LOGARITHMIC REGRET IN PREFERENCE LEARNING VIA OPTIMISTIC PAC-BAYESIAN PARTICLE ENSEMBLES

Anonymous authors
Paper under double-blind review

ABSTRACT

The remarkable sample efficiency of preference-based reinforcement learning, which underpins the alignment of large language models with human feedback (RLHF), presents a significant theoretical puzzle. Existing analyses often rely on idealized assumptions, such as infinite-particle ensembles or exact, full-batch gradients, that are disconnected from the practical realities of deployed algorithms. This paper closes this theory-practice gap. We introduce a unified optimistic PAC-Bayesian framework that distills the statistical essence of complex, multi-stage RLHF pipelines into a single, provably efficient online learning algorithm. Our central result is a high-probability regret bound of $\mathbf{O}(d_{\text{eluder}} \log T)$ for a rich, nonlinear class of reward models, demonstrating that logarithmic regret is achievable even when using *finite* ensembles and noisy *stochastic gradient* updates. This unified theory provides an explanation for the sample efficiency of pairwise preference optimization, extends naturally to full Markov Decision Processes, and establishes a theoretical foundation for the empirical success of methods like RLHF.

1 INTRODUCTION

The alignment of large language models (LLMs) through preference-based learning has become a cornerstone of modern artificial intelligence, enabling the development of systems that are helpful, harmless, and attuned to human intent (Ouyang et al., 2022; Bai et al., 2022; Dong et al., 2024). A striking empirical observation in this domain is the profound *sample efficiency* of these alignment pipelines. Practitioners routinely steer billion-parameter models toward complex desired behaviors using on the order of only tens of thousands of pairwise human preferences (Rafailov et al., 2023; Christiano et al., 2017). This efficiency stands in stark contrast to the sheer dimensionality of the models and suggests that the correct theoretical target for regret should exhibit a near-logarithmic dependence on the number of interaction rounds, T. While classical online learning analyses for expressive function classes typically yield regret bounds of $\mathbf{O}(\sqrt{T})$ (Russo & Van Roy, 2013; 2014), the empirical reality of RLHF motivates a much sharper theoretical goal. This leads to a pivotal open question: *Can we provide a rigorous theoretical explanation for the sample efficiency of practical preference-based alignment pipelines that yields sharp, near-logarithmic regret guarantees?*

The standard practical pipeline for Reinforcement Learning from Human Feedback (RLHF) is a complex, multi-stage process (Ouyang et al., 2022; Bai et al., 2022). It typically begins with Supervised Fine-Tuning (SFT) on a high-quality dataset, proceeds to the training of a separate reward model on collected human preference data, and culminates in policy optimization via an algorithm like PPO against that static reward model. This multi-stage pipeline, while empirically successful, presents a formidable challenge for unified theoretical analysis, as theoretical work often focuses on specific stages in isolation.

In this work, we move beyond analyzing the pipeline's components separately and instead propose a more fundamental theoretical model, the **Optimistic Langevin Ensemble (OLE)**, that captures the statistical core of preference-based learning in a single, cohesive online process. By analyzing this unified algorithm, we explain the sample efficiency of existing complex pipelines and provide a principled blueprint for a more theoretically grounded approach to alignment.

Bridging the empirical-theoretical divide requires that our unified model remains faithful to the realities of practical implementations. We identify four critical gaps[1] that must be addressed:

- **Gap 1: Mean-Field vs. Finite Ensembles.** Theoretical analyses often study a mean-field (infinite-particle) posterior flow for analytical tractability (Jordan et al., 1998; Sznitman, 2006), whereas practical implementations maintain a (often small) *finite* ensemble of reward models.
- **Gap 2: Exact vs. Stochastic Gradients.** Continuous-time or full-batch gradient derivations obscure the fact that all large-scale implementations rely on noisy mini-batch updates.
- **Gap 3: Continuous-Time vs. Discrete-Time Dynamics.** Mathematical tools like Wasserstein gradient flows offer an elegant continuous-time perspective (Ambrosio et al., 2008), but deployed algorithms operate in discrete time with a finite step size η.
- **Gap 4: Intractable vs. Tractable Uncertainty.** The principle of optimism requires an upper confidence bound on the true reward, but the exact Bayesian posterior uncertainty is intractable for deep neural networks. Practical algorithms rely on computationally feasible proxies, such as ensemble variance.

In this work, we develop an *optimistic PAC-Bayesian particle* framework for preference-based reinforcement learning that resolves these four gaps within our unified OLE model. Our framework is designed to be faithful to the algorithms used in practice while providing sharp, meaningful performance guarantees. We prove that such procedures attain a cumulative regret that scales as $\mathsf{O}(d_{\text{eluder}} \log T)$, where d_{eluder} is the eluder dimension of the function class (Russo & Van Roy, 2013; Li et al., 2022). Our analysis achieves this by coupling a PAC-Bayesian control of generalization (McAllester, 1999; Catoni, 2007) with concentration inequalities for stochastic dynamics (Freedman, 1975) and Wasserstein stability bounds for particle approximations (Fournier & Guillin, 2015), thereby addressing the four gaps within a single, cohesive theory.

Positioning and Scope. Our work is complementary to the important and emerging body of theory on *KL-regularized* bandits and RL, which has also achieved logarithmic regret guarantees but in the distinct setting of *numeric rewards* and often under additional structural assumptions like data coverage (Zhao et al., 2024; 2025b). We, in contrast, focus on the more foundational problem of learning from *pairwise preference feedback*, which is the canonical setup for RLHF and DPO where a reward model is itself learned from human comparisons (Christiano et al., 2017; Bradley & Terry, 1952; Luce et al., 1959). Our analysis is algorithm-native, deriving guarantees directly from a PAC-Bayesian treatment of particle ensembles, rather than from the specific optimization landscape of a KL-regularized objective. Conceptually, our approach is related to optimism-in-the-face-of-uncertainty and to feel-good Thompson sampling (Zhang, 2022), but our setting, estimators, and guarantees are novel. A comprehensive survey and detailed comparisons appear in Appendix B.

Table 1: Our work achieves logarithmic regret for pairwise preference feedback with general function approximation in a framework that models practical algorithmic constraints.

Setting	Feedback Model	Key Assumptions	Regret (Leading Term)
This work (OLE)	**Pairwise Preference**	**Realizable + Eluder Dim.**	$\mathbf{O}(\boldsymbol{d_{\text{eluder}} \log T})$
KL-Reg. Bandits (Zhao et al., 2025a)	Numeric Reward	Realizable + Eluder Dim.	$\mathsf{O}(d \log T)$
Preference RL (Wang et al., 2023)	Pairwise Preference	Realizable	$\mathsf{O}(\sqrt{T})$
Dueling Bandits (Yue et al., 2012)	Pairwise Preference	Tabular/Linear	$\mathsf{O}(\log T)$ or $\mathsf{O}(\sqrt{T})$
Optimistic Bandits (Russo & Van Roy, 2014)	Numeric Reward	Realizable + Eluder Dim.	$\mathsf{O}(d_{\text{eluder}} \log T)$

We summarize our main results, which provide a comprehensive theoretical account of preference-based learning.

- **Unified PAC-Bayesian Particle Analysis with Logarithmic Regret.** For preference-based contextual bandits, we analyze a practical algorithm using finite ensembles and mini-batch SGD. We prove that, with high probability, the cumulative regret is bounded by $\mathrm{Regret}(T) = \mathsf{O}(d_{\text{eluder}} \log T) \quad +$ lower-order terms for discretization, finite ensembles, and mini-batching, where the leading term captures the statistical cost of exploration, and the lower-order terms explicitly quantify the practical algorithmic costs.

[1] More discussion on the four gaps in Appendix Section A.3.

- **Optimistic Langevin Ensembles.** We introduce and analyze an optimistic Langevin-style ensemble update that provides exploration bonuses online and connects to standard preference optimization methods in the offline limit. Our analysis combines PAC-Bayesian inequalities with martingale concentration to provide non-asymptotic stability and concentration bounds.
- **Extension to Markov Decision Processes.** We extend our framework to preference-based RL with dynamics (e.g., discounted MDPs), obtaining analogous near-logarithmic regret guarantees. This complements results for numeric-reward MDPs (Zhao et al., 2025a) while operating in the more fundamental pairwise feedback regime.
- **Practical Implications.** Our bounds provide a direct theoretical explanation for the sample efficiency of methods like RLHF and DPO (Rafailov et al., 2023) and offer principled guidance for setting hyperparameters. We also show how parameter-efficient fine-tuning methods like LoRA (Hu et al., 2022) naturally lead to a small eluder dimension, connecting our theory to the practice of large-scale model alignment.

2 Problem Setup and Structural Assumptions

This section formally establishes the mathematical foundation[2] for our analysis. We begin by defining the preference-based contextual bandit model and the notion of cumulative preference regret. We then introduce the key structural assumptions[3] on the underlying reward function class that enable efficient, low-regret learning.

2.1 The Preference-Based Contextual Bandit Model

We consider an online learning problem that unfolds over T rounds. At each round $t \in \{1, \dots, T\}$, the environment presents a context $x_t \in \mathcal{X}$. The learning agent then selects a pair of actions to be compared, typically to maximize information gain about the optimal action. The agent receives feedback in the form of a pairwise preference. This process models the core interaction loop in RLHF, where a context might be a user prompt and the actions are different model-generated responses (Ouyang et al., 2022; Christiano et al., 2017).

Underlying this preference feedback is a latent, unknown reward function $r^* : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. This function represents the true, unobserved quality or utility of an action y in a context x. The observed preferences are stochastic manifestations of this latent function. We model this relationship using the standard and widely adopted Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce et al., 1959). Given a pair of actions (y_w, y_ℓ), the probability that y_w is preferred over y_ℓ (denoted $y_w \succ y_\ell$) in context x is given by a logistic link function:

$$p(y_w \succ y_\ell \mid x) = \sigma\left(r^*(x, y_w) - r^*(x, y_\ell)\right). \tag{2.1}$$

The preference likelihood in Equation (2.1) is the Bradley–Terry–Luce model (Bradley & Terry, 1952; Mosteller, 1951; Luce et al., 1959), where $\sigma(z) = (1 + e^{-z})^{-1}$ is the sigmoid function. This model is central to many preference-based learning algorithms, including Direct Preference Optimization (DPO) (Rafailov et al., 2023), and forms the basis of our likelihood-based objective.

The agent's goal is to learn a policy π that, for any given context x, selects actions that have high latent reward $r^*(x, y)$. The performance of the agent is measured by the *cumulative preference regret*, which quantifies the total opportunity cost incurred over T rounds. Let y_t be the action selected by the agent's policy at round t in context x_t, and let $y_t^* = \arg\max_{y \in \mathcal{Y}} r^*(x_t, y)$ be the optimal action for that context. The regret at round t is the difference in expected reward between the optimal action and the chosen action. The cumulative regret over T rounds is defined as:

$$\mathrm{Regret}(T) = \sum_{t=1}^{T} \left(r^*(x_t, y_t^*) - r^*(x_t, y_t)\right). \tag{2.2}$$

We will use Equation (2.2) as our formal notion of cumulative regret throughout the paper. The objective is to design an algorithm whose cumulative regret grows as slowly as possible with T. A logarithmic growth rate, $\mathrm{Regret}(T) = \mathbf{O}(\log T)$, is the theoretical ideal, indicating extremely efficient learning.

[2] Frequently used symbols are summarized in Table 2 in Appendix Section A.

[3] An assumption checklist appears in Table 3 in Appendix Section A.

2.2 Structural Assumptions on the Reward Class

To enable tractable learning from preference data alone, we impose a set of structural assumptions on the class of possible reward functions $\mathcal{R}$. These assumptions are standard in the theoretical analysis of learning with function approximation (Foster & Rakhlin, 2023) and are chosen to be as general as possible while still permitting strong performance guarantees.

Assumption 2.1. *We assume that the true latent reward function r^* belongs to a known, parameterized function class $\mathcal{R} = \{r_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$. This means there exists a true parameter vector $\theta^* \in \Theta$ such that $r^*(x, y) = r_{\theta^*}(x, y)$ for all (x, y). The parameter space Θ is assumed to be a compact set, implying a bounded norm $\|\theta\| \leq B$ for all $\theta \in \Theta$.*

This is a common starting point for theoretical analysis, allowing us to focus on the learning problem without the additional complication of model misspecification (Azar et al., 2024).

Assumption 2.2 (Lipschitz Continuity)**.** *We assume that the reward function parameterization is smooth. Specifically, the function class is L-Lipschitz with respect to the parameters: for all $\theta, \theta' \in \Theta$ and all (x, y), we have:*

$$|r_\theta(x, y) - r_{\theta'}(x, y)| \leq L\|\theta - \theta'\|_2. \tag{2.3}$$

This assumption is satisfied by many practical models, including neural networks with bounded weights and smooth activation functions. It is a crucial property that ensures that small changes in the parameter space lead to correspondingly small changes in the reward space, which is essential for generalization, optimization stability, and for relating parameter-space uncertainty to function-space uncertainty (Zhang, 2023).

Assumption 2.3. *This is the most critical assumption for enabling efficient exploration and achieving logarithmic regret. We assume that the function class $\mathcal{R}$ has a finite eluder dimension (Russo & Van Roy, 2013; 2014).*

Eluder dimension. We adopt the ϵ-eluder dimension $d_{\mathrm{eluder}}(\mathcal{R}, \epsilon)$ as the intrinsic complexity controlling regret in our analysis. For completeness, a concise definition together with its variance–information connection appears in Appendix D.2. Moreover, for LoRA-parameterized reward classes we establish sharp eluder control; see Proposition D.4 in Appendix D.3.

3 PAC-Bayesian Generalization and Wasserstein Gradient Flow

This section connects PAC-Bayesian generalization objective to a Wasserstein gradient-flow (WGF) description of the learning dynamics. We (i) motivate a PAC-Bayes objective as the optimization target, (ii) introduce a smoothed/projected–KL device that yields a sharpened bound suitable for particle posteriors, and (iii) show that steepest descent of this objective in the 2-Wasserstein geometry yields a Langevin diffusion and the associated Fokker–Planck (continuity) equation. Full statements with constants and all proofs are deferred to Section C and Section E.

Let $S = \{z_i\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, parameter space $\Theta \subseteq \mathbb{R}^d$, prior Π on Θ, posterior $\mu \in \mathcal{P}(\Theta)$, and per-example loss $\ell_\theta(z) \in [0, 1]$ that is L-Lipschitz in θ for each z. We write $\hat{L}_S(\mu) := \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\theta\sim\mu}\,\ell_\theta(z_i)$ and $\mathrm{Risk}\mu_{\mathcal{D}} := \mathbb{E}_{z\sim\mathcal{D}}\mathbb{E}_{\theta\sim\mu}\,\ell_\theta(z)$. For a Markov kernel S on Θ, $\mathsf{S}_{\#}\mu$ denotes the push forward, and the *projected KL* is

$$D_{\mathrm{KL}\mathsf{S}}(\mu\|\Pi) \;:=\; D_{\mathrm{KL}}\big(\mathsf{S}_{\#}\mu \,\|\, \mathsf{S}_{\#}\Pi\big),$$

which satisfies $D_{\mathrm{KL}\mathsf{S}}(\mu\|\Pi) \leq D_{\mathrm{KL}}(\mu\|\Pi)$ by data processing (see Theorem C.2 and Section C).

A classical PAC-Bayes inequality for a posterior μ independent of S reads

$$\mathrm{Risk}_{\mathcal{D}}(\mu) \leq \hat{L}_S(\mu) + \sqrt{\frac{D_{\mathrm{KL}}(\mu\|P) + \ln\frac{2\sqrt{m}}{\delta}}{2m}}. \tag{3.1}$$

This suggests optimizing the right-hand side by trading empirical fit against complexity. Introducing an inverse-temperature parameter $\beta > 0$ yields the variational objective

$$J_{\mathrm{PAC}}(\mu) \;=\; \hat{L}_S(\mu) \;+\; \beta\, D_{\mathrm{KL}}(\mu \,\|\, \Pi), \tag{3.2}$$

which is the *free energy* associated with empirical risk and prior regularization.

3.1 SMOOTHED/PROJECTED–KL PAC-BAYES BOUND

We now state the smoothed/projected variant that will be used both for theory (to control finite-particle posteriors) and for algorithms (to motivate noise schedules). The definition is given here, while the full theorem and constants appear in Section C.

Definition 3.1 (Projected/Smoothed KL)**.** *For $\mu, \Pi \in \mathcal{P}(\Theta)$ and any smoothing kernel (confer Definition C.1) S, define the* projected (smoothed) KL *by*

$$D_{\mathrm{KL}\mathsf{S}}(\mu\|\Pi) \;:=\; D_{\mathrm{KL}}\big(\mathsf{S}_{\#}\mu \,\|\, \mathsf{S}_{\#}\Pi\big).$$

By data processing for f-divergences, $D_{\mathrm{KL}\mathsf{S}}(\mu\|\Pi) \le D_{\mathrm{KL}}(\mu\|\Pi)$ when the right-hand side is finite. For the Gaussian kernel, we write $D_{\mathrm{KLS}_h}(\mu\|\Pi) := D_{\mathrm{KL}}(\mathsf{S}_{h,\#}\mu\|\mathsf{S}_{h,\#}\Pi)$.

Theorem 3.2 (PAC-Bayes via smoothing)**.** *Assume $\ell_\theta(z) \in [0,1]$ is L-Lipschitz in θ. Let $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$ be an N-particle posterior and let S_h denote Gaussian smoothing with variance $h^2 I_d$. For any prior Π independent of S and any $h > 0$, with probability at least $1-\delta$,*

$$\mathrm{Risk}{\mu^N}_{\mathcal{D}} \;\le\; \mathrm{Risk}{\mu^N}_{S} \;+\; Lh\,\mathbb{E}\|Z\| \;+\; \sqrt{\frac{D_{\mathrm{KLS}_h}(\mu^N\|\Pi) + \ln(2m/\delta)}{2m}},$$

where $Z \sim \mathcal{N}(0, I_d)$ so $\mathbb{E}\|Z\| \le \sqrt{d}$. Moreover, if $\Pi = \mathcal{N}(\theta_0, \sigma_0^2 I_d)$ then

$$D_{\mathrm{KLS}_h}(\mu^N\|\Pi) \;\le\; \frac{1}{2N(\sigma_0^2+h^2)}\sum_{i=1}^N \|\theta_i - \theta_0\|^2 \;+\; \frac{d}{2}\,\phi\Big(\tfrac{h^2}{\sigma_0^2+h^2}\Big), \text{with } \phi(\rho) = \rho - 1 - \ln\rho.$$

.

3.2 OPTIMIZATION DYNAMICS AS A WASSERSTEIN GRADIENT FLOW

Interpreting Equation (3.2) as a free-energy functional on $\mathcal{P}(\Theta)$, the 2-Wasserstein gradient flow of J_{PAC} is the continuity equation

$$\partial_t \mu_t \;=\; \nabla_\theta \cdot \big(\mu_t\, \nabla_\theta V[\mu_t]\big), \tag{3.3}$$

where $V[\mu]$ is any C^1 potential whose gradient equals the Wasserstein gradient of J_{PAC} at μ. Concretely, one may take

$$\nabla_\theta V[\mu](\theta) \;=\; \nabla_\theta\, \mathbb{E}_{z\sim S}\, \ell_\theta(z) \;+\; \beta\, \nabla_\theta\big(\log\mu(\theta) - \log\Pi(\theta)\big),$$

so that equation 3.3 coincides with the Fokker–Planck equation of the Langevin diffusion

$$d\theta(t) \;=\; -\nabla_\theta V[\mu_t]\big(\theta(t)\big)\, dt \;+\; \sqrt{2\beta}\, dW(t), \tag{3.4}$$

see, e.g., Jordan et al. (1998); Ambrosio et al. (2008); Villani (2008). Thus, gradient-based training of the free energy J_{PAC} admits an exact continuum description as WGF.

A first-order time discretization of equation 3.4 (Euler–Maruyama) with step size $\eta > 0$ yields the particle update

$$\theta_{k+1} \;=\; \theta_k \;-\; \eta\, \nabla_\theta V[\mu_k](\theta_k) \;+\; \sqrt{2\eta\beta}\, \xi_k, \text{with } \xi_k \sim \mathcal{N}(0, I_d)$$

Replacing full gradients with mini-batch estimates recovers SGLD. This principled discretizations exposes and quantifies the approximation gaps that drive our regret analysis (precise bounds in Section E):**Finite-ensemble gap** (Monte Carlo drift error): $\mathbf{O}\big(\sqrt{\sum_t v_t^2/N_t}\big)$. **Stochastic-gradient gap** (mini-batch noise): $\mathbf{O}\big(\sqrt{\sum_t \sigma_t^2/B_t}\big)$. **Discretization gap** (time stepping): $\mathbf{O}(\eta T)$. These terms map exactly onto the four sources of error isolated in the Introduction.

4 THE OPTIMISTIC LANGEVIN ENSEMBLE (OLE) ALGORITHM

This section translates the theoretical framework developed in the preceding sections into a concrete, self-contained algorithm for preference-based contextual bandits. The algorithm, which we call the **Optimistic Langevin Ensemble (OLE)**, instantiates the discretized Wasserstein gradient flow perspective. It maintains a finite ensemble of reward models, updates them using stochastic Langevin dynamics, and makes decisions using an optimistic selection rule based on ensemble

statistics. The specific variant for online contextual bandits is termed Optimistic Thompson Sampling with Langevin Ensembles (O-TSLE).

The OLE algorithm operates in rounds. At each round t, it leverages its current posterior belief about the reward function, represented by an ensemble of particles, to optimistically select an action. It then observes the resulting preference feedback and updates its posterior belief using a Langevin step. Pseudo-code of additional variants are provided in Appendix G.1, such as for online contextual bandits and MDP scenarios.

Algorithm 1: Optimistic Langevin Ensemble (OLE): Generic Template

Input: Prior Π_0; step sizes $\{\eta_t\}$; ensemble sizes $\{N_t\}$; batch sizes $\{B_t\}$; optimism schedule $\{\kappa_t\}$

for $t = 1, 2, \ldots, T$ **do**
- Observe context x_t;
- `// Optimistic Selection`
- Compute ensemble mean $\hat{r}_t(x_t, y)$ and variance $\widehat{\mathrm{Var}}_t(x_t, y)$ for all $y \in \mathcal{Y}$;
- Construct optimistic index: $I_t(x_t, y) \leftarrow \hat{r}_t(x_t, y) + \kappa_t \sqrt{\widehat{\mathrm{Var}}_t(x_t, y)}$;
- Select action pair $(y_t^{(w)}, y_t^{(\ell)})$ based on maximizing information gain using $\{I_t(x_t, y)\}_{y \in \mathcal{Y}}$;
- Receive preference feedback, forming data batch $\mathcal{D}_t$;
- `// Posterior Update (SGLD)`
- Compute mini-batch gradient $\widehat{\nabla}_t$ of $J_{\mathrm{PAC}}(\theta) = \hat{L}_{\mathcal{D}_t}(\theta) + \beta D_{\mathrm{KL}}(\delta_\theta \| \Pi_{t-1})$;
- **for** $i = 1, \ldots, N_t$ **do**
  - Draw Gaussian noise $\xi_t^{(i)} \sim \mathcal{N}(0, I)$;
  - $\theta_{t+1}^{(i)} \leftarrow \theta_t^{(i)} - \eta_t \, \widehat{\nabla}_t J_{\mathrm{PAC}}(\theta_t^{(i)}) + \sqrt{2\eta_t \beta}\, \xi_t^{(i)}$;

The core components of the algorithm are as follows:

- **Ensemble Maintenance:** The algorithm's belief about the true reward parameter θ^* is represented by an ensemble of N_t particles, $\{\theta_t^{(i)}\}_{i=1}^{N_t}$. This ensemble serves as a Monte Carlo approximation of the posterior distribution μ_t. At the start of learning ($t = 0$), these particles are drawn from a prior distribution Π_0.
- **Langevin Update Step:** This is the learning step of the algorithm. After receiving new preference data $\mathcal{D}_t$, each particle in the ensemble is updated using one step of Stochastic Gradient Langevin Dynamics (SGLD). The gradient is computed with respect to the PAC-Bayesian objective J_{PAC} on a mini-batch of the new data. This update moves the particles towards regions of the parameter space that better explain the observed preferences, while the injected Gaussian noise ensures that the ensemble continues to represent a distribution and does not collapse to a single point.
- **Optimistic Selection Rule:** This is the exploration mechanism of the algorithm and the component that addresses the fourth implementation gap (intractable uncertainty). To make decisions that efficiently balance exploration and exploitation, the agent needs an upper confidence bound (UCB) on the true, unknown reward function r^*. Computing the exact Bayesian UCB is intractable for complex models. The OLE algorithm therefore uses a computationally feasible proxy based on the statistics of its particle ensemble. For each candidate action y in the current context x_t, it computes an optimistic index:

$$I_t(x_t, y) = \hat{r}_t(x_t, y) + \kappa_t \cdot \sqrt{\widehat{\mathrm{Var}}_t(x_t, y)}. \tag{4.1}$$

  The exploration bonus in Equation (4.1) follows the eluder-dimension view of exploration (Russo & Van Roy, 2013; 2014) and yields the desired logarithmic-regret scaling (Hazan et al., 2007).

Here, $\hat{r}_t(x_t, y)$ is the mean reward predicted by the ensemble, serving as the best guess for the true reward. $\widehat{\mathrm{Var}}_t(x_t, y)$ is the variance of the reward predictions across the ensemble, which serves as a proxy for the posterior uncertainty about the reward of that action. The parameter κ_t is an optimism coefficient that controls the weight given to this uncertainty, effectively determining how much the agent prioritizes exploration. The agent then selects a pair of actions to query for a preference based on these optimistic indices, typically choosing a pair that is expected to be most informative for

resolving the current uncertainty. While the exact Bayesian posterior uncertainty is intractable for complex models, we will show in our analysis (Section 5) that the ensemble variance serves as a theoretically sound proxy. This is because of a fundamental duality between variance and information gain , which ensures that exploring regions of high ensemble variance leads to an efficient reduction of uncertainty about the true reward function, thereby enabling logarithmic regret.

5 Regret Analysis

This section presents the main theoretical result of the paper: a unified, high-probability regret bound for the Optimistic Langevin Ensemble (OLE) algorithm. The bound demonstrates that the algorithm achieves a cumulative regret that scales logarithmically with the time horizon T, plus explicit, sublinear terms that quantify the costs of the practical approximations corresponding to the "four gaps." This result provides a rigorous theoretical explanation for the remarkable sample efficiency of preference-based learning. Full proofs are in Appendix Section E.

Our main theorem bounds the cumulative preference regret of the OLE algorithm. It shows that the regret is controlled by the intrinsic complexity of the reward function class, as measured by the eluder dimension, and by the parameters governing the algorithmic approximations.

Theorem 5.1. *Let Assumptions 2.1 (Realizability), 2.2 (Lipschitz Continuity), and 2.3 (Finite Eluder Dimension) hold. For any $\delta \in (0,1)$, consider the* OLE *algorithm run for T rounds with step sizes $\{\eta_t\}$, ensemble sizes $\{N_t\}$, mini-batch sizes $\{B_t\}$, and an optimism schedule $\kappa_t = C_0\sqrt{\log(T/\delta)}$ for a suitable constant C_0. Let v_t^2 be an upper bound on the conditional variance of the Monte Carlo estimate of the optimistic value, and let σ_t^2 be an upper bound on the conditional variance of the mini-batch gradient estimator. Then with probability at least $1-\delta$, the cumulative regret satisfies:*

$$\mathrm{Regret}(T) \le \underbrace{C_1\, d_{\mathrm{eluder}} \log T}_{\textit{Exploration Cost}} + C_2 \left(\underbrace{\sum_{t=1}^{T} \eta_t}_{\textit{Discretization}} + \underbrace{\mathbf{O}\left(\sqrt{\sum_{t=1}^{T} \frac{v_t^2}{N_t}} \right)}_{\textit{Finite Ensemble}} + \underbrace{\mathbf{O}\left(\sqrt{\sum_{t=1}^{T} \frac{\sigma_t^2}{B_t}} \right)}_{\textit{Stochastic Gradient}} \right), \quad (5.1)$$

where C_1 and C_2 are absolute constants. The eluder dimension d_{eluder} is evaluated at a precision scale ϵ that decreases with t, such as $\epsilon_t = 1/(1+t)$.

Remark 5.2 (On tightness of the leading term)**.** *Up to polylogarithmic factors, the $\mathbf{O}(d_{\mathrm{eluder}} \log T)$ leading term in our regret bound matches known lower bounds and optimal algorithms for contextual bandits with rich (e.g., generalized linear) function classes, where the eluder dimension governs sample complexity (Russo & Van Roy, 2013; 2014). In particular, the $\log T$ factor is information-theoretically unavoidable even in parametric bandit settings with well-specified models (Hazan et al., 2007).*

This bound provides a comprehensive picture of the algorithm's performance and completes the narrative arc of bridging the four gaps. Each term has a precise interpretation:

- **The Exploration Term:** $C_1 d_{\mathrm{eluder}} \log T$. This is the leading-order term and represents the fundamental statistical cost of exploration. Its logarithmic dependence on the horizon T is the key result, confirming that the algorithm learns extremely efficiently. The cost scales linearly with the eluder dimension d_{eluder}, which captures the intrinsic complexity of the learning problem. This term arises directly from the use of an optimistic exploration strategy.
- **The Discretization Error:** $\sum_{t=1}^{T} \eta_t$. This term quantifies the cost of Gap 3: approximating the continuous-time Wasserstein gradient flow with a discrete-time algorithm. It represents the cumulative bias from the Euler-Maruyama discretization. For a constant step size η, this error is $\mathbf{O}(\eta T)$. However, as shown in the corollary below, this term can be made negligible by using a decreasing step size schedule.
- **The Finite-Ensemble Error:** $\mathbf{O}(\sqrt{\sum_{t=1}^{T} v_t^2/N_t})$. This term quantifies the cost of Gap 1: approximating the true posterior distribution with a finite ensemble of N_t particles. It represents the accumulated Monte Carlo estimation error. The term grows sub-linearly in T and decreases as the ensemble size N_t increases, explicitly characterizing the trade-off between computational cost and statistical accuracy.

- **The Stochastic Gradient Error:** $\mathbf{O}(\sqrt{\sum_{t=1}^{T} \sigma_t^2 / B_t})$. This term quantifies the cost of Gap 2: using noisy mini-batch gradients instead of exact full-batch gradients. It represents the accumulated noise from the stochastic optimization process. Like the ensemble error, it grows sub-linearly and decreases as the mini-batch size B_t increases.

In the idealized limit where $\eta_t \to 0$, $N_t \to \infty$, and $B_t \to \infty$, all three lower-order terms vanish, and we are left with a purely logarithmic regret bound, $\mathrm{Regret}(T) = \mathbf{O}(d_{\mathrm{eluder}} \log T)$. Our theorem provides the first analysis that makes this trade-off explicit for preference-based RL.

Corollary 5.3. *If the step sizes and resource allocation schedules are chosen such that $\sum_{t=1}^{T} \eta_t = \mathbf{O}(1)$, $\sum_{t=1}^{T} v_t^2 / N_t = \mathbf{O}(1)$, and $\sum_{t=1}^{T} \sigma_t^2 / B_t = \mathbf{O}(1)$, then under the assumptions of Theorem 5.1, the cumulative regret is:*

$$\mathrm{Regret}(T) = \mathbf{O}\left(d_{\mathrm{eluder}} \log T\right). \tag{5.2}$$

This corollary shows that by using standard schedules, such as a decreasing step size $\eta_t \propto 1/t$ and geometrically increasing ensemble and batch sizes, the approximation errors can be rendered into constant, lower-order terms, achieving the theoretical ideal.

Remark 5.4. *As discussed in Section 2, the eluder dimension can be related to the intrinsic dimensionality of the learning task. For models fine-tuned with low-rank adaptation (LoRA), the eluder dimension d_{eluder} is controlled not by the total number of parameters d, but by the much smaller intrinsic rank d_* (Hu et al., 2022; Yang et al., 2023). Consequently, the regret bounds in Theorem 5.1 and Corollary 5.3 scale as $\mathbf{O}(d_* \log T)$. This provides a direct and rigorous theoretical explanation for the empirical observation that parameter-efficient fine-tuning methods can achieve high sample efficiency even on massive models.*

6 Extensions to Markov Decision Processes

To demonstrate the versatility and power of our theoretical framework, we extend the analysis from the contextual bandit setting to the more general and challenging setting of Markov Decision Processes (MDPs). This extension requires handling temporal dependencies, long-term credit assignment, and the propagation of uncertainty through Bellman updates. We show that our optimistic PAC-Bayesian ensemble approach can be naturally adapted to both finite-horizon and discounted MDPs, yielding analogous logarithmic regret guarantees. Proofs in Appendix Section F.

6.1 Setup for Preference-Based MDPs

A finite-horizon MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, H, P, r^*, \rho_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, H is the horizon, P are the transition dynamics, r^* is the latent reward function, and ρ_0 is the initial state distribution. In the preference-based RL setting, the agent does not observe the numeric rewards $r^*(s, a)$. Instead, it receives preference feedback, typically comparing entire trajectories or state-action pairs. The agent's objective is to learn a policy $\pi = \{\pi_h\}_{h=1}^{H}$ that maximizes the expected cumulative latent reward.

To enable value-based learning algorithms, we require an additional structural assumption beyond those for the bandit case.

Assumption 6.1. *We assume the function class for the action-value function (Q-function) is approximately closed under the Bellman optimality operator. That is, for any Q-function in our class, applying one step of Bellman backup results in a function that is still close to (or within) the class (Agarwal et al., 2023; Jin et al., 2021). This is a standard assumption in the theory of RL with function approximation, ensuring that the value functions produced during learning remain representable within our chosen model class.*

6.2 The O-TDLE Algorithm for MDPs

We adapt our OLE algorithm to the MDP setting, resulting in a method we call Optimistic TD with Langevin Ensembles (O-TDLE). The core idea remains the same: maintain an ensemble of models to represent the posterior distribution and use optimistic exploration. The key difference is that the ensemble now represents the Q-function, and the updates are driven by temporal difference errors.

The O-TDLE algorithm (detailed in Algorithm 5)proceeds in episodes. At each step h within an episode, the agent is in state s_h. It uses its ensemble of Q-function models, $\{Q_{\theta^{(i)}}\}_{i=1}^{N}$, to compute an optimistic index for each action $a \in \mathcal{A}$:

$$I_h(s_h, a) = \hat{Q}_h(s_h, a) + \kappa_h \cdot \sqrt{\widehat{\mathrm{Var}}_h(Q(s_h, a))}, \tag{6.1}$$

where $\hat{Q}_h$ and $\widehat{\mathrm{Var}}_h$ are the mean and variance of the Q-value predictions across the ensemble. The agent then selects the action $a_h = \arg\max_{a \in \mathcal{A}} I_h(s_h, a)$. After executing the action and observing the next state s_{h+1}, the agent collects preference data (e.g., by comparing the executed trajectory segment to a reference—such as a SFT model). This data is then used to perform an SGLD update on the ensemble parameters $\{\theta^{(i)}\}$, using a loss derived from a Bellman-style TD error consistent with the preference feedback.

6.3 Regret Analysis for MDPs

We prove that the O-TDLE algorithm achieves a logarithmic regret bound in the MDP setting. The bound now includes a polynomial dependence on the horizon H, which is expected as errors can propagate and compound over the steps of an episode.

Theorem 6.2. *Under Assumptions 2.1-2.3 and 6.1, the* O-TDLE *algorithm, run for T episodes, achieves a cumulative regret that satisfies, with high probability:*

$$\mathrm{Regret}(T) = \mathbf{O}\left(H^2 \cdot d_{\mathrm{eluder}} \cdot \log T\right) + \textit{lower-order approximation terms}. \tag{6.2}$$

The lower-order terms for discretization, finite-ensemble, and stochastic gradient errors have a similar structure to the bandit case, now summed over all steps and episodes.

Remark 6.3 (On the H-dependence)**.** *Our bound incurs an H^2 factor in the leading term, which is standard for episodic finite-horizon analyses under function approximation. Improving the H-dependence typically requires stronger structural assumptions (e.g., linear MDPs or Bellman completeness with additional mixing/realizability properties) or refined variance decompositions; see, e.g., Azar et al. (2024); Jin et al. (2021).*

Our proof for the MDP setting employs a powerful policy decomposition technique, inspired by recent advances in the analysis of KL-regularized RL with numeric rewards Zhao et al. (2025a). This technique allows us to reduce the multi-step credit assignment problem to a sequence of bandit-like analyses, to which our core optimistic exploration argument can be applied. The novelty of our approach lies in adapting this tool to the preference-based feedback setting and integrating it within our PAC-Bayesian particle ensemble framework. A similar analysis can be performed for the infinite-horizon discounted MDP setting, yielding a regret bound with a polynomial dependence on the effective horizon $(1-\gamma)^{-1}$.

7 Conclusion, Limitations, and Future Work

In this work, we developed a unified optimistic PAC-Bayesian framework for preference-based learning that closes several critical gaps between theory and practice. Our analysis provides the first theoretical explanation for the sample efficiency of modern alignment pipelines by establishing a near-logarithmic regret bound, $\mathbf{O}(d_{\mathrm{eluder}} \log T)$, that explicitly accounts for the algorithmic costs of using finite ensembles, stochastic gradients, and discrete-time updates. Our framework provides a firm theoretical foundation for the empirical success of methods like DPO (Rafailov et al., 2023) and connects the complexity of exploration to the intrinsic dimensionality of parameter-efficient fine-tuning (Aghajanyan et al., 2020; Hu et al., 2022).

Limitations and Future works. Our theoretical guarantees rely on standard but strong structural assumptions. The realizability assumption, which posits that the true reward function lies within the model class, is a significant idealization for complex models like LLMs, which are likely to be misspecified (Foster & Rakhlin, 2023). Similarly, our extension to MDPs requires Bellman completeness, a condition known to be restrictive for reinforcement learning with general function approximation (Agarwal et al., 2023; Golowich & Moitra, 2024; Wu et al., 2024). Finally, the decoupled structure of our regret bound opens the door to designing adaptive algorithms that can dynamically schedule computational resources, such as ensemble and mini-batch sizes, to optimally balance the statistical and computational trade-offs inherent in practical alignment.

Ethics Statement

This work is theoretical, focusing on the algorithmic foundations of preference learning for the alignment of large language models. As with any alignment methodology, the practical application of our framework carries potential risks. These include *over-optimization* to the learned reward model, which may not perfectly capture nuanced human intent, and the potential for malicious *reward hacking*. We emphasize that our algorithms are designed for statistical and computational efficiency in optimizing a given preference model; they do not define the values inherent in that model. The collection and curation of the preference data that serves as the source of these values must be approached with care to respect privacy and mitigate the encoding and amplification of societal biases. Appropriate guardrails, diverse data sourcing, and multi-faceted evaluation of aligned models remain necessary to mitigate unintended consequences.

The Use of Large Language Models

In this work, the authors used generative AI tools (ChatGPT-5) to aid in and polish the writing of this paper. We use the following prompt to check the language section by section (including abstract): "Check the following statement, examine if the narrative is professional and understandable for broader audience in the area of machine learning community, and examine if the language meets native speaker standard. If not, generate feedback on how should I modify my narratives." All LLM-generated content was thoroughly reviewed and verified by the authors prior to inclusion. Research design, critical analyses, and all final decisions were carried out independently by the authors.

Reproducibility Statement

This work is entirely theoretical. To ensure the reproducibility of our results, we provide complete and self-contained proofs for all theorems, propositions, and lemmas in the appendix. The appendix also contains detailed pseudocode for our proposed algorithms (Appendix G), a full discussion of the structural assumptions (Appendix A), and guidance on the hyperparameter schedules required to achieve the stated regret bounds. All cross-references within the document are hyperlinked for ease of navigation.

References

Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 987–1063. PMLR, 2023.

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Pierre Alquier. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, 2 edition, 2008.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

Olivier Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

Dylan J Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*, 2023.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Noah Golowich and Ankur Moitra. Linear bellman completeness suffices for efficient online reinforcement learning with few actions. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1939–1981. PMLR, 2024.

Benjamin Guedj. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A pac-bayesian link between generalisation and flat minima. *arXiv preprint arXiv:2402.08508*, 2024.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Gene Li, Pritish Kamath, Dylan J Foster, and Nati Srebro. Understanding the eluder dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750, 2022.

Linshan Liu, Mateusz B. Majka, and Łukasz Szpruch. Polyak–łojasiewicz inequality on the space of measures and convergence of mean-field birth-death processes. *Applied Mathematics & Optimization*, 87(2):48, 2023.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew G Wilson. Pac-bayes compression bounds so tight that they can explain generalization. *Advances in Neural Information Processing Systems*, 35:31459–31473, 2022.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

David A. McAllester. PAC-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pp. 164–170, 1999.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Frederick Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1):3–9, 1951.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

A. Pacchiano, A. Saha, and J. Lee. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Taiji Suzuki, Denny Wu, and Atsushi Nitanda. Mean-field langevin dynamics: Time-space discretization, stochastic gradient, and variance reduction. In *NeurIPS*, 2023.

Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pp. 165–251. Springer, 2006.

Louis L Thurstone. A law of comparative judgment. In *Scaling*, pp. 81–92. Routledge, 2017.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:76006–76032, 2023.

Runzhe Wu, Ayush Sekhari, Akshay Krishnamurthy, and Wen Sun. Computationally efficient rl under linear bellman completeness for deterministic dynamics. *arXiv preprint arXiv:2406.11810*, 2024.

W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*, 2023.

Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

T Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and rlhf. *arXiv preprint arXiv:2411.04625*, 2024.

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic regret for online kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025a.

Qingyue Zhao, Kaixuan Ji, Heyang Zhao, Tong Zhang, and Quanquan Gu. Nearly optimal sample complexity of offline kl-regularized contextual bandits under single-policy concentrability. *arXiv e-prints*, pp. arXiv–2502, 2025b.

APPENDIX CONTENTS

- **Section A**: Notation used throughout and additional background definitions (including the formal eluder definition and its variance–information link).
- **Section B**: Extended related work.
- **Section C**: Canonical smoothed/projected–KL PAC-Bayes bound with full proofs.
- **Section D**: Technical lemmas (variance–information inequality, discretization, stochastic-gradient control, Monte Carlo concentration).
- **Section E**: Complete statements and proofs of the unified regret theorem and supporting results.
- **Section F**: Full proofs for finite-horizon and discounted MDP extensions.
- **Section G**: Implementation notes and additional pseudocode.

A NOTATION AND ADDITIONAL BACKGROUND

This appendix provides the complete theoretical underpinnings for the results presented in the main paper. We begin by establishing a unified notational system and providing a deeper discussion of the foundational concepts that motivate our work. This ensures the appendix is self-contained and accessible to readers with background in machine learning.

A.1 NOTATION

We summarize the most frequently used symbols throughout the paper and this appendix in Table 2 for ease of reference. This consistent notation is crucial for maintaining clarity throughout the complex derivations that follow.

Table 2: Notation used throughout the paper and appendix.

Symbol	Meaning
$\mathcal{X}, \mathcal{Y}$	Context and candidate/output spaces
$\mathcal{S}, \mathcal{A}$	State and action spaces (for MDPs)
$r^*(\cdot)$	Ground-truth latent reward function, parameterized by θ^*
π, π_t	Policy (at round t)
Π_t, μ_t	Posterior distribution over parameters θ at round t
μ_t^N	Empirical measure of the N-particle ensemble at time t
N_t, B_t, η_t	Ensemble size, mini-batch size, and step size at round t
d_{eluder}	Eluder dimension of the reward function class $\mathcal{R}$
γ	Discount factor (for discounted MDPs)
β	Inverse temperature in the PAC-Bayesian objective and SGLD updates
κ_t	Optimism/bonus coefficient at round t
$\mathrm{Regret}(T)$	Cumulative preference regret up to time T
$\mathrm{W}_2(\cdot, \cdot)$	The 2-Wasserstein distance between two probability measures

A.2 ASSUMPTION CHECKLIST

How to read Table 3. Each row states an assumption, its informal meaning, and where it is used (by theorem/lemma label). This helps map the dependency structure of the proofs and where each assumption is required.

How to read Table 4. We separate the bandit and MDP settings and indicate which assumptions are active in each, together with any setting-specific constants.

Table 3: Assumptions at a glance: informal summary and usage.

Name	Informal content	Used in
Realizability	Rewards lie in the model class	Theorem 5.1
Lipschitzness	Gradients/Loss are L-Lipschitz in parameters	Theorem 3.2, Theorem 5.1
Eluder finiteness	Finite eluder dimension of the class	Exploration term in Theorem 5.1
Block smoothness	Stability of discretized flow	Theorem D.7
Martingale control	Freedman-style concentration	Theorem D.10

Table 4: Assumptions at a glance (by setting).

Setting	Assumptions
Bandits / Contextual preference	Theorems 2.1 to 2.3
Finite-horizon MDPs	Theorem 6.1

A.3 Detailed Discussion of Theoretical Gaps

The introduction highlighted four critical gaps between idealized theory and practical RLHF implementations. Here, we elaborate on why each gap presents a formidable theoretical challenge and how their interplay necessitates a unified analysis.

- **Gap 1 (Finite Ensembles vs. Mean-Field):** Many theoretical analyses of particle-based systems, especially those leveraging tools from optimal transport (Jordan et al., 1998; Ambrosio et al., 2008), operate in the mean-field limit where the number of particles $N \to \infty$. In this limit, the empirical distribution of particles converges to the solution of a deterministic partial differential equation (the Fokker-Planck equation), a phenomenon known as propagation of chaos (Sznitman, 2006). However, practical implementations use small, finite ensembles (N is often less than 10). This introduces a non-trivial Monte Carlo sampling error at each step, as the interaction term in the particle dynamics depends on the empirical measure, not the true mean-field distribution. Our analysis must quantify this error and ensure it does not accumulate uncontrollably.
- **Gap 2 (Stochastic vs. Exact Gradients):** Large-scale model training is computationally infeasible without mini-batch stochastic gradients. While the noise introduced by mini-batching is zero-mean, its cumulative effect over T rounds is a significant source of error. The variance of this noise depends on the batch size B_t and the local curvature of the loss landscape. A rigorous analysis cannot simply assume gradients are exact; it must employ tools like martingale concentration inequalities to bound the accumulated deviation caused by this stochasticity.
- **Gap 3 (Discrete-Time vs. Continuous-Time):** The Wasserstein gradient flow perspective provides a powerful, continuous-time picture of the ideal optimization path. However, algorithms are implemented with a discrete step size η_t. The standard method for discretizing the underlying Langevin SDE is the Euler-Maruyama scheme. This introduces a discretization bias at each step, and the cumulative bias can grow linearly with T if not carefully controlled, potentially overwhelming the desired logarithmic regret term. Our analysis must explicitly account for this weak error and show how to manage it with a proper step-size schedule.
- **Gap 4 (Tractable vs. Intractable Uncertainty):** The principle of optimism requires an upper confidence bound on the true reward function. For complex models like neural networks, the true Bayesian posterior variance is intractable to compute. Practical algorithms use the variance of predictions across the finite ensemble as a proxy for uncertainty. While intuitive, it is not a priori guaranteed that this ensemble variance is a valid upper bound on the true posterior uncertainty. A central part of our theoretical contribution is to formally justify this proxy and prove that it is sufficient to drive efficient exploration.

A crucial point is the interdependence of these gaps. The noise from stochastic gradients (Gap 2) can interact with and amplify the discretization error (Gap 3). The quality of the finite-ensemble approximation (Gap 1) directly determines the reliability of the uncertainty proxy used for exploration (Gap 4). A successful theory, therefore, cannot analyze these in isolation. Our unified framework is

designed to bound the sum of these interacting error terms, demonstrating that their interplay does not lead to a catastrophic amplification of regret.

A.4 Contributions to Formal Results Map

To provide a clear roadmap for the reader, Table 5 explicitly links the main contributions of this work to the formal theorems and proofs contained within this appendix. This table serves as a guide to verifying each of our central claims.

Table 5: Map of contributions to their formal statements and proofs in the appendix.

Contribution	**Formal Statement (Proof Location)**
Unified PAC-Bayesian particle theory	Theorem D.1 (App. D.1)
Unified regret bound for bandits	Theorem 5.1 (App. E.2)
Finite-sample error decomposition	Lemmas D.7, D.8, D.9, D.10 (App. D.4)
Extension to finite-horizon MDPs	Theorem 6.2 (App. F.2)
Extension to discounted MDPs	Section F.1 (App. F.3)
Eluder dimension for LoRA	Theorem D.4 (App. D.3)
Well-tuned schedules corollary	Theorem 5.3 (App. E.2)
Algorithmic pseudocode	OLE/OTSLE/OTDLE (App. G.1; Algs. 2, 3, 5)

B Extended Related Work

Our work connects to and builds upon several distinct but related lines of research in machine learning theory and practice.

RLHF and Direct Preference Optimization. The modern paradigm of aligning LLMs was established by large-scale RLHF pipelines (Ouyang et al., 2022; Bai et al., 2022; Dong et al., 2024), which combine preference data collection, reward modeling, and policy optimization. More recent direct preference optimization methods, such as DPO and its variants (Rafailov et al., 2023; Meng et al., 2024), have streamlined this process and demonstrated strong empirical performance. Our work provides a foundational theoretical explanation for the remarkable sample efficiency observed in these practical systems, showing that near-logarithmic regret is achievable.

Preference Learning, Dueling Bandits, and RL with Preferences. The problem of learning from comparative feedback has a long history, rooted in foundational statistical models like the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce et al., 1959; Thurstone, 2017). In the online setting, this problem is formalized as the *dueling bandits* problem, for which a rich body of literature provides sample complexity guarantees, typically achieving $\mathbf{O}(\sqrt{T})$ regret in general settings and $\mathbf{O}(\log T)$ in more restricted tabular or linear cases (Yue & Joachims, 2009; Yue et al., 2012). Extensions to reinforcement learning with preferences have been studied, but these analyses often yield sub-optimal $\mathbf{O}(\sqrt{T})$ regret for general function classes (Wang et al., 2023; Pacchiano et al., 2021). Our work is the first to establish a near-logarithmic regret bound for preference-based RL with general non-linear function approximation.

KL-Regularized Bandits and RL (Numeric Rewards). Our work is complementary to the important and emerging body of theory on KL-regularized bandits and RL, which has also achieved logarithmic regret guarantees but in the distinct setting of *numeric rewards* (Xiong et al., 2024; Zhao et al., 2024; 2025a;b) and often under additional structural assumptions like data coverage. While this parallel line of work provides deep insights into policy optimization given a numeric reward, our work addresses the more foundational problem of learning the reward function itself from *pairwise preference feedback*. This is the canonical setup for RLHF and DPO, where the reward model is the primary object to be learned from human comparisons. Our analysis is therefore algorithm-native, deriving guarantees directly from a PAC-Bayesian treatment of particle ensembles, rather than from the specific optimization landscape of a KL-regularized objective.

PAC-Bayes, Optimism, and Thompson Sampling. Our theoretical approach is built on the foundations of PAC-Bayesian learning theory, which provides powerful, high-probability generalization bounds for randomized predictors (McAllester, 1999; Catoni, 2007; Alquier, 2021; Guedj, 2019). Recent work has shown the power of PAC-Bayesian analysis for explaining generalization in deep learning (Lotfi et al., 2022; Haddouche et al., 2024). We combine these tools with the classical principle of optimism-in-the-face-of-uncertainty from the bandit literature (Hazan et al., 2007). The complexity of exploration in our framework is measured by the eluder dimension (Russo & Van Roy, 2013; 2014), a concept central to achieving logarithmic regret in benign regimes. Our optimistic posterior update mechanism is conceptually related to feel-good Thompson sampling (Zhang, 2022), but is tailored to the preference-based setting and analyzed via PAC-Bayesian tools.

Particle Approximations and Optimal-Transport Tools. To rigorously analyze the behavior of our finite-ensemble algorithm, we interpret its dynamics as a discretization of a Wasserstein gradient flow on the space of probability measures (Jordan et al., 1998). We control the approximation error introduced by the finite number of particles using tools from optimal transport theory and the study of empirical measures (Ambrosio et al., 2008; Villani, 2008; Fournier & Guillin, 2015; Sznitman, 2006). The analysis of the stochastic gradient and discretization errors is informed by the literature on the convergence of stochastic-gradient Langevin-type methods (Liu et al., 2023; Suzuki et al., 2023), allowing us to derive explicit, non-asymptotic lower-order terms in our regret bound.

In summary, prior analyses for preference-based learning typically achieve $\mathbf{O}(\sqrt{T})$ regret for general function classes. In parallel, analyses of KL-regularized learning with numeric rewards have achieved $\mathbf{O}(\log T)$ regret, sometimes under strong assumptions. Our work is the first to deliver a near-logarithmic regret bound for the fundamental problem of *pairwise preference feedback* within a framework that is faithful to the practical algorithms used in RLHF, thereby closing a critical gap between theory and practice.

C Smoothed/Projected–KL PAC-Bayes and WGF: Full Statements and Proofs

This section consolidates all technical material supporting Section 3. We (i) formalize the projected–KL device and prove the smoothed PAC-Bayes theorem with full constants, (ii) record the standard calculus linking the free energy to a Wasserstein gradient flow and the associated Langevin/Fokker–Planck dynamics, and (iii) point to where the end-to-end regret proofs and allocation/scheduling lemmas are proved in Section E.

C.1 Projected–KL smoothing and basic properties

We recall the projected divergence used in the main text.

Definition C.1 (Smoothing kernel and pushforward)**.** *Let $(\Theta, \mathcal{B})$ be a measurable parameter space. A* smoothing kernel *is a Markov kernel $\mathsf{S} : \Theta \times \mathcal{B} \to [0,1]$, i.e., for each $\theta \in \Theta$, $\mathsf{S}(\theta, \cdot)$ is a probability measure and for each $A \in \mathcal{B}$, $\theta \mapsto \mathsf{S}(\theta, A)$ is measurable. For a probability measure $\mu \in \mathcal{P}(\Theta)$, its pushforward by S is*

$$(\mathsf{S}_{\#}\mu)(A) \;:=\; \int_{\Theta} \mathsf{S}(\theta, A)\, \mu(d\theta), \qquad A \in \mathcal{B}.$$

When $\Theta = \mathbb{R}^d$ and $h > 0$, the Gaussian smoothing kernel *is $\mathsf{S}_h(\theta, \cdot) := \mathcal{N}\big(\theta, h^2 I_d\big)$, in which case $\mathsf{S}_{h,\#}\mu = \mu * \mathcal{N}(0, h^2 I_d)$ is the usual Gaussian convolution. We write $\mathsf{S}_h := \mathsf{S}_h$ for brevity.*

Definition C.2 (Projected/Smoothed KL)**.** *For $\mu, \Pi \in \mathcal{P}(\Theta)$ and any smoothing kernel S, define the* projected (smoothed) KL *by*

$$D_{\mathrm{KL}\mathsf{S}}(\mu \| \Pi) \;:=\; D_{\mathrm{KL}}\big(\mathsf{S}_{\#}\mu \,\|\, \mathsf{S}_{\#}\Pi\big).$$

By data processing for f-divergences, $D_{\mathrm{KL}\mathsf{S}}(\mu\|\Pi) \le D_{\mathrm{KL}}(\mu\|\Pi)$ when the right-hand side is finite. For the Gaussian kernel of Definition C.1, we write $D_{\mathrm{KL}\mathsf{S}_h}(\mu\|\Pi) := D_{\mathrm{KL}}(\mathsf{S}_{h,\#}\mu \| \mathsf{S}_{h,\#}\Pi)$.

C.2 SMOOTHED/PROJECTED–KL PAC-BAYES BOUND: FULL STATEMENT AND PROOF

We now give the full version of Theorem 3.2 including constants and a convenient specialization for Gaussian priors.

Theorem C.3 (PAC-Bayes via smoothing; full)**.** *Assume $\ell_\theta(z) \in [0,1]$ is L-Lipschitz in θ for each z. Let $S = \{z_i\}_{i=1}^m \overset{i.i.d.}{\sim} \mathcal{D}$, and let $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$ be any N-particle posterior (possibly data-dependent). For any prior Π independent of S, any $h > 0$, and any $\delta \in (0,1)$, with probability at least $1-\delta$ over S,*

$$\mathrm{Risk}\mu^N{}_{\mathcal{D}} \;\le\; \mathrm{Risk}\mu^N{}_{S} \;+\; Lh\,\mathbb{E}\|Z\| \;+\; \sqrt{\frac{D_{\mathrm{KL}\mathsf{S}_h}(\mu^N\|\Pi) + \ln(2m/\delta)}{2m}},$$

where $Z \sim \mathcal{N}(0, I_d)$ so that $\mathbb{E}\|Z\| \le \sqrt{d}$. Moreover, if $\Pi = \mathcal{N}(\theta_0, \sigma_0^2 I_d)$, then

$$D_{\mathrm{KL}\mathsf{S}_h}(\mu^N\|\Pi) \;\le\; \frac{1}{2N(\sigma_0^2+h^2)}\sum_{i=1}^N \|\theta_i-\theta_0\|^2 \;+\; \frac{d}{2}\,\phi\Big(\tfrac{h^2}{\sigma_0^2+h^2}\Big),$$

with $\phi(\rho) = \rho - 1 - \ln\rho$.

Proof. Apply a standard PAC-Bayes bound for bounded losses (e.g., *empirical Bernstein*/McAllester-style) to the *smoothed* posterior $\mathsf{S}_{h,\#}\mu^N$ and prior $\mathsf{S}_{h,\#}\Pi$:

$$\mathrm{Risk}\mathsf{S}_{h,\#}\mu^N{}_{\mathcal{D}} \;\le\; \mathrm{Risk}\mathsf{S}_{h,\#}\mu^N{}_{S} \;+\; \sqrt{\frac{D_{\mathrm{KL}}\big(\mathsf{S}_{h,\#}\mu^N\|\mathsf{S}_{h,\#}\Pi\big) + \ln(2m/\delta)}{2m}}.$$

Lipschitzness and Gaussian smoothing yield the bias control $\mathrm{Risk}\mu^N{}_{\mathcal{D}} \le \mathrm{Risk}\mathsf{S}_{h,\#}\mu^N{}_{\mathcal{D}} + Lh\,\mathbb{E}\|Z\|$ and $\mathrm{Risk}\mathsf{S}_{h,\#}\mu^N{}_{S} \le \mathrm{Risk}\mu^N{}_{S} + Lh\,\mathbb{E}\|Z\|$, whence

$$\mathrm{Risk}\mu^N{}_{\mathcal{D}} \;\le\; \mathrm{Risk}\mu^N{}_{S} \;+\; Lh\,\mathbb{E}\|Z\| \;+\; \sqrt{\frac{D_{\mathrm{KL}\mathsf{S}_h}(\mu^N\|\Pi) + \ln(2m/\delta)}{2m}},$$

using $D_{\mathrm{KL}\mathsf{S}_h}(\mu^N\|\Pi) = D_{\mathrm{KL}}(\mathsf{S}_{h,\#}\mu^N\|\mathsf{S}_{h,\#}\Pi)$ (definition) and $\mathbb{E}\|Z\| \le \sqrt{d}$. For the Gaussian-prior specialization, compute the KL between Gaussians:

$$D_{\mathrm{KL}}\Big(\mathcal{N}(\theta_i, h^2 I_d)\,\Big\|\,\mathcal{N}\big(\theta_0, (\sigma_0^2+h^2)I_d\big)\Big) \;=\; \frac{\|\theta_i-\theta_0\|^2}{2(\sigma_0^2+h^2)} \;+\; \frac{d}{2}\phi\Big(\tfrac{h^2}{\sigma_0^2+h^2}\Big),$$

and average over $i = 1, \ldots, N$. This proves the claim. □

Gaussian prior specialization. If $\Pi = \mathcal{N}(\theta_0, \sigma_0^2 I_d)$ and $\mu^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$, then

$$D_{\mathrm{KL}\mathsf{S}_h}(\mu^N\|\Pi) \;=\; \frac{1}{N}\sum_{i=1}^N D_{\mathrm{KL}}\Big(\mathcal{N}(\theta_i, h^2 I_d)\,\Big\|\,\mathcal{N}\big(\theta_0, (\sigma_0^2+h^2)I_d\big)\Big)$$

with

$$D_{\mathrm{KL}}\Big(\mathcal{N}(\theta_i, h^2 I_d)\,\Big\|\,\mathcal{N}\big(\theta_0, (\sigma_0^2+h^2)I_d\big)\Big) = \frac{\|\theta_i-\theta_0\|^2}{2(\sigma_0^2+h^2)} + \frac{d}{2}\,\phi\Big(\tfrac{h^2}{\sigma_0^2+h^2}\Big), \quad \phi(\rho) = \rho - 1 - \ln\rho.$$

C.3 WASSERSTEIN GRADIENT-FLOW CALCULUS USED IN THE MAIN TEXT

For completeness we state the standard correspondence used in Section 3. Consider the free-energy functional $J_{\mathrm{PAC}}(\mu) = \hat{L}_S(\mu) + \beta D_{\mathrm{KL}}(\mu\|\Pi)$ on $\mathcal{P}(\Theta)$. Its 2-Wasserstein gradient flow is given by the continuity equation

$$\partial_t \mu_t \;=\; \nabla_\theta\big(\mu_t \nabla_\theta \delta J_{\mathrm{PAC}}/\delta\mu(\theta)\big) \quad \text{with} \quad \frac{\delta J_{\mathrm{PAC}}}{\delta\mu}(\theta) = \mathbb{E}_{z\sim S}\,\ell_\theta(z) + \beta\big(\log\mu(\theta) - \log\Pi(\theta)\big) + c_t,$$

which matches the Fokker–Planck equation in Equation (3.3) and the Langevin SDE in Equation (3.4) (up to the irrelevant additive constant c_t). See Jordan et al. (1998); Ambrosio et al. (2008); Villani (2008) for full details and the JKO scheme. We omit repetition of these standard proofs to avoid redundancy.

Where to find the end-to-end regret analysis. The budget allocation across episodes/iterations and the root-time Monte Carlo accumulation lemmas used for our final regret bounds are proved once in Section E (see "Restatement of Main Theorems" therein). This avoids duplicating those results here while keeping this appendix focused on the PAC-Bayes smoothing and the WGF calculus.

D Technical Lemmas and Auxiliary Results

This section gathers technical lemmas (variance–information coupling, discretization, stochastic gradients, Monte Carlo concentration) used by Section E.

We start by recalling the PAC-Bayesian objective and its connection to the Wasserstein gradient flow, and then proceed to rigorously analyze each source of approximation error.

D.1 PAC-Bayesian Generalization and the Learning Objective

The PAC-Bayesian framework provides high-probability bounds on the generalization error of randomized predictors (McAllester, 1999; Catoni, 2007). A standard result, adapted to our setting (Alquier, 2021; Guedj, 2019), states that for any prior distribution P on Θ, any posterior Q, and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the draw of a dataset S of size m:

$$\mathbb{E}_{\theta\sim Q} \leq \mathbb{E}_{\theta\sim Q} + \sqrt{\frac{D_{\mathrm{KL}}(Q\|P) + \ln(m/\delta)}{2m}}, \tag{D.1}$$

where $L_{\mathcal{D}}$ is the true expected loss and $\hat{L}_S$ is the empirical loss. This motivates minimizing the right-hand side, which is equivalent to minimizing the regularized objective functional:

$$J_{\mathrm{PAC}}(\mu) = \hat{L}_S(\mu) + \beta\, D_{\mathrm{KL}}(\mu\|P). \tag{D.2}$$

The Langevin update step in our OLE algorithm is precisely a noisy gradient step on this functional, where μ is represented by the particle ensemble.

Theorem D.1. *The posterior distribution Π_t maintained by the idealized (continuous-time, infinite-particle) Langevin dynamics minimizes the PAC-Bayesian functional $J_{\mathrm{PAC}}(\mu)$ over the space of probability measures. The finite-ensemble, discrete-time, stochastic-gradient implementation approximates this ideal posterior, and its generalization error is controlled by the sum of the PAC-Bayesian objective and the approximation error terms.*

Proof. The proof follows from the variational characterization of the Fokker-Planck equation as the Wasserstein gradient flow of the free energy functional, which in our case is $J_{\mathrm{PAC}}(\mu)$ (Jordan et al., 1998). The practical algorithm is a numerical approximation of this flow, and its deviation from the ideal posterior is bounded by the lemmas in Section D.4. □

D.2 Eluder Dimension and the Variance-Information Bound

The key to bounding the exploration cost is the eluder dimension (Russo & Van Roy, 2013; 2014).

Definition D.2. *A sequence of context-action pairs $(x_1, y_1), \ldots, (x_k, y_k)$ is ϵ-independent for a function class $\mathcal{R}$ if for every $i \in \{1, \ldots, k\}$, there exist two functions $r_1, r_2 \in \mathcal{R}$ such that $|r_1(x_j, y_j) - r_2(x_j, y_j)| \leq \epsilon$ for all $j < i$, but $|r_1(x_i, y_i) - r_2(x_i, y_i)| > \epsilon$. The ϵ-eluder dimension, $d_{\mathrm{eluder}}(\mathcal{R}, \epsilon)$, is the length of the longest such sequence.*

A low eluder dimension means that after a few queries, any two functions consistent with the observations must be close everywhere, enabling efficient learning. This complexity measure is connected to regret via the following lemma.

Lemma D.3. *Under the Bradley-Terry-Luce model for preferences, for any posterior distribution μ over parameters, the mutual information gained from observing a preference for a pair (y_w, y_l) in context x is bounded by the variance of the reward predictions:*

$$I\big(\theta^*; (y_w, y_l) \mid x, \mathcal{F}_{t-1}\big) \geq C\left(\mathrm{Var}_{\theta\sim\mu}\big(r_\theta(x, y_w)\big) + \mathrm{Var}_{\theta\sim\mu}\big(r_\theta(x, y_l)\big)\right),$$

for a universal constant $C > 0$.

Proof. The proof relates the mutual information to the expected KL-divergence between the conditional likelihoods $p(\cdot \mid x, \theta)$ and the marginal likelihood $p(\cdot \mid x) = \int p(\cdot \mid x, \theta) d\mu(\theta)$. For the logistic link function, the KL-divergence can be lower-bounded by the squared difference of the logits, which in turn relates to the variance of the reward predictions under μ. A detailed derivation can be found in related contexts (Russo & Van Roy, 2014). □

D.3 Sharp Eluder-Dimension Control for LoRA-Based Models

A key argument for the practical relevance of our theory is that the eluder dimension for massive models is not as large as their parameter count might suggest, especially when using parameter-efficient fine-tuning methods like LoRA (Hu et al., 2022).

Proposition D.4. *Consider a reward function class $\mathcal{R}$ parameterized by a large neural network with weights $W_0 \in \mathbb{R}^{d \times d'}$. Let the fine-tuning be restricted to a LoRA update $W = W_0 + AB$, where $A \in \mathbb{R}^{d \times d_*}$, $B \in \mathbb{R}^{d_* \times d'}$, and $d_* \ll d, d'$. The trainable parameters are the entries of A and B. Under standard smoothness assumptions on the network architecture, the eluder dimension of this class scales as $d_{\text{eluder}}(\mathcal{R}, \epsilon) = \mathbf{O}(d_*(d + d') \log(1/\epsilon))$, not with the full parameter count $d \times d'$.*

Assumption D.5 (Blockwise Lipschitzness for LoRA layers)**.** *For each modified layer $\ell \in [L]$ with base weight $W_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$ and low-rank update $A_\ell B_\ell^\top$ with rank r_ℓ, we assume the reward (or preference log-likelihood) is L_ℓ-Lipschitz in each block parameter and smooth in the base activations, uniformly over the input domain. That is, for all admissible inputs, perturbations $(\Delta A_\ell, \Delta B_\ell)$ satisfy*

$$\left|\mathcal{R}\big(W_\ell + (A_\ell + \Delta A_\ell)(B_\ell + \Delta B_\ell)^\top\big) - \mathcal{R}\big(W_\ell + A_\ell B_\ell^\top\big)\right| \le L_\ell\big(\|\Delta A_\ell\|_F + \|\Delta B_\ell\|_F\big).$$

Corollary D.6 (Intrinsic dimension under blockwise Lipschitz LoRA)**.** *Under Theorem D.5, the eluder dimension of the LoRA-parameterized reward class satisfies, for any $\epsilon \in (0, 1]$,*

$$d_{\text{eluder}}(\epsilon; \mathcal{R}_{\textit{LoRA}}) \;\le\; C\left(\sum_{\ell=1}^{L} r_\ell\big(m_\ell + n_\ell - r_\ell\big)\right) \log\frac{C'}{\epsilon},$$

for universal positive constants C, C'. In particular, the effective intrinsic dimension scales with the rank budget rather than the ambient parameter count, aligning with empirical observations on parameter-efficient fine-tuning (Hu et al., 2022; Aghajanyan et al., 2020).

Proof. **Step 1 (Model class and parameterization).** Let $\mathcal{R}$ denote the LoRA-parameterized reward class obtained by freezing a base network and adding, in each layer $\ell \in [L]$, a rank-r_ℓ update of the form $U_\ell V_\ell^\top$ with $U_\ell \in \mathbb{R}^{m_\ell \times r_\ell}$, $V_\ell \in \mathbb{R}^{n_\ell \times r_\ell}$. Assumption Theorem D.5 ensures *blockwise Lipschitzness*: for any two parameter tuples Θ, Θ',

$$\sup_{(x,y)} \left|r_\Theta(x, y) - r_{\Theta'}(x, y)\right| \;\le\; \sum_{\ell=1}^{L} L_\ell \left\| [U_\ell, V_\ell] - [U'_\ell, V'_\ell] \right\|_F.$$

Step 2 (Covering numbers for low-rank blocks). Fix radii R_ℓ so that $\|(U_\ell, V_\ell)\|_F \le R_\ell$ for all admissible parameters (w.l.o.g. finite by compactness assumptions). For each block ℓ, the parameter set lives on a smooth manifold of dimension $d_\ell = r_\ell(m_\ell + n_\ell - r_\ell)$. Standard volumetric bounds give an ϵ_ℓ-net of size at most $(CR_\ell/\epsilon_\ell)^{d_\ell}$ in Frobenius norm. By the blockwise Lipschitzness, an (ϵ_ℓ/L_ℓ)-cover in parameters induces an ϵ_ℓ-cover in function sup-norm. Taking the product over blocks and distributing a total accuracy ϵ across blocks (e.g., $\epsilon_\ell = \epsilon/L$) yields the *function-class* covering bound

$$\mathcal{N}(\epsilon, \mathcal{R}, \|\cdot\|_\infty) \;\le\; \prod_{\ell=1}^{L} \left(\frac{C_\ell}{\epsilon}\right)^{d_\ell} \;=\; \left(\frac{C}{\epsilon}\right)^{\sum_{\ell=1}^{L} d_\ell}, \tag{D.3}$$

for constants C_ℓ depending on (L_ℓ, R_ℓ) and a universal $C = \prod_\ell C_\ell$. (See, e.g., standard covering-number bounds for low-rank matrix manifolds.)

Step 3 (From covering numbers to eluder dimension). By the growth-function argument of Russo & Van Roy (2013; 2014) (see also Lemma Theorem D.3), for any $\epsilon \in (0, 1]$ there exists a universal $C' > 0$ such that

$$d_{\text{eluder}}(\mathcal{R}, \epsilon) \;\le\; C' \sup_{\delta \in [\epsilon, 1]} \log \mathcal{N}(\delta, \mathcal{R}, \|\cdot\|_\infty). \tag{D.4}$$

Combining equation D.3 and equation D.4 gives

$$d_{\text{eluder}}(\mathcal{R}, \epsilon) \;\le\; C'\Big(\sum_{\ell=1}^{L} d_\ell\Big)\,\log\frac{C}{\epsilon} \;=\; C'\Big(\sum_{\ell=1}^{L} r_\ell\big(m_\ell + n_\ell - r_\ell\big)\Big)\,\log\frac{C}{\epsilon},$$

which is precisely the claimed bound (absorbing constants into C, C').

Step 4 (Interpretation). The dependence is *intrinsic*: it scales with the low-rank degrees of freedom and is independent of the ambient widths except through the block dimensions (m_ℓ, n_ℓ) and Lipschitz constants L_ℓ. This matches the intuition that parameter-efficient fine-tuning reduces the exploration burden. □

Proof. The proof follows from the observation that the reward function $r_{A,B}(x, y)$ is a smooth function of the low-rank matrices A and B. The effective number of parameters is $d_*(d + d')$. Applying standard covering number arguments for Lipschitz function classes to this lower-dimensional parameter space yields the stated bound on the eluder dimension. This result formalizes the intuition that the intrinsic dimensionality of the fine-tuning task is what governs the exploration complexity (Aghajanyan et al., 2020; Li et al., 2022). □

D.4 Detailed Approximation Error Proofs

Here we provide the detailed proofs for the lemmas that quantify the three sources of algorithmic approximation error.

Lemma D.7. *Assume the drift of the mean-field Langevin SDE is L-Lipschitz. Let θ_t be the continuous-time process and θ_t^η be its Euler-Maruyama discretization with step size η. The cumulative weak error in estimating the expected reward, $\sum_{t=1}^{T} |\mathbb{E}[r_{\theta_t}] - \mathbb{E}[r_{\theta_t^\eta}]|$, is bounded by $\mathbf{O}(\eta T)$.*

Proof. Let $\mathcal{R}$ be the reward class realized by a base network with weight matrices $\{W_\ell\}_{\ell=1}^{L}$ and LoRA updates $W_\ell \mapsto W_\ell + U_\ell V_\ell^\top$ with $U_\ell \in \mathbb{R}^{m_\ell \times r_\ell}$, $V_\ell \in \mathbb{R}^{n_\ell \times r_\ell}$. Assume the network is L_ℓ-Lipschitz w.r.t. the Frobenius norm on each block: for any parameter tuples Θ, Θ',

$$\sup_{(x,y)} |r_\Theta(x, y) - r_{\Theta'}(x, y)| \le \sum_{\ell=1}^{L} L_\ell \|[U_\ell, V_\ell] - [U'_\ell, V'_\ell]\|_F.$$

(i) Covering numbers. Restrict parameters to $\|U_\ell\|_F \le R_\ell$, $\|V_\ell\|_F \le R_\ell$; by compactness this is w.l.o.g. for learning. The admissible parameter set for block ℓ lies on a smooth manifold of dimension $d_\ell = r_\ell(m_\ell + n_\ell - r_\ell)$. Volumetric bounds yield an ϵ_ℓ-net of size at most $(CR_\ell/\epsilon_\ell)^{d_\ell}$ in Frobenius norm. By blockwise Lipschitzness, the induced function class has sup-norm cover of size at most $(C'/\epsilon)^{d_\ell}$ per block when we allocate $\epsilon_\ell = \epsilon/L$. Taking the product over blocks gives

$$\mathcal{N}_\infty(\epsilon, \mathcal{R}) \;\le\; \Big(\frac{C''}{\epsilon}\Big)^{\sum_{\ell=1}^{L} d_\ell}.$$

(ii) From covers to eluder dimension. Let $d = \sum_\ell d_\ell$. The growth function bound of Russo & Van Roy (2013; 2014) implies that if $\log \mathcal{N}_\infty(\epsilon, \mathcal{R}) \le d \log(C''/\epsilon)$, then for $\epsilon \in (0, 1)$,

$$d_{\text{eluder}}(\mathcal{R}, \epsilon) \;\le\; C_1\, d\, \log\Big(\frac{C_2}{\epsilon}\Big).$$

(iii) Conclusion. Combining (i) and (ii) yields

$$d_{\text{eluder}}(\mathcal{R}, \epsilon) \;\le\; C\Big(\sum_{\ell=1}^{L} r_\ell(m_\ell + n_\ell - r_\ell)\Big)\,\log\Big(\frac{C'}{\epsilon}\Big),$$

which is the claimed bound. □

Lemma D.8 (Finite-Particle Approximation Error)**.** *Let μ_t be the mean-field law and μ_t^N be the empirical measure of N particles. For any L-Lipschitz function ϕ, the error in estimating its expectation is bounded in probability: $|\int \phi d\mu_t^N - \int \phi d\mu_t| = \mathbf{O}(1/\sqrt{N})$ (Fournier & Guillin, 2015).*

Proof. This follows from classical results on the convergence rate of the empirical measure in Wasserstein distance and the duality between Wasserstein distance and expectations of Lipschitz functions. The error from approximating the interaction term in the SGLD update accumulates, leading to the term in the final regret bound. □

Lemma D.9. *Let $\hat{g}_t(\theta)$ be an unbiased mini-batch gradient estimator of the true gradient $g_t(\theta)$ with conditional variance $\mathrm{Var}(\hat{g}_t - g_t \mid \mathcal{F}_{t-1}) \le \sigma_t^2/B_t$. The cumulative error from the noise sequence $\xi_t = \eta_t(\hat{g}_t - g_t)$ is bounded with high probability by $\mathbf{O}(\sqrt{\sum_{t=1}^T \eta_t^2 \sigma_t^2/B_t})$.*

Proof. Let $\widehat{g}_t(\theta)$ be an unbiased mini-batch estimator of the population gradient $g_t(\theta)$ with $\mathbb{E}[\widehat{g}_t(\theta) \mid \mathcal{F}_{t-1}] = g_t(\theta)$ and conditional covariance $\mathbb{E}\big[\|\widehat{g}_t(\theta) - g_t(\theta)\|^2 \mid \mathcal{F}_{t-1}\big] \le \sigma_t^2/B_t$. Consider the parameter update $\theta_{t+1} = \theta_t - \eta_t \widehat{g}_t(\theta_t) +$ (other terms) and track the noise contribution to the PAC objective $J(\theta)$ through the descent lemma. Define the noise martingale $\zeta_t := \langle \nabla J(\theta_t), \widehat{g}_t(\theta_t) - g_t(\theta_t)\rangle$ with $\mathbb{E}[\zeta_t \mid \mathcal{F}_{t-1}] = 0$. Then

$$\sum_{t=1}^T \eta_t \, \zeta_t$$

is a martingale with predictable quadratic variation bounded by

$$\sum_{t=1}^T \eta_t^2 \, \mathbb{E}[\zeta_t^2 \mid \mathcal{F}_{t-1}] \;\le\; \sum_{t=1}^T \eta_t^2 \, \|\nabla J(\theta_t)\|^2 \, \frac{\sigma_t^2}{B_t} \;\le\; G^2 \sum_{t=1}^T \eta_t^2 \frac{\sigma_t^2}{B_t},$$

where G bounds $\|\nabla J(\theta)\|$ on the iterates (ensured by standard coercivity/compacity arguments in our setting). Applying Freedman's inequality (or Azuma–Hoeffding with conditional variances) yields, with probability at least $1-\delta$,

$$\Big|\sum_{t=1}^T \eta_t \, \zeta_t\Big| \;\le\; c_1 \, G \sqrt{\log \tfrac{2}{\delta}} \, \sqrt{\sum_{t=1}^T \eta_t^2 \frac{\sigma_t^2}{B_t}} \;+\; c_2 \, G \, \log \tfrac{2}{\delta} \, \max_t \eta_t \frac{\sigma_t}{\sqrt{B_t}},$$

establishing the stated $\mathbf{O}\big(\sqrt{\sum_t \eta_t^2 \sigma_t^2/B_t}\big)$ high-probability control on the cumulative stochastic-gradient error. □

Lemma D.10 (Finite-Ensemble Monte Carlo Error)**.** *Let the Monte Carlo error in estimating the optimistic index be $\xi_t = \hat{I}_t - I_t$, with $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathrm{Var}(\xi_t \mid \mathcal{F}_{t-1}) \le v_t^2/N_t$. The cumulative error $\sum_{t=1}^T \xi_t$ is bounded with high probability by $\mathbf{O}(\sqrt{\sum_{t=1}^T v_t^2/N_t})$.*

Proofs of Theorem D.9 and Theorem D.10. Both proofs rely on the same core argument. The error sequences $\{\xi_t\}$ in both cases are martingale difference sequences with respect to the filtration $\mathcal{F}_{t-1}$. We can therefore apply a concentration inequality for martingales. Freedman's inequality is particularly well-suited as it handles predictable, time-varying variance bounds (Freedman, 1975). Let $S_T = \sum_{t=1}^T \xi_t$. Let $V_T = \sum_{t=1}^T \mathbb{E}[\xi_t^2 \mid \mathcal{F}_{t-1}]$ be the predictable quadratic variation. Freedman's inequality states that for any $u, v > 0$:

$$\Pr(S_T \ge u \text{ and } V_T \le v) \le \exp\left(-\frac{u^2/2}{v + cu/3}\right)$$

where c is a uniform bound on $|\xi_t|$. Setting v to be the sum of our variance bounds (e.g., $v = \sum v_t^2/N_t$) and solving for u for a given probability level δ yields the stated $\mathbf{O}(\sqrt{\cdot})$ bounds. □

E Main Theorems: Full Statements and Proofs

This section contains the *full* proofs of the main results; it relies on auxiliary tools in Sections C and D.

E.1 Restatement of Main Theorems

Let Assumptions 2.1, 2.2, and 2.3 hold. For any $\delta \in (0,1)$, consider the OLE algorithm run for T rounds with step sizes $\{\eta_t\}$, ensemble sizes $\{N_t\}$, mini-batch sizes $\{B_t\}$, and an optimism schedule $\kappa_t = C_0\sqrt{\log(T/\delta)}$ for a suitable constant C_0. Let v_t^2 be an upper bound on the conditional variance of the Monte Carlo estimate of the optimistic value, and let σ_t^2 be an upper bound on the conditional variance of the mini-batch gradient estimator. Then with probability at least $1-\delta$, the cumulative regret satisfies:

$$\mathrm{Regret}(T) \le \underbrace{C_1\, d_{\mathrm{eluder}} \log T}_{\text{Exploration Cost}} + \underbrace{C_2 \sum_{t=1}^{T} \eta_t}_{\text{Discretization}} + \underbrace{\mathbf{O}\left(\sqrt{\sum_{t=1}^{T} \frac{v_t^2}{N_t}}\right)}_{\text{Finite Ensemble}} + \underbrace{\mathbf{O}\left(\sqrt{\sum_{t=1}^{T} \frac{\sigma_t^2}{B_t}}\right)}_{\text{Stochastic Gradient}}, \tag{E.1}$$

where C_1 and C_2 are absolute constants depending on model parameters like the Lipschitz constant L. The eluder dimension d_{eluder} is evaluated at a precision scale that decreases with t.

E.2 Proof of the Unified Regret Bound (Section E.1)

The proof proceeds by decomposing the instantaneous regret at each round and then bounding the sum of each component over the horizon T.

Step 1: Instantaneous Regret Decomposition. The regret at a single round t is $r^*(x_t, y_t^*) - r^*(x_t, y_t)$, where y_t is the action chosen by the policy induced by the optimistic index I_t. Let $\hat{r}_t(x,y)$ be the ensemble mean and $\widehat{\mathrm{Var}}_t(x,y)$ be the ensemble variance. The optimistic index is $I_t(x,y) = \hat{r}_t(x,y) + \kappa_t\sqrt{\widehat{\mathrm{Var}}_t(x,y)}$. The chosen action y_t is one part of a pair selected to maximize information gain, which implies it is a point of high optimistic value. For simplicity of analysis, we consider the regret of selecting $y_t = \arg\max_y I_t(x_t, y)$. The regret can be decomposed as:

$$r^*(x_t, y_t^*) - r^*(x_t, y_t) = \underbrace{\left(I_t(x_t, y_t^*) - r^*(x_t, y_t)\right)}_{\text{(I)}} - \underbrace{\left(I_t(x_t, y_t^*) - r^*(x_t, y_t^*)\right)}_{\text{(II)}} \tag{E.2}$$

$$\le \underbrace{\left(I_t(x_t, y_t) - \hat{r}_t(x_t, y_t)\right)}_{\text{Optimism Term}} + \underbrace{\left(\hat{r}_t(x_t, y_t) - r^*(x_t, y_t)\right)}_{\text{Estimation Error}} - \text{(II)} \tag{E.3}$$

Term (I) is upper-bounded by $I_t(x_t, y_t)$ since $y_t^* \in \mathcal{Y}$. With a properly chosen optimism coefficient κ_t, the optimistic index serves as a high-probability upper confidence bound on the true reward. Let $\mathcal{E}_t$ be the event that $r^*(x,y) \le I_t(x,y)$ for all (x,y). On this event, Term (II) is non-negative. The total error comes from the sum of optimism terms and the rounds where $\mathcal{E}_t$ fails.

Step 2: Bounding the Sum of Optimism Gaps. This is the core of the exploration analysis. The sum of the optimism terms is $\sum_{t=1}^{T} \kappa_t\sqrt{\widehat{\mathrm{Var}}_t(x_t, y_t)}$. By Cauchy-Schwarz, this is bounded by $\sqrt{T\sum_{t=1}^{T} \kappa_t^2 \widehat{\mathrm{Var}}_t(x_t, y_t)}$. A more refined analysis, following the potential function method of optimistic algorithms (Russo & Van Roy, 2014; Hazan et al., 2007), bounds the sum of variances directly. A key result, adapted to our setting in Theorem D.3, is the variance-information duality, which states that high variance implies high information gain:

$$I(\theta^*; \text{feedback}_t \mid \mathcal{F}_{t-1}) \ge C \cdot \widehat{\mathrm{Var}}_t(x_t, y_t)$$

for some constant C. Summing over t and using the chain rule for mutual information, we get:

$$\sum_{t=1}^{T} \widehat{\mathrm{Var}}_t(x_t, y_t) \le C^{-1} \sum_{t=1}^{T} I(\theta^*; \text{feedback}_t \mid \mathcal{F}_{t-1}) = C^{-1} I(\theta^*; \text{feedback}_{1:T})$$

The total information gain from T observations about a parameter in a class with eluder dimension d_{eluder} is bounded by $\mathbf{O}(d_{\mathrm{eluder}} \log T)$ (Russo & Van Roy, 2013). This implies that $\sum_{t=1}^{T} \widehat{\mathrm{Var}}_t(x_t, y_t) = \mathbf{O}(d_{\mathrm{eluder}} \log T)$. This directly bounds the cumulative optimism, leading to the leading term in our regret bound.

Step 3: Bounding the Cumulative Approximation Errors. The remaining terms in the regret come from the estimation error, which is influenced by the three practical gaps. We decompose the total estimation error into components corresponding to each gap and bound their sum.

$$\sum_{t=1}^{T}(\hat{r}_t - r^*) = \sum_{t=1}^{T}(\hat{r}_t - \bar{r}_t) + \sum_{t=1}^{T}(\bar{r}_t - r^*)$$

where $\bar{r}_t$ is the mean prediction of the ideal, continuous-time, infinite-particle posterior.

- The term $\sum(\bar{r}_t - r^*)$ is a martingale difference sequence whose cumulative sum is controlled by PAC-Bayesian generalization bounds, which are implicitly handled by the information-theoretic argument above.
- The term $\sum(\hat{r}_t - \bar{r}_t)$ captures the approximation errors. We can decompose this further into errors from discretization, finite particles, and stochastic gradients.

The cumulative effect of these errors on the regret is controlled by the following lemmas, which are proven in Appendix D.4:

- **Discretization Error:** The bias from using a finite step size η_t accumulates. Theorem D.7 shows this contributes a term of order $\mathbf{O}(\sum_{t=1}^{T}\eta_t)$ to the regret.
- **Finite-Ensemble and Stochastic Gradient Error:** The errors from using a finite ensemble and mini-batch gradients form martingale difference sequences. We apply Freedman's inequality for martingales (Theorem D.10 and Theorem D.9, based on (Freedman, 1975)) to bound their cumulative sum. This yields the high-probability bounds of $\mathbf{O}(\sqrt{\sum v_t^2/N_t})$ and $\mathbf{O}(\sqrt{\sum \sigma_t^2/B_t})$, respectively.

Step 4: Combine. Combining the bounds on the exploration term and the three approximation error terms via a union bound over all high-probability events yields the final unified regret bound as stated in the theorem. The structure of the bound reveals a fundamental decoupling: the exploration cost is a statistical quantity determined by the problem's intrinsic complexity (d_{eluder}), while the other terms are algorithmic costs determined by the allocation of computational resources (N_t, B_t, η_t). This provides a clear path for practitioners: the logarithmic term is the best one can hope for, while the other terms can be systematically reduced by investing more computation. □

F Full Proofs for the MDP Extension

This section extends our analysis to the more general setting of Markov Decision Processes (MDPs), demonstrating the robustness of our framework. We provide full proofs for both finite-horizon and discounted MDPs.

F.1 MDP Regret Bounds

Under Assumptions 2.1-2.3 and 6.1, the O-TDLE algorithm, run for T episodes, achieves a cumulative regret that satisfies, with high probability:

$$\text{Regret}(T) = \mathbf{O}\left(H^2 \cdot d_{\text{eluder}} \cdot \log T\right) + \text{lower-order approximation terms}, \tag{F.1}$$

where the lower-order terms have a similar structure to the bandit case, summed over all $T \times H$ steps.

Under the same assumptions, for an infinite-horizon discounted MDP, the O-DQLE algorithm run for T steps achieves a cumulative regret that satisfies, with high probability:

$$\text{Regret}(T) = \mathbf{O}\left(\frac{d_{\text{eluder}}}{(1-\gamma)^3} \cdot \log T\right) + \text{lower-order approximation terms}. \tag{F.2}$$

F.2 Proof for Finite-Horizon MDPs (Section F.1)

The proof requires adapting the regret decomposition to handle temporal dependencies. A naive application of the value-difference lemma can lead to errors compounding exponentially in the horizon H. To avoid this, we employ a more sophisticated policy decomposition technique.

Step 1: Regret Decomposition via Policy Hybrids. Let π_e be the (non-stationary) policy learned by the algorithm in episode e, and let π^* be the optimal policy. The regret in episode e is $V_1^{\pi^*}(s_1) - V_1^{\pi_e}(s_1)$. We introduce a sequence of H hybrid policies $\{\pi^{(h)}\}_{h=1}^{H}$. Policy $\pi^{(h)}$ follows the learned policy π_e for the first $h-1$ steps, and then switches to the optimal policy π^* from step h onwards. The total regret can be written as a telescoping sum:

$$V_1^{\pi^*}(s_1) - V_1^{\pi_e}(s_1) = \sum_{h=1}^{H} \left(V_1^{\pi^{(h)}}(s_1) - V_1^{\pi^{(h+1)}}(s_1) \right)$$

where $\pi^{(1)} = \pi^*$ and $\pi^{(H+1)} = \pi_e$. The key insight is that the difference $V_1^{\pi^{(h)}} - V_1^{\pi^{(h+1)}}$ depends only on the deviation of π_e from π^* at step h. This effectively reduces the multi-step credit assignment problem to a sequence of H single-step analyses.

Step 2: Bounding the Single-Step Deviations. For each term in the sum, we have:

$$V_1^{\pi^{(h)}} - V_1^{\pi^{(h+1)}} = \mathbb{E}_{s_h \sim d_h^{\pi_e}} \left[Q_h^{\pi^*}(s_h, \pi_h^*(s_h)) - Q_h^{\pi^*}(s_h, \pi_{e,h}(s_h)) \right]$$

where $d_h^{\pi_e}$ is the state distribution at step h under policy π_e. This is now a bandit-like regret term, where the "reward" is the optimal Q-function $Q_h^{\pi^*}$. Under the Bellman completeness assumption (Theorem 6.1), our optimistic Q-value estimates $I_h(s,a)$ serve as high-probability upper bounds on $Q_h^{\pi^*}(s,a)$. We can therefore apply the same eluder-dimension-based argument from the bandit setting to bound the sum of these single-step deviations over all episodes. The sum of variances is bounded by $\mathbf{O}(H \cdot d_{\text{eluder}} \log T)$, and the regret picks up an additional factor of H from the sum over the hybrid policies, leading to the H^2 dependence.

Step 3: Bounding Approximation Errors. The approximation errors from discretization, finite ensembles, and stochastic gradients are summed over all $T \times H$ steps. The martingale concentration arguments still apply, leading to lower-order terms of the form $\mathbf{O}(\sqrt{TH(\cdot)})$. With appropriate scheduling of N_e and B_e, these can be controlled. □

F.3 Proof for Discounted MDPs (Section F.1)

The proof for the discounted case follows a similar structure, but the regret decomposition is adapted. We use the performance difference lemma for discounted MDPs:

$$V^{\pi^*} - V^{\pi} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}} \left[Q^{\pi^*}(s, \pi^*(s)) - Q^{\pi^*}(s, \pi(s)) \right]$$

where d^{π} is the stationary state distribution under π. The analysis then proceeds by bounding the advantage of the optimal policy at each step. The optimism argument again bounds the sum of variances by $\mathbf{O}(d_{\text{eluder}} \log T)$. The factors of $(1-\gamma)^{-1}$ arise from the discounted sums and the stationary distribution, leading to the final regret bound. The dependence is $(1-\gamma)^{-3}$ due to one factor from the value difference, one from the concentration of the stationary distribution, and one from the effective horizon in the variance sum. □

G Implementation Details and Additional Pseudocode

This section provides the necessary details of pseudocode for the proposed algorithms and a discussion of hyperparameter schedules that achieve the optimal regret rates.

G.1 Complete Pseudocode

The following algorithms formalize the procedures analyzed in this paper. Algorithm 2 provides the generic template, Algorithm 4 and Algorithm 3 specifies the contextual bandit variant online contextual bandit variant respectively, and Algorithm 5 details the extension to MDPs using temporal-difference learning.

Algorithm 2: Optimistic Langevin Ensemble (OLE): Generic Template

Input: Prior Π_0; step sizes $\{\eta_t\}$; ensemble sizes $\{N_t\}$; batch sizes $\{B_t\}$; optimism schedule $\{\kappa_t\}$

for $t = 1, 2, \ldots, T$ **do**
- Observe context x_t;
- `// Optimistic Selection`
- Compute ensemble mean $\hat{r}_t(x_t, y)$ and variance $\widehat{\mathrm{Var}}_t(x_t, y)$ for all $y \in \mathcal{Y}$;
- Construct optimistic index: $I_t(x_t, y) \leftarrow \hat{r}_t(x_t, y) + \kappa_t \sqrt{\widehat{\mathrm{Var}}_t(x_t, y)}$;
- Select action pair $(y_t^{(w)}, y_t^{(\ell)})$ based on maximizing information gain using $\{I_t(x_t, y)\}_{y \in \mathcal{Y}}$;
- Receive preference feedback, forming data batch $\mathcal{D}_t$;
- `// Posterior Update (SGLD)`
- Compute mini-batch gradient $\widehat{\nabla}_t$ of $J_{\mathrm{PAC}}(\theta) = \hat{L}_{\mathcal{D}_t}(\theta) + \beta D_{\mathrm{KL}}(\delta_\theta \| \Pi_{t-1})$;
- **for** $i = 1, \ldots, N_t$ **do**
  - Draw Gaussian noise $\xi_t^{(i)} \sim \mathcal{N}(0, I)$;
  - $\theta_{t+1}^{(i)} \leftarrow \theta_t^{(i)} - \eta_t \widehat{\nabla}_t J_{\mathrm{PAC}}(\theta_t^{(i)}) + \sqrt{2\eta_t \beta}\, \xi_t^{(i)}$;

Algorithm 3: Optimistic Thompson Sampling with Langevin Ensembles (O-TSLE)

Input: Prior Π_0, step size η, particles N_t, batch size B_t, optimism schedule κ_t.

for $t = 1, 2, \ldots, T$ **do**
- Draw $\{\theta_t^{(i)}\}_{i=1}^{N_t}$ by 1 SGLD step from Π_{t-1} using B_t samples;
- Compute predictive mean $\hat{r}_t(y)$ and uncertainty $\hat{\sigma}_t(y)$ over candidates $y \in \mathcal{Y}$;
- Select action $y_t \in \arg\max_y \ \hat{r}_t(y) + \kappa_t \hat{\sigma}_t(y)$;
- Observe (pairwise) feedback at y_t and update posterior to Π_t (PAC-Bayes loss);

Algorithm 4: Optimistic Langevin Ensemble (OLE) — Contextual Bandit Variant (O-TSLE)

Input: Prior Π_0; step sizes $\{\eta_t\}$; ensemble sizes $\{N_t\}$; batch sizes $\{B_t\}$; optimism schedule $\{\kappa_t\}$

for $t = 1, 2, \ldots, T$ **do**
- Observe context x_t;
- `// Optimistic Selection`
- Compute ensemble mean and variance for all $y \in \mathcal{Y}$:
- $\hat{r}_t(x_t, y) \leftarrow \frac{1}{N_t} \sum_{i=1}^{N_t} r_{\theta_t^{(i)}}(x_t, y)$;
- $\widehat{\mathrm{Var}}_t(x_t, y) \leftarrow \frac{1}{N_t - 1} \sum_{i=1}^{N_t} (r_{\theta_t^{(i)}}(x_t, y) - \hat{r}_t(x_t, y))^2$;
- Construct optimistic index: $I_t(x_t, y) \leftarrow \hat{r}_t(x_t, y) + \kappa_t \sqrt{\widehat{\mathrm{Var}}_t(x_t, y)}$;
- Select action pair $(y_t^{(w)}, y_t^{(\ell)})$ to query, based on maximizing information gain using $\{I_t(x_t, y)\}_{y \in \mathcal{Y}}$;
- Receive preference feedback for the selected pair, forming data batch $\mathcal{D}_t$;
- `// Posterior Update`
- Compute mini-batch gradient $\widehat{\nabla}_t$ of J_{PAC} using $\mathcal{D}_t$ (batch size B_t);
- **for** $i = 1, \ldots, N_t$ **do**
  - Draw Gaussian noise $\xi_t^{(i)} \sim \mathcal{N}(0, I)$;
  - **Langevin step:** $\theta_{t+1}^{(i)} \leftarrow \theta_t^{(i)} - \eta_t \widehat{\nabla}_t J_{\mathrm{PAC}}(\theta_t^{(i)}) + \sqrt{2\eta_t \beta}\, \xi_t^{(i)}$

Algorithm 5: Optimistic TD with Langevin Ensembles (O-TDLE) for MDPs

Input: Prior Π_0 on Q-function parameters; step sizes $\{\eta_e\}$; ensemble sizes $\{N_e\}$; batch sizes $\{B_e\}$; optimism schedule $\{\kappa_h\}$

for *episode* $e = 1, 2, \ldots, T$ **do**
  Initialize state s_1;
  for *step* $h = 1, 2, \ldots, H$ **do**
    `// Optimistic Action Selection`
    Compute ensemble mean $\hat{Q}_{e,h}(s_h, a)$ and variance $\widehat{\mathrm{Var}}_{e,h}(s_h, a)$ for all $a \in \mathcal{A}$;
    Select action $a_h = \arg\max_{a \in \mathcal{A}} \left(\hat{Q}_{e,h}(s_h, a) + \kappa_h \sqrt{\widehat{\mathrm{Var}}_{e,h}(s_h, a)} \right)$;
    Execute a_h, observe next state s_{h+1} and collect preference data for the transition;
  `// Posterior Update (after episode)`
  Form a batch of transitions and preferences $\mathcal{D}_e$ from the episode;
  Compute TD targets $y_h = r(s_h, a_h) + \gamma \max_{a'} \hat{Q}_{e,H}(s_{h+1}, a')$ (using ensemble mean);
  Compute mini-batch gradient $\widehat{\nabla}_e$ of a TD-based loss on $\mathcal{D}_e$ regularized by $D_{\mathrm{KL}}(\cdot \| \Pi_{e-1})$;
  Update all particles $\{\theta_e^{(i)}\}$ to $\{\theta_{e+1}^{(i)}\}$ using one or more SGLD steps with gradient $\widehat{\nabla}_e$;

G.2 Discussion of Hyperparameter Schedules

Corollary 5.3 states that if the algorithmic parameters are scheduled appropriately, the lower-order approximation error terms in the regret bound become asymptotically negligible, leaving a purely logarithmic regret. Here we specify schedules that achieve this.

- **Step Size (η_t):** To ensure the cumulative discretization error $\sum \eta_t$ remains bounded, a decreasing step size schedule is required. A standard choice is $\eta_t = \eta_0/t$ or $\eta_t = \eta_0/\sqrt{t}$. With such schedules, the sum converges or grows slower than any linear function, making the $\mathbf{O}(\sum \eta_t)$ term sub-leading.
- **Ensemble Size (N_t) and Batch Size (B_t):** To control the finite-ensemble and stochastic gradient errors, whose cumulative sums scale as $\mathbf{O}(\sqrt{\sum 1/N_t})$ and $\mathbf{O}(\sqrt{\sum 1/B_t})$ respectively (assuming bounded variances), we need the sums $\sum 1/N_t$ and $\sum 1/B_t$ to be bounded. This can be achieved by increasing N_t and B_t over time. For example, setting $N_t = \lceil N_0 \log(t+1) \rceil$ and $B_t = \lceil B_0 \log(t+1) \rceil$ would suffice. A practical alternative is an episodic schedule where N_t and B_t are increased (e.g., doubled) at the start of geometrically spaced episodes. This ensures the approximation errors are effectively "paid for" by the logarithmic exploration term.

These schedules demonstrate that our theory provides an asymptotic guarantee, and offers concrete, practical guidance for algorithm design, directly connecting the theoretical results to the desired performance outcome.