No Coin Left Behind: Maximizing Strategic Surplus Against No-Regret Dynamics

Yiheng Su
University of Wisconsin–Madison
ysu24@cs.wisc.edu

Emmanouil-Vasileios Vlatakis-Gkaragkounis
University of Wisconsin–Madison
vlatakis@wisc.edu

Abstract

We investigate the *strategic surplus* obtainable against a *Follow-the-Regularized-Leader (FTRL)* learner with constant step size η in $n \times m$ two-player zero-sum games played over T rounds against a clairvoyant optimizer. In contrast with prior analysis, we show that the extraction of such regret-scale surplus is an inherent feature of the FTRL family, rather than an artifact of specific instantiations. First, for a fixed max-min optimizer, we establish a sweeping law of order $\Omega(N_{\mathrm{sub}}/\eta)$, proving that utility surplus scales with the number of the learner's suboptimal actions N and vanishes in their absence. Second, for an alternating optimizer, a surplus of $\Omega(\eta T/\operatorname{poly}(n,m))$ can be guaranteed regardless of the equilibrium structure, with high probability, in random games. Our analysis uncovers a sharp geometric dichotomy: *non-steep* regularizers allow the optimizer to realize the maximal transient surplus via finite-time elimination of suboptimal actions, whereas *steep* regularizers introduce a vanishing tail correction that can delay surplus saturation. Finally, we discuss whether this leverage persists under bilateral payoff uncertainty and propose a susceptibility measure quantifying which regularizers are most vulnerable to learner-aware strategic steering.

1 Introduction

The study of learning algorithms in multi-agent systems has traditionally been viewed through a *defensive* lens. For instance, in the archetypal setting of repeated zero-sum games, algorithms like *Multiplicative Weight Updates (MWU)* [3] are celebrated for their robustness: they guarantee that a learner secures an average payoff at least equal to the minimax value of the game, effectively neutralizing any adversary in the long run. This "no-regret" property has established such dynamics as the bedrock of modern equilibrium computation from high-frequency trading [27] to solving imperfect-information large games like Poker [9, 30].

However, a parallel line of inquiry at the intersection of economics and learning theory suggests that this is only half the story. Pioneering work by Braverman et al. [8], Cai et al. [12] demonstrated that in auction settings, a strategic seller can exploit a no-regret buyer to extract nearly the entire social surplus. More recently, this perspective has evolved into an emerging theory of "steering" learning agents, with applications ranging from contract design [18] to Bayesian persuasion [23]. The unifying insight across these works is that when the symmetry of information is broken—specifically, when a "clairvoyant" optimizer faces a learner with a known update rule—the interaction morphs from a standard game into a *Principal-Agent* control problem [22]. Hence, from this viewpoint, the learner's regret is no longer merely a vanishing error term; it becomes a structured and potentially harvestable reservoir of additional utility [4, 10].

The $\Omega(T)$ and $o(T)$ exploitation regimes. Two distinct scales of exploitation have emerged in the literature. Early results showed that against mean-based algorithms, a strategic optimizer can

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

secure an $\Omega(T)$ gain above the Stackelberg payoff [8, 12, 18]. More recent work, however, identifies settings in which the maximum attainable excess utility shrinks to $o(T)$, together with intriguing asymmetries between the corresponding lower and upper bounds [5, 22]. The latter results reveal that even the sublinear regime is not merely a technical footnote: the order of a cumulative gain–$\Theta(\sqrt{T})$ or $\Theta(T^{2/3})$– represents different levels of persistent finite-horizon cost of learning.

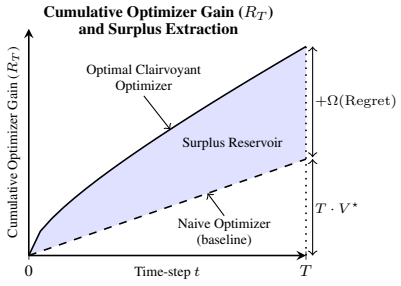
Given their central role, zero-sum games have offered a natural testbed for these questions. Recent work [5] demonstrates that learner-aware optimization can exceed the von Neumann benchmark $T \cdot V^*$ even in this classical setting—yet results remain tied to specific algorithms such as Hedge, or to particular update geometries. Consequently, a unified account of exploitability across the full family of no-regret dynamics is still lacking, which brings us to the following central question:

Is the learner's vulnerability an artifact of the complex economic setting, or is it a structural inevitability encoded in the algorithm itself?

In this work, we argue for the latter. To substantiate this thesis, we anchor our investigation in the canonical setting of repeated zero-sum games between an *FTRL learner* and a *clairvoyant optimizer*. We focus on the *Follow-the-Regularized-Leader* (FTRL) framework—the standard paradigm for no-regret learning—as it unifies algorithms like Multiplicative Weights, 1/2-Tsallis or Online Gradient Descent under a single geometric umbrella. For clarity, we work primarily in the *full-feedback* model. Appendix H provides concentration bounds comparing sampled payoffs with their mixed strategy counterparts along arbitrary predictable mixed-strategy trajectories, while Appendix I reports exploratory bandit-feedback experiments with explicit exploration.

The Surplus Reservoir. To formalize the interaction, consider the optimizer's cumulative reward Rew_T over T rounds against the FTRL learner. Standard no-regret bounds confine this reward within a tight envelope:

$$\boxed{T \cdot V^\star \;\le\; \mathrm{Rew}_T \;\le\; T \cdot V^\star + \mathrm{Regret}^\eta_{\mathcal{L}}(T)} \qquad (\star)$$



where $T \cdot V^\star$ is the cumulative von Neumann value and $\mathrm{Regret}^\eta_{\mathcal{L}}(T)$ the learner's regret under its update rule $\mathcal{L}$ (e.g., FTRL) with step size η. The lower bound is trivially achieved by any max-min strategy. Maximizing the upper bound, however, is not: obtaining both the learner's maximal regret and the game's cumulative value requires the optimizer to actively exploit the learner's specific update rule. Recall that the folklore analysis shows that $\mathrm{Regret}^\eta_{\mathcal{L}}(T) \le A/\eta + B\eta T$. This leads to the crux of our work, which we term *maximal extractability*:

Can a strategic optimizer actively steer *the learner's dynamics so as to consistently saturate this reservoir of extractable surplus such that* $\mathrm{Rew}_T \;\approx\; T \cdot V^\star + \Omega(\max\{1/\eta,\; \eta T\})$.

For two-player zero-sum games, we answer affirmatively by constructing explicit optimizer strategies—equivalently, adversarial loss sequences—that force the learner to incur regret at least $\Omega(\max\{C_1/\eta,\; C_2\,\eta\,T\})$, for game-dependent constants $C_1, C_2 > 0$, matching the classical upper bound up to constant factors. Hence, our results constitute *instance-dependent adversarial regret lower bounds*: tight, game-interpretable certificates that the surplus reservoir is fully saturatable. Our analysis reveals that this saturability is governed by two structural features of the learner's algorithm: the *suboptimality geometry* of the best-response polytope, and the *curvature* of the regularizer—which together determine how much inertia the optimizer can exploit.

1.1 Our Contributions

Formally, we provide a comprehensive characterization for FTRL learners, showing that surplus saturation is inextricably linked to the boundary geometry of the regularizer. As a preliminary step, we develop a toolkit that links discrete-time FTRL updates to their continuous-time counterparts, yielding sharp best-response convergence rates for generic regularizers. (Lemma 3.6 & Corollary E.3)

Our contributions divide naturally into three results of increasing depth:

1. Passive Utility-Extraction: The "Inverse-Rate Law" Fixed Strategy. We begin with the warm-up case of a fixed optimizer strategy—the most natural choice being the max-min strategy. Although static play might seem easily learnable and thus harmless, we show the opposite. Since the learner does not know the game a priori, she must explore all actions before identifying best responses. Every suboptimal action incurs a convergence delay the optimizer passively collects. We establish the Inverse-Rate Law: the surplus scales as $\Theta(|N_{\text{sub}}|/\eta)$, where $|N_{\text{sub}}|$ counts actions that are not best responses to the optimizer's fixed strategy. At step size $\eta = O(T^{-\beta})$ this yields surplus $\Theta(T^{\beta})$ (Fig. 1). The boundary case is instructive: when all learner actions are best responses (e.g., Rock-Paper-Scissors against the max-min strategy, Fig. 2), fixed-strategy surplus vanishes entirely, since the learner merely oscillates among equally optimal actions. (Theorem 4.1)

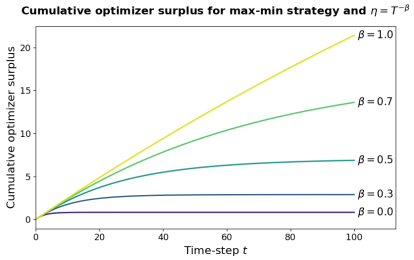


Figure 1: Fixed max-min strategy, Game 1*. Surplus $\Theta(T^{\beta})$, confirming the *Inverse-Rate Law*.

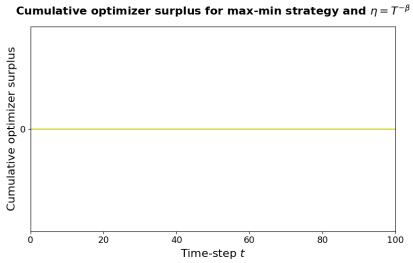


Figure 2: Fixed max-min strategy, Game 2†. All curves coincide at 0: surplus vanishes.

2. Active Utility-Extraction: The "Trap" Dynamic Strategy. However, the vanishing of fixed-strategy surplus in fully mixed games is not a dead end. We show the optimizer can recover nonzero surplus by *actively engineering* the learner's trajectory. The key structural condition is that at least two best responses to the optimizer's max-min support are not ordered by weak domination; this prevents the learner from settling and keeps her oscillating perpetually. While this condition was assumed in prior work [5], its prevalence and quantitative bite were left unexamined. Rather than taking it as given, we turn to random games: we prove that the condition holds with high probability. Our resulting *alternating trap strategy*—reminiscent of trap constructions in auctions [8] and free-fall phases in contracts [18]—guarantees surplus $\Omega(\eta T)$ with high probability (Theorem 4.3). At step size $\eta = O(T^{-\beta})$, this yields surplus $\Theta(T^{1-\beta})$ (Fig. 3).

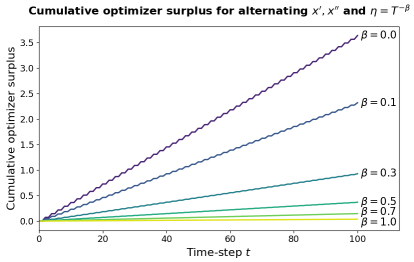


Figure 3: Alternating trap, Game 2†. Surplus $\Omega(T^{1-\beta})$ for all $\beta < 1$.

Results 1 and 2 are complementary relative to a fixed max-min strategy. Result 1 characterizes the fixed-strategy surplus, which vanishes when $N_{\text{sub}} = 0$. When this surplus vanishes, active extraction is guaranteed under the stated non-identification and anti-degeneracy conditions, which hold with high probability in the random-game model. On instances where both quantitative mechanisms are available, balancing the $1/\eta$ and ηT scales yields $\eta = \Theta(T^{-1/2})$, recovering the classical $\Theta(\sqrt{T})$ benchmark.

3. The Price of Learning Best Responses. Due to space constraints, we defer the complementary learner-side view of saturability, based on the cost of identifying a best response in Appendix G.This yields a sharp dichotomy: non-steep regularizers (e.g., Euclidean) reach an ε-best response in $O(1/\varepsilon)$ time, whereas steep regularizers (e.g., entropic) require $\omega(1/\varepsilon)$. The reason is structural: their infinite boundary gradients—the very feature that grants them robustness near the boundary—also prevent rapid convergence to pure strategies. (Theorem 4.4)

2 Preliminaries

We consider a two-player zero-sum game played between an *Optimizer* and a *Learner* over a time horizon T, which may be discrete ($t \in \{0, \dots, T-1\}$) or continuous ($t \in [0, T]$). The optimizer chooses strategies $x(t) \in \Delta^n$ and the learner chooses $y(t) \in \Delta^m$, where Δ^d denotes the probability simplex in $\mathbb{R}^d$. The game is defined by a payoff matrix $A \in \mathbb{R}^{n \times m}$. At time t, the optimizer receives payoff $u_o(x, y) = x(t)^\top A y(t)$ and the learner receives $u_\ell(x, y) = x(t)^\top B y(t)$, where $B = -A$. Additionally, we denote by $\text{Val}(A) := \max_{x \in \Delta^n} \min_{y \in \Delta^m} x^\top A y$ the *minimax value* of the game.

*__Game 1__: $A = \begin{pmatrix} 1 & -1 & 1 & 2 \\ -1 & 1 & 1 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix}$, $x^\star_{\max\min} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $|N_{\text{sub}}| = 2$, fixed-strategy surplus $\Theta(2/\eta)$.

†__Game 2__: $A_{\text{RPS}} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix}$, $x^\star_{\max\min} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, $|N_{\text{sub}}| = 0$, alternating-trap surplus $\Omega(\eta T)$.

Next, we briefly review two essential concepts from convex analysis[1] that underpin our results:

- **Fenchel (convex) conjugate.** Let $\mathrm{h} : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ be a *proper* convex function[2]. The convex conjugate $\mathrm{h}^* : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is defined by $\mathrm{h}^*(y) := \sup_{x \in \mathrm{dom}(\mathrm{h})} \big\{\langle x, y\rangle - \mathrm{h}(x)\big\}, y \in \mathbb{R}^m$.
- **Bregman divergence.** Let $\mathrm{h} : \Delta^m \to \mathbb{R} \cup \{+\infty\}$ be a strictly convex function that is differentiable on the *relative interior* of the simplex[3], denoted by $\mathrm{ri}(\Delta^m)$. The Bregman divergence induced by h is defined as $\mathrm{D}_{\mathrm{h}}(x, x') := \mathrm{h}(x) - \mathrm{h}(x') - \langle \nabla \mathrm{h}(x'), x - x'\rangle, x \in \mathrm{dom}(\mathrm{h}),\ x' \in \mathrm{ri}(\Delta^m)$.

►**The Learner. (Alice)** The learner employs a fixed update *Follow-the-Regularized-Leader (FTRL)* rule with a constant step size $\eta > 0$.[4] She maintains a cumulative payoff vector, or *score*, denoted by $\mathrm{S}(t) \in \mathbb{R}^m$. Initially $\mathrm{S}(0) = 0$. The score accumulates her observed utility gradients:

$$\mathrm{S}(t) = \begin{cases} \sum_{\tau=1}^{t-1} B^\top x(\tau) & \text{(Discrete time)}, \\ \int_0^t B^\top x(\tau)\, d\tau & \text{(Continuous time)}. \end{cases} \qquad \text{(Cumulative Reward)}$$

The learner's strategy $y(t)$ is derived from this score via a regularized projection. Let $\mathrm{h} : \Delta^m \to \mathbb{R}$ be a strictly convex, continuous *regularizer*. We define the *choice map* $Q_{\mathrm{h}} : \mathbb{R}^m \to \Delta^m$ as

$$\text{Learner plays strategy} \quad y(t) = Q_{\mathrm{h}}\big(\eta\, \mathrm{S}(t)\big) := \arg\max_{y \in \Delta^m} \big\{\langle \eta\, \mathrm{S}(t), y\rangle - \mathrm{h}(y)\big\}. \qquad \text{(Choice Map)}$$

Intuitively, FTRL is a "smoothed" best-response: as $\eta \to \infty$ (or $\mathrm{h} \to 0$), the learner approaches the exact best-response dynamics. In this work, we focus on the standard *separable* regularizers of the form $\mathrm{h}(y) = \sum_{i=1}^m \theta(y_i)$, where the kernel $\theta : [0, 1] \to \mathbb{R}$ is strictly convex and differentiable on $(0, 1]$. In this setting, the choice map decouples, as follows:

$$y(t)_i = \phi\big(\lambda(t) + \eta\, \mathrm{S}(t)_i\big), \quad \text{subject to } \sum_i y(t)_i = 1, \tag{1}$$

where $\phi = \mathrm{Trunc}_{[0,1]}\{(\theta')^{-1}\}$ is the inverse of the kernel's derivative and $\lambda(t)$ is a normalization scalar.

Examples of FTRL methods (see Appendix C.4 for details)

⋆**Example 1.** ***Entropic*:** The negative entropy $\mathrm{h}(y) = \sum_i y_i \ln y_i$ leads to the *logit* choice map $y(t)_i \propto \exp(\eta\, \mathrm{S}_i(t))$, yielding the classic Exponential Weights, cf. [6].

⋆**Example 2.** ***Euclidean*:** The quadratic penalty $\mathrm{h}(y) = \frac{1}{2}\|y\|_2^2$ induces the standard *projection* map $y(t) = \arg\min_{y \in \Delta^m} \|y - \eta\, \mathrm{S}(t)\|_2$, corresponding to Online Gradient Descent, cf. [29].

⋆**Example 3.** ***Tsallis*:** For $q \in (0, 1)$, the negative Tsallis regularizer is $\mathrm{h}(y) = \frac{1}{q-1}\sum_i (y_i^q - y_i)$, cf. [1]. The associated choice map is characterized by $q\, y(t)_i^{q-1} = 1 + (q-1)\eta\big(\mathrm{S}_i(t) - \tau\big)$, where $\tau \in \mathbb{R}$ is the normalization scalar chosen so that $\sum_i y(t)_i = 1$. As $q \to 1$, this regularizer converges to the negative entropy.

The geometry of θ near the boundary induces a fundamental dichotomy in the learner's behavior:

Definition 1 (Steep vs. Non-Steep Regularizers)**.** *A regularizer is **steep** if $|\theta'(y)| \to \infty$ as $y \to 0^+$ (e.g., Negative Entropy/Negative Tsallis). It is **non-steep** if $\theta'(0^+)$ is finite (e.g., Euclidean).*

Remark 1. *This distinction is critical: steep regularizers force $y(t)$ to strictly stay in the interior of the simplex (since $\phi(-\infty) = 0$), whereas non-steep regularizers allow the learner to assign* exact zero probability *to actions in finite time, enabling the "finite-time elimination" of suboptimal strategies.*

Remark 2. *The map ϕ, introduced in [15], is an analytic workhorse that lifts our analysis beyond the Hedge-specific calculations of [5]. For any regularizer θ, ϕ encodes its local curvature in a single object: the weights $1/\theta''(y_i(r))$ emerge naturally, driving the sharp coordinate-wise bounds of Lemma D.5 in fixed strategy and the curvature terms in the alternating-strategy analysis in Lemma F.8. This is precisely what enables matching upper and lower bounds across the full FTRL family.*

[1] For an extensive convex-analysis recap, closed-form derivations, and the worked examples, see Appendix C.

[2] By *proper*, we mean that the effective domain $\mathrm{dom}(\mathrm{h}) := \{x \in \mathbb{R}^m \mid \mathrm{h}(x) < +\infty\}$ is nonempty and $\mathrm{h}(x) > -\infty$ everywhere. In our context, functions are usually defined on the simplex Δ^m and are set to $+\infty$ outside it.

[3] Recall that the relative interior $\mathrm{ri}(C)$ is the interior of a set C relative to its affine hull. For the probability simplex, this corresponds to the set of strictly positive probability distributions, $\mathrm{ri}(\Delta^m) = \{x \in \Delta^m \mid x_i > 0, \forall i\}$.

[4] Although we use a constant step size, we do not treat it as independent of the horizon. Instead, we work in an asymptotic regime where $\eta = \eta(T) \to 0$ as $T \to \infty$ while $\eta T \to \infty$; for instance, $\eta = \Theta(1/\sqrt{T})$ satisfies these conditions.

► **The Optimizer. (Bob)** The optimizer is a strategic agent with *strategic foresight*. Unlike the learner, who reacts to past payoffs, the optimizer knows the game matrix A and the learner's update rule. The optimizer's goal is to select a trajectory $\{x(t)\}_t$ to maximize his total payoff $\int x(t)^\top A y(t)\, dt$. We summarize the interaction protocol below:

Protocol: Repeated Game with FTRL Learner

For each round $t = 1, \ldots, T$:

1. **Learner updates:** The learner computes $y(t) = Q_{\mathrm{h}}(\eta\, \mathrm{S}(t))$.
2. **Optimizer plays:** The optimizer selects $x(t) \in \Delta^n$ (potentially using knowledge of $y(t)$).
3. **Payoffs & Feedback:** Players receive payoffs $u_{o/\ell}(x(t), y(t))$. The learner observes the feedback vector $B^\top x(t)$.
4. **State update:** The learner updates the score $\mathrm{S}(t) = \mathrm{S}(t-1) + B^\top x(t)$.

Feedback Models. Our primary analysis assumes *full-feedback*, where the learner observes the entire vector $B^\top x(t)$. Appendix H provides concentration bounds comparing sampled payoffs with their mixed-strategy counterparts along arbitrary predictable mixed-strategy trajectories. The bandit experiments in Appendix I are exploratory, and the explanations discussed in Appendix J are hypotheses. Bandit-feedback analogues of our extraction guarantees remain open.

3 Our Toolbox

3.1 A Discrete-Continuous Bridge

To understand the long-run behavior in our setting, we establish a fundamental link between the optimizer's discrete and continuous objectives. We adapt the potential-based discrete–continuous analysis of Kwon and Mertikopoulos [19] to the optimizer's reward. We show that the discrete reward tracks a continuous-time benchmark, deviating only by a *discretization gap* governed by the Bregman divergence of the learner's dual updates. We begin by defining the optimizer's total reward in both the continuous-time and discrete-time settings, each of which depends on the learner's strategy as determined by FTRL. For any optimizer's strategy $x(t)\colon [0,T] \to \Delta^n$:

$$\mathrm{R}_{\mathrm{cont}}(x(t),\, \mathrm{S}(0),\, T) = \int_0^T \left\langle x(t), A \cdot Q_{\mathrm{h}}\left(\eta \left[\mathrm{S}(0) + \int_0^t B^\top x(\tau)\; d\tau\right]\right)\right\rangle dt \tag{2}$$

$$\mathrm{R}_{\mathrm{disc}}(x(t),\, \mathrm{S}(0),\, T) = \sum_{t=0}^{T-1} \left\langle x(t), A \cdot Q_{\mathrm{h}}\left(\eta \left[\mathrm{S}(0) + \sum_{s=0}^{t-1} B^\top x(s)\right]\right)\right\rangle \tag{3}$$

We define the optimizer's optimal reward in both continuous and discrete time settings as follows:

$$\mathrm{R}_{\mathrm{cont}}{}^*(\mathrm{S}(0),\, T) \coloneqq \max_{x(t)} \mathrm{R}_{\mathrm{cont}}(x(t),\, \mathrm{S}(0),\, T) \text{ and } \mathrm{R}_{\mathrm{disc}}{}^*(\mathrm{S}(0),\, T) \coloneqq \max_{x(t)} \mathrm{R}_{\mathrm{disc}}(x(t),\, \mathrm{S}(0),\, T).$$

In continuous time, $x(t)$ is defined for $t \in [0,T]$. In discrete time, $x(t)$ denotes the sequence $\{x(0), x(1), \ldots, x(T-1)\}$. To analyze these reward functions, we introduce the learner's *potential function*: $\Phi_\eta(Z) \coloneqq \mathrm{h}^*(\eta Z)/\eta$ for any $Z \in \mathbb{R}^m$. We defer the proofs of this part in Appendix D.1.

►**The Continuous Benchmark**

Our first observation reveals that in continuous time, the optimizer's reward is fully determined by the initial and final states of the learner, rendering the intermediate trajectory irrelevant:

Lemma 3.1. *For any continuous optimizer strategy $x_c : [0,T] \to \Delta^n$, the total reward is given exactly by the drop in the learner's potential:* $\mathrm{R}_{\mathrm{cont}}(x_c(t),\, \mathrm{S}(0),\, T) = \Phi_\eta\big(\mathrm{S}(0)\big) - \Phi_\eta\big(\mathrm{S}(T)\big)$.[5]

Remark 3 (Path Independence). *Crucially, Lemma 3.1 implies that the continuous reward depends on the strategy $x_c(t)$ only through its time-average $\bar{x} \coloneqq \frac{1}{T}\int_0^T x_c(t)\, dt$. Since $\mathrm{S}(T) = \mathrm{S}(0) + TB^\top \bar{x}$, any two strategies with the same mean yield identical rewards. This collapses the infinite-dimensional variational problem of finding the optimal trajectory $x_c(t)$ into a finite-dimensional*

[5]The proof follows since $\nabla\Phi_\eta(Z) = Q_{\mathrm{h}}(\eta Z)$, so by the chain rule $\frac{d}{dt}\Phi_\eta(\mathrm{S}(t)) = \big\langle \nabla\Phi_\eta(\mathrm{S}(t)), \dot{\mathrm{S}}(t)\big\rangle = \big\langle y(t),\, B^\top x_c(t)\big\rangle$. Integrating on $[0,T]$ gives the result. We defer the full proof for the continuous-time case to Theorem D.1 in Appendix D.1. The range for the total reward is characterized in Proposition D.3, which recovers the reward sandwich ($\star$) in the continuous-time setting.

convex minimization over constant strategies $x \in \Delta^n$:

$$\mathrm{R_{cont}}^*(\mathrm{S}(0),\, T) \;=\; \Phi_\eta\big(\mathrm{S}(0)\big) - \min_{x\in\Delta^n} \Phi_\eta\big(\mathrm{S}(0) + TB^\top x\big). \tag{4}$$

Efficient Computation & Correction to Prior Work. The reduction of Remark 3 allows for efficient optimization via Frank-Wolfe. Setting $G(x) \coloneqq \Phi_\eta(\mathrm{S}(0) + TB^\top x)$, one verifies that G is convex and β-smooth with $\beta = (\eta T\|A\|_{\mathrm{op}})^2/\alpha$, yielding the following complexity.[6,7]

Corollary 3.2 (Frank-Wolfe Complexity [11, Theorem 3.8])**.** *An ϵ-approximate strategy $\hat{x}$ with $\mathrm{R_{cont}}^*(\mathrm{S}(0),\, T) - \mathrm{R_{cont}}(\hat{x},\, \mathrm{S}(0),\, T) \le \epsilon$ can be computed in time $\mathcal{O}\big(nm\, R^2\eta T^2\|A\|_{\mathrm{op}}^2/(\alpha\epsilon)\big)$, where R is the diameter of Δ^n.*

►The Discretization Gap

In discrete time, the optimizer cannot perfectly integrate the learner's potential.[8] The error induced by the step size η, as we will see, manifests as a strictly non-negative "bonus" for the optimizer, captured by the Bregman divergence.

Theorem 3.3 (Discrete-Continuous Decomposition)**.** *Let $x_d(t)$ be a discrete strategy and $\tilde{x}(t) = x_d(\lfloor t \rfloor)$ its piecewise-constant extension. The discrete reward satisfies:*

$$\mathrm{R_{disc}}(x_d(t),\, \mathrm{S}(0),\, T) \;=\; \underbrace{\mathrm{R_{cont}}(\tilde{x}(t),\, \mathrm{S}(0),\, T)}_{\textit{Continuous Benchmark}} \;+\; \underbrace{\frac{1}{\eta}\sum_{t=0}^{T-1} \mathrm{D}_{\mathrm{h}^*}\big(\eta\,\mathrm{S}(t+1),\, \eta\,\mathrm{S}(t)\big)}_{\textit{Discretization Gap} \,:=\mathrm{DG}(x_d,T)\ge 0}. \tag{5}$$

Remark 4. *The non-negativity of $\mathrm{D}_{\mathrm{h}^*}$ in Theorem 3.3 implies that the discrete optimizer always outperforms the continuous benchmark: $\mathrm{R_{disc}}^*(0,\, T) \ge \mathrm{R_{cont}}^*(0,\, T)$. Conversely, maximizing the decomposition (5) yields an upper bound controlled by the maximum cumulative divergence:*

$$\mathrm{R_{disc}}^*(0,\, T) \;\le\; \mathrm{R_{cont}}^*(0,\, T) \;+\; \sup_{x_d(t)} \mathrm{DG}(x_d(t), T) \;\textit{ and }\; \mathrm{DG}(x_d(t), T) \;\le\; \frac{\eta}{2\alpha}\sum_{t=0}^{T-1}\Big\|A^\top x_d(t)\Big\|_*^2, \tag{6}$$

where h is α-strongly convex with respect to $\|\cdot\|$.

3.2 Surplus Toolbox: Benchmarks, Decomposition, and Decay

To set the stage for our formal notion of surplus, we introduce a few definitions and remarks.

$*$ ***Learner's Viewpoint*****. (See Appendix D.3.1 for details and proof)** By path independence (Remark 3) in continuous time, it suffices to consider a fixed optimizer strategy $\hat{x} \in \Delta^n$. Let $v \coloneqq A^\top \hat{x}$ be the vector of expected payoffs. We denote the *best-response value* as $u_o^* \coloneqq \min_i v_i$, the *suboptimality gaps* as $\delta_i \coloneqq v_i - u_o^* \ge 0$, and the *best-response set* as $\mathrm{br}(\hat{x}) \coloneqq \{i \mid \delta_i = 0\}$ and $k \coloneqq |\mathrm{br}(\hat{x})|$. The following lemma characterizes the learner's trajectory.

Lemma 3.4 (Learner's limit)**.** *$y(t)$ converges to $y^* \in \mathrm{br}(\hat{x})$ as $t \to \infty$. Moreover, strict convexity of h implies that y^* is unique. When $\hat{x} = x^\star$ is max-min, we have that $y^* \in \arg\min_{y\in\mathrm{conv}(\mathrm{br}(x^\star))} \mathrm{h}(y)$, so the payoff converges to the value $\mathrm{Val}(A)$ although y^* need not be a min-max strategy.*

$*$ ***Optimizer's Viewpoint*****. (See Appendix D.3.2 for details and proof)** Returning to the optimizer's payoff, it is driven by the learner's transient mass on suboptimal actions.

Lemma 3.5 (Optimizer's payoff)**.** *$u_o(\hat{x}, y(t)) = u_o^* + \sum_{i\notin\mathrm{br}(\hat{x})} \delta_i\, y_i(t)$, for all $t \ge 0$.*

[6] From Proposition C.1, if h is α-strongly convex then h^* is $(1/\alpha)$-smooth. The extra $(\eta T\|A\|_{\mathrm{op}})^2$ factor arises from the chain rule through the affine argument. See Appendix D.1.1.

[7] *Correction to [5]: Assos et al. [5, Proposition 5] analyze the MWU setting assuming smoothness constant $\beta = 1$. Our analysis shows $\beta = \Theta((\eta T\|A\|_{\mathrm{op}})^2)$, which is not $O(1)$ for large horizons. Correcting for this, the Frank-Wolfe complexity is $\mathcal{O}(nm\eta T^2\|A\|_{\mathrm{op}}^2/(\alpha\varepsilon))$ rather than $\mathcal{O}(nm/\varepsilon)$. Recall that even if $\eta = T^{-\beta}$ for some $\beta \in [0,1]$, the runtime will be heavily horizon-dependent.*

[8] *The challenges of optimal control differ substantially in continuous and discrete time. The continuous optimum $\mathrm{R_{cont}}^*(x_c(t),\, T)$ is obtained from a single static convex program in Equation (4). In contrast, the discrete optimum $\mathrm{R_{disc}}^*(x_d(t),\, T)$ maximizes a cumulative Bregman sum, and thus becomes a sequential decision problem: the optimizer must balance minimizing the terminal potential with maximizing the path-dependent divergence term, i.e., volatility harvesting. Since the continuous optimum provides a lower bound, and since we later show that asymptotically simple fixed or alternating strategies already achieve the optimal upper bounds, we do not study the exact computation of global or local optima here. For this perspective, see [2, 10].*

$*$ ***The Cumulative Surplus.*** Then, we define the *cumulative surplus* as the optimizer's excess payoff above the game value.

Definition 2. *For a horizon $T > 0$,* $\mathrm{SURPLUS}(T) := \int_0^T \big(u_o(x(t), y(t)) - \mathrm{Val}(A)\big)\, dt$. *When the optimizer plays a fixed strategy $\hat{x}$, Lemma 3.5 yields the decomposition*

$$\mathrm{SURPLUS}(T) = \underbrace{T\,(u_o^* - \mathrm{Val}(A))}_{\mathrm{VAG}(T)} + \underbrace{\int_0^T \sum_{i \notin \mathrm{br}(\hat{x})} \delta_i\, y_i(t)\, dt}_{\mathrm{LAG}(T)} \qquad \text{(Surplus Decomposition)}$$

The ***Value Gap*** $\mathrm{VAG}(T)$ *is a linear penalty (since $u_o^* \le \mathrm{Val}(A)$) that vanishes iff $\hat{x}$ is max-min. The* ***Learner Approximation Gap*** $\mathrm{LAG}(T) \ge 0$ *captures the profit extracted from the learner's suboptimal actions before convergence.*

Remark 5. *Surplus Decomposition separates a static* cost of commitment, $\mathrm{VAG}(T)$, *from dynamic* gains from learning, $\mathrm{LAG}(T)$. *The former vanishes iff $\hat{x}$ is max-min. The latter measures the transient profit extracted from the learner's suboptimal actions. Hence, under perfect knowledge, controlling surplus amounts to bounding* $\mathrm{LAG}(T)$, *namely, how quickly the regularizer drives $y_i(t)$ to zero.*

The following structural lemma offers a unified analysis for generic separable regularizers, acting as our "speedometer," linking the geometry of the regularizer (through the inverse map ϕ) to the decay rate of non-best-response probabilities.[9] The proof of it is in Appendix D.3.3.

Lemma 3.6 (Pointwise FTRL Bounds)**.** *Fix $\hat{x}$, step size $\eta > 0$, and a separable regularizer* h. *The FTRL response satisfies $y_i(t) = \phi(\lambda(t) - \eta t\, v_i)$ for a normalization scalar $\lambda(t)$. Then for any $t \ge 0$:*

$\star$***Optimal Actions:*** *For all $i \in \mathrm{br}(\hat{x})$, $y_i(t)$ are identical a_t with $\frac{1}{m} \le a_t \le \frac{1}{k}$.*

$\star$***Suboptimal Actions:*** *For all $j \notin \mathrm{br}(\hat{x})$: $\phi\big(\theta'(1/m) - \eta t\, \delta_j\big) \le y_j(t) \le \phi\big(\theta'(1/k) - \eta t\, \delta_j\big)$.*

Additionally, steep regularizers yield asymptotic decay ($y_j(t) \to 0$), whereas non-steep ones eliminate suboptimal actions in finite time $t \ge t_\theta^* := \frac{\theta'(1/k) - \theta'(0^+)}{\eta \delta_{\min}}$.

4 Main Results

4.1 Passive Extraction: The Inverse-Rate Law

For our first result, fix a constant optimizer strategy $\hat{x}$. Let $\mathrm{br}(\hat{x})$ be the set of best responses, $k = |\mathrm{br}(\hat{x})|$, $m - k = N_{\mathrm{sub}}$ the number of suboptimal actions, and $\delta_{\min} \le \delta_i \le \delta_{\max}$ the payoff gaps of suboptimal actions. Define the *dual potential* $V(u) := \theta^*(\theta'(u)) = u\,\theta'(u) - \theta(u)$[10], and the positive potential drops $\overline{\Delta V}(T), \underline{\Delta V}(T)$:

$$\overline{\Delta V}(T) := \tfrac{\delta_{\max}}{\delta_{\min}} \Big[V(1/k) - \theta^*\big(\max\{\theta'(1/k) - \eta \delta_{\min} T,\, \theta'(0^+)\}\big) \Big], \tag{7}$$

$$\underline{\Delta V}(T) := \tfrac{\delta_{\min}}{\delta_{\max}} \Big[V(1/m) - \theta^*\big(\max\{\theta'(1/m) - \eta \delta_{\max} T,\, \theta'(0^+)\}\big) \Big]. \tag{8}$$

Theorem 4.1 (Inverse-Rate Law, Continuous Time)**.** *The cumulative surplus* $\mathrm{LAG}(T)$ *against a fixed strategy satisfies*

$$\frac{N_{\mathrm{sub}}}{\eta} \cdot \underline{\Delta V}(T) \;\le\; \mathrm{LAG}(T) \;\le\; \frac{N_{\mathrm{sub}}}{\eta} \cdot \overline{\Delta V}(T). \qquad \text{(Exploitation Bound)}$$

The discrete-time version holds up to an upper-bound additive $O(N_{\mathrm{sub}} \delta_{\max}/k)$ term (Corollary E.3). The asymptotic regime is dictated by $\theta'(0^+)$, yielding a sharp dichotomy.

[9] Specializing Lemma 3.6 to standard regularizers, in the supplement we show that for suboptimal weights $i \notin \mathrm{br}(\hat{x})$: **(Negative Entropy)** yields *exponential decay* $y_i(t) \sim e^{-\eta t \delta_i}$; **(Euclidean)** quadratic penalty yields *finite-time decay* $y_i(t) \sim [\mathrm{const} - \eta t \delta_i]_+$; **(Negative Tsallis)** for $q \in (0,1)$ yields *polynomial decay* $y_i(t) \sim (\mathrm{const} + \eta t \delta_i)^{-1/(1-q)}$, while $q > 1$ allows sparse solutions similar to the Euclidean case.

[10] For standard kernels: Euclidean $V(u) = \frac{1}{2}u^2$, Tsallis-q $V(u) = u^q$, entropic $V(u) = u$ (up to constants).

Theorem 4.2 (Steep vs. Non-Steep, Continuous Time). *Let* $V_{\mathrm{bdry}} := \theta^*(\theta'(0^+))$. *As* $\eta T \to \infty$*:*

(A) Non-steep *(*$\theta'(0^+) > -\infty$*,* $V_{\mathrm{bdry}} < \infty$*): suboptimal actions are eliminated in finite time* $T^* < \infty$ *and surplus* saturates *to a constant in* $\frac{N_{\mathrm{sub}}}{\eta}\left[\frac{\delta_{\min}}{\delta_{\max}}(V(1/m) - V_{\mathrm{bdry}}),\ \frac{\delta_{\max}}{\delta_{\min}}(V(1/k) - V_{\mathrm{bdry}})\right]$.

(B) Steep *(*$\theta'(0^+) = -\infty$*,* $V_{\mathrm{bdry}} = 0$*): the learner approaches the boundary only asymptotically and surplus has an* infinite tail*, converging to* $\frac{N_{\mathrm{sub}}}{\eta}\left[\frac{\delta_{\min}}{\delta_{\max}}V(1/m),\ \frac{\delta_{\max}}{\delta_{\min}}V(1/k)\right] \pm o(1)$.

We provide an expanded discussion in Appendix E. The formal statements and proofs of Theorems 4.1 and 4.2 are presented as Theorems E.1 and E.2, respectively, with concrete examples available in Appendix E.3. Therefore, if the optimizer simply plays max-min strategy $\hat{x} = x^\star$, he will receive as surplus exactly $\mathrm{LAG}(T)$.

4.2 Active Extraction: The Alternating Trap

As discussed in the introduction, a fixed strategy fails whenever $N_{\mathrm{sub}} = 0$ (e.g., RPS against the max-min). Here, we show that the optimizer can recover again optimal surplus by *actively engineering* the learner's trajectory.

Theorem 4.3 (Discrete Time). *Let* $A \in \mathbb{R}^{n\times m}$ *have i.i.d.* $\mathrm{Unif}[-1,1]$ *entries. Fix any* $\delta \in (0, \frac{1}{m})$ *and* α*-strongly convex separable regularizer* h *with curvature bound* $M = \sup_{a\in[\delta,1-\delta]}\theta''(a)$*, and step size* $\eta \le \frac{\alpha}{c_{\|\cdot\|}\sup_{u\in[-1,1]^m}\|u\|_*}(\frac{1}{m} - \delta)$ *where* $c_{\|\cdot\|} = \sup_{u\neq 0}(\|u\|_\infty/\|u\|)$*. Then with probability at least* $1 - \frac{n!m!}{(n+m-1)!} - \frac{1}{nm}$*, there exists an* alternating strategy $x_d(t)$ *such that for every horizon* T

$$\mathrm{R}_{\mathrm{disc}}(x_d(t),\, 0,\, T) \;\ge\; T\,\mathrm{Val}(A) \;+\; \eta T \cdot C_A^{\mathrm{h}},\ \textit{where}\ C_A^{\mathrm{h}} \;:=\; \frac{1}{2M(nm)^6}$$

Remark 6. *At a high level, the following presentation has the same general blueprint as in Assos et al. [5, Proposition 11], but several technical ingredients must be refined substantially in order to extend the analysis to arbitrary regularizers. The first issue is to obtain a quantitative lower bound ensuring that two distinct learner best responses to the max-min strategy are not ordered by weak dominance. Rather than taking such a condition as an assumption, we show that, with high probability, a uniform polynomial lower bound holds for random zero-sum games. This is a critical ingredient: although a related property is implicitly used in [5], the relevant game-dependent magnitude is never made explicit there. A second key ingredient is a uniform control of the curvature of the optimization landscape under an arbitrary regularizer. This is precisely where our proof departs from prior entropy-specific arguments. In this way, the negative-entropy case emerges as a special instance of a broader proof template that applies uniformly to all regularizers under consideration.*

Our Proof Sketch. The formal statement and proof are provided in Appendix F. Below, we provide a proof sketch. The *key idea* is to construct two perturbed strategies x', x'' with $(x' + x'')/2 = x^\star$ that flip the strict preference between two best responses $i \neq j$, locking the learner in perpetual oscillation:

Step 1: Two high-probability events on A. With high probability over the random payoff matrix A, there exist a max-min optimizer strategy $x^\star$ such that two distinct learner best responses $i \neq j \in \mathrm{br}(x^\star)$, and an index $\ell \in \mathrm{supp}(x^\star)$ such that $e_\ell^\top A e_i = A_{\ell i} \neq A_{\ell j} = e_\ell^\top A e_j$. We record this support separation property as the event $\mathcal{A}_{nm}$ (Assumption 1) and show that it holds with high probability in Lemma F.2. We also work on an anti-degeneracy event $\mathcal{E}_{\mathrm{gap}}$, which guarantees a uniform polynomial lower bound on $|A_{pq} - A_{pq'}|$ whenever $A_{pq} \neq A_{pq'}$. Conditioning on $\mathcal{A}_{nm} \cap \mathcal{E}_{\mathrm{gap}}$, we fix such $x^\star$, $i \neq j$, and ℓ for the remainder of the proof.

Step 2: An alternating perturbation construction. Using the fixed $(x^\star, i, j, \ell)$, we construct two perturbed strategies x', x'' satisfying $(x' + x'')/2 = x^\star$, such that the strict preference between i and j flips $(A^\top x')_i > (A^\top x')_j, (A^\top x'')_j > (A^\top x'')_i$. The perturbation is defined in (40) and its properties are proved in Lemma F.6. We then define an alternating optimizer strategy $x_d(t)$ that plays x' on even rounds and x'' on odd rounds. This keeps the even-round time-average at $x^\star$ while repeatedly switching the unique best response between i and j. Let $v := A^\top x^\star$ and $v' := A^\top x'$. Using $(x' + x'')/2 = x^\star$ and the max-min property of $x^\star$, we lower bound the optimizer's payoff over each pair of rounds $(2s, 2s+1)$ by

$$2\,\mathrm{Val}(A) \;+\; \Big(\langle v', y_{2s}\rangle - \langle v', y_{2s+1}\rangle\Big). \tag{9}$$

Thus, it suffices to prove a uniform lower bound on the one-step difference $\langle v', y_{2s}\rangle - \langle v', y_{2s+1}\rangle$.

Step 3. Interpolation and a variance identity. To control the difference term, we interpolate between the two consecutive FTRL iterates $y(r) = Q_{\mathrm{h}}\big(-2\eta s\, v - r\,\eta\, v'\big), r \in [0,1]$, so that $\langle v', y_{2s}\rangle - \langle v', y_{2s+1}\rangle = -\int_0^1 \frac{d}{dr}\langle v', y(r)\rangle\, dr$. Differentiating the KKT conditions along this path yields the variance identity

$$\frac{d}{dr}\langle v', y(r)\rangle \;=\; -\eta\, W_r\ \mathrm{Var}_{\pi_r}(Z_r),$$

where $w_r(\ell) \propto 1/\theta''(y_\ell(r))$, $W_r := \sum_\ell w_r(\ell)$, $\pi_r(\ell) := w_r(\ell)/W_r$, and $Z_r := v'_I$ for $I \sim \pi_r$ (curvature-weighted sampling). A two-point lower bound on variance then isolates the distinguished coordinates $i \neq j$:

$$W_r\ \mathrm{Var}_{\pi_r}(Z_r) \;\geq\; \frac{(v'_i - v'_j)^2}{\theta''(y_i(r)) + \theta''(y_j(r))}. \tag{10}$$

We formalize this step in Lemma F.8.

Step 4. Uniform control of curvature. We next show that along the interpolation, the learner keeps nontrivial mass on the two distinguished actions i and j: both $y_i(r)$ and $y_j(r)$ stay uniformly bounded away from 0 by the choice of our step size η (Lemma F.9). This prevents the curvature $\theta''(y_i(r))$ and $\theta''(y_j(r))$ from blowing up along the path and stay uniformly upper bounded by some constant M. Therefore, the denominator in (10) is at most $2M$, which yields a constant lower bound

$$\langle v', y_{2s}\rangle - \langle v', y_{2s+1}\rangle = \int_0^1 \eta\, W_r\ \mathrm{Var}(Z_r)\, dr \geq \eta\frac{(v'_i - v'_j)^2}{2M} > 0. \tag{11}$$

Step 5. Sum the uniform one-step lower bound. We lower bound the payoff gap by combining the perturbation property from Lemma F.6, $v'_i - v'_j \;=\; x_\ell\big(A_{\ell i} - A_{\ell j}\big)$, with the anti-degeneracy event $\mathcal{E}_{\mathrm{gap}}$ from Step 1, which yields $v'_i - v'_j \geq C_A > 0$. Together with the uniform curvature bound M from Step 4, this gives the one-step lower bound $\langle v', y_{2s}\rangle - \langle v', y_{2s+1}\rangle \geq \eta C_A^{\mathrm{h}}$ as stated in (11), where $C_A^{\mathrm{h}} = (C_A)^2/(2M)$. Substituting into (9) and summing over $T/2$ pairs yields

$$\sum_{t=0}^{T-1} x_d(t)^\top A y_t \;\geq\; T\,\mathrm{Val}(A) + \eta T\, C_A^{\mathrm{h}}.$$

The stated success probability follows by a union bound over $\mathcal{A}_{nm} \cap \mathcal{E}_{\mathrm{gap}}$.

4.3 The Price of ε-Best-Response

Results 1 and 2 quantify how much surplus a clairvoyant optimizer can extract. We close by inverting this perspective: *how much surplus must the learner inevitably surrender simply to identify her approximate ε-best response?*

Under classical minimax analysis, the choice of regularizer is largely immaterial—all strongly convex choices yield $\Theta(\sqrt{T})$ regret, with the entropic regularizer preferred on the simplex for its dimension dependence. This apparent equivalence breaks once we view learning as *information acquisition* under adversarial asymmetry. The optimizer's surplus is not merely extracted; it is a *tuition fee* the learner pays to filter noise and converge to y^*. We formalize the minimum such fee from learner's perspective as the *Price of Best-Response*:

$$\mathcal{C}(\varepsilon) \;:=\; \inf_{\eta,\,t}\ \sup_{x_d(\cdot),\,A}\ \Big\{\,\mathrm{SURPLUS}_{\mathrm{h}}^{(\eta,t)}\ \Big|\ \|y(t) - y^*\|_1 \leq \varepsilon\,\Big\}. \tag{PBR}$$

In this regime, we provide the following results:

Theorem 4.4 (Discrete Time)**.** *For any sufficiently small $\varepsilon > 0$, $\mathcal{C}(\varepsilon) \;=\; \Omega(\theta'(1/k) - \theta'(\frac{\varepsilon}{2(m-k)}))$.*

Observation 4.5 (Euclidean vs. Entropic tuition). *As $\varepsilon \to 0$: Euclidean regularization ($\theta'(u) = u$) yields $\mathcal{C}_{\mathrm{Euc}}(\varepsilon) = \Theta(1)$—a* finite*, one-time fee, independent of precision. Entropic regularization ($\theta'(u) \approx \log u$) yields $\mathcal{C}_{\mathrm{Ent}}(\varepsilon) = \Theta(\log(1/\varepsilon)) \to +\infty$—an* unbounded *rent that grows as the learner seeks higher precision.*

This reveals a fundamental "paradox": the very property that makes entropic regularization robust—infinite boundary gradients that prevent over-committing to pure strategies—is precisely what makes it *infinitely saturatable* in the limit. *Strict safety and efficient best-response identification are structurally incompatible.* The full analysis is in Appendix G. Appendix H separately provides concentration bounds comparing sampled payoffs with their mixed-strategy counterparts along arbitrary predictable mixed-strategy trajectories.

Epilogue. Taken together, our three results describe a single phenomenon from complementary angles: *how a clairvoyant optimizer extracts utility against an FTRL learner, and how the regularizer governs that extraction*. The experiments in Appendix I indicate that the predicted rates are consistent with the predicted scalings. Natural next steps include adaptive step sizes and movement-constrained learners, partial-feedback models such as bandit observations against optimal learning rates, and extensions beyond full-information zero-sum games to richer strategic settings. Another intriguing direction is to understand how much surplus can still be recovered when the optimizer knows only the learner's algorithmic profile, but not the payoff matrix A. A broader discussion is deferred to Appendix J.

Acknowledgments and Disclosure of Funding

Acks. We thank Jon Schneider for insightful discussions during the preliminary stages of this work. Emmanouil V. Vlatakis-Gkaragkounis acknowledges support from the faculty startup funds provided by the University of Wisconsin–Madison.

Competing Interests. The authors declare no competing financial interests related to this work.

References

[1] Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 28, 2015. URL https://arxiv.org/abs/1512.04152.

[2] Nivasini Ananthakrishnan, Yuval Dagan, and Kunhe Yang. Learning local stackelberg equilibria from repeated interactions with a learning agent, 2025. URL https://arxiv.org/abs/2510.22471.

[3] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012. URL https://theoryofcomputing.org/articles/v008a006/.

[4] Eshwar Ram Arunachaleswaran, Natalie Collina, and Jon Schneider. Learning to play against unknown opponents. In *Proceedings of the 26th ACM Conference on Economics and Computation*, EC '25, page 478–504, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400719431. doi: 10.1145/3736252.3742575. URL https://doi.org/10.1145/3736252.3742575.

[5] Angelos Assos, Yuval Dagan, and Constantinos Daskalakis. Maximizing utility in multi-agent environments by anticipating the behavior of other learners. *Advances in Neural Information Processing Systems*, 37:38769–38798, 2024. URL https://arxiv.org/abs/2407.04889.

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331, 1995. doi: 10.1109/SFCS.1995.492488. URL https://ieeexplore.ieee.org/document/492488.

[7] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375. URL https://doi.org/10.1137/S0097539701398375.

[8] Mark Braverman, Jieming Mao, Jon Schneider, and Matt Weinberg. Selling to a no-regret buyer. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, EC '18, page 523–538, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450358293. doi: 10.1145/3219166.3219233. URL https://doi.org/10.1145/3219166.3219233.

[9] Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365 (6456):885–890, 2019. doi: 10.1126/science.aay2400. URL https://www.science.org/doi/abs/10.1126/science.aay2400.

[10] William Brown, Jon Schneider, and Kiran Vodrahalli. Is learning in games good for the learners? *Advances in Neural Information Processing Systems*, 36:54228–54249, 2023. URL https://arxiv.org/abs/2305.19496.

[11] Sébastien Bubeck. Convex optimization: Algorithms and complexity, 2015. URL https://arxiv.org/abs/1405.4980.

[12] Linda Cai, S. Matthew Weinberg, Evan Wildenhain, and Shirley Zhang. Selling to multiple no-regret buyers. In *Web and Internet Economics: 19th International Conference, WINE 2023, Shanghai, China, December 4–8, 2023, Proceedings*, page 113–129, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-48973-0. doi: 10.1007/978-3-031-48974-7_7. URL https://doi.org/10.1007/978-3-031-48974-7_7.

[13] Natalie Collina, Aaron Roth, and Han Shao. Efficient prior-free mechanisms for no-regret agents. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 511–541, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707049. doi: 10.1145/3670865.3673460. URL https://doi.org/10.1145/3670865.3673460.

[14] Yuan Deng, Jon Schneider, and Balasubramanian Sivan. Strategizing against no-regret learners. In *Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1909.13861.

[15] Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. *Advances in Neural Information Processing Systems*,

34:22655–22666, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bf40f0ab4e5e63171dd16036913ae828-Abstract.html.

[16] Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information, 2021. URL https://arxiv.org/abs/2101.04667.

[17] Samuel Karlin. *Mathematical Methods and Theory in Games, Programming and Economics*, volume 1. Addison-Wesley, Reading, Mass., 1959.

[18] Yoav Kolumbus, Jon Schneider, Inbal Talgam-Cohen, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Joshua R Wang, and SM Weinberg. Contracting with a learning agent. *Advances in Neural Information Processing Systems*, 37:77366–77408, 2024. URL https://arxiv.org/abs/2401.16198.

[19] Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *Journal of Dynamics and Games*, 4(2):125–148, 2017. ISSN 2164-6066. doi: 10.3934/jdg.2017008. URL https://arxiv.org/abs/1401.6956.

[20] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020. URL https://tor-lattimore.com/downloads/book/book.pdf.

[21] Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Bias no more: high-probability data-dependent regret bounds for adversarial bandits and mdps, 2020. URL https://arxiv.org/abs/2006.08040.

[22] Tao Lin and Yiling Chen. Generalized principal-agent problem with a learning agent, 2025. URL https://arxiv.org/abs/2402.09721.

[23] Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. Strategizing against learners in bayesian games. In *Conference on Learning Theory*, pages 5221–5252. PMLR, 2022. URL https://arxiv.org/abs/2205.08562.

[24] Brendan O'Donoghue, Tor Lattimore, and Ian Osband. Matrix games with bandit feedback, 2021. URL https://arxiv.org/abs/2006.05145.

[25] R. Tyrrell Rockafellar and Roger J. B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer-Verlag, Berlin Heidelberg, 1998. ISBN 978-3-540-62772-2. doi: 10.1007/978-3-642-02431-3. URL https://sites.math.washington.edu/~rtr/papers/rtr169-VarAnalysis-RockWets.pdf.

[26] Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4(2):107–194, February 2012. ISSN 1935-8237. doi: 10.1561/2200000018. URL https://doi.org/10.1561/2200000018.

[27] Steve Yang, Mark Paddrik, Roy Hayes, Andrew Todd, Andrei Kirilenko, Peter Beling, and William Scherer. Behavior based learning in identifying high frequency trading strategies. In *2012 IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 1–8, 2012. doi: 10.1109/CIFEr.2012.6327783. URL https://ieeexplore.ieee.org/document/6327783.

[28] Junyao Zhao. From no-regret to strategically robust learning in repeated auctions, 2026. URL https://arxiv.org/abs/2601.03853.

[29] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 928–935. AAAI Press, 2003. ISBN 1577351894. URL https://dl.acm.org/doi/10.5555/3041838.3041955.

[30] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'07, page 1729–1736, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520. URL https://dl.acm.org/doi/10.5555/2981562.2981779.

A Broader Impact and Limitations

Broader Impact. This work contributes to the mathematical foundations of multi-agent learning, with implications for the design, evaluation, and governance of learning algorithms in strategic environments.

On the positive side, our results provide *explicit, quantitative* warnings to practitioners who deploy FTRL-based algorithms—including Multiplicative Weights, Online Gradient Descent, and Tsallis-regularized methods—in settings where an informed counterpart may anticipate their dynamics. The Inverse-Rate Law (Section 4.1) and the Price of ε-Best-Response (Section 4.3) together furnish a concrete diagnostic checklist: how many suboptimal actions does the learner maintain, what step size is used, and how steep is the regularizer? These are, in principle, measurable quantities, and our bounds translate them into explicit surplus estimates. In particular, the geometric dichotomy between steep and non-steep regularizers offers actionable guidance: when high-precision convergence is required in the presence of a strategic counterpart, non-steep regularizers (e.g., Euclidean projection) incur a *finite* structural cost, whereas steep regularizers (e.g., entropic/MWU) incur a diverging one.

Such awareness is especially consequential in domains where algorithmic decisions interact repeatedly with strategic human or institutional actors, including automated market-making and algorithmic trading [27], repeated procurement and contract design [18], and multi-agent reinforcement learning systems in which one agent has partial knowledge of another's policy class. In such settings, understanding the structural cost of learning—not just worst-case regret or asymptotic convergence—is essential for responsible system design.

On the negative side, we acknowledge that our results also provide a *constructive* toolkit: the alternating trap strategy of Theorem 4.3 and the fixed-strategy surplus bounds of Theorem 4.1 could, in principle, be used by a strategic agent to improve its own payoff at the expense of a learning counterpart. This ethical concern is especially relevant in high-stakes settings, where even modest but repeated utility extraction may lead to unfair or systematically asymmetric outcomes for less informed participants. We believe, however, that transparency about such vulnerabilities—rather than obscuring them—is the more responsible path, because it enables scrutiny, robustification, and, where appropriate, institutional or regulatory safeguards. The present work is purely theoretical and does not introduce a deployable system or recommend adversarial use in practice.

Limitations. We highlight four boundaries of the current analysis that represent natural directions for future work.

1. **Zero-sum and separable structure.** Our main results (Theorems 4.1 and 4.3) are established for two-player zero-sum matrix games with separable regularizers. While this is the canonical adversarial benchmark, extension to *general-sum* games or *non-separable* regularizers (e.g., matrix-entropy or spectral regularizers) requires new tools beyond those developed here.
2. **Constant step size.** Our results are established for a *constant* step size η throughout the horizon T. Although adaptive or decaying schedules, such as $\eta_t = \Theta(1/\sqrt{t})$, are theoretically attractive due to their noise-damping and stabilizing effects, constant or non-vanishing step sizes remain central in practice because of their simplicity, scalability, and strong finite-time behavior. In particular, vanishing schedules may incur long transients, whereas constant step sizes often reach near-stationary regimes more quickly. Extending our framework to adaptive and decaying schedules remains an important direction for future work.
3. **Full-information feedback.** The main body of the paper operates under the *full-information* (full-feedback) model, where the learner observes the entire payoff vector at each round. Appendix H provides only an observation-level concentration comparison for arbitrary predictable mixed-strategy trajectories. It does not establish extraction guarantees for FTRL learners updating from bandit estimates. Such guarantees remain open.
4. **Clairvoyance assumption.** The optimizer is assumed to know the learner's update rule $\mathcal{L}$, step size η, and regularizer θ exactly. This "white-box" model is natural for theoretical lower bounds but may be unrealistic in practice. Robustness of the surplus extraction results to partial or approximate knowledge of the learner's algorithm is not addressed here.

None of these limitations undermines the core contribution: showing that surplus saturation is a *structural property* of the FTRL family in zero-sum games, encoded in the geometry of the regularizer rather than in any external economic mechanism.

B Related Work

In this section, we provide a detailed overview of the literature surrounding the strategic anticipation of learning algorithms, placing our contributions within the broader landscape of mechanism design and game theory.

Strategic Anticipation in Mechanism Design and Principal–Agent Problems. The paradigm of exploiting a learning agent's predictability originated in the field of algorithmic mechanism design. The seminal work of Braverman et al. [8] and subsequently Cai et al. [12] established that in repeated auctions, a strategic seller can leverage the "mean-based" property of standard no-regret algorithms (such as EXP3 and MWU) to extract revenue approaching the entire social surplus. This line of work highlights that without specific format constraints (e.g., no overbidding), the learner's regret guarantees are insufficient to protect their utility. Collina et al. [13] further refined this by addressing adversarial states and policy-regret goals, proposing "stable" policies to mitigate alignment assumptions.

Beyond auctions, this perspective has been generalized to Principal–Agent interactions. Lin and Chen [22] provide a unified framework for general Stackelberg settings, reducing the repeated interaction to a one-shot approximate best response problem and bounding the principal's utility based on the agent's regret (or swap-regret) guarantees. In the specific context of dynamic contracting, Kolumbus et al. [18] demonstrate how a principal can induce a "free-fall" in the agent's action space by switching from a linear to a zero-reward contract, effectively securing rewards at zero cost.

Strategizing in Repeated Games. Parallel to mechanism design, recent research has examined rational optimizers in general repeated games. Deng et al. [14] analyze the limits of what an optimizer can force against a no-regret learner, showing that while "mean-based" learners can be exploited to yield strictly more than the Stackelberg value, learners with no-swap-regret can effectively cap the optimizer's payoff. In the Bayesian setting, Mansour et al. [23] extend this inquiry to games with private types, introducing the concept of "polytope swap regret" to characterize the minimum payoff an optimizer can guarantee against a learning opponent.

Comparison with our work. The works above share a common thread: in rich economic environments—auctions, contracts, Bayesian games—a clairvoyant principal can steer a learning agent to yield substantially more than any static benchmark, often achieving linear surplus $\Omega(T)$ above the Stackelberg value. Our work departs from this literature in two respects:

- First, we operate in the simplest possible strategic environment: a *zero-sum matrix game*, where the Stackelberg value coincides with the minimax von Neumann value $V^\star$ and no asymmetric economic structure is present. This is precisely the setting where one might expect learning to be safest—and yet we show that structural surplus persists.
- Second, and more importantly, we focus on the *sublinear regime*. The reward sandwich $T \cdot V^\star \leq \mathrm{Rew}_T \leq T \cdot V^\star + \mathrm{Regret}^\eta_{\mathcal{L}}(T)$ caps the optimizer's gain at $o(T)$ when the learner employs standard no-regret algorithms with $\eta = \Theta(T^{-1/2})$. Rather than asking whether linear surplus is achievable—it is not, in this setting—we ask whether a clairvoyant optimizer can *saturate* the sublinear ceiling $\Theta(\sqrt{T})$. This is a strictly harder and more delicate question, requiring a fine-grained understanding of how the regularizer's geometry governs the learner's convergence rate. Our answer is affirmative, and the saturation conditions are encoded entirely in the structure of the FTRL algorithm itself—not in any external economic mechanism.

In other words, both the works on general-sum games that obtain surplus $\Omega(T)$ and the works on zero-sum games that obtain surplus $\Omega(T^{1/2})$ revolve around the same central question:

How can one optimally anticipate a no-regret learner so as to maximize one's own utility?

The $o(T)$ Regime. It is important to note that $\Omega(T)$ exploitation is not inevitable under all conditions. Zhao [28] identifies conditions in single-item auctions where monotone bidding strategies combined with gradient feedback render learners "strategically robust," preventing the auctioneer from exceeding Myerson's optimal revenue. The authors provided results where saturated surplus

is tightly $\Theta(\sqrt{T})$. The most closely related thread to our work involves optimizers that explicitly anticipate the learner's update rule. Assos et al. [5] adopt this "white-box" view, designing optimal strategies against Replicator Dynamics (the continuous-time analogue of MWU) and providing discrete-time guarantees for MWU. Our work aligns with this "control-theoretic" approach but generalizes the scope significantly. Instead of focusing solely on MWU/Replicator dynamics, we develop a unified theory for the entire FTRL family. Crucially, we introduce the geometric dichotomy between Steep and Non-Steep regularizers, showing that maximal extractability is not merely a feature of multiplicative updates but a structural consequence of boundary curvature.

Technical comparison with Assos et al. [5] Assos et al. establish a strong zero-sum result for MWU/Replicator Dynamics, which corresponds to the negative-entropy regularizer in our FTRL framework. The argument is fundamentally tied to the entropy structure of MWU and does not provide a theory for general FTRL regularizers. In contrast, our analysis works directly with the FTRL choice map $y(t) = Q_{\mathrm{h}}(\eta\, \mathrm{S}(t))$, and therefore covers entropy/MWU, Tsallis, Euclidean, and other separable regularizers within a single framework. This also lets us expose how the regularizer geometry changes the optimizer's surplus. In particular, while the analysis of Assos et al. does not compare different no-regret learners, our $\phi = (\theta')^{-1}$-based proof shows that the surplus is governed by the curvature and boundary behavior of h: steep regularizers such as entropy and Tsallis keep all actions in the interior and approach the boundary only asymptotically, whereas non-steep regularizers such as Euclidean may eliminate suboptimal actions in finite time.

A second difference concerns how the fixed-strategy surplus is organized. Assos et al.'s proof of Proposition 7 contains an implicit separation between the contribution of best-response coordinates and the exponentially vanishing contribution of non-best-response coordinates in the MWU formula. However, this separation is used inside the entropy-specific closed-form calculation, rather than being formulated as a general structural decomposition for arbitrary constant optimizer strategies and general FTRL choice maps. Our analysis makes this separation explicit through the decomposition from Equation (Surplus Decomposition)

$$\mathrm{SURPLUS}(T) = \mathrm{VAG}(T) + \mathrm{LAG}(T).$$

This decomposition holds for every constant optimizer strategy $\hat{x}$. When $\hat{x} = x^\star$ is max-min, the value-gap term vanishes, so the surplus is exactly the learner-approximation gap. By contrast, if $\hat{x}$ is not max-min, then the value-gap term is negative and linear in T. Since the learner-approximation gap is only $O(1/\eta)$, this negative linear term dominates whenever $1/\eta = o(T)$. Thus, among fixed strategies, the max-min strategy is the only one that can sustain positive surplus asymptotically, and the remaining surplus comes precisely from the learner's transient mass on suboptimal actions.

Finally, for alternating optimizer strategies, Assos et al. prove an $\Omega(\eta T)$ lower bound under a structural condition, with a game-dependent constant that may depend on the payoff matrix A. Our contribution is not only to extend the proof architecture beyond MWU, but also to quantify when this alternating construction is meaningful. In their result, the constant hidden in the $\Omega(\eta T)$ term is game-dependent, and the argument does not verify that this constant is bounded away from zero for a typical game, nor does it provide an explicit rate. In contrast, we show that for random zero-sum games the required alternating-trap condition holds with high probability, and we obtain an explicit polynomial lower bound on the corresponding game-dependent constant. Thus the lower bound is not merely qualitative: the $\Omega(\eta T)$ term has a nondegenerate coefficient under high-probability random-game events. The randomization in our theorem is used only to certify these structural events.

Dynamics-Specific Analysis. The technical foundation for our discrete-continuous bridge is rooted in the work of Kwon and Mertikopoulos [19], who formalize the disparity between continuous and discrete time to provide a unified treatment of FTRL regret through potential-based decompositions. Regarding the stability of the learning process itself, Giannou et al. [16] provide a comprehensive analysis of FTRL dynamics in multi-player games, establishing an equivalence between the stochastic stability of equilibria and the strictness of best responses. While their work focuses on convergence and stability rather than adversarial surplus extraction, it provides the dynamical foundations upon which our surplus-extraction analysis is built.

C Details Omitted from Section 2 (Preliminaries)

C.1 Definitions and Notation

Definition 3 (Simplex). *For a positive integer $d \in \mathbb{N}$, we define the probability simplex in $\mathbb{R}^d$ as*

$$\Delta^d \;:=\; \{z \in \mathbb{R}^d : \sum_{i=1}^{d} z_i = 1,\; z_i \geq 0,\; \forall i \in [d]\}.$$

Definition 4 (Best-response set). *For $x \in \Delta^n$, define the learner's best-response set*

$$\mathrm{br}(x) \;:=\; \arg\min_{i \in [m]} \langle x,\, Ae_i \rangle \;\subseteq\; [m].$$

This definition is equivalent to the one used in Section 3.2,

$$\mathrm{br}(x) \;:=\; \{i \in [m] \mid \delta_i = 0\},$$

since $\delta_i = 0$ holds exactly for learner actions i attaining the minimum payoff against x.

Definition 5 (Support). *For a mixed strategy $x \in \Delta^n$, define its support as the index set*

$$\mathrm{supp}(x) \;:=\; \{k \in [n] : x_k > 0\} \;\subseteq [n].$$

C.2 Omitted Convex Analysis Preliminaries

First, we introduce the definitions of strong convexity and smoothness.

Definition 6. *Let $f \colon \Delta^m \to \mathbb{R}$ be a differentiable function and let $\|\cdot\|$ be a norm on the space. For $\alpha > 0$ and $\beta > 0$, we define:*

1. *f is α-strongly convex with respect to $\|\cdot\|$ if, for all $y_1, y_2 \in \Delta^m$,*

$$f(y_1) \geq f(y_2) + \langle \nabla f(y_2),\, y_1 - y_2 \rangle + \frac{\alpha}{2}\|y_1 - y_2\|^2.$$

2. *f is β-smooth with respect to $\|\cdot\|$ if, for all $y_1, y_2 \in \Delta^m$,*

$$f(y_1) \leq f(y_2) + \langle \nabla f(y_2),\, y_1 - y_2 \rangle + \frac{\beta}{2}\|y_1 - y_2\|^2.$$

Strong convexity is equivalent to smoothness of the conjugate.

Proposition C.1 (Rockafellar and Wets [25, Proposition 12.60]). *Let $f \colon \Delta^m \to \mathbb{R} \cup \{+\infty\}$ be proper and lower semi-continuous. Then for $\alpha > 0$, the following are equivalent:*

- *f is α-strongly convex with respect to $\|\cdot\|$.*
- *f^* is differentiable and ∇f^* is $1/\alpha$-Lipschitz.*
- *f^* is $\beta = 1/\alpha$-strongly smooth with respect to the dual norm $\|\cdot\|_*$ on ${\Delta^m}_*$.*

Therefore, since our regularizers are proper and lower semicontinuous by definition, they satisfy the assumptions of Proposition C.1.

Proposition C.2 (Shalev-Shwartz [26, Section 2.7]). *Under the assumption of Proposition C.1, the choice map Q_f is given by the gradient of the conjugate:*

$$Q_f(y) = \nabla f^*(y), \qquad \forall y \in {\Delta^m}_*.$$

We next introduce the Bregman divergence induced by a convex potential, which will serve as our basic geometric tool when we move from continuous to discrete time doing discretization.

Definition 7 (Bregman Divergence). *Let $f \colon \Delta^m \to \mathbb{R}$ be a differentiable convex function. The Bregman divergence induced by f is defined as*

$$\mathrm{D}_f(y_1, y_2) \coloneqq f(y_1) - f(y_2) - \langle \nabla f(y_2),\, y_1 - y_2 \rangle,$$

for all $y_1, y_2 \in \Delta^m$.

Much of our analysis will take place in the dual space ${\Delta^m}_*$. In this context, it is natural to consider the Bregman divergence induced by the conjugate function f^*, which measures deviation in ${\Delta^m}_*$ relative to its linearization. Building on the equivalence between strong convexity and smoothness, we now derive a useful upper bound on the Bregman divergence of the conjugate.

Proposition C.3. *If f is α-strongly convex with respect to $\|\cdot\|$, then for all $z_1, z_2 \in {\Delta^m}_*$,*

$$\mathrm{D}_{f^*}(z_1, z_2) \le \frac{1}{2\alpha}\|z_1 - z_2\|_*^2.$$

Proof. Since f is α-strongly convex, Proposition C.1 implies that f^* is $\beta = 1/\alpha$-strongly smooth. By the definition of Bregman divergence and smoothness (Definition 6),

$$\mathrm{D}_{f^*}(z_1, z_2) = f^*(z_1) - f^*(z_2) - \langle \nabla f^*(z_2),\, z_1 - z_2\rangle \le \frac{1}{2\alpha}\|z_1 - z_2\|_*^2.$$

□

C.3 KKT Derivation of the Kernel Choice Map

Fix $z \in \mathbb{R}^m$ and recall (Choice Map)

$$Q_{\mathrm{h}}(z) \;\coloneqq\; \operatorname*{arg\,max}_{y \in \Delta^m} \{\langle z, y\rangle - \mathrm{h}(y)\}. \tag{12}$$

Assume h is a separable regularizer on Δ^m with kernel θ.

KKT conditions. Introduce $\lambda \in \mathbb{R}$ for the equality constraint $\sum_i y_i = 1$ and $\nu \in \mathbb{R}^m_+$ for the inequalities $y_i \ge 0$. The Lagrangian is

$$\mathcal{L}(y, \lambda, \nu) \;\coloneqq\; \langle z, y\rangle - \mathrm{h}(y) + \lambda\Big(1 - \sum_{i=1}^m y_i\Big) + \sum_{i=1}^m \nu_i y_i. \tag{13}$$

At the maximizer $y^* = Q_{\mathrm{h}}(z)$, the KKT system is

$$y^* \in \Delta^m, \quad \nu \in \mathbb{R}^m_+, \tag{14}$$

$$0 \in \partial_{y_i}\mathcal{L}(y^*, \lambda, \nu) \iff \theta'(y_i^*) = z_i + \lambda + \nu_i, \qquad i \in [m], \tag{15}$$

$$\nu_i\, y_i^* = 0, \qquad i \in [m]. \tag{16}$$

Since h is strictly convex on Δ^m, the maximizer $y^* = Q_{\mathrm{h}}(z)$ is unique. By (16), if $y_i^* > 0$ then $\nu_i = 0$ and $\theta'(y_i^*) = \lambda + z_i$. If $y_i^* = 0$, then $\nu_i \ge 0$ and (15) implies $z_i + \lambda \le \theta'(0^+)$. If $y_i^* = 1$, then (15) implies $z_i + \lambda \ge \theta'(1)$. Recalling the definition of $\phi = \mathrm{Trunc}[(\theta')^{-1}]$, (15)–(16) yield Equation (1) we need:

$$(Q_{\mathrm{h}}(z))_i \;=\; \phi(\lambda(z) + z_i), \qquad i \in [m], \tag{17}$$

for some scalar $\lambda(z) \in \mathbb{R}$, determined by the simplex constraint

$$\sum_{i=1}^m \phi(\lambda(z) + z_i) \;=\; 1. \tag{18}$$

Existence and uniqueness of $\lambda(z)$. Since θ is strongly convex, ϕ is strictly increasing and continuous, the function

$$g(\lambda) \;\coloneqq\; \sum_{i=1}^m \phi(\lambda + z_i)$$

is strictly increasing and continuous, with $\lim_{\lambda\to-\infty} g(\lambda) = 0$ and $\lim_{\lambda\to+\infty} g(\lambda) = m$. Therefore, by the intermediate value theorem, (18) admits at least one solution. Uniqueness follows from strict monotonicity of g at the solution.

C.4 Examples: Euclidean, Negative Entropy, and Negative Tsallis Kernels

We collect three standard separable regularizers $\mathrm{h}(y) = \sum_{i=1}^m \theta(y_i)$ on Δ^m and their associated one-dimensional (1). For each example we list θ, θ', the boundary slopes $\theta'(0^+)$, the resulting θ^* and ϕ.

C.4.1 Euclidean Regularizer.

We take the quadratic kernel $\theta(y_i) \coloneqq \frac{1}{2}y_i^2$ for $y_i \in [0,1]$, and the associated regularizer $\mathrm{h}(y) \coloneqq \frac{1}{2}\|y\|_2^2$. Its derivative satisfies $\theta'(y_i) = y_i$ and $\theta'(0^+) = 0$, so h is non-steep at the boundary $y_i = 0$. The conjugate is obtained by maximizing a concave quadratic over $y_i \in [0,1]$:

$$\theta^*(z_i) \;=\; \sup_{y_i\in[0,1]} \left\{ z_i y_i - \tfrac{1}{2}y_i^2 \right\} \;=\; \begin{cases} 0, & z_i < 0, \\ \frac{1}{2}z_i^2, & 0 \le z_i \le 1, \\ z_i - \frac{1}{2}, & z_i > 1, \end{cases} \tag{19}$$

and its derivative recovers the kernel choice map

$$\phi(z_i) \;=\; (\theta^*)'(z_i) \;=\; \Pi(z_i) \;\coloneqq\; \operatorname*{arg\,min}_{y_i\in[0,1]} \|y_i - z_i\|^2 \;=\; \begin{cases} 0, & z_i < 0, \\ z_i, & 0 \le z_i \le 1, \\ 1, & z_i > 1. \end{cases} \tag{20}$$

C.4.2 Negative Entropy Regularizer.

We take the entropy kernel $\theta(y_i) \coloneqq y_i \log y_i$ for $y_i \in (0,1]$ with $\theta(0) \coloneqq 0$, and the associated regularizer $\mathrm{h}(y) \coloneqq \sum_{i=1}^m y_i \log y_i$. Its derivative satisfies $\theta'(y_i) = 1 + \log y_i$ and $\theta'(0^+) = -\infty$, so h is steep at the boundary $y_i = 0$. The conjugate is obtained by maximizing $z_i y_i - y_i \log y_i$ over $y_i \in [0,1]$:

$$\theta^*(z_i) \;=\; \sup_{y_i\in[0,1]} \left\{ z_i y_i - y_i \log y_i \right\} \;=\; \begin{cases} \exp(z_i - 1), & z_i \le 1, \\ z_i, & z_i > 1, \end{cases} \tag{21}$$

and its derivative recovers the kernel choice map

$$\phi(z_i) \;=\; (\theta^*)'(z_i) \;=\; \begin{cases} \exp(z_i - 1), & z_i \le 1, \\ 1, & z_i > 1. \end{cases} \tag{22}$$

C.4.3 Negative Tsallis Regularizer ($q \in (0,1)$).

We take the negative Tsallis kernel $\theta(y_i) \coloneqq \frac{y_i^q - y_i}{q-1}$ for $y_i \in [0,1]$, and the associated regularizer $\mathrm{h}(y) \coloneqq \sum_{i=1}^m \frac{y_i^q - y_i}{q-1}$. Its derivative satisfies $\theta'(y_i) = \frac{q\, y_i^{q-1} - 1}{q-1}$ and $\theta'(0^+) = -\infty$, so h is steep at the boundary $y_i = 0$. The conjugate is obtained by maximizing a concave function over $y_i \in [0,1]$

$$\theta^*(z_i) \;=\; \sup_{y_i\in[0,1]} \left\{ z_i y_i - \theta(y_i) \right\} \;=\; \begin{cases} \left(\frac{1+(q-1)z_i}{q}\right)^{\frac{q}{q-1}}, & z_i \le 1, \\ z_i, & z_i > 1, \end{cases} \tag{23}$$

and its derivative recovers the kernel choice map:

$$\phi(z_i) \;=\; (\theta^*)'(z_i) \;=\; \begin{cases} \left(\frac{1+(q-1)z_i}{q}\right)^{\frac{1}{q-1}}, & z_i \le 1, \\ 1, & z_i > 1. \end{cases} \tag{24}$$

D Details Omitted from Section 3 (Toolbox)

D.1 Completed Proof of Lemma 3.1 and Theorem 3.3

We combine these statements into a single theorem, restate it here, and prove it.

Theorem D.1 (Lemma 3.1 & Theorem 3.3)**.** *Fix a constant step size $\eta > 0$ and let h be a regularizer on Δ^m.*

- ***Continuous time.*** *Let the continuous historical reward $\mathrm{S}(t)$ be defined as in* (Cumulative Reward)*. Then, for any continuous strategy $x_c(t) : [0,T] \to \Delta^n$,*

$$\mathrm{R}_{\mathrm{cont}}(x_c(t),\, \mathrm{S}(0),\, T) = \frac{1}{\eta}\, \mathrm{h}^*\big(\eta\, \mathrm{S}(0)\big) - \frac{1}{\eta}\, \mathrm{h}^*\big(\eta\, \mathrm{S}(T)\big). \tag{25}$$

- ***Discrete time.*** *Let the discrete historical reward* $\mathrm{S}(t)$ *be defined as in* (Cumulative Reward)*. Then, for any discrete strategy* $x_d(t)\colon \{0, 1, \dots, T-1\} \to \Delta^n$*,*

$$\mathrm{R}_{\mathrm{disc}}(x_d(t),\, \mathrm{S}(0),\, T) = \frac{1}{\eta}\, \mathrm{h}^*\big(\eta\, \mathrm{S}(0)\big) - \frac{1}{\eta}\, \mathrm{h}^*\big(\eta\, \mathrm{S}(T)\big) + \frac{1}{\eta} \sum_{t=0}^{T-1} \mathrm{D}_{\mathrm{h}^*}(\eta\, \mathrm{S}(t+1),\, \eta\, \mathrm{S}(t)). \tag{26}$$

Proof. Recall that the FTRL choice at dual state z is $y(z) = Q_{\mathrm{h}}(z) = \nabla\, \mathrm{h}^*(z)$. To analyze these reward functions, define the learner's potential

$$\Phi_\eta(Z) \coloneqq \frac{1}{\eta}\, \mathrm{h}^*(\eta Z), \qquad Z \in \mathbb{R}^m. \tag{27}$$

Then, by the chain rule, $\nabla \Phi_\eta(Z) = \nabla\, \mathrm{h}^*(\eta Z) = Q_{\mathrm{h}}(\eta Z)$. Thus, changes in the potential Φ_η track the reward accumulated along the trajectory.

Continuous time. Suppose $x_c(t)\colon [0, T] \to \Delta^n$ is a strategy for the continuous case. The continuous reward is defined as the integral of the instantaneous payoff

$$\mathrm{R}_{\mathrm{cont}}(x_c(t),\, \mathrm{S}(0),\, T) = \int_0^T \langle x_c(t), A \cdot Q_{\mathrm{h}}(\eta\, \mathrm{S}(t))\rangle\, dt.$$

Since $\dot{\mathrm{S}}(t) = B^\top x_c(t)$ and $A = -B$, we rewrite

$$\langle x_c(t), A \cdot Q_{\mathrm{h}}(\eta\, \mathrm{S}(t))\rangle = \langle A^\top x_c(t), Q_{\mathrm{h}}(\eta\, \mathrm{S}(t))\rangle = \langle -\dot{\mathrm{S}}(t), Q_{\mathrm{h}}(\eta\, \mathrm{S}(t))\rangle.$$

Using $Q_{\mathrm{h}}(\eta\, \mathrm{S}(t)) = \nabla \Phi_\eta(\mathrm{S}(t))$, we obtain

$$\langle -\dot{\mathrm{S}}(t), \nabla \Phi_\eta(\mathrm{S}(t))\rangle = -\frac{d}{dt} \Phi_\eta(\mathrm{S}(t)).$$

Therefore, by the Fundamental Theorem of Calculus,

$$\mathrm{R}_{\mathrm{cont}}(x_c(t),\, \mathrm{S}(0),\, T) = -\int_0^T \frac{d}{dt} \Phi_\eta(\mathrm{S}(t))\, dt = \Phi_\eta(\mathrm{S}(0)) - \Phi_\eta(\mathrm{S}(T)) \tag{28}$$

$$= \frac{1}{\eta}\, \mathrm{h}^*(\eta\, \mathrm{S}(0)) - \frac{1}{\eta}\, \mathrm{h}^*(\eta\, \mathrm{S}(T)). \tag{29}$$

Discrete time. Suppose $x_d(t)\colon \{0, 1, \dots, T-1\} \to \Delta^n$ is a strategy for the discrete case. The reward at time step t is

$$\langle x_d(t), A \cdot Q_{\mathrm{h}}(\eta\, \mathrm{S}(t))\rangle.$$

Since $\mathrm{S}(t+1) = \mathrm{S}(t) + B^\top x_d(t)$ and $A = -B$, we have

$$A^\top x_d(t) = \mathrm{S}(t) - \mathrm{S}(t+1).$$

Substituting this into the reward expression yields

$$\mathrm{R}_{\mathrm{disc}}(x_d(t),\, \mathrm{S}(t),\, 1) = \langle \mathrm{S}(t) - \mathrm{S}(t+1), Q_{\mathrm{h}}(\eta\, \mathrm{S}(t))\rangle = \langle \mathrm{S}(t) - \mathrm{S}(t+1), \nabla \Phi_\eta(\mathrm{S}(t))\rangle.$$

Applying the standard identity for the Bregman divergence from Definition 7 gives

$$\langle \mathrm{S}(t) - \mathrm{S}(t+1), \nabla \Phi_\eta(\mathrm{S}(t))\rangle = \Phi_\eta(\mathrm{S}(t)) - \Phi_\eta(\mathrm{S}(t+1)) + \mathrm{D}_{\Phi_\eta}(\mathrm{S}(t+1), \mathrm{S}(t)).$$

Substituting this back into the reward equation gives

$$\mathrm{R}_{\mathrm{disc}}(x_d(t),\, \mathrm{S}(t),\, 1) = \Phi_\eta(\mathrm{S}(t)) - \Phi_\eta(\mathrm{S}(t+1)) + \mathrm{D}_{\Phi_\eta}(\mathrm{S}(t+1), \mathrm{S}(t)).$$

Summing from $t = 0$ to $T-1$, the first two terms form a telescoping sum:

$$\sum_{t=0}^{T-1} \mathrm{R}_{\mathrm{disc}}(x_d(t),\, \mathrm{S}(t),\, 1) = \Phi_\eta(\mathrm{S}(0)) - \Phi_\eta(\mathrm{S}(T)) + \sum_{t=0}^{T-1} \mathrm{D}_{\Phi_\eta}(\mathrm{S}(t+1), \mathrm{S}(t))$$

$$= \frac{1}{\eta}\, \mathrm{h}^*(\eta\, \mathrm{S}(0)) - \frac{1}{\eta}\, \mathrm{h}^*(\eta\, \mathrm{S}(T)) + \sum_{t=0}^{T-1} \frac{1}{\eta}\, \mathrm{D}_{\mathrm{h}^*}(\eta\, \mathrm{S}(t+1), \eta\, \mathrm{S}(t)),$$

because $\mathrm{D}_{\Phi_\eta}(Z', Z) = \frac{1}{\eta}\, \mathrm{D}_{\mathrm{h}^*}(\eta Z', \eta Z)$. □

D.1.1 Completed Proof of Smoothness and Corollary 3.2

Claim D.2. *The function G is β-smooth with respect to $\|\cdot\|$, with*

$$\beta \;=\; \frac{(\eta\, T\, \|B^\top\|_{\mathrm{op}})^2}{\alpha}, \qquad \|B^\top\|_{\mathrm{op}} \coloneqq \sup_{x\neq 0} \frac{\|B^\top x\|_*}{\|x\|}.$$

Proof. Define the affine map

$$f(x) \coloneqq \eta\big(\,\mathrm{S}(0) + T B^\top x\big),$$

so that $G(x) = \mathrm{h}^*(f(x))$ and $\nabla f(x) = \eta\, T\, B^\top$. By the chain rule,

$$\nabla G(x) = (\eta\, T\, B^\top)^\top \nabla\, \mathrm{h}^*\big(f(x)\big).$$

Since h is α-strongly convex with respect to $\|\cdot\|$, its conjugate h^* is $(1/\alpha)$-smooth with respect to the dual norm $\|\cdot\|_*$. Thus, for any $x, x' \in \Delta^n$,

$$\begin{aligned}
\|\nabla G(x) - \nabla G(x')\|_* &= \Big\|(\eta\, T\, B^\top)^\top\Big(\nabla\, \mathrm{h}^*(f(x)) - \nabla\, \mathrm{h}^*(f(x'))\Big)\Big\|_* \\
&\le \eta\, T\, \|B^\top\|_{\mathrm{op}} \Big\|\nabla\, \mathrm{h}^*(f(x)) - \nabla\, \mathrm{h}^*(f(x'))\Big\|_* \\
&\le \eta\, T\, \|B^\top\|_{\mathrm{op}}\, \frac{1}{\alpha}\, \|f(x) - f(x')\|_* \\
&= \eta\, T\, \|B^\top\|_{\mathrm{op}}\, \frac{1}{\alpha}\, \big\|\eta\, T\, B^\top (x - x')\big\|_* \\
&\le \frac{(\eta\, T\, \|B^\top\|_{\mathrm{op}})^2}{\alpha}\, \|x - x'\|.
\end{aligned}$$

□

Corollary 3.2 (Frank-Wolfe Complexity [11, Theorem 3.8])**.** *An ϵ-approximate strategy $\hat{x}$ with $\mathrm{R_{cont}}^*(\mathrm{S}(0),\, T) - \mathrm{R_{cont}}(\hat{x},\, \mathrm{S}(0),\, T) \le \epsilon$ can be computed in time $\mathcal{O}\big(nm\, R^2 \eta T^2 \|A\|_{\mathrm{op}}^2/(\alpha\epsilon)\big)$, where R is the diameter of Δ^n.*

Proof. We apply the Frank-Wolfe algorithm to minimizing the objective function $G(x)$. According to Bubeck [11, Theorem 3.8], applying Frank–Wolfe to a β-smooth convex function over a domain with diameter R yields iterates $\{x_k\}_{k\ge 0}$ satisfying the convergence rate

$$G(x_k) - G(x^*) \;\le\; \frac{2R^2\,\beta}{k+1} \;=\; \frac{2R^2\,(\eta\, T)^2\, \|A\|_{\mathrm{op}}^2}{\alpha\,(k+1)},$$

where $x^* \in \arg\min_{x\in\Delta^n} G(x)$.

Recall from Lemma 3.1 (and Remark 3) that the optimizer's continuous reward is given by

$$\mathrm{R_{cont}}(x,\, \mathrm{S}(0),\, T) = \frac{1}{\eta}\, \mathrm{h}^*(\eta\, \mathrm{S}(0)) - \frac{1}{\eta} G(x).$$

Therefore, the suboptimality gap in terms of reward is directly proportional to the suboptimality gap of the function G, scaled by $1/\eta$

$$\mathrm{R_{cont}}^*(\mathrm{S}(0),\, T) - \mathrm{R_{cont}}(x_k,\, \mathrm{S}(0),\, T) \;=\; \frac{G(x_k) - G(x^*)}{\eta}.$$

To achieve a target accuracy of ϵ in the reward, we require

$$\frac{G(x_k) - G(x^*)}{\eta} \le \epsilon \implies G(x_k) - G(x^*) \le \epsilon\eta.$$

Substituting the convergence rate of Frank-Wolfe

$$\frac{2R^2\,(\eta\, T)^2\, \|A\|_{\mathrm{op}}^2}{\alpha\,(k+1)} \le \epsilon\eta.$$

Solving for the number of iterations k

$$k+1 \;\geq\; \frac{2R^2\,\eta^2\,T^2\,\|A\|_{\mathrm{op}}^2}{\alpha\,\epsilon\,\eta} \;=\; \frac{2R^2\,\eta\,T^2\,\|A\|_{\mathrm{op}}^2}{\alpha\,\epsilon}.$$

Thus, the number of iterations required is $k = \mathcal{O}\Big(\frac{R^2\,\eta\,T^2\,\|A\|_{\mathrm{op}}^2}{\alpha\,\epsilon}\Big)$. Since each Frank-Wolfe iteration involves a Linear Minimization Oracle over the simplex (or strategy space), which takes $\mathcal{O}(nm)$ time, the total runtime is

$$\text{Total Time} = k \times \mathcal{O}(nm) = \mathcal{O}\left(nm\,\frac{R^2\,\eta\,T^2\,\|A\|_{\mathrm{op}}^2}{\alpha\,\epsilon}\right).$$

□

D.2 Optimizer Reward Sandwich

Below we present a self-contained proof of the sandwich property for the reward of the optimizer. This recovers (⋆) from the Introduction. Proposition D.3 is the continuous-time analogue of the standard FTRL "regret ≤ regularizer diameter/η" bound.

Proposition D.3. *Let* h *be a regularizer on* Δ^m*. Fix a constant step size* $\eta > 0$ *and horizon* $T \in \mathbb{N}$*. The optimizer's optimal continuous-time reward is bounded as*

$$T \cdot \mathrm{Val}(A) \;\leq\; \mathrm{R_{cont}}^*(0,\,T) \;\leq\; T \cdot \mathrm{Val}(A) + \frac{(\mathrm{h_{max}} - \mathrm{h_{min}})}{\eta}. \tag{30}$$

Proof. The optimal continuous-time reward is trivially lower bounded by $T \cdot \mathrm{Val}(A)$ since the optimizer can always play a max-min strategy throughout $[0, T]$.

For the upper bound, assume $\mathrm{S}(0) = 0$. Then

$$\mathrm{h}^*(0) = \sup_{y\in\Delta^m}\{-\mathrm{h}(y)\} = -\inf_{y\in\Delta^m}\mathrm{h}(y) = -\mathrm{h_{min}}.$$

Starting from the definition of $\mathrm{R_{cont}}^*(\mathrm{S}(0),\,T)$:

$$\mathrm{R_{cont}}^*(\mathrm{S}(0),\,T) \;=\; \max_{x\in\Delta^n}\mathrm{R_{cont}}(x,\,\mathrm{S}(0),\,T) = \Phi_\eta\big(\mathrm{S}(0)\big) - \min_{x\in\Delta^n}\Phi_\eta\big(\mathrm{S}(0) + TB^\top x\big).$$

$$\mathrm{R_{cont}}^*(\mathrm{S}(0),\,T) \leq -\frac{\mathrm{h_{min}}}{\eta} - \min_{x\in\Delta^n}\frac{1}{\eta}\,\mathrm{h}^*\big(\eta\big(\mathrm{S}(0) + TB^\top x\big)\big).$$

By definition of h^* and $A = -B$,

$$\begin{aligned}
-\min_{x\in\Delta^n}\frac{1}{\eta}\,\mathrm{h}^*\big(\eta\big(\mathrm{S}(0) + TB^\top x\big)\big) &= -\frac{1}{\eta}\min_{x\in\Delta^n}\sup_{y\in\Delta^m}\big\{\eta\,\langle \mathrm{S}(0) + TB^\top x, y\rangle - \mathrm{h}(y)\big\} \\
&\leq -\frac{1}{\eta}\sup_{y\in\Delta^m}\min_{x\in\Delta^n}\big\{\eta\,\langle \mathrm{S}(0) + TB^\top x, y\rangle - \mathrm{h}(y)\big\} \\
&= \min_{y\in\Delta^m}\left\{-\langle \mathrm{S}(0), y\rangle + T\max_{x\in\Delta^n}\langle A^\top x, y\rangle + \frac{1}{\eta}\,\mathrm{h}(y)\right\} \\
&\leq -\min_{y\in\Delta^m}\langle \mathrm{S}(0), y\rangle + T\cdot\mathrm{Val}(A) + \frac{\mathrm{h_{max}}}{\eta}.
\end{aligned}$$

Plugging back yields

$$\mathrm{R_{cont}}^*(\mathrm{S}(0),\,T) \leq -\frac{\mathrm{h_{min}}}{\eta} - \min_{y\in\Delta^m}\langle \mathrm{S}(0), y\rangle + T\cdot\mathrm{Val}(A) + \frac{\mathrm{h_{max}}}{\eta},$$

and $\mathrm{S}(0) = 0$ concludes the proof. □

D.3 Details Omitted from Section 3.2

D.3.1 Learner's Viewpoint

We begin with a basic fact about FTRL when the optimizer plays a fixed constant strategy. For simplicity, assume the initial reward is 0.

Observation D.4. *If the optimizer plays a constant strategy $\hat{x}$, the learner's FTRL responses can be written, for all $t > 0$, as*

$$y(t) \;:=\; Q_{\mathrm{h}}\big(\eta\, t\, B^{\top}\hat{x}\big) = \arg\max_{y\in\Delta^m}\Big\{\langle \eta\, t\, B^{\top}\hat{x}, y\rangle - \mathrm{h}(y)\Big\} = \arg\min_{y\in\Delta^m}\Big\{\langle \hat{x}, Ay\rangle + \frac{1}{\eta\, t}\,\mathrm{h}(y)\Big\}. \tag{31}$$

From the last expression, we see that the regularization term vanishes as $t \to \infty$. Therefore, we expect to see the learner's FTRL responses converge to a best response to $\hat{x}$.

This observation will be a major tool of our analysis. We state this formally below.

Lemma D.5 (Lemma 3.4)**.** *Fix a constant strategy $\hat{x} \in \Delta^n$. Then the learner's FTRL response $y(t)$ converges to $\hat{y}^*$ as $t \to \infty$, where the limit is the unique point*

$$\hat{y}^* \;=\; \arg\min_{y\in \mathrm{conv}\{e_i : i\in \mathrm{br}(\hat{x})\}} \mathrm{h}(y).$$

Proof. We start from Equation (31). For $t > 0$,

$$y(t) = \arg\min_{y\in\Delta^m}\Big\{\langle \hat{x}, Ay\rangle + \frac{1}{\eta t}\,\mathrm{h}(y)\Big\}.$$

Let $\lambda_t = \frac{1}{\eta t}$ and define

$$\psi_t(y) := \langle \hat{x}, Ay\rangle + \lambda_t\,\mathrm{h}(y), \qquad \psi(y) := \langle \hat{x}, Ay\rangle, \qquad y \in \Delta^m.$$

Let

$$F = \arg\min_{y\in\Delta^m}\psi(y) = \arg\min_{y\in\Delta^m}\langle \hat{x}, Ay\rangle = \arg\min_{y\in \mathrm{conv}\{\, e_i : i\in \mathrm{br}(\hat{x})\,\}} \langle \hat{x}, Ay\rangle$$

since ψ is linear over the simplex. Define $\hat{y}^* := \arg\min_{y\in F}\mathrm{h}(y)$, which is unique because h is strictly convex and F is convex.

Fix any sequence $t_k \to \infty$. By compactness of Δ^m, the sequence $\{y(t_k)\}$ has a convergent subsequence. Assume $y(t_k) \to y_\infty$. Because Δ^m is compact and h is continuous, we have

$$\sup_{y\in\Delta^m} |\psi_{t_k}(y) - \psi(y)| = \sup_{y\in\Delta^m} \lambda_{t_k}|\,\mathrm{h}(y)| \to 0,$$

so $\psi_{t_k} \to \psi$ uniformly on Δ^m, hence ψ_{t_k} is epi-convergence to ψ. Since $y(t_k) \in \arg\min_{y\in\Delta^m}\psi_{t_k}(y)$ for each k, Rockafellar and Wets [25, Theorem 7.33] implies that every cluster point of $\{y(t_k)\}$ belongs to $\arg\min_{y\in\Delta^m}\psi(y) = F$. In particular, $y_\infty \in F$.

Since $y(t_k)$ minimizes ψ_{t_k}, taking $y = \hat{y}^*$ gives

$$\psi_{t_k}\big(y(t_k)\big) \le \psi_{t_k}(\hat{y}^*) \quad\Rightarrow\quad \psi\big(y(t_k)\big) + \lambda_{t_k}\,\mathrm{h}\big(y(t_k)\big) \le \psi(\hat{y}^*) + \lambda_{t_k}\,\mathrm{h}(\hat{y}^*).$$

Because $\hat{y}^* \in F$, we have $\psi(\hat{y}^*) \le \psi(y(t_k))$, hence

$$0 \le \psi\big(y(t_k)\big) - \psi(\hat{y}^*) \le \lambda_{t_k}\big(\mathrm{h}(\hat{y}^*) - \mathrm{h}\big(y(t_k)\big)\big),$$

which implies $\mathrm{h}(y(t_k)) \le \mathrm{h}(\hat{y}^*)$ for all k. By continuity of h,

$$\mathrm{h}(y_\infty) = \lim_{k\to\infty}\mathrm{h}\big(y(t_k)\big) \le \mathrm{h}(\hat{y}^*).$$

But $y_\infty \in F$ and $\hat{y}^*$ minimizes h over F, so $\mathrm{h}(\hat{y}^*) \le \mathrm{h}(y_\infty)$, and thus $\mathrm{h}(y_\infty) = \mathrm{h}(\hat{y}^*)$. By uniqueness of the minimizer of h on F, we conclude $y_\infty = \hat{y}^*$.

We have shown that every sequence $t_k \to \infty$ admits a subsequence along which $y(t_k) \to \hat{y}^*$; hence $y(t) \to \hat{y}^*$ as $t \to \infty$. □

The following remark completes the learner-viewpoint analysis for Lemma 3.4.

Remark 7. If the optimizer plays a max-min strategy $x^\star$, then by Lemma D.5, the learner converges to $y^* \in \arg\min_{y\in\mathrm{conv}\{e_i : i\in\mathrm{br}(x^\star)\}}\mathrm{h}(y)$, and moreover $\langle x^\star, Ay^*\rangle$ equals the value of the game $\mathrm{Val}(A)$. Since we do not necessarily have that $x^\star$ is a best response to y^*, y^* is not guaranteed to be a min-max strategy for the learner.

D.3.2 Optimizer's Viewpoint

The optimizer plays a fixed constant strategy $\hat{x}$. By Lemma D.5, the learner's response $y(t)$ converges to the unique regularizer-selected best response $\hat{y}^*$ with $u_o(\hat{x}, \hat{y}^*) = u_o^*$. The instantaneous surplus above the best-response value u_o^* at time t is $u_o(\hat{x}, y(t)) - u_o^*$. We restate and prove the gap decomposition below.

Lemma 3.5 (Optimizer's payoff)**.** $u_o(\hat{x}, y(t)) = u_o^* + \sum_{i \notin \mathrm{br}(\hat{x})} \delta_i \, y_i(t)$*, for all* $t \geq 0$*.*

Proof. By linearity, $u_o(\hat{x}, y(t)) = \langle v, y(t) \rangle = \sum_{i=1}^m (u_o^* + \delta_i) y_i(t) = u_o^* \sum y_i(t) + \sum \delta_i y_i(t)$. Since $y(t) \in \Delta^m$ and $\delta_i = 0$ for $i \in \mathrm{br}(\hat{x})$, the result follows. □

D.3.3 Completed Proof of Lemma 3.6

Lemma 3.6 (Pointwise FTRL Bounds)**.** *Fix* $\hat{x}$*, step size* $\eta > 0$*, and a separable regularizer* h*. The FTRL response satisfies* $y_i(t) = \phi(\lambda(t) - \eta t \, v_i)$ *for a normalization scalar* $\lambda(t)$*. Then for any* $t \geq 0$*:*

⋆***Optimal Actions:*** *For all* $i \in \mathrm{br}(\hat{x})$*,* $y_i(t)$ *are identical* a_t *with* $\frac{1}{m} \leq a_t \leq \frac{1}{k}$*.*

⋆***Suboptimal Actions:*** *For all* $j \notin \mathrm{br}(\hat{x})$*:* $\phi\big(\theta'(1/m) - \eta t \, \delta_j\big) \;\leq\; y_j(t) \;\leq\; \phi\big(\theta'(1/k) - \eta t \, \delta_j\big)$*.*

Additionally, steep regularizers yield asymptotic decay ($y_j(t) \to 0$)*, whereas non-steep ones eliminate suboptimal actions in finite time* $t \geq t_\theta^* := \frac{\theta'(1/k) - \theta'(0^+)}{\eta \delta_{\min}}$*.*

Proof. From Equation (31), $y(t) = Q_{\mathrm{h}}\big(\eta \, t \, B^\top \hat{x}\big) = Q_{\mathrm{h}}\big(-\eta \, t \, v\big)$. The KKT condition in (1) gives $y_i(t) = \phi(\lambda(t) - \eta t \, v_i)$ with $\sum_{i=1}^m y_i(t) = 1$ for some scalar $\lambda(t)$.

(i) Since $v_i = u_o^*$ for all $i \in \mathrm{br}(\hat{x})$, then $y_i(t) = \phi(\lambda(t) - \eta t \, u_o^*)$ for all $i \in \mathrm{br}(\hat{x})$, which is independent of i. Define $a_t := \phi(\lambda(t) - \eta t \, u_o^*) \geq 0$. Then, $y_i(t) = a_t$ for all $i \in \mathrm{br}(\hat{x})$.

Then we prove the bounds on a_t. Write the normalization as

$$1 = \sum_{i \in \mathrm{br}(\hat{x})} a_t + \sum_{j \notin \mathrm{br}(\hat{x})} y_j(t) = k a_t + \sum_{j \notin \mathrm{br}(\hat{x})} y_j(t) \;\geq\; k a_t,$$

so $a_t \leq 1/k$. Also, for $j \notin \mathrm{br}(\hat{x})$, by monotonicity of ϕ we have

$$y_j(t) = \phi(\lambda(t) - \eta t \, v_j) = \phi(\lambda(t) - \eta t (u_o^* + \delta_j)) \leq \phi(\lambda(t) - \eta t \, u_o^*) = a_t.$$

Therefore, $1 = \sum_{i=1}^m y_i(t) \leq m a_t$, so $a_t \geq 1/m$.

(ii) For $j \notin \mathrm{br}(\hat{x})$,

$$y_j(t) = \phi\big(\lambda(t) - \eta t (u_o^* + \delta_j)\big) = \phi\big(\theta'(a_t) - \eta t \, \delta_j\big).$$

Since $a_t \in [1/m, 1/k]$ and θ' is increasing, $\theta'(a_t) \in [\theta'(1/m), \theta'(1/k)]$. Applying the increasing map ϕ yields the claimed bounds. □

E Details Omitted from Section 4.1 (The Inverse-Rate Law)

We are now in a position to state our first main result. Assuming a fixed optimizer strategy, we establish that surplus is governed by the interplay between: a) the *strategic landscape* (specifically, the mass of learner's actions that are *not* best responses to optimizer) and b) the *geometry* of the regularizer's conjugate θ^*. The following meta-theorem encapsulates the essence of our findings:

Optimizer's surplus is fundamentally the ratio of strategic suboptimality to the learning speed.

E.1 The Cost of Learning

Our result establishes a fundamental sweeping condition: surplus against a fixed strategy arises *solely* from the presence of suboptimal actions (non-best responses). If the learner is already fully aligned with the optimizer's strategy, the cumulative lag vanishes. Conversely, any misalignment incurs an inescapable surplus cost scaling with the inverse step size.

Meta-Theorem. Consider a learner with step size η playing against a fixed optimizer strategy. Let N_{sub} be the number of suboptimal actions (non-best responses).

- **Perfect Alignment:** If all learner actions are best responses ($N_{sub} = 0$), there is **no surplus**.

$$\mathrm{LAG}(T) = 0$$

- **Cost of Misalignment:** If there exist suboptimal actions ($N_{sub} > 0$), the cumulative surplus is proportional to the number of bad choices divided by the step size, corrected by the boundary geometry:

$$\mathrm{LAG}(T) \approx \Theta\left(\frac{N_{sub}}{\eta(T)}\right) - o_{\eta T,\mathrm{h}}(1) \cdot \mathbb{I}(\text{steep}).$$

(Note: While non-steep regularizers saturate quickly, steep regularizers retain a slowly decaying residual potential, strictly reducing the realized surplus at finite T.)

E.2 Surplus in Continuous-Time FTRL Dynamics

Regularizer Geometry via a Dual Potential To turn the preceding meta-theorem into precise bounds, we need a scalar measure of how "costly" it is for the learner to maintain probability mass on a coordinate. We capture this effect through a *dual potential* associated with the regularizer.

Definition 8. *Let $\theta : [0,1] \to \mathbb{R}$ be the kernel of a separable regularizer and θ^* its convex conjugate. We define the* dual potential $V : [0,1] \to \mathbb{R}$ *by* $V(u) \coloneqq \theta^*(\theta'(u)) = u\,\theta'(u) - \theta(u)$.

Remark 8. For standard kernels, V admits simple closed forms: *Euclidean* regularizer yields $V(u) = \frac{1}{2}u^2$ (up to constants), *Tsallis* regularizer with parameter $q \neq 1$ yields $V(u) = u^q$, and the *entropic* case arises in the limit $q \to 1$, giving $V(u) = u$ (again up to irrelevant constants).

We now provide the formal statement of Theorem 4.1. While the main text provided a simplified version for intuition, the statement below specifies the exact bounds in terms of the potential energy change ΔV and the discretization factor N_{sub}.

Theorem E.1 (Formal Version of Theorem 4.1)**.** *Fix a constant optimizer strategy $\hat{x}$. Let $\delta_{\min}, \delta_{\max}$ be the minimum and maximum payoff gaps of the suboptimal actions. For any horizon $T \geq 0$,* $\mathrm{LAG}(T)$ *is bounded by*

$$\frac{m-k}{\eta} \cdot \underline{\Delta V}(T) \;\leq\; \mathrm{LAG}(T) \;\leq\; \frac{m-k}{\eta} \cdot \overline{\Delta V}(T), \qquad \text{(Surplus Bound)}$$

where $\underline{\Delta V}(T), \overline{\Delta V}(T)$ represent the potential drops of form $\Delta V(T) \sim V_{start} - V_{end}(T)$:

$$\underline{\Delta V}(T) := \frac{\delta_{\min}}{\delta_{\max}}\left[V(1/m) - \theta^*\Big(\max\{\theta'(1/m) - \eta\delta_{\max}T,\ \theta'(0^+)\}\Big)\right]. \tag{32}$$

$$\overline{\Delta V}(T) := \frac{\delta_{\max}}{\delta_{\min}}\left[V(1/k) - \theta^*\Big(\max\{\theta'(1/k) - \eta\delta_{\min}T,\ \theta'(0^+)\}\Big)\right]. \tag{33}$$

Proof. Start from (Surplus Decomposition)

$$\mathrm{SURPLUS}(T) = \mathrm{VAG}(T) \;+\; \int_0^T \sum_{i \notin \mathrm{br}(\hat{x})} \delta_i\, y_i(t)\, dt.$$

For $i \notin \mathrm{br}(\hat{x})$, $\delta_{\min} \leq \delta_i \leq \delta_{\max}$, hence

$$\delta_{\min} \sum_{i \notin \mathrm{br}(\hat{x})} y_i(t) \;\leq\; \sum_{i \notin \mathrm{br}(\hat{x})} \delta_i\, y_i(t) \;\leq\; \delta_{\max} \sum_{i \notin \mathrm{br}(\hat{x})} y_i(t).$$

From Lemma 3.6 with monotonicity of ϕ,

$$\phi\big(\theta'(1/m) - \eta t\,\delta_{\max}\big) \leq y_i(t) \leq \phi\big(\theta'(1/k) - \eta t\,\delta_{\min}\big).$$

Therefore

$$\sum_{i \notin \mathrm{br}(\hat{x})} \delta_i\, y_i(t) \;\geq\; \delta_{\min}(m-k)\,\phi\big(\theta'(1/m) - \eta t\,\delta_{\max}\big) =: \underline{g}(t), \tag{34}$$

$$\sum_{i \notin \mathrm{br}(\hat{x})} \delta_i\, y_i(t) \;\leq\; \delta_{\max}(m-k)\,\phi\big(\theta'(1/k) - \eta t\,\delta_{\min}\big) =: \overline{g}(t). \tag{35}$$

For the lower bound, let $u = \theta'(1/m) - \eta\delta_{\max} t$. Then $dt = -du/(\eta\delta_{\max})$ and

$$\int_0^T \phi\big(\theta'(1/m) - \eta\delta_{\max} t\big)\, dt = \frac{1}{\eta\delta_{\max}} \int_{\theta'(1/m)-\eta\delta_{\max} T}^{\theta'(1/m)} \phi(u)\, du.$$

From definition of ϕ, we have that $\phi(u) = 0$ for $u \le \theta'(0^+)$, this equals

$$\begin{aligned}\int_0^T \underline{g}(t)\, dt &= \frac{m-k}{\eta}\frac{\delta_{\min}}{\delta_{\max}} \int_{\max\{\theta'(1/m)-\eta\delta_{\max} T,\ \theta'(0^+)\}}^{\theta'(1/m)} \phi(u)\, du \\ &= \frac{m-k}{\eta}\frac{\delta_{\min}}{\delta_{\max}} \Big(\theta^*(\theta'(1/m)) - \theta^*(\max\{\theta'(1/m) - \eta\delta_{\max} T,\ \theta'(0^+)\})\Big), \end{aligned} \tag{36}$$

which yields the stated $\underline{\Delta V}(T)$.

For the upper bound, do the same with $u = \theta'(1/k) - \eta\delta_{\min} t$ to obtain $\overline{\Delta V}(T)$ by integrating $\int_0^T \overline{g}(t)\, dt$ such that

$$\begin{aligned}\int_0^T \overline{g}(t)\, dt &= \frac{m-k}{\eta}\frac{\delta_{\max}}{\delta_{\min}} \int_{\max\{\theta'(1/k)-\eta\delta_{\min} T,\ \theta'(0^+)\}}^{\theta'(1/k)} \phi(u)\, du \\ &= \frac{m-k}{\eta}\frac{\delta_{\max}}{\delta_{\min}} \Big(\theta^*(\theta'(1/k)) - \theta^*(\max\{\theta'(1/k) - \eta\delta_{\min} T,\ \theta'(0^+)\})\Big), \end{aligned} \tag{37}$$

which yields the upper bound. □

Regime Analysis: Steep vs. Non-Steep The asymptotic behavior of surplus is dictated by the regularizer's boundary behavior— leading to two distinct regimes: if $|\theta'(0^+)| < \infty$ (non-steep), suboptimal actions are eliminated in finite time and surplus saturates and if $|\theta'(0^+)| = \infty$ (steep), the learner approaches the boundary only asymptotically, leaving a persistent tail. We provide the formal statement of Theorem 4.2 below, which was described informally in the main text.

Theorem E.2 (Formal Version of Theorem 4.2)**.** *Fix a separable regularizer with kernel θ, step size $\eta > 0$, and a fixed optimizer strategy $\hat{x}$ inducing gaps $\{\delta_i\}$, with $\delta_{\min} \le \delta_i \le \delta_{\max}$ for all $i \notin \mathrm{br}(\hat{x})$. Let $V_{\mathrm{bdry}} := \theta^*(\theta'(0^+)) \in [0, +\infty]$ be the boundary energy. Then, for large horizons $\eta T \to \infty$, $\mathrm{LAG}(T)$ satisfies the following dichotomy.*

- ***(A) Non-steep regularizers (finite boundary slope).*** *If $\theta'(0^+) > -\infty$ (equivalently, $V_{\mathrm{bdry}} < \infty$), then suboptimal actions are eliminated in finite time $T^* = t^*_\theta < \infty$ and surplus* saturates*. In particular, for all $T \ge T^*$,*

$$\mathrm{LAG}(T) \in \frac{m-k}{\eta}\left[\frac{\delta_{\min}}{\delta_{\max}}\Big(V(1/m) - V_{\mathrm{bdry}}\Big),\ \frac{\delta_{\max}}{\delta_{\min}}\Big(V(1/k) - V_{\mathrm{bdry}}\Big)\right].$$

- ***(B) Steep regularizers (infinite boundary slope).*** *If $\theta'(0^+) = -\infty$ (so $V_{\mathrm{bdry}} = 0$ under the normalization $\theta^*(-\infty) = 0$), then the learner approaches the boundary only asymptotically and surplus has an* infinite tail*. As $\eta T \to \infty$,*

$$\mathrm{LAG}(T) \in \frac{m-k}{\eta}\left[\frac{\delta_{\min}}{\delta_{\max}} V\big(1/m\big),\ \frac{\delta_{\max}}{\delta_{\min}} V\big(1/k\big)\right] \pm o(1).$$

Proof. In part (A), assume $\theta'(0^+) > -\infty$. For $\underline{\Delta V}(T)$ in (32), the inner maximum equals $\theta'(0^+)$ whenever $\theta'(1/m) - \eta\,\delta_{\max} T \le \theta'(0^+)$, i.e., for all

$$T \ge T_1 := \frac{\theta'(1/m) - \theta'(0^+)}{\eta\,\delta_{\max}}.$$

Thus, for all $T \ge T_1$,

$$\theta^*\Big(\max\{\theta'(1/m) - \eta\delta_{\max} T,\ \theta'(0^+)\}\Big) = \theta^*\big(\theta'(0^+)\big).$$

Similarly, for $\overline{\Delta V}(T)$ in (33), the maximum equals $\theta'(0^+)$ once

$$T \ge T_2 := \frac{\theta'(1/k) - \theta'(0^+)}{\eta\,\delta_{\min}},$$

so for all $T \geq T_2$,

$$\theta^*\Big(\max\{\theta'(1/k) - \eta\delta_{\min}T,\ \theta'(0^+)\}\Big) \ = \ \theta^*\big(\theta'(0^+)\big).$$

Since $T_2 \geq T_1$, for all $T \geq T_2$ both maxima equal $\theta'(0^+)$, yielding the bounds in part (A).

In part (B), assume $\theta'(0^+) = -\infty$ and $\theta^*(-\infty) = 0$. By Theorem E.1,

$$\underline{\Delta V}(T) = \frac{\delta_{\min}}{\delta_{\max}}\Big(\theta^*\big(\theta'(1/m)\big) - \theta^*\big(\theta'(1/m) - \eta\delta_{\max}T\big)\Big),$$
$$\overline{\Delta V}(T) = \frac{\delta_{\max}}{\delta_{\min}}\Big(\theta^*\big(\theta'(1/k)\big) - \theta^*\big(\theta'(1/k) - \eta\delta_{\min}T\big)\Big).$$

As $\eta T \to \infty$, we have $\theta'(1/m) - \eta\delta_{\max}T$ and $\theta'(1/k) - \eta\delta_{\min}T$ go to $-\infty$. Hence,

$$\theta^*\big(\theta'(1/m) - \eta\delta_{\max}T\big) = o_{\eta T}(1), \qquad \theta^*\big(\theta'(1/k) - \eta\delta_{\min}T\big) = o_{\eta T}(1),$$

which yields the bounds in part (B). □

E.3 Examples of Potential Drops, $\mathrm{VAG}(T)$, $\mathrm{LAG}(T)$

Theorems E.1 and E.2 yield the following takeaways. Define the lower-bound potential endpoint

$$\underline{V_{\text{end}}}(T) \ := \ \theta^*\left(\max\left\{\theta'\Big(\frac{1}{m}\Big) - \eta\,\delta_{\max}\,T,\ \theta'(0^+)\right\}\right).$$

If $\theta'(0^+) = -\infty$ (steep), then the maximum is attained by the first term for all $T \geq 0$. If $\theta'(0^+) > -\infty$ (non-steep), by Lemma 3.6, there exists $T^* = t^*_\theta < \infty$ such that the maximum is attained by the first term for $T \leq T^*$ and by the second term for $T \geq T^*$. The same steep/non-steep dichotomy applies to the corresponding upper-bound terms. Using the kernel choice maps in Example C.4, we now make the bounds in Theorems E.1 and E.2 explicit.

(a) *Negative entropy.* By (21) and $\theta'(y) = 1 + \log y$,

$$\underline{V_{\text{start}}} = \theta^*\Big(\theta'\Big(\frac{1}{m}\Big)\Big) = \frac{1}{m}, \qquad \underline{V_{\text{end}}}(T) = \theta^*\Big(\theta'\Big(\frac{1}{m}\Big) - \eta\delta_{\max}T\Big) = \frac{1}{m}e^{-\eta\delta_{\max}T},$$
$$\overline{V_{\text{start}}} = \theta^*\Big(\theta'\Big(\frac{1}{k}\Big)\Big) = \frac{1}{k}, \qquad \overline{V_{\text{end}}}(T) = \theta^*\Big(\theta'\Big(\frac{1}{k}\Big) - \eta\delta_{\min}T\Big) = \frac{1}{k}e^{-\eta\delta_{\min}T}.$$

Thus,

$$\underline{\Delta V}(T) = \underline{V_{\text{start}}} - \underline{V_{\text{end}}}(T) = \frac{1}{m}\big(1 - e^{-\eta\delta_{\max}T}\big), \quad \overline{\Delta V}(T) = \overline{V_{\text{start}}} - \overline{V_{\text{end}}}(T) = \frac{1}{k}\big(1 - e^{-\eta\delta_{\min}T}\big).$$

By Theorem E.1, bringing back to (Surplus Decomposition), for all $T \geq 0$,

$$\mathrm{SURPLUS}(T) \ \geq \ \mathrm{VAG}(T) \ + \ \frac{m-k}{\eta}\cdot\frac{\delta_{\min}}{\delta_{\max}}\cdot\frac{1}{m}\big(1 - e^{-\eta\delta_{\max}T}\big),$$
$$\mathrm{SURPLUS}(T) \ \leq \ \mathrm{VAG}(T) \ + \ \frac{m-k}{\eta}\cdot\frac{\delta_{\max}}{\delta_{\min}}\cdot\frac{1}{k}\big(1 - e^{-\eta\delta_{\min}T}\big).$$

Moreover, as $\eta T \to \infty$,

$$\mathrm{SURPLUS}(T) \ \geq \ \mathrm{VAG}(T) \ + \ \frac{m-k}{\eta}\cdot\frac{\delta_{\min}}{\delta_{\max}}\cdot\frac{1}{m} \ + \ o_{\eta T}(1),$$
$$\mathrm{SURPLUS}(T) \ \leq \ \mathrm{VAG}(T) \ + \ \frac{m-k}{\eta}\cdot\frac{\delta_{\max}}{\delta_{\min}}\cdot\frac{1}{k} \ + \ o_{\eta T}(1).$$

(b) *Negative Tsallis ($q \in (0,1)$).* By (23) and $\theta'(y) = \frac{q\,y^{q-1}-1}{q-1}$,

$$\underline{V_{\text{start}}} = \theta^*\Big(\theta'\Big(\frac{1}{m}\Big)\Big) = m^{-q}, \quad \underline{V_{\text{end}}}(T) = \theta^*\Big(\theta'\Big(\frac{1}{m}\Big) - \eta\delta_{\max}T\Big) = \Big[m^{1-q} + \frac{1-q}{q}\,\eta T\,\delta_{\max}\Big]^{-\frac{q}{1-q}},$$
$$\overline{V_{\text{start}}} = \theta^*\Big(\theta'\Big(\frac{1}{k}\Big)\Big) = k^{-q}, \quad \overline{V_{\text{end}}}(T) = \theta^*\Big(\theta'\Big(\frac{1}{k}\Big) - \eta\delta_{\min}T\Big) = \Big[k^{1-q} + \frac{1-q}{q}\,\eta T\,\delta_{\min}\Big]^{-\frac{q}{1-q}}.$$

Therefore,

$$\underline{\Delta V}(T) = m^{-q} - \left[m^{1-q} + \frac{1-q}{q}\,\eta T\,\delta_{\max}\right]^{-\frac{q}{1-q}},$$
$$\overline{\Delta V}(T) = k^{-q} - \left[k^{1-q} + \frac{1-q}{q}\,\eta T\,\delta_{\min}\right]^{-\frac{q}{1-q}}.$$

By Theorem E.1, substituting into (Surplus Decomposition), for all $T \geq 0$,

$$\mathrm{SURPLUS}(T) \;\geq\; \mathrm{VAG}(T) \;+\; \frac{m-k}{\eta}\cdot\frac{\delta_{\min}}{\delta_{\max}}\cdot\left(m^{-q} - \left[m^{1-q} + \frac{1-q}{q}\,\eta T\,\delta_{\max}\right]^{-\frac{q}{1-q}}\right),$$
$$\mathrm{SURPLUS}(T) \;\leq\; \mathrm{VAG}(T) \;+\; \frac{m-k}{\eta}\cdot\frac{\delta_{\max}}{\delta_{\min}}\cdot\left(k^{-q} - \left[k^{1-q} + \frac{1-q}{q}\,\eta T\,\delta_{\min}\right]^{-\frac{q}{1-q}}\right).$$

As $\eta T \to \infty$,

$$\mathrm{SURPLUS}(T) \;\geq\; \mathrm{VAG}(T) \;+\; \frac{m-k}{\eta}\cdot\frac{\delta_{\min}}{\delta_{\max}}\cdot m^{-q} \;+\; o_{\eta T}(1),$$
$$\mathrm{SURPLUS}(T) \;\leq\; \mathrm{VAG}(T) \;+\; \frac{m-k}{\eta}\cdot\frac{\delta_{\max}}{\delta_{\min}}\cdot k^{-q} \;+\; o_{\eta T}(1).$$

(c) *Euclidean.* $\theta'(0^+) = 0$ and $\theta^*(z_i) = \frac{1}{2}z_i^2$ by (19). By Lemma 3.6, suboptimal actions are eliminated after $t^*_\theta = \frac{1}{k\,\eta\,\delta_{\min}}$. Accordingly, the endpoints take the clipped forms

$$\underline{V_{\text{end}}}(T) = \theta^*\left(\max\left\{\theta'\Big(\frac{1}{m}\Big) - \eta\delta_{\max}T,\ \theta'(0^+)\right\}\right) = \frac{1}{2}\left(\max\left\{\frac{1}{m} - \eta\delta_{\max}T,\ 0\right\}\right)^2,$$
$$\overline{V_{\text{end}}}(T) = \theta^*\left(\max\left\{\theta'\Big(\frac{1}{k}\Big) - \eta\delta_{\min}T,\ \theta'(0^+)\right\}\right) = \frac{1}{2}\left(\max\left\{\frac{1}{k} - \eta\delta_{\min}T,\ 0\right\}\right)^2.$$

Moreover,

$$\underline{V_{\text{start}}} = \theta^*\Big(\theta'\Big(\frac{1}{m}\Big)\Big) = \frac{1}{2m^2}, \qquad \overline{V_{\text{start}}} = \theta^*\Big(\theta'\Big(\frac{1}{k}\Big)\Big) = \frac{1}{2k^2}.$$

Case 1: $0 \leq T \leq t^*_\theta$. In this regime the maximum is attained by the first term, so

$$\underline{V_{\text{end}}}(T) = \frac{1}{2}\Big(\frac{1}{m} - \eta\delta_{\max}T\Big)^2, \qquad \overline{V_{\text{end}}}(T) = \frac{1}{2}\Big(\frac{1}{k} - \eta\delta_{\min}T\Big)^2,$$

and hence

$$\underline{\Delta V}(T) = \frac{1}{2m^2} - \frac{1}{2}\Big(\frac{1}{m} - \eta\delta_{\max}T\Big)^2, \qquad \overline{\Delta V}(T) = \frac{1}{2k^2} - \frac{1}{2}\Big(\frac{1}{k} - \eta\delta_{\min}T\Big)^2.$$

By Theorem E.1 and substituting into (Surplus Decomposition), for all $0 \leq T \leq t^*_\theta$,

$$\mathrm{SURPLUS}(T) \;\geq\; \mathrm{VAG}(T) \;+\; \frac{m-k}{\eta}\cdot\frac{\delta_{\min}}{\delta_{\max}}\cdot\left(\frac{1}{2m^2} - \frac{1}{2}\Big(\frac{1}{m} - \eta\delta_{\max}T\Big)^2\right),$$
$$\mathrm{SURPLUS}(T) \;\leq\; \mathrm{VAG}(T) \;+\; \frac{m-k}{\eta}\cdot\frac{\delta_{\max}}{\delta_{\min}}\cdot\left(\frac{1}{2k^2} - \frac{1}{2}\Big(\frac{1}{k} - \eta\delta_{\min}T\Big)^2\right).$$

Case 2: $T > t^*_\theta$. In this regime both endpoints clip to the boundary:

$$\underline{V_{\text{bdry}}} = \overline{V_{\text{bdry}}} = \theta^*(0) = 0.$$

By Theorem E.2 and (Surplus Decomposition), for all $T > t^*_\theta$,

$$\mathrm{SURPLUS}(T) \;\geq\; \mathrm{VAG}(T) \;+\; \frac{m-k}{\eta}\cdot\frac{\delta_{\min}}{\delta_{\max}}\cdot\frac{1}{2m^2},$$
$$\mathrm{SURPLUS}(T) \;\leq\; \mathrm{VAG}(T) \;+\; \frac{m-k}{\eta}\cdot\frac{\delta_{\max}}{\delta_{\min}}\cdot\frac{1}{2k^2}.$$

E.4 Surplus in Discrete-Time FTRL Algorithm

While continuous-time dynamics offer clean geometric insights, practical algorithms like FTRL operate in discrete steps. A natural question arises: *Does the inverse-rate law survive after discretization?* We show that the answer is affirmative: the discrete-time surplus is simply the continuous-time surplus plus a bounded error term. Consider the standard discrete-time FTRL update with step size η. The optimizer's cumulative surplus over T rounds is defined as:

$$\mathrm{SURPLUS}^{\mathrm{d}}(T) \;:=\; \sum_{t=0}^{T-1}\big(u_o(\hat{x}, y(t)) - \mathrm{Val}(A)\big) \;=\; \mathrm{VAG}(T) \;+\; \underbrace{\sum_{t=0}^{T-1}\sum_{i\notin \mathrm{br}(\hat{x})} \delta_i\, y_i(t)}_{\mathrm{LAG}^{\mathrm{d}}(T)}, \tag{38}$$

where $\mathrm{LAG}^{\mathrm{d}}(T)$ is the discrete version of the cumulative lag. The following corollary establishes that $\mathrm{LAG}^{\mathrm{d}}(T)$ tracks its continuous counterpart $\mathrm{LAG}(T)$ closely, preserving the $\mathcal{O}((m-k)/\eta)$ scaling.

Corollary E.3 (Discrete-Time Robustness)**.** *Fix a constant optimizer strategy $\hat{x}$ and step size $\eta > 0$. The discrete-time cumulative lag $\mathrm{LAG}^{\mathrm{d}}(T)$ satisfies the continuous-time bounds of Theorem E.1 up to an additive constant:*

$$\frac{m-k}{\eta}\cdot \underline{\Delta V}(T) \;\le\; \mathrm{LAG}^{\mathrm{d}}(T) \;\le\; \frac{m-k}{\eta}\cdot \overline{\Delta V}(T) \;+\; \frac{m-k}{k}\delta_{\max}. \tag{39}$$

Remark 9 (Robustness of Geometry). *This result confirms that the "Inverse-Rate Law" is the dominant force in surplus of FTRL methodology. The discretization error is merely $\mathcal{O}(1)$, whereas the geometric cost scales with $1/\eta$. Therefore, the classification of regularizers into **Steep** (infinite tail) and **Non-Steep** (finite saturation) derived in Theorem E.2 applies essentially unchanged to the discrete FTRL algorithm.*

Proof. From Lemma 3.6, for every $j \notin \mathrm{br}(\hat{x})$ and every integer $t \ge 0$,

$$\phi\big(\theta'(1/m) - \eta t\,\delta_j\big) \;\le\; y_j(t) \;\le\; \phi\big(\theta'(1/k) - \eta t\,\delta_j\big).$$

Following the proof of Theorem E.1. Recall $\underline{g}(t)$ and $\overline{g}(t)$ be as in (34) and (35). Since $t \mapsto \theta'(1/m) - \eta t\,\delta_j$ and $t \mapsto \theta'(1/k) - \eta t\,\delta_j$ are decreasing and ϕ is increasing, both $\underline{g}$ and $\overline{g}$ are non-increasing. Hence, for all $T \ge 1$,

$$\sum_{t=0}^{T-1}\underline{g}(t) \;\ge\; \int_0^T \underline{g}(t)\,dt, \qquad \sum_{t=0}^{T-1}\overline{g}(t) \;\le\; \overline{g}(0) + \int_0^T \overline{g}(t)\,dt \le \frac{m-k}{k}\,\delta_{\max} + \int_0^T \overline{g}(t)\,dt.$$

Therefore, we have

$$\mathrm{VAG}(T) + \int_0^T \underline{g}(t)\,dt \le \mathrm{SURPLUS}^{\mathrm{d}}(T) \le \mathrm{VAG}(T) + \frac{m-k}{k}\,\delta_{\max} + \int_0^T \overline{g}(t)\,dt.$$

Then the rest of the proof follows the same steps as in Theorem E.1 to evaluate the integrals and cancel $\mathrm{VAG}(T)$. □

F Details omitted from Section 4.2 (The Alternating Trap)

Rather than committing to a fixed strategy $\hat{x}$, can the optimizer choose an $x_d(t)$ that achieves significantly larger reward against FTRL over T rounds than any fixed $\hat{x}$? In this appendix we answer this question for random zero-sum games and any strongly convex separable regularizer. Specifically, we consider random games with payoff matrix $A \in \mathbb{R}^{n\times m}$ whose entries are i.i.d. with $A_{ij} \sim \mathrm{Unif}[-1,1]$. We show that, with high probability over A, there exists an optimizer strategy $x_d(t)$ that guarantees a per-round surplus $\Omega\big(\eta/\mathrm{poly}(nm)\big)$ uniformly over all horizons T, against FTRL with any separable regularizer on Δ^m for some sufficiently small step size $\eta > 0$.

We restate Theorem 4.3 in the main text.

Theorem F.1. *Let $A \in \mathbb{R}^{n\times m}$ have i.i.d.* $\mathrm{Unif}[-1,1]$ *entries. Fix any $\delta \in (0, 1/m)$, α-strongly convex separable regularizer* h *with curvature bound $M = \sup_{a\in[\delta,1-\delta]} \theta''(a)$, and step size*

$$\eta \;\le\; \frac{\alpha}{c_{\|\cdot\|}\sup_{u\in[-1,1]^m}\|u\|_*}\Big(\frac{1}{m}-\delta\Big).$$

Then with probability at least

$$1 - \frac{n!m!}{(n+m-1)!} - \frac{1}{nm},$$

there exists an alternating trap strategy $x_d(t)$ *such that for every horizon T,*

$$\mathrm{R}_{\mathrm{disc}}(x_d(t),\,0,\,T) \;\ge\; T\,\mathrm{Val}(A) \;+\; \eta T\cdot C_A^{\mathrm{h}}, \qquad C_A^{\mathrm{h}} \;\coloneqq\; \frac{1}{2M(nm)^6}.$$

Remark. This stepsize scaling is standard for a gradient method: $\eta = O(\mu/L)$ (strong convexity over smoothness).

We prove Theorem F.1 by following the same sequence of steps as the proof sketch in Section 4.2.

Step 1: Two high-probability events on A. We condition on two events that hold with high probability under the i.i.d. $\mathrm{Unif}[-1,1]$ model. The first is the non-identification-on-support event $\mathcal{A}_{nm}$ from Assumption 1, whose probability is lower bounded in Lemma F.2. The second is the entry-gap anti-degeneracy event $\mathcal{E}_{\text{gap}}$, proved in Lemma F.5.

Non-identification-on-support event We rely on the following structural event, adapted from Assos et al. [5, Assumption 1]. It ensures that, at some max-min optimizer strategy, the learner has at least two best responses that can be distinguished on the optimizer's support.

Assumption 1. There exists a max-min strategy $x^\star$ for the optimizer and two learner best responses $i, j \in \mathrm{br}(x^\star)$ such that they do not identify on $\mathrm{supp}(x^\star)$. Equivalently, there exists a support $\ell \in \mathrm{supp}(x^\star)$ such that $e_\ell^\top A e_i = A_{\ell i} \neq A_{\ell j} = e_\ell^\top A e_j$.

Define the event $\mathcal{A}_{nm} := \{A :$ Assumption 1 holds$\}$. The next lemma shows that this event holds with high probability for random matrix games.

Lemma F.2. *Let $A \in \mathbb{R}^{n\times m}$ have i.i.d. continuous entries $A_{ij} \sim \mathrm{Unif}[-1,1]$. Then*

$$\mathbb{P}\big[\textit{Assumption 1 holds}\big] \;\ge\; 1 - \frac{n!\,m!}{(n+m-1)!}.$$

The assumption holds with high probability as n, m grow.

Proof. Define the events

$$\Pi_{nm} := \{A :\ \text{the game has no pure Nash equilibrium}\},$$
$$\mathcal{D}_{nm} := \{A :\ \exists (i,j) \neq (i',j') \text{ with } A_{ij} = A_{i'j'}\}.$$

Claim F.3. $\mathbb{P}(\mathcal{D}_{nm}) = 0$.

Proof. Since the entries are i.i.d. continuous, for any fixed distinct index pairs $(i,j) \neq (i',j')$ we have $\mathbb{P}(A_{ij} = A_{i'j'}) = 0$. Taking a finite union over all such pairs yields $\mathbb{P}(\mathcal{D}_{nm}) = 0$. □

Claim F.4. $\Pi_{nm} \cap \mathcal{D}_{nm}^c \subseteq \mathcal{A}_{nm}$.

Proof. Fix $A \in \Pi_{nm} \cap \mathcal{D}_{nm}^c$ and let (x, y) be any Nash equilibrium.

We claim y cannot be pure. Suppose for contradiction that $y = e_j$. Because $A \in \mathcal{D}_{nm}^c$, the entries in column j are all distinct, hence the maximizing row $i^\star = \arg\max_i A_{ij}$ is unique. Therefore any best response of the row player to $y = e_j$ must be $x = e_{i^\star}$. Since (x, y) is a Nash equilibrium, $y = e_j$

must also be a best response to $x = e_{i^\star}$. Thus $(e_{i^\star}, e_j)$ is a pure Nash equilibrium, contradicting $A \in \Pi_{nm}$. Hence y is mixed and $|\operatorname{supp}(y)| \geq 2$.

Pick two distinct indices $i_1 \neq i_2$ in $\operatorname{supp}(y)$. Then $e_{i_1}, e_{i_2} \in \operatorname{br}(x)$. Choose any $k \in \operatorname{supp}(x)$. Because $A \in \mathcal{D}_{nm}^c$, all entries of A are distinct, so in particular $A_{ki_1} \neq A_{ki_2}$. Thus the two best responses e_{i_1} and e_{i_2} do not identify on $\operatorname{supp}(x)$, which is exactly Assumption 1. Hence $A \in \mathcal{A}_{nm}$, proving $\Pi_{nm} \cap \mathcal{D}_{nm}^c \subseteq \mathcal{A}_{nm}$. □

A classical formula for i.i.d. continuous payoff matrices (see Karlin [17], p. 79) gives

$$\mathbb{P}(\text{there exists a pure Nash equilibrium}) = \frac{n!\,m!}{(n+m-1)!}, \qquad \text{so } \mathbb{P}(\Pi_{nm}) = 1 - \frac{n!\,m!}{(n+m-1)!}.$$

Using Claim F.4 $\Pi_{nm} \cap \mathcal{D}_{nm}^c \subseteq \mathcal{A}_{nm}$ and Claim F.3 $\mathbb{P}(\mathcal{D}_{nm}) = 0$,

$$\mathbb{P}(\mathcal{A}_{nm}) \geq \mathbb{P}(\Pi_{nm} \cap \mathcal{D}_{nm}^c) = \mathbb{P}(\Pi_{nm}) - \mathbb{P}(\Pi_{nm} \cap \mathcal{D}_{nm}) \geq \mathbb{P}(\Pi_{nm}) - \mathbb{P}(\mathcal{D}_{nm}) = \mathbb{P}(\Pi_{nm}),$$

which yields the stated lower bound. □

Anti-degeneracy gap event for entries We isolate an anti-degeneracy event ensuring that no two entries in the same row are "too close". This lets us lower bound the distinguishing gap $A_{\ell i} - A_{\ell j}$ uniformly whenever $A_{\ell i} > A_{\ell j}$.

Lemma F.5. *Let $A \in \mathbb{R}^{n \times m}$ have i.i.d. entries $A_{\ell i} \sim \operatorname{Unif}[-1, 1]$. Define*

$$\gamma := \frac{2}{n^2 m^3}, \qquad \mathcal{E}_{\text{gap}} := \left\{ A : \min_{\ell \in [n]} \min_{\substack{i,j \in [m] \\ i \neq j}} |A_{\ell i} - A_{\ell j}| \geq \gamma \right\}.$$

Then $\mathbb{P}(\mathcal{E}_{\text{gap}}) \geq 1 - \frac{1}{nm}$.

Proof. Fix $\ell \in [n]$ and $i < j$ and set $X := A_{\ell i}$, $Y := A_{\ell j}$. Then X, Y are independent $\operatorname{Unif}[-1, 1]$. Consider the event $\mathcal{E}_{\ell ij} := \{|X - Y| < \gamma\}$. Conditioning on $Y = y$,

$$\mathbb{P}(\mathcal{E}_{\ell ij} \mid Y = y) = \mathbb{P}(y - \gamma < X < y + \gamma \mid Y = y) \leq \gamma,$$

since X is uniform on an interval of length 2 and $(y - \gamma, y + \gamma) \cap [-1, 1]$ has length at most 2γ. Taking expectation over Y gives $\mathbb{P}(\mathcal{E}_{\ell ij}) \leq \gamma$.

There are at most $n\binom{m}{2} \leq n \cdot \frac{m^2}{2}$ unordered triples (ℓ, i, j) with $i < j$. By a union bound,

$$\mathbb{P}(\exists (\ell, i < j) : |A_{\ell i} - A_{\ell j}| < \gamma) \leq n \binom{m}{2} \cdot \gamma \leq n \cdot \frac{m^2}{2} \cdot \gamma = \frac{1}{nm}.$$

Therefore, with probability at least $1 - \frac{1}{nm}$, for all $\ell \in [n]$ and all $i \neq j$, $|A_{\ell i} - A_{\ell j}| \geq \gamma$, i.e., $\mathcal{E}_{\text{gap}}$ holds. □

Define the good event $\mathcal{E} := \mathcal{A}_{nm} \cap \mathcal{E}_{\text{gap}}$. By Lemmas F.2 and F.5, we have the union bound

$$\mathbb{P}(\mathcal{E}) = 1 - \mathbb{P}(\mathcal{E}^c) \geq 1 - \mathbb{P}(\mathcal{E}^c) \geq 1 - \mathbb{P}({\mathcal{A}_{nm}}^c) - \mathbb{P}({\mathcal{E}_{\text{gap}}}^c) \geq 1 - \frac{n!\,m!}{(n+m-1)!} - \frac{1}{nm},$$

which is the stated success probability in Theorem F.1.

Step 2: An alternating perturbation construction. We assume $A \in \mathcal{E}$. On the non-identification event $\mathcal{A}_{nm}$, fix an optimizer max-min strategy $x^\star$ and two distinct learner best responses $i, j \in \operatorname{br}(x^\star)$. By definition of $\mathcal{A}_{nm}$, there exists a support $\ell \in \operatorname{supp}(x^\star)$ such that $A_{\ell i} \neq A_{\ell j}$. On the gap event $\mathcal{E}_{\text{gap}}$, this separation is uniformly bounded away from 0 on the support: for every $\ell \in \operatorname{supp}(x^\star)$, $|A_{\ell i} - A_{\ell j}| \geq \gamma$. Let's pick $\ell \in \operatorname{supp}(x^\star)$ be a maximizer of $x^\star$ over its support, i.e., ${x^\star}_\ell = \max_{p \in \operatorname{supp}(x^\star)} {x^\star}_p$. Define two perturbed strategies

$$x' := (1 - {x^\star}_\ell) x^\star + {x^\star}_\ell e_\ell, \qquad x'' := (1 + {x^\star}_\ell) x^\star - {x^\star}_\ell e_\ell. \tag{40}$$

Since x' is a convex combination of $x^\star$ and e_ℓ, we have $x' \in \Delta^n$. For x'', note that $x''_\ell = ({x^\star}_\ell)^2 \geq 0$, $x''_i = (1 + {x^\star}_\ell) {x^\star}_i \geq 0$ for $i \neq \ell$, and $\sum_i x''_i = 1$. Hence $x'' \in \Delta^n$. The next lemma shows some useful properties of x' and x''.

Lemma F.6. *The construction (40) has the following properties:*

1. $(x' + x'')/2 = x^\star$.
2. $x'^\top A e_i > x'^\top A e_j$ *and* $x''^\top A e_i < x''^\top A e_j$.

Proof. *1.* Expanding the definition of x' and x'',

$$\frac{x' + x''}{2} = \frac{\big((1 - x^\star{}_\ell)x^\star + x^\star{}_\ell e_\ell\big) + \big((1 + x^\star{}_\ell)x^\star - x^\star{}_\ell e_\ell\big)}{2} = x^\star.$$

2. By Assumption 1, $A_{\ell i} \neq A_{\ell j}$. WLOG, swap i and j so that $A_{\ell i} > A_{\ell j}$. Since $i, j \in \mathrm{br}(x^\star)$, we have $x^{\star\top} A e_i = x^{\star\top} A e_j$. Expanding,

$$x'^\top A e_i = \big((1 - x^\star{}_\ell)x^\star + x^\star{}_\ell e_\ell\big)^\top A e_i = (1 - x^\star{}_\ell)\, x^{\star\top} A e_i + x^\star{}_\ell A_{\ell i},$$

and similarly $x''^\top A e_j = (1 + x^\star{}_\ell)\, x^{\star\top} A e_j - x^\star{}_\ell A_{\ell j}$. Subtracting and using $x^{\star\top} A e_i = x^{\star\top} A e_j$,

$$x'^\top A e_i - x'^\top A e_j = x^\star{}_\ell (A_{\ell i} - A_{\ell j}) > 0.$$

The same expansion yields

$$x''^\top A e_i - x''^\top A e_j = -x^\star{}_\ell (A_{\ell i} - A_{\ell j}) < 0,$$

so $x'^\top A e_i > x'^\top A e_j$ and $x''^\top A e_i < x''^\top A e_j$. □

We use x' and x'' to define an alternating optimizer strategy whose time-average equals the max-min strategy $x^\star$ while flipping the relative advantage between i and j. Define

$$x_d(t) := \begin{cases} x', & t \text{ even}, \\ x'', & t \text{ odd}. \end{cases} \tag{41}$$

Recall that learner plays FTRL with (Choice Map)

$$y_t = Q_{\mathrm{h}}\Big(\eta \sum_{\tau=0}^{t-1} B^\top x_d(\tau)\Big) = \arg\max_{y \in \Delta^m} \Big\{ \Big\langle \eta \sum_{\tau=0}^{t-1} B^\top x_d(\tau),\, y \Big\rangle - \mathrm{h}(y) \Big\}.$$

Define $v' = A^\top x'$ and $v = A^\top x^\star$. Since $(x' + x'')/2 = x^\star$, we have $A^\top x'' = 2v - v'$. For $t = 2s$, using $A = -B$,

$$y_{2s} = Q_{\mathrm{h}}\big(-2\eta s\, v\big), \qquad y_{2s+1} = Q_{\mathrm{h}}\big(-2\eta s\, v - \eta v'\big).$$

Over the pair of rounds $(2s, 2s+1)$, the optimizer's payoff is

$$x'^\top A y_{2s} + x''^\top A y_{2s+1} = \langle v', y_{2s} \rangle + \langle 2v - v', y_{2s+1} \rangle.$$

Since $x^\star$ is max-min, we have $\langle v, y \rangle = x^{\star\top} A y \geq \mathrm{Val}(A)$ for all $y \in \Delta^m$, and in particular $\langle 2v, y_{2s+1} \rangle \geq 2\mathrm{Val}(A)$. Therefore,

$$x'^\top A y_{2s} + x''^\top A y_{2s+1} \geq 2\mathrm{Val}(A) + \langle v', y_{2s} \rangle - \langle v', y_{2s+1} \rangle. \tag{42}$$

Step 3. Interpolation and a variance identity. Fix an integer $s \geq 0$ and define the interpolation

$$y(r) := Q_{\mathrm{h}}\big(-2\eta s\, v - r\, \eta\, v'\big), \qquad r \in [0, 1]. \tag{Interpolation}$$

Claim F.7. $y(\cdot)$ *is* $(\eta \|v'\|_* / \alpha)$*-Lipschitz on* $[0, 1]$.

Proof. By Proposition C.1, since h is α-strongly convex with respect to $\|\cdot\|$, the map Q_{h} is $1/\alpha$-Lipschitz. Hence, for any $r_1, r_2 \in [0, 1]$,

$$\|y(r_1) - y(r_2)\| \leq \frac{1}{\alpha} \|(r_1 - r_2)\eta v'\| = \frac{\eta \|v'\|_*}{\alpha} |r_1 - r_2|.$$

Thus $y(\cdot)$ is $(\eta \|v'\|_* / \alpha)$-Lipschitz on $[0, 1]$. □

Therefore, $y(\cdot)$ is absolutely continuous and differentiable for almost every $r \in [0,1]$ by Rademacher's theorem. Consequently, $\langle v', y(\cdot)\rangle$ is absolutely continuous, and the fundamental theorem of calculus yields

$$\langle v', y_{2s}\rangle - \langle v', y_{2s+1}\rangle = \langle v', y(0)\rangle - \langle v', y(1)\rangle = -\int_0^1 \frac{d}{dr}\langle v', y(r)\rangle \, dr. \tag{43}$$

Lemma F.8. *Fix $s \geq 0$ and define $y(r)$ as in (Interpolation). For each $r \in [0,1]$, let*

$$w_r(i) \coloneqq \begin{cases} \dfrac{1}{\theta''(y_i(r))}, & i \in \operatorname{supp}(y(r)), \\ 0, & \textit{otherwise}, \end{cases} \qquad W_r \coloneqq \sum_{i=1}^m w_r(i), \qquad \pi_r(i) \coloneqq \frac{w_r(i)}{W_r}. \tag{44}$$

Let $I_r \sim \pi_r$ and define $Z_r \coloneqq v'_{I_r}$. Then, for almost every $r \in [0,1]$,

$$\frac{d}{dr}\,\langle v', y(r)\rangle \;=\; -\eta\, W_r\, \operatorname{Var}(Z_r). \tag{45}$$

Moreover, for any $i, j \in \operatorname{supp}(y(r))$ and $i \neq j$, we have

$$W_r\, \operatorname{Var}(Z_r) \;\geq\; \frac{(v'_i - v'_j)^2}{\theta''(y_i(r)) + \theta''(y_j(r))}. \tag{46}$$

Proof. Fix $r \in [0,1]$ at which $y(\cdot)$ is differentiable. The KKT conditions (1) for Q_{h} imply that there exists a scalar $\lambda(r)$ such that for every $i \in \operatorname{supp}(y(r))$,

$$\theta'(y_i(r)) = \lambda(r) - 2\eta s\, v_i - r\eta\, v'_i.$$

Differentiating with respect to r yields, for every $i \in \operatorname{supp}(y(r))$,

$$\theta''(y_i(r))\,\frac{d}{dr}y_i(r) = \frac{d}{dr}\lambda(r) - \eta v'_i, \qquad \text{equivalently } \frac{d}{dr}y_i(r) = \frac{\frac{d}{dr}\lambda(r) - \eta v'_i}{\theta''(y_i(r))}. \tag{47}$$

For $i \notin \operatorname{supp}(y(r))$, we have $y_i(r) = 0$, and our definition set $w_r(i) = 0$. Therefore, the identity

$$\frac{d}{dr}y_i(r) = w_r(i)\big(\frac{d}{dr}\lambda(r) - \eta v'_i\big)$$

holds for all $i \in [m]$ at this r.

Since $y(r) \in \Delta^m$ for all r, we have $\sum_i y_i(r) = 1$ and thus $\sum_i \frac{d}{dr}y_i(r) = 0$. Using (47) and $w_r(i) = 1/\theta''(y_i(r))$,

$$0 = \sum_{i=1}^m \frac{d}{dr}y_i(r) = \sum_{i=1}^m w_r(i)\big(\lambda'(r) - \eta\, v'_i\big),$$

so

$$\frac{d}{dr}\lambda(r) = \eta\,\frac{\sum_{i=1}^m w_r(i)\, v'_i}{\sum_{i=1}^m w_r(i)} = \eta \sum_{i=1}^m \pi_r(i)\, v'_i = \eta\, \mathbb{E}[Z_r].$$

Substituting back gives

$$\frac{d}{dr}y_i(r) = \eta\, w_r(i)\big(\mathbb{E}[Z_r] - v'_i\big).$$

Therefore,

$$\frac{d}{dr}\langle v', y(r)\rangle = \sum_{i=1}^m v'_i \frac{d}{dr}y_i(r) = \eta \sum_{i=1}^m v'_i w_r(i)\big(\mathbb{E}[Z_r] - v'_i\big) = -\eta \sum_{i=1}^m w_r(i)\big(v'_i - \mathbb{E}[Z_r]\big)^2.$$

Since

$$\sum_i w_r(i)\big(v'_i - \mathbb{E}[Z_r]\big)^2 = W_r \sum_i \pi_r(i)\big(v'_i - \mathbb{E}[Z_r]\big)^2 = W_r \operatorname{Var}(Z_r),$$

we obtain (45).

Fix $i \neq j$ and $i, j \in \mathrm{supp}(y(r))$, and let $E_{ij} = \{I_r \in \{i, j\}\}$ with indicator $\mathbf{1}_{E_{ij}}$. By the law of total variance,

$$\mathrm{Var}(Z_r) \;\geq\; \mathbb{E}[\mathrm{Var}(Z_r \mid \mathbf{1}_{E_{ij}})] \;\geq\; \mathbb{P}(E_{ij})\ \mathrm{Var}(Z_r \mid E_{ij}).$$

Moreover, $\mathbb{P}(E_{ij}) = \pi_r(i) + \pi_r(j)$ and, conditional on E_{ij}, Z_r takes values v'_i and v'_j with probabilities $\pi_r(i)/(\pi_r(i) + \pi_r(j))$ and $\pi_r(j)/(\pi_r(i) + \pi_r(j))$. Hence

$$\mathrm{Var}(Z_r \mid E_{ij}) = \frac{\pi_r(i)\pi_r(j)}{(\pi_r(i) + \pi_r(j))^2}\,(v'_i - v'_j)^2,$$

and therefore

$$\mathrm{Var}(Z_r) \;\geq\; \frac{\pi_r(i)\,\pi_r(j)}{\pi_r(i) + \pi_r(j)}\,(v'_i - v'_j)^2.$$

Multiplying by W_r and using $\pi_r(i) = w_r(i)/W_r$ gives

$$W_r\,\mathrm{Var}(Z_r) \geq W_r \cdot \frac{\pi_r(i)\pi_r(j)}{\pi_r(i) + \pi_r(j)}\,(v'_i - v'_j)^2 = \frac{w_r(i)w_r(j)}{w_r(i) + w_r(j)}\,(v'_i - v'_j)^2.$$

Substituting $w_r(i) = 1/\theta''(y_i(r))$ yields (46). □

Step 4. Uniform bound of curvature. We next show that along the interpolation path $y(r)$, the coordinates corresponding to best-response indices remain uniformly lower bounded away from 0 for all $r \in [0, 1]$. This yields a uniform upper bound on the curvature $\theta''(y_p(r))$ for all $p \in \mathrm{br}(x^\star)$ along the path. In particular, applying Lemma F.8 to the two best-response indices $i, j \in \mathrm{br}(x^\star)$ from Assumption 1, integrating (45) over $r \in [0, 1]$ and using (46) gives

$$\langle v', y_{2s}\rangle - \langle v', y_{2s+1}\rangle \geq \eta\,(v'_i - v'_j)^2 \int_0^1 \frac{dr}{\theta''(y_i(r)) + \theta''(y_j(r))}. \tag{48}$$

It therefore remains to show that the best-response coordinates $y_i(r)$ and $y_j(r)$ remain uniformly lower bounded away from 0 for all $r \in [0, 1]$. The next lemma establishes this bound for all $p \in \mathrm{br}(x^\star)$.

Lemma F.9. *Fix an integer $s \geq 0$ and a max-min strategy $x^\star$. Assume that h is α-strongly convex on Δ^m with respect to $\|\cdot\|$. Define $y(r)$ as in (Interpolation). Then, for all $r \in [0, 1]$ and all $p \in \mathrm{br}(x^\star)$,*

$$y_p(r) \;\geq\; \max\Big\{0,\ \frac{1}{m} - \frac{\eta}{\alpha}\,\|v'\|_*\Big\}. \tag{49}$$

In particular, for any $\delta \in (0, 1/m)$, if

$$0 \;<\; \eta \;\leq\; \frac{\alpha}{\|v'\|_*}\Big(\frac{1}{m} - \delta\Big), \tag{50}$$

then $y_p(r) \geq \delta$ for all $r \in [0, 1]$ and all $p \in \mathrm{br}(x^\star)$.

Proof. Recall (Interpolation) for any fixed $s \geq 0$

$$y(r) = Q_{\mathrm{h}}\big(-2\eta s\, v - r\,\eta v'\big), \qquad r \in [0, 1].$$

At even rounds $2s$, $y(0) = Q_{\mathrm{h}}\big(-2\eta s\, v\big)$, which is the FTRL response to a constant max-min strategy. By Lemma 3.6, we have the uniform lower bound $y_p(0) \geq 1/m$ for all $p \in \mathrm{br}(x^\star)$.

We have shown that $y(\cdot)$ is Lipschitz with constant $\eta\,\|v'\|_*/\alpha$ from Claim F.7. Therefore,

$$\|y(r) - y(0)\| = \|Q_{\mathrm{h}}\big(-2\eta s\, v\big) - Q_{\mathrm{h}}\big(-2\eta s\, v - r\eta v'\big)\| \leq \frac{r\eta}{\alpha}\|v'\|_*.$$

Let $c_{\|\cdot\|} := \sup_{u \neq 0}(\|u\|_\infty / \|u\|)$, so that $\|u\|_\infty \leq c_{\|\cdot\|}\,\|u\|$ for all $u \in \mathbb{R}^m$. For any coordinate $i \in [m]$,

$$y_i(r) \geq [y_i(0) - \|y(r) - y(0)\|_\infty]_+ \geq \Big[y_i(0) - c_{\|\cdot\|} \cdot \frac{\eta}{\alpha}\,\|v'\|_*\Big]_+,$$

and therefore for all $p \in \mathrm{br}(x^\star)$

$$y_p(r) \geq \frac{1}{m} - c_{\|\cdot\|} \cdot \frac{\eta}{\alpha}\,\|v'\|_*.$$

To guarantee $y_p(r) \geq \delta$ for all $r \in [0,1]$ and all $p \in \mathrm{br}(x^\star)$, it suffices that

$$\frac{1}{m} - c_{\|\cdot\|} \cdot \frac{\eta}{\alpha} \|v'\|_* \geq \delta \qquad \Longleftrightarrow \qquad \eta \leq \frac{\alpha}{c_{\|\cdot\|}\|v'\|_*}\Big(\frac{1}{m} - \delta\Big).$$

□

Moreover, since $A \in [-1,1]^{n\times m}$ and $x' \in \Delta^n$, each coordinate of $v' = A^\top x'$ lies in $[-1,1]$. Hence $v' \in [-1,1]^m$ and

$$\|v'\|_* \leq \sup_{u\in[-1,1]^m} \|u\|_*.$$

Therefore, it suffices that

$$\eta \leq \frac{\alpha}{c_{\|\cdot\|} \sup_{u\in[-1,1]^m}\|u\|_*}\Big(\frac{1}{m} - \delta\Big),$$

which is the step size condition in Theorem 4.3.

With Lemma F.9, we can now uniformly upper bound the curvature terms $\theta''(y_i(r))$ and $\theta''(y_j(r))$ in the integral.

Corollary F.10. *Under the same assumptions as in Lemma F.9, fix any $s > 0$ and $\delta \in (0, 1/m)$, and choose*

$$\eta \in \Big(0,\ \frac{\alpha}{c_{\|\cdot\|}\|v'\|_*}\Big(\frac{1}{m} - \delta\Big)\Big].$$

Let $y(r)$ be defined in (Interpolation). Then, for all $r \in [0,1]$ and all $p \in \mathrm{br}(x^\star)$,

$$0 \ < \ \theta''(y_p(r)) \ \leq \ \sup_{t\in[\delta,\,1-\delta]} \theta''(t) \ \coloneqq \ M \ < \ \infty. \tag{51}$$

Proof. By Lemma F.9, for all $r \in [0,1]$ and all $p \in \mathrm{br}(x^\star)$, we have $y_p(r) \geq \delta$. By Assumption 1, there exists $q \in \mathrm{br}(x^\star)$ with $q \neq p$. Since $y(r) \in \Delta^m$, we have $y_p(r) + y_q(r) \leq 1$, hence $y_p(r) \leq 1 - \delta$ and $y_q(r) \leq 1 - \delta$. Therefore $y_p(r) \in [\delta, 1-\delta]$ for all $r \in [0,1]$ and all $p \in \mathrm{br}(x^\star)$. Because θ'' is continuous and strongly convex on $(0,1]$, it is bounded on the compact interval $[\delta, 1-\delta]$, so (51) holds with

$$0 < M = \sup_{t\in[\delta,1-\delta]} \theta''(t) < \infty.$$

□

Fix any $\delta \in (0, 1/m)$ and choose any $\eta \in \Big(0,\ \frac{\alpha}{\|v'\|_*}\Big(\frac{1}{m} - \delta\Big)\Big]$. By Corollary F.10, for all $r \in [0,1]$ and any $i, j \in \mathrm{br}(x^\star)$,

$$\theta''(y_i(r)) + \theta''(y_j(r)) \leq 2M \qquad \text{for all } r \in [0,1].$$

Hence,

$$\int_0^1 \frac{dr}{\theta''(y_i(r)) + \theta''(y_j(r))} \ \geq \ \int_0^1 \frac{dr}{2M} \ = \ \frac{1}{2M}.$$

Plugging this into (48) yields the one-step bound

$$\langle v', y_{2s}\rangle - \langle v', y_{2s+1}\rangle \ \geq \ \eta\,(v'_i - v'_j)^2 \int_0^1 \frac{dr}{\theta''(y_i(r)) + \theta''(y_j(r))} \ \geq \ \frac{\eta}{2M}\,(v'_i - v'_j)^2.$$

Step 5. Sum the uniform one-step lower bound With our construction of $x_d(t)$ in (40), we know that $x^\star{}_\ell = \max_{a_\ell \in \mathrm{supp}(x^\star)} x^\star{}_\ell$. Since $x^\star$ is a probability distribution over n actions, we have $x^\star{}_\ell \geq 1/|\mathrm{supp}(x^\star)| \geq 1/n$. Since $A \in \mathcal{E}_{\mathrm{gap}}$, we have $|A_{\ell i} - A_{\ell j}| \geq \gamma = \frac{2}{n^2m^3}$. Since for all $i \in [m]$, $v'_i = (A^\top x')_i = (1 - x^\star{}_\ell)(A^\top x^\star)_i + x^\star{}_\ell A_{\ell i}$ and $i, j \in \mathrm{br}(x^\star)$, we have

$$\begin{aligned} v'_i - v'_j = (1 - x^\star{}_\ell)\big((A^\top x^\star)_i - (A^\top x^\star)_j\big) + x^\star{}_\ell(A_{\ell i} - A_{\ell j}) &= x^\star{}_\ell(A_{\ell i} - A_{\ell j}) \\ &\geq \frac{1}{n} \cdot \frac{2}{n^2m^3} = \frac{2}{n^3m^3}. \end{aligned}$$

Combining all the pieces, for every $s \geq 0$, we have each two-round payoff bounded as

$$\begin{aligned} x'^{\top} A y_{2s} + x''^{\top} A y_{2s+1} &\geq 2\mathrm{Val}(A) + \langle v', y_{2s} \rangle - \langle v', y_{2s+1} \rangle \\ &\geq 2\mathrm{Val}(A) + \frac{\eta}{2M}(v'_i - v'_j)^2 \\ &\geq 2\mathrm{Val}(A) + \frac{\eta}{2M} \cdot \left(\frac{2}{n^3 m^3}\right)^2 \\ &= 2\mathrm{Val}(A) + \frac{2\eta}{M n^6 m^6}. \end{aligned}$$

For even T, the optimizer's total reward satisfies

$$\begin{aligned} \sum_{t=0}^{T-1} x_d(t)^{\top} A y_t &= \sum_{s=0}^{T/2-1} \left(x'^{\top} A y_{2s} + x''^{\top} A y_{2s+1} \right) \\ &\geq \frac{T}{2}\left(2\mathrm{Val}(A) + \frac{2\eta}{M n^6 m^6}\right) \\ &= T\mathrm{Val}(A) + \frac{\eta T}{M n^6 m^6}. \end{aligned} \tag{52}$$

For odd T, set the last-round optimizer strategy to $x^\star$, which guarantees payoff at least $\mathrm{Val}(A)$. Then

$$\begin{aligned} \sum_{t=0}^{T-1} x_d(t)^{\top} A y_t &= \sum_{s=0}^{(T-1)/2-1} \left(x'^{\top} A y_{2s} + x''^{\top} A y_{2s+1} \right) + x^{\star\top} A y_{T-1} \\ &\geq \frac{T-1}{2}\left(2\mathrm{Val}(A) + \frac{2\eta}{M n^6 m^6}\right) + \mathrm{Val}(A) \\ &= T\mathrm{Val}(A) + \frac{\eta(T-1)}{M n^6 m^6}. \end{aligned} \tag{53}$$

This completes the proof of Theorem F.1.

G Details Omitted from Section 4.3 (The Price of ε-Best-Response)

We conclude our work with a natural question about no-regret algorithmic design:

"How much surplus must the learner inevitably surrender simply to identify her approximate ε-best response?"

We formalize the minimum such fee from learner's perspective as the Price of Best-Response.

Definition 9. Let $\gamma(\eta, t) = \|y(t) - y^*_{x^*}\|_1$ be the learner's distance from the optimal strategy, and let $\mathrm{SURPLUS}_{\mathrm{h}}^{(\eta,t)}$ be the cumulative surplus up to time t. We define the *cost curve* $\mathcal{C}(\varepsilon)$ as:

$$\mathcal{C}(\varepsilon) \;:=\; \inf_{\eta,t} \sup_{x_d(\cdot),A} \left\{ \mathrm{SURPLUS}_{\mathrm{h}}^{(\eta,t)} \;\middle|\; \gamma(\eta,t) \leq \varepsilon \right\}. \qquad \text{(Price of Best-Response (PBR))}$$

This quantity represents the immutable lower bound on cumulative surplus that any FTRL learner must incur to guarantee ε-convergence against the worst-case stabilizing environment.

While a full classification of all regularizers and optimizer choices is beyond the scope of this work, restricting our attention to the separable case yields Theorem 4.4.

Theorem 4.4 (Discrete Time)**.** *For any sufficiently small $\varepsilon > 0$, $\mathcal{C}(\varepsilon) \;=\; \Omega(\theta'(1/k) - \theta'(\frac{\varepsilon}{2(m-k)}))$.*

Before proving Theorem 4.4, we state a useful corollary of Lemma 3.6 that characterizes the ℓ_1-distance between the learner's response and the best-response under a fixed optimizer strategy.

Corollary G.1. *Fix a constant optimizer strategy $\hat{x}$. Let h be a separable regularizer on Δ^m, and fix a constant step size $\eta > 0$. Under FTRL against $\hat{x}$, the learner's response at each $t \geq 0$ satisfies*

$$\|y(t) - \hat{y}^*\|_1 \;=\; 2 \sum_{j \notin \mathrm{br}(\hat{x})} y_j(t), \qquad \hat{y}^* = \mathrm{unif}(\mathrm{br}(\hat{x})). \tag{54}$$

Moreover, we have the two-sided bound

$$2(m-k)\,\phi\big(\theta'(1/m) - \eta t\,\delta_{\max}\big) \;\leq\; \|y(t) - \hat{y}^*\|_1 \;\leq\; 2(m-k)\,\phi\big(\theta'(1/k) - \eta t\,\delta_{\min}\big). \tag{55}$$

Proof. By Lemma 3.6(i), there exists $a_t \geq 0$ such that $y_i(t) = a_t$ for all $i \in \mathrm{br}(\hat{x})$. Since $\sum_{i=1}^m y_i(t) = 1$, we have $\sum_{j \notin \mathrm{br}(\hat{x})} y_j(t) = 1 - k a_t$, and $a_t \leq 1/k$. Hence

$$\|y(t) - \hat{y}^*\|_1 = \sum_{i \in \mathrm{br}(\hat{x})} \big|a_t - \tfrac{1}{k}\big| + \sum_{j \notin \mathrm{br}(\hat{x})} y_j(t) = k\Big(\tfrac{1}{k} - a_t\Big) + \sum_{j \notin \mathrm{br}(\hat{x})} y_j(t) = 2 \sum_{j \notin \mathrm{br}(\hat{x})} y_j(t),$$

which proves (54). The two-sided bound (55) follows by summing the coordinate-wise bounds in Lemma 3.6(ii) over $j \notin \mathrm{br}(\hat{x})$, using monotonicity of ϕ, and applying the definitions of $\delta_{\min}$ and $\delta_{\max}$. □

Then, we are ready to prove Theorem 4.4 under the same assumption of Corollary G.1.

Proof of Theorem 4.4. Fix a matrix A in the high-probability event of Theorem F.1. Let $x^\star$ be a max-min strategy, and let x', x'' be the two optimizer strategies from (40). Define $x_d(\cdot)$ to alternate between x' and x'' as in (41).

Fix $\eta > 0$ and a horizon $t \geq 1$. If t is even, play $x_d(0), \ldots, x_d(t-1)$; if t is odd, play $x_d(0), \ldots, x_d(t-2)$ and then $x^\star$ at time $t-1$. In both cases,

$$\frac{1}{t} \sum_{\tau=0}^{t-1} x_d(\tau) \;=\; x^\star, \qquad \sum_{\tau=0}^{t-1} A^\top x_d(\tau) \;=\; t\,A^\top x^\star.$$

Let $y^* := \mathrm{unif}(\mathrm{br}(x^\star))$. By (Choice Map),

$$y(t) \;=\; Q_{\mathrm{h}}\big(-\eta t\,A^\top x^\star\big), \qquad \gamma(\eta, t) \;=\; \|y(t) - y^*\|_1.$$

By Theorem F.1, the cumulative surplus under this $x_d(\cdot)$ satisfies

$$\mathrm{SURPLUS}_{\mathrm{h}}^{(\eta,t)} \;\geq\; \eta(t-1)\,C_A^{\mathrm{h}}(n, m).$$

Applying Corollary G.1 with $\hat{x} = x^\star$ yields

$$\gamma(\eta, t) \;=\; 2 \sum_{j \notin \mathrm{br}(x^\star)} y_j(t), \qquad \gamma(\eta, t) \;\geq\; 2(m-k)\,\phi\big(\theta'(a_t) - \eta t\,\delta_{\max}\big),$$

where $k = |\,\mathrm{br}(x^\star)|$, and a_t is the common mass on $\mathrm{br}(x^\star)$.

Assume $\gamma(\eta, t) \leq \varepsilon < 2(m-k)$. Then $\varepsilon/(2(m-k)) \in (0, 1]$, and using $\phi^{-1}(u) = \theta'(u)$ on $(0, 1]$,

$$\eta t \;\geq\; \frac{\theta'(a_t) - \theta'\big(\varepsilon/(2(m-k))\big)}{\delta_{\min}}.$$

Moreover, $\gamma(\eta, t) \leq \varepsilon$ implies $\sum_{j \notin \mathrm{br}(x^\star)} y_j(t) \leq \varepsilon/2$, hence $\sum_{i \in \mathrm{br}(x^\star)} y_i(t) \geq 1 - \varepsilon/2$, i.e., $k a_t \geq 1 - \varepsilon/2$ and

$$a_t \;\geq\; \frac{1 - \varepsilon/2}{k}.$$

Since θ' is increasing, we obtain the necessary condition

$$\eta t \;\geq\; \frac{\theta'\big((1-\varepsilon/2)/k\big) - \theta'\big(\varepsilon/(2(m-k))\big)}{\delta_{\max}}.$$

Using the definition

$$\mathcal{C}(\varepsilon) \;\coloneqq\; \inf_{\eta,t} \sup_{x_d(\cdot),\,A} \{\mathrm{SURPLUS}_{\mathrm{h}}^{(\eta,t)} : \gamma(\eta,t) \le \varepsilon\},$$

we lower bound the inner supremum by evaluating it at the specific pair (x_d, A) constructed above:

$$\sup_{x_d(\cdot),\,A} \{\mathrm{SURPLUS}_{\mathrm{h}}^{(\eta,t)} : \varepsilon(\eta,t) \le \varepsilon\} \;\ge\; \mathrm{SURPLUS}_{\mathrm{h}}^{(\eta,t)}(x_d, A) \;\ge\; \eta(t-1)\, C_A^{\mathrm{h}}(n,m).$$

Taking $\inf_{\eta,t}$ and using $t-1 \ge t/2$ for $t \ge 2$ gives

$$\mathcal{C}(\varepsilon) \;\ge\; \frac{C_A^{\mathrm{h}}(n,m)}{2\delta_{\max}} \left(\theta'\Big(\frac{1-\varepsilon/2}{k}\Big) - \theta'\Big(\frac{\varepsilon}{2(m-k)}\Big) \right), \qquad \forall \varepsilon \in (0, 2(m-k)).$$

Since θ' is continuous near $1/k$, we have $\theta'((1-\varepsilon/2)/k) = \theta'(1/k) + O(\varepsilon)$; thus, for all sufficiently small ε,

$$\mathcal{C}(\varepsilon) \;=\; \Omega\Big(\theta'(1/k) - \theta'\big(\varepsilon/(2(m-k))\big)\Big),$$

as claimed. □

H From Mixed-Strategy Payoffs to Realized Payoffs

This appendix compares realized payoffs with their mixed-strategy counterparts along the same arbitrary predictable strategy process $\{x(t), y(t)\}_{t=1}^T$. The results apply, in particular, to strategies generated by bandit-feedback algorithms satisfying the assumptions below.

H.1 Mixed-Strategy Learner Regret

At each round $t = 1, \dots, T$, the players choose mixed strategies $x(t) \in \Delta^n$ and $y(t) \in \Delta^m$. The learner's mixed-strategy regret is

$$\begin{aligned} \mathrm{Regret}_y(T) &\coloneqq \sum_{t=1}^T x(t)^\top A\, y(t) - \min_{y \in \Delta^m} \sum_{t=1}^T x(t)^\top A\, y \\ &= \sum_{t=1}^T x(t)^\top A\, y(t) - \min_{j \in [m]} \sum_{t=1}^T x(t)^\top A\, e_j. \end{aligned} \qquad \text{(Learner's Regret)}$$

This compares the learner's mixed play to the best fixed learner action in hindsight, evaluated against the opponent optimizer's sequence $\{x(t)\}_{t=1}^T$.

H.2 Realized Learner Regret

At each round $t = 1, \dots, T$, the players choose mixed strategies $x(t) \in \Delta^n$ and $y(t) \in \Delta^m$, and then sample actions $i_t \sim x(t)$ and $j_t \sim y(t)$ independently of each other given the past. The realized payoff is A_{i_t, j_t}. Define the learner's realized regret by

$$\widehat{\mathrm{Regret}}_y(T) \coloneqq \sum_{t=1}^T A_{i_t,j_t} - \min_{j \in [m]} \sum_{t=1}^T A_{i_t,j} = \sum_{t=1}^T A_{i_t,j_t} - \min_{j \in [m]} \sum_{t=1}^T A_{i_t,j}.$$

(Learner's Realized Regret)

Related work on adversarial bandits and repeated games is closely connected to our setting. O'Donoghue et al. [24] analyze repeated zero-sum matrix games with bandit-feedback and observed opponent actions, Abernethy et al. [1] study regularization-based bandit methods, and Lee et al. [21] give high-probability guarantees against adaptive adversaries.

Proposition H.1. *Let $A \in [-1,1]^{n \times m}$. At each round $t = 1, \dots, T$, the optimizer and learner choose mixed strategies $x(t) \in \Delta^n$ and $y(t) \in \Delta^m$ using arbitrary (possibly adaptive) mechanisms. Define the pre-sampling history*

$$\mathcal{F}_{t-1} \coloneqq \sigma\Big(\{x(\tau), y(\tau), i_\tau, j_\tau\}_{\tau=1}^{t-1},\, x(t), y(t)\Big),$$

so that $x(t)$ and $y(t)$ are $\mathcal{F}_{t-1}$-measurable. At round t, sample $i_t \sim x(t)$ and $j_t \sim y(t)$ conditionally independently given $\mathcal{F}_{t-1}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\widehat{\textit{Regret}}_y(T) \le \textit{Regret}_y(T) + \mathcal{O}\left(\sqrt{T \log \frac{m}{\delta}}\right). \tag{56}$$

Proof. Decompose (Learner's Realized Regret) as

$$\begin{aligned}\widehat{\text{Regret}}_y(T) &= \text{Regret}_y(T) + \sum_{t=1}^{T}\left(A_{i_t,j_t} - x(t)^\top A\, y(t)\right) + \left(\min_{j\in[m]}\sum_{t=1}^{T} x(t)^\top A\, e_j - \min_{j\in[m]}\sum_{t=1}^{T} A_{i_t,j}\right) \\ &\le \text{Regret}_y(T) + \underbrace{\sum_{t=1}^{T}\left(A_{i_t,j_t} - x(t)^\top A\, y(t)\right)}_{(I)} + \underbrace{\max_{j\in[m]}\sum_{t=1}^{T}\left(x(t)^\top A\, e_j - A_{i_t,j}\right)}_{(II)}.\end{aligned}$$

Term (I) is a martingale difference sequence with increments in $[-2,2]$. By Azuma–Hoeffding, with probability at least $1-\delta/2$,

$$(I) \le 2\sqrt{2T\log\frac{2}{\delta}}. \tag{57}$$

For term (II), for each fixed $j \in [m]$, the sum $\sum_{t=1}^{T}\left(x(t)^\top A e_j - A_{i_t,j}\right)$ is a martingale difference sequence with increments in $[-2,2]$. Applying Azuma–Hoeffding and a union bound over $j \in [m]$, with probability at least $1-\delta/2$,

$$(II) \le 2\sqrt{2T\log\frac{2m}{\delta}}. \tag{58}$$

Combining the two bounds, with probability at least $1-\delta$,

$$\widehat{\text{Regret}}_y(T) \le \text{Regret}_y(T) + 2\sqrt{2T\log\frac{2}{\delta}} + 2\sqrt{2T\log\frac{2m}{\delta}} = \text{Regret}_y(T) + \mathcal{O}\left(\sqrt{T\log\frac{m}{\delta}}\right).$$

□

H.3 From Regret Bounds to Realized Reward

The definition of mixed-strategy regret and the minimax value imply

$$\sum_{t=1}^{T} x(t)^\top A\, y(t) \le T\cdot \text{Val}(A) + \text{Regret}_y(T).$$

We have the following proposition relating the optimizer's realized reward to the realized learner regret.

Proposition H.2. *Under the same setting as Proposition H.1, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} A_{i_t,j_t} \;\le\; T\cdot\text{Val}(A) \;+\; \textit{Regret}_y(T) \;+\; \mathcal{O}\left(\sqrt{T\log\left(\frac{1}{\delta}\right)}\right).$$

Proof. Observe that

$$\begin{aligned}\sum_{t=1}^{T} A_{i_t,j_t} &= \sum_{t=1}^{T} x(t)^\top A\, y(t) + \sum_{t=1}^{T}\left(A_{i_t,j_t} - x(t)^\top A\, y(t)\right) \\ &\le T\cdot\text{Val}(A) + \text{Regret}_y(T) + \sum_{t=1}^{T}\left(A_{i_t,j_t} - x(t)^\top A\, y(t)\right) \\ &\le T\cdot\text{Val}(A) + \text{Regret}_y(T) + 2\sqrt{2T\log\frac{1}{\delta}},\end{aligned}$$

with probability at least $1-\delta$, where the last inequality follows from (57). □

I Experimental Evaluation

We design experiments to illustrate both passive and active utility extraction by the optimizer in the full-feedback setting. We further include bandit-feedback experiments to examine how these phenomena change when the learner observes only realized payoffs. Unlike the full-feedback setting, the bandit experiments do not exhibit the same clean qualitative patterns, likely due to the additional sampling noise and the variance introduced by importance-weighted feedback. Since our theoretical analysis is restricted to full feedback, we leave a rigorous understanding of the bandit-feedback setting as an open question for future work. All experiments were run on a standard laptop (MacBook M1 Pro). The full-feedback experiments took less than one minute to run, whereas the bandit-feedback experiments took around three hours for each game.

I.1 Setup.

We evaluate the FTRL dynamics induced by three regularizers: negative entropy, Euclidean, and negative Tsallis with $q = \frac{1}{2}$. We consider both the full-feedback and bandit-feedback settings. In each setting, we study three zero-sum games and two optimizer strategies: the max-min strategy and the alternating strategy defined from Equation (40).

For all experiments, the learner plays

$$y(t) = Q_{\mathrm{h}}\big(\eta\, \mathrm{S}(t)\big), \qquad \eta = T^{-\beta}, \qquad \beta \in \{0, 0.3, 0.5, 0.7, 1.0\}.$$

Full-Feedback Update. In the full-feedback setting, after the optimizer plays $x(t)$, the learner observes the full reward vector $-A^\top x(t)$ and updates

$$\mathrm{S}(t+1) = \mathrm{S}(t) - A^\top x(t).$$

Bandit-Feedback Update. Appendix H studies realized-payoff concentration for sampled actions, treating the mixed strategies as given. The experiments below use a different, algorithmic bandit FTRL update: the learner must update its score from a one-coordinate importance-weighted estimate of the reward vector. Suppose first that the learner samples $j_t \sim y(t)$, while the optimizer samples $i_t \sim x(t)$. The learner observes only the realized reward $-A_{i_t,j_t}$, and forms the importance-weighted estimator (see Lattimore and Szepesvári [20, Section 11.2] for background on importance-weighted estimators in the EXP3 algorithm for adversarial bandits):

$$\hat{r}(t) = \frac{-A_{i_t,j_t}}{y(t)_{j_t}}\, e_{j_t}. \tag{59}$$

This estimator is unbiased conditionally on the past $\mathbb{E}[\hat{r}(t) \mid \mathcal{F}_{t-1}] = -A^\top x(t)$.

However, without explicit exploration, the estimator is only well-defined and unbiased on the support of $y(t)$. If a non-steep FTRL, like Euclidean regularizer, assigns zero probability to an action, then that action is never sampled, and the bandit feedback no longer provides an unbiased estimate of the full reward vector. To avoid this issue, we also consider the explicitly explored sampling distribution

$$q(t) = (1-\gamma)y(t) + \gamma u, \qquad u = \frac{1}{m}\mathbf{1},$$

where $\gamma \in [0,1]$ is the exploration parameter. In this case, the learner samples $j_t \sim q(t)$ and uses

$$\hat{r}_\gamma(t) = \frac{-A_{i_t,j_t}}{q(t)_{j_t}}\, e_{j_t}. \tag{60}$$

Since $q(t)_j \geq \gamma/m$ for every action j, the denominator is bounded away from zero whenever $\gamma > 0$. Moreover,

$$\mathbb{E}[\hat{r}_\gamma(t) \mid \mathcal{F}_{t-1}] = -A^\top x(t),$$

so the estimator remains unbiased for the full reward vector (see Auer et al. [7] for details on explicit exploration in EXP3). The score vector is then updated by $\mathrm{S}(t+1) = \mathrm{S}(t) + \hat{r}_\gamma(t)$ in the experiment.

We consider the following three zero-sum games.

Game 1. The optimizer payoff matrix is

$$A = \begin{pmatrix} 1 & -1 & 1 & 2 \\ -1 & 1 & 1 & 0 \\ 0 & 0 & -1 & 0 \end{pmatrix},$$

and the learner payoff matrix is $-A$. A max-min strategy is $x^\star = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, with $\mathrm{br}(x^\star) = \{1, 2\}$, game value 0, and $N_{\mathrm{sub}} = 4 - 2 = 2$. For the alternating strategy, we use Equation (40) with $\ell = 1$, yielding

$$x' = \left(\tfrac{5}{9}, \tfrac{2}{9}, \tfrac{2}{9}\right), \qquad x'' = \left(\tfrac{1}{9}, \tfrac{4}{9}, \tfrac{4}{9}\right).$$

Game 2. The Rock–Paper–Scissors payoff matrix is

$$A_{\mathrm{RPS}} = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{pmatrix},$$

and the learner payoff matrix is $-A_{\mathrm{RPS}}$. The max-min strategy is $x^\star = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, with $\mathrm{br}(x^\star) = \{1, 2, 3\}$, game value 0, and $N_{\mathrm{sub}} = 3 - 3 = 0$. We use the same alternating construction as in Game 1.

Game 3. The third game is a random zero-sum game $A_{\mathrm{rand}} \in [-1, 1]^{30 \times 30}$, whose entries are drawn independently and uniformly from $[-1, 1]$. We compute a max-min strategy $x^\star$ and the game value $\mathrm{Val}(A_{\mathrm{rand}})$ via linear programming. We then choose $\ell \in \mathrm{supp}(x^\star)$ maximizing $x^\star_\ell$ over the support, and construct the alternating strategy using Equation (40).

I.2 Full-Feedback Experiments

In the full-feedback experiments, for each triple (β, h, A) and each optimizer strategy, we run one trajectory from the initialization $\mathrm{S}(0) = 0$. The horizon T is chosen separately for each game and is specified in the corresponding figure. We report the cumulative optimizer surplus

$$\sum_{t=1}^{T} x(t)^\top A\, y(t) - T\,\mathrm{Val}(A).$$

Rock–Paper–Scissors: isolating the alternating effect Figure 4 illustrates the contrast between passive and active utility extraction in Rock–Paper–Scissors. When the optimizer repeatedly plays the max-min strategy $x^\star = (1/3, 1/3, 1/3)$, the cumulative surplus is identically zero for all three regularizers. This is the boundary case of the Inverse-Rate Law: every learner action is a best response to $x^\star$, so $N_{\mathrm{sub}} = 0$, and there is no passive surplus for the optimizer to collect.

The right column shows that the disappearance of fixed-strategy surplus does not make the game unexploitable. Under the alternating trap strategy, the optimizer alternates between two perturbations whose average is $x^\star$. Thus the value benchmark is preserved over each pair of rounds, while the perturbations reverse the payoff ordering between two learner best responses. Because the FTRL iterate cannot track this reversal instantaneously, its regularized inertia creates a positive payoff gap across each pair of rounds. Summing these gains yields the active-extraction rate predicted by Theorem 4.3, namely $\Omega(\eta T)$.

Since we set $\eta = T^{-\beta}$, the active surplus scales as $T^{1-\beta}$, so smaller β produces steeper growth. The small sawtooth pattern is the finite-time signature of the alternating trap. Across rows, negative entropy, negative Tsallis with $q = 1/2$, and Euclidean regularizer all exhibit positive surplus under the alternating optimizer. This supports the main message that active extraction is not specific to MWU, but is a structural phenomenon for FTRL.

The three regularizers exhibit the same predicted $\Omega(\eta T)$ order, but Euclidean regularizer has a noticeably larger empirical constant at this horizon. This is consistent with the curvature dependence in the variance lower bound, Equation (46) of Lemma F.8. Euclidean regularizer has constant curvature $\theta'' = 1$, whereas negative entropy and negative Tsallis with $q = 1/2$ have curvatures $\theta''(y_i) = 1/y_i$ and $\theta''(y_i) = \frac{1}{2} y_i^{-3/2}$, respectively. This explains why the Euclidean curve can have a larger empirical slope even though all three regularizers follow the same $\Omega(\eta T)$ mechanism.

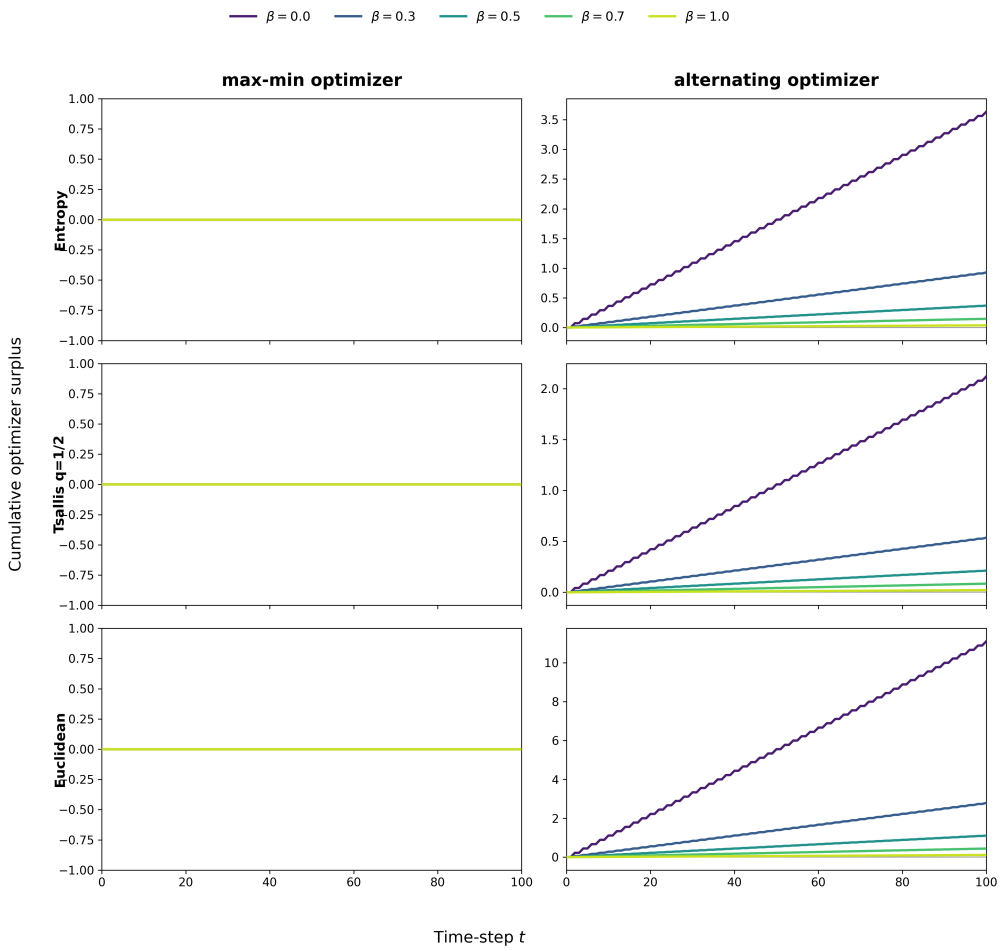


Figure 4: Full-feedback results for RPS. We plot cumulative optimizer surplus under the fixed max-min optimizer (left) and alternating trap optimizer (right) for Negative Entropy, Negative Tsallis $q = 1/2$, and Euclidean regularizers. Curves use $\eta = T^{-\beta}$, $\beta \in \{0, 0.3, 0.5, 0.7, 1.0\}$, with $T = 100$. For the fixed strategy, all surplus curves are identically zero.

Game 1: coexistence of passive and active surplus Figure 5 applies the same fixed-versus-alternating comparison to Game 1. Unlike RPS, the max-min optimizer strategy has $N_{\text{sub}} = 2$.Hence the fixed max-min optimizer already extracts positive surplus. This is the passive mechanism in Theorem 4.1: the learner initially assigns mass to suboptimal actions, and the optimizer collects the resulting convergence lag at scale $\Theta(N_{\text{sub}}/\eta)$. Since $\eta = T^{-\beta}$, this passive contribution scales as T^{β}, so larger β produces larger fixed-strategy surplus.

The right column uses the same alternating construction as in Figure 4, so we do not repeat the mechanism here. In Game 1, this active term is added to the passive surplus already present under the fixed max-min strategy. Thus the total surplus reflects two competing scales: the passive $\Theta(N_{\text{sub}}/\eta) = \Theta(T^{\beta})$ term and the active $\Omega(\eta T) = \Omega(T^{1-\beta})$ term. Small β emphasizes the active contribution, while large β emphasizes the passive inverse-rate contribution, making the fixed and alternating curves closer on the raw scale.

The regularizer-dependent constants follow the same curvature explanation discussed for RPS. Euclidean regularizer has a larger empirical constant under the alternating strategy when $\beta = 0$. Under the fixed max-min strategy, the optimizer obtains less surplus against the Euclidean regularizer, consistent with the discussion of steep and non-steep geometries and the price of best response in Section 4.3.

Overall, Game 1 shows that the passive and active mechanisms are complementary. Suboptimal learner actions generate the passive $\Theta(N_{\text{sub}}/\eta)$ surplus, while alternating play can further exploit FTRL inertia through the $\Omega(\eta T)$ trap.

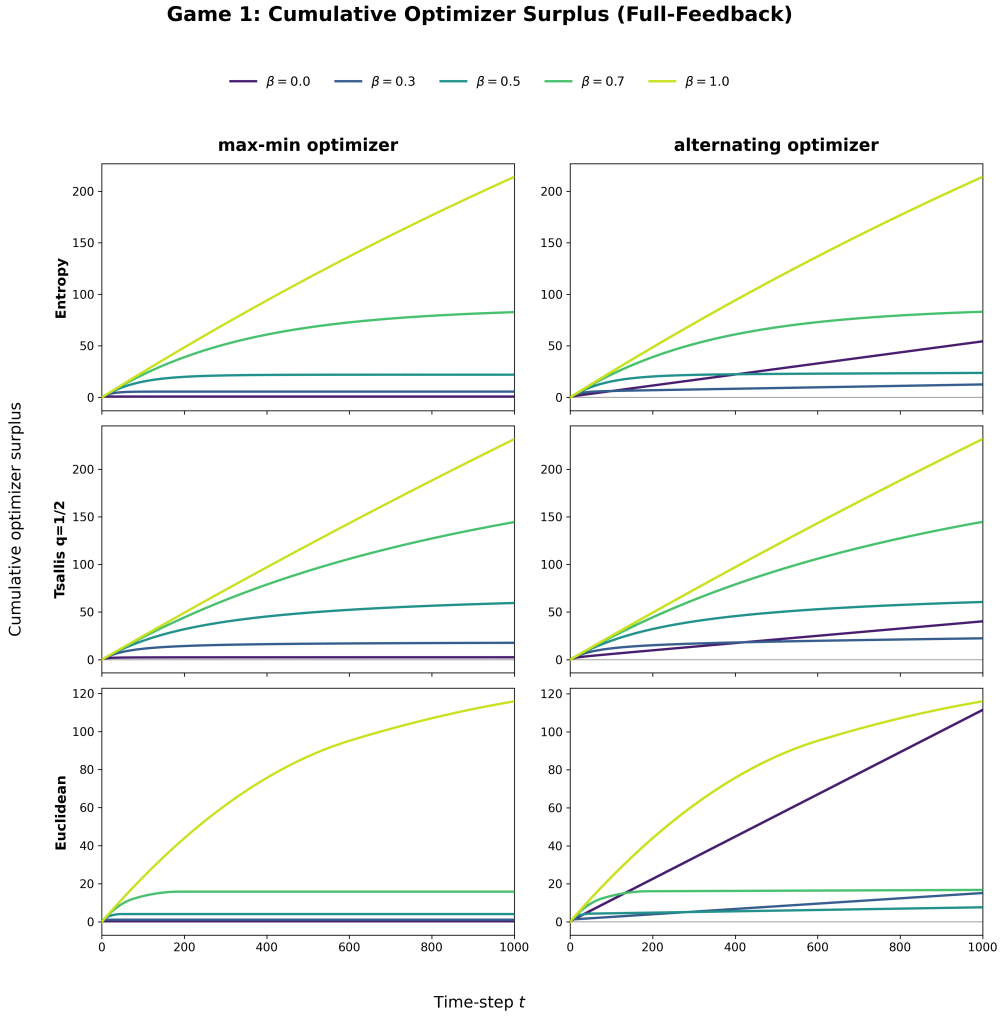


Figure 5: Full-feedback results for Game 1. We plot cumulative optimizer surplus under the fixed max-min optimizer (left) and alternating trap optimizer (right) for Negative Entropy, Negative Tsallis $q = 1/2$, and Euclidean regularizers. Curves use $\eta = T^{-\beta}$, $\beta \in \{0, 0.3, 0.5, 0.7, 1.0\}$, with $T = 1000$.

Random games: surplus extraction beyond toy examples Figure 6 repeats the fixed-versus-alternating comparison for a random 30×30 zero-sum game. As in Game 1, the fixed max-min optimizer already obtains positive surplus, indicating that the random instance contains learner actions that are suboptimal against $x^\star$. The passive contribution therefore follows the Inverse-Rate Law, scaling as $\Theta(N_{\text{sub}}/\eta) = \Theta(T^\beta)$.

The alternating part use the same active trap mechanism discussed above. In this random game, the alternating optimizer adds an $\Omega(\eta T) = \Omega(T^{1-\beta})$ contribution on top of the passive surplus. Thus small β emphasizes the active term, while large β emphasizes the passive inverse-rate term.

The regularizer-dependent constants are especially visible in this higher-dimensional instance. Under the alternating strategy, Euclidean regularizer has a much larger empirical constant, particularly when $\beta = 0$. This is consistent with the curvature explanation: in a 30×30 game, the entropy and Tsallis iterates can place smaller mass on best-response actions, making $\theta''(y_i) = 1/y_i$ and

$\theta''(y_i) = \frac{1}{2} y_i^{-3/2}$ larger relative to the constant Euclidean curvature $\theta'' = 1$ than in the previous games. Under the fixed max-min strategy, the optimizer obtains less surplus against Euclidean regularizer for all values of β, consistent with the steep versus non-steep geometry discussion and the price of best response in Section 4.3.

Overall, the random-game experiment shows that the passive and active mechanisms are not artifacts of the handcrafted examples. Passive surplus comes from suboptimal learner actions, active surplus comes from the alternating trap, and the regularizers mainly differ through geometry-dependent constants.

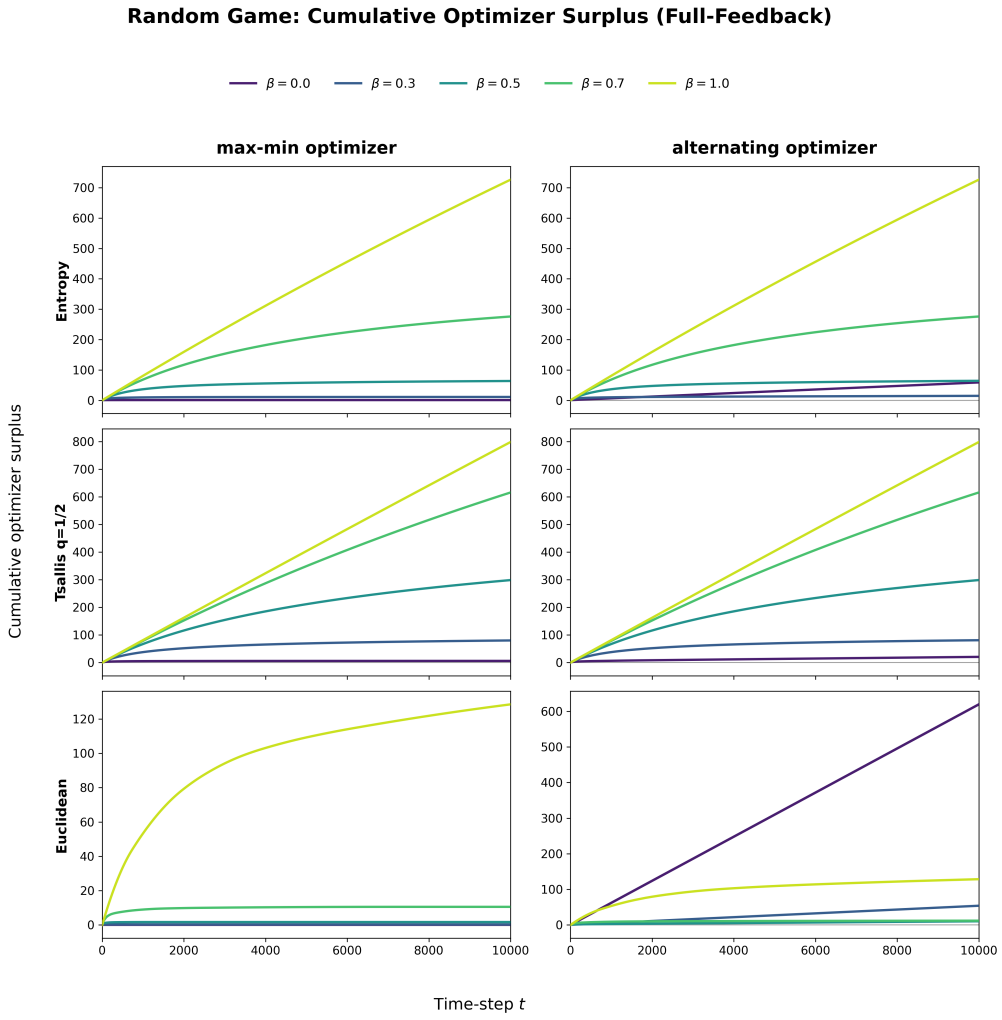


Figure 6: Full-feedback results for random Game 3. We plot cumulative optimizer surplus under the fixed max-min optimizer (left) and alternating trap optimizer (right) for Negative Entropy, Negative Tsallis $q = 1/2$, and Euclidean regularizers. Curves use $\eta = T^{-\beta}$, $\beta \in \{0, 0.3, 0.5, 0.7, 1.0\}$, with $T = 10000$.

I.3 Bandit-Feedback Experiments

In the bandit experiments with explicit exploration, for each triple $(\beta, \mathrm{h}, A, \gamma)$ and each optimizer strategy, we run 500 independent trials with horizon $T = 5000$, again with initialization $\mathrm{S}(0) = 0$. For each trial, we record the cumulative realized optimizer surplus

$$\sum_{t=1}^{T} A_{i_t, j_t} - T \operatorname{Val}(A).$$

Across trials, we summarize the distribution at each time t by the median together with the 10th and 90th percentiles because of realized action draws and importance-weighted feedback. We set $\gamma \approx 0.02, 0.025, 0.11$ for the three games below, respectively, following the standard EXP3 tuning in Auer et al. [7, Corollary 3.2]

$$\gamma = \sqrt{\frac{m \ln m}{(e-1)T}},$$

where m is the number of learner actions and T is the horizon. We use the same exploration rate for all regularizers as a standard variance-control choice for the importance-weighted bandit estimator.

Rock–Paper–Scissors under bandit-feedback Figure 7 shows that, under bandit feedback, the fixed max-min optimizer produces no visible surplus. This mirrors the full-feedback RPS behavior in the left column of Figure 4, where the fixed-strategy surplus is identically zero. The wide percentile bands reflect the additional randomness from realized action draws and importance-weighted feedback, rather than passive utility extraction.

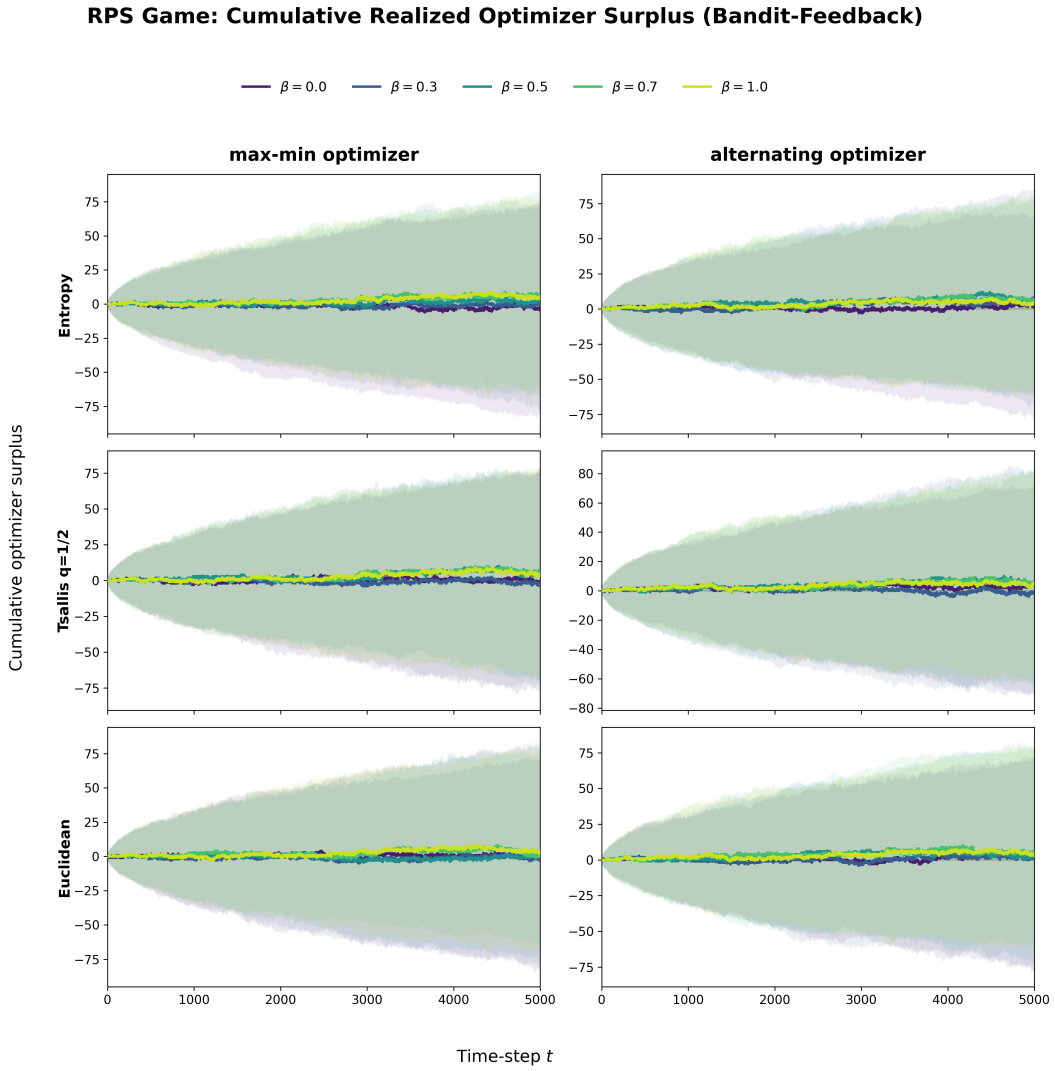


Figure 7: Bandit-feedback results for RPS. We plot cumulative realized optimizer surplus under the fixed max-min optimizer (left) and the alternating trap optimizer (right) for Negative Entropy, Negative Tsallis with $q = 1/2$, and Euclidean regularizer (top to bottom). Curves correspond to step sizes $\eta = T^{-\beta}$, $\beta \in \{0, 0.3, 0.5, 0.7, 1.0\}$, with $T = 5000$ and $\gamma = 0.02$. For each setting, we run 500 independent trials and plot the median together with the 10th and 90th percentiles.

The alternating trap optimizer exhibits a less clean pattern under bandit-feedback than in the full-feedback experiment. In contrast to the right column of Figure 4, the median realized surplus remains close to zero for all three regularizers and all tested step sizes, while the percentile bands are large. A rigorous analysis of whether utility extraction persists in expectation is left for future work.

Game 1 under bandit-feedback We next repeat the Game 1 experiment under bandit-feedback. The median realized surplus is clearly positive under both optimizer strategies, and its dependence on β mainly follows the inverse-rate law already visible in the full-feedback experiment: larger β produces larger median surplus. This is consistent with Game 1 having suboptimal learner actions against $x^\star$, which already generate passive utility extraction under the fixed max-min optimizer.

The 10th–90th percentile bands are wider for smaller β, indicating that sampling noise has a larger effect when the step size is larger. Under the alternating optimizer, the median curves remain broadly similar to the fixed max-min curves, rather than showing the clean additional separation seen in the right column of Figure 5. Thus, Figure 8 still displays the passive utility extraction, but it does not provide a clean additional active extraction under alternating play.

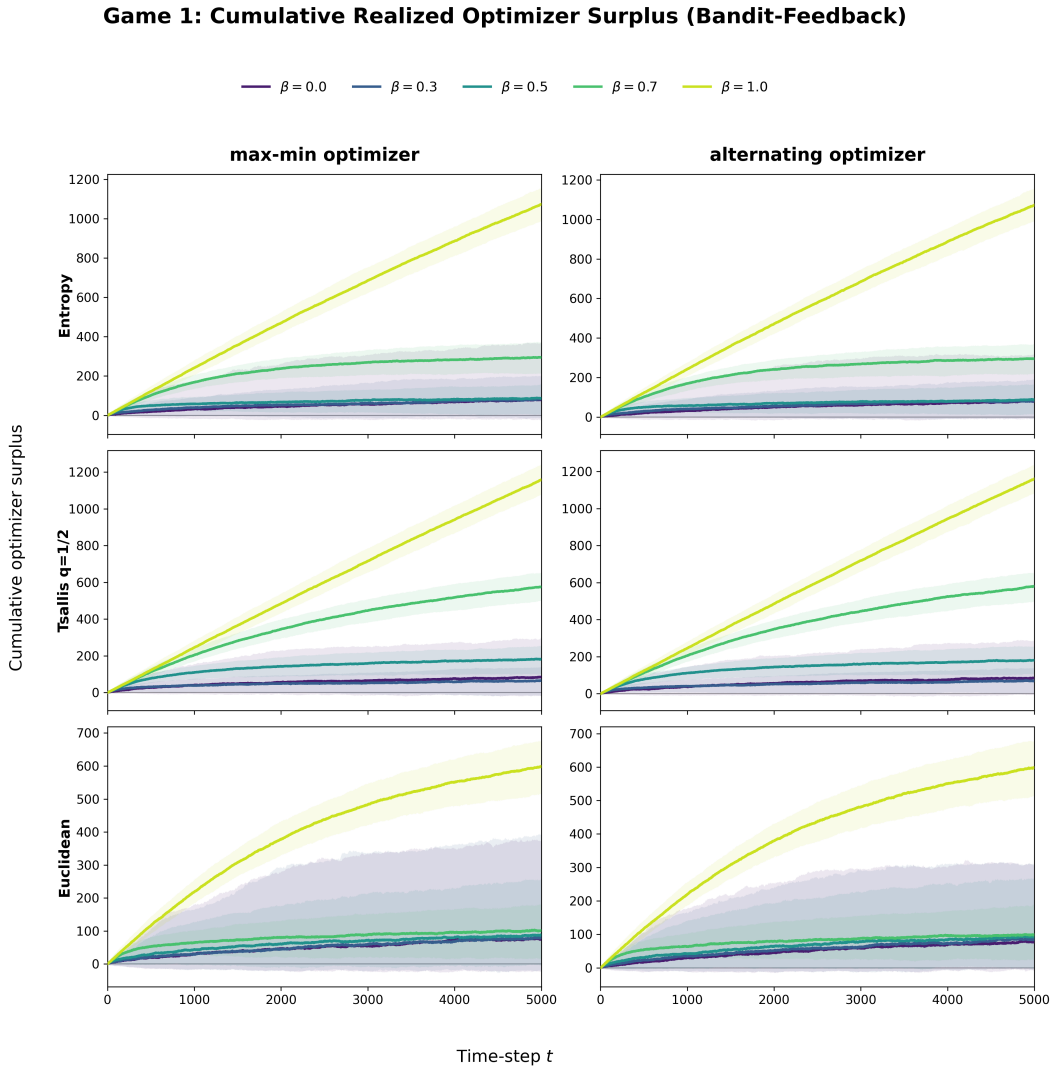


Figure 8: Bandit-feedback results for Game 1. We plot the cumulative realized optimizer surplus under the fixed max-min optimizer (left) and the alternating trap optimizer (right) for negative entropy, negative Tsallis with $q = 1/2$, and Euclidean regularizer (top to bottom). Curves correspond to step sizes $\eta = T^{-\beta}$, $\beta \in \{0, 0.3, 0.5, 0.7, 1.0\}$, with $T = 5000$ and $\gamma = 0.025$. For each parameter setting we run 500 independent trials and plot the median together with the 10th and 90th percentiles.

Random game under bandit-feedback Figure 9 shows the same qualitative behavior as Game 1 under bandit-feedback. The median realized surplus still follows the inverse-rate law. However, the alternating optimizer does not produce a clean additional extraction.

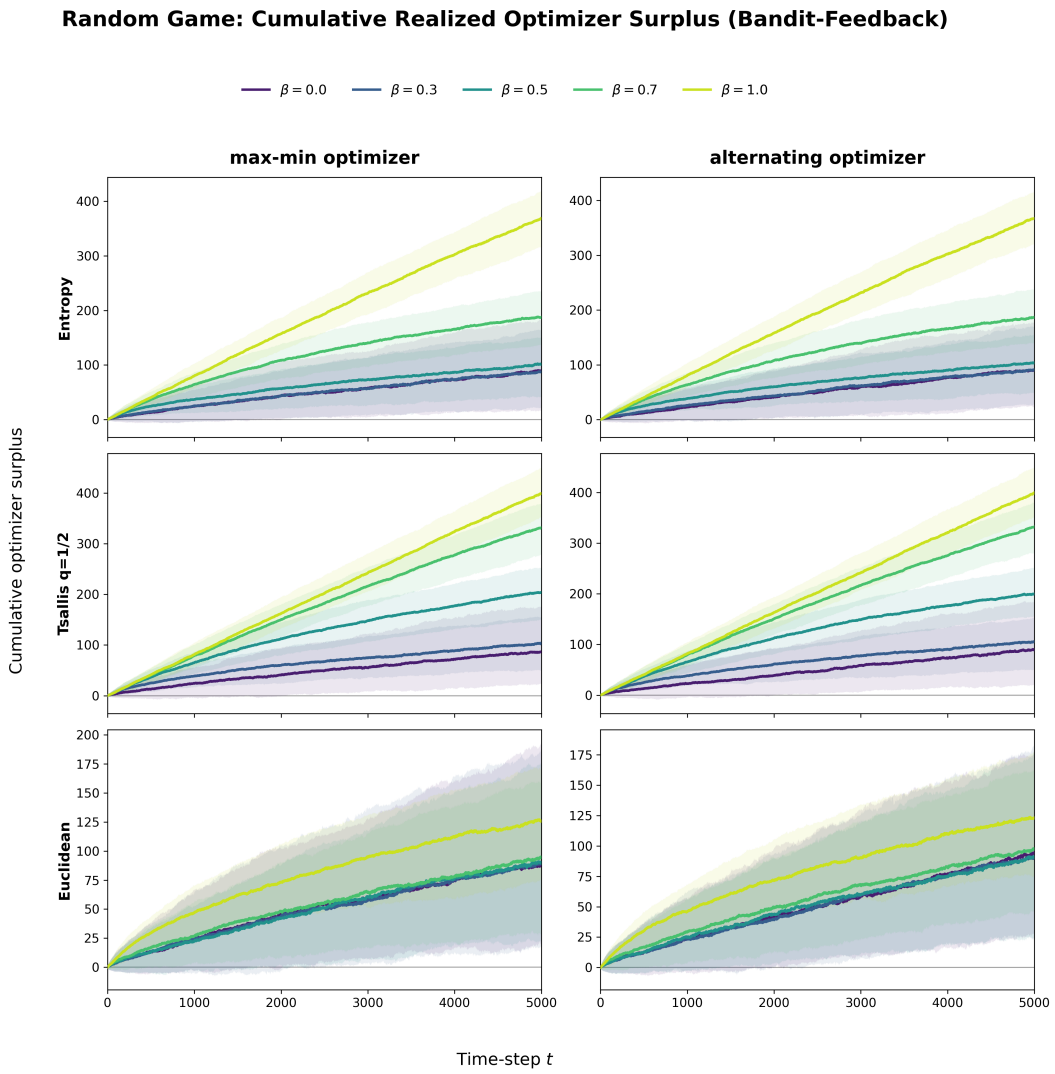


Figure 9: Bandit-feedback results for random game. We plot cumulative realized optimizer surplus under the fixed max-min optimizer (left) and the alternating trap optimizer (right) for negative entropy, negative Tsallis with $q = 1/2$, and Euclidean regularizer (top to bottom). Curves correspond to step sizes $\eta = T^{-\beta}$, $\beta \in \{0, 0.3, 0.5, 0.7, 1.0\}$, with $T = 5000$ and $\gamma = 0.11$. For each parameter setting we run 500 independent trials and plot the median together with the 10th and 90th percentiles.

Overall, the full-feedback experiments match the theoretical predictions: fixed max-min play exhibits the inverse-rate passive surplus when suboptimal learner actions are present, while alternating play produces the additional active-extraction effect. In the bandit-feedback experiments, the median realized surplus still reflects the inverse-rate law in games with suboptimal actions, but the alternating effect is not cleanly visible at the median level. We also note that the bandit surplus curves include the cost of explicit exploration, which is absent from the full-feedback experiments and may further affect the observed trajectories. Whether active utility extraction persists for the expected realized reward under bandit-feedback remains an important direction for future work.

J Future Directions: An Extended Discussion

The framework developed in this paper — pricing the FTRL learner's regret as a transferable resource extracted by a clairvoyant optimizer, and reading the regularizer's geometry as the mediator of that transfer — opens several natural research directions. We organize them around the three themes flagged in the main-text epilogue: the learner's flexibility, partial feedback, and structure beyond full knowledge. For each, we describe the question, what we believe the right form of the answer should be, and which technical pieces of the present work do (and do not) carry over.

J.1 The Learner's Flexibility

Our results assume two simplifying choices on the learner's side: a constant step size η, and an unconstrained ability to revise the mixed strategy from one round to the next. Relaxing either assumption changes the picture in a structured way, and the two relaxations interact in ways we find conceptually appealing.

Adaptive and decaying schedules. The inverse-rate law $\Theta(N_{\mathrm{sub}}/\eta)$ and the alternating-trap rate $\Omega(\eta T/\mathrm{poly}(n,m))$ balance precisely at $\eta = \Theta(T^{-1/2})$, recovering the classical $\sqrt{T}$ benchmark — but as a *consequence* of the static–dynamic duality rather than as a free design choice. For an adaptive or vanishing schedule $\{\eta_t\}_{t=1}^{T}$, the natural conjecture is that η should be replaced by an effective rate derived from a *cumulative learning-rate budget*. The cleanest candidate is $H_T := \sum_{t=1}^{T} \eta_t$, yielding a static-style bound of the form $\Omega(N_{\mathrm{sub}} \cdot T/H_T)$ in terms of the time-averaged rate $\overline{\eta}_T = H_T/T$, and a trap-style bound of the form $\Omega(H_T/\mathrm{poly}(n,m))$. Under the AdaGrad-style schedule $\eta_t = \Theta(1/\sqrt{t})$, $H_T = \Theta(\sqrt{T})$, again recovering the benchmark. Establishing this rigorously requires extending our continuous-time decomposition (Theorem 3.3) to time-inhomogeneous mirror flows, where the standard energy arguments acquire an additional drift term proportional to $\dot{\eta}_t$. The technical question is whether this drift is benign — preserving the inverse-rate structure with H_T replacing ηT — or whether it introduces a genuinely new failure mode for steep regularizers, whose mirror map is most sensitive precisely in the regimes where η_t shrinks fastest.

Switching-constrained and lazy learners. A more realistic class of learners cannot revise their mixed strategy arbitrarily fast — either because of an explicit *switching cost* per unit movement in y, or because the update rule itself enforces a small step (e.g., projected lazy mirror descent, or FTRL with proximal regularization). This constraint cuts both ways. On one hand, the alternating trap of Theorem 4.3 relies on the learner's response cycling between two best responses within a constant number of rounds; if switching is penalized, the trap is weakened or destroyed entirely, and the dynamic surplus drops below $\Omega(\eta T)$. On the other hand, the inverse-rate law *sharpens*: a learner that cannot abandon suboptimal actions quickly accrues a longer transient on each one. We conjecture the static surplus grows from $\Theta(N_{\mathrm{sub}}/\eta)$ to $\Theta(N_{\mathrm{sub}}/(\eta \cdot \kappa))$, where $\kappa \in (0,1]$ is a switching-cost parameter measuring the maximum per-round movement allowed. The borderline case — where the learner's switching budget is comparable to the optimizer's alternation period — is the regime in which we expect the most interesting phase transition, and possibly a third extraction mechanism that interpolates between static and alternating play.

Joint adaptivity. Combining the two relaxations above is the most realistic — and most challenging — variant. A learner who tunes η_t adaptively while paying a switching cost effectively chooses a trajectory in a two-dimensional design space, and the optimizer's optimal extraction strategy presumably depends on which axis dominates. We expect a clean trade-off curve here, parameterized by the ratio of the switching budget to the cumulative learning-rate budget H_T.

J.2 Partial Feedback

Appendix H compares sampled payoffs with their mixed-strategy counterparts along arbitrary predictable trajectories via an Azuma-style concentration argument. The behaviour at the optimal step size — say, EXP3 with $\eta_T = \Theta(\sqrt{\log n/(nT)})$ — is genuinely unsettled, and the experiments in Appendix I are themselves ambiguous on this point: under the alternating-trap optimizer, the median realized surplus collapses toward zero across all three regularizers and all tested step sizes, while the percentile bands remain wide. Two competing hypotheses are consistent with this evidence.

The variance hypothesis. The importance-weighting noise of $\hat{r}_t = r_t \cdot \mathbf{1}[i_t = i]/y_t(i)$ inflates the second moment of the learner's gradient estimate by a factor of $1/y_{\min}$, where $y_{\min}$ is the minimum probability the learner places on any action. Under this hypothesis, the same $o(T)$ leak persists in expectation, but is masked at the median level by variance that scales with $1/y_{\min}$. The leak should re-emerge over horizons $T \gtrsim 1/y_{\min}^2$, i.e., once the optimizer has had enough rounds for the law of large numbers to dominate the per-round noise. A clean test would be to run the bandit experiment at exponentially longer horizons and look for a crossover at which the median surplus separates from zero.

The cancellation hypothesis. Alternatively, the unbiasedness of $\hat{g}_t$ together with the learner's optimally-tuned regularization actually averages the leak away in expectation, and the true bandit surplus is $o(\sqrt{T})$ or smaller. This would mark a genuine separation between the full-feedback and bandit settings: the geometric vulnerability of FTRL would be a property of *deterministic* gradient access, not of the regularizer alone. Distinguishing the two hypotheses requires either a tight Azuma-style upper bound matching our lower bound at the optimal rate, or a constructive optimizer strategy that survives unbiased noise — both of which appear to demand new technical tools beyond the concentration arguments developed here. Our prior is mildly toward the variance hypothesis, since the inverse-rate mechanism is fundamentally about transient suboptimality, and unbiasedness alone does not eliminate transients; but we view the question as open.

Bandits with side information. A softer variant, intermediate between full and bandit feedback, is the case where the learner observes the realized payoff plus a coarse signal about unplayed actions — for example, a noisy unbiased estimate of the full gradient with bounded variance. We expect our extraction rates to interpolate smoothly between the two extremes as a function of the signal's variance, and this regime may be technically more tractable than pure bandit feedback while still capturing the realistic case of an analyst who has access to historical aggregates but not full counterfactuals.

J.3 Structure Beyond Full Knowledge

The third theme groups together two extensions that look superficially different — going beyond two-player zero-sum games, and weakening the optimizer's payoff knowledge — but which share a common technical core: in both, the optimizer must operate without one of the structural primitives our current proofs rely on.

Fine grained analysis beyond two-player zero-sum. Our analysis leans on three structural features of the zero-sum bilinear setting: the reward sandwich $T \cdot V^\star \;\le\; \mathrm{Rew}_T \;\le\; T \cdot V^\star + \mathrm{Regret}(T)$ that pins the surplus into the $o(T)$ regime, the bilinear form $x^\top A y$ that decouples the two players' contributions, and the simplex geometry of mixed strategies that supports our regularizer dichotomy. Each of these breaks differently in the natural extensions. *General-sum games* preserve the simplex geometry but lose the sandwich, so the right benchmark is the Stackelberg value V^{Stack} rather than the minimax value V^*; the inverse-rate law must be reformulated as extraction *above the Stackelberg payoff*, and the alternating trap may yield linear surplus when the two values separate. *Network and polymatrix games* preserve bilinearity locally but introduce cross-coupling between learners, opening the possibility of *indirect* extraction, where the optimizer steers one learner through another — a phenomenon with no analogue in the bilateral setting. *Extensive-form games* replace the simplex with the sequence-form polytope, and the regularizer's geometry interacts with the tree structure in ways not captured by our separability assumption. We conjecture the inverse-rate law survives in network games (with N_{sub} summed across players, weighted by the network's spectral structure) and in extensive-form games under dilated regularizers, but with constants that depend on the depth of the tree and possibly degrade exponentially in pathological branching patterns.

The partially-informed optimizer. Perhaps the most operationally interesting variant: in realistic learning markets, the optimizer rarely has full knowledge of the payoff matrix A, but often does know — or can fingerprint from a few rounds of observation — the learner's *algorithmic profile*: the regularizer family, the step-size schedule, and the feedback model. We conjecture that a substantial fraction of our surplus is recoverable from the profile alone, with A entering only as a lower-order term. Formally, let $S^\star(A, \mathcal{L})$ denote the optimal extractable surplus under full payoff knowledge

against a learner $\mathcal{L}$, and define the profile-only worst-case surplus

$$S^{\dagger}(\mathcal{L}) \;:=\; \sup_{\sigma}\; \inf_{A}\; \mathbb{E}\big[\, S(\sigma; A, \mathcal{L}) \,\big],$$

where σ ranges over optimizer strategies that depend only on the learner's profile and the observed history of play. The conjecture is that $S^{\dagger}(\mathcal{L}) \geq (1 - o(1)) \cdot S^{\star}(A, \mathcal{L})$ for a $1 - o(1)$ fraction of payoff matrices drawn from a natural prior, with the gap closing as the optimizer accumulates observations. A constructive version would proceed in two phases: a short *profiling* phase in which the optimizer estimates the active best-response correspondence from the learner's trajectory, followed by an *extraction* phase in which the alternating trap of Section 3 is deployed once two non-dominated best responses have been identified. A clean such reduction — say, $\widetilde{O}(n^2 m)$ rounds of profile-only exploration sufficient to recover the trap rate up to log factors — would unify our static and dynamic results into a single anytime algorithm, and would represent, in our view, the natural endpoint of the program initiated here.

Profile uncertainty. A symmetric variant is the case where the optimizer knows A but is uncertain about the learner's profile — for example, the regularizer family or the step size is drawn from a small set and revealed only through play. Here we expect a robust extraction rate that interpolates between the worst-case and best-case profiles, with the optimizer paying a logarithmic exploration overhead to identify the active profile before deploying the matching trap. This is technically closer to a partial-monitoring problem than to the full-knowledge analysis of the present work, and we expect the right tools to come from the literature on online learning against unknown opponents rather than from FTRL theory itself.

J.4 A Closing Thought

Across all three themes, we conjecture that the inverse-rate law and the alternating trap are robust under *relaxations of the learner's side* (slower updates, bandit feedback, structurally constrained games), but require new ideas under *relaxations of the optimizer's side* (unknown payoffs, unknown profiles). This asymmetry is, we think, the right one to expect — the optimizer's leverage in our framework comes from its informational advantage, and chipping away at that advantage is precisely what makes the questions hard. We hope the framework here proves a useful starting point for chipping back.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction restrict the claims to repeated two-player zero-sum games with FTRL learners and a clairvoyant optimizer, while Section 2 states the full-feedback default and separable-regularizer scope. The fixed-strategy inverse-rate law, random-game alternating-trap lower bound, steep/non-steep dichotomy, and realized-payoff concentration comparisons are matched by Theorems 4.1, 4.3, and 4.4, together with Appendices F, G, and H.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions for the FTRL learner, feedback model, regularizer class, step-size regime, and random-game events are stated in Section 2 and in the corresponding theorem statements. Complete proofs are provided in the appendix for the reward decomposition, inverse-rate law, alternating-trap result, price-of-best-response result, and realized-payoff concentration comparisons, with main-text proof sketches for the deferred arguments.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The main claims are theoretical, and Appendix I discloses the supporting experiment setup: the update rules, games or random-game distribution, regularizers, step-size grid, horizons, initialization, optimizer strategies, and cumulative-surplus metric.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the experiment implementation and instructions for reproducing the reported figures in the supplementary material, with the experimental setup described in Appendix I. The experiments use synthetically specified games or random games, so no external dataset is required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experiments are simulation studies rather than train/test evaluations, so no data splits are required. Appendix I specifies the games, regularizers, feedback models, optimizer strategies, step-size grid, horizons, initialization, update rules, and cumulative-surplus metric needed to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: For the stochastic bandit experiments, we run multiple independent trials and report the distribution of cumulative surplus across trials using median curves with percentile bands. Appendix I specifies the number of trials, the source of randomness, and how the reported bands are computed.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: These details are provided in Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is a theoretical and simulation-based study of learning dynamics in zero-sum games, uses no human subjects, personal data, or sensitive datasets, and does not deploy or evaluate systems in real-world decision-making settings.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in Appendix A.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models with high misuse risk. The experiments use synthetically specified zero-sum games and simulation code for theoretical learning dynamics.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses synthetically specified games and original simulation code, and cites the external algorithms, theoretical tools, and prior work on which the analysis builds. No third-party datasets or pretrained models are used.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The only newly released assets are the simulation code and synthetic experiment specifications, which are submitted as supplementary material. No new dataset or pretrained model is introduced.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, study participants, or crowdsourcing, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.