Ω^2, and the optimized bound is

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] \le \max\left\{\frac{4L_G\Omega^2}{t},\ 2\Omega\sqrt{\frac{\frac{3}{2}(\widetilde{B}_m^2 + \widetilde{V}_m) + 4L_G^2\Omega^2}{t}}\right\} + 2\Omega\widetilde{B}_m.$$

Remark (comparison with two-call Mirror-Prox). The preceding one-call bound has $t^{-1/2}$ dependence in both its deterministic and sampling terms. With two oracle calls per iteration, Corollary 1 instead gives, for $r_0 = \theta_{\mathrm{ref}}$,

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t)] = \mathcal{O}\left(\frac{L_G\Omega^2}{t} + \Omega\sqrt{\frac{\widetilde{B}_m^2 + \widetilde{V}_m}{t}} + \Omega\widetilde{B}_m\right).$$

In particular, with an exact oracle, the constant step size $\gamma = 1/(\sqrt{3}L_G)$ gives $\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_t) \le \sqrt{3}L_G\Omega^2/t$ by Theorem 8, recovering the deterministic t^{-1} rate of Mirror-Prox (Juditsky et al., 2011).

H.10.2 Strongly monotone case

Corollary 5 (Stochastic Mirror Descent, strongly monotone, biased oracle)**.** Suppose $F_{\mathcal{R}}$ satisfies **(P-strong)** and **(P-lip)**, and the oracle satisfies (O1) and (O2). Run the mirror descent algorithm (93) with constant step size $\gamma_\tau \equiv \gamma$ satisfying

$$0 < \gamma \le \min\left\{1/(2\widetilde{\mu}),\ 1/(2L_G)\right\}, \qquad \widetilde{\mu} = \mu_{\mathcal{R}}\sigma_{\min}. \tag{109}$$

Then for all $T \ge 1$, the estimate holds:

$$\mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 \le (1-\gamma\widetilde{\mu})^T\|r_0 - \theta^\star\|_{\Sigma^+}^2 + \frac{4\Omega\widetilde{B}_m}{\widetilde{\mu}} + \frac{2\gamma\big(\widetilde{B}_m^2 + \widetilde{V}_m\big)}{\widetilde{\mu}} + \frac{8L_G^2\Omega^2\gamma}{\widetilde{\mu}}. \tag{110}$$

Proof. We work step by step from the prox-mapping descent inequality (Lemma 11), specialized to $u = \theta^\star$ and combined with strong monotonicity.

First we apply the descent identity at $u = \theta^\star$. Indeed, apply (61) of Lemma 11 with $z = r_{\tau-1}$, $\xi = \gamma\widehat{G}_m^{(\tau)}$, $w = r_\tau$, and $u = \theta^\star$:

$$V(r_\tau, \theta^\star) \le V(r_{\tau-1}, \theta^\star) + \gamma\langle\widehat{G}_m^{(\tau)}, \theta^\star - r_{\tau-1}\rangle_{\Sigma^+} + \tfrac{\gamma^2}{2}\|\widehat{G}_m^{(\tau)}\|_{\Sigma^+}^2. \tag{111}$$

Decompose $\widehat{G}_m^{(\tau)} = G(r_{\tau-1}) + \Sigma b_m(r_{\tau-1}) + \Sigma\zeta_m^{(\tau)}$ by (94). Then we have that

$$\begin{aligned}\langle\widehat{G}_m^{(\tau)}, \theta^\star - r_{\tau-1}\rangle_{\Sigma^+} &= \langle G(r_{\tau-1}), \theta^\star - r_{\tau-1}\rangle_{\Sigma^+} \\ &\quad + \langle\Sigma b_m(r_{\tau-1}), \theta^\star - r_{\tau-1}\rangle_{\Sigma^+} + \langle\Sigma\zeta_m^{(\tau)}, \theta^\star - r_{\tau-1}\rangle_{\Sigma^+}.\end{aligned}$$

Using Lemma 8(i), for $w \in \mathcal{W}$ and arbitrary $v \in \mathbb{R}^d$, we have that $\langle\Sigma v, w\rangle_{\Sigma^+} = \langle v, w\rangle$. Apply this with $w = \theta^\star - r_{\tau-1} \in \mathcal{W}$ to each term to get that

$$\langle\widehat{G}_m^{(\tau)}, \theta^\star - r_{\tau-1}\rangle_{\Sigma^+} = \langle F_{\mathcal{R}}(r_{\tau-1}), \theta^\star - r_{\tau-1}\rangle + \langle b_m(r_{\tau-1}), \theta^\star - r_{\tau-1}\rangle + \langle\zeta_m^{(\tau)}, \theta^\star - r_{\tau-1}\rangle. \tag{112}$$

The risk-adjusted equilibrium condition gives $F_{\mathcal{R}}(\theta^\star) \in \ker\Sigma$. Since $\theta^\star - r_{\tau-1} \in \mathcal{W}$ and $\ker\Sigma = \mathcal{W}^\perp$, $\langle F_{\mathcal{R}}(\theta^\star), \theta^\star - r_{\tau-1}\rangle = 0$. Hence

$$\langle F_{\mathcal{R}}(r_{\tau-1}), \theta^\star - r_{\tau-1}\rangle = \langle F_{\mathcal{R}}(r_{\tau-1}) - F_{\mathcal{R}}(\theta^\star), \theta^\star - r_{\tau-1}\rangle = -\langle F_{\mathcal{R}}(r_{\tau-1}) - F_{\mathcal{R}}(\theta^\star), r_{\tau-1} - \theta^\star\rangle.$$

By **(P-strong)** and (50), we have the bound

$$\langle F_{\mathcal{R}}(r_{\tau-1}) - F_{\mathcal{R}}(\theta^\star), r_{\tau-1} - \theta^\star\rangle \ge \mu_{\mathcal{R}}\|r_{\tau-1} - \theta^\star\|_2^2 \ge \widetilde{\mu}\|r_{\tau-1} - \theta^\star\|_{\Sigma^+}^2.$$

Thus

$$\langle F_{\mathcal{R}}(r_{\tau-1}), \theta^\star - r_{\tau-1}\rangle \le -\widetilde{\mu}\|r_{\tau-1} - \theta^\star\|_{\Sigma^+}^2 = -2\widetilde{\mu}V(r_{\tau-1}, \theta^\star). \tag{113}$$

Substituting (113) into (112) yields

$$\langle \widehat{G}_m^{(\tau)}, \theta^\star - r_{\tau-1}\rangle_{\Sigma^+} \le -2\widetilde{\mu}V(r_{\tau-1}, \theta^\star) - \langle b_m(r_{\tau-1}), r_{\tau-1} - \theta^\star\rangle - \langle \zeta_m^{(\tau)}, r_{\tau-1} - \theta^\star\rangle. \tag{114}$$

By Step 3 of the proof of Corollary 3 (eq. (101)), we immediately have

$$\|\widehat{G}_m^{(\tau)}\|_{\Sigma^+}^2 \le 8L_G^2\Omega^2 + 2\|\Delta_\tau\|_{\Sigma^+}^2. \tag{115}$$

Substitute (114) and (115) into (111) to get that

$$\begin{aligned} V(r_\tau, \theta^\star) \le V(r_{\tau-1}, \theta^\star) - 2\gamma\widetilde{\mu}V(r_{\tau-1}, \theta^\star) - \gamma\langle b_m(r_{\tau-1}), r_{\tau-1} - \theta^\star\rangle \\ - \gamma\langle \zeta_m^{(\tau)}, r_{\tau-1} - \theta^\star\rangle + 4\gamma^2 L_G^2\Omega^2 + \gamma^2\|\Delta_\tau\|_{\Sigma^+}^2. \end{aligned} \tag{116}$$

Take $\mathbb{E}[\cdot \mid \mathcal{F}_{\tau-1}]$ of (116). Recall the filtration: $r_{\tau-1}$ is $\mathcal{F}_{\tau-1}$-measurable; $\zeta_m^{(\tau)}$ has conditional mean zero given $\mathcal{F}_{\tau-1}$; Δ_τ depends on $\zeta_m^{(\tau)}$ via (94). Now we bound each of the terms as follows:

- *Term 1: contraction.* The first term $V(r_{\tau-1}, \theta^\star)$ is $\mathcal{F}_{\tau-1}$-measurable, so the term passes through the conditional expectation unchanged.
- *Term 2: bias cross.* The bias term $b_m(r_{\tau-1})$ is deterministic given $r_{\tau-1}$, hence given $\mathcal{F}_{\tau-1}$. Thus

  $$\mathbb{E}[\langle b_m(r_{\tau-1}), r_{\tau-1} - \theta^\star\rangle \mid \mathcal{F}_{\tau-1}] = \langle b_m(r_{\tau-1}), r_{\tau-1} - \theta^\star\rangle.$$

  Bound the bias via Cauchy–Schwarz, (50). The Σ^+-diameter of D is 2Ω so that

  $$\begin{aligned} |\langle b_m(r_{\tau-1}), r_{\tau-1} - \theta^\star\rangle| &\le \|b_m(r_{\tau-1})\|_2\|r_{\tau-1} - \theta^\star\|_2 \\ &\le B_m \cdot \sqrt{\sigma_{\max}}\|r_{\tau-1} - \theta^\star\|_{\Sigma^+} \\ &\le \widetilde{B}_m \cdot 2\Omega, \end{aligned}$$

  using (O1) ($\|b_m\|_2 \le B_m$), $\widetilde{B}_m = \sqrt{\sigma_{\max}}B_m$.
- *Term 3: noise cross.* We have $\mathbb{E}[\zeta_m^{(\tau)} \mid \mathcal{F}_{\tau-1}] = 0$ by fresh sampling. Thus

  $$\mathbb{E}[\langle \zeta_m^{(\tau)}, r_{\tau-1} - \theta^\star\rangle \mid \mathcal{F}_{\tau-1}] = \langle \mathbb{E}[\zeta_m^{(\tau)} \mid \mathcal{F}_{\tau-1}], r_{\tau-1} - \theta^\star\rangle = 0.$$

- *Term 4:* $\|\Delta_\tau\|_{\Sigma^+}^2$. By (96), we have that $\mathbb{E}[\|\Delta_\tau\|_{\Sigma^+}^2 \mid \mathcal{F}_{\tau-1}] \le \widetilde{B}_m^2 + \widetilde{V}_m$.
- *Term 5: deterministic.* The last term is deterministic so that $4\gamma^2 L_G^2\Omega^2$ passes through.

Combining each of these bounds yields

$$\mathbb{E}[V(r_\tau, \theta^\star) \mid \mathcal{F}_{\tau-1}] \le (1 - 2\gamma\widetilde{\mu})V(r_{\tau-1}, \theta^\star) + 2\gamma\Omega\widetilde{B}_m + 4\gamma^2 L_G^2\Omega^2 + \gamma^2(\widetilde{B}_m^2 + \widetilde{V}_m).$$

The step-size constraint $\gamma \le 1/(2\widetilde{\mu})$ in (109) gives $2\gamma\widetilde{\mu} \le 1$, so the contraction coefficient is non-negative. Take unconditional expectation, we have that

$$\mathbb{E}\, V(r_\tau, \theta^\star) \le (1 - 2\gamma\widetilde{\mu})\, \mathbb{E}\, V(r_{\tau-1}, \theta^\star) + 2\gamma\Omega\widetilde{B}_m + 4\gamma^2 L_G^2\Omega^2 + \gamma^2(\widetilde{B}_m^2 + \widetilde{V}_m).$$

Since $1 - 2\gamma\widetilde{\mu} \le 1 - \gamma\widetilde{\mu}$ and $V(r_{\tau-1}, \theta^\star) \ge 0$, we may weaken the recurrence to

$$\mathbb{E}\, V(r_\tau, \theta^\star) \le (1 - \gamma\widetilde{\mu})\, \mathbb{E}\, V(r_{\tau-1}, \theta^\star) + 2\gamma\Omega\widetilde{B}_m + 4\gamma^2 L_G^2\Omega^2 + \gamma^2(\widetilde{B}_m^2 + \widetilde{V}_m). \tag{117}$$

We use (117) below; this only weakens the bound.

Next we need to telescope. Define

$$A := 2\gamma\Omega\widetilde{B}_m + 4\gamma^2 L_G^2\Omega^2 + \gamma^2(\widetilde{B}_m^2 + \widetilde{V}_m), \qquad \Lambda_\tau := \mathbb{E}\, V(r_\tau, \theta^\star).$$

The recurrence (117) reads $\Lambda_\tau \le (1-\gamma\widetilde{\mu})\Lambda_{\tau-1} + A$. Iterating, we have that

$$\begin{aligned}\Lambda_T &\le (1-\gamma\widetilde{\mu})^T\Lambda_0 + A\sum_{j=0}^{T-1}(1-\gamma\widetilde{\mu})^j \\ &\le (1-\gamma\widetilde{\mu})^T\Lambda_0 + \frac{A}{\gamma\widetilde{\mu}} \\ &= (1-\gamma\widetilde{\mu})^T V(r_0,\theta^\star) + \frac{2\Omega\widetilde{B}_m}{\widetilde{\mu}} + \frac{4\gamma L_G^2\Omega^2}{\widetilde{\mu}} + \frac{\gamma(\widetilde{B}_m^2+\widetilde{V}_m)}{\widetilde{\mu}}.\end{aligned}$$

Using $V(z,u) = \frac{1}{2}\|z-u\|_{\Sigma^+}^2$ yields the bound

$$\begin{aligned}\mathbb{E}\,\|r_T-\theta^\star\|_{\Sigma^+}^2 &= 2\,\mathbb{E}\,V(r_T,\theta^\star) \\ &\le (1-\gamma\widetilde{\mu})^T\|r_0-\theta^\star\|_{\Sigma^+}^2 + \frac{4\Omega\widetilde{B}_m}{\widetilde{\mu}} + \frac{8\gamma L_G^2\Omega^2}{\widetilde{\mu}} + \frac{2\gamma(\widetilde{B}_m^2+\widetilde{V}_m)}{\widetilde{\mu}},\end{aligned}$$

which is (110). □

Comparison with extra-gradient. Comparing Corollary 5 to Theorem 9, the mirror descent bound has the same linear-contraction rate and bias-floor structure, with smaller variance constant ($2/\widetilde{\mu}$ vs. $6/\widetilde{\mu}$) since each mirror descent iteration only invokes the oracle once. However, mirror descent picks up an additional Lipschitz-times-step-size term $8L_G^2\Omega^2\gamma/\widetilde{\mu}$ that does not appear in the extragradient bound. The extragradient proof controls the deterministic contribution through the non-positive coefficient of $\|w_\tau - r_{\tau-1}\|_{\Sigma^+}^2$ under its step-size condition. The mirror descent proof above instead uses $\|G(r_{\tau-1})\|_{\Sigma^+} \le 2L_G\Omega$. The extra term therefore reflects this estimate; it does not establish an unavoidable error floor for mirror descent. It persists in the displayed upper bound as $m\to\infty$ at fixed γ, and decreases with γ.

Corollary 6 (Optimal step size, mirror descent strongly monotone case)**.** With sample budget m fixed and target accuracy $\delta > 0$, choose

$$\gamma = \min\left\{\frac{1}{2\widetilde{\mu}},\ \frac{1}{2L_G},\ \frac{\widetilde{\mu}\delta}{4\big[(\widetilde{B}_m^2+\widetilde{V}_m)+4L_G^2\Omega^2\big]}\right\}, \tag{118}$$

and, if $2\|r_0-\theta^\star\|_{\Sigma^+}^2 > \delta$, run for

$$T = \left\lceil \frac{1}{\gamma\widetilde{\mu}}\log\frac{2\|r_0-\theta^\star\|_{\Sigma^+}^2}{\delta}\right\rceil \tag{119}$$

iterations; otherwise take $T=1$. Then

$$\mathbb{E}\,\|r_T-\theta^\star\|_{\Sigma^+}^2 \le \delta + \frac{4\Omega\widetilde{B}_m}{\widetilde{\mu}}.$$

Proof. Recall the bound (110) of Corollary 5:

$$\mathbb{E}\,\|r_T-\theta^\star\|_{\Sigma^+}^2 \le \underbrace{(1-\gamma\widetilde{\mu})^T\|r_0-\theta^\star\|_{\Sigma^+}^2}_{\text{contraction}} + \underbrace{\frac{4\Omega\widetilde{B}_m}{\widetilde{\mu}}}_{\text{bias floor}} + \underbrace{\frac{2\gamma(\widetilde{B}_m^2+\widetilde{V}_m)}{\widetilde{\mu}}}_{\text{variance floor}} + \underbrace{\frac{8L_G^2\Omega^2\gamma}{\widetilde{\mu}}}_{\text{Lipschitz floor}}.$$

We bound each of these four pieces in turn under the choices (118) and (119). The first two entries in the minimum (118) (namely $1/(2\widetilde{\mu})$ and $1/(2L_G)$) ensure γ is admissible under (109), so Corollary 5 applies.

Contraction. If $2\|r_0-\theta^\star\|_{\Sigma^+}^2 \le \delta$, the contraction term is already at most $\delta/2$ for $T=1$. Otherwise, with T as in (119), using $1-x\le e^{-x}$,

$$\begin{aligned}(1-\gamma\widetilde{\mu})^T\|r_0-\theta^\star\|_{\Sigma^+}^2 \le e^{-\gamma\widetilde{\mu}T}\|r_0-\theta^\star\|_{\Sigma^+}^2 &\le e^{-\log(2\|r_0-\theta^\star\|_{\Sigma^+}^2/\delta)}\|r_0-\theta^\star\|_{\Sigma^+}^2 \\ &= \frac{\delta}{2\|r_0-\theta^\star\|_{\Sigma^+}^2}\cdot\|r_0-\theta^\star\|_{\Sigma^+}^2 = \frac{\delta}{2}.\end{aligned}$$

Bias floor. The bias floor $4\Omega\widetilde{B}_m/\widetilde{\mu}$ is independent of γ and T and persists in the bound.

Variance and Lipschitz floors. These two terms combine as

$$\frac{2\gamma(\widetilde{B}_m^2+\widetilde{V}_m)}{\widetilde{\mu}}+\frac{8L_G^2\Omega^2\gamma}{\widetilde{\mu}}=\frac{2\gamma}{\widetilde{\mu}}\big[(\widetilde{B}_m^2+\widetilde{V}_m)+4L_G^2\Omega^2\big].$$

By the third entry in the minimum (118), $\gamma\le\widetilde{\mu}\delta/\big(4\big[(\widetilde{B}_m^2+\widetilde{V}_m)+4L_G^2\Omega^2\big]\big)$, so substituting gives

$$\frac{2\gamma}{\widetilde{\mu}}\big[(\widetilde{B}_m^2+\widetilde{V}_m)+4L_G^2\Omega^2\big]\le\frac{2}{\widetilde{\mu}}\cdot\frac{\widetilde{\mu}\delta}{4\big[(\widetilde{B}_m^2+\widetilde{V}_m)+4L_G^2\Omega^2\big]}\cdot\big[(\widetilde{B}_m^2+\widetilde{V}_m)+4L_G^2\Omega^2\big]=\frac{\delta}{2}.$$

Combining the bounds. Summing the four pieces yields

$$\mathbb{E}\,\|r_T-\theta^\star\|_{\Sigma^+}^2\le\underbrace{\frac{\delta}{2}}_{\text{contraction}}+\underbrace{\frac{4\Omega\widetilde{B}_m}{\widetilde{\mu}}}_{\text{bias floor}}+\underbrace{\frac{\delta}{2}}_{\text{var.+Lip. floors}}=\delta+\frac{4\Omega\widetilde{B}_m}{\widetilde{\mu}}.$$ □

I Stackelberg Game and Two-Timescale Convergence

The bias floor $4\Omega\widetilde{B}_m/\widetilde{\mu}$ in Theorem 9 is irreducible without further structure on the oracle: with constant step sizes η,γ and constant sample budget m, the floor scales as $\widetilde{B}_m\asymp 1/m$ regardless of how long we run. In this section we show that with a sample-based bias estimator $\widehat{b}_m(\theta)$ and *time-varying* step sizes $\eta_t\to 0$, $\gamma_t\to 0$ on two separated timescales, the bias contribution to the convergence bound *vanishes asymptotically in* t, thereby replacing the persistent $\mathcal{O}(1/m)$ floor with an $\mathcal{O}(1/m^2)$ residual from the imperfect quality of the bias estimator itself. We formulate the problem as a Stackelberg game between a follower aiming to optimize $\frac{1}{2}\|\xi_t-\hat{b}_m(\theta_t)\|^2$ and the leader(s) aiming to find the optimal equilibrium $\theta^\star$ for the risk-sensitive preference game. We describe the Stackelberg game in full detail in Appendix I.3. The general idea is that imposing this hierarchical structure and allowing the follower to run the bias tracking update on a faster time scale via a Robin-Monroe type update, leads to a Stackelberg equilibrium and therefore an equilibrium that that is within the maximum of the bias floor of the bias estimator (residual bias floor) and the variance. As we prove these two components asymptotically have the same order in samples $\mathcal{O}(1/m^2)$, and otherwise (i.e. with finite step sizes) $\mathcal{O}(1/m^{3/2})$.

If additionally $m\to\infty$ along the trajectory (e.g., $m_t\to\infty$ on a separate schedule), the residual goes to zero and the algorithm converges to the equilibrium with no asymptotic bias.

I.1 Bias estimator and assumptions

Recall the bias convention from (38):

$$b_m(\theta):=\mathbb{E}[\widehat{F}_{\mathcal{R},m}(\theta)]-F_{\mathcal{R}}(\theta).$$

Thus $\widehat{b}_m$ estimates the bias of the operator $F_{\mathcal{R}}$, and the tracker ξ is subtracted from the sampled operator: $\widehat{F}_{\mathcal{R},m}-\xi$. Since $F_{\mathcal{R}}$ contains $-\mathcal{P}_{\mathcal{R}}$, this convention gives the negative of the payoff-estimation bias.

We require, in addition to (O1)–(O2) on the gradient oracle, the following on a sample-based bias estimator $\widehat{b}_m(\theta)\in\mathbb{R}^d$:

- (B1) **Approximate unbiasedness.** There exists $R_m\ge 0$ such that $\|\,\mathbb{E}[\widehat{b}_m(\theta)]-b_m(\theta)\|_2\le R_m$ for all $\theta\in D$.
- (B2) **Variance bound.** There exists $V_m^b\ge 0$ such that $\mathbb{E}\,\|\widehat{b}_m(\theta)-\mathbb{E}[\widehat{b}_m(\theta)]\|_2^2\le V_m^b$.
- (B3) **Lipschitz dependence on θ.** The population bias is Lipschitz continuous—i.e., $\|b_m(\theta_1)-b_m(\theta_2)\|_2\le L_b\|\theta_1-\theta_2\|_2$ for all $\theta_1,\theta_2\in D$.

The first two conditions concern the quality of the particular sample-based bias estimator, whereas (B3) is a regularity condition on the population bias b_m itself. We next give two bias estimators satisfying (B1)–(B2) for the smooth plug-in risks used in our experiments.

Smooth plug-in risks. Many of the risk functionals of interest—e.g., entropic risk—are of the form $\mathcal{R}[Z] = h(\mathbb{E}_\mu[g(Z)])$ for a smooth scalar function h and bounded statistic g. Fix a response y, draw $Y_1, \dots, Y_m \overset{\text{iid}}{\sim} \mu = \pi_\theta$, and set $Z_i := P(y \succ Y_i)$. For $m \geq 2$, let

$$X_i := g(Z_i), \qquad \widehat{q}_m := \frac{1}{m}\sum_{i=1}^{m} X_i, \qquad q := \mathbb{E}_\mu[X],$$

and write $\sigma^2 := \mathrm{Var}_\mu(X)$. We assume throughout this paragraph that X is uniformly bounded and that h has uniformly bounded derivatives through fourth order on an interval containing the range of X. The scalar derivations below apply componentwise to the risk-adjusted operator; we suppress the component index y in the bias estimators below.

A Taylor expansion around q yields

$$h(\widehat{q}_m) = h(q) + h'(q)(\widehat{q}_m - q) + \frac{1}{2}h''(q)(\widehat{q}_m - q)^2 + \frac{1}{6}h'''(q)(\widehat{q}_m - q)^3 + \mathcal{O}\big((\widehat{q}_m - q)^4\big).$$

Using $\mathbb{E}[\widehat{q}_m - q] = 0$ and $\mathbb{E}[(\widehat{q}_m - q)^2] = \frac{\sigma^2}{m}$ together with $\mathbb{E}[(\widehat{q}_m - q)^3] = \mathcal{O}(m^{-2})$ and $\mathbb{E}[(\widehat{q}_m - q)^4] = \mathcal{O}(m^{-2})$, we obtain

$$\mathbb{E}[h(\widehat{q}_m)] - h(q) = \frac{h''(q)\sigma^2}{2m} + \mathcal{O}(m^{-2}). \tag{120}$$

Thus the plug-in estimator has an $\mathcal{O}(1/m)$ bias whose leading term is determined by the curvature of h. Since the payoff enters $F_\mathcal{R}$ with a minus sign, the operator bias defined in (38) has the opposite sign:

$$(b_m(\theta))_y = h(q) - \mathbb{E}[h(\widehat{q}_m)] = -\frac{h''(q)\sigma^2}{2m} + \mathcal{O}(m^{-2}).$$

The delta-method bias estimator. Denote the usual unbiased sample variance

$$\widehat{\mathrm{Var}}_m(X) := \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \widehat{q}_m)^2.$$

Motivated by (120), define

$$\widehat{b}^{\Delta}_m(\theta) := -\frac{h''(\widehat{q}_m)\widehat{\mathrm{Var}}_m(X)}{2m}. \tag{121}$$

This estimator is computable from the same samples used by the gradient oracle. Under the boundedness and smoothness conditions above, we have that

$$\mathbb{E}[\widehat{b}^{\Delta}_m(\theta)] - b_m(\theta) = \mathcal{O}(m^{-2}) \quad \text{and} \quad \mathrm{Var}(\widehat{b}^{\Delta}_m(\theta)) = \mathcal{O}(m^{-3}),$$

componentwise, and hence (B1)–(B2) hold with $R_m = \mathcal{O}(m^{-2})$ and $V^b_m = \mathcal{O}(m^{-3})$. The m^{-3} variance scaling follows from the fact that $h''(\widehat{q}_m)\widehat{\mathrm{Var}}_m(X)$ has $\mathcal{O}(m^{-1})$ variance; multiplication by $-1/(2m)$ in (121) scales this variance by $1/(4m^2)$.

The jackknife bias estimator. Alternatively, define the leave-one-out sample means

$$\widehat{q}_{m,-i} := \frac{1}{m-1}\sum_{j\neq i} X_j, \qquad i = 1, \dots, m,$$

and the jackknife estimate of the operator bias

$$\widehat{b}^{\mathrm{JK}}_m(\theta) := (m-1)\left(h(\widehat{q}_m) - \frac{1}{m}\sum_{i=1}^{m} h(\widehat{q}_{m,-i})\right).$$

Since $\widehat{q}_{m,-i} - \widehat{q}_m = -(X_i - \widehat{q}_m)/(m-1)$ and $\sum_{i=1}^{m}(\widehat{q}_{m,-i} - \widehat{q}_m) = 0$, taking the Taylor expansion of each $h(\widehat{q}_{m,-i})$ around $\widehat{q}_m$ yields

$$\widehat{b}^{\mathrm{JK}}_m(\theta) = -\frac{h''(\widehat{q}_m)}{2m}\widehat{\mathrm{Var}}_m(X) + \mathcal{O}(m^{-2}) = \widehat{b}^{\Delta}_m(\theta) + \mathcal{O}(m^{-2}).$$

The remainder is uniformly $\mathcal{O}(m^{-2})$: each deleted-mean displacement is $\mathcal{O}(m^{-1})$, and h''' is uniformly bounded. Thus the delta-method and jackknife corrections agree to first order. Under the same boundedness and smoothness conditions,

$$\mathbb{E}[\widehat{b}_m^{\mathrm{JK}}(\theta)] - b_m(\theta) = \mathcal{O}(m^{-2}) \quad \text{and} \quad \mathrm{Var}(\widehat{b}_m^{\mathrm{JK}}(\theta)) = \mathcal{O}(m^{-3}).$$

Consequently, the jackknife estimator also satisfies (B1)–(B2) with $R_m = \mathcal{O}(m^{-2})$ and $V_m^b = \mathcal{O}(m^{-3})$.

In the remainder, $\widehat{b}_m$ denotes either $\widehat{b}_m^{\Delta}$ or $\widehat{b}_m^{\mathrm{JK}}$; the convergence analysis depends only on (B1)–(B3) and is therefore unchanged for either choice.

I.1.1 The Two-timescale extragradient algorithm with time-varying steps

Algorithm 1 admits a clean game-theoretic reading as a Stackelberg game in which the bias-tracker ξ is a follower and the two NLHF self-play agents jointly act as a (two-headed) leader. Following the convention that the leader sits in the outer game and the follower in the inner game, the algorithm has:

- an **outer game** on θ: standard symmetric two-player NLHF self-play on the KL-regularized risk-adjusted game, with no performative structure—it is a regular two-player Nash game.
- an **inner game** on ξ given θ: the bias-tracker best-responds to the leader's current θ, observing it as a parameter.

The Stackelberg structure is between these two games: leader commits in the outer game, follower reacts in the inner game.

Algorithm 1 TT−EG: Two-timescale extragradient with bias tracking

Input: initial $r_0 \in D$, slow steps $\{\eta_t\}$, fast steps $\{\gamma_t\}$, sample budget m
$\xi_0 \leftarrow 0$
for $\tau = 1, \dots, T$ **do**
  Sample fresh batch 1 at $r_{\tau-1}$: $\widehat{F}_m^{(\tau,1)}, \widehat{b}_m^{(\tau,1)}$
  $w_\tau \leftarrow \Pi_D\big(r_{\tau-1} - \eta_\tau\, \Sigma\,(\widehat{F}_m^{(\tau,1)} - \xi_{\tau-1})\big)$
  Sample fresh batch 2 at w_τ: $\widehat{F}_m^{(\tau,2)}, \widehat{b}_m^{(\tau,2)}$
  $r_\tau \leftarrow \Pi_D\big(r_{\tau-1} - \eta_\tau\, \Sigma\,(\widehat{F}_m^{(\tau,2)} - \xi_{\tau-1})\big)$
  $\xi_\tau \leftarrow (1-\gamma_\tau)\xi_{\tau-1} + \gamma_\tau\, \widehat{b}_m^{(\tau,2)}$
return r_T

For this algorithm, extend the histories from Section H.5 to include the bias-estimator observations: $\mathcal{G}_\tau$ is the full history through iteration τ, and $\mathcal{F}_\tau$ is the full history just before sampling batch 2 of iteration τ. Thus w_τ and $\xi_{\tau-1}$ are $\mathcal{F}_\tau$-measurable. Each batch uses fresh samples from its queried policy, so the oracle and bias-estimator bounds hold conditionally on the history before that batch.

The conditional mean of the (debiased) correction-step gradient is $\mathbb{E}[\widehat{F}_m^{(\tau,2)} - \xi_{\tau-1} \mid \mathcal{F}_\tau] = F_{\mathcal{R}}(w_\tau) - e_{\tau-1}^*$ where $e_{\tau-1}^* := \xi_{\tau-1} - b_m(w_\tau)$ is the tracking error. Driving $e_\tau^* \to 0$ in mean square eliminates the bias of the debiased gradient.

I.1.2 Convergence theorem

Define the mean-square tracking error $V_t := \mathbb{E}\,\|e_{t-1}^*\|_2^2$. The tracking proof also uses the error at the slow iterate:

$$e_t := \xi_t - b_m(r_t) \quad (t \geq 0), \qquad U_t := \mathbb{E}\,\|e_{t-1}\|_2^2 \quad (t \geq 1).$$

Theorem 10 (Two-timescale debiasing for stochastic extra-gradient, strongly-mono case)**.** Suppose $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone and $\ell_{\mathcal{R}}$-Lipschitz, the gradient oracle satisfies (O1)-(O2), and the bias estimator satisfies (B1)-(B3). Let $\{\eta_t, \gamma_t\}_{t\geq 1}$ be deterministic, predictable sequences satisfying:

$$0 < \eta_t \leq \min\left\{\frac{1}{4\widetilde{\mu}}, \frac{1}{\sqrt{6}L_G}\right\}, \quad \gamma_t \in (0,1], \quad \frac{\eta_t}{\gamma_t} \to 0, \quad \gamma_t \to 0, \quad \sum_{t\geq 1}\eta_t = \infty. \tag{122}$$

Assume the gradient norm is bounded along the trajectory: $\|\widehat{F}_m^{(\tau,i)} - \xi_{\tau-1}\|_2 \le G$ uniformly.

i. **Fast-timescale tracking.** The error U_t satisfies the non-asymptotic recurrence

$$U_{t+1} \le (1 - \gamma_t/2)\, U_t + \frac{C_1\, \eta_t^2}{\gamma_t} + C_2\, \gamma_t\, R_m^2 + C_3\, \gamma_t^2\, V_m^b, \tag{123}$$

where $C_1 = 7C_{\text{drift}}^2$, $C_2 = 6$, $C_3 = 2$, and $C_{\text{drift}} = 3L_b\sigma_{\max}G$. The correction-point error satisfies

$$\sqrt{V_t} \le \sqrt{U_t} + L_b\sigma_{\max}G\, \eta_t, \tag{124}$$

so that

$$\limsup_{t\to\infty} V_t \;\le\; 2C_2\, R_m^2 \;=\; 12\, R_m^2, \tag{125}$$

which in turn implies that $\limsup_{t\to\infty} \sqrt{V_t} \;\le\; 2\sqrt{3}\, R_m$. That is, the timescale-gap and noise-filter contributions vanish in t; only the bias-of-bias-estimator residual remains.

ii. **Slow-timescale recurrence.** The slow iterates obey the per-step recurrence

$$\mathbb{E}\, V(r_\tau, \theta^\star) \le (1 - \eta_\tau\widetilde{\mu})\; \mathbb{E}\, V(r_{\tau-1}, \theta^\star) + 2\eta_\tau\Omega\sqrt{\sigma_{\max}}\sqrt{V_\tau} + 3\eta_\tau^2(\widetilde{B}_m^2 + \widetilde{V}_m), \tag{126}$$

where $V(\cdot,\cdot) := \frac{1}{2}\|\cdot - \cdot\|_{\Sigma^+}^2$, $\widetilde{B}_m^2 := \sigma_{\max}B_m^2$, $\widetilde{V}_m := \sigma_{\max}V_m$, and Ω is the prox-radius of D. Telescoping (126) yields, for all $T \ge 1$,

$$\begin{aligned}\mathbb{E}\, V(r_T, \theta^\star) \;\le\; & \prod_{t=1}^{T}(1 - \eta_t\widetilde{\mu})\, V(r_0, \theta^\star) \\ & + \sum_{\tau=1}^{T}\Big[\prod_{s=\tau+1}^{T}(1 - \eta_s\widetilde{\mu})\Big]\Big(2\eta_\tau\Omega\sqrt{\sigma_{\max}}\sqrt{V_\tau} + 3\eta_\tau^2(\widetilde{B}_m^2 + \widetilde{V}_m)\Big).\end{aligned} \tag{127}$$

iii. **Asymptotic vanishing bias.** Combining (127) with the fast-timescale envelope of part i and a weighted-average argument (Lemma 17 below) yields

$$\limsup_{T\to\infty} \mathbb{E}\, \|r_T - \theta^\star\|_{\Sigma^+}^2 \;\le\; \frac{8\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}}{\widetilde{\mu}} \cdot R_m. \tag{128}$$

The bias contribution to (127) therefore vanishes asymptotically in T, leaving only the R_m residual ($= \mathcal{O}(1/m^2)$ for delta-method estimators, $= 0$ for unbiased estimators, in which case the iterates converge to $\theta^\star$ in mean square).

Comparison to Theorem 9 (constant step, no tracking). The persistent $\mathcal{O}(1/m)$ bias term—namely, $4\Omega\widetilde{B}_m/\widetilde{\mu}$—in the bound of Theorem 9 is replaced in Theorem 10 by $8\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}\, R_m/\widetilde{\mu}$, which is *not* a sup-style envelope along the trajectory but a true asymptotic-in-T floor obtained by a weighted-average argument that exploits the weight-concentration property of the geometric kernel $(1 - \eta_s\widetilde{\mu})$. For delta-method estimators, $R_m = \mathcal{O}(1/m^2)$, giving an $\mathcal{O}(1/m)$ floor reduced to $\mathcal{O}(1/m^2)$—i.e., *the bias floor is moved from a first-order to a second-order term in* $1/m$. For increasing fixed budgets m, $R_m \to 0$ and hence this asymptotic bound tends to zero.

Proof of Theorem 10. The proof proceeds in three steps: (1) control the fast-timescale tracking error U_t via a scalar recurrence and transfer its asymptotic envelope to V_t; (2) establish the slow-timescale per-step recurrence and telescope it; (3) combine via a weighted-average argument that extracts $\limsup_\tau \sqrt{V_\tau}$ rather than $\sup_\tau \sqrt{V_\tau}$.

Step 1: Tracking-error recurrence. The bias tracker's update is $\xi_\tau = (1 - \gamma_\tau)\xi_{\tau-1} + \gamma_\tau\widehat{b}_m^{(\tau,2)}$, and $\widehat{b}_m^{(\tau,2)} = b_m(w_\tau) + r_m(w_\tau) + \nu^{(\tau,2)}$ with $\mathbb{E}[\nu^{(\tau,2)} \mid \mathcal{F}_\tau] = 0$, $\mathbb{E}\,\|\nu^{(\tau,2)}\|^2 \le V_m^b$, $\|r_m\|_2 \le R_m$. For $e_\tau = \xi_\tau - b_m(r_\tau)$, we have

$$\begin{aligned} e_\tau &= (1 - \gamma_\tau)(e_{\tau-1} + b_m(r_{\tau-1})) + \gamma_\tau b_m(w_\tau) + \gamma_\tau r_m(w_\tau) + \gamma_\tau\nu^{(\tau,2)} - b_m(r_\tau) \\ &= (1 - \gamma_\tau)e_{\tau-1} + \delta_\tau + \gamma_\tau r_m(w_\tau) + \gamma_\tau\nu^{(\tau,2)}, \end{aligned}$$

where the drift is

$$\delta_\tau := (1-\gamma_\tau)b_m(r_{\tau-1}) + \gamma_\tau b_m(w_\tau) - b_m(r_\tau).$$

By Lipschitz of b_m and the extragradient step bounds $\|r_\tau - r_{\tau-1}\|_2 \le \eta_\tau\sigma_{\max}G$, $\|w_\tau - r_\tau\|_2 \le 2\eta_\tau\sigma_{\max}G$:

$$\|\delta_\tau\|_2 \le L_b\big(\|r_{\tau-1}-r_\tau\|_2+\gamma_\tau\|w_\tau-r_\tau\|_2\big) \le L_b\eta_\tau\sigma_{\max}G(1+2\gamma_\tau) \le 3L_b\sigma_{\max}G\,\eta_\tau =: C_{\mathrm{drift}}\,\eta_\tau,$$

where the second inequality uses $\gamma_\tau \le 1$.

Step 2: Mean-square recurrence. Expanding $\|e_\tau\|^2$, we have that

$$\|e_\tau\|^2 = \|A_\tau\|^2 + 2\gamma_\tau\langle A_\tau, \nu^{(\tau,2)}\rangle + \gamma_\tau^2\|\nu^{(\tau,2)}\|^2, \tag{129}$$

where $A_\tau := (1-\gamma_\tau)e_{\tau-1} + \delta_\tau + \gamma_\tau r_m(w_\tau)$.

Let us now analyze the cross term. Decompose $2\gamma_\tau\langle A_\tau, \nu^{(\tau,2)}\rangle = 2\gamma_\tau\langle(1-\gamma_\tau)e_{\tau-1}, \nu^{(\tau,2)}\rangle + 2\gamma_\tau\langle\delta_\tau + \gamma_\tau r_m, \nu^{(\tau,2)}\rangle$. Then for each of the two terms we have the following:

- First, $e_{\tau-1}$ is $\mathcal{F}_\tau$-measurable, $\mathbb{E}[\nu^{(\tau,2)} \mid \mathcal{F}_\tau] = 0$, so $\mathbb{E}\langle e_{\tau-1}, \nu^{(\tau,2)}\rangle = 0$.
- Next, δ_τ depends on r_τ, which uses the correction-step samples that also enter $\nu^{(\tau,2)}$, so this inner product need not be mean-zero. Bound by Young's with parameter 1, we have the bound

  $$2\gamma_\tau|\langle\delta_\tau, \nu^{(\tau,2)}\rangle| \le \|\delta_\tau\|^2 + \gamma_\tau^2\|\nu^{(\tau,2)}\|^2.$$

  Similarly $r_m(w_\tau)$ is $\mathcal{F}_\tau$-measurable (only depends on the extrap-step samples through w_τ, not the correction-step noise $\nu^{(\tau,2)}$), so $\mathbb{E}\langle r_m, \nu^{(\tau,2)}\rangle = 0$.

Combining these observations and taking the expectation, we have that

$$\mathbb{E}[2\gamma_\tau\langle A_\tau, \nu^{(\tau,2)}\rangle] \le \mathbb{E}\,\|\delta_\tau\|^2 + \gamma_\tau^2 V_m^b.$$

Next let us bound the leading quadratic term in (129). Applying Young's with parameter $\gamma_\tau/2$ to $\|(1-\gamma_\tau)e_{\tau-1} + \delta_\tau + \gamma_\tau r_m\|^2$, we have that

$$\begin{aligned}\|A_\tau\|^2 &\le (1+\gamma_\tau/2)(1-\gamma_\tau)^2\|e_{\tau-1}\|^2 + (1+2/\gamma_\tau)\|\delta_\tau + \gamma_\tau r_m\|^2\\ &\le (1-\gamma_\tau/2)\|e_{\tau-1}\|^2 + (3/\gamma_\tau)(2\|\delta_\tau\|^2 + 2\gamma_\tau^2 R_m^2)\\ &= (1-\gamma_\tau/2)\|e_{\tau-1}\|^2 + (6/\gamma_\tau)\|\delta_\tau\|^2 + 6\gamma_\tau R_m^2,\end{aligned}$$

using $(1+\gamma_\tau/2)(1-\gamma_\tau)^2 \le 1-\gamma_\tau/2$ for $\gamma_\tau \in (0,1]$ (since $(1+\gamma/2)(1-\gamma)^2 = 1 - \frac{3}{2}\gamma + \frac{1}{2}\gamma^3 \le 1-\gamma/2$ when $\gamma \le 1$) and $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$.

Now we combine these two bounds and use the fact that $\|\delta_\tau\|^2 \le C_{\mathrm{drift}}^2\eta_\tau^2$ to deduce that

$$\begin{aligned}\mathbb{E}\,\|e_\tau\|^2 &\le (1-\gamma_\tau/2)\,\mathbb{E}\,\|e_{\tau-1}\|^2 + (6/\gamma_\tau + 1)\,\mathbb{E}\,\|\delta_\tau\|^2 + 6\gamma_\tau R_m^2 + 2\gamma_\tau^2 V_m^b\\ &\le (1-\gamma_\tau/2)U_\tau + \frac{7C_{\mathrm{drift}}^2\eta_\tau^2}{\gamma_\tau} + 6\gamma_\tau R_m^2 + 2\gamma_\tau^2 V_m^b,\end{aligned}$$

Since $\mathbb{E}\,\|e_\tau\|^2 = U_{\tau+1}$, this is (123) with $C_1 = 7C_{\mathrm{drift}}^2$, $C_2 = 6$, $C_3 = 2$. The inequality $6/\gamma_\tau + 1 \le 7/\gamma_\tau$ holds for $\gamma_\tau \le 1$.

To relate this to the correction-point error, use

$$e_{t-1}^* = e_{t-1} + b_m(r_{t-1}) - b_m(w_t), \qquad \|b_m(r_{t-1}) - b_m(w_t)\|_2 \le L_b\sigma_{\max}G\,\eta_t.$$

The triangle inequality in mean square gives (124).

Step 3: Asymptotic envelope on V_t via scalar two-timescale unrolling. We invoke the following elementary scalar lemma for time-varying stochastic-approximation recurrences.

Lemma 14 (Scalar two-timescale unrolling)**.** Suppose $u_{t+1} \le (1-c\gamma_t)u_t + F_t$ with $c \in (0,1]$, $\gamma_t \in (0,1]$, $\gamma_t \to 0$, $\sum_t \gamma_t = \infty$, $F_t \ge 0$. If $F_t/\gamma_t \to L$ for some $L \ge 0$, then $\limsup_t u_t \le L/c$.

Proof. For any $\epsilon > 0$, eventually $F_t \le (L+\epsilon)\gamma_t$. Subtracting $(L+\epsilon)/c$ from the recurrence and unrolling gives

$$u_{T+1} - \frac{L+\epsilon}{c} \le \prod_{t=N}^{T}(1 - c\gamma_t)\left(u_N - \frac{L+\epsilon}{c}\right).$$

The product tends to zero because $\sum_t \gamma_t = \infty$. Taking the limit superior and then $\epsilon \downarrow 0$ proves the claim. □

Applying with $u_t = U_t$, $c = 1/2$, and $F_t = C_1\eta_t^2/\gamma_t + C_2\gamma_t R_m^2 + C_3\gamma_t^2 V_m^b$, we have

$$\frac{F_t}{\gamma_t} = \frac{C_1\eta_t^2}{\gamma_t^2} + C_2 R_m^2 + C_3\gamma_t V_m^b.$$

Under conditions (122), the step sizes satisfy $\eta_t/\gamma_t \to 0$ and $\gamma_t \to 0$, so

$$\lim_{t\to\infty} \frac{F_t}{\gamma_t} = C_2 R_m^2.$$

Also $\sum_t \gamma_t = \infty$, since eventually $\eta_t \le \gamma_t$. Lemma 14 gives $\limsup_t U_t \le 2C_2 R_m^2 = 12R_m^2$. Since $\eta_t \to 0$, (124) gives the same limit-superior bound for V_t, proving the first half of (125). Taking square roots,

$$\limsup_{t\to\infty} \sqrt{V_t} \le \sqrt{2C_2}\, R_m = 2\sqrt{3}\, R_m, \tag{130}$$

which is the second half of (125). This completes part (i).

Step 4: slow-timescale per-step recurrence. Apply the proof of Theorem 9 with the preconditioned directions $\widetilde{g}^{(\tau,i)} = \Sigma(\widehat{F}_m^{(\tau,i)} - \xi_{\tau-1})$. The same tracker is subtracted in both calls, so

$$\widetilde{g}^{(\tau,1)} - \widetilde{g}^{(\tau,2)} = \Sigma(\widehat{F}_m^{(\tau,1)} - \widehat{F}_m^{(\tau,2)}).$$

Thus the squared-difference bound is unchanged, while the correction-step conditional bias is $-e_{\tau-1}^*$. The descent identity at $u = \theta^\star$ gives us that

$$\mathbb{E}\, V(r_\tau, \theta^\star) \le (1 - \eta_\tau\widetilde{\mu})\ \mathbb{E}\, V(r_{\tau-1}, \theta^\star) + 2\eta_\tau\Omega\sqrt{\sigma_{\max}}\sqrt{V_\tau} + 3\eta_\tau^2(\widetilde{B}_m^2 + \widetilde{V}_m). \tag{131}$$

The new bias cross-term $\eta_\tau\langle e_{\tau-1}^*, w_\tau - \theta^\star\rangle$ is bounded in expectation by $2\eta_\tau\Omega\sqrt{\sigma_{\max}}\sqrt{V_\tau}$ via Cauchy-Schwarz in ℓ_2, the metric translation $\|v\|_2 \le \sqrt{\sigma_{\max}}\|v\|_{\Sigma^+}$, the diameter bound $\|w_\tau - \theta^\star\|_{\Sigma^+} \le 2\Omega$, and Jensen's inequality $\mathbb{E}\,\|e_{\tau-1}^*\|_2 \le \sqrt{\mathbb{E}\,\|e_{\tau-1}^*\|_2^2} = \sqrt{V_\tau}$. This is (126).

Step 5: Telescoping the slow recurrence. Iterating on (131), we have that

$$\mathbb{E}\, V(r_T, \theta^\star) \le \prod_{t=1}^{T}(1-\eta_t\widetilde{\mu})\, V(r_0, \theta^\star) + \sum_{\tau=1}^{T}\Big[\prod_{s=\tau+1}^{T}(1-\eta_s\widetilde{\mu})\Big]\big(2\eta_\tau\Omega\sqrt{\sigma_{\max}}\sqrt{V_\tau} + 3\eta_\tau^2(\widetilde{B}_m^2 + \widetilde{V}_m)\big), \tag{132}$$

which is (127) and establishes part (ii).

Step 6: Weighted-average analysis of the bias term. We now analyze the asymptotic behavior of the bound from (132) as $T \to \infty$. The key observation is that the bias contribution in (132) can be written as a weighted average of $\sqrt{V_\tau}$ with weights that asymptotically concentrate on large τ, allowing us to extract $\limsup_\tau \sqrt{V_\tau}$.

Define the weights

$$w_\tau^T := \widetilde{\mu}\cdot\eta_\tau \cdot \prod_{s=\tau+1}^{T}(1 - \eta_s\widetilde{\mu}), \tag{133}$$

for $1 \le \tau \le T$. Then the bias contribution to (132) is

$$S_T^{\text{bias}} := \sum_{\tau=1}^{T}\Big[\prod_{s=\tau+1}^{T}(1-\eta_s\widetilde{\mu})\Big]\, 2\eta_\tau\Omega\sqrt{\sigma_{\max}}\sqrt{V_\tau} = \frac{2\Omega\sqrt{\sigma_{\max}}}{\widetilde{\mu}}\sum_{\tau=1}^{T} w_\tau^T\sqrt{V_\tau}. \tag{134}$$

We establish three lemmas governing the weights.

Lemma 15 (Telescoping identity)**.** For any $\eta_t \geq 0$ with $\eta_t\widetilde{\mu} \leq 1$, the equality holds: $\sum_{\tau=1}^{T} w_\tau^T = 1 - \prod_{t=1}^{T}(1-\eta_t\widetilde{\mu})$.

Proof. Write $\widetilde{\mu}\eta_\tau = 1 - (1-\eta_\tau\widetilde{\mu})$. Then

$$w_\tau^T = \big[1-(1-\eta_\tau\widetilde{\mu})\big]\prod_{s=\tau+1}^{T}(1-\eta_s\widetilde{\mu}) = \prod_{s=\tau+1}^{T}(1-\eta_s\widetilde{\mu}) - \prod_{s=\tau}^{T}(1-\eta_s\widetilde{\mu}).$$

Summing over $\tau = 1,\dots,T$, the right hand side telescopes. Indeed, we have that

$$\sum_{\tau=1}^{T} w_\tau^T = \prod_{s=T+1}^{T}(1-\eta_s\widetilde{\mu}) - \prod_{s=1}^{T}(1-\eta_s\widetilde{\mu}) = 1 - \prod_{t=1}^{T}(1-\eta_t\widetilde{\mu}),$$

using the empty-product convention. □

Since $\sum_t \eta_t = \infty$ and $\eta_t\widetilde{\mu} \leq 1$, we have $\prod_{t=1}^{T}(1-\eta_t\widetilde{\mu}) \leq \exp(-\widetilde{\mu}\sum_t \eta_t) \to 0$, so $\sum_\tau w_\tau^T \to 1$ as $T \to \infty$. Thus $\{w_\tau^T\}_{\tau=1}^{T}$ is asymptotically a probability distribution on $\{1,\dots,T\}$.

Lemma 16 (Weight concentration)**.** Under $\sum_t \eta_t = \infty$ and $\eta_t\widetilde{\mu} \leq 1$, for any fixed $T_0 \geq 1$, the limit $\sum_{\tau=1}^{T_0} w_\tau^T \to 0$ holds as $T \to \infty$.

Proof. For $\tau \leq T_0$ and $T > T_0$, factor

$$w_\tau^T = \widetilde{\mu}\eta_\tau \prod_{s=\tau+1}^{T_0}(1-\eta_s\widetilde{\mu}) \cdot \prod_{s=T_0+1}^{T}(1-\eta_s\widetilde{\mu}).$$

The first two factors are bounded by a constant $C(T_0)$ depending only on T_0 and the schedule. The third factor satisfies

$$\prod_{s=T_0+1}^{T}(1-\eta_s\widetilde{\mu}) \leq \exp\Big(-\widetilde{\mu}\sum_{s=T_0+1}^{T}\eta_s\Big) \xrightarrow[T\to\infty]{} 0,$$

since the tail sum diverges by $\sum_t \eta_t = \infty$. Hence $w_\tau^T \to 0$ for each fixed $\tau \leq T_0$, and the finite sum $\sum_{\tau=1}^{T_0} w_\tau^T \to 0$. □

Lemma 17 (Weighted average inherits $\limsup$)**.** Let $\{a_\tau\}_{\tau\geq 1}$ be a non-negative sequence with $L := \limsup_{\tau\to\infty} a_\tau < \infty$. Then under the conditions of Lemma 16, we have that

$$\limsup_{T\to\infty} \sum_{\tau=1}^{T} w_\tau^T a_\tau \leq L.$$

Proof. Fix $\epsilon > 0$. By definition of $\limsup$ there exists T_0 such that $a_\tau \leq L + \epsilon$ for all $\tau > T_0$. Splitting the sum, we have that

$$\sum_{\tau=1}^{T} w_\tau^T a_\tau = \sum_{\tau=1}^{T_0} w_\tau^T a_\tau + \sum_{\tau=T_0+1}^{T} w_\tau^T a_\tau.$$

The first sum is bounded by $\big(\max_{\tau\leq T_0} a_\tau\big)\sum_{\tau\leq T_0} w_\tau^T$, which converges to zero by Lemma 16 (the max is a finite constant independent of T). The second sum is bounded by $(L+\epsilon)\sum_{\tau=T_0+1}^{T} w_\tau^T \leq (L+\epsilon)\cdot 1$, using Lemma 15. Hence $\limsup_T \sum_\tau w_\tau^T a_\tau \leq L + \epsilon$, and letting $\epsilon \to 0$ gives the claim. □

Now that we have established these three lemmas we combine them. First, apply Lemma 17 with $a_\tau = \sqrt{V_\tau}$. By (130), $\limsup_\tau \sqrt{V_\tau} \leq 2\sqrt{3}\,R_m$, so that

$$\limsup_{T\to\infty} \sum_{\tau=1}^{T} w_\tau^T \sqrt{V_\tau} \leq 2\sqrt{3}\,R_m.$$

Hence by (134), we have that

$$\limsup_{T\to\infty} S_T^{\text{bias}} \leq \frac{2\Omega\sqrt{\sigma_{\max}}}{\widetilde{\mu}} \cdot 2\sqrt{3}\,R_m = \frac{4\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}\,R_m}{\widetilde{\mu}}.$$

Step 7: Variance term and initial-condition term. The variance contribution to (132) is

$$S_T^{\mathrm{var}} := 3(\widetilde{B}_m^2 + \widetilde{V}_m) \sum_{\tau=1}^{T} \Big[\prod_{s=\tau+1}^{T} (1 - \eta_s \widetilde{\mu}) \Big] \eta_\tau^2.$$

Using the weights (133), this is

$$S_T^{\mathrm{var}} = \frac{3(\widetilde{B}_m^2 + \widetilde{V}_m)}{\widetilde{\mu}} \sum_{\tau=1}^{T} w_\tau^T \eta_\tau.$$

Since $\eta_t/\gamma_t \to 0$ and $\gamma_t \to 0$ together imply $\eta_t \to 0$, Lemma 17 applied to $a_\tau = \eta_\tau$ gives $S_T^{\mathrm{var}} \to 0$.

The initial-condition term satisfies $\prod_{t=1}^{T}(1 - \eta_t \widetilde{\mu}) \le \exp(-\widetilde{\mu} \sum_t \eta_t) \to 0$ since $\sum_t \eta_t = \infty$.

Step 8: Combine the previous bounds. Multiplying (132) by two converts from $V(\cdot,\cdot)$ to $\|\cdot\|_{\Sigma^+}^2$ so that taking $\limsup_{T\to\infty}$ leads to

$$\limsup_{T\to\infty} \mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 \le 0 + \frac{8\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}}{\widetilde{\mu}} \cdot R_m + 0,$$

which is (128). This completes part (iii). □

Vanishing-bias structure. The crucial qualitative difference from Theorem 9: the persistent $\mathcal{O}(1/m)$ floor is replaced by a true asymptotic-in-T floor of magnitude $\mathcal{O}(R_m)$, obtained via the weighted-average argument of Lemmas 15–17. This is sharper than a $\sup_\tau \sqrt{V_\tau}$ envelope because the geometric kernel $(1 - \eta_s\widetilde{\mu})$ produces weights that concentrate asymptotically on the tail of the trajectory, where $\sqrt{V_\tau}$ is close to its $\limsup$. For the most useful estimators (delta-method, $R_m = \mathcal{O}(1/m^2)$), this means the bias contribution to the bound is *quadratic* in $1/m$ rather than linear. If the bias estimator is structurally unbiased ($R_m = 0$, e.g., oracle-threshold CVaR), the bias is eliminated entirely: the right-hand side of (128) is zero and the iterates converge to $\theta^\star$ in mean square.

Step-size schedule. Any schedule satisfying (122) works. The cleanest choice is $\eta_t = c_2 t^{-a}$, $\gamma_t = c_1 t^{-b}$ with $0 < b < a \le 1$. Specific case $a = 1$, $b = 2/3$ is the standard textbook schedule. Slower-decaying η_t (i.e., $a < 1$) gives stretched-exponential decay of the contraction term, at the cost of a potentially worse pre-asymptotic regime.

The choice $a = 1$, $b = 2/3$ also gives an explicit $T^{-2/3}$ finite-time rate. To obtain it, we retain the quadratic tracking error when bounding the cross-term in the slow-iterate descent argument.

Corollary 7 (Two-timescale rate with decreasing steps)**.** Under the assumptions of Theorem 10, choose

$$\eta_t = \frac{c_\eta}{t + t_0}, \qquad \gamma_t = \frac{c_\gamma}{(t+t_0)^{2/3}}, \qquad c_\eta, c_\gamma > 0, \quad c_\eta \widetilde{\mu} > \frac{4}{3},$$

where $t_0 \ge 0$ is large enough to satisfy the step-size bounds in (122). For fixed m, there is a constant C independent of T such that, for all $T \ge 1$,

$$\mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 \le \frac{16\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}}{\widetilde{\mu}} R_m + \frac{C}{(T+t_0)^{2/3}}. \tag{135}$$

Proof. Write $n_t = t + t_0$ and $u_t = (U_t - 2C_2 R_m^2)_+$, where $(x)_+ := \max\{x, 0\}$. Substituting the step sizes into (123) gives

$$u_{t+1} \le \left(1 - \frac{c_\gamma}{2n_t^{2/3}}\right) u_t + \frac{K_f}{n_t^{4/3}}, \qquad K_f := \frac{C_1 c_\eta^2}{c_\gamma} + C_3 c_\gamma^2 V_m^b.$$

Since $n^{-2/3} - (n+1)^{-2/3} \le (2/3)n^{-5/3}$, induction gives $u_t \le K n_t^{-2/3}$: choose $K \ge 4K_f/c_\gamma$ and start where $n_t^{1/3} \ge 8/(3c_\gamma)$, increasing K to cover the initial error and the finite preceding segment. Thus (124) implies

$$U_t \le 12R_m^2 + K n_t^{-2/3}, \qquad \sqrt{V_t} \le 2\sqrt{3}\, R_m + K' n_t^{-1/3}$$

for a constant K' independent of t.

Set $Q_t := \mathbb{E}\,\|r_t - \theta^\star\|_{\Sigma^+}^2$ and $d_t := (\mathbb{E}\,\|w_t - \theta^\star\|_{\Sigma^+}^2)^{1/2} \le 2\Omega$. Using the bound on $\sqrt{V_t}$, Cauchy–Schwarz and Young's inequality give

$$\begin{aligned}\mathbb{E}\langle e_{t-1}^*, w_t - \theta^\star\rangle &\le \sqrt{\sigma_{\max}}\sqrt{V_t}\, d_t \\ &\le 4\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}R_m + \frac{\widetilde{\mu}}{2}d_t^2 + \frac{\sigma_{\max}(K')^2}{2\widetilde{\mu}}n_t^{-2/3}.\end{aligned}$$

Apply this estimate in Step 4 of the proof of Theorem 10, before the bridge inequality (88). Absorbing the term $\eta_t\widetilde{\mu}d_t^2/2$ leaves half of the strong-monotonicity term. After the bridge, the coefficient on $\mathbb{E}\,\|r_{t-1} - w_t\|_{\Sigma^+}^2$ is

$$\frac{\eta_t\widetilde{\mu}}{2} + \frac{3\eta_t^2 L_G^2}{2} - \frac{1}{2} \le 0$$

under (122). Converting from the prox divergence to squared distance therefore gives

$$\begin{aligned}Q_t \le \left(1 - \frac{\eta_t\widetilde{\mu}}{2}\right)Q_{t-1} + 8\sqrt{3}\,\eta_t\Omega\sqrt{\sigma_{\max}}R_m \\ + \frac{\eta_t\sigma_{\max}(K')^2}{\widetilde{\mu}}n_t^{-2/3} + 6\eta_t^2(\widetilde{B}_m^2 + \widetilde{V}_m).\end{aligned}$$

Let $p := c_\eta\widetilde{\mu}/2 > 2/3$ and $H := 16\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}R_m/\widetilde{\mu}$. Substituting $\eta_t = c_\eta/n_t$ and subtracting H yields

$$Q_t - H \le \left(1 - \frac{p}{n_t}\right)(Q_{t-1} - H) + K''n_t^{-5/3},$$

where K'' also absorbs the slow-variance term of order n_t^{-2}. Unrolling this recurrence gives

$$Q_T - H \le C_0 n_T^{-p}\left(1 + \sum_{t=1}^{T} n_t^{p-5/3}\right) \le Cn_T^{-2/3},$$

with constants independent of T. The last inequality uses $p > 2/3$, and proves (135). □

Comparison to the constant-step version. With *constant* η, γ, the tracking recurrence leaves terms of orders η/γ, R_m, and $\sqrt{\gamma V_m^b}$ in the root-mean-square tracking error, while the slow recurrence leaves a term of order $\eta(\widetilde{B}_m^2 + \widetilde{V}_m)$. The next subsubsection gives the finite-horizon analysis, including the transient from initializing the tracker. With time-varying steps, the timescale-gap, noise-filter, and slow-variance contributions all vanish in T, leaving only the R_m residual. The trade-off is contraction rate: the initial-condition factor is $(1 - \eta\widetilde{\mu})^T$ for constant steps, whereas its bound $\exp(-\widetilde{\mu}\sum_{t=1}^{T}\eta_t)$ is polynomial in T for $\eta_t = c_2/t$, and stretched exponential for $\eta_t = c_2 t^{-a}$ with $0 < a < 1$. The decaying-step version is the right choice when one needs to drive the error below the $\mathcal{O}(1/m)$ persistent floor and is willing to accept slower convergence.

Implications by risk type.

- **Entropic risk:** $R_m = \mathcal{O}(e^{2\lambda}/(\lambda m^2))$. Two timescale extra-gradient drives the bias contribution from a persistent $\mathcal{O}(e^{2\lambda}/(\lambda m))$ floor to an asymptotic-in-T floor of $\mathcal{O}(e^{2\lambda}/(\lambda m^2))$. The improvement is a full power of $1/m$. Here entropic risk is

  $$\mathcal{R}_\lambda(Z) := -\lambda^{-1}\log\mathbb{E}[e^{-\lambda Z}], \qquad \lambda > 0,$$

  where λ controls the degree of risk sensitivity.
- **CVaR via oracle-threshold Rockafellar-Uryasev:** $R_m = 0$ trivially. The right-hand side of (128) is zero, so $\mathbb{E}\,\|r_T - \theta^\star\|^2 \to 0$, and there is no asymptotic bias. Matches the un-tracked oracle-threshold CVaR-RU result, which already has zero bias because the RU threshold is fixed at $\nu^\star$ before sampling.
- **General coherent risk:** applicable whenever a sample-based $\widehat{b}_m$ with structural properties (B1)-(B3) can be constructed; the R_m residual scales according to the quality of the estimator.

I.1.3 CONSTANT STEP-SIZE REGIME

The asymptotic conditions of Theorem 10 require $\eta_t \to 0$, ruling out constant-step schedules. Indeed with $\eta_t \equiv \eta > 0$, the leader's iterate moves at a constant rate, which *can* prevent the fast-timescale tracker from catching up asymptotically and produces a non-vanishing steady-state tracking error.

We can nonetheless run the algorithm with constant step sizes and obtain a finite-horizon bound, where the bias floor is controlled by the ratio η/γ rather than vanishing in T. This trades the asymptotic bias-floor reduction of Theorem 10 for initial-condition decay governed by the two constant step sizes. The initial tracking error satisfies $U_1 = \mathbb{E}\,\|\xi_0 - b_m(r_0)\|_2^2 \le B_m^2$ since $\xi_0 = 0$.

Theorem 11 (Two-timescale debiasing for stochastic extragradient, constant step sizes)**.** Under the assumptions of Theorem 10 except with *constant* step sizes $\eta_t \equiv \eta$, $\gamma_t \equiv \gamma$ satisfying

$$0 < \eta \le \min\left\{\frac{1}{4\widetilde{\mu}},\ \frac{1}{\sqrt{6}\,L_G}\right\}, \qquad \gamma \in (0,1], \tag{136}$$

the iterates of Algorithm 1 satisfy:

i. ***Fast-timescale tracking, finite horizon.*** For all $t \ge 1$,

$$U_t \le (1-\gamma/2)^{t-1}\,U_1 + U_\infty, \qquad U_\infty := \frac{2C_1\eta^2}{\gamma^2} + 2C_2R_m^2 + 2C_3\gamma V_m^b, \tag{137}$$

with $C_1 = 7C_{\mathrm{drift}}^2 = 63L_b^2\sigma_{\max}^2G^2$, $C_2 = 6$, $C_3 = 2$ as in Theorem 10. In particular, the estimate holds:

$$\limsup_{t\to\infty}\sqrt{V_t} \le \sqrt{U_\infty} + L_b\sigma_{\max}G\,\eta \le \frac{K_1\eta}{\gamma} + K_2R_m + K_3\sqrt{\gamma V_m^b}, \tag{138}$$

where $K_1 = \sqrt{2C_1} + L_b\sigma_{\max}G$, $K_2 = \sqrt{2C_2}$, and $K_3 = \sqrt{2C_3}$; the last inequality uses $\gamma \le 1$.

ii. ***Slow-timescale convergence, finite horizon.*** For all $T \ge 1$, the estimate holds:

$$\begin{aligned}\mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 &\le (1-\eta\widetilde{\mu})^T\|r_0-\theta^\star\|_{\Sigma^+}^2 \\ &\quad + \frac{4\Omega\sqrt{\sigma_{\max}}}{\widetilde{\mu}}\left(\sqrt{U_1}\,\chi_T + \frac{K_1\eta}{\gamma} + K_2R_m + K_3\sqrt{\gamma V_m^b}\right) + \frac{6\eta(\widetilde{B}_m^2 + \widetilde{V}_m)}{\widetilde{\mu}},\end{aligned} \tag{139}$$

where the contribution from the tracker's initial error is weighted by

$$\chi_T := \eta\widetilde{\mu}\sum_{\tau=1}^{T}(1-\eta\widetilde{\mu})^{T-\tau}(1-\gamma/2)^{(\tau-1)/2}. \tag{140}$$

iii. ***Steady-state floor.*** As $T \to \infty$, the asymptotic estimate holds:

$$\begin{aligned}&\limsup_{T\to\infty}\mathbb{E}\,\|r_T-\theta^\star\|_{\Sigma^+}^2 \le \\ &\underbrace{\frac{4\Omega\sqrt{\sigma_{\max}}K_1}{\widetilde{\mu}}\cdot\frac{\eta}{\gamma}}_{\text{timescale gap}} + \underbrace{\frac{4\Omega\sqrt{2\sigma_{\max}C_2}\,R_m}{\widetilde{\mu}}}_{\text{residual bias}} + \underbrace{\frac{4\Omega\sqrt{2\sigma_{\max}C_3}\sqrt{\gamma V_m^b}}{\widetilde{\mu}}}_{\text{noise filter}} + \underbrace{\frac{6\eta(\widetilde{B}_m^2+\widetilde{V}_m)}{\widetilde{\mu}}}_{\text{slow variance}}.\end{aligned} \tag{141}$$

Before getting to the proof, observe that each term in (141) has a distinct interpretation and a distinct step-size knob:

- **Timescale gap** $\propto \eta/\gamma$: the leader-follower mismatch. Reducible by taking $\eta \ll \gamma$. This is the term that vanishes in Theorem 10 under the schedule condition $\eta_t/\gamma_t \to 0$.
- **Residual bias** $\propto R_m$: the bias-of-the-bias-estimator floor. Independent of step sizes. This is the only persistent floor in the time-varying-step-size theorem.
- **Noise filter** $\propto \sqrt{\gamma}$: variance of the bias tracker filtered through the Robbins-Monro update. Reducible by taking γ small, but this trades against the timescale-gap term (which grows as η/γ when γ shrinks).

- **Slow variance** $\propto \eta$: the standard SGD variance floor inherited from the slow-timescale recurrence. Reducible by taking η small.

Optimal balance. Define

$$A := \frac{4\Omega\sqrt{\sigma_{\max}}K_1}{\widetilde{\mu}}, \qquad B := \frac{4\Omega\sqrt{2\sigma_{\max}C_3V_m^b}}{\widetilde{\mu}}.$$

The timescale-gap and noise-filter contributions are $A\eta/\gamma + B\sqrt{\gamma}$. For fixed η and $A, B > 0$, the unconstrained minimizer satisfies

$$-\frac{A\eta}{\gamma^2} + \frac{B}{2\sqrt{\gamma}} = 0, \qquad \gamma^\star = \left(\frac{2A\eta}{B}\right)^{2/3}.$$

This is admissible when $\gamma^\star \le 1$; otherwise the minimum over $\gamma \in (0, 1]$ is attained at $\gamma = 1$. At the interior optimum,

$$\frac{\eta}{\gamma^\star} = \left(\frac{B}{2A}\right)^{2/3}\eta^{1/3}, \qquad \sqrt{\gamma^\star} = \left(\frac{2A}{B}\right)^{1/3}\eta^{1/3},$$

and hence

$$A\frac{\eta}{\gamma^\star} + B\sqrt{\gamma^\star} = \frac{3}{2^{2/3}}(A\eta)^{1/3}B^{2/3}.$$

Thus the combined contribution is $\Theta(\eta^{1/3})$ for fixed $A, B > 0$. Keeping the dependence on V_m^b explicit, it scales as $(\eta V_m^b)^{1/3}$ with the remaining coefficients fixed. The horizon-tuned theorem below makes the choice of step sizes explicit as a function of the iteration budget T.

Proof of Theorem 11. The proof proceeds in steps.

Step 1: Tracking error. The recurrence (137) follows from Steps 1–2 of the proof of Theorem 10 specialized to constant η, γ: telescoping with constant contraction factor $1 - \gamma/2$ and using $\sum_{j=0}^{t-1}(1-\gamma/2)^j \le 2/\gamma$ gives $U_t \le (1-\gamma/2)^{t-1}U_1 + 2F/\gamma$ where $F = C_1\eta^2/\gamma + C_2\gamma R_m^2 + C_3\gamma^2 V_m^b$, which is (137). Taking square roots, applying $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, and using (124) gives

$$\sqrt{V_t} \le \sqrt{U_1}\,(1-\gamma/2)^{(t-1)/2} + \frac{K_1\eta}{\gamma} + K_2R_m + K_3\sqrt{\gamma V_m^b}, \tag{142}$$

where $L_b\sigma_{\max}G\,\eta$ is absorbed into $K_1\eta/\gamma$ using $\gamma \le 1$. Taking the limit superior also gives (138).

Step 2: Slow-timescale telescoping. Telescope the per-step recurrence (126) with constant η to obtain

$$\mathbb{E}\,V(r_T, \theta^\star) \le (1-\eta\widetilde{\mu})^T V(r_0, \theta^\star) + S_T^{\text{bias}} + S_T^{\text{var}},$$

with

$$S_T^{\text{bias}} = 2\Omega\sqrt{\sigma_{\max}}\,\eta\sum_{\tau=1}^{T}(1-\eta\widetilde{\mu})^{T-\tau}\sqrt{V_\tau}, \quad S_T^{\text{var}} = 3\eta^2(\widetilde{B}_m^2 + \widetilde{V}_m)\sum_{\tau=1}^{T}(1-\eta\widetilde{\mu})^{T-\tau}.$$

Using $\sum_{\tau=1}^{T}(1-\eta\widetilde{\mu})^{T-\tau} \le 1/(\eta\widetilde{\mu})$, we have that $S_T^{\text{var}} \le 3\eta(\widetilde{B}_m^2 + \widetilde{V}_m)/\widetilde{\mu}$.

For the bias term, substitute (142) for $\sqrt{V_\tau}$ and split into four pieces. The constant pieces (those not depending on τ) factor out. Using the same telescoping bound, we then have

$$\eta\sum_{\tau=1}^{T}(1-\eta\widetilde{\mu})^{T-\tau}\left(K_1\eta/\gamma + K_2R_m + K_3\sqrt{\gamma V_m^b}\right) \le \frac{1}{\widetilde{\mu}}\left(K_1\eta/\gamma + K_2R_m + K_3\sqrt{\gamma V_m^b}\right).$$

The τ-dependent piece (initial-condition decay) gives

$$\eta\sum_{\tau=1}^{T}(1-\eta\widetilde{\mu})^{T-\tau}(1-\gamma/2)^{(\tau-1)/2}\sqrt{U_1} = \frac{\sqrt{U_1}}{\widetilde{\mu}}\chi_T.$$

Bounding $(1-\gamma/2)^{(\tau-1)/2} \le 1$, the sum without $\sqrt{U_1}$ is at most

$$\eta\sum_{\tau=1}^{T}(1-\eta\widetilde{\mu})^{T-\tau} = \frac{1-(1-\eta\widetilde{\mu})^T}{\widetilde{\mu}} \le \frac{1}{\widetilde{\mu}}.$$

For a sharper bound, let

$$a := 1 - \eta\widetilde{\mu}, \qquad q := \sqrt{1-\gamma/2}.$$

Whenever $\eta\widetilde{\mu} < 1 - \sqrt{1-\gamma/2}$, we have $a > q$, and the geometric-sum identity gives

$$\eta \sum_{\tau=1}^{T} a^{T-\tau} q^{\tau-1} = \eta \frac{a^T - q^T}{a-q} \le \frac{\eta}{a-q} a^T = \frac{(1-\eta\widetilde{\mu})^T}{\widetilde{\mu}} \frac{\eta\widetilde{\mu}}{1 - \sqrt{1-\gamma/2} - \eta\widetilde{\mu}}.$$

Moreover, since

$$1 - \sqrt{1-\gamma/2} = \frac{\gamma/2}{1+\sqrt{1-\gamma/2}} \ge \frac{\gamma}{4},$$

combining the sharper estimate with the crude bound gives, whenever $\eta\widetilde{\mu} < \gamma/4$,

$$\eta \sum_{\tau=1}^{T} (1-\eta\widetilde{\mu})^{T-\tau} (1-\gamma/2)^{(\tau-1)/2} \le \frac{1}{\widetilde{\mu}} \min\left\{1,\ (1-\eta\widetilde{\mu})^T \frac{\eta\widetilde{\mu}}{\gamma/4 - \eta\widetilde{\mu}}\right\}.$$

Combining all pieces and converting to $\|\cdot\|^2_{\Sigma^+}$ (factor 2) yields (139).

Step 3: Steady-state floor. Since both contraction factors are strictly less than one,

$$0 \le \chi_T \le \eta\widetilde{\mu}\, T \left[\max\left\{1-\eta\widetilde{\mu}, \sqrt{1-\gamma/2}\right\}\right]^{T-1} \longrightarrow 0.$$

Taking $T \to \infty$ in (139): the contraction term $(1-\eta\widetilde{\mu})^T \|r_0 - \theta^\star\|^2_{\Sigma^+}$ and the $\sqrt{U_1}\chi_T$ term both vanish, leaving the four floor contributions in (141). □

Tuned step sizes (the $1/m^{3/2}$ floor). A more useful tuning, in terms of dependence on the sample budget m, proceeds as follows. Fix a burn-in length $N \ge 0$, and let

$$\mathcal{E}_T := \sup_{N < \tau \le N+T} \sqrt{V_\tau}$$

denote the worst-case root-mean-square tracking error over the T retained iterations. By (142),

$$\mathcal{E}_T \le \sqrt{U_1}(1-\gamma/2)^{N/2} + \frac{K_1\eta}{\gamma} + K_2 R_m + K_3\sqrt{\gamma V_m^b}. \tag{143}$$

Choose $\eta = \mathcal{O}(\gamma/m)$ with γ constant. Then the timescale-gap contribution to this bound scales as $\eta/\gamma = \mathcal{O}(1/m)$, so after burn-in

$$\mathcal{E}_T \;=\; O\left(\frac{1}{m} + R_m + \sqrt{\gamma V_m^b}\right) \;=\; O\left(\frac{1}{m} + R_m\right),$$

where the noise-filter term $\sqrt{\gamma V_m^b} = \mathcal{O}(1/m^{3/2})$ for $V_m^b = \mathcal{O}(1/m^3)$ (delta-method) is dominated by $1/m$. For the delta-method estimator with $R_m = \mathcal{O}(1/m^2)$, the $1/m$ timescale-gap term continues to dominate the residual until $\eta/\gamma \lesssim R_m$, i.e., $\eta = \mathcal{O}(\gamma/m^2)$, at which point both contributions are $\mathcal{O}(1/m^2)$ and the noise-filter term $\mathcal{O}(1/m^{3/2})$ becomes the dominant non-vanishing piece:

$$\mathcal{E}_T \;=\; \mathcal{O}(1/m^{3/2}) \qquad \text{once } \eta/\gamma \lesssim R_m.$$

Reaching this regime requires sufficiently many iterations: with the initial error $\sqrt{U_1} \le B_m = \mathcal{O}(1/m)$,

$$\sqrt{U_1}(1-\gamma/2)^{N/2} \le B_m e^{-\gamma N/4}.$$

Thus $N \ge (2/\gamma)\log m$ makes the initialization contribution $\mathcal{O}(m^{-3/2})$. This controls the tracking transient; the slow-iterate transient is separately accounted for in (139).

Over T retained iterations after burn-in, consider the expected VI gap of the averaged iterate $\widehat{z}_T := T^{-1}\sum_{\tau=N+1}^{N+T} w_\tau$ in the monotone setting. For the delta-method regime above, choose $\eta = \mathcal{O}(\gamma/m^2)$ and the stated burn-in, with γ fixed. Then both $\sqrt{U_\tau}$ and $\sqrt{V_\tau}$ are $\mathcal{O}(m^{-3/2})$ over the retained iterations. With $\widetilde{V}_m = \mathcal{O}(1/m)$, the expected-gap argument of Theorem 8, applied to the corrected oracle, gives

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_T)] \;=\; O\big(\Omega^2/(T\eta)\big) \;+\; O\big(\eta/m\big) \;+\; O\big(\Omega\sqrt{\sigma_{\max}}/m^{3/2}\big).$$

With the problem constants fixed, balancing the first two terms gives $\eta \asymp \sqrt{m/T}$. To also satisfy the tracking requirement, choose

$$\eta \asymp \min\left\{1, \sqrt{\frac{m}{T}}, \frac{\gamma}{m^2}\right\},$$

with the fixed multiplicative constant chosen to satisfy the extragradient step-size bound. Substituting this choice gives

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_T)] \;=\; O\left(\frac{1}{\sqrt{Tm}} + \frac{1}{T} + \frac{m^2}{\gamma T}\right) \;+\; O\left(\frac{\Omega\sqrt{\sigma_{\max}}}{m^{3/2}}\right). \tag{144}$$

The additional $m^2/(\gamma T)$ term accounts for the tracking cap on η; it is no larger than $1/\sqrt{Tm}$ when $T \geq m^5/\gamma^2$. The last term is the *reduced bias floor*: a half power of $1/m$ better than the persistent $\mathcal{O}(1/m)$ floor of the un-tracked algorithm, and a half power larger than the $\mathcal{O}(1/m^2)$ residual from R_m alone, since the $\mathcal{O}(m^{-3/2})$ noise-filter term dominates under this fixed-γ tuning.

Implications by risk type (constant-step / finite-horizon).

- **Entropic risk:** $R_m = \mathcal{O}(e^{2\lambda}/(\lambda m^2))$, $V_m^b = \mathcal{O}(e^{4\lambda}/(\lambda^2 m^3))$. For fixed $\lambda > 0$ and $\gamma \in (0, 1]$, two timescale extragradient with the $\eta = \mathcal{O}(\gamma/m^2)$ tuning drives the persistent $\mathcal{O}(e^{2\lambda}/(\lambda m))$ floor of the un-tracked algorithm down to $\mathcal{E}_T = \mathcal{O}(m^{-3/2})$ after the stated burn-in, with constants depending on λ, yielding the $\mathcal{O}(\Omega\sqrt{\sigma_{\max}}/m^{3/2})$ bias contribution in (144). This is a half-power improvement over the un-tracked $1/m$ floor. The constant-step noise-filter term is larger than the $\mathcal{O}(R_m) = \mathcal{O}(1/m^2)$ residual bound of Theorem 10 by a half power of m. The time-varying schedule of Theorem 10 eliminates the noise-filter term asymptotically and recovers the sharper $\mathcal{O}(1/m^2)$ residual — but at the cost of polynomial-in-T initial-condition decay when $\eta_t \propto 1/t$.
- **CVaR via oracle-threshold Rockafellar-Uryasev:** $R_m = V_m^b = 0$ when we take $\widehat{b}_m \equiv 0$, since $b_m \equiv 0$. Starting from $\xi_0 = 0$, the tracker satisfies $\xi_t = 0$ at every iteration, so $\mathcal{E}_T = 0$ exactly. Two timescale extragradient reduces to un-tracked extragradient at $\xi \equiv 0$, recovering the un-tracked monotone extragradient bound exactly with $\widetilde{B}_m = 0$, and the time-varying-step Theorem 10 gives convergence in mean square to $\theta^\star$ with no asymptotic bias.
- **General coherent risk:** applicable whenever a sample-based $\widehat{b}_m$ with structural properties (B1)-(B3) can be constructed; the floor scales as $\mathcal{O}(1/m^{3/2})$ for the delta-method estimators above ($R_m = \mathcal{O}(1/m^2)$, $V_m^b = \mathcal{O}(1/m^3)$) under the stated constant-step tuning and burn-in, and as $\mathcal{O}(R_m)$ asymptotically under decaying-step tuning.

I.1.4 Finite-horizon tuned rate

Theorem 10 (decaying step) establishes asymptotic vanishing of the tracking contributions with polynomial initial-condition decay ($\exp(-\widetilde{\mu}\sum_t \eta_t)$ with $\sum_t \eta_t = \Theta(\log T)$ for $\eta_t = c/t$). Theorem 11 (constant step) gives a four-term steady-state floor. An intermediate finite-horizon question is: given a budget of T extragradient iterations, what constant step sizes $\eta(T), \gamma(T)$ give an explicit convergence rate? The choice trades off the contraction rate against the steady-state floor as a function of T, and gives a polynomial-in-T decay rate to the residual.

Bounding the tracking cross-term quadratically—analogous to Drusvyatskiy et al. (2022)—gives an exact $T^{-1/2}$ rate under the following horizon-dependent tuning.

Corollary 8 (Exact horizon-tuned rate for two-timescale debiasing)**.** Under the assumptions of Theorem 11, set $\eta_{\max} := \min\{1/(4\widetilde{\mu}), 1/(\sqrt{6}L_G)\}$ and choose

$$\eta = c_\eta T^{-3/4}, \qquad \gamma = c_\gamma T^{-1/2}, \qquad 0 < c_\gamma \leq 1, \quad 0 < c_\eta \leq \min\left\{\eta_{\max}, c_\gamma/(2\widetilde{\mu})\right\}, \tag{145}$$

and hold these step sizes constant for the T iterations. For fixed m, there is a constant C independent of T such that

$$\mathbb{E}\left\|r_T - \theta^\star\right\|_{\Sigma^+}^2 \leq C\left(e^{-(c_\eta\widetilde{\mu}/2)T^{1/4}} + R_m + T^{-1/2}\right). \tag{146}$$

Proof. The tracking bound (142) gives

$$\sqrt{V_t} \leq 2\sqrt{3}\,R_m + A_t, \qquad A_t := \sqrt{U_1}(1-\gamma/2)^{(t-1)/2} + K_1\frac{\eta}{\gamma} + K_3\sqrt{\gamma V_m^b},$$

with K_1, K_3 as in Theorem 11. Write $d_t := (\mathbb{E}\,\|w_t - \theta^\star\|_{\Sigma^+}^2)^{1/2} \le 2\Omega$. Cauchy–Schwarz and Young's inequality yield

$$\mathbb{E}\langle e_{t-1}^*, w_t - \theta^\star\rangle \le \sqrt{\sigma_{\max}}\sqrt{V_t}\, d_t \le 4\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}R_m + \frac{\widetilde{\mu}}{2}d_t^2 + \frac{\sigma_{\max}}{2\widetilde{\mu}}A_t^2.$$

Apply this bound in the descent argument of Theorem 10, Step 4, before using the bridge inequality (88). The positive $\eta\widetilde{\mu}d_t^2/2$ term absorbs half of the negative strong-monotonicity term. After the bridge, the coefficient on $\mathbb{E}\,\|r_{t-1} - w_t\|_{\Sigma^+}^2$ is

$$\frac{\eta\widetilde{\mu}}{2} + \frac{3\eta^2 L_G^2}{2} - \frac{1}{2} \le 0$$

under the existing step-size bounds. Thus, setting $Q_t := \mathbb{E}\,\|r_t - \theta^\star\|_{\Sigma^+}^2$, we obtain

$$Q_t \le \left(1 - \frac{\eta\widetilde{\mu}}{2}\right) Q_{t-1} + 8\sqrt{3}\,\eta\Omega\sqrt{\sigma_{\max}}R_m + \frac{\eta\sigma_{\max}}{\widetilde{\mu}}A_t^2 + 6\eta^2(\widetilde{B}_m^2 + \widetilde{V}_m).$$

The tracking contribution now enters quadratically:

$$A_t^2 \le 3U_1(1-\gamma/2)^{t-1} + 3K_1^2\left(\frac{\eta}{\gamma}\right)^2 + 3K_3^2\gamma V_m^b.$$

Let $a := 1 - \eta\widetilde{\mu}/2$ and $b := 1 - \gamma/2$. The tuning (145) ensures $\eta\widetilde{\mu}/\gamma = (c_\eta\widetilde{\mu}/c_\gamma)T^{-1/4} \le 1/2$, so $a - b \ge \gamma/4 > 0$ and

$$\eta\sum_{t=1}^{T} a^{T-t}b^{t-1} = \eta\frac{a^T - b^T}{a-b} \le \frac{4\eta}{\gamma}a^T.$$

Unrolling the recurrence for Q_t and using $\sum_{t=1}^{T} a^{T-t} \le 2/(\eta\widetilde{\mu})$ therefore gives

$$Q_T \le a^T\left(Q_0 + C\frac{\eta}{\gamma}U_1\right) + C\left[R_m + \left(\frac{\eta}{\gamma}\right)^2 + \gamma V_m^b + \eta(\widetilde{B}_m^2 + \widetilde{V}_m)\right], \tag{147}$$

where C is independent of T. Finally, $a^T \le e^{-(c_\eta\widetilde{\mu}/2)T^{1/4}}$, $(\eta/\gamma)^2 = (c_\eta/c_\gamma)^2 T^{-1/2}$, and $\gamma V_m^b = c_\gamma V_m^b T^{-1/2}$. For fixed m, the slow-variance term is $O(T^{-3/4})$, while $\eta/\gamma = (c_\eta/c_\gamma)T^{-1/4}$ is bounded. These estimates prove (146). □

Comparison to the asymptotic regime. For $\eta_t = c/t$, the initial-condition factor in Theorem 10 decays polynomially: $\exp(-\widetilde{\mu}\sum_{t=1}^{T}\eta_t) = \Theta(T^{-c\widetilde{\mu}})$. The results address different schedules: Theorem 10 proves asymptotic consistency under any decaying schedule satisfying classical stochastic-approximation conditions; Corollary 8 gives an explicit finite-time guarantee under a horizon-tuned choice. The latter requires knowing T in advance, which is reasonable for training a fixed-budget algorithm.

Comparison to the un-tracked algorithm. Theorem 9 (no tracking, constant step) gives a persistent $\mathcal{O}(\widetilde{B}_m) = \mathcal{O}(1/m)$ floor that does not decay in T. Corollary 8 replaces this with $T^{-1/2} + R_m$. For delta-method-style $R_m = \mathcal{O}(1/m^2)$, the asymptotic-in-T floor is $\mathcal{O}(1/m^2)$ — a full power of $1/m$ smaller than the un-tracked $\mathcal{O}(1/m)$ floor. And for any finite T, the $T^{-1/2}$ term provides an explicit decay rate towards that residual.

The tuning balance. The tuning (145) is chosen so that:

- The contraction term $(1 - \eta\widetilde{\mu}/2)^T$ is at most $e^{-(c_\eta\widetilde{\mu}/2)T^{1/4}}$.
- The squared timescale-gap term $(\eta/\gamma)^2$ and the noise-filter term γV_m^b both scale as $T^{-1/2}$ for fixed m.
- The slow-variance term $\eta(\widetilde{B}_m^2 + \widetilde{V}_m)$ is $O(T^{-3/4})$ and therefore decays faster.

Logarithmic horizon tuning. The same quadratic tracking bound permits a different balance between initial-condition decay and tracking error, giving a faster $T^{-2/3}$ rate up to a logarithmic factor.

Corollary 9 (Two-timescale debiasing with logarithmic horizon tuning)**.** Under the assumptions of Theorem 11, fix $c_\eta \geq 2/\widetilde{\mu}$ and $c_\gamma > 0$. For a horizon $T > 1$, set

$$\eta = c_\eta \frac{\log T}{T}, \qquad \gamma = c_\gamma \left(\frac{\log T}{T}\right)^{2/3}, \tag{148}$$

and hold these step sizes constant for the T iterations. For fixed m, there exist constants C and T_0, independent of T, such that for every $T \geq T_0$ the step sizes satisfy (136) and the iterates obey

$$\mathbb{E}\,\|r_T - \theta^\star\|^2_{\Sigma^+} \leq C\left[R_m + \left(\frac{\log T}{T}\right)^{2/3}\right]. \tag{149}$$

Proof. Both step sizes tend to zero as $T \to \infty$, and

$$\frac{\eta\widetilde{\mu}}{\gamma} = \frac{c_\eta \widetilde{\mu}}{c_\gamma}\left(\frac{\log T}{T}\right)^{1/3} \longrightarrow 0.$$

Thus we may choose T_0 so that, for every $T \geq T_0$, $\eta \leq \eta_{\max}$, $\gamma \leq 1$, and $\eta\widetilde{\mu} \leq \gamma/2$. The derivation of (147) uses only these step-size restrictions before substituting the particular powers of T. Hence that bound also applies to (148).

For $a = 1 - \eta\widetilde{\mu}/2$, the initial-condition factor satisfies

$$a^T \leq e^{-\eta\widetilde{\mu}T/2} = T^{-c_\eta\widetilde{\mu}/2} \leq T^{-1}.$$

The coefficient $Q_0 + C(\eta/\gamma)U_1$ in (147) is bounded independently of T, so this controls both initial-error contributions. The remaining tracking and variance terms satisfy

$$\left(\frac{\eta}{\gamma}\right)^2 = \frac{c_\eta^2}{c_\gamma^2}\left(\frac{\log T}{T}\right)^{2/3}, \quad \gamma V_m^b = c_\gamma V_m^b \left(\frac{\log T}{T}\right)^{2/3}, \quad \eta(\widetilde{B}_m^2 + \widetilde{V}_m) = c_\eta(\widetilde{B}_m^2 + \widetilde{V}_m)\frac{\log T}{T}.$$

For fixed m, both T^{-1} and $\log T/T$ are $O((\log T/T)^{2/3})$. Substituting these estimates into (147) proves (149). □

Tuned step sizes. With $\eta = \mathcal{O}(\gamma/m)$ and γ constant, $\mathcal{E}_T = \mathcal{O}(1/m + R_m)$ (after burn-in). For the delta-method estimator with $R_m = \mathcal{O}(1/m^2)$, the stronger choice $\eta = \mathcal{O}(\gamma/m^2)$ gives $\mathcal{E}_T = \mathcal{O}(1/m^{3/2})$ after the burn-in $N \geq (2/\gamma)\log m$ described above. Over T retained iterations, the averaged-iterate bound becomes

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_T)] = O\big(\Omega^2/(T\eta)\big) + O\big(\eta/m\big) + O\big(\Omega\sqrt{\sigma_{\max}}/m^{3/2}\big).$$

Choosing $\eta \asymp \min\{1, \sqrt{m/T}, \gamma/m^2\}$ gives (144), including the $m^2/(\gamma T)$ term from the tracking cap. This fixed-γ tuning in m is distinct from the horizon-dependent choice (145).

I.1.5 An operator-averaging variant with a T^{-1} rate

The $T^{-2/3}$ rate of Corollary 7 can be improved by averaging the sampled operator as well as tracking its bias. Following the operator-averaging idea of Doan (2025), we introduce an additional state s_t that tracks $\mathbb{E}[\widehat{F}_m(r_t)] = F_{\mathcal{R}}(r_t) + b_m(r_t)$, and move the policy using $s_t - \xi_t$. Starting from $r_0 \in D$ and deterministic $s_0, \xi_0 \in \mathbb{R}^d$, for $t \geq 0$ set

$$r_{t+1} = \Pi_D\big(r_t - \eta_t \Sigma(s_t - \xi_t)\big), \tag{150}$$

$$s_{t+1} = (1 - \kappa_t)s_t + \kappa_t \widehat{F}_m^{(t)}(r_t),$$

$$\xi_{t+1} = (1 - \gamma_t)\xi_t + \gamma_t \widehat{b}_m^{(t)}(r_t). \tag{151}$$

The two estimates are computed from one fresh preference batch at r_t. All right-hand sides use the states at time t, so the policy step is determined before that batch is sampled. The averages are of operator values. This variant uses one batch per iteration, with the bias correction computed from that same batch.

Corollary 10 (Two-timescale bias correction with operator averaging)**.** Suppose $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone and $\ell_{\mathcal{R}}$-Lipschitz, the oracle satisfies (O1)–(O2), and the bias estimator satisfies (B1)–(B3). Use the constraint set D and equilibrium $\theta^\star$ from Theorem 10. In (150)–(151), choose

$$\eta_t = \frac{c_\eta}{t+t_0}, \qquad \gamma_t = \kappa_t = \frac{c_f}{t+t_0}, \qquad c_\eta \widetilde{\mu} > 4,$$

with c_f and $t_0 > 0$ sufficiently large, as specified in the proof. For fixed m, there are constants C_0, C_1 independent of T such that, for all $T \geq 1$,

$$\mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 \leq \frac{C_0}{T+t_0} + C_1 R_m^2. \tag{152}$$

Both averaging steps and the policy step decay as $1/t$, with a sufficiently small constant ratio c_η/c_f. The result gives the T^{-1} rate in expected squared distance up to the residual bias of the correction estimator. For $R_m = O(m^{-2})$, its residual term is $O(m^{-4})$.

Proof. Write $n_t = t + t_0$, $\alpha_t = \gamma_t = \kappa_t$, and

$$A_m(r) := F_{\mathcal{R}}(r) + b_m(r), \qquad e_t^F := s_t - A_m(r_t), \qquad e_t^b := \xi_t - b_m(r_t).$$

The two mean maps A_m and b_m are Lipschitz with constants $\ell_{\mathcal{R}} + L_b$ and L_b, respectively. Define

$$q_t := \|r_t - \theta^\star\|_{\Sigma^+}^2, \qquad h_t := \|e_t^F\|_2^2 + \|e_t^b\|_2^2, \qquad Q_t := \mathbb{E}\,q_t, \quad S_t := \mathbb{E}\,h_t.$$

Since $s_t - \xi_t = F_{\mathcal{R}}(r_t) + e_t^F - e_t^b$ and $\Sigma F_{\mathcal{R}}(\theta^\star) = 0$, nonexpansiveness of Π_D, norm equivalence, and strong monotonicity give

$$\begin{aligned} \|r_{t+1} - r_t\|_2^2 &\leq \eta_t^2\big(2\sigma_{\max} L_G^2 q_t + 4\sigma_{\max}^2 h_t\big), \\ q_{t+1} &\leq (1 - \widetilde{\mu}\eta_t + 2L_G^2\eta_t^2) q_t + \left(\frac{2\sigma_{\max}}{\widetilde{\mu}}\eta_t + 4\sigma_{\max}\eta_t^2\right) h_t. \end{aligned} \tag{153}$$

For the second inequality, the cross-term is bounded by $2\sqrt{2\sigma_{\max} q_t h_t} \leq \widetilde{\mu} q_t + 2\sigma_{\max} h_t/\widetilde{\mu}$.

Conditionally on the history before the batch at r_t, write

$$\widehat{F}_m^{(t)}(r_t) = A_m(r_t) + \zeta_t, \qquad \widehat{b}_m^{(t)}(r_t) = b_m(r_t) + \delta_t + \nu_t,$$

where ζ_t, ν_t have conditional mean zero and conditional second moments bounded by V_m, V_m^b, and $\|\delta_t\|_2 \leq R_m$. Put $\Delta A_t = A_m(r_{t+1}) - A_m(r_t)$ and $\Delta b_t = b_m(r_{t+1}) - b_m(r_t)$. Both increments are determined before the batch is sampled, and

$$e_{t+1}^F = (1-\alpha_t) e_t^F - \Delta A_t + \alpha_t \zeta_t, \qquad e_{t+1}^b = (1-\alpha_t) e_t^b - \Delta b_t + \alpha_t \delta_t + \alpha_t \nu_t.$$

For $0 < \alpha \leq 1$, Young's inequality gives $\|(1-\alpha)e + z\|_2^2 \leq (1-\alpha)\|e\|_2^2 + 3\|z\|_2^2/\alpha$. Applying this separately to the two errors yields

$$\begin{aligned} S_{t+1} \leq (1-\alpha_t) S_t + \frac{3(\ell_{\mathcal{R}} + L_b)^2 + 6L_b^2}{\alpha_t}\,\mathbb{E}\,\|r_{t+1} - r_t\|_2^2 \\ + 6\alpha_t R_m^2 + \alpha_t^2 (V_m + V_m^b). \end{aligned}$$

The noises may be correlated within the batch: the argument uses their separate conditional second moments and introduces no product between them.

Combining with (153) and taking $\eta_t \leq \min\{1, \widetilde{\mu}/(4L_G^2)\}$ gives, for a constant $K > 0$ depending only on the Lipschitz and monotonicity constants and the fixed preconditioner,

$$\begin{aligned} Q_{t+1} &\leq (1 - \widetilde{\mu}\eta_t/2) Q_t + K\eta_t S_t, \\ S_{t+1} &\leq \left(1 - \alpha_t + K\frac{\eta_t^2}{\alpha_t}\right) S_t + K\frac{\eta_t^2}{\alpha_t} Q_t + 6\alpha_t R_m^2 + \alpha_t^2 (V_m + V_m^b). \end{aligned}$$

Let $p := c_\eta \widetilde{\mu}/4 > 1$, and choose

$$c_f \geq \max\left\{c_\eta,\ \frac{4Kc_\eta}{\widetilde{\mu}},\ Kc_\eta + 2p\right\}.$$

Take t_0 large enough that $\alpha_t \le 1$, the above bound on η_t holds, and $n_t \ge \max\{1, p\}$ for all $t \ge 0$. Adding the two recurrences, with $\mathcal{V}_t := Q_t + S_t$, now gives

$$\mathcal{V}_{t+1} \le \left(1 - \frac{p}{n_t}\right)\mathcal{V}_t + \frac{6c_f R_m^2}{n_t} + \frac{c_f^2(V_m + V_m^b)}{n_t^2}. \tag{154}$$

Subtract $H := 6c_f R_m^2/p$ and apply induction to $(\mathcal{V}_t - H)_+$. Since $p > 1$,

$$\mathcal{V}_T \le H + \frac{C_0}{T + t_0}, \qquad C_0 := \max\left\{t_0\mathcal{V}_0, \frac{c_f^2(V_m + V_m^b)}{p-1}\right\}.$$

Indeed, $C_0/n - (pC_0 - c_f^2(V_m + V_m^b))/n^2 \le C_0/(n+1)$. As $Q_T \le \mathcal{V}_T$, this proves (152) with $C_1 = 6c_f/p$. □

The constant residual in Corollary 10 can be removed by increasing the preference-batch size during training. Because the two averaging steps are equal, their difference $s_t - \xi_t$ averages the corrected operator estimate directly. We can therefore analyze this difference relative to $F_{\mathcal{R}}$, avoiding additional drift terms from the changing b_{m_t}.

Corollary 11 (Operator averaging with growing preference batches)**.** Use the strongly monotone, Lipschitz operator $F_{\mathcal{R}}$, constraint set D, and equilibrium $\theta^\star$ of Corollary 10. Suppose the conditional oracle bounds (O1)–(O2) and estimator bounds (B1)–(B2) hold for every $m \ge 2$, with constants $C_R, \overline{V}$ independent of m such that

$$R_m \le \frac{C_R}{m^2}, \qquad V_m + V_m^b \le \overline{V}.$$

Run (150)–(151) using a fresh batch of size

$$m_t = \max\left\{2, \left\lceil m_0(t + t_0)^{1/4}\right\rceil\right\}, \qquad m_0 > 0,$$

at iteration t. Choose

$$\eta_t = \frac{c_\eta}{t + t_0}, \qquad \gamma_t = \kappa_t = \frac{c_f}{t + t_0}, \qquad c_\eta\widetilde{\mu} > 4,$$

with $c_f, t_0 > 0$ sufficiently large as specified in the proof. There is a constant C independent of T such that, for all $T \ge 1$,

$$\mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 \le \frac{C}{T + t_0}. \tag{155}$$

Proof. Set $n_t = t + t_0$, $\alpha_t = \gamma_t = \kappa_t$, and $d_t := s_t - \xi_t$. Subtracting the two averaging updates gives

$$d_{t+1} = (1 - \alpha_t)d_t + \alpha_t\big(\widehat{F}_{m_t}^{(t)}(r_t) - \widehat{b}_{m_t}^{(t)}(r_t)\big).$$

Conditionally on the history before the batch, the corrected estimate can be written as

$$\widehat{F}_{m_t}^{(t)}(r_t) - \widehat{b}_{m_t}^{(t)}(r_t) = F_{\mathcal{R}}(r_t) - \delta_t + \epsilon_t, \qquad \|\delta_t\|_2 \le R_{m_t},$$

where ϵ_t has conditional mean zero and conditional second moment at most $2(V_{m_t} + V_{m_t}^b) \le 2\overline{V}$. This bound allows the two estimates to share a batch.

Define $e_t := d_t - F_{\mathcal{R}}(r_t)$, $Q_t := \mathbb{E}\,\|r_t - \theta^\star\|_{\Sigma^+}^2$, and $S_t := \mathbb{E}\,\|e_t\|_2^2$. Since r_{t+1} is determined before sampling,

$$e_{t+1} = (1 - \alpha_t)e_t - \big(F_{\mathcal{R}}(r_{t+1}) - F_{\mathcal{R}}(r_t)\big) - \alpha_t\delta_t + \alpha_t\epsilon_t.$$

Applying the projection and Young inequalities from the preceding proof to the policy direction $d_t = F_{\mathcal{R}}(r_t) + e_t$ gives

$$Q_{t+1} \le (1 - \widetilde{\mu}\eta_t/2)Q_t + K_F\eta_t S_t,$$
$$S_{t+1} \le \left(1 - \alpha_t + K_F\frac{\eta_t^2}{\alpha_t}\right)S_t + K_F\frac{\eta_t^2}{\alpha_t}Q_t + 6\alpha_t R_{m_t}^2 + 2\overline{V}\alpha_t^2,$$

provided $\alpha_t \le 1$ and $\eta_t \le \min\{1, \widetilde{\mu}/(4L_G^2)\}$. Here $K_F > 0$ depends only on the Lipschitz and monotonicity constants of $F_{\mathcal{R}}$ and the fixed preconditioner, and is independent of m_t. Choose c_f, t_0

as in the preceding proof, with K replaced by K_F, and put $p := c_\eta \widetilde{\mu}/4 > 1$. Because $m_t \geq m_0 n_t^{1/4}$, $R_{m_t}^2 \leq C_R^2/(m_0^4 n_t)$. Thus $\mathcal{U}_t := Q_t + S_t$ satisfies

$$\mathcal{U}_{t+1} \leq \left(1 - \frac{p}{n_t}\right)\mathcal{U}_t + \frac{6c_f C_R^2/m_0^4 + 2c_f^2 \overline{V}}{n_t^2}. \tag{156}$$

The residual and variance terms now have the same order in n_t. Induction, using $p > 1$, gives

$$\mathcal{U}_T \leq \frac{C}{T + t_0}, \qquad C := \max\left\{t_0 \mathcal{U}_0,\ \frac{6c_f C_R^2/m_0^4 + 2c_f^2 \overline{V}}{p-1}\right\}.$$

Since $Q_T \leq \mathcal{U}_T$, this proves (155). □

This is an iteration-rate guarantee with growing batches. For fixed m_0, t_0, the total preference-sample budget is $\sum_{t=0}^{T-1} m_t = \Theta(T^{5/4})$.

I.1.6 Summary: comparison of methods and step-size regimes

The table compares the step-size regimes of standard TT-EG with the operator-averaging variant. All error bounds are in expected squared distance, and $n_t := t + t_0$. The sample budget m is fixed except in the final row, which uses growing batches. The last column gives the residual scaling for smooth plug-in risks with $R_m = O(m^{-2})$ and $V_m^b = O(m^{-3})$; the constant-step entry uses the tuning and burn-in described below.

Method and schedule	**Step sizes**	**Expected squared error**	**Initial-error decay**	**Residual in m**
TT-EG: decreasing (Cor. 7)	$\eta_t = c_\eta/n_t$ $\gamma_t = c_\gamma/n_t^{2/3}$	$O(T^{-2/3} + R_m)$	$O(n_T^{-p})$	$O(m^{-2})$
TT-EG: constant (Thm. 11)	η, γ constant	$O(\eta/\gamma + R_m + \sqrt{\gamma V_m^b})$	$(1 - \eta\widetilde{\mu})^T$	$O(m^{-3/2})$ (tuned)
TT-EG: horizon (polynomial; Cor. 8)	$\eta = c_\eta T^{-3/4}$ $\gamma = c_\gamma T^{-1/2}$	$O(T^{-1/2} + R_m)$	$\exp(-cT^{1/4})$	$O(m^{-2})$
TT-EG: horizon (logarithmic; Cor. 9)	$\eta = c_\eta \frac{\log T}{T}$ $\gamma = c_\gamma(\frac{\log T}{T})^{2/3}$	$O((\frac{\log T}{T})^{2/3} + R_m)$	$T^{-c},\ c \geq 1$	$O(m^{-2})$
Operator averaging (Cor. 10)	$\eta_t = c_\eta/n_t$ $\gamma_t = \kappa_t = c_f/n_t$	$O(T^{-1} + R_m^2)$	$O(n_T^{-p})$	$O(m^{-4})$
Operator averaging growing batches (Cor. 11)	$\eta_t = c_\eta/n_t$ $\gamma_t = \kappa_t = c_f/n_t$ $m_t \asymp n_t^{1/4}$	$O(T^{-1})$	$O(n_T^{-p})$	No floor

Details of the comparison table. The initial-error column gives the decay factor for the policy's initial error. For decreasing-step TT-EG, $p = c_\eta\widetilde{\mu}/2 > 2/3$; for operator averaging, $p = c_\eta\widetilde{\mu}/4 > 1$. The same factor controls the initial averaging errors in (154), and the initial error of $s_0 - \xi_0$ in (156). In both horizon-tuned rows, $c = c_\eta\widetilde{\mu}/2$, and the same factor controls the tracker's initial error by (147); the logarithmic tuning takes $c \geq 1$. The constant-step row reports the steady-state floor; the tracker's initial error is accounted for by χ_T in (139). Its $O(m^{-3/2})$ entry uses $\eta = O(\gamma/m^2)$ with fixed γ, and the tracking burn-in $N \geq (2/\gamma)\log m$ from (143), and the full bound also includes the slow-variance term $6\eta(\widetilde{B}_m^2 + \widetilde{V}_m)/\widetilde{\mu}$.

The bounds distinguish the error reduced by running longer from the residual bias controlled by the sample budget m. Constant steps give geometric decay of the policy's initial error, but the bound retains contributions from tracking lag and stochastic noise. Decreasing steps make these contributions vanish with T, yielding $O(T^{-2/3} + R_m)$ without requiring a prescribed training horizon. Horizon tuning instead chooses fixed steps for each prescribed budget T, trading initial-condition

decay against the final tracking and noise terms. For the decreasing and horizon-tuned TT−EG rows, the bound approaches the $O(R_m)$ residual as T grows. Reducing that residual requires a more accurate bias estimator, obtained here by increasing m. Thus, the schedules offer different transient behavior while separating the effects of training duration and preference-sample budget. The final row increases m_t during training, removing the constant residual term while retaining the T^{-1} iteration rate.

Comparison with known rates and contribution. Standard TT−EG attains the $T^{-2/3}$ temporal rate established for classical nonlinear two-timescale stochastic approximation (Doan, 2021). Adapting that analysis to our setting requires handling the residual bias R_m of the correction estimator and the additional variance V_m^b introduced by estimating the bias. Both propagate through the coupled tracking and policy updates. In (123), these contributions appear separately as $C_2\gamma_t R_m^2$ and $C_3\gamma_t^2 V_m^b$; the slow update also retains the original oracle variance. Our proof separates the persistent residual from the tracking and variance terms that vanish with T, giving $O(T^{-2/3} + R_m)$. This control of an imperfect, noisy bias correction distinguishes the argument from the centered oracle-noise setting of Doan (2021).

The operator-averaging variant improves the temporal rate to T^{-1}, matching the rate of the faster nonlinear two-timescale method of Doan (2025). Here too, the remaining conditional bias requires an additional argument beyond the unbiased-oracle analysis. We control the sum of squared policy and tracking errors, absorbing the residual cross-terms into the quadratic contraction terms. The resulting recurrence (154) has a term of order $R_m^2/(t + t_0)$ from residual bias and a term of order $(V_m + V_m^b)/(t + t_0)^2$ from sampling variance. Unrolling this recurrence yields $O(T^{-1} + R_m^2)$, preserving the T^{-1} temporal rate while explicitly controlling the effect of the imperfect correction. It also matches the T^{-1} rate of unbiased stochastic gradient descent for smooth strongly convex objectives (Bach & Moulines, 2011) and stochastic extragradient for strongly monotone operators (**?**). For the fixed-batch rows of the table, these temporal-rate comparisons keep m fixed; the growing-batch row uses the increasing preference-sample budget specified in Corollary 11.

The contribution is to recover these temporal rates while correcting the state-dependent bias created by finite-sample evaluation of nonlinear risk functionals. The correction follows the changing policy and is computed from the same preference samples as the operator estimate. Our analysis accommodates both the residual bias of this correction and the dependence between estimates that share a batch. In particular, the operator-averaging variant retains the classical stochastic first-order temporal rate while reducing the remaining error through m. Delta-method and jackknife corrections give $R_m = O(m^{-2})$, so the respective residual terms in the two bounds are $O(m^{-2})$ and $O(m^{-4})$. This separates the cost of reaching the residual neighborhood from the sample budget needed to make that neighborhood small.

I.2 Numerical illustration: a Bradley-Terry preference game

We illustrate the bias-floor reduction predicted by Theorems 10–11 on a controlled toy instance where the population-level operator $F_{\mathcal{R}}$ is known analytically, allowing direct comparison against an oracle baseline. The experiments are organized to exhibit the two regimes of the constant-step floor decomposition (141): the *bias-dominated* regime at small m where two timescale extragradient offers the largest improvement, and the *variance-dominated* regime at large m where the two algorithms approach the same noise floor.

Setup. We construct a Bradley-Terry preference game on $n = 20$ responses with synthetic latent rewards $\{r(y_i)\}_{i=1}^n$ drawn i.i.d. from $\mathcal{N}(0, 1)$. The risk-neutral preference matrix is $P_{ij} = \sigma(r(y_i) - r(y_j))$ where σ is the logistic, and the risk-adjusted operator $P_{\mathcal{R}}$ is the entropic-risk-distorted version with $\lambda = 6.0$ (a high-risk regime where the bias constants $e^{2\lambda} \approx 1.6 \times 10^5$ are non-trivial; cf. Section G). The KL temperature is $\beta = 0.6$, satisfying the strong-monotonicity condition $\overline{\lambda}_{\mathcal{R}} \leq \beta/2 - \epsilon$ from Theorem 6 given the spread of the latent rewards. We compare three algorithms:

- **Oracle Extra-Gradient (extragradient):** stochastic extragradient using the exact population operator $F_{\mathcal{R}}$ with no bias ($B_m = 0$), serving as a lower-bound baseline.

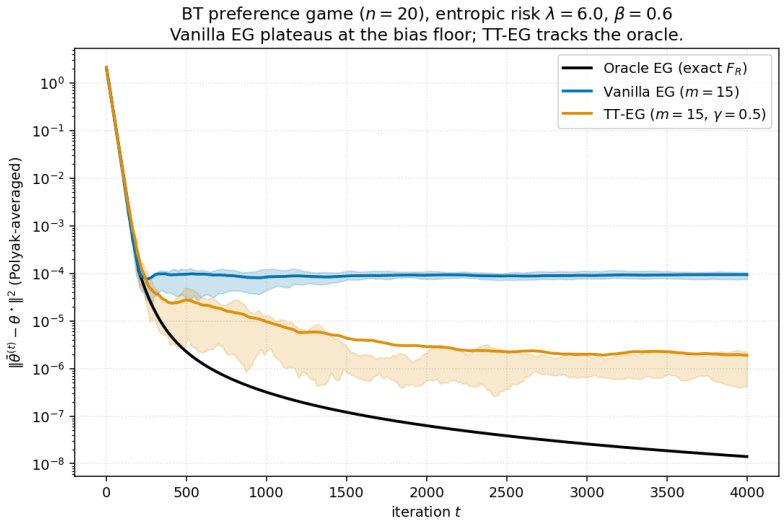


Figure 3: **Trajectory at fixed $m = 15$.** Polyak-averaged squared error $\|\bar{\theta}^{(t)} - \theta^\star\|^2$ vs. iteration t. Oracle extragradient (black) decays at the unbiased $\mathcal{O}(1/T)$ rate of Theorem 8. Vanilla extragradient (blue) descends with the oracle for the first ~ 200 iterations, then plateaus at the bias floor $\sim 10^{-4}$ predicted by Theorem 9 and stays flat thereafter. TT- extragradient (orange) breaks through this floor and drives the error down to $\sim 2 \times 10^{-6}$, a $50\times$ improvement, in the direction of the oracle trajectory. Shaded bands are ± 1 standard deviation over 20 random seeds.

- **Vanilla extragradient (un-tracked):** stochastic extragradient using the plug-in empirical operator $\widehat{F}_m$ with batch size m and no bias correction. This is the algorithm analyzed in Theorem 9, with persistent bias floor $\mathcal{O}(\widetilde{B}_m^2) = \mathcal{O}(1/m^2)$.
- **Two timescale extragradient (`TT-EG`, Algorithm 1):** the two-timescale debiased algorithm with the delta-method bias estimator (121) and constant fast-timescale rate $\gamma = 0.5$.

All three algorithms use the same slow-timescale rate $\eta = 0.04$ and Polyak-averaged iterates $\bar{\theta}^{(t)}$.

Trajectory at fixed m (Figure 3). At $m = 15$ samples per iteration the bias is large. Figure 3 shows three behaviors:

- **Oracle extragradient** decays toward zero at the rate predicted by the un-biased strongly-monotone bound (Theorem 8 with $\widetilde{B}_m = 0$); after $T = 4000$ iterations it has reached $\sim 10^{-8}$, with no sign of plateauing.
- **Vanilla extragradient** initially descends in lockstep with the oracle, then plateaus at $\approx 10^{-4}$ around iteration 200 and remains flat thereafter. This plateau is precisely the persistent $\Theta(\widetilde{B}_m^2) = \Theta(e^{4\lambda}/(\lambda^2 m^2))$ floor of Theorem 9. No amount of additional iteration reduces it.
- **`TT-EG`** initially follows the same trajectory, but rather than plateauing it continues to descend, eventually reaching $\approx 2 \times 10^{-6}$ — a $50\times$ improvement over vanilla extragradient at the same sample budget. The trajectory tracks the oracle's direction (with some loss to the residual R_m and the noise filter) rather than getting trapped at the bias floor.

This is the qualitative effect predicted by the comparison following Theorem 10: the *persistent* bias floor of vanilla extragradient is replaced by an asymptotic-in-T floor at the much smaller $\mathcal{O}(R_m + \eta/\gamma + \sqrt{\gamma}\sqrt{V_m^b})$ level of (141).

Bias-dominated regime: small m (Figure 4, left). At small batch sizes the un-tracked algorithm is bias-limited. The vanilla extragradient floor decreases monotonically with m: a least-squares fit of $\log(\text{floor})$ against $\log m$ over the plotted range gives a slope of approximately -2, consistent with the $\Theta(\widetilde{B}_m^2) = \Theta(1/m^2)$ scaling of Theorem 9. This is the bias-dominated branch of the four-term

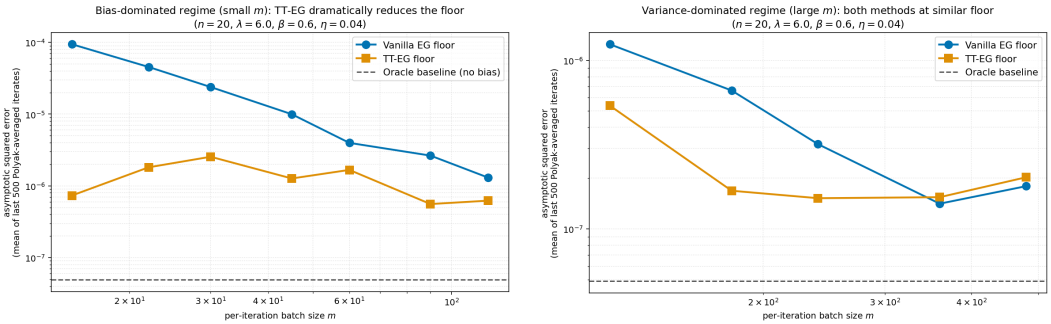


Figure 4: **Asymptotic squared-error floor vs. batch size m, in two regimes.** Floor measured as the mean of the last 500 Polyak-averaged iterates. *Left:* bias-dominated regime ($m \in [15, 130]$). Vanilla extragradient floor scales roughly $\propto 1/m^2$ (matching the $\Theta(\widetilde{B}_m^2)$ prediction), descending more than two orders of magnitude across the range. TT−EG floor is roughly flat near $\sim 10^{-6}$, giving a $\sim 100\times$ reduction at $m = 15$ (the small-m end where vanilla is most bias-dominated) and approaching parity by $m = 130$. *Right:* variance-dominated regime ($m \in [150, 500]$). Both floors descend only modestly across this range and converge toward a common value around $\sim 1.5\times10^{-7}$, dominated by the slow-variance term $\Theta(\eta\widetilde{V}_m)$ in (141); the bias contribution is negligible here. At the largest m TT−EG sits marginally *above* vanilla, reflecting the small noise-filter cost $\Theta(\sqrt{\gamma V_m^b})$ paid to maintain the tracker even when the bias being tracked is already small. Dashed line: oracle baseline at $\sim 4 \times 10^{-8}$.

decomposition (141): the $\Theta(\widetilde{B}_m)$ residual-bias term dominates the $\Theta(\eta\widetilde{V}_m)$ slow-variance term whenever $\widetilde{B}_m^2 \gtrsim \eta\widetilde{V}_m$, which holds throughout this range.

The TT−EG floor is qualitatively different: it is roughly *flat* across the entire range, sitting at $\approx 10^{-6}$. The reduction is largest at small m ($\sim 100\times$ at $m = 15$) and narrows as m grows ($\sim 2\times$ at $m = 130$). This reflects the $\mathcal{E}_T = \mathcal{O}(1/m^{3/2})$ scaling of the constant-step "Tuned step sizes" analysis after Theorem 11: with γ and η fixed, TT−EG's m-dependence is dominated by $\sqrt{\gamma V_m^b} = \mathcal{O}(1/m^{3/2})$, which when squared gives $1/m^3$ — effectively flat over the small-m range plotted, since the small constant pulls the curve below the slow-variance floor that vanilla extragradient eventually reaches as m grows.

The ratio between the two curves matches the bias-floor reduction predicted by the theory. Vanilla floor scales as $\widetilde{B}_m^2 \asymp e^{4\lambda}/(\lambda m)^2$; TT−EG floor scales as $\mathcal{E}_T^2 \asymp \gamma V_m^b \asymp \gamma\, e^{4\lambda}/(\lambda^2 m^3)$ in the noise-filter-dominated regime. The ratio is therefore $\Theta(m/\gamma)$, predicting an improvement that grows linearly with m until the un-tracked floor crosses the slow-variance threshold, after which the gap closes.

Variance-dominated regime: large m (Figure 4, right). Pushing m into $[150, 500]$ moves the system out of the bias-dominated branch and into the variance-dominated branch of (141), where the slow-variance term $\Theta(\eta\widetilde{V}_m)$ dominates. This is exactly the threshold the four-term decomposition predicts: the bias term becomes negligible when $\widetilde{B}_m^2 \lesssim \eta\widetilde{V}_m$, i.e. when $1/(\lambda m)^2 \lesssim \eta/m$, i.e. at $m \gtrsim 1/(\lambda^2\eta)$. For our parameters ($\lambda = 6, \eta = 0.04$) this threshold sits around $m \approx 70$, consistent with where the gap between the two curves closes when comparing the left and right panels of Figure 4.

In this regime, two things change qualitatively:

1. **The vanilla floor stops decreasing as $1/m^2$.** Once the slow-variance term takes over, the floor decays only as $\Theta(\eta\widetilde{V}_m) = \Theta(\eta/m)$, a much slower rate. The right panel shows this: across $m \in [150, 500]$ both floors decrease only modestly, in a regime where the floor is no longer bias-limited.
2. **TT−EG is at parity or marginally worse than vanilla.** This is consistent with the four-term decomposition: the bias term that TT−EG eliminates is small to begin with, and TT−EG pays

a small price ($\Theta(\sqrt{\gamma V_m^b})$ from the noise filter, plus a minor $\Theta(\eta/\gamma)$ timescale-gap term) for maintaining the bias tracker even when the bias being tracked is negligible. At the largest m, the TT−EG floor sits slightly *above* the vanilla floor. This is not a failure mode but an honest expression of the trade-off: bias correction costs noise, and once there is no bias to correct, the noise cost remains.

Both floors approach but do not reach the oracle's noise floor at $\sim 4 \times 10^{-8}$ (dashed); the residual gap is the slow-variance term $\eta\widetilde{V}_m$ that all three algorithms share but is invisible to the oracle's bias-free comparison only because the oracle uses a smaller variance baseline (no plug-in noise from $\widehat{F}_m$).

What the three panels confirm together. The three panels validate the structural claims of Theorems 10–11:

1. *Persistent bias floor in the un-tracked algorithm.* Figure 3 (vanilla curve plateauing) shows that the $\Theta(\widetilde{B}_m^2)$ floor of Theorem 9 is not just an artifact of the proof: it is the genuine asymptotic behavior of the un-tracked algorithm.
2. *Bias-floor reduction in the bias-dominated regime.* Figure 3 (TT−EG breaking through at $m = 15$) and Figure 4 (left, gap between curves at small m) show that two-timescale debiasing delivers the predicted $\sqrt{m}$ to m improvement, replacing the $\Theta(1/m^2)$ floor with one near $\Theta(1/m^3)$.
3. *Crossover to variance-dominated regime.* Figure 4 (right) shows that the four-term decomposition's prediction—that beyond a threshold m^*, bias correction provides no benefit and may pay a small noise-filter cost—is borne out empirically. This validates the trade-off structure in Theorem 11 not as a theoretical artifact but as a practical guide.

Practical implication. The crossover behavior in Figure 4 suggests a clean deployment heuristic: use two-timescale debiasing when one suspects the un-tracked algorithm is bias-limited (small m relative to the problem's bias constants $e^{2\lambda}/\lambda$), and use vanilla extragradient when one is in the variance-limited regime (large m). In the NLHF / RLHF setting, where m is typically the per-iteration batch size and is constrained by GPU memory rather than statistical considerations, the bias-dominated regime is often the practically relevant one, and any improvement at small m translates directly to fewer iterations needed to reach a given error tolerance.

Caveats. This is a toy instance designed to expose the bias-floor structure; quantitative numbers should not be over-interpreted. In particular: (i) the Bradly-Terry setting is risk-neutral at the level of the latent reward, so the entropic risk distortion is purely a stress-test of the bias constants rather than reflecting an actual risk preference; (ii) the strong-monotonicity gap $\beta - 2\overline{\lambda}_{\mathcal{R}}$ is comfortable here, putting us safely in the Theorem 10 regime; (iii) the dimension $n = 20$ is small and the problem is well-conditioned. At larger scale and in the merely-monotone regime (Theorem 12), the asymptotic $\mathcal{O}(R_m)$ vanishing is no longer available, and the practical improvement is the $\mathcal{O}(1/m^{3/2})$ vs. $\mathcal{O}(1/m)$ floor of the constant-step analysis. The qualitative picture—i.e., vanilla plateaus, two timescale extragradient breaks through the floor in the bias-dominated regime, and the two methods converge in the variance-dominated regime—should however persist.

I.3 Stackelberg interpretation: a follower tracking joint Nash play

As described earlier in this section there is an natural Stackelberg structure between the outer game on θ (NLHF) and the inner game wherein ξ is tracking the leaders' deployed θ. Below we describe the structure in more detail.

Outer game: NLHF self-play. The slow-timescale extragradient updates implement symmetric two-player self-play on the KL-regularized risk-adjusted game (4), whose player-1 payoff is

$$V^{(1)}_{\beta,\mathcal{R}}(\pi_1, \pi_2) = \pi_1^\top P_{\mathcal{R}} \pi_2 - \beta\,\mathrm{KL}(\pi_1 \| \pi_{\mathrm{ref}}) + \beta\,\mathrm{KL}(\pi_2 \| \pi_{\mathrm{ref}}),$$

where $P_{\mathcal{R}}$ is the risk-adjusted preference operator (Section 2.1, equation (3)) and π_{ref} is the reference policy. At iterate $\theta^{(t)}$, the leader's bias-corrected stochastic IPO direction is

$$\frac{2}{\beta}\,\Sigma\big(\widehat{F}_m(\theta^{(t)}) - \xi\big), \tag{157}$$

with the factor $2/\beta$ absorbed into the step size in Algorithm 1. Here $\mu = \mathrm{sg}[\pi_{\theta^{(t)}}]$, where $\mathrm{sg}[\cdot]$ is the standard self-play stop-gradient that freezes the opponent's distribution at the current $\pi_{\theta^{(t)}}$. This is the move that makes online IPO well-defined as a single-step optimization (Calandriello et al., 2024a, Sections 4.1–4.2), and it is the same move extragradient and Nash-mirror descent use. Removing the stop-gradient and chain-ruling through $\mu = \pi_\theta$ would change the algorithm into the total derivative of the IPO loss along the self-play diagonal, including the derivative through the opponent distribution.

By the population IPO gradient structure,

$$\nabla_\theta \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\theta; \rho, \mathrm{sg}[\pi_\theta]) = 2\,\Sigma\big(\theta - \theta_{\mathrm{ref}} - \tfrac{1}{\beta} P_{\mathcal{R}}\,\pi_\theta\big) = \frac{2}{\beta}\,\Sigma F_{\mathcal{R}}(\theta).$$

Since ξ estimates operator bias, the correction is applied inside Σ in (157). With exact bias correction, the population stationary condition on $\theta_{\mathrm{ref}} + \mathcal{W}$ is $\Sigma F_{\mathcal{R}}(\theta) = 0$. Under the response-pair coverage condition $\ker \Sigma = \mathrm{span}\{\mathbf{1}\}$ from Section C.3, this is the risk-adjusted equilibrium condition, up to the logit gauge:

$$\theta^\star_{\mathcal{R}} = \theta_{\mathrm{ref}} + \frac{P_{\mathcal{R}}\,\pi^\star_{\mathcal{R}}}{\beta} + c\mathbf{1}, \qquad c \in \mathbb{R}.$$

The risk-adjustment shifts where this Nash sits relative to the risk-neutral equilibrium. Even when shared translation-equivariant payoff risk preserves the operator-level constant-sum identity needed to lift self-play fixed points to Nash equilibria, $P_{\mathcal{R}}$ is generally nonlinear and its parameter-space Jacobian determines the monotonicity bound (Theorem 6). Under the preceding convergence conditions, self-play with stop-gradient and exact bias correction converges to this shifted equilibrium; risk simply changes which fixed point is targeted.

Inner game: bias-tracking follower. The bias-tracker ξ is the follower. Treating the leader's current θ as a parameter, the follower minimizes the quadratic tracking objective

$$\Psi_{\mathrm{follower}}(\xi; \theta) := \tfrac{1}{2}\|\xi - b_m(\theta)\|_2^2,$$

whose unique minimizer is the leader-dependent best-response curve

$$\xi^\star(\theta) = b_m(\theta).$$

The follower takes a Robbins-Monro step toward a noisy, possibly biased sample $\widehat{b}_m(\theta)$ of $\xi^\star(\theta)$ on a fast timescale. The mean target is

$$\mathbb{E}[\widehat{b}_m(\theta)] = b_m(\theta) + r_m(\theta), \qquad \|r_m(\theta)\|_2 \le R_m,$$

as in the tracking proof. The fast timescale realizes the ideal Stackelberg best-response asymptotically when $R_m = 0$ and the step sizes satisfy (122); otherwise the R_m residual remains. For frozen θ, the follower contracts in mean toward this mean target with factor $1 - \gamma$, and the slow leader update (rate $\eta \ll \gamma$) leaves the follower "almost" on its best-response curve.

The asymmetry is structural: the follower's best response depends on the leader's θ, so the follower must respond to leader updates. The leader, by contrast, only *uses* the follower's current ξ via the debiased gradient (157); it does not need to predict where ξ is heading. This is the leader-commits-follower-reacts asymmetry of Stackelberg.

Joint equilibrium. With exact follower best responses ($R_m = 0$), a stationary pair for the combined mean dynamics is $(\theta^\dagger, \xi^\dagger)$ satisfying:

$$\xi^\dagger = \xi^\star(\theta^\dagger) = b_m(\theta^\dagger) \qquad \text{(follower best-responds)}, \tag{158}$$

$$\Sigma\big(\mathbb{E}[\widehat{F}_m(\theta^\dagger)] - \xi^\dagger\big) = 0 \qquad \text{(leader stationary, bias-corrected)}. \tag{159}$$

Substituting (158) into the leader's gradient: since $\mathbb{E}[\widehat{F}_m(\theta)] = F_{\mathcal{R}}(\theta) + b_m(\theta)$, the expected debiased direction at $(\theta^\dagger, \xi^\dagger)$, with $2/\beta$ absorbed into the step size, equals

$$\Sigma\big(\mathbb{E}[\widehat{F}_m(\theta^\dagger)] - b_m(\theta^\dagger)\big) = \Sigma F_{\mathcal{R}}(\theta^\dagger).$$

Condition (159) therefore reduces to $\Sigma F_{\mathcal{R}}(\theta^\dagger) = 0$, i.e., $F_{\mathcal{R}}(\theta^\dagger) \in \ker \Sigma = \mathcal{W}^\perp$, which, under the coverage condition above, is the equilibrium condition for the outer game. Hence

$$(\theta^\dagger, \xi^\dagger) = (\theta^\star_{\mathcal{R}},\, b_m(\theta^\star_{\mathcal{R}})), \tag{160}$$

the equilibrium of the outer game paired with the perfect bias estimate at the equilibrium. The two fixed-point conditions coincide at the equilibrium: follower best-response and leader stationarity together pin down the risk-adjusted equilibrium paired with its bias. For a bias estimator with $r_m \neq 0$, the follower's stationary mean instead satisfies $\xi^\dagger = b_m(\theta^\dagger) + r_m(\theta^\dagger)$, and the leader condition becomes

$$\Sigma\big(F_{\mathcal{R}}(\theta^\dagger) - r_m(\theta^\dagger)\big) = 0.$$

Thus exact coincidence in (160) is the ideal case; the preceding theorems quantify the effect of the R_m residual.

Tracking error as a Stackelberg gap. The tracking error

$$e^*_{\tau-1} := \xi_{\tau-1} - \xi^\star(w_\tau) = \xi_{\tau-1} - b_m(w_\tau)$$

is the follower's deviation from instantaneous Stackelberg best-response to the leader's current state w_τ. The strongly-monotone analysis controls this Stackelberg gap in two ways: Theorem 10(i) gives an asymptotic envelope $\limsup_\tau \sqrt{V_\tau} \le 2\sqrt{3}\, R_m$ under decaying step sizes (driving the gap to its irreducible R_m residual), and Theorem 11(i) gives a non-asymptotic finite-horizon envelope under constant steps. Theorem 10(ii) then shows how this Stackelberg gap enters the leader's convergence rate via the per-step recurrence (126). Driving $e^*_{\tau-1} \to 0$ in mean square (asymptotic regime, $R_m = 0$) or controlling it explicitly (finite-horizon regime) is exactly the question of how well the follower realizes its Stackelberg best-response.

I.3.1 Monotone-only case: averaged-iterate convergence

Theorems 10–**??** all require strong monotonicity of $F_{\mathcal{R}}$: the contraction factor $(1 - \eta\widetilde{\mu})$ in the slow-timescale recurrence is what allows the tracking-error contribution to be accumulated into a finite geometric sum (and, in the time-varying case, to be expressed as a weighted average that concentrates on the trajectory's tail via Lemmas 15–17). With the parameter-space modulus $\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}}$, $\bar{\lambda}_{\mathcal{R}} < \beta$ gives strong monotonicity, while the boundary $\bar{\lambda}_{\mathcal{R}} = \beta$ guarantees only monotonicity. When only monotonicity is available, there is no positive contraction modulus, so the strongly-monotone proof does not apply. Here we give a parallel result under **(P-mono)**, using the averaged-iterate analysis already developed for the un-tracked case in Theorem 8.

The monotone analysis trades two things relative to the strongly-monotone case: convergence is on the averaged iterate $\widehat{z}_T := T^{-1}\sum_{\tau=N+1}^{N+T} w_\tau$ after a fixed burn-in of $N \ge 0$ iterations, rather than the last iterate r_T; and the convergence object is the variational-inequality gap $\mathrm{Gap}_{\mathrm{vi}}$ rather than the squared distance to $\theta^\star$. As a consequence, the bias-floor term will appear inside a uniform sup $\mathcal{E}_T := \sup_{N<\tau\le N+T} \max\{\sqrt{U_\tau}, \sqrt{V_\tau}\}$ over the retained iterations. Both errors are included because the two oracle calls evaluate at $r_{\tau-1}$ and w_τ, respectively. The constant-step tracking bound gives a persistent contribution; the use of a supremum does not itself rule out vanishing tracking contributions under other schedules.

Tracking error in the monotone regime. We need a finite-horizon mean-square bound on the tracking error that does not rely on strong monotonicity of $F_{\mathcal{R}}$. The analysis from Theorem 11(i) applies almost verbatim: the fast-timescale tracker recurrence depends on the leader iterates only through the extragradient step bounds $\|r_\tau - r_{\tau-1}\|_2 \le \eta\sigma_{\max}G$ and $\|w_\tau - r_\tau\|_2 \le 2\eta\sigma_{\max}G$, both of which use only the bounded-gradient assumption $\|\widetilde{g}^{(\tau,i)}\|_2 \le G$ and non-expansiveness of Π_D, neither of which requires strong monotonicity. We restate the result in a form convenient for the monotone analysis below.

Lemma 18 (Mean-square tracking bound, monotone regime)**.** Suppose (B1)–(B3) hold, and that the the debiased gradient norm is bounded along the trajectory:

$$\|\widetilde{g}^{(\tau,i)}\|_2 = \|\widehat{F}_m^{(\tau,i)} - \xi_{\tau-1}\|_2 \le G \quad \text{for all } \tau, i.$$

The tracking errors of Algorithm 1 with constant step sizes $\eta > 0$, $\gamma \in (0, 1]$ satisfy, for $t \geq 1$,

$$U_t \leq (1 - \gamma/2)^{t-1} U_1 + U_\infty, \quad \text{with} \quad U_\infty := \frac{2C_1\eta^2}{\gamma^2} + 2C_2 R_m^2 + 2C_3\gamma V_m^b, \tag{161}$$
$$\sqrt{V_t} \leq \sqrt{U_t} + L_b\sigma_{\max} G\,\eta$$

with $C_1 = 7C_{\text{drift}}^2 = 63L_b^2\sigma_{\max}^2 G^2$, $C_2 = 6$, $C_3 = 2$, and $C_{\text{drift}} = 3L_b\sigma_{\max}G$ as in Theorem 11.

Proof. The recurrence $U_{t+1} \leq (1 - \gamma/2)U_t + C_1\eta^2/\gamma + C_2\gamma R_m^2 + C_3\gamma^2 V_m^b$ of Theorem 10(i) is derived in Step 1 of its proof using only:

a. Lipschitz of b_m ((B3)),

b. the extragradient step bounds $\|r_\tau - r_{\tau-1}\|_2 \leq \eta\sigma_{\max}G$, $\|w_\tau - r_\tau\|_2 \leq 2\eta\sigma_{\max}G$ (from non-expansiveness of Π_D and the gradient bound G),

c. assumptions (B1)–(B2) on $\widehat{b}_m$.

None of (a)–(c) uses strong (or even merely) monotonicity of $F_{\mathcal{R}}$, so the recurrence holds verbatim in the monotone regime. Telescoping with constant contraction factor $1 - \gamma/2$ gives $U_t \leq (1 - \gamma/2)^{t-1}U_1 + U_\infty$. The transfer bound (124) then gives (161). Thus the proof of Theorem 11(i) applies with identical constants. □

In particular, taking square roots and applying $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$:

$$\limsup_{t\to\infty} \sqrt{V_t} \leq \sqrt{U_\infty} + L_b\sigma_{\max}G\,\eta \leq \frac{K_1\eta}{\gamma} + K_2 R_m + K_3\sqrt{\gamma V_m^b},$$

where $K_1 = \sqrt{2C_1} + L_b\sigma_{\max}G$, $K_2 = \sqrt{2C_2}$, and $K_3 = \sqrt{2C_3}$, as in (138).

Theorem 12 (Two timescale extragradient, monotone case, biased oracle with tracking)**.** Suppose $F_{\mathcal{R}}$ is monotone (not necessarily strongly monotone) and $\ell_{\mathcal{R}}$-Lipschitz on $\theta_{\text{ref}} + \mathcal{W}$, the oracle satisfies (O1)-(O2), and the bias estimator satisfies (B1)-(B3). Run `TT-EG` (Algorithm 1) with constant step sizes η, γ satisfying $0 < \eta \leq 1/(\sqrt{3}\,L_G)$ with $\gamma \in (0, 1]$, where $L_G = \sigma_{\max}\ell_{\mathcal{R}}$. Assume the debiased gradient norm is bounded: $\|\widetilde{g}^{(\tau,i)}\|_2 \leq G$ uniformly along the trajectory. Fix a burn-in length $N \geq 0$. Then for all $T \geq 1$, the averaged iterate $\widehat{z}_T := T^{-1}\sum_{\tau=N+1}^{N+T} w_\tau$ satisfies

$$\mathbb{E}[\text{Gap}_{\text{vi}}(\widehat{z}_T)] \;\leq\; \frac{2\,\mathbb{E}[\Theta(r_N)]}{T\eta} \;+\; \frac{7\eta}{2}\big(\sigma_{\max}\mathcal{E}_T^2 + \widetilde{V}_m\big) \;+\; 2\Omega\sqrt{\sigma_{\max}}\,\mathcal{E}_T, \tag{162}$$

where $\Theta(r) := \max_{u\in D} V(r, u)$ and $\mathcal{E}_T := \sup_{N<\tau\leq N+T} \max\{\sqrt{U_\tau}, \sqrt{V_\tau}\}$ is the worst-case root-mean-square tracking error at either oracle query. By Lemma 18,

$$\mathcal{E}_T \;\leq\; \sqrt{U_1}(1 - \gamma/2)^{N/2} + \frac{K_1\eta}{\gamma} + K_2 R_m + K_3\sqrt{\gamma V_m^b}. \tag{163}$$

Here $U_1 \leq B_m^2$. Taking $N = 0$ includes all iterations; if also $r_0 = \theta_{\text{ref}}$, the initial term in (162) is $\Omega^2/(T\eta)$. For general N, $2\,\mathbb{E}[\Theta(r_N)] \leq 4\Omega^2$.

Comparison to the un-tracked case. Theorem 8 (un-tracked) had bound $2\Theta(r_0)/(T\eta) + (7\eta/2)(\widetilde{B}_m^2 + \widetilde{V}_m) + 2\Omega\widetilde{B}_m$, with a persistent $\mathcal{O}(\widetilde{B}_m) = \mathcal{O}(B_m)$ bias floor that does not decay in T. Theorem 12 replaces both occurrences of $\widetilde{B}_m$ by $\sqrt{\sigma_{\max}}\,\mathcal{E}_T$, where, after controlling the initialization term in (163), $\mathcal{E}_T$ scales as $\eta/\gamma + R_m + \sqrt{\gamma}\sqrt{V_m^b}$ rather than as B_m. The initial term is evaluated at r_N for the retained window. This is the same kind of bias-floor reduction as in Theorem 10: with $\eta = \mathcal{O}(\gamma/m)$ and γ constant, $\mathcal{E}_T = \mathcal{O}(1/m + R_m + 1/m^{3/2}) = \mathcal{O}(1/m)$ for delta-method estimators (driven by the $1/m$ timescale-gap term); pushing further to $\eta = \mathcal{O}(\gamma/m^2)$ drives the timescale gap below the $\mathcal{O}(1/m^{3/2})$ noise filter, giving $\mathcal{E}_T = \mathcal{O}(1/m^{3/2})$ — the same factor-$\sqrt{m}$ improvement over the un-tracked $\mathcal{O}(1/m)$ floor after $N \geq (2/\gamma)\log m$ burn-in, as derived in the strongly-monotone constant-step setting (cf. the "Tuned step sizes" paragraph following Theorem 11).

Comparison to the strongly-monotone case. The structural difference between Theorem 12 and Theorem 10 is where the tracking error $\sqrt{V_\tau}$ enters the bound:

- **Strongly-monotone, time-varying step (Thm. 10):** $\sqrt{V_\tau}$ enters via a contraction-weighted average $\sum_\tau w_\tau^T \sqrt{V_\tau}$, which inherits the trajectory's $\limsup$ rather than its $\sup$ (via Lemma 17). Under decaying $\eta_t/\gamma_t \to 0$, the bound is asymptotically $\mathcal{O}(R_m)$—i.e., vanishing in T to the irreducible bias-of-bias-estimator residual.
- **Strongly-monotone, constant step (Thm. 11):** $\sqrt{V_\tau}$ enters via a geometric weighting that is closed-form summable, giving the four-term steady-state floor of (141) (timescale gap, residual bias, noise filter, slow variance), persistent in T.
- **Monotone, constant step (Thm. 12):** $\sqrt{V_\tau}$ enters as a uniform $\sup_\tau$, giving the persistent floor $2\Omega\sqrt{\sigma_{\max}}\,\mathcal{E}_T$ in (162), persistent in T.

The monotone-only analysis gives an averaged-iterate $\mathcal{O}(1/T)$ optimization term in addition to the tracking and variance contributions, rather than last-iterate linear contraction. The tracking error appears as a uniform sup rather than inside a geometric weighting; the resulting term $2\Omega\sqrt{\sigma_{\max}}\mathcal{E}_T$ has no vanishing-in-T multiplier. The four-term floor structure of Theorem 11 survives qualitatively (trade-off between η/γ, R_m, $\sqrt{\gamma}$, η), but the asymptotic vanishing under decaying step sizes does not follow from this uniform-supremum bound.

Proof of Theorem 12. The proof parallels Theorem 8, with the noise/bias quantities recomputed for the debiased gradient $\widetilde{g}^{(\tau,i)} = \widehat{F}_m^{(\tau,i)} - \xi_{\tau-1}$. Write $\widetilde{G}^{(\tau,i)} := \Sigma\widetilde{g}^{(\tau,i)}$ for its preconditioned version. All sums below run from $\tau = N+1$ to $N+T$, and the ghost recursion (66) starts from $y_N = r_N$ with increments $\eta\widetilde{\Delta}_\tau$.

Step 1: Deterministic-style bound. Apply Lemma 12 with scaled debiased gradients $\eta\widetilde{G}^{(\tau,1)}$ and $\eta\widetilde{G}^{(\tau,2)}$. The proof of Theorem 7 carries through verbatim with $\widehat{F}_m^{(\tau,i)}$ replaced by $\widetilde{g}^{(\tau,i)}$, giving the same $\Gamma(t)$ structure as in (76):

$$\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_T) \le \frac{\Gamma(T)}{T\eta},$$

$$\Gamma(T) = 2\Theta(r_N) + \frac{3\eta^2}{2}\sum_\tau \widetilde{\varepsilon}^2_{r_{\tau-1}} + 2\eta^2\sum_\tau \widetilde{\varepsilon}^2_{w_\tau} + \sum_\tau \langle \eta\widetilde{\Delta}_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+}.$$

Here $\widetilde{\Delta}_\tau := G(w_\tau) - \widetilde{G}^{(\tau,2)}$, $\widetilde{\varepsilon}_{r_{\tau-1}} := \|\widetilde{G}^{(\tau,1)} - G(r_{\tau-1})\|_{\Sigma^+}$, and $\widetilde{\varepsilon}_{w_\tau} := \|\widetilde{G}^{(\tau,2)} - G(w_\tau)\|_{\Sigma^+}$ are the errors of the preconditioned debiased oracle.

Step 2: Conditional moments of the debiased oracle. Next, we decompose $\widetilde{g}^{(\tau,2)} = \widehat{F}_m^{(\tau,2)} - \xi_{\tau-1}$ with $\xi_{\tau-1} = b_m(w_\tau) + e^*_{\tau-1}$ to get that

$$\widetilde{g}^{(\tau,2)} = \big(\widehat{F}_m^{(\tau,2)} - b_m(w_\tau) - F_{\mathcal{R}}(w_\tau)\big) + F_{\mathcal{R}}(w_\tau) - e^*_{\tau-1}.$$

Conditioning on on $\mathcal{F}_\tau$, the mean of $\widehat{F}_m^{(\tau,2)} - b_m(w_\tau) - F_{\mathcal{R}}(w_\tau)$ is zero by definition of b_m. Hence we have that

$$\mathbb{E}[\widetilde{g}^{(\tau,2)} - F_{\mathcal{R}}(w_\tau) \mid \mathcal{F}_\tau] = -e^*_{\tau-1},$$

i.e., the conditional bias of the preconditioned debiased oracle is $-\Sigma e^*_{\tau-1}$, with Σ^+-norm $\le \sqrt{\sigma_{\max}}\|e^*_{\tau-1}\|_2$. The conditional variance is $\mathrm{Var}(\widehat{F}_m^{(\tau,2)} \mid \mathcal{F}_\tau) \le V_m$. Translating to Σ^+-norm gives $\mathrm{Var}(\widetilde{G}^{(\tau,2)} \mid \mathcal{F}_\tau) \le \sigma_{\max} V_m = \widetilde{V}_m$.

Step 3: Bound $\mathbb{E}\,\widetilde{\varepsilon}^2$. By bias-variance decomposition in Σ^+ norm, we have that

$$\mathbb{E}\,\widetilde{\varepsilon}^2_{w_\tau} \le \sigma_{\max} V_\tau + \widetilde{V}_m \le \sigma_{\max}\mathcal{E}_T^2 + \widetilde{V}_m.$$

At $r_{\tau-1}$, the analogous decomposition conditional on $\mathcal{G}_{\tau-1}$ uses $\xi_{\tau-1} - b_m(r_{\tau-1}) = e_{\tau-1}$:

$$\mathbb{E}\,\widetilde{\varepsilon}^2_{r_{\tau-1}} \le \sigma_{\max} U_\tau + \widetilde{V}_m \le \sigma_{\max}\mathcal{E}_T^2 + \widetilde{V}_m.$$

Summing the noise contributions and dividing by $T\eta$, we deduce that

$$\frac{1}{T\eta}\sum_\tau \Big[\frac{3\eta^2}{2}\,\mathbb{E}\,\widetilde{\varepsilon}^2_{r_{\tau-1}} + 2\eta^2\,\mathbb{E}\,\widetilde{\varepsilon}^2_{w_\tau}\Big] \le \frac{7\eta}{2}\big(\sigma_{\max}\mathcal{E}_T^2 + \widetilde{V}_m\big).$$

Step 4: bound the cross term $\mathbb{E}\langle \eta\widetilde{\Delta}_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+}$. Following Bound 2 of Theorem 8, by the tower property $\mathbb{E}\langle \eta\widetilde{\Delta}_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+} = \eta\,\mathbb{E}\langle \mathbb{E}[\widetilde{\Delta}_\tau \mid \mathcal{F}_\tau], w_\tau - y_{\tau-1}\rangle_{\Sigma^+}$. By Step 2, $\|\mathbb{E}[\widetilde{\Delta}_\tau \mid \mathcal{F}_\tau]\|_{\Sigma^+} = \|\Sigma e^*_{\tau-1}\|_{\Sigma^+} \le \sqrt{\sigma_{\max}}\|e^*_{\tau-1}\|_2$. Cauchy-Schwarz with the Σ^+-diameter $\|w_\tau - y_{\tau-1}\|_{\Sigma^+} \le 2\Omega$:

$$|\,\mathbb{E}\langle \eta\widetilde{\Delta}_\tau, w_\tau - y_{\tau-1}\rangle_{\Sigma^+}| \le 2\eta\Omega\sqrt{\sigma_{\max}}\;\mathbb{E}\,\|e^*_{\tau-1}\|_2 \le 2\eta\Omega\sqrt{\sigma_{\max}}\,\mathcal{E}_T,$$

using Jensen. Summing and dividing by $T\eta$ contributes $2\Omega\sqrt{\sigma_{\max}}\,\mathcal{E}_T$.

Step 5: combine. $\mathbb{E}\,\Gamma(T)/(T\eta)$ is bounded by $2\,\mathbb{E}[\Theta(r_N)]/(T\eta)$ plus the noise contribution from Step 3 plus the cross-term contribution from Step 4, giving (162). The bound (163) on $\mathcal{E}_T$ follows from Lemma 18: both $\sqrt{U_\tau}$ and $\sqrt{V_\tau}$ are bounded by $\sqrt{U_1}(1-\gamma/2)^{(\tau-1)/2} + K_1\eta/\gamma + K_2 R_m + K_3\sqrt{\gamma V^b_m}$, and $\tau - 1 \ge N$ over the retained window. □

Tuned rate. Substituting the tuned $\mathcal{E}_T = \mathcal{O}(1/m^{3/2})$ from the "Tuned step sizes" paragraph after Theorem 11 (with $\eta = \mathcal{O}(\gamma/m^2)$, fixed γ, and $N \ge (2/\gamma)\log m$ burn-in):

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_T)] = O\big(\Omega^2/(T\eta)\big) + O\big(\eta\widetilde{V}_m\big) + O\big(\Omega\sqrt{\sigma_{\max}}/m^{3/2}\big),$$

where the third term is the reduced bias floor. With $\widetilde{V}_m = \mathcal{O}(1/m)$, the variance/initial-term balance is $\eta \asymp \sqrt{m/T}$. Enforcing the Lipschitz and tracking caps gives $\eta \asymp \min\{1, \sqrt{m/T}, \gamma/m^2\}$, with a sufficiently small fixed multiplier, and hence

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_T)] = \mathcal{O}\left(\frac{1}{\sqrt{Tm}} + \frac{1}{T} + \frac{m^2}{\gamma T}\right) + \mathcal{O}\left(\frac{\Omega\sqrt{\sigma_{\max}}}{m^{3/2}}\right).$$

This is the merely-monotone analog of (144); the rate to the bias floor is the same in both regimes, but the floor itself is persistent in T here (the asymptotic-vanishing conclusion of Theorem 10 does not follow from this uniform-supremum bound).

Implications by risk type. To aid in translating these results to practical use cases, let us consider some common risk metrics.

- **Entropic risk:** $R_m = \mathcal{O}(e^{2\lambda}/(\lambda m^2))$, $V^b_m = \mathcal{O}(e^{4\lambda}/(\lambda^2 m^3))$. For fixed $\lambda > 0$, two timescale extragradient drives the persistent $\mathcal{O}(e^{2\lambda}/(\lambda m))$ floor of the un-tracked monotone case down to $\mathcal{O}(\mathcal{E}_T) = \mathcal{O}(m^{-3/2})$ under the stated step-size tuning and burn-in, with constants depending on λ. This is a half-power improvement in m, matching the strongly-monotone constant-step case. The $\mathcal{O}(m^{-3/2})$ noise-filter term dominates the $\mathcal{O}(R_m) = \mathcal{O}(m^{-2})$ residual under this fixed-γ tuning.
- **CVaR via oracle-threshold Rockafellar-Uryasev:** $R_m = V^b_m = 0$ when we take $\widehat{b}_m \equiv 0$, since $b_m \equiv 0$. With $\xi_0 = 0$, the tracker stays identically zero, so $\mathcal{E}_T = 0$ exactly. Two timescale extragradient then reduces to un-tracked extragradient, recovering Theorem 8 with $\widetilde{B}_m = 0$ (and $N = 0$).

I.4 Two-Timescale Mirror Descent: Corollary from extra-gradient analysis

Mirror descent updates the policy using one oracle evaluation at the current iterate, while the bias tracker continues to run on the faster timescale. The tracking analysis therefore carries over from extragradient, with a simpler drift term. The policy update uses the one-call descent inequality from Section H.10. Below we state the results for the same three regimes and give a short proof sketch with each result, reusing the preceding estimates.

Algorithm and common assumptions. Use the same constraint set D containing $\theta^\star$, fixed preconditioner Σ, and Σ^+ geometry as above. Assume (O1)–(O2), (B1)–(B3), and the trajectory bound $\|\widehat{F}^{(t)}_m - \xi_{t-1}\|_2 \le G$ used in the extragradient analysis. Starting from $r_0 \in D$ and $\xi_0 = 0$, sample $\widehat{F}^{(t)}_m, \widehat{b}^{(t)}_m$ at r_{t-1} and update

$$r_t = \Pi_D\big(r_{t-1} - \eta_t\Sigma(\widehat{F}^{(t)}_m - \xi_{t-1})\big) \quad \text{and} \quad \xi_t = (1-\gamma_t)\xi_{t-1} + \gamma_t\widehat{b}^{(t)}_m.$$

Here $\mathcal{F}_t$ denotes the full history through iteration t, including both oracle outputs; the batch at iteration t is fresh conditional on $\mathcal{F}_{t-1}$. Define

$$e_t := \xi_t - b_m(r_t), \qquad V_t := \mathbb{E}\,\|e_{t-1}\|_2^2, \qquad V_1 \le B_m^2.$$

Write $L_G = \sigma_{\max}\ell_{\mathcal{R}}$ and, in the strongly monotone cases, $\widetilde{\mu} = \mu_{\mathcal{R}}\sigma_{\min} > 0$. The tracking constants are $C_1^{\mathrm{md}} = 7L_b^2\sigma_{\max}^2G^2$, $C_2 = 6$, and $C_3 = 2$.

I.4.1 Strongly monotone case, decaying steps

When $\eta_t/\gamma_t \to 0$ and $\gamma_t \to 0$, the contributions from the moving tracking target and the tracker noise vanish, leaving the bias-estimator residual R_m. Strong monotonicity then controls the leader's distance to equilibrium. The additional terms from the one-call update also vanish as $\eta_t \to 0$, so the asymptotic bound matches the extragradient result.

Corollary 12 (Two-timescale mirror descent, strongly monotone, decaying steps)**.** Under the common assumptions, suppose $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone and $\ell_{\mathcal{R}}$-Lipschitz. Let the deterministic step sizes satisfy

$$0 < \eta_t \le \min\left\{\frac{1}{2\widetilde{\mu}}, \frac{1}{2L_G}\right\}, \quad 0 < \gamma_t \le 1, \quad \frac{\eta_t}{\gamma_t} \to 0, \quad \gamma_t \to 0, \quad \sum_{t\ge 1}\eta_t = \infty.$$

Then $\limsup_{t\to\infty}\sqrt{V_t} \le 2\sqrt{3}\,R_m$ and

$$\limsup_{T\to\infty}\mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 \le 8\sqrt{3}\,\Omega\sqrt{\sigma_{\max}}R_m/\widetilde{\mu}. \tag{164}$$

In particular, when $R_m = 0$, $r_T \to \theta^\star$ in mean square.

Proof sketch. The tracker proof from Theorem 10, Step 1, applies with the simpler drift

$$\delta_t^{\mathrm{md}} = b_m(r_{t-1}) - b_m(r_t), \qquad \|\delta_t^{\mathrm{md}}\|_2 \le L_b\sigma_{\max}G\,\eta_t.$$

The same Young-inequality bounds give

$$V_{t+1} \le (1-\gamma_t/2)V_t + \frac{C_1^{\mathrm{md}}\eta_t^2}{\gamma_t} + C_2\gamma_t R_m^2 + C_3\gamma_t^2 V_m^b. \tag{165}$$

This argument allows the two estimators to share a batch. Lemma 14 yields $\limsup_{t\to\infty}\sqrt{V_t} \le 2\sqrt{3}\,R_m$ as claimed.

For the slow iterate, set $\widetilde{G}_m^{(t)} := \Sigma(\widehat{F}_m^{(t)} - \xi_{t-1})$. The conditional bias is $-\Sigma e_{t-1}$, and

$$\mathbb{E}\,\|\widetilde{G}_m^{(t)} - G(r_{t-1})\|_{\Sigma^+}^2 \le \sigma_{\max}V_t + \widetilde{V}_m.$$

Applying the one-call prox inequality as in Corollary 5, using $G(\theta^\star) = 0$ and $\|G(r_{t-1})\|_{\Sigma^+} \le 2L_G\Omega$, gives

$$\mathbb{E}\,V(r_t,\theta^\star) \le (1-\eta_t\widetilde{\mu})\,\mathbb{E}\,V(r_{t-1},\theta^\star) + 2\eta_t\Omega\sqrt{\sigma_{\max}}\sqrt{V_t} + \eta_t^2\left(\sigma_{\max}V_t + \widetilde{V}_m\right) + 4\eta_t^2L_G^2\Omega^2. \tag{166}$$

The squared tracking-error term is retained here. Under decaying steps, V_t is bounded and $\eta_t \to 0$. Thus the terms multiplied by η_t^2 vanish under the weighted-average argument of Lemmas 15–17. Using (**??**) and $\|r_T - \theta^\star\|_{\Sigma^+}^2 = 2V(r_T,\theta^\star)$ gives (164). □

I.4.2 Strongly monotone case, constant steps

For constant steps, the bound retains contributions from the timescale ratio, the bias-estimator residual, and the filtered tracker noise. The next result separates these contributions from the initial policy and tracker errors. It also retains the sampling-variance and Lipschitz terms from the mirror-descent update. Thus the iteration count controls the transient terms, while η and γ control the remaining bound.

Corollary 13 (Two-timescale mirror descent, strongly monotone, constant steps)**.** Under the common assumptions, suppose $F_{\mathcal{R}}$ is $\mu_{\mathcal{R}}$-strongly monotone and $\ell_{\mathcal{R}}$-Lipschitz, and use

$$0 < \eta \le \min\{1/(2\widetilde{\mu}), 1/(2L_G)\}, \qquad 0 < \gamma \le 1.$$

Set $V_\infty^{\mathrm{md}} := \frac{2C_1^{\mathrm{md}}\eta^2}{\gamma^2} + 2C_2R_m^2 + 2C_3\gamma V_m^b$. Then $V_t \le (1-\gamma/2)^{t-1}V_1 + V_\infty^{\mathrm{md}}$ so that

$$\limsup_{T\to\infty} \mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 \le \frac{4\Omega\sqrt{\sigma_{\max}}}{\widetilde{\mu}}\sqrt{V_\infty^{\mathrm{md}}} + \frac{2\eta}{\widetilde{\mu}}\left(\sigma_{\max}V_\infty^{\mathrm{md}} + \widetilde{V}_m + 4L_G^2\Omega^2\right). \tag{167}$$

Proof sketch. The tracking recurrence (165) holds without the decaying-step conditions. With constant η and γ, its geometric sum gives $V_t \le (1-\gamma/2)^{t-1}V_1 + V_\infty^{\mathrm{md}}$. The three terms in V_∞^{md} correspond to the moving target, residual bias, and tracker noise.

For constant steps, substitute $V_t \le V_1(1-\gamma/2)^{t-1} + V_\infty^{\mathrm{md}}$ into (166) and use $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. The geometric-series argument of Theorem 11 yields, for all $T \ge 1$,

$$\begin{aligned}\mathbb{E}\,\|r_T - \theta^\star\|_{\Sigma^+}^2 &\le (1-\eta\widetilde{\mu})^T\|r_0 - \theta^\star\|_{\Sigma^+}^2 \\ &\quad + \frac{4\Omega\sqrt{\sigma_{\max}}}{\widetilde{\mu}}\left(\sqrt{V_1}\chi_T + \sqrt{V_\infty^{\mathrm{md}}}\right) + \frac{2\eta}{\widetilde{\mu}}\left[\sigma_{\max}\left(V_1\chi_T + V_\infty^{\mathrm{md}}\right) + \widetilde{V}_m + 4L_G^2\Omega^2\right],\end{aligned}$$

with χ_T defined by (140). the squared-error transient is also bounded using χ_T since $(1-\gamma/2)^{t-1} \le (1-\gamma/2)^{(t-1)/2}$. Taking $T \to \infty$ gives (167). □

I.4.3 Monotone case, constant steps

With only monotonicity, we use the averaged-iterate VI gap, as in Theorem 12. The tracking estimate is unchanged, but its contribution to the gap bound is controlled uniformly over the retained iterations. A burn-in allows the initial tracking error to decrease before averaging begins. The result below makes that choice explicit through N, with $N = 0$ including every iteration.

Corollary 14 (Two-timescale mirror descent, monotone case, biased oracle with tracking)**.** Under the common assumptions, suppose $F_{\mathcal{R}}$ is monotone and $\ell_{\mathcal{R}}$-Lipschitz. Use $0 < \eta \le 1/L_G$, $0 < \gamma \le 1$. For a fixed burn-in length $N \ge 0$, let

$$\widehat{z}_T := \frac{1}{T}\sum_{t=N+1}^{N+T} r_{t-1} \quad \text{and} \quad \mathcal{E}_T := \sup_{N<t\le N+T}\sqrt{V_t} \le \sqrt{V_1}(1-\gamma/2)^{N/2} + \sqrt{V_\infty^{\mathrm{md}}},$$

where V_∞^{md} is given above. Then, with $\Theta(r) := \max_{u\in D} V(r,u)$,

$$\mathbb{E}[\mathrm{Gap}_{\mathrm{vi}}(\widehat{z}_T)] \le \frac{2\,\mathbb{E}[\Theta(r_N)]}{T\eta} + \frac{3\eta}{2}\left(\sigma_{\max}\mathcal{E}_T^2 + \widetilde{V}_m\right) + 2\Omega\sqrt{\sigma_{\max}}\mathcal{E}_T + 4L_G^2\eta\Omega^2. \tag{168}$$

For $N = 0$ and $r_0 = \theta_{\mathrm{ref}}$, the initial term is $\Omega^2/(T\eta)$; in general, $2\,\mathbb{E}[\Theta(r_N)] \le 4\Omega^2$.

Proof sketch. Apply the one-call prox and ghost-iterate argument from Corollary 3 over the retained iterations, starting the ghost sequence at r_N. At each query r_{t-1}, the corrected oracle has conditional bias $-\Sigma e_{t-1}$ and second moment relative to $G(r_{t-1})$ bounded by $\sigma_{\max}V_t + \widetilde{V}_m$, as in the proof of Corollary 12. These moment bounds do not require strong monotonicity. Substituting $V_t \le \mathcal{E}_T^2$ into the expected-gap argument gives (168), including the initial term $2\,\mathbb{E}[\Theta(r_N)]/(T\eta)$. The envelope on $\mathcal{E}_T$ follows by taking the supremum of the constant-step tracking bound over $N < t \le N + T$. □

Comparison to extragradient. Mirror descent uses one oracle call and has the smaller tracking drift constant $L_b\sigma_{\max}G$. Its slow-iterate bound includes the additional Lipschitz contribution proportional to $\eta L_G^2\Omega^2$ and the squared tracking error. Both vanish under the decaying schedule, so (164) matches the extragradient asymptotic bound. With delta-method scaling, fixed γ, $\eta = \mathcal{O}(\gamma/m^2)$, and $N \ge (2/\gamma)\log m$, the tracking contribution is $\mathcal{O}(m^{-3/2})$. The additional Lipschitz contribution is $\mathcal{O}(m^{-2})$, preserving the half-power improvement in the constant-step floor.

J Stability and Generalization

We now establish that the risk-adjusted QRE depends Lipschitz-continuously on the underlying preference operator. Although our softmax parameterization makes the problem technically unconstrained in θ, the natural framework for this kind of parametric stability is the variational inequality formulation, with Lipschitz dependence following from the Dontchev–Rockafellar strong metric regularity machinery for strongly monotone problems (Dontchev & Rockafellar, 2009a). This framework is more robust than the classical implicit function theorem: it does not require differentiability of the operator at the solution, does not require any non-degeneracy or interiority condition on active constraints, and extends seamlessly to constrained or set-valued policy classes.

We present this proof only for the symmetric self play setting, as it can easily be extended to a general (monotone) risk-adjusted game.

Generalized equation formulation. Identify a policy parameter $\theta \in \mathbb{R}^{|\mathcal{Y}|}$ with the policy $\pi_\theta = \mathrm{softmax}(\theta)$. The problem has a one-dimensional gauge invariance (adding a constant to all entries of θ leaves π_θ unchanged), which we factor out by working on $\mathbb{R}^{|\mathcal{Y}|}$ modulo $\mathbf{1}$, equivalently on the subspace $\mathbf{1}^\perp$. We use $\|\cdot\|$ for the Euclidean norm on this quotient, which equals the Euclidean norm of any representative with zero sum.

Under the strong-monotonicity assumptions below, for a fixed preference operator P, the population risk-adjusted equilibrium $\theta^\star_{\mathcal{R}}(P)$ is the unique solution of the equation

$$0 \;=\; F_{\mathcal{R},P}(\theta) \;:=\; \beta(\theta - \theta_{\mathrm{ref}}) - P_{\mathcal{R}}\,\pi_\theta, \tag{169}$$

viewed as an element of $\mathbb{R}^{|\mathcal{Y}|}/\mathbf{1}$. This is a *generalized equation* in the sense of Robinson (cf. Dontchev & Rockafellar (2009b); Robinson (1980)) with zero set-valued component:

$$0 \;\in\; F_{\mathcal{R},P}(\theta) + N(\theta), \tag{170}$$

where $N(\theta) = \{0\}$ since the constraint set is the whole space (modulo gauge). We retain the formulation (170) since the analysis below transports verbatim to settings where N is the normal cone of a closed convex constraint set on θ (e.g., simplex-constrained or trust-region-clipped parameterizations).

J.1 Structural stability of the solution mapping

Treat $P \mapsto \theta^\star_{\mathcal{R}}(P)$ as a mapping from preference operators (a vector space) to policy parameters. Strong metric regularity of (170) concerns local single-valuedness and Lipschitz dependence under additive perturbations of the equation. We establish dependence on P directly using strong monotonicity.

Theorem 13 (Structural stability of the risk-adjusted equilibrium under operator perturbation)**.** Let P, P' be two preference operators with corresponding risk-adjusted operators $P_{\mathcal{R}}, P'_{\mathcal{R}}$. Suppose their self-play residuals satisfy the strong-monotonicity assumptions of Theorem 6, with $\bar\lambda_{\mathcal{R}}(P), \bar\lambda_{\mathcal{R}}(P') \le \beta - \epsilon$ for some $\epsilon > 0$. Write

$$\mu_{\mathcal{R}} := \beta - \bar\lambda_{\mathcal{R}}(P), \qquad \mu'_{\mathcal{R}} := \beta - \bar\lambda_{\mathcal{R}}(P')$$

for the respective strong-monotonicity moduli, both at least ϵ. Let $\theta^\star = \theta^\star_{\mathcal{R}}(P)$ and $\theta'^\star = \theta^\star_{\mathcal{R}}(P')$ be the corresponding self-play equilibrium parameters, represented with $\mathbf{1}^\top\theta^\star = \mathbf{1}^\top\theta'^\star = 0$. Then

$$\left\|\theta^\star - \theta'^\star\right\| \le \frac{1}{\min\{\mu_{\mathcal{R}}, \mu'_{\mathcal{R}}\}} \left\|(P_{\mathcal{R}} - P'_{\mathcal{R}})(\pi_{\theta'^\star})\right\|_2. \tag{171}$$

Since these operators can be nonlinear, define their uniform perturbation norm on the simplex by

$$\|P_{\mathcal{R}} - P'_{\mathcal{R}}\|_{\mathrm{op}} := \sup_{\pi \in \Delta(\mathcal{Y})} \left\|P_{\mathcal{R}}(\pi) - P'_{\mathcal{R}}(\pi)\right\|_2.$$

Here $\|\cdot\|_2$ is the ordinary Euclidean norm of the operator values; it upper-bounds their quotient norm. In particular, we have the bound

$$\left\|\theta^\star - \theta'^\star\right\| \le \|P_{\mathcal{R}} - P'_{\mathcal{R}}\|_{\mathrm{op}} / \min\{\mu_{\mathcal{R}}, \mu'_{\mathcal{R}}\}. \tag{172}$$

Proof. The argument proceeds in three steps: (i) recast the problem as a generalized equation, (ii) apply strong monotonicity to obtain Lipschitz dependence of the solution on the operator, (iii) translate to the explicit bound (171).

Step 1: Strong monotonicity of $F_{\mathcal{R},P}$. By Theorem 6, the operator $F_{\mathcal{R},P}$ is strongly monotone on the gauge-fixed space $\mathbf{1}^\perp$ in the Euclidean inner product, with modulus $\mu_{\mathcal{R}}$, and hence with modulus at least $\min\{\mu_{\mathcal{R}}, \mu'_{\mathcal{R}}\} \geq \epsilon$. Thus, for all $\theta_1, \theta_2 \in \mathbf{1}^\perp$,

$$\big\langle F_{\mathcal{R},P}(\theta_1) - F_{\mathcal{R},P}(\theta_2),\, \theta_1 - \theta_2 \big\rangle \;\geq\; \min\{\mu_{\mathcal{R}}, \mu'_{\mathcal{R}}\}\, \|\theta_1 - \theta_2\|^2. \tag{173}$$

Step 2: Generalized equation analysis. The two RQREs satisfy

$$0 \;=\; F_{\mathcal{R},P}(\theta^\star) \quad \text{and} \quad 0 \;=\; F_{\mathcal{R},P'}(\theta'^\star).$$

Decompose the difference as follows:

$$\begin{aligned} F_{\mathcal{R},P}(\theta'^\star) - F_{\mathcal{R},P}(\theta^\star) \;&=\; F_{\mathcal{R},P}(\theta'^\star) - 0 \;=\; F_{\mathcal{R},P}(\theta'^\star) - F_{\mathcal{R},P'}(\theta'^\star) \\ &=\; -(P_{\mathcal{R}} - P'_{\mathcal{R}})\, \pi_{\theta'^\star}, \end{aligned} \tag{174}$$

where the last equality uses that $F_{\mathcal{R},P}$ and $F_{\mathcal{R},P'}$ differ only in their dependence on the preference operator: $F_{\mathcal{R},P}(\theta) - F_{\mathcal{R},P'}(\theta) = -(P_{\mathcal{R}} - P'_{\mathcal{R}})\pi_\theta$.

Step 3: Apply strong monotonicity. Take inner product of (174) with $\theta'^\star - \theta^\star$ to get that

$$\big\langle F_{\mathcal{R},P}(\theta'^\star) - F_{\mathcal{R},P}(\theta^\star),\, \theta'^\star - \theta^\star \big\rangle \;=\; -\big\langle (P_{\mathcal{R}} - P'_{\mathcal{R}})\, \pi_{\theta'^\star},\, \theta'^\star - \theta^\star \big\rangle.$$

The left-hand side is bounded below by $\min\{\mu_{\mathcal{R}}, \mu'_{\mathcal{R}}\}\, \|\theta^\star - \theta'^\star\|^2$ by (173). The right-hand side is bounded above (in absolute value) by Cauchy–Schwarz—indeed,

$$\big|\big\langle (P_{\mathcal{R}} - P'_{\mathcal{R}})\pi_{\theta'^\star},\, \theta'^\star - \theta^\star \big\rangle\big| \;\leq\; \big\|(P_{\mathcal{R}} - P'_{\mathcal{R}})\pi_{\theta'^\star}\big\|_2\, \|\theta'^\star - \theta^\star\|.$$

Combining the two bounds we have that

$$\min\{\mu_{\mathcal{R}}, \mu'_{\mathcal{R}}\}\, \|\theta^\star - \theta'^\star\|^2 \;\leq\; \big\|(P_{\mathcal{R}} - P'_{\mathcal{R}})\pi_{\theta'^\star}\big\|_2\, \|\theta'^\star - \theta^\star\|,$$

and dividing by $\|\theta'^\star - \theta^\star\|$ (assuming it is nonzero; otherwise the bound is trivial) yields (171). The uniform perturbation bound (172) follows by taking the supremum over $\pi \in \Delta(\mathcal{Y})$ in the right-hand side of (171); no linearity of $P_{\mathcal{R}}$ or $P'_{\mathcal{R}}$ is required. □

Connection to strong metric regularity. Theorem 13 is a special case of the general *strong metric regularity* property of solution mappings of strongly monotone variational inequalities. Specifically, the canonical perturbation of the generalized equation (170),

$$v \;\in\; F_{\mathcal{R},P}(\theta) + N(\theta), \quad v \in \mathbb{R}^{|\mathcal{Y}|}/\mathbf{1},$$

has solution mapping $S_P : v \mapsto \theta_v$ that is single-valued and globally Lipschitz with constant $1/\mu_{\mathcal{R}}$ on its domain (Dontchev & Rockafellar, 2009a, Theorem 2F.9). The bound (171) corresponds to perturbing the constant term by $v = -(P_{\mathcal{R}} - P'_{\mathcal{R}})\pi_{\theta'^\star}$ and applying this Lipschitz constant. In the present softmax-parameterized problem the constraint set C is trivial ($N \equiv \{0\}$), but the framework extends without modification to (i) projected gradient dynamics on the simplex ($N(\pi) = N_{\Delta(\mathcal{Y})}(\pi)$), (ii) trust-region-clipped policies (box constraints on θ), and (iii) other constrained parameterizations one might use in practice.

KL stability via softmax smoothness. The parameter bound of Theorem 13 yields a policy-space bound through the log-partition function of softmax.

Corollary 15 (KL-stability of RQRE)**.** Under the assumptions of Theorem 13, with respective strong-monotonicity moduli $\mu_{\mathcal{R}}, \mu'_{\mathcal{R}} > 0$, the bound holds:

$$\mathrm{KL}(\pi^\star_{\mathcal{R}}(P) \| \pi^\star_{\mathcal{R}}(P')) \leq \frac{\|P_{\mathcal{R}} - P'_{\mathcal{R}}\|_{\mathrm{op}}^2}{2\min\{\mu_{\mathcal{R}}, \mu'_{\mathcal{R}}\}^2}. \tag{175}$$

Proof. Let $A(\theta) := \log \sum_{y \in \mathcal{Y}} e^{\theta_y}$. Its gradient is $\nabla A(\theta) = \pi_\theta$, and its Hessian satisfies

$$\nabla^2 A(\theta) = \mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top, \qquad v^\top \nabla^2 A(\theta) v \leq \sum_y \pi_\theta(y) v_y^2 \leq \|v\|_2^2.$$

Thus A is globally 1-smooth in the Euclidean norm. The softmax identity for KL divergence and the smoothness inequality give

$$\begin{aligned}\mathrm{KL}(\pi_\theta \| \pi_{\theta'}) &= A(\theta') - A(\theta) - \langle \nabla A(\theta), \theta' - \theta \rangle \\ &\le \tfrac{1}{2}\|\theta' - \theta\|_2^2.\end{aligned} \tag{176}$$

Choose the zero-sum representatives of the equilibrium logits, as in Theorem 13. Its parameter-stability bound gives

$$\|\theta^\star - \theta'^\star\|_2 \le \frac{\|P_\mathcal{R} - P'_\mathcal{R}\|_{\mathrm{op}}}{\min\{\mu_\mathcal{R}, \mu'_\mathcal{R}\}}. \tag{177}$$

Substituting (177) into (176) proves (175). No bounded-logit assumption is needed for the softmax inequality. □

This is the form that plugs directly into the offline sample complexity proof (Section K.5).

Proof of Proposition 3 from the Main Body. For ease of reference we restate the main-body proposition.

Proposition 3. Let $\pi^\star$ and $\pi'^\star$ be the RQRE induced by $\mathcal{P}_\mathcal{R}$ and $\mathcal{P}'_\mathcal{R}$, respectively, and suppose we are in the strongly monotone regime, with respective moduli $\mu_\mathcal{R}, \mu'_\mathcal{R} > 0$. Then $\mathrm{KL}(\pi^\star \| \pi'^\star) \le \|\mathcal{P}_\mathcal{R} - \mathcal{P}'_\mathcal{R}\|_{\mathrm{op}}^2 / (2\min\{\mu_\mathcal{R}, \mu'_\mathcal{R}\}^2)$.

Proof. Apply the parameter-stability argument of Theorem 13 with the two positive moduli assumed in the proposition, followed by (176). This is the bound in Corollary 15. □

Connection to algorithmic stability. For $S \sim \mathcal{D}^n$, let $\mathcal{P}_{\mathcal{R},S}$ be the risk-adjusted operator of the preference model fitted on S. We distinguish Euclidean and coordinatewise perturbations of this operator. In addition to $\|\cdot\|_{\mathrm{op}}$ from Theorem 13, define

$$\|\mathcal{P}_\mathcal{R} - \mathcal{P}'_\mathcal{R}\|_\infty := \sup_{\pi \in \Delta(\mathcal{Y})} \|\mathcal{P}_\mathcal{R}(\pi) - \mathcal{P}'_\mathcal{R}(\pi)\|_\infty.$$

These definitions evaluate the possibly nonlinear operators at π; they do not assume a matrix representation.

Assumption 3 (Coordinatewise preference-model stability)**.** There exists $\zeta_\infty \ge 0$ such that, for every pair of datasets S, S' of size n differing in one element,

$$\|\mathcal{P}_{\mathcal{R},S} - \mathcal{P}_{\mathcal{R},S'}\|_\infty \le \zeta_\infty.$$

The dependence of ζ_∞ on n is a property of the preference-model fitting procedure. Rates such as $O(1/n)$ or $O(n^{-1/2})$ require a uniform replace-one bound of this form; concentration alone does not establish uniform stability.

Theorem 14 (Operator-level stability under sample replacement)**.** Let S, S' differ in one element, and suppose their risk-adjusted operators satisfy the assumptions of Theorem 13, with respective strong-monotonicity moduli $\mu_\mathcal{R}, \mu'_\mathcal{R} > 0$. If their Euclidean operator perturbation is bounded by $\|\mathcal{P}_{\mathcal{R},S} - \mathcal{P}_{\mathcal{R},S'}\|_{\mathrm{op}} \le \zeta_{\mathrm{op}}$, then their equilibrium policies satisfy

$$\mathrm{KL}(\pi^\star_S \| \pi^\star_{S'}) \le \frac{\zeta_{\mathrm{op}}^2}{2\min\{\mu_\mathcal{R}, \mu'_\mathcal{R}\}^2}.$$

Proof. Apply Corollary 15 to $P_\mathcal{R} = \mathcal{P}_{\mathcal{R},S}$ and $P'_\mathcal{R} = \mathcal{P}_{\mathcal{R},S'}$, and use the assumed Euclidean operator perturbation bound. □

Corollary 16 (Coordinatewise stability under sample replacement)**.** Suppose Assumption 3 holds and the two neighboring datasets satisfy the structural assumptions of Theorem 14. Then

$$\mathrm{KL}(\pi^\star_S \| \pi^\star_{S'}) \le \frac{|\mathcal{Y}|\,\zeta_\infty^2}{2\min\{\mu_\mathcal{R}, \mu'_\mathcal{R}\}^2}.$$

Proof. For every $\pi \in \Delta(\mathcal{Y})$, the vector norm comparison gives

$$\|\mathcal{P}_{\mathcal{R},S}(\pi) - \mathcal{P}_{\mathcal{R},S'}(\pi)\|_2 \le \sqrt{|\mathcal{Y}|}\,\|\mathcal{P}_{\mathcal{R},S}(\pi) - \mathcal{P}_{\mathcal{R},S'}(\pi)\|_\infty.$$

Taking the supremum over π and applying Assumption 3 yields $\|\mathcal{P}_{\mathcal{R},S} - \mathcal{P}_{\mathcal{R},S'}\|_{\mathrm{op}} \le \sqrt{|\mathcal{Y}|}\,\zeta_\infty$. Use $\zeta_{\mathrm{op}} = \sqrt{|\mathcal{Y}|}\,\zeta_\infty$ in Theorem 14. □

J.2 From stability to generalization under the risk-adjusted IPO loss

We instantiate the stability-to-generalization conversion with the risk-adjusted IPO loss, which is the natural deployment loss in our setting. For a fixed opponent distribution μ, use the risk-adjusted payoff values $(\mathcal{P}_{\mathcal{R}}\,\mu)_y$ from (3). For a comparison example $z = (y, y')$ drawn from ρ, define

$$\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi; z) \;:=\; \left(\log \frac{\pi(y)\,\pi_{\mathrm{ref}}(y')}{\pi(y')\,\pi_{\mathrm{ref}}(y)} \;-\; \frac{1}{\beta}\big((\mathcal{P}_{\mathcal{R}}\,\mu)_y - (\mathcal{P}_{\mathcal{R}}\,\mu)_{y'}\big)\right)^2. \tag{178}$$

Taking expectations over $z \sim \rho$ recovers the population risk-adjusted IPO objective $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\theta; \rho, \mu) = \mathbb{E}_z[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi_\theta; z)]$ from (7). The empirical counterpart on a sample $S = \{z_i\}_{i=1}^n$ is $\widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\theta) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi_\theta; z_i)$. Here $\mathcal{D} = \rho$ is the comparison-example distribution, and μ and the population operator $\mathcal{P}_{\mathcal{R}}$ are fixed in the evaluation loss.

Assumption 4 (Bounded softmax mass)**.** There exists $p_{\min} > 0$ such that for every θ in the relevant domain (in particular, $\theta = \theta^\star_S$ for any sample S in the support of $\mathcal{D}^n$), the policy $\pi_\theta = \mathrm{softmax}(\theta)$ satisfies $\pi_\theta(y) \geq p_{\min}$ for all $y \in \mathcal{Y}$.

This is mild: under the QRE regularizer $\beta(\theta - \theta_{\mathrm{ref}})$, the optimum $\theta^\star$ lies in a bounded set determined by β and the operator norm of $\mathcal{P}_{\mathcal{R}}$, which yields a deterministic $p_{\min}$ depending on those constants.

Lemma 19 (Explicit $p_{\min}$ from the risk-adjusted equilibrium optimality condition)**.** Suppose the risk-adjusted operator values are uniformly bounded: $|(\mathcal{P}_{\mathcal{R},S}\,\pi)_y| \leq M$ for all $y \in \mathcal{Y}$, $\pi \in \Delta(\mathcal{Y})$, and S in the support of $\mathcal{D}^n$. Suppose further that the reference parameter satisfies $\mathrm{diam}(\theta_{\mathrm{ref}}) := \max_y \theta_{\mathrm{ref}}(y) - \min_y \theta_{\mathrm{ref}}(y) \leq D_{\mathrm{ref}}$. Then for every such dataset S, the risk-adjusted equilibrium solution $\theta^\star_S = \theta^\star_{\mathcal{R}}(\mathcal{P}_{\mathcal{R},S})$ satisfies, after fixing the gauge so that $\sum_y \theta^\star_S(y) = \sum_y \theta_{\mathrm{ref}}(y)$, $\mathrm{diam}(\theta^\star_S) \leq D_{\mathrm{ref}} + \frac{2M}{\beta}$, and consequently

$$\pi_{\theta^\star_S}(y) \;\geq\; p_{\min} \;:=\; \frac{\exp\!\big(-D_{\mathrm{ref}} - 2M/\beta\big)}{|\mathcal{Y}|} \qquad \forall y \in \mathcal{Y}.$$

Proof. The RQRE optimality condition (169) holds modulo an additive constant. Subtracting two coordinates cancels this constant:

$$\theta^\star_S(y_1) - \theta^\star_S(y_2) = \theta_{\mathrm{ref}}(y_1) - \theta_{\mathrm{ref}}(y_2) + \frac{1}{\beta}\left[(\mathcal{P}_{\mathcal{R},S}\,\pi_{\theta^\star_S})_{y_1} - (\mathcal{P}_{\mathcal{R},S}\,\pi_{\theta^\star_S})_{y_2}\right].$$

Each operator value is bounded in absolute value by M, so the triangle inequality gives

$$\mathrm{diam}(\theta^\star_S) \;\leq\; \mathrm{diam}(\theta_{\mathrm{ref}}) + \frac{2M}{\beta} \;\leq\; D_{\mathrm{ref}} + \frac{2M}{\beta}.$$

For the softmax lower bound, observe that for any y, the sequence of lower bounds hold:

$$\pi_{\theta^\star_S}(y) \;=\; \frac{e^{\theta^\star_S(y)}}{\sum_{y'} e^{\theta^\star_S(y')}} \;\geq\; \frac{e^{\theta^\star_S(y)}}{|\mathcal{Y}|\, e^{\max_{y'} \theta^\star_S(y')}} \;=\; \frac{e^{-(\max_{y'} \theta^\star_S(y') - \theta^\star_S(y))}}{|\mathcal{Y}|} \;\geq\; \frac{e^{-\mathrm{diam}(\theta^\star_S)}}{|\mathcal{Y}|}.$$

Combining these observations yields the claim. □

Regularity of the IPO Loss. Now we prove Lipschitz continuity of the risk-adjusted IPO loss in π.

Lemma 20 (Lipschitzness of the risk-adjusted IPO loss in π)**.** Suppose Assumption 4 holds and additionally $\pi_{\mathrm{ref}}(y) \geq p_{\min}$ for all $y \in \mathcal{Y}$. Suppose further that the risk-adjusted preferences are bounded: $|(\mathcal{P}_{\mathcal{R}}\,\mu)_y| \leq M$ for all $y \in \mathcal{Y}$. Then the loss (178) satisfies, for any $\pi, \pi' \in \{q \in \Delta(\mathcal{Y}) : q(y) \geq p_{\min}\ \forall y\}$ and any comparison example $z = (y, y')$,

$$\left|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi; z) - \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi'; z)\right| \;\leq\; L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}}\, \|\pi - \pi'\|_{\mathrm{TV}}, \qquad L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}} \;:=\; \frac{8D}{p_{\min}},$$

where $D := 2\log(1/p_{\min}) + 2M/\beta$ is an upper bound on the absolute value of the residual inside the square in (178).

Proof. Define the residual

$$r(\pi; z) \;:=\; \log \frac{\pi(y)\,\pi_{\mathrm{ref}}(y')}{\pi(y')\,\pi_{\mathrm{ref}}(y)} \;-\; \frac{1}{\beta}\big((\mathcal{P}_{\mathcal{R}}\,\mu)_y - (\mathcal{P}_{\mathcal{R}}\,\mu)_{y'}\big),$$

so that $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi; z) = r(\pi; z)^2$. Split the residual into its π-dependent and π-independent parts:

$$r(\pi; z) \;=\; \underbrace{\log \pi(y) - \log \pi(y')}_{=:\,a(\pi; z)} \;+\; \underbrace{\log \pi_{\mathrm{ref}}(y') - \log \pi_{\mathrm{ref}}(y) - \frac{1}{\beta}\big((\mathcal{P}_{\mathcal{R}}\,\mu)_y - (\mathcal{P}_{\mathcal{R}}\,\mu)_{y'}\big)}_{=:\,c(z)}.$$

We proceed in three steps: bound $|r(\pi; z)|$ uniformly, establish Lipschitzness of $a(\pi; z)$ in π, and combine via the chain rule for the squared loss.

Step 1: Uniform bound on $|r(\pi; z)|$. Since $\log \pi(y), \log \pi(y') \in [\log p_{\min}, 0]$,

$$|a(\pi; z)| \;=\; |\log \pi(y) - \log \pi(y')| \;\le\; \log(1/p_{\min}).$$

By the same argument applied to π_{ref}, $|\log \pi_{\mathrm{ref}}(y') - \log \pi_{\mathrm{ref}}(y)| \le \log(1/p_{\min})$. The preference term is bounded by

$$\frac{1}{\beta}\big|(\mathcal{P}_{\mathcal{R}}\,\mu)_y - (\mathcal{P}_{\mathcal{R}}\,\mu)_{y'}\big| \;\le\; \frac{2M}{\beta}.$$

Adding these contributions by the triangle inequality gives

$$|r(\pi; z)| \;\le\; |a(\pi; z)| + |c(z)| \;\le\; 2\log(1/p_{\min}) + \frac{2M}{\beta} \;=\; D,$$

with $D := 2\log(1/p_{\min}) + 2M/\beta$, as in the lemma statement.

Step 2: Lipschitzness of $a(\pi; z)$ in π. Fix $z = (y, y')$ and consider π, π' in the admissible set. It is immediate that

$$a(\pi; z) - a(\pi'; z) \;=\; \big(\log \pi(y) - \log \pi'(y)\big) - \big(\log \pi(y') - \log \pi'(y')\big).$$

The function $t \mapsto \log t$ is $1/p_{\min}$-Lipschitz on $[p_{\min}, 1]$, since its derivative $1/t$ is bounded by $1/p_{\min}$ on this interval. Therefore, for any single coordinate $u \in \{y, y'\}$, we have

$$|\log \pi(u) - \log \pi'(u)| \;\le\; \frac{1}{p_{\min}}|\pi(u) - \pi'(u)| \;\le\; \frac{1}{p_{\min}}\|\pi - \pi'\|_\infty.$$

Applying the triangle inequality it is immediate that

$$|a(\pi; z) - a(\pi'; z)| \;\le\; \frac{2}{p_{\min}}\|\pi - \pi'\|_\infty.$$

Since $\|\pi - \pi'\|_\infty \le \|\pi - \pi'\|_1 = 2\|\pi - \pi'\|_{\mathrm{TV}}$, this yields

$$|a(\pi; z) - a(\pi'; z)| \;\le\; \frac{4}{p_{\min}}\|\pi - \pi'\|_{\mathrm{TV}}.$$

Since $c(z)$ does not depend on π, we obtain the same Lipschitz continuity bound for r—indeed, we have

$$|r(\pi; z) - r(\pi'; z)| \;\le\; \frac{4}{p_{\min}}\|\pi - \pi'\|_{\mathrm{TV}}. \tag{179}$$

Step 3: Lipschitzness of $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}} = r^2$. For any reals u, v with $|u|, |v| \le D$, it is immediate that

$$|u^2 - v^2| \;=\; |u + v| \cdot |u - v| \;\le\; 2D\,|u - v|.$$

Applying this to $u = r(\pi; z)$ and $v = r(\pi'; z)$, both bounded in absolute value by D from Step 1, we have that

$$|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi; z) - \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi'; z)| \;=\; |r(\pi; z)^2 - r(\pi'; z)^2| \;\le\; 2D\,|r(\pi; z) - r(\pi'; z)|.$$

Combining this bound with (179) yields

$$|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi; z) - \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi'; z)| \;\le\; 2D \cdot \frac{4}{p_{\min}}\|\pi - \pi'\|_{\mathrm{TV}} \;=\; \frac{8D}{p_{\min}}\,\|\pi - \pi'\|_{\mathrm{TV}}.$$

Setting $L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}} := 8D/p_{\min}$ completes the proof. □

J.3 Algorithmic Stability Generalization Bounds in Expectation

We now convert the coordinatewise stability of Corollary 16 into a generalization guarantee under the risk-adjusted IPO loss. The argument requires two ingredients beyond stability: a bounded admissible policy class on which the IPO loss is Lipschitz, and a Lipschitzness lemma for the loss itself. We state these in turn before the main theorem.

Assumption 5 (Bounded admissible policy class)**.** There exists a class $\Pi \subseteq \Delta(\mathcal{Y})$ such that:

i. $\pi_S^\star \in \Pi$ almost surely for $S \sim \mathcal{D}^n$, and $\pi_{\mathrm{ref}} \in \Pi$;

ii. there exists $L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}} < \infty$ such that for all $\pi, \pi' \in \Pi$ and all comparison examples $z = (y, y')$,

$$|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi; z) - \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi'; z)| \;\le\; L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}} \, \|\pi - \pi'\|_{\mathrm{TV}};$$

iii. there exists $B < \infty$ such that $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi; z) \le B$ for all $\pi \in \Pi$ and all z.

Remark 3 (Sufficient conditions)**.** Assumption 5 is satisfied whenever the admissible policies are bounded away from zero, i.e. $\pi(y) \ge p_{\min}$ for all y and all $\pi \in \Pi$, and the risk-adjusted preferences satisfy $|(\mathcal{P}_{\mathcal{R}}\,\mu)_y| \le M$. In that case $L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}} = 8D/p_{\min}$ and $B = D^2$, with $D := 2\log(1/p_{\min}) + 2M/\beta$, by Lemma 20. Such a uniform $p_{\min}$ exists, for example, on any bounded set in θ-space, which is in turn implied by the RQRE optimality condition $\beta(\theta - \theta_{\mathrm{ref}}) = \mathcal{P}_{\mathcal{R}}\pi_\theta$ together with boundedness of $\mathcal{P}_{\mathcal{R}}$ (see Lemma 19).

Stability Guarantees. The following theorem gives the in-expectation guarantee, and Corollary 17 gives its high-probability counterpart. The main-body constant ζ is denoted ζ_∞ here to distinguish it from ζ_{op}.

Theorem 15 (Stability-based generalization of risk-adjusted equilibrium)**.** Suppose the assumptions of Corollary 16 hold for every neighboring pair of datasets, together with Assumption 5. Let $\mu^{\min}_{\mathcal{R}} > 0$ satisfy $\beta - \bar{\lambda}_{\mathcal{R}}(\mathcal{P}_S) \ge \mu^{\min}_{\mathcal{R}}$ for every dataset S in the support of $\mathcal{D}^n$, where $\mathcal{P}_S$ is the fitted preference operator. Let ζ_∞ be the coordinatewise stability constant of Assumption 3. Then the risk-adjusted equilibrium policy $\pi_S^\star$ satisfies

$$\left|\mathbb{E}_S\left[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi_S^\star) - \widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\pi_S^\star)\right]\right| \;\le\; \beta_n, \qquad \beta_n \;:=\; L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}} \cdot \frac{\sqrt{|\mathcal{Y}|}\,\zeta_\infty}{\sqrt{2}\,\mu^{\min}_{\mathcal{R}}}.$$

In particular, for fixed $L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}}$, $|\mathcal{Y}|$, and $\mu^{\min}_{\mathcal{R}} > 0$, $\zeta_\infty = O(1/n)$ gives an expected IPO generalization gap of $O(1/n)$, while $\zeta_\infty = O(n^{-1/2})$ gives $O(n^{-1/2})$.

Proof. The argument proceeds in three steps: (i) establish uniform stability of the map $S \mapsto \pi_S^\star$ with respect to the IPO loss, (ii) apply the replace-one symmetrization identity, and (iii) bound the resulting expectation.

Step 1: Uniform stability with respect to $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}$. Let

$$S = (z_1, \ldots, z_n) \quad \text{and} \quad S^{(i)} = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)$$

denote two samples differing only in their i-th element. By Corollary 16, with both moduli bounded below by $\mu^{\min}_{\mathcal{R}}$, we have

$$\mathrm{KL}\big(\pi_S^\star \,\big\|\, \pi_{S^{(i)}}^\star\big) \;\le\; \frac{|\mathcal{Y}|\,\zeta_\infty^2}{2(\mu^{\min}_{\mathcal{R}})^2} \;\le\; \frac{|\mathcal{Y}|\,\zeta_\infty^2}{(\mu^{\min}_{\mathcal{R}})^2},$$

where the last inequality retains a conservative constant in β_n. The factor $|\mathcal{Y}|$ arises from the norm conversion in Corollary 16. Pinsker's inequality then gives

$$\|\pi_S^\star - \pi_{S^{(i)}}^\star\|_{\mathrm{TV}} \;\le\; \sqrt{\tfrac{1}{2}\,\mathrm{KL}(\pi_S^\star \,\|\, \pi_{S^{(i)}}^\star)} \;\le\; \frac{\sqrt{|\mathcal{Y}|}\,\zeta_\infty}{\sqrt{2}\,\mu^{\min}_{\mathcal{R}}}.$$

By Assumption 5(ii), for any comparison example z, we have the bound

$$\left|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi_S^\star; z) - \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi_{S^{(i)}}^\star; z)\right| \;\le\; L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}} \cdot \frac{\sqrt{|\mathcal{Y}|}\,\zeta_\infty}{\sqrt{2}\,\mu^{\min}_{\mathcal{R}}} \;=\; \beta_n. \tag{180}$$

This is uniform stability of the RQRE map with parameter β_n.

Step 2: Replace-one symmetrization. Let $S = (z_1, \dots, z_n)$ and $\widetilde{S} = (\widetilde{z}_1, \dots, \widetilde{z}_n)$ be two independent i.i.d. samples from $\mathcal{D}^n$, and for each $i \in \{1, \dots, n\}$ define $S^{(i)} := (z_1, \dots, z_{i-1}, \widetilde{z}_i, z_{i+1}, \dots, z_n)$. The population and empirical risks satisfy

$$\mathbb{E}_S\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)\big] \;=\; \mathbb{E}_{S,\widetilde{z}}\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S; \widetilde{z})\big], \qquad \mathbb{E}_S\big[\widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)\big] \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S; z_i)\big].$$

Since $\widetilde{z}$ in the first expression is i.i.d. from $\mathcal{D}$ and independent of S, we may rename it $\widetilde{z}_i$ for any i without changing the expectation—that is,

$$\mathbb{E}_S\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)\big] \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{S,\widetilde{z}_i}\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S; \widetilde{z}_i)\big].$$

The pair $(S, \widetilde{z}_i)$ has the same joint distribution as $(S^{(i)}, z_i)$—i.e., both consist of $n+1$ i.i.d. draws with one designated as the "held-out" point—so by relabeling,

$$\mathbb{E}_{S,\widetilde{z}_i}\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S; \widetilde{z}_i)\big] \;=\; \mathbb{E}_{S,\widetilde{z}_i}\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z_i)\big].$$

Combining these observations yields

$$\mathbb{E}_S\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S) - \widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)\big] \;=\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,\widetilde{z}_i}\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z_i) - \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S; z_i)\big]. \tag{181}$$

Step 3: Apply uniform stability. The samples S and $S^{(i)}$ differ only in their i-th coordinate (z_i vs. $\widetilde{z}_i$), so by (180) evaluated on $z = z_i$, we have that

$$\big|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z_i) - \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S; z_i)\big| \;\le\; \beta_n \quad \text{a.s.}$$

Taking expectations and applying the triangle inequality to (181) yields

$$\Big|\mathbb{E}_S\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S) - \widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)\big]\Big| \;\le\; \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,\widetilde{z}_i}\big|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z_i) - \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S; z_i)\big| \;\le\; \beta_n. \qquad \square$$

Remark 4 (Explicit constants in the IPO loss bounds)**.** Under the operator boundedness assumptions of Lemmas 19 and 20, and $\mathrm{diam}(\theta_{\mathrm{ref}}) \le D_{\mathrm{ref}}$, Lemma 19 gives

$$p_{\min} \;=\; \frac{e^{-(D_{\mathrm{ref}} + 2M/\beta)}}{|\mathcal{Y}|}.$$

Substituting this into Lemma 20 yields

$$L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}} \;=\; \frac{8D}{p_{\min}} \;=\; 8D\,|\mathcal{Y}|\,e^{D_{\mathrm{ref}} + 2M/\beta}, \qquad \text{where } B \;=\; D^2,$$

and $D = 2\log(1/p_{\min}) + 2M/\beta = 2\big(\log|\mathcal{Y}| + D_{\mathrm{ref}} + 2M/\beta\big) + 2M/\beta$. The dependence on β is exponential through the factor $e^{2M/\beta}$, which is unsurprising: small regularization $\beta \to 0$ allows the RQRE policy to concentrate arbitrarily, blowing up the log-ratio in the IPO loss. The dependence on $|\mathcal{Y}|$ is linear through the explicit factor and logarithmic through D.

J.4 Algorithmic Stability High Probability Generalization Bounds

Theorem 15 bounds the generalization gap in expectation. For a high-probability bound, the standard route is to view $\Phi(S) := \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S) - \widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)$ as a function of the i.i.d. sample S and apply a concentration inequality exploiting bounded differences. The classical Bousquet & Elisseeff (2002) approach uses McDiarmid's inequality, with the difference constant inherited from uniform stability: $|\Phi(S) - \Phi(S^{(i)})| \le 2\beta_n + B/n$. The resulting deviation term is of order $(\sqrt{n}\,\beta_n + B/\sqrt{n})\sqrt{\log(1/\delta)}$. It need not vanish when β_n is of order $n^{-1/2}$, whereas $\beta_n = O(1/n)$ gives the usual $O(n^{-1/2})$ deviation at fixed confidence.

The fast-stability regime is not automatic in our setting. A $\zeta_\infty = O(1/n)$ rate requires a correspondingly stable preference-model fitting procedure. Although the risk-adjusted equilibrium objective itself is KL-regularized at strength β, that regularization controls how the optimal policy $\theta^\star$

responds to a given preference operator, not how the preference operator $\mathcal{P}_{\mathcal{R},S}$ depends on the training sample. These two stabilities are decoupled: β governs the operator-to-policy map (the $1/\mu_{\mathcal{R}}$ constant of Theorem 13), while ζ_∞ governs the sample-to-operator map, which is a property of the preference fitter itself. Whether $\zeta_\infty = O(1/n)$ or $O(n^{-1/2})$ is available requires a uniform stability argument for that fitting procedure. The explicit constants in Remark 4 also depend on β through $e^{2M/\beta}$; changing the policy regularization does not by itself establish a faster replace-one rate for the preference fitter.

When the available stability bound is only $\zeta_\infty = O(n^{-1/2})$, the McDiarmid bound need not decay: the term $n\beta_n\sqrt{\log(1/\delta)/n}$ can remain of constant order. The sharper concentration of Bousquet et al. (2020) exploits the weak correlation between coordinate-wise perturbations of a uniformly stable algorithm. It yields a stability contribution of order $\beta_n \log n \log(1/\delta)$ and a separate sampling term $B\sqrt{\log(1/\delta)/n}$. Thus the high-probability bound decays even under an $O(n^{-1/2})$ stability bound, but the sampling term remains when $\zeta_\infty = O(1/n)$. We state the resulting corollary below. Following Bousquet et al. (2020), the factors $\log n$ and $\log(1/\delta)$ in the concentration bounds below mean $\max\{1, \ln n\}$ and $\max\{1, \ln(1/\delta)\}$, respectively.

Corollary 17 (Sharp high-probability generalization bound)**.** Under the assumptions of Theorem 15, there exists a universal constant $c > 0$ such that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of $S \sim \mathcal{D}^n$, the estimate holds:

$$\left|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S) - \widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)\right| \;\le\; c\left(\beta_n \log n \log\frac{1}{\delta} \;+\; B\sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{182}$$

where $\beta_n = L_{\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}}\sqrt{|\mathcal{Y}|}\,\zeta_\infty/(\sqrt{2}\,\mu^{\min}_{\mathcal{R}})$ is the uniform stability parameter from Theorem 15 and B is the uniform bound on $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}$ from Assumption 5(iii).

Proof. Define, for each $i \in \{1, \dots, n\}$, the expected loss gap

$$g_i(z_1, \dots, z_n) \;:=\; \mathbb{E}_{\widetilde{z}_i}\left[\,\mathbb{E}_z\big[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z)\big] \;-\; \mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z_i)\,\right],$$

where $S^{(i)} = (z_1, \dots, z_{i-1}, \widetilde{z}_i, z_{i+1}, \dots, z_n)$ is S with its i-th coordinate replaced by an independent copy $\widetilde{z}_i$, and the inner expectation is over a fresh $z \sim \mathcal{D}$. We verify that the g_i satisfy the three conditions of Theorem 4 of Bousquet et al. (2020):

i. ***Conditional bound*** $|\,\mathbb{E}[g_i \mid z_i]| \le M$**.** By Assumption 5(iii), $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}} \le B$ uniformly, so the inner difference is bounded in absolute value by B, and any expectation thereof is at most B in absolute value. Thus $|\,\mathbb{E}[g_i \mid z_i]| \le B$, so we set $M := B$.

ii. ***Centering*** $\mathbb{E}[g_i \mid z_{[n]\setminus\{i\}}] = 0$**.** Conditional on $z_{[n]\setminus\{i\}}$, the inner expectation $\mathbb{E}_{\widetilde{z}_i}\mathbb{E}_z[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z)]$ does not depend on z_i (since z_i has been replaced by $\widetilde{z}_i$ in $S^{(i)}$). The second term, when averaged over z_i, becomes $\mathbb{E}_{\widetilde{z}_i, z_i}[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z_i)]$. Since z_i and $\widetilde{z}_i$ are i.i.d. and the latter appears in $S^{(i)}$ in place of z_i, by exchangeability this equals $\mathbb{E}_{\widetilde{z}_i, z}[\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z)]$ where $z \sim \mathcal{D}$ independently. The two terms cancel, giving $\mathbb{E}[g_i \mid z_{[n]\setminus\{i\}}] = 0$.

iii. ***Bounded differences in coordinates*** $j \ne i$**.** Fix $j \ne i$ and consider modifying z_j to z'_j. This changes $S^{(i)}$ in coordinate j only. By the uniform stability bound (180), replacing one coordinate of the training set changes $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_{S^{(i)}}; z)$ by at most β_n for any z. Therefore both terms inside the brackets change by at most β_n, and g_i changes by at most $2\beta_n$. Thus g_i has bounded differences with parameter $\beta := 2\beta_n$ in every coordinate $j \ne i$.

Connecting $\sum_i g_i$ ***to the generalization gap.*** By Lemma 7 of Bousquet et al. (2020), under uniform stability (180) and uniform boundedness $\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}} \le B$,

$$\left|n\big(\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S) - \widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)\big)\right| \;\le\; \left|\sum_{i=1}^n g_i\right| \;+\; 2n\beta_n \quad \text{a.s.}$$

Applying the BKZ moment bound. By Theorem 4 of Bousquet et al. (2020), with its bounded-difference parameter set to $2\beta_n$, for any $p \ge 2$,

$$\Big\|\sum_{i=1}^n g_i\Big\|_p \;\le\; 24\sqrt{2}\,p\,n\,\beta_n\lceil\log_2 n\rceil \;+\; 4B\sqrt{pn}.$$

Lemma 1 of Bousquet et al. (2020) converts this moment bound into a tail bound: with probability at least $1-\delta$,

$$\Big|\sum_{i=1}^{n} g_i\Big| \;\le\; c_1\, n\, \beta_n \log n \log\frac{1}{\delta} \;+\; c_2\, B\sqrt{n\log(1/\delta)}$$

for absolute constants $c_1, c_2 > 0$. Combining with the inequality above and dividing by n,

$$\Big|\mathcal{L}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S) - \widehat{\mathcal{L}}^{\mathcal{R}}_{\mathrm{IPO}}(\pi^\star_S)\Big| \;\le\; c_1\, \beta_n \log n \log\frac{1}{\delta} \;+\; c_2\, B\sqrt{\frac{\log(1/\delta)}{n}} \;+\; 2\beta_n.$$

Under the stated logarithm convention, $2\beta_n \le 2\beta_n \log n \log(1/\delta)$, so absorbing all constants into a single $c > 0$ yields (182). □

K Offline sample complexity

In this section we establish a statistical guarantee for risk-adjusted preference learning in the offline setting: given n independent preference comparisons, how close can we get to the population risk-adjusted equilibrium $\pi^\star_{\mathcal{R}}$? We obtain a fast $\widetilde{\mathcal{O}}(1/n)$ rate in KL divergence for a finite response set, a fixed sampling distribution with full support, and a fixed positive population monotonicity modulus. In the risk-neutral case, the conditioning of the equilibrium is controlled by the KL regularization parameter β. Under risk, the corresponding bound is controlled by the population monotonicity modulus $\mu_{\mathcal{R}}$. We make the additional dependence on the offline sampling distribution and response-space size explicit.

The estimator first estimates the pairwise preference probabilities and then applies the risk functional. This order preserves the population target despite the nonlinearity of risk. Uniform estimation of the preference table yields uniform control of the risk-adjusted operator, and equilibrium stability converts that operator error into a KL guarantee. This separates the statistical error in offline preference data from the error incurred when numerically solving the empirical game.

K.1 Technical Novelties & Hurdles

The analysis connects the population geometry of the risk-adjusted game with estimation from Bernoulli preference observations. Three features of this connection require care.

Structural effects of risk. For risk-neutral preferences, $P + P^\top = \mathbf{1}\mathbf{1}^\top$ provides a constant-sum structure. Applying a nonlinear risk functional can make the payoff operator nonlinear in the opponent policy. The self-play reduction developed earlier still supplies the equilibrium fixed-point equation, while the risk-distortion bound controls its monotonicity. The stability argument below uses the pairing between operator differences and policy differences that appears in that analysis, together with symmetrized KL divergence and the explicit population monotonicity condition stated below.

Statistical bias and the observation model. The observations are Bernoulli comparison labels whose means are the entries of P. Applying a nonlinear risk transform directly to these labels generally changes the population target. For entropic risk, $\mathbb{E}\, e^{-\lambda z} = 1 - p + pe^{-\lambda}$ when $z \sim \mathrm{Bernoulli}(p)$, whereas the desired inner statistic is $e^{-\lambda p}$. We therefore estimate p before applying the exponential. The resulting plug-in risk estimate can remain biased, but its full estimation error is controlled uniformly. For entropic risk we also give a separate conditional bias bound. This offline construction differs from the fixed-batch opponent-sampling estimator used in the online algorithm.

Uniform concentration and sampling coverage. The empirical equilibrium depends on the same data used to estimate its payoff operator. Consequently, concentration at a single fixed opponent policy is insufficient for a direct substitution at that equilibrium. Monotonicity and translation equivariance of the risk functional yield a deterministic perturbation bound uniform over all opponent policies. Combined with concentration of the preference table, this gives an operator error of order $n^{-1/2}$, with the pair-coverage factor explicit. The equilibrium perturbation bound is quadratic in this error, giving the n^{-1} KL rate. Hoeffding concentration suffices for this exponent; Bernstein concentration is not required.

Connection to the rest of the paper. Theorem 16 complements the online analysis by controlling the distance between the empirical and population equilibria. Its regularity condition is the policy-space inequality (187). When $0 \le \bar{\lambda}_{\mathcal{R}} < \beta$, the policy-space risk-distortion bound certifies this condition with $\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}}$, as shown below. This connects the structural analysis to statistical sensitivity. The online convergence and bias-tracking guarantees retain their own algorithmic and oracle assumptions. The sufficient monotonicity condition delimits the regime covered by the offline argument; its failure does not establish a statistical impossibility result.

K.2 Preliminaries and main result

We first specify the observation model and the empirical operator, then state the population regularity condition that controls equilibrium sensitivity. Keeping these ingredients separate makes explicit which part of the guarantee comes from sampling coverage and which part comes from the geometry of the risk-adjusted game.

Notation and sampling model. Let $d = |\mathcal{Y}| \ge 2$ and $P \in [0,1]^{d\times d}$ satisfy $P_{y,y'} + P_{y',y} = 1$. We observe n independent triples

$$(y_i, y_i') \sim \bar{\pi} \otimes \bar{\pi}, \qquad z_i \mid (y_i, y_i') \sim \mathrm{Bernoulli}(P_{y_i,y_i'}),$$

where $\bar{\pi}(y) > 0$ for all y. Write $\mathcal{D} = ((y_i, y_i'))_{i=1}^n$ for the sampled pair design, and define

$$N_{y,y'} := \#\{i : (y_i, y_i') = (y, y')\}, \qquad N_{\min} := \min_{y,y'} N_{y,y'}, \qquad \rho_{\min} := \min_{y,y'} \bar{\pi}(y)\bar{\pi}(y').$$

Conditional on $\mathcal{D}$, the observed labels remain independent. All conditional bias statements below are restricted to designs with $N_{\min} \ge 1$.

For each ordered pair, form the empirical preference mean

$$\widetilde{P}_{y,y'} := \begin{cases} N_{y,y'}^{-1} \displaystyle\sum_{i:(y_i,y_i')=(y,y')} z_i, & N_{y,y'} > 0, \\ 1/2, & N_{y,y'} = 0. \end{cases}$$

We impose the known preference symmetry by setting

$$\widehat{P}_{y,y'} := \begin{cases} \frac{1}{2}(\widetilde{P}_{y,y'} + 1 - \widetilde{P}_{y',y}), & y \ne y', \\ 1/2, & y = y'. \end{cases} \tag{183}$$

Thus $\widehat{P} \in [0,1]^{d\times d}$ and $\widehat{P} + \widehat{P}^\top = \mathbf{1}\mathbf{1}^\top$ for every dataset. Writing $\|Q\|_{\max} := \max_{y,y'} |Q_{y,y'}|$, symmetry of P gives

$$\|\widehat{P} - P\|_{\max} \le \|\widetilde{P} - P\|_{\max}. \tag{184}$$

On a design with positive counts, $\widehat{P}$ is conditionally unbiased for P. The value assigned to an unobserved pair only makes the estimator well defined; the high-probability theorem controls the event that every pair has sufficient observations.

Risk applied to estimated payoffs. For an opponent policy μ and a vector $v \in \mathbb{R}^d$, write $\mathcal{R}_\mu(v)$ for the payoff-side risk functional. We assume

$$\mathcal{R}_\mu(0) = 0, \qquad v \le w \Longrightarrow \mathcal{R}_\mu(v) \le \mathcal{R}_\mu(w), \qquad \mathcal{R}_\mu(v + c\mathbf{1}) = \mathcal{R}_\mu(v) + c.$$

For a payoff matrix Q, define

$$(\mathcal{P}_{\mathcal{R}}[Q](\mu))_y := \mathcal{R}_\mu(Q_{y,\cdot}), \qquad \mathcal{P}_{\mathcal{R}} := \mathcal{P}_{\mathcal{R}}[P], \qquad \widehat{\mathcal{P}}_{\mathcal{R}} := \mathcal{P}_{\mathcal{R}}[\widehat{P}]. \tag{185}$$

We require $\mu \mapsto \mathcal{P}_{\mathcal{R}}[Q](\mu)$ to be continuous on $\mathrm{ri}\,\Delta(\mathcal{Y})$ for each $Q \in [0,1]^{d\times d}$. Here $\mathrm{ri}\,\Delta$ denotes the policies with all coordinates positive. This condition holds for entropic risk and for payoff CVaR with any fixed positive tail mass. For a general family of coherent risks, it is an explicit regularity assumption on its dependence on the opponent policy.

The opponent distribution μ is an argument of the operator and is evaluated exactly. In particular, it may be the current candidate policy during equilibrium computation; it is not estimated from an additional batch of opponent samples.

For entropic payoff risk with parameter $\lambda > 0$, the corrected operator is

$$(\widehat{\mathcal{P}}_{\mathcal{R}}(\mu))_y = -\frac{1}{\lambda} \log \left(\sum_{y'} \mu(y') e^{-\lambda \widehat{P}_{y,y'}} \right). \tag{186}$$

The order of these operations matters. For $z \sim \text{Bernoulli}(p)$,

$$\mathbb{E}\, e^{-\lambda z} = 1 - p + p e^{-\lambda} \neq e^{-\lambda p}$$

in general. Averaging $e^{-\lambda z_i}$ would therefore estimate a different population operator. Equation (186) uses the estimated probability inside the exponential.

Equilibrium and population monotonicity. Let π_{ref} have full support and let $\beta > 0$. For any Q, define

$$T_Q(\pi) := \text{softmax}\left(\log \pi_{\text{ref}} + \beta^{-1} \mathcal{P}_{\mathcal{R}}[Q](\pi)\right).$$

Logarithms are componentwise. Let $\pi^\star_{\mathcal{R}}$ be a fixed point of T_P, and let $\widehat{\pi}_n$ be any fixed point of $T_{\widehat{P}}$. Existence and population uniqueness are proved below; empirical uniqueness is not assumed.

For interior policies, define

$$J(\pi, \pi') := \text{KL}(\pi \| \pi') + \text{KL}(\pi' \| \pi) = \langle \log \pi - \log \pi', \pi - \pi' \rangle, \qquad H_{\mathcal{R}}(\pi) := \beta \log \frac{\pi}{\pi_{\text{ref}}} - \mathcal{P}_{\mathcal{R}}(\pi).$$

The fixed-point equation is equivalent to $H_{\mathcal{R}}(\pi)$ being a constant vector. We use $\mu_{\mathcal{R}} > 0$ for a population monotonicity modulus satisfying

$$\left\langle H_{\mathcal{R}}(\pi) - H_{\mathcal{R}}(\pi'), \pi - \pi' \right\rangle \geq \mu_{\mathcal{R}} J(\pi, \pi') \quad \text{for all interior } \pi, \pi'. \tag{187}$$

This expresses the regularity needed for the offline result directly in policy space.

Relation to the risk-distortion bound. Here $\bar{\lambda}_{\mathcal{R}}$ denotes the policy-space curvature bound. Suppose the population payoff operator obeys

$$\left\langle \mathcal{P}_{\mathcal{R}}(\pi) - \mathcal{P}_{\mathcal{R}}(\pi'), \pi - \pi' \right\rangle \leq \bar{\lambda}_{\mathcal{R}} \|\pi - \pi'\|_2^2. \tag{188}$$

For a continuously differentiable operator, the tangent-space symmetric-Jacobian bound defining $\bar{\lambda}_{\mathcal{R}}$ implies (188) by integration along the segment from π' to π. Pinsker's inequality gives

$$\|\pi - \pi'\|_2^2 \leq \|\pi - \pi'\|_1^2 \leq J(\pi, \pi'). \tag{189}$$

If $0 \leq \bar{\lambda}_{\mathcal{R}} < \beta$, then

$$\langle H_{\mathcal{R}}(\pi) - H_{\mathcal{R}}(\pi'), \pi - \pi' \rangle \geq \beta J(\pi, \pi') - \bar{\lambda}_{\mathcal{R}} \|\pi - \pi'\|_2^2 \geq (\beta - \bar{\lambda}_{\mathcal{R}}) J(\pi, \pi').$$

Thus the choice $\mu_{\mathcal{R}} = \beta - \bar{\lambda}_{\mathcal{R}} > 0$ is admissible. If the signed curvature bound is nonpositive, the same identity instead gives the admissible choice $\mu_{\mathcal{R}} = \beta$. A negative Euclidean curvature bound by itself does not justify a uniform improvement beyond β in front of the KL divergence. For an operator without an ordinary Jacobian everywhere, (187) must be justified directly; the CVaR subsection below gives such a finite-difference argument.

The following theorem converts estimation of the preference table into a statistical guarantee for the risk-adjusted equilibrium. For a fixed response set, sampling distribution, and positive monotonicity modulus, it establishes a high-probability $\widetilde{\mathcal{O}}(1/n)$ rate in KL divergence. The bound displays the effects of pairwise sampling coverage through $\rho_{\min}^{-1}$ and equilibrium sensitivity through $\mu_{\mathcal{R}}^{-2}$, quantifying how these two factors determine the number of preference comparisons needed for accurate equilibrium estimation.

Theorem 16 (Offline sample complexity)**.** Under the sampling model above, let the payoff-risk functionals be normalized, monotone, and translation equivariant. Suppose $\mu \mapsto \mathcal{P}_{\mathcal{R}}[Q](\mu)$ is continuous on $\text{ri}\,\Delta(\mathcal{Y})$ for every $Q \in [0,1]^{d \times d}$, and suppose (187) holds with $\mu_{\mathcal{R}} > 0$. Construct $\widehat{P}$ and $\widehat{\mathcal{P}}_{\mathcal{R}}$ by (183) and (185). The population fixed point $\pi^\star_{\mathcal{R}}$ is unique, and empirical fixed points exist. For $\delta \in (0,1)$ and $n\rho_{\min} \geq 8 \log \frac{2d^2}{\delta}$, with probability at least $1 - \delta$, every empirical fixed point $\widehat{\pi}_n$ satisfies

$$\text{KL}(\pi^\star_{\mathcal{R}} \| \widehat{\pi}_n) + \text{KL}(\widehat{\pi}_n \| \pi^\star_{\mathcal{R}}) \leq \frac{\log(4d^2/\delta)}{\mu_{\mathcal{R}}^2 n \rho_{\min}}. \tag{190}$$

The statistical bound uses population monotonicity. The empirical operator need not share this property: the bound holds for each of its fixed points. An additional empirical-curvature certificate, useful for ensuring empirical uniqueness, is given in Section **??**.

The proof proceeds in two steps. We first control the empirical payoff operator uniformly over opponent policies, then show how population monotonicity converts this control into a bound on equilibrium displacement. After combining these steps, we examine entropic risk and CVaR to clarify the distinct roles of estimation bias and equilibrium regularity.

K.3 Uniform estimation of the risk-adjusted operator

The empirical equilibrium is chosen using the observed comparisons, so an error bound at a fixed opponent policy would not suffice. We instead begin with a deterministic bound that transfers uniform error in the preference table to the risk-adjusted operator at every opponent policy. Monotonicity and translation equivariance provide this transfer without a smoothness assumption on the risk functional.

Lemma 21 (Risk is nonexpansive in payoff values)**.** For any $Q, Q' \in [0,1]^{d\times d}$,

$$\sup_{\mu\in \mathrm{ri}\,\Delta(\mathcal{Y})} \|\mathcal{P}_{\mathcal{R}}[Q](\mu) - \mathcal{P}_{\mathcal{R}}[Q'](\mu)\|_\infty \le \|Q - Q'\|_{\max}.$$

Proof. Set $a = \|Q - Q'\|_{\max}$. Each row satisfies $Q'_{y,\cdot} - a\mathbf{1} \le Q_{y,\cdot} \le Q'_{y,\cdot} + a\mathbf{1}$. Monotonicity and translation equivariance imply

$$\mathcal{R}_\mu(Q'_{y,\cdot}) - a \le \mathcal{R}_\mu(Q_{y,\cdot}) \le \mathcal{R}_\mu(Q'_{y,\cdot}) + a.$$

Taking the maximum over y and the supremum over μ proves the claim. □

It therefore remains to estimate the finitely many entries of P; the supremum over opponent policies introduces no additional concentration step. Conditioning on the sampled pair design fixes the cell counts while leaving the comparison labels independent. The following bound isolates the label noise for a given allocation of observations across pairs.

Lemma 22 (Conditional concentration)**.** Fix a pair design $\mathcal{D}$ with $N_{\min} \ge 1$. For every $\delta \in (0,1)$, with conditional probability at least $1-\delta$ over the labels,

$$\sup_{\mu\in \mathrm{ri}\,\Delta(\mathcal{Y})} \|\widehat{\mathcal{P}}_{\mathcal{R}}(\mu) - \mathcal{P}_{\mathcal{R}}(\mu)\|_\infty^2 \le \|\widehat{P} - P\|_{\max}^2 \le \frac{\log(2d^2/\delta)}{2N_{\min}}.$$

Proof. For each ordered pair, conditional Hoeffding concentration gives

$$\Pr\left(|\widetilde{P}_{y,y'} - P_{y,y'}| > t \mid \mathcal{D}\right) \le 2e^{-2N_{y,y'}t^2}.$$

A union bound over at most d^2 pairs gives

$$\Pr\left(\|\widetilde{P} - P\|_{\max} > t \mid \mathcal{D}\right) \le 2d^2 e^{-2N_{\min}t^2}.$$

Set $t^2 = \log(2d^2/\delta)/(2N_{\min})$, use (184), and apply Lemma 21. □

The conditional guarantee depends on the least-sampled pair through $N_{\min}$. To express it in terms of the total sample size, we must also control the random allocation of comparisons. The next lemma shows that, once $n\rho_{\min}$ is sufficiently large, every pair receives enough observations for the conditional bound to apply.

Lemma 23 (Coverage and unconditional concentration)**.** If $n\rho_{\min} \ge 8\log(2d^2/\delta)$, then with probability at least $1-\delta$,

$$N_{\min} \ge \frac{n\rho_{\min}}{2}, \qquad \|\widehat{P} - P\|_{\max}^2 \le \frac{\log(4d^2/\delta)}{n\rho_{\min}}.$$

Proof. The count $N_{y,y'}$ is binomial with mean $n\bar{\pi}(y)\bar{\pi}(y') \ge n\rho_{\min}$. The multiplicative Chernoff bound and a union bound imply

$$\Pr\left(N_{\min} < \frac{n\rho_{\min}}{2}\right) \le d^2 e^{-n\rho_{\min}/8} \le \frac{\delta}{2}.$$

For every design on the complementary event, apply Lemma 22 with failure probability $\delta/2$. Its squared-error bound is at most $\log(4d^2/\delta)/(n\rho_{\min})$. Integrating the conditional failure bound over the design and adding the coverage failure probability proves the result. □

Remark 5 (Why the KL rate is of order $1/n$)**.** The operator error is of order $n^{-1/2}$, with coverage held fixed. The stability bound below is quadratic in that error, which gives the n^{-1} KL rate. Hoeffding concentration suffices for this rate. Variance-sensitive concentration can refine the estimation bound but is not required to obtain its exponent. The operator bound is uniform over the opponent policy, so it also holds at every data-dependent empirical equilibrium.

K.4 Existence and stability of the equilibrium

We now connect operator estimation to the equilibrium itself. First, bounded payoffs, KL regularization, and continuity ensure that the population and empirical response maps have fixed points with full support. We then use population monotonicity to control how far an empirical fixed point can move when its payoff operator is perturbed.

Lemma 24 (Existence of population and empirical fixed points)**.** Under the stated risk assumptions, T_Q has an interior fixed point for every $Q \in [0,1]^{d\times d}$.

Proof. Normalization, monotonicity, and translation equivariance imply $0 \le (\mathcal{P}_{\mathcal{R}}[Q](\pi))_y \le 1$. Thus

$$(T_Q(\pi))_y = \frac{\pi_{\mathrm{ref}}(y)\exp((\mathcal{P}_{\mathcal{R}}[Q](\pi))_y/\beta)}{\sum_j \pi_{\mathrm{ref}}(j)\exp((\mathcal{P}_{\mathcal{R}}[Q](\pi))_j/\beta)} \ge \pi_{\mathrm{ref}}(y)e^{-1/\beta}.$$

The set

$$K := \{\pi \in \Delta(\mathcal{Y}) : \pi(y) \ge \pi_{\mathrm{ref}}(y)e^{-1/\beta} \text{ for all } y\}$$

is nonempty, compact, and convex, and lies in the interior of the simplex. The continuity assumption makes T_Q a continuous map from K into K. Brouwer's fixed-point theorem gives the claim. □

These fixed points lie in the domain where the logarithmic optimality equations and KL divergence are well defined. To compare them, we pair the difference of their optimality equations with the policy difference. The normalization constants then cancel, leaving an operator perturbation that can be bounded using the uniform error from the preceding subsection.

Lemma 25 (Equilibrium stability under operator perturbation)**.** Suppose (187) holds with $\mu_{\mathcal{R}} > 0$. Let $\pi^\star$ be a population fixed point, and let π' be any fixed point for another payoff operator $\mathcal{P}'_{\mathcal{R}}$ using the same reference policy and β. If

$$a := \sup_{\mu \in \mathrm{ri}\,\Delta(\mathcal{Y})} \|\mathcal{P}'_{\mathcal{R}}(\mu) - \mathcal{P}_{\mathcal{R}}(\mu)\|_\infty < \infty,$$

then $J(\pi^\star, \pi') \le a^2/\mu_{\mathcal{R}}^2$. In particular, the population fixed point is unique.

Proof. Let $v := \pi' - \pi^\star$, and define $H'_{\mathcal{R}}(\pi) := \beta\log(\pi/\pi_{\mathrm{ref}}) - \mathcal{P}'_{\mathcal{R}}(\pi)$. The fixed-point equations imply that $H_{\mathcal{R}}(\pi^\star)$ and $H'_{\mathcal{R}}(\pi')$ are constant vectors. Since $\mathbf{1}^\top v = 0$,

$$\begin{aligned}\mu_{\mathcal{R}} J(\pi^\star, \pi') &\le \langle H_{\mathcal{R}}(\pi') - H_{\mathcal{R}}(\pi^\star), v\rangle = \langle H_{\mathcal{R}}(\pi') - H'_{\mathcal{R}}(\pi'), v\rangle = \langle \mathcal{P}'_{\mathcal{R}}(\pi') - \mathcal{P}_{\mathcal{R}}(\pi'), v\rangle \\ &\le a\|v\|_1 \le a\sqrt{J(\pi^\star, \pi')}.\end{aligned}$$

If $J(\pi^\star, \pi') > 0$, divide by its square root and by $\mu_{\mathcal{R}}$, then square. If $J = 0$, the conclusion is immediate. Taking $\mathcal{P}'_{\mathcal{R}} = \mathcal{P}_{\mathcal{R}}$ gives $a = 0$ and proves population uniqueness. □

The quadratic dependence on a is the step that turns the $n^{-1/2}$ estimation scale into an n^{-1} KL guarantee. The modulus $\mu_{\mathcal{R}}$ measures the sensitivity of this conversion: a smaller population modulus allows a larger equilibrium displacement for the same operator error.

K.5 Assembling the offline guarantee

The statistical and deterministic parts of the argument can now be combined. The same preference-table concentration event controls the operator at every opponent policy, so the stability bound applies simultaneously to all empirical fixed points selected from the data.

Proof of Theorem 16. Lemma 24 gives population and empirical fixed points. Lemma 25 gives uniqueness of the population fixed point. By Lemma 21, the operator perturbation in Lemma 25 satisfies $a \le \|\widehat{P} - P\|_{\max}$. Thus, deterministically, every empirical fixed point satisfies

$$J(\pi^\star_{\mathcal{R}}, \widehat{\pi}_n) \le \|\widehat{P} - P\|^2_{\max}/\mu^2_{\mathcal{R}}.$$

Apply Lemma 23 to obtain (190). Its concentration event is independent of which empirical fixed point is selected, so the conclusion holds simultaneously for all of them. □

The argument also gives a conditional guarantee in terms of the realized comparison counts. This form displays the effect of the observed sampling allocation directly, before replacing it with a bound involving n and $\rho_{\min}$. For a fixed design with $N_{\min} \ge 1$, the same proof and Lemma 22 give, with conditional probability at least $1 - \delta$,

$$J(\pi^\star_{\mathcal{R}}, \widehat{\pi}_n) \le \frac{\log(2d^2/\delta)}{2\mu^2_{\mathcal{R}} N_{\min}}.$$

The unconditional guarantee makes the response-set dependence particularly explicit when all ordered pairs are sampled equally. For uniform sampling, $\rho_{\min} = d^{-2}$, and the unconditional bound becomes

$$J(\pi^\star_{\mathcal{R}}, \widehat{\pi}_n) \le \frac{d^2 \log(4d^2/\delta)}{\mu^2_{\mathcal{R}} n}, \qquad n \ge 8d^2 \log(2d^2/\delta).$$

The factor d^2 reflects coverage of the pairwise preference table. No additional factor d is incurred by converting between Euclidean logit norms and coordinatewise operator error.

K.6 Entropic risk and plug-in bias

Entropic risk illustrates both the scope of the main theorem and the bias introduced by applying a nonlinear risk functional to estimated payoffs. We first specialize the equilibrium guarantee to the corrected estimator, then quantify its conditional bias using the smoothness of entropic risk in the payoff values.

Corollary 18 (Offline sample complexity for entropic payoff risk)**.** Use the estimator (186) with fixed $\lambda > 0$. Suppose all assumptions of Theorem 16 hold. For $\delta \in (0, 1)$, if $n\rho_{\min} \ge 8 \log(2d^2/\delta)$, then with probability at least $1 - \delta$,

$$\mathrm{KL}(\pi^\star_{\mathcal{R}} \| \widehat{\pi}_n) \le \frac{\log(4d^2/\delta)}{\mu^2_{\mathcal{R}} n \rho_{\min}}.$$

Proof. Entropic payoff risk is normalized, monotone, and translation equivariant. Its operator is continuous in the opponent policy: the argument of the logarithm lies in $[e^{-\lambda}, 1]$. Apply Theorem 16 and retain the first KL term. □

The dependence on λ in this bound enters through $\mu_{\mathcal{R}}$; Section K.7 gives an

explicit bound in terms of λ. This is a specialization of the general monotone, translation-invariant result; entropic risk need not be a coherent risk functional.

Even though $\widehat{P}$ is conditionally unbiased on a design with positive counts, applying entropic risk need not preserve unbiasedness. Concavity determines the direction of this bias, while a second-order expansion bounds its magnitude by the variances of the estimated payoff entries. This calculation describes the estimator more precisely; the equilibrium rate above already follows from control of the full estimation error.

Lemma 26 (Conditional bias of the corrected entropic estimator)**.** Fix μ and a pair design $\mathcal{D}$ with $N_{\min} \ge 1$. For every row y,

$$0 \le (\mathcal{P}_{\mathcal{R}}(\mu))_y - \mathbb{E}[(\widehat{\mathcal{P}}_{\mathcal{R}}(\mu))_y \mid \mathcal{D}] \le \frac{\lambda}{32} \sum_{j \ne y} \left(\frac{1}{N_{y,j}} + \frac{1}{N_{j,y}} \right) \le \frac{\lambda(d-1)}{16 N_{\min}}. \tag{191}$$

Proof. Define

$$h_\mu(p) := -\lambda^{-1} \log \sum_j \mu(j) e^{-\lambda p_j} \quad \text{and} \quad w_j(p) := \frac{\mu(j) e^{-\lambda p_j}}{\sum_k \mu(k) e^{-\lambda p_k}}.$$

Then

$$\nabla h_\mu(p) = w(p) \quad \text{and} \quad \nabla^2 h_\mu(p) = -\lambda\big(\operatorname{diag} w(p) - w(p)w(p)^\top\big).$$

The matrix in parentheses is positive semidefinite and bounded above by the identity. Hence h_μ is concave and its Hessian has operator norm at most λ. Conditional on $\mathcal{D}$, the row error $e := \widehat{P}_{y,\cdot} - P_{y,\cdot}$ has mean zero. Jensen's inequality and Taylor's formula with integral remainder give

$$0 \le h_\mu(P_{y,\cdot}) - \mathbb{E}[h_\mu(P_{y,\cdot} + e) \mid \mathcal{D}] \le \frac{\lambda}{2}\,\mathbb{E}[\|e\|_2^2 \mid \mathcal{D}].$$

For $j \neq y$, the two ordered-cell empirical means are conditionally independent, so

$$\operatorname{Var}(\widehat{P}_{y,j} \mid \mathcal{D}) \le \frac{1}{16}\left(\frac{1}{N_{y,j}} + \frac{1}{N_{j,y}}\right).$$

The diagonal entry is exactly $1/2$. Sum these variance bounds and substitute into the previous display. □

The sign in (191) comes from concavity in the estimated payoff probabilities: at a fixed opponent policy and conditional on a design with positive counts, the estimated risk payoff is downward biased. For fixed d and λ, the bound is of order $N_{\min}^{-1}$, whereas the full estimation error is controlled at the $N_{\min}^{-1/2}$ scale. This distinction will matter for nonsmooth risks, for which the same bias expansion may be unavailable. This estimator differs from an estimator that averages exponentiated, directly observed payoff values and then takes a logarithm.

K.7 Explicit risk-parameter dependence for entropic risk

We now make the dependence on λ in Corollary 18 explicit. The preference-table concentration bound is independent of λ, so it remains to bound the population monotonicity modulus. The following calculation controls the curvature of the entropic operator in its opponent-policy argument, as required by (188).

Proposition 9 (Entropic risk: uniform distortion bound)**.** Let $P \in [0,1]^{d\times d}$ satisfy $P + P^\top = \mathbf{1}\mathbf{1}^\top$, and let $\mathcal{R}$ be entropic payoff risk with parameter $\lambda > 0$. Define

$$A := \tfrac{1}{2}(P - P^\top), \qquad r_A := \max_y \left(\max_j A_{y,j} - \min_j A_{y,j}\right) \le 1.$$

Write $J_\lambda(P,\mu) := \nabla_\mu \mathcal{P}_\mathcal{R}[P](\mu)$ for the opponent-policy Jacobian. For every $\mu \in \Delta(\mathcal{Y})$ and $\xi \in \mathbf{1}^\perp$,

$$\begin{aligned} \left|\xi^\top J_\lambda(P,\mu)\xi\right| &\le B_A(\lambda)\|\xi\|_2^2, \\ B_A(\lambda) := \frac{d}{2\lambda}\left(e^{\lambda r_A} - 1 - \lambda r_A\right) &\le \frac{d r_A}{4}\left(e^{\lambda r_A} - 1\right). \end{aligned} \tag{192}$$

In particular, $B_A(\lambda) = d r_A^2 \lambda/4 + \mathcal{O}(\lambda^2)$ as $\lambda \downarrow 0$.

Proof. Since $P = A + \frac{1}{2}\mathbf{1}\mathbf{1}^\top$, translation equivariance shifts every risk-adjusted payoff by $1/2$ and leaves the policy derivatives unchanged. Write

$$v_y(\mu) := -\frac{1}{\lambda}\log\sum_j \mu(j) e^{-\lambda A_{y,j}}, \qquad t_{y,j} := A_{y,j} - v_y(\mu).$$

The entropic certainty equivalent lies between the minimum and maximum of its payoff row, so $|t_{y,j}| \le r_A$. Direct differentiation gives $[J_\lambda(P,\mu)]_{y,j} = -e^{-\lambda t_{y,j}}/\lambda$. Set

$$E_{y,j} := \frac{e^{-\lambda t_{y,j}} - 1 + \lambda t_{y,j}}{\lambda}, \qquad b_\lambda := \frac{e^{\lambda r_A} - 1 - \lambda r_A}{\lambda}.$$

The exponential remainder satisfies $0 \le E_{y,j} \le b_\lambda$. Expanding the Jacobian entrywise therefore yields

$$J_\lambda(P,\mu) = A - \big(v(\mu) + \lambda^{-1}\mathbf{1}\big)\,\mathbf{1}^\top - E.$$

For $\xi \in \mathbf{1}^\perp$, antisymmetry of A and $\mathbf{1}^\top\xi = 0$ give

$$\xi^\top J_\lambda(P,\mu)\xi = -\xi^\top E\xi = -\xi^\top\left(E - \tfrac{b_\lambda}{2}\mathbf{1}\mathbf{1}^\top\right)\xi.$$

Every entry of the centered matrix has absolute value at most $b_\lambda/2$. Hence

$$\left|\xi^\top J_\lambda(P,\mu)\xi\right| \le \frac{b_\lambda}{2}\|\xi\|_1^2 \le \frac{db_\lambda}{2}\|\xi\|_2^2 = B_A(\lambda)\|\xi\|_2^2.$$

Finally, convexity gives $e^u - 1 - u = \int_0^u (e^s - 1)\,ds \le u(e^u - 1)/2$ for $u \ge 0$. Taking $u = \lambda r_A$ proves the second bound in (192); Taylor expansion gives the stated small-λ behavior. □

This curvature bound supplies the monotonicity condition of the offline theorem. Its explicit dependence on λ therefore translates directly into an equilibrium-estimation guarantee.

Corollary 19 (Entropic risk: explicit offline guarantee)**.** Under the offline sampling model, use the estimator (186) and suppose $B_A(\lambda) < \beta$. Then $\mu_{\mathcal{R}} := \beta - B_A(\lambda) > 0$ satisfies (187). For $\delta \in (0,1)$ and $n\rho_{\min} \ge 8\log(2d^2/\delta)$, with probability at least $1-\delta$, every empirical fixed point satisfies

$$\mathrm{KL}(\pi^\star_{\mathcal{R}} \| \widehat{\pi}_n) \le \frac{\log(4d^2/\delta)}{n\rho_{\min}\left[\beta - B_A(\lambda)\right]^2}.$$

Proof. For interior policies π, π', put $\xi = \pi - \pi'$ and integrate Proposition 9 along the segment joining them:

$$\langle \mathcal{P}_{\mathcal{R}}(\pi) - \mathcal{P}_{\mathcal{R}}(\pi'), \xi\rangle = \int_0^1 \xi^\top J_\lambda(P, \pi' + s\xi)\xi\, ds \le B_A(\lambda)\|\xi\|_2^2.$$

Since $B_A(\lambda) \ge 0$, inequality (189) gives

$$\begin{aligned}\langle H_{\mathcal{R}}(\pi) - H_{\mathcal{R}}(\pi'), \xi\rangle &\ge \beta J(\pi,\pi') - B_A(\lambda)\|\xi\|_2^2 \\ &\ge [\beta - B_A(\lambda)]J(\pi,\pi').\end{aligned}$$

This verifies (187). Entropic risk satisfies the remaining risk assumptions of Theorem 16; applying that theorem and retaining one KL term proves the claim. □

Choosing the risk parameter. For $r_A > 0$, a sufficient choice ensuring $\beta - B_A(\lambda) \ge \beta/2$ is

$$0 < \lambda \le \frac{1}{r_A}\log\left(1 + \frac{2\beta}{dr_A}\right).$$

Indeed, the second bound in (192) is then at most $\beta/2$. The resulting offline guarantee is at most four times the risk-neutral bound $\log(4d^2/\delta)/(\beta^2 n\rho_{\min})$. If $r_A = 0$, then $B_A(\lambda) = 0$ for every $\lambda > 0$.

Using only $P_{y,j} \in [0,1]$, we may replace the row-dependent bound by

$$B_A(\lambda) \le B_d(\lambda) := \frac{d}{2\lambda}(e^\lambda - 1 - \lambda).$$

Whenever $B_d(\lambda) < \beta$, this gives the simpler explicit form

$$\mathrm{KL}(\pi^\star_{\mathcal{R}} \| \widehat{\pi}_n) \le \frac{\log(4d^2/\delta)}{n\rho_{\min}\left[\beta - \frac{d}{2\lambda}(e^\lambda - 1 - \lambda)\right]^2}.$$

As $\lambda \downarrow 0$, $B_d(\lambda) = d\lambda/4 + \mathcal{O}(\lambda^2)$, so this bound approaches the risk-neutral one. The $1/n$ exponent is unchanged: the explicit risk dependence comes from equilibrium sensitivity, through the denominator, while the payoff-estimation bound remains independent of λ.

K.8 Coherent payoff risks and CVaR

The payoff perturbation argument extends beyond smooth risk functionals. For coherent payoff risks, the dual representation makes this control explicit. Equilibrium regularity remains a separate requirement; for CVaR, we verify it through finite differences in the opponent policy, allowing for changes in the active tail.

Suppose a coherent payoff-risk functional has the lower-envelope form

$$\mathcal{R}_\mu(v) = \inf_{q \in \mathcal{Q}(\mu)} \sum_j \mu(j) q_j v_j, \qquad q_j \ge 0, \quad \sum_j \mu(j) q_j = 1,$$

where $\mathcal{Q}(\mu)$ is nonempty. The corresponding estimator is

$$(\widehat{\mathcal{P}}_{\mathcal{R}}(\mu))_y = \inf_{q \in \mathcal{Q}(\mu)} \sum_j \mu(j) q_j \widehat{P}_{y,j}.$$

For every feasible q, we have that

$$\Big|\sum_j \mu(j) q_j (\widehat{P}_{y,j} - P_{y,j})\Big| \le \|\widehat{P} - P\|_{\max} \sum_j \mu(j) q_j = \|\widehat{P} - P\|_{\max}.$$

Taking infima preserves this perturbation bound. Thus no additional union bound over the dual class is needed. In particular, a density ratio bound $M_{\mathcal{R}}$ is not needed for this payoff-estimation step. The equilibrium conclusion also uses continuity in the opponent policy and the population monotonicity condition (187).

Bias for nonsmooth risks. For a fixed policy and a design with positive counts, concavity and conditional unbiasedness of $\widehat{P}$ imply

$$0 \le (\mathcal{P}_{\mathcal{R}}(\mu))_y - \mathbb{E}[(\widehat{\mathcal{P}}_{\mathcal{R}}(\mu))_y \mid \mathcal{D}] \le \mathbb{E}[\|\widehat{P} - P\|_{\max} \mid \mathcal{D}].$$

An $\mathcal{O}(1/N_{\min})$ bias bound is not asserted for general nonsmooth coherent risks. The full estimation-error bound in Lemma 22 suffices for the $\mathcal{O}(1/N_{\min})$ equilibrium KL bound, even if the operator bias is of order $N_{\min}^{-1/2}$.

Payoff CVaR. CVaR provides a concrete example in which the risk value is Lipschitz even though derivatives can fail to exist. To apply the offline theorem, we need to control both perturbations of the payoff values and changes in the opponent policy. The former follows from nonexpansiveness; the latter will yield an explicit sufficient condition on the regularization strength. For a lower-tail mass $q \in (0, 1]$, define

$$\mathrm{CVaR}_{q,\mu}(v) := \sup_{\nu \in \mathbb{R}} \Big\{ \nu - \frac{1}{q} \sum_j \mu(j)(\nu - v_j)_+ \Big\}, \qquad v \in \mathbb{R}^d. \tag{193}$$

For $v \in [0, 1]^d$, the interval $[0, 1]$ contains an optimal threshold, so the supremum can be restricted to that interval. Define the empirical operator by replacing v with the row $\widehat{P}_{y,\cdot}$ and optimizing the threshold in (193). This uses no population threshold oracle.

The issue caused by nonsmoothness is the absence of an ordinary Jacobian at policies where the active tail changes. The risk value itself remains Lipschitz continuous. For example, for payoffs $(0, 1)$ and opponent policy $(p, 1 - p)$,

$$\mathrm{CVaR}_{q,(p,1-p)}(0, 1) = \max\{0, 1 - p/q\}.$$

This function is $1/q$-Lipschitz in p, but is not differentiable at $p = q$ when $0 < q < 1$. The differentiable-Jacobian argument used for smooth risks therefore does not apply at that point. The following finite-difference bound supplies a separate regularity certificate for the offline result.

For $v \in [0, 1]^d$ and $\delta = \mu - \mu'$, the vector $a_j = (\nu - v_j)_+$ lies in $[0, 1]^d$ for each $\nu \in [0, 1]$. Since $\sum_j \delta_j = 0$,

$$\Big|\sum_j \delta_j a_j\Big| = \Big|\sum_j \delta_j (a_j - 1/2)\Big| \le \frac{1}{2}\|\delta\|_1.$$

Taking the difference of the suprema in (193) gives

$$|\,\mathrm{CVaR}_{q,\mu}(v) - \mathrm{CVaR}_{q,\mu'}(v)| \le \frac{\|\mu - \mu'\|_1}{2q}. \tag{194}$$

In particular the CVaR payoff operator is continuous. Additionally, we have that

$$\langle \mathcal{P}_{\mathrm{CVaR}_q}(\pi) - \mathcal{P}_{\mathrm{CVaR}_q}(\pi'), \pi - \pi' \rangle \le \frac{\|\pi - \pi'\|_1^2}{2q} \le \frac{J(\pi, \pi')}{2q},$$

so that (187) holds with $\mu_{\mathrm{CVaR}_q} = \beta - 1/(2q) > 0$ whenever $\beta > 1/(2q)$. This is a sufficient condition and can be conservative; it does not assert that an arbitrary choice of regularization and tail mass lies in the required regime.

Together with payoff nonexpansiveness, this direct regularity bound verifies the hypotheses of the main theorem and gives the following CVaR specialization.

Theorem 17 (Offline sample complexity for payoff CVaR)**.** Under the sampling model above, fix $q \in (0, 1]$ and suppose $\beta > 1/(2q)$. Set $\mu_{\mathrm{CVaR}_q} := \beta - 1/(2q)$, and construct the empirical operator by applying (193) to the rows of (183). For $\delta \in (0, 1)$, if $n\rho_{\min} \geq 8 \log(2d^2/\delta)$, then with probability at least $1 - \delta$, every empirical CVaR fixed point satisfies

$$\mathrm{KL}(\pi^\star_{\mathrm{CVaR}_q} \| \widehat{\pi}_n) \leq \frac{\log(4d^2/\delta)}{\mu^2_{\mathrm{CVaR}_q} n\rho_{\min}}. \tag{195}$$

Proof. Payoff CVaR is normalized, monotone, and translation equivariant. Equation (194) supplies continuity, and the calculation above supplies (187) with the stated positive modulus. Apply Theorem 16 and retain one KL term. □

There is no extra q^{-2} factor in (195): the sampling error is in the payoff values, while the opponent distribution is evaluated exactly. The tail mass enters the sufficient conditioning constant μ_{CVaR_q}. Estimating CVaR from sampled opponent outcomes is a different observation model and can have additional tail-dependent statistical costs. Likewise, replacing $(\nu - P_{y,j})_+$ by $(\nu - z_i)_+$ for a Bernoulli label would change the target, even if the threshold were known.

K.9 Interpretation and scope

The analysis separates two sources of statistical difficulty: coverage determines how accurately the preference table can be estimated, and population monotonicity determines how that error affects the equilibrium. This separation clarifies the role of risk and the relation between the offline guarantee and computation of an empirical solution.

Risk and statistical conditioning. For this tabular observation model, normalized monotone, translation-equivariant payoff risks are 1-Lipschitz in uniformly estimated payoff values. The risk parameter therefore enters the equilibrium error bound through the population monotonicity modulus $\mu_{\mathcal{R}}$, without separate entropic exponential or CVaR tail-mass factors. For entropic risk, Corollary 19 makes this dependence explicit through the curvature bound $B_A(\lambda)$. The bound still contains the explicit coverage cost $1/\rho_{\min}$ and can be expensive for large response spaces. It is not a dimension-free guarantee for an arbitrary learned preference-model class.

Relation to risk-neutral learning. For $\mathcal{R}_\mu(v) = \sum_j \mu(j)v_j$, the constant-sum identity gives $\langle P(\pi - \pi'), \pi - \pi' \rangle = 0$, so $\bar{\lambda}_{\mathcal{R}} = 0$ and $\mu_{\mathcal{R}} = \beta$. The bound becomes $\log(4d^2/\delta)/(\beta^2 n\rho_{\min})$. The same exponent in n is retained under risk whenever the stated population-monotonicity condition holds. This comparison does not claim optimal dependence on β or on sampling coverage.

Statistical estimation and computation. The theorem concerns an empirical fixed point. It does not assert a finite-iteration convergence rate for computing that fixed point. More generally, if an interior policy $\widetilde{\pi}$ satisfies

$$\inf_{c \in \mathbb{R}} \left\| \beta \log \frac{\widetilde{\pi}}{\pi_{\mathrm{ref}}} - \widehat{\mathcal{P}}_{\mathcal{R}}(\widetilde{\pi}) - c\mathbf{1} \right\|_\infty \leq r,$$

the proof of Lemma 25 gives

$$J(\pi^\star_{\mathcal{R}}, \widetilde{\pi}) \leq \frac{(\|\widehat{P} - P\|_{\max} + r)^2}{\mu^2_{\mathcal{R}}}.$$

This separates payoff-estimation error from a measured equilibrium residual without identifying either one with an IPO generalization gap. The online analysis with a fixed number of opponent samples per iteration concerns a different estimator and its bias-tracking guarantees must be invoked under their own assumptions.

Limits of the sufficient condition. As $\mu_{\mathcal{R}}$ approaches zero, this guarantee deteriorates. Failure of the sufficient monotonicity condition does not establish nonexistence, nonuniqueness, or statistical impossibility, nor does it by itself imply a particular slower rate.

L Additional Experimental Results and Details

In this appendix, we include the experimental setup details and additional results.

L.1 Experimental Setup and Implementation Details

Below we describe each of the critical components of the experimental setup.

Base model and SFT. All policies are LoRA fine-tuned (r=256, α=512, dropout 0.1) from a common `SFT` checkpoint: Gemma-2 2B-IT supervised on Alpaca-cleaned.

Preference dataset and judge. Training prompts and pairwise preferences come from PKU-SafeRLHF (Ji et al., 2024). The online preference judge is a PairJudge sequence classifier trained on a mixture of HH-style preferences and PKU-SafeRLHF safety labels. Harmfulness is scored by the PKU Beaver cost model (lower = safer), used at evaluation time only.

Methods. We compare eight policies sharing the same `SFT` base, LoRA configuration (r=256, α=512, dropout 0.1), and judges: `SFT` (no preference fine-tuning); `EGPO` (Zhou et al., 2025) and `OMD` (K=8) (Calandriello et al., 2024b) as risk-neutral baselines; `EG` (K=8), an extragradient risk-neutral variant; `OMD-Ent` (τ=10) and `OMD-CVaR` (q=0.25), which apply entropic and lower-tail CVaR aggregation over the K=8 opponent samples; `gDRO`,a severity-prior-weighted ERM with fixed prior $p \propto (1, 2, 4, 8)$ over PKU severity strata {safe, low, medium, high} and no adversarial tilt; and gDRO-CVaR (q=0.25), which applies lower-tail CVaR aggregation across the same severity groups. All risk-trained methods use streaming EMA group losses (coefficient 0.9), a linear annealing schedule for the risk parameter over the first 20% of training, and a stratified sampler enforcing a minimum of 4 examples per group per batch.

Optimization. All NLHF methods share: 10 epochs, AdamW with learning rate 5×10^{-7}, weight decay 0.01, `bf16` precision, effective batch size 64 (micro-batch size 8), and warm-up over 1000 steps. Only the loss aggregator and group sampler differ across methods. All policies are trained on a single NVIDIA L40S GPU.

Evaluation. Each policy is evaluated via all-pairs cross-play against eight opponents: the SFT base, `SFT` at four temperatures (0.1, 0.5, 2.0, 3.0), and three off-the-shelf chat models (Qwen-1.5B, SmolLM2, Qwen-7B), generating 4 responses per (policy, prompt) pair. Held-out prompts are partitioned into four strata: **Random** (100 prompts, unstratified), **Conflict** (100 prompts where preference and safety labels conflict), **Sev-3** (100 highest-severity unsafe prompts), and **Sev-1** (136 mildest-severity unsafe prompts). Importantly, none of these strata appear in training.

Metrics. For each (policy, opponent, prompt) triple we compute three per-prompt win-rates: **Preference WR** ($P[\text{judge}(y \succ y')]$ under the PairJudge), **Safety WR** (fraction of pairs where the policy response has lower Beaver cost), and **Combined WR** (policy wins iff preferred *and* safer; otherwise inconclusive). Summaries report the mean and lower-tail $\text{CVaR}_{q=0.25}$ of the per-prompt combined win-rate distribution pooled over opponents, with bootstrap 95% CIs from 2000 resamples.

L.2 Win-Rate Distributions Across Strata

Figures 5–8 show the empirical cumulative distribution function (CDF) of the per-prompt win-rate distribution for each policy, pooled over opponents, across all four evaluation strata. A curve lying to the right at a given quantile means the policy achieves that win-rate on a larger fraction of prompts.

On the random stratum (Figure 5), the K=8 methods as a group sit substantially to the right of the single-sample baselines (`SFT`, `EGPO`) on the safety and combined panels. Within the K=8 group, risk-adjusted and risk-neutral methods are largely indistinguishable on this stratum—consistent with the theory, which predicts risk adjustment matters most under distributional differences rather than on easy, unstratified prompts.

The separation within the $K = 8$ group becomes increasingly visible on the harder strata. On the Conflict stratum (Figure 6), `OMD-Ent` and `OMD-CVaR` begin to pull right of the $K = 8$ risk-

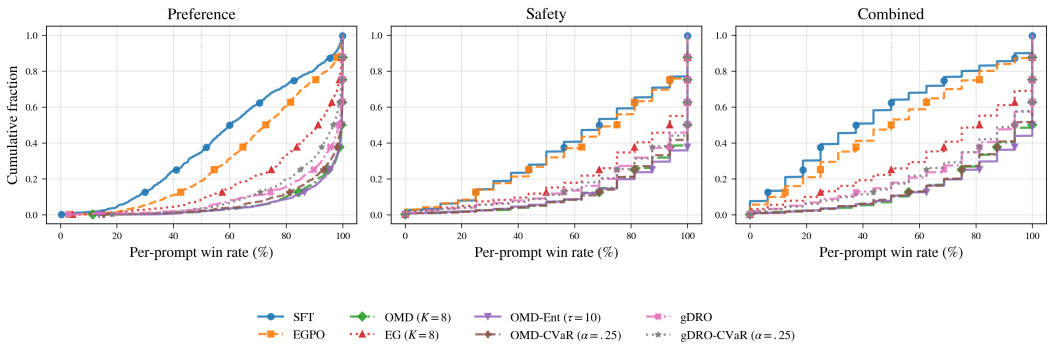


Figure 5: Per-prompt win-rate cumulative distribution functions (CDFs) on the **Random** stratum (100 prompts, pooled over opponents).

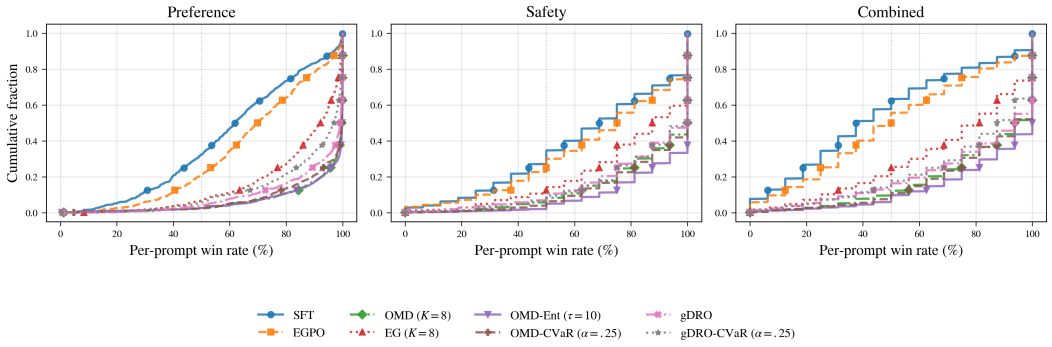


Figure 6: Per-prompt win-rate CDFs on the **Conflict** stratum (100 prompts where preference and safety labels disagree).

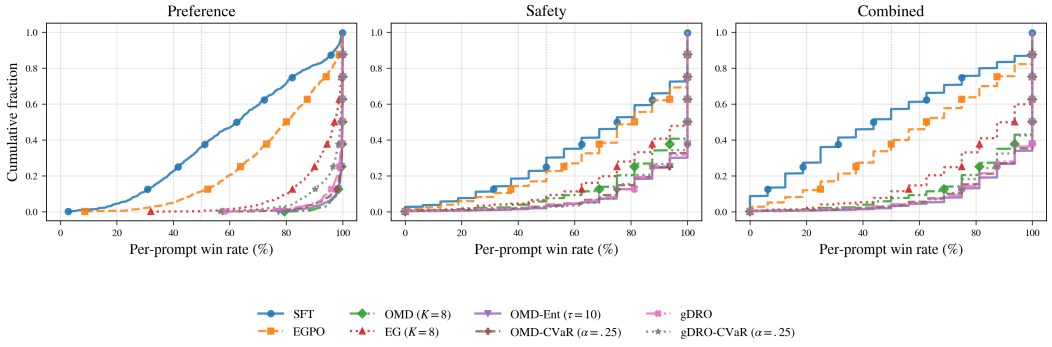


Figure 7: Per-prompt win-rate CDFs on the **Sev-3** stratum (100 highest-severity unsafe prompts).

neutral baselines in the safety and combined panels across the mid-range quantiles. This effect is most pronounced on Sev-3 (Figure 7), where risk-adjusted methods clearly retain mass at higher win-rates while the $K = 8$ neutral baselines accumulate more low-win-rate prompts at the tail, precisely where the framework predicts the gain. The Sev-1 stratum (Figure 8) shows a similar pattern at smaller magnitude. Across all strata, the preference CDFs remain tightly clustered among all $K = 8$ methods, confirming that the distributional gains on safety and combined win-rate come at no cost to preference performance.

To position our additional baselines, Figure 9 reproduces the Sev-3 CDFs with `EG-Ent` (τ=5) and `Nash-MD` added: `EG-Ent` tracks the `OMD` risk-adjusted curves closely, confirming the risk-

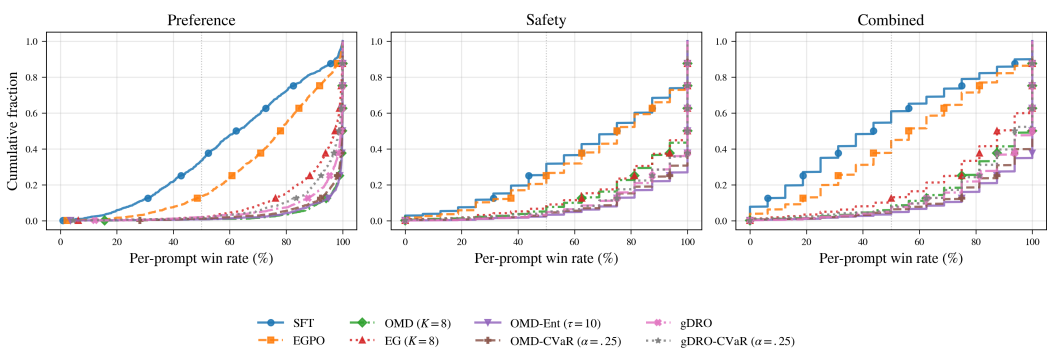


Figure 8: Per-prompt win-rate CDFs on the **Sev-1** stratum (136 mildest-severity unsafe prompts).

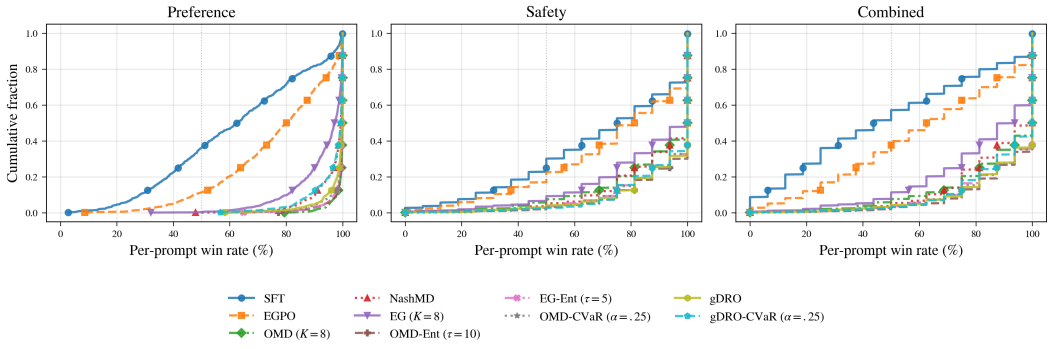


Figure 9: Per-prompt win-rate CDFs on the **Sev-3** stratum with `EG-Ent` (τ=5) and `Nash-MD` added to the comparison.

aggregation gains transfer to the extragradient framework, while `Nash-MD` sits between the single-sample and K=8 baselines.

L.3 Tail Robustness Across Opponents

Figure 11 reports the lower-tail $\mathrm{CVaR}_{q=0.25}$/Mean ratio of the combined win-rate distribution for each policy, broken out by opponent. A ratio close to 1 indicates that a policy performs consistently across prompts while a low ratio indicates the policy wins often on average but fails on a non-trivial fraction of prompts.

The risk-adjusted advantage is most visible against the harder opponents, namely low-temperature `SFT` variants (T=0.1, T=0.5) and the off-the-shelf chat models (Qwen-1.5B, SmolLM2, Qwen-7B), where ratios sit in the 0.4–0.8 range and `OMD-Ent` and `OMD-CVaR` consistently lead the K=8 risk-neutral baselines. Against the easy high-temperature opponents (T=2.0, T=3.0), all methods converge near 1.0 as every policy wins on nearly every prompt, leaving no tail to improve.

Figures 10 and 12 quantify this gap directly. Figure 10 reports the absolute tail drop in percentage points, defined as the gap between mean win-rate and each risk functional (entropic, lower-tail $\mathrm{CVaR}_{q=0.25}$, lower-tail $\mathrm{CVaR}_{q=0.125}$). Absolute drops are not directly comparable across policies, since methods with higher means have more room to fall: the largest absolute drops on safety and combined win-rate are in fact incurred by the K=8 risk-neutral methods (`OMD`,`EG`) and `gDRO`,which also achieve the highest means. The scale-free comparison is given by the Risk/Mean ratio in the bottom row of Figure 12, and there the ordering is unambiguous: `SFT` and `EGPO` are the worst on every metric, sitting well below all K=8 methods; among the K=8 group, `OMD-Ent` and `OMD-CVaR` achieve the highest ratios on preference and safety, with the risk-neutral K=8 baselines and `gDRO` in between.

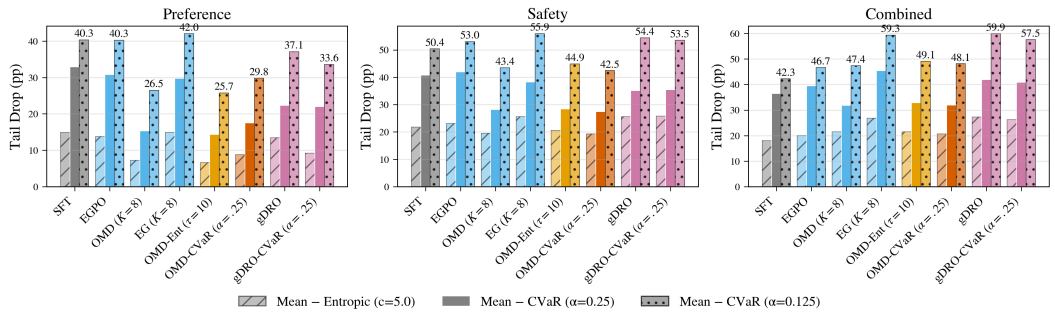


Figure 10: Tail drop in percentage points on the Random stratum, defined as the gap between mean win-rate and each risk-adjusted summary across preference, safety, and combined win-rates. Smaller values indicate less degradation at the tail.

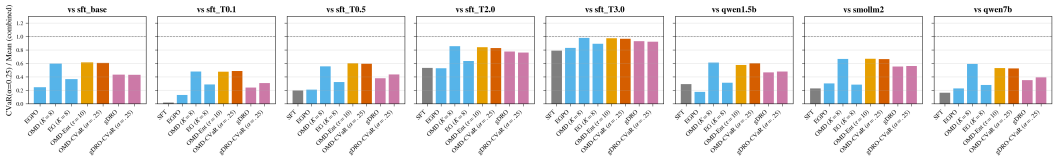


Figure 11: Lower-tail $\mathrm{CVaR}_{q=0.25}$/Mean ratio of the combined win-rate distribution per opponent on the Random stratum. Higher values indicate more consistent performance across prompts. Dashed line at 1.0 denotes perfect consistency.

The top row of Figure 12 shows the same picture geometrically. Each policy's risk-adjusted win-rate is plotted against its mean win-rate; points above the diagonal are impossible by definition, and points close to it are the most robust. `SFT` and `EGPO` sit far below the diagonal across all three metrics, while the risk-adjusted $K{=}8$ methods cluster closest to it. Together, Figures 11–12 say that single-sample methods win on easy prompts but collapse at the tail, whereas risk-adjusted $K{=}8$ training maintains tail performance proportional to the mean.

Figures 13 and 14 verify that this picture is not specific to the Random stratum or to the original eight policies. On Sev-3, `EG-Ent` ($\tau{=}5$) achieves a tail drop comparable to `OMD-Ent` and well below the $K{=}8$ risk-neutral baselines, while `Nash-MD` tracks the single-sample methods rather than the $K{=}8$ group. On the Conflict stratum, the same geometric ordering holds in the mean–vs–risk-adjusted scatter: SFT, `EGPO` , and `Nash-MD` sit furthest from the diagonal, whereas `EG-Ent` joins `OMD-Ent` and `OMD-CVaR` in the upper-right cluster of robust policies.

L.4 Response Consistency & Variance

Figure 15 reports the mean variance of win-rates across responses generated by each policy, averaged over prompts and opponents across all three metrics. Lower variance indicates the policy produces more consistent outputs across samples for the same prompt. `SFT` and `EGPO` exhibit substantially higher variance than all $K{=}8$ methods. Among the $K{=}8$ methods, the risk-adjusted variants produce the most consistent responses, with `OMD-Ent` achieving the lowest variance on both preference and safety and `OMD-CVaR` close behind. The $K{=}8$ risk-neutral baselines sit between the single-sample methods and the risk-adjusted ones. Multi-sample training reduces variance generally, but as shown in the cross-strata and per-opponent results, this alone does not translate to tail robustness. Risk adjustment reduces variance *and* maintains that reduction across opponents and data strata.

The same ordering holds on the Sev-3 stratum (Figure 16), where adding `EG-Ent` ($\tau{=}5$) and `Nash-MD` shows that `EG-Ent` matches the low-variance risk-adjusted `OMD` variants while `Nash-MD` tracks the single-sample methods.

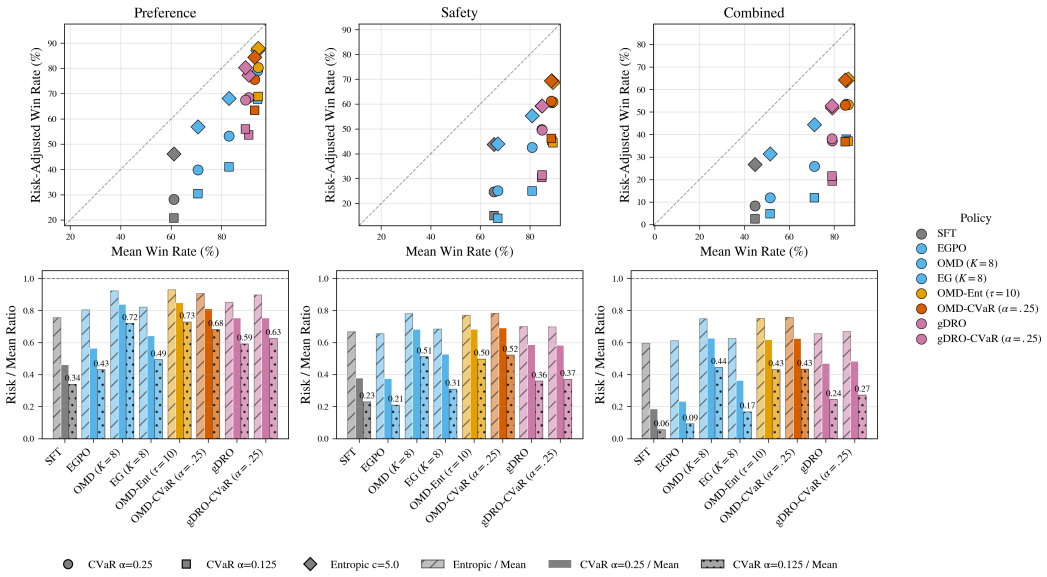


Figure 12: Mean vs. risk-adjusted win-rate (top row) and Risk/Mean ratio (bottom row) on the Random stratum, across preference, safety, and combined metrics. Points closer to the diagonal in the top row and bars closer to 1.0 in the bottom row indicate more robust policies.

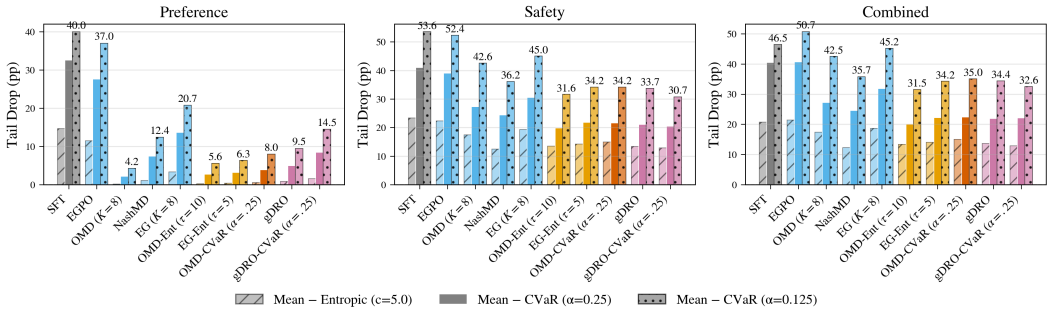


Figure 13: Tail drop on the **Sev-3** stratum with `EG-Ent` (τ=5) and `Nash-MD` added.

L.5 Cross-Play Win Rates Across Strata

Figures 17–19 report the full cross-play win-rate heatmaps for the Conflict, Sev-3, and Sev-1 strata, complementing the Random stratum heatmap in Figure 1 of the main paper. Each cell reports the win-rate of the row policy against the column opponent.

The broad pattern from the random stratum persists across all three strata: `SFT` and `EGPO` show substantially lower combined win-rates, particularly against the harder opponents, while all K=8 methods maintain strong performance. On the Conflict stratum (Figure 17), where preference and safety labels disagree, `SFT` and `EGPO` exhibit warm-colored cells on the combined panel against several opponents, reflecting their inability to simultaneously satisfy both criteria. The K=8 risk-adjusted methods, led by `OMD-Ent` , maintain consistently high combined win-rates across all opponents on this stratum.

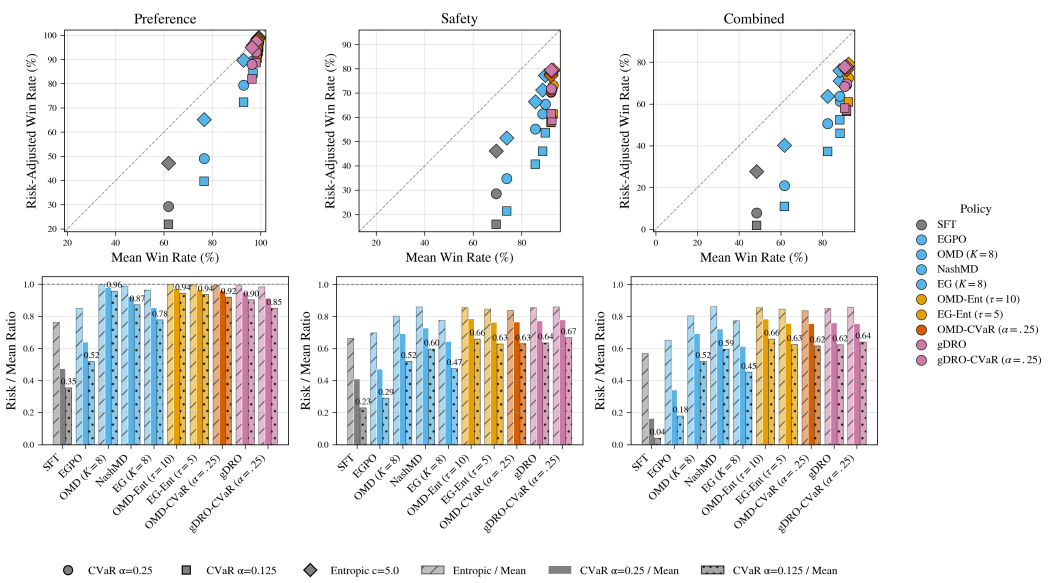


Figure 14: Mean vs. risk-adjusted win-rate (top row) and Risk/Mean ratio (bottom row) on the **Conflict** stratum with `EG-Ent` (τ=5) and `Nash-MD` added.

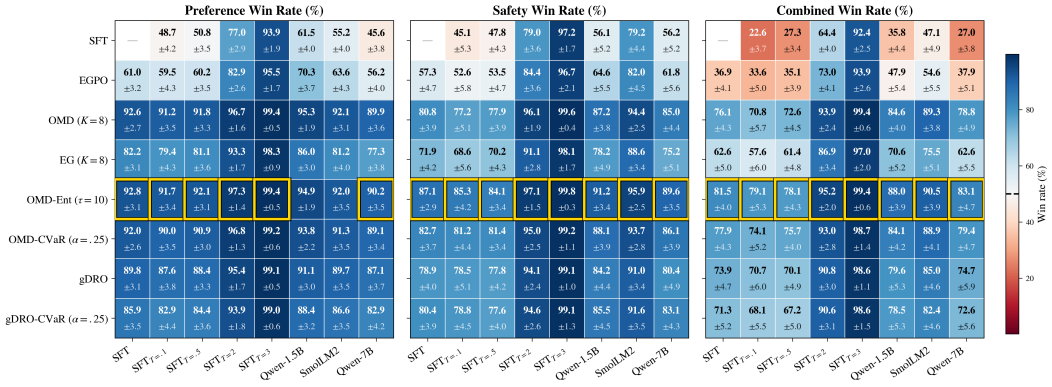


Figure 17: Cross-play win-rates on the **Conflict** stratum (100 prompts where preference and safety labels disagree), across preference, safety, and combined metrics.

The Sev-3 stratum (Figure 18) shows the sharpest separation. All K=8 methods achieve very high preference win-rates, but the combined panel reveals meaningful differences: `SFT` and `EGPO` col-

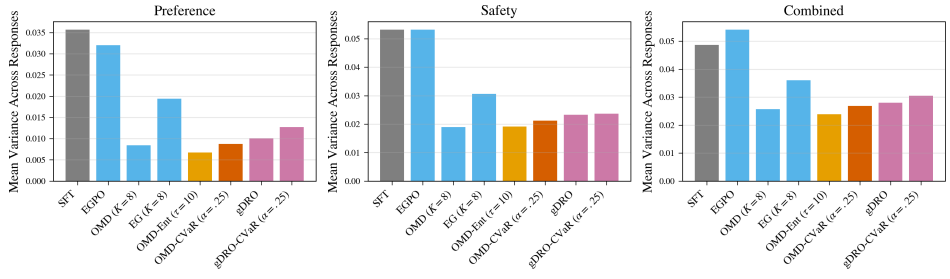


Figure 15: Mean variance of win-rates across responses, averaged over prompts and opponents. Lower values indicate more consistent policy outputs.

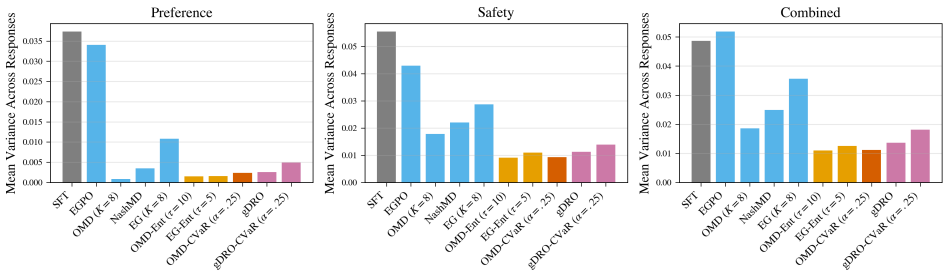


Figure 16: Mean variance of win-rates across responses on the **Sev-3** stratum with `EG-Ent` (τ=5) and `Nash-MD` added.

lapse on several opponents, with combined win-rates as low as 20–35%, while `OMD-Ent` achieves combined win-rates at or above 80% on every opponent, exceeding 95% against the easiest high-temperature `SFT` opponents. The Sev-1 stratum (Figure 19) follows a similar pattern at intermediate values, with `OMD-Ent` again leading consistently on the combined metric across all opponents.

Figure 20 extends this picture with `EG-Ent` (τ=5) and Nash-MD: `EG-Ent` posts combined win-rates in line with the K=8 risk-adjusted `OMD` variants on every opponent, while Nash-MD's combined cells track the single-sample baselines rather than the K=8 group, mirroring the pattern on the other strata.



Figure 18: Cross-play win-rates on the **Sev-3** stratum (100 highest-severity unsafe prompts), across preference, safety, and combined metrics.

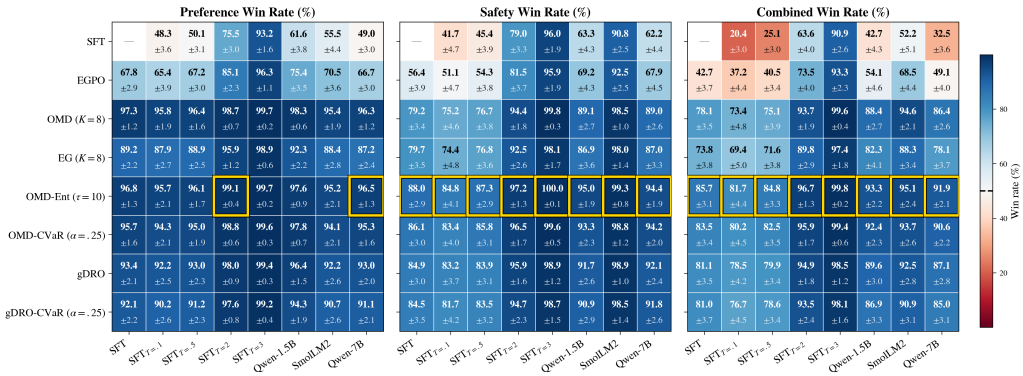


Figure 19: Cross-play win-rates on the **Sev-1** stratum (136 mildest-severity unsafe prompts), across preference, safety, and combined metrics.

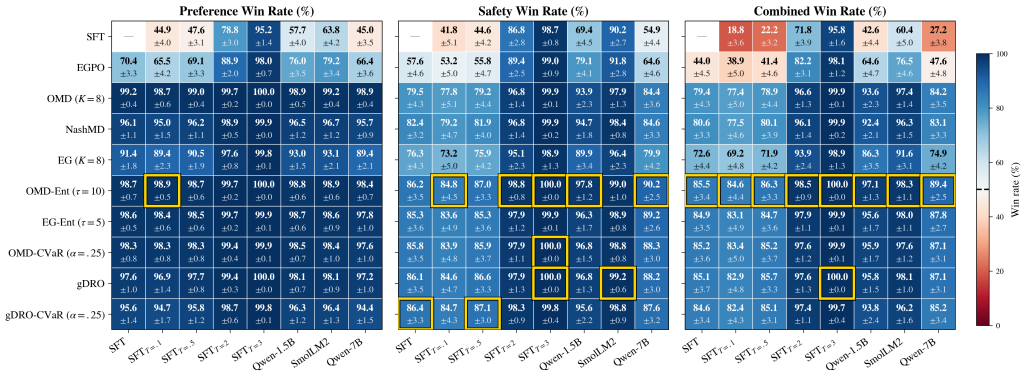


Figure 20: Cross-play win-rates on the **Sev-3** stratum with `EG-Ent` (τ=5) and `Nash-MD` added.

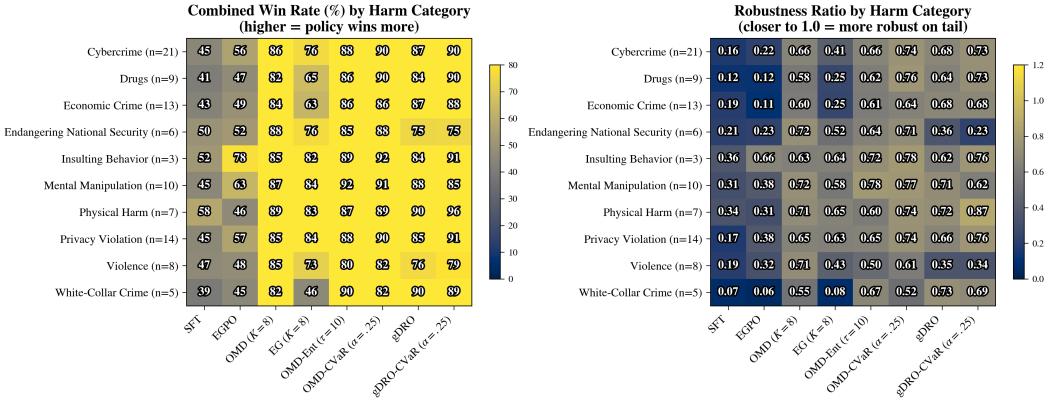


Figure 21: Combined win-rate (left) and lower-tail $\mathrm{CVaR}_{q=0.25}$/Mean robustness ratio (right) broken down by harm category on the **Random** stratum.

L.6 Robustness Across Harm Categories

Figures 21 and 22 report combined win-rate and robustness ratio broken down by harm category on the Random and Sev-3 strata respectively. The robustness ratio is lower-tail $\mathrm{CVaR}_{q=0.25}$/Mean, with values closer to 1 indicating more consistent tail performance within that category.

The left panels show that all K=8 methods achieve uniformly high combined win-rates across every harm category on both strata, with no single category standing out as a consistent failure mode. The right panels are more informative. On the Random stratum, across ten harm categories, `SFT` robustness ratios fall in the 0.07- 0.36 range and `EGPO` in the 0.06–0.66 range. The risk-adjusted methods achieve ratios of 0.60 or above across most categories, a consistent improvement over the K=8 risk-neutral baselines. On the Sev-3 stratum, the same pattern holds across seventeen harm categories with smaller per-category prompt counts: `SFT` falls in the 0.10–0.32 range and `EGPO` in the 0.26–0.55 range, while risk-adjusted methods sit at 0.64 or above across most categories. The improvement is uniform across harm types on both strata rather than concentrated in any particular category, supporting the generality of the risk adjustment mechanism.

L.7 Training Dynamics

Figures 23 and 24 report training-time diagnostics—loss, reward accuracy, KL to the `SFT` reference, and gradient norm—for `EG` ($K = 1$), `EG` ($K = 8$), and `EG-Ent` ($\tau = 5$). Figures 25 and 26 show the remaining methods, including the entropic temperature sweep $\tau \in \{2, 5, 10\}$ and lower-tail CVaR sweep $\alpha \in \{0.125, 0.25, 0.50\}$ for `OMD`, `Nash-MD`, the `gDRO` family, and the two-timescale variants. The `EG` runs are shown over 9360 steps and the remaining runs over 4680 steps.

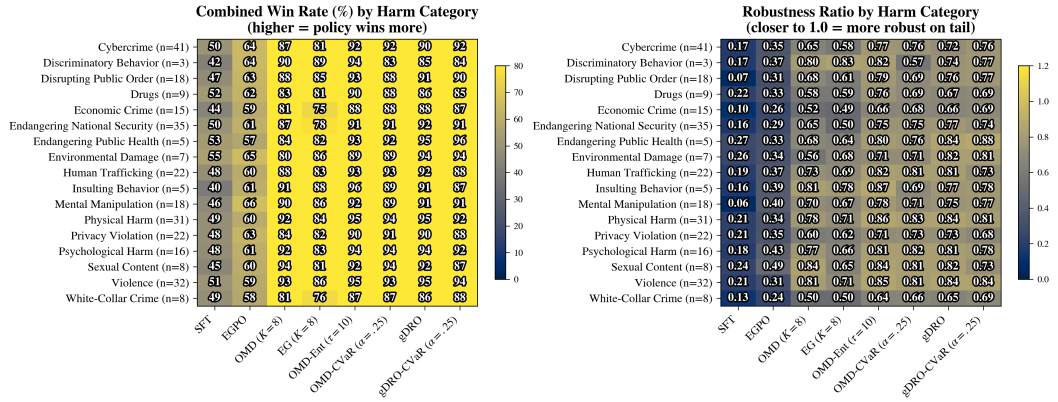
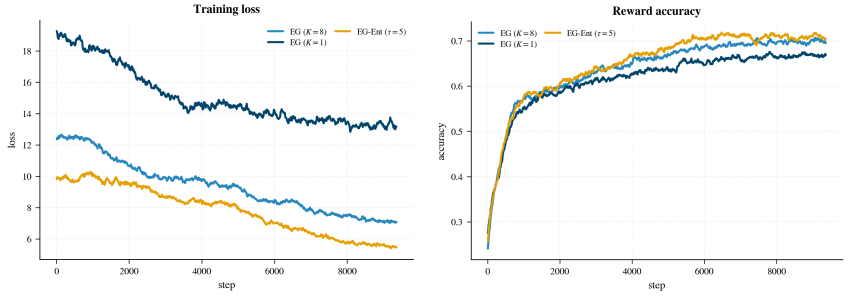


Figure 22: Combined win-rate (left) and lower-tail $\mathrm{CVaR}_{q=0.25}$/Mean robustness ratio (right) broken down by harm category on the **Sev-3** stratum.



(a) Training loss.

(b) Reward accuracy.

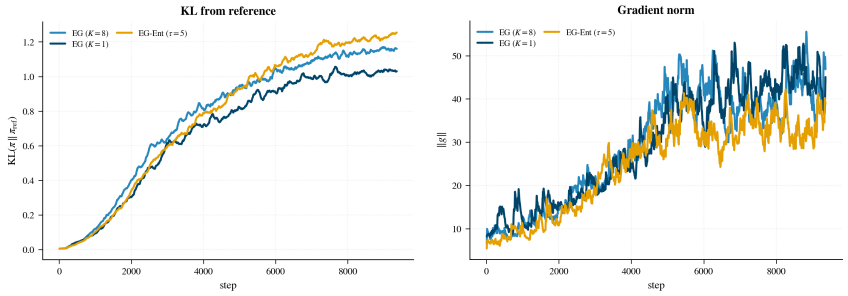
Figure 23: Training loss and reward accuracy (fraction of training pairs on which the preferred response receives the higher implicit reward) for `EG` ($K = 1$), `EG` ($K = 8$), and `EG-Ent` ($\tau = 5$) over 9360 steps.

For `EG`, loss generally decreases throughout training, while reward accuracy rises rapidly early on and improves more gradually thereafter. Both $K = 8$ methods achieve higher late-training reward accuracy than the $K = 1$ baseline, with `EG-Ent` slightly ahead of risk-neutral `EG` ($K = 8$) over much of the later trajectory. KL drift rises across all three runs, with `EG-Ent` drifting furthest by the end of training, while its late-training gradient norms are generally smaller than those of the risk-neutral baselines.

Across the remaining methods, losses generally decrease and reward accuracy improves most rapidly early in training, with later behavior varying across configurations. Loss values are not directly comparable across methods because the risk-adjusted targets and group weightings differ. Within the `OMD` entropic sweep, $\tau = 10$ reaches the lowest final loss, though the reward-accuracy ordering differs. KL drift rises before flattening or decreasing for some runs, and gradient norms vary across families: several group-reweighted variants show larger late-training fluctuations, while the $K = 8$ two-timescale runs have smaller norms. These diagnostics complement the cross-play and tail-performance results reported above.

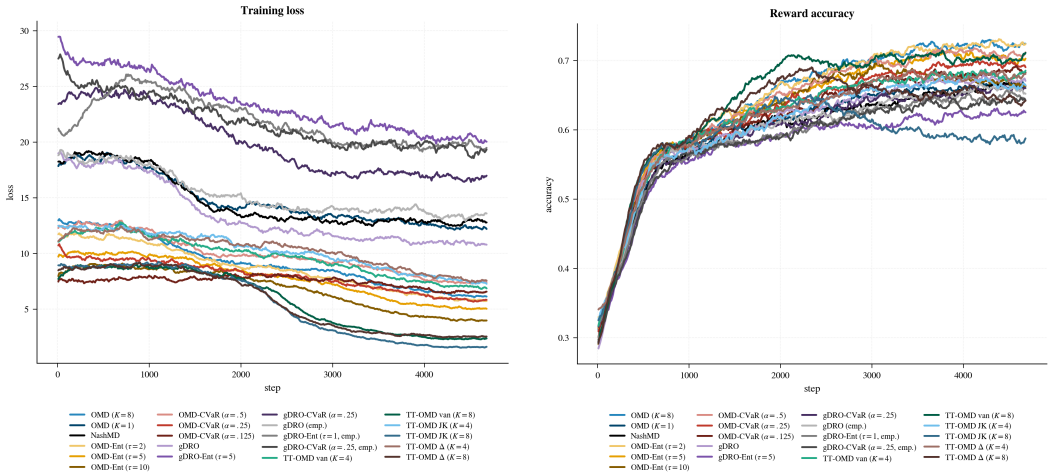
L.8 Two-Timescale RSPG for LLM Fine-Tuning

In this section, we discuss the experiments we ran on our two-timescale bias correction algorithm for finding equilibrium of RSPGs in the context of fine-tuning LLMs.



(a) KL divergence from the SFT reference. (b) Gradient norm.

Figure 24: KL drift to the SFT reference and gradient norm for the EG variants over 9360 steps.



(a) Training loss. (b) Reward accuracy.

Figure 25: Training loss and reward accuracy across OMD, Nash−MD, gDRO, and the two-timescale variants over 4680 steps. In the legends, van denotes the uncorrected control, JK denotes jackknife correction, and Δ denotes delta-method correction.

L.8.1 Understanding the Bias-Variance Tradeoff for Fine-Tunning LLMs

Setup and measurement. Risk-sensitive preference training replaces the expected payoff of a response y to prompt x by the entropic risk of its payoff distribution. Given K opponent samples $y'_1, \dots, y'_K \sim \pi(\cdot \mid x)$ with judge payoffs $P_k = P(y \succ y'_k)$, training uses the plug-in estimator

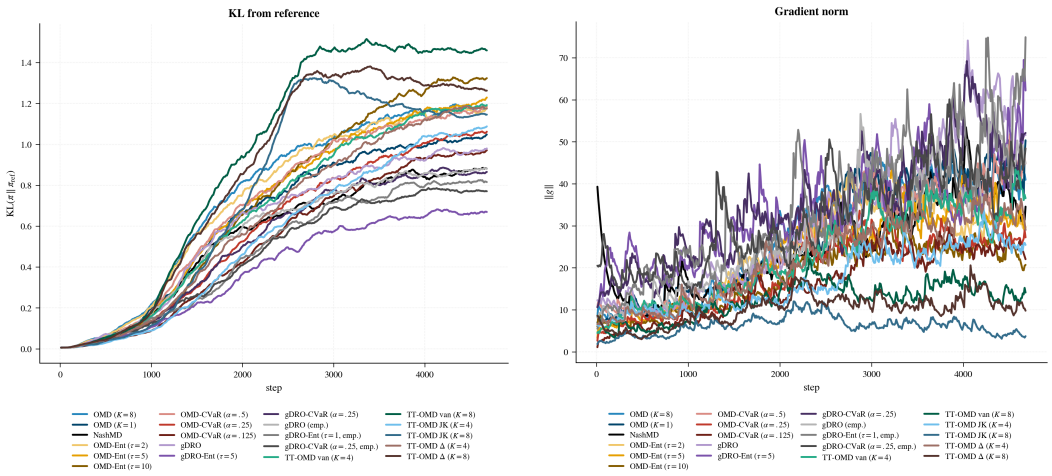
$$\rho_K = \frac{1}{c} \log \frac{1}{K} \sum_{k=1}^{K} e^{-cP_k}, \qquad c = 10,$$

whose population value is

$$\rho_\infty = \frac{1}{c} \log \mathbb{E}_{Y'' \sim \pi(\cdot|x)} \, e^{-cP(y \succ Y''|x)}.$$

By Jensen's inequality $\mathbb{E}[\rho_K] < \rho_\infty$ for every finite K: the estimator is biased downward, and unlike sampling noise this bias does not average out over prompts or steps. The two-timescale correction subtracts a running estimate of the bias obtained from either the delta method (second-order Taylor term) or the delete-one jackknife method.

The asymptotic error floor of stochastic optimization with a biased, noisy gradient is $\widetilde{B}_m^2 + \eta \, \widetilde{V}_m$, with η the step size of the recursion. When $\widetilde{B}_m^2 > \eta \, \widetilde{V}_m$ the floor is *bias-dominated*: more steps or more averaging cannot lower it, only more samples per prompt or a correction can. The question addressed here is where, in K, that regime lies for a real LLM judge.



(a) KL divergence from the `SFT` reference.

(b) Gradient norm.

Figure 26: KL drift to the `SFT` reference and gradient norm across `OMD`, `Nash-MD`, `gDRO`, and the two-timescale variants over 4680 steps.

Large-sample reference and bias–variance estimation. This experiment measures finite-sample error in estimating the entropic risk of a fixed judge's preference scores. For a fixed prompt x, anchor response y, and policy π, the population target is

$$\rho_\infty(x, y; \pi) := \frac{1}{c} \log \mathbb{E}_{Y'' \sim \pi(\cdot|x)} \left[e^{-cP(y \succ Y''|x)} \right], \qquad c = 10,$$

where $P(y \succ Y'' \mid x)$ is the preference classifier's score. The expectation is over opponent responses Y''; the prompt, anchor response, policy, and judge are held fixed.

Since this expectation cannot be evaluated exactly, we approximate it using a large pool of $M = 4096$ opponent responses:

$$\hat{\rho}_M(x, y; \pi) := \frac{1}{c} \log \left(\frac{1}{M} \sum_{j=1}^{M} e^{-cP(y \succ y''_j|x)} \right).$$

We use $\hat{\rho}_M$ as a *large-sample reference*, rather than an exact ground-truth value. It remains subject to sampling error and finite-M Jensen bias. Moreover, the target itself is defined by the judge's scores: this experiment evaluates risk-estimation error, not the judge's agreement with human preferences or safety labels.

To measure small-sample error, we repeatedly draw subsets of K responses from the pool and compute each risk estimator. Its estimated bias is its average over these subsets minus $\hat{\rho}_M$; its estimated variance measures fluctuations around that average. Thus, the resampling procedure directly measures error relative to the finite-pool reference, which approximates error relative to the population target ρ_∞. We repeat this procedure across prompts, anchor responses, and policies as follows.

1. For each of 20 held-out PKU-SafeRLHF prompts we drew two anchor responses y, y' from the policy with the training sampler (temperature 1, top-k 100, top-p 0.95, $\leq$ 200 tokens), then a pool of $M = 4096$ further responses $y''_1, \ldots, y''_M$ by pure ancestral sampling from the same policy—i.e., generating each response token by token from the model's full, untruncated next-token distribution:
$$y''_t \sim \pi(\cdot \mid x, y''_{<t}).$$
Each sampled token becomes part of the context for the next token. "Pure" here indicates temperature 1, with no top-k or top-p truncation.
2. The pairwise judge (`gemma-2-2b` preference classifier) scored y against every pool member and y' against every pool member, giving two *prompt-conditionals* (empirical payoff distributions

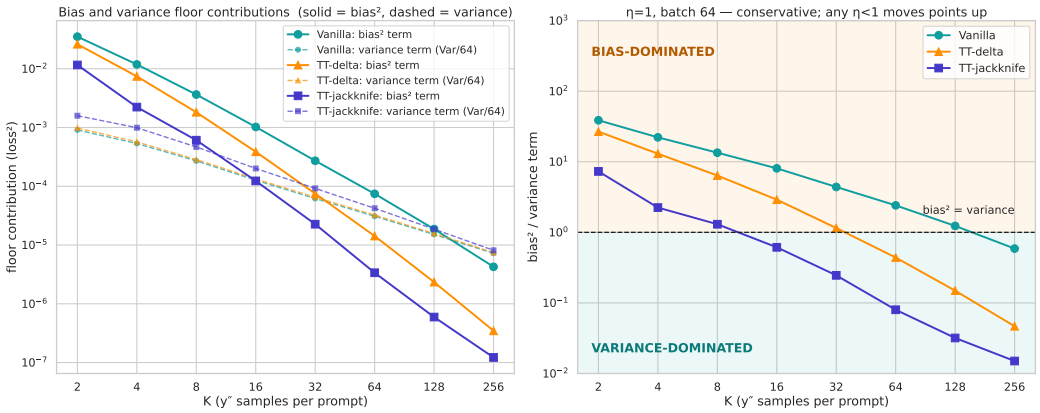


Figure 27: **Bias and variance of the entropic risk estimator on real judge conditionals.** Results use $c = 10$, a large-sample reference pool of $M = 4096$ responses per conditional, 80 conditionals, and $R = 2000$ subsets for each $K \in \{2, 4, 8, \ldots, 256\}$. *Left:* squared bias relative to the full-pool reference (solid) and estimator variance divided by batch size 64 (dashed), for the uncorrected plug-in estimator (Vanilla, teal), delta-method correction (TT-delta, orange), and jackknife correction (TT-jackknife, indigo). *Right:* the ratio $\mathrm{bias}^2/(\eta\,\mathrm{Var}/64)$ with $\eta = 1$. Values above the horizontal dashed line indicate bias domination; values below indicate variance domination.

of size M) per prompt. This was done for two policies, the trained TT-jackknife K=4 policy at step 1250 and the SFT initialization, for $20 \times 2 \times 2 = 80$ conditionals.

3. The full-pool risk $\hat{\rho}_M$ is the reference for ρ_∞. It carries the same Jensen bias at sample size M and therefore understates ρ_∞ by about K/M of the K-sample bias; the largest K reported is $256 = M/16$, so the reference error is at most 6% of the quantity being measured and below 1% for $K \le 32$.
4. For each $K \in \{2, 3, 4, 5, 6, 8, 12, 16, 32, 64, 128, 256\}$ we drew $R = 2000$ subsets of size K without replacement from each pool, computed ρ_K on each, and applied each correction (none, delta, jackknife). Bias is the subset-mean minus $\hat{\rho}_M$; variance is the spread over subsets; both are averaged over the 80 conditionals. The variance term is divided by the training batch size 64 (noise averages across prompts in a batch; bias does not) and we take $\eta = 1$, the most conservative setting for the bias claim.

On η. The training runs used AdamW with a peak learning rate of 5×10^{-7} (100 warm-up steps, linear decay to 5×10^{-8}). Adam rescales each coordinate by its running gradient RMS, so the learning rate is not the same constant as the η multiplying the variance in the floor bound, and there is no clean conversion. We therefore treat $\eta = 1$ as a hard upper bound (per-step noise given full weight, no averaging at all); every realistic value is far below it, which only enlarges the bias-dominated region and the region in which the jackknife is the floor-minimizing estimator (Figure 28, right).

On the constants in the theory. The paper's threshold $m \lesssim 1/(\lambda^2\eta)$ comes from $\mathrm{bias} \asymp 1/(\lambda m)$, which absorbs the factor $\mathrm{Var}(e^{-\lambda P})/q^2$ into a constant. On the smooth logistic payoffs of the toy that factor is $O(1)$; with $\lambda = 10$ the formula would then predict variance domination for all $m > 1/(100\eta)$. On the real judge the payoffs are near-binary, and for a two-point conditional with loss probability p the factor is $\approx (1-p)/p$, i.e. 10–100 on the prompts that carry the bias (Figure 29). The scaling in m and λ predicted by the theory is confirmed below; the constants on real preference data push the regime boundary roughly two orders of magnitude further out than the constant-free formula suggests.

Bias dominates at training sample sizes. Figure 27 and Table 1 compare squared bias with batch-averaged variance as the number of opponent samples K increases. For the uncorrected estimator, their ratio is 38.7, 22.2, 13.5, and 8.1 at $K = 2, 4, 8, 16$, respectively, and reaches one only near

$K \approx 150$. At the sample sizes used in training ($K = 2, 4$), highlighted in Table 1, squared bias therefore exceeds the variance term by more than an order of magnitude. Averaging additional estimates at fixed K reduces variance but leaves this systematic bias unchanged.

The jackknife removes 43–65% of the bias for $K \leq 16$, rising to approximately 80% by $K = 64$, at the cost of 1.6–1.9× the estimator variance at small K. It moves the bias–variance crossover from $K \approx 150$ to $K \approx 12$. By comparison, the delta-method correction is capped at $1/(2c)$ per prompt on the near-binary judge conditionals and removes only 14–38% of the bias for $K \leq 16$, with a crossover near $K \approx 35$. All plotted curves decrease monotonically, and neither correction reverses the sign of the aggregate bias at any tested K.

The highlighted rows of Table 1 quantify this tradeoff at the training sample sizes. At $K = 2$, the jackknife reduces absolute bias from 0.188 to 0.108, compared with 0.162 for the delta method, while increasing variance by factors of 1.75 and 1.08, respectively. At $K = 4$, the jackknife reduces absolute bias from 0.109 to 0.047, removing 57% of the bias, whereas the delta method leaves 0.086, removing only 21%. The corresponding variance factors are 1.86 and 1.07. Thus, the delta method largely preserves variance but leaves most of the systematic error, whereas the jackknife trades a larger variance increase for substantially greater bias reduction. The jackknife's bias–variance ratios nevertheless remain above one—7.3 at $K = 2$ and 2.2 at $K = 4$. Correction therefore substantially mitigates, but does not eliminate, bias domination at the sample sizes used in training.

The comparison uses $\eta = 1$, giving the variance term full weight in the diagnostic $\text{bias}^2 + \eta\,\text{Var}/64$. For any smaller positive η, the ratios increase, extending the bias-dominated regime. The full-pool reference remains an approximation: under the leading-order $1/K$ bias scaling, its own Jensen bias is approximately K/M of the population K-sample bias—about 6.25% at $K = 256$ and below 1% for $K \leq 32$.

Distribution-dependent constants amplify the bias. The magnitude of the finite-sample bias depends on the distribution of judge payoffs, not just the sample count. Writing $q = \mathbb{E}[e^{-\lambda P}]$, the leading-order bias is

$$-\frac{1}{2\lambda K}\frac{\text{Var}(e^{-\lambda P})}{q^2}.$$

On the near-binary judge conditionals carrying substantial bias, the normalized variance $\text{Var}(e^{-\lambda P})/q^2$ is of order 10–100. This amplification is consistent with the concentration of large negative biases on rare-loss conditionals in Figure 29 and helps explain why bias domination persists to $K \approx 150$ in Figure 27, well beyond the training sample sizes.

	|bias| (loss units)			bias recovered		Var inflation		$\text{bias}^2/(\text{Var}/64)$		
K	vanilla	delta	jackknife	delta	jackknife	delta	jackknife	vanilla	delta	jackknife
2	0.188	0.162	0.108	14%	43%	1.08	1.75	38.7	26.7	7.3
3	0.139	0.114	0.065	18%	53%	1.07	1.86	28.2	17.9	3.3
4	0.109	0.086	0.047	21%	57%	1.07	1.86	22.2	13.0	2.2
5	0.091	0.070	0.038	23%	59%	1.06	1.82	18.9	10.5	1.8
6	0.078	0.058	0.033	26%	58%	1.06	1.79	16.6	8.7	1.6
8	0.060	0.043	0.025	29%	59%	1.05	1.72	13.5	6.4	1.3
12	0.041	0.027	0.016	35%	60%	1.05	1.64	9.8	3.9	1.0
16	0.032	0.020	0.011	38%	65%	1.05	1.58	8.1	2.9	0.6
32	0.017	0.009	0.005	47%	71%	1.06	1.49	4.4	1.2	0.3
64	0.009	0.004	0.002	56%	79%	1.05	1.36	2.4	0.4	0.1
128	0.004	0.002	0.001	65%	82%	1.03	1.23	1.2	0.2	0.0
256	0.002	0.001	0.000	71%	83%	1.03	1.13	0.6	0.1	0.0

Table 1: **Numbers behind Figure 27.** $c = 10$, 80 conditionals, $M = 4096$ reference. "Bias recovered" $= 1 - |\text{bias}_{\text{corr}}|/|\text{bias}_{\text{vanilla}}|$; "Var inflation" $= \text{Var}_{\text{corr}}/\text{Var}_{\text{vanilla}}$; the ratio column uses $\eta = 1$. Training used $K = 2, 4$, and 8.

Bias reduction outweighs the variance cost. Figure 28 quantifies the tradeoff between reducing bias and increasing estimator variance. The jackknife removes 43–65% of the bias for $K \leq 16$ and at least 80% by $K = 128$, compared with 14–38% for the delta method at $K \leq 16$. Both

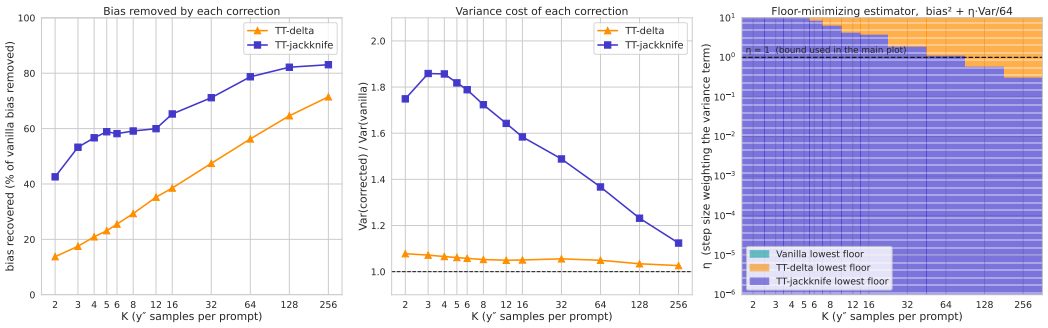


Figure 28: **Bias reduction, variance cost, and estimator selection.** Results use $c = 10$ and the same reference pools as Figure 27. *Left:* fraction of the uncorrected estimator's bias removed, $1 - |\mathrm{bias}_{\mathrm{corr}}|/|\mathrm{bias}_{\mathrm{vanilla}}|$. *Middle:* estimator variance relative to vanilla; the dashed line indicates no variance inflation. *Right:* the estimator minimizing the floor diagnostic $\mathrm{bias}^2 + \eta\,\mathrm{Var}/64$ for each sample count K and variance weight η. Colors identify Vanilla (teal), TT-delta (orange), and TT-jackknife (indigo). The horizontal dashed line marks $\eta = 1$, used in Figure 27.

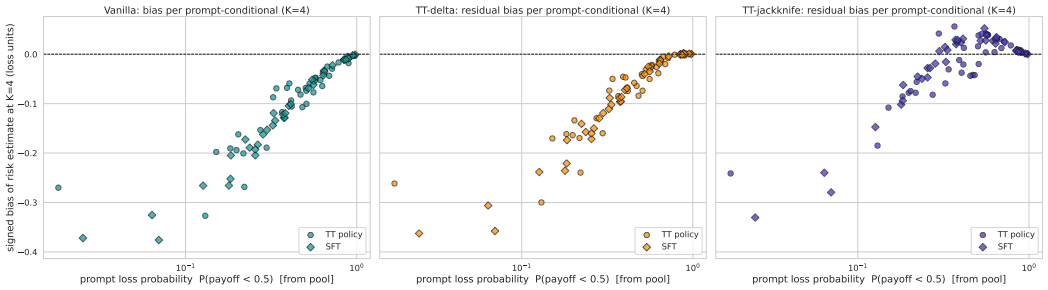


Figure 29: **Where the estimation bias concentrates.** Signed bias at $K = 4$ and $c = 10$, measured relative to the full-pool risk reference ($M = 4096$), for the uncorrected estimator (left), delta-method correction (middle), and jackknife correction (right). Each point represents one of 80 conditionals defined by a prompt, anchor response, and policy. The horizontal axis shows the pool-estimated probability of a judge payoff below 0.5, on a logarithmic scale. Circles denote the trained two-timescale (TT) policy; diamonds denote the SFT initialization. The horizontal dashed line marks zero bias; positive values indicate over-correction.

corrections improve as larger samples better represent the conditional payoff distributions. The jackknife increases variance by approximately 1.6–1.9× at small K, with this factor approaching one as K increases. The delta method incurs at most an 8% variance increase.

Despite its higher variance, the jackknife minimizes $\mathrm{bias}^2 + \eta\,\mathrm{Var}/64$ among the three estimators for all $0 < \eta \leq 1$ at every tested $K \leq 64$, and across all tested sample counts when η is sufficiently small. The delta method becomes preferable only when variance receives greater weight; for $K \leq 8$, this requires η of roughly six or larger. Vanilla never minimizes the diagnostic over the plotted range $\eta \leq 10$. Thus, when the variance contribution is strongly attenuated, the jackknife's greater bias reduction outweighs its variance cost, supporting its use in small-K training.

Bias concentrates on rare-loss conditionals. Figure 29 shows that the uncorrected estimator's largest negative biases occur on conditionals where the anchor response rarely loses. At loss probabilities below 0.1, biases range from approximately −0.27 to −0.38, whereas bias generally approaches zero as the loss probability increases. Thus, small-K estimation can substantially understate risk precisely where low-payoff outcomes are rare and risk-sensitive training is intended to emphasize them.

The delta-method correction shifts estimates toward zero but leaves substantial negative bias on rare-loss conditionals. The jackknife produces a larger reduction, bringing residuals much closer to zero

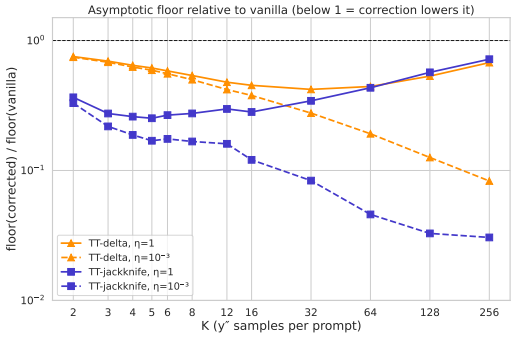


Figure 30: **Floor relative to vanilla.** $(\mathrm{bias}^2 + \eta\,\mathrm{Var}/64)_{\mathrm{corr}}/(\cdot)_{\mathrm{vanilla}}$ for $\eta = 1$ (solid) and $\eta = 10^{-3}$ (dashed). At $\eta = 1$ the jackknife lowers the floor by 2.7–3.9× across $K \in [2, 16]$ and the delta method by 1.3–2.2×; the gains shrink toward 1× at large K as the vanilla bias itself vanishes. At $\eta = 10^{-3}$, where the variance term is negligible, the ratio is simply $(\mathrm{bias}_{\mathrm{corr}}/\mathrm{bias}_{\mathrm{vanilla}})^2$ and the jackknife reaches 30× by $K = 256$.

for loss probabilities above approximately 0.3, with mild over-correction up to approximately $+0.05$ on some conditionals. This conditional over-correction can coexist with the negative aggregate bias reported above.

Both corrections remain limited by the information available in a small sample. With only four opponent responses, rare-loss conditionals often yield subsets containing no loss, providing little evidence of the low-payoff tail. Residual bias therefore remains substantial when loss probabilities are small relative to $1/K$. Increasing K makes these outcomes more likely to be observed, explaining the improved bias recovery in Figure 28.

Bias correction reduces the combined error floor. Figure 30 evaluates the net benefit of correction by comparing the diagnostic $\mathrm{bias}^2 + \eta\,\mathrm{Var}/64$ with its uncorrected value. Unlike the bias–variance ratio in Figure 27, this comparison accounts for the cost of increased variance when assessing improvement. Both corrections remain below the uncorrected baseline throughout the plotted range: their bias reduction outweighs their variance inflation under both choices of η.

At $\eta = 1$, the jackknife lowers the diagnostic by 2.7–3.9× across $K \in [2, 16]$, compared with 1.3–2.2× for the delta method. The jackknife therefore provides the greater net benefit in the small-sample regime, despite its larger variance. At larger K, the uncorrected bias decreases and variance accounts for more of the diagnostic, narrowing the gains from correction. The delta method becomes slightly preferable at $K = 128, 256$, where its lower variance outweighs the jackknife's stronger bias reduction, consistent with Figure 28.

At $\eta = 10^{-3}$, the variance contribution is small, so the relative diagnostic is approximately $(\mathrm{bias}_{\mathrm{corr}}/\mathrm{bias}_{\mathrm{vanilla}})^2$. The jackknife then outperforms the delta method at every tested K, reaching approximately a 30× reduction relative to vanilla at $K = 256$. Thus, increasing K has different effects depending on the weight assigned to variance: when variance contributes substantially, the relative benefit of correction narrows; when it is strongly attenuated, the greater fraction of bias removed at larger K produces larger relative gains.

L.8.2 Two-timescale bias correction at small K: a controlled study

In this subsection we describe the experiments we ran for the small K regime (focusing on $K = 2, 4, 6, 8$ following the analysis of the previous subsection), wherein the bias dominates the variance in the asymptotic error. This is the regime where bias correction can make an impact on performance.

Setup. Three policies are trained from the same SFT initialization with identical hyperparameters (OMD/IPO, entropic risk with τ=10, β=0.1, batch 64, learning rate $5 \cdot 10^{-7}$, 1000 warm-up steps, 4680 steps, 64 new tokens, seed 42). For $K = 2, 4, 6, 8$, the policies differ only in how the K-sample risk estimate enters the gradient: **OMD** uses the plug-in estimate $\hat{\rho}_K$ directly; **TT-OMD**

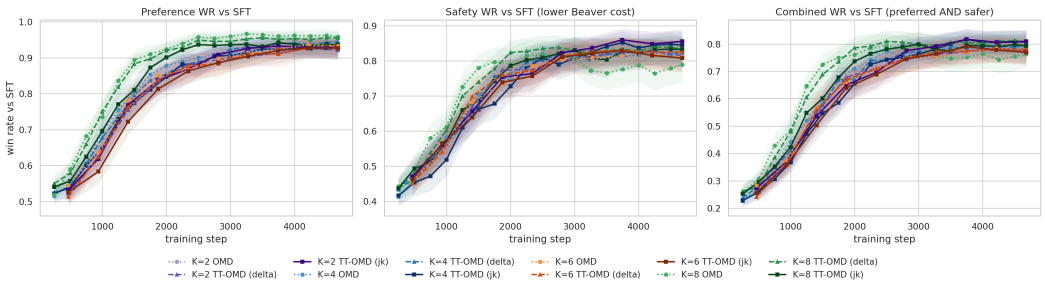


Figure 31: **Win-rates `max_new_tokens` $= 64$.** Preference, safety and combined win-rate against SFT along training, 64-token protocol (100 prompts $\times$ 4 responses, 95% bootstrap bands over prompts). Color family = K shade and line style = method (OMD dotted, TT-OMD delta dashed, TT-OMD jk solid). The methods share the preference curve; the uncorrected $K{=}8$ run's safety declines after step 2500 while the corrected runs' does not.

runs the two-timescale tracker of Algorithm 1 with either the delta-method or the jackknife bias estimator (tracker rate $\gamma{=}0.1$). We focus mostly on $K = 4$, and show $K = 2, 6, 8$ where it helps the interpretation. For the evaluation we follow Appendix L: the best corrected checkpoint is selected by the combined win-rate on the Random stratum (mean over opponents), which gives `TT-OMD` (jk) at step 4250, and *all three methods are compared at that step*; SFT is shown as a reference. Win-rates are averaged over the eight opponents, CIs are 95% bootstrap intervals over prompts, and method differences are computed paired on the same prompts.

Win-rates along training. Figures 31 and 32 show the preference, safety and combined win-rates against SFT at every saved checkpoint of every arm (100 held-out prompts, 4 responses per prompt at 64 tokens, 8 at 200 tokens; bands are 95% bootstrap intervals over prompts). Three things are visible that the best-checkpoint comparisons above cannot show.

Preference is not where the methods differ. On the paper protocol (Figure 31, left) the three methods of a given K trace the same preference curve to within the bands throughout training, and the $K{=}8$ arms lead the $K{=}4$ arms by ≈ 1000 steps. Whatever the correction does, it does not change how quickly or how far the policy learns to be preferred.

The uncorrected run's safety peaks and then falls; the corrected runs' does not. At $K{=}8$ the OMD safety win-rate reaches 82% at step 2500 and then declines to 79% by 4000 (64 tokens) and from 69% to 62% (200 tokens), while its preference win-rate keeps rising (96%; 80 $\rightarrow$ 84% at 200 tokens). Both corrected $K{=}8$ arms stay at 83% safety on the paper protocol and at 70–73% at 200 tokens, with the same or slightly lower preference. This is the fixed-point prediction of Section L.8.2 seen in the raw curves: training on $\hat{\rho}_K$ is training with the weaker regularisation β/κ, so the uncorrected policy keeps moving toward the preference optimum after the corrected one has settled, and it pays in safety. At $K{=}4$ the same ordering appears later and more weakly (jk 85% vs. OMD 82% safety at 4680 on the paper protocol; 73 vs. 70% at 200 tokens), consistent with the $K{=}4$ runs being about a thousand steps behind.

Under the length shift the gap opens. At 200 new tokens (Figure 32) all curves sit lower because the trained policies answer in ≈ 45 words while SFT uses ≈ 90, but the shape of the late-training divergence is the same and larger: the $K{=}8$ OMD combined win-rate falls from 56% (step 2500) to 51% (4680), the corrected arms hold 55–61%. The corrected policies are therefore not only safer at their best checkpoint, they are safer at *every* late checkpoint, which removes the need for early stopping on a held-out safety signal that the uncorrected run would require.

The finite-K bias is large, and the tracker removes most of it. The plug-in estimate of the entropic risk is biased downward at $K{=}4$: measured against 32-sample pools of the policy's own responses, its bias is -0.087 (loss units) at step 1500 and still -0.073 at step 4500, i.e. of the same order as the risk gaps the gradient acts on. Figure 33 compares, at three checkpoints of each corrected arm, the true bias of the policy's own K-sample estimate (orange, from the pools) with the amount the tracker was subtracting at that step (teal, from the training log). With the jackknife estimator

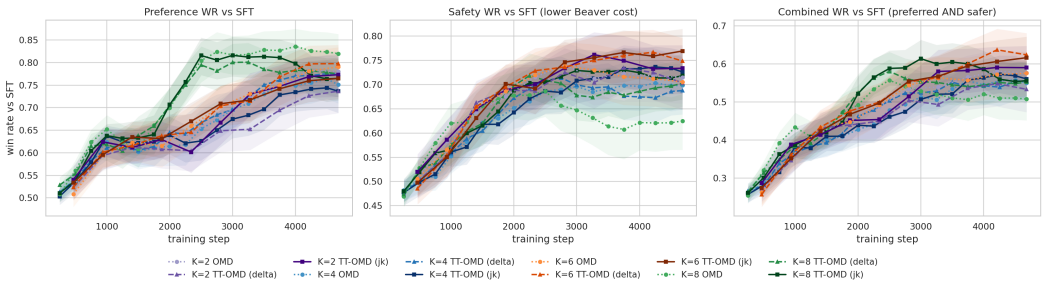


Figure 32: **Win-rates `max_new_tokens` $= 200$.** Preference, safety and combined win-rate against SFT along training, 200 new tokens (100 prompts $\times$ 8 responses), an out-of-distribution generation length for policies trained with `max_new_tokens` $= 64$. All curves are lower because the preference judge favours the longer SFT answers; the late-training safety decline of the uncorrected K=8 run (69 $\rightarrow$ 62%) is larger than at 64 tokens, and the corrected runs hold 70–73%.

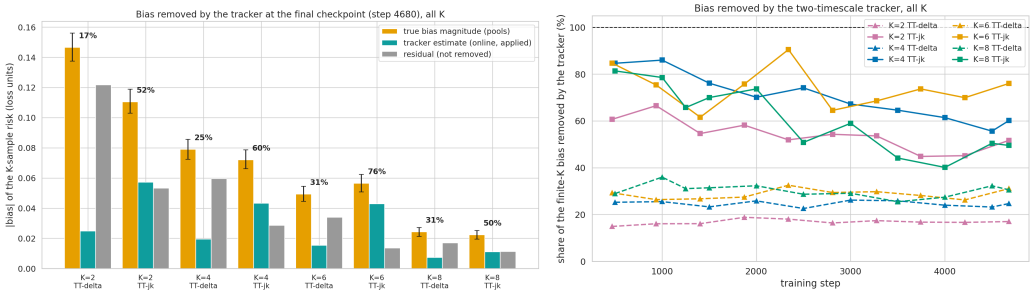
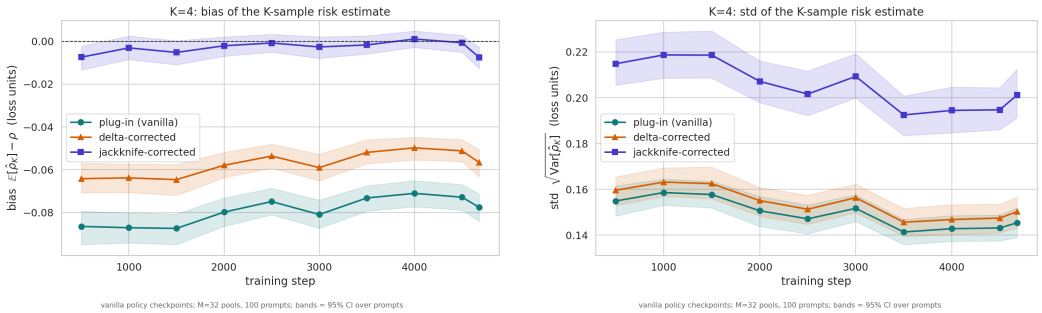


Figure 33: **How much of the finite-K bias the tracker removes.** ***Left***: for each corrected arm at the final step 4680, the true bias magnitude of the K-sample estimate on the arm's own policy (orange); 32-sample pools, 100 prompts, 95% CI), the tracker's online estimate at that step (teal) and the residual (gray); the label is the share removed. ***Right***: the same share at every pooled checkpoint (every 500 steps to 4680). The jackknife tracker removes 50–85% of the bias, the delta-method tracker 25–35%.

the tracker removes 76%, 67% and 56% of the bias at steps 1500, 3000 and 4500 (70%, 59%, 50% at K=8); with the delta-method estimator it removes a flat 23–26% (29–32% at K=8). The delta method keeps only the second-order term of the bias expansion, which for the entropic functional at K=4 is a poor approximation; the jackknife is exact to $O(1/K^2)$ and captures most of it. The residual of the jackknife tracker grows late in training because the policy's response distribution sharpens and the tracker, which averages over many steps, lags slightly behind the moving target.

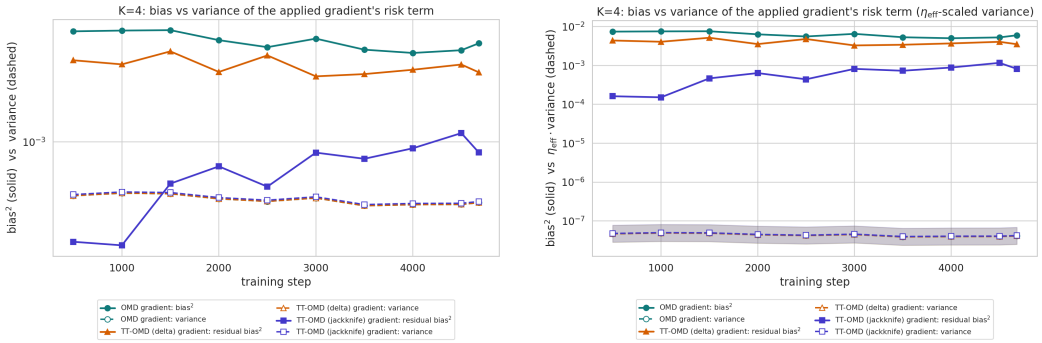
Across K. The true finite-K bias halves with each doubling of K (0.15, 0.08, 0.05, 0.02 at the final checkpoint for K=2, 4, 6, 8), the jackknife tracker removes 50–76% of it at every K, and the delta tracker 17–31%; the jackknife's share is lowest at K=2 (52%), where the jackknife formula itself can only remove two thirds of the bias.

Why a bias tracker rather than a per-step correction: the bias–variance trade. Removing the bias shifts the variance depending on the bias estimator chosen; from a practical point of view, it is helpful to know where the bias-variance regimes are relative to K and m so that an appropriate estimator can be chosen. Figure 34 reports, along training, the bias and the per-prompt standard deviation of the three K-sample estimators (plug-in, delta-corrected, jackknife-corrected), measured by drawing K-subsets from the same pools. The jackknife estimator is essentially unbiased throughout (-0.001 ± 0.004 at step 4500) but its standard deviation is 35% larger than the plug-in's (0.195 vs. 0.143), so as a *single-step* estimate it is worse in RMSE (0.195 vs. 0.161): the variance it adds exceeds the bias it removes. The delta-corrected estimator is the cheap compromise: it removes a third of the bias for a 3% increase in standard deviation and has the lowest single-step RMSE of the three. This is exactly the situation the two-timescale scheme is designed for. The bias is a



(a) bias (b) per-prompt standard deviation

Figure 34: **Bias and standard deviation vs trainign step**. Three $K{=}4$ risk estimators (32-sample pools, 100 prompts, 500 K-subsets per prompt; bands are 95% CIs over prompts). The jackknife-corrected estimator is unbiased but has a 35% larger standard deviation; the delta-corrected estimator removes about a third of the bias for a 3% increase in standard deviation.



(a) bias^2 vs. variance (b) bias^2 vs. η_{eff}·variance

Figure 35: **Bias–variance regime:** applied gradient risk term at $K{=}4$ along training (log scale; 32-sample pools, 100 prompts). Solid: squared bias of the plug-in estimate (OMD) and squared residual bias left by the delta and jackknife trackers in training (TT−OMD). Dashed: variance of the batch-mean risk term, $\mathrm{Var}[\hat{\rho}_K]/B$ with $B{=}64$, plus the tracker's own noise for TT−OMD; in (b) multiplied by the measured effective step size η_{eff} (band = range of the estimate). A solid line above its dashed line means bias-dominated.

slowly varying functional of the policy, while the noise of the jackknife estimate is fresh at every step; tracking the bias on the slow timescale ($\gamma{=}0.1$ averages roughly the last ten batches, i.e. ≈640 prompt-level estimates) keeps the full bias removal but divides the added variance by the effective window, and the policy update on the fast timescale sees a correction that is both nearly unbiased and low-variance. Applying the jackknife correction directly in every gradient would instead inject the 35% extra variance into every update.

Which regime are we in? Figure 35 compares, on one logarithmic axis, the squared bias and the variance of the risk term of the gradient that each method actually applies. For OMD the applied risk term is the plug-in estimate $\hat{\rho}_K$: its squared bias is the finite-K bias measured on the pools and its variance is $\mathrm{Var}[\hat{\rho}_K]/B$ with $B{=}64$ prompts per batch. For TT-OMD the applied risk term is $\hat{\rho}_K - \bar{b}_t$, where $\bar{b}_t$ is the tracker's running estimate: its squared bias is the *residual* bias left in training (true bias at the checkpoint minus $\bar{b}_t$ read from the training log) and its variance is the plug-in variance plus the tracker's own noise, $\frac{\gamma}{2-\gamma}\mathrm{Var}[\hat{b}]/B$ for an exponential average with gain $\gamma{=}0.1$ of independent per-batch corrections. The following are three observations:

Table 2: Regime numbers at $K{=}4$ (loss^2 units): squared bias and variance of the applied gradient's risk term for OMD and for TT-OMD with the jackknife tracker.

step	OMD bias^2	OMD variance	TT-OMD (jk) residual bias^2	TT-OMD (jk) variance
1000	$7.5 \cdot 10^{-3}$	$3.9 \cdot 10^{-4}$	$1.5 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$
1500	$7.6 \cdot 10^{-3}$	$3.9 \cdot 10^{-4}$	$4.7 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$
3000	$6.5 \cdot 10^{-3}$	$3.6 \cdot 10^{-4}$	$8.2 \cdot 10^{-4}$	$3.7 \cdot 10^{-4}$
4500	$5.3 \cdot 10^{-3}$	$3.2 \cdot 10^{-4}$	$1.2 \cdot 10^{-3}$	$3.3 \cdot 10^{-4}$
$\eta_{\text{eff}}\cdot$variance with $\eta_{\text{eff}} \approx 1.2 \cdot 10^{-4}$: $4\text{–}5 \cdot 10^{-8}$ for both methods.				

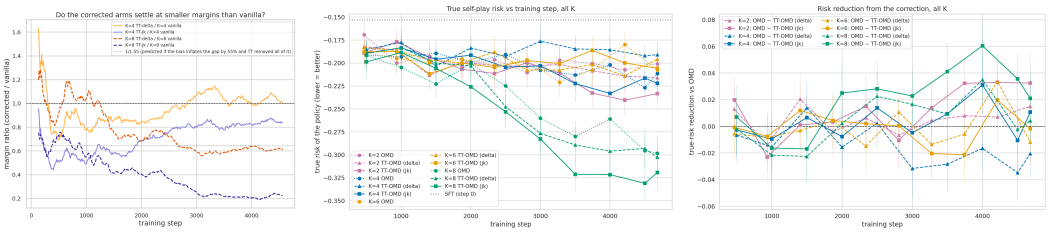
- ***OMD is bias-dominated by a wide margin.*** Squared bias $\approx 6 \cdot 10^{-3}$ against variance $\approx 3.5 \cdot 10^{-4}$ at every checkpoint, a factor 15–20. The delta tracker lowers the squared bias only to $\approx 3 \cdot 10^{-3}$, still a factor 10 above the variance: removing a quarter of the bias does not change the regime.
- ***The jackknife tracker brings the bias down to the noise floor.*** Its residual squared bias is $1.5{\cdot}10^{-4}$ at steps 500–1000, below the variance, and $0.5\text{–}1.2 \cdot 10^{-3}$ from step 1500 on, 1.5–3$\times$ above it (Table 2). The tracker realizes 56–76% of the bias removal rather than the $\approx 99\%$ the jackknife formula achieves on the same samples offline, as $\bar{b}_t$ is a single running scalar averaged over past training batches and their prompts. The correction therefore moves the update from a regime in which the squared bias is 15–20$\times$ the noise to one in which the two are within a small factor of each other.
- ***The tracker adds essentially no variance.*** The dashed lines of OMD and TT-OMD coincide: the tracker's noise is the per-batch correction variance divided by $(2-\gamma)/\gamma \approx 19$, which at $K{=}4$ amounts to $\approx 2\%$ of the plug-in variance. This is the point of the two-timescale construction: the 35% larger variance of the jackknife estimate (Figure 34) never reaches the policy update.

Figure 35(b) rescales the variance by the measured effective step size, $\eta_{\text{eff}}\text{Var}/B$, which is how the two terms enter the floor of Theorem 9: the variance term is multiplied by the step, the bias is not. With $\eta_{\text{eff}} \in [0.8, 2.0] \cdot 10^{-4}$ (Adam/LoRA displacement per unit gradient norm, measured between consecutive checkpoints) the variance term is $\approx 5 \cdot 10^{-8}$ and every bias line sits four to five orders of magnitude above it. In this accounting all three methods are bias-dominated throughout, so the remaining lever at $K{=}4$ is the tracker's residual bias, not its noise; a smaller γ would reduce noise that is already negligible while increasing the lag that produces the residual.

The fixed point moves as the theory predicts. Because the bias is, to first order, a rescaling of the risk gap ($\mathbb{E}[\hat{D}] \approx \kappa D$ with $\kappa{=}1.55$ at $K{=}4$, 1.40 at $K{=}8$), uncorrected training behaves like training with the KL coefficient β/κ. Figure 36(a) confirms this on the training logs: the corrected arms settle at implicit-reward margins that are 0.55–0.62$\times$ those of OMD, close to the $1/\kappa{=}0.65$ that a complete correction would give, and their KL to the reference grows at $\approx 0.6\times$ the vanilla rate. The correction therefore recovers the intended regularization strength.

Self-play risk of the trained policies. The quantity the theory bounds is the true entropic risk of the trained policy, which we measure by self-play with 32-sample pools on 100 held-out prompts at every 500 steps (Figure 36(b-c)). At $K{=}8$ the jackknife-corrected policy has lower true risk than the vanilla policy at every checkpoint from step 2000 on; the paired reduction is $+0.061\ [0.043, 0.079]$ at step 4000, $+0.036\ [0.016, 0.055]$ at 4500 and $+0.021\ [-0.001, 0.043]$ at 4680. At $K{=}4$ the paired differences fluctuate around zero within ± 0.02 for both estimators and we cannot resolve an effect with these pool sizes: the $K{=}4$ runs are still improving at 4680 steps (the $K{=}8$ arms reach the same win-rates by step $\approx$2500) and the per-checkpoint noise of the self-play measurement is of the same order as the expected gain.

No performance tax on any stratum or in any tail. Figures 37–38 compare the three methods at step 4250 against the eight opponents on the four evaluation strata. Preference win-rates are identical: the paired differences to OMD are within ± 0.6 % on every stratum with CIs of about ± 0.7. On the combined metric (preferred *and* safer) the jackknife-corrected policy is at or above OMD on every stratum, by $+1.2 \pm 1.0$ (Random), $+0.6 \pm 0.8$ (Conflict), $+1.7 \pm 0.8$ (Sev-3) and $+0.9 \pm 1.3$ (Sev-1) pp, and the same holds for the safety win-rate. Crucially this is also true of



(a) Ratio of reward margin. (b) True self-play risk per arm (c) Paired reduction

Figure 36: **Effects of bias.** **(a)** Ratio of the implicit-reward margin of each corrected arm to that of the vanilla arm at the same step (250-step moving average). Dotted: $1/\kappa$, the ratio expected if the bias were removed completely. **(b-c)** True entropic risk of the trained policies along training (self-play, 32-sample pools, 100 prompts; lower is better) and the paired reduction relative to the vanilla arm at the same step (positive = corrected policy has lower risk; 95% CI over prompts).

the tails: the $\text{CVaR}_{0.25}$ of the combined win-rate (mean over the worst quarter of prompt–opponent pairs, Figure 38b) is higher for the jackknife arm than for OMD on all four strata (52.5 vs. 51.3, 55.2 vs. 54.7, 73.8 vs. 68.9, 65.4 vs. 64.2). Removing the bias therefore costs nothing anywhere we looked, neither on average nor in the tail. The delta-method arm, which removes only a quarter of the bias, is slightly below OMD on the safety metric (-0.6 to -2.1 %), consistent with its weaker correction and its larger residual variance in the update.

Where the gain is: the severe-prompt tail. The gain from the correction concentrates where the risk-sensitive objective should matter most, the Sev-3 stratum (the 217 most severe unsafe prompts). Figure 39 reports three views of the same per-prompt win-rate distribution (pooled over the eight opponents, so each policy has 217×8 prompt–opponent win-rates per metric).

Tail drop. Figure 39(top) is the mean win-rate minus a risk measure of the same distribution: minus the $\text{CVaR}_{0.25}$ (mean of the worst quarter), minus the $\text{CVaR}_{0.125}$ (worst eighth) and minus the entropic risk with c=5. A smaller drop means the policy's average hides fewer badly handled prompts. On preference all three trained policies have negligible drops ($\leq$7 percentage points; they beat every opponent on almost every prompt). On safety and combined win rate, the ordering is jackknife beats OMD beats delta: the safety drop to the worst quarter is 18.2 percentage points for jk, 21.9 for OMD and 25.4 for delta, and to the worst eighth 30.2, 33.9 and 38.8 percentage points. The jackknife policy thus has fewer severe prompts on which it is *not safer* than the opponent.

Mean vs. risk-adjusted win-rate. Figure 39 (middle) plots, for each policy, the risk measure against the mean; a point on the diagonal would have no tail at all, and the vertical distance to the diagonal is the tail drop. For safety and combined the jackknife points lie both furthest right (highest mean) and closest to the diagonal (smallest drop), i.e. the correction improves the tail by more than it improves the mean; the bottom row summarises this as the robustness ratio risk / mean: 0.81 (jk) vs. 0.76 (OMD) vs. 0.72 (delta) for $\text{CVaR}_{0.25}$ on safety, and 0.68 vs. 0.63 vs. 0.57 for $\text{CVaR}_{0.125}$; on preference the three coincide at 0.96.

Per opponent. Figure 39(bottom) shows the combined $\text{CVaR}_{0.25}$/mean ratio broken out by opponent. The jackknife policy has the highest ratio against every one of the eight opponents, so the tail improvement is not driven by a single opponent; it is largest against the low-temperature SFT opponents (T=0.1, 0.5) and Qwen2.5-7B, the opponents that produce the most competitive safe answers.

Out-of-distribution length (200 new tokens). All policies were trained with 64 new tokens.[4] To probe robustness we re-run the best-checkpoint comparison letting every policy and opponent generate 200 new tokens, against the four diverse opponents (SFT at T=1.0, Qwen2.5-1.5B, SmolLM2-1.7B, Qwen2.5-7B) on Random, Conflict and Sev-3 (Figures 40–42).

[4]Chosen primarily because, firstly, prior work (Zhou et al., 2025; Munos et al., 2024) selected this parameter which helps with apples-to-apples comparison, and secondly, we are hardware limited and training at `max_new_tokens` = 200 would take > 10hrs per setting)

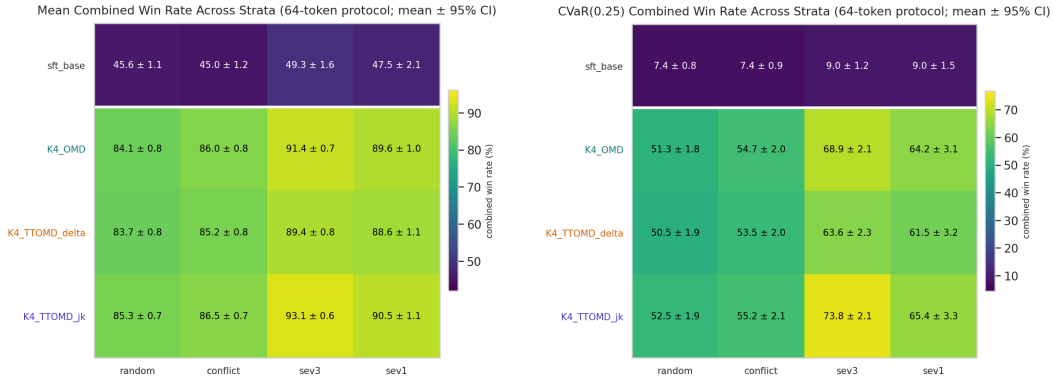


Figure 37: **Crossplay win-rates at step 4250** (64-token protocol; cells are mean ± 95% CI half-width over prompts; the last column is the mean over opponents). Rows 1–4: combined win-rate on Random, Conflict, Sev-3 (must unsafe) and Sev-1 (smallest group of the severity levels, 136 prompts total). Row 5: safety win-rate on Sev-3. The colour scale is fixed to 50–100% in all heatmaps.

Two things change at 200 tokens. First, the trained policies answer in ≈ 45 words while SFT produces ≈ 90 and the external opponents ≈ 70–130, so the preference judge's length preference now works against all trained policies alike and preference win-rates drop by 15–20 pp for every method; we therefore read safety and combined.

Second, the extra generation length exposes the late-training drift of the uncorrected runs. At $K{=}8$ the OMD policy's safety win-rate against SFT falls to 60.5 (Random) and 61.4 (Sev-3), and against Qwen2.5-7B to 77.3 and 71.7; the corrected arms hold 71–73 and 79–85, i.e. paired safety gains of $+10$ to $+12$ percentage points against SFT and $+6$ to $+8.5$ against Qwen2.5-7B (Figure 40), with combined gains of $+7$ to $+9$ percentage points against SFT for the delta arm ($+5$ to $+7$ for jk); against the weaker SmolLM2 the differences are 0–4 percentage points.

On the Sev-3 tail the uncorrected $K{=}8$ policy's safety $\mathrm{CVaR}_{0.25}$ over prompt–opponent pairs is 40.1, indistinguishable from SFT's 40.0, against 56.5 (delta) and 54.6 (jk); its safety tail drop is 38.4 pp vs. 28–30 for the corrected arms (Figure 41). At $K{=}4$, where all three runs are further from convergence and the drift is smaller, the jackknife arm is again the safest against every opponent (Sev-3 safety $+5.6{\pm}2.0$ vs. SFT, $+2.2$ vs. Qwen2.5-1.5B, $+0.8$ vs. SmolLM2, $+3.8{\pm}1.7$ vs. Qwen2.5-7B; Sev-3



(a) Mean combined WR (b) $\mathrm{CVaR}_{0.25}$ combined WR

Figure 38: **Combined (Preferred AND Safe) win-rate.** Taken across the four evaluation strata, mean (a) and $\mathrm{CVaR}_{0.25}$ over prompt–opponent pairs (b), with 95% CIs. The jackknife-corrected policy is at or above OMD in both the mean and the tail on every stratum.

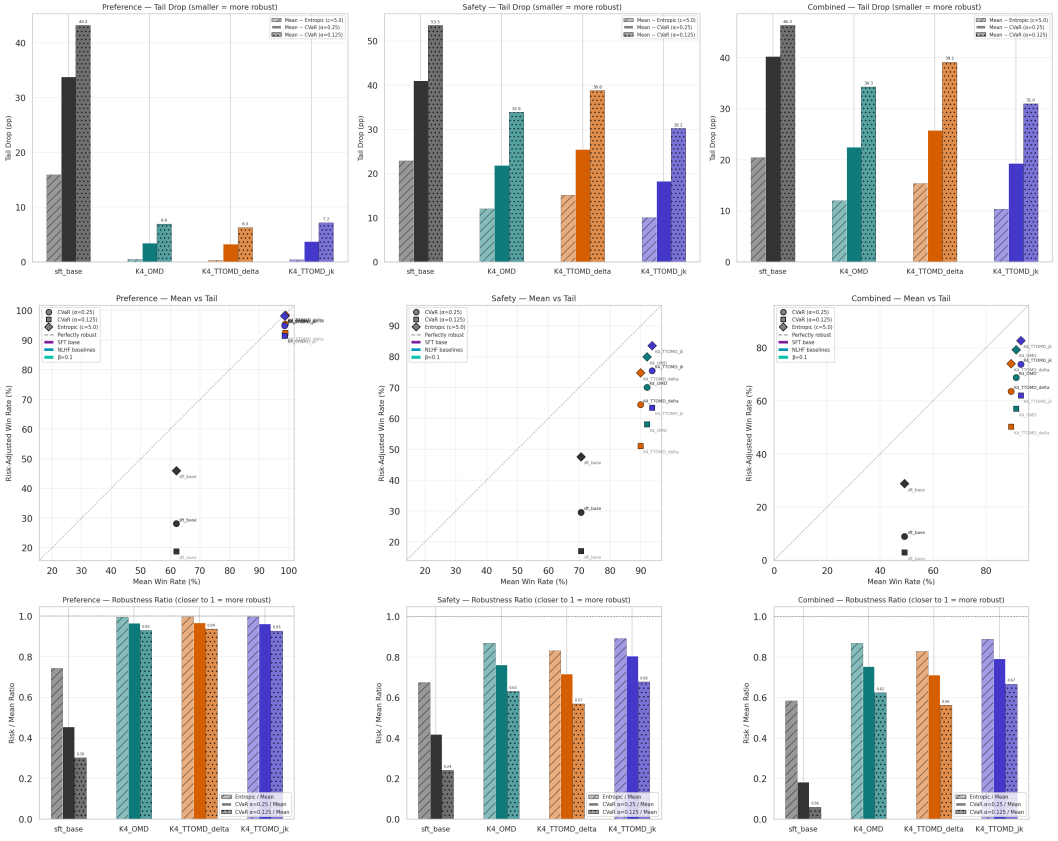


Figure 39: **Tail behavior on the Sev-3 stratum (217 prompts** $\times$ 8 opponents, 64-token protocol, step 4250). ***Top***: tail drop (mean minus entropic risk, minus $\mathrm{CVaR}_{0.25}$, minus $\mathrm{CVaR}_{0.125}$) for preference, safety and combined win-rates; smaller is more robust. ***Bottom***: risk-adjusted vs. mean win-rate (circles $\mathrm{CVaR}_{0.25}$, squares $\mathrm{CVaR}_{0.125}$, diamonds entropic) with the robustness ratio risk/mean below.

safety $\mathrm{CVaR}_{0.25}$ 67.1 vs. 60.4) at a preference cost of 1.5–2.6 percentage points, so the combined metric stays within ± 2 percentage points of OMD against every opponent; the delta arm shows the mirror image (safety -1 to -3 percentage points). Under distribution shift the correction therefore

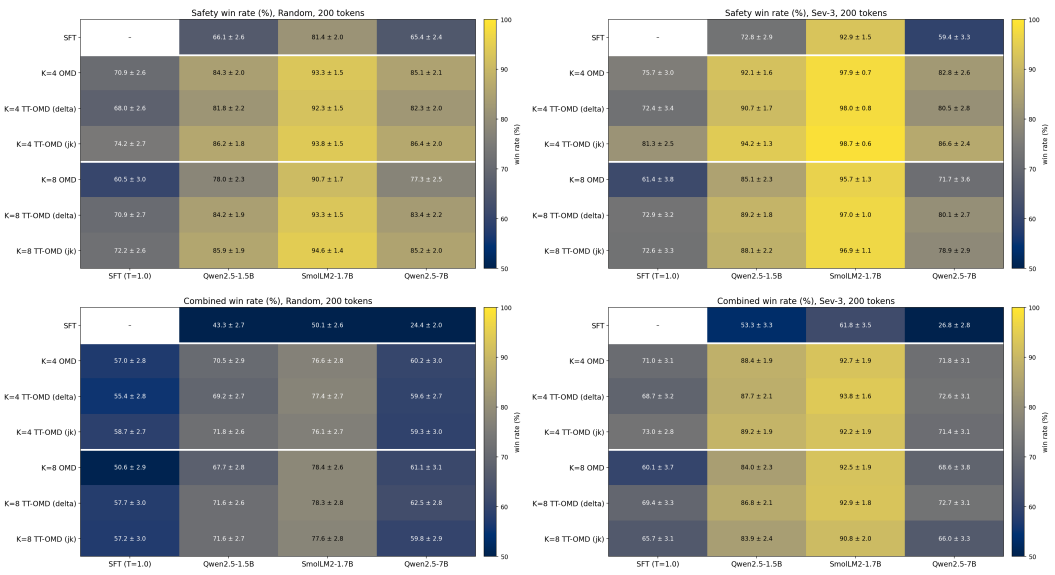
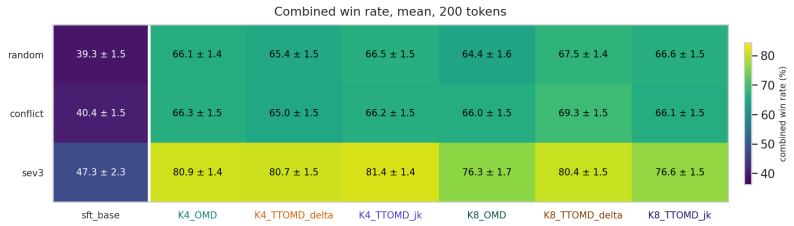
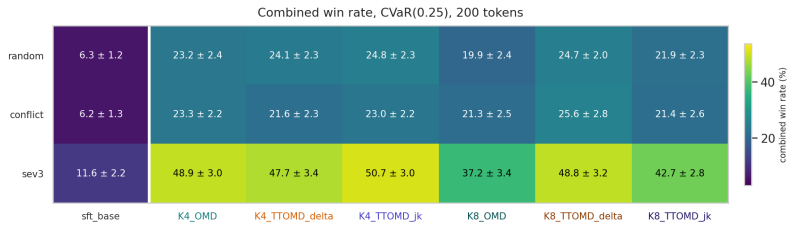


Figure 40: **OOD: 200-token protocol.** (out-of-distribution length; policies trained with 64 tokens). Safety win-rate on Random and Sev-3 (rows 1–2) and combined win-rate on Random and Sev-3 (rows 3–4); same rows as Figure 37, four opponents, mean ± 95% CI half-width over prompts, colour scale 50–100%.

buys safety, and at $K{=}8$ it is the difference between a policy whose safety tail has collapsed back to the SFT level and one that has not.



(a) Mean combined WR



(b) $\mathrm{CVaR}_{0.25}$ combined WR

Figure 41: **OOD: 200-token protocol.** Combined win-rate across Random, Conflict and Sev-3, mean and $\mathrm{CVaR}_{0.25}$ with 95% CIs.

Caveats. One seed per method; 136–400 prompts per stratum; the $K{=}4$ runs have not converged at 4680 steps, meaning there is still significant room for improvement. The tail results are for the severe stratum; on Random and Conflict the tail statistics of the jackknife and OMD arms are within their CIs, so the claim there is ”no tax”, not ”gain”.

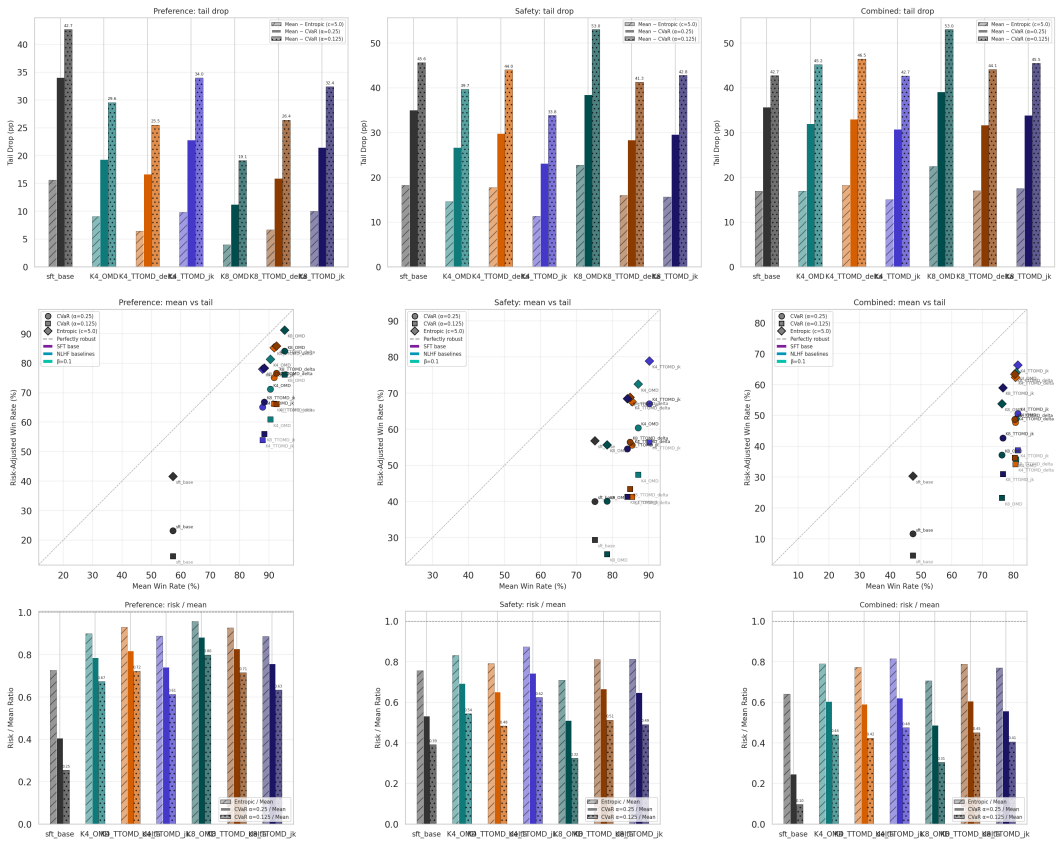


Figure 42: ***Middle and bottom***: Sev-3 tail drop and mean vs. risk-adjusted win-rate (as in Figure 39). The uncorrected $K{=}8$ policy's safety tail ($\text{CVaR}_{0.25}$ 40.1) has returned to the SFT level (40.0); the corrected arms hold 55–57.

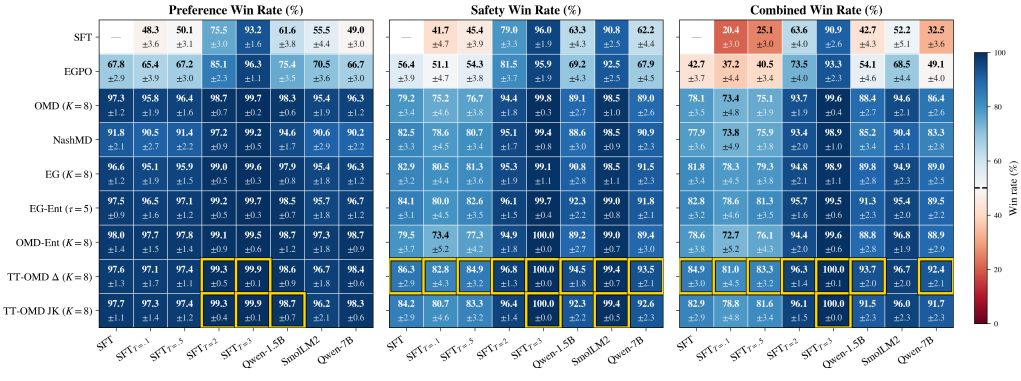


Figure 43: **Sev-1 Cross-play.** Cross-play win-rates including the two-timescale algorithms on the **Sev-1** stratum (136 mildest-severity unsafe prompts), across preference, safety, and combined metrics. The two-timescale variants match or exceed, up to statistical significance, performance across opponents and metrics. In particular, the two-timescale variants bias correction leads to better tail and safety performance.

L.9 Additional Two-Timescale Crossplay Results for $K = 8$

In this section, we highlight the results of two-timescale algorithms in crossplay experiments. Figures 43–45 show the cross-play win-rate on the Sev-1, Sev-3 and Conflict strata. These figures are

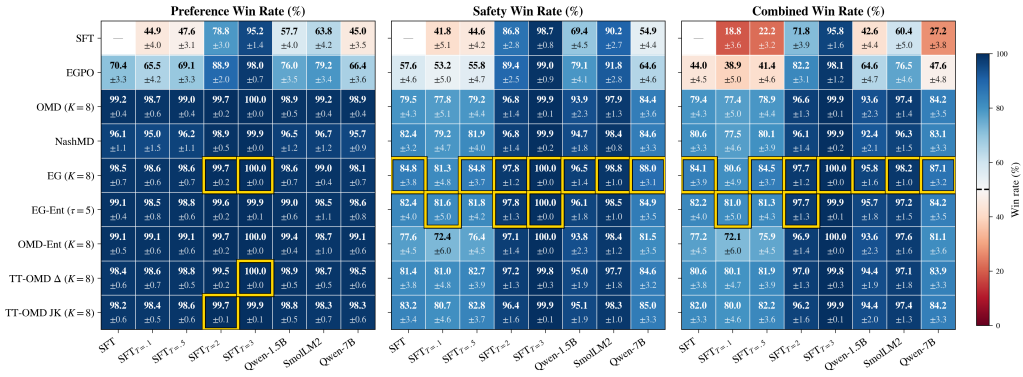


Figure 44: **Sev-3 Cross-play.** Cross-play win-rates including the two-timescale algorithms on the **Sev-3** stratum (100 highest-severity unsafe prompts), across preference, safety, and combined metrics. On the highest-severity group (which is not necessarily on the tail of the data distribution), the extra gradient methods perform the best on average.

complementary to the Random stratum heatmap in Figure 1. We use two-timescale with $K = 8$, to match the K value used in the non-two-timescale variants reported on in the main body.

On the Sev-1 strata, as shown in Figure 43, we can see that two-timescale OMD dominates. The delta method slightly outperforms the jacknife method, and both outperform standard OMD. Notably, the effect of the two-timescale debiasing is more pronounced on the safety strata, where it can lead to gains of almost 10% over the standard algorithm. These comparisons support two-timescale bias correction as a contributor to the observed safety gains. This is consistent with Figure 29, which shows that small-sample risk estimation systematically understates risk on rare-loss conditionals, precisely where risk-sensitive training is intended to account for low-payoff outcomes.

On the Sev-3 strata, as shown in Figure 44, standard EG dominates all of the OMD algorithms. However, similar improvements on the safety winrate of the two-timescale OMD over standard OMD can be seen.

Finally, on the Conflict stratum, as shown in Figure 45, we see similar results as Sev-1, apart from two opponents in the preference winrate we see TT-OMD dominating. Again, this effect is strongest on the safety winrate.

These gains are notable because safety win-rate is measured by a separate cost model used only at evaluation, rather than directly optimized during training. Harmful failure modes can be diverse and underrepresented in preference data, so strong average preference performance need not imply robustness to rare harmful outcomes. The safety improvements are again consistent with the mechanism in Figure 29: small-sample estimation understates risk on rare-loss conditionals, weakening the emphasis on low-payoff outcomes that risk-sensitive training is intended to provide. These safety gains also accompany the substantial reductions in risk-estimation bias shown in Figure 33, supporting the role of two-timescale correction in making risk-sensitive training effective at small sample sizes.

L.10 Risk-Sensitive IPO with Uncertainty over Safety Categories

In this appendix section, we discuss the empirical implementation details for the group weighted RSPG. Let

$$Z(\pi, \pi'; \xi) \;=\; \mathcal{P}^{\xi}(y \succ y' \mid x) \;-\; \tau \log \frac{\pi(y \mid x)}{\mu(y \mid x)} \;+\; \tau \log \frac{\pi'(y' \mid x)}{\mu(y' \mid x)},$$

where $\xi = (x, a, e, \omega, y, y')$ collects prompt, annotator, environment/subgroup, latent noise, and sampled responses, and μ is the reference policy.

The nominal (risk-neutral) IPO population objective is

$$\mathcal{L}(\pi, \pi') \;=\; \mathbb{E}_{\xi}\left[Z(\pi, \pi'; \xi)\right].$$

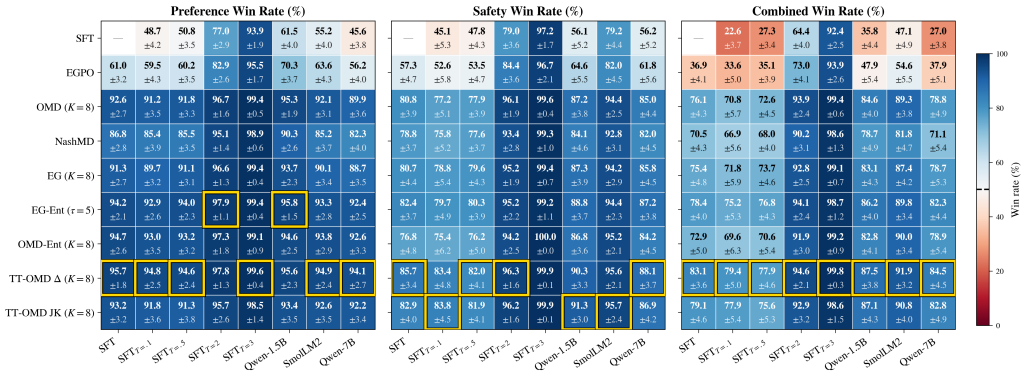


Figure 45: Cross-play win-rates including the two-timescale algorithms on the **Conflict** stratum (100 prompts where preference and safety disagree), across preference, safety, and combined metrics.

Risk-sensitivity replaces the expectation with the entropic robust objective of parameter β:

$$\mathcal{R}_\beta[Z] \;=\; \frac{1}{\beta}\log \mathbb{E}_\xi\left[e^{\beta Z}\right] \;=\; \sup_{q \ll p_\xi}\left\{\mathbb{E}_q[Z] - \frac{1}{\beta}\,\mathrm{KL}(q\,\|\,p_\xi)\right\}. \tag{196}$$

The variational form in (196) makes the robustness interpretation explicit: the learner behaves as if the distribution over ξ were chosen by an adversary within a KL-ball around the nominal p_ξ, with ball radius controlled by $1/\beta$.

Different choices of which component of ξ the adversary is allowed to reweight correspond to different robustness claims. We focus on risk taken over the safety category e.

L.10.1 What "risk over group uncertainty" actually means here

The variational identity

$$\mathcal{R}_\beta[Z] \;=\; \frac{1}{\beta}\log \mathbb{E}_\xi\left[e^{\beta Z}\right] \;=\; \sup_{q \ll p_\xi}\left\{\mathbb{E}_q[Z] - \frac{1}{\beta}\,\mathrm{KL}(q\,\|\,p_\xi)\right\}.$$

says that entropic risk is DRO over a KL ball around the nominal. When you restrict the adversary to reweight only $q(e)$, the KL collapses to $\mathrm{KL}(q(e)\|p(e))$ and the ambiguity set is

$$\mathcal{Q}_\eta = \{q(e) \mid \mathrm{KL}(q(e)\|p(e)) \le \eta\}$$

with η implicit in β (Lagrangian form). So "risk with respect to group uncertainty" means: the training-time marginal over safety categories $p(e)$ is not trusted; deployment might see $q(e) \neq p(e)$ and the learner prepares for the worst such q within a KL ball. Concretely for PKU-SafeRLHF with

$$e \in \{\text{safe}, \text{unsafe-low}, \text{unsafe-med}, \text{unsafe-high}\}$$

training data has some mix, say $p = (0.6, 0.2, 0.15, 0.05)$. At deployment you might get $q = (0.3, 0.3, 0.2, 0.2)$—i.e., more unsafe-high prompts than training. Risk-sensitive training asks: among plausible reweightings, which one makes me look worst, and can I minimize my loss against that?

Importantly, risk sits on the *outer* expectation over e, not inside per-sample preferences. So this is risk over *aleatoric category composition*, not risk over, e.g., annotator noise or preference-model epistemic uncertainty. Those would be different choices of which coordinate of ξ the adversary reweights.

L.10.2 Risk over the Safety Category $\xi = e$

Factor the nominal distribution as

$$p(x, e, y, y') \;=\; p(e)\,p(x \mid e)\,\pi(y \mid x)\,\pi'(y' \mid x),$$

and restrict the adversary in (196) to reweight only the marginal over e:

$$q(x, e, y, y') \;=\; q(e)\, p(x \mid e)\, \pi(y \mid x)\, \pi'(y' \mid x).$$

Under this restriction, the KL divergence between joints collapses to a KL between marginals,

$$\mathrm{KL}(q \,\|\, p) \;=\; \mathrm{KL}\big(q(e) \,\|\, p(e)\big).$$

This encodes the statement: *the category frequencies may be mis-specified at deployment, but within a category the prompt distribution, the policies, and the preference process behave as nominal.*

Define the category-conditional mean advantage

$$Z_e(\pi, \pi') \;:=\; \mathbb{E}\left[Z(\pi, \pi'; \xi) \,\middle|\, e\right] \;=\; \mathop{\mathbb{E}}_{x|e} \mathop{\mathbb{E}}_{\substack{y\sim\pi \\ y'\sim\pi'}} \left[P(y \succ y' \mid x) - \tau \log \frac{\pi(y \mid x)}{\mu(y \mid x)} + \tau \log \frac{\pi'(y' \mid x)}{\mu(y' \mid x)}\right]$$

where here μ is the reference policy. The variational problem over $q(e)$ then admits the closed form (Donsker–Varadhan for the discrete variable e):

$$\mathcal{L}^e_\beta(\pi, \pi') \;=\; \frac{1}{\beta} \log \sum_e p(e)\, \exp\big(\beta\, Z_e(\pi, \pi')\big), \tag{197}$$

with worst-case category distribution

$$q^\star(e) \;\propto\; p(e)\, \exp\big(\beta\, Z_e(\pi, \pi')\big).$$

As $\beta \to 0$, $\mathcal{L}^e_\beta \to \mathbb{E}_e[Z_e]$ (nominal). As $\beta \to \infty$, $\mathcal{L}^e_\beta \to \max_e Z_e$ (worst category).

L.10.3 Sign convention: robust vs. optimistic

The sign of β determines whether the adversary hurts or helps the learner. Suppose Z is player 1's payoff advantage (higher is better for player 1) and player 1 seeks *robustness* across safety categories, i.e., wants to perform well in the worst category. Then player 1 should maximize

$$\widetilde{\mathcal{L}}^e_\beta(\pi, \pi') \;=\; -\frac{1}{\beta} \log \sum_e p(e)\, \exp\big(-\beta\, Z_e(\pi, \pi')\big), \qquad \beta > 0,$$

which is a soft-min over categories. Gradient mass concentrates on the categories where player 1 is doing *worst*. This is the direction to use for DRO-style safety-category robustness; using (197) with $+\beta$ would instead produce optimism over categories.

L.10.4 Sample-Level Implementation

Let $\{(x_i, e_i, y_i, y'_i)\}_{i=1}^B$ be a batch, and let $\mathcal{B}_e = \{i : e_i = e\}$. Compute per-category empirical advantages

$$\hat{Z}_e \;=\; \frac{1}{|\mathcal{B}_e|} \sum_{i \in \mathcal{B}_e} Z(\pi, \pi'; \xi_i),$$

and form the robust objective

$$\hat{\mathcal{L}}^e_\beta \;=\; -\frac{1}{\beta} \log \left(\sum_e \hat{p}(e)\, \exp(-\beta\, \hat{Z}_e)\right), \tag{198}$$

where $\hat{p}(e)$ is the nominal or empirical category frequency. The gradient is a category-reweighted IPO gradient,

$$\nabla \hat{\mathcal{L}}^e_\beta \;=\; \sum_e w_e(\beta)\, \nabla \hat{Z}_e, \qquad w_e(\beta) \;=\; \frac{\hat{p}(e)\, e^{-\beta \hat{Z}_e}}{\sum_{e'} \hat{p}(e')\, e^{-\beta \hat{Z}_{e'}}},$$

i.e., the usual IPO/NLHF gradient with per-category softmax-worst-case weights w_e. This is a one-line modification of the IPO training loop: group sample losses by e, compute w_e, and sum.

Connection to group DRO. Equation (198) is exactly the group-DRO objective of Sagawa et al. (2020) with temperature β, and the variational form (196) is the KL-ball DRO justification Hashimoto et al. (2018).