LAST-ITERATE CONVERGENCE OF OPTIMISTIC PRIMAL–DUAL METHODS FOR KL-REGULARIZED CONSTRAINED RLHF

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning from Human Feedback (RLHF) plays a significant role in aligning Large Language Models (LLMs) with human preferences. RLHF with expected reward constraints can be formulated as a KL-regularized saddle-point problem, in which simultaneous primal–dual updates can exhibit oscillatory dynamics. We study an optimistic predictor–corrector method that applies optimism to both the policy and the constraint multipliers. In the distributional policy space, we prove geometric last-iterate convergence to the regularized saddle-point set. For parameterized policies, we give an inexact-oracle extension whose residual separates primal solver error and comparator coverage under an additional uniform relative-density bound. Experiments with a PPO-style surrogate on PKU-SafeRLHF-30K illustrate reduced oscillation and improved final-iterate robustness over the implemented PPO-Lagrangian baseline.

1 INTRODUCTION

Although large language models (LLMs) have demonstrated remarkable capabilities across a wide range of language tasks Zhang et al. (2024); Elshin et al. (2024); Wang & Chen (2023), they can also exhibit harmful behaviors such as generating misleading content, producing toxic outputs, and leaking sensitive data Guerreiro et al. (2023); Zhang et al. (2025b); Wen et al. (2023); Feretzakis & Verykios (2024). Aligning LLMs with human preferences that jointly emphasize helpfulness and safety is critical.

In practice, preferences involve multiple attributes, such as helpfulness, conciseness, factuality, and harmlessness, and these attributes are often not perfectly aligned and can even conflict with one another Sorensen et al. (2024). Standard reinforcement learning from human feedback (RLHF) typically optimizes a single scalar reward, which can obscure trade-offs among these attributes Ziegler et al. (2019); Stiennon et al. (2020). This observation naturally motivates constrained RLHF, where the model is optimized for helpfulness while explicitly enforcing safety-related constraints. For example, prior work Dai et al. (2024); Huang et al. (2022); Du et al. (2025) maximizes the helpfulness reward subject to lower-bound constraints on safety-related expected rewards.

Most existing constrained RLHF algorithms based on Lagrangian relaxation cast the problem as a saddle-point game between a policy and a set of non-negative dual variables Dai et al. (2024); Huang et al. (2024); Zhang et al. (2025a). We describe these methods through a common primal–dual template that specifies how they approximately solve the policy subproblem and update the constraint multipliers. This unified perspective exposes several limitations of existing approaches. Simultaneous primal–dual updates Dai et al. (2024) can exhibit unstable saddle-point dynamics and fail to converge in the last iterate, even in simple bilinear problems. Such behavior is problematic in safe RLHF, where the final training iterate is often the model deployed. Multi-shot methods can mitigate this instability when their inner solves are sufficiently accurate, but can be computationally expensive, as they require repeatedly solving the primal policy optimization problem to near optimality for each dual update Zhang et al. (2025a). Meanwhile, one-shot dualization-based approaches exploit the closed-form characterization of the optimal distributional policy for fixed dual variables Huang et al. (2024). This characterization enables efficient dual optimization, but implementing the resulting policy with a neural model can introduce representation and optimization errors. These observations

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

raise the following question: *Can we design an iterative algorithm for constrained RLHF that is single-loop, provably last-iterate convergent, and supports parameterized policies?*

To address this question, we study an optimistic primal–dual method for KL-regularized constrained RLHF. Optimistic primal–dual methods are known to stabilize saddle-point dynamics and admit last-iterate convergence guarantees Ding et al. (2023). Motivated by this observation, we augment both primal and dual updates with optimistic steps. The method first forms primal and dual predictors and then uses them to compute corrected updates.

Our main contributions are threefold. First, we present a unified primal–dual framework for constrained RLHF that characterizes Lagrangian-based algorithms in terms of their primal oracle and dual update Dai et al. (2024); Huang et al. (2024); Zhang et al. (2025a). Within this framework, we develop an optimistic primal–dual (OPD) algorithm for KL-regularized constrained RLHF. Second, we establish last-iterate guarantees for OPD in both distributional and parameterized policy spaces. Under the stated assumptions, we prove geometric last-iterate convergence to the optimal policy in distribution space. For parameterized policies, we derive an explicit last-iterate error bound that separates the geometric decay of the initial error from the accumulated effects of inexact primal updates. This bound quantifies how the accuracy of the policy updates determines the accuracy of the final iterate. Third, we evaluate OPD on the PKU-SafeRLHF-30K alignment benchmark. In the main comparison, our PPO-based implementation exhibits reduced oscillations and a more favorable final helpfulness–safety trade-off than the PD baseline. A more detailed discussion of related work is provided in Section A.

2 Preliminaries on Constrained RLHF

2.1 Constrained RLHF Problem and Lagrangian Method

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the sets of prompts and responses, respectively. We represent a language model as a stochastic policy $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$ that maps each prompt to a distribution over responses, where $\Delta(\mathcal{Y})$ is the probability simplex on $\mathcal{Y}$. Let π_{ref} denote the reference policy obtained by supervised fine-tuning (SFT). For the distribution-space analysis, we take $\mathcal{D}$ to have finite support and $\mathcal{Y}$ to be a finite feasible response set (for example, token sequences truncated at a fixed maximum length). We identify $\mathcal{X}$ with $\mathrm{supp}(\mathcal{D})$ for the theoretical analysis, so $\mathcal{D}(x) > 0$ on $\mathcal{X}$. Hence $\Pi := \prod_{x \in \mathrm{supp}(\mathcal{D})} \Delta(\mathcal{Y})$ is a compact convex set. We consider multiple preference objectives indexed by $\mathcal{K} = \mathcal{S} \cup \mathcal{H}$, where objectives in $\mathcal{S}$ are optimized and those in $\mathcal{H}$, referred to as *constrained rewards*, are enforced via the inequality $\mathbb{E}[R_j(x, y)] \geq 0$ for all $j \in \mathcal{H}$. Without loss of generality, we absorb constraint thresholds into the rewards: an original threshold constraint $\mathbb{E}[R_j] \geq b_j$ is reduced to this form by the shift $R_j \leftarrow R_j - b_j$.

For each objective $k \in \mathcal{K}$, we assume access to a reward model R_k learned from a human preference dataset $\{(x_k^i, y_k^{i,w}, y_k^{i,l})\}_{i=1}^N$ via Bradley–Terry maximum-likelihood estimation Ouyang et al. (2022), where $(y_k^{i,w}, y_k^{i,l})$ are the preferred and less-preferred responses annotated for prompt x_k^i.

The goal of RLHF is to optimize a policy with respect to the learned reward signals while regularizing it to remain close to a reference policy. Let $\mathbf{w} \in \mathbb{R}_+^{|\mathcal{S}|}$, with $\sum_{k \in \mathcal{S}} w_k = 1$, encode the desired trade-off among the soft objectives in $\mathcal{S}$. The constrained multi-objective RLHF problem is

$$\begin{aligned} \max_{\pi} \quad & \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi(\cdot|x)} \left[\sum_{k \in \mathcal{S}} w_k R_k(x, y) \right] - \beta \mathrm{KL}\left(\pi(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)\right) \right] \\ \text{s.t.} \quad & \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi(\cdot|x)} \left[R_j(x, y) \right] \geq 0, \quad \forall j \in \mathcal{H}, \end{aligned}$$

where $\mathcal{D}$ denotes the prompt distribution and $\beta > 0$ controls the weight of KL regularization to the reference policy π_{ref}.

A standard approach to solving constrained RLHF problems is the Lagrangian method. For each constrained reward $j \in \mathcal{H}$, we introduce a non-negative Lagrange multiplier $\lambda_j \geq 0$, and collect them into the dual vector $\boldsymbol{\lambda} := (\lambda_j)_{j \in \mathcal{H}} \in \mathbb{R}_+^{|\mathcal{H}|}$. Given the preference weights $\mathbf{w}$ over the soft objectives, we define the aggregated reward

$$S_{\boldsymbol{\lambda}}(x, y) := \sum_{k \in \mathcal{S}} w_k R_k(x, y) + \sum_{j \in \mathcal{H}} \lambda_j R_j(x, y). \tag{1}$$

Algorithm 1 Universal Lagrangian Alignment Framework

Require: Prompt distribution $\mathcal{D}$; reward models $\{R_k(x,y)\}_{k\in\mathcal{K}}$; soft weights $\{w_j\}_{j\in\mathcal{S}}$; reference policy π_{ref}; KL coefficient β; inverse dual stepsize $\eta_\lambda > 0$; initial θ_0, $\boldsymbol{\lambda}_0 \geq 0$;

1: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

2: **Primal update:** $\pi_{\theta_{t+1}} \leftarrow \textsc{PrimalOracle}(\pi_{\theta_t}, \boldsymbol{\lambda}_t, \pi_{\text{ref}}, \mathcal{D}, \beta)$.

3: **Dual update:** $\boldsymbol{g}_t \leftarrow \textsc{GradEst}(\theta_{t+1}, \boldsymbol{\lambda}_t, \pi_{\text{ref}}, \mathcal{D})$, $\boldsymbol{\lambda}_{t+1} \leftarrow \left[\boldsymbol{\lambda}_t - \frac{1}{\eta_\lambda}\boldsymbol{g}_t\right]_+$.

4: **end for**

5: **Return** $\boldsymbol{\lambda}_T$ and θ_T.

The resulting Lagrangian of the constrained multi-objective RLHF problem is

$$\mathcal{L}(\pi, \boldsymbol{\lambda}) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi(\cdot|x)}\left[S_{\boldsymbol{\lambda}}(x,y)\right] - \beta\, \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}(\pi(\cdot|x) \,\|\, \pi_{\text{ref}}(\cdot|x))\right].$$

The corresponding saddle-point problem is

$$\min_{\boldsymbol{\lambda}\geq 0}\ \max_{\pi}\ \mathcal{L}(\pi, \boldsymbol{\lambda}). \tag{2}$$

2.2 A UNIVERSAL SAFE RLHF FRAMEWORK

For practical implementation, we parameterize the policy as π_θ with parameters $\theta \in \Theta$. Algorithm 1 summarizes a common framework for constrained RLHF algorithms based on Lagrangian relaxation.

For each iteration, the framework alternates between a primal update and a dual update. (1) The primal update is abstracted as a PRIMALORACLE, which aims to maximize the Lagrangian objective for a given dual variable. The oracle may perform one or more policy-gradient steps, use an iterative inner solver to approximately optimize the parameterized policy, or return an exact closed-form solution in distribution space Huang et al. (2024). (2) The dual update uses GRADEST to estimate the expected rewards of the constrained objectives under the current policy, followed by a projected gradient step on the dual variable. Existing alignment methods correspond to different choices of the primal oracle and dual gradient estimator Dai et al. (2024); Du et al. (2025); Liu et al. (2024); Zhang et al. (2025a); Huang et al. (2024), as detailed in Section A.2.

3 OPTIMISTIC PRIMAL–DUAL METHOD

Last-iterate stability is a central concern in constrained RLHF, where training produces a final policy for deployment. In Algorithm 1, the primal update optimizes the policy and the dual update adjusts the constraint multipliers. In the distributional policy space, the KL regularizer makes the Lagrangian strongly concave in the primal variable but only linear in the dual multipliers. The resulting saddle-point problem lies outside the strongly-convex–strongly-concave setting in which gradient descent–ascent admits linear last-iterate convergence guarantees Zamani et al. (2024). This motivates a convergence analysis tailored to the strongly-concave–linear structure of KL-regularized constrained RLHF.

Example: Failure of Last-Iterate Convergence in a Bilinear Saddle-Point Problem We consider a simple convex–concave bilinear problem $\min_{\mathbf{y}} \max_{\mathbf{x}} \mathbf{x}^\top \mathbf{A}\mathbf{y}$, where $\mathbf{A} \in \mathbb{R}^{M\times M}$ is a full-rank matrix whose singular values are $[\sigma_1, \cdots, \sigma_M]$. The standard primal–dual gradient method gives $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha\mathbf{A}\mathbf{y}_t$ and $\mathbf{y}_{t+1} = \mathbf{y}_t - \alpha\mathbf{A}^\top\mathbf{x}_t$, where $\alpha > 0$ is the stepsize. Let $z_t = (\mathbf{x}_t^\top, \mathbf{y}_t^\top)^\top$. Then the update can be written as a linear iteration

$$z_{t+1} = (\mathbf{I} - \alpha\mathbf{J})z_t, \quad \mathbf{J} = \begin{bmatrix} 0 & -\mathbf{A} \\ \mathbf{A}^\top & 0 \end{bmatrix}.$$

The matrix $\mathbf{J}$ has imaginary eigenvalues $\pm i\sigma_i$. Hence, $(\mathbf{I} - \alpha\mathbf{J})$ has eigenvalues $1 \pm i\alpha\sigma_i$ whose magnitudes are larger than 1, implying that the last iterates do not contract toward the saddle point due to the saddle-point problem's inherently rotational structure Daskalakis & Panageas (2018).

This example illustrates how coupling between primal and dual updates can produce rotational dynamics. For KL-regularized constrained RLHF, we analyze this coupling together with the

Algorithm 2 OPD in Distribution Space

Require: Prompt distribution $\mathcal{D}$; Reward models $\{R_k(x,y)\}_{k\in\mathcal{K}}$; soft weights $\{w_j\}_{j\in\mathcal{S}}$; reference policy π_{ref}; KL coefficient β; primal proximal coefficient $\eta_\theta > 0$; dual proximal coefficient $\eta_\lambda > 0$; dual projection set Λ; initial $\hat{\pi}_0 = \pi_{\mathrm{ref}}$ and $\hat{\boldsymbol{\lambda}}_0 \in \Lambda$; initialize the predictors $\pi_{-1} := \hat{\pi}_0$ and $\boldsymbol{\lambda}_{-1} := \hat{\boldsymbol{\lambda}}_0$;

1: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
2: Primal predictor: $\pi_t = \arg\max_\pi \left(\mathcal{L}(\pi, \boldsymbol{\lambda}_{t-1}) - \mathbb{E}_{x\sim\mathcal{D}}\left[\eta_\theta \mathrm{KL}\left(\pi(\cdot|x)\|\hat{\pi}_t(\cdot|x)\right)\right]\right)$.
3: Dual predictor: $\boldsymbol{\lambda}_t = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \boldsymbol{\lambda}^\top \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_{t-1}(\cdot|x)}\left[\left(R_j(x,y)\right)_{j\in\mathcal{H}}\right] + \eta_\lambda \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_t\|_2^2$.
4: Primal corrector: $\hat{\pi}_{t+1} = \arg\max_\pi \mathcal{L}(\pi, \boldsymbol{\lambda}_t) - \mathbb{E}_{x\sim\mathcal{D}}\left[\eta_\theta \mathrm{KL}\left(\pi(\cdot|x)\|\hat{\pi}_t(\cdot|x)\right)\right]$.
5: Dual corrector: $\hat{\boldsymbol{\lambda}}_{t+1} = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \boldsymbol{\lambda}^\top \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_t(\cdot|x)}\left[\left(R_j(x,y)\right)_{j\in\mathcal{H}}\right] + \eta_\lambda \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_t\|_2^2$.
6: **end for**
7: **Return** $\hat{\boldsymbol{\lambda}}_T$ and $\hat{\pi}_T$.

curvature supplied by KL regularization. Policy parameterization further complicates the analysis because the Lagrangian need not remain concave in the policy parameters.

Motivated by these challenges, we adopt an optimistic primal–dual (OPD) method with predictor–corrector updates for both the policy and the constraint multipliers. We first establish geometric last-iterate convergence to the optimal policy and the optimal multiplier set in distribution space. We then derive last-iterate error bounds for parameterized policies under approximation and relative-density assumptions.

3.1 OPD in Distribution Space

We present the optimistic primal–dual algorithm in the distributional policy space in Algorithm 2. The algorithm projects the dual variable onto a bounded set $\Lambda \subset \mathbb{R}_+^{|\mathcal{H}|}$, which keeps the dual iterates in a compact region while still containing every optimal dual multiplier. The explicit construction of Λ depends on a Slater-gap radius and is given in (3) after we introduce the assumptions.

OPD first computes the predictor iterates $(\pi_t, \boldsymbol{\lambda}_t)$ and then uses them to form the corrected iterates $(\hat{\pi}_{t+1}, \hat{\boldsymbol{\lambda}}_{t+1})$. We analyze the convergence of this predictor–corrector scheme below.

To establish the convergence guarantee, we make the following assumptions. Assumption 3.1 is Slater's condition. Together with bounded rewards and a full-support reference in the finite setting, it ensures strong duality and existence of a saddle point, as shown in Lemma B.4. Slater's condition is standard in the analysis of constrained optimization and primal–dual methods Huang et al. (2024); Zhang et al. (2025a); Du et al. (2025). Assumption 3.2 assumes that all reward models are uniformly bounded, which is a common condition in the RLHF literature Du et al. (2025). Assumption 3.3 requires the reference policy to assign nonzero probability to every feasible action. For LLM policies parameterized by softmax distributions, token probabilities are strictly positive over the modeled action set.

Assumption 3.1 (Slater's condition)**.** There exists a policy $\bar{\pi} \in \Pi$ and a constant $\xi > 0$ such that $\mathbb{E}_{x\sim\mathcal{D}, y\sim\bar{\pi}}\left[R_j(x,y)\right] \geq \xi, \forall j \in \mathcal{H}$.

Assumption 3.2 (Bounded rewards)**.** There exists $R_{\max} > 0$ such that $\left|R_k(x,y)\right| \leq R_{\max}$ for all $k \in \mathcal{K}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$.

Assumption 3.3 (Full support of the reference policy)**.** Assume that the reference policy satisfies $\pi_{\mathrm{ref}}(y \mid x) > 0$ for every $x \in \mathrm{supp}(\mathcal{D})$ and every $y \in \mathcal{Y}$.

We initialize $\hat{\pi}_0 = \pi_{\mathrm{ref}}$. Because rewards and dual multipliers are finite on the bounded projection set, the closed-form KL-mirror update is strictly positive wherever π_{ref} is positive. Thus every predictor and corrector has the same support as π_{ref}, which ensures that all KL terms in the proof are finite.

We now specify the dual projection set Λ used in Algorithm 2. Let $\mathcal{J}_0(\pi) := \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi(\cdot|x)}[\sum_{k\in\mathcal{S}} w_k R_k(x,y)] - \beta\mathbb{E}_{x\sim\mathcal{D}}[\mathrm{KL}(\pi(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))]$ denote the KL-regularized soft-

objective value, and $\mathcal{J}_{0,\max} := \sup_{\pi\in\Pi} \mathcal{J}_0(\pi)$. Following Lemma B.4, we set

$$B_\lambda := \frac{\mathcal{J}_{0,\max} - \mathcal{J}_0(\bar{\pi})}{\xi}, \qquad \Lambda := \left\{ \boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{H}|} : \|\boldsymbol{\lambda}\|_1 \le B_\lambda \right\}, \tag{3}$$

where $\bar{\pi}$ and ξ are from Assumption 3.1. Every optimal dual multiplier satisfying the Slater-gap bound in Lemma B.4 belongs to Λ, so this projection does not remove the optimal dual set.

Constants in the convergence bound. Suppose Assumptions 3.1 to 3.3 hold in the finite setting above and $|\mathcal{H}| \ge 1$. Let $\pi^\star$ be the unique distribution-space optimum and $\Lambda^\star \subseteq \Lambda$ its nonempty set of optimal dual multipliers. For $\boldsymbol{\lambda} \in \Lambda$, write

$$\operatorname{dist}_2(\boldsymbol{\lambda}, \Lambda^\star) := \inf_{\boldsymbol{\lambda}^\star \in \Lambda^\star} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2.$$

Set $a := \sqrt{|\mathcal{H}|} R_{\max}$, and $\eta := 3a$, and define

$$L_0 := \exp\left(\frac{2(1 + B_\lambda) R_{\max}}{\beta} \right). \tag{4}$$

For $R_{\mathcal{H}} := (R_j)_{j\in\mathcal{H}}$, define

$$\Gamma := \mathbb{E}_{x\sim\mathcal{D}} \operatorname{Cov}_{y\sim\pi_{\mathrm{ref}}(\cdot|x)}(R_{\mathcal{H}}(x,y)), \qquad g^\star := \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi^\star(\cdot|x)} R_{\mathcal{H}}(x,y).$$

Fix $\boldsymbol{\lambda}^\circ \in \Lambda^\star$ and let $H_{\pi_{\mathrm{ref}}} > 0$ be any finite constant satisfying

$$\operatorname{dist}_2(\boldsymbol{\lambda}, \Lambda^\star) \le H_{\pi_{\mathrm{ref}}} \Big(\|\Gamma^{1/2}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\circ)\|_2 + |(g^\star)^\top \boldsymbol{\lambda}| \Big), \boldsymbol{\lambda} \in \Lambda. \tag{5}$$

Here $\Gamma^{1/2}$ is the positive semidefinite square root of Γ. Such a constant exists by Hoffman's error bound for the linear system characterizing the optimal multipliers. It converts residuals in the optimality conditions into a bound on the distance to $\Lambda^\star$.

For $L \ge L_0$, define

$$\begin{aligned} C_A(L) &:= 2L\beta + \frac{2L\eta^2}{a+\beta} + \frac{L\eta^2}{a} + \frac{\|\Gamma\|_{\mathrm{op}}}{a}, \\ C_g &:= 8B_\lambda^2 a^2 \left(\frac{1}{\beta} + \frac{1}{a+\beta} \right) + \frac{24 B_\lambda^2 \eta^2}{a}, \\ \kappa(L) &:= \frac{\eta}{\beta} + 2\eta H_{\pi_{\mathrm{ref}}}^2 \big(C_A(L) + C_g \big) + \frac{a}{a+\beta} + 1. \end{aligned} \tag{6}$$

Here $\|\Gamma\|_{\mathrm{op}}$ denotes the Euclidean operator norm. These constants depend on the problem and the chosen density bound L, and are independent of the iteration index.

Theorem 3.4. *Suppose Assumptions 3.1 to 3.3 hold in the finite setting above and $|\mathcal{H}| \ge 1$. Initialize $\hat{\pi}_0 = \pi_{\mathrm{ref}}$ and set $\eta_\theta = \eta_\lambda = \eta := 3\sqrt{|\mathcal{H}|} R_{\max}$. For $0 \le t \le T$, the exact population iterates of Algorithm 2 satisfy*

$$K(\pi^\star, \hat{\pi}_t) + \operatorname{dist}_2^2(\hat{\boldsymbol{\lambda}}_t, \Lambda^\star) \le \rho^t \Phi_0 / \eta,$$

where $K(p,q) := \mathbb{E}_x \mathrm{KL}(p(\cdot \mid x) \| q(\cdot \mid x))$, $\pi^\star$ is the unique distribution-space optimum, $\Lambda^\star$ is its nonempty optimal multiplier set in Λ, and

$$\operatorname{dist}_2(\boldsymbol{\lambda}, \Lambda^\star) := \inf_{\boldsymbol{\lambda}^\star \in \Lambda^\star} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\|_2.$$

Here $\Phi_0 = \eta K(\pi^\star, \pi_{\mathrm{ref}}) + \eta \operatorname{dist}_2^2(\hat{\boldsymbol{\lambda}}_0, \Lambda^\star)$, and $\rho := \kappa(L_0)/(1 + \kappa(L_0)) \in (0,1)$, with L_0 and $\kappa(L)$ defined in (4) and (6).

The proof of Theorem 3.4 is given in Section C.

Theorem 3.4 establishes a linear last-iterate convergence guarantee for OPD in the policy distribution space. Specifically, the last iterate $(\hat{\pi}_t, \hat{\boldsymbol{\lambda}}_t)$ admits a geometrically decreasing error bound at rate $\rho < 1$, with the policy error measured by the KL divergence to the optimal policy $\pi^\star$ and the dual error measured by the squared ℓ_2 distance to the optimal dual set $\Lambda^\star$. The bound directly controls the final policy iterate. The result applies even though the Lagrangian is linear, rather than strongly convex, in the dual variable.

Algorithm 3 OPD in Parameter Space

Require: Prompt distribution $\mathcal{D}$; Reward models $\{R_k(x,y)\}_{k\in\mathcal{K}}$; soft weights $\{w_j\}_{j\in\mathcal{S}}$; reference policy π_{ref}; KL coefficient β; primal proximal coefficient $\eta_\theta > 0$; dual proximal coefficient $\eta_\lambda > 0$; dual projection set Λ; initial $\hat{\theta}_0 \in \Theta$ with $\pi_{\hat{\theta}_0} = \pi_{\mathrm{ref}}$ and $\hat{\boldsymbol{\lambda}}_0 \in \Lambda$; initialize the predictors $\theta_{-1} := \hat{\theta}_0$ and $\boldsymbol{\lambda}_{-1} := \hat{\boldsymbol{\lambda}}_0$;

1: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

2: Primal predictor: $\theta_t = \mathrm{Proj}_\Theta\left(\hat{\theta}_t + \frac{1}{\eta_\theta+\beta} F(\hat{\theta}_t)^\dagger \nabla_\theta \mathcal{L}(\pi_{\hat{\theta}_t}, \boldsymbol{\lambda}_{t-1})\right)$.

3: Dual predictor: $\boldsymbol{\lambda}_t = \mathrm{Proj}_\Lambda\left(\hat{\boldsymbol{\lambda}}_t - \frac{1}{2\eta_\lambda}\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\theta_{t-1}}(\cdot|x)}\left[(R_j(x,y))_{j\in\mathcal{H}}\right]\right)$.

4: Primal corrector: $\hat{\theta}_{t+1} = \mathrm{Proj}_\Theta\left(\hat{\theta}_t + \frac{1}{\eta_\theta+\beta} F(\hat{\theta}_t)^\dagger \nabla_\theta \mathcal{L}(\pi_{\hat{\theta}_t}, \boldsymbol{\lambda}_t)\right)$.

5: Dual corrector: $\hat{\boldsymbol{\lambda}}_{t+1} = \mathrm{Proj}_\Lambda\left(\hat{\boldsymbol{\lambda}}_t - \frac{1}{2\eta_\lambda}\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\theta_t}(\cdot|x)}\left[(R_j(x,y))_{j\in\mathcal{H}}\right]\right)$.

6: **end for**

7: **Return** $\hat{\boldsymbol{\lambda}}_T$ and $\pi_{\hat{\theta}_T}$.

3.2 OPD in Parameter Space

3.2.1 OPD Updates in the Parameterized Policy Space

For a parameterized policy π_θ, $\theta \in \Theta$, the Lagrangian need not be concave in θ, and the closed-form policy updates in Algorithm 2 may lie outside the parameterized policy class. We therefore adopt a gradient-based optimistic primal–dual method in the parameter space.

We denote the parameterized counterparts of π_t and $\hat{\pi}_t$ by π_{θ_t} and $\pi_{\hat{\theta}_t}$, respectively. We approximate the KL divergence by its second-order Taylor expansion around $\hat{\theta}_t$, assuming a twice continuously differentiable, strictly positive parameterization. Specifically, when θ_t is sufficiently close to $\hat{\theta}_t$, we have

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}\left(\pi_{\theta_t}(\cdot|x)\|\pi_{\hat{\theta}_t}(\cdot|x)\right)\right] \approx \frac{1}{2}(\theta_t - \hat{\theta}_t)^\top F(\hat{\theta}_t)(\theta_t - \hat{\theta}_t),$$

where $F(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\nabla_\theta \log \pi_\theta(y|x) \nabla_\theta \log \pi_\theta(y|x)^\top\right]$ denotes the Fisher information matrix. To accommodate possible rank deficiency, we use the Moore–Penrose pseudoinverse $F(\theta)^\dagger$. Based on this local KL approximation, we formulate the parameter-space OPD updates using natural policy gradient (NPG) steps. For full tabular logits, these NPG steps reproduce the exact distributional KL-mirror updates, as stated in Remark 3.5.

The OPD procedure in the parameterized policy space is summarized in Algorithm 3. As in distribution space, each iteration computes primal and dual predictors followed by corrector updates.

Both primal gradients in Algorithm 3 are evaluated at $\hat{\theta}_t$, using $\boldsymbol{\lambda}_{t-1}$ for the predictor and $\boldsymbol{\lambda}_t$ for the corrector. Both dual projections use $\hat{\boldsymbol{\lambda}}_t$ as their base point, with constraint rewards evaluated under $\pi_{\theta_{t-1}}$ and π_{θ_t}, respectively. The coefficients η_θ and η_λ correspond to the primal and dual step lengths $(\eta_\theta + \beta)^{-1}$ and $(2\eta_\lambda)^{-1}$ in these updates.

The following remark establishes when the parameter-space updates in Algorithm 3 exactly recover the distribution-space updates in Algorithm 2 under a tabular softmax parameterization.

Remark 3.5 (Policy-Level Equivalence for Tabular Softmax). In the tabular case, take Θ to be the full logit space $\mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$, or any parameter set whose projection leaves the induced policy unchanged. Run Algorithms 2 and 3 with the same problem data, coefficients, dual projection set, and initial multipliers, and with $\pi_{\hat{\theta}_0} = \hat{\pi}_0 = \pi_{\mathrm{ref}}$. For exact population gradients, Fisher matrices, and dual updates, the two algorithms generate the same policies and dual iterates. The equivalence is at the level of policies: for all t, $\pi_{\theta_t} = \pi_t$ and $\pi_{\hat{\theta}_{t+1}} = \hat{\pi}_{t+1}$.

Full tabular logits represent every strictly positive policy. Under the conditions of Remark 3.5, tabular NPG therefore reproduces the exact distribution-space KL-mirror updates. The detailed derivation is given in Section D.

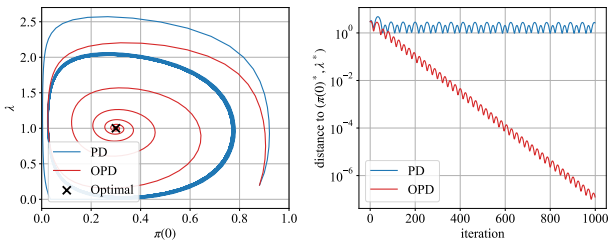


Figure 1: Comparison of OPD and PD under a softmax tabular parameterization in a single-state, two-action RLHF toy problem. At the stated stepsizes, OPD (red) approaches the optimum and PD (blue) exhibits persistent oscillations over the displayed iterations.

3.2.2 A Toy RLHF Example Illustrating the Stability of OPD

We consider a minimal RLHF-style constrained optimization problem with a single state ($|\mathcal{X}| = 1$) and two actions ($|\mathcal{Y}| = 2$), denoted by y_0 and y_1. Since there is only one state, we omit the dependence on x in the following.

We consider two reward models. The first reward R_s represents the objective to be maximized, while the second reward R_h corresponds to a safety-related constraint. We set $R_s(y_0) = 1$ and $R_s(y_1) = 0$, so that the expected reward under a policy π is simply $\pi(y_0)$. For the constrained reward, we choose $R_h(y_0) = -0.7$ and $R_h(y_1) = 0.3$, which induces the constraint $-0.7\,\pi(y_0) + 0.3\,(1 - \pi(y_0)) \geq 0$, or equivalently $\pi(y_0) \leq 0.3$.

We select the reference policy as $\pi_{\mathrm{ref}}(y_0) = 0.3$. The resulting optimization problem is to maximize

$$\pi(y_0) - \beta\,\mathrm{KL}(\pi(\cdot)\,\|\,\pi_{\mathrm{ref}}(\cdot)) \quad \text{subject to} \quad \pi(y_0) \leq 0.3,$$

with $\beta = 0.05$. It is easy to verify that the optimal policy coincides with the reference policy π_{ref}.

To avoid explicit projection onto the policy simplex, we adopt a softmax parameterization $\pi(y_0) = 1/(1 + \exp(\theta))$. We set the effective primal stepsize $\alpha = (\eta_\theta + \beta)^{-1} = 0.6$ and the effective dual stepsize $(2\eta_\lambda)^{-1} = 0.6$. Figure 1 compares the trajectories of OPD and standard primal–dual updates under this parameterization.

As shown in Figure 1, the OPD trajectory approaches the optimal primal–dual solution set, with the plotted distance to the optimum exhibiting approximately geometric decay. In contrast, standard PD updates exhibit persistent oscillations around the optimum. These trajectories illustrate the stabilizing effect of optimistic updates in this constrained RLHF example.

3.2.3 Theoretical Results

Let $\Pi_\Theta = \{\pi_\theta : \theta \in \Theta\} \subseteq \Pi$ be the parameterized policy class. We assume $\pi_\theta(y \mid x) > 0$ for every $\theta \in \Theta$, $x \in \mathrm{supp}(\mathcal{D})$, and $y \in \mathcal{Y}$. We further assume that the parameter domain $\Theta \subset \mathbb{R}^d$ is closed and convex, so that the projection operator $\mathrm{Proj}_\Theta(\cdot)$ used in the updates is well-defined.

We measure primal and dual errors relative to the distribution-space optimum $\pi^\star$ and its optimal multiplier set $\Lambda^\star$, respectively, as in Theorem 3.4. The policy $\pi^\star$ may lie outside Π_Θ. We use the projection set (3), initialize $\pi_{\hat{\theta}_0} = \pi_{\mathrm{ref}}$, and assume exact population dual updates as in Algorithm 3. We write $\pi_t = \pi_{\theta_t}$ and $\hat{\pi}_t = \pi_{\hat{\theta}_t}$. The following assumption compares each implemented primal update with its ideal distribution-space counterpart.

Assumption 3.6 (Approximate realization of KL-mirror primal updates)**.** For iteration t, define the ideal distribution-space KL-mirror predictor and corrector by

$$\pi_t^\dagger := \arg\max_{\pi \in \Pi} \{\mathcal{L}(\pi, \boldsymbol{\lambda}_{t-1}) - \eta_\theta \mathbb{E}_{x \sim \mathcal{D}}[\mathrm{KL}(\pi(\cdot \mid x) \| \hat{\pi}_t(\cdot \mid x))]\}\,,$$

$$\hat{\pi}_{t+1}^\dagger := \arg\max_{\pi \in \Pi} \{\mathcal{L}(\pi, \boldsymbol{\lambda}_t) - \eta_\theta \mathbb{E}_{x \sim \mathcal{D}}[\mathrm{KL}(\pi(\cdot \mid x) \| \hat{\pi}_t(\cdot \mid x))]\}\,.$$

Here Π is the full distributional policy space, and both updates use the current implemented policy $\hat{\pi}_t = \pi_{\hat{\theta}_t}$ as their KL base point. There are nonnegative, possibly time-varying budgets $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$

such that the predictor obeys

$$\mathbb{E}_x \mathrm{KL}(\pi_{\theta_t} \| \pi_t^\dagger) \le \varepsilon_{1,t}, \qquad \mathbb{E}_x \Big[\mathrm{KL}(\hat{\pi}_{t+1} \| \pi_{\theta_t}) - \mathrm{KL}(\hat{\pi}_{t+1} \| \pi_t^\dagger) \Big] \le \varepsilon_{2,t},$$

and the corrector obeys

$$\mathbb{E}_x \mathrm{KL}(\pi_{\hat{\theta}_{t+1}} \| \hat{\pi}_{t+1}^\dagger) \le \varepsilon_{1,t}, \qquad \mathbb{E}_x \Big[\mathrm{KL}(\pi^\star \| \pi_{\hat{\theta}_{t+1}}) - \mathrm{KL}(\pi^\star \| \hat{\pi}_{t+1}^\dagger) \Big] \le \varepsilon_{2,t}.$$

In Assumption 3.6, $\varepsilon_{1,t}$ bounds the KL discrepancy between the implemented and ideal primal updates. The second quantity, $\varepsilon_{2,t}$, bounds the increase in KL distance from the comparison policy when the ideal update is replaced by the implemented update. The comparison policy is $\hat{\pi}_{t+1}$ for the predictor and $\pi^\star$ for the corrector.

Assumption 3.7 (Uniform relative-density lower bound)**.** For a fixed finite $L \ge L_0$ with L_0 defined in (4), every implemented predictor π_t and corrector $\hat{\pi}_t$ satisfies $\pi_t / \pi_{\mathrm{ref}} \ge L^{-1}$ and $\hat{\pi}_t / \pi_{\mathrm{ref}} \ge L^{-1}$ pointwise.

Assumption 3.7 bounds the probability of each response relative to its reference probability, uniformly over iterations. The exact KL-mirror updates satisfy this bound with $L = L_0$. For inexact parameterized updates, it is an additional assumption.

Corollary 3.8. *Suppose the assumptions of Theorem 3.4 and Assumptions 3.6 and 3.7 hold, with the comparison policy, initialization, and exact dual updates specified above. Set* $\eta_\theta = \eta_\lambda = \eta = 3\sqrt{|\mathcal{H}|} R_{\max}$. *For the implemented iterates, define the energy*

$$\Phi_t := \eta K(\pi^\star, \hat{\pi}_t) + \eta \, \mathrm{dist}_2^2(\hat{\boldsymbol{\lambda}}_t, \Lambda^\star) + a K(\hat{\pi}_t, \pi_{t-1}) + a \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2.$$

Initialize the predictors as $\pi_{-1} = \hat{\pi}_0$ *and* $\boldsymbol{\lambda}_{-1} = \hat{\boldsymbol{\lambda}}_0$. *For* $0 \le t \le T$,

$$\Phi_t \le \rho_{\mathrm{in}}^t \Phi_0 + \sum_{s=0}^{t-1} \rho_{\mathrm{in}}^{t-1-s} \big[2\rho_{\mathrm{in}} \mathrm{gap}_s + (1 - \rho_{\mathrm{in}}) \mathcal{R}_s \big],$$

where $\mathrm{gap}_s = (\eta + \beta)(\varepsilon_{1,s} + \varepsilon_{2,s})$, $\mathcal{R}_s = 4\eta H_{\pi_{\mathrm{ref}}}^2 L (\eta + \beta)^2 \varepsilon_{1,s}$, *and the contraction factor for the inexact primal updates is* $\rho_{\mathrm{in}} := \frac{2\kappa(L)}{1 + 2\kappa(L)} \in (0, 1)$. *Here the subscript* in *denotes inexactness;* $H_{\pi_{\mathrm{ref}}}$ *and* $\kappa(L)$ *are specified in* (5) *and* (6)*, respectively. In particular,*

$$K(\pi^\star, \hat{\pi}_t) + \mathrm{dist}_2^2(\hat{\boldsymbol{\lambda}}_t, \Lambda^\star) \le \Phi_t / \eta.$$

The proof of Corollary 3.8 is given in Section E.

The per-iteration error term $\mathrm{gap}_t = (\eta_\theta + \beta)(\varepsilon_{1,t} + \varepsilon_{2,t})$ contains no explicit factor involving the inverse minimum response probability. In the bound of Corollary 3.8, errors at iteration s are weighted by $\rho_{\mathrm{in}}^{t-1-s}$, so their contribution decays geometrically with subsequent iterations.

Under the assumptions of Corollary 3.8, the initial error decays geometrically and the primal-update errors determine the remaining error bound. For a run satisfying these assumptions with the same constants for all $t \ge 0$, if both $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ tend to zero, then $K(\pi^\star, \hat{\pi}_t) + \mathrm{dist}_2^2(\hat{\boldsymbol{\lambda}}_t, \Lambda^\star)$ tends to zero. Uniformly bounded approximation errors yield a bound on the limit superior of this quantity.

4 Computational Experiments

We evaluate the stability and performance of the PPO-based OPD implementation on helpfulness and harmlessness. We ask two questions: (1) Does OPD produce more stable training trajectories than the PD baseline? (2) Does this stability translate into a better helpfulness–safety trade-off at evaluation?

Datasets and Reward Models We conduct our experiments on the PKU-SafeRLHF-30K preference dataset Dai et al. (2024), which contains approximately 27K training prompts and 3K test prompts, each paired with a preferred and a less-preferred response. In addition to preference labels, the dataset provides safety annotations, where preferences are determined jointly based on helpfulness and harmlessness. We adopt the TinyLlama-1.1B model as the reference policy throughout our experiments. We also initialize the reward and constrained-reward models from TinyLlama-1.1B and train them on the preference dataset following Dai et al. (2024).

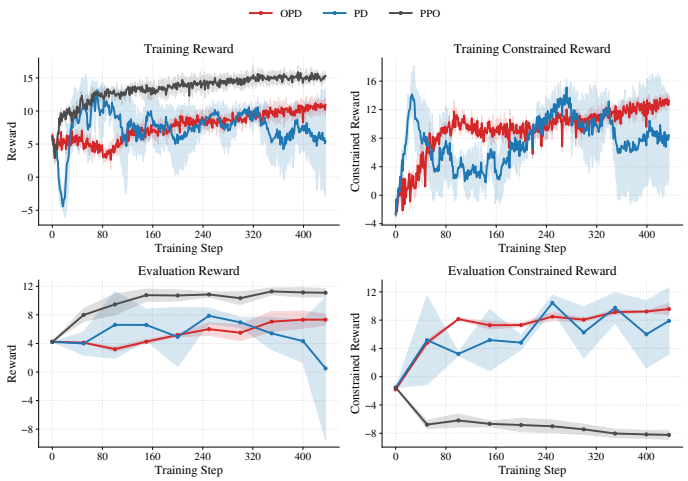


Figure 2: Comparison of OPD, PD, and plain PPO on reward and constrained reward during training and evaluation. Curves show the mean over multiple seeds and shaded regions denote one standard deviation.

Implementation of OPD and Baselines Our implementation applies predictor and corrector steps to the policy and dual variable within a clipped-PPO training loop. The two policy steps use the same rollout batch, and the corrector gradient is evaluated at the predicted policy. The dual updates operate on $\log \lambda$, use an exponentially smoothed batch cost, and constrain λ to $[0.5, 10]$. The PD baseline shares the model and PPO pipeline and uses a standard primal–dual update with exponentially smoothed cost. Plain PPO serves as an unconstrained reference for the helpfulness–safety trade-off.

Hyperparameters and Evaluation Protocol We set the actor and critic learning rates to 5×10^{-6} and 2×10^{-6}, respectively. Both OPD and PD use a dual stepsize of 0.05 for the updates to $\log \lambda$. We use this configuration to examine the stability of OPD and PD under relatively aggressive updates. We impose a lower bound of 0 on the expected constrained reward to express the safety requirement in terms of the learned reward model. We run each method with three random seeds. Solid curves show the mean across seeds, and shaded regions indicate one standard deviation. Evaluation is performed at periodic checkpoints on a held-out subset of test prompts from PKU-SafeRLHF-30K.

Results In Figure 2, OPD exhibits smaller oscillations than PD and attains a more favorable final reward–constraint trade-off in the reported runs. Plain PPO obtains higher reward with lower constrained-reward scores. These observations support the use of predictor–corrector updates in this experimental setting. The learning-rate study in Section F further examines sensitivity to the training configuration.

5 Conclusion

We present a unified primal–dual framework for constrained RLHF and develop an optimistic primal–dual algorithm. For KL-regularized objectives in distribution space, we establish geometric last-iterate convergence. For parameterized policies, we quantify how inexact primal updates affect last-iterate accuracy under the stated approximation and relative-density assumptions. On PKU-SafeRLHF-30K, the PPO-based implementation exhibits reduced oscillations and a more favorable final helpfulness–safety trade-off than the PD baseline in the reported comparison.

AI use statement

In this work, we used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, providing feedback on research methodology or experiments

and implementing methods. We have not used generative AI tools for generating synthetic data sets, proposing or refining hypotheses, designing research methodology or experiments and implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, interpreting results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Akhil Agnihotri, Rahul Jain, Deepak Ramachandran, and Zheng Wen. Multi-objective preference optimization: Improving human alignment of generative models. *arXiv preprint arXiv:2505.10892*, 2025.

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

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in neural information processing systems*, 31, 2018.

Guneet S Dhillon, Xingjian Shi, Yee Whye Teh, and Alex Smola. L3ms–lagrange large language models. *arXiv preprint arXiv:2410.21533*, 2024.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *Advances in Neural Information Processing Systems*, 36:66138–66200, 2023.

Dongsheng Ding, Kaiqing Zhang, Jiali Duan, Tamer Basar, and Mihailo R Jovanovic. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. *Journal of Machine Learning Research*, 26(256):1–76, 2025.

Yihan Du, Seo Taek Kong, and R Srikant. Primal-dual direct preference optimization for constrained llm alignment. *arXiv preprint arXiv:2510.05703*, 2025.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, et al. From general llm to translation: How we dramatically improve translation quality using human evaluation data for llm finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 247–252, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Georgios Feretzakis and Vassilios S Verykios. Trustworthy ai: Securing sensitive data in large language models. *AI*, 5(4):2773–2800, 2024.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023.

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952. URL `https://nvlpubs.nist.gov/nistpubs/jres/049/4/v49.n04.a05.pdf`.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

Sandy Huang, Abbas Abdolmaleki, Giulia Vezzani, Philemon Brakel, Daniel J Mankowitz, Michael Neunert, Steven Bohez, Yuval Tassa, Nicolas Heess, Martin Riedmiller, et al. A constrained multi-objective reinforcement learning framework. In *Conference on Robot Learning*, pp. 883–893. PMLR, 2022.

Xinmeng Huang, Shuo Li, Edgar Dobriban, Osbert Bastani, Hamed Hassani, and Dongsheng Ding. One-shot safety alignment for large language models via optimal dualization. *Advances in Neural Information Processing Systems*, 37:84350–84383, 2024.

Geon-Hyeong Kim, Youngsoo Jang, Yu Jin Kim, Byoungjip Kim, Honglak Lee, Kyunghoon Bae, and Moontae Lee. Safedpo: A simple approach to direct preference optimization with enhanced safety. *arXiv preprint arXiv:2505.20065*, 2025.

Zixuan Liu, Xiaolin Sun, and Zizhan Zheng. Enhancing llm safety via constrained direct preference optimization. *arXiv preprint arXiv:2403.02475*, 2024.

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

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

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Akifumi Wachi, Thien Tran, Rei Sato, Takumi Tanabe, and Youhei Akimoto. Stepwise alignment for constrained language model policy optimization. *Advances in Neural Information Processing Systems*, 37:104471–104520, 2024.

Jianxun Wang and Yixiang Chen. A review on code generation with llms: Application and evaluation. In *2023 IEEE International Conference on Medical Artificial Intelligence (MedAI)*, pp. 284–289. IEEE, 2023.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. Unveiling the implicit toxicity in large language models. *arXiv preprint arXiv:2311.17391*, 2023.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 56276–56297, 2024.

Moslem Zamani, Hadi Abbaszadehpeivasti, and Etienne de Klerk. Convergence rate analysis of the gradient descent–ascent method for convex–concave saddle-point problems. *Optimization Methods and Software*, 39(5):967–989, 2024.

Botong Zhang, Shuo Li, Ignacio Hounie, Osbert Bastani, Dongsheng Ding, and Alejandro Ribeiro. Alignment of large language models with constrained learning. *arXiv preprint arXiv:2505.19387*, 2025a.

Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*, 2024.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pp. 1–46, 2025b.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10586–10613, 2024.

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

A Related Works and Unified Framework for Constrained RLHF

A.1 Related Works

This section summarizes the related work in the LLMs safety alignment.

Constrained Alignment for LLMs There is a growing body of work that formulates safe RLHF as a constrained optimization problem, where helpfulness is maximized subject to safety-related constraints Dai et al. (2024); Huang et al. (2024); Zhang et al. (2025a). A representative approach is modeling the safety violations via an expected cost constraint and solving the resulting constrained objective using iterative primal-dual updates Dai et al. (2024). However, the primal-dual methods can be computationally expensive and may suffer from training instability and hyperparameter sensitivity Huang et al. (2024). To mitigate these issues, Huang et al. (2024) leverages the closed-form structure of the optimal distribution induced by fixed dual variables, and optimizes a smooth dual objective to eliminate repeated primal-dual policy iterations. This dualization-based method leads to more stable training in practice. In contrast, Zhang et al. (2025a) studies constrained alignment in the parameterized LLM policy space and develops an iterative dual-based alignment method that alternates between maximizing the Lagrangian over the LLM policy parameters and performing dual descent updates. Dhillon et al. (2024) proposes an interior point method and uses a relaxed log-barrier function to enforce constraints, thereby avoiding the oscillation between primal and dual variables. Existing approaches stabilize training via simplifying the dual problem with the closed-form structure of the optimal policy distribution, solving near-optimal primal subproblems for each dual variable, or enforcing the constraints with the interior-point methods. The question studied here is a last-iterate guarantee for the particular optimistic KL-mirror updates under a fixed KL-to-reference objective, together with a conditional inexact-oracle extension.

RL-free Based Safety Alignment In parallel, a line of RL-free methods has been proposed for preference alignment, which bypasses explicit reward model learning and policy optimization via reinforcement learning, and instead directly optimize the policy using preference data Rafailov et al. (2023); Azar et al. (2024); Ethayarajh et al. (2024); Hong et al. (2024); Yang et al. (2024). Building upon these approaches, several recent works propose constrained preference alignment in an RL-free manner Liu et al. (2024); Wachi et al. (2024); Du et al. (2025); Kim et al. (2025). Among them, some methods still adopt a primal-dual perspective and iteratively update both the policy and the dual variables, while using Direct Policy Optimization (DPO)-style objectives as the primal optimizer Liu et al. (2024); Du et al. (2025). To implicitly control the trade-off between reward and safety using only reward and cost preference datasets, existing approaches either reweight or reconstruct preference data according to the current dual variables Liu et al. (2024), or perform separate preference optimization on reward and cost datasets under a Lagrangian formulation Du et al. (2025). In contrast, Wachi et al. (2024) avoids iterative dual updates and instead performs constrained alignment that evaluates multiple fixed dual values. Kim et al. (2025) proposes a heuristic yet lightweight approach that enforces safety by directly reordering preference pairs: responses that violate safety guidelines are automatically relabeled as the worse one, enabling safety-aware alignment without explicit dual optimization. While these RL-free approaches improve efficiency and empirical stability, they either rely on heuristic trade-off control or lack a principled analysis with respect to last-iterate convergence under safety constraints.

General Multi-objective Preference Optimization Beyond single-constraint formulations, several works study alignment from a multi-objective perspective. Some approaches aim to approximate Pareto-optimal policies by optimizing scalarization over multiple objectives with respect to a given preference vector Zhou et al. (2024). Other works vary the threshold of constraints to construct a Pareto front, where a primary objective is optimized subject to secondary objectives satisfying varying bounds, enabling flexible trade-offs between helpfulness and safety Agnihotri et al. (2025). In addition, recent work explores context-dependent preference modeling, where alignment objectives dynamically vary with user intent or task context, as exemplified by reward-in-context approaches Yang et al. (2024). While these methods emphasize flexible and expressive preference modeling, they typically do not analyze the optimization dynamics of constrained saddle-point formulations, particularly under expectation-based safety constraints.

A.2 Detailed Explaination of the Unified Framework

We now provide a detailed explanation of how different existing constrained RLHF methods can be recovered as special cases of the universal primal-dual framework described in Algorithm 1.

- Finite-step primal–dual updates. When the primal oracle performs a finite number of stochastic gradient steps, the algorithm reduces to the class of coupled primal–dual methods used in safe RLHF and constrained DPO Dai et al. (2024); Du et al. (2025); Liu et al. (2024). In this regime, the primal policy is updated by a small number of stochastic gradient steps under a non-stationary objective induced by the evolving dual variable. Convex–concave structure alone does not imply last-iterate convergence of these simultaneous updates; guarantees also depend on the algorithm and its assumptions.
- Approximate multi-shot variants. Some recent works decouple the optimization by introducing an outer-loop dual update and an inner-loop primal optimization that approximately maximizes the Lagrangian for a fixed dual variable Zhang et al. (2025a). While this reduces interference between primal and dual updates, the inner-loop problem remains non-convex in the parameter space and is only solved approximately, which prevents these methods from being interpreted as exact primal oracles.
- Exact dualization and one-shot alignment. In contrast, one-shot methods are obtained by analytically eliminating the primal variable in the distribution space, which yields an explicit, closed-form, and convex dual objective Huang et al. (2024). They can be viewed as a degenerate instantiation of the universal framework: the primal oracle returns the closed-form optimal distributional policy for a given dual variable, so no iterative primal updates are required during dual optimization. Therefore, each iteration reduces to a pure dual update step.

B Useful Lemmas

Lemma B.1 (Hölder's inequality)**.** $\|fg\|_1 \le \|f\|_p \|g\|_q$ *for* $1 \le p, q \le \infty$ *and* $\frac{1}{p} + \frac{1}{q} = 1$.

Lemma B.2 (Pinsker's inequality (discrete form))**.** *Let* $p = [p_1, \cdots, p_d]^\top$ *and* $q = [q_1, \cdots, q_d]^\top$ *be probability vectors on a finite set, and assume* $\mathrm{KL}(p\|q) < \infty$*. Then*

$$\|p - q\|_1 \le \sqrt{2\mathrm{KL}(p\|q)}.$$

Lemma B.3 (Young's inequality)**.** *For* $\delta > 0$ *and* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$*,*

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \le \frac{\delta}{2}\|\mathbf{u}\|_2^2 + \frac{1}{2\delta}\|\mathbf{v}\|_2^2.$$

Lemma B.4. *(Ding et al. (2025) Lemma 3(b)) In the finite policy setting, suppose Assumptions 3.1 to 3.3 hold and* $\beta > 0$*. A primal–dual saddle point exists, and every optimal multiplier* $\boldsymbol{\lambda}^\star \in \mathbb{R}_+^{|\mathcal{H}|}$ *satisfies*

$$\|\boldsymbol{\lambda}^\star\|_1 \le \frac{1}{\xi}\mathbb{E}_{x\sim\mathcal{D}}\left[\left(\mathbb{E}_{y\sim\pi^*(\cdot|x)}\left[\sum_{j\in\mathcal{S}} w_j R_j(x,y)\right] - \beta\mathrm{KL}(\pi^*(\cdot|x)||\pi_{\mathrm{ref}}(\cdot|x))\right) - \left(\mathbb{E}_{y\sim\bar{\pi}(\cdot|x)}\left[\sum_{j\in\mathcal{S}} w_j R_j(x,y)\right] - \beta\mathrm{KL}(\bar{\pi}(\cdot|x)||\pi_{\mathrm{ref}}(\cdot|x))\right)\right].$$

Remark B.5. Define the KL-regularized soft-objective value

$$\mathcal{J}_0(\pi) := \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi(\cdot|x)}\left[\sum_{k\in\mathcal{S}} w_k R_k(x,y)\right] - \beta\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}(\pi(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right].$$

By Lemma B.4, we have

$$\|\boldsymbol{\lambda}^\star\|_1 \le \frac{\mathcal{J}_0(\pi^\star) - \mathcal{J}_0(\bar{\pi})}{\xi} \le \frac{\mathcal{J}_{0,\max} - \mathcal{J}_0(\bar{\pi})}{\xi}, \qquad \mathcal{J}_{0,\max} := \sup_{\pi\in\Pi} \mathcal{J}_0(\pi).$$

Define $B_\lambda = \frac{\mathcal{J}_{0,\max} - \mathcal{J}_0(\bar{\pi})}{\xi}$. If we set $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{H}|} \mid \|\boldsymbol{\lambda}\|_1 \le B_\lambda\}$, then the optimality of $\boldsymbol{\lambda}^\star$ is not affected by projection. Moreover, since $\sum_{k\in\mathcal{S}} w_k = 1$, Assumption 3.2 implies $\mathcal{J}_{0,\max} \le R_{\max}$. If $K_{\bar{\pi}} := \mathbb{E}_{x\sim\mathcal{D}}[\mathrm{KL}(\bar{\pi}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))]$ is finite, a fully explicit sufficient radius is

$$B_\lambda \le \frac{2R_{\max} + \beta K_{\bar{\pi}}}{\xi}.$$

Lemma B.6. *Let $\mathcal{X} = \mathrm{supp}(\mathcal{D})$ and $\mathcal{Y}$ be finite, let $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and suppose $\beta > 0$ and $\pi_{\mathrm{ref}}(y \mid x) > 0$ for every (x, y). Consider the optimization problem over all conditional distributions*

$$\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}[\mathbb{E}_{y\sim\pi(\cdot|x)}[S(x,y)] - \beta \mathrm{KL}\left(\pi(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)\right)] \tag{7}$$

The solution can be written as

$$\pi^\star(y|x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\tfrac{1}{\beta}S(x,y)\right),$$

where $Z(x)$ is the normalization factor $Z(x) = \sum_y \pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}S(x,y)\right)$.

Proof. Since the objective decomposes over x, the maximizer can be found pointwise in x. For a fixed x, we can rewrite the optimization problem as

$$\max_{\pi(\cdot|x)\in\Delta} \sum_y \pi(y|x)S(x,y) - \beta \sum_y \pi(y|x)\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)}, \tag{8}$$

For finite S, $\beta > 0$, and a full-support reference, the optimizer below is strictly positive. Here $\Delta = \{\mathbf{x} | \|\mathbf{x}\|_1 = 1 \text{ and } \mathbf{x} \ge 0\}$

Introduce a multiplier $\eta(x)$ for the normalization constraint. The Lagrangian for Equation (8) is

$$\mathcal{L}_x(\pi_x, \eta) = \sum_y \pi(y|x)S(x,y) - \beta\sum_y \pi(y|x)\log\frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} + \eta(x)(\sum_y \pi(y|x) - 1). \tag{9}$$

Taking the derivative w.r.t. $\pi(y|x)$ and setting to zero gives, for every y in the support,

$$0 = \frac{\partial\mathcal{L}_x}{\partial\pi(y|x)} = S(x,y) - \beta(\log\pi(y|x) - \log\pi_{\mathrm{ref}}(y|x) + 1) + \eta(x). \tag{10}$$

Rearranging Equation (10) yields

$$\begin{aligned} \log\pi(y|x) &= \log\pi_{\mathrm{ref}}(y|x) + \frac{1}{\beta}(S(x,y) + \eta(x) - \beta). \\ \Rightarrow \quad \pi^\star(y|x) &= \pi_{\mathrm{ref}}(y|x)\exp\left(\tfrac{1}{\beta}(S(x,y) + \eta(x) - \beta)\right), \end{aligned} \tag{11}$$

Taking the derivative w.r.t. $\eta(x)$ and setting to zero gives $\sum_y \pi^\star(y|x) = 1$. Therefore, we have $\pi^\star(y|x) = \frac{1}{Z(x)}\pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}S(x,y)\right)$, where $Z(x) = \sum_y \pi_{\mathrm{ref}}(y|x)\exp\left(\frac{1}{\beta}S(x,y)\right)$.

For every probability vector $\pi(\cdot \mid x)$, substitution gives

$$\mathbb{E}_{y\sim\pi(\cdot|x)}S(x,y) - \beta\mathrm{KL}(\pi(\cdot \mid x)\|\pi_{\mathrm{ref}}(\cdot \mid x)) = \beta\log Z(x) - \beta\mathrm{KL}(\pi(\cdot \mid x)\|\pi^\star(\cdot \mid x)).$$

Nonnegativity of KL verifies that this strictly positive candidate is the unique global maximizer, including comparison with boundary policies. □

Lemma B.7 (Three-point identity for Bregman divergences)**.** *Let f be strictly convex and continuous on a convex set Ω, and continuously differentiable on its relative interior. Define $D_f(u,v) = f(u) - f(v) - \langle\nabla f(v), u - v\rangle$ for $u \in \Omega$ and $v \in \mathrm{ri}\,\Omega$. Then, for $x \in \Omega$ and $y, z \in \mathrm{ri}\,\Omega$ (using derivatives on the affine hull),*

$$D_f(x,z) - D_f(x,y) - D_f(y,z) = \langle\nabla_y D_f(y,z), x - y\rangle. \tag{12}$$

Proof. By the definition of the Bregman divergence, we have

$$D_f(x,z) = f(x) - f(z) - \langle \nabla f(z), x - z \rangle ,$$
$$D_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle ,$$
$$D_f(y,z) = f(y) - f(z) - \langle \nabla f(z), y - z \rangle .$$

Subtracting the latter two from the first gives

$$\begin{aligned}
&D_f(x,z) - D_f(x,y) - D_f(y,z) \\
&= (f(x) - f(z) - \langle \nabla f(z), x - z \rangle) - (f(x) - f(y) - \langle \nabla f(y), x - y \rangle) - (f(y) - f(z) - \langle \nabla f(z), y - z \rangle) \\
&= - \langle \nabla f(z), x - z \rangle + \langle \nabla f(y), x - y \rangle + \langle \nabla f(z), y - z \rangle \\
&= \langle \nabla f(y), x - y \rangle - \langle \nabla f(z), x - y \rangle \\
&= \langle \nabla f(y) - \nabla f(z), x - y \rangle .
\end{aligned}$$

Also we have

$$\begin{aligned}
\langle \nabla_y D_f(y,z), x - y \rangle &= \langle \nabla_y \left(f(y) - f(z) - \langle \nabla f(z), y - z \rangle \right), x - y \rangle \\
&= \langle \nabla f(y) - \nabla f(z), x - y \rangle
\end{aligned}$$

This concludes the proof. □

Lemma B.8. *Let $h : \Omega \to \mathbb{R}$ be a function that is: a) strictly convex, b) continuously differentiable, c) defined on a closed convex set Ω, and $D_h(u,v), \forall u, v$ be the Bregman divergence defined on h. Fix $\mathbf{x}_{\text{old}} \in \Omega$. Let $f(\mathbf{x}) = \langle g, \mathbf{x} \rangle - \eta D_h(\mathbf{x}, \mathbf{x}_{\text{old}})$. For $\eta > 0$, suppose the following maximizer exists and define*

$$\mathbf{x}_{\text{new}} = \arg\max_{\mathbf{x} \in \Omega} f(\mathbf{x}). \tag{13}$$

Then for any $\mathbf{x}' \in \Omega$,

$$\langle g, \mathbf{x}_{\text{new}} - \mathbf{x}' \rangle \geq \eta \left(-D_h(\mathbf{x}', \mathbf{x}_{\text{old}}) + D_h(\mathbf{x}', \mathbf{x}_{\text{new}}) + D_h(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{old}}) \right). \tag{14}$$

Proof. Note that

$$\begin{aligned}
f(\mathbf{x}) &= \langle g, \mathbf{x} \rangle - \eta D_h(\mathbf{x}, \mathbf{x}_{\text{old}}) \\
&= \langle g, \mathbf{x} \rangle - \eta (h(\mathbf{x}) - h(\mathbf{x}_{\text{old}}) - \langle \nabla h(\mathbf{x}_{\text{old}}), \mathbf{x} - \mathbf{x}_{\text{old}} \rangle) \\
&= \langle g + \eta \nabla h(\mathbf{x}_{\text{old}}), \mathbf{x} \rangle - \eta h(\mathbf{x}) + \eta h(\mathbf{x}_{\text{old}}) - \eta \langle \nabla h(\mathbf{x}_{\text{old}}), \mathbf{x}_{\text{old}} \rangle .
\end{aligned}$$

The first term is linear with $\mathbf{x}$, and the second term is strictly concave with $\mathbf{x}$ since $h(\cdot)$ is strictly convex, and the last two terms are constants with $\mathbf{x}$. Therefore, $f(\mathbf{x})$ is strictly concave with $\mathbf{x}$. As $f(\cdot)$ is differentiable on Ω and $\mathbf{x}_{\text{new}} = \arg\max_{\mathbf{x} \in \Omega} f(\mathbf{x})$, we have

$$\langle \nabla f(\mathbf{x}_{\text{new}}), \mathbf{x}_{\text{new}} - \mathbf{x}' \rangle \geq 0.$$

Substituting $f(\mathbf{x}) = \langle g, \mathbf{x} \rangle - \eta D_h(\mathbf{x}, \mathbf{x}_{\text{old}})$ into the above inequality, we have

$$\langle g - \eta \nabla D_h(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{old}}), \mathbf{x}_{\text{new}} - \mathbf{x}' \rangle \geq 0. \tag{15}$$

Let $x = \mathbf{x}'$, $y = \mathbf{x}_{\text{new}}$, and $z = \mathbf{x}_{\text{old}}$ and $D_f(\cdot, \cdot) = D_h(\cdot, \cdot)$ in Lemma B.7, we have

$$\langle \nabla D_h(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{old}}), \mathbf{x}' - \mathbf{x}_{\text{new}} \rangle = D_h(\mathbf{x}', \mathbf{x}_{\text{old}}) - D_h(\mathbf{x}', \mathbf{x}_{\text{new}}) - D_h(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{old}}).$$

Substituting the above equation to the LHS of Equation (15), we get

$$\langle g, \mathbf{x}_{\text{new}} - \mathbf{x}' \rangle + \eta \left(D_h(\mathbf{x}', \mathbf{x}_{\text{old}}) - D_h(\mathbf{x}', \mathbf{x}_{\text{new}}) - D_h(\mathbf{x}_{\text{new}}, \mathbf{x}_{\text{old}}) \right) \geq 0.$$

This completes the proof. □

Lemma B.9 (Three-point inequality with KL regularization)**.** *Let $\Delta(\mathcal{Y})$ be the probability simplex, and $\pi_{\text{old}}, \pi_{\text{ref}} \in \Delta(\mathcal{Y})$. Let $f(\pi) = \langle g, \pi \rangle - \eta \text{KL}(\pi \| \pi_{\text{old}}) - \beta \text{KL}(\pi \| \pi_{\text{ref}})$. Assume the two supports have nonempty intersection. Define $\Pi_{\text{eff}} = \{\pi \in \Delta(\mathcal{Y}) | \text{supp}(\pi) \subseteq \text{supp}(\pi_{\text{old}}) \cap \text{supp}(\pi_{\text{ref}})\}$. Given step sizes $\eta > 0$ and $\beta > 0$, consider the update*

$$\pi_{\text{new}} = \arg\max_{\pi \in \Delta(\mathcal{Y})} f(\pi). \tag{16}$$

Then for any $\pi' \in \Pi_{\text{eff}}$,

$$\begin{aligned}
&\langle g, \pi_{\text{new}} - \pi' \rangle - \beta \left(\text{KL}(\pi_{\text{new}} \| \pi_{\text{ref}}) - \text{KL}(\pi' \| \pi_{\text{ref}}) \right) \\
&\geq \eta \left(-\text{KL}(\pi' \| \pi_{\text{old}}) + \text{KL}(\pi' \| \pi_{\text{new}}) + \text{KL}(\pi_{\text{new}} \| \pi_{\text{old}}) \right) + \beta \text{KL}(\pi' \| \pi_{\text{new}}).
\end{aligned} \tag{17}$$

Proof. If $\mathrm{supp}(\pi) \not\subseteq \mathrm{supp}(\pi_{\mathrm{old}}) \cap \mathrm{supp}(\pi_{\mathrm{ref}})$, at least one of $\mathrm{KL}(\pi\|\pi_{\mathrm{old}})$ and $\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})$ is infinite. Since $\eta, \beta > 0$, the finite-valued domain of f is therefore Π_{eff}.

Moreover, we can rearrange $f(\pi)$ as

$$f(\pi) = \langle g + \eta \log \pi_{\mathrm{old}} + \beta \log \pi_{\mathrm{ref}}, \pi \rangle - (\eta + \beta) h(\pi),$$

where $h(\pi)$ is the negative entropy of π. Since the first term is linear with π, and $h(\pi)$ is strictly convex with π and $\eta + \beta > 0$, we have $f(\pi)$ is a strictly concave function with π.

On the common support, f is differentiable in the relative interior of Π_{eff}. By the optimality of π_{new} in Equation (16), for every $\pi' \in \Pi_{\mathrm{eff}}$,

$$\langle \nabla f(\pi_{\mathrm{new}}), \pi_{\mathrm{new}} - \pi' \rangle \geq 0.$$

Calculating $\nabla f(\pi_{\mathrm{new}})$ and rearranging the above equation, we have

$$\langle g - \eta \nabla_{\pi_{\mathrm{new}}} \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{old}}) - \beta \nabla_{\pi_{\mathrm{new}}} \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{ref}}), \pi_{\mathrm{new}} - \pi' \rangle \geq 0. \tag{18}$$

The Bregman divergence generated by $h(p)$ is $D_h(u, v) = \mathrm{KL}(u\|v)$. Applying Lemma B.7, we have

$$\begin{aligned}
\langle \nabla_{\pi_{\mathrm{new}}} \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{old}}), \pi' - \pi_{\mathrm{new}} \rangle =& \mathrm{KL}(\pi'\|\pi_{\mathrm{old}}) - \mathrm{KL}(\pi'\|\pi_{\mathrm{new}}) - \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{old}}), \\
\langle \nabla_{\pi_{\mathrm{new}}} \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{ref}}), \pi' - \pi_{\mathrm{new}} \rangle =& \mathrm{KL}(\pi'\|\pi_{\mathrm{ref}}) - \mathrm{KL}(\pi'\|\pi_{\mathrm{new}}) - \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{ref}}),
\end{aligned}$$

where the first equation is letting $x = \pi'$, $y = \pi_{\mathrm{new}}$, $z = \pi_{\mathrm{old}}$, and $f(p) = h(p)$, and the second equation is letting $x = \pi'$, $y = \pi_{\mathrm{new}}$, $z = \pi_{\mathrm{ref}}$, and $f(p) = h(p)$. Substituting the above equations into the LHS of Equation (18), we have

$$\begin{aligned}
&\langle g - \eta \nabla_{\pi_{\mathrm{new}}} \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{old}}) - \beta \nabla_{\pi_{\mathrm{new}}} \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{ref}}), \pi_{\mathrm{new}} - \pi' \rangle \\
&\quad = \langle g, \pi_{\mathrm{new}} - \pi' \rangle + \eta \left(\mathrm{KL}(\pi'\|\pi_{\mathrm{old}}) - \mathrm{KL}(\pi'\|\pi_{\mathrm{new}}) - \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{old}}) \right) \\
&\qquad + \beta \left(\mathrm{KL}(\pi'\|\pi_{\mathrm{ref}}) - \mathrm{KL}(\pi'\|\pi_{\mathrm{new}}) - \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{ref}}) \right) \\
&\quad \geq 0.
\end{aligned} \tag{19}$$

Rearranging this equation, we have

$$\begin{aligned}
&\langle g, \pi_{\mathrm{new}} - \pi' \rangle - \beta \left(\mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{ref}}) - \mathrm{KL}(\pi'\|\pi_{\mathrm{ref}}) \right) \\
&\geq \eta \left(-\mathrm{KL}(\pi'\|\pi_{\mathrm{old}}) + \mathrm{KL}(\pi'\|\pi_{\mathrm{new}}) + \mathrm{KL}(\pi_{\mathrm{new}}\|\pi_{\mathrm{old}}) \right) + \beta \mathrm{KL}(\pi'\|\pi_{\mathrm{new}}).
\end{aligned}$$

This concludes the proof. □

Lemma B.10. *Work on the finite prompt–response domain with a full-support reference policy, and let $\Pi_\Theta \subseteq \Pi$ be a class of strictly positive policies. Let $g(x, y)$ be finite for every (x, y).*

Let $\pi_{\mathrm{old}} := \pi_{\theta_{\mathrm{old}}} \in \Pi_\Theta$ and $\eta, \beta > 0$. For each prompt x and probability vector $v \in \Delta(\mathcal{Y})$, define

$$\begin{aligned}
f_x(v) &:= \langle g(x, \cdot), v \rangle - \eta \mathrm{KL}(v\|\pi_{\mathrm{old}}(\cdot \mid x)) - \beta \mathrm{KL}(v\|\pi_{\mathrm{ref}}(\cdot \mid x)), \\
\pi^\dagger_{\mathrm{new}}(\cdot \mid x) &:= \arg\max_{v \in \Delta(\mathcal{Y})} f_x(v).
\end{aligned}$$

The comparator $\pi' \in \Pi$ is any policy satisfying the bounds below. In the remaining formulas, inner products and KL divergences are evaluated pointwise in x before taking expectation over $x \sim \mathcal{D}$.

Let $\pi_{\theta_{\mathrm{new}}} \in \Pi_\Theta$ be the implemented parameterized primal update corresponding to $\pi^\dagger_{\mathrm{new}}$.

For nonnegative budgets $\varepsilon_{1,t}, \varepsilon_{2,t}$, assume

$$\begin{aligned}
\mathbb{E}_x \mathrm{KL}(\pi_{\theta_{\mathrm{new}}}\|\pi^\dagger_{\mathrm{new}}) &\leq \varepsilon_{1,t}, \\
\mathbb{E}_x \left[\mathrm{KL}(\pi'\|\pi_{\theta_{\mathrm{new}}}) - \mathrm{KL}(\pi'\|\pi^\dagger_{\mathrm{new}}) \right] &\leq \varepsilon_{2,t}.
\end{aligned}$$

These are the two error bounds in Assumption 3.6 for this update and comparator.

Then the following inequality holds for that comparator π':

$$\begin{aligned}
&\mathbb{E}_{x \sim \mathcal{D}} \left[\left\langle g - \eta \nabla_{\pi_{\theta_{\mathrm{new}}}} \mathrm{KL}(\pi_{\theta_{\mathrm{new}}}\|\pi_{\mathrm{old}}) - \beta \nabla_{\pi_{\theta_{\mathrm{new}}}} \mathrm{KL}(\pi_{\theta_{\mathrm{new}}}\|\pi_{\mathrm{ref}}), \pi_{\theta_{\mathrm{new}}} - \pi' \right\rangle \right] \\
&\quad \geq -\mathrm{gap}_t.
\end{aligned}$$

where $\mathrm{gap}_t = (\eta + \beta)(\varepsilon_{1,t} + \varepsilon_{2,t})$.

Proof. Fix x and write $q(\cdot|x) = \pi_{\theta_{\text{new}}}(\cdot|x)$. Define

$$H^\dagger(x,\cdot) := g(x,\cdot) - \eta\nabla_{\pi^\dagger_{\text{new}}}\mathrm{KL}(\pi^\dagger_{\text{new}}(\cdot|x)\|\pi_{\text{old}}(\cdot|x)) - \beta\nabla_{\pi^\dagger_{\text{new}}}\mathrm{KL}(\pi^\dagger_{\text{new}}(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)).$$

Since π_{old} and π_{ref} have full support, the KL-mirror update $\pi^\dagger_{\text{new}}$ is an interior optimizer over the probability simplex. The KKT condition therefore implies that $H^\dagger(x,y)$ is constant over y for each fixed x. Hence, for any distributions $u(\cdot|x)$ and $v(\cdot|x)$,

$$\left\langle H^\dagger(x,\cdot), u(\cdot|x) - v(\cdot|x)\right\rangle = 0.$$

Using

$$\nabla_q\mathrm{KL}(q\|\pi_{\text{old}}) - \nabla_{\pi^\dagger_{\text{new}}}\mathrm{KL}(\pi^\dagger_{\text{new}}\|\pi_{\text{old}}) = \log\frac{q}{\pi^\dagger_{\text{new}}},$$

and the same identity with π_{ref} in place of π_{old}, we obtain

$$\begin{aligned}
&\langle g - \eta\nabla_q\mathrm{KL}(q\|\pi_{\text{old}}) - \beta\nabla_q\mathrm{KL}(q\|\pi_{\text{ref}}), q - \pi'\rangle \\
&= \left\langle H^\dagger(x,\cdot), q - \pi'\right\rangle + (\eta+\beta)\left\langle \log\frac{\pi^\dagger_{\text{new}}}{q}, q - \pi'\right\rangle \\
&= (\eta+\beta)\left\langle \log\frac{\pi^\dagger_{\text{new}}}{q}, q - \pi'\right\rangle.
\end{aligned}$$

Expanding the KL divergences gives the identity

$$\begin{aligned}
\left\langle \log\frac{\pi^\dagger_{\text{new}}}{q}, q - \pi'\right\rangle = &-\mathrm{KL}(q\|\pi^\dagger_{\text{new}}) \\
&- \left[\mathrm{KL}(\pi'\|q) - \mathrm{KL}(\pi'\|\pi^\dagger_{\text{new}})\right].
\end{aligned}$$

Taking expectation and applying the two error bounds gives $\mathbb{E}_x\langle\nabla f_x(q), q - \pi'\rangle \geq -(\eta+\beta)(\varepsilon_{1,t} + \varepsilon_{2,t})$.

We define

$$\mathrm{gap}_t = (\eta+\beta)(\varepsilon_{1,t} + \varepsilon_{2,t}). \tag{20}$$

This proves the desired inequality. □

Remark B.11 (Sufficient conditions for the approximation errors). Write $q = \pi_{\theta_{\text{new}}}$ and $\pi^\dagger = \pi^\dagger_{\text{new}}$. Here $\mathbb{E}_p$ denotes expectation under the joint distribution $\mathcal{D}(x)p(y \mid x)$, and $\chi^2(\pi'\|q) := \mathbb{E}_q[(\pi'/q - 1)^2]$. If $\|\log(q/\pi^\dagger)\|_\infty \leq \varepsilon_\infty$, both error budgets may be chosen as ε_∞. Alternatively, Cauchy–Schwarz gives

$$\left|\mathbb{E}_{\pi'}\left[\log\frac{\pi^\dagger}{q}\right]\right| \leq \sqrt{1+\chi^2(\pi'\|q)}\left(\mathbb{E}_q\left[\log^2\frac{\pi^\dagger}{q}\right]\right)^{1/2}.$$

Thus, if $e := (\mathbb{E}_q[\log^2(\pi^\dagger/q)])^{1/2}$ and $\chi^2(\pi'\|q) \leq C$, one may choose $\varepsilon_{1,t} = e$ and $\varepsilon_{2,t} = \sqrt{1+C}\,e$. The first follows from $\mathrm{KL}(q\|\pi^\dagger) \leq e$; the second follows from the displayed change-of-measure bound. These conditions quantify both the update error and the comparator's coverage under q.

Corollary B.12. *Under the hypotheses of Lemma B.10, the implemented update and its comparator $\pi' \in \Pi$ satisfy the following three-point inequality. All KL divergences are finite because π_{old}, $\pi_{\theta_{\text{new}}}$, and π_{ref} are strictly positive.*

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[\langle g, \pi_{\theta_{\text{new}}} - \pi'\rangle - \beta\left(\mathrm{KL}(\pi_{\theta_{\text{new}}}\|\pi_{\text{ref}}) - \mathrm{KL}(\pi'\|\pi_{\text{ref}})\right)\right] \\
&\geq \mathbb{E}_{x\sim\mathcal{D}}\left[\eta\left(-\mathrm{KL}(\pi'\|\pi_{\text{old}}) + \mathrm{KL}(\pi'\|\pi_{\theta_{\text{new}}}) + \mathrm{KL}(\pi_{\theta_{\text{new}}}\|\pi_{\text{old}})\right) + \beta\mathrm{KL}(\pi'\|\pi_{\theta_{\text{new}}})\right] \\
&\quad - \mathrm{gap}_t,
\end{aligned}$$

where gap_t *is defined in Equation* (20)*.*

Proof. By Lemma B.10, we have

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle g-\eta\nabla_{\pi_{\theta_{\text{new}}}}\mathrm{KL}(\pi_{\theta_{\text{new}}}\|\pi_{\text{old}})-\beta\nabla_{\pi_{\theta_{\text{new}}}}\mathrm{KL}(\pi_{\theta_{\text{new}}}\|\pi_{\text{ref}}),\pi_{\theta_{\text{new}}}-\pi'\right\rangle\right] \geq -\mathrm{gap}_t.$$

Applying Lemma B.7 with f set to the negative entropy function, we have

$$\left\langle\nabla_{\pi_{\theta_{\text{new}}}}\mathrm{KL}(\pi_{\theta_{\text{new}}}\|\pi_{\text{old}}),\pi'-\pi_{\theta_{\text{new}}}\right\rangle=\mathrm{KL}(\pi'\|\pi_{\text{old}})-\mathrm{KL}(\pi'\|\pi_{\theta_{\text{new}}})-\mathrm{KL}(\pi_{\theta_{\text{new}}}\|\pi_{\text{old}}),$$
$$\left\langle\nabla_{\pi_{\theta_{\text{new}}}}\mathrm{KL}(\pi_{\theta_{\text{new}}}\|\pi_{\text{ref}}),\pi'-\pi_{\theta_{\text{new}}}\right\rangle=\mathrm{KL}(\pi'\|\pi_{\text{ref}})-\mathrm{KL}(\pi'\|\pi_{\theta_{\text{new}}})-\mathrm{KL}(\pi_{\theta_{\text{new}}}\|\pi_{\text{ref}}),$$

The identities also hold when π' has zero coordinates, by continuity with the convention $0\log 0=0$. Substituting the above identities into the previous inequality concludes the proof. □

Lemma B.13 (Reference-weighted log ratios)**.** *Let p, q, π_{ref} be policies on the finite prompt–response domain, with $\pi_{\text{ref}}(y \mid x) > 0$ for every (x, y). Let $\|z\|^2_{\pi_{\text{ref}}} = \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_{\text{ref}}(\cdot|x)}z(x,y)^2$, and $C_{\pi_{\text{ref}}}z = z - \mathbb{E}_{y\sim\pi_{\text{ref}}(\cdot|x)}z(x,y)$. For $L \geq 1$, if $p/\pi_{\text{ref}}, q/\pi_{\text{ref}} \geq 1/L$ pointwise, then*

$$\|C_{\pi_{\text{ref}}}\log(p/q)\|^2_{\pi_{\text{ref}}} \leq 2L\mathbb{E}_x\mathrm{KL}(p(\cdot \mid x)\|q(\cdot \mid x)).$$

The same bound holds with the KL orientation reversed.

Proof. Fix a prompt, write $z = \log(p/q)$, and define $\psi(s) = \log\sum_y q(y)e^{sz(y)}$ and $p_s = p^s q^{1-s}e^{-\psi(s)}$ for $0 \leq s \leq 1$. Hölder's inequality implies $e^{\psi(s)} \leq 1$, so $p_s \geq \pi_{\text{ref}}/L$. Thus $\psi''(s) = \mathrm{Var}_{p_s}(z) \geq L^{-1}\mathrm{Var}_{\pi_{\text{ref}}}(z)$, using $\mathrm{Var}_p(z) = \inf_c \mathbb{E}_p(z-c)^2$. Since $\psi(0) = \psi(1) = 0$, integration by parts gives

$$\mathrm{KL}(p\|q) = \psi'(1) = \int_0^1 s\psi''(s)\,ds \geq \mathrm{Var}_{\pi_{\text{ref}}}(z)/(2L).$$

Average over prompts and exchange p, q for the reverse orientation. □

C Proof of Theorem 3.4

All prompts in this appendix belong to $\mathrm{supp}(\mathcal{D})$. An unbold λ in S_λ or $\mathcal{L}(\pi,\lambda)$ denotes the same multiplier vector $\boldsymbol{\lambda}$, where the coordinates are λ_j.

The aggregated reward in Equation (1) combines the soft objectives and the constraint rewards:

$$S_\lambda(x,y) = \sum_{k\in\mathcal{S}} w_k R_k(x,y) + \sum_{j\in\mathcal{H}} \lambda_j R_j(x,y),$$

Define the corresponding value function as

$$V^\pi_{S_\lambda}(x) := \mathbb{E}_{y\sim\pi(\cdot|x)}\big[S_\lambda(x,y)\big]. \tag{21}$$

The Lagrangian associated with the constrained MO-RLHF problem can then be written as

$$\begin{aligned}\mathcal{L}(\pi,\lambda) &= \mathcal{J}_0(\pi) + \sum_{j\in\mathcal{H}}\lambda_j\mathbb{E}_{x\sim\mathcal{D},y\sim\pi(\cdot|x)}\big[R_j(x,y)\big]\\ &= \mathbb{E}_{x\sim\mathcal{D}}\Big[V^\pi_{S_\lambda}(x) - \beta\mathrm{KL}\big(\pi(\cdot \mid x)\|\pi_{\text{ref}}(\cdot \mid x)\big)\Big].\end{aligned}$$

The Lagrangian problem is

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^{|\mathcal{H}|}_+}\max_{\pi\in\Pi}\mathcal{L}(\pi,\boldsymbol{\lambda}).$$

Note that Π is a compact set (the finite product of probability simplices), and the Lagrangian is continuous in the policy, so the primal maximization attains an optimum. Moreover, by Lemma B.4 and Equation (3), there exists an optimal dual variable $\boldsymbol{\lambda}^\star \in \Lambda$. Under Assumption 3.1, strong duality holds and the optimal saddle-point $(\pi^\star, \boldsymbol{\lambda}^\star)$ exists. Fix an optimal primal policy $\pi^\star$ that appears in such a saddle point and define

$$\Lambda^\star := \{\boldsymbol{\lambda}\in\Lambda :\ (\pi^\star,\boldsymbol{\lambda}) \text{ is an optimal saddle point}\}.$$

The set $\Lambda^\star$ is nonempty and compact. In the one-step analysis below, $\boldsymbol{\lambda}^\star$ denotes an arbitrary element of $\Lambda^\star$.

Since $\pi^\star = \operatorname{argmax}_\pi \mathcal{L}(\pi, \boldsymbol{\lambda}^\star)$, we have $\mathcal{L}(\pi^\star, \boldsymbol{\lambda}^\star) \geq \mathcal{L}(\pi, \boldsymbol{\lambda}^\star)$ for any $\pi \in \Pi$. Similarly, since $\boldsymbol{\lambda}^\star \in \operatorname{argmin}_{\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{H}|}} \mathcal{L}(\pi^\star, \boldsymbol{\lambda})$, we have $\mathcal{L}(\pi^\star, \boldsymbol{\lambda}) \geq \mathcal{L}(\pi^\star, \boldsymbol{\lambda}^\star)$ for any $\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{H}|}$. Combining these two inequalities together, for any $\pi \in \Pi$ and $\boldsymbol{\lambda} \in \mathbb{R}_+^{|\mathcal{H}|}$, we have

$$\mathcal{L}(\pi^\star, \boldsymbol{\lambda}) - \mathcal{L}(\pi, \boldsymbol{\lambda}^\star) = \underbrace{\mathcal{L}(\pi^\star, \boldsymbol{\lambda}) - \mathcal{L}(\pi^\star, \boldsymbol{\lambda}^\star)}_{\geq 0} + \underbrace{\mathcal{L}(\pi^\star, \boldsymbol{\lambda}^\star) - \mathcal{L}(\pi, \boldsymbol{\lambda}^\star)}_{\geq 0} \geq 0 \tag{22}$$

Let $\pi = \pi_t$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t$ and substituting the definition of $\mathcal{L}(\pi, \boldsymbol{\lambda})$ into the LHS of the above inequality, we have

$$\begin{aligned}
\mathcal{L}(\pi^\star, \boldsymbol{\lambda}_t) - \mathcal{L}(\pi_t, \boldsymbol{\lambda}^\star) =& \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi^\star}_{S_{\boldsymbol{\lambda}_t}}(x) - \beta \mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] \\
&- \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{S_{\boldsymbol{\lambda}^\star}}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] \\
=& \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi^\star}_{S_{\boldsymbol{\lambda}_t}}(x) - \beta \mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] \\
&\underbrace{- \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{S_{\boldsymbol{\lambda}_t}}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right]}_{\mathrm{A}} \\
&+ \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{S_{\boldsymbol{\lambda}_t}}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] \\
&\underbrace{- \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{S_{\boldsymbol{\lambda}^\star}}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right]}_{\mathrm{B}}
\end{aligned} \tag{23}$$

C.1 Upper bound of term A

We can rewrite term A as:

$$\begin{aligned}
\mathrm{A} =& \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi^\star}_{S_{\boldsymbol{\lambda}_t}}(x) - \beta \mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{S_{\boldsymbol{\lambda}_t}}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] \\
=& \mathbb{E}_{x\sim\mathcal{D}}\left[\left(V^{\pi^\star}_{S_{\boldsymbol{\lambda}_t}}(x) - V^{\pi_t}_{S_{\boldsymbol{\lambda}_t}}(x)\right) - \beta\left(\mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right)\right] \\
\overset{(a)}{=}& \mathbb{E}_{x\sim\mathcal{D}}\left[\langle \pi^\star(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot)\rangle - \beta\left(\mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right)\right] \\
\overset{(b)}{=}& \mathbb{E}_{x\sim\mathcal{D}}\left[\langle \pi^\star(\cdot|x) - \hat{\pi}_{t+1}(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot)\rangle - \beta\left(\mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right)\right. \\
&\quad + \langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_{t-1}}(x,\cdot)\rangle - \beta\left(\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right) \\
&\quad \left. + \langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot) - S_{\boldsymbol{\lambda}_{t-1}}(x,\cdot)\rangle\right]
\end{aligned} \tag{24}$$

where (a) is because the action space is discrete and $V^\pi_{S_{\boldsymbol{\lambda}_t}}(x) = \sum_y \pi(y|x) S_{\boldsymbol{\lambda}_t}(x,y) = \langle \pi(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot)\rangle$ for any $\pi \in \Pi$, and (b) is because adding and subtracting the same term keeps the equality. Recall the $\hat{\pi}_{t+1}$ update shown in the optimistic policy gradient primal-dual method Algorithm 2 ,

$$\hat{\pi}_{t+1} = \arg\max_\pi \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}\left[S_{\boldsymbol{\lambda}_t}(x,y)\right] - \beta \mathrm{KL}\left(\pi(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)\right) - \eta_\theta \mathrm{KL}\left(\pi(\cdot|x)\|\hat{\pi}_t(\cdot|x)\right)\right],$$

Since the optimization decouples across x, for fixed x we can write

$$\begin{aligned}
\hat{\pi}_{t+1}(\cdot|x) =& \operatorname{argmax}_\pi \mathbb{E}_{y\sim\pi(\cdot|x)}\left[S_{\boldsymbol{\lambda}_t}(x,y)\right] - \beta \mathrm{KL}\left(\pi(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)\right) - \eta_\theta \mathrm{KL}\left(\pi(\cdot|x)\|\hat{\pi}_t(\cdot|x)\right) \\
=& \operatorname{argmax}_\pi \langle \pi(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot)\rangle - \beta \mathrm{KL}\left(\pi(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)\right) - \eta_\theta \mathrm{KL}\left(\pi(\cdot|x)\|\hat{\pi}_t(\cdot|x)\right)
\end{aligned}$$

Initialize $\pi_{-1} = \hat{\pi}_0 = \pi_{\mathrm{ref}}$ and $\boldsymbol{\lambda}_{-1} = \hat{\boldsymbol{\lambda}}_0 \in \Lambda$, as in Algorithm 2. The reference has full support. The KL-mirror predictor and corrector have the closed forms $\pi_t(y|x) \propto \hat{\pi}_t(y|x)^{\frac{\eta_\theta}{\eta_\theta+\beta}} \pi_{\mathrm{ref}}(y|x)^{\frac{\beta}{\eta_\theta+\beta}} \exp\left(\frac{S_{\lambda_{t-1}}(x,y)}{\eta_\theta+\beta}\right)$ and $\hat{\pi}_{t+1}(y|x) \propto \hat{\pi}_t(y|x)^{\frac{\eta_\theta}{\eta_\theta+\beta}} \pi_{\mathrm{ref}}(y|x)^{\frac{\beta}{\eta_\theta+\beta}} \exp\left(\frac{S_{\lambda_t}(x,y)}{\eta_\theta+\beta}\right)$ for any t. By iteration, we have $\hat{\pi}_t$ and π_t have the same support as π_{ref} for any t. Therefore, $\mathrm{supp}(\pi_t) \cap \mathrm{supp}(\pi_{\mathrm{ref}}) = \mathrm{supp}(\hat{\pi}_t) \cap \mathrm{supp}(\pi_{\mathrm{ref}}) = \mathrm{supp}(\pi_{\mathrm{ref}})$.

Since π_{ref} spans the action space by Assumption 3.3, $\pi^\star$ is covered by π_{ref}. Therefore $\text{supp}\,(\pi^\star) \subseteq \text{supp}\,(\pi_t) \cap \text{supp}\,(\pi_{\text{ref}})$. Using Lemma B.9 and letting $\eta = \eta_\theta$, $g = S_{\boldsymbol{\lambda}_t}(x,\cdot)$, $\pi_{\text{old}} = \hat{\pi}_t(\cdot|x)$, $\pi_{\text{new}} = \hat{\pi}_{t+1}(\cdot|x)$, and $\pi' = \pi^\star(\cdot|x)$, we have

$$\begin{aligned}
&\langle S_{\boldsymbol{\lambda}_t}(x,\cdot), \hat{\pi}_{t+1}(\cdot|x) - \pi^\star(\cdot|x)\rangle - \beta\left(\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)) - \mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\text{ref}}(\cdot|x))\right)\\
&\geq \eta_\theta\Big(-\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + \mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x))\\
&\qquad + \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x))\Big) + \beta\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)).
\end{aligned}$$

Putting a negative sign on both sides, we have

$$\begin{aligned}
&\langle \pi^\star(\cdot|x) - \hat{\pi}_{t+1}(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot)\rangle - \beta\left(\mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)) - \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_{\text{ref}}(\cdot|x))\right)\\
&\leq \eta_\theta\Big(\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - \mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x))\\
&\qquad - \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x))\Big) - \beta\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x))\\
&= \eta_\theta\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x))\\
&\quad - \eta_\theta\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x))
\end{aligned} \tag{25}$$

Similarly, the update of π_t in the optimistic update is

$$\pi_t = \arg\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}\left[S_{\boldsymbol{\lambda}_{t-1}}(x,y)\right] - \beta\mathrm{KL}\left(\pi(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)\right) - \eta_\theta\mathrm{KL}\left(\pi(\cdot|x)\|\hat{\pi}_t(\cdot|x)\right)\right]$$

let $\eta = \eta_\theta$, $g = S_{\boldsymbol{\lambda}_{t-1}}(x,\cdot)$, $\pi_{\text{old}} = \hat{\pi}_t(\cdot|x)$, $\pi_{\text{new}} = \pi_t(\cdot|x)$, and $\pi' = \hat{\pi}_{t+1}(\cdot|x)$ in Lemma B.9, we have

$$\begin{aligned}
&\langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_{t-1}}(x,\cdot)\rangle - \beta\left(\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)) - \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\text{ref}}(\cdot|x))\right)\\
&\leq \eta_\theta\Big(\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))\\
&\qquad - \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x))\Big) - \beta\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))\\
&= \eta_\theta\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) - \eta_\theta\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)).
\end{aligned} \tag{26}$$

Let $C > 0$ be a constant. For the last term in the RHS of Equation (24), we derive the upper bound as

$$\begin{aligned}
&\langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot) - S_{\boldsymbol{\lambda}_{t-1}}(x,\cdot)\rangle\\
&= \left\langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), \sum_{j\in\mathcal{H}}(\lambda_{t,j} - \lambda_{t-1,j})R_j(x,\cdot)\right\rangle\\
&\overset{(a)}{\leq} \|\hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x)\|_1 \left\|\sum_{j\in\mathcal{H}}(\lambda_{t,j} - \lambda_{t-1,j})R_j(x,\cdot)\right\|_\infty\\
&\overset{(b)}{\leq} \|\hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x)\|_1\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1 R_{\max}\\
&\leq \sqrt{2\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1 R_{\max}\\
&\quad \text{(By Pinsker's inequality in Lemma B.2)}\\
&\leq C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) + \frac{R_{\max}^2}{2C}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1^2\\
&\quad \text{(By AM-GM inequality } \frac{x^2}{2C} + \frac{y^2C}{2} \geq xy \text{ with } C>0)\\
&\leq C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) + \frac{|\mathcal{H}|R_{\max}^2}{2C}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_2^2\\
&\quad \text{(By } \|\mathbf{x}\|_1^2 \leq d\|\mathbf{x}\|_2^2, \forall \mathbf{x}\in\mathbb{R}^d)\\
&\overset{(c)}{\leq} C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))\\
&\quad + \frac{|\mathcal{H}|R_{\max}^2}{C}\left(\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2\right),
\end{aligned} \tag{27}$$

where (a) is by Hölder's inequality shown in Lemma B.1 and letting $f(\cdot) = \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x)$, $g(\cdot) = \sum_{j\in\mathcal{H}}(\lambda_{t,j} - \lambda_{t-1,j})R_j(x,\cdot)$, $p = 1$, and $q = \infty$. (b) is by

$$\begin{aligned}
\left\|\sum_{j\in\mathcal{H}}(\lambda_{t,j} - \lambda_{t-1,j})R_j(x,\cdot)\right\|_\infty \leq& \sum_{j\in\mathcal{H}}\|(\lambda_{t,j} - \lambda_{t-1,j})R_j(x,\cdot)\|_\infty \\
&\text{(by triangle inequality)} \\
=& \sum_{j\in\mathcal{H}}|\lambda_{t,j} - \lambda_{t-1,j}|\,\|R_j(x,\cdot)\|_\infty \\
\leq& \left(\sum_{j\in\mathcal{H}}|\lambda_{t,j} - \lambda_{t-1,j}|\right)\max_j\|R_j(x,\cdot)\|_\infty \\
=&\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1\max_j\|R_j(x,\cdot)\|_\infty \\
&\text{(by definition of 1-norm)} \\
\leq&\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1 R_{\max} \\
&\text{(by Assumption 3.2 that } |R_j(x,y)| \leq R_{\max} \text{ for any} \\
&j \in \mathcal{S}\cup\mathcal{H} \text{ and } (x,y) \text{ pair).}
\end{aligned}$$

(c) is because

$$\begin{aligned}
\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 =&\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t + \hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 \\
=&\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 + 2\langle\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t, \hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\rangle \\
\leq&2\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + 2\|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 \\
&\text{(by Young's inequality with } p = q = 2, \text{ i.e., } \langle\mathbf{x},\mathbf{y}\rangle \leq \frac{1}{2}(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2))
\end{aligned}$$

Substituting Equation (25), Equation (26), and Equation (27) into the RHS of Equation (24), we have

$$\begin{aligned}
\mathrm{A} \leq&\mathbb{E}_{x\sim\mathcal{D}}\,[\eta_\theta\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)) - \eta_\theta\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x)) \\
&+\eta_\theta\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) - \eta_\theta\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) \\
&+C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) + \frac{|\mathcal{H}|R_{\max}^2}{C}\left(\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2\right)\Big] \\
=&\mathbb{E}_{x\sim\mathcal{D}}\,[\eta_\theta\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)) \\
&-(\eta_\theta+\beta)\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) - \eta_\theta\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) \\
&+C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))] + \frac{|\mathcal{H}|R_{\max}^2}{C}\left(\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2\right)
\end{aligned} \tag{28}$$

C.2 Upper bound of term B

Similarly, we rewrite the term B as

$$\begin{aligned}
\mathrm{B} =&\mathbb{E}_{x\sim\mathcal{D}}\left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - \beta\mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{S_{\boldsymbol{\lambda}^\star}}^{\pi_t}(x) - \beta\mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] \\
=&\mathbb{E}_{x\sim\mathcal{D}}\left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - V_{S_{\boldsymbol{\lambda}^\star}}^{\pi_t}(x)\right] \\
=&\mathbb{E}_{x\sim\mathcal{D}}\left[\left(V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - V_{S_{\hat{\boldsymbol{\lambda}}_{t+1}}}^{\pi_t}(x)\right) + \left(V_{S_{\hat{\boldsymbol{\lambda}}_{t+1}}}^{\pi_t}(x) - V_{S_{\boldsymbol{\lambda}^\star}}^{\pi_t}(x)\right)\right. \\
&\left.- \left(V_{S_{\boldsymbol{\lambda}_t}}^{\pi_{t-1}}(x) - V_{S_{\hat{\boldsymbol{\lambda}}_{t+1}}}^{\pi_{t-1}}(x)\right) + \left(V_{S_{\boldsymbol{\lambda}_t}}^{\pi_{t-1}}(x) - V_{S_{\hat{\boldsymbol{\lambda}}_{t+1}}}^{\pi_{t-1}}(x)\right)\right]
\end{aligned} \tag{29}$$

Define $R_{\mathcal{H}} = \left[R_{j_1}, R_{j_2}, \cdots, R_{j_{|\mathcal{H}|}}\right] \in \mathbb{R}^{|\mathcal{H}|}$, where $j_1 < j_2 < \cdots < j_{|\mathcal{H}|}$, and $j_k \in \mathcal{H}$ for any integer $1 \leq k \leq |\mathcal{H}|$. Define $V_{R_{\mathcal{H}}}^{\pi}(x) = \mathbb{E}_{y\sim\pi(\cdot|x)}R_{\mathcal{H}}(x,y) \in \mathbb{R}^{|\mathcal{H}|}$. By the definition of $V_{S_\lambda}^{\pi}(x)$,

we have

$$
\begin{aligned}
V_{S_\lambda}^{\pi}(x) =& \mathbb{E}_{y\sim\pi(\cdot|x)} S_\lambda(x,y) \quad \text{(by definition of } V_{S_\lambda}^{\pi}(x)\text{)} \\
=& \mathbb{E}_{y\sim\pi(\cdot|x)} \left[\sum_{j\in\mathcal{S}} w_j R_j(x,y) + \sum_{j\in\mathcal{H}} \lambda_j R_j(x,y)\right] \quad \text{(by definition of } S_\lambda\text{)} \\
=& \mathbb{E}_{y\sim\pi(\cdot|x)} \left[\sum_{j\in\mathcal{S}} w_j R_j(x,y)\right] + \sum_{j\in\mathcal{H}} \lambda_j \mathbb{E}_{y\sim\pi(\cdot|x)} \left[R_j(x,y)\right] \\
=& \mathbb{E}_{y\sim\pi(\cdot|x)} \left[\sum_{j\in\mathcal{S}} w_j R_j(x,y)\right] + \lambda^T V_{R_\mathcal{H}}^{\pi}(x) \quad \text{(by definition of } V_{R_\mathcal{H}}^{\pi}\text{)}
\end{aligned}
$$

Plugging the above expression of $V_{S_\lambda}^{\pi}(x)$ into term B, we have

$$
\begin{aligned}
\mathrm{B} =& \mathbb{E}_{x\sim\mathcal{D}}\Big[\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T V_{R_\mathcal{H}}^{\pi_t}(x) + \left(\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star\right)^T V_{R_\mathcal{H}}^{\pi_t}(x) \\
& - \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T V_{R_\mathcal{H}}^{\pi_{t-1}}(x) + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\Big] \\
=& \mathbb{E}_{x\sim\mathcal{D}}\Big[\left(\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star\right)^T V_{R_\mathcal{H}}^{\pi_t}(x) + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T V_{R_\mathcal{H}}^{\pi_{t-1}}(x) \\
& + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(V_{R_\mathcal{H}}^{\pi_t}(x) - V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\right)\Big] \\
=& \left(\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star\right)^T \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_\mathcal{H}}^{\pi_t}(x)\right] + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\right] \\
& + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \Big(\mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_\mathcal{H}}^{\pi_t}(x)\right] \\
& \qquad - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\right]\Big)
\end{aligned} \tag{30}
$$

By Lemma B.4 and Equation (3), $\boldsymbol{\lambda}^\star \in \Lambda$. Recall the dual actual update in Algorithm 2. In vector form, it is

$$
\begin{aligned}
\hat{\boldsymbol{\lambda}}_{t+1} &= \arg\min_{\boldsymbol{\lambda}\in\Lambda} \boldsymbol{\lambda}^T \mathbb{E}_{x\sim\mathcal{D}}\big[V_{R_\mathcal{H}}^{\pi_t}(x)\big] + \eta_\lambda \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_t\|_2^2 \\
&= \arg\max_{\boldsymbol{\lambda}\in\Lambda} -\boldsymbol{\lambda}^T \mathbb{E}_{x\sim\mathcal{D}}\big[V_{R_\mathcal{H}}^{\pi_t}(x)\big] - \eta_\lambda \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_t\|_2^2.
\end{aligned}
$$

All dual iterates lie in Λ by construction. Let $\Omega = \Lambda$, $g = -\mathbb{E}_{x\sim\mathcal{D}}\big[V_{R_\mathcal{H}}^{\pi_t}(x)\big]$, $\eta = \eta_\lambda$, $h(\mathbf{x}) = \|\mathbf{x}\|_2^2$, $D_h(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2$, $\mathbf{x}' = \boldsymbol{\lambda}^\star$, $\mathbf{x}_{\text{new}} = \hat{\boldsymbol{\lambda}}_{t+1}$, $\mathbf{x}_{\text{old}} = \hat{\boldsymbol{\lambda}}_t$ in Lemma B.8, then we have

$$
\langle -\mathbb{E}_{x\sim\mathcal{D}}\big[V_{R_\mathcal{H}}^{\pi_t}(x)\big], \hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star\rangle \geq \eta_\lambda \left(-\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 + \|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2\right).
$$

Putting a negative sign on both sides, we have

$$
\langle \hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star, \mathbb{E}_{x\sim\mathcal{D}}\big[V_{R_\mathcal{H}}^{\pi_t}(x)\big]\rangle \leq \eta_\lambda \left(\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2\right). \tag{31}
$$

Similarly, the optimistic dual update in Algorithm 2 is

$$
\boldsymbol{\lambda}_t = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \boldsymbol{\lambda}^T \mathbb{E}_{x\sim\mathcal{D}}\big[V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\big] + \eta_\lambda \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_t\|_2^2,
$$

Applying Lemma B.8 by setting $g = -\mathbb{E}_{x\sim\mathcal{D}}\big[V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\big]$, $\eta = \eta_\lambda$, $h(\mathbf{x}) = \|\mathbf{x}\|_2^2$, $D_h(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2$, $\mathbf{x}' = \hat{\boldsymbol{\lambda}}_{t+1}$, $\mathbf{x}_{\text{new}} = \boldsymbol{\lambda}_t$, $\mathbf{x}_{\text{old}} = \hat{\boldsymbol{\lambda}}_t$, and $\Omega = \Lambda$, we have

$$
\langle \boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}, \mathbb{E}_{x\sim\mathcal{D}}\big[V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\big]\rangle \leq \eta_\lambda \left(\|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}_t\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2\right). \tag{32}
$$

We upper bound the last term of B as

$$
\begin{aligned}
&\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(\mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_t}(x)\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x)\right]\right) \\
&= \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(V_{R_{\mathcal{H}}}^{\pi_t}(x) - V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x)\right)\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(\langle \pi_t(\cdot|x), R_{\mathcal{H}}(x,\cdot)\rangle - \langle \pi_{t-1}(\cdot|x), R_{\mathcal{H}}(x,\cdot)\rangle\right)\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(by } V_{R_{\mathcal{H}}}^{\pi}(x) = \mathbb{E}_{y\sim\pi(\cdot|x)}\left[R_{\mathcal{H}}(x,y)\right] = \langle \pi(\cdot|x), R_{\mathcal{H}}(x,\cdot)\rangle) \\
&= \mathbb{E}_{x\sim\mathcal{D}}\left[\langle \pi_t(\cdot|x) - \pi_{t-1}(\cdot|x), \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T R_{\mathcal{H}}(x,\cdot)\rangle\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle \pi_t(\cdot|x) - \pi_{t-1}(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \hat{\lambda}_{t+1,j}) R_j(x,\cdot)\right\rangle\right]
\end{aligned}
$$

Define constants $C_1 > 0$ and $C_2 > 0$. Fixing x, we have

$$
\begin{aligned}
&\left\langle \pi_t(\cdot|x) - \pi_{t-1}(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \hat{\lambda}_{t+1,j}) R_j(x,\cdot)\right\rangle \\
&= \left\langle \pi_t(\cdot|x) - \hat{\pi}_t(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \hat{\lambda}_{t+1,j}) R_j(x,\cdot)\right\rangle + \left\langle \hat{\pi}_t(\cdot|x) - \pi_{t-1}(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \hat{\lambda}_{t+1,j}) R_j(x,\cdot)\right\rangle \\
&\le C_1 \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + \frac{|\mathcal{H}| R_{\max}^2}{2C_1}\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\
&\quad + C_2 \mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x)) + \frac{|\mathcal{H}| R_{\max}^2}{2C_2}\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\
&\quad \text{(By derivations of Equation (27))} \\
&= C_1 \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2 \mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x)) \\
&\quad + |\mathcal{H}| R_{\max}^2 \left(\frac{1}{2C_1} + \frac{1}{2C_2}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2
\end{aligned}
$$

Substituting the above inequality into $\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(\mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_t}(x)\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x)\right]\right)$, we have

$$
\begin{aligned}
&\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(\mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_t}(x)\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x)\right]\right) \\
&\le \mathbb{E}_{x\sim\mathcal{D}}\left[C_1 \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2 \mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\right] + |\mathcal{H}| R_{\max}^2 \left(\frac{1}{2C_1} + \frac{1}{2C_2}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2
\end{aligned} \tag{33}
$$

Combining Equation (31), Equation (32), and Equation (33), we can upper bound B as

$$
\begin{aligned}
\mathrm{B} \le& \eta_\lambda \left(\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2\right) \\
&+ \eta_\lambda \left(\|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}_t\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2\right) \\
&+ \mathbb{E}_{x\sim\mathcal{D}}\left[C_1 \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2 \mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\right] \\
&+ |\mathcal{H}| R_{\max}^2 \left(\frac{1}{2C_1} + \frac{1}{2C_2}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\
=& \eta_\lambda \left(\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2\right) \\
&+ \mathbb{E}_{x\sim\mathcal{D}}\left[C_1 \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2 \mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\right] \\
&+ |\mathcal{H}| R_{\max}^2 \left(\frac{1}{2C_1} + \frac{1}{2C_2}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2
\end{aligned} \tag{34}
$$

C.3 Combining A and B

Substituting Equation (28) and Equation (34) into the RHS of Equation (23), we get

$$
\begin{aligned}
&\mathcal{L}(\pi^\star, \boldsymbol{\lambda}_t) - \mathcal{L}(\pi_t, \boldsymbol{\lambda}^\star) \\
&\le \mathbb{E}_{x\sim\mathcal{D}}\left[\eta_\theta \mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))\right. \\
&\qquad \left. - \eta_\theta \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))\right] \\
&\quad + \frac{|\mathcal{H}|R_{\max}^2}{C}\left(\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2\right) \\
&\quad + \eta_\lambda\left(\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2\right) \\
&\quad + \mathbb{E}_{x\sim\mathcal{D}}\left[C_1\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2\mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\right] \\
&\quad + |\mathcal{H}|R_{\max}^2\left(\frac{1}{2C_1}+\frac{1}{2C_2}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\
&\le \mathbb{E}_{x\sim\mathcal{D}}\Big[\eta_\theta \mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)) \\
&\qquad - (\eta_\theta+\beta-C)\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) \\
&\qquad - (\eta_\theta - C_1)\,\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2\mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\Big] \\
&\quad - \left(\eta_\lambda - \frac{|\mathcal{H}|R_{\max}^2}{C}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 \\
&\quad + \frac{|\mathcal{H}|R_{\max}^2}{C}\|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 \\
&\quad + \eta_\lambda\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \eta_\lambda\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\
&\quad - \left(\eta_\lambda - |\mathcal{H}|R_{\max}^2\left(\frac{1}{2C_1}+\frac{1}{2C_2}\right)\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2,
\end{aligned}
$$

where terms sharing the same color can be combined. Recall Equation (22) that $\mathcal{L}(\pi^\star, \boldsymbol{\lambda}_t) - \mathcal{L}(\pi_t, \boldsymbol{\lambda}^\star) \ge 0$. Substituting this into the above equation and rearranging the equation, we have

$$
\begin{aligned}
&(\eta_\theta+\beta)\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x))\right] + (\eta_\theta+\beta-C)\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))\right] \\
&+ \eta_\lambda\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 + \left(\eta_\lambda - |\mathcal{H}|R_{\max}^2\left(\frac{1}{2C_1}+\frac{1}{2C_2}\right)\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\
&\le \eta_\theta\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x))\right] + C_2\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\right] \\
&\quad + \eta_\lambda\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \frac{|\mathcal{H}|R_{\max}^2}{C}\|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 \\
&\quad - (\eta_\theta - C_1)\,\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x))\right] - \left(\eta_\lambda - \frac{|\mathcal{H}|R_{\max}^2}{C}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2
\end{aligned}
\tag{35}
$$

Geometric contraction. We first express (35) as a descent inequality. We then bound the distance to the optimal multiplier set by the stationarity and complementary-slackness residuals, and control both residuals by the descent terms.

Write $K(p,q) := \mathbb{E}_x\mathrm{KL}(p(\cdot\mid x)\|q(\cdot\mid x))$, $a = \sqrt{|\mathcal{H}|}R_{\max} > 0$, and $\eta = 3a$. Set $C = C_1 = C_2 = a$ and $\eta_\theta = \eta_\lambda = \eta$ in (35). Define

$$
\begin{aligned}
\Phi_t = \eta K(\pi^\star, \hat{\pi}_t) + \eta\,\mathrm{dist}_2^2(\hat{\boldsymbol{\lambda}}_t, \Lambda^\star) \\
+ aK(\hat{\pi}_t, \pi_{t-1}) + a\|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2.
\end{aligned}
\tag{36}
$$

For this step abbreviate $q = \hat{\pi}_t$, $p = \pi_t$, $u = \hat{\pi}_{t+1}$, $v = \hat{\boldsymbol{\lambda}}_t$, $l = \boldsymbol{\lambda}_t$, and $w = \hat{\boldsymbol{\lambda}}_{t+1}$, and put

$$
k = K(\pi^\star, u), \quad U = K(u,p), \quad V = K(p,q), \quad b_1 = \|w-l\|_2^2, \quad d_1 = \|l-v\|_2^2.
$$

Choosing a closest optimal multiplier to v in the preceding inequality, and bounding the next distance to the set by distance to this same multiplier, gives

$$
\Phi_{t+1} + D_t \le \Phi_t, \qquad D_t = \beta k + (a+\beta)U + 2aV + ab_1 + 2ad_1. \tag{37}
$$

This also holds at $t = 0$, with the predictor initialization in Algorithm 2.

Let $R_s = \sum_{j\in\mathcal{S}} w_j R_j$, and $M = (1 + B_\lambda)R_{\max}$. The exact update has the normalized form

$$P(q, \boldsymbol{\lambda})(y \mid x) \propto \pi_{\mathrm{ref}}(y \mid x) \left(\frac{q(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}\right)^{\eta/(\eta+\beta)} \exp\left(\frac{S_{\boldsymbol{\lambda}}(x, y)}{\eta + \beta}\right).$$

Indeed, this follows from Lemma B.6. Writing $\hat{\pi}_t \propto \pi_{\mathrm{ref}} e^{z_t}$ with $z_0 = 0$ permits $z_{t+1} = (\eta z_t + S_{\boldsymbol{\lambda}_t})/(\eta + \beta)$. Since $|S_{\boldsymbol{\lambda}}| \le M$ on Λ, induction gives $\|z_t\|_\infty \le M/\beta$. The same bound holds for the predictor tilt and for the optimal tilt $S_{\boldsymbol{\lambda}^\star}/\beta$. Thus, for every $t \ge 0$ and $\nu \in \{\pi^\star, \pi_t, \hat{\pi}_t\}$,

$$L_0^{-1} \le \nu(y \mid x)/\pi_{\mathrm{ref}}(y \mid x) \le L_0, \qquad L_0 = \exp(2M/\beta). \tag{38}$$

This is a bound relative to the reference, derived from the exact updates. It does not require a minimum reference probability.

Using the norm and centering operator in Lemma B.13, set

$$A\boldsymbol{\lambda} = C_{\pi_{\mathrm{ref}}} \sum_{j\in\mathcal{H}} \lambda_j R_j, \qquad b = C_{\pi_{\mathrm{ref}}}[\beta \log(\pi^\star/\pi_{\mathrm{ref}}) - R_s], \qquad g^\star = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi^\star(\cdot|x)} R_{\mathcal{H}}(x, y).$$

Gibbs stationarity and complementary slackness characterize the optimal set exactly as

$$\Lambda^\star = \{\boldsymbol{\lambda} \in \Lambda : A\boldsymbol{\lambda} = b,\ (g^\star)^\top \boldsymbol{\lambda} = 0\}. \tag{39}$$

Necessity follows from optimality; conversely these equalities, together with primal feasibility and $\boldsymbol{\lambda} \ge 0$, imply the saddle inequalities. Define

$$\Gamma = \mathbb{E}_x \operatorname{Cov}_{y\sim\pi_{\mathrm{ref}}(\cdot|x)}(R_{\mathcal{H}}(x, y)), \qquad L_A = \sqrt{\|\Gamma\|_{\mathrm{op}}} \le a.$$

For $\boldsymbol{\lambda}^\circ \in \Lambda^\star$, $\|A\boldsymbol{\lambda} - b\|_{\pi_{\mathrm{ref}}} = \|\Gamma^{1/2}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^\circ)\|_2$. Hoffman's bound for this finite linear system (Hoffman, 1952) therefore supplies a finite constant $H_{\pi_{\mathrm{ref}}} > 0$ such that

$$\operatorname{dist}_2(\boldsymbol{\lambda}, \Lambda^\star) \le H_{\pi_{\mathrm{ref}}} \big(\|A\boldsymbol{\lambda} - b\|_{\pi_{\mathrm{ref}}} + |(g^\star)^\top \boldsymbol{\lambda}|\big), \quad \boldsymbol{\lambda} \in \Lambda. \tag{40}$$

The inequality residuals defining Λ vanish on this domain. This argument permits singular Γ and nonunique optimal multipliers.

Gibbs stationarity of the corrector gives $Al - b = C_{\pi_{\mathrm{ref}}}[\beta \log(u/\pi^\star) + \eta \log(u/q)]$. Split $\log(u/q) = \log(u/p) + \log(p/q)$ and apply Lemma B.13 with any $L \ge L_0$ to obtain

$$\|Aw - b\|_{\pi_{\mathrm{ref}}} \le \sqrt{L}\big(\beta\sqrt{2k} + \eta\sqrt{2U} + \eta\sqrt{2V}\big) + L_A\sqrt{b_1}. \tag{41}$$

Write $g(p) = \mathbb{E}_{x\sim\mathcal{D}, y\sim p(\cdot|x)} R_{\mathcal{H}}(x, y)$. Bounded rewards and Pinsker's inequality give $\|g(p) - g(q)\|_2 \le a\sqrt{2K(p, q)}$. The optimality condition for $w = \operatorname{Proj}_\Lambda(v - g(p)/(2\eta))$, tested at any $\boldsymbol{\lambda}^\star \in \Lambda^\star$, gives

$$\begin{aligned} 0 \le (g^\star)^\top w &\le 2B_\lambda\big(\|g(p) - g^\star\|_2 + 2\eta\|w - v\|_2\big) \\ &\le 2B_\lambda\big(a\sqrt{2k} + a\sqrt{2U} + 2\eta\sqrt{b_1} + 2\eta\sqrt{d_1}\big). \end{aligned}$$

Here $g^\star \ge 0$, $(g^\star)^\top \boldsymbol{\lambda}^\star = 0$, and the diameter of Λ is at most $2B_\lambda$. Weighted Cauchy–Schwarz, with the weights in D_t, now yields $\|Aw - b\|^2_{\pi_{\mathrm{ref}}} \le C_A(L)D_t$ and $|(g^\star)^\top w|^2 \le C_g D_t$, where

$$\begin{aligned} C_A(L) &= 2L\beta + \frac{2L\eta^2}{a + \beta} + \frac{L\eta^2}{a} + \frac{L_A^2}{a}, \\ C_g &= 8B_\lambda^2 a^2 \left(\frac{1}{\beta} + \frac{1}{a + \beta}\right) + \frac{24B_\lambda^2\eta^2}{a}, \\ \kappa(L) &= \frac{\eta}{\beta} + 2\eta H^2_{\pi_{\mathrm{ref}}}(C_A(L) + C_g) + \frac{a}{a + \beta} + 1. \end{aligned} \tag{42}$$

Combining (40) with these bounds and (36) proves

$$\Phi_{t+1} \le \kappa(L_0)D_t, \qquad \rho = \frac{\kappa(L_0)}{1 + \kappa(L_0)} \in (0, 1). \tag{43}$$

Combining $\Phi_{t+1} \le \kappa(L_0) D_t$ with $D_t \le \Phi_t - \Phi_{t+1}$ gives $(1+\kappa(L_0))\Phi_{t+1} \le \kappa(L_0)\Phi_t$. Induction yields $\Phi_t \le \rho^t \Phi_0$. The initialization $\pi_{-1} = \hat{\pi}_0 = \pi_{\mathrm{ref}}$ and $\boldsymbol{\lambda}_{-1} = \hat{\boldsymbol{\lambda}}_0$ gives

$$\Phi_0 = \eta K(\pi^\star, \pi_{\mathrm{ref}}) + \eta \operatorname{dist}_2^2(\hat{\boldsymbol{\lambda}}_0, \Lambda^\star).$$

Finally, the nonnegativity of the two memory terms implies

$$K(\pi^\star, \hat{\pi}_t) + \operatorname{dist}_2^2(\hat{\boldsymbol{\lambda}}_t, \Lambda^\star) \le \Phi_t/\eta \le \rho^t \Phi_0/\eta,$$

which is the claimed bound, including $t = 0$.

D Proof of Remark 3.5

In this section, we prove the policy-level equivalence stated in Remark 3.5. All prompts are in $\mathrm{supp}(\mathcal{D})$, so $\mathcal{D}(x) > 0$. Gradients and Fisher matrices are population quantities. We work in the full tabular logit space, so the primal projection in Algorithm 3 is inactive. The same argument also applies whenever the projection is policy-preserving. Because tabular logits are not identifiable, all statements below compare induced policies rather than raw parameter vectors: adding a state-dependent constant to all logits at a fixed state leaves the policy unchanged. Consider the unprojected NPG update

$$\theta_+ = \theta + \frac{1}{\eta_\theta + \beta} F(\theta)^\dagger \nabla_\theta \mathcal{L}(\pi_\theta, \lambda), \tag{44}$$

and the corresponding distribution-space KL-mirror update

$$\pi_+ = \arg\max_{\pi} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathbb{E}_{y\sim\pi(\cdot|x)} \left[S_\lambda(x,y) \right] - \beta \mathrm{KL}\left(\pi(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)\right) - \eta_\theta \mathrm{KL}\left(\pi(\cdot|x) \| \pi_\theta(\cdot|x)\right) \right], \tag{45}$$

where $\mathcal{L}(\pi_\theta, \lambda) = \mathbb{E}_{x\sim\mathcal{D}} \left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} S_\lambda(x,y) - \beta \mathrm{KL}(\pi_\theta(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right]$ and $S_\lambda(x,y) = \sum_{k\in\mathcal{S}} w_k R_k(x,y) + \sum_{j\in\mathcal{H}} \lambda_j R_j(x,y)$. If a projection is applied and satisfies

$$\pi_{\mathrm{Proj}_\Theta(\theta_+)} = \pi_{\theta_+},$$

then the same policy-level equivalence remains valid; otherwise exact equivalence is not claimed.

We first prove the equality of these two single-step policy updates, then apply it to the predictor and corrector of OPD.

Define

$$V_\lambda^\pi(x) := \mathbb{E}_{y\sim\pi(\cdot|x)} \Big[S_\lambda(x,y) - \beta \log \frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} \Big], \tag{46}$$

$$A_\lambda^\pi(x,y) := S_\lambda(x,y) - \beta \log \frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} - V_\lambda^\pi(x). \tag{47}$$

We can rewrite $\nabla_\theta \mathcal{L}(\pi_\theta, \lambda)$ as

$$\begin{aligned} \nabla_\theta \mathcal{L}(\pi_\theta, \lambda) =& \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[\left(S_\lambda(x,y) - \beta \log \tfrac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)} \right) \nabla_\theta \log \pi_\theta(y|x) \right] \\ =& \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[A_\lambda^{\pi_\theta}(x,y) \nabla_\theta \log \pi_\theta(y|x) \right], \end{aligned} \tag{48}$$

The first equality follows by differentiating the reward and KL terms: the additional constant $-\beta$ from the KL derivative vanishes because $\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} \nabla_\theta \log \pi_\theta(y \mid x) = 0$. The same identity removes the prompt-dependent baseline $V_\lambda^{\pi_\theta}(x)$ in the second equality.

The partial derivative with respect to $\theta_{x,y}$ is

$$\begin{aligned} \frac{\partial \mathcal{L}(\pi_\theta, \lambda)}{\partial \theta_{x,y}} &= \mathbb{E}_{x'\sim\mathcal{D}, y'\sim\pi_\theta(\cdot|x')} \left[A_\lambda^{\pi_\theta}(x',y') \frac{\partial}{\partial \theta_{x,y}} \log \pi_\theta(y'|x') \right] \\ &\overset{(a)}{=} \mathbb{E}_{y'\sim\pi_\theta(\cdot|x)} \left[A_\lambda^{\pi_\theta}(x,y') \left(\mathbb{I}\{y' = y\} - \pi_\theta(y|x) \right) \right] \mathcal{D}(x) \\ &= \mathcal{D}(x)\pi_\theta(y|x) \left(A_\lambda^{\pi_\theta}(x,y) - \mathbb{E}_{y'\sim\pi_\theta(\cdot|x)} [A_\lambda^{\pi_\theta}(x,y')] \right) \\ &\overset{(b)}{=} \mathcal{D}(x)\pi_\theta(y|x) A_\lambda^{\pi_\theta}(x,y), \end{aligned}$$

where (a) is because

$$\frac{\partial}{\partial \theta_{x,y}} \log \pi_\theta(y'|x') = \mathbb{I}\{x' = x\} \left(\mathbb{I}\{y' = y\} - \pi_\theta(y|x)\right),$$

and (b) is because $\mathbb{E}_{y' \sim \pi_\theta(\cdot|x)}[A_\lambda^{\pi_\theta}(x, y')] = 0$.

Let $\mathbf{e}_{x,y} \in \mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$ with only the position $\theta_{x,y}$ has element 1 and all other elements are 0, and $\pi_{\theta,x} \in \mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$ with only the positions $\theta_{x,y}$ has value $\pi_\theta(y|x)$ and all other elements are 0. Then we rewrite $F(\theta)$ as

$$\begin{aligned}
F(\theta) =& \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \left[\nabla_\theta \log \pi_\theta(y|x) \nabla_\theta \log \pi_\theta(y|x)^\top\right] \\
=& \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \left[\left(\mathbf{e}_{x,y} - \pi_{\theta,x}\right)\left(\mathbf{e}_{x,y} - \pi_{\theta,x}\right)^\top\right] \\
=& \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \left[\mathbf{e}_{x,y}\mathbf{e}_{x,y}^\top - \pi_{\theta,x}\mathbf{e}_{x,y}^\top - \mathbf{e}_{x,y}\pi_{\theta,x}^\top + \pi_{\theta,x}\pi_{\theta,x}^\top\right] \\
=& \mathbb{E}_{x \sim \mathcal{D}} \left[\mathrm{diag}(\pi_{\theta,x}) - \pi_{\theta,x}\pi_{\theta,x}^\top - \pi_{\theta,x}\pi_{\theta,x}^\top + \pi_{\theta,x}\pi_{\theta,x}^\top\right] \\
=& \mathbb{E}_{x \sim \mathcal{D}} \left[\mathrm{diag}(\pi_{\theta,x}) - \pi_{\theta,x}\pi_{\theta,x}^\top\right],
\end{aligned}$$

where the second equality uses the partial derivative of θ of $\log \pi_\theta(y|x)$.

We now characterize the natural-gradient direction $\mathbf{d} = F(\theta)^\dagger \nabla_\theta \mathcal{L}(\pi_\theta, \lambda)$, where $F(\theta)^\dagger$ is the Moore–Penrose pseudoinverse of $F(\theta)$. Each Fisher block has the constants as its null space, and the policy gradient sums to zero within each prompt block, so it lies in the Fisher range. In other words, $F(\theta)\mathbf{d} = \nabla_\theta \mathcal{L}(\pi_\theta, \lambda)$. Let $\bar{d}_x := \pi_{\theta,x}^\top \mathbf{d} = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[\mathbf{d}_{x,y}]$. Consider the (x, y)-th coordinate of the LHS, we have

$$\begin{aligned}
\left[F(\theta)\mathbf{d}\right]_{x,y} =& \mathcal{D}(x) \left[\left(\mathrm{diag}(\pi_{\theta,x}) - \pi_{\theta,x}\pi_{\theta,x}^\top\right)\mathbf{d}\right]_{x,y} \\
=& \mathcal{D}(x) \left[\mathrm{diag}(\pi_{\theta,x})\mathbf{d}\right]_{x,y} - \mathcal{D}(x) \left[\pi_{\theta,x}\pi_{\theta,x}^\top \mathbf{d}\right]_{x,y} \\
=& \mathcal{D}(x)\pi_\theta(y|x) \left(\mathbf{d}_{x,y} - \bar{d}_x\right),
\end{aligned}$$

Comparing with $\frac{\partial \mathcal{L}(\pi_\theta, \lambda)}{\partial \theta_{x,y}} = \mathcal{D}(x)\pi_\theta(y|x)A_\lambda^{\pi_\theta}(x, y)$, we have $\mathbf{d}_{x,y} = A_\lambda^{\pi_\theta}(x, y) + c(x)$, where $c(x) := \bar{d}_x$ is constant across responses at prompt x. Plugging $\mathbf{d}_{x,y}$ into Equation (44), we have $\theta_{+,x,y} = \theta_{x,y} + \frac{1}{\eta_\theta + \beta}\left(A_\lambda^{\pi_\theta}(x, y) + c(x)\right)$.

Since $c(x)$ does not depend on y, it cancels in the softmax normalization:

$$\begin{aligned}
\pi_{\theta_+}(y|x) &= \frac{\exp(\theta_{+,x,y})}{\sum_{y' \in \mathcal{Y}} \exp(\theta_{+,x,y'})} = \frac{\exp(\theta_{x,y}) \exp(\frac{1}{\eta_\theta + \beta}(A_\lambda^{\pi_\theta}(x, y) + c(x)))}{\sum_{y' \in \mathcal{Y}} \exp(\theta_{x,y'}) \exp(\frac{1}{\eta_\theta + \beta}(A_\lambda^{\pi_\theta}(x, y') + c(x)))} \\
&= \frac{\pi_\theta(y|x) \exp(\frac{1}{\eta_\theta + \beta} A_\lambda^{\pi_\theta}(x, y))}{\sum_{y' \in \mathcal{Y}} \pi_\theta(y'|x) \exp(\frac{1}{\eta_\theta + \beta} A_\lambda^{\pi_\theta}(x, y'))}.
\end{aligned}$$

That is,

$$\pi_{\theta_+}(y|x) \propto \pi_\theta(y|x) \exp\left(\frac{1}{\eta_\theta + \beta} A_\lambda^{\pi_\theta}(x, y)\right).$$

As shown in Equation (45)

$$\pi_+ = \arg\max_\pi \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi(\cdot|x)} \left[S_\lambda(x, y)\right] - \beta \mathrm{KL}\left(\pi(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)\right) - \eta_\theta \mathrm{KL}\left(\pi(\cdot|x) \| \pi_\theta(\cdot|x)\right)\right].$$

Solving this maximization over each simplex $\Delta(\mathcal{Y})$ gives the softmax solution

$$\begin{aligned}
\pi_+(y|x) \propto& \exp\left(\tfrac{1}{\eta_\theta + \beta}\left(S_\lambda(x, y) + \beta \log \pi_{\mathrm{ref}}(y|x) + \eta_\theta \log \pi_\theta(y|x)\right)\right) \\
\propto& \pi_\theta(y|x) \exp\left(\tfrac{1}{\eta_\theta + \beta}\left(S_\lambda(x, y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right)\right) \\
\propto& \pi_\theta(y|x) \exp\left(\frac{1}{\eta_\theta + \beta} A_\lambda^{\pi_\theta}(x, y)\right),
\end{aligned}$$

where the last equality is by the definition of $A_\lambda^{\pi_\theta}(x, y)$.

The two normalized policies coincide, so $\pi_{\theta_+} = \pi_+$. For the OPD predictor, take $\theta = \hat{\theta}_t$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{t-1}$; for the corrector, take $\theta = \hat{\theta}_t$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t$. Starting from matching policy and dual initializations, induction gives matching predictors, correctors, and dual iterates, because both algorithms use the same population dual updates. A policy-preserving projection leaves this conclusion unchanged.

E Proof of Corollary 3.8

Explicit form of the inexact primal updates in Assumption 3.6. We first specify the ideal distribution-space KL-mirror updates $\pi_t^\dagger$ and $\hat{\pi}_{t+1}^\dagger$ used in Assumption 3.6. They are the exact maximizers of the two KL-regularized inner subproblems associated with the primal optimistic and primal actual updates in Algorithm 3, namely

$$\pi_t^\dagger \in \arg\max_{\pi\in\Pi} \Big\{ \mathcal{L}(\pi, \boldsymbol{\lambda}_{t-1}) - \eta_\theta \, \mathbb{E}_{x\sim\mathcal{D}} \big[\mathrm{KL}(\pi(\cdot|x) \| \pi_{\hat{\theta}_t}(\cdot|x))\big] \Big\},$$
$$\hat{\pi}_{t+1}^\dagger \in \arg\max_{\pi\in\Pi} \Big\{ \mathcal{L}(\pi, \boldsymbol{\lambda}_t) - \eta_\theta \, \mathbb{E}_{x\sim\mathcal{D}} \big[\mathrm{KL}(\pi(\cdot|x) \| \pi_{\hat{\theta}_t}(\cdot|x))\big] \Big\}.$$

For full tabular logits and a policy-preserving projection, Remark 3.5 identifies the implemented policies with these maximizers for exact population updates, so both error budgets are zero. For general parameterized updates, we use Assumption 3.6 with comparator $\hat{\pi}_{t+1}$ for the predictor and comparator $\pi^\star$ for the corrector.

Setup. Use the finite distribution-space problem and its saddle pair $(\pi^\star, \boldsymbol{\lambda}^\star)$ from Theorem 3.4, with $\boldsymbol{\lambda}^\star \in \Lambda^\star$ arbitrary. Write $\pi_t = \pi_{\theta_t}$ and $\hat{\pi}_t = \pi_{\hat{\theta}_t}$. The comparator $\pi^\star$ is the distribution-space optimum and need not belong to the parameterized class. The exact saddle inequalities imply

$$\mathcal{L}(\pi^\star, \boldsymbol{\lambda}) - \mathcal{L}(\pi_\theta, \boldsymbol{\lambda}^\star) \geq 0 \quad (\theta \in \Theta, \ \boldsymbol{\lambda} \geq 0). \tag{49}$$

Let $\pi = \pi_t$ (equivalently, $\theta = \theta_t$) and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_t$ and substituting the definition of $\mathcal{L}(\pi, \boldsymbol{\lambda})$ into the LHS of the above inequality, we have

$$\begin{aligned}
\mathcal{L}(\pi^\star, \boldsymbol{\lambda}_t) - \mathcal{L}(\pi_t, \boldsymbol{\lambda}^\star) =& \mathbb{E}_{x\sim\mathcal{D}} \left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi^\star}(x) - \beta \mathrm{KL}(\pi^\star(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right] \\
& - \mathbb{E}_{x\sim\mathcal{D}} \left[V_{S_{\boldsymbol{\lambda}^\star}}^{\pi_t}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right] \\
=& \mathbb{E}_{x\sim\mathcal{D}} \left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi^\star}(x) - \beta \mathrm{KL}(\pi^\star(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right] \\
& \underbrace{- \mathbb{E}_{x\sim\mathcal{D}} \left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right]}_{\mathrm{A}} \\
& + \mathbb{E}_{x\sim\mathcal{D}} \left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right] \\
& \underbrace{- \mathbb{E}_{x\sim\mathcal{D}} \left[V_{S_{\boldsymbol{\lambda}^\star}}^{\pi_t}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right]}_{\mathrm{B}}
\end{aligned} \tag{50}$$

E.1 Upper bound of term A

We can rewrite term A as:

$$\begin{aligned}
\mathrm{A} =& \mathbb{E}_{x\sim\mathcal{D}} \left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi^\star}(x) - \beta \mathrm{KL}(\pi^\star(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right] - \mathbb{E}_{x\sim\mathcal{D}} \left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - \beta \mathrm{KL}(\pi_t(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right] \\
=& \mathbb{E}_{x\sim\mathcal{D}} \left[\left(V_{S_{\boldsymbol{\lambda}_t}}^{\pi^\star}(x) - V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) \right) - \beta \left(\mathrm{KL}(\pi^\star(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_t(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right) \right] \\
\overset{(a)}{=}& \mathbb{E}_{x\sim\mathcal{D}} \left[\langle \pi^\star(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_t}(x, \cdot) \rangle - \beta \left(\mathrm{KL}(\pi^\star(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_t(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right) \right] \\
\overset{(b)}{=}& \mathbb{E}_{x\sim\mathcal{D}} \left[\langle \pi^\star(\cdot|x) - \hat{\pi}_{t+1}(\cdot|x), S_{\boldsymbol{\lambda}_t}(x, \cdot) \rangle - \beta \left(\mathrm{KL}(\pi^\star(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right) \right. \\
& \quad + \langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_{t-1}}(x, \cdot) \rangle - \beta \left(\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_t(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)) \right) \\
& \quad \left. + \langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_t}(x, \cdot) - S_{\boldsymbol{\lambda}_{t-1}}(x, \cdot) \rangle \right]
\end{aligned} \tag{51}$$

where (a) is because the action space is discrete and $V^{\pi}_{S_{\boldsymbol{\lambda}_t}}(x) = \sum_y \pi(y|x) S_{\boldsymbol{\lambda}_t}(x,y) = \langle \pi(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot)\rangle$ for any $\pi \in \Pi$, and (b) is because adding and subtracting the same term keeps the equality.

As we consider the NPG update in the parameterized space, π_t and $\hat{\pi}_{t+1}$ are obtained from the corresponding updates in Algorithm 3. Using Corollary B.12 with $\eta = \eta_\theta$, $g = S_{\boldsymbol{\lambda}_t}(x,\cdot)$, $\theta_{\rm old} = \hat{\theta}_t$ (so $\pi_{\theta_{\rm old}} = \hat{\pi}_t$), $\theta_{\rm new} = \hat{\theta}_{t+1}$ (so $\pi_{\theta_{\rm new}} = \hat{\pi}_{t+1}$), and $\pi' = \pi^\star(\cdot|x)$, we have

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[\langle S_{\boldsymbol{\lambda}_t}(x,\cdot), \hat{\pi}_{t+1}(\cdot|x) - \pi^\star(\cdot|x)\rangle - \beta\left(\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_{\rm ref}(\cdot|x)) - \mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\rm ref}(\cdot|x))\right)\right] \\
&\geq \mathbb{E}_{x\sim\mathcal{D}}\Big[\eta_\theta\Big(-\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + \mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)) \\
&\qquad + \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x))\Big) + \beta\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x))\Big] \\
&\quad -\mathrm{gap}_t.
\end{aligned}
$$

Putting a negative sign on both sides, we have

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[\langle \pi^\star(\cdot|x) - \hat{\pi}_{t+1}(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot)\rangle - \beta\left(\mathrm{KL}(\pi^\star(\cdot|x)\|\pi_{\rm ref}(\cdot|x)) - \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_{\rm ref}(\cdot|x))\right)\right] \\
&\leq \mathbb{E}_{x\sim\mathcal{D}}\Big[\eta_\theta\Big(\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - \mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)) \\
&\qquad - \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x))\Big) - \beta\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x))\Big] \\
&\quad +\mathrm{gap}_t \\
&= \mathbb{E}_{x\sim\mathcal{D}}\Big[\eta_\theta\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)) \\
&\qquad - \eta_\theta\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x))\Big] + \mathrm{gap}_t
\end{aligned}
\tag{52}
$$

Applying Corollary B.12 again with $\eta = \eta_\theta$, $g = S_{\boldsymbol{\lambda}_{t-1}}(x,\cdot)$, $\theta_{\rm old} = \hat{\theta}_t$ (so $\pi_{\theta_{\rm old}} = \hat{\pi}_t$), $\theta_{\rm new} = \theta_t$ (so $\pi_{\theta_{\rm new}} = \pi_t$), and $\pi' = \hat{\pi}_{t+1}(\cdot|x)$, we have

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[\langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_{t-1}}(x,\cdot)\rangle - \beta\left(\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_{\rm ref}(\cdot|x)) - \mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\rm ref}(\cdot|x))\right)\right] \\
&\leq \mathbb{E}_{x\sim\mathcal{D}}\Big[\eta_\theta\Big(\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) \\
&\qquad - \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x))\Big) - \beta\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))\Big] \\
&\quad +\mathrm{gap}_t \\
&= \mathbb{E}_{x\sim\mathcal{D}}\Big[\eta_\theta\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) \\
&\qquad - \eta_\theta\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x))\Big] \\
&\quad +\mathrm{gap}_t.
\end{aligned}
\tag{53}
$$

Let $C > 0$ be a constant. For the last term in the RHS of Equation (51), we derive the upper bound as

$$
\begin{aligned}
&\langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), S_{\boldsymbol{\lambda}_t}(x,\cdot) - S_{\boldsymbol{\lambda}_{t-1}}(x,\cdot)\rangle \\
&= \left\langle \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \lambda_{t-1,j}) R_j(x,\cdot) \right\rangle \\
&\overset{(a)}{\leq} \|\hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x)\|_1 \left\| \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \lambda_{t-1,j}) R_j(x,\cdot) \right\|_\infty \\
&\overset{(b)}{\leq} \|\hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x)\|_1 \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1 R_{\max} \\
&\leq \sqrt{2\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))} \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1 R_{\max} \\
&\quad \text{(By Pinsker's inequality in Lemma B.2)} \\
&\leq C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) + \frac{R_{\max}^2}{2C} \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1^2 \\
&\quad \text{(By AM-GM inequality } \frac{x^2}{2C} + \frac{y^2 C}{2} \geq xy \text{ with } C > 0) \\
&\leq C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) + \frac{|\mathcal{H}| R_{\max}^2}{2C} \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 \\
&\quad \text{(By } \|\mathbf{x}\|_1^2 \leq d\|\mathbf{x}\|_2^2, \forall \mathbf{x} \in \mathbb{R}^d) \\
&\overset{(c)}{\leq} C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) \\
&\quad + \frac{|\mathcal{H}| R_{\max}^2}{C} \left(\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 \right),
\end{aligned}
\tag{54}
$$

where (a) is by Hölder's inequality shown in Lemma B.1 and letting $f(\cdot) = \hat{\pi}_{t+1}(\cdot|x) - \pi_t(\cdot|x)$, $g(\cdot) = \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \lambda_{t-1,j}) R_j(x,\cdot)$, $p = 1$, and $q = \infty$. (b) is by

$$
\begin{aligned}
\left\| \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \lambda_{t-1,j}) R_j(x,\cdot) \right\|_\infty &\leq \sum_{j\in\mathcal{H}} \|(\lambda_{t,j} - \lambda_{t-1,j}) R_j(x,\cdot)\|_\infty \\
&\quad \text{(by triangle inequality)} \\
&= \sum_{j\in\mathcal{H}} |\lambda_{t,j} - \lambda_{t-1,j}| \, \|R_j(x,\cdot)\|_\infty \\
&\leq \left(\sum_{j\in\mathcal{H}} |\lambda_{t,j} - \lambda_{t-1,j}| \right) \max_j \|R_j(x,\cdot)\|_\infty \\
&= \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1 \max_j \|R_j(x,\cdot)\|_\infty \\
&\quad \text{(by definition of 1-norm)} \\
&\leq \|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_1 R_{\max} \\
&\quad \text{(by Assumption 3.2 that } |R_j(x,y)| \leq R_{\max} \text{ for any} \\
&\quad j \in \mathcal{S} \cup \mathcal{H} \text{ and } (x,y) \text{ pair).}
\end{aligned}
$$

(c) is because

$$
\begin{aligned}
\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 &= \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t + \hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 \\
&= \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 + 2\langle \boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t, \hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\rangle \\
&\leq 2\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + 2\|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2 \\
&\quad \text{(by Young's inequality with } p = q = 2, \text{ i.e., } \langle \mathbf{x}, \mathbf{y}\rangle \leq \frac{1}{2}(\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2))
\end{aligned}
$$

Substituting Equation (52), Equation (53), and Equation (54) into the RHS of Equation (51), we have

$$\begin{aligned}
\mathrm{A} \leq & \mathbb{E}_{x\sim\mathcal{D}}\left[\eta_\theta \mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x)) - \eta_\theta \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x))\right. \\
& + \eta_\theta \mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) - \eta_\theta \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) \\
& \left. + C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) + \frac{|\mathcal{H}|R_{\max}^2}{C}\left(\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2\right)\right] + 2\mathrm{gap}_t \\
= & \mathbb{E}_{x\sim\mathcal{D}}\left[\eta_\theta \mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_t(\cdot|x)) - (\eta_\theta+\beta)\mathrm{KL}(\pi^\star(\cdot|x)\|\hat{\pi}_{t+1}(\cdot|x))\right. \\
& - (\eta_\theta+\beta)\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x)) - \eta_\theta \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) \\
& \left. + C\mathrm{KL}(\hat{\pi}_{t+1}(\cdot|x)\|\pi_t(\cdot|x))\right] + \frac{|\mathcal{H}|R_{\max}^2}{C}\left(\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2\right) + 2\mathrm{gap}_t
\end{aligned} \tag{55}$$

E.2 Upper bound of term B

Similarly, we rewrite the term B as

$$\begin{aligned}
\mathrm{B} = & \mathbb{E}_{x\sim\mathcal{D}}\left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - \beta\mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{S_{\boldsymbol{\lambda}^\star}}^{\pi_t}(x) - \beta\mathrm{KL}(\pi_t(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\right] \\
= & \mathbb{E}_{x\sim\mathcal{D}}\left[V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - V_{S_{\boldsymbol{\lambda}^\star}}^{\pi_t}(x)\right] \\
= & \mathbb{E}_{x\sim\mathcal{D}}\left[\left(V_{S_{\boldsymbol{\lambda}_t}}^{\pi_t}(x) - V_{S_{\hat{\boldsymbol{\lambda}}_{t+1}}}^{\pi_t}(x)\right) + \left(V_{S_{\hat{\boldsymbol{\lambda}}_{t+1}}}^{\pi_t}(x) - V_{S_{\boldsymbol{\lambda}^\star}}^{\pi_t}(x)\right)\right. \\
& \left. - \left(V_{S_{\boldsymbol{\lambda}_t}}^{\pi_{t-1}}(x) - V_{S_{\hat{\boldsymbol{\lambda}}_{t+1}}}^{\pi_{t-1}}(x)\right) + \left(V_{S_{\boldsymbol{\lambda}_t}}^{\pi_{t-1}}(x) - V_{S_{\hat{\boldsymbol{\lambda}}_{t+1}}}^{\pi_{t-1}}(x)\right)\right]
\end{aligned} \tag{56}$$

Define $R_{\mathcal{H}} = \left[R_{j_1}, R_{j_2}, \cdots, R_{j_{|\mathcal{H}|}}\right] \in \mathbb{R}^{|\mathcal{H}|}$, where $j_1 < j_2 < \cdots < j_{|\mathcal{H}|}$, and $j_k \in \mathcal{H}$ for any integer $1 \leq k \leq |\mathcal{H}|$. Define $V_{R_{\mathcal{H}}}^{\pi}(x) = \mathbb{E}_{y\sim\pi(\cdot|x)} R_{\mathcal{H}}(x,y) \in \mathbb{R}^{|\mathcal{H}|}$. By the definition of $V_{S_\lambda}^{\pi}(x)$, we have

$$\begin{aligned}
V_{S_\lambda}^{\pi}(x) = & \mathbb{E}_{y\sim\pi(\cdot|x)} S_\lambda(x,y) \quad \text{(by definition of } V_{S_\lambda}^{\pi}(x)\text{)} \\
= & \mathbb{E}_{y\sim\pi(\cdot|x)}\left[\sum_{j\in\mathcal{S}} w_j R_j(x,y) + \sum_{j\in\mathcal{H}} \lambda_j R_j(x,y)\right] \quad \text{(by definition of } S_\lambda\text{)} \\
= & \mathbb{E}_{y\sim\pi(\cdot|x)}\left[\sum_{j\in\mathcal{S}} w_j R_j(x,y)\right] + \sum_{j\in\mathcal{H}} \lambda_j \mathbb{E}_{y\sim\pi(\cdot|x)}\left[R_j(x,y)\right] \\
= & \mathbb{E}_{y\sim\pi(\cdot|x)}\left[\sum_{j\in\mathcal{S}} w_j R_j(x,y)\right] + \lambda^T V_{R_{\mathcal{H}}}^{\pi}(x) \quad \text{(by definition of } V_{R_{\mathcal{H}}}^{\pi}\text{)}
\end{aligned}$$

Plugging the above expression of $V_{S_\lambda}^{\pi}(x)$ into term B, we have

$$\begin{aligned}
\mathrm{B} = & \mathbb{E}_{x\sim\mathcal{D}}\Big[\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T V_{R_{\mathcal{H}}}^{\pi_t}(x) + \left(\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star\right)^T V_{R_{\mathcal{H}}}^{\pi_t}(x) \\
& - \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x) + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x)\Big] \\
= & \mathbb{E}_{x\sim\mathcal{D}}\Big[\left(\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star\right)^T V_{R_{\mathcal{H}}}^{\pi_t}(x) + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x) \\
& + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(V_{R_{\mathcal{H}}}^{\pi_t}(x) - V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x)\right)\Big] \\
= & \left(\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star\right)^T \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_t}(x)\right] + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x)\right] \\
& + \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \Big(\mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_t}(x)\right] \\
& - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_{\mathcal{H}}}^{\pi_{t-1}}(x)\right]\Big)
\end{aligned} \tag{57}$$

Recall the dual actual update in the parameterized space shown in Algorithm 3. In vector form, it is

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_{t+1} &= \arg\min_{\boldsymbol{\lambda}\in\Lambda} \boldsymbol{\lambda}^T \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{R_{\mathcal{H}}}(x)\right] + \eta_\lambda \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_t\|_2^2 \\ &= \arg\max_{\boldsymbol{\lambda}\in\Lambda} -\boldsymbol{\lambda}^T \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{R_{\mathcal{H}}}(x)\right] - \eta_\lambda \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_t\|_2^2.\end{aligned}$$

Under the distribution-space assumptions of Theorem 3.4, Lemma B.4 gives $\boldsymbol{\lambda}^\star \in \Lambda$, and all dual iterates lie in Λ by construction. Let $g = -\mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{R_{\mathcal{H}}}(x)\right]$, $\eta = \eta_\lambda$, $h(\mathbf{x}) = \|\mathbf{x}\|_2^2$, $D_h(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2$, $\mathbf{x}' = \boldsymbol{\lambda}^\star$, $\mathbf{x}_{\text{new}} = \hat{\boldsymbol{\lambda}}_{t+1}$, $\mathbf{x}_{\text{old}} = \hat{\boldsymbol{\lambda}}_t$, and $\Omega = \Lambda$ in Lemma B.8, we have

$$\langle -\mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{R_{\mathcal{H}}}(x)\right], \hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star\rangle \geq \eta_\lambda \left(-\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 + \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 + \|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2\right).$$

Putting negative sign on both sides, we have

$$\langle \hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}^\star, \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{R_{\mathcal{H}}}(x)\right]\rangle \leq \eta_\lambda \left(\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2\right). \tag{58}$$

Similarly, the corresponding optimistic dual update in Algorithm 3 is

$$\boldsymbol{\lambda}_t = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \boldsymbol{\lambda}^T \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_{t-1}}_{R_{\mathcal{H}}}(x)\right] + \eta_\lambda \|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}_t\|_2^2,$$

Applying Lemma B.8 by setting $g = -\mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_{t-1}}_{R_{\mathcal{H}}}(x)\right]$, $\eta = \eta_\lambda$, $h(\mathbf{x}) = \|\mathbf{x}\|_2^2$, $D_h(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2^2$, $\mathbf{x}' = \hat{\boldsymbol{\lambda}}_{t+1}$, $\mathbf{x}_{\text{new}} = \boldsymbol{\lambda}_t$, $\mathbf{x}_{\text{old}} = \hat{\boldsymbol{\lambda}}_t$, and $\Omega = \Lambda$, we have

$$\langle \boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}, \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_{t-1}}_{R_{\mathcal{H}}}(x)\right]\rangle \leq \eta_\lambda \left(\|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}_t\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2\right). \tag{59}$$

We upper bound the last term of B as

$$\begin{aligned}&\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(\mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_t}_{R_{\mathcal{H}}}(x)\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V^{\pi_{t-1}}_{R_{\mathcal{H}}}(x)\right]\right) \\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(V^{\pi_t}_{R_{\mathcal{H}}}(x) - V^{\pi_{t-1}}_{R_{\mathcal{H}}}(x)\right)\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(\langle \pi_t(\cdot|x), R_{\mathcal{H}}(x,\cdot)\rangle - \langle \pi_{t-1}(\cdot|x), R_{\mathcal{H}}(x,\cdot)\rangle\right)\right] \\ &\qquad\qquad \text{(by } V^{\pi}_{R_{\mathcal{H}}}(x) = \mathbb{E}_{y\sim\pi(\cdot|x)}\left[R_{\mathcal{H}}(x,y)\right] = \langle \pi(\cdot|x), R_{\mathcal{H}}(x,\cdot)\rangle\text{)} \\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[\langle \pi_t(\cdot|x) - \pi_{t-1}(\cdot|x), \left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T R_{\mathcal{H}}(x,\cdot)\rangle\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}}\left[\left\langle \pi_t(\cdot|x) - \pi_{t-1}(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \hat{\lambda}_{t+1,j}) R_j(x,\cdot)\right\rangle\right]\end{aligned}$$

Define constants $C_1 > 0$ and $C_2 > 0$. Fixing x, we have

$$\begin{aligned}&\left\langle \pi_t(\cdot|x) - \pi_{t-1}(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \hat{\lambda}_{t+1,j}) R_j(x,\cdot)\right\rangle \\ &= \left\langle \pi_t(\cdot|x) - \hat{\pi}_t(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \hat{\lambda}_{t+1,j}) R_j(x,\cdot)\right\rangle + \left\langle \hat{\pi}_t(\cdot|x) - \pi_{t-1}(\cdot|x), \sum_{j\in\mathcal{H}} (\lambda_{t,j} - \hat{\lambda}_{t+1,j}) R_j(x,\cdot)\right\rangle \\ &\leq C_1 \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + \frac{|\mathcal{H}| R_{\max}^2}{2C_1} \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\ &\quad + C_2 \mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x)) + \frac{|\mathcal{H}| R_{\max}^2}{2C_2} \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\ &\quad \text{(By derivations of Equation (54))} \\ &= C_1 \mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2 \mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x)) \\ &\quad + |\mathcal{H}| R_{\max}^2 \left(\frac{1}{2C_1} + \frac{1}{2C_2}\right) \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2\end{aligned}$$

Substituting the above inequality into $\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(\mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_\mathcal{H}}^{\pi_t}(x)\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\right]\right)$, we have

$$\begin{aligned}
&\left(\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\right)^T \left(\mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_\mathcal{H}}^{\pi_t}(x)\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[V_{R_\mathcal{H}}^{\pi_{t-1}}(x)\right]\right) \\
&\leq \mathbb{E}_{x\sim\mathcal{D}}\left[C_1\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2\mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\right] + |\mathcal{H}|R_{\max}^2\left(\frac{1}{2C_1}+\frac{1}{2C_2}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2
\end{aligned} \tag{60}$$

Combining Equation (58), Equation (59), and Equation (60), we can upper bound B as

$$\begin{aligned}
\mathrm{B} \leq& \eta_\lambda\left(\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2\right) \\
&+ \eta_\lambda\left(\|\hat{\boldsymbol{\lambda}}_{t+1} - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\hat{\boldsymbol{\lambda}}_{t+1} - \boldsymbol{\lambda}_t\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2\right) \\
&+ \mathbb{E}_{x\sim\mathcal{D}}\left[C_1\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2\mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\right] \\
&+ |\mathcal{H}|R_{\max}^2\left(\frac{1}{2C_1}+\frac{1}{2C_2}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 \\
=& \eta_\lambda\left(\|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_t\|_2^2 - \|\boldsymbol{\lambda}^\star - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2 - \|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_t\|_2^2\right) \\
&+ \mathbb{E}_{x\sim\mathcal{D}}\left[C_1\mathrm{KL}(\pi_t(\cdot|x)\|\hat{\pi}_t(\cdot|x)) + C_2\mathrm{KL}(\hat{\pi}_t(\cdot|x)\|\pi_{t-1}(\cdot|x))\right] \\
&+ |\mathcal{H}|R_{\max}^2\left(\frac{1}{2C_1}+\frac{1}{2C_2}\right)\|\boldsymbol{\lambda}_t - \hat{\boldsymbol{\lambda}}_{t+1}\|_2^2
\end{aligned} \tag{61}$$

E.3 Combining A and B

Let $a = \sqrt{|\mathcal{H}|}R_{\max}$, $\eta = 3a$, and $K(p,q) = \mathbb{E}_x\mathrm{KL}(p(\cdot \mid x)\|q(\cdot \mid x))$. For the implemented policies, define

$$\begin{aligned}
\Phi_t :=& \eta K(\pi^\star, \hat{\pi}_t) + \eta\,\mathrm{dist}_2^2(\hat{\boldsymbol{\lambda}}_t, \Lambda^\star) \\
&+ aK(\hat{\pi}_t, \pi_{t-1}) + a\|\hat{\boldsymbol{\lambda}}_t - \boldsymbol{\lambda}_{t-1}\|_2^2.
\end{aligned}$$

At iteration t, abbreviate $q = \hat{\pi}_t$, $p = \pi_t$, $u = \hat{\pi}_{t+1}$, $v = \hat{\boldsymbol{\lambda}}_t$, $l = \boldsymbol{\lambda}_t$, and $w = \hat{\boldsymbol{\lambda}}_{t+1}$. Set

$$\begin{gathered}
k = K(\pi^\star, u), \quad U = K(u,p), \quad V = K(p,q), \quad b_1 = \|w-l\|_2^2, \quad d_1 = \|l-v\|_2^2, \\
D_t := \beta k + (a+\beta)U + 2aV + ab_1 + 2ad_1.
\end{gathered}$$

Substitute $C = C_1 = C_2 = a$ and $\eta_\theta = \eta_\lambda = \eta$ in (55) and (61). Choose $\boldsymbol{\lambda}^\star$ closest to v in $\Lambda^\star$, and bound $\mathrm{dist}_2(w, \Lambda^\star)$ by $\|w - \boldsymbol{\lambda}^\star\|_2$. The nonnegative saddle gap in (49) then gives

$$\Phi_{t+1} + D_t \leq \Phi_t + 2\mathrm{gap}_t, \qquad \mathrm{gap}_t = (\eta+\beta)(\varepsilon_{1,t} + \varepsilon_{2,t}). \tag{62}$$

There are two primal oracle calls, each contributing at most gap_t. The dual updates are exact population projections, and hence contribute no approximation term.

To control the distance to $\Lambda^\star$, put $u^\dagger = \hat{\pi}_{t+1}^\dagger$ and $\sigma_t = \|C_{\pi_{\mathrm{ref}}}\log(u/u^\dagger)\|_{\pi_{\mathrm{ref}}}$. Here, for any function f on prompt–response pairs,

$$\begin{gathered}
(C_{\pi_{\mathrm{ref}}}f)(x,y) = f(x,y) - \mathbb{E}_{z\sim\pi_{\mathrm{ref}}(\cdot|x)}f(x,z), \\
\|f\|_{\pi_{\mathrm{ref}}}^2 = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\mathrm{ref}}(\cdot|x)}f(x,y)^2.
\end{gathered}$$

Assumption 3.7 gives $u/\pi_{\mathrm{ref}} \geq L^{-1}$ and $q/\pi_{\mathrm{ref}} \geq L^{-1}$. Let $M = (1+B_\lambda)R_{\max}$ and $\omega = \eta/(\eta+\beta)$. The ideal corrector has the normalized form

$$\frac{u^\dagger(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} = \frac{(q(y \mid x)/\pi_{\mathrm{ref}}(y \mid x))^\omega e^{S_l(x,y)/(\eta+\beta)}}{Z_t(x)},$$

where $Z_t(x)$ is the expectation of the numerator under $y \sim \pi_{\mathrm{ref}}(\cdot \mid x)$. Since $|S_l| \leq M$, Jensen's inequality bounds $Z_t(x)$ by $e^{M/(\eta+\beta)}\mathbb{E}_{\pi_{\mathrm{ref}}}(q/\pi_{\mathrm{ref}})^\omega \leq e^{M/(\eta+\beta)}$ at each prompt. Thus

$$u^\dagger/\pi_{\mathrm{ref}} \geq L^{-\omega}e^{-2M/(\eta+\beta)} \geq L^{-1},$$

where the last inequality uses $\log L \geq 2M/\beta$. Lemma B.13 and Assumption 3.6 give $\sigma_t^2 \leq 2L\varepsilon_{1,t}$.

Recall the affine stationarity map and optimal constraint values:

$$A\boldsymbol{\lambda} = C_{\pi_{\mathrm{ref}}} \sum_{j\in\mathcal{H}} \lambda_j R_j, \qquad b = C_{\pi_{\mathrm{ref}}} \left[\beta \log(\pi^\star/\pi_{\mathrm{ref}}) - \sum_{j\in\mathcal{S}} w_j R_j \right],$$
$$g(p) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim p(\cdot|x)} R_{\mathcal{H}}(x,y), \qquad g^\star = g(\pi^\star).$$

Stationarity of $u^\dagger$ gives the perturbed identity

$$Al - b = C_{\pi_{\mathrm{ref}}} \left[\beta \log \frac{u}{\pi^\star} + \eta \log \frac{u}{q} - (\eta+\beta) \log \frac{u}{u^\dagger} \right].$$

Split $\log(u/q) = \log(u/p) + \log(p/q)$ and apply Lemma B.13. All three implemented policies have density at least $1/L$ relative to π_{ref}, and $\pi^\star$ has density at least $1/L_0 \geq 1/L$ by (38). Thus

$$\begin{aligned} \|Aw - b\|_{\pi_{\mathrm{ref}}} &\leq \sqrt{L}\big(\beta\sqrt{2k} + \eta\sqrt{2U} + \eta\sqrt{2V}\big) + L_A\sqrt{b_1} + (\eta+\beta)\sigma_t \\ &\leq \sqrt{C_A(L) D_t} + (\eta+\beta)\sigma_t, \end{aligned}$$

where $L_A = \sqrt{\|\Gamma\|_{\mathrm{op}}}$ and $C_A(L), C_g, \kappa(L)$ are the constants in (6). The projection $w = \mathrm{Proj}_\Lambda(v - g(p)/(2\eta))$ gives

$$0 \leq (g^\star)^\top w \leq 2B_\lambda\big(a\sqrt{2k} + a\sqrt{2U} + 2\eta\sqrt{b_1} + 2\eta\sqrt{d_1}\big) \leq \sqrt{C_g D_t}.$$

This follows by testing projection optimality at an element of $\Lambda^\star$, as in the exact proof; the inequality uses only bounded rewards and exact dual projection. Applying the Hoffman bound (40) and $(x+y)^2 \leq 2x^2 + 2y^2$ yields

$$\mathrm{dist}_2^2(w, \Lambda^\star) \leq 4H_{\pi_{\mathrm{ref}}}^2 (C_A(L) + C_g) D_t + 2H_{\pi_{\mathrm{ref}}}^2 (\eta+\beta)^2 \sigma_t^2.$$

It follows that

$$\Phi_{t+1} \leq \kappa_{\mathrm{in}} D_t + \mathcal{R}_t, \quad \kappa_{\mathrm{in}} = 2\kappa(L), \quad \mathcal{R}_t = 4\eta H_{\pi_{\mathrm{ref}}}^2 L(\eta+\beta)^2 \varepsilon_{1,t}. \tag{63}$$

Indeed, the remaining energy terms satisfy $\eta k + aU + ab_1 \leq [\eta/\beta + a/(a+\beta) + 1] D_t$. Together with $\sigma_t^2 \leq 2L\varepsilon_{1,t}$, this gives the displayed coefficient $2\kappa(L)$ and remainder $\mathcal{R}_t$. Combining this inequality with (62) gives

$$\Phi_{t+1} \leq \rho_{\mathrm{in}} \Phi_t + 2\rho_{\mathrm{in}} \mathrm{gap}_t + (1-\rho_{\mathrm{in}}) \mathcal{R}_t, \qquad \rho_{\mathrm{in}} = \frac{\kappa_{\mathrm{in}}}{1+\kappa_{\mathrm{in}}}. \tag{64}$$

Iterating this inequality from $t = 0$ gives

$$\Phi_t \leq \rho_{\mathrm{in}}^t \Phi_0 + \sum_{s=0}^{t-1} \rho_{\mathrm{in}}^{t-1-s} \big[2\rho_{\mathrm{in}} \mathrm{gap}_s + (1-\rho_{\mathrm{in}}) \mathcal{R}_s \big].$$

The initialization $\pi_{-1} = \hat{\pi}_0 = \pi_{\mathrm{ref}}$ and $\boldsymbol{\lambda}_{-1} = \hat{\boldsymbol{\lambda}}_0$ makes the two memory terms in Φ_0 zero. Since $\eta[K(\pi^\star, \hat{\pi}_t) + \mathrm{dist}_2^2(\hat{\boldsymbol{\lambda}}_t, \Lambda^\star)] \leq \Phi_t$, this proves the corollary.

For a run satisfying the same assumptions for every $t \geq 0$, if $\mathrm{gap}_t \leq \bar{g}$ and $\mathcal{R}_t \leq \overline{\mathcal{R}}$, the geometric series gives

$$\limsup_{t\to\infty} \Phi_t \leq \frac{2\rho_{\mathrm{in}}}{1-\rho_{\mathrm{in}}} \bar{g} + \overline{\mathcal{R}} = 2\kappa_{\mathrm{in}} \bar{g} + \overline{\mathcal{R}}.$$

If both approximation budgets tend to zero, define $e_s = 2\rho_{\mathrm{in}} \mathrm{gap}_s + (1-\rho_{\mathrm{in}}) \mathcal{R}_s$, so $e_s \to 0$. For fixed N, the part of the error sum with $s < N$ tends to zero, while its remaining part is at most $\sup_{s\geq N} e_s/(1-\rho_{\mathrm{in}})$. Letting $t \to \infty$ and then $N \to \infty$ proves $\Phi_t \to 0$.

Table 1: Computational and memory overhead of PD and OPD.

Algorithm	Median step time (ms)	Peak memory (GB)
PD	714	65.4
OPD	991	82.1

F Ablation Experiments

F.1 Running Time and Memory Consumption Comparison

We adopt the TinyLlama-1.1B model as the reference policy throughout our experiments. For model-based alignment, we directly use the corresponding reference model to train the reward and constrained reward with Safe-RLHF Dai et al. (2024). All experiments are conducted on a single NVIDIA H200 GPU. Each run takes approximately 1–2 hours.

OPD introduces additional memory and computation compared with simpler baselines. The dominant overhead comes from the predictor–corrector structure: OPD requires two actor forward–backward passes per optimization step (one predictor pass at the current iterate and one corrector pass at the predicted point), whereas the PD baseline performs one.

We also note that in our practical implementation, the optimistic and actual updates can reuse the same rollout batch. Therefore, OPD does not necessarily require an additional round of online response generation, and the dominant extra cost is the additional backward computation rather than doubled sample collection.

To quantify this cost, we directly profiled the each step function on a single H200 GPU using a TinyLlama-1.1B actor, batch size 16, and response length 256. For each algorithm, we report the median step time and peak allocated memory averaged over 5 steps. As shown in Table 1, OPD increases per-step wall-clock time by about 38 percent.

Note that the extra cost is worth it in practically constrained RLHF settings, because OPD is clearly more stable than PD. As shown in the figure, with relatively aggressive learning rates (actor/critic learning rates 5×10^{-6} and 10^{-6}), PD often shows large swings during training and may end up with policies that have low reward and still break the safety constraints. In contrast, OPD trains much more smoothly and more consistently reaches policies with high reward that also satisfy the constraints.

F.2 Varied Learning Rates

We further evaluate the sensitivity of PD and OPD to the actor and critic learning rates. Specifically, we consider three learning-rate settings: $(5\times10^{-6}, 2\times10^{-6})$, $(10^{-5}, 5\times10^{-6})$, and $(2\times10^{-5}, 10^{-5})$. Figure 3 shows the reward and constraint reward during training. Across these settings, PD exhibits higher sensitivity to the learning rate and shows noticeable fluctuations in both reward and constraint reward. This instability becomes more pronounced as the learning rate increases. In contrast, OPD produces more stable training trajectories across all three settings and consistently maintains favorable reward and constraint reward values.

Table 2 reports the corresponding evaluation reward and constraint reward at different training checkpoints. Although PD can achieve higher scores at some intermediate checkpoints, its evaluation performance varies substantially across training and learning-rate settings. Under the largest learning-rate setting, $(2 \times 10^{-5}, 10^{-5})$, PD eventually collapses to a policy with a negative final evaluation reward. By contrast, OPD achieves higher final reward and constraint reward than PD across all three learning-rate settings. These results are consistent with the training curves in Figure 3 and indicate that OPD is less sensitive to the choice of learning rate.

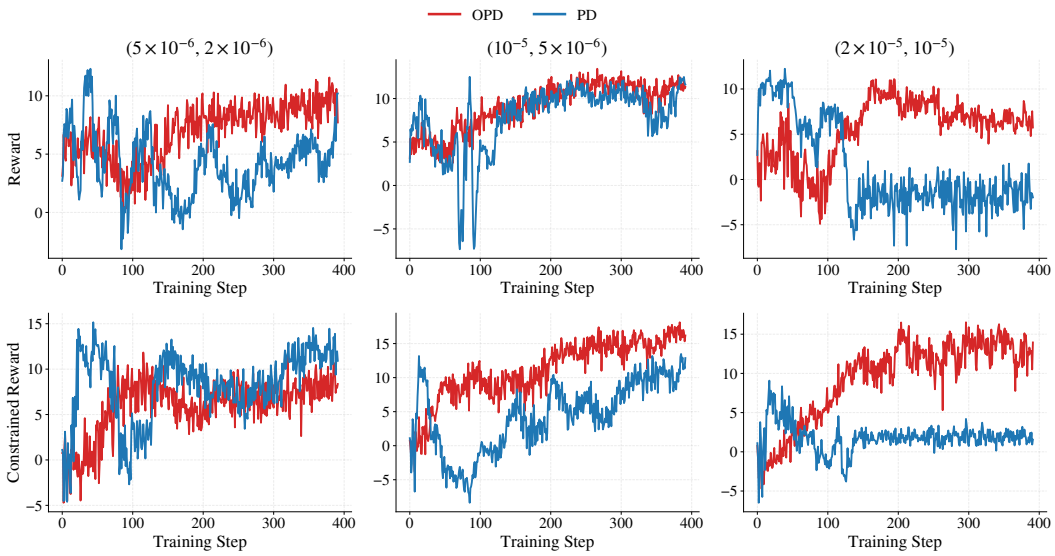


Figure 3: OPD and PD comparison with varied learning rate in the training phase

Table 2: Evaluation reward and constraint reward under different learning-rate settings. We report results at checkpoints 0, 100, 200, 300, and 392. Higher values are better.

Actor/Critic LR	Algo.	Step 0	Step 100	Step 200	Step 300	Step 392
Evaluation Reward						
$(5\times10^{-6}, 2\times10^{-6})$	OPD	4.05	2.51	**4.72**	**6.09**	**7.92**
	PD	4.08	**5.12**	3.89	1.96	6.83
$(10^{-5}, 5\times10^{-6})$	OPD	4.05	**7.08**	8.52	**9.15**	**9.10**
	PD	4.08	1.93	**9.51**	8.15	5.25
$(2\times10^{-5}, 10^{-5})$	OPD	4.05	4.78	**7.35**	**7.19**	**6.57**
	PD	4.08	**6.83**	-1.75	-1.75	-1.75
Evaluation Constraint Reward						
$(5\times10^{-6}, 2\times10^{-6})$	OPD	-1.47	**5.73**	3.49	5.24	**8.62**
	PD	-1.54	-0.51	**7.61**	**7.88**	6.64
$(10^{-5}, 5\times10^{-6})$	OPD	-1.47	**8.75**	**9.84**	**12.14**	**13.70**
	PD	-1.54	-0.69	6.04	5.97	3.73
$(2\times10^{-5}, 10^{-5})$	OPD	-1.47	**6.98**	**16.81**	**14.81**	**11.88**
	PD	-1.54	-2.84	1.82	1.82	1.82