# WHEN DOES MOMENTUM HELP IN REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Momentum is a classical tool for accelerating convex optimization and a standard component of modern large language model training. Yet recent empirical work has questioned the usefulness of first-moment momentum in reinforcement learning with verifiable rewards (RLVR). We study this question through a bandit formulation of RLVR with bias-corrected exponential moving average (EMA) momentum, where the EMA coefficient $\beta$ controls how quickly the weights on past gradients decay. When policy gradients are evaluated exactly, we prove that increasing $\beta$ within a range that prevents overshoot strictly accelerates optimization. When gradients are instead estimated from sampled responses, a positive expected gain can be obscured by sampling fluctuations. We show that as the learning rate $\eta$ decreases, the expected gain with fixed $\beta$ vanishes faster than these fluctuations, whereas increasing $\beta$ to keep $\eta/(1-\beta)$ within the safe regime preserves a non-vanishing gain that eventually dominates them. Bandit simulations quantitatively validate these predictions. RLVR experiments on Qwen3-4B-Base using SGD with EMA and AdamW support our principle of choosing larger coefficients in practical LLM post-training. On Qwen2.5-32B with AdamW, simply increasing $\beta_1$ improves mean validation accuracy by nearly 2 percentage points over the conventional choice, demonstrating the principle's effectiveness at a larger model scale.

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has become an important approach to improving the reasoning capabilities of large language models (LLMs) (Lambert et al., 2024; DeepSeek-AI et al., 2025; Yu et al., 2025). In RLVR, responses are sampled from the current policy and evaluated using verifiable rewards, so the training data distribution evolves as the policy changes. This differs from pretraining and supervised fine-tuning, which optimize next-token prediction (NTP) on given text sequences (Brown et al., 2020; Ouyang et al., 2022). Despite these differences, RLVR commonly adopts Adam (Kingma & Ba, 2015) or AdamW (Loshchilov & Hutter, 2019), together with conventional first-moment momentum settings (Zhang et al., 2025; Mukherjee et al., 2026). The design and configuration of optimizers for RLVR have received relatively limited attention.

One important aspect of optimizer design and configuration is the use of first-moment momentum, implemented in Adam and AdamW as an exponential moving average (EMA) of past gradients. While EMA has shown empirical benefits in NTP training (Chen et al., 2023; Zhao et al., 2025), Mukherjee et al. (2026) report that momentum does not consistently improve performance in their RLVR experiments and can even reduce it under a conventional choice of the momentum coefficient. These observations motivate a closer examination of whether EMA can benefit RLVR and what factors influence its effectiveness.

To understand when EMA benefits RLVR and how to choose its coefficient, we adopt a KL-regularized response-level bandit formulation (Xiong et al., 2024). We analyze natural policy gradient (NPG, Kakade, 2001) with a softmax policy parameterization and bias-corrected EMA. NPG provides a tractable theoretical framework for studying policy improvement while controlling changes in the policy distribution, a principle shared by TRPO and PPO (Schulman et al., 2015; 2017; Shani et al., 2020). NPG-based analysis has also provided practical guidance for LLM post-training (Gai et al., 2026). Our contributions are as follows.



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

**Deterministic acceleration and stochastic gains.** We first prove that, with exact gradients, increasing the EMA coefficient within a safe regime strictly accelerates optimization, which appears to contrast with the limited benefits reported in prior RLVR experiments (Mukherjee et al., 2026). A deterministic gain, however, need not translate into a consistent improvement when gradients are estimated from sampled responses, as sampling fluctuations may obscure the benefit. We characterize the expectation and variance of the objective gain and prove that, for a fixed EMA coefficient, the expected gain remains positive for sufficiently small learning rates but vanishes faster than the sampling fluctuations as the learning rate decreases (Theorems 4.2 and 5.3).

**Principled selection of momentum coefficients.** We next show that scaling the EMA coefficient $\beta$ with the learning rate $\eta$ so that $\eta/(1-\beta)$ is a positive constant within the safe regime yields a nonvanishing expected gain that eventually dominates sampling fluctuations (Theorem 5.4). We further prove that an appropriately increased coefficient allows a smaller learning rate to achieve comparable expected performance under the same update and sample budgets, with asymptotically lower variance of the objective value (Theorem 5.5).

**Validation in bandits and LLM post-training.** We quantitatively validate the predicted gain and fluctuation scales in numerical bandit experiments. GRPO experiments (Shao et al., 2024) on Qwen3-4B-Base (Yang et al., 2025) support our theoretical predictions in practical LLM post-training with both SGD with EMA and the widely used AdamW optimizer: larger coefficients improve performance and, when paired with smaller learning rates, yield more stable training with comparable or slightly improved performance. We further demonstrate the principle's effectiveness at a larger model scale: on Qwen2.5-32B (Qwen et al., 2024), simply increasing AdamW's first-moment coefficient improves mean validation accuracy by nearly 2 percentage points over the conventional choice.

## 2 Related work

**Optimization and momentum in RLVR.** Recent work revisits optimizer design for RLVR through momentum-based normalization (Zhang et al., 2025), spectral transformations (Fan et al., 2026; Shi et al., 2026; Zhu et al., 2026), and parameter-space exploration (Venkatkrishna et al., 2026). Focusing on optimizer momentum, Fontana et al. (2026) analyze how accumulated AdamW momentum can continue driving updates after clipping suppresses the current policy-gradient contribution during repeated optimization of the same batch. Most directly related to our motivation, Mukherjee et al. (2026) find that SGD can match or outperform AdamW and report no consistent benefit from momentum. Motivated by these empirical findings, we study how the benefit of ordinary EMA momentum in RLVR depends on the momentum coefficient, learning rate, and sampling noise.

**Momentum in reinforcement learning.** Prior work studies momentum for accelerating policy optimization and reducing gradient-estimation error in conventional RL. Chen et al. (2024) establish accelerated convergence for exact-gradient policy optimization under tabular softmax parameterization using Nesterov momentum with a restart rule. Momentum-based variance reduction has also been studied for policy gradient (Huang et al., 2020; Ding et al., 2022) and natural policy gradient (Feng et al., 2024). Huang et al. (2022) develop Bregman policy optimization with basic momentum and a variance-reduced variant. This literature primarily concerns convergence and sample complexity in conventional RL, with reported experiments on control and Atari benchmarks. We study bias-corrected averaging of past natural gradients in a KL-regularized RLVR bandit model, comparing objective values with and without momentum and examining the practical implications through GRPO experiments with language models.

**Momentum in supervised learning and stochastic optimization.** Analyses of online learning and general stochastic optimization characterize regimes in which momentum behaves similarly to SGD under appropriate learning-rate scaling (Yuan et al., 2016; Wang et al., 2024). In language-model pretraining, AdEMAMix combines fast and slow gradient averages to exploit longer gradient memory (Pagliardini et al., 2025), while ADana combines time-varying momentum schedules with a direct gradient term and damping (Ferbach et al., 2026). These pretraining studies concern next-token prediction on given text sequences. Our focus is momentum in RLVR, where reward feedback is obtained from responses sampled from the evolving policy.

# 3 PRELIMINARIES

In this section, we introduce the model, algorithms, and notation used throughout the paper, providing the foundation for our theoretical analysis.

## 3.1 MODEL AND PARAMETERIZATION

**Basic Setup.** Let $x \in \mathcal{X}$ be a given input prompt, and let $y \in \mathcal{Y}$ be a corresponding response generated by the language model. The model is parameterized as a policy $\pi$ characterized by $\theta$, from which responses are sampled $y \sim \pi_\theta(\cdot \mid x)$. Each prompt-response pair $(x, y)$ is evaluated by a verifier and receives a reward $r(x, y)$. We consider the case of deterministic binary rewards: $r(x, y)$ is 1 if and only if the response $y$ is a correct answer to the prompt $x$, and 0 otherwise. Let $\mathbf{r} \in \{0, 1\}^{|\mathcal{Y}|}$ denote the vector of rewards for all possible responses to the prompt $x$. We maximize the expected reward while penalizing divergence from a fixed reference policy $\pi_{\text{ref}}$, using the objective

$$J(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[r(x, y)] - \gamma \, \mathrm{KL}(\pi_\theta(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x)), \tag{3.1}$$

where $\mathrm{KL}(\pi_\theta(\cdot \mid x) \| \pi_{\text{ref}}(\cdot \mid x)) := \sum_{y \in \mathcal{Y}} \pi_\theta(y \mid x) \log(\pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x))$ is the Kullback-Leibler divergence between the policy $\pi_\theta(\cdot \mid x)$ and the reference policy, and $\gamma > 0$ is a regularization coefficient. Assume $\pi_{\text{ref}}$ is a full-support policy. To exclude degenerate cases, we assume that the initial policy is suboptimal and that the response set contains both correct and incorrect responses. The objective $J(\theta)$ encourages higher expected reward while penalizing deviations from a fixed reference policy, such as one obtained through supervised fine-tuning.

**Policy parameterization.** We adopt the standard softmax parameterization to model response generation. Given a prompt $x$, we parameterize the conditional distribution over responses using a logit vector $\theta \in \mathbb{R}^{|\mathcal{Y}|}$ as $\pi_\theta(y \mid x) = e^{\theta(x,y)} / \sum_{y' \in \mathcal{Y}} e^{\theta(x,y')}$, where $\theta(x, y)$ is the logit score for response $y$ given prompt $x$. This response-level softmax parameterization provides a tractable abstraction of autoregressive generation in modern LLMs, whose token-level policies use softmax distributions (Ziegler et al., 2019; Zhong et al., 2025).

## 3.2 TRAINING DYNAMICS

**Natural policy gradient.** Our theoretical analysis focuses on the natural policy gradient (NPG) method, which preconditions the policy gradient using the Fisher information matrix (Kakade, 2001). Let $\theta_t$ denote the parameter vector at iteration $t$ and $\pi_{\theta_t}(\cdot \mid x)$ the induced policy. When policy gradients are evaluated exactly, the NPG update is given by:

$$\theta_{t+1} = \theta_t + \eta \mathcal{F}^\dagger(\theta_t) \nabla J(\theta_t), \qquad \mathcal{F}(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[ \nabla \log \pi_\theta(y \mid x) \nabla \log \pi_\theta(y \mid x)^\top \right]. \tag{3.2}$$

Here $\eta > 0$ represents the learning rate, $\mathcal{F}(\theta)$ is the Fisher information matrix, and $\mathcal{F}^\dagger(\theta)$ is its Moore–Penrose pseudoinverse. When policy gradients are estimated from sampled responses, we replace the gradient $\nabla J(\cdot)$ with its unbiased estimator $\widehat{\nabla J}(\cdot)$, which uses a batch of sampled responses to approximate the expected reward, while the KL gradient and the Fisher information matrix are computed exactly. Specifically, let $\widehat{\theta}_t$ be the parameter vector under the sampled NPG update. The update is given by:

$$\widehat{\theta}_{t+1} = \widehat{\theta}_t + \eta \mathcal{F}^\dagger(\widehat{\theta}_t) \widehat{\nabla J}(\widehat{\theta}_t).$$

NPG not only enjoys favorable convergence guarantees under softmax parameterization (Agarwal et al., 2021; Cen et al., 2022), but also connects to modern RL algorithms used in LLM post-training. It can be viewed as maximizing a linearized objective while controlling the KL divergence between consecutive policies (Shani et al., 2020; Cai et al., 2020; Zhong & Zhang, 2023), sharing the principle of limiting policy changes with TRPO (Schulman et al., 2015) and with PPO and GRPO, which are used in LLM post-training (Schulman et al., 2017; Ouyang et al., 2022; Shao et al., 2024). We study NPG to shed light on the training dynamics of RL algorithms used in modern LLM post-training.

**Proposition 3.1.** *Under the softmax parameterization, the policy Jacobian equals the Fisher information matrix; specifically, $\partial \pi_\theta(\cdot \mid x)/\partial \theta^\top = \mathcal{F}(\theta)$. Thus, NPG directions under the exact and*

*sampled settings satisfy*

$$\mathcal{F}^{\dagger}(\theta)\nabla J(\theta) = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{Y}|}\right)\left[\mathbf{r} - \gamma(\theta - \log \pi_{\mathrm{ref}}(\cdot \mid x))\right], \tag{3.3}$$

$$\mathcal{F}^{\dagger}(\widehat{\theta})\widehat{\nabla J}(\widehat{\theta}) = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{Y}|}\right)\left[\frac{1}{B}\sum_{b=1}^{B}\frac{r(x, y_b)}{\pi_{\widehat{\theta}}(y_b \mid x)}\mathbf{e}_{y_b} - \gamma(\widehat{\theta} - \log \pi_{\mathrm{ref}}(\cdot \mid x))\right]. \tag{3.4}$$

*Here $B$ denotes the batch size used to estimate the policy gradient, and $y_1, \ldots, y_B \overset{\text{iid}}{\sim} \pi_{\widehat{\theta}}(\cdot \mid x)$ are the sampled responses. Let $\mathbf{I}$ be the identity matrix, $\mathbf{1}$ the all-ones vector, and $\mathbf{e}_y$ the one-hot vector corresponding to response $y$.*

**Exponential Moving Average (EMA) momentum.** In classical optimization, Polyak's heavy-ball momentum can accelerate convergence under suitable curvature and parameter conditions (Polyak, 1964), and global convergence guarantees are available for smooth convex objectives (Ghadimi et al., 2015). EMA is its normalized form and is widely used in modern LLM training optimizers, such as Adam and AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). EMA maintains a first-moment momentum $m_t$, where the EMA coefficient $\beta \in (0, 1)$ controls the contribution of the current gradient to the parameter update. With exact gradients, the NPG update with EMA is given by

$$m_{t+1} = \beta m_t + (1 - \beta)\mathcal{F}^{\dagger}(\theta_t^{\beta})\nabla J(\theta_t^{\beta}), \qquad \theta_{t+1}^{\beta} = \theta_t^{\beta} + \eta\frac{m_{t+1}}{1 - \beta^{t+1}}. \tag{3.5}$$

The initial condition is $m_0 = 0$ and $\theta_t^{\beta}$ denotes the parameter vector updated by EMA with coefficient $\beta$ at iteration $t$. In the stochastic setting, we replace $\nabla J(\cdot)$ with its unbiased estimator and denote the corresponding parameter vector by $\widehat{\theta}_t^{\beta}$.

Without EMA, the current natural gradient is applied in full to the current parameter update. With EMA, a gradient's contribution is spread across the current and subsequent updates, while each update uses a weighted sum of the current and past natural gradients. Thus, EMA can be viewed as a temporal reallocation of gradient contributions to parameter updates. To isolate this reallocation from changes in the total gradient weight, we normalize $m_t$ by $1 - \beta^t$ for $t \geq 1$ before using it to update the parameters. This makes the gradient weights in each update sum to one, following the first-moment bias correction used in Adam and AdamW.

Let $\omega_\beta(t, t')$ denote the weight of the gradient at time $t'$ in the update at time $t$, under EMA with coefficient $\beta$. For $\beta \in (0, 1)$, (3.5) can be rewritten as

$$\theta_{t+1}^{\beta} = \theta_t^{\beta} + \eta\sum_{t'=0}^{t}\omega_\beta(t, t')\mathcal{F}^{\dagger}(\theta_{t'}^{\beta})\nabla J(\theta_{t'}^{\beta}), \qquad \omega_\beta(t, t') = \frac{(1 - \beta)\beta^{t-t'}}{1 - \beta^{t+1}}.$$

The stochastic update uses the same weights. In particular, we define $\omega_0(t, t') = 0$ for all $t' \neq t$ and $\omega_0(t, t) = 1$, which recovers the standard NPG update without EMA. For consistency, we use the same notation with $\beta = 0$ to denote the corresponding quantities for vanilla NPG without EMA.

## 4 Deterministic case: acceleration with exact gradients

In this section, we consider deterministic optimization with exact policy gradients, where the natural-gradient direction is given by (3.3). Maximizing $J$ in (3.1) yields the unique optimal policy $\pi^{\star}(y \mid x) \propto \pi_{\mathrm{ref}}(y \mid x)\exp(r(x, y)/\gamma)$. For an arbitrary initial parameter $\theta_0$, define

$$\theta^{\star} := \log \pi_{\mathrm{ref}}(\cdot \mid x) + \frac{\mathbf{r}}{\gamma} + \frac{\mathbf{1}^{\top}[\theta_0 - \log \pi_{\mathrm{ref}}(\cdot \mid x) - \mathbf{r}/\gamma]\,\mathbf{1}}{|\mathcal{Y}|},$$

which is one of the maximizers of $J$ and induces $\pi^{\star}$. Here $\mathbf{r}$ is the reward vector and $|\mathcal{Y}|$ is the number of possible responses. We compare the exact NPG iterates obtained with EMA coefficient $\beta$ against those obtained without EMA ($\beta = 0$). Both runs are initialized at $\theta_0$ with $m_0 = 0$ and use the same learning rate $\eta$ and number of updates $T$.

Following the update in (3.5), we recursively define a family of functions $\{f_t : [0, 1) \to \mathbb{R}\}_{t \geq 0}$ by $f_0(\beta) = 1$ and $f_{t+1}(\beta) = f_t(\beta) - \eta\gamma\sum_{t'=0}^{t}\omega_\beta(t, t')f_{t'}(\beta)$ for $t \geq 0$. In particular, for vanilla NPG, $f_t(0) = (1 - \eta\gamma)^t$. The following lemma shows that these functions characterize the natural policy gradient and the parameter error at each iteration.

**Lemma 4.1** (Exact dynamics). *The exact NPG iterates in (3.5) satisfy, for every $t \geq 0$,*

$$\mathcal{F}^{\dagger}(\theta_t^{\beta})\nabla J(\theta_t^{\beta}) = f_t(\beta)\mathcal{F}^{\dagger}(\theta_0)\nabla J(\theta_0), \qquad \theta_t^{\beta} - \theta^{\star} = f_t(\beta)(\theta_0 - \theta^{\star}).$$

The proof is in Appendix B. Lemma 4.1 shows that the exact natural gradients at all iterations are collinear and that the parameter error $\theta_t^{\beta} - \theta^{\star}$ remains a scalar multiple of $\theta_0 - \theta^{\star}$. As long as $f_{t'}(\beta) > 0$ for all $0 \leq t' \leq t$, the iterates up to time $t$ do not overshoot $\theta^{\star}$. Over these iterations, the iterates $\theta_{t'}^{\beta}$ approach $\theta^{\star}$, so the natural gradients point in the same direction and decrease in magnitude. EMA averages the current natural gradient with earlier, larger ones, which can accelerate optimization. The following theorem gives a sufficient range that prevents overshoot and establishes a strict benefit from increasing $\beta$. The proof is in Appendix B.

**Theorem 4.2** (Strict acceleration from momentum). *Assume $0 < \eta\gamma < 1$ and $T \geq 2$. For any $\beta_1, \beta_2$ in the range $[0, ((1-\eta\gamma)/(1+\eta\gamma))^2]$ with $\beta_1 < \beta_2$, we have $J(\theta_T^{\beta_2}) > J(\theta_T^{\beta_1})$. In particular, every positive $\beta$ in this range satisfies $J(\theta_T^{\beta}) > J(\theta_T^{0})$.*

Thus, within the stated range, EMA accelerates deterministic optimization at the same learning rate and update count through a temporal reallocation of natural gradients. This raises two questions: *Why can a strictly positive exact-gradient benefit fail to yield a reliable improvement in sampled training? How can this benefit be reliably realized and exploited in sampled training?*

## 5 Stochastic case: sampling fluctuations and expected gain

We now study the stochastic setting, where natural gradients are estimated from sampled responses, introducing noise into the updates. To analyze training dynamics at small learning rates, we fix the problem, the batch size $B \geq 1$, and the optimization time $\tau = \eta T > 0$, and let $\eta = \tau/T \to 0$. We compare $J(\widehat{\theta}_T^{\beta})$ and $J(\widehat{\theta}_T^{0})$ obtained with and without EMA. The two runs share the same settings, including the initial parameter $\theta_0$, zero initial momentum, learning rate $\eta$, batch size $B$, and update count $T$, while using independent samples even when $\beta = 0$. Expectations and variances are over training randomness. The objective $J$ is evaluated exactly.

By Proposition 3.1, the sampled natural-gradient direction is conditionally unbiased. Taking expectations in the EMA updates then gives $\mathbb{E}[\widehat{\theta}_t^{\beta}] = \theta_t^{\beta}$. To characterize the expected objective, we must also account for sampling fluctuations. Let $\Sigma_T^{\beta}$ denote a covariance coefficient characterizing the accumulated sampling noise along the exact trajectory $\theta_t^{\beta}$. It is a nonnegative weighted sum of noise covariance matrices evaluated along this trajectory, where the noise comes from estimating the reward vector $\mathbf{r}$ using a single sampled response. This coefficient is a deterministic positive semidefinite matrix, independent of the realized samples; its precise definition is given in Appendix C. Using this coefficient, the following lemma expresses the expected objective difference as the exact difference plus a sampling correction.

**Lemma 5.1** (Expected objective difference). *Fix $B \geq 1$ and $\tau > 0$. For any $\beta \in [0, 1)$, as $\eta = \tau/T \to 0$, the expected objective difference satisfies*

$$\mathbb{E}\Big[J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})\Big] = J(\theta_T^{\beta}) - J(\theta_T^{0}) + \frac{\eta}{2B}\operatorname{tr}\Big[\nabla^2 J(\theta_T^{\beta})\Sigma_T^{\beta} - \nabla^2 J(\theta_T^{0})\Sigma_T^{0}\Big] + O((\eta/B)^{3/2}).$$

*The remainder is uniform in $\beta \in [0, 1)$, and $\sup_{\beta\in[0,1)} \|\Sigma_T^{\beta}\|_{\mathrm{op}}$ is bounded independently of $\eta$.*

Establishing the expansion in Lemma 5.1 requires controlling sampling fluctuations throughout training. This is challenging because the sampled natural-gradient direction weights rewards by inverse sampling probabilities: a sampled response with positive reward and very small probability can produce a large noise term. To control this noise, we construct an auxiliary parameter update process whose distance from the exact iterates is bounded. Since the exact trajectory is bounded, the sampling probabilities in this auxiliary process stay bounded away from zero, and its sampling noise is bounded. During training, the probability that the auxiliary process and the original sampled process differ is sufficiently small; therefore, we can characterize the expected objective difference using the auxiliary process. A proof sketch follows and Appendix C gives the full proof.

**Proof sketch.** Let $\varepsilon_t^\beta$ denote the estimation noise in the sampled natural gradient at $\widehat{\theta}_t^\beta$. We construct an auxiliary process driven by the same noise sequence, initialized at $\widetilde{\theta}_0^\beta = \theta_0$ and defined by

$$\widetilde{\varepsilon}_t^\beta = \mathbf{1}\Big\{\max_{0\le s\le t} \|\widetilde{\theta}_s^\beta - \theta_s^\beta\|_\infty \le 1\Big\}\varepsilon_t^\beta, \quad \widetilde{\theta}_{t+1}^\beta = \widetilde{\theta}_t^\beta + \eta \sum_{t'=0}^{t} \omega_\beta(t,t') \left[\mathcal{F}^\dagger(\widetilde{\theta}_{t'}^\beta)\nabla J(\widetilde{\theta}_{t'}^\beta) + \widetilde{\varepsilon}_{t'}^\beta\right].$$

The two processes coincide through the first exit from the unit neighborhood of the exact path. After this exit, the exact direction and EMA memory continue to evolve, while no further sampling noise enters the auxiliary process. The construction gives $\max_{t\le T} \|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_\infty = O(1)$ for every sample path. Since the exact trajectory is bounded, this implies $\inf_{t\le T, y\in\mathcal{Y}} \pi_{\widetilde{\theta}_t^\beta}(y \mid x) = \Omega(1)$ and $\max_{t<T} \|\widetilde{\varepsilon}_t^\beta\|_\infty = O(1)$. These bounds allow us to estimate the covariance and third moment of the auxiliary parameter fluctuations. A second-order Taylor expansion, together with a comparison of the noise covariances along the auxiliary and exact paths, then yields

$$\mathbb{E}[J(\widetilde{\theta}_T^\beta)] = J(\theta_T^\beta) + \frac{\eta}{2B}\operatorname{tr}\left[\nabla^2 J(\theta_T^\beta)\Sigma_T^\beta\right] + O((\eta/B)^{3/2}).$$

For each fixed update and coordinate, the weighted noise contributions have zero conditional mean, so their partial sums form a martingale. Using a concentration bound and a union bound, we show that the probability of path separation is sufficiently small, uniformly in $\beta$. Since $J$ is bounded, we obtain

$$\left|\mathbb{E}[J(\widehat{\theta}_T^\beta)] - \mathbb{E}[J(\widetilde{\theta}_T^\beta)]\right| \le 2\sup_\theta |J(\theta)|\, \mathbb{P}(\exists\, t \le T : \widehat{\theta}_t^\beta \ne \widetilde{\theta}_t^\beta) = o((\eta/B)^{3/2}).$$

Combining these estimates and subtracting the expansions for $\beta$ and 0 proves Lemma 5.1.

Applying the same analysis to $J^2$ also allows us to quantify the fluctuations in the objective difference through its variance.

**Proposition 5.2** (Variance of the objective difference)**.** *Under the conditions of Lemma 5.1, the independent runs satisfy the following expansion uniformly in $\beta \in [0,1)$:*

$$\operatorname{Var}\left[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)\right] = \frac{\eta}{B}\left[\nabla J(\theta_T^\beta)^\top \Sigma_T^\beta \nabla J(\theta_T^\beta) + \nabla J(\theta_T^0)^\top \Sigma_T^0 \nabla J(\theta_T^0)\right] + O((\eta/B)^{3/2}).$$

*Under the nondegeneracy condition, specifically, $\liminf_{\eta\to 0} \nabla J(\theta_T^0)^\top \Sigma_T^0 \nabla J(\theta_T^0) > 0$, uniformly over $\beta \in [0,1)$, we have*

$$\operatorname{Var}\left[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)\right] = \Theta(\eta/B).$$

The proof is in Appendix D. Proposition 5.2 shows that the variance of the objective difference remains of order $\eta/B$ regardless of the EMA coefficient. Lemma 5.1 shows that the expected difference is characterized by the exact objective difference. Thus, when the exact gain is small relative to the standard-deviation scale $\sqrt{\eta/B}$, training fluctuations make a consistent benefit from EMA difficult to observe.

The critical metric is $\rho = \eta/(1-\beta)$. Intuitively, EMA averages natural gradients over an effective memory window of $1/(1-\beta)$ steps. Since each exact update moves the parameters by $O(\eta)$, the averaged gradients span a parameter-displacement scale of $O(\eta/(1-\beta))$, which controls their difference from the current gradient. When this ratio is small, gradients vary little within the window, so EMA behaves similarly to the update without momentum and the difference vanishes. The next two subsections quantify the cases of a vanishing or fixed positive ratio.

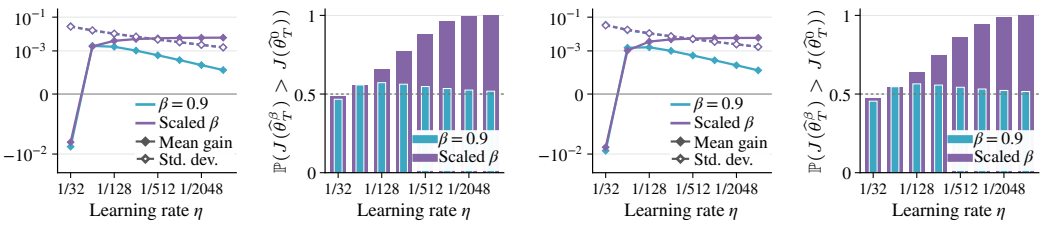
### 5.1 A FIXED EMA COEFFICIENT GIVES A VANISHING EXPECTED GAIN

**Theorem 5.3** (Fixed EMA coefficient)**.** *Fix $\beta \in (0,1)$, $B \ge 1$, $\tau > 0$, and suppose that the nondegeneracy condition in Proposition 5.2 holds. As $\eta = \tau/T \to 0$, the expected objective gain satisfies*

$$\mathbb{E}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)] = C(\tau,\gamma)\frac{\eta\beta}{1-\beta} + O(\eta^{3/2}),$$

*where $C(\tau,\gamma)$ is a positive constant depending only on $\tau$, $\gamma$, and the fixed problem; it is independent of $\beta$. The probability that EMA yields a higher objective value satisfies*

$$\lim_{\eta\to 0} \mathbb{P}\big(J(\widehat{\theta}_T^\beta) > J(\widehat{\theta}_T^0)\big) = 1/2.$$




(a) Gain and Noise. (b) Win Rate. (c) Gain and Noise. (d) Win Rate.

Figure 1: Numerical bandit results with 3 reward-one and 7 reward-zero arms under uniform initialization (1(a), 1(b)) and random policy initialization (1(c), 1(d)).

The proof is in Appendix E. Theorem 5.3 shows that fixed $\beta$ gives a positive expected gain for sufficiently small $\eta$, but the gain is only of order $\eta$, which matches the variance order in Proposition 5.2 with fixed batch size. Sampling fluctuations therefore overwhelm the gain, making the two algorithms difficult to distinguish in individual runs. This offers a possible explanation for the limited benefits of momentum reported in prior RLVR experiments (Mukherjee et al., 2026).

### 5.2 Scaling the EMA coefficient with the learning rate

Beyond Theorem 5.3, we further show that $\eta/(1-\beta) \to 0$ forces the expected gain to vanish, even when $\beta$ depends on $\eta$; the full proof is given in Appendix F. To satisfy the sufficient condition for avoiding overshoot in Theorem 4.2, this ratio must remain bounded. Thus, a nonvanishing expected gain requires $\eta/(1-\beta) = \Theta(1)$. For simplicity, we set $\beta$ so that $\eta/(1-\beta) = \rho$, where $\rho > \eta$ is a positive constant. In Appendix G, we prove that $\rho \in (0, 1/(4\gamma)]$ ensures the sufficient condition in Theorem 4.2. The following theorem compares the objective values of two independent runs with and without EMA in this setting.

**Theorem 5.4** (EMA coefficient scaling with the learning rate)**.** *Fix $B \geq 1$, $\tau > 0$, and $0 < \rho \leq 1/(4\gamma)$, and suppose that the nondegeneracy condition in Proposition 5.2 holds. Set $\beta = 1 - \eta/\rho$. As $\eta = \tau/T \to 0$, the expected objective gain satisfies*

$$\mathbb{E}[J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})] = C(\tau, \gamma, \rho)\beta\rho + O(\eta),$$

*where $C(\tau, \gamma, \rho)$ is a positive constant depending only on $\tau$, $\gamma$, $\rho$, and the fixed problem. It is independent of $\eta$ and $\beta$. The probability that EMA yields a higher objective value satisfies*

$$\mathbb{P}\big(J(\widehat{\theta}_T^{\beta}) > J(\widehat{\theta}_T^{0})\big) = 1 - O(\eta/B).$$

The proof is in Appendix G. When $\rho = \eta/(1-\beta)$ is held at a fixed positive value in the safe regime, Theorem 5.4 gives a nonvanishing expected gain, while Proposition 5.2 shows that the standard deviation of the objective difference is of order $\sqrt{\eta/B}$. As $\eta \to 0$, the expected gain therefore dominates sampling fluctuations, and EMA's advantage becomes consistently observable in the stochastic case. These findings suggest that choosing $\beta$ substantially above the conventional fixed value can yield greater benefits in practice.

### 5.3 Reducing variability across runs with EMA

We next investigate whether EMA can help reduce variability across runs while achieving comparable expected performance under the same training budget. A smaller learning rate can reduce sampling fluctuations, but it may also slow optimization over a given number of updates. The deterministic acceleration established in Section 4 suggests that increasing the EMA coefficient may compensate for this slower progress. We compare two independent EMA runs with settings $(\eta_1, \beta_1)$ and $(\eta_2, \beta_2)$, using the same initialization, batch size $B$, and update count $T$. We write $\theta_T^{(i)}$ and $\widehat{\theta}_T^{(i)}$ for the parameters obtained with exact and sampled gradients, respectively, under setting $i = 1, 2$.

**Theorem 5.5** (Comparable expected performance with lower variance)**.** *Fix $B > 1$ and $\tau > 0$. For any $(\eta_1, \beta_1)$ with $\eta_1 = \tau/T$ and $T \geq 2$, and any $\rho_2$ satisfying $0 < \rho_1 = \eta_1/(1-\beta_1) < \rho_2 \leq 1/(4\gamma)$,*

*there exist* $0 < \eta_2 < \eta_1$ *and* $\beta_1 < \beta_2 < 1$ *such that* $\eta_2/(1-\beta_2) = \rho_2$ *and* $\theta_T^{(2)} = \theta_T^{(1)}$. *As* $\eta_1 \to 0$ *with* $\rho_1, \rho_2$ *fixed, the expected objective difference and the limiting variance ratio satisfy*

$$\mathbb{E}[J(\widehat{\theta}_T^{(2)}) - J(\widehat{\theta}_T^{(1)})] = O(\eta_1/B), \qquad \lim_{\tau\to\infty}\lim_{\eta_1\to 0} \frac{\mathrm{Var}[J(\widehat{\theta}_T^{(2)})]}{\mathrm{Var}[J(\widehat{\theta}_T^{(1)})]} = \frac{1+\sqrt{1-4\gamma\rho_2}}{1+\sqrt{1-4\gamma\rho_1}} < 1.$$

The proof is in Appendix H. Increasing the EMA coefficient compensates for the smaller learning rate by matching the parameters obtained with exact gradients after the same number of updates. With sampled gradients, the expected objective values differ by $O(\eta_1/B)$, while the variance ratio converges to a constant below one in the stated asymptotic regime. Thus, EMA can help reduce variability across runs while achieving comparable expected performance under the same update count $T$ and sample budget $BT$.

## 6 Experiments

We use numerical bandit experiments to quantitatively validate the predictions of Proposition 5.2 and Theorems 5.3 and 5.4. Then we turn to RLVR experiments on Qwen3-4B-Base to test our theoretical predictions in a practical LLM post-training setting. Finally, we apply our principle of increasing the EMA coefficient to Qwen2.5-32B to assess its applicability as model size increases.

### 6.1 Validating theoretical predictions in bandits

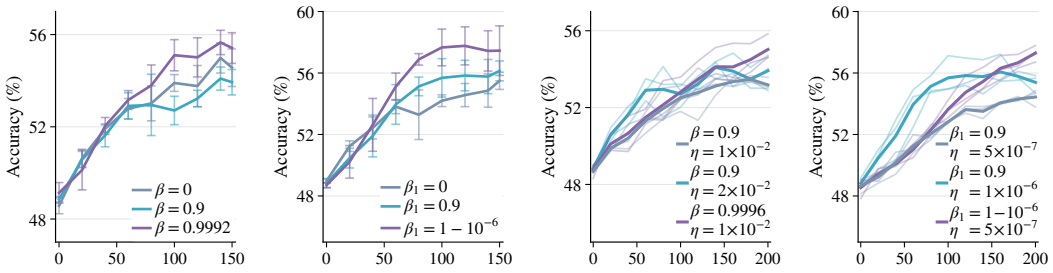
**Setup.** We consider a 10-arm bandit in which 3 arms have reward 1 and the remaining 7 have reward 0. The reference policy is set equal to the initial policy, with KL penalty coefficient $\gamma = 1$, batch size $B = 1$, and optimization time $\tau = 1$. Training uses $T = 32, 64, \ldots, 4096$ updates with learning rate $\eta = 1/T$. We compare sampled NPG with fixed $\beta = 0.9$ or scaled $\beta = 1 - 4\eta$ against the no-EMA baseline. Each configuration uses 50,000 independent training replicates, with $J$ evaluated exactly after training, to estimate the expectation and standard deviation of the objective gap $J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})$ and the probability of outperforming an independent no-EMA run. To measure the gain relative to these fluctuations, we compare SNR of the objective gap, defined as $\mathrm{SNR} := |\mathbb{E}[J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})]| / \mathrm{Std}[J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})]$. Both uniform and random policy initializations are considered. Additional experiments with different reward configurations are provided in Appendix I.1.

**Results.** As the learning rate $\eta$ decreases, Figure 1 shows that the standard deviation scales as $\sqrt{\eta}$ for small $\eta$ under both momentum schedules, as predicted by Proposition 5.2. With fixed $\beta = 0.9$, the mean gain remains positive for small $\eta$ but scales as $\eta$, decreasing faster than the standard deviation. With scaled $\beta = 1 - 4\eta$, it approaches a positive constant, consistent with Theorems 5.3 and 5.4. These trends hold under both uniform and random initialization. At $\eta = 1/4096$, fixed $\beta$ gives SNRs of 0.0453 and 0.0406 under uniform and random initialization, respectively, while scaled $\beta$ gives 3.7647 and 3.3982. (More details are reported in Table 1.) The probability of EMA outperforming an independent no-EMA run approaches $1/2$ with fixed $\beta$ and 1 with scaled $\beta$.

### 6.2 Practical LLM post-training

**Setup.** We conduct RLVR experiments on Qwen3-4B-Base with GRPO (Shao et al., 2024), using the DAPO-17k-processed dataset with 1,000 problems held out for validation. We vary $\beta$ in SGD with bias-corrected EMA and, to support our claim in a more practical setting, also study AdamW by varying $\beta_1$ while fixing $\beta_2 = 0.999$. Setting $\beta = 0$ recovers vanilla SGD, while $\beta_1 = 0$ gives an RMSProp-style optimizer. All configurations share the same data and evaluation protocol. We use three training generation seeds per configuration to show that our findings are consistent. Details of the experimental setup are provided in Appendix I.2.

**Benefits of larger momentum coefficients.** We compare momentum coefficients within each optimizer: $\beta \in \{0, 0.9, 0.9992\}$ for SGD with EMA at $\eta = 2 \times 10^{-2}$ (Figure 2(a)), where 0.9992 is chosen by $\eta/(1-\beta) = 1/(4\gamma)$, matching the upper bound suggested by Theorem 5.4; $\beta_1 \in \{0, 0.9, 1 - 10^{-6}\}$ for AdamW at $\eta = 1 \times 10^{-6}$ (Figure 2(b)). For both optimizers, the tuned larger momentum coefficient substantially improves validation performance over the conventional value of 0.9 and the baseline without first-moment momentum. These results support our claim that choosing momentum coefficients substantially above the conventional value of 0.9 can yield greater benefits in practical RLVR training.




(a) SGD with EMA. (b) AdamW. (c) SGD with EMA. (d) AdamW.

Figure 2: RLVR experiments on Qwen3-4B-Base with GRPO. The horizontal axis shows the training step. 2(a) and 2(b) compare momentum coefficients at a fixed learning rate; lines show mean validation accuracy and error bars show $\pm 1$ sample standard deviation across three seeds. 2(c) and 2(d) compare a smaller learning rate with a larger momentum coefficient against a larger learning rate with the conventional momentum coefficient; light curves show individual runs and dark curves their means.

**More stable training without hurting performance.** For SGD with EMA, we compare $(\eta, \beta) = (1\times10^{-2}, 0.9996)$ with $(2\times10^{-2}, 0.9)$ and include $(1 \times 10^{-2}, 0.9)$ as a baseline (Figure 2(c)). Here, $\beta = 0.9996$ is chosen to satisfy $\eta/(1-\beta) = 1/(4\gamma)$. For AdamW, we compare $(\eta, \beta_1) = (5 \times 10^{-7}, 1 - 10^{-6})$ with $(1 \times 10^{-6}, 0.9)$ and include $(5 \times 10^{-7}, 0.9)$ as a baseline (Figure 2(d)). Under the same update and sample budgets, using a larger momentum coefficient at the smaller learning rate preserves the low variability in final validation accuracy across seeds while achieving slightly better performance than runs using the larger learning rate with conventional momentum. These results show that pairing a larger momentum coefficient with a smaller learning rate makes training more stable while maintaining or even slightly improving performance.

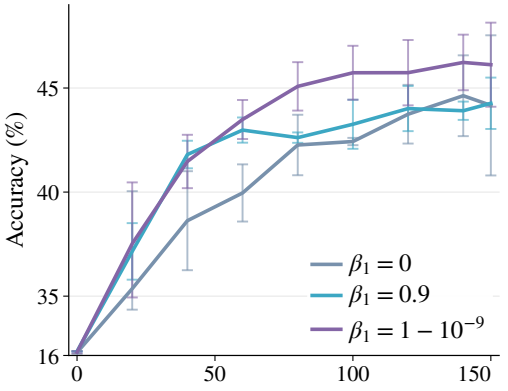



Figure 3: RLVR experiment on Qwen2.5-32B with AdamW. The horizontal axis shows the training step. Lines show means and error bars show $\pm 1$ sample standard deviation across three seeds.

### 6.3 Extending to a larger model

We apply the same GRPO setup to Qwen2.5-32B and increase the momentum coefficient in AdamW from 0.9 to $1 - 10^{-9}$, using a learning rate of $2 \times 10^{-6}$. Simply increasing the momentum coefficient $\beta_1$ substantially improves validation performance. At step 150, the conventional $\beta_1 = 0.9$ achieves a mean accuracy of $44.3\%$, similar to $44.2\%$ for the baseline without first-moment momentum, whereas $\beta_1 = 1 - 10^{-9}$ reaches $46.1\%$, nearly 2 percentage points higher than the conventional setting. Results are shown in Figure 3.

## 7 Conclusion

In this work, we explain why momentum's benefit can be obscured by sampling noise and derive a principle for recovering and exploiting it. We develop a theoretical model and prove that EMA strictly improves the objective value under exact gradients within a range that prevents overshoot. With sampled gradients, fixing $\beta$ as $\eta$ decreases makes the expected gain vanish faster than sampling fluctuations, whereas suitably increasing $\beta$ within the safe regime produces a consistent improvement. GRPO experiments on LLMs confirm the practical benefits of larger momentum coefficients. These findings suggest using coefficients substantially above the conventional $\beta = 0.9$ to better exploit the first-moment information already maintained by the optimizer.

### AI USE STATEMENT

We used generative AI to assist with some mathematical derivations, literature searches, proof checking, language editing, and LaTeX formatting. The authors carefully reviewed all AI-assisted material and independently verified the mathematical arguments. The authors take full responsibility for this paper.

## REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4): 2563–2578, 2022.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yen-Ju Chen, Nai-Chieh Huang, Ching-Pei Lee, and Ping-Chun Hsieh. Accelerated policy gradient: On the convergence rates of the Nesterov momentum for reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6881–6949, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global optimum convergence of momentum-based policy gradient. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1910–1934, 2022.

Chongyu Fan, Gaowen Liu, Mingyi Hong, Ramana Rao Kompella, and Sijia Liu. Rethinking Muon beyond pretraining: Spectral failures and high-pass remedies for VLA and RLVR. *arXiv preprint arXiv:2605.19282*, 2026.

Jie Feng, Ke Wei, and Jinchi Chen. Global convergence of natural policy gradient with hessian-aided momentum variance reduction. *Journal of Scientific Computing*, 101(2):44, 2024.

Damien Ferbach, Courtney Paquette, Gauthier Gidel, Katie Everett, and Elliot Paquette. Logarithmic-time schedules for scaling language models with momentum. *arXiv preprint arXiv:2602.05298*, 2026.

Aleksandar Fontana, Marco Simoni, Giulio Rossolini, Paolo Mori, and Andrea Saracino. On the hidden objective biases of group-based reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 109–121. Association for Computational Linguistics, 2026.

Jingchu Gai, Guanning Zeng, Huaqing Zhang, Han Zhong, Yige Hong, Andrej Risteski, and Aditi Raghunathan. Demystifying entropy control in LLM RL training: Theoretical analysis and dynamic scheduling. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pp. 310–315. IEEE, 2015.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4422–4433, 2020.

Feihu Huang, Shangqian Gao, and Heng Huang. Bregman gradient policy optimization. In *International Conference on Learning Representations*, 2022.

Sham M. Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Sagnik Mukherjee, Lifan Yuan, Pavan Jayasinha, Dilek Hakkani-Tür, and Hao Peng. Do we need adam? surprisingly strong and sparse reinforcement learning with SGD in LLMs. In *Forty-third International Conference on Machine Learning*, 2026.

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

Matteo Pagliardini, Pierre Ablin, and David Grangier. The AdEMAMix optimizer: Better, faster, older. In *International Conference on Learning Representations*, 2025.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015.

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

Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5668–5675, 2020.

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

Kexuan Shi, Hanxuan Li, Zeju Qiu, Yandong Wen, Simon Buchholz, and Weiyang Liu. Pion: A spectrum-preserving optimizer via orthogonal equivalence transformation. *arXiv preprint arXiv:2605.12492*, 2026.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Vatsal Venkatkrishna, Nico Daheim, and Iryna Gurevych. Parameter exploration for RLVR via variational learning. *arXiv preprint arXiv:2608.09805*, 2026.

Runzhe Wang, Sadhika Malladi, Tianhao Wang, Kaifeng Lyu, and Zhiyuan Li. The marginal value of momentum for small learning rate SGD. In *International Conference on Learning Representations*, 2024.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54715–54754, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244, 2025.

Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016.

Huishuai Zhang, Bohan Wang, and Luoxin Chen. AdamS: Momentum itself can be a normalizer for LLM pretraining and post-training. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10719–10738. Association for Computational Linguistics, 2025.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for autoregressive language models. In *International Conference on Learning Representations*, 2025.

Han Zhong and Tong Zhang. A theoretical analysis of optimistic proximal policy optimization in linear markov decision processes. *Advances in Neural Information Processing Systems*, 36: 73666–73690, 2023.

Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. DPO meets PPO: Reinforced token optimization for RLHF. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78498–78521, 2025.

Hanqing Zhu, Wenyan Cong, Zhizhou Sha, Sagnik Mukherjee, Xinyuan Song, David González-Martínez, Xiaoxia Wu, Yuandong Tian, Shiwei Liu, David Z. Pan, and Zhangyang "Atlas" Wang. ISO: An RLVR-native optimization stack. *arXiv preprint arXiv:2607.19331*, 2026.

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

## A Proof of Proposition 3.1

*Proof.* **Policy Jacobian and Fisher matrix.** Differentiating the softmax parameterization, we have

$$\frac{\partial \pi_\theta(y \mid x)}{\partial \theta(x, y')} = \pi_\theta(y \mid x)\big(\mathbf{1}_{\{y=y'\}} - \pi_\theta(y' \mid x)\big), \quad \nabla \log \pi_\theta(y \mid x) = \mathbf{e}_y - \pi_\theta(\cdot \mid x),$$

where $\mathbf{e}_y$ is the one-hot vector corresponding to $y \in \mathcal{Y}$. Substituting the second identity into the definition of the Fisher information matrix in (3.2) and comparing with the first gives

$$\mathcal{F}(\theta) = \operatorname{diag}\big(\pi_\theta(\cdot \mid x)\big) - \pi_\theta(\cdot \mid x)\pi_\theta(\cdot \mid x)^\top = \frac{\partial \pi_\theta(\cdot \mid x)}{\partial \theta^\top}. \tag{A.1}$$

Since the policy has full support, $\mathcal{F}(\theta)$ is positive semidefinite with null space $\operatorname{span}\{\mathbf{1}\}$. Hence $\mathcal{F}^\dagger(\theta)\mathcal{F}(\theta) = \mathbf{I} - \mathbf{1}\mathbf{1}^\top/|\mathcal{Y}|$.

**Exact and sampled natural gradients.** By the chain rule, differentiating (3.1) gives

$$\begin{aligned}
\nabla J(\theta) &= \Big(\frac{\partial \pi_\theta(\cdot \mid x)}{\partial \theta^\top}\Big)^\top \mathbf{r} - \gamma \sum_{y \in \mathcal{Y}} \Big(\log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} + 1\Big)\nabla \pi_\theta(y \mid x) \\
&= \mathcal{F}(\theta)\left[\mathbf{r} - \gamma\left(\log \pi_\theta(\cdot \mid x) - \log \pi_{\mathrm{ref}}(\cdot \mid x) + \mathbf{1}\right)\right] \\
&= \mathcal{F}(\theta)\left[\mathbf{r} - \gamma\big(\theta - \log \pi_{\mathrm{ref}}(\cdot \mid x)\big)\right].
\end{aligned} \tag{A.2}$$

The second equality follows from (A.1). The last uses $\log \pi_\theta(\cdot \mid x) = \theta - \log(\sum_{y \in \mathcal{Y}} \exp(\theta(x, y)))\mathbf{1}$ and $\mathcal{F}(\theta)\mathbf{1} = 0$. For the sampled gradient, the KL term is evaluated exactly, so

$$\begin{aligned}
\widehat{\nabla J}(\widehat{\theta}) &= \frac{1}{B}\sum_{b=1}^{B} r(x, y_b)\nabla \log \pi_{\widehat{\theta}}(y_b \mid x) - \gamma \nabla \operatorname{KL}\big(\pi_{\widehat{\theta}}(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x)\big) \\
&= \mathcal{F}(\widehat{\theta})\left[\frac{1}{B}\sum_{b=1}^{B} \frac{r(x, y_b)}{\pi_{\widehat{\theta}}(y_b \mid x)}\mathbf{e}_{y_b} - \gamma\big(\widehat{\theta} - \log \pi_{\mathrm{ref}}(\cdot \mid x)\big)\right].
\end{aligned} \tag{A.3}$$

The second equality uses $\mathcal{F}(\widehat{\theta})\mathbf{e}_{y_b}/\pi_{\widehat{\theta}}(y_b \mid x) = \mathbf{e}_{y_b} - \pi_{\widehat{\theta}}(\cdot \mid x)$. Premultiplying (A.2) and (A.3) by the corresponding Fisher pseudoinverse proves (3.3) and (3.4). □

## B Proofs of Lemma 4.1 and Theorem 4.2

**Optimality of $\theta^\star$.** The definition of $\theta^\star$ gives $\pi_{\theta^\star}(y \mid x) \propto \pi_{\mathrm{ref}}(y \mid x)e^{r(x,y)/\gamma}$. Substituting this identity into (3.1), we obtain

$$J(\theta^\star) - J(\theta) = \gamma \operatorname{KL}\big(\pi_\theta(\cdot \mid x) \| \pi_{\theta^\star}(\cdot \mid x)\big) \geq 0.$$

Equality holds exactly when $\pi_\theta = \pi_{\theta^\star}$, which verifies the unique optimal policy used in Section 4.

*Proof of Lemma 4.1.* Substituting the definition of $\theta^\star$ into (3.3) gives

$$\mathcal{F}^\dagger(\theta_t^\beta)\nabla J(\theta_t^\beta) = \gamma\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^\top}{|\mathcal{Y}|}\right)(\theta^\star - \theta_t^\beta).$$

By Proposition 3.1, every natural-gradient direction has coordinate sum zero. Starting from $m_0 = 0$, the EMA update therefore gives $\mathbf{1}^\top m_t = 0$ at every step, so the parameter update preserves $\mathbf{1}^\top \theta_t^\beta = \mathbf{1}^\top \theta_0$. By construction, $\mathbf{1}^\top \theta^\star = \mathbf{1}^\top \theta_0$ as well. Thus $\mathbf{1}^\top(\theta^\star - \theta_t^\beta) = 0$, and the preceding identity simplifies to

$$\mathcal{F}^\dagger(\theta_t^\beta)\nabla J(\theta_t^\beta) = \gamma(\theta^\star - \theta_t^\beta). \tag{B.1}$$

**Scalar residual dynamics.** We prove the parameter identity by induction. It holds at $t = 0$ since $f_0(\beta) = 1$. If it holds through iteration $t$, unrolling (3.5) yields

$$\begin{aligned}
\theta_{t+1}^\beta - \theta^\star &= \theta_t^\beta - \theta^\star - \eta\gamma \sum_{t'=0}^{t} \omega_\beta(t, t')(\theta_{t'}^\beta - \theta^\star) \\
&= \Big[f_t(\beta) - \eta\gamma \sum_{t'=0}^{t} \omega_\beta(t, t') f_{t'}(\beta)\Big](\theta_0 - \theta^\star) = f_{t+1}(\beta)(\theta_0 - \theta^\star).
\end{aligned}$$

The last equality is the defining recurrence of $f_t$. Substituting the parameter identity into (B.1) proves the natural-gradient identity. □

*Proof of Theorem 4.2.* **Positive residuals in the safe range.** We first prove that $f_t(\beta) > 0$ for all $t \geq 0$ when $0 < \eta\gamma < 1$ and $0 \leq \beta \leq ((1-\eta\gamma)/(1+\eta\gamma))^2$. Separating the current term in the defining recurrence of $f_t$ and substituting the preceding update gives, for $t \geq 1$,

$$\begin{aligned} f_{t+1}(\beta) &= f_t(\beta) - \frac{\eta\gamma}{\sum_{t'=0}^{t}\beta^{t'}}\Big[f_t(\beta) + \beta\sum_{t'=0}^{t-1}\beta^{t-1-t'}f_{t'}(\beta)\Big] \\ &= f_t(\beta) - \frac{\eta\gamma}{\sum_{t'=0}^{t}\beta^{t'}}f_t(\beta) + \frac{\beta\sum_{t'=0}^{t-1}\beta^{t'}}{\sum_{t'=0}^{t}\beta^{t'}}\big[f_t(\beta) - f_{t-1}(\beta)\big] \\ &= (1-\eta\gamma)f_t(\beta) + \frac{\beta\sum_{t'=0}^{t-1}\beta^{t'}}{\sum_{t'=0}^{t}\beta^{t'}}\big[(1+\eta\gamma)f_t(\beta) - f_{t-1}(\beta)\big]. \end{aligned}$$

Dividing it by $f_t(\beta)$ gives

$$\frac{f_{t+1}(\beta)}{f_t(\beta)} = 1 - \eta\gamma + \frac{\beta\sum_{t'=0}^{t-1}\beta^{t'}}{\sum_{t'=0}^{t}\beta^{t'}}\left(1 + \eta\gamma - \frac{f_{t-1}(\beta)}{f_t(\beta)}\right). \tag{B.2}$$

We prove $\sqrt{\beta} \leq f_t(\beta)/f_{t-1}(\beta) \leq 1 - \eta\gamma$ by induction. Since $f_1(\beta)/f_0(\beta) = 1 - \eta\gamma$, the claim holds at $t = 1$. If it holds at $t$, then

$$0 \leq \frac{\beta\sum_{t'=0}^{t-1}\beta^{t'}}{\sum_{t'=0}^{t}\beta^{t'}} < \beta, \qquad 1 + \eta\gamma - \frac{f_{t-1}(\beta)}{f_t(\beta)} \leq -\frac{(\eta\gamma)^2}{1-\eta\gamma} < 0.$$

Thus (B.2) gives the upper bound. For the lower bound,

$$\begin{aligned} \frac{f_{t+1}(\beta)}{f_t(\beta)} - \sqrt{\beta} &\geq 1 - \eta\gamma + \beta\left(1 + \eta\gamma - \frac{1}{\sqrt{\beta}}\right) - \sqrt{\beta} \\ &= (1-\sqrt{\beta})\big[(1-\eta\gamma) - (1+\eta\gamma)\sqrt{\beta}\big] \geq 0. \end{aligned}$$

The last inequality follows from the assumed range of $\beta$. Consequently $f_{t+1}(\beta) \geq \sqrt{\beta}f_t(\beta)$, completing the induction. Since $f_0(\beta) = 1 > 0$, we have $f_t(\beta) > 0$ for all $t \geq 0$. At $\beta = 0$, $f_t(0) = (1-\eta\gamma)^t > 0$.

**Strict decrease with the EMA coefficient.** Differentiating (B.2) with respect to $\beta$ gives

$$\partial_\beta\left(\frac{f_{t+1}(\beta)}{f_t(\beta)}\right) = \frac{\beta\sum_{t'=0}^{t-1}\beta^{t'}}{\sum_{t'=0}^{t}\beta^{t'}}\left(\frac{f_{t-1}(\beta)}{f_t(\beta)}\right)^2\partial_\beta\left(\frac{f_t(\beta)}{f_{t-1}(\beta)}\right) + \frac{\sum_{t'=1}^{t}t'\beta^{t'-1}}{(\sum_{t'=0}^{t}\beta^{t'})^2}\left(1 + \eta\gamma - \frac{f_{t-1}(\beta)}{f_t(\beta)}\right).$$

The first coefficient is nonnegative, and the second term is strictly negative for $t \geq 1$. Starting from $\partial_\beta(f_1(\beta)/f_0(\beta)) = \partial_\beta(1-\eta\gamma) = 0$, induction therefore gives $\partial_\beta(f_{t+1}(\beta)/f_t(\beta)) < 0$. Differentiating $f_T(\beta) = \prod_{t=1}^{T} f_t(\beta)/f_{t-1}(\beta)$ gives

$$\partial_\beta f_T(\beta) = f_T(\beta)\sum_{t=1}^{T}\frac{f_{t-1}(\beta)}{f_t(\beta)}\partial_\beta\left(\frac{f_t(\beta)}{f_{t-1}(\beta)}\right) < 0, \qquad T \geq 2.$$

Applying the ratio lower bound at the largest admissible $\beta$ and using this monotonicity gives

$$\left(\frac{1-\eta\gamma}{1+\eta\gamma}\right)^T \leq f_T(\beta) \leq (1-\eta\gamma)^T < 1. \tag{B.3}$$

Hence $0 < f_T(\beta_2) < f_T(\beta_1) < 1$ whenever $\beta_1 < \beta_2$ in the stated range.

**Strict improvement of the objective.** Lemma 4.1 and (A.2) give

$$\begin{aligned} \partial_\beta J(\theta_T^\beta) &= \partial_\beta f_T(\beta)(\theta_0 - \theta^\star)^\top \nabla J(\theta_T^\beta) \\ &= -\gamma f_T(\beta)\partial_\beta f_T(\beta)(\theta_0 - \theta^\star)^\top \mathcal{F}(\theta_T^\beta)(\theta_0 - \theta^\star) > 0. \end{aligned}$$

The scalar factor $-\gamma f_T(\beta)\partial_\beta f_T(\beta)$ is positive. The quadratic form is also positive since $\theta_0 - \theta^\star$ is nonzero and lies outside the null space $\mathrm{span}\{\mathbf{1}\}$ of $\mathcal{F}(\theta_T^\beta)$. □

## C Proof of Lemma 5.1

The proof follows the auxiliary process in the main text: bound its noise, estimate its moments, and transfer the objective expansion to the original process. Throughout this section, the bounds are uniform over $\beta \in [0, 1)$, with the problem, $B$, and $\tau$ fixed.

*Proof.* **Sampling noise and its propagation.** Let $y^{\beta}_{t,b}$ be the $b$-th response at update $t$. Conditional on the sample history, the batch consists of independent responses from $\pi_{\widehat{\theta}^{\beta}_t}(\cdot \mid x)$. Let $\varepsilon^{\beta}_{t,b}$ denote the estimation noise from the $b$-th response at update $t$. Then we have

$$\varepsilon^{\beta}_{t,b} = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\top}}{|\mathcal{Y}|}\right)\left[\frac{r(x, y^{\beta}_{t,b})}{\pi_{\widehat{\theta}^{\beta}_t}(y^{\beta}_{t,b} \mid x)}\mathbf{e}_{y^{\beta}_{t,b}} - \mathbf{r}\right], \qquad \varepsilon^{\beta}_{t} = \frac{1}{B}\sum_{b=1}^{B}\varepsilon^{\beta}_{t,b}. \tag{C.1}$$

Since the sampled directions have zero coordinate sum and the initial momentum is zero, the updates preserve $\mathbf{1}^{\top}\widehat{\theta}^{\beta}_t = \mathbf{1}^{\top}\theta_0 = \mathbf{1}^{\top}\theta^{\star}$. Therefore, the sampled direction at update $t$ in Proposition 3.1 can be written as $\gamma(\theta^{\star} - \widehat{\theta}^{\beta}_t) + \varepsilon^{\beta}_t$. Each $\varepsilon^{\beta}_{t,b}$ has conditional mean zero, so the expected direction is $\gamma(\theta^{\star} - \mathbb{E}[\widehat{\theta}^{\beta}_t])$. Thus, after taking expectations, the sampled EMA updates obey the same recurrence as the exact updates. With the same initial parameter and zero initial momentum, this gives $\mathbb{E}[\widehat{\theta}^{\beta}_t] = \theta^{\beta}_t$ by induction.

Let $\Psi_{\beta}(t, j)$ denote the weight of the noise $\varepsilon^{\beta}_j$ in the difference between the sampled and exact parameters, so that $\widehat{\theta}^{\beta}_t - \theta^{\beta}_t = \sum_{j<t}\Psi_{\beta}(t, j)\varepsilon^{\beta}_j$. Subtracting the exact EMA update from the sampled EMA update gives

$$\widehat{\theta}^{\beta}_{t+1} - \theta^{\beta}_{t+1} = \widehat{\theta}^{\beta}_t - \theta^{\beta}_t + \eta\sum_{k=0}^{t}\omega_{\beta}(t, k)\left[-\gamma(\widehat{\theta}^{\beta}_k - \theta^{\beta}_k) + \varepsilon^{\beta}_k\right].$$

The noise at update $j$ first affects the parameter at time $j + 1$, so $\Psi_{\beta}(t, j) = 0$ for $t \le j$. Substituting the noise expansion above into this update and collecting the terms multiplying $\varepsilon^{\beta}_j$ gives

$$\Psi_{\beta}(t + 1, j) = \Psi_{\beta}(t, j) - \gamma\eta\sum_{k=j}^{t}\omega_{\beta}(t, k)\Psi_{\beta}(k, j) + \eta\omega_{\beta}(t, j), \qquad t \ge j. \tag{C.2}$$

The last term captures the contribution of $\varepsilon^{\beta}_j$ retained in the EMA memory. The sum captures how this noise changes later parameter values and hence the exact natural-gradient directions evaluated at those values. According to (C.2), $\Psi_{\beta}(t, j)$ is a deterministic function of $\eta, \gamma, \beta, t, j$ and does not depend on the realized samples. The covariance coefficient in Lemma 5.1 is

$$\Sigma^{\beta}_T = \frac{1}{\eta}\sum_{t<T}\Psi_{\beta}(T, t)^2\left[\operatorname{diag}\left(\frac{\mathbf{r}}{\pi_{\theta^{\beta}_t}(\cdot \mid x)}\right) - \mathbf{r}\mathbf{r}^{\top}\right]. \tag{C.3}$$

The division inside diag is elementwise. The bracket is the covariance of $r(x, y)\mathbf{e}_y/\pi_{\theta^{\beta}_t}(y \mid x)$ under $y \sim \pi_{\theta^{\beta}_t}(\cdot \mid x)$, using $r(x, y)^2 = r(x, y)$. Thus $\Sigma^{\beta}_T$ is deterministic and positive semidefinite. The response coefficients $\Psi_{\beta}(t, j)$ are bounded in the next lemma. The proof is delayed to Section C.1.

**Lemma C.1** (Uniform response bounds)**.** *Let $f_t$ be defined in Section 4 and $\Psi_{\beta}(t, j)$ by (C.2). For $\beta \in [0, 1)$ and integers $t \ge 0$ with $t\eta \le \tau$, the following bounds hold, with $0 \le j < t$ in the pointwise bound on $\Psi_{\beta}(t, j)$:*

$$|f_t(\beta)| \le e^{\gamma\tau}, \qquad |\Psi_{\beta}(t, j)| \le e^{\gamma\tau}\eta\left(1 + \sum_{k=j+1}^{t}\frac{1}{k}\right),$$
$$\sum_{j<t}|\Psi_{\beta}(t, j)| \le \tau e^{\gamma\tau}, \qquad \sum_{j<t}\Psi_{\beta}(t, j)^2 \le 5e^{2\gamma\tau}t\eta^2. \tag{C.4}$$

**The auxiliary process.** Use the same noise sequence as the original process and initialize $\widetilde{\theta}_0^\beta = \theta_0$. As in the main text, define

$$\widetilde{\varepsilon}_t^\beta = \mathbf{1}\Big\{\max_{0\le s\le t}\|\widetilde{\theta}_s^\beta - \theta_s^\beta\|_\infty \le 1\Big\}\varepsilon_t^\beta, \qquad \widetilde{\theta}_{t+1}^\beta = \widetilde{\theta}_t^\beta + \eta\sum_{t'=0}^{t}\omega_\beta(t,t')\big[\mathcal{F}^\dagger(\widetilde{\theta}_{t'}^\beta)\nabla J(\widetilde{\theta}_{t'}^\beta) + \widetilde{\varepsilon}_{t'}^\beta\big].$$

For the bounds below, let $\zeta^\beta = \inf\{t \ge 1 : \|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_\infty > 1\}$ denote the first exit time, taking $\zeta^\beta = \infty$ if no exit occurs. The two processes agree before exit; afterward, only fresh noise is suppressed. Since the stopping indicator is known before sampling, $\widetilde{\varepsilon}_t^\beta$ has conditional mean zero, and taking expectations in the auxiliary updates yields $\mathbb{E}[\widetilde{\theta}_t^\beta] = \theta_t^\beta$.

The bounds on probabilities and noise can be made explicit. Define $M = \|\theta^\star\|_\infty + e^{\gamma\tau}\|\theta_0 - \theta^\star\|_\infty + 1$ and $L = 2(|\mathcal{Y}|e^{2M} + 1)$. Lemma 4.1 gives $\theta_t^\beta = \theta^\star + f_t(\beta)(\theta_0 - \theta^\star)$, and Lemma C.1 gives $|f_t(\beta)| \le e^{\gamma\tau}$ for $t \le T$. Therefore,

$$\|\theta_t^\beta\|_\infty \le \|\theta^\star\|_\infty + |f_t(\beta)|\|\theta_0 - \theta^\star\|_\infty \le M - 1.$$

On $\{t < \zeta^\beta\}$, the two processes agree and $\|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_\infty \le 1$, so $\|\widetilde{\theta}_t^\beta\|_\infty \le \|\theta_t^\beta\|_\infty + 1 \le M$. Then we have $\pi_{\widetilde{\theta}_t^\beta}(y \mid x) \ge 1/(|\mathcal{Y}|\exp(2M))$. According to the definition of the sampling noise in (C.1), on $\{t < \zeta^\beta\}$, we have $\|\varepsilon_{t,b}^\beta\|_\infty \le 2\big(1/\pi_{\widetilde{\theta}_t^\beta}(y_{t,b}^\beta \mid x) + 1\big) \le 2\big(|\mathcal{Y}|e^{2M} + 1\big) = L$. Applying (C.2) to the auxiliary process and using (C.4) gives for every sample path

$$\max_{t\le T}\|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_\infty \le \max_{t\le T}\sum_{j<t}|\Psi_\beta(t,j)|\|\widetilde{\varepsilon}_j^\beta\|_\infty \le L\tau e^{\gamma\tau}. \tag{C.5}$$

In particular, the auxiliary parameters remain in a fixed bounded region even after exit.

**Exit probability.** Fix a time $1 \le t \le T$. We first bound the first coordinate of the parameter difference. Reveal responses one at a time, in update order and within each batch. The partial sums of

$$(\widetilde{\theta}_t^\beta)_1 - (\theta_t^\beta)_1 = \sum_{j<t}\sum_{b=1}^{B}\frac{\Psi_\beta(t,j)}{B}\mathbf{1}_{\{j<\zeta^\beta\}}(\varepsilon_{j,b}^\beta)_1$$

form a martingale: the indicator and parameter are known before the sample is revealed, and its conditional mean is zero. Let $\mathscr{H}_{j,b-1}$ denote the information available before response $b$ of batch $j$ is revealed. Applying Freedman's inequality (Tropp, 2011, Theorem 1.1) to both signs gives, for every $u > 0$,

$$\mathbb{P}\big(|(\widetilde{\theta}_t^\beta)_1 - (\theta_t^\beta)_1| > u\big) \le 2\exp\left[-\frac{Bu^2}{2\eta\big\{5\tau e^{2\gamma\tau}|\mathcal{Y}|L^2 + e^{\gamma\tau}L(2+\log T)u/3\big\}}\right]. \tag{C.6}$$

Here we used the following bounds on each increment and the sum of conditional second moments, obtained from Lemma C.1:

$$\left|\frac{\Psi_\beta(t,j)}{B}\mathbf{1}_{\{j<\zeta^\beta\}}(\varepsilon_{j,b}^\beta)_1\right| \le \frac{L}{B}|\Psi_\beta(t,j)| \le e^{\gamma\tau}L(2+\log T)\frac{\eta}{B},$$

$$\sum_{j<t}\sum_{b=1}^{B}\mathbb{E}\left[\left(\frac{\Psi_\beta(t,j)}{B}\mathbf{1}_{\{j<\zeta^\beta\}}(\varepsilon_{j,b}^\beta)_1\right)^2 \,\middle|\, \mathscr{H}_{j,b-1}\right] \le \frac{|\mathcal{Y}|L^2}{B}\sum_{j<t}\Psi_\beta(t,j)^2 \le 5\tau e^{2\gamma\tau}|\mathcal{Y}|L^2\frac{\eta}{B}.$$

The same bound holds for every coordinate. The processes agree through the update causing the first exit. Denote $C = 10\tau e^{2\gamma\tau}|\mathcal{Y}|L^2 + 4e^{\gamma\tau}L/3$. Taking $u = 1$ and a union bound over updates and coordinates gives

$$\mathbb{P}(\exists\, t \le T : \widehat{\theta}_t^\beta \ne \widetilde{\theta}_t^\beta) \le \mathbb{P}(\zeta^\beta \le T) \le 2|\mathcal{Y}|T\exp\left[-\frac{B}{C\eta\log(eT)}\right]. \tag{C.7}$$

The bound is uniform in $\beta$.

**Moment bounds.** We next bound the moments of the auxiliary parameter fluctuations. Since the noise increments have conditional mean zero, cross terms between distinct increments vanish in

expectation, giving

$$\mathbb{E}\|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_2^2 = \frac{1}{B^2}\sum_{j<t}\sum_{b=1}^{B}\Psi_\beta(t,j)^2\mathbb{E}\Big[\mathbf{1}_{\{j<\zeta^\beta\}}\|\varepsilon_{j,b}^\beta\|_2^2\Big]$$
$$\le \frac{|\mathcal{Y}|L^2}{B}\sum_{j<t}\Psi_\beta(t,j)^2 \le 5\tau e^{2\gamma\tau}|\mathcal{Y}|L^2\frac{\eta}{B}. \tag{C.8}$$

We also need a third-moment bound to control the remainder in the second-order expansion of $J$ below. Using the constant $C$ above, (C.6) implies

$$\mathbb{P}\big(|(\widetilde{\theta}_t^\beta)_1 - (\theta_t^\beta)_1| > u\big) \le 2\exp\!\left(-\frac{Bu^2}{2C\eta}\right) + 2\exp\!\left(-\frac{Bu}{2C\eta\log(eT)}\right).$$

Integrating the tail probability against $3u^2\,\mathrm{d}u$ gives the third moment. If the Euclidean norm exceeds $u$, at least one of the $|\mathcal{Y}|$ coordinates has magnitude above $u/\sqrt{|\mathcal{Y}|}$. A union bound and the coordinate tail bounds therefore give

$$\max_{t\le T}\mathbb{E}\|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_2^3 \le 6|\mathcal{Y}|\int_0^\infty u^2\left[\exp\!\left(-\frac{Bu^2}{2C|\mathcal{Y}|\eta}\right) + \exp\!\left(-\frac{Bu}{2C\sqrt{|\mathcal{Y}|}\eta\log(eT)}\right)\right]\mathrm{d}u$$
$$= O\big((\eta/B)^{3/2} + (\eta\log(eT)/B)^3\big) = O((\eta/B)^{3/2}). \tag{C.9}$$

The last equality uses $\sqrt{\eta/B}\log(eT) \to 0$.

**Covariance along the exact trajectory.** We next identify the quadratic term in the expansion of $J$. We compare the single-response covariance matrices in (C.3) along the auxiliary and exact trajectories. New sampling noise enters the auxiliary process only before exit, so the expected replacement error is

$$\mathbb{E}\left\|\mathbf{1}_{\{t<\zeta^\beta\}}\left[\operatorname{diag}\!\left(\frac{\mathbf{r}}{\pi_{\widetilde{\theta}_t^\beta}(\cdot\mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right] - \left[\operatorname{diag}\!\left(\frac{\mathbf{r}}{\pi_{\theta_t^\beta}(\cdot\mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right]\right\|_{\mathrm{op}}. \tag{C.10}$$

On $\{t < \zeta^\beta\}$, the two $\mathbf{r}\mathbf{r}^\top$ terms cancel. Since rewards are binary, the remaining diagonal matrix has operator norm at most the largest difference between the inverse probabilities. Both parameters have norm at most $M$, so the identity $\pi_\theta(y\mid x)^{-1} = \sum_{y'} e^{\theta(x,y')-\theta(x,y)}$ and the mean value theorem give

$$\max_y\big|\pi_{\widetilde{\theta}_t^\beta}(y\mid x)^{-1} - \pi_{\theta_t^\beta}(y\mid x)^{-1}\big| \le 2|\mathcal{Y}|e^{2M}\|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_\infty.$$

On $\{t \ge \zeta^\beta\}$, the auxiliary covariance is zero, so the difference is bounded by the exact-trajectory covariance:

$$\left\|\operatorname{diag}\!\left(\frac{\mathbf{r}}{\pi_{\theta_t^\beta}(\cdot\mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right\|_{\mathrm{op}} \le |\mathcal{Y}|(e^{2M}+1).$$

Splitting the expectation between these two events bounds the replacement error in (C.10) by

$$2|\mathcal{Y}|e^{2M}\mathbb{E}\big[\mathbf{1}_{\{t<\zeta^\beta\}}\|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_\infty\big] + |\mathcal{Y}|(e^{2M}+1)\mathbb{P}(\zeta^\beta \le t)$$
$$\le 2|\mathcal{Y}|e^{2M}\sqrt{\mathbb{E}\|\widetilde{\theta}_t^\beta - \theta_t^\beta\|_2^2} + |\mathcal{Y}|(e^{2M}+1)\mathbb{P}(\zeta^\beta \le T)$$
$$\le 2|\mathcal{Y}|e^{2M}\sqrt{5\tau e^{2\gamma\tau}|\mathcal{Y}|L^2\eta/B} + |\mathcal{Y}|(e^{2M}+1)\mathbb{P}(\zeta^\beta \le T). \tag{C.11}$$

The first inequality uses Cauchy–Schwarz, and the second uses (C.8).

The Hessian $\nabla^2 J(\theta_T^\beta)$ is uniformly bounded along the exact trajectory and satisfies $\nabla^2 J(\theta_T^\beta)\mathbf{1} = 0$ by the shift invariance of $J$. Thus the projection in (C.1) does not change the corresponding quadratic form. Conditional independence within each batch contributes a factor $1/B$, giving

$$\mathbb{E}\big[(\widetilde{\varepsilon}_t^\beta)^\top\nabla^2 J(\theta_T^\beta)\widetilde{\varepsilon}_t^\beta \mid \mathscr{H}_{t,0}\big] = \frac{\mathbf{1}_{\{t<\zeta^\beta\}}}{B}\operatorname{tr}\left[\nabla^2 J(\theta_T^\beta)\left(\operatorname{diag}\!\left(\frac{\mathbf{r}}{\pi_{\widetilde{\theta}_t^\beta}(\cdot\mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right)\right].$$

Subtracting the corresponding exact-trajectory term and taking expected absolute values gives

$$\begin{aligned}
&\mathbb{E}\left|\mathbb{E}[(\widetilde{\varepsilon}_t^\beta)^\top \nabla^2 J(\theta_T^\beta)\widetilde{\varepsilon}_t^\beta \mid \mathscr{H}_{t,0}] - \frac{1}{B}\operatorname{tr}\left[\nabla^2 J(\theta_T^\beta)\left(\operatorname{diag}\left(\frac{\mathbf{r}}{\pi_{\theta_t^\beta}(\cdot \mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right)\right]\right| \\
&\quad \le \frac{|\mathcal{Y}|}{B}\|\nabla^2 J(\theta_T^\beta)\|_{\mathrm{op}}\mathbb{E}\left\|\mathbf{1}_{\{t<\zeta^\beta\}}\left[\operatorname{diag}\left(\frac{\mathbf{r}}{\pi_{\widetilde{\theta}_t^\beta}(\cdot \mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right] - \left[\operatorname{diag}\left(\frac{\mathbf{r}}{\pi_{\theta_t^\beta}(\cdot \mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right]\right\|_{\mathrm{op}} \\
&\quad = O\left(\frac{1}{B}\left[\sqrt{\eta/B} + \mathbb{P}(\zeta^\beta \le T)\right]\right). \qquad \text{(C.12)}
\end{aligned}$$

The inequality bounds the trace by $|\mathcal{Y}|$ times the product of the operator norms. The last equality uses the upper bound in (C.11) and the boundedness of the Hessian. Substituting the noise expansion of $\widetilde{\theta}_T^\beta - \theta_T^\beta$ into the quadratic term, we use the zero conditional mean of the noise to remove the cross terms between distinct updates:

$$\begin{aligned}
&\mathbb{E}\big[(\widetilde{\theta}_T^\beta - \theta_T^\beta)^\top \nabla^2 J(\theta_T^\beta)(\widetilde{\theta}_T^\beta - \theta_T^\beta)\big] \\
&\quad = \sum_{t<T} \Psi_\beta(T,t)^2 \mathbb{E}[(\widetilde{\varepsilon}_t^\beta)^\top \nabla^2 J(\theta_T^\beta)\widetilde{\varepsilon}_t^\beta] \\
&\quad = \frac{1}{B}\sum_{t<T} \Psi_\beta(T,t)^2 \operatorname{tr}\left[\nabla^2 J(\theta_T^\beta)\left(\operatorname{diag}\left(\frac{\mathbf{r}}{\pi_{\theta_t^\beta}(\cdot \mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right)\right] \\
&\quad + \sum_{t<T} \Psi_\beta(T,t)^2 O\left(\frac{1}{B}\left[\sqrt{\eta/B} + \mathbb{P}(\zeta^\beta \le T)\right]\right) \qquad \text{(C.13)} \\
&\quad = \frac{\eta}{B}\operatorname{tr}[\nabla^2 J(\theta_T^\beta)\Sigma_T^\beta] + O((\eta/B)^{3/2}). \qquad \text{(C.14)}
\end{aligned}$$

Line (C.13) uses the bound from (C.12). Line (C.14) substitutes the definition of $\Sigma_T^\beta$ in (C.3) and uses $\sum_{t<T}\Psi_\beta(T,t)^2 = O(\eta)$. Using (C.4) and the covariance bound above in (C.3), we obtain $\sup_{\beta\in[0,1)}\|\Sigma_T^\beta\|_{\mathrm{op}} \le 5\tau e^{2\gamma\tau}|\mathcal{Y}|(e^{2M}+1)$, where the upper bound is independent of $\eta$ and $\beta$.

**Objective expansion and transfer.** All derivatives of $J$ through order three are bounded on the region containing the exact and auxiliary paths. Taylor's formula and the third-moment bound in (C.9) give an expected remainder bounded by

$$\frac{1}{6}\sup_{\|\vartheta\|_\infty \le M + L\tau e^{\gamma\tau}} \|\nabla^3 J(\vartheta)\|_{\mathrm{op}}\mathbb{E}\|\widetilde{\theta}_T^\beta - \theta_T^\beta\|_2^3 = O((\eta/B)^{3/2}).$$

The bounded region includes every segment joining the exact and auxiliary endpoints. Using $\mathbb{E}[\widetilde{\theta}_T^\beta] = \theta_T^\beta$ to remove the linear term, (C.14) for the quadratic term, and (C.9) for the remainder, we obtain

$$\mathbb{E}[J(\widetilde{\theta}_T^\beta)] = J(\theta_T^\beta) + \frac{\eta}{2B}\operatorname{tr}[\nabla^2 J(\theta_T^\beta)\Sigma_T^\beta] + O((\eta/B)^{3/2}).$$

The objective is globally bounded: $-\gamma\log(1/\min_y \pi_{\mathrm{ref}}(y \mid x)) \le J \le 1$. Consequently, (C.7) gives

$$|\mathbb{E}[J(\widehat{\theta}_T^\beta)] - \mathbb{E}[J(\widetilde{\theta}_T^\beta)]| \le 2\sup_\theta |J(\theta)|\,\mathbb{P}(\zeta^\beta \le T) = o((\eta/B)^{3/2}).$$

Thus each original run satisfies

$$\mathbb{E}[J(\widehat{\theta}_T^\beta)] = J(\theta_T^\beta) + \frac{\eta}{2B}\operatorname{tr}[\nabla^2 J(\theta_T^\beta)\Sigma_T^\beta] + O((\eta/B)^{3/2}). \qquad \text{(C.15)}$$

Subtracting this expansion at $\beta$ and 0 proves the lemma. □

### C.1 Proof of Lemma C.1

*Proof.* **Residual and single-noise response.** By the defining recurrence of $f_t(\beta)$ in Section 4, the triangle inequality gives

$$|f_{t+1}(\beta)| = \left| f_t(\beta) - \gamma\eta \sum_{k=0}^{t} \omega_\beta(t,k) f_k(\beta) \right| \le |f_t(\beta)| + \gamma\eta \sum_{k=0}^{t} \omega_\beta(t,k) |f_k(\beta)|$$
$$\le \max_{0\le k\le t} |f_k(\beta)| + \gamma\eta \max_{0\le k\le t} |f_k(\beta)| \sum_{k=0}^{t} \omega_\beta(t,k) = (1+\gamma\eta) \max_{0\le k\le t} |f_k(\beta)|,$$

where we used $\omega_\beta(t,k) \ge 0$ and $\sum_{k=0}^{t} \omega_\beta(t,k) = 1$. The same upper bound also covers all earlier $|f_k(\beta)|$, so

$$\max_{0\le k\le t+1} |f_k(\beta)| \le (1+\gamma\eta) \max_{0\le k\le t} |f_k(\beta)|.$$

With $f_0(\beta) = 1$, we have $|f_t(\beta)| \le \max_{0\le k\le t} |f_k(\beta)| \le (1+\gamma\eta)^t \le e^{\gamma\eta t} \le e^{\gamma\tau}$.

Next, we bound $\Psi_\beta(t,j)$. Fixing $j$ and applying the triangle inequality to (C.2) yields

$$|\Psi_\beta(s+1,j)| \le |\Psi_\beta(s,j)| + \gamma\eta \sum_{k=j}^{s} \omega_\beta(s,k) |\Psi_\beta(k,j)| + \eta\omega_\beta(s,j)$$
$$\le (1+\gamma\eta) \max_{j\le r\le s} |\Psi_\beta(r,j)| + \eta\omega_\beta(s,j).$$

The second inequality uses $\omega_\beta(s,k) \ge 0$ for all $k$ and $\sum_{k=j}^{s} \omega_\beta(s,k) \le 1$. This also bounds $\max_{j\le r\le s+1} |\Psi_\beta(r,j)|$. Starting from $\Psi_\beta(j,j) = 0$ and iterating gives

$$|\Psi_\beta(t,j)| \le \eta \sum_{k=j}^{t-1} (1+\gamma\eta)^{t-1-k} \omega_\beta(k,j) \le e^{\gamma\tau}\eta \sum_{k=j}^{t-1} \omega_\beta(k,j).$$

For $0 < \beta < 1$, using $\sum_{\ell=0}^{k} \beta^\ell \ge (k+1)\beta^k$ gives

$$\sum_{k=j}^{t-1} \omega_\beta(k,j) = \sum_{k=j}^{t-1} \frac{\beta^{k-j}}{\sum_{\ell=0}^{k} \beta^\ell} = \sum_{k=j}^{t-1} \left[ (1-\beta)\beta^{k-j} + \frac{\beta^{2k+1-j}}{\sum_{\ell=0}^{k} \beta^\ell} \right]$$
$$\le 1 - \beta^{t-j} + \sum_{k=j}^{t-1} \frac{\beta^{k+1-j}}{k+1} \le 1 + \sum_{k=j+1}^{t} \frac{1}{k}.$$

For $\beta = 0$, this sum equals 1, so the same bound holds. Combining these inequalities proves the claimed bound on $|\Psi_\beta(t,j)|$.

**Summed responses.** Summing over $j$ and interchanging the two sums gives

$$\sum_{j=0}^{t-1} |\Psi_\beta(t,j)| \le e^{\gamma\tau}\eta \sum_{j=0}^{t-1} \sum_{k=j}^{t-1} \omega_\beta(k,j) = e^{\gamma\tau}\eta \sum_{k=0}^{t-1} \sum_{j=0}^{k} \omega_\beta(k,j) = e^{\gamma\tau}\eta \sum_{k=0}^{t-1} 1 = e^{\gamma\tau} t\eta \le e^{\gamma\tau}\tau.$$

For the squares, expanding the harmonic sums yields $\sum_{j=0}^{t-1} \left(1 + \sum_{k=j+1}^{t} 1/k\right)^2 = 5t - \sum_{k=1}^{t} \frac{1}{k} \le 5t$. Combining this identity with the response envelope proves (C.4). □

## D Proof of Proposition 5.2

*Proof.* **Individual variances.** Apply the expansion proved in Appendix C to $J^2$. Since $\nabla^2(J^2) = 2J\nabla^2 J + 2\nabla J \nabla J^\top$, it gives

$$\mathbb{E}[J(\widehat{\theta}_T^\beta)^2] = J(\theta_T^\beta)^2 + \frac{\eta}{B} J(\theta_T^\beta) \operatorname{tr}[\nabla^2 J(\theta_T^\beta) \Sigma_T^\beta] + \frac{\eta}{B} \nabla J(\theta_T^\beta)^\top \Sigma_T^\beta \nabla J(\theta_T^\beta) + O((\eta/B)^{3/2}).$$

Subtracting the square of (C.15) cancels the Hessian term and yields, uniformly in $\beta$,

$$\mathrm{Var}[J(\widehat{\theta}_T^\beta)] = \frac{\eta}{B}\nabla J(\theta_T^\beta)^\top \Sigma_T^\beta \nabla J(\theta_T^\beta) + O((\eta/B)^{3/2}). \tag{D.1}$$

**Independent comparison.** The two variances add by independence, proving the expansion in Proposition 5.2. Its two leading quadratic forms are nonnegative and uniformly bounded. Under the stated nondegeneracy condition, the no-EMA term is bounded away from zero as $\eta \to 0$. □

## E Proof of Theorem 5.3

The mean comparison requires one more order of accuracy than Lemma 5.1 alone: the two leading sampling corrections cancel when $\beta$ is fixed. The winning probability then follows from a normal limit, since the mean is smaller than the standard deviation.

*Proof.* **Exact objective gain.** We first record the approximations needed below; their proof is in the next subsection.

**Lemma E.1** (Approximations for a fixed EMA coefficient). *Let $f_t$ be defined in Section 4 and $\Sigma_T^\beta$ by (C.3). For fixed $\beta \in [0,1)$ and $\tau > 0$, as $\eta = \tau/T \to 0$, for all $0 \le t \le T$, the following approximations hold*

$$f_t(\beta) = \left(1 - \gamma\eta - \gamma^2\eta^2\frac{\beta}{1-\beta}\right)^t + O(\eta^2), \qquad \|\Sigma_T^\beta - \Sigma_T^0\|_{\mathrm{op}} = O(\sqrt{\eta}). \tag{E.1}$$

The first estimate, together with $f_T(0) = (1-\gamma\eta)^T$, gives

$$\begin{aligned} f_T(0) - f_T(\beta) &= (1-\gamma\eta)^T - \left(1 - \gamma\eta - \gamma^2\eta^2\frac{\beta}{1-\beta}\right)^T + O(\eta^2) \\ &= T\gamma^2\eta^2\frac{\beta}{1-\beta}(1-\gamma\eta)^{T-1} + O(T^2\eta^4 + \eta^2) \\ &= \gamma^2\tau e^{-\gamma\tau}\frac{\eta\beta}{1-\beta} + O(\eta^2). \end{aligned}$$

The second line is Taylor's formula for the power $a \mapsto a^T$; both bases lie in $(0,1)$ for sufficiently small $\eta$. The last line uses $T\eta = \tau$ and $(1-\gamma\eta)^{T-1} = e^{-\gamma\tau} + O(\eta)$. Along the exact line, (A.2) yields for $a > 0$,

$$\frac{\mathrm{d}}{\mathrm{d}a}J(\theta^\star + a(\theta_0 - \theta^\star)) = -\gamma a\,\mathrm{Var}_{y\sim\pi_{\theta^\star + a(\theta_0-\theta^\star)}(\cdot|x)}[\theta_0(x,y) - \theta^\star(x,y)] < 0. \tag{E.2}$$

Lemma E.1 and $T\eta = \tau$ give $f_T(0) = e^{-\gamma\tau} + O(\eta)$ and $f_T(\beta) = e^{-\gamma\tau} + O(\eta)$. Integrating (E.2) from $a = f_T(\beta)$ to $a = f_T(0)$ gives

$$\begin{aligned} J(\theta_T^\beta) - J(\theta_T^0) &= \gamma\int_{f_T(\beta)}^{f_T(0)} a\,\mathrm{Var}_{y\sim\pi_{\theta^\star + a(\theta_0-\theta^\star)}(\cdot|x)}[\theta_0(x,y) - \theta^\star(x,y)]\,\mathrm{d}a \\ &= \gamma e^{-\gamma\tau}\,\mathrm{Var}_{y\sim\pi_{\theta^\star + e^{-\gamma\tau}(\theta_0-\theta^\star)}}[\theta_0(x,y) - \theta^\star(x,y)][f_T(0) - f_T(\beta)] + O(\eta^2) \\ &= C(\tau,\gamma)\frac{\eta\beta}{1-\beta} + O(\eta^2). \end{aligned} \tag{E.3}$$

The integrand varies by $O(\eta)$ over an interval of length $O(\eta)$, which explains the second-line remainder. Thus the constant suppressed in the main theorem is explicitly

$$C(\tau,\gamma) = \gamma^3\tau e^{-2\gamma\tau}\,\mathrm{Var}_{y\sim\pi_{\theta^\star + e^{-\gamma\tau}(\theta_0-\theta^\star)}(\cdot|x)}[\theta_0(x,y) - \theta^\star(x,y)].$$

This expression specifies its dependence on the problem, $\tau$, and $\gamma$; it is independent of $\eta$, $\beta$, and $B$. The coefficient is strictly positive under the nondegeneracy condition, which implies $\theta_0 \neq \theta^\star$.

**Cancellation of sampling corrections.** Using $f_t(0) = (1-\gamma\eta)^t$, Lemma E.1 and the mean value theorem give $|f_t(\beta) - f_t(0)| \le t\gamma^2\eta^2\beta/(1-\beta) + O(\eta^2) = O(\eta)$ uniformly for $0 \le t \le T$, since

$t\eta \le \tau$ and both bases lie in $(0,1)$ for sufficiently small $\eta$. Lemma 4.1 then gives $\|\theta_t^\beta - \theta_t^0\|_\infty = |f_t(\beta) - f_t(0)|\|\theta_0 - \theta^\star\|_\infty$, and hence $\max_{t\le T}\|\theta_t^\beta - \theta_t^0\|_\infty = O(\eta)$. The Hessian is Lipschitz on the bounded region containing these paths. In particular, its Lipschitz constant is bounded by the supremum of $\|\nabla^3 J\|_{\mathrm{op}}$ on that region, so $\|\nabla^2 J(\theta_T^\beta) - \nabla^2 J(\theta_T^0)\|_{\mathrm{op}} = O(\eta)$. Adding and subtracting $\nabla^2 J(\theta_T^0)\Sigma_T^\beta$ gives

$$\begin{aligned}&\left|\mathrm{tr}[\nabla^2 J(\theta_T^\beta)\Sigma_T^\beta - \nabla^2 J(\theta_T^0)\Sigma_T^0]\right| \\ &\quad\le |\mathcal{Y}|\|\nabla^2 J(\theta_T^\beta) - \nabla^2 J(\theta_T^0)\|_{\mathrm{op}}\|\Sigma_T^\beta\|_{\mathrm{op}} + |\mathcal{Y}|\|\nabla^2 J(\theta_T^0)\|_{\mathrm{op}}\|\Sigma_T^\beta - \Sigma_T^0\|_{\mathrm{op}} \\ &\quad= O(\eta) + O(\sqrt{\eta}).\end{aligned}$$

Lemma 5.1 and (E.3) therefore yield

$$\begin{aligned}\mathbb{E}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)] &= C(\tau,\gamma)\frac{\eta\beta}{1-\beta} + O(\eta^2) + O(\eta^{3/2}/B) + O((\eta/B)^{3/2}) \\ &= C(\tau,\gamma)\frac{\eta\beta}{1-\beta} + O(\eta^{3/2}).\end{aligned}$$

**Probability of improvement.** The mean-to-standard-deviation ratio alone does not determine the winning probability. We also need the following normal limit, proved in the next subsection by a direct characteristic-function comparison.

**Lemma E.2** (Asymptotic normality of the objective difference)**.** *Under the assumptions of Theorem 5.3,*

$$\frac{J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0) - \mathbb{E}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)]}{\mathrm{Std}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)]} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1). \tag{E.4}$$

The mean expansion and Proposition 5.2 give $\mathbb{E}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)]/\mathrm{Std}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)] = O(\sqrt{B\eta}) \to 0$. Combining this estimate with the normal limit in Lemma E.2 gives $\mathbb{P}(J(\widehat{\theta}_T^\beta) > J(\widehat{\theta}_T^0)) \to 1/2$. □

### E.1 Proofs of the auxiliary lemmas

*Proof of Lemma E.1.* **Residual expansion.** The scalar recursion and Lemma C.1 give $|f_{t+1}(\beta) - f_t(\beta)| \le \gamma\eta e^{\gamma\tau}$. The mean lag of the geometric weights is at most $\beta/(1-\beta)$, so their average differs from $f_t(\beta)$ by $O(\eta)$. Summing the updates over $k$ preceding steps gives

$$f_{t-k}(\beta) - f_t(\beta) = \gamma\eta \sum_{i=t-k}^{t-1}\sum_{j=0}^{i}\omega_\beta(i,j)f_j(\beta) = \gamma\eta k f_t(\beta) + O(\eta^2(k+k^2)).$$

For the second equality, replace each average by $f_i(\beta)$ with error $O(\eta)$, then use $|f_i(\beta) - f_t(\beta)| \le \gamma\eta e^{\gamma\tau}(t-i)$. Average this identity with weights proportional to $\beta^k$. Their second moments are bounded for fixed $\beta$, and their mean is $\beta/(1-\beta) - (t+1)\beta^{t+1}/(1-\beta^{t+1})$; explicitly, their second moments are at most $\beta(1+\beta)/(1-\beta)^2$. Therefore

$$\sum_{j=0}^{t}\omega_\beta(t,j)f_j(\beta) - f_t(\beta) = \gamma\eta f_t(\beta)\left[\frac{\beta}{1-\beta} - \frac{(t+1)\beta^{t+1}}{1-\beta^{t+1}}\right] + O(\eta^2),$$

$$f_{t+1}(\beta) = \left(1 - \gamma\eta - \gamma^2\eta^2\frac{\beta}{1-\beta}\right)f_t(\beta) + O(\eta^3 + \eta^2(t+1)\beta^{t+1}).$$

The startup terms have total size $O(\eta^2)$ because $\sum_{t\ge0}(t+1)\beta^{t+1} = \beta/(1-\beta)^2$. The remaining errors have the same order because $T\eta^3 = \tau\eta^2$. For sufficiently small $\eta$, the coefficient multiplying $f_t$ lies in $(0,1)$, so propagating these errors cannot increase their total absolute size. Iteration proves the first estimate in (E.1). It also gives $\max_{t\le T}|f_t(\beta) - f_t(0)| = O(\eta)$.

**Covariance comparison.** Summing the geometric weights gives, for $t > j$,

$$\begin{aligned}\sum_{k=j}^{t-1}\omega_\beta(k,j) &= 1 - \beta^{t-j} + \sum_{k=j}^{t-1}\frac{(1-\beta)\beta^{2k-j+1}}{1-\beta^{k+1}} \\ &= 1 - \beta^{t-j} + O(\beta^{j+1}).\end{aligned}$$

The last sum is nonnegative and at most $\beta^{j+1}/(1-\beta^2)$, uniformly in $t$. Sum the response recursion (C.2) over time and interchange its finite sums:

$$\Psi_\beta(t,j) = \eta \sum_{k=j}^{t-1} \omega_\beta(k,j) - \gamma\eta \sum_{k=j+1}^{t-1} \left( \sum_{\ell=k}^{t-1} \omega_\beta(\ell,k) \right) \Psi_\beta(k,j).$$

At $\beta = 0$, the parenthesized sums equal one and $\Psi_0(t,j) = \eta(1-\gamma\eta)^{t-j-1}$. Subtracting the two identities and dividing by $\eta$ yields

$$\left| \frac{\Psi_\beta(t,j) - \Psi_0(t,j)}{\eta} \right| \le C_\beta(\beta^{t-j} + \beta^{j+1} + \eta) + C_\beta\gamma\eta \sum_{k=j+1}^{t-1} \left| \frac{\Psi_\beta(k,j) - \Psi_0(k,j)}{\eta} \right|.$$

Here the weight errors accumulated against the no-EMA response are $O(\eta)$ because both geometric tails have bounded sums. Discrete Gronwall now gives

$$\left| \frac{\Psi_\beta(t,j) - \Psi_0(t,j)}{\eta} \right| \le C_\beta(\beta^{t-j} + \beta^{j+1} + \eta), \qquad \eta \sum_{j<T} \left| \frac{\Psi_\beta(T,j) - \Psi_0(T,j)}{\eta} \right|^2 = O(\eta). \tag{E.5}$$

For the first bound, the direct weight error contributes the two geometric terms; accumulated errors against the no-EMA response contribute $O(\eta)$. Gronwall preserves the constant terms, and summing its decaying forcing adds only $O(\eta)$.

The covariance matrices in (C.3) are bounded and Lipschitz on the exact paths. Subtract their two weighted sums, use $|a^2 - b^2| = |a-b||a+b|$, and apply Cauchy–Schwarz with (C.4) and (E.5):

$$\|\Sigma_T^\beta - \Sigma_T^0\|_{\mathrm{op}} \le C_\beta \left[ \left( \eta \sum_{j<T} \left| \frac{\Psi_\beta(T,j) - \Psi_0(T,j)}{\eta} \right|^2 \right)^{1/2} + \max_{j \le T} \|\theta_j^\beta - \theta_j^0\|_\infty \right] = O(\sqrt{\eta}).$$

This proves the second estimate. □

*Proof of Lemma E.2.* **Martingale contributions.** For this proof, write $s_T = \mathrm{Std}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)]$; by Proposition 5.2, $s_T^2 = \Theta(\eta/B)$. For the EMA run, define the normalized single-response contribution and its deterministic comparison variance by

$$Y_{t,b}^\beta = \frac{\Psi_\beta(T,t)}{B s_T} \mathbf{1}_{\{t<\zeta^\beta\}} \nabla J(\theta_T^\beta)^\top \varepsilon_{t,b}^\beta,$$

$$\nu_{t,b}^\beta = \frac{\Psi_\beta(T,t)^2}{B^2 s_T^2} \mathrm{Var}_{y\sim\pi_{\theta_t^\beta}(\cdot|x)} \left[ \frac{r(x,y)}{\pi_{\theta_t^\beta}(y \mid x)} \frac{\partial J}{\partial \theta(x,y)}(\theta_T^\beta) \right].$$

Use the corresponding quantities for the no-EMA run, with the same denominator $s_T$. Reveal the independent runs consecutively, negate the no-EMA contributions, and index the resulting arrays by $Y_\ell, \nu_\ell$. Their partial sums form a martingale. By (C.2), their total sum is

$$\sum_\ell Y_\ell = \frac{\nabla J(\theta_T^\beta)^\top(\widetilde{\theta}_T^\beta - \theta_T^\beta) - \nabla J(\theta_T^0)^\top(\widetilde{\theta}_T^0 - \theta_T^0)}{s_T}.$$

The projection in the single-response errors does not change these scalar contributions because $\mathbf{1}^\top \nabla J(\theta) = 0$. By (C.3) and Proposition 5.2,

$$\sum_\ell \nu_\ell = \frac{\eta}{B s_T^2} \left[ \nabla J(\theta_T^\beta)^\top \Sigma_T^\beta \nabla J(\theta_T^\beta) + \nabla J(\theta_T^0)^\top \Sigma_T^0 \nabla J(\theta_T^0) \right] = 1 + O(\sqrt{\eta/B}). \tag{E.6}$$

**Conditional variance and higher moments.** By (E.5), $\max_{t<T}(|\Psi_\beta(T,t)| + |\Psi_0(T,t)|) \le C_\beta\eta$. Combining this estimate with the bounded noise in the auxiliary process gives

$$\max_\ell |Y_\ell| \le C_\beta\sqrt{\eta/B}, \qquad \max_\ell \nu_\ell \le C_\beta\eta/B, \qquad \sum_\ell \nu_\ell^2 = O(\eta/B). \tag{E.7}$$

The single-response covariance bound used to derive (C.12) also controls the conditional variance of each noise contribution along an endpoint gradient. Both endpoint gradients are uniformly bounded. Multiplying the bound by $\Psi_\beta(T,t)^2/(B^2 s_T^2)$ for the EMA run and by $\Psi_0(T,t)^2/(B^2 s_T^2)$ for the no-EMA run, then summing over all responses, gives

$$\sum_\ell \mathbb{E}\left|\mathbb{E}[Y_\ell^2 \mid \mathscr{H}_{\ell-1}] - \nu_\ell\right| \le C_\beta\left[\sqrt{\eta/B} + \mathbb{P}(\zeta^\beta \le T) + \mathbb{P}(\zeta^0 \le T)\right]$$
$$= O(\sqrt{\eta/B}), \tag{E.8}$$

where $\mathscr{H}_{\ell-1}$ contains all previously revealed responses in the concatenated array. Thus $\sum_\ell \mathbb{E}Y_\ell^2 = 1 + O(\sqrt{\eta/B})$. Multiplying this bounded sum by the maximum increment in (E.7) also gives $\sum_\ell \mathbb{E}|Y_\ell|^3 = O(\sqrt{\eta/B})$.

**Direct normal approximation.** Replace the increments successively by independent centered Gaussians with variances $\nu_\ell$, independent of both runs. For each real $z$, conditional Taylor expansion through order two and telescoping give

$$\left|\mathbb{E}e^{iz\sum_\ell Y_\ell} - e^{-z^2\sum_\ell \nu_\ell/2}\right| \le \frac{z^2}{2}\sum_\ell \mathbb{E}\left|\mathbb{E}[Y_\ell^2 \mid \mathscr{H}_{\ell-1}] - \nu_\ell\right|$$
$$+ \frac{|z|^3}{6}\sum_\ell \mathbb{E}|Y_\ell|^3 + \frac{z^4}{8}\sum_\ell \nu_\ell^2 = O_z(\sqrt{\eta/B}).$$

The first-order terms vanish by conditional centering. In each replacement, earlier increments remain unchanged and later terms are independent Gaussians, so the argument does not require independence of the actual increments. Equations (E.6)–(E.8) show that the characteristic function converges to $e^{-z^2/2}$. Hence $\sum_\ell Y_\ell \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$.

**Returning to the sampled objectives.** For each auxiliary objective, the first-order Taylor remainder has expected absolute value at most $\frac{1}{2}\sup\|\nabla^2 J\|_{\mathrm{op}}\mathbb{E}\|\widetilde{\theta}_T^\beta - \theta_T^\beta\|_2^2 = O(\eta/B)$, where the supremum is over the bounded region in (C.5). Dividing by $s_T$ makes this remainder $O(\sqrt{\eta/B})$ in $L^1$. The exit bound (C.7) transfers the normal limit to the original objectives. Finally, Lemma 5.1 gives

$$\frac{\left|\mathbb{E}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)] - [J(\theta_T^\beta) - J(\theta_T^0)]\right|}{s_T} = O(\sqrt{\eta/B}).$$

Replacing exact centering by mean centering therefore leaves the limit unchanged, proving (E.4). □

## F Vanishing expected gain as $\eta/(1-\beta) \to 0$

We prove the claim preceding Theorem 5.4 for arbitrary deterministic $\beta = \beta(\eta) \in [0,1)$.

**Lemma F.1** (Vanishing expected gain). *Fix $B \ge 1$ and $\tau > 0$, and consider the two runs in Section 5. For any deterministic $\beta = \beta(\eta) \in [0,1)$, if $\eta/(1-\beta) \to 0$ as $\eta = \tau/T \to 0$, then*

$$\mathbb{E}[J(\widehat{\theta}_T^\beta) - J(\widehat{\theta}_T^0)] \longrightarrow 0.$$

*Proof.* **Exact trajectory difference.** Lemma C.1 and the scalar update give $|f_{t+1}(\beta) - f_t(\beta)| \le \gamma\eta e^{\gamma\tau}$. Since the geometric weights have mean lag at most $\beta/(1-\beta)$,

$$\left|\sum_{j=0}^t \omega_\beta(t,j) f_j(\beta) - f_t(\beta)\right| \le \gamma e^{\gamma\tau}\frac{\eta\beta}{1-\beta}.$$

Rewrite the scalar update as the no-EMA update plus this error. For sufficiently small $\eta$, $0 < 1-\gamma\eta < 1$; iterating the difference from $f_t(0) = (1-\gamma\eta)^t$ gives

$$\max_{t\le T}|f_t(\beta) - f_t(0)| \le \gamma^2\tau e^{\gamma\tau}\frac{\eta\beta}{1-\beta}, \qquad |J(\theta_T^\beta) - J(\theta_T^0)| \le C\frac{\eta\beta}{1-\beta}.$$

The second bound uses Lemma 4.1 and the bounded gradient of $J$ on the exact paths. The constants are independent of $\beta$.

**Expected gain.** Lemma 5.1 contributes only $O(\eta/B)$, uniformly in $\beta$. Hence, the expected difference can be bounded by

$$\left|\mathbb{E}[J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})]\right| \le C\left(\frac{\eta\beta}{1-\beta} + \frac{\eta}{B}\right).$$

The upper bound goes to zero when $\eta/(1-\beta)$ and $\eta$ both go to zero. In addition, the sufficient condition in Theorem 4.2 implies $\eta/(1-\beta) \le (1+\gamma\eta)^2/(4\gamma)$. Under this condition, a gain bounded away from zero therefore requires $\eta/(1-\beta) = \Theta(1)$. □

## G Proof of Theorem 5.4

We identify the limiting exact gain at fixed $\rho$, then use Lemma 5.1 and Proposition 5.2 to control the sampled comparison.

*Proof.* **Continuous residual and discrete approximation.** Here $\beta = 1 - \eta/\rho$ varies with $\eta$, so the fixed-coefficient approximations do not apply directly. For fixed $\rho > 0$, we introduce a continuous residual $f(s;\rho)$ at optimization time $s \ge 0$, corresponding to $f_t(\beta)$ at $s = t\eta$. It describes the remaining displacement from $\theta^\star$ through $\theta(s;\rho) = \theta^\star + f(s;\rho)(\theta_0 - \theta^\star)$ and is defined by

$$\partial_s f(s;\rho) = -\gamma \frac{\int_0^s e^{-(s-u)/\rho} f(u;\rho)\,\mathrm{d}u}{\rho(1-e^{-s/\rho})}, \qquad f(0;\rho) = 1. \tag{G.1}$$

The normalized average at $s = 0$ is interpreted by continuity. The following approximation is proved in the next subsection.

**Lemma G.1** (Continuous approximation of the exact path)**.** *Let $f_t$ be defined in Section 4 and $f(s;\rho)$ by (G.1). For $0 < \eta < \rho$, $\beta = 1 - \eta/\rho$, and $T\eta \le \tau$,*

$$\max_{t \le T} |f_t(\beta) - f(t\eta;\rho)| \le \frac{7}{2}\gamma e^{\gamma\tau}(e^{\gamma\tau} - 1)\eta. \tag{G.2}$$

For the no-EMA path, $(1-\gamma\eta)^T = e^{-\gamma\tau} + O(\eta)$. The approximation above and smoothness of $J$ then give

$$J(\theta_T^{\beta}) - J(\theta_T^{0}) = J(\theta(\tau;\rho)) - J(\theta^\star + e^{-\gamma\tau}(\theta_0 - \theta^\star)) + O(\eta). \tag{G.3}$$

**Positive gain in the safe range.** For $0 < \rho \le 1/(4\gamma)$ and $\eta < \rho$,

$$0 < \beta = 1 - \eta/\rho \le 1 - 4\gamma\eta \le \left(\frac{1-\gamma\eta}{1+\gamma\eta}\right)^2.$$

Thus Theorem 4.2 applies. Passing to the limit in the lower bound of (B.3) gives $f(s;\rho) > 0$. The integral in (G.1) is positive, so $f(s;\rho)$ is strictly decreasing. Its average over preceding times is therefore strictly larger than its current value, implying $\partial_s f(s;\rho) < -\gamma f(s;\rho)$ for $s > 0$. Hence

$$0 < f(s;\rho) < e^{-\gamma s}, \qquad s > 0. \tag{G.4}$$

By (E.2), $J(\theta^\star + a(\theta_0 - \theta^\star))$ is strictly decreasing in $a > 0$. Hence (G.4) gives $J(\theta(\tau;\rho)) > J(\theta^\star + e^{-\gamma\tau}(\theta_0 - \theta^\star))$. The coefficient in the main theorem is therefore

$$\begin{aligned} C(\tau,\gamma,\rho) &:= \frac{J(\theta(\tau;\rho)) - J(\theta^\star + e^{-\gamma\tau}(\theta_0 - \theta^\star))}{\rho} \\ &= \frac{\gamma}{\rho}\int_{f(\tau;\rho)}^{e^{-\gamma\tau}} a\,\mathrm{Var}_{y\sim\pi_{\theta^\star + a(\theta_0-\theta^\star)}(\cdot|x)}[\theta_0(x,y) - \theta^\star(x,y)]\,\mathrm{d}a > 0. \end{aligned}$$

This specifies its dependence on $\rho$, $\tau$, $\gamma$, and the fixed problem; it does not depend on $\eta$, $\beta$, or $B$. Using $\rho = \beta\rho + \eta$ in (G.3) gives $J(\theta_T^{\beta}) - J(\theta_T^{0}) = C(\tau,\gamma,\rho)\beta\rho + O(\eta)$.

**Expected gain and winning probability.** Lemma 5.1 gives, more explicitly,

$$\begin{aligned}\mathbb{E}[J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})] &= J(\theta_T^{\beta}) - J(\theta_T^{0}) + O(\eta/B) \\ &= C(\tau,\gamma,\rho)\beta\rho + O(\eta) + O(\eta/B).\end{aligned}$$

The $O(\eta)$ error is the discrete-path approximation, and the $O(\eta/B)$ term is the sampling correction. At each fixed positive $\rho$, the expected gain tends to $C(\tau,\gamma,\rho)\rho > 0$. Thus it is at least half this limit for sufficiently small $\eta$. Proposition 5.2 and Chebyshev's inequality imply

$$\mathbb{P}\Big(J(\widehat{\theta}_T^{\beta}) \le J(\widehat{\theta}_T^{0})\Big) \le \frac{\mathrm{Var}[J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})]}{\big(\mathbb{E}[J(\widehat{\theta}_T^{\beta}) - J(\widehat{\theta}_T^{0})]\big)^2} \le \frac{4C}{C(\tau,\gamma,\rho)^2\rho^2}\frac{\eta}{B} = O(\eta/B).$$

Here $C$ bounds the variance coefficient and depends only on the fixed problem and $\tau$; the winning-probability bound may depend on the fixed positive $\rho$ through its denominator.

**Effect of increasing the memory ratio.** Fix $0 < \rho_1 < \rho_2 \le 1/(4\gamma)$ and set $\beta_i = 1 - \eta/\rho_i$, $i = 1, 2$. We compare the expected objective values of two runs with the same $\eta, T, B$. We first show $f(s;\rho_2) < f(s;\rho_1)$ by comparing their logarithmic decay rates $h_\rho(s) = -\partial_s \log f(s;\rho)$. The inequality $\partial_s f(s;\rho) < -\gamma f(s;\rho)$, established in the derivation of (G.4), gives $h_\rho(s) > \gamma$ for $s > 0$. Differentiating (G.1) gives

$$h_\rho'(s) = h_\rho(s)^2 - \frac{h_\rho(s) - \gamma}{\rho(1 - e^{-s/\rho})}, \qquad h_\rho(0) = \gamma. \tag{G.5}$$

Define $\mu(s) = (e^{s/\rho_2} - 1) f(s;\rho_1) f(s;\rho_2)$. Differentiating and substituting (G.5), using $\partial_s f(s;\rho_i) = -h_{\rho_i}(s) f(s;\rho_i)$, gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\big[\mu(s)(h_{\rho_2}(s) - h_{\rho_1}(s))\big] &= \mu'(s)(h_{\rho_2}(s) - h_{\rho_1}(s)) + \mu(s)(h_{\rho_2}'(s) - h_{\rho_1}'(s)) \\ &= \mu(s)(h_{\rho_1}(s) - \gamma)\Big[\frac{1}{\rho_1(1 - e^{-s/\rho_1})} - \frac{1}{\rho_2(1 - e^{-s/\rho_2})}\Big] > 0.\end{aligned}$$

The last inequality uses $\mu(s) > 0$, $h_{\rho_1}(s) > \gamma$, and the strict increase of $\rho(1 - e^{-s/\rho})$ with $\rho$. Since $\mu(s)(h_{\rho_2}(s) - h_{\rho_1}(s)) \to 0$ as $s \to 0$ and $\mu(s) > 0$, we have $h_{\rho_2}(s) > h_{\rho_1}(s)$ for $s > 0$. Integrating gives $f(s;\rho_2) < f(s;\rho_1)$. By (E.2), $J(\theta(\tau;\rho_2)) > J(\theta(\tau;\rho_1))$. Subtracting the two expansions in Lemma 5.1 and applying Lemma G.1 gives

$$\begin{aligned}\mathbb{E}[J(\widehat{\theta}_T^{\beta_2}) - J(\widehat{\theta}_T^{\beta_1})] &= J(\theta_T^{\beta_2}) - J(\theta_T^{\beta_1}) + O(\eta/B) \\ &= \underbrace{J(\theta(\tau;\rho_2)) - J(\theta(\tau;\rho_1))}_{>0} + O(\eta) + O(\eta/B).\end{aligned}$$

For fixed $\rho_1, \rho_2$, the underbraced difference is a positive constant independent of $\eta$. Hence the larger ratio gives a strictly larger expected objective value for sufficiently small $\eta = \tau/T$. □

### G.1 Proof of Lemma G.1

*Proof.* **Existence and time regularity.** The normalized averaging operator in (G.1) has norm one on each preceding time interval. Iterating its composition with $\gamma\int_0^s$ gives norm at most $(\gamma s)^k/k!$ at iteration $k$. The resulting Volterra series converges uniquely and bounds both $f$ and its normalized average by $e^{\gamma\tau}$ on $[0,\tau]$. It follows that $f$ is Lipschitz with constant $\gamma e^{\gamma\tau}$. The average has the same Lipschitz bound: it can be written as

$$\int_0^1 f\big(\rho\log(1 + z(e^{s/\rho} - 1));\rho\big)\,\mathrm{d}z,$$

and the derivative in $s$ of the inner time lies in $[0,1]$.

**Comparison of discrete and continuous averages.** The continuous averaging weights describe an exponential age of mean $\rho$, conditioned to $[0,s]$. Its quantile at $z \in [0,1]$ is $-\rho\log(1 - z + ze^{-s/\rho})$. These quantiles increase with $\rho$, and the derivative of the conditional mean is

$$\frac{\partial}{\partial\rho}\left(\rho - \frac{s}{e^{s/\rho} - 1}\right) = 1 - \left(\frac{s/\rho}{2\sinh(s/(2\rho))}\right)^2 \in [0,1].$$

Using the same quantile therefore matches the ages before and after a change in the exponential mean, with expected absolute difference at most the change in that mean.

The discrete average uses ages $0, \eta, \dots, t\eta$ with probabilities proportional to $1, \beta, \dots, \beta^t$. These are obtained by rounding down an exponential age of mean $-\eta/\log\beta$, conditioned to $[0, (t+1)\eta]$. The elementary logarithm bounds give

$$\rho - \eta = \beta\rho \le -\frac{\eta}{\log\beta} \le \rho.$$

The continuous average at $t\eta$ uses an exponential age of mean $\rho$, conditioned to $[0, t\eta]$. The two ages can be matched with expected absolute difference $3\eta$, uniformly in $\rho > \eta$ and $t\eta \le \tau$. Changing the mean costs at most $\eta$ by the preceding calculation. The derivative of a truncated exponential quantile with respect to the truncation time lies in $[0, 1]$, so changing that time costs at most $\eta$; rounding costs another $\eta$.

The Lipschitz bound therefore gives

$$\left|\sum_{j=0}^{t} \omega_\beta(t,j) f_j(\beta) - \frac{\int_0^{t\eta} e^{-(t\eta-u)/\rho} f(u;\rho)\,\mathrm{d}u}{\rho(1-e^{-t\eta/\rho})}\right| \le \max_{j\le t} |f_j(\beta) - f(j\eta;\rho)| + 3\gamma e^{\gamma\tau}\eta,$$

with both averages equal to one at $t = 0$. The continuous time-cell quadrature error is at most $\gamma^2 e^{\gamma\tau}\eta^2/2$ by the Lipschitz bound for the average. Subtracting the two updates and iterating the resulting inequality gives

$$\max_{j\le t} |f_j(\beta) - f(j\eta;\rho)| \le \frac{7}{2}\gamma^2 e^{\gamma\tau}\eta^2 \sum_{j=0}^{t-1} (1+\gamma\eta)^j \le \frac{7}{2}\gamma e^{\gamma\tau}(e^{\gamma\tau}-1)\eta.$$

This proves (G.2), with no dependence on $\rho$ in the constant. □

## H Proof of Theorem 5.5

We first match the exact endpoints at finite step size. The mean comparison then follows from the sampling expansion; the variance comparison requires the continuous response and its limit at long optimization times. In this section, $f_T(\eta, \beta)$ displays the learning-rate argument because the two runs use different rates.

*Proof.* **Finite-step endpoint matching.** The assumptions give $0 < \eta_1 \le \rho_1 < \rho_2 \le 1/(4\gamma)$ and

$$0 \le \beta_1 < 1 - \eta_1/\rho_2 \le 1 - 4\gamma\eta_1 \le \left(\frac{1-\gamma\eta_1}{1+\gamma\eta_1}\right)^2.$$

By the residual comparison in the proof of Theorem 4.2, $f_T(\eta_1, 1-\eta_1/\rho_2) < f_T(\eta_1, \beta_1) < 1$. For fixed $T$, the recurrence extends continuously to $(\eta, \beta) = (0, 1)$, where its value is 1: write $\omega_\beta(t,j) = \beta^{t-j}/\sum_{k=0}^{t}\beta^k$ to remove the apparent singularity. The intermediate value theorem on $[0, \eta_1]$ therefore gives a deterministic $\eta_2 \in (0, \eta_1)$ such that

$$f_T(\eta_2, 1-\eta_2/\rho_2) = f_T(\eta_1, \beta_1).$$

Set $\beta_2 = 1 - \eta_2/\rho_2$. Then $\beta_2 > \beta_1$, and Lemma 4.1 gives $\theta_T^{(2)} = \theta_T^{(1)}$.

**Limiting optimization times and expected objectives.** Fix $\tau, \rho_1, \rho_2$. By the strict time and memory monotonicity proved in Appendix G, there is a unique $\tau_2 \in (0, \tau)$ such that

$$f(\tau_2;\rho_2) = f(\tau;\rho_1), \qquad \theta(\tau_2;\rho_2) = \theta(\tau;\rho_1). \tag{H.1}$$

Combining Lemma G.1 with the exact endpoint matching gives $f(T\eta_2;\rho_2) = f(\tau;\rho_1) + O(\eta_1) = f(\tau_2;\rho_2) + O(\eta_1)$. The strict monotonicity of $f(\cdot;\rho_2)$ then implies

$$T\eta_2 \longrightarrow \tau_2, \qquad \frac{\eta_2}{\eta_1} = \frac{T\eta_2}{\tau} \longrightarrow \frac{\tau_2}{\tau} \quad \text{as } \eta_1 \to 0.$$

Applying the individual expansion (C.15) to both runs and using $\theta_T^{(2)} = \theta_T^{(1)}$ gives

$$\mathbb{E}[J(\widehat{\theta}_T^{(2)}) - J(\widehat{\theta}_T^{(1)})] = O(\eta_1/B) + O(\eta_2/B) = O(\eta_1/B).$$

This proves the expected objective comparison.

**Continuous covariance coefficients.** To compare the variance coefficients, consider the propagation of a single noise input in continuous optimization time. Let $s_1 > 0$ be the noise input time and $s_2 \geq s_1$ the observation time. For one parameter coordinate, let $\Delta\theta(s_2)$ and $\Delta m(s_2)$ denote its effects on the parameter and the EMA momentum before bias correction. With $\beta = 1 - \eta/\rho$, the continuous limit of the EMA updates after this input is

$$\frac{\mathrm{d}}{\mathrm{d}s_2}\Delta\theta(s_2) = \frac{\Delta m(s_2)}{1 - e^{-s_2/\rho}}, \qquad \frac{\mathrm{d}}{\mathrm{d}s_2}\Delta m(s_2) = -\frac{\gamma}{\rho}\Delta\theta(s_2) - \frac{1}{\rho}\Delta m(s_2).$$

Let $\Phi_\rho(s_2, s_1)$ be the $2 \times 2$ transition matrix for $(\Delta\theta, \rho\Delta m)^\top$ from $s_1$ to $s_2$. The two equations above give

$$\partial_{s_2}\Phi_\rho(s_2, s_1) = \begin{pmatrix} 0 & [\rho(1 - e^{-s_2/\rho})]^{-1} \\ -\gamma & -1/\rho \end{pmatrix} \Phi_\rho(s_2, s_1), \qquad \Phi_\rho(s_1, s_1) = \mathbf{I}_2. \tag{H.2}$$

Let $\Psi(s_2, s_1; \rho)$ denote the continuous-time limit of $\Psi_\beta(T, j)/\eta$, with $\Psi_\beta$ defined in (C.2), as $\eta \to 0$, $T\eta \to s_2$, and $j\eta \to s_1$. Since noise enters the second state component, we have

$$\Psi(s_2, s_1; \rho) = \mathbf{e}_1^\top \Phi_\rho(s_2, s_1)\mathbf{e}_2,$$

where $\mathbf{e}_1, \mathbf{e}_2$ are the coordinate vectors of this two-dimensional state. We define the continuous-time counterpart of the covariance coefficient $\Sigma_T^\beta$ in (C.3) by

$$\Sigma(s_2; \rho) = \int_0^{s_2} \Psi(s_2, s_1; \rho)^2 \left[\mathrm{diag}\left(\frac{\mathbf{r}}{\pi_{\theta(s_1;\rho)}(\cdot \mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right] \mathrm{d}s_1.$$

Its limit as the optimization time $s_2 \to \infty$ is denoted by

$$\Sigma_\infty = \frac{1}{2\gamma}\left[\mathrm{diag}\left(\frac{\mathbf{r}}{\pi_{\theta^\star}(\cdot \mid x)}\right) - \mathbf{r}\mathbf{r}^\top\right].$$

The following lemma establishes both limits and controls quadratic forms even when their directions vary with the endpoint. Its proof is in the next subsection.

**Lemma H.1** (Limits of the covariance coefficient)**.** *Fix $0 < \rho \leq 1/(4\gamma)$. If $\eta \to 0$, $T\eta \to s_2 > 0$, and $\beta = 1 - \eta/\rho$, then $\Sigma_T^\beta \to \Sigma(s_2; \rho)$. Moreover,*

$$\Sigma(s_2; \rho) \longrightarrow \Sigma_\infty, \qquad \sup_{v:\, v^\top\Sigma_\infty v > 0} \left|\frac{v^\top\Sigma(s_2; \rho)v}{v^\top\Sigma_\infty v} - 1\right| \longrightarrow 0 \quad (s_2 \to \infty). \tag{H.3}$$

**Variance ratio at matched endpoints.** Since the exact endpoints satisfy $\theta_T^{(1)} = \theta_T^{(2)}$, Lemma G.1 and the continuity of $\nabla J$ give $\nabla J(\theta_T^{(1)}) = \nabla J(\theta_T^{(2)}) \to \nabla J(\theta(\tau; \rho_1))$ as $\eta_1 \to 0$. To obtain the limiting variance ratio from the leading variance terms, we first verify that the limiting quadratic forms are nonzero for all sufficiently large fixed $\tau$. As shown in the proof of Lemma H.1, $\Sigma(\tau; \rho_1)$ and $\Sigma(\tau_2; \rho_2)$ are positive definite on the reward-1 subspace and zero on its orthogonal complement. It therefore suffices to find a nonzero reward-1 coordinate of the common limiting gradient. Along the exact path $\theta(\tau; \rho_1) = \theta^\star + f(\tau; \rho_1)(\theta_0 - \theta^\star)$, fix any reward-1 response $y$. Since $f(\tau; \rho_1) > 0$ and the policy has full support, (A.2) shows that its gradient coordinate vanishes exactly when $\theta_0(x, y) - \theta^\star(x, y)$ equals the average of $\theta_0 - \theta^\star$ under $\pi_{\theta(\tau;\rho_1)}$. The derivative of this average with respect to $\tau$ is

$$\partial_\tau f(\tau; \rho_1)\ \mathrm{Var}_{y' \sim \pi_{\theta(\tau;\rho_1)}(\cdot|x)}[\theta_0(x, y') - \theta^\star(x, y')] < 0,$$

because $\partial_\tau f(\tau; \rho_1) < 0$ and the variance is positive by nonoptimal initialization and full support. The average is therefore strictly decreasing, so the chosen gradient coordinate can vanish at most once for $\tau > 0$ and is nonzero for all sufficiently large $\tau$. Both quadratic forms are consequently strictly positive for these $\tau$.

For such a fixed $\tau$, Lemma H.1 implies that the two discrete quadratic coefficients converge to these strictly positive limits. Thus (D.1) gives, as $\eta_1 \to 0$,

$$\mathrm{Var}[J(\widehat{\theta}_T^{(i)})] = \frac{\eta_i}{B} \nabla J(\theta_T^{(i)})^\top \Sigma_T^{\beta_i} \nabla J(\theta_T^{(i)}) \big(1 + o(1)\big), \qquad i = 1, 2,$$

where each covariance coefficient uses its own learning rate. Taking the ratio of these nonzero leading terms and using $\eta_2/\eta_1 \to \tau_2/\tau$ yields

$$\lim_{\eta_1 \to 0} \frac{\mathrm{Var}[J(\widehat{\theta}_T^{(2)})]}{\mathrm{Var}[J(\widehat{\theta}_T^{(1)})]} = \frac{\tau_2}{\tau} \frac{\nabla J(\theta(\tau;\rho_1))^\top \Sigma(\tau_2;\rho_2) \nabla J(\theta(\tau;\rho_1))}{\nabla J(\theta(\tau;\rho_1))^\top \Sigma(\tau;\rho_1) \nabla J(\theta(\tau;\rho_1))}. \tag{H.4}$$

Both $\tau$ and $\tau_2$ tend to infinity as the common residual tends to zero. The relative bound in Lemma H.1 makes the quotient of quadratic forms in (H.4) tend to one, even though the common gradient tends to zero. It remains to determine $\tau_2/\tau$.

**Long-time residual decay.** Use the logarithmic rate $h_\rho$ from (G.5). For each $s_2 > 0$, define the quadratic polynomial

$$Q_{s_2}(x) = x^2 - \frac{x - \gamma}{\rho(1 - e^{-s_2/\rho})}, \qquad h_\rho'(s_2) = Q_{s_2}(h_\rho(s_2)).$$

The smaller root $r_\rho(s_2)$ of $Q_{s_2}(x) = 0$ is strictly increasing in $s_2$ and satisfies

$$r_\rho(s_2) = \frac{2\gamma}{1 + \sqrt{1 - 4\gamma\rho(1 - e^{-s_2/\rho})}}, \qquad \lim_{s_2 \to \infty} r_\rho(s_2) = \frac{2\gamma}{1 + \sqrt{1 - 4\gamma\rho}}.$$

We denote this limit by $\kappa(\rho)$. Near zero, $h_\rho(s_2) = \gamma + \gamma^2 s_2/2 + O(s_2^2)$ and $r_\rho(s_2) = \gamma + \gamma^2 s_2 + O(s_2^2)$, so $h_\rho < r_\rho$ initially. If $h_\rho(s_*) = r_\rho(s_*)$ at a first contact time $s_* > 0$, then $(h_\rho - r_\rho)'(s_*) \geq 0$. But $(h_\rho - r_\rho)'(s_*) = Q_{s_*}(r_\rho(s_*)) - r_\rho'(s_*) = -r_\rho'(s_*) < 0$, which contradicts the inequality above. Thus $h_\rho(s_2) < r_\rho(s_2)$ for every $s_2 > 0$. The quadratic is positive below its smaller root, so $h_\rho$ is increasing and bounded above by $\kappa(\rho)$. We now show that the limit of $h_\rho$ is $\kappa(\rho)$. Let $L = \lim_{s_2 \to \infty} h_\rho(s_2) \leq \kappa(\rho)$. By (G.5), $h_\rho'(s_2) \to L^2 - (L - \gamma)/\rho$. This limit must be zero, since a nonzero limiting derivative would prevent convergence to $L$. Thus $L^2 - (L - \gamma)/\rho = 0$, and $L \leq \kappa(\rho)$ selects the smaller root, giving $L = \kappa(\rho)$. Therefore

$$\lim_{s_2 \to \infty} \left( -\frac{\log f(s_2;\rho)}{s_2} \right) = \lim_{s_2 \to \infty} \frac{1}{s_2} \int_0^{s_2} h_\rho(s_1)\, \mathrm{d}s_1 = \kappa(\rho).$$

This argument includes $\rho = 1/(4\gamma)$, where the two limiting roots coincide. Taking logarithms of the matched residuals in (H.1) now gives

$$\lim_{\tau \to \infty} \frac{\tau_2}{\tau} = \lim_{\tau \to \infty} \frac{-\log f(\tau;\rho_1)/\tau}{-\log f(\tau_2;\rho_2)/\tau_2} = \frac{1 + \sqrt{1 - 4\gamma\rho_2}}{1 + \sqrt{1 - 4\gamma\rho_1}} < 1.$$

Substituting into (H.4) proves the double-limit variance ratio in Theorem 5.5. □

### H.1 Proof of Lemma H.1

*Proof.* **Discrete-to-continuous response.** Without drift feedback, the continuous parameter response is $\int_{s_1}^{s_2} e^{-(v-s_1)/\rho}/[\rho(1 - e^{-v/\rho})]\, \mathrm{d}v$. The integrand is at most $\rho^{-1} e^{-(v-s_1)/\rho} + v^{-1}$. The Gronwall argument used for Lemma C.1 therefore gives

$$|\Psi(s_2, s_1; \rho)| \leq e^{\gamma s_2}[1 + \log(s_2/s_1)], \qquad 0 < s_1 \leq s_2. \tag{H.5}$$

For the discrete response, divide the parameter and scaled-memory coordinates by $\eta$. An innovation at update $j$ gives the initial vector $(1/\sum_{k=0}^{j} \beta^k, 1)^\top$ at update $j + 1$; subsequent updates multiply it by

$$\begin{pmatrix} 1 - \gamma\eta/\sum_{k=0}^{n} \beta^k & \beta/\sum_{k=0}^{n} \beta^k \\ -\gamma\eta & \beta \end{pmatrix}.$$

For $n\eta$ bounded away from zero, this is the identity plus $\eta$ times the coefficient matrix in (H.2), with error $O(\eta^2)$. Indeed, $\eta \sum_{k=0}^{n} \beta^k = \rho[1 - (1 - \eta/\rho)^{n+1}] = \rho(1 - e^{-n\eta/\rho}) + O(\eta)$. The initial

response tends to $\mathbf{e}_2$. Comparing the matrix products with the continuous transitions thus gives, for each $a > 0$,

$$\sup_{a \le j\eta < T\eta} |\Psi_\beta(T, j)/\eta - \Psi(T\eta, j\eta; \rho)| = O_a(\eta) \quad \text{when } T\eta \to s_2 > 0.$$

Lemma G.1 and a Riemann sum identify the covariance contribution from $j\eta \ge a$. For earlier samples, Lemma C.1 gives

$$\limsup_{\eta \to 0} \eta \sum_{j\eta < a} |\Psi_\beta(T, j)/\eta|^2 \le C \int_0^a [1 + \log(s_2/s_1)]^2 \, \mathrm{d}s_1.$$

The isolated $j = 0$ term is $O(\eta \log^2(1/\eta))$. The continuous response has the same integrable envelope by (H.5). The reward covariance matrices are bounded along these paths, so letting $a \to 0$ proves $\Sigma_T^\beta \to \Sigma(s_2; \rho)$.

**Covariance at long optimization times.** The coefficient matrix in (H.2) converges exponentially to

$$A_\infty = \begin{pmatrix} 0 & 1/\rho \\ -\gamma & -1/\rho \end{pmatrix}.$$

Its eigenvalues solve $\lambda^2 + \lambda/\rho + \gamma/\rho = 0$ and have negative real parts, including at the repeated root. Thus $\|e^{A_\infty v}\|_{\mathrm{op}} \le Ce^{-cv}$ for some $c > 0$. The difference from the time-dependent matrix has only one nonzero entry, $e^{-s_2/\rho}/[\rho(1 - e^{-s_2/\rho})]$. Variation of constants therefore gives, for $s_2 \ge s_1 \ge 1$,

$$e^{c(s_2 - s_1)} \|\Phi_\rho(s_2, s_1)\|_{\mathrm{op}} \le C + C \int_{s_1}^{s_2} \frac{e^{-u/\rho}}{\rho(1 - e^{-u/\rho})} e^{c(u - s_1)} \|\Phi_\rho(u, s_1)\|_{\mathrm{op}} \, \mathrm{d}u.$$

The coefficient in this integral is integrable on $[1, \infty)$. Gronwall's inequality, absorbing its integral into $C$, gives

$$\|\Phi_\rho(s_2, s_1)\|_{\mathrm{op}} \le Ce^{-c(s_2 - s_1)}, \qquad s_2 \ge s_1 \ge 1. \tag{H.6}$$

Innovations before time 1 have finite total response energy. The first coordinate of $\Phi_\rho(1, s_1)\mathbf{e}_2$ is bounded by (H.5); its second coordinate equals

$$e^{-(1 - s_1)/\rho} - \gamma \int_{s_1}^1 e^{-(1 - v)/\rho} \Psi(v, s_1; \rho) \, \mathrm{d}v.$$

Both are bounded by $C[1 + \log(1/s_1)]$, whose square is integrable. Composing with $\Phi_\rho(s_2, 1)$ shows that their covariance contribution vanishes exponentially.

For later innovations, write their age as $v = s_2 - s_1$. For each fixed $v$, $\Phi_\rho(s_2, s_2 - v) \to e^{A_\infty v}$ and $\theta(s_2 - v; \rho) \to \theta^\star$. Equation (H.6) provides an integrable dominating bound, so the limiting covariance equals the reward covariance at $\theta^\star$ times $\int_0^\infty (\mathbf{e}_1^\top e^{A_\infty v} \mathbf{e}_2)^2 \, \mathrm{d}v$. To evaluate this energy, write $S = \int_0^\infty e^{A_\infty v} \mathbf{e}_2 \mathbf{e}_2^\top e^{A_\infty^\top v} \, \mathrm{d}v$ for this two-dimensional integral. Integrating the derivative of its integrand, whose boundary term at infinity vanishes, gives

$$A_\infty S + S A_\infty^\top = -\mathbf{e}_2 \mathbf{e}_2^\top.$$

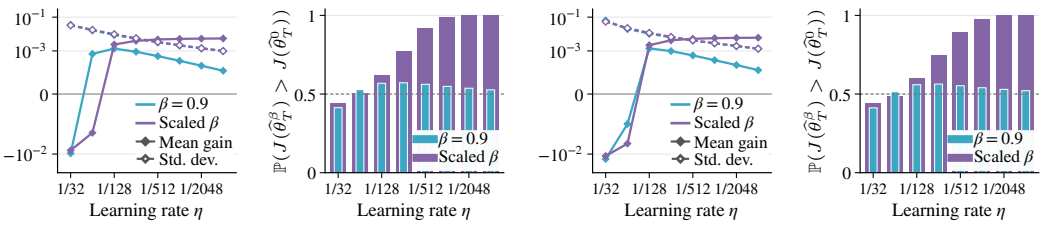
The $(1, 1)$, $(2, 2)$, and $(1, 2)$ entries, respectively, yield $2S_{12}/\rho = 0$, $-2\gamma S_{12} - 2S_{22}/\rho = -1$, and $S_{22}/\rho - \gamma S_{11} - S_{12}/\rho = 0$. Hence

$$S = \begin{pmatrix} 1/(2\gamma) & 0 \\ 0 & \rho/2 \end{pmatrix}.$$

Its upper-left entry gives the required scalar energy, proving $\Sigma(s_2; \rho) \to \Sigma_\infty$.

**Relative control in changing directions.** All reward covariance matrices have the same support, namely the span of $\{\mathbf{e}_y : r(x, y) = 1\}$. Weighted Cauchy–Schwarz gives, for every vector $v$,

$$\begin{aligned} v^\top \left[ \mathrm{diag}\left( \frac{\mathbf{r}}{\pi_\theta(\cdot \mid x)} \right) - \mathbf{r}\mathbf{r}^\top \right] v &= \sum_{y:\, r(x,y)=1} \frac{v_y^2}{\pi_\theta(y \mid x)} - \left( \sum_{y:\, r(x,y)=1} v_y \right)^2 \\ &\ge \left( 1 - \sum_{y:\, r(x,y)=1} \pi_\theta(y \mid x) \right) \sum_{y:\, r(x,y)=1} \frac{v_y^2}{\pi_\theta(y \mid x)}. \end{aligned}$$




(a) Gain and Noise. (b) Win Probability. (c) Gain and Noise. (d) Win Probability.

Figure 4: Numerical bandit results with 5 reward-one and 5 reward-zero arms under uniform initialization (4(a), 4(b)) and random policy initialization (4(c), 4(d)).

Since both reward values occur and policies have full support, these matrices are positive definite on that span and vanish on its orthogonal complement. The same holds for $\Sigma_\infty$ and $\Sigma(s_2;\rho)$, whose response weight is nonzero near its endpoint. The smallest eigenvalue of $\Sigma_\infty$ restricted to this fixed span is positive. Dividing the operator-norm error $\|\Sigma(s_2;\rho) - \Sigma_\infty\|_{\mathrm{op}}$ by this eigenvalue bounds the relative error of every nonzero quadratic form uniformly. This proves (H.3). □

# I Experimental Details

## I.1 Bandit Results

**Settings.** We consider a 10-armed bandit problem and maximize

$$J(\theta) = \sum_{y=1}^{10} \pi_\theta(y \mid x) r(x,y) - \mathrm{KL}(\pi_\theta(\cdot \mid x) \| \pi_{\theta_0}(\cdot \mid x)) \,.$$

We set $\gamma = 1$, $B = 1$, and $\tau = 1$, use $\eta = 1/T$ for $T = 32, 64, \ldots, 4096$, and compare fixed $\beta = 0.9$ and scaled $\beta = 1 - 4\eta$ with $\beta = 0$. Each configuration uses 50,000 independent training replicates, with $J$ evaluated exactly after training. We vary the reward vector and policy initialization across four settings:

- **3:7, uniform.** Reward vector $(1,1,1,0,0,0,0,0,0,0)$; initial policy $\pi_{\theta_0}(\cdot \mid x) = (0.1, \ldots, 0.1)$. Results appear in Figures 1(a) and 1(b).
- **3:7, random.** Reward vector $(1,1,1,0,0,0,0,0,0,0)$; initial policy $\pi_{\theta_0}(\cdot \mid x) \approx (0.0388, 0.0713, 0.1502, 0.3932, 0.0100, 0.0459, 0.0635, 0.0928, 0.0891, 0.0453)$. Results appear in Figures 1(c) and 1(d).
- **5:5, uniform.** Reward vector $(1,1,1,1,1,0,0,0,0,0)$; initial policy $\pi_{\theta_0}(\cdot \mid x) = (0.1, \ldots, 0.1)$. Results appear in Figures 4(a) and 4(b).
- **5:5, random.** Reward vector $(1,1,1,1,1,0,0,0,0,0)$; initial policy $\pi_{\theta_0}(\cdot \mid x) \approx (0.1320, 0.1596, 0.0361, 0.0342, 0.0221, 0.0846, 0.2399, 0.0667, 0.1895, 0.0353)$. Results appear in Figures 4(c) and 4(d).

Random initializations are drawn separately from $\mathrm{Dirichlet}(1, \ldots, 1)$ and are held fixed across learning rates, EMA coefficients, and training replicates.

**Results.** All four settings exhibit the same trends as $\eta$ decreases: the standard deviation scales as $\sqrt{\eta}$ under both schedules, while the mean gain decays as $\eta$ with fixed $\beta$ and approaches a positive constant with scaled $\beta$. Consequently, the winning probability approaches $1/2$ in the fixed setting and 1 in the scaled setting. These results support the robustness of the theoretical predictions across reward vectors and policy initializations. Table 1 reports the SNR defined in Section 6.1 for each setting and learning rate.

| Setting | $\beta$ | Learning rate $\eta$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $1/32$ | $1/64$ | $1/128$ | $1/256$ | $1/512$ | $1/1024$ | $1/2048$ | $1/4096$ |
| 3:7, uniform | 0.9 | 0.1271 | 0.1236 | 0.1732 | 0.1530 | 0.1183 | 0.0876 | 0.0632 | 0.0453 |
| | $1-4\eta$ | 0.0719 | 0.1118 | 0.3720 | 0.7229 | 1.1844 | 1.7900 | 2.6258 | 3.7647 |
| 3:7, random | 0.9 | 0.1809 | 0.0904 | 0.1536 | 0.1393 | 0.1071 | 0.0796 | 0.0573 | 0.0406 |
| | $1-4\eta$ | 0.1220 | 0.0585 | 0.3178 | 0.6486 | 1.0712 | 1.6238 | 2.3713 | 3.3982 |
| 5:5, uniform | 0.9 | 0.2625 | 0.0402 | 0.1546 | 0.1726 | 0.1524 | 0.1226 | 0.0931 | 0.0673 |
| | $1-4\eta$ | 0.1849 | 0.0310 | 0.2455 | 0.7047 | 1.4036 | 2.3824 | 3.7419 | 5.5492 |
| 5:5, random | 0.9 | 0.3162 | 0.0083 | 0.1343 | 0.1539 | 0.1310 | 0.1002 | 0.0746 | 0.0534 |
| | $1-4\eta$ | 0.2428 | 0.1064 | 0.1862 | 0.6293 | 1.2251 | 1.9919 | 3.0391 | 4.4557 |

Table 1: SNR of the objective difference between independent EMA and no-EMA runs. Each ratio gives the numbers of reward-one and reward-zero arms, respectively.

### I.2 Large language model experiments

**Settings.** We use Qwen3-4B-Base and Qwen2.5-32B models using the verl training framework. The parameters are set as follows.

| **Parameter** | **Value** | **Parameter** | **Value** |
|---|---|---|---|
| Training dataset | DAPO-17k (English) | Generations per prompt | 8 |
| Prompts per batch | 16 | Updates per RL step | 1 |
| Max prompt length | 2,048 | Max response length | 4,096 |
| Rollout temperature | 1.0 | Rollout top-$p$ | 1.0 |
| Clip ratio | 0.2 | Gradient clipping norm | 1.0 |
| KL coefficient | 0.01 | Training generation seeds | 1001, 2001, 3001 |

Table 2: Shared training settings for the LLM experiments.

**Prompts and rewards.** We use the original prompts provided in the DAPO-17k English dataset. As in DAPO's implementation, we restrict answer extraction to the final 300 characters of each response. A response receives a reward of 1 if the extracted answer matches the ground truth, and 0 otherwise. The same rule applies to responses truncated at the length limit; truncation alone does not force a zero reward.

**Validation.** To assess the optimization performance, we evaluate on held-out data drawn from the same distribution as the training data. We construct this validation set by randomly holding out 1,000 problems from the DAPO-17k English dataset and excluding them from training. Evaluation is performed every 20 optimization steps, with an additional evaluation at step 150 for runs ending at that step. For each problem, we sample eight responses at a temperature of 0.7 and top-$p$ of 1.0, with a maximum response length of 4,096 tokens. Each response receives a binary correctness score from the answer verifier. We report the average score over all 8,000 responses, equivalently the mean per-problem accuracy. All runs use the same held-out problems.