Generalisation of RLHF under Reward Shift and Clipped KL Regularisation

Kenton Tang[1] **Yuzhu Chen**[2] **Fengxiang He**[1]

[1]University of Edinburgh
[2]University of Science and Technology of China

Abstract

Alignment and adaptation in large language models heavily rely on reinforcement learning from human feedback (RLHF); yet, theoretical understanding of its generalisability remains premature, especially when the learned reward could shift, and the KL control is estimated and clipped. To address this issue, we develop generalisation theory for RLHF that explicitly accounts for (1) *reward shift*: reward models are trained on preference data from earlier or mixed behaviour policies while RLHF optimises the current policy on its own rollouts; and (2) *clipped KL regularisation*: the KL regulariser is estimated from sampled log-probability ratios and then clipped for stabilisation, resulting in an error to RLHF. We present generalisation bounds for RLHF, suggesting that the generalisation error stems from a sampling error from prompts and rollouts, a reward shift error, and a KL clipping error. We also discuss special cases of (1) initialising RLHF parameters with a uniform prior over a finite space, and (2) training RLHF by stochastic gradient descent, as an Ornstein-Uhlenbeck process. The theory yields practical implications in (1) optimal KL clipping threshold, and (2) budget allocation in prompts, rollouts, and preference data.

1 INTRODUCTION

Reinforcement learning from human feedback (RLHF) has become a central method for steering large language models (LLMs) towards better reflecting human preferences [Christiano et al., 2017, Stiennon et al., 2020], task requirements [Ouyang et al., 2022, Chung et al., 2024], safety constraints [Bai et al., 2022a,b], amongst many others. Despite the empirical success of RLHF, the theoretical understanding of its generalisability remains largely absent.

To address this issue, this paper presents generalisation bounds for RLHF. To note, we analyse the post-trained policy in deployment, rather than studying online reinforcement learning during interactions with the environment.

A typical RLHF algorithm consists of two coupled components: (1) a reward model trained from preference data and serving as a proxy for human judgment, and (2) a policy model optimised to maximise the reward model. Most RLHF algorithms additionally employ Kullback-Leibler (KL) regularisation to keep the policy close to a reference model, typically, a supervised fine-tuned (SFT) model [Ziegler et al., 2019], for improving stability and limiting distribution shift [Schulman et al., 2015]. These induce two major challenges that significantly complicate analysis, as follows.

Reward shift. The reward model is usually trained on preference data collected from an *earlier* behaviour policy or a *mixture* of policies [Christiano et al., 2017]. However, the policy model is evaluated and optimised on rollouts drawn from the *current* distribution of responses. As the policy improves or drifts, it can move into regions where the reward model is less reliable, creating a feedback loop in which reward-model error is amplified in optimisation [Gao et al., 2023]. This calls for the potential RLHF generalisation theory to account for *reward shift* between the data used to train the reward model and the rollout distribution induced by the current policy.

Clipped KL regularisation. KL regularisation is usually assumed to be computed as a population expectation in theoretical treatments [Schulman et al., 2015]. In practice, however, the KL control is computed from sampled sequences through log probability ratios; empirical analyses have shown that the choice of estimator and implementation details can materially affect optimisation stability [Shah et al., 2025]. A common stabilisation is to clip the per-sample log ratio, in order to control rare trajectories whose likelihood ratios are extreme, echoing clipping used in PPO [Schulman et al., 2017, Lambert, 2025]. This clipped KL regularisation further introduces an error.

Motivated by these, we develop generalisation theory for RLHF that explicitly accounts for both: the reward is *learned* and *shifting*, rather than given and fixed; and the KL regularisation is *estimated* and *clipped*, rather than an exact population quantity. Based on a change-of-measure decomposition and employing PAC-Bayes tools [McAllester, 1999, Seeger, 2002, Catoni, 2007], our analysis yields high-probability generalisation bounds for the learned, data-dependent policy that decompose the generalisation error into three distinct, interpretable sources: (1) a *sampling error*, induced by the two-stage sampling of observing finitely many prompts and estimating expectations from limited Monte Carlo rollouts, (2) a *reward shift error*, capturing the gap between the learned reward and the (implicit) target reward, and the additional error induced when the policy-driven rollout distribution differs from the reward model's training distribution, (3) a *KL clipping error*, characterising the deviation from the clipped KL regulariser.

A good theory has practical implications. Our theory suggests: (1) *optimal KL clipping threshold*: the theory indicates that the KL log-ratio clipping threshold τ controls the bias-variance trade-off, since clipping reduces sampling noise while introducing an objective mismatch that does not vanish asymptotically. Our theory further provides advice on how to strike a good balance; (2) *budget allocation across prompts, rollouts, and preference labels*: our generalisation bounds separate the impacts of prompts, rollouts per prompt, and preference labels, thereby guiding budget allocation across prompts and rollouts, and preference data collection.

2 RELATED WORK

Optimisation theory of RLHF Efforts have been made in theoretically studying RLHF from an optimisation perspective. Zhu et al. [2023] analyse RLHF based on pairwise and list-wise comparisons, and characterise how the reward model error can induce suboptimal policies, motivating conservative strategies under coverage assumptions. Similarly, Zhan et al. [2024] provide finite-sample guarantees for offline RLHF that depend on a concentrability coefficient quantifying the coverage of the target policy by the offline data. Xiong et al. [2024] establish finite-sample guarantees for KL-regularised RLHF, in the offline, online, and hybrid regimes.

Reward shift The literature has seen empirical studies on the impact of reward shift. Gleave and Irving [2022] empirically study uncertainty estimation for reward models, highlighting that reward models can be unreliable out of distribution. Gao et al. [2023] empirically characterise reward model over-optimisation by measuring how the proxy-oracle gap grows when a policy is optimised against a learned proxy reward and evaluated under a stronger oracle reward. In addition, RewardBench provides a complementary evaluation resource for quantifying reward model behaviour on challenging and out-of-distribution comparisons [Lambert et al., 2025].

Clipped KL regularisation As an empirical work, Shah et al. [2025] provide an extensive analysis showing that several commonly used estimators for KL regularisation can produce biased gradients, which can affect optimisation and stability. Liu et al. [2025] analyse KL regularisation implementations in RLHF and characterise when common choices are principled or biased, including off-policy bias that arises when importance weighting is neglected.

Concurrent paper A concurrent work, released on 23 Jan 2026, provides interesting results on the generalisation of RLHF, under linear reward model assumption, through the algorithmic stability framework [Li et al., 2026]. Our work is more general, formulated for RLHF pipelines beyond linear reward; instead, the reward in this paper is learned from preference data and shifts with policy updates. Moreover, our paper allows the KL control to be estimated from sampled log ratios and clipped for stabilisation, while Li et al. [2026] formulate the KL penalty as an exact conditional KL divergence term in the objective, without sample-based KL estimation or clipping.

3 PRELIMINARIES

RLHF Given a prompt $x \in \mathcal{X}$, a policy π specifies a conditional distribution $\pi(\cdot \mid x)$ over responses $y \in \mathcal{Y}$. We denote the post-trained policy by π_θ with parameter $\theta \in \Theta$, and denote π_{ref} as a fixed reference policy. Evaluation uses prompts drawn from a distribution ρ, while preference data for reward modelling may come from a different prompt distribution ρ_{label} because of prompt shift.

Suppose the target reward is $r^\star : \mathcal{X} \times \mathcal{Y} \to [0,1]$. A reward model is a proxy $\hat{r}_\phi : \mathcal{X} \times \mathcal{Y} \to [0,1]$, with parameter $\phi \in \Phi$; the pointwise error is $e_\phi(x,y) = \hat{r}_\phi(x,y) - r^\star(x,y)$. Training the reward model uses a data collection distribution, defined as $D_{\mathrm{train}}(x,y) = \rho_{\mathrm{label}}(x)\pi_{\mathrm{ref}}(y \mid x)$, where we also use π_{ref} as the behaviour policy for reward-data collection. In practice, this policy can be a mixture, and we write $\pi_{\mathrm{ref}}(\cdot \mid x) := \sum_{m=1}^{M} c_m \pi^{(m)}(\cdot \mid x)$, reflecting the standard practice of collecting preference rankings over diverse behaviour policy mixtures. Moreover, the policy-induced distribution is defined as $D_\theta(x,y) = \rho(x)\,\pi_\theta(y \mid x)$.

The reward model is evaluated on the training distribution by the mean-squared error $L^{(2)}_{\mathrm{train}}(\phi)$, defined by

$$\mathbb{E}_{(X,Y)\sim D_{\mathrm{train}}}\left[\left(\hat{r}_\phi(X,Y) - r^\star(X,Y)\right)^2\right]. \qquad (1)$$

This quantity is an oracle-risk term defined with respect to $r^\star$, and is typically not directly observable from pairwise preference labels in practice.

Clipped KL regularisation Let $\beta > 0$ denote the regularisation strength, and $\ell_\theta(x, y) = \log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)$ denote the exact log ratio. This log ratio is the per-sample quantity that appears when the KL control is implemented from sampled rollouts, which refer to the response sequence generated by sampling sequentially from the policy conditioned on the prompt. Its conditional expectation strictly recovers the standard reference KL divergence (Definition 2, Appendix B) within the population objective. In particular, for every prompt x, we have $\mathbb{E}_{Y\sim\pi_\theta(\cdot|x)}[\ell_\theta(x, Y)] = \mathrm{KL}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x))$.

In post-training, $\ell_\theta(x, y)$ can have a large magnitude on rare samples, which can significantly increase the variance of empirical KL-related quantities and destabilise optimisation unless additional control is imposed [Shah et al., 2025, Lambert, 2025]. To stabilise KL control while keeping the target objective explicit, a popular approach is clipping with threshold $\tau > 0$: $\ell^\tau_\theta(x, y) = \mathrm{clip}(\ell_\theta(x, y), -\tau, \tau)$ [Schulman et al., 2017, Lambert, 2025]. Correspondingly, the clipped population objective $J^{r,\tau}(\theta)$ is given by

$$\mathbb{E}_{X\sim\rho}\mathbb{E}_{Y\sim\pi_\theta(\cdot|X)}[r(X, Y) - \beta\, \ell^\tau_\theta(X, Y)]\,. \tag{2}$$

Generalisation The population objective is

$$J^r(\theta) = \mathbb{E}_{X\sim\rho}\mathbb{E}_{Y\sim\pi_\theta(\cdot|X)}[r(X, Y) - \beta\, \ell_\theta(X, Y)]\,,$$

where $\ell_\theta(x, y) = \log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)$ is the exact log ratio. Evaluating a policy relies on finite prompts and rollouts. Let $x_1, \ldots, x_n$ be independent prompts drawn from ρ. For each x_i, let $y_{i,1}, \ldots, y_{i,K}$ denote K independent rollouts drawn from $\pi_\theta(\cdot \mid x_i)$. The resulting empirical objective is

$$\widehat{J}^{r,\tau}_{n,K}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}\sum_{j=1}^{K}[r(x_i, y_{i,j}) - \beta\, \ell^\tau_\theta(x_i, y_{i,j})]\,.$$

For brevity, $\widehat{J}^{\phi,\tau}_{n,K}(\theta)$ denotes $\widehat{J}^{\hat{r}_\phi,\tau}_{n,K}(\theta)$; $J^\star(\theta)$ denotes $J^{r^\star}(\theta)$; $J^\phi(\theta)$ denotes $J^{\hat{r}_\phi}(\theta)$; $J^{\phi,\tau}(\theta)$ denotes $J^{\hat{r}_\phi,\tau}(\theta)$.

The generalisability can be quantified by the generalisation error, defined to be the discrepancy between the empirical and population objectives: $\left|\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^\star(\theta)\right|$.

4 MAIN RESULTS

This section presents our theoretical results.

4.1 DECOMPOSING GENERALISATION ERROR

We decompose the generalisation error into three components: (1) a *sampling error*, induced by prompts and rollouts, which is present even if the following two terms do not exist; (2) a *reward shift error*, induced by reward shift under the same *exact* KL regulariser; and (3) a *KL clipping error*, induced by the objective mismatch induced by estimating and clipping the KL penalty.

Lemma 1 (Generalisation error decomposition)**.** *Given parameters $\theta \in \Theta$ and $\phi \in \Phi$, and clipping threshold $\tau > 0$. Then, we have the following decomposition,*

$$\begin{aligned}&\left|\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^\star(\theta)\right| \\ &\le \underbrace{\left|\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^{\phi,\tau}(\theta)\right|}_{\textit{sampling error}} + \underbrace{\left|J^\phi(\theta) - J^\star(\theta)\right|}_{\textit{reward shift error}} \\ &\quad + \underbrace{\left|J^{\phi,\tau}(\theta) - J^\phi(\theta)\right|}_{\textit{KL clipping error}}\end{aligned} \tag{3}$$

4.2 SAMPLING ERROR BOUND

We first study the sampling error $\left|\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^{\phi,\tau}(\theta)\right|$. We define $\widehat{J}^{r,\tau}_{n,\infty}(\theta)$ as the conditional expectation of $\widehat{J}^{r,\tau}_{n,K}(\theta)$, given the prompts $x_{1:n}$. Equivalently, it is the value one would obtain by averaging infinitely many rollouts per prompt while keeping the same finite set of prompts. The estimator $\widehat{J}^{\phi,\tau}_{n,K}(\theta)$ thus has a two-stage structure: (1) prompts are sampled from ρ, leading to a deviation $\left|\widehat{J}^{r,\tau}_{n,\infty}(\theta) - J^{r,\tau}(\theta)\right|$, and (2) rollouts are sampled from $\pi_\theta(\cdot \mid x)$, conditional on each prompt, inducing a deviation $\left|\widehat{J}^{r,\tau}_{n,K}(\theta) - \widehat{J}^{r,\tau}_{n,\infty}(\theta)\right|$.

We first bound the rollout sampling error as follows.

Lemma 2 (Rollout sampling error bound)**.** *Given parameter $\theta \in \Theta$, reward $r : \mathcal{X}\times\mathcal{Y} \to [0, 1]$, clipping threshold $\tau > 0$, and confidence level $\delta \in (0, 1)$, with probability at least $1 - \delta$ over rollouts, conditional on prompts $x_{1:n}$, we have*

$$\left|\widehat{J}^{r,\tau}_{n,K}(\theta) - \widehat{J}^{r,\tau}_{n,\infty}(\theta)\right| \le (1 + 2\beta\tau)\sqrt{\frac{\log(2/\delta)}{2nK}}.$$

Proof sketch Loss clipping ensures that $\ell^\tau_\theta(x, y) \in [-\tau, \tau]$ by construction, which is a standard stabilisation approach in reinforcement learning [Mnih et al., 2015, Schulman et al., 2017]. Combining that the reward function satisfies $r(x, y) \in [0, 1]$, for each per-rollout contribution, $r(x, y) - \beta\, \ell^\tau_\theta(x, y)$ lies in an interval of length $1 + 2\beta\tau$. Given the prompts $x_{1:n}$, the rollouts are independent across both i and j. Applying Hoeffding's inequality (Lemma 8) to the average over the nK rollout terms yields Lemma 2. Detailed proofs are in Appendix C.2.

Remark 1. *Lemma 2 controls the Monte Carlo error from using finitely many rollouts per prompt. The bound decays at rate $O(nK)^{-1/2}$ as the number of rollouts per prompt K increases (assuming β and τ are independent of n and K). The factor $1 + 2\beta\tau$ comes from the range of the per-rollout contribution. Clipping is the mechanism that makes this range finite without imposing any artificial uniform bound on the exact log ratio ℓ_θ.*

Lemma 3 (Prompt sampling error bound)**.** *Under the same conditions of Lemma 2, with probability at least $1-\delta$ over prompts $x_{1:n}$, we have*

$$\left|\widehat{J}^{r,\tau}_{n,\infty}(\theta) - J^{r,\tau}(\theta)\right| \le (1+2\beta\tau)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof sketch Treating $\widehat{J}^{\phi,\tau}_{n,\infty}(\theta)$ as a function of the sampled prompts only, it is an average of n independent bounded terms, each term being the conditional expectation over rollouts for a fixed prompt. Applying Hoeffding's inequality again yields Lemma 3. Detailed proofs are in Appendix C.2.

Remark 2. *Lemma 3 suggests that the prompt sampling error decays at rate $O(n^{-1/2})$, and the corresponding bound does not depend on the number of rollouts per prompt K, similarly, assuming β and τ are independent of n and K. It isolates the deviation induced purely by finite prompt sampling. Even an arbitrarily accurate estimate of each conditional expectation over rollouts cannot compensate for having too few evaluation prompts, because the population objective is defined as an expectation over ρ.*

Combining the two lemmas leads to the following lemma on the sampling error.

Lemma 4 (Sampling error bound)**.** *Under the same conditions of Lemma 2, with probability at least $1-\delta$ over both prompts and rollouts, the sampling error satisfies*

$$\begin{aligned}&\left|\widehat{J}^{r,\tau}_{n,K}(\theta) - J^{r,\tau}(\theta)\right| \\ &\quad\le (1+2\beta\tau)\left(\sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{\log(4/\delta)}{2nK}}\right).\end{aligned} \tag{4}$$

Remark 3. *In addition to the noise induced by prompts and rollouts, a penalty term carries the additional factor $2\beta\tau$ because the clipped log ratio ranges in $[-\tau,\tau]$. Consequently, increasing τ enlarges the range of each rollout penalty term, and the resulting concentration bound is looser.*

4.3 REWARD SHIFT ERROR BOUND

This subsection studies the reward shift error $\left|J^{\phi}(\theta) - J^{\star}(\theta)\right|$. To characterise the reward shift error in transfers from D_{train} to D_θ, we use a χ^2 coverage coefficient, defined below, based on χ^2 divergence (see Definition 3 in Appendix B). χ^2 coverage coefficient is standard in the literature of importance weighting and covariate shift analyses; see, e.g., Sugiyama et al. [2007], Owen [2013].

Definition 1 (χ^2 coverage coefficient)**.** *Suppose that D_θ is absolutely continuous with respect to D_{train}. The χ^2 coverage coefficient is defined to be*

$$\mathcal{C}(\theta) := \sqrt{1+\chi^2(D_\theta \| D_{\text{train}})}, \tag{5}$$

where $\chi^2(\cdot\|\cdot)$ is χ^2 divergence.

Remark 4. *Intuitively, $\mathcal{C}(\theta)$ measures how far the policy-induced distribution departs from the distribution used to train the reward model. It acts as an amplification factor when we upper bound the reward shift error.*

Because $J^{\phi}(\theta)$ and $J^{\star}(\theta)$ share the same exact KL regulariser, the KL penalty cancels in the difference; consequently, only the reward model error remains. Defining $e_\phi(x,y) = \hat{r}_\phi(x,y) - r^\star(x,y)$, we have $J^{\phi}(\theta) - J^{\star}(\theta) = \mathbb{E}_{(X,Y)\sim D_\theta}[e_\phi(X,Y)]$, so the problem is to control the reward model error under the deployment distribution D_θ using information available under the training distribution D_{train}. This step requires a coverage condition, stated below, when deriving the change-of-measure bound; it yields the same coefficient $\mathcal{C}(\theta)$ defined in eq. (5).

Assumption 1 (Absolute continuity and finite coverage)**.** *The policy-induced distribution D_θ is absolutely continuous with respect to the reward model training distribution D_{train}. Moreover, the χ^2 divergence $\chi^2(D_\theta\|D_{\text{train}})$ is finite, so the coverage coefficient $\mathcal{C}(\theta)$ in eq. (5) is finite.*

Remark 5. *Assumption 1 is the standard condition that makes a change of measure from D_θ back to D_{train} legitimate [Sugiyama et al., 2007, Shimodaira, 2000, Precup et al., 2000]. It ensures that the density ratio $\frac{\mathrm{d}D_\theta}{\mathrm{d}D_{\text{train}}}$ exists and has a finite second moment, which is required for the Cauchy-Schwarz step in Lemma 5 [Owen, 2013]. The coefficient $\mathcal{C}(\theta)$ plays the role of an amplification factor, which quantifies how strongly the reward model error can be magnified when the policy visits regions that are rare under the data used for reward modelling.*

Our theory also relies on the following mild assumption.

Assumption 2 (Bounded training error)**.** *The squared training error $L^{(2)}_{\text{train}}(\phi)$ defined in eq. (1) is bounded.*

Remark 6. *This assumption does not assert that the reward model is accurate everywhere; instead, it provides a baseline level of accuracy on the distribution where preference supervision is available. The coverage coefficient explains how the baseline can degrade under deployment.*

We then prove the reward shift bound.

Lemma 5 (Reward shift error bound)**.** *Under Assumptions 1 and 2, we have*

$$\left|J^{\phi}(\theta) - J^{\star}(\theta)\right| \le \mathcal{C}(\theta)\sqrt{L^{(2)}_{\text{train}}(\phi)}.$$

Proof sketch To relate $\left|J^{\phi}(\theta) - J^{\star}(\theta)\right|$ to the training distribution, we rewrite the expectation under D_θ as an importance-weighted expectation under D_{train}. Assumption 1 ensures that the required density ratio exists and has finite second moment. Applying the χ^2 change-of-measure

bound (Lemma 11 in Appendix B) yields a product of two square roots: the first factor under the square root is exactly $1+\chi^2(D_\theta\|D_{\mathrm{train}})$, the second moment of the density ratio under D_{train}, whose square root therefore produces $\mathcal{C}(\theta)$; and the second factor under the square root is $L^{(2)}_{\mathrm{train}}(\phi)$ by definition, the second moment of the reward model error under D_{train}. This yields Lemma 5. Detailed proofs are in Appendix C.3.

Remark 7. *Lemma 5 characterises two ingredients that play different roles: (1) the term $L^{(2)}_{\mathrm{train}}(\phi)$ measures reward model error only on the reward model training distribution D_{train}; and (2) the coefficient $\mathcal{C}(\theta)$ measures how far D_θ moves away from that training distribution. It also quantifies how much training error can be amplified when moving to deployment.*

When prompt shifts and policy shifts are qualitatively different, it is useful to further factorise the coverage coefficient. The next lemma interprets the source of shifts in practice.

Lemma 6 (Coverage factorisation)**.** *Suppose $\rho \ll \rho_{\mathrm{label}}$ and $\pi_\theta(\cdot \mid x) \ll \pi_{\mathrm{ref}}(\cdot \mid x)$, for any prompt x with $\rho_{\mathrm{label}}(x) > 0$. Define*

$$\mathcal{C}_{\mathrm{prompt}} = \left(\mathbb{E}_{X\sim\rho_{\mathrm{label}}}\left[\left(\frac{\rho(X)}{\rho_{\mathrm{label}}(X)}\right)^2\right]\right)^{1/2},$$

and define $\mathcal{C}_{\mathrm{pol}}(\theta)$ by

$$\sup_{x\in\mathcal{X}:\rho_{\mathrm{label}}(x)>0}\left(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}\left[\left(\frac{\pi_\theta(Y\mid x)}{\pi_{\mathrm{ref}}(Y\mid x)}\right)^2\right]\right)^{1/2}.$$

If both $\mathcal{C}_{\mathrm{prompt}}$ and $\mathcal{C}_{\mathrm{pol}}(\theta)$ are bounded, we have $\mathcal{C}(\theta) \le \mathcal{C}_{\mathrm{prompt}}\,\mathcal{C}_{\mathrm{pol}}(\theta)$.

Remark 8. *Lemma 6 separates mismatch in prompts from mismatch in policies. The coefficient $\mathcal{C}_{\mathrm{prompt}}$ depends only on the shift between ρ and ρ_{label}. The coefficient $\mathcal{C}_{\mathrm{pol}}(\theta)$ depends only on how far π_θ departs from π_{ref} on the support of ρ_{label}. This separation is valuable when diagnosing failures in reward modelling and post training, because the two sources of shift have different operational causes and different mitigation strategies.*

4.4 KL CLIPPING ERROR BOUND

We now bound the KL clipping error $\left|J^{\phi,\tau}(\theta) - J^{\phi}(\theta)\right|$. This is the only place where the systematic mismatch created by clipping enters the analysis. Clipping is beneficial in the sampling bounds because it makes each rollout contribution bounded. Meanwhile, clipping may bias the objective used in practice, thereby introducing a systematic mismatch between the optimised objective and the target objective.

To state this mismatch cleanly, we only require an integrability condition on the exact log ratio in deployment.

Assumption 3 (Integrability of exact log ratio)**.** *The exact log ratio is integrable in deployment, i.e., $\mathbb{E}_{(X,Y)\sim D_\theta}[|\ell_\theta(X,Y)|] < \infty$.*

Remark 9. *Assumption 3 is mild. It allows heavy tails in ℓ_θ while still ensuring that the exact objective is well defined. Under this assumption, clipping is analysed as an explicit bias-inducing modification of the penalty; the KL clipping error term in Lemma 7, $\beta\,\mathbb{E}_{(X,Y)\sim D_\theta}[|\ell_\theta(X,Y) - \ell^\tau_\theta(X,Y)|]$, is an objective mismatch term that does not vanish asymptotically as the number of evaluation prompts or rollouts increases. It is strictly weaker than assuming ℓ_θ is uniformly bounded, and it matches the intent of treating clipping as an algorithmic choice rather than as a structural property of the policy class. Similar integrability conditions are standard in analysing truncation-based stabilisation and importance sampling; see, e.g., Ionides [2008], Owen and Zhou [2000].*

We then prove the following bound on the bias induced by the surrogate.

Lemma 7 (KL clipping error bound)**.** *Under Assumption 3, we have*

$$\begin{aligned}&\left|J^{\phi,\tau}(\theta) - J^{\phi}(\theta)\right| \\ &\qquad \le \beta\,\mathbb{E}_{(X,Y)\sim D_\theta}[|\ell_\theta(X,Y) - \ell^\tau_\theta(X,Y)|]\,.\end{aligned} \tag{6}$$

Proof sketch $\left|J^{\phi,\tau}(\theta) - J^{\phi}(\theta)\right|$ is not an estimation error; it compares two population objectives under the same learned reward, with the only difference being whether the penalty uses ℓ^τ_θ or ℓ_θ. Expanding definitions shows that the reward contributions cancel and only the penalty difference remains. Taking absolute values and applying the triangle inequality yields Lemma 7. Detailed proofs are in Appendix C.3.

Remark 10. *The right-hand side of eq.* (6) *measures clipping bias directly as the expected amount of truncation under the deployment distribution. This term can remain nonzero even with infinite evaluation data, which reflects the fact that clipping is an objective mismatch rather than an estimation error. It is small when the policy rarely produces extreme log ratios under D_θ, and it can be large when the policy places substantial mass in regions where the exact log ratio has heavy tails. This is why the final bound contains a term that depends on the tail behaviour of the exact log ratio under D_θ and does not involve n or K.*

4.5 FIXED-POLICY GENERALISATION BOUND

We now combine the results on sampling error, reward shift error, and KL clipping error into a single statement for a fixed policy parameter θ, as follows.

Theorem 1 (Fixed-policy generalisation bound)**.** *Under Assumptions 1, 2, and 3, with probability at least* $1-\delta$ *over the evaluation prompts and rollouts, the following holds,*

$$\begin{aligned}&\left|\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^\star(\theta)\right| \\ &\quad\le \underbrace{(1+2\beta\tau)\left(\sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{\log(4/\delta)}{2nK}}\right)}_{\textit{sampling error}} \\ &\quad + \underbrace{\mathcal{C}(\theta)\sqrt{L^{(2)}_{\mathrm{train}}(\phi)}}_{\textit{reward shift error}} \\ &\quad + \underbrace{\beta\,\mathbb{E}_{(X,Y)\sim D_\theta}[|\ell_\theta(X,Y) - \ell^\tau_\theta(X,Y)|]}_{\textit{KL clipping error}}.\end{aligned} \tag{7}$$

4.6 DATA-DEPENDENT PAC-BAYES BOUND

The fixed-policy theorem treats θ as pre-fixed. In practice, θ is often selected after observing data. This section fixes the gap by employing PAC-Bayes theory that extends our analysis to data-dependent selection [McAllester, 1999, Seeger, 2002, Catoni, 2007]. Specifically, we provide a bound that holds simultaneously for all posteriors Q over Θ, at the cost of a complexity term that measures how far Q deviates from a prior P on Θ. Define

$$\widehat{J}^{\phi,\tau}_{n,K}(Q) = \mathbb{E}_{\theta\sim Q}\left[\widehat{J}^{\phi,\tau}_{n,K}(\theta)\right], \quad J^\star(Q) = \mathbb{E}_{\theta\sim Q}[J^\star(\theta)].$$

Then, we have the following data-dependen PAC-Bayes bound.

Theorem 2 (Data-dependent generalisation bound)**.** *Let* P *be any prior distribution on* Θ *that is independent of the evaluation prompts and rollouts. For any posterior* Q *on* Θ*, suppose Assumptions 1, 2, and 3 hold for any* θ *in the support of* Q*. Then, with probability at least* $1-\delta$ *over the evaluation prompts and rollouts, the following inequality holds simultaneously for all such posteriors* Q*,*

$$\begin{aligned}&\left|\widehat{J}^{\phi,\tau}_{n,K}(Q) - J^\star(Q)\right| \\ &\quad\le (1+2\beta\tau)\Bigg(\sqrt{\frac{\mathrm{KL}(Q\|P) + \log(8/\delta)}{2n}} \\ &\qquad\qquad + \sqrt{\frac{\mathrm{KL}(Q\|P) + \log(8/\delta)}{2nK}}\Bigg) \\ &\qquad\qquad \textit{sampling error} \\ &\quad + \underbrace{\mathbb{E}_{\theta\sim Q}[\mathcal{C}(\theta)]\,\sqrt{L^{(2)}_{\mathrm{train}}(\phi)}}_{\textit{reward shift error}} \\ &\quad + \underbrace{\beta\,\mathbb{E}_{\theta\sim Q}\left[\mathbb{E}_{(X,Y)\sim D_\theta}[|\ell_\theta(X,Y) - \ell^\tau_\theta(X,Y)|]\right]}_{\textit{KL clipping error}}.\end{aligned}$$

Remark 11. *Comparing with Theorem 1, Theorem 2 replaces the fixed* θ *with an average over* $\theta\sim Q$*. The complexity term* $\mathrm{KL}(Q\|P)$ *appears only inside the sampling error, as the price paid for making the guarantee hold uniformly over data-dependent choices of* Q*.*

5 SPECIAL CASES

The PAC-Bayes bound in Theorem 2 contains a complexity term $\mathrm{KL}(Q\|P)$, which measures how strongly the data-dependent posterior Q departs from the data-independent prior P. This subsection discusses two operational instantiations of $\mathrm{KL}(Q\|P)$ that are common in practice.

5.1 INITIALISATION BY UNIFORM PRIOR OVER FINITE CANDIDATE CLASS

Let $M \ge 2$ be an integer, $\theta^{(1)},\dots,\theta^{(M)} \in \Theta$ be a collection of candidate parameters specified independently of the evaluation sample used to construct $\widehat{J}^{\phi,\tau}_{n,K}$. Suppose $\Theta_M := \{\theta^{(1)},\dots,\theta^{(M)}\}$ and restrict both P and Q to be distributions on Θ_M. Suppose the prior is uniform on Θ_M; i.e., $P(\theta^{(m)}) = 1/M$ for any m. This non-informative prior is standard in finite model selection [Seeger, 2002].

Corollary 1 (KL bound for uniform prior over finite candidate class)**.** *Under the conditions above,* $\mathrm{KL}(Q\|P) \le \log M$*. In particular, if* Q *is the Dirac distribution supported on a data-selected checkpoint* $\theta^{(\widehat{m})}$*, we have* $\mathrm{KL}(Q\|P) = \log M$*.*

Remark 12. *Corollary 1 yields a direct interpretation of model selection in the PAC-Bayes bound. Evaluating* M *fixed checkpoints and selecting one using the evaluation sample incurs an additional sampling error cost in Theorem 2, controlled by* $\log M$ *via the quantity* $\mathrm{KL}(Q\|P)$*.*

5.2 TRAINING RLHF BY SGD AS ORNSTEIN-UHLENBECK PROCESS

Stochastic gradient descent (SGD), and its variants, are popular optimisers [Robbins and Monro, 1951]. Suppose the parameter space is $\mathbb{R}^d$ for some $d \in \mathbb{N}$. Assume the prior is Gaussian, i.e., $P = \mathcal{N}(\theta_0, \Lambda)$ for some $\theta_0 \in \mathbb{R}^d$ and some symmetric positive definite matrix $\Lambda \in \mathbb{R}^{d\times d}$.

We employ a standard local diffusion approximation for constant-step-size SGD. Near a locally stable optimum, late-stage SGD iterates follow an Ornstein-Uhlenbeck (OU) process [Uhlenbeck and Ornstein, 1930]:

$$d\theta_t = -H(\theta_t - \hat{\theta})\,dt + \sqrt{\varepsilon}\,\Sigma_g^{1/2}\,dW_t,$$

where W_t is a d-dimensional Brownian motion, $H \succ 0$ is the local Hessian at $\hat{\theta}$, and $\Sigma_g \succ 0$ is the local gradient-noise covariance. We make the following assumptions, standard for this local OU approximation; see, e.g., Mandt et al. [2017], He et al. [2019], Chen et al. [2023].

Assumption 4. *Assume the optimisation problem has a locally stable optimum $\hat{\theta} \in \mathbb{R}^d$; i.e., within a neighbourhood of $\hat{\theta}$, the objective admits a quadratic approximation with Hessian $H \succ 0$ and the gradient noise covariance is approximately constant and equal to $\Sigma_g \succ 0$. In addition, the matrices H and Σ_g commute; i.e., $H\Sigma_g = \Sigma_g H$ holds. Moreover, there exist constants $0 < m \leq M < \infty$ such that the local curvature spectrum satisfies $mI \preceq H \preceq MI$.*

Under the assumption above, the OU process admits a stationary Gaussian law $\mathcal{N}(\hat{\theta}, \Sigma)$, where $\Sigma \succ 0$ satisfies the continuous Lyapunov equation $H\Sigma + \Sigma H = \varepsilon\Sigma_g$. Accordingly, we approximate the PAC-Bayes posterior induced by late-stage SGD iterates by $Q_{\mathrm{SGD}} := \mathcal{N}(\hat{\theta}, \Sigma)$.

Corollary 2 (KL bound for SGD as Ornstein-Uhlenbeck process)**.** *In the special case above, the PAC-Bayes complexity term admits the upper bound*

$$\begin{aligned}&\mathrm{KL}(Q_{\mathrm{SGD}} \| P)\\ &\leq \frac{1}{2}\Big((\hat{\theta}-\theta_0)^\top \Lambda^{-1}(\hat{\theta}-\theta_0) + \frac{\varepsilon}{2m}\operatorname{tr}(\Lambda^{-1}\Sigma_g) - d \\ &\quad + \operatorname{logdet}(\Lambda) - \operatorname{logdet}(\Sigma_g) - d\log\Big(\frac{\varepsilon}{2M}\Big)\Big).\end{aligned} \tag{8}$$

A detailed proof is given in Appendix C.7.2.

Remark 13. *Corollary 2 yields a locally valid, optimiser-explicit bound for $\mathrm{KL}(Q\|P)$ via the stationary diffusion approximation of constant-step-size SGD. This secondary specialisation of Theorem 2 imposes no additional structural assumptions on the main RLHF analysis. The diffusion perspective and the associated Ornstein-Uhlenbeck approximation are discussed in detail by Mandt et al. [2017].*

6 PRACTICAL IMPLICATIONS

The discussion below translates our theory into concrete, practical algorithm design recommendations.

6.1 OPTIMAL KL CLIPPING THRESHOLD

Lemma 4 includes a factor $1+2\beta\tau$, indicating that a smaller τ tightens the sampling deviations that arise from finite n and K. Lemma 12 in Appendix B gives the corresponding KL-specific concentration bound for the clipped log-ratio average, whose deviation also scales linearly with τ. Meanwhile, clipping changes the regularised objective and introduces a systematic mismatch that does not vanish with more evaluation samples, as formalised in Lemma 7.

Therefore, τ acts as a bias-variance trade-off hyperparameter rather than a purely stabilising tweak. Aggressive clipping reduces Monte Carlo noise but increases objective mismatch; weak clipping preserves the exact KL objective but exposes training to high-variance log-ratio estimates.

For brevity, we define

$$\begin{aligned}\alpha_{n,K,\delta} :=& \sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{\log(4/\delta)}{2nK}},\\ T_\theta(\tau) :=& \mathbb{E}_{(X,Y)\sim D_\theta}\big[(|\ell_\theta(X,Y)| - \tau)_+\big].\end{aligned}$$

Since $|\ell_\theta - \ell_\theta^\tau| = (|\ell_\theta| - \tau)_+$ under symmetric clipping, the τ-dependent part of eq. (7) is thus

$$B_\theta(\tau) := (1 + 2\beta\tau)\,\alpha_{n,K,\delta} \;+\; \beta\, T_\theta(\tau).$$

Let $\tau^\star$ be any minimiser of $\tau \mapsto B_\theta(\tau)$ over $\tau \geq 0$. We have the following corollary.

Corollary 3 (Optimal KL clipping threshold)**.** *For any parameters $\theta \in \Theta$ and $\phi \in \Phi$, regularisation coefficient $\beta > 0$, confidence level $\delta \in (0,1)$, and integers $n \geq 1$ and $K \geq 1$, if $2\alpha_{n,K,\delta} < 1$, $\tau^\star$ satisfies*

$$\begin{aligned}&\Pr_{(X,Y)\sim D_\theta}\big(|\ell_\theta(X,Y)| > \tau^\star\big) \;\leq\\ &\qquad 2\,\alpha_{n,K,\delta} \;\leq\; \Pr_{(X,Y)\sim D_\theta}\big(|\ell_\theta(X,Y)| \geq \tau^\star\big),\end{aligned}$$

if, in addition, $\Pr_{(X,Y)\sim D_\theta}(|\ell_\theta(X,Y)| = \tau^\star) = 0$, we have

$$\Pr_{(X,Y)\sim D_\theta}\big(|\ell_\theta(X,Y)| > \tau^\star\big) \;=\; 2\,\alpha_{n,K,\delta},$$

and, equivalently, $\tau^\star$ is the $(1 - 2\alpha_{n,K,\delta})$-quantile of $|\ell_\theta(X,Y)|$ under D_θ. Otherwise, if $2\alpha_{n,K,\delta} \geq 1$, we have $\tau^\star = 0$ is a minimizer of $\tau \mapsto B_\theta(\tau)$ over $\tau \geq 0$.

Detailed proofs are in Appendix C.4.

Remark 14. *Corollary 3 suggests choosing τ so that the clipping fraction $\Pr(|\ell_\theta| > \tau)$ matches the target level $2\alpha_{n,K,\delta}$. As the evaluation budget (n or K) increases, $\alpha_{n,K,\delta}$ decreases. Consequently, the target clipping fraction decreases and the recommended threshold τ increases. This quantile-based rule automatically relaxes clipping as Monte Carlo error diminishes.*

Threshold calibration Practitioners often treat the clipping threshold τ as a static hyperparameter that requires manual tuning. Corollary 3 instead yields a direct, budget-aware calibration rule. Given an evaluation batch $\{(x_i, y_{i,j})\}_{i\leq n,\, j\leq K}$, compute the log-ratio magnitudes $u_{i,j} := |\ell_\theta(x_i, y_{i,j})|$. If $2\alpha_{n,K,\delta} \geq 1$, set $\widehat{\tau} := 0$. Otherwise, set $\widehat{\tau}$ to the empirical $(1 - 2\alpha_{n,K,\delta})$-quantile of $\{u_{i,j}\}$. Algorithmically, this theory-guided rule balances the bias–variance trade-off by clipping approximately the top $2\alpha_{n,K,\delta}$ fraction of extreme log-ratios in the batch, thereby reducing reliance on heuristic hyperparameter sweeps.

Theorems 1–2 treat τ as fixed; when τ is selected from the evaluation sample (e.g., by an empirical quantile rule), the resulting procedure should be viewed as a practical calibration heuristic unless additional uniformity or sample-splitting arguments are used.

6.2 BUDGET ALLOCATION ACROSS PROMPTS, ROLLOUTS, AND PREFERENCE DATA

Given a fixed computational budget, practitioners often face an allocation trade-off among prompts, rollouts per prompt, and preference data. This subsection provides theoretically grounded guidelines for this budget distribution.

6.2.1 Uniform-cost baseline

Suppose rollouts share the same cost and the sampling budget is bounded by $nK \leq B$ for some $B > 0$. Substituting $n = B/K$ into the leading-order sampling terms of Lemma 4 reveals that the upper bound is minimised at $K^\star = 1$. Therefore, under a uniform-cost model, the range-based concentration bound strongly favours allocating budget to prompt coverage rather than additional rollouts per prompt. A detailed derivation is given in Appendix C.8.

6.2.2 Prefill and decode cost model

In LLM inference, sampling costs are typically asymmetric across prompts and rollouts [Pope et al., 2023]. Evaluating a new prompt requires a forward pass over prompt tokens to construct an attention cache, whereas additional rollouts reuse this cache, primarily incurring incremental decoding costs [Kwon et al., 2023]. We model this asymmetry by separating a prefill and a decode cost, imposing the constraint: $B \geq n\, c_{\text{prefill}} + nK\, c_{\text{decode}}$. Substituting $n = B/(c_{\text{prefill}} + Kc_{\text{decode}})$ into the dominant sampling structure isolates a one-dimensional objective in K. Treating $K \geq 1$ as a continuous variable in the leading-order proxy from Lemma 4 yields the following optimal allocation rule.

Corollary 4 (Optimal rollout allocation)**.** *The continuous proxy minimiser over $K \geq 1$ satisfies*

$$K^\star \;=\; \max\left\{1, \left(\frac{c_{\text{prefill}}}{c_{\text{decode}}}\right)^{2/3}\right\}.$$

In practice, one may take $K = \lfloor K^\star \rceil$, and then set n by the budget constraint.

Detailed proofs are in Appendix C.8.

Remark 15. *The expression for $K^\star$ depends only on the ratio $c_{\text{prefill}}/c_{\text{decode}}$ because the shared range multiplier $1 + 2\beta\tau$ does not affect the minimiser. The $2/3$ power law implies that the optimal number of rollouts per prompt grows sublinearly with $c_{\text{prefill}}/c_{\text{decode}}$.*

Variance-aware refinement The range-based sampling simplification above is conservative because it does not separate prompt-level variability from rollout-level variability. A refinement is to use a two-stage variance decomposition. Let Z denote a per-rollout contribution in the empirical objective (see the variables $Z_{i,j}$ in the proof of Lemma 2 in Appendix C.2), and define

$$\sigma^2_{\text{prompt}} := \mathrm{Var}(\mathbb{E}[Z \mid X])\,, \quad \sigma^2_{\text{rollout}} := \mathbb{E}[\mathrm{Var}(Z \mid X)]\,.$$

Corollary 5. *Under the same cost constraint $B \geq n\, c_{\text{prefill}} + nK\, c_{\text{decode}}$, optimising the resulting variance proxy yields an allocation rule of the form*

$$K^\star \;\approx\; \max\left\{1, \sqrt{\frac{c_{\text{prefill}}}{c_{\text{decode}}} \cdot \frac{\sigma^2_{\text{rollout}}}{\sigma^2_{\text{prompt}}}}\right\}.$$

A proof is given in Appendix C.8.

6.2.3 Preference data

Beyond prompts and rollouts, the reward shift error introduces an additional budget consideration. By Lemma 5, this term depends on the reward-model training error $L^{(2)}_{\text{train}}(\phi)$ and the coverage coefficient $\mathcal{C}(\theta)$. Preference data collection therefore affects the bound in two ways. Increasing relevant preference data can improve reward-model fit on the training distribution, and collecting data closer to the policy-induced distribution can reduce the mismatch captured by $\mathcal{C}(\theta)$. These observations provide guidance for preference data collection through their effect on the reward shift term, although the present analysis does not derive an explicit allocation rule in terms of the number of preference labels. This implication is most relevant when the sampling terms are no longer the dominant terms in the bound.

7 CONCLUSIONS

Alignment and adaptation in large language models (LLMs) are now driven by reinforcement learning from human feedback (RLHF), but a rigorous theory of how RLHF generalises is still underdeveloped, particularly when the reward could shift, and a KL clipping regularisation is implemented. To address this gap, we develop generalisation theory for RLHF that explicitly models two key practical effects: (1) distribution shift between the data used to train the reward model and the policy-induced distribution encountered at deployment, and (2) statistical noise introduced by empirical estimation of the clipped KL regulariser. We prove high-probability generalisation bounds that decompose the generalisation error into interpretable components, including sampling error from both prompts and rollouts, reward shift error, and KL clipping error. Our theory suggests optimal KL clipping threshold rules, quantitative budget allocation guidance on prompts and rollouts, and guidance for preference data collection through the reward shift term.

References

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint*, 2022b.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes–Monograph Series*. Institute of Mathematical Statistics, 2007. doi: 10.1214/074921707000000391.

Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. Stochastic collapse: How gradient noise attracts sgd dynamics towards simpler subnetworks. *Advances in Neural Information Processing Systems*, 36:35027–35063, 2023.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024. URL https://jmlr.org/papers/v25/23-0870.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Adam Gleave and Geoffrey Irving. Uncertainty estimation for language reward models. *arXiv preprint*, 2022.

Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in neural information processing systems*, 32, 2019.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Edward L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008. doi: 10.1198/106186008X320456.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, mar 1951. doi: 10.1214/aoms/1177729694.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace, editors, *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Nathan Lambert. Reinforcement learning from human feedback, 2025. URL https://arxiv.org/abs/2504.12501. RLHF Book; online version also available at https://rlhfbook.com/.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.96. URL https://aclanthology.org/2025.findings-naacl.96/.

Zhaochun Li, Mingyang Yi, Yue Wang, Shisheng Cui, and Yong Liu. Towards a theoretical understanding to the generalization of RLHF. *arXiv preprint*, 2026.

Kezhao Liu, Jason Klein Liu, Mingtao Chen, and Yiming Liu. Rethinking KL regularization in RLHF: From value estimation to gradient optimization. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2510.01555.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134): 1–35, 2017.

David A. McAllester. Pac-bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT)*, pages 164–170, 1999. doi: 10.1145/307400.307435.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236.

Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, Cambridge, MA, 2022. URL https://probml.github.io/pml-book/book1.html.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2203.02155.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. doi: 10.1080/01621459.2000.10473909.

Art B. Owen. *Monte Carlo theory, methods and examples*. Self-published, 2013. URL https://artowen.su.domains/mc/.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. In *Proceedings of Machine Learning and Systems*, 2023. URL https://proceedings.mlsys.org/paper_files/paper/2023/hash/c4be71ab8d24cdfb45e3d06dbfca2780-Abstract-mlsys2023.html.

Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 759–766, 2000. URL https://www.incompleteideas.net/papers/PSS-00.pdf.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

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

Matthias Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

Vedant Shah, Johan Obando-Ceron, Vineet Jain, Brian Bartoldson, Bhavya Kailkhura, Sarthak Mittal, Glen Berseth, Pablo Samuel Castro, Yoshua Bengio, Nikolay Malkin, Moksh Jain, Siddarth Venkatraman, and Aaron Courville. A comedy of estimators: On kl regularization in rl training of llms. *arXiv preprint*, 2025.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/S0378-3758(00)00115-4.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2009.01325.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794.

G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Physical Review*, 36(5):823–841, 1930. doi: 10.1103/PhysRev.36.823.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54715–54754. PMLR, 2024. URL https://proceedings.mlr.press/v235/xiong24a.html.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tVMPfEGT2w.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

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

Generalisation of RLHF under Reward Shift and Clipped KL Regularisation (Supplementary Material)

Kenton Tang[1] **Yuzhu Chen**[2] **Fengxiang He**[1]

[1]University of Edinburgh
[2]University of Science and Technology of China

A NOTATION

Table 1: Notation

Symbol	Meaning
$\mathcal{X}, \mathcal{Y}$	Prompt space and response space.
(x, y)	A prompt-response pair.
ρ	Prompt distribution used for post-training / evaluation.
ρ_{label}	Prompt distribution used for collecting preference data (reward modelling).
$\pi(\cdot \mid x)$	A policy: conditional distribution over responses given prompt x.
π_θ	Post-trained policy, parameterised by θ.
Θ	Policy parameter space.
π_{ref}	Reference policy (typically an SFT model).
θ	Parameters of the policy π_θ.
Φ	Reward-model parameter space.
ϕ	Parameters of the learned reward model $\hat{r}_\phi$.
$r^\star : \mathcal{X} \times \mathcal{Y} \to [0, 1]$	Target (oracle) reward function.
$\hat{r}_\phi : \mathcal{X} \times \mathcal{Y} \to [0, 1]$	Learned reward model with parameters ϕ.
$e_\phi(x, y)$	Reward-model error, typically $e_\phi(x, y) = \hat{r}_\phi(x, y) - r^\star(x, y)$.
D_{train}	Joint distribution for reward-model training, e.g. $D_{\text{train}}(x, y) = \rho_{\text{label}}(x)\, \pi_{\text{ref}}(y \mid x)$.
D_θ	Policy-induced joint distribution, $D_\theta(x, y) = \rho(x)\, \pi_\theta(y \mid x)$.
$L^{(2)}_{\text{train}}(\phi)$	Reward-model MSE on D_{train}: $\mathbb{E}_{(X,Y)\sim D_{\text{train}}}[e_\phi(X, Y)^2]$.
$\chi^2(D_\theta \Vert D_{\text{train}})$	Chi-square divergence measuring coverage / shift from D_{train} to D_θ.
$C(\theta)$	Coverage coefficient, typically $C(\theta) = \sqrt{1 + \chi^2(D_\theta \Vert D_{\text{train}})}$.
C_{prompt}	Prompt-shift component of coverage (in a factorisation of $C(\theta)$).
$C_{\text{pol}}(\theta)$	Policy-shift component of coverage (in a factorisation of $C(\theta)$).
$\beta > 0$	KL-regularisation strength (penalty coefficient).
$\ell_\theta(x, y)$	Log-ratio, $\ell_\theta(x, y) = \log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)$.
$\tau > 0$	Clipping threshold for log-ratios.
$\ell^\tau_\theta(x, y)$	Clipped log-ratio, $\ell^\tau_\theta(x, y) = \text{clip}(\ell_\theta(x, y), -\tau, \tau)$.
$\text{KL}(\pi_\theta(\cdot \mid x) \Vert \pi_{\text{ref}}(\cdot \mid x))$	Reference KL at prompt x (population expectation of $\ell_\theta(x, Y)$ under $Y \sim \pi_\theta(\cdot \mid x)$).
$J_r(\theta)$	Population objective under reward r: $\mathbb{E}_{X\sim\rho, Y\sim\pi_\theta(\cdot\mid X)}[r(X, Y) - \beta\, \ell_\theta(X, Y)]$.
$J_{r,\tau}(\theta)$	Clipped population objective: replace ℓ_θ by ℓ^τ_θ in $J_r(\theta)$.
$J^\star(\theta)$	Target objective, typically $J^\star(\theta) = J_{r^\star}(\theta)$.
$J^\phi(\theta)$	Learned-reward objective, typically $J^\phi(\theta) = J_{\hat{r}_\phi}(\theta)$.
$J^{\phi,\tau}(\theta)$	Learned-reward clipped objective, typically $J^{\phi,\tau}(\theta) = J_{\hat{r}_\phi,\tau}(\theta)$.

Symbol	Meaning
$\widehat{J}^{r,\tau}_{n,K}(\theta)$	Empirical objective using n prompts and K rollouts per prompt (reward r, clipping τ).
$\widehat{J}^{r,\tau}_{n,\infty}(\theta)$	Conditional (infinite-rollout) analogue: expectation over rollouts given the n sampled prompts.
n	Number of sampled prompts.
K	Number of rollouts per prompt.
P	Prior distribution over Θ (PAC-Bayes).
Q	Posterior distribution over Θ (PAC-Bayes).
$\mathrm{KL}(Q\|P)$	PAC-Bayes complexity term.
$\delta \in (0,1)$	Confidence parameter for high-probability bounds.

B DEFINITIONS AND LEMMAS

Definition 2 (KL divergence [Kullback and Leibler, 1951])**.** *Suppose that P is absolutely continuous with respect to Q. The* KL *divergence is defined by*

$$\mathrm{KL}(P\|Q) := \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

Lemma 8 (Hoeffding's inequality [Hoeffding, 1963])**.** *Let $Z_1, \ldots, Z_N$ be independent random variables. Assume there exist constants $a \le b$ such that $a \le Z_i \le b$ almost surely for every i. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} Z_i\right]| \le (b-a)\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Lemma 9 (Hoeffding's lemma [Boucheron et al., 2013])**.** *Let Z be a random variable and assume $a \le Z \le b$ almost surely. Then, for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\big[\exp\big(\lambda(Z - \mathbb{E}[Z])\big)\big] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Lemma 10 (Change of measure [Catoni, 2007])**.** *Let P and Q be distributions on Θ such that $\mathrm{KL}(Q\|P) < \infty$. Let $F : \Theta \to \mathbb{R}$ satisfy $\mathbb{E}_{\theta\sim P}[\exp(F(\theta))] < \infty$. Then,*

$$\mathbb{E}_{\theta\sim Q}[F(\theta)] \le \mathrm{KL}(Q\|P) + \log \mathbb{E}_{\theta\sim P}[\exp(F(\theta))]\,.$$

Proof. Let p and q denote densities of P and Q with respect to a common reference. By definition, $\mathrm{KL}(Q\|P) = \mathbb{E}_Q[\log(q/p)]$.

Start from the identity

$$\mathbb{E}_Q[F] = \mathbb{E}_Q[\log(e^F)].$$

Insert the density ratio p/q inside the logarithm:

$$\mathbb{E}_Q[F] = \mathbb{E}_Q\left[\log\left(e^F \frac{p}{q}\right)\right] + \mathbb{E}_Q\left[\log\left(\frac{q}{p}\right)\right].$$

The second term is exactly $\mathrm{KL}(Q\|P)$. For the first term, Jensen's inequality gives

$$\mathbb{E}_Q\left[\log\left(e^F \frac{p}{q}\right)\right] \le \log \mathbb{E}_Q\left[e^F \frac{p}{q}\right] = \log \mathbb{E}_P[e^F].$$

Substituting these two relations into the previous display yields

$$\mathbb{E}_Q[F] \le \mathrm{KL}(Q\|P) + \log \mathbb{E}_P[e^F],$$

which is the claimed inequality. □

Definition 3 (χ^2 divergence [Tsybakov, 2009])**.** *Suppose that D_θ is absolutely continuous with respect to D_{train}. The χ^2 divergence is defined by*

$$\chi^2(D_\theta\|D_{\mathrm{train}}) := \mathbb{E}_{(X,Y)\sim D_{\mathrm{train}}}\left[\left(\frac{D_\theta(X,Y)}{D_{\mathrm{train}}(X,Y)} - 1\right)^2\right].$$

Lemma 11 (χ^2 change of measure)**.** *Let P and Q be distributions on a common space and assume $Q \ll P$. Let $w = \frac{dQ}{dP}$ and assume $\chi^2(Q\|P) < \infty$. If f satisfies $\mathbb{E}_{Z\sim P}[f(Z)^2] < \infty$, we have*

$$|\mathbb{E}_{Z\sim Q}[f(Z)]| \le \sqrt{1+\chi^2(Q\|P)}\,\sqrt{\mathbb{E}_{Z\sim P}[f(Z)^2]}.$$

Proof. Because $Q \ll P$, the density ratio $w = \frac{dQ}{dP}$ exists and the expectation under Q can be written as

$$\mathbb{E}_{Z\sim Q}[f(Z)] = \mathbb{E}_{Z\sim P}[w(Z)f(Z)].$$

Applying Cauchy–Schwarz to the right-hand side gives

$$|\mathbb{E}_P[wf]| \le \sqrt{\mathbb{E}_P[w^2]}\,\sqrt{\mathbb{E}_P[f^2]}.$$

It remains to express $\mathbb{E}_P[w^2]$ in terms of $\chi^2(Q\|P)$. By definition,

$$\chi^2(Q\|P) = \mathbb{E}_P[(w-1)^2] = \mathbb{E}_P[w^2] - 2\mathbb{E}_P[w] + 1.$$

Also $\mathbb{E}_P[w] = 1$, since $w = dQ/dP$ integrates to 1 under P. Substituting $\mathbb{E}_P[w] = 1$ into the previous identity yields $\mathbb{E}_P[w^2] = 1 + \chi^2(Q\|P)$. Plugging this into the Cauchy–Schwarz bound gives

$$|\mathbb{E}_{Z\sim Q}[f(Z)]| \le \sqrt{1+\chi^2(Q\|P)}\,\sqrt{\mathbb{E}_{Z\sim P}[f(Z)^2]},$$

which completes the proof. □

Lemma 12 (Monte Carlo estimation of the clipped log ratio)**.** *Under the same conditions of Lemma 2, with probability at least $1-\delta$ over the evaluation prompts and rollouts,*

$$\left|\widehat{\kappa}^{\tau}_{n,K}(\theta) - \kappa^{\tau}(\theta)\right| \le 2\tau\left(\sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{\log(4/\delta)}{2nK}}\right). \tag{9}$$

Lemma 13 (KL divergence between Gaussian distributions [Murphy, 2022])**.** *Let $Q = \mathcal{N}(\mu_Q, \Sigma_Q)$ and $P = \mathcal{N}(\mu_P, \Sigma_P)$ be Gaussian distributions on $\mathbb{R}^d$, where $\Sigma_Q \succ 0$ and $\Sigma_P \succ 0$. Then,*

$$\mathrm{KL}(Q\|P) = \frac{1}{2}\Big(\operatorname{tr}(\Sigma_P^{-1}\Sigma_Q) + (\mu_Q-\mu_P)^\top\Sigma_P^{-1}(\mu_Q-\mu_P) - d + \log\frac{\det(\Sigma_P)}{\det(\Sigma_Q)}\Big). \tag{10}$$

C PROOFS

C.1 ERROR DECOMPOSITION

Proof of Lemma 1. Let $\theta \in \Theta$ and $\phi \in \Phi$ be arbitrary, and let $\tau > 0$ be an arbitrary clipping threshold. The argument is a purely algebraic decomposition in which two intermediate population objectives are inserted between the empirical surrogate objective and the target objective.

Consider the difference $\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^\star(\theta)$. Add and subtract the intermediate quantities $J^{\phi,\tau}(\theta)$ and $J^{\phi}(\theta)$ to obtain

$$\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^\star(\theta) = \widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^{\phi,\tau}(\theta) + J^{\phi,\tau}(\theta) - J^{\phi}(\theta) + J^{\phi}(\theta) - J^\star(\theta).$$

Taking absolute values and applying the triangle inequality gives

$$|\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^\star(\theta)| \le |\widehat{J}^{\phi,\tau}_{n,K}(\theta) - J^{\phi,\tau}(\theta)| + |J^{\phi,\tau}(\theta) - J^{\phi}(\theta)| + |J^{\phi}(\theta) - J^\star(\theta)|.$$

This is exactly the inequality stated in Lemma 1. □

C.2 STATISTICAL ERROR

Proof of Lemma 2. Let $\theta \in \Theta$ be an arbitrary policy parameter. Let $r : \mathcal{X} \times \mathcal{Y} \to [0,1]$ be an arbitrary reward function, let $\tau > 0$ be an arbitrary clipping threshold, and let $\delta \in (0,1)$ be an arbitrary confidence level.

The goal is to control the Monte Carlo deviation arising from drawing only K rollouts per prompt, while conditioning on the realized prompts. Let $x_1, \ldots, x_n$ denote the realized prompts. For each $i \in \{1, \ldots, n\}$ and each rollout index $j \in \{1, \ldots, K\}$, define the per-rollout contribution

$$Z_{i,j} := r(x_i, y_{i,j}) - \beta\, \ell_\theta^\tau(x_i, y_{i,j}).$$

By the definition of the empirical objective, one can rewrite

$$\widehat{J}_{n,K}^{r,\tau}(\theta) = \frac{1}{nK} \sum_{i=1}^{n} \sum_{j=1}^{K} Z_{i,j}.$$

Next define the conditional expectation of the empirical objective given the prompts. For each fixed prompt x_i, conditional on x_i the rollout $y_{i,j}$ is distributed as $\pi_\theta(\cdot \mid x_i)$, hence

$$\mathbb{E}[Z_{i,j} \mid x_i] = \mathbb{E}_{Y \sim \pi_\theta(\cdot|x_i)}[r(x_i, Y)] - \beta\, \mathbb{E}_{Y \sim \pi_\theta(\cdot|x_i)}[\ell_\theta^\tau(x_i, Y)].$$

Averaging these conditional expectations over i yields the infinite-rollout analogue

$$\widehat{J}_{n,\infty}^{r,\tau}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Z_{i,1} \mid x_i].$$

By construction,

$$\mathbb{E}\Big[\widehat{J}_{n,K}^{r,\tau}(\theta) \mid x_{1:n}\Big] = \widehat{J}_{n,\infty}^{r,\tau}(\theta).$$

To apply Hoeffding's inequality, it remains to verify a uniform bound on each $Z_{i,j}$. Because $r(x_i, y_{i,j}) \in [0,1]$ and $\ell_\theta^\tau(x_i, y_{i,j}) \in [-\tau, \tau]$, it follows that

$$-\beta\tau \le Z_{i,j} \le 1 + \beta\tau,$$

so the interval width is $1 + 2\beta\tau$.

Conditional on the prompts $x_{1:n}$, the rollouts are independent across all index pairs (i,j). Therefore the collection $\{Z_{i,j}\}_{i \le n,\, j \le K}$ is independent conditional on $x_{1:n}$. Applying Lemma 8 to the average of these nK bounded independent random variables, with failure probability δ, gives that with probability at least $1 - \delta$ over the rollouts conditional on $x_{1:n}$,

$$|\widehat{J}_{n,K}^{r,\tau}(\theta) - \mathbb{E}\Big[\widehat{J}_{n,K}^{r,\tau}(\theta) \mid x_{1:n}\Big]| \le (1 + 2\beta\tau)\sqrt{\frac{\log(2/\delta)}{2nK}}.$$

Replacing the conditional expectation by $\widehat{J}_{n,\infty}^{r,\tau}(\theta)$ yields

$$|\widehat{J}_{n,K}^{r,\tau}(\theta) - \widehat{J}_{n,\infty}^{r,\tau}(\theta)| \le (1 + 2\beta\tau)\sqrt{\frac{\log(2/\delta)}{2nK}},$$

which is the conclusion of Lemma 2. □

Proof of Lemma 3. Let $\theta \in \Theta$ be an arbitrary policy parameter. Let $r : \mathcal{X} \times \mathcal{Y} \to [0,1]$ be an arbitrary reward function, let $\tau > 0$ be an arbitrary clipping threshold, and let $\delta \in (0,1)$ be an arbitrary confidence level.

This lemma controls the deviation due only to sampling finitely many prompts, after taking the conditional expectation over rollouts. Define, for each prompt $x \in \mathcal{X}$,

$$g_\theta^{r,\tau}(x) = \mathbb{E}_{Y \sim \pi_\theta(\cdot|x)}[r(x, Y)] - \beta\, \mathbb{E}_{Y \sim \pi_\theta(\cdot|x)}[\ell_\theta^\tau(x, Y)].$$

Because $r(\cdot,\cdot) \in [0,1]$ and $\ell_\theta^\tau(\cdot,\cdot) \in [-\tau, \tau]$ pointwise, the first expectation lies in $[0,1]$ and the second expectation lies in $[-\tau, \tau]$. Consequently, for every x,

$$-\beta\tau \le g_\theta^{r,\tau}(x) \le 1 + \beta\tau,$$

so the interval width is again $1 + 2\beta\tau$.

By definition,

$$\widehat{J}^{r,\tau}_{n,\infty}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g^{r,\tau}_{\theta}(x_i), \qquad J^{r,\tau}(\theta) = \mathbb{E}_{X\sim\rho}[g^{r,\tau}_{\theta}(X)].$$

Since $x_1, \ldots, x_n$ are independent draws from ρ, the sequence $g^{r,\tau}_{\theta}(x_1), \ldots, g^{r,\tau}_{\theta}(x_n)$ consists of i.i.d. random variables bounded in an interval of width $1 + 2\beta\tau$. Applying Lemma 8 with $N = n$ and failure probability δ yields that with probability at least $1 - \delta$ over the prompts,

$$|\widehat{J}^{r,\tau}_{n,\infty}(\theta) - J^{r,\tau}(\theta)| \le (1 + 2\beta\tau)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

This is precisely the statement of Lemma 3. □

Proof of Lemma 4. Let $\theta \in \Theta$ be an arbitrary policy parameter. Let $r : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be an arbitrary reward function, let $\tau > 0$ be an arbitrary clipping threshold, and let $\delta \in (0, 1)$ be an arbitrary confidence level.

The proof combines the two previous concentration statements by enforcing that they hold on a common high-probability event, and then applying a triangle inequality.

Define the rollout concentration event

$$\mathcal{E}_{\text{roll}} := \left\{ |\widehat{J}^{r,\tau}_{n,K}(\theta) - \widehat{J}^{r,\tau}_{n,\infty}(\theta)| \le (1 + 2\beta\tau)\sqrt{\frac{\log(4/\delta)}{2nK}} \right\}.$$

Lemma 2 applied with confidence parameter $\delta/2$ implies that, conditional on $x_{1:n}$,

$$\Pr(\mathcal{E}_{\text{roll}} \mid x_{1:n}) \ge 1 - \delta/2.$$

Define the prompt concentration event

$$\mathcal{E}_{\text{prompt}} := \left\{ |\widehat{J}^{r,\tau}_{n,\infty}(\theta) - J^{r,\tau}(\theta)| \le (1 + 2\beta\tau)\sqrt{\frac{\log(4/\delta)}{2n}} \right\}.$$

Lemma 3 applied with confidence parameter $\delta/2$ yields

$$\Pr(\mathcal{E}_{\text{prompt}}) \ge 1 - \delta/2.$$

Let $\mathcal{E}_{\text{stat}} := \mathcal{E}_{\text{roll}} \cap \mathcal{E}_{\text{prompt}}$. By the union bound,

$$\Pr(\mathcal{E}_{\text{stat}}) \ge 1 - \delta.$$

Assume that $\mathcal{E}_{\text{stat}}$ holds. Then, the triangle inequality gives

$$\begin{aligned} |\widehat{J}^{r,\tau}_{n,K}(\theta) - J^{r,\tau}(\theta)| &\le |\widehat{J}^{r,\tau}_{n,K}(\theta) - \widehat{J}^{r,\tau}_{n,\infty}(\theta)| + |\widehat{J}^{r,\tau}_{n,\infty}(\theta) - J^{r,\tau}(\theta)| \\ &\le (1 + 2\beta\tau)\left(\sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{\log(4/\delta)}{2nK}}\right), \end{aligned}$$

which is exactly the inequality claimed in Lemma 4. □

Proof of Lemma 12. Let $\theta \in \Theta$ be an arbitrary policy parameter, let $\tau > 0$ be an arbitrary clipping threshold, and let $\delta \in (0, 1)$ be an arbitrary confidence level. Recall that $x_1, \ldots, x_n$ are independent draws from ρ, and that, conditional on each x_i, the rollouts $y_{i,1}, \ldots, y_{i,K}$ are independent draws from $\pi_\theta(\cdot \mid x_i)$. Define the per-rollout clipped log ratio

$$Z_{i,j} := \ell^{\tau}_{\theta}(x_i, y_{i,j}),$$

so that, by the definition of $\widehat{\kappa}^{\tau}_{n,K}(\theta)$,

$$\widehat{\kappa}^{\tau}_{n,K}(\theta) = \frac{1}{nK} \sum_{i=1}^{n} \sum_{j=1}^{K} Z_{i,j}.$$

Because $\ell^{\tau}_{\theta}(x, y) = \mathrm{clip}(\ell_{\theta}(x, y), -\tau, \tau)$ by definition, it follows that $Z_{i,j} \in [-\tau, \tau]$ almost surely for all (i, j), and therefore each $Z_{i,j}$ is bounded in an interval of width 2τ.

To make the two-stage sampling structure explicit, introduce the conditional infinite-rollout analogue

$$\widehat{\kappa}^{\tau}_{n,\infty}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Z_{i,1} \mid x_i] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Y \sim \pi_{\theta}(\cdot|x_i)}[\ell^{\tau}_{\theta}(x_i, Y)] \,.$$

By construction, conditional on the realized prompts $x_{1:n}$, the random variables $\{Z_{i,j}\}_{i \le n,\, j \le K}$ are independent, and moreover

$$\mathbb{E}\big[\widehat{\kappa}^{\tau}_{n,K}(\theta) \mid x_{1:n}\big] = \widehat{\kappa}^{\tau}_{n,\infty}(\theta).$$

Applying Lemma 8 to the average of the nK bounded independent random variables $\{Z_{i,j}\}$, conditional on $x_{1:n}$ and with failure probability $\delta/2$, yields that with probability at least $1 - \delta/2$ over the rollouts conditional on $x_{1:n}$,

$$\big|\widehat{\kappa}^{\tau}_{n,K}(\theta) - \widehat{\kappa}^{\tau}_{n,\infty}(\theta)\big| \le 2\tau \sqrt{\frac{\log(4/\delta)}{2nK}}.$$

It remains to control the deviation due to sampling only finitely many prompts. Define the prompt-level functional

$$h^{\tau}_{\theta}(x) := \mathbb{E}_{Y \sim \pi_{\theta}(\cdot|x)}[\ell^{\tau}_{\theta}(x, Y)] \,.$$

Since $\ell^{\tau}_{\theta}(x, Y) \in [-\tau, \tau]$ almost surely under $Y \sim \pi_{\theta}(\cdot \mid x)$, it follows that $h^{\tau}_{\theta}(x) \in [-\tau, \tau]$ for every x, and thus $h^{\tau}_{\theta}(X)$ is bounded in an interval of width 2τ when $X \sim \rho$. By the definition of $\widehat{\kappa}^{\tau}_{n,\infty}(\theta)$,

$$\widehat{\kappa}^{\tau}_{n,\infty}(\theta) = \frac{1}{n} \sum_{i=1}^{n} h^{\tau}_{\theta}(x_i).$$

Moreover, by the definition of $D_{\theta}(x, y) = \rho(x)\pi_{\theta}(y \mid x)$, the clipped population average can be written as

$$\kappa^{\tau}(\theta) = \mathbb{E}_{(X,Y) \sim D_{\theta}}[\ell^{\tau}_{\theta}(X, Y)] = \mathbb{E}_{X \sim \rho}[h^{\tau}_{\theta}(X)] \,.$$

Since $x_1, \ldots, x_n$ are independent draws from ρ, the sequence $h^{\tau}_{\theta}(x_1), \ldots, h^{\tau}_{\theta}(x_n)$ consists of i.i.d. random variables bounded in an interval of width 2τ. Applying Lemma 8 with $N = n$ and failure probability $\delta/2$ yields that with probability at least $1 - \delta/2$ over the prompts,

$$\big|\widehat{\kappa}^{\tau}_{n,\infty}(\theta) - \kappa^{\tau}(\theta)\big| \le 2\tau \sqrt{\frac{\log(4/\delta)}{2n}}.$$

Finally, consider the event on which both of the preceding inequalities hold. By the union bound, this event has probability at least $1 - \delta$ over the joint draw of prompts and rollouts. On this event, the triangle inequality implies

$$\begin{aligned} \big|\widehat{\kappa}^{\tau}_{n,K}(\theta) - \kappa^{\tau}(\theta)\big| &\le \big|\widehat{\kappa}^{\tau}_{n,K}(\theta) - \widehat{\kappa}^{\tau}_{n,\infty}(\theta)\big| + \big|\widehat{\kappa}^{\tau}_{n,\infty}(\theta) - \kappa^{\tau}(\theta)\big| \\ &\le 2\tau \left(\sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{\log(4/\delta)}{2nK}} \right), \end{aligned}$$

which is exactly the claimed bound in (9). □

C.3 REWARD SHIFT AND SURROGATE BIAS

Proof of Lemma 5. Let $\theta \in \Theta$ and $\phi \in \Phi$ be arbitrary parameters. The proof begins by expressing the objective gap as an expectation of reward-model error under the deployment distribution, and then transferring this expectation back to the reward-model training distribution via a density ratio.

By definition,

$$J^{\phi}(\theta) = \mathbb{E}_{X\sim\rho}\mathbb{E}_{Y\sim\pi_\theta(\cdot|X)}[\hat{r}_\phi(X,Y)] - \beta\,\mathbb{E}_{X\sim\rho}\mathrm{KL}(\pi_\theta(\cdot \mid X)\|\pi_{\mathrm{ref}}(\cdot \mid X)),$$

and

$$J^{\star}(\theta) = \mathbb{E}_{X\sim\rho}\mathbb{E}_{Y\sim\pi_\theta(\cdot|X)}[r^{\star}(X,Y)] - \beta\,\mathbb{E}_{X\sim\rho}\mathrm{KL}(\pi_\theta(\cdot \mid X)\|\pi_{\mathrm{ref}}(\cdot \mid X)).$$

The KL regularization terms coincide, so they cancel after subtraction, giving

$$J^{\phi}(\theta) - J^{\star}(\theta) = \mathbb{E}_{X\sim\rho}\mathbb{E}_{Y\sim\pi_\theta(\cdot|X)}[\hat{r}_\phi(X,Y) - r^{\star}(X,Y)].$$

Introduce the pointwise reward-model error $e_\phi(x,y) = \hat{r}_\phi(x,y) - r^{\star}(x,y)$. Using the joint distribution $D_\theta(x,y) = \rho(x)\pi_\theta(y \mid x)$, the preceding display can be rewritten as

$$J^{\phi}(\theta) - J^{\star}(\theta) = \mathbb{E}_{(X,Y)\sim D_\theta}[e_\phi(X,Y)].$$

Assume that $D_\theta \ll D_{\mathrm{train}}$ and define the density ratio

$$w_\theta(x,y) := \frac{D_\theta(x,y)}{D_{\mathrm{train}}(x,y)}.$$

Then, the expectation under D_θ can be written under D_{train} as

$$\mathbb{E}_{(X,Y)\sim D_\theta}[e_\phi(X,Y)] = \mathbb{E}_{(X,Y)\sim D_{\mathrm{train}}}[w_\theta(X,Y)e_\phi(X,Y)].$$

Applying Cauchy–Schwarz yields

$$|\mathbb{E}_{D_{\mathrm{train}}}[w_\theta e_\phi]| \le \sqrt{\mathbb{E}_{D_{\mathrm{train}}}[w_\theta^2]}\,\sqrt{\mathbb{E}_{D_{\mathrm{train}}}[e_\phi^2]}.$$

The second factor is exactly $\sqrt{L^{(2)}_{\mathrm{train}}(\phi)}$ by the definition of $L^{(2)}_{\mathrm{train}}(\phi)$. For the first factor, note that $\mathbb{E}_{D_{\mathrm{train}}}[w_\theta] = 1$ and

$$\chi^2(D_\theta\|D_{\mathrm{train}}) = \mathbb{E}_{D_{\mathrm{train}}}\left[(w_\theta - 1)^2\right] = \mathbb{E}_{D_{\mathrm{train}}}[w_\theta^2] - 1.$$

Consequently, $\mathbb{E}_{D_{\mathrm{train}}}[w_\theta^2] = 1 + \chi^2(D_\theta\|D_{\mathrm{train}})$. Substituting these identities into the Cauchy–Schwarz bound gives

$$|J^{\phi}(\theta) - J^{\star}(\theta)| \le \sqrt{1 + \chi^2(D_\theta\|D_{\mathrm{train}})}\,\sqrt{L^{(2)}_{\mathrm{train}}(\phi)}.$$

By the definition of $\mathcal{C}(\theta)$ in eq. (5), this is

$$|J^{\phi}(\theta) - J^{\star}(\theta)| \le \mathcal{C}(\theta)\,\sqrt{L^{(2)}_{\mathrm{train}}(\phi)},$$

which is the statement of Lemma 5. □

Proof of Lemma 6. Let $\theta \in \Theta$ be arbitrary. Assume that $\rho \ll \rho_{\mathrm{label}}$ and that $\pi_\theta(\cdot \mid x) \ll \pi_{\mathrm{ref}}(\cdot \mid x)$ for every x with $\rho_{\mathrm{label}}(x) > 0$. Under these conditions, $D_\theta \ll D_{\mathrm{train}}$ holds and the density ratio

$$w_\theta(x,y) := \frac{D_\theta(x,y)}{D_{\mathrm{train}}(x,y)}$$

is well defined on the support of D_{train}.

By definition,

$$\mathcal{C}(\theta)^2 = 1 + \chi^2(D_\theta\|D_{\mathrm{train}}) = \mathbb{E}_{(X,Y)\sim D_{\mathrm{train}}}\left[w_\theta(X,Y)^2\right].$$

Using $D_{\mathrm{train}}(x,y) = \rho_{\mathrm{label}}(x)\pi_{\mathrm{ref}}(y \mid x)$ and $D_\theta(x,y) = \rho(x)\pi_\theta(y \mid x)$, one obtains the factorization

$$w_\theta(x,y) = \frac{\rho(x)}{\rho_{\mathrm{label}}(x)} \cdot \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}.$$

Substituting this expression into the definition of $\mathcal{C}(\theta)^2$ and taking expectation under D_{train} yields

$$\mathcal{C}(\theta)^2 = \mathbb{E}_{X\sim\rho_{\mathrm{label}}}\left[\left(\frac{\rho(X)}{\rho_{\mathrm{label}}(X)}\right)^2 \mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|X)}\left[\left(\frac{\pi_\theta(Y \mid X)}{\pi_{\mathrm{ref}}(Y \mid X)}\right)^2\right]\right].$$

By the definition of $\mathcal{C}_{\mathrm{pol}}(\theta)$, the inner expectation is bounded above by $\mathcal{C}_{\mathrm{pol}}(\theta)^2$ for each x in the support of ρ_{label}. Therefore,

$$\mathcal{C}(\theta)^2 \le \mathcal{C}_{\mathrm{pol}}(\theta)^2 \mathbb{E}_{X\sim\rho_{\mathrm{label}}}\left[\left(\frac{\rho(X)}{\rho_{\mathrm{label}}(X)}\right)^2\right] = \mathcal{C}_{\mathrm{pol}}(\theta)^2\,\mathcal{C}_{\mathrm{prompt}}^2.$$

Taking square roots yields $\mathcal{C}(\theta) \le \mathcal{C}_{\mathrm{prompt}}\,\mathcal{C}_{\mathrm{pol}}(\theta)$. □

Proof of Lemma 7. Let $\theta \in \Theta$ and $\phi \in \Phi$ be arbitrary parameters, and let $\tau > 0$ be an arbitrary clipping threshold. The argument is an identity at the level of population objectives, followed by a standard absolute-value bound.

By definition of the clipped objective,

$$J^{\phi,\tau}(\theta) = \mathbb{E}_{X\sim\rho}\mathbb{E}_{Y\sim\pi_\theta(\cdot|X)}[\hat{r}_\phi(X,Y)] - \beta\,\mathbb{E}_{X\sim\rho}\mathbb{E}_{Y\sim\pi_\theta(\cdot|X)}[\ell_\theta^\tau(X,Y)].$$

Using $D_\theta(x,y) = \rho(x)\pi_\theta(y \mid x)$, this can be written as

$$J^{\phi,\tau}(\theta) = \mathbb{E}_{(X,Y)\sim D_\theta}[\hat{r}_\phi(X,Y)] - \beta\,\mathbb{E}_{(X,Y)\sim D_\theta}[\ell_\theta^\tau(X,Y)].$$

For the exact objective, recall that

$$\mathrm{KL}(\pi_\theta(\cdot \mid x)\|\pi_{\mathrm{ref}}(\cdot \mid x)) = \mathbb{E}_{Y\sim\pi_\theta(\cdot|x)}[\ell_\theta(x,Y)].$$

Substituting this identity into the definition of $J^\phi(\theta)$ yields

$$J^\phi(\theta) = \mathbb{E}_{(X,Y)\sim D_\theta}[\hat{r}_\phi(X,Y)] - \beta\,\mathbb{E}_{(X,Y)\sim D_\theta}[\ell_\theta(X,Y)].$$

Subtracting the two displays gives the exact identity

$$J^{\phi,\tau}(\theta) - J^\phi(\theta) = \beta\,\mathbb{E}_{(X,Y)\sim D_\theta}\big[\ell_\theta(X,Y) - \ell_\theta^\tau(X,Y)\big].$$

Taking absolute values and using $|\mathbb{E}[U]| \le \mathbb{E}[|U|]$ yields

$$|J^{\phi,\tau}(\theta) - J^\phi(\theta)| \le \beta\,\mathbb{E}_{(X,Y)\sim D_\theta}\big[|\ell_\theta(X,Y) - \ell_\theta^\tau(X,Y)|\big],$$

which is precisely the inequality asserted in Lemma 7. □

C.4 UNIFIED FIXED-POLICY BOUND

Proof of Theorem 1. Let $\theta \in \Theta$ and $\phi \in \Phi$ be arbitrary, and let $\tau > 0$ and $\delta \in (0,1)$ be arbitrary. Assume the conditions stated in Theorem 1, so that Lemmas 4, 5, and 7 are applicable.

Lemma 1 provides the deterministic decomposition

$$|\widehat{J}_{n,K}^{\phi,\tau}(\theta) - J^\star(\theta)| \le |\widehat{J}_{n,K}^{\phi,\tau}(\theta) - J^{\phi,\tau}(\theta)| + |J^{\phi,\tau}(\theta) - J^\phi(\theta)| + |J^\phi(\theta) - J^\star(\theta)|.$$

To control the first term, apply Lemma 4 with $r = \hat{r}_\phi$. With probability at least $1-\delta$ over the evaluation prompts and rollouts,

$$|\widehat{J}_{n,K}^{\phi,\tau}(\theta) - J^{\phi,\tau}(\theta)| \le (1+2\beta\tau)\left(\sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{\log(4/\delta)}{2nK}}\right).$$

The remaining two terms are controlled deterministically. Lemma 7 gives

$$|J^{\phi,\tau}(\theta) - J^{\phi}(\theta)| \le \beta\, \mathbb{E}_{(X,Y)\sim D_\theta}[|\ell_\theta(X,Y) - \ell^\tau_\theta(X,Y)|]\,,$$

and Lemma 5 gives

$$|J^{\phi}(\theta) - J^{\star}(\theta)| \le \mathcal{C}(\theta)\,\sqrt{L^{(2)}_{\mathrm{train}}(\phi)}.$$

Substituting these three bounds into the decomposition yields the inequality stated in Theorem 1. □

Proof of Corollary 3. Let $\theta \in \Theta$ be an arbitrary policy parameter, let $\phi \in \Phi$ be an arbitrary reward-model parameter, let $\beta > 0$ be an arbitrary regularization coefficient, let $\delta \in (0,1)$ be an arbitrary confidence level, and let $n \ge 1$ and $K \ge 1$ be arbitrary integers. Define

$$\alpha_{n,K,\delta} := \sqrt{\frac{\log(4/\delta)}{2n}} + \sqrt{\frac{\log(4/\delta)}{2nK}}, \qquad B_\theta(\tau) := (1 + 2\beta\tau)\alpha_{n,K,\delta} + \beta T_\theta(\tau),$$

where

$$T_\theta(\tau) := \mathbb{E}_{(X,Y)\sim D_\theta}\big[(|\ell_\theta(X,Y)| - \tau)_+\big].$$

Let $(X,Y) \sim D_\theta$ and define the nonnegative random variable $Z := |\ell_\theta(X,Y)|$. With this notation one has $T_\theta(\tau) = \mathbb{E}[(Z-\tau)_+]$, so the function of interest can be written as

$$B_\theta(\tau) = (1 + 2\beta\tau)\alpha_{n,K,\delta} + \beta\, \mathbb{E}[(Z-\tau)_+].$$

The next step is to relate the one-sided derivatives of $\tau \mapsto \mathbb{E}[(Z-\tau)_+]$ to the tail probabilities of Z. For every $z \ge 0$ and every $\tau \ge 0$, the identity

$$(z-\tau)_+ \;=\; \int_\tau^\infty \mathbf{1}\{z > t\}\, dt$$

holds, because the integrand equals 1 precisely on the interval $t \in [\tau, z)$ when $z > \tau$, and otherwise it is identically zero. Applying this identity with $z = Z$ and using Tonelli's theorem, which is applicable because the integrand is nonnegative, yields the representation

$$\mathbb{E}[(Z-\tau)_+] = \int_\tau^\infty \Pr(Z > t)\, dt.$$

Let $\tau \ge 0$ and let $h > 0$. Using the integral representation at τ and at $\tau + h$ gives

$$\mathbb{E}[(Z-(\tau+h))_+] - \mathbb{E}[(Z-\tau)_+] = -\int_\tau^{\tau+h} \Pr(Z > t)\, dt.$$

Since the function $t \mapsto \Pr(Z > t)$ is nonincreasing, one has

$$h\,\Pr(Z > \tau + h) \;\le\; \int_\tau^{\tau+h} \Pr(Z > t)\, dt \;\le\; h\,\Pr(Z > \tau).$$

Dividing by h and combining with the previous display yields

$$-\Pr(Z > \tau) \;\le\; \frac{\mathbb{E}[(Z-(\tau+h))_+] - \mathbb{E}[(Z-\tau)_+]}{h} \;\le\; -\Pr(Z > \tau + h).$$

Letting $h \downarrow 0$ and using the monotone convergence $\Pr(Z > \tau + h) \to \Pr(Z > \tau)$ yields the right derivative identity

$$\frac{d}{d\tau^+}\mathbb{E}[(Z-\tau)_+] \;=\; -\Pr(Z > \tau).$$

Let $\tau > 0$ and let $h \in (0, \tau)$. Using the integral representation at τ and at $\tau - h$ gives

$$\mathbb{E}[(Z-\tau)_+] - \mathbb{E}[(Z-(\tau-h))_+] = -\int_{\tau-h}^{\tau} \Pr(Z > t)\, dt.$$

Since $t \mapsto \Pr(Z > t)$ is nonincreasing, one has

$$h \Pr(Z > \tau) \;\le\; \int_{\tau-h}^{\tau} \Pr(Z > t)\, dt \;\le\; h \Pr(Z > \tau - h).$$

Dividing by h and combining with the previous display yields

$$-\Pr(Z > \tau - h) \;\le\; \frac{\mathbb{E}[(Z-\tau)_+] - \mathbb{E}[(Z-(\tau-h))_+]}{h} \;\le\; -\Pr(Z > \tau).$$

Letting $h \downarrow 0$ and using the monotone convergence $\Pr(Z > \tau - h) \to \Pr(Z \ge \tau)$ yields the left derivative identity

$$\frac{d}{d\tau^-}\mathbb{E}[(Z-\tau)_+] \;=\; -\Pr(Z \ge \tau).$$

It now follows that B_θ has one-sided derivatives for every $\tau \ge 0$, and these derivatives satisfy

$$B'_\theta(\tau^+) = 2\beta\alpha_{n,K,\delta} - \beta \Pr(Z > \tau), \qquad B'_\theta(\tau^-) = 2\beta\alpha_{n,K,\delta} - \beta \Pr(Z \ge \tau) \quad \text{for every } \tau > 0.$$

Let $\tau^\star$ be any minimizer of $\tau \mapsto B_\theta(\tau)$ over $\tau \ge 0$. If $\tau^\star > 0$, the minimality of $\tau^\star$ implies that the left derivative is nonpositive and the right derivative is nonnegative, so $B'_\theta((\tau^\star)^-) \le 0 \le B'_\theta((\tau^\star)^+)$ holds. Substituting the one-sided derivative expressions yields

$$\Pr(Z > \tau^\star) \;\le\; 2\alpha_{n,K,\delta} \;\le\; \Pr(Z \ge \tau^\star).$$

If $\tau^\star = 0$, the minimality of $\tau^\star$ implies $0 \le B'_\theta(0^+)$, and therefore $\Pr(Z > 0) \le 2\alpha_{n,K,\delta}$ holds. If $2\alpha_{n,K,\delta} < 1$, the inequality $2\alpha_{n,K,\delta} \le \Pr(Z \ge 0) = 1$ holds as well, and this yields the same two-sided condition with $\tau^\star = 0$.

Finally, if $2\alpha_{n,K,\delta} \ge 1$, for every $\tau > 0$, one has

$$B'_\theta(\tau^-) = 2\beta\alpha_{n,K,\delta} - \beta \Pr(Z \ge \tau) \ge 2\beta\alpha_{n,K,\delta} - \beta \ge 0,$$

and therefore B_θ is nondecreasing on $(0, \infty)$, which implies that $\tau^\star = 0$ is a minimizer over $\tau \ge 0$. Recalling that $Z = |\ell_\theta(X, Y)|$ with $(X, Y) \sim D_\theta$, the stated conditions are exactly

$$\Pr_{(X,Y)\sim D_\theta}\big(|\ell_\theta(X,Y)| > \tau^\star\big) \;\le\; 2\alpha_{n,K,\delta} \;\le\; \Pr_{(X,Y)\sim D_\theta}\big(|\ell_\theta(X,Y)| \ge \tau^\star\big),$$

and when $\Pr(Z = \tau^\star) = 0$ the two inequalities collapse to the equality $\Pr(Z > \tau^\star) = 2\alpha_{n,K,\delta}$, which is equivalent to the quantile statement. □

C.5 PAC-BAYES AUXILIARY BOUNDS

Lemma 14 (PAC-Bayes bound for prompt sampling [McAllester, 1999, Seeger, 2002])**.** *Let P be a prior distribution on Θ, let $\tau > 0$ and $\delta \in (0, 1)$ be given, and let $r : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be a given reward function. With probability at least $1 - \delta$ over $x_1, \ldots, x_n \sim \rho$, the following inequality holds simultaneously for all posteriors Q on Θ:*

$$|J^{r,\tau}(Q) - \widehat{J}^{r,\tau}_{n,\infty}(Q)| \le (1 + 2\beta\tau)\sqrt{\frac{\mathrm{KL}(Q\|P) + \log(4/\delta)}{2n}}.$$

Proof. Let $\lambda > 0$ be arbitrary. For a given parameter value $\theta \in \Theta$, consider a single prompt draw $X \sim \rho$. As in the prompt-sampling argument in Lemma 3, the quantity $g^{r,\tau}_\theta(X)$ lies in the interval $[-\beta\tau,\ 1 + \beta\tau]$. Consequently, the centered random variable $J^{r,\tau}(\theta) - g^{r,\tau}_\theta(X)$ is almost surely bounded in an interval of width $1 + 2\beta\tau$. Applying Lemma 9 yields

$$\mathbb{E}_{X\sim\rho} \exp\big(\lambda\big(J^{r,\tau}(\theta) - g^{r,\tau}_\theta(X)\big)\big) \le \exp\left(\frac{\lambda^2(1+2\beta\tau)^2}{8}\right).$$

Now let $x_1, \ldots, x_n$ be i.i.d. draws from ρ. Using independence and the definition

$$\widehat{J}^{r,\tau}_{n,\infty}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g^{r,\tau}_\theta(x_i),$$

it follows that

$$\mathbb{E}\exp\Big(\lambda\big(J^{r,\tau}(\theta) - \widehat{J}^{r,\tau}_{n,\infty}(\theta)\big)\Big) \le \exp\left(\frac{\lambda^2(1+2\beta\tau)^2}{8n}\right).$$

Taking expectation with respect to $\theta \sim P$ and applying Markov's inequality yields that, with probability at least $1-\delta/2$ over $x_{1:n}$,

$$\mathbb{E}_{\theta\sim P}\exp\Big(\lambda\big(J^{r,\tau}(\theta) - \widehat{J}^{r,\tau}_{n,\infty}(\theta)\big)\Big) \le \frac{2}{\delta}\exp\left(\frac{\lambda^2(1+2\beta\tau)^2}{8n}\right).$$

On this event, Lemma 10 can be applied with

$$F(\theta) = \lambda\big(J^{r,\tau}(\theta) - \widehat{J}^{r,\tau}_{n,\infty}(\theta)\big).$$

For every posterior Q on Θ, this gives

$$\lambda\big(J^{r,\tau}(Q) - \widehat{J}^{r,\tau}_{n,\infty}(Q)\big) \le \mathrm{KL}(Q\|P) + \log\frac{2}{\delta} + \frac{\lambda^2(1+2\beta\tau)^2}{8n}.$$

Optimizing over $\lambda > 0$ yields the one-sided bound

$$J^{r,\tau}(Q) - \widehat{J}^{r,\tau}_{n,\infty}(Q) \le (1+2\beta\tau)\sqrt{\frac{\mathrm{KL}(Q\|P)+\log(2/\delta)}{2n}}.$$

Applying the same argument to the opposite deviation $\widehat{J}^{r,\tau}_{n,\infty}(Q) - J^{r,\tau}(Q)$ and taking a union bound yields the stated two-sided inequality with $\log(4/\delta)$. □

Lemma 15 (PAC-Bayes bound for rollout sampling [Catoni, 2007])**.** *Let P be a prior distribution on Θ, let $\tau > 0$ and $\delta \in (0,1)$ be given, and let $r : \mathcal{X}\times\mathcal{Y} \to [0,1]$ be a given reward function. With probability at least $1-\delta$ over the rollouts conditional on $x_{1:n}$, the following inequality holds simultaneously for all posteriors Q on Θ:*

$$|\widehat{J}^{r,\tau}_{n,\infty}(Q) - \widehat{J}^{r,\tau}_{n,K}(Q)| \le (1+2\beta\tau)\sqrt{\frac{\mathrm{KL}(Q\|P)+\log(4/\delta)}{2nK}}.$$

Proof. Condition on the realized prompts $x_{1:n}$, and let $\lambda > 0$ be arbitrary. For each index pair (i,j), define

$$Z_{i,j}(\theta) := r(x_i, y_{i,j}) - \beta\,\ell^\tau_\theta(x_i, y_{i,j}).$$

For every $\theta \in \Theta$, the bounds $r \in [0,1]$ and $\ell^\tau_\theta \in [-\tau,\tau]$ imply

$$-\beta\tau \le Z_{i,j}(\theta) \le 1 + \beta\tau.$$

Conditional on $(x_{1:n},\theta)$, the rollouts are independent across all pairs (i,j).

Define the deviation

$$\Delta(\theta) := \widehat{J}^{r,\tau}_{n,\infty}(\theta) - \widehat{J}^{r,\tau}_{n,K}(\theta).$$

By construction, $\widehat{J}^{r,\tau}_{n,K}(\theta)$ is the average of the nK random variables $Z_{i,j}(\theta)$, and $\widehat{J}^{r,\tau}_{n,\infty}(\theta)$ is their conditional expectation given $x_{1:n}$. Applying Lemma 9 to the average of bounded independent terms yields

$$\mathbb{E}\big[\exp\big(\lambda\Delta(\theta)\big) \mid x_{1:n},\theta\big] \le \exp\left(\frac{\lambda^2(1+2\beta\tau)^2}{8nK}\right).$$

Taking expectation over $\theta \sim P$ and applying Markov's inequality implies that, with probability at least $1-\delta/2$ over rollouts conditional on $x_{1:n}$,

$$\mathbb{E}_{\theta\sim P}\left[\exp\big(\lambda\Delta(\theta)\big) \mid x_{1:n}\right] \le \frac{2}{\delta}\exp\left(\frac{\lambda^2(1+2\beta\tau)^2}{8nK}\right).$$

On this event, Lemma 10 applied with $F(\theta) = \lambda\Delta(\theta)$ yields that, for every posterior Q,

$$\lambda\big(\widehat{J}^{r,\tau}_{n,\infty}(Q) - \widehat{J}^{r,\tau}_{n,K}(Q)\big) \le \mathrm{KL}(Q\|P) + \log\frac{2}{\delta} + \frac{\lambda^2(1+2\beta\tau)^2}{8nK}.$$

Optimizing over $\lambda > 0$ gives

$$\widehat{J}^{r,\tau}_{n,\infty}(Q) - \widehat{J}^{r,\tau}_{n,K}(Q) \le (1 + 2\beta\tau)\sqrt{\frac{\mathrm{KL}(Q\|P) + \log(2/\delta)}{2nK}}.$$

Applying the same argument to the deviation $-\Delta(\theta)$ and taking a union bound yields the stated two-sided inequality with $\log(4/\delta)$. □

C.6 PAC-BAYES MAIN BOUND

Proof of Theorem 2. Let $\phi \in \Phi$ be arbitrary, and let $\tau > 0$ and $\delta \in (0, 1)$ be given. Let P denote the prior that appears in Theorem 2. The proof proceeds by combining two PAC-Bayes concentration inequalities with the deterministic reward-shift and clipping-bias bounds, and then substituting these ingredients into the same three-term decomposition used in the fixed-policy case.

Apply Lemma 15 with reward $r = \hat{r}_\phi$ and confidence level $\delta/2$. Apply Lemma 14 with reward $r = \hat{r}_\phi$ and confidence level $\delta/2$. By a union bound, with probability at least $1 - \delta$ over prompts and rollouts, both inequalities hold simultaneously for all posteriors Q on Θ.

On this event, for every posterior Q,

$$|\widehat{J}^{\phi,\tau}_{n,K}(Q) - \widehat{J}^{\phi,\tau}_{n,\infty}(Q)| \le (1 + 2\beta\tau)\sqrt{\frac{\mathrm{KL}(Q\|P) + \log(8/\delta)}{2nK}},$$
$$|\widehat{J}^{\phi,\tau}_{n,\infty}(Q) - J^{\phi,\tau}(Q)| \le (1 + 2\beta\tau)\sqrt{\frac{\mathrm{KL}(Q\|P) + \log(8/\delta)}{2n}}.$$

Combining these two bounds via the triangle inequality yields

$$|\widehat{J}^{\phi,\tau}_{n,K}(Q) - J^{\phi,\tau}(Q)| \le (1 + 2\beta\tau)\left(\sqrt{\frac{\mathrm{KL}(Q\|P) + \log(8/\delta)}{2n}} + \sqrt{\frac{\mathrm{KL}(Q\|P) + \log(8/\delta)}{2nK}}\right).$$

The remaining two contributions follow by averaging pointwise bounds over $\theta \sim Q$. Taking expectation in Lemma 7 yields

$$|J^{\phi,\tau}(Q) - J^{\phi}(Q)| \le \beta\, \mathbb{E}_{\theta\sim Q}\left[\mathbb{E}_{(X,Y)\sim D_\theta}[|\ell_\theta(X,Y) - \ell^\tau_\theta(X,Y)|]\right].$$

Taking expectation in Lemma 5 yields

$$|J^{\phi}(Q) - J^{\star}(Q)| \le \mathbb{E}_{\theta\sim Q}[\mathcal{C}(\theta)]\,\sqrt{L^{(2)}_{\mathrm{train}}(\phi)}.$$

Finally, apply the same add-and-subtract decomposition used in Lemma 1 directly to $\widehat{J}^{\phi,\tau}_{n,K}(Q) - J^\star(Q)$, and then substitute the three bounds established above to obtain the stated inequality. On the same event of probability at least $1 - \delta$, this gives the inequality stated in Theorem 2, and the statement holds simultaneously for all posteriors Q because the concentration step was uniform over Q. □

C.7 PROOFS FOR PAC-BAYES SPECIAL CASES

C.7.1 Finite candidate class and checkpoint selection

Proof of Corollary 1. Let $M \ge 2$ be an integer, and let $\Theta_M = \{\theta^{(1)}, \ldots, \theta^{(M)}\}$ be the finite set of candidate parameters described in the statement of the corollary. Let P denote the uniform distribution on Θ_M, so that $P(\theta^{(m)}) = 1/M$ holds for every $m \in \{1, \ldots, M\}$. Let Q be an arbitrary distribution supported on the same finite set Θ_M.

For each $m \in \{1, \ldots, M\}$, define

$$p_m := P(\theta^{(m)}) = \frac{1}{M}, \qquad q_m := Q(\theta^{(m)}),$$

so that $q_m \geq 0$ holds for every m and $\sum_{m=1}^{M} q_m = 1$ holds by the definition of a probability mass function. By the definition of the Kullback–Leibler divergence on a finite set, one has

$$\mathrm{KL}(Q\|P) = \sum_{m=1}^{M} q_m \log \frac{q_m}{p_m}.$$

Substituting the identity $p_m = 1/M$ into the preceding display yields

$$\mathrm{KL}(Q\|P) = \sum_{m=1}^{M} q_m \log(q_m M) = \log M + \sum_{m=1}^{M} q_m \log q_m,$$

where the final equality follows because $\sum_{m=1}^{M} q_m = 1$ allows the factor $\log M$ to be separated from the summation.

It therefore remains to control the quantity $\sum_{m=1}^{M} q_m \log q_m$. For every index $m \in \{1, \ldots, M\}$, the probability value q_m lies in the interval $[0, 1]$, and therefore one has $\log q_m \leq 0$ whenever $q_m > 0$, which implies that $q_m \log q_m \leq 0$ whenever $q_m > 0$. When $q_m = 0$, the contribution $q_m \log q_m$ is interpreted as 0, which is consistent with the limiting identity $\lim_{t \downarrow 0} t \log t = 0$. Consequently, every term in the sum $\sum_{m=1}^{M} q_m \log q_m$ is less than or equal to 0, and hence

$$\sum_{m=1}^{M} q_m \log q_m \leq 0.$$

Substituting this inequality into the identity above gives

$$\mathrm{KL}(Q\|P) = \log M + \sum_{m=1}^{M} q_m \log q_m \leq \log M.$$

Finally, consider the special case in which Q is the Dirac distribution concentrated on a single element $\theta^{(\widehat{m})} \in \Theta_M$. In that case one has $q_{\widehat{m}} = 1$ and $q_m = 0$ for all $m \neq \widehat{m}$. Substituting these values into the definition $\mathrm{KL}(Q\|P) = \sum_{m=1}^{M} q_m \log \frac{q_m}{p_m}$ shows that the only nonzero contribution is the term indexed by $\widehat{m}$, and therefore

$$\mathrm{KL}(Q\|P) = 1 \cdot \log \frac{1}{1/M} = \log M.$$

This proves the final statement of the corollary. □

C.7.2 OU–SGD special case for the PAC-Bayes complexity term

Lemma 16 (Bounds for the stationary covariance in the OU approximation)**.** *Let $H \in \mathbb{R}^{d \times d}$ be symmetric and positive definite, let $\Sigma_g \in \mathbb{R}^{d \times d}$ be symmetric and positive definite, and let $\varepsilon > 0$. Assume that $\Sigma \in \mathbb{R}^{d \times d}$ is symmetric and satisfies the matrix equation*

$$H\Sigma + \Sigma H = \varepsilon \Sigma_g.$$

Assume also that H and Σ_g commute, meaning that $H\Sigma_g = \Sigma_g H$ holds. Assume finally that there exist constants $0 < m \leq M < \infty$ such that $mI \preceq H \preceq MI$. Then, Σ satisfies the two-sided bound

$$\frac{\varepsilon}{2M} \Sigma_g \preceq \Sigma \preceq \frac{\varepsilon}{2m} \Sigma_g. \tag{11}$$

Proof. Throughout the proof, for symmetric matrices A and B, the notation $A \preceq B$ means that $v^\top A v \leq v^\top B v$ holds for every vector $v \in \mathbb{R}^d$. This definition is convenient because it reduces the verification of a matrix inequality to the verification of an ordinary inequality that holds uniformly over all vectors.

Define the matrix-valued function

$$F(t) := e^{-tH} \Sigma e^{-tH} \qquad \text{for } t \geq 0.$$

Since H is symmetric, the matrix exponential e^{-tH} is well-defined for every $t \geq 0$, and the map $t \mapsto F(t)$ is differentiable. Differentiating and using the product rule yields

$$\frac{d}{dt}F(t) = (-He^{-tH})\Sigma e^{-tH} + e^{-tH}\Sigma(-He^{-tH}) = -e^{-tH}(H\Sigma + \Sigma H)e^{-tH}.$$

Substituting the identity $H\Sigma + \Sigma H = \varepsilon\, \Sigma_g$ gives

$$\frac{d}{dt}F(t) = -\varepsilon\, e^{-tH}\Sigma_g e^{-tH}.$$

Integrating the preceding identity from 0 to T gives

$$F(T) - F(0) = -\varepsilon \int_0^T e^{-tH}\Sigma_g e^{-tH}\, dt.$$

Since $F(0) = \Sigma$, rearranging yields

$$\Sigma = F(T) + \varepsilon \int_0^T e^{-tH}\Sigma_g e^{-tH}\, dt.$$

Because H is positive definite, there exists a constant $m_0 > 0$ such that $H \succeq m_0 I$, and therefore the operator norm satisfies $\|e^{-tH}\|_2 \leq e^{-tm_0}$ for every $t \geq 0$. This inequality implies $\|F(T)\|_2 = \|e^{-TH}\Sigma e^{-TH}\|_2 \leq \|e^{-TH}\|_2^2\|\Sigma\|_2 \leq e^{-2Tm_0}\|\Sigma\|_2$, and hence $F(T)$ converges to the zero matrix as $T \to \infty$. Taking the limit $T \to \infty$ yields the integral identity

$$\Sigma = \varepsilon \int_0^\infty e^{-tH}\Sigma_g e^{-tH}\, dt.$$

It remains to compare $e^{-tH}\Sigma_g e^{-tH}$ to scalar multiples of Σ_g in the Loewner order. The assumption $mI \preceq H \preceq MI$ means that every eigenvalue of H lies in the interval $[m, M]$. Consequently, every eigenvalue of e^{-2tH} lies in the interval $[e^{-2tM},\, e^{-2tm}]$, and this implies the inequalities

$$e^{-2tM} I \preceq e^{-2tH} \preceq e^{-2tm} I \qquad \text{for every } t \geq 0.$$

The commutativity condition $H\Sigma_g = \Sigma_g H$ implies that Σ_g commutes with the matrix exponential e^{-tH} for every $t \geq 0$. Therefore one has

$$e^{-tH}\Sigma_g e^{-tH} = \Sigma_g e^{-tH} e^{-tH} = \Sigma_g e^{-2tH}.$$

Since $\Sigma_g \succ 0$, the matrix square root $\Sigma_g^{1/2}$ exists and is symmetric and positive definite. Applying the congruence transformation with $\Sigma_g^{1/2}$ to the Loewner inequalities above yields

$$\Sigma_g^{1/2}\big(e^{-2tM} I\big)\Sigma_g^{1/2} \preceq \Sigma_g^{1/2} e^{-2tH}\Sigma_g^{1/2} \preceq \Sigma_g^{1/2}\big(e^{-2tm} I\big)\Sigma_g^{1/2} \qquad \text{for every } t \geq 0.$$

Using $\Sigma_g^{1/2} I \Sigma_g^{1/2} = \Sigma_g$ and the scalar factors in the two outer terms gives

$$e^{-2tM}\Sigma_g \preceq \Sigma_g^{1/2} e^{-2tH}\Sigma_g^{1/2} \preceq e^{-2tm}\Sigma_g \qquad \text{for every } t \geq 0.$$

The commutativity condition implies that $\Sigma_g^{1/2}$ commutes with e^{-tH} and therefore also commutes with e^{-2tH}, which yields

$$\Sigma_g^{1/2} e^{-2tH}\Sigma_g^{1/2} = e^{-2tH}\Sigma_g = e^{-tH}\Sigma_g e^{-tH}.$$

Substituting this identity into the preceding display yields

$$e^{-2tM}\Sigma_g \preceq e^{-tH}\Sigma_g e^{-tH} \preceq e^{-2tm}\Sigma_g \qquad \text{for every } t \geq 0.$$

Substituting these two bounds into the integral representation of Σ yields

$$\varepsilon \int_0^\infty e^{-2tM}\Sigma_g\, dt \preceq \Sigma \preceq \varepsilon \int_0^\infty e^{-2tm}\Sigma_g\, dt.$$

Evaluating the scalar integrals gives

$$\varepsilon \int_0^\infty e^{-2tM}\, dt = \frac{\varepsilon}{2M}, \qquad \varepsilon \int_0^\infty e^{-2tm}\, dt = \frac{\varepsilon}{2m},$$

and substituting these values proves eq. (11). □

Proof of Corollary 2. Assume the parameter space is $\mathbb{R}^d$ and the prior is $P = \mathcal{N}(\theta_0, \Lambda)$ with $\Lambda \succ 0$. Furthermore, assume the posterior induced by SGD with constant step size $\varepsilon > 0$ is approximated by the stationary Ornstein-Uhlenbeck law $Q_{\mathrm{SGD}} = \mathcal{N}(\hat{\theta}, \Sigma)$.

By the local quadratic approximation of the objective, the covariance Σ satisfies the continuous Lyapunov equation $H\Sigma + \Sigma H = \varepsilon\, \Sigma_g$, where $\Sigma_g \succ 0$ is the gradient noise covariance and $H \succ 0$ is the objective Hessian at the optimum $\hat{\theta}$. We assume that H and Σ_g commute, and that the matrix H is symmetric and satisfies $mI \preceq H \preceq MI$ for some constants $0 < m \leq M < \infty$.

Apply Lemma 13 with $\mu_Q = \hat{\theta}$, $\Sigma_Q = \Sigma$, $\mu_P = \theta_0$, and $\Sigma_P = \Lambda$. This yields

$$\mathrm{KL}(Q_{\mathrm{SGD}} \| P) = \frac{1}{2}\Big(\mathrm{tr}(\Lambda^{-1}\Sigma) + (\hat{\theta} - \theta_0)^\top \Lambda^{-1}(\hat{\theta} - \theta_0) - d + \log \frac{\det(\Lambda)}{\det(\Sigma)} \Big). \tag{12}$$

The remaining task is to upper bound the trace term and to upper bound the logarithmic determinant ratio in a way that makes the dependence on ε, Σ_g, and the constants m and M explicit.

First, apply Lemma 16, which gives $\Sigma \preceq \frac{\varepsilon}{2m}\Sigma_g$. Since $\Lambda^{-1} \succ 0$, this inequality implies $\Lambda^{-1/2}\Sigma\Lambda^{-1/2} \preceq \frac{\varepsilon}{2m}\Lambda^{-1/2}\Sigma_g\Lambda^{-1/2}$, and taking traces yields

$$\mathrm{tr}(\Lambda^{-1}\Sigma) = \mathrm{tr}(\Lambda^{-1/2}\Sigma\Lambda^{-1/2}) \leq \frac{\varepsilon}{2m}\,\mathrm{tr}(\Lambda^{-1/2}\Sigma_g\Lambda^{-1/2}) = \frac{\varepsilon}{2m}\,\mathrm{tr}(\Lambda^{-1}\Sigma_g).$$

Second, apply Lemma 16 again, which also gives $\Sigma \succeq \frac{\varepsilon}{2M}\Sigma_g$. This inequality implies that the eigenvalues of Σ dominate the eigenvalues of $\frac{\varepsilon}{2M}\Sigma_g$ when both collections are arranged in nondecreasing order, and therefore the product of the eigenvalues of Σ is at least the product of the eigenvalues of $\frac{\varepsilon}{2M}\Sigma_g$. Consequently,

$$\det(\Sigma) \geq \det\Big(\frac{\varepsilon}{2M}\Sigma_g\Big) = \Big(\frac{\varepsilon}{2M}\Big)^d \det(\Sigma_g),$$

where the last equality uses the basic scaling rule for determinants. Taking logarithms and rearranging yields

$$\log \frac{\det(\Lambda)}{\det(\Sigma)} \leq \log\det(\Lambda) - \log\det(\Sigma_g) - d\log\Big(\frac{\varepsilon}{2M}\Big).$$

Substituting the preceding two bounds into eq. (12) yields the claimed inequality eq. (8), and this completes the proof. □

C.8 BUDGET ALLOCATION

Derivation of the uniform-cost baseline $K^\star = 1$. Assume that the sampling budget satisfies $nK \leq B$ for some $B > 0$, and assume that each rollout has the same cost so that the constraint depends only on the product nK. Consider the leading-order sampling structure in Lemma 4 and ignore multiplicative constants that do not depend on K. The resulting proxy has the form

$$\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{nK}}.$$

Under the constraint $nK \leq B$, one may take $n = B/K$ without loss of generality for minimizing the proxy over $K \geq 1$. Substituting $n = B/K$ yields

$$\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{nK}} = \sqrt{\frac{K}{B}} + \frac{1}{\sqrt{B}}.$$

The second term does not depend on K, and the first term is strictly increasing in K for $K \geq 1$. Therefore the proxy is minimized by the smallest admissible value of K, which is $K^\star = 1$. □

Proof of Corollary 4. Let $B > 0$, $c_{\mathrm{prefill}} > 0$, and $c_{\mathrm{decode}} > 0$ be given. Assume the budget constraint

$$B \geq n\, c_{\mathrm{prefill}} + nK\, c_{\mathrm{decode}},$$

and consider the leading-order sampling structure induced by Lemma 4. As in the statement, treat K as a continuous variable with $K \geq 1$ and ignore multiplicative constants and logarithmic factors that do not depend on K. The sampling proxy can be written in the form

$$E(n, K) = \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{nK}}.$$

Under the constraint, the choice

$$n = \frac{B}{c_{\text{prefill}} + K c_{\text{decode}}}$$

saturates the budget and maximizes n for a given K, hence it minimizes $E(n, K)$ for that K. Substituting this expression for n gives an objective that depends only on K,

$$E(K) = \sqrt{\frac{c_{\text{prefill}} + K c_{\text{decode}}}{B}} \left(1 + \frac{1}{\sqrt{K}}\right).$$

Since B is constant, minimizing $E(K)$ over $K \geq 1$ is equivalent to minimizing the squared objective

$$F(K) := \left(c_{\text{prefill}} + K c_{\text{decode}}\right) \left(1 + \frac{1}{\sqrt{K}}\right)^2.$$

Expanding the square gives

$$F(K) = \left(c_{\text{prefill}} + K c_{\text{decode}}\right) \left(1 + \frac{2}{\sqrt{K}} + \frac{1}{K}\right) = \left(c_{\text{prefill}} + K c_{\text{decode}}\right) + 2\left(c_{\text{prefill}} + K c_{\text{decode}}\right) K^{-1/2} + \left(c_{\text{prefill}} + K c_{\text{decode}}\right) K^{-1}.$$

Differentiating term by term yields

$$F'(K) = c_{\text{decode}} + 2\left(c_{\text{decode}} K^{-1/2} - \frac{1}{2}\left(c_{\text{prefill}} + K c_{\text{decode}}\right) K^{-3/2}\right) + \left(c_{\text{decode}} K^{-1} - \left(c_{\text{prefill}} + K c_{\text{decode}}\right) K^{-2}\right).$$

Simplifying this expression gives

$$F'(K) = c_{\text{decode}} + c_{\text{decode}} K^{-1/2} - c_{\text{prefill}} K^{-3/2} - c_{\text{prefill}} K^{-2}.$$

Multiplying by K^2 yields an equivalent first-order condition

$$K^2 F'(K) = c_{\text{decode}} K^2 + c_{\text{decode}} K^{3/2} - c_{\text{prefill}} K^{1/2} - c_{\text{prefill}}.$$

Let $u = \sqrt{K}$, so that $K = u^2$ and $K^{3/2} = u^3$. The condition $F'(K) = 0$ is equivalent to

$$c_{\text{decode}} u^4 + c_{\text{decode}} u^3 - c_{\text{prefill}} u - c_{\text{prefill}} = 0,$$

and the polynomial factors as

$$c_{\text{decode}} u^3 (u + 1) - c_{\text{prefill}} (u + 1) = (u + 1)\left(c_{\text{decode}} u^3 - c_{\text{prefill}}\right).$$

Since $u = \sqrt{K} \geq 1$, one has $u + 1 > 0$, so any interior stationary point satisfies $c_{\text{decode}} u^3 = c_{\text{prefill}}$. Therefore

$$u = \left(\frac{c_{\text{prefill}}}{c_{\text{decode}}}\right)^{1/3}, \qquad K = u^2 = \left(\frac{c_{\text{prefill}}}{c_{\text{decode}}}\right)^{2/3}.$$

This is the interior stationary point. Because the optimisation domain is $K \geq 1$, the continuous proxy minimiser is

$$K^\star = \max\left\{1, \left(\frac{c_{\text{prefill}}}{c_{\text{decode}}}\right)^{2/3}\right\}.$$

□

Proof of Corollary 5. Let Z denote the per-rollout contribution used in the empirical objective. Assume that the estimator averages Z over n independent prompts and K independent rollouts per prompt. Define the two-stage variance quantities

$$\sigma^2_{\text{prompt}} := \text{Var}(\mathbb{E}[Z \mid X]), \qquad \sigma^2_{\text{rollout}} := \mathbb{E}[\text{Var}(Z \mid X)].$$

The standard variance decomposition for a two-stage average yields the proxy

$$V(n, K) = \frac{\sigma^2_{\text{prompt}}}{n} + \frac{\sigma^2_{\text{rollout}}}{nK}.$$

Assume the budget constraint

$$B \geq n\, c_{\text{prefill}} + nK\, c_{\text{decode}},$$

and substitute the saturated choice $n = B/(c_{\text{prefill}} + Kc_{\text{decode}})$. This yields

$$V(K) = \frac{c_{\text{prefill}} + Kc_{\text{decode}}}{B} \left(\sigma^2_{\text{prompt}} + \frac{\sigma^2_{\text{rollout}}}{K}\right).$$

Since B is constant, minimizing $V(K)$ over $K \geq 1$ is equivalent to minimizing

$$G(K) := \left(c_{\text{prefill}} + Kc_{\text{decode}}\right) \left(\sigma^2_{\text{prompt}} + \frac{\sigma^2_{\text{rollout}}}{K}\right).$$

Expanding gives

$$G(K) = c_{\text{prefill}}\sigma^2_{\text{prompt}} + c_{\text{prefill}}\frac{\sigma^2_{\text{rollout}}}{K} + c_{\text{decode}}K\sigma^2_{\text{prompt}} + c_{\text{decode}}\sigma^2_{\text{rollout}}.$$

Differentiating yields

$$G'(K) = -c_{\text{prefill}}\frac{\sigma^2_{\text{rollout}}}{K^2} + c_{\text{decode}}\sigma^2_{\text{prompt}}.$$

Setting $G'(K) = 0$ gives

$$c_{\text{decode}}\sigma^2_{\text{prompt}} = c_{\text{prefill}}\frac{\sigma^2_{\text{rollout}}}{K^2},$$

which implies

$$K^2 = \frac{c_{\text{prefill}}}{c_{\text{decode}}} \cdot \frac{\sigma^2_{\text{rollout}}}{\sigma^2_{\text{prompt}}}.$$

Taking square roots yields the interior stationary point

$$K = \sqrt{\frac{c_{\text{prefill}}}{c_{\text{decode}}} \cdot \frac{\sigma^2_{\text{rollout}}}{\sigma^2_{\text{prompt}}}}.$$

Because the optimisation domain is $K \geq 1$, the continuous proxy minimiser is

$$K^\star = \max\left\{1, \sqrt{\frac{c_{\text{prefill}}}{c_{\text{decode}}} \cdot \frac{\sigma^2_{\text{rollout}}}{\sigma^2_{\text{prompt}}}}\right\}.$$

□