CONSTRAINED BILEVEL REINFORCEMENT LEARNING VIA APPROXIMATE LOWER-LEVEL RESPONSES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study constrained bilevel reinforcement learning when the lower-level problem is a constrained Markov decision process with a generally parameterized policy. The exact constrained solution map can be nonunique and need not possess the stability or differentiability required by conventional hypergradient formulas. We therefore formulate the bilevel problem through an approximate lower-level response characterized directly by expected reward suboptimality and constraint violation. This response is produced by a finite-sample stochastic primal–dual natural policy-gradient method whose natural-gradient direction is approximated by vanilla stochastic gradient descent. To differentiate the upper objective induced by this approximate response, we introduce an unbiased two-point encoding of every reward and constraint signal and apply a score-function identity to the probability law of the complete lower-level training procedure. The resulting estimator does not differentiate through the full sequence of lower-level policy, critic, and projected dual updates. We establish lower-level reward and feasibility guarantees, upper-level stationarity, and an end-to-end environment-interaction complexity of $\widetilde{\mathcal{O}}(\delta_\ell^{-6}\epsilon_o^{-2})$, where δ_ℓ is the controllable numerical lower-level accuracy and ϵ_o is the target upper-level stationarity error. Setting $\delta_\ell = \epsilon_o = \epsilon$ gives $\widetilde{\mathcal{O}}(\epsilon^{-8})$.

1 INTRODUCTION

Bilevel optimization provides a natural framework for learning and decision-making problems in which an upper-level decision influences a nested learning or control problem. In reinforcement learning (RL), this structure arises in hyperparameter optimization, meta-learning, reward design, inverse RL, and preference-based learning. We study the more challenging setting of constrained bilevel RL, where the lower-level policy must additionally satisfy safety, resource, fairness, or performance constraints and the upper-level variable modifies the reward or constraint specification of a lower-level constrained Markov decision process (CMDP).

The problem is particularly difficult with generally parameterized policies because upper-level gradients require stability of the lower-level solution set with respect to the upper-level variable. Existing nonconvex bilevel analyses obtain hypergradient formulas under Polyak–Łojasiewicz conditions, error bounds, or related stability assumptions (Chen et al., 2024; Kwon et al., 2024). Such regularity is difficult to establish in parameterized constrained RL: the primal optimizer and Lagrange multiplier may be nonunique, the solution set may vary nonsmoothly with the upper-level variable, and differentiating the KKT or primal–dual system requires strong regularity of the constrained solution map.

We therefore replace the exact constrained optimizer with an approximate lower-level response characterized directly by reward suboptimality and constraint violation. We construct it using a finite-sample stochastic primal–dual natural policy-gradient method whose natural-gradient direction is approximated by vanilla stochastic gradient descent. The resulting response is random because of trajectory sampling, stochastic gradient estimation, and random iterate selection; this randomness belongs to the response-construction procedure rather than the bilevel formulation itself.

To differentiate the upper objective induced by the approximate lower-level response, we introduce a two-point encoding of every sampled reward and constraint signal used during lower-level training. This encoding preserves the conditional means required by the lower-level algorithm while transfer-

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.

ring the dependence on the upper-level parameter to the probability law of the complete stochastic training trace. Once the encoded trace is fixed, the returned policy parameter is fixed as well. Applying the standard score-function identity (Schulman et al., 2015) to this encoded training law, Lemma 10 (in Appendix) derives an upper gradient without differentiating through the lower-level policy, critic, and dual iterates. To make this hypergradient usable for finite-sample optimization, Lemma 17 (in Appendix) proves that the upper objective induced by the approximate response is smooth, while Lemma 18 (in Appendix) shows that the hypergradient admits an implementable unbiased estimator with controlled second moment. Together, these results turn the two-point encoding construction into a tractable stochastic first-order method for the induced upper objective, without requiring differentiability of an exact constrained lower-level solution map.

Our main contributions are summarized as follows.

- **Approximate constrained bilevel formulation.** We formulate constrained bilevel RL through an approximate lower-level response characterized directly by reward suboptimality and constraint violation, avoiding the need to differentiate an exact constrained solution map.
- **Two-point encoding and training-law gradient.** We introduce a two-point encoding that preserves the conditional means required by the lower-level learner while transferring upper-parameter dependence to the law of its complete stochastic training trace. Applying likelihood-ratio differentiation to this law avoids differentiation through projected primal–dual updates and requires neither a lower-level PL condition nor inverse-Hessian computation.
- **End-to-end convergence.** We establish upper-level stationarity together with lower-level reward-suboptimality and constraint-violation guarantees. The resulting expected environment-interaction complexity is $\widetilde{\mathcal{O}}(\delta_\ell^{-6}\epsilon_o^{-2})$, where δ_ℓ is the controllable numerical lower-level accuracy. It becomes $\widetilde{\mathcal{O}}(\epsilon^{-8})$ when $\delta_\ell = \epsilon_o = \epsilon$.

2 Related Work

2.1 Bilevel Optimization and Reinforcement Learning

Bilevel methods obtain upper gradients through implicit differentiation with Hessian or inverse-Hessian information (Ghadimi & Wang, 2018; Chen et al., 2021; 2022; Yang et al., 2021), or through iterative differentiation of a finite lower trajectory (Ji et al., 2021). Hessian-free, fully first-order, penalty-, and value-function-based methods reduce or avoid this computation (Li et al., 2022; Sow et al., 2022; Kwon et al., 2023; Liu et al., 2022; 2021; Shen & Chen, 2023; Kwon et al., 2024). These ideas have been extended to bilevel RL and RLHF through stochastic approximation, actor–critic, penalty, and sample-complexity analyses (Hong et al., 2023; Shen et al., 2025; Thoma et al., 2024; Gaur et al., 2025; Yang et al., 2025; Chakraborty et al., 2024), but primarily for unconstrained lower-level policy optimization. Appendix A provides the detailed literature discussion.

2.2 Constrained Bilevel Optimization

Constrained bilevel methods use implicit hypergradients, projected stochastic updates, Clarke-subdifferential approximations, and first-order or proximal-Lagrangian reformulations (Tsaknakis et al., 2022; Xiao et al., 2023; Xu & Zhu, 2023; Kornowski et al., 2024; Yao et al., 2024b;a). Their nonasymptotic guarantees generally retain convexity or solution-map regularity; the nonconvex value-function approach of Liu et al. (2023) is asymptotic. In constrained RL, nonuniqueness and changing active constraints obstruct implicit hypergradients. ITD instead differentiates through the primal–dual trajectory, including projected dual updates, tying the derivative to that unrolled procedure. We control an approximate response through reward and feasibility guarantees and differentiate its stochastic training law.

3 PROBLEM FORMULATION

3.1 REINFORCEMENT LEARNING AND CONSTRAINED MDPS

We consider a discounted constrained Markov decision process written as the tuple $\mathcal{M}_v = (\mathcal{S}, \mathcal{A}, P, \rho, \gamma, r_x, c_y)$, where $r_x(s,a) := r(s,a;x)$ and $c_y(s,a) := c(s,a;y)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the transition kernel, ρ is the initial-state distribution, and $\gamma \in [0,1)$ is the discount factor. The upper-level variable is $v = (x,y) \in \mathbb{R}^{d_v}$, where $x \in \mathbb{R}^{d_x}$ parameterizes the reward function and $y \in \mathbb{R}^{d_y}$ parameterizes the constraint function. Importantly, P and ρ do not depend on v. For a stationary randomized policy π, the discounted reward and constraint returns are defined jointly in Eq. 1.

$$J_r(\pi, x) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t; x) \right], \qquad J_c(\pi, y) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t; y) \right]. \tag{1}$$

For a fixed upper-level parameter $v = (x,y)$, the lower-level CMDP is $J_r^\star(v) := \max_\pi \{J_r(\pi, x) : J_c(\pi, y) \geq 0\}$. We optimize over a differentiable parameterized policy class $\{\pi_\theta : \theta \in \mathbb{R}^{d_\theta}\}$, which is not assumed to contain an exact optimizer of the CMDP. For brevity, write $J_r(\theta, x) := J_r(\pi_\theta, x)$ and $J_c(\theta, y) := J_c(\pi_\theta, y)$.

3.2 APPROXIMATE CONSTRAINED BILEVEL REINFORCEMENT LEARNING

The exact formulation obtained by substituting an optimal constrained policy into the upper objective would require analyzing the solution correspondence $v \mapsto \arg\max_\pi \{J_r(\pi, x) : J_c(\pi, y) \geq 0\}$. For parameterized constrained RL, this correspondence can be nonunique and nonsmooth. Rather than imposing differentiability of this exact solution map, we define the bilevel problem through a stochastic approximate lower-level response.

For each upper-level parameter v, let $Z \sim P_v$ denote the complete random trace generated by the specified lower-level stochastic procedure, including all training randomness and the final random iterate selection. The trace law P_v depends on v, whereas $\theta(Z)$ denotes the deterministic policy parameter returned for a realized trace Z and has no additional explicit dependence on v. Thus the lower-level dependence is represented as $v \longrightarrow P_v(Z) \longrightarrow Z \longrightarrow \theta(Z)$, rather than as a deterministic differentiable response map. For every $v \in \mathbb{R}^{d_v}$, we characterize this stochastic response through the reward-suboptimality and feasibility requirements in Eq. 2 and Eq. 3. Here, ϵ_ℓ is an abstract response tolerance and $C_r, C_c > 0$ are constants.

$$\mathbb{E}_{Z \sim P_v} \left[J_r^\star(v) - J_r\left(\pi_{\theta(Z)}, x\right) \right] \leq C_r \epsilon_\ell, \tag{2}$$

$$\mathbb{E}_{Z \sim P_v} \left[-J_c\left(\pi_{\theta(Z)}, y\right) \right] \leq C_c \epsilon_\ell, \tag{3}$$

Let $G(v, \theta)$ denote the upper-level objective. The approximate constrained bilevel problem induced by this stochastic response is defined in Eq. 4.

$$\min_{v \in \mathbb{R}^{d_v}} \Phi_{\epsilon_\ell}(v) := \mathbb{E}_{Z \sim P_v} \left[G(v, \theta(Z)) \right]. \tag{4}$$

Accordingly, the complete dependency for the upper objective is $v \longrightarrow P_v(Z) \longrightarrow Z \longrightarrow \theta(Z) \longrightarrow G(v, \theta(Z))$. The requirements in Eq. 2–Eq. 3 characterize the lower-level response and do not restrict the feasible domain of the upper-level variable v.

The formulation does not require the specified lower-level algorithm and parameterized policy class to attain every arbitrarily prescribed ϵ_ℓ; the subsequent analysis characterizes the response accuracy attainable by the proposed lower-level procedure, including the compatible-function approximation error. We next construct the lower-level stochastic procedure and the corresponding upper-level gradient estimator.

4 PROPOSED APPROACH

Two-point encoding defines the lower-level estimators that drive the primal–dual algorithm. The resulting trace defines a training-distribution score and hence an upper gradient obtained from the law of the complete training procedure, rather than from derivatives of its individual updates.

4.1 Two-Point Encoding of Reward and Constraint Signals

Assumption 1 (Upper-level signal regularity)**.** *There exist finite $R, C, R_1, C_1, R_2, C_2 > 0$ such that, uniformly in the arguments, the bounds in Eq. 5 hold.*

$$|r(s,a;x)| \le R, \quad |c(s,a;y)| \le C, \quad \|\nabla_x r(s,a;x)\| \le R_1, \quad \|\nabla_y c(s,a;y)\| \le C_1, \\ \|\nabla^2_{xx} r(s,a;x)\| \le R_2, \qquad \|\nabla^2_{yy} c(s,a;y)\| \le C_2. \tag{5}$$

Boundedness and smooth dependence of parameterized reward and constraint signals on the upper-level variables are standard regularity conditions in bilevel and configurable reinforcement-learning analyses (Thoma et al., 2024; Shen et al., 2025; Gaur et al., 2025). The bounded signals make the encoding probabilities well defined, while the uniform first- and second-order bounds control the encoding scores and their Hessians. Appendix E.1.1 and Remark 2 provide a concrete normalized function class satisfying these conditions.

Direct differentiation through the projected dual update would require differentiating the projection operator. We instead encode the sampled scalar signals and differentiate the probability law of the complete lower-level training procedure, avoiding differentiation through the sequence of lower-level optimization updates. For any scalar signal $a(v)$ satisfying $|a(v)| \le A$, generate $B \in \{-1, +1\}$ according to the distribution in Eq. 6.

$$\mathbb{P}_v(B = b) = \frac{1}{2} + \frac{b\, a(v)}{4A}, \qquad b \in \{-1, +1\}. \tag{6}$$

The corresponding decoded signal is $\widetilde{a} = 2AB$, which satisfies $\mathbb{E}_v[\widetilde{a}] = a(v)$. The score of the encoded signal is given in Eq. 7.

$$s_a(B; v) = \nabla_v \log \mathbb{P}_v(B) = \frac{B \nabla_v a(v)}{2A + B a(v)}. \tag{7}$$

We conditionally independently encode every sampled reward and constraint signal used by the lower-level algorithm. Under Assumption 1, the decoded signals are $\widetilde{r}(s,a;x) = 2RB^r$ and $\widetilde{c}(s,a;y) = 2CB^c$, where B^r and B^c are generated by Eq. 6 with $A = R$ and $A = C$, respectively. For the j-th encoded signal, we denote the corresponding encoding signs by B^r_j and B^c_j. Because the decoded signals are unbiased, this transformation does not change the conditional means of the stochastic critic and constraint estimators used by the lower-level algorithm.

4.2 Lower-Level Primal-Dual Natural Policy Gradient

Assumption 2 (Uniform Slater condition)**.** *There exist $c_0 : \mathbb{R}^{d_v} \to (0, \infty)$ and a constant $\underline{c} > 0$, independent of v, such that every $v = (x, y)$ admits a stationary policy $\pi_{\mathrm{sl}}(v)$ satisfying $J_c(\pi_{\mathrm{sl}}(v), y) \ge c_0(v) \ge \underline{c}$.*

Strict feasibility through a Slater condition is standard in primal–dual analyses of CMDPs and is used to control the optimal dual multiplier (Bai et al., 2023; Mondal & Aggarwal, 2024; 2026). Because the constraint function in our bilevel formulation depends on the upper-level parameter y, we require the Slater margin to hold uniformly over y. This uniform margin permits a fixed interval $\Lambda = [0, \lambda_{\max}]$, with $\lambda_{\max} \ge 4R/((1-\gamma)\underline{c})$, for every v; Appendix C.1.5 gives the multiplier bound and a sufficient construction.

For a fixed upper-level parameter $v = (x, y)$, define $J_L(\theta, \lambda; v) := J_r(\theta, x) + \lambda J_c(\theta, y)$ for $\lambda \in \Lambda$. Write $\psi_\theta(s,a) := \nabla_\theta \log \pi_\theta(a \mid s)$, let $\nu^{\pi_\theta}_\rho$ denote the normalized discounted state–action occupancy measure, and define the Lagrangian advantage by $A^{\pi_\theta}_{L,\lambda} := A^{\pi_\theta}_r + \lambda A^{\pi_\theta}_c$. With discounted Fisher matrix $F_\rho(\theta)$, the natural-gradient direction $\omega^\star_{\theta,\lambda} := F_\rho(\theta)^{-1} \nabla_\theta J_L(\theta, \lambda; v)$ equivalently minimizes the compatible regression objective

$$\mathcal{L}_\rho(\omega; \theta, \lambda, v) := \frac{1}{2} \mathbb{E}_{(s,a) \sim \nu^{\pi_\theta}_\rho} \left[\left(\psi_\theta(s,a)^\top \omega - \frac{A^{\pi_\theta}_{L,\lambda}(s,a;v)}{1-\gamma} \right)^2 \right]. \tag{8}$$

The critic estimator uses reset access to generate the independent continuations required by Appendix Algorithm 3. At lower iteration k, initialize $\omega_{k,0} = 0$ and apply N vanilla-SGD critic steps

using the encoded stochastic gradient:

$$\omega_{k,n+1} = \omega_{k,n} - \beta \widehat{\nabla}_\omega \mathcal{L}_\rho(\omega_{k,n}; \theta_k, \lambda_k, v), \qquad n = 0, \dots, N-1. \tag{9}$$

With $\widehat{\omega}_k := \omega_{k,N}$, the primal and dual updates are $\theta_{k+1} = \theta_k + \eta\widehat{\omega}_k$ and $\lambda_{k+1} = \Pi_{[0,\lambda_{\max}]}[\lambda_k - \zeta \widehat{J}_{c,\rho}(\theta_k; y)]$. After K iterations, $R \sim \mathrm{Unif}\{0, \dots, K-1\}$ is sampled independently and the algorithm returns the approximate response $\widehat{\theta} = \theta_R = \theta(Z)$. Appendix Algorithms 2–3 give the complete lower-level procedure, encoded oracle, score accumulation, and sampling details.

4.3 Score of the Complete Lower-Level Training Procedure

Recall that $Z \sim P_v$ denotes the complete random trace of one execution of the lower-level algorithm, including all environment trajectories, encoded reward and constraint signals, critic samples, and the final random iterate selection. The complete training score, formalized in Lemma 13 (in Appendix), is defined in Eq. 10. Let $\mathcal{I}_r(Z)$ and $\mathcal{I}_c(Z)$ denote the sets of reward- and constraint-encoding events in the trace, and let $s_{r,j}$ and $s_{c,j}$ denote the corresponding instances of Eq. 7, embedded in the x- and y-coordinate blocks of v, respectively.

$$S(Z; v) = \sum_{j \in \mathcal{I}_r(Z)} s_{r,j} + \sum_{j \in \mathcal{I}_c(Z)} s_{c,j}. \tag{10}$$

Once the complete encoded trace Z is fixed, the lower-level response $\theta(Z)$ is fixed and has no additional explicit dependence on v. The law P_v retains the dependence on the upper-level parameter, and $\mathbb{E}_v$ denotes expectation under this law. This gives the following exact gradient expression for the upper objective induced by the approximate lower-level response.

Theorem 1 (Training-distribution gradient)**.** *Under Assumptions 1 and 3, $\Phi_{\epsilon_\ell}(v) := \mathbb{E}_{Z \sim P_v}[G(v, \theta(Z))]$ is differentiable and, for any deterministic scalar baseline $c(v)$, satisfies the identity in Eq. 11.*

$$\nabla \Phi_{\epsilon_\ell}(v) = \mathbb{E}_v \left[\nabla_v G(v, \theta(Z)) + (G(v, \theta(Z)) - c(v)) S(Z; v) \right]. \tag{11}$$

Here, $S(Z; v)$ is the sum of the score functions of all reward and constraint encoding variables generated during the lower-level training run.

The identity follows from the training-distribution score argument formalized in Lemma 10 (in Appendix), using the same notation $Z \sim P_v$, $\theta(Z)$, and $S(Z; v)$.

Algorithm 1 Constrained Bilevel Reinforcement Learning with an Approximate Lower-Level Response

Require: Initial upper-level parameter v_0, upper-level stepsize α, number of upper-level iterations T_o, common lower-level primal initialization θ_0, fixed dual initialization $\lambda_0 = 0$, and lower-level parameters $K, N, \eta, \zeta, \beta, \lambda_{\max}$.

for $t = 0, \dots, T_o - 1$ **do**
  Run the lower-level primal-dual procedure in Section 4.2 at v_t.
  Obtain its stochastic policy parameter $\widehat{\theta}_t = \theta(Z_t)$ and the complete encoding score $S_t = S(Z_t; v_t)$.
  Set $d_t = \widehat{\theta}_t - \theta_0$ and $B_{G,t} = \max\{1, \lceil \|d_t\|^2 \rceil\}$.
  for $m = 1, \dots, B_{G,t}$ **do**
    Sample $U_{t,m} \sim \mathrm{Unif}[0, 1]$ independently.
    Obtain a fresh upper-gradient sample $\widehat{G}_{\theta,t,m} = \widehat{\nabla}_\theta G(v_t, \theta_0 + U_{t,m} d_t)$.
    Set $Y_{t,m} = \langle \widehat{G}_{\theta,t,m}, d_t \rangle$.
  end for
  Form $\widehat{\delta}_{G,t} = B_{G,t}^{-1} \sum_{m=1}^{B_{G,t}} Y_{t,m}$.
  Obtain an independent upper partial-gradient sample $\widehat{G}_{v,t} = \widehat{\nabla}_v G(v_t, \widehat{\theta}_t)$.
  Form $\widehat{g}_t = \widehat{G}_{v,t} + \widehat{\delta}_{G,t} S_t$.
  Update $v_{t+1} = v_t - \alpha \widehat{g}_t$.
end for
Independently sample $R_o \sim \mathrm{Unif}\{0, \dots, T_o - 1\}$.
Run one fresh lower-level procedure at v_{R_o} to obtain a trace Z_{out} and $\widehat{\theta}_{\mathrm{out}} = \theta(Z_{\mathrm{out}})$.
return $(v_{R_o}, \widehat{\theta}_{\mathrm{out}})$.

4.4 Stochastic Upper-Level Gradient

Assumption 3 (Upper-objective regularity). *The objective satisfies $G_{\min} \le G(v,\theta) \le G_{\max}$, $\|\nabla_v G(v,\theta)\| \le L_{G,0}$, and $\|\nabla_\theta G(v,\theta)\| \le M_G$, while $\nabla_v G(v,\theta)$ is L_G-Lipschitz in v. Fresh upper-gradient oracles satisfy the unbiasedness and variance conditions in Eq. 12.*

$$\begin{aligned} \mathbb{E}[\widehat{\nabla}_v G(v,\theta) \mid v,\theta] &= \nabla_v G(v,\theta), \qquad \mathbb{E}[\widehat{\nabla}_\theta G(v,\theta) \mid v,\theta] = \nabla_\theta G(v,\theta), \\ \mathbb{E}[\|\widehat{\nabla}_v G - \nabla_v G\|^2 \mid v,\theta] &\le \sigma_v^2, \qquad \mathbb{E}[\|\widehat{\nabla}_\theta G - \nabla_\theta G\|^2 \mid v,\theta] \le \sigma_\theta^2. \end{aligned} \tag{12}$$

Smoothness of the upper-level objective together with unbiased, bounded-variance stochastic gradient oracles is standard in stochastic bilevel and nonconvex optimization analyses (Hu et al., 2023; Gong et al., 2024). The additional uniform boundedness conditions control the direct-gradient and score-function terms in our upper estimator; a concrete oracle construction and detailed verification are given in Appendix E.1.2.

For the estimator, choose the baseline $c(v) = G(v,\theta_0)$, where θ_0 is the fixed initialization common to all lower-level runs. To estimate $G(v,\theta(Z)) - G(v,\theta_0)$ using only upper-gradient oracle calls, define $d = \theta(Z) - \theta_0$. Under Assumption 3, write $V_\theta := M_G^2 + \sigma_\theta^2$ and set the implementable batch size $B_G(Z) := \max\{1, \lceil \|d\|^2 V_\theta \rceil\}$. Draw conditionally independent $U_m \sim \mathrm{Unif}[0,1]$ and fresh upper-gradient oracle samples for $m = 1, \dots, B_G(Z)$. The fundamental theorem of calculus gives the identity in Eq. 13.

$$G(v,\theta(Z)) - G(v,\theta_0) = \int_0^1 \langle \nabla_\theta G(v, \theta_0 + ud), d \rangle \, du. \tag{13}$$

Accordingly, we define the value-difference estimator in Eq. 14.

$$\widehat{\delta}_G := \frac{1}{B_G(Z)} \sum_{m=1}^{B_G(Z)} \left\langle \widehat{\nabla}_\theta G(v, \theta_0 + U_m d), d \right\rangle. \tag{14}$$

Conditionally on Z, this estimator is unbiased for the value difference and satisfies $\mathrm{Var}(\widehat{\delta}_G \mid Z) \le V_\theta$. Moreover, Lemma 15 gives $\mathbb{E}[B_G(Z)] \le 2 + \mathbb{E}\|d\|^2 V_\theta = O(1+K)$, so under the theorem's choice $K = \lceil \delta_\ell^{-2} \rceil$, each upper-level iteration uses $O(\delta_\ell^{-2})$ upper θ-gradient calls in expectation, in addition to one v-gradient call. Under the trajectory-based Bradley–Terry oracle, every call uses two trajectories of length at most H; this cost is included in Theorem 3. The resulting stochastic upper gradient, using an independent fresh unbiased sample $\widehat{\nabla}_v G(v,\theta(Z))$, is given in Eq. 15.

$$\widehat{g}(v) = \widehat{\nabla}_v G(v,\theta(Z)) + \widehat{\delta}_G \, S(Z;v). \tag{15}$$

The upper-level parameter is then updated as $v_{t+1} = v_t - \alpha \widehat{g}_t$, as summarized in Algorithm 1. A key feature of Eq. 15 is that it requires neither an inverse Hessian of a constrained saddle system nor differentiation through the full sequence of lower-level updates, including the projected dual update.

5 Convergence Analysis

We first establish that the proposed lower-level algorithm satisfies the approximate-response requirements in Eq. 2–Eq. 3. We then analyze the upper objective Φ_{ϵ_ℓ} induced by this response, establish upper-level stationarity, and combine the two guarantees. The proof sketch in Section 5.5 explains how the length of the lower-level training procedure enters the final environment-interaction complexity. Full proofs and detailed interaction accounting are deferred to the appendix.

5.1 Remaining Assumptions

The convergence analysis operates under the upper-level signal, Uniform Slater, and upper-objective regularity assumptions introduced in Section 4, together with the additional assumptions below.

Assumption 4 (Continuous-state CMDP regularity). *The state space $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ is a Borel state space, and the action space $\mathcal{A}$ is finite. The initial distribution ρ and each transition kernel $P(\cdot \mid s,a)$ admit densities $\rho(s)$ and $p(\cdot \mid s,a)$, respectively, with respect to Lebesgue measure. The transition kernel is weakly continuous: for every bounded continuous function $f : \mathcal{S} \to \mathbb{R}$ and every $a \in \mathcal{A}$, the mapping $s \mapsto \int_{\mathcal{S}} f(s')p(s' \mid s,a)\,ds'$ is continuous. For every fixed upper-level parameter $v = (x,y)$ and every $a \in \mathcal{A}$, the mappings $r(\cdot,a;x)$ and $c(\cdot,a;y)$ are continuous in the state.*

Weak continuity of the transition kernel together with state regularity of the one-stage signals is standard in continuous-space MDP and reinforcement-learning theory (Kara et al., 2023; Kara & Yüksel, 2026), including constrained discounted MDPs (Feinberg et al., 2020). These conditions support the existence and occupancy-measure arguments used here. The additional density requirement permits a density-valued presentation of discounted occupancies and is satisfied by common continuous-state systems with nondegenerate additive noise; Appendix C.1.1 gives a sufficient construction and detailed justification.

Assumption 5 (Policy regularity)**.** *The policies have full action support, $\pi_\theta(a \mid s) > 0$, and the common initialization satisfies $\pi_{\theta_0}(a \mid s) \geq p_0$ uniformly for some $p_0 > 0$. There exist $G_\pi, L_\pi > 0$ satisfying the bounds in Eq. 16.*

$$\|\nabla_\theta \log \pi_\theta(a \mid s)\| \leq G_\pi, \qquad \|\nabla_\theta \log \pi_\theta(a \mid s) - \nabla_\theta \log \pi_{\theta'}(a \mid s)\| \leq L_\pi \|\theta - \theta'\|. \tag{16}$$

Full support and the score bounds control likelihood-ratio and policy-gradient terms, while the initialization condition controls the initial KL comparison. Bounded-score assumptions are common in finite-sample policy-gradient analyses (Bedi et al., 2022; Papini et al., 2022); bounded and Lipschitz score conditions are used, for example, by Fatkhullin et al. (2023) and Zhang et al. (2023). Appendix C.1.2 gives further discussion.

Assumption 6 (Fisher non-degeneracy)**.** *There exists $\mu_F > 0$ such that $F_\rho(\theta) \succeq \mu_F I$ for every θ; under policy regularity, $F_\rho(\theta) \preceq G_\pi^2 I$.*

Uniform Fisher non-degeneracy is standard in analyses of natural and general-parameterization policy-gradient methods (Liu et al., 2020; Ding et al., 2022; Fatkhullin et al., 2023; Mondal & Aggarwal, 2024). It makes the natural-gradient direction well-defined and makes the compatible-regression objective strongly convex, which is needed for geometric critic-bias decay; see Appendix C.1.3.

Let d_ρ^π denote the normalized discounted state-occupancy measure of policy π, let $\nu_\rho^\pi(s,a) := d_\rho^\pi(s)\pi(a \mid s)$ denote its state-action occupancy measure, and let A_r^π and A_c^π denote its reward and constraint advantages, respectively.

Assumption 7 (Compatible-function approximation)**.** *For an optimal stationary policy $\pi^\star$, define $A_{L,\lambda}^{\pi_\theta} := A_r^{\pi_\theta} + \lambda A_c^{\pi_\theta}$ and the transferred compatible-function loss in Eq. 17.*

$$\mathcal{L}_{\nu_\rho^{\pi^\star}}(\omega, \theta, \lambda) := \frac{1}{2} \int_{\mathcal{S}} d_\rho^{\pi^\star}(s) \sum_{a \in \mathcal{A}} \pi^\star(a \mid s) \left[\nabla_\theta \log \pi_\theta(a \mid s)^\top \omega - \frac{A_{L,\lambda}^{\pi_\theta}(s,a;v)}{1-\gamma} \right]^2 ds. \tag{17}$$

Uniformly over θ and $\lambda \in [0, \lambda_{\max}]$, the exact natural-gradient direction satisfies $\mathcal{L}_{\nu_\rho^{\pi^\star}}(\omega_{\theta,\lambda}^\star, \theta, \lambda) \leq \epsilon_{\text{bias}}/2$.

Compatible-function approximation assumptions are standard in analyses of natural policy-gradient methods with general policy parameterization (Agarwal et al., 2020; Liu et al., 2020) and have been used directly in parameterized CMDPs (Bai et al., 2023; Mondal & Aggarwal, 2024; 2026). The term ϵ_{bias} is the resulting approximation floor and measures the representational mismatch between the Lagrangian advantage and the score-function feature class; see Appendix C.1.4.

5.2 Lower-Level Convergence

The two-point encoding preserves the conditional means of the sampled reward and constraint signals, so the stochastic critic gradient remains conditionally unbiased. For the vanilla-SGD critic recursion in Eq. 9, let $\omega_k^\star := \omega_{\theta_k,\lambda_k}^\star$ and let $\mathbb{E}_k[\cdot]$ denote expectation conditional on the history at the start of lower iteration k. If R_ω uniformly bounds $\|\omega_k^\star\|$, Lemma 5 (in Appendix) gives the conditional-bias bound

$$\|\mathbb{E}_k[\widehat{\omega}_k] - \omega_k^\star\| \leq R_\omega \exp(-\beta \mu_F N). \tag{18}$$

The same lemma provides a uniform second-moment bound despite the stochastic critic noise floor. Thus $N = \Theta(\log K)$ makes the first-order critic bias $O(K^{-1/2})$, matching the primal–dual optimization scale. Combining this estimate with Lemmas 6 and 7 (in Appendix) gives the response guarantee.

Theorem 2 (Approximate lower-level response guarantee)**.** *Suppose Assumptions 1, 2, and 4–7 hold. Choose $\eta = \eta_0/\sqrt{K}$, $\zeta = \zeta_0/\sqrt{K}$ for fixed $\eta_0, \zeta_0 > 0$, choose the critic stepsize β to satisfy the condition in Lemma 5 (in Appendix), take $N \geq \max\{1, \lceil(\beta\mu_F)^{-1}\log(R_\omega\sqrt{K})\rceil\}$, and let $R \sim \mathrm{Unif}\{0, \ldots, K-1\}$. Let $Z \sim P_v$ denote the complete trace of this lower run, including the selection of R, so that $\theta_R = \theta(Z)$. The bounds in Eq. 19 and Eq. 20 hold for problem-dependent constants $C_r, C_c > 0$.*

$$\mathbb{E}_{Z\sim P_v}[J_r^\star(v) - J_r(\theta(Z), x)] \leq \sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_r}{\sqrt{K}}, \tag{19}$$

$$\mathbb{E}_{Z\sim P_v}[-J_c(\theta(Z), y)] \leq C_{\mathrm{sl}}\sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_c}{\sqrt{K}}. \tag{20}$$

Here $C_{\mathrm{sl}} = 2/\lambda_{\max}$. The approximation error ϵ_{bias} may depend on $\lambda_{\max}$.

A detailed proof of the theorem is provided in Appendix C.

Let $\delta_\ell := K^{-1/2}$ and $\epsilon_\ell := \max\{\sqrt{\epsilon_{\mathrm{bias}}}, \delta_\ell\}$. With $\overline{C}_r := 1 + C_r$ and $\overline{C}_c := C_{\mathrm{sl}} + C_c$, Theorem 2 gives reward and constraint errors bounded by $\overline{C}_r\epsilon_\ell$ and $\overline{C}_c\epsilon_\ell$. Thus $K = \Theta(\delta_\ell^{-2})$ and $N = \Theta(\log(\delta_\ell^{-1}))$ produce the required response; increasing K does not remove the approximation floor.

5.3 Upper-Level Stationarity

For the lower-level choices in Theorem 2, define the lower-run scale $Q_\ell := K(N+1)/(1-\gamma)$. Lemma 17 (in Appendix) proves that $\Phi_{\epsilon_\ell}(v) = \mathbb{E}_{Z\sim P_v}[G(v, \theta(Z))]$ is L_Φ-smooth with $L_\Phi = O(1 + Q_\ell)$. Lemma 18 (in Appendix) proves that the estimator in Eq. 15 satisfies

$$\mathbb{E}[\widehat{g}(v)] = \nabla\Phi_{\epsilon_\ell}(v), \qquad \mathbb{E}[\|\widehat{g}(v)\|^2] \leq V_\Phi, \qquad V_\Phi = O(1 + Q_\ell). \tag{21}$$

Together, these bounds determine the stepsize and iteration complexity of the upper-level update. Section 5.5 explains why both quantities inherit the same dependence on Q_ℓ.

Lemma 1 (Upper-level stationarity)**.** *By Lemma 17 (in Appendix), Φ_{ϵ_ℓ} is L_Φ-smooth. Suppose Eq. 21 holds uniformly in v. For $\epsilon_o \in (0, 1]$, define $\overline{L}_\Phi := \max\{1, L_\Phi\}$, $\overline{V}_\Phi := \max\{1, V_\Phi\}$, and $\Delta_G := G_{\max} - G_{\min}$. Choose $\alpha = \epsilon_o/(\overline{L}_\Phi\overline{V}_\Phi)$ and $T_o = \max\{1, \lceil 2\Delta_G\overline{L}_\Phi\overline{V}_\Phi/\epsilon_o^2\rceil\}$, and let $R_o \sim \mathrm{Unif}\{0, \ldots, T_o - 1\}$. The resulting stationarity bound is given in Eq. 22.*

$$\mathbb{E}[\|\nabla\Phi_{\epsilon_\ell}(v_{R_o})\|^2] \leq \epsilon_o. \tag{22}$$

Lemma 19 (in Appendix) provides the detailed proof of this result.

5.4 End-to-End Constrained Bilevel Guarantee

Combining upper-level stationarity with an independent fresh lower-level run ensures that the reported upper-level iterate is paired with a policy satisfying the reward and feasibility guarantees.

Theorem 3 (End-to-end constrained bilevel guarantee)**.** *Suppose the conditions of Theorem 2 and Lemma 1 hold uniformly in v. For a prescribed numerical lower-level accuracy $\delta_\ell \in (0, 1]$, choose $K = \lceil\delta_\ell^{-2}\rceil$ and choose N according to Theorem 2. Define the effective response accuracy by $\epsilon_\ell := \max\{\sqrt{\epsilon_{\mathrm{bias}}}, \delta_\ell\}$ and let $(v_{R_o}, \widehat{\theta}_{\mathrm{out}})$ be the final output, where $\widehat{\theta}_{\mathrm{out}}$ is obtained from an independent lower run at v_{R_o}. Then*

$$\mathbb{E}[\|\nabla\Phi_{\epsilon_\ell}(v_{R_o})\|^2] \leq \epsilon_o, \tag{23}$$

$$\mathbb{E}[J_r^\star(v_{R_o}) - J_r(\widehat{\theta}_{\mathrm{out}}, x_{R_o})] \leq \overline{C}_r\epsilon_\ell, \tag{24}$$

$$\mathbb{E}[-J_c(\widehat{\theta}_{\mathrm{out}}, y_{R_o})] \leq \overline{C}_c\epsilon_\ell. \tag{25}$$

If each upper-gradient oracle call uses at most $2H$ environment transitions, then, for fixed primitive problem constants, the expected total interaction count satisfies $N_{\mathrm{env}} = \widetilde{\mathcal{O}}(\delta_\ell^{-6}\epsilon_o^{-2})$. In particular, setting $\delta_\ell = \epsilon_o = \epsilon$ gives $N_{\mathrm{env}} = \widetilde{\mathcal{O}}(\epsilon^{-8})$.

A detailed proof is provided in Appendix Theorem 5.

5.5 Proof Sketch of the End-to-End Sample Complexity

Step 1: Constructing an Approximate Lower-Level Response. Theorem 2 bounds both reward suboptimality and constraint violation by the compatible-approximation floor plus $O(K^{-1/2})$. Consequently, numerical accuracy δ_ℓ requires $K = \Theta(\delta_\ell^{-2})$ and $N = \Theta(\log(\delta_\ell^{-1}))$, up to fixed constants and logarithmic factors. Define $Q_\ell := K(N+1)/(1-\gamma) = \widetilde{\mathcal{O}}(\delta_\ell^{-2})$. Here $K(N+1)$ counts the encoded critic calls in a complete lower run, while $(1-\gamma)^{-1}$ is the expected trajectory-length scale of each call. Thus Q_ℓ measures the expected scale of one complete lower-level training run: producing one δ_ℓ-accurate response already costs $\widetilde{\mathcal{O}}(\delta_\ell^{-2})$ environment interactions. The primal–dual details establishing this response are deferred to Appendix Theorem 4.

Step 2: Replacing Solution-Map Sensitivity with Training-Law Sensitivity. Once the complete encoded trace Z is fixed, every sampled signal and update is fixed, so the realized output $\theta(Z)$ has no additional explicit dependence on v. The lower-level dependence on v is instead carried by the trace law P_v. Differentiating this law represents the indirect upper sensitivity through the complete training score $S(Z; v) := \nabla_v \log P_v(Z)$, rather than through a derivative of a constrained optimizer or differentiation through the full sequence of lower-level updates. Lemma 10 (in Appendix) formally establishes the resulting gradient identity for $\Phi_{\epsilon_\ell}(v) = \mathbb{E}_{Z\sim P_v}[G(v, \theta(Z))]$.

Step 3: Tracing Lower-Training Length Through the Upper-Level Analysis. The complete training score is a sum of score increments from the encoded signals generated throughout the lower run. These increments accumulate over all $K(N+1)$ calls, so the score second moment and the corresponding log-density Hessian moment grow with Q_ℓ; Lemmas 13 and 14 (in Appendix) make this accumulation precise. This growth has two distinct consequences. First, differentiating the score-function representation of Φ_{ϵ_ℓ} introduces both score products and log-density Hessian terms. Their moment bounds imply $L_\Phi = O(1 + Q_\ell)$, as established by Lemma 17 (in Appendix). Second, the stochastic upper-gradient estimator multiplies a value-difference estimate by the complete training score. Its second moment therefore inherits the score second moment, giving $V_\Phi = O(1 + Q_\ell)$, as established by Lemma 18 (in Appendix). Avoiding differentiation through the full sequence of lower-level updates does not remove dependence on the length of the lower training procedure; under the training-distribution construction, that dependence reappears through the moments of the complete training score.

Step 4: Composing the Interaction Complexity. Using these smoothness and second-moment bounds, Lemma 19 (in Appendix) yields $T_o = O(L_\Phi V_\Phi \epsilon_o^{-2}) = O((1+Q_\ell)^2 \epsilon_o^{-2}) = \widetilde{\mathcal{O}}(\delta_\ell^{-4}\epsilon_o^{-2})$. Every upper-level iteration executes one lower response costing $O(Q_\ell)$. Including the upper-oracle calls and the final fresh lower run, the detailed accounting in Theorem 5 (in Appendix) gives $N_{\mathrm{env}} = O(Q_\ell^3 \epsilon_o^{-2}) = \widetilde{\mathcal{O}}(\delta_\ell^{-6}\epsilon_o^{-2})$. The factor δ_ℓ^{-6} therefore decomposes into three factors of δ_ℓ^{-2}: one for a lower response, one for score-induced smoothness, and one for the upper-gradient second moment. This decomposition is not an information-theoretic lower bound. Within the present construction, improving the δ_ℓ dependence requires removing at least one factor: reducing the cost of a lower response, preventing upper smoothness from growing with the complete training score, or constructing an upper-gradient estimator whose second moment has weaker score-length dependence.

6 Conclusion

We developed a constrained bilevel RL framework that replaces the potentially nonunique exact constrained solution map with a performance-controlled approximate lower-level response. A stochastic primal–dual natural policy-gradient method constructs this response, while two-point encoding yields an unbiased score-function upper gradient by differentiating the distribution of the complete training procedure rather than unrolling the critic, policy, or projected dual updates. We established lower-level reward and feasibility guarantees, stationarity of the induced upper objective, and an expected environment-interaction complexity of $\widetilde{\mathcal{O}}(\delta_\ell^{-6}\epsilon_o^{-2})$, where δ_ℓ is the controllable numerical lower-level accuracy. The effective response accuracy is $\epsilon_\ell = \max\{\sqrt{\epsilon_{\mathrm{bias}}}, \delta_\ell\}$ and therefore retains the compatible-function approximation floor. When $\delta_\ell = \epsilon_o = \epsilon$, the interaction complexity becomes $\widetilde{\mathcal{O}}(\epsilon^{-8})$.

AI USE STATEMENT

Generative AI was used to format author-provided manuscript text in the ICLR LaTeX template and to draft the title and abstract. The authors reviewed the AI-assisted edits and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in Markov decision processes. In *Proceedings of the Thirty-Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 64–66. PMLR, 2020.

Qinbo Bai, Amrit Singh Bedi, and Vaneet Aggarwal. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6737–6744, 2023.

Amrit Singh Bedi, Souradip Chakraborty, Anjaly Parayil, Brian M. Sadler, Pratap Tokekar, and Alec Koppel. On the hidden biases of policy mirror ascent in continuous action spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1716–1731. PMLR, 2022.

Souradip Chakraborty, Amrit Bedi, Alec Koppel, Huazheng Wang, Dinesh Manocha, Mengdi Wang, and Furong Huang. PARL: A unified framework for policy alignment in reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 947–980. PMLR, 2024.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34: 25294–25307, 2021.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2466–2488. PMLR, 2022.

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

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9827–9869. PMLR, 2023.

Eugene A. Feinberg, Anna Jaśkiewicz, and Andrzej S. Nowak. Constrained discounted markov decision processes with Borel state spaces. *Automatica*, 111:108582, 2020.

Lucy L Gao, Jane J Ye, Haian Yin, Shangzhi Zeng, and Jin Zhang. Moreau envelope based difference-of-weakly-convex reformulation and algorithm for bilevel programs. *arXiv preprint arXiv:2306.16761*, 2023.

Mudit Gaur, Utsav Singh, Amrit Singh Bedi, Raghu Pasupathy, and Vaneet Aggarwal. On the sample complexity bounds of bilevel reinforcement learning. *Advances in Neural Information Processing Systems*, 38, 2025.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Xiaochuan Gong, Jie Hao, and Mingrui Liu. An accelerated algorithm for stochastic bilevel optimization under unbounded smoothness. *Advances in Neural Information Processing Systems*, 37, 2024.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Yifan Hu, Jie Wang, Yao Xie, Andreas Krause, and Daniel Kuhn. Contextual stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36, 2023.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Liuyuan Jiang, Quan Xiao, Victor M Tenorio, Fernando Real-Rojas, Antonio G Marques, and Tianyi Chen. A primal-dual-assisted penalty approach to bilevel optimization with coupled constraints. *arXiv preprint arXiv:2406.10148*, 2024.

Ali D. Kara and Serdar Yüksel. Approximations and learning for continuous state and action MDPs under average cost criteria. *Journal of Machine Learning Research*, 27(68):1–50, 2026.

Ali Devran Kara, Naci Saldi, and Serdar Yüksel. Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity. *Journal of Machine Learning Research*, 24(199):1–34, 2023.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in neural information processing systems*, 34:30271–30283, 2021.

Prashant Khanduri, Ioannis Tsaknakis, Yihua Zhang, Jia Liu, Sijia Liu, Jiawei Zhang, and Mingyi Hong. Linearly constrained bilevel optimization: A smoothed implicit gradient approach. In *International Conference on Machine Learning*, pp. 16291–16325. PMLR, 2023.

Guy Kornowski, Swati Padmanabhan, Kai Wang, Jimmy Zhang, and Suvrit Sra. First-order methods for linearly constrained bilevel optimization. *Advances in Neural Information Processing Systems*, 37:141417–141460, 2024.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pp. 18083–18113. PMLR, 2023.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. In *The Twelfth International Conference on Learning Representations*, 2024.

Junyi Li, Bin Gu, and Heng Huang. A fully single loop algorithm for bilevel optimization without hessian inverse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7426–7434, 2022.

Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in neural information processing systems*, 35: 17248–17262, 2022.

Risheng Liu, Xuan Liu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A value-function-based interior-point method for non-convex bi-level optimization. In *International conference on machine learning*, pp. 6882–6892. PMLR, 2021.

Risheng Liu, Xuan Liu, Shangzhi Zeng, Jin Zhang, and Yixuan Zhang. Value-function-based sequential minimization for bi-level optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15930–15948, 2023.

Risheng Liu, Zhu Liu, Wei Yao, Shangzhi Zeng, and Jin Zhang. Moreau envelope for nonconvex bilevel optimization: A single-loop and hessian-free solution strategy. In *International Conference on Machine Learning*, pp. 31566–31596. PMLR, 2024.

Yanli Liu, Kaiqing Zhang, Tamer Başar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Washim Uddin Mondal and Vaneet Aggarwal. Sample-efficient constrained reinforcement learning with general parameterization. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Washim Uddin Mondal and Vaneet Aggarwal. Last-iterate convergence of general parameterized policies in constrained MDPs. *Transactions on Machine Learning Research*, 2026.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *Machine Learning*, 111(11):4081–4137, 2022.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International conference on machine learning*, pp. 30992–31015. PMLR, 2023.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and rlhf. *Journal of Machine Learning Research*, 26(114):1–49, 2025.

Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the convergence theory for hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022.

Vinzenz Thoma, Barna Pásztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. Contextual bilevel reinforcement learning for incentive alignment. *Advances in Neural Information Processing Systems*, 37:127369–127435, 2024.

Ioannis Tsaknakis, Prashant Khanduri, and Mingyi Hong. An implicit gradient-type method for linearly constrained bilevel problems. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5438–5442. IEEE, 2022.

Quan Xiao, Han Shen, Wotao Yin, and Tianyi Chen. Alternating projected sgd for equality-constrained bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 987–1023. PMLR, 2023.

Siyuan Xu and Minghui Zhu. Efficient gradient approximation method for constrained bilevel optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12509–12517, 2023.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.

Yan Yang, Bin Gao, and Ya-xiang Yuan. Bilevel reinforcement learning via the development of hyper-gradient without lower-level convexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 4780–4788. PMLR, 2025.

Wei Yao, Haian Yin, Shangzhi Zeng, and Jin Zhang. Overcoming lower-level constraints in bilevel optimization: A novel approach with regularized gap functions. *arXiv preprint arXiv:2406.01992*, 2024a.

Wei Yao, Chengming Yu, Shangzhi Zeng, and Jin Zhang. Constrained bi-level optimization: Proximal lagrangian value function approach and hessian-free algorithm. In *International Conference on Learning Representations*, volume 2024, pp. 20429–20460, 2024b.

Shenao Zhang, Boyi Liu, Zhaoran Wang, and Tuo Zhao. Model-based reparameterization policy gradient methods: Theory and practical algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68391–68419, 2023.

A Extended Related Work

A.1 Bilevel Optimization

Bilevel optimization has been widely studied in the context of hyperparameter optimization, meta-learning, and hierarchical decision-making. Ghadimi & Wang (2018) were among the first to establish sample-complexity guarantees for hypergradient-based bilevel algorithms under convexity assumptions on the lower-level problem. Chen et al. (2021) subsequently developed a general stochastic bilevel optimization framework with applications to stochastic min–max optimization and reinforcement learning, while Chen et al. (2022) established nonasymptotic convergence guarantees for a single-timescale stochastic bilevel method. Momentum-based approaches further improved the iteration complexity of bilevel optimization (Yang et al., 2021; Khanduri et al., 2021). These methods, however, rely on Hessian-based computations and can therefore be computationally intensive.

To alleviate the cost of Hessian inversion, several works developed Hessian-free or fully first-order bilevel algorithms (Li et al., 2022; Sow et al., 2022; Kwon et al., 2023; Liu et al., 2022). Penalty-based approaches further simplify algorithm design by avoiding explicit hypergradient calculations; convergence guarantees under smoothness and regularity assumptions were established in (Shen & Chen, 2023; Kwon et al., 2024). An alternative line of work introduced value-function-based reformulations of bilevel problems (Liu et al., 2021). Collectively, these advances provide an important foundation for the development and analysis of bilevel RL algorithms.

More recently, envelope-based techniques have been introduced to address nonsmoothness and weak convexity. Gao et al. (2023) proposed a Moreau-envelope-based reformulation under a convex lower-level assumption, whereas Liu et al. (2024) used the Moreau envelope to design a single-loop, Hessian-free method for nonconvex bilevel optimization. Despite this progress, most existing analyses continue to require smoothness, convexity, or other strong regularity conditions on the lower-level objective.

A.2 Bilevel Reinforcement Learning

Bilevel RL has recently emerged as a useful modeling paradigm for hierarchical structures arising in meta-RL, incentive design, and reinforcement learning from human feedback (RLHF). Early work developed stochastic approximation frameworks for bilevel RL, often combining two-timescale updates with policy evaluation and policy improvement. In particular, Hong et al. (2023) established finite-time complexity guarantees for a two-timescale bilevel framework and demonstrated its application to actor–critic methods.

Penalty-based formulations tailored to RL and RLHF were developed by Shen et al. (2025), who established convergence guarantees under smoothness and regularity assumptions. Gaur et al. (2025) studied the sample complexity of bilevel RL, characterized the additional difficulties caused by nested policy optimization, and provided improved sample-complexity guarantees for unconstrained bilevel RL and bilevel optimization. Contextual bilevel RL was explored in (Thoma et al., 2024), which developed a hypergradient-based method under a softmax policy parameterization.

Several recent works have sought to relax lower-level convexity assumptions. Yang et al. (2025) developed hypergradient-based bilevel RL algorithms without lower-level convexity, while Chakraborty et al. (2024) proposed a unified bilevel framework for policy alignment in RLHF. Nevertheless, most existing bilevel RL analyses assume an unconstrained lower-level problem and do not establish how their methods or guarantees extend to constrained policy optimization.

A.3 Constrained Bilevel Optimization

Another active line of research considers bilevel optimization with explicit lower-level constraints. Early approaches used implicit hypergradients or projected stochastic methods under strong convexity of the lower-level problem (Tsaknakis et al., 2022; Xiao et al., 2023). Xu & Zhu (2023) combined hypergradient approximations with the Clarke subdifferential for problems with convex lower-level objectives and equality and inequality constraints. Khanduri et al. (2023) also developed a hypergradient-based approach for linearly constrained bilevel optimization and established convergence under weak convexity of the upper-level objective.

The preceding methods rely on computationally demanding Hessian-based hypergradient calculations. Kornowski et al. (2024) developed a first-order algorithm for linearly constrained bilevel problems with a strongly convex lower-level objective. Jiang et al. (2024) proposed a first-order penalty method motivated by a primal–dual formulation for convex lower-level constraints, while retaining strong convexity of the lower-level objective. Proximal-Lagrangian and regularized-gap approaches inspired by envelope constructions were introduced in (Yao et al., 2024b;a) to remove the lower-level strong-convexity requirement.

Although these methods substantially broaden the scope of constrained bilevel optimization, their analyses generally retain convexity assumptions on the lower-level objective or constraints. Liu et al. (2023) removed the lower-level convexity assumption through a value-function-based sequential minimization method, but did not provide a nonasymptotic convergence guarantee. Thus, existing constrained bilevel optimization theory does not directly yield a generally parameterized policy-gradient algorithm for constrained bilevel RL, whose lower-level policy-optimization problem is nonconvex. In this setting, the constrained solution set can be nonunique and unstable as the upper-level variable changes, so the corresponding solution map need not satisfy the regularity required by existing techniques. We therefore introduce an approximate lower-level formulation characterized directly by reward suboptimality and constraint violation. Our work develops a sample-efficient constrained bilevel RL framework based on approximate lower-level responses and a training-distribution score-function gradient, avoiding differentiation through the constrained lower-level optimizer or its primal–dual updates.

B Supporting Algorithms

B.1 Lower-Level Primal-Dual Natural Policy Gradient

For a fixed upper-level parameter $v = (x, y)$, the lower-level routine approximately solves the CMDP by applying primal–dual updates to the Lagrangian $J_L(\theta, \lambda; v) = J_r(\theta, x) + \lambda J_c(\theta, y)$. At primal–dual iteration k, it approximates the natural-gradient direction $F_\rho(\theta_k)^{-1}\nabla_\theta J_L(\theta_k, \lambda_k; v)$ by running N vanilla-SGD iterations on the quadratic critic objective in Eq. 8. The final critic iterate $\widehat{\omega}_k$ is used directly in the policy update $\theta_{k+1} = \theta_k + \eta\widehat{\omega}_k$.

The multiplier is updated by projected descent using an independent encoded constraint-return estimate: $\lambda_{k+1} = \Pi_{[0,\lambda_{\max}]}[\lambda_k - \zeta\widehat{J}_{c,\rho}(\theta_k; y)]$. Every critic or constraint-oracle call also returns the score contribution of its encoded reward and constraint signals. These contributions are accumulated into the complete training score S. After K iterations, the routine samples $R \sim \mathrm{Unif}\{0, \ldots, K-1\}$ and returns the stochastic policy parameter $\widehat{\theta} = \theta_R = \theta(Z)$ associated with the realized trace Z, together with S. Thus, Algorithm 2 constructs the approximate lower-level response required by the bilevel formulation and provides the score used by the upper-level gradient estimator.

Algorithm 2 Lower-Level Primal-Dual Natural Policy Gradient with Vanilla SGD

Require: Upper-level parameter $v = (x, y)$, initial primal parameter θ_0, dual initialization $\lambda_0 = 0$, initial distribution ρ_0, numbers of primal-dual and critic iterations K and N, stepsizes η, ζ, and β, and a fixed dual interval $\Lambda = [0, \lambda_{\max}]$ with $\lambda_{\max} \geq \frac{4R}{(1-\gamma)\underline{c}}$.

Initialize the cumulative encoding score $S \leftarrow 0$.
for $k = 0, \ldots, K-1$ **do**
  Initialize $\omega_{k,0} = 0$.
  for $n = 0, \ldots, N-1$ **do**
    Call Algorithm 3 with $(\theta_k, \omega_{k,n}, \lambda_k, v)$ to obtain $\widehat{\nabla}_\omega \mathcal{L}_\rho(\omega_{k,n}; \theta_k, \lambda_k, v)$ and score $S_{k,n}$.
    Accumulate $S \leftarrow S + S_{k,n}$.
    Update $\omega_{k,n+1} = \omega_{k,n} - \beta \widehat{\nabla}_\omega \mathcal{L}_\rho(\omega_{k,n}; \theta_k, \lambda_k, v)$.
  end for
  Set $\widehat{\omega}_k = \omega_{k,N}$.
  Independently call Algorithm 3 at $(\theta_k, 0, \lambda_k, v)$ to obtain $\widehat{J}_{c,\rho}(\theta_k; y)$ and score S_k^c.
  Accumulate $S \leftarrow S + S_k^c$.
  Update $\theta_{k+1} = \theta_k + \eta \widehat{\omega}_k$ and $\lambda_{k+1} = \Pi_{[0,\lambda_{\max}]}[\lambda_k - \zeta \widehat{J}_{c,\rho}(\theta_k; y)]$.
end for
Independently sample $R \sim \mathrm{Unif}\{0, \ldots, K-1\}$.
return $\widehat{\theta} = \theta_R = \theta(Z)$ and the complete score S.

B.2 Encoded Critic and Constraint Oracle

Algorithm 3 uses geometric-horizon sampling. We assume that the simulator can initialize an occupancy-sampling trajectory from $\rho_0 = \rho$ and reset to a state $\widehat{s}$ sampled by that trajectory. Conditional on $(\widehat{s}, \widehat{a})$ and the preceding history, it can generate independent continuations with fresh transition, action-sampling, and encoding randomness. The Q-continuation fixes its first action to $\widehat{a}$, whereas the V-continuation draws a fresh first action from $\pi_\theta(\cdot \mid \widehat{s})$.

Let $\psi_\theta(s, a) := \nabla_\theta \log \pi_\theta(a \mid s)$. For each queried reward and constraint signal, independently sample the two-point signs using Eq. 6 and decode

$$\widetilde{r}_j = 2RB_j^r, \qquad \widetilde{c}_j = 2CB_j^c. \tag{26}$$

Their score increments, embedded in the x and y coordinates of v, are

$$s_{r,j} = \begin{bmatrix} \dfrac{B_j^r \nabla_x r(s_j, a_j; x)}{2R + B_j^r r(s_j, a_j; x)} \\ 0_{d_y} \end{bmatrix} \in \mathbb{R}^{d_v}, \tag{27}$$

$$s_{c,j} = \begin{bmatrix} 0_{d_x} \\ \dfrac{B_j^c \nabla_y c(s_j, a_j; y)}{2C + B_j^c c(s_j, a_j; y)} \end{bmatrix} \in \mathbb{R}^{d_v}. \tag{28}$$

Algorithm 3 Encoded Unbiased Critic and Constraint Sampling

Require: Policy parameter θ, critic parameter ω, dual variable λ, upper-level parameter $v = (x, y)$, discount factor γ, initial distribution $\rho_0 = \rho$, and simulator reset access as described above.

Initialize the local score $S_{\mathrm{loc}} \leftarrow 0$.
Sample $T_0 \sim \mathrm{Geo}(1-\gamma)$, $s_0 \sim \rho_0$, and $a_0 \sim \pi_\theta(\cdot \mid s_0)$.
for $j = 0, \dots, T_0$ **do**
  Query $c_j = c(s_j, a_j; y)$ and sample $B_j^c \in \{-1, +1\}$ with $\mathbb{P}(B_j^c = b) = \frac{1}{2} + \frac{bc_j}{4C}$, $b \in \{-1, +1\}$.
  Decode $\widetilde{c}_j = 2CB_j^c$ and accumulate $S_{\mathrm{loc}} \leftarrow S_{\mathrm{loc}} + s_{c,j}$.
  if $j < T_0$ **then**
    Execute a_j, observe $s_{j+1} \sim P(\cdot \mid s_j, a_j)$, and sample $a_{j+1} \sim \pi_\theta(\cdot \mid s_{j+1})$.
  end if
end for
Set $\widehat{J}_{c,\rho}(\theta; y) = \sum_{j=0}^{T_0} \widetilde{c}_j$ and $(\widehat{s}, \widehat{a}) = (s_{T_0}, a_{T_0})$.
Independently sample $T_Q \sim \mathrm{Geo}(1-\gamma)$, reset to $\widehat{s}$, and set $s_0 = \widehat{s}$ and $a_0 = \widehat{a}$.
for $j = 0, \dots, T_Q$ **do**
  Query $r_j = r(s_j, a_j; x)$ and $c_j = c(s_j, a_j; y)$.
  Independently encode them as $\widetilde{r}_j = 2RB_j^r$ and $\widetilde{c}_j = 2CB_j^c$.
  Accumulate $S_{\mathrm{loc}} \leftarrow S_{\mathrm{loc}} + s_{r,j} + s_{c,j}$.
  if $j < T_Q$ **then**
    Execute a_j, observe $s_{j+1} \sim P(\cdot \mid s_j, a_j)$, and sample $a_{j+1} \sim \pi_\theta(\cdot \mid s_{j+1})$.
  end if
end for
Set $\widehat{Q}_r^{\pi_\theta}(\widehat{s}, \widehat{a}) = \sum_{j=0}^{T_Q} \widetilde{r}_j$ and $\widehat{Q}_c^{\pi_\theta}(\widehat{s}, \widehat{a}) = \sum_{j=0}^{T_Q} \widetilde{c}_j$.
Independently sample $T_V \sim \mathrm{Geo}(1-\gamma)$, reset to $\widehat{s}$, set $s_0 = \widehat{s}$, and sample $a_0 \sim \pi_\theta(\cdot \mid s_0)$.
for $j = 0, \dots, T_V$ **do**
  Query and independently encode the reward and constraint as $\widetilde{r}_j$ and $\widetilde{c}_j$.
  Accumulate $S_{\mathrm{loc}} \leftarrow S_{\mathrm{loc}} + s_{r,j} + s_{c,j}$.
  if $j < T_V$ **then**
    Execute a_j, observe $s_{j+1} \sim P(\cdot \mid s_j, a_j)$, and sample $a_{j+1} \sim \pi_\theta(\cdot \mid s_{j+1})$.
  end if
end for
Set $\widehat{V}_r^{\pi_\theta}(\widehat{s}) = \sum_{j=0}^{T_V} \widetilde{r}_j$ and $\widehat{V}_c^{\pi_\theta}(\widehat{s}) = \sum_{j=0}^{T_V} \widetilde{c}_j$.
Form $\widehat{A}_r^{\pi_\theta} = \widehat{Q}_r^{\pi_\theta} - \widehat{V}_r^{\pi_\theta}$, $\widehat{A}_c^{\pi_\theta} = \widehat{Q}_c^{\pi_\theta} - \widehat{V}_c^{\pi_\theta}$, and $\widehat{A}_{L,\lambda}^{\pi_\theta} = \widehat{A}_r^{\pi_\theta} + \lambda \widehat{A}_c^{\pi_\theta}$.
Compute $\widehat{F}_\rho(\theta) = \psi_\theta(\widehat{s}, \widehat{a}) \psi_\theta(\widehat{s}, \widehat{a})^\top$ and $\widehat{H}_\rho(\theta, \lambda) = \widehat{A}_{L,\lambda}^{\pi_\theta} \psi_\theta(\widehat{s}, \widehat{a})$.
Form $\widehat{\nabla}_\omega \mathcal{L}_\rho(\omega; \theta, \lambda, v) = \widehat{F}_\rho(\theta)\omega - (1-\gamma)^{-1} \widehat{H}_\rho(\theta, \lambda)$.
return $\widehat{J}_{c,\rho}(\theta; y)$, $\widehat{\nabla}_\omega \mathcal{L}_\rho(\omega; \theta, \lambda, v)$, and S_{loc}.

The decoded signals are conditionally unbiased, so the critic-gradient and constraint-return estimators retain the means required by the lower-level analysis. Meanwhile, conditioned on the complete encoded trace, the returned policy has no additional explicit dependence on v.

C Proof of Theorem 2

C.1 Justification for Assumptions for Theorem 2

C.1.1 Discussion on Assumption 4

Weak continuity of the controlled transition kernel and continuity of the one-stage reward and constraint functions are standard regularity conditions in continuous-space MDP and reinforcement-learning theory. Kara et al. (2023) study Q-learning on general state and action spaces under weak continuity of the transition kernel together with continuity and boundedness of the one-stage cost, while Kara & Yüksel (2026) use weakly continuous dynamics in approximation and learning results for continuous-space average-cost MDPs. In the constrained discounted setting, Feinberg et al. (2020) study Borel-state CMDPs under a weak-continuity regime. These references provide precedent for the weak-continuity and state-regularity component of our assumption, rather than for every condition imposed here. Our Euclidean state space and finite action space specialize the more general Borel-space formulations in these works.

The additional density condition is not fundamental to existence of an optimal policy, but is imposed so that the initial law, transition kernels, and discounted occupancies can be represented by ordinary densities

$$\rho(s), \qquad p(s' \mid s, a), \qquad d_\rho^\pi(s),$$

and the expressions below can be written as Lebesgue integrals. For each action, the state continuity imposed here and differentiability with respect to the upper-level variable imposed in Assumption 1 make the signals separately continuous in their Euclidean arguments and hence jointly Borel measurable, as needed below.

As a concrete example, let

$$\mathcal{S} = \mathbb{R}^{d_s}, \qquad \mathcal{A} = \{1, \dots, m\},$$

and consider the additive-noise dynamics

$$s_{t+1} = f_a(s_t) + \xi_t^a, \qquad \xi_t^a \sim \mathcal{N}(0, \Sigma_a), \qquad \Sigma_a \succ 0,$$

where each $f_a : \mathbb{R}^{d_s} \to \mathbb{R}^{d_s}$ is continuous. The transition kernel then has the Lebesgue density

$$p(s' \mid s, a) = \mathcal{N}(s'; f_a(s), \Sigma_a).$$

If the initial state is also absolutely continuous, for example

$$s_0 \sim \mathcal{N}(m_0, \Sigma_0), \qquad \Sigma_0 \succ 0,$$

then ρ admits a Lebesgue density as well. To verify weak continuity, let h be bounded and continuous and let $s_n \to s$. Under a common realization of ξ^a, continuity of f_a gives

$$f_a(s_n) + \xi^a \longrightarrow f_a(s) + \xi^a \qquad \text{almost surely.}$$

Bounded convergence therefore yields

$$\mathbb{E}[h(f_a(s_n) + \xi^a)] \longrightarrow \mathbb{E}[h(f_a(s) + \xi^a)],$$

or equivalently,

$$\int_{\mathcal{S}} h(s')p(s' \mid s_n, a)\, ds' \longrightarrow \int_{\mathcal{S}} h(s')p(s' \mid s, a)\, ds'.$$

Thus, if $r(\cdot, a; x)$ and $c(\cdot, a; y)$ are continuous in the state, all state-regularity requirements in Assumption 4 hold. The uniform boundedness and derivative requirements with respect to the upper-level parameter are handled separately by Assumption 1.

C.1.2 Discussion on Assumption 5

Uniform boundedness of the policy score is a common regularity condition in finite-sample policy-gradient analysis; see, for example, Bedi et al. (2022), Papini et al. (2022), and Fatkhullin et al. (2023). Fatkhullin et al. also impose Lipschitz continuity of the score, as does Assumption 5.1 of Zhang et al. (2023). These references provide precedent specifically for the bounded- and Lipschitz-score conditions, rather than for every component of Assumption 5.

The uniform initialization bound is an explicit requirement on the chosen policy class and initialization; it is not automatic on a noncompact state space. A state-independent uniform initial policy satisfies it whenever representable by the class.

C.1.3 Discussion on Assumption 6

Uniform positive definiteness of the Fisher information matrix is used in global and finite-sample analyses of policy-gradient and natural policy-gradient methods (Liu et al., 2020; Ding et al., 2022; Fatkhullin et al., 2023). It is also assumed in the general-parameterization CMDP analysis of Mondal & Aggarwal (2024). Besides ensuring that the natural gradient $F_\rho(\theta)^{-1}\nabla J(\theta)$ is well-defined, the condition makes the quadratic compatible-function objective μ_F-strongly convex, which yields the geometric critic convergence used below. Prior work also discusses concrete Fisher-nondegenerate classes: Gaussian policies with fixed positive-definite covariance and a suitably full-rank mean parameterization, as well as broader subclasses of exponential-family policies; see Fatkhullin et al. (2023) and Section 8 of Ding et al. (2022). These examples require structural conditions on the parameterization and do not imply Fisher non-degeneracy for arbitrary Gaussian or exponential-family policies.

C.1.4 Discussion on Assumption 7

The transferred form above evaluates the squared compatible-function residual under the occupancy distribution of an optimal CMDP policy. Closely matching assumptions appear in general-parameterization CMDP analyses (Bai et al., 2023; Mondal & Aggarwal, 2024; 2026), building on compatible-function approximation conditions used in unconstrained natural policy-gradient analyses (Agarwal et al., 2020; Liu et al., 2020). Accordingly, ϵ_{bias} records the expressivity error of the score-function feature class. In settings identified by prior work, the error vanishes for complete parameterizations such as direct or softmax policies, with analogous exact-representation results for linear MDPs; sufficiently rich classes can instead make the error small (Bai et al., 2023; Mondal & Aggarwal, 2024). For a restricted policy class, $\epsilon_{\text{bias}} > 0$ explicitly retains the irreducible representational mismatch rather than hiding it in the optimization error.

Since the Lagrangian advantage $A^{\pi_\theta}_{L,\lambda} = A^{\pi_\theta}_r + \lambda A^{\pi_\theta}_c$ depends explicitly on λ, enlarging the dual interval may increase the compatible-function approximation constant required to hold uniformly over $[0, \lambda_{\max}]$. Consequently, the factor $2/\lambda_{\max}$ appearing later in the feasibility bound should be interpreted as a normalization factor rather than as a mechanism for arbitrarily suppressing the approximation floor. Because $\lambda_{\max}$ is fixed through the uniform Slater condition, the resulting ϵ_{bias} is treated as a problem-dependent constant and may inherit dependence on the Slater margin.

The loss in Eq. 17 is a nonnegative squared residual evaluated under the optimal policy's occupancy density. The loss in Eq. 17 is the transferred counterpart of the critic objective in Eq. 8, obtained by evaluating the same squared residual under the optimal-policy occupancy rather than the current-policy occupancy. The factor $(1-\gamma)^{-1}$ is consistent with the normalized occupancy convention and the natural-gradient definition in Section 4.2.

C.1.5 Discussion on Assumption 2

Ordinary Slater condition. For a fixed CMDP, existence of an interior feasible policy with a positive margin is standard in primal–dual constrained-RL analyses (Bai et al., 2023; Mondal & Aggarwal, 2024; 2026). It supports strong duality and yields a finite bound on the optimal Lagrange multiplier. These references impose Slater conditions for fixed CMDPs; they do not require a margin uniform over a bilevel upper-level parameter.

Uniformity over the upper-level parameter. Our assumption is stronger because the constraint function changes with y. The required uniformity is nevertheless compatible with unconstrained upper-level updates when strict feasibility is enforced structurally through the constraint parameterization rather than through projection of the upper-level variable. A simple sufficient construction is

$$c(s, a; y) = c_0(s, a) + \Delta c(s, a; y),$$

where the perturbation induced by the upper-level parameter is uniformly bounded:

$$|\Delta c(s, a; y)| \le \delta, \qquad \forall (s, a, y).$$

Suppose there exists a stationary policy π_{safe} such that

$$J_{c_0}(\pi_{\text{safe}}) \ge \underline{c} + \frac{\delta}{1-\gamma}$$

for some $\underline{c} > 0$. Then, for every $y \in \mathbb{R}^{d_y}$,

$$J_c(\pi_{\text{safe}}, y) \ge J_{c_0}(\pi_{\text{safe}}) - \frac{\delta}{1-\gamma} \ge \underline{c}.$$

Hence the same policy is strictly feasible with a uniform margin over the entire upper-level parameter space.

Such a construction is natural when y is unconstrained but enters the constraint function through a bounded map, such as a sigmoid or $\tanh$ parameterization. Assumption 2 can therefore be interpreted as a structural requirement ensuring global strict feasibility, rather than as requiring an explicit projection of the upper-level iterate onto a compact set.

C.2 Supporting Lemma for Theorem 2

For a stationary policy π, define its normalized discounted state and state-action occupancy densities by

$$d_\rho^\pi(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_t^\pi(s), \qquad \nu_\rho^\pi(s,a) := d_\rho^\pi(s)\pi(a \mid s), \tag{29}$$

where p_t^π is the density of s_t under $s_0 \sim \rho$ and policy π. These densities satisfy

$$d_\rho^\pi(s') = (1-\gamma)\rho(s') + \gamma\sum_{a\in\mathcal{A}}\int_{\mathcal{S}} p(s' \mid s,a)\nu_\rho^\pi(s,a)\,ds. \tag{30}$$

For the fixed upper-level parameter $v = (x, y)$, the returns are linear in the occupancy density:

$$\begin{aligned} J_r(\pi,x) &= \frac{1}{1-\gamma}\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\nu_\rho^\pi(s,a)r(s,a;x)\,ds, \\ J_c(\pi,y) &= \frac{1}{1-\gamma}\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\nu_\rho^\pi(s,a)c(s,a;y)\,ds. \end{aligned} \tag{31}$$

When v is fixed, $J_g(\pi)$ denotes the corresponding reward or constraint return for $g \in \{r, c\}$.

Lemma 2 (Continuous-state occupancy measures and CMDP duality)**.** *Under Assumptions 4 and 2, and the boundedness part of Assumption 1, for each fixed v the unparameterized CMDP has an optimal stationary randomized policy $\pi^\star$ and a finite optimal multiplier $\lambda^\star(v) \geq 0$. Writing $J_L(\pi,\lambda;v) := J_r(\pi,x) + \lambda J_c(\pi,y)$ for all $\lambda \geq 0$, one has*

$$J_r^\star(v) = \min_{\lambda\geq 0}\max_{\pi} J_L(\pi,\lambda;v) = \max_{\pi} J_L(\pi,\lambda^\star(v);v), \tag{32}$$

and the saddle inequalities

$$J_L(\pi,\lambda^\star(v);v) \leq J_L(\pi^\star,\lambda^\star(v);v) = J_r^\star(v) \leq J_L(\pi^\star,\lambda;v) \tag{33}$$

hold for every stationary randomized π and every $\lambda \geq 0$. Moreover, any average, including an expectation, of discounted occupancy measures is the occupancy measure of a stationary randomized policy. Neither that policy nor $\pi^\star$ is required to belong to the parameterized policy class.

Proof. Only in this proof, we identify densities with their corresponding Borel probability measures. In particular,

$$\rho(E) := \int_E \rho(s)\,ds, \qquad P(E \mid s,a) := \int_E p(s' \mid s,a)\,ds', \qquad \nu_\rho^\pi(E\times\{a\}) := \int_E \nu_\rho^\pi(s,a)\,ds$$

for each Borel set $E \subseteq \mathcal{S}$. Let $\mathcal{D}_\rho$ be the set of probability measures ν on $\mathcal{S}\times\mathcal{A}$ whose state marginal $\nu_{\mathcal{S}}$ satisfies the discounted flow equation

$$\nu_{\mathcal{S}}(E) = (1-\gamma)\rho(E) + \gamma\sum_{a\in\mathcal{A}}\int_{\mathcal{S}} P(E \mid s,a)\,\nu(\mathrm{d}s,\{a\}) \tag{34}$$

for every Borel $E \subseteq \mathcal{S}$. Every probability measure with occupancy density given by Eq. 29 satisfies this equation, by Eq. 30. Conversely, disintegration on the Borel state space gives $\nu(\mathrm{d}s,\{a\}) = \nu_{\mathcal{S}}(\mathrm{d}s)\pi(a \mid s)$ for a Borel stationary randomized policy, defined arbitrarily on a $\nu_{\mathcal{S}}$-null set. For this fixed policy, the right-hand side of the state-marginal flow equation is a contraction with factor γ in total variation. Its unique solution is d_ρ^π, so $\nu = \nu_\rho^\pi$. Thus $\mathcal{D}_\rho$ is exactly the set of stationary-policy occupancy measures. Its defining equations are linear, so it is convex and is also closed under expectations of random occupancy measures, by integration of Eq. 34.

We next prove weak compactness without assuming compactness of $\mathcal{S}$. For each fixed time t, the family of state laws reachable from ρ is uniformly tight over all policies. This follows by induction. The initial law ρ is tight because it is a Borel probability measure on a Borel subset of Euclidean space. For the induction step, choose a compact set carrying arbitrarily high probability uniformly for the time-t laws. On this compact set, weak continuity of P and finiteness of $\mathcal{A}$ make the family of next-state kernels weakly compact, hence uniformly tight. A compact set carrying high probability

for all these kernels then carries high probability for all time-$(t+1)$ laws, up to the probability of leaving the first compact set. This proves the induction claim. The occupancy mass from times $t \geq m$ is γ^m, uniformly in the policy. Combining this tail bound with tightness at the finitely many times $0, \ldots, m-1$ proves uniform tightness of $\mathcal{D}_\rho$. Finally, the flow equation tested against any bounded continuous f is preserved under weak limits, since $Pf(s,a)$ is bounded continuous on $\mathcal{S} \times \mathcal{A}$. Such test functions determine Borel probability measures, so $\mathcal{D}_\rho$ is weakly closed. Prokhorov's theorem now gives weak compactness. Although densities need not be preserved under arbitrary weak limits, they are preserved within this flow-constrained set: for any Lebesgue-null set E, the density assumptions give $\rho(E) = P(E \mid s,a) = 0$ for all (s,a), and Eq. 34 implies $\nu_{\mathcal{S}}(E) = 0$. Each action component is dominated by the state marginal, so every $\nu \in \mathcal{D}_\rho$ admits a joint occupancy density. Consequently, this measure-based compactness argument applies to precisely the density-based policy model used elsewhere in the paper.

For fixed v, define the set of attainable constraint–reward pairs

$$\mathcal{K}_v := \left\{ \left(\frac{\int c(s,a;y)\,\mathrm{d}\nu}{1-\gamma}, \frac{\int r(s,a;x)\,\mathrm{d}\nu}{1-\gamma} \right) : \nu \in \mathcal{D}_\rho \right\}.$$

State continuity from Assumption 4 and boundedness from Assumption 1 make $\mathcal{K}_v$ compact, and linearity makes it convex. Its subset with nonnegative first coordinate is nonempty by Slater feasibility and compact. Maximizing the second coordinate therefore attains $J_r^\star(v)$; disintegration gives an optimal stationary randomized policy $\pi^\star$.

For completeness, separate the convex set $\mathcal{K}_v$ from the disjoint open convex set $\{(q,z) : q > 0,\ z > J_r^\star(v)\}$. The separating-hyperplane theorem gives coefficients $b_c, b_r \geq 0$, not both zero, such that

$$b_c q + b_r z \leq b_r J_r^\star(v) \qquad \text{for every } (q,z) \in \mathcal{K}_v.$$

Nonnegativity follows because the open set is unbounded in each positive coordinate direction. Slater feasibility provides a point of $\mathcal{K}_v$ with $q > 0$, which rules out $b_r = 0$. Thus $\lambda^\star(v) := b_c/b_r$ is finite and nonnegative, and

$$J_r(\pi,x) + \lambda^\star(v) J_c(\pi,y) \leq J_r^\star(v) \qquad \text{for every } \pi.$$

At $\pi^\star$, feasibility and optimality give the reverse inequality, hence equality and complementary slackness $\lambda^\star(v) J_c(\pi^\star, y) = 0$. For any $\lambda \geq 0$, feasibility gives $\max_\pi J_L(\pi,\lambda;v) \geq J_L(\pi^\star,\lambda;v) \geq J_r^\star(v)$. These observations prove Eq. 32 and Eq. 33, including dual attainment. □

The saddle point above belongs to the unparameterized policy–multiplier problem. No existence or convexity assertion is made for a saddle point in (θ,λ) coordinates.

The uniform margin allows one dual projection interval to be used throughout the bilevel algorithm. Let $\lambda^\star(v) \geq 0$ be an optimal multiplier provided by Lemma 2. By Eq. 32,

$$J_r^\star(v) = \max_\pi \{ J_r(\pi,x) + \lambda^\star(v) J_c(\pi,y) \} \geq J_r(\pi_{\mathrm{sl}}(v),x) + \lambda^\star(v) c_0(v).$$

Since $|r(s,a;x)| \leq R$ uniformly by Assumption 1, this gives

$$0 \leq \lambda^\star(v) \leq \frac{J_r^\star(v) - J_r(\pi_{\mathrm{sl}}(v),x)}{c_0(v)} \leq \frac{2R}{(1-\gamma)c_0(v)} \leq \frac{2R}{(1-\gamma)\underline{c}}. \tag{35}$$

Accordingly, fix a finite endpoint satisfying

$$\lambda_{\max} \geq \frac{4R}{(1-\gamma)\underline{c}}. \tag{36}$$

Here $R > 0$ is the reward-magnitude bound. Then $\lambda_{\max} \geq 2\lambda^\star(v)$ for every v, as required by the Slater decoupling argument. Neither $\pi_{\mathrm{sl}}(v)$ nor $c_0(v)$ is used in the lower updates; the algorithm uses only the fixed dual interval. In particular, its endpoint has no explicit dependence on v.

For the remainder of the proof, let

$$b_k := \nabla_\theta J_L(\theta_k, \lambda_k; v), \qquad F_k := F_\rho(\theta_k),$$

and

$$\omega_k^\star = F_k^{-1} b_k.$$

By Assumptions 5–6, there exists a constant $R_\omega < \infty$ such that

$$\|\omega_k^\star\| \le R_\omega \tag{37}$$

uniformly in k.

Properties of the Encoded Lower-Level Oracle. We first show that the signal encoding required by the upper-level analysis preserves the conditional means of the population quantities used by the lower-level critic and dual updates.

Lemma 3 (Unbiasedness of two-point encoding)**.** *Let $a(v)$ be a scalar signal satisfying $|a(v)| \le A$, and let $B \in \{-1, +1\}$ satisfy*

$$\mathbb{P}_v(B = b) = \frac{1}{2} + \frac{b\, a(v)}{4A}, \qquad b \in \{-1, +1\}.$$

If $\widetilde{a} = 2AB$, then

$$\mathbb{E}[\widetilde{a} \mid a(v)] = a(v).$$

Consequently, replacing every sampled reward and constraint signal in the geometric-horizon lower-level oracle by its independently decoded two-point signal preserves the conditional means of the constraint-return and critic-gradient estimators:

$$\mathbb{E}_k[\widehat{J}_{c,\rho}(\theta_k; y)] = J_{c,\rho}(\theta_k; y),$$

and, for every critic vector ω,

$$\mathbb{E}_{k,n}\left[\widehat{\nabla}_\omega \mathcal{L}_\rho(\omega; \theta_k, \lambda_k, v)\right] = \nabla_\omega \mathcal{L}_\rho(\omega; \theta_k, \lambda_k, v).$$

Here $\mathbb{E}_k[\cdot]$ denotes expectation conditional on the lower history up to policy iteration k, and $\mathbb{E}_{k,n}[\cdot]$ further conditions on the critic history up to inner iteration n.

Proof. By definition,

$$\mathbb{E}[\widetilde{a}] = 2A\left[\mathbb{P}(B = 1) - \mathbb{P}(B = -1)\right].$$

Substituting the encoding probabilities gives

$$\mathbb{E}[\widetilde{a}] = 2A\left[\left(\frac{1}{2} + \frac{a(v)}{4A}\right) - \left(\frac{1}{2} - \frac{a(v)}{4A}\right)\right] = a(v).$$

Condition on an arbitrary realized state-action trajectory. Every decoded reward and constraint signal is conditionally unbiased for its corresponding raw signal. Since the return, Q-function, V-function, and advantage estimators used by the critic oracle are linear combinations of these signals, iterated expectation gives

$$\mathbb{E}[\widehat{A}_r^{\pi_\theta}(s, a) \mid s, a] = A_r^{\pi_\theta}(s, a)$$

and

$$\mathbb{E}[\widehat{A}_c^{\pi_\theta}(s, a) \mid s, a] = A_c^{\pi_\theta}(s, a).$$

Therefore,

$$\mathbb{E}[\widehat{A}_{L,\lambda}^{\pi_\theta}(s, a) \mid s, a] = A_{L,\lambda}^{\pi_\theta}(s, a).$$

The Fisher estimator does not depend on the reward or constraint signals. Hence

$$\mathbb{E}[\widehat{F}_\rho(\theta)] = F_\rho(\theta).$$

Combining the two identities with

$$\widehat{\nabla}_\omega \mathcal{L}_\rho = \widehat{F}_\rho(\theta)\omega - \frac{1}{1-\gamma}\widehat{H}_\rho(\theta, \lambda)$$

yields

$$\mathbb{E}[\widehat{\nabla}_\omega \mathcal{L}_\rho(\omega; \theta, \lambda, v)] = F_\rho(\theta)\omega - \nabla_\theta J_L(\theta, \lambda; v),$$

which is precisely $\nabla_\omega \mathcal{L}_\rho(\omega; \theta, \lambda, v)$. The constraint-return identity follows analogously. □

The encoding increases the variance of the lower oracle but does not change its mean. The following lemma records the moment property needed below.

Lemma 4 (Second moment of the stochastic critic error)**.** *There exist finite constants σ_0^2, σ_1^2, independent of k, n, such that*

$$\mathbb{E}_{k,n}\left[\left\|\widehat{\nabla}_\omega \mathcal{L}_\rho(\omega;\theta_k,\lambda_k,v) - \nabla_\omega \mathcal{L}_\rho(\omega;\theta_k,\lambda_k,v)\right\|^2\right] \le \sigma_0^2 + \sigma_1^2\|\omega-\omega_k^\star\|^2.$$

Moreover, the encoded constraint-return estimator satisfies

$$\mathbb{E}_k\left[\widehat{J}_{c,\rho}(\theta_k;y)^2\right] \le C_J^2$$

for a finite constant C_J.

Proof. Write

$$\widehat{g}_k(\omega) = \widehat{F}_k\omega - \widehat{b}_k, \qquad g_k(\omega) = F_k\omega - b_k.$$

Since

$$\omega = (\omega-\omega_k^\star) + \omega_k^\star,$$

we have

$$\widehat{g}_k(\omega) - g_k(\omega) = (\widehat{F}_k - F_k)(\omega-\omega_k^\star) + \left(\widehat{F}_k\omega_k^\star - \widehat{b}_k\right),$$

where

$$F_k\omega_k^\star - b_k = 0.$$

Therefore,

$$\|\widehat{g}_k(\omega) - g_k(\omega)\|^2 \le 2\|\widehat{F}_k - F_k\|^2\|\omega-\omega_k^\star\|^2 + 2\|\widehat{F}_k\omega_k^\star - \widehat{b}_k\|^2.$$

Assumption 5 gives

$$\|\widehat{F}_k\| \le G_\pi^2, \qquad \|F_k\| \le G_\pi^2,$$

and consequently

$$\|\widehat{F}_k - F_k\| \le 2G_\pi^2.$$

Thus the first term is bounded by

$$8G_\pi^4\|\omega-\omega_k^\star\|^2.$$

The decoded signals are bounded by $2R$ and $2C$, respectively. The geometric horizon has finite moments of every fixed order for $\gamma < 1$. Together with the bounded score function and Eq. 37, this implies the existence of a finite uniform constant $\sigma_\star^2$ satisfying

$$\mathbb{E}_k\left[\|\widehat{F}_k\omega_k^\star - \widehat{b}_k\|^2\right] \le \sigma_\star^2.$$

Hence the claim follows with, for example,

$$\sigma_1^2 = 8G_\pi^4, \qquad \sigma_0^2 = 2\sigma_\star^2.$$

The same bounded-signal and geometric-horizon argument yields a finite uniform second-moment bound C_J^2 for $\widehat{J}_{c,\rho}(\theta_k;y)$. □

Vanilla-SGD Natural-Gradient Estimation. We now establish the vanilla-SGD analogue of the accelerated critic analysis in Mondal & Aggarwal (2024b).

Lemma 5 (Bias and mean-square error of the vanilla-SGD critic)**.** *Fix a lower iteration k. Let*

$$\omega_{k,n+1} = \omega_{k,n} - \beta\widehat{\nabla}_\omega \mathcal{L}_\rho(\omega_{k,n};\theta_k,\lambda_k,v), \qquad \omega_{k,0} = 0.$$

Suppose

$$0 < \beta \le \min\left\{\frac{1}{G_\pi^2}, \frac{\mu_F}{\mu_F^2+\sigma_1^2}\right\}.$$

Then the last iterate

$$\widehat{\omega}_k = \omega_{k,N}$$

satisfies

$$\|\mathbb{E}_k[\widehat{\omega}_k] - \omega_k^\star\| \le e^{-\beta\mu_F N}R_\omega, \tag{38}$$

and

$$\mathbb{E}_k\left[\|\widehat{\omega}_k - \omega_k^\star\|^2\right] \le (1-\beta\mu_F)^N R_\omega^2 + \frac{\beta\sigma_0^2}{\mu_F}. \tag{39}$$

Consequently,

$$\mathbb{E}_k\left[\|\widehat{\omega}_k\|^2\right] \le W_\omega^2, \tag{40}$$

where

$$W_\omega^2 := 4R_\omega^2 + \frac{2\beta\sigma_0^2}{\mu_F}. \tag{41}$$

Proof. Define

$$e_{k,n} = \omega_{k,n} - \omega_k^\star.$$

Since

$$\nabla_\omega \mathcal{L}_\rho(\omega;\theta_k,\lambda_k,v) = F_k(\omega - \omega_k^\star),$$

the SGD recursion can be written as

$$e_{k,n+1} = (I - \beta F_k)e_{k,n} - \beta\xi_{k,n},$$

where

$$\xi_{k,n} = \widehat{\nabla}_\omega \mathcal{L}_\rho(\omega_{k,n};\theta_k,\lambda_k,v) - \nabla_\omega \mathcal{L}_\rho(\omega_{k,n};\theta_k,\lambda_k,v).$$

By Lemma 3,

$$\mathbb{E}_{k,n}[\xi_{k,n}] = 0.$$

We first analyze the conditional mean. Since $e_{k,n}$ is measurable with respect to the critic history before drawing the fresh sample at iteration n,

$$\mathbb{E}_k[e_{k,n+1}] = (I - \beta F_k)\mathbb{E}_k[e_{k,n}].$$

Iterating gives

$$\mathbb{E}_k[e_{k,N}] = (I - \beta F_k)^N e_{k,0}.$$

Since

$$\mu_F I \preceq F_k \preceq G_\pi^2 I$$

and $\beta \le G_\pi^{-2}$,

$$\|I - \beta F_k\| \le 1 - \beta\mu_F.$$

Therefore,

$$\|\mathbb{E}_k[e_{k,N}]\| \le (1-\beta\mu_F)^N \|e_{k,0}\|.$$

Using $\omega_{k,0} = 0$ and $\|\omega_k^\star\| \le R_\omega$,

$$\|\mathbb{E}_k[\widehat{\omega}_k] - \omega_k^\star\| \le (1-\beta\mu_F)^N R_\omega \le e^{-\beta\mu_F N} R_\omega,$$

which proves Eq. 38.

For the mean-square error,

$$\mathbb{E}_{k,n}\|e_{k,n+1}\|^2 = \|(I-\beta F_k)e_{k,n}\|^2 + \beta^2 \mathbb{E}_{k,n}\|\xi_{k,n}\|^2,$$

because the cross term vanishes conditionally. Applying Lemma 4,

$$\mathbb{E}_{k,n}\|e_{k,n+1}\|^2 \le (1-\beta\mu_F)^2\|e_{k,n}\|^2 + \beta^2\left(\sigma_0^2 + \sigma_1^2\|e_{k,n}\|^2\right).$$

Hence

$$\mathbb{E}_{k,n}\|e_{k,n+1}\|^2 \le \left[1 - 2\beta\mu_F + \beta^2(\mu_F^2 + \sigma_1^2)\right]\|e_{k,n}\|^2 + \beta^2\sigma_0^2.$$

The stepsize condition

$$\beta(\mu_F^2 + \sigma_1^2) \le \mu_F$$

implies

$$\mathbb{E}_{k,n}\|e_{k,n+1}\|^2 \le (1-\beta\mu_F)\|e_{k,n}\|^2 + \beta^2\sigma_0^2.$$

Taking total expectation and iterating yields

$$\mathbb{E}_k\|e_{k,N}\|^2 \le (1-\beta\mu_F)^N\|e_{k,0}\|^2 + \beta^2\sigma_0^2\sum_{j=0}^{N-1}(1-\beta\mu_F)^j.$$

Using

$$\sum_{j=0}^{N-1}(1-\beta\mu_F)^j \le \frac{1}{\beta\mu_F}$$

and $\|e_{k,0}\| \le R_\omega$ gives

$$\mathbb{E}_k\|e_{k,N}\|^2 \le (1-\beta\mu_F)^N R_\omega^2 + \frac{\beta\sigma_0^2}{\mu_F},$$

which proves Eq. 39.

Finally,

$$\|\widehat{\omega}_k\|^2 \le 2\|\widehat{\omega}_k - \omega_k^\star\|^2 + 2\|\omega_k^\star\|^2.$$

Using Eq. 39 and $\|\omega_k^\star\| \le R_\omega$ yields

$$\mathbb{E}_k\|\widehat{\omega}_k\|^2 \le 4R_\omega^2 + \frac{2\beta\sigma_0^2}{\mu_F} = W_\omega^2.$$

□

Remark 1. *The distinction between Eq. 38 and Eq. 39 is important. The conditional bias decreases geometrically with the number N of critic iterations, whereas the mean-square error may converge only to a constant noise floor. This is sufficient for the lower-level CMDP analysis because the conditional bias appears as a first-order error term, while the second moment of $\widehat{\omega}_k$ is multiplied by the policy stepsize η.*

Local-to-Global Lagrangian Comparison

Define

$$D_k = \int_{\mathcal{S}} d_\rho^{\pi^\star}(s) D_{\mathrm{KL}}\left(\pi^\star(\cdot \mid s) \| \pi_{\theta_k}(\cdot \mid s)\right) ds.$$

Since $\mathcal{A}$ is finite, the KL divergence here is the finite action sum, weighted by the state occupancy density and integrated with ds. Assumption 5 ensures

$$0 \le D_0 \le \log(1/p_0),$$

uniformly in v and in the optimal comparator $\pi^\star$. The bounded log-policy gradients also keep these integrals finite at every finite iterate.

Lemma 6 (Local-to-global comparison)**.** *Under Assumptions 5–7, the iterates of Algorithm 2 satisfy*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[J_L(\pi^\star,\lambda_k;v) - J_L(\theta_k,\lambda_k;v)\right] \le \sqrt{\epsilon_{\mathrm{bias}}} + \frac{G_\pi}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\|\mathbb{E}_k[\widehat{\omega}_k] - \omega_k^\star\|\right] + \frac{L_\pi\eta}{2K}\sum_{k=0}^{K-1}\mathbb{E}\|\widehat{\omega}_k\|^2 + \frac{D_0}{\eta K}. \tag{42}$$

Proof. The L_π-smoothness of $\log\pi_\theta(a|s)$ gives

$$\log\pi_{\theta_{k+1}}(a|s) \ge \log\pi_{\theta_k}(a|s) + \eta\nabla_\theta\log\pi_{\theta_k}(a|s)^\top\widehat{\omega}_k - \frac{L_\pi\eta^2}{2}\|\widehat{\omega}_k\|^2.$$

Taking expectation over $s \sim d_\rho^{\pi^\star}$ and $a \sim \pi^\star(\cdot|s)$ yields

$$D_{k+1} \le D_k - \eta\mathbb{E}_{\nu_\rho^{\pi^\star}}\left[\nabla_\theta\log\pi_{\theta_k}(a|s)^\top\widehat{\omega}_k\right] + \frac{L_\pi\eta^2}{2}\|\widehat{\omega}_k\|^2.$$

Conditioning on the history up to iteration k,

$$\mathbb{E}_k[D_{k+1}] \le D_k - \eta\mathbb{E}_{\nu_\rho^{\pi^\star}}\left[\nabla_\theta\log\pi_{\theta_k}(a|s)^\top\mathbb{E}_k[\widehat{\omega}_k]\right] + \frac{L_\pi\eta^2}{2}\mathbb{E}_k\|\widehat{\omega}_k\|^2.$$

Add and subtract $\omega_k^\star$. Assumption 5 gives

$$\left|\mathbb{E}_{\nu_\rho^{\pi^\star}}\left[\nabla_\theta\log\pi_{\theta_k}(a|s)^\top\left(\mathbb{E}_k[\widehat{\omega}_k] - \omega_k^\star\right)\right]\right| \le G_\pi\left\|\mathbb{E}_k[\widehat{\omega}_k] - \omega_k^\star\right\|.$$

The performance-difference identity together with the compatible-function approximation assumption implies

$$J_L(\pi^\star, \lambda_k; v) - J_L(\theta_k, \lambda_k; v) \le \mathbb{E}_{\nu_\rho^{\pi^\star}} \left[\nabla_\theta \log \pi_{\theta_k}(a|s)^\top \omega_k^\star\right] + \sqrt{\epsilon_{\text{bias}}}.$$

Substituting the previous two inequalities gives

$$\eta \mathbb{E}_k \left[J_L(\pi^\star, \lambda_k; v) - J_L(\theta_k, \lambda_k; v)\right] \le D_k - \mathbb{E}_k[D_{k+1}] + \eta\sqrt{\epsilon_{\text{bias}}}$$

$$+\eta G_\pi \left\|\mathbb{E}_k[\widehat{\omega}_k] - \omega_k^\star\right\| + \frac{L_\pi \eta^2}{2} \mathbb{E}_k \|\widehat{\omega}_k\|^2.$$

Summing over $k = 0, \dots, K-1$, taking total expectations, using $D_K \ge 0$, and dividing by ηK proves Eq. 42. □

Combining Lemmas 5 and 6 gives the following immediate result.

Corollary 1 (Average Lagrangian convergence)**.** *Under the assumptions above,*

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[J_L(\pi^\star, \lambda_k; v) - J_L(\theta_k, \lambda_k; v)\right] \le \sqrt{\epsilon_{\text{bias}}} + G_\pi R_\omega e^{-\beta \mu_F N} + \frac{L_\pi \eta}{2} W_\omega^2 + \frac{D_0}{\eta K}. \quad (43)$$

Proof. Apply Eq. 38 and Eq. 40 uniformly in k to Eq. 42. □

For convenience, define

$$\Delta_{K,N} := \sqrt{\epsilon_{\text{bias}}} + G_\pi R_\omega e^{-\beta \mu_F N} + \frac{L_\pi \eta}{2} W_\omega^2 + \frac{D_0}{\eta K}. \quad (44)$$

Projected Dual Recursion

The projected Lagrange-multiplier update is

$$\lambda_{k+1} = \Pi_{[0, \lambda_{\max}]} \left[\lambda_k - \zeta \widehat{J}_{c,\rho}(\theta_k; y)\right].$$

Lemma 7 (Dual cross-term bounds)**.** *Suppose* $\lambda_0 = 0$*. Then*

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\lambda_k J_c(\theta_k, y)\right] \le \frac{\zeta C_J^2}{2}, \quad (45)$$

and

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[(\lambda_k - \lambda_{\max}) J_c(\theta_k, y)\right] \le \frac{\lambda_{\max}^2}{2\zeta K} + \frac{\zeta C_J^2}{2}. \quad (46)$$

Proof. By non-expansiveness of Euclidean projection and $\lambda_0 = 0$,

$$\lambda_{k+1}^2 \le \left(\lambda_k - \zeta \widehat{J}_{c,\rho}(\theta_k; y)\right)^2.$$

Therefore,

$$\lambda_{k+1}^2 - \lambda_k^2 \le -2\zeta \lambda_k \widehat{J}_{c,\rho}(\theta_k; y) + \zeta^2 \widehat{J}_{c,\rho}(\theta_k; y)^2.$$

Summing and using $\lambda_K^2 \ge 0$ gives

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\left[\lambda_k \widehat{J}_{c,\rho}(\theta_k; y)\right] \le \frac{\zeta}{2K} \sum_{k=0}^{K-1} \mathbb{E}\left[\widehat{J}_{c,\rho}(\theta_k; y)^2\right].$$

Lemma 3 and conditional independence imply

$$\mathbb{E}[\lambda_k \widehat{J}_{c,\rho}(\theta_k; y)] = \mathbb{E}[\lambda_k J_c(\theta_k, y)].$$

Lemma 4 then yields Eq. 45.

For the second claim, use non-expansiveness relative to the feasible point $\lambda_{\max}$:

$$|\lambda_{k+1} - \lambda_{\max}|^2 \le |\lambda_k - \zeta \widehat{J}_{c,\rho}(\theta_k; y) - \lambda_{\max}|^2.$$

Expanding gives

$$|\lambda_{k+1} - \lambda_{\max}|^2 \le |\lambda_k - \lambda_{\max}|^2 - 2\zeta(\lambda_k - \lambda_{\max})\widehat{J}_{c,\rho}(\theta_k; y) + \zeta^2 \widehat{J}_{c,\rho}(\theta_k; y)^2.$$

Summing, using $|\lambda_0 - \lambda_{\max}|^2 = \lambda_{\max}^2$ and dropping the nonnegative terminal distance yields

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[(\lambda_k - \lambda_{\max}) J_c(\theta_k, y)] \le \frac{\lambda_{\max}^2}{2\zeta K} + \frac{\zeta C_J^2}{2}.$$

□

Reward and Constraint Convergence

Lemma 8 (Reward suboptimality)**.** *The lower iterates satisfy*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[J_r^\star(v) - J_r(\theta_k, x)\right] \le \Delta_{K,N} + \frac{\zeta C_J^2}{2}. \tag{47}$$

Proof. Since $\pi^\star$ is feasible,

$$J_c(\pi^\star, y) \ge 0.$$

Therefore,

$$J_L(\pi^\star, \lambda_k; v) - J_L(\theta_k, \lambda_k; v) \ge J_r^\star(v) - J_r(\theta_k, x) - \lambda_k J_c(\theta_k, y).$$

Rearranging, averaging, and applying Corollary 1 gives

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[J_r^\star(v) - J_r(\theta_k, x)] \le \Delta_{K,N} + \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\lambda_k J_c(\theta_k, y)].$$

Apply Eq. 45. □

We next recall the standard Slater-based decoupling result.

Lemma 9 (Slater decoupling)**.** *For the fixed upper-level parameter v, write $\lambda^\star = \lambda^\star(v)$ for an optimal multiplier supplied by Lemma 2. Suppose $C_{\text{pen}} > 0$, $C_{\text{pen}} \ge 2\lambda^\star$, and $\delta \ge 0$. If a stationary randomized policy $\bar{\pi}$ satisfies*

$$J_r^\star(v) - J_r(\bar{\pi}, x) + C_{\text{pen}}[-J_c(\bar{\pi}, y)] \le \delta,$$

then

$$-J_c(\bar{\pi}, y) \le \frac{2\delta}{C_{\text{pen}}}.$$

Proof. If $J_c(\bar{\pi}, y) \ge 0$, the claim is immediate. Suppose $J_c(\bar{\pi}, y) < 0$ and set

$$\tau = J_c(\bar{\pi}, y).$$

Define

$$V(\tau) = \max_{\pi} \{J_r(\pi, x) : J_c(\pi, y) \ge \tau\}.$$

Since $\bar{\pi}$ is feasible for this relaxed problem,

$$J_r(\bar{\pi}, x) \le V(\tau).$$

By weak duality for the relaxed problem and the attained strong-duality identity Eq. 32 for the original CMDP,

$$V(\tau) \le J_r^\star(v) - \lambda^\star \tau.$$

Therefore,

$$J_r^\star(v) - J_r(\bar{\pi}, x) \ge \lambda^\star \tau.$$

Since $\tau < 0$,

$$(C_{\mathrm{pen}} - \lambda^\star)(-\tau) \le J_r^\star(v) - J_r(\bar{\pi}, x) + C_{\mathrm{pen}}(-\tau) \le \delta.$$

Because $C_{\mathrm{pen}} \ge 2\lambda^\star$,

$$C_{\mathrm{pen}} - \lambda^\star \ge \frac{C_{\mathrm{pen}}}{2},$$

and hence

$$-J_c(\bar{\pi}, y) = -\tau \le \frac{2\delta}{C_{\mathrm{pen}}}.$$

□

C.3 Proof of Theorem 2

Theorem 4 (Approximate lower-level response with vanilla SGD)**.** *Suppose Assumptions 4–7 and 2 hold, and suppose the boundedness part of Assumption 1 holds. Fix $\lambda_{\max}$ according to Eq. 36. Let Algorithm 2 use*

$$\eta = \frac{\eta_0}{\sqrt{K}}, \qquad \zeta = \frac{\zeta_0}{\sqrt{K}},$$

for positive constants η_0, ζ_0, choose the critic stepsize β to satisfy the condition in Lemma 5, and choose the number of critic iterations so that

$$N \ge \max\left\{1, \left\lceil \frac{1}{\beta\mu_F} \log\left(R_\omega \sqrt{K}\right) \right\rceil \right\}. \tag{48}$$

Then there exist problem-dependent constants $C_r, C_c > 0$, independent of K and N, such that

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[J_r^\star(v) - J_r(\theta_k, x)\right] \le \sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_r}{\sqrt{K}}, \tag{49}$$

and

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[-J_c(\theta_k, y)\right] \le C_{\mathrm{sl}}\sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_c}{\sqrt{K}}, \tag{50}$$

where one may take $C_{\mathrm{sl}} = 2/\lambda_{\max}$. Thus C_{sl} is independent of v; the dependence of the bounds on Slater feasibility enters through the common lower bound $\underline{c}$ and the fixed dual endpoint $\lambda_{\max}$, rather than a separately chosen endpoint for each v.

If

$$R \sim \mathrm{Unif}\{0, \dots, K-1\}$$

is selected independently after the lower run and $Z \sim P_v$ denotes the resulting complete trace, then $\theta_R = \theta(Z)$ and

$$\mathbb{E}\left[J_r^\star(v) - J_r(\theta_R, x)\right] \le \sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_r}{\sqrt{K}}, \tag{51}$$

and

$$\mathbb{E}\left[-J_c(\theta_R, y)\right] \le C_{\mathrm{sl}}\sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_c}{\sqrt{K}}. \tag{52}$$

Proof. By Eq. 48,

$$R_\omega e^{-\beta\mu_F N} \le \frac{1}{\sqrt{K}}.$$

Substituting $\eta = \eta_0/\sqrt{K}$ into Eq. 44 gives

$$\Delta_{K,N} \le \sqrt{\epsilon_{\mathrm{bias}}} + \frac{G_\pi}{\sqrt{K}} + \frac{L_\pi \eta_0 W_\omega^2}{2\sqrt{K}} + \frac{D_0}{\eta_0\sqrt{K}}.$$

Hence there exists $C_L > 0$ such that

$$\Delta_{K,N} \le \sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_L}{\sqrt{K}}.$$

Lemma 8 and $\zeta = \zeta_0/\sqrt{K}$ imply

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[J_r^\star(v) - J_r(\theta_k, x)] \le \sqrt{\epsilon_{\text{bias}}} + \frac{1}{\sqrt{K}}\left(C_L + \frac{\zeta_0 C_J^2}{2}\right).$$

This proves Eq. 49.

For feasibility, combining Corollary 1 with Eq. 46 gives

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[J_r^\star(v) - J_r(\theta_k, x) + \lambda_{\max}(-J_c(\theta_k, y))\right] \le \Delta_{K,N} + \frac{\lambda_{\max}^2}{2\zeta K} + \frac{\zeta C_J^2}{2}.$$

With $\zeta = \zeta_0/\sqrt{K}$,

$$\frac{\lambda_{\max}^2}{2\zeta K} + \frac{\zeta C_J^2}{2} = \frac{1}{\sqrt{K}}\left(\frac{\lambda_{\max}^2}{2\zeta_0} + \frac{\zeta_0 C_J^2}{2}\right).$$

Hence the right-hand side is

$$\sqrt{\epsilon_{\text{bias}}} + O(K^{-1/2}).$$

For this fixed v, form the deterministic expected average occupancy density

$$\bar{\nu}(s,a) := \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\nu_\rho^{\pi_{\theta_k}}(s,a)], \qquad \bar{d}(s) := \sum_{a\in\mathcal{A}}\bar{\nu}(s,a).$$

Define $\bar{\pi}(a \mid s) = \bar{\nu}(s,a)/\bar{d}(s)$ wherever $\bar{d}(s) > 0$, and choose the uniform action distribution elsewhere. The averaged density satisfies the linear flow identity Eq. 30. By Lemma 2, it is the occupancy density of $\bar{\pi}$. Boundedness, Fubini's theorem, and linearity of the returns in Eq. 31 give

$$J_g(\bar{\pi}) = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[J_g(\theta_k)], \qquad g \in \{r, c\}.$$

Thus the preceding expected combined-gap bound is a deterministic bound for $\bar{\pi}$, to which the decoupling lemma applies directly. By Eq. 35–Eq. 36, the fixed endpoint satisfies $\lambda_{\max} \ge 2\lambda^\star(v)$. Applying Lemma 9 with $C_{\text{pen}} = \lambda_{\max}$ yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[-J_c(\theta_k, y)] \le C_{\text{sl}}\sqrt{\epsilon_{\text{bias}}} + \frac{C_c}{\sqrt{K}}$$

for a finite problem-dependent C_c. This proves Eq. 50.

Finally, because R is uniform and independent,

$$\mathbb{E}[J_r^\star(v) - J_r(\theta_R, x)] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[J_r^\star(v) - J_r(\theta_k, x)],$$

and

$$\mathbb{E}[-J_c(\theta_R, y)] = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[-J_c(\theta_k, y)].$$

The bounds Eq. 51 and Eq. 52 follow immediately. □

Corollary 2 (Lower-level iteration and sample complexity)**.** *For any target numerical accuracy $\delta_\ell > 0$, choose*

$$K = \Theta(\delta_\ell^{-2})$$

and

$$N = \Theta(\log(\delta_\ell^{-1})).$$

Define the effective response accuracy by $\epsilon_\ell := \max\{\sqrt{\epsilon_{\text{bias}}}, \delta_\ell\}$ and let $\overline{C}_r := 1 + C_r$ and $\overline{C}_c := C_{\text{sl}} + C_c$. Then

$$\mathbb{E}\left[J_r^\star(v) - J_r(\theta_R, x)\right] \le \overline{C}_r\epsilon_\ell,$$

and

$$\mathbb{E}\left[-J_c(\theta_R, y)\right] \le \overline{C}_c \epsilon_\ell.$$

Since each stochastic critic call uses a geometric-horizon trajectory with expected length $O((1-\gamma)^{-1})$, the total number of lower-level environment transitions is

$$\widetilde{\mathcal{O}}\left(\frac{KN}{1-\gamma}\right) = \widetilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)\delta_\ell^2}\right),$$

up to the additional polynomial dependence on $(1-\gamma)^{-1}$ contained in the problem-dependent constants and stepsizes.

Proof. The iteration choices follow directly from Theorem 4. Since $K = \Theta(\delta_\ell^{-2})$, we have $K^{-1/2} = O(\delta_\ell)$, while the concrete choice $N = \Theta(\log K) = \Theta(\log(\delta_\ell^{-1}))$ satisfies Eq. 48. The effective bounds follow because both $\sqrt{\epsilon_{\text{bias}}}$ and δ_ℓ are bounded by ϵ_ℓ.

Each critic iteration invokes one geometric-horizon stochastic oracle whose expected trajectory length is proportional to $(1-\gamma)^{-1}$. Summing over the N critic calls at each of the K primal-dual iterations gives the stated order, with logarithmic factors absorbed into $\widetilde{\mathcal{O}}(\cdot)$. □

D Proof of Theorem 1

Exact Gradient of the Approximate-Response Objective.

The lower initialization, iteration counts, step sizes, dual interval, and decoding bounds are fixed independently of v for each prescribed numerical accuracy δ_ℓ and corresponding effective response accuracy $\epsilon_\ell = \max\{\sqrt{\epsilon_{\text{bias}}}, \delta_\ell\}$.

Lemma 10 (Training-distribution gradient)**.** *Under Assumptions 1 and 3, the objective*

$$\Phi_{\epsilon_\ell}(v) = \mathbb{E}_{Z\sim P_v}\left[G(v, \theta(Z))\right]$$

is differentiable and, for any deterministic scalar baseline $c(v)$, satisfies

$$\nabla\Phi_{\epsilon_\ell}(v) = \mathbb{E}_v\left[\nabla_v G(v,\theta(Z)) + \big(G(v,\theta(Z)) - c(v)\big)S(Z;v)\right]. \tag{53}$$

Here $S(Z;v)$ is the sum of the score functions of all reward and constraint encoding variables generated during the lower-level training run.

Proof. Let

$$\mathcal{T} = (T_1, \ldots, T_{3K(N+1)})$$

collect all geometric horizons sampled during one complete execution of the lower-level algorithm. The number of horizon variables is deterministic, and every T_i is sampled independently of the upper-level parameter v. Therefore,

$$\Phi_{\epsilon_\ell}(v) = \mathbb{E}_{\mathcal{T}}\left[\Phi_{\mathcal{T}}(v)\right], \tag{54}$$

where

$$\Phi_{\mathcal{T}}(v) := \mathbb{E}_v\left[G(v,\theta(Z)) \mid \mathcal{T}\right].$$

Fix an arbitrary realization

$$\mathcal{T} = t.$$

Conditional on this realization, the lower-level training procedure has a finite number of state-action samples and therefore a finite number of reward and constraint encoding events.

Let $N_r(t)$ and $N_c(t)$ denote the resulting numbers of reward and constraint encoding events, respectively, and define

$$M(t) := N_r(t) + N_c(t).$$

Enumerate these $M(t)$ encoding variables as

$$B_1, \ldots, B_{M(t)}$$

in chronological order. For each encoding event j, let

$$q_j(B_j; v)$$

denote the conditional probability of the realized encoding sign given the training history immediately before that encoding event.

By construction, the lower-level algorithm receives only the decoded signals

$$\widetilde{r} = 2RB^r, \qquad \widetilde{c} = 2CB^c.$$

Hence, once the complete sampled trace Z is fixed, all lower-level critic, policy, and dual iterates are fixed functions of that trace. In particular, $\theta(Z)$ has no additional explicit dependence on v.

The initial-state distribution, transition kernel, geometric-horizon distribution, final random-index distribution, and all other auxiliary randomness are independent of v. Furthermore, conditional on a fixed trace, the policy probabilities are evaluated at policy parameters reconstructed from the fixed decoded signals and are therefore also fixed with respect to v.

Consequently, the dependence of the conditional trace distribution on v arises only through the encoding probabilities. Thus its score is

$$\nabla_v \log P_v(Z \mid \mathcal{T} = t) = \sum_{j=1}^{M(t)} \nabla_v \log q_j(B_j; v) = S(Z; v). \tag{55}$$

Because $M(t) < \infty$ for every fixed t, standard likelihood-ratio differentiation gives

$$\nabla \Phi_t(v) = \nabla_v \mathbb{E}_v \left[G(v, \theta(Z)) \mid \mathcal{T} = t \right] \tag{56}$$

$$= \mathbb{E}_v \left[\nabla_v G(v, \theta(Z)) + G(v, \theta(Z)) S(Z; v) \mid \mathcal{T} = t \right]. \tag{57}$$

It remains to remove the conditioning on the geometric horizons. By Assumption 3,

$$\|\nabla_v G(v, \theta)\| \le L_{G,0}, \qquad |G(v, \theta)| \le C_G,$$

where

$$C_G := \max\{|G_{\min}|, |G_{\max}|\}.$$

Moreover, each encoding score satisfies

$$\|s_{r,j}\| \le \frac{R_1}{R}, \qquad \|s_{c,j}\| \le \frac{C_1}{C}.$$

Hence, for

$$\kappa := \max\left\{ \frac{R_1}{R}, \frac{C_1}{C} \right\},$$

we have

$$\|S(Z; v)\| \le \kappa M(\mathcal{T}).$$

Therefore,

$$\|\nabla \Phi_{\mathcal{T}}(v)\| \le L_{G,0} + C_G \kappa M(\mathcal{T}).$$

The total number of encoding events is bounded by

$$M(\mathcal{T}) \le 2 \sum_{i=1}^{3K(N+1)} (T_i + 1).$$

Since

$$\mathbb{E}[T_i + 1] = \frac{1}{1 - \gamma},$$

we have

$$\mathbb{E}[M(\mathcal{T})] < \infty.$$

Thus

$$L_{G,0} + C_G \kappa M(\mathcal{T})$$

is integrable with respect to the distribution of $\mathcal{T}$. We may therefore differentiate Eq. 54 under the expectation over $\mathcal{T}$. Using Eq. 57,

$$\nabla\Phi_{\epsilon_\ell}(v) = \mathbb{E}_{\mathcal{T}}\left[\nabla\Phi_{\mathcal{T}}(v)\right] \tag{58}$$

$$= \mathbb{E}_v\left[\nabla_v G(v,\theta(Z)) + G(v,\theta(Z))S(Z;v)\right]. \tag{59}$$

To make the last expectation explicit, set $m := 3K(N+1)$. The unconditional trace law is the total-probability mixture

$$P_v(Z) = \sum_{t\in\mathbb{N}_0^m} \mathbb{P}(\mathcal{T}=t)P_v(Z \mid \mathcal{T}=t), \qquad \mathbb{P}(\mathcal{T}=t) = \prod_{i=1}^{m}(1-\gamma)\gamma^{t_i},$$

where the first equality is understood as an equality of probability measures. Thus the second equality in Eq. 59 is precisely the law of total expectation applied to the conditional expression in Eq. 57. Since the distribution of $\mathcal{T}$ is independent of v, it contributes no term to the score of P_v.

Finally, every individual encoding score has conditional mean zero. Therefore, by iterated expectation,

$$\mathbb{E}_v[S(Z;v)] = 0.$$

Hence, for any deterministic scalar $c(v)$,

$$\mathbb{E}_v[c(v)S(Z;v)] = c(v)\mathbb{E}_v[S(Z;v)] = 0.$$

Subtracting this zero term from Eq. 59 yields

$$\nabla\Phi_{\epsilon_\ell}(v) = \mathbb{E}_v\left[\nabla_v G(v,\theta(Z)) + \big(G(v,\theta(Z)) - c(v)\big)S(Z;v)\right],$$

which proves the result. □

E Proof of Theorem 3

We now analyze the convergence of Algorithm 1. The upper-level objective is induced by the distribution of the complete stochastic lower-level training procedure. In contrast to an iterative-differentiation approach, we do not differentiate through the policy, critic, or projected dual updates of the lower-level algorithm.

For a fixed upper-level parameter v, let $Z \sim P_v$ denote the complete random trace of Algorithm 2, including all environment trajectories, encoded reward and constraint signals, critic samples, and the final random lower-iterate selection. Here P_v is the trace law induced by the upper-level parameter. Let

$$\theta(Z)$$

denote the resulting policy parameter. The upper objective is

$$\Phi_{\epsilon_\ell}(v) := \mathbb{E}_{Z\sim P_v}\left[G(v,\theta(Z))\right], \tag{60}$$

where the parameters of the lower-level algorithm are chosen to achieve lower numerical accuracy δ_ℓ, inducing effective response accuracy $\epsilon_\ell = \max\{\sqrt{\epsilon_{\text{bias}}}, \delta_\ell\}$.

Once the complete encoded trace Z is fixed, the output $\theta(Z)$ is fixed. All dependence on v through the lower-level algorithm therefore enters through the probabilities of the encoded reward and constraint signals.

E.1 Justification for Assumptions for Theorem 3

E.1.1 Discussion on Assumption 1

Remark 2 (A normalized architecture satisfying Assumption 1)**.** *The signal conditions above are stated for a general parameterized function-approximation class. As one concrete sufficient example, let the unconstrained parameter be $x \in \mathbb{R}^{d_x}$, define the bounded effective parameter*

$$w_r = \varphi_r(x), \qquad \varphi_r(x) = M_r\tanh(x), \qquad M_r > 0,$$

where the hyperbolic tangent is applied componentwise, and set

$$r(s,a;x) = \overline{r}(s,a;\varphi_r(x)).$$

Suppose that $\overline{r}(s,a;w_r) = f_r(\phi(s,a);w_r)$ is a finite-depth feedforward neural network with twice continuously differentiable activations, with all trainable weights and biases included in w_r. Here $\phi(s,a)$ is a fixed, finite-dimensional bounded feature map, continuous in s for each $a \in \mathcal{A}$.

Although the state space $\mathcal{S} = \mathbb{R}^{d_s}$ is not compact, the feature-range closure

$$\mathcal{K}_{\mathrm{feat}} := \overline{\{\phi(s,a) : s \in \mathcal{S},\ a \in \mathcal{A}\}}$$

is compact. Moreover,

$$\overline{\varphi_r(\mathbb{R}^{d_x})} \subseteq [-M_r, M_r]^{d_x}.$$

Joint continuity of f_r, $\nabla_{w_r} f_r$, and $\nabla^2_{w_r w_r} f_r$ on the compact feature–parameter domain $\mathcal{K}_{\mathrm{feat}} \times [-M_r, M_r]^{d_x}$ therefore yields uniform bounds, valid for all states, actions, and effective parameters in this cube:

$$|\overline{r}| \le \overline{C}_{r,0}, \qquad \|\nabla_{w_r}\overline{r}\| \le \overline{C}_{r,1}, \qquad \|\nabla^2_{w_r w_r}\overline{r}\| \le \overline{C}_{r,2},$$

for finite positive constants $\overline{C}_{r,0}, \overline{C}_{r,1}, \overline{C}_{r,2}$. In addition,

$$\|D\varphi_r(x)\| \le M_r, \qquad \|D^2\varphi_r(x)\|_{\mathrm{op}} \le 2M_r,$$

where the second norm is the bilinear operator norm induced by Euclidean norms. The chain rule consequently gives, uniformly in (s,a,x),

$$|r(s,a;x)| \le \overline{C}_{r,0}, \qquad \|\nabla_x r(s,a;x)\| \le M_r\overline{C}_{r,1},$$

and

$$\|\nabla^2_{xx} r(s,a;x)\| \le M_r^2\overline{C}_{r,2} + 2M_r\overline{C}_{r,1}.$$

An identical construction $w_c = \varphi_c(y) = M_c\tanh(y)$ with $M_c > 0$ and $c(s,a;y) = \overline{c}(s,a;\varphi_c(y))$ yields the analogous constants $\overline{C}_{c,0}, \overline{C}_{c,1}, \overline{C}_{c,2}$. Thus valid choices in Assumption 1 are

$$R = \overline{C}_{r,0}, \quad R_1 = M_r\overline{C}_{r,1}, \quad R_2 = M_r^2\overline{C}_{r,2} + 2M_r\overline{C}_{r,1},$$
$$C = \overline{C}_{c,0}, \quad C_1 = M_c\overline{C}_{c,1}, \quad C_2 = M_c^2\overline{C}_{c,2} + 2M_c\overline{C}_{c,1}.$$

The assumption does not restrict the analysis to neural networks, while smooth networks with bounded effective parameters and bounded features provide a non-vacuous class satisfying its signal bounds without projecting the unconstrained parameters x and y. The continuous feature map also ensures state continuity of these signals. This construction justifies signal regularity only; the uniform Slater condition remains a separate requirement on the constraint family.

E.1.2 Discussion on Assumption 3

The boundedness part of Assumption 3 follows from Assumption 1 when G averages Bradley–Terry negative log-likelihoods for uniformly bounded-length reward and constraint sums, as shown next.

Lemma 11 (Bounded Bradley–Terry objective for finite reward and constraint sums)**.** *Suppose $|r(s,a;x)| \le R$ and $|c(s,a;y)| \le C$ uniformly, and every compared trajectory segment $\tau = ((s_t,a_t))_{t=0}^{h(\tau)-1}$ has length $h(\tau) \le H$, where H is a fixed finite positive integer independent of (v,θ). Define its reward and constraint preference scores by*

$$u_x^r(\tau) := \sum_{t=0}^{h(\tau)-1} r(s_t,a_t;x), \qquad u_y^c(\tau) := \sum_{t=0}^{h(\tau)-1} c(s_t,a_t;y).$$

For a compared pair (τ_1,τ_2) and labels $Y_r, Y_c \in \{0,1\}$, where a value of one means that the first trajectory is preferred under the corresponding criterion, let

$$\Delta_r := u_x^r(\tau_1) - u_x^r(\tau_2), \qquad \Delta_c := u_y^c(\tau_1) - u_y^c(\tau_2),$$

and define

$$\ell_r := \log(1+e^{\Delta_r}) - Y_r\Delta_r, \qquad \ell_c := \log(1+e^{\Delta_c}) - Y_c\Delta_c, \qquad \ell_{\mathrm{BT}} := \frac{1}{2}(\ell_r + \ell_c).$$

If $G(v,\theta)$ is an empirical average or an expectation of ℓ_{BT} over such jointly labeled pairs, then, uniformly in (v,θ),

$$\frac{1}{2}\left[\log\bigl(1+e^{-2HR}\bigr)+\log\bigl(1+e^{-2HC}\bigr)\right] \le G(v,\theta) \le \frac{1}{2}\left[\log\bigl(1+e^{2HR}\bigr)+\log\bigl(1+e^{2HC}\bigr)\right]. \tag{61}$$

The distribution of the compared pairs may depend on (v,θ), provided the same length, reward, and constraint bounds hold throughout its support. In particular, valid constants for the boundedness part of Assumption 3 are

$$G_{\min}=0, \qquad G_{\max}=\frac{1}{2}\left[\log\bigl(1+e^{2HR}\bigr)+\log\bigl(1+e^{2HC}\bigr)\right], \qquad \Delta_G = G_{\max}. \tag{62}$$

Proof. The triangle inequality and the uniform signal bounds give

$$|u_x^r(\tau)| \le \sum_{t=0}^{h(\tau)-1} |r(s_t,a_t;x)| \le h(\tau)R \le HR, \qquad |u_y^c(\tau)| \le h(\tau)C \le HC.$$

Consequently, $|\Delta_r| \le 2HR$ and $|\Delta_c| \le 2HC$. For either value of its binary label, each logistic loss equals $\log(1+e^z)$ for some $z \in [-|\Delta|,|\Delta|]$. Monotonicity of this function gives the two componentwise bounds, whose average is Eq. 61. Empirical averaging or expectation preserves them. The lower bound is positive, so replacing it by zero yields Eq. 62. □

Lemma 11 alone justifies only the boundedness part of Assumption 3. The smoothness and gradient-oracle conditions are established next for a specified policy-generated sampling model. The common length bound H is essential to this argument: an undiscounted sum over an almost surely finite but unbounded random horizon, such as the geometric horizons used by the lower-level critic, need not have a uniformly bounded score. If G is an unnormalized sum of a fixed number of preference losses rather than their average, its bounds are multiplied by that number.

Policy-generated preference samples. Fix a positive integer H. At any queried (v,θ), suppose we can sample two independent trajectories τ_1,τ_2 under π_θ, starting from the fixed initial distribution ρ. Each length $h_i \in \{1,\dots,H\}$ is fixed in advance or sampled independently from a fixed law independent of (v,θ). The transition kernel is the parameter-independent kernel P of the problem formulation. For each pair, suppose we can sample labels $(Y_r,Y_c) \in \{0,1\}^2$ from a fixed Borel measurable conditional law $\mathcal{Q}(\cdot \mid \tau_1,\tau_2)$, where $Y_r=1$ means that the first trajectory is reward-preferred and $Y_c=1$ means that it is preferred under the constraint or safety criterion. This joint label law has no additional dependence on (v,θ) and need not itself follow a Bradley–Terry model. Each oracle call uses fresh trajectories and fresh labels, conditionally independently of all previous calls.

Using the reward- and constraint-sum scores from Lemma 11, define

$$\Delta_r = u_x^r(\tau_1)-u_x^r(\tau_2), \qquad p_r=\frac{1}{1+e^{-\Delta_r}}, \qquad \ell_r=\log(1+e^{\Delta_r})-Y_r\Delta_r,$$

and

$$\Delta_c = u_y^c(\tau_1)-u_y^c(\tau_2), \qquad p_c=\frac{1}{1+e^{-\Delta_c}}, \qquad \ell_c=\log(1+e^{\Delta_c})-Y_c\Delta_c.$$

Let $\ell_{\mathrm{BT}} := (\ell_r+\ell_c)/2$. Let

$$G(v,\theta) := \mathbb{E}_{\tau_1,\tau_2\sim\pi_\theta,\,(Y_r,Y_c)\sim\mathcal{Q}(\cdot|\tau_1,\tau_2)}[\ell_{\mathrm{BT}}]. \tag{63}$$

Thus, at fixed θ, the sampling law is independent of v. Define the pair policy score

$$S_\theta^{\mathrm{pair}} := \sum_{i=1}^{2}\sum_{t=0}^{h_i-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}).$$

This is the score of the upper trajectory pair, not the lower training score $S(Z;v)$. Assume the reward and constraint values and their respective x- and y-derivatives can be evaluated on the sampled trajectories. Define the single-pair estimators

$$\widehat{\nabla}_v G(v,\theta) = \frac{1}{2}\begin{bmatrix}(p_r-Y_r)\nabla_x\Delta_r\\(p_c-Y_c)\nabla_y\Delta_c\end{bmatrix}, \qquad \widehat{\nabla}_\theta G(v,\theta) = \ell_{\mathrm{BT}}\, S_\theta^{\mathrm{pair}} \tag{64}$$

Using the policy-score bound G_π from Assumption 5, let

$$L_{\mathrm{BT}} := \frac{1}{2}\left[\log(1+e^{2HR}) + \log(1+e^{2HC})\right].$$

Lemma 12 (Unbiasedness and bounded variance of Bradley–Terry gradients)**.** *Under Assumptions 5 and 1 and the sampling model above, the estimators in Eq. 64 satisfy, uniformly in* (v,θ)*,*

$$\begin{aligned}\mathbb{E}[\widehat{\nabla}_v G(v,\theta) \mid v,\theta] &= \nabla_v G(v,\theta),\\ \mathbb{E}[\widehat{\nabla}_\theta G(v,\theta) \mid v,\theta] &= \nabla_\theta G(v,\theta),\end{aligned} \tag{65}$$

and

$$\begin{aligned}\mathbb{E}\Big[\|\widehat{\nabla}_v G(v,\theta) - \nabla_v G(v,\theta)\|^2 \mid v,\theta\Big] &\le H^2(R_1^2 + C_1^2),\\ \mathbb{E}\Big[\|\widehat{\nabla}_\theta G(v,\theta) - \nabla_\theta G(v,\theta)\|^2 \mid v,\theta\Big] &\le (2HG_\pi L_{\mathrm{BT}})^2.\end{aligned} \tag{66}$$

Proof. Condition on the queried (v,θ) and the history before the oracle call. The new data have the sampling law specified above. For a fixed labeled pair, direct differentiation gives

$$\nabla_x \ell_{\mathrm{BT}} = \frac{1}{2}(p_r - Y_r)\nabla_x \Delta_r, \qquad \nabla_y \ell_{\mathrm{BT}} = \frac{1}{2}(p_c - Y_c)\nabla_y \Delta_c.$$

Since the pair and label law is independent of v at fixed θ, differentiating its expectation gives the first estimator in Eq. 64. This interchange is justified by $|p_r - Y_r|, |p_c - Y_c| \le 1$, $\|\nabla_x \Delta_r\| \le 2HR_1$, and $\|\nabla_y \Delta_c\| \le 2HC_1$.

Let $f_\theta(\tau_1, \tau_2, Y_r, Y_c)$ denote the joint density/mass function of the sampled trajectories, lengths, and labels. On its support, all factors except the action probabilities are independent of θ, so

$$\nabla_\theta \log f_\theta(\tau_1, \tau_2, Y_r, Y_c) = S_\theta^{\mathrm{pair}}.$$

The loss has no explicit θ-dependence at a fixed labeled pair. Consequently,

$$\nabla_\theta G(v,\theta) = \mathbb{E}[\ell_{\mathrm{BT}} S_\theta^{\mathrm{pair}}],$$

which proves unbiasedness of the second estimator. For completeness, full action support gives a common support, and the bounded policy score implies the local likelihood-ratio bound

$$\frac{f_{\theta'}(\tau_1, \tau_2, Y_r, Y_c)}{f_\theta(\tau_1, \tau_2, Y_r, Y_c)} \le \exp(2HG_\pi \|\theta' - \theta\|)$$

on that support. Together with $\ell_{\mathrm{BT}} \le L_{\mathrm{BT}}$ and $\|S_\theta^{\mathrm{pair}}\| \le 2HG_\pi$, this supplies local domination for the differentiation step. The conditional label probabilities need not be evaluated or differentiated to implement either estimator.

The preceding componentwise gradient bounds, $\ell_{\mathrm{BT}} \le L_{\mathrm{BT}}$, and $\|S_\theta^{\mathrm{pair}}\| \le 2HG_\pi$ give

$$\|\widehat{\nabla}_v G(v,\theta)\| \le H\sqrt{R_1^2 + C_1^2}, \qquad \|\widehat{\nabla}_\theta G(v,\theta)\| \le 2HG_\pi L_{\mathrm{BT}}. \tag{67}$$

Taking expectations bounds the true gradient norms. For either unbiased vector estimator X,

$$\mathbb{E}[\|X - \mathbb{E}[X \mid v,\theta]\|^2 \mid v,\theta] \le \mathbb{E}[\|X\|^2 \mid v,\theta],$$

which proves Eq. 66.

Finally, at a fixed labeled pair,

$$\nabla_{xx}^2 \ell_{\mathrm{BT}} = \frac{1}{2}\left[p_r(1-p_r)\nabla_x \Delta_r \nabla_x \Delta_r^\top + (p_r - Y_r)\nabla_{xx}^2 \Delta_r\right].$$

Using $p_r(1-p_r) \le 1/4$ and $\|\nabla_{xx}^2 \Delta_r\| \le 2HR_2$ gives

$$\|\nabla_{xx}^2 \ell_{\mathrm{BT}}\| \le \frac{1}{2}\left(H^2 R_1^2 + 2HR_2\right).$$

Analogously,

$$\|\nabla_{yy}^2 \ell_{\mathrm{BT}}\| \le \frac{1}{2}\left(H^2 C_1^2 + 2HC_2\right), \qquad \nabla_{xy}^2 \ell_{\mathrm{BT}} = 0.$$

The sampling law is independent of v at fixed θ. Averaging these uniform block-Hessian bounds establishes the claimed Lipschitz constant for $\nabla_v G$. □

The proof also shows that the following are valid constants in Assumption 3:

$$\begin{aligned} L_{G,0} &= H\sqrt{R_1^2 + C_1^2}, & M_G &= 2HG_\pi L_{\mathrm{BT}}, \\ \sigma_v^2 &= H^2(R_1^2 + C_1^2), & \sigma_\theta^2 &= (2HG_\pi L_{\mathrm{BT}})^2, \\ L_G &= \frac{1}{2}\max\left\{H^2R_1^2 + 2HR_2, H^2C_1^2 + 2HC_2\right\}. \end{aligned} \tag{68}$$

Together with the value bounds in Lemma 11, this verifies Assumption 3 for Eq. 63.

Lemma 12 applies at every queried policy parameter, including the interpolated parameters in Eq. 13. Each oracle call in this implementation uses two fresh rollouts of length at most H and obtains reward-preference and constraint-preference labels; separate calls use independent data as required by the algorithm. The lemma concerns the policy-generated expectation Eq. 63. For a fixed, parameter-independent dataset with reward and constraint scores depending on (x, y), minibatch v-gradients instead estimate the empirical objective, whose θ-partial gradient is zero. If pair selection or the conditional label law has additional parameter dependence, its derivative contributions must also be included; the estimators above do not automatically cover that case.

Define

$$V_\theta := M_G^2 + \sigma_\theta^2. \tag{69}$$

E.2 Supporting Lemmas for Theorem 3

Score Bounds for the Random-Length Lower Trace. For a reward encoding, define the score already embedded in the x-coordinate block of $v = (x, y)$ by

$$s_r = \begin{bmatrix} \dfrac{B^r \nabla_x r(s, a; x)}{2R + B^r r(s, a; x)} \\ 0_{d_y} \end{bmatrix} \in \mathbb{R}^{d_v},$$

and for a constraint encoding define the score already embedded in the y-coordinate block by

$$s_c = \begin{bmatrix} 0_{d_x} \\ \dfrac{B^c \nabla_y c(s, a; y)}{2C + B^c c(s, a; y)} \end{bmatrix} \in \mathbb{R}^{d_v}.$$

The corresponding log-probability Hessians, embedded in the same coordinate blocks, are denoted by $H_r, H_c \in \mathbb{R}^{d_v \times d_v}$.

Let

$$N_r(Z), \qquad N_c(Z)$$

be the total numbers of reward and constraint encoding events in one complete lower-level training run. Define

$$\bar{n}_r := \sup_v \mathbb{E}_v[N_r(Z)], \qquad \bar{n}_c := \sup_v \mathbb{E}_v[N_c(Z)]. \tag{70}$$

Define

$$\bar{I} := \frac{\bar{n}_r}{3}\left(\frac{R_1}{R}\right)^2 + \frac{\bar{n}_c}{3}\left(\frac{C_1}{C}\right)^2, \tag{71}$$

and

$$\bar{J} := \bar{n}_r\left[\frac{R_2}{R} + \left(\frac{R_1}{R}\right)^2\right] + \bar{n}_c\left[\frac{C_2}{C} + \left(\frac{C_1}{C}\right)^2\right]. \tag{72}$$

Lemma 13 (Moments of the complete training score)**.** *Under Assumption 1, define*

$$S(Z; v) := \sum_{j=1}^{N_r(Z)} s_{r,j} + \sum_{j=1}^{N_c(Z)} s_{c,j}, \qquad H_{\log}(Z; v) := \sum_{j=1}^{N_r(Z)} H_{r,j} + \sum_{j=1}^{N_c(Z)} H_{c,j}.$$

Then

$$\mathbb{E}_v[S(Z; v)] = 0, \tag{73}$$

and

$$\mathbb{E}_v\left[\|S(Z;v)\|^2\right] \le \bar{I}. \tag{74}$$

Moreover,

$$\mathbb{E}_v\left[\|H_{\log}(Z;v)\|\right] \le \bar{J}. \tag{75}$$

Proof. Fix v. Enumerate all reward and constraint encoding events in one lower training run in chronological order, with reward encodings preceding constraint encodings when both are queried at the same step. Let $N_{\mathrm{enc}} := N_r(Z) + N_c(Z)$ be the total number of such events. For $j \ge 1$, define the vector $D_j \in \mathbb{R}^{d_v}$ by

$$D_j := \begin{cases} s_r^{(j)}, & \text{if event } j \text{ is a reward encoding,} \\ s_c^{(j)}, & \text{if event } j \text{ is a constraint encoding,} \\ 0_{d_v}, & \text{if } j > N_{\mathrm{enc}}. \end{cases}$$

Here $s_r^{(j)}$ and $s_c^{(j)}$ are the already embedded scores in Eq. 27 and Eq. 28, evaluated at the state–action pair and encoded sign of event j. Thus j indexes the combined sequence of encoding events, rather than policy or critic iterations. With the coordinate embedding above and zero padding after termination, the complete score is

$$S(Z;v) = \sum_{j=1}^{N_{\mathrm{enc}}} D_j = \sum_{j\ge 1} D_j.$$

Let $\mathcal{F}_{j-1}$ contain the training history immediately before encoding event j, including whether the run has already terminated and, if it has not, the event type and queried state–action pair. After termination, retain the terminal history. These sigma-algebras form an increasing filtration, and D_i is $\mathcal{F}_{j-1}$-measurable whenever $i < j$. The occurrence and type of event j are known before drawing its encoded sign; no conditioning on future encoding outcomes is involved.

Conditional mean of each increment. For an active event j, write $a(v)$ for its scalar signal and $A > 0$ for its fixed encoding bound, with $|a(v)| \le A$. Conditional on $\mathcal{F}_{j-1}$, the queried pair is fixed and

$$\mathbb{P}_v(B = b \mid \mathcal{F}_{j-1}) = \frac{2A + b\,a(v)}{4A}, \qquad s_a(b;v) = \frac{b\nabla_v a(v)}{2A + b\,a(v)}, \qquad b \in \{-1,+1\}.$$

The derivative holds the queried pair and preceding trace fixed, and the bound A is independent of v. Since $2A + b\,a(v) \ge A > 0$, direct summation over the two possible signs gives

$$\begin{aligned} \mathbb{E}_v[s_a(B;v) \mid \mathcal{F}_{j-1}] &= \sum_{b\in\{-1,+1\}} \frac{2A + b\,a(v)}{4A} \frac{b\nabla_v a(v)}{2A + b\,a(v)} \\ &= \frac{\nabla_v a(v)}{4A} \sum_{b\in\{-1,+1\}} b = 0. \end{aligned}$$

The coordinate-embedded reward and constraint score definitions above preserve this identity. On terminated runs $D_j = 0$, so in all cases

$$\mathbb{E}_v[D_j \mid \mathcal{F}_{j-1}] = 0, \qquad \mathbb{E}_v[D_j] = \mathbb{E}_v[\mathbb{E}_v[D_j \mid \mathcal{F}_{j-1}]] = 0. \tag{76}$$

Mean of the complete random-length score. There are $K(N+1)$ oracle calls, each using finitely many geometric-horizon trajectories with finite expected lengths. Thus $\mathbb{E}_v[N_r] + \mathbb{E}_v[N_c] < \infty$; the precise counts are given in Lemma 14. Assumption 1 and the positive denominator bounds imply

$$\|s_r^{(j)}\| \le \frac{R_1}{R}, \qquad \|s_c^{(j)}\| \le \frac{C_1}{C}.$$

Consequently, Tonelli's theorem gives

$$\sum_{j\ge 1} \mathbb{E}_v\|D_j\| = \mathbb{E}_v\left[\sum_{j\ge 1} \|D_j\|\right] \le \frac{R_1}{R}\mathbb{E}_v[N_r] + \frac{C_1}{C}\mathbb{E}_v[N_c] < \infty.$$

This absolute integrability justifies interchanging expectation and the infinite, zero-padded sum. Using Eq. 76, we obtain

$$\mathbb{E}_v[S(Z;v)] = \sum_{j\geq 1} \mathbb{E}_v[D_j] = \sum_{j\geq 1} \mathbb{E}_v[\mathbb{E}_v[D_j \mid \mathcal{F}_{j-1}]] = 0,$$

which proves Eq. 73.

Second moment of the complete score. For an active scalar encoding with $\|\nabla_v a(v)\| \leq A_1$, the conditional second moment is

$$\begin{aligned}\mathbb{E}_v[\|s_a(B;v)\|^2 \mid \mathcal{F}_{j-1}] &= \sum_{b\in\{-1,+1\}} \frac{2A + b\,a(v)}{4A} \frac{\|\nabla_v a(v)\|^2}{(2A + b\,a(v))^2} \\ &= \frac{\|\nabla_v a(v)\|^2}{4A}\left(\frac{1}{2A + a(v)} + \frac{1}{2A - a(v)}\right) \\ &= \frac{\|\nabla_v a(v)\|^2}{4A^2 - a(v)^2} \leq \frac{A_1^2}{3A^2}.\end{aligned}$$

Let I_j^r and I_j^c be the indicators that event j is a reward or constraint encoding, respectively, with both zero after termination. They are $\mathcal{F}_{j-1}$-measurable. Applying the preceding bound to each event type gives

$$\mathbb{E}_v[\|D_j\|^2 \mid \mathcal{F}_{j-1}] \leq \frac{I_j^r}{3}\left(\frac{R_1}{R}\right)^2 + \frac{I_j^c}{3}\left(\frac{C_1}{C}\right)^2.$$

Since $\sum_j I_j^r = N_r$ and $\sum_j I_j^c = N_c$, taking expectations and summing yields

$$\sum_{j\geq 1} \mathbb{E}_v\|D_j\|^2 \leq \frac{\mathbb{E}_v[N_r]}{3}\left(\frac{R_1}{R}\right)^2 + \frac{\mathbb{E}_v[N_c]}{3}\left(\frac{C_1}{C}\right)^2 \leq \bar{I} < \infty. \tag{77}$$

For $i < j$, the earlier increment D_i is known given $\mathcal{F}_{j-1}$. The tower property and Eq. 76 therefore imply

$$\mathbb{E}_v[D_i^\top D_j] = \mathbb{E}_v\big[D_i^\top \mathbb{E}_v[D_j \mid \mathcal{F}_{j-1}]\big] = 0.$$

Thus the increments are orthogonal in expectation; independence is not required. For the finite partial sum $S_m := \sum_{j=1}^m D_j$, expansion of the squared norm gives

$$\mathbb{E}_v\|S_m\|^2 = \sum_{j=1}^m \mathbb{E}_v\|D_j\|^2 + 2\sum_{1\leq i<j\leq m} \mathbb{E}_v[D_i^\top D_j] = \sum_{j=1}^m \mathbb{E}_v\|D_j\|^2.$$

To pass to the full random-length score, note that for $n > m$,

$$\mathbb{E}_v\|S_n - S_m\|^2 = \sum_{j=m+1}^n \mathbb{E}_v\|D_j\|^2 \leq \sum_{j>m} \mathbb{E}_v\|D_j\|^2 \longrightarrow 0 \quad \text{as } m \to \infty,$$

by Eq. 77. Hence S_m converges in L^2. Since $N_{\text{enc}} < \infty$ almost surely, these partial sums also eventually equal $S(Z;v)$ on each realized finite run, identifying the L^2 limit with $S(Z;v)$. Therefore,

$$\mathbb{E}_v\|S(Z;v)\|^2 = \lim_{m\to\infty} \mathbb{E}_v\|S_m\|^2 = \sum_{j\geq 1} \mathbb{E}_v\|D_j\|^2 \leq \bar{I},$$

which proves Eq. 74.

Log-Hessian bound. Consider one active encoding event with scalar signal $a(v)$, encoding bound A, and realized sign $B \in \{-1, +1\}$. Its conditional probability given the preceding history is

$$p_v(B) = \frac{2A + Ba(v)}{4A}.$$

Holding the preceding history, event type, and queried state–action pair fixed, its log-probability Hessian is

$$H_a(B;v) = \nabla_v^2 \log p_v(B) = \frac{B\nabla^2 a(v)}{2A + Ba(v)} - \frac{\nabla a(v)\nabla a(v)^\top}{(2A + Ba(v))^2}.$$

The second term follows because differentiating the denominator contributes $B\nabla a(v)$ and $B^2 = 1$. Under $|a(v)| \le A$, $\|\nabla_v a(v)\| \le A_1$, and $\|\nabla_v^2 a(v)\| \le A_2$, we have

$$2A + Ba(v) \ge A,$$

and

$$\|\nabla a(v)\nabla a(v)^\top\| = \|\nabla a(v)\|^2.$$

The triangle inequality therefore gives the deterministic eventwise bound

$$\|H_a(B; v)\| \le \frac{A_2}{A} + \left(\frac{A_1}{A}\right)^2.$$

For a reward encoding, substitute $(A, A_1, A_2) = (R, R_1, R_2)$; for a constraint encoding, substitute $(A, A_1, A_2) = (C, C_1, C_2)$. Embedding these Hessians in the corresponding x- or y-coordinate block does not change their operator norms. Hence, on every realized lower trace,

$$\begin{aligned}\|H_{\log}(Z; v)\| &\le \sum_{j=1}^{N_r(Z)} \|H_{r,j}\| + \sum_{j=1}^{N_c(Z)} \|H_{c,j}\| \\ &\le N_r(Z)\left[\frac{R_2}{R} + \left(\frac{R_1}{R}\right)^2\right] + N_c(Z)\left[\frac{C_2}{C} + \left(\frac{C_1}{C}\right)^2\right].\end{aligned}$$

Taking expectations, using $\mathbb{E}_v[N_r(Z)] \le \bar{n}_r$ and $\mathbb{E}_v[N_c(Z)] \le \bar{n}_c$, and recalling the definition of $\bar{J}$ gives

$$\begin{aligned}\mathbb{E}_v\|H_{\log}(Z; v)\| \le \bar{n}_r &\left[\frac{R_2}{R} + \left(\frac{R_1}{R}\right)^2\right] \\ &+ \bar{n}_c\left[\frac{C_2}{C} + \left(\frac{C_1}{C}\right)^2\right] = \bar{J},\end{aligned}$$

which proves Eq. 75. □

Lemma 14 (Encoding counts for the geometric-horizon lower-level algorithm)**.** *Suppose the geometric random variables in Algorithm 3 have distribution*

$$\mathbb{P}(T = t) = (1-\gamma)\gamma^t, \qquad t = 0, 1, \ldots.$$

One call to Algorithm 3 uses, in expectation,

$$\frac{2}{1-\gamma}$$

reward encodings and

$$\frac{3}{1-\gamma}$$

constraint encodings.

Since Algorithm 2 makes $N+1$ such calls at each of its K primal-dual iterations,

$$\bar{n}_r = \frac{2K(N+1)}{1-\gamma}, \tag{78}$$

and

$$\bar{n}_c = \frac{3K(N+1)}{1-\gamma}. \tag{79}$$

In particular,

$$\bar{I} = O\left(\frac{K(N+1)}{1-\gamma}\right), \qquad \bar{J} = O\left(\frac{K(N+1)}{1-\gamma}\right), \tag{80}$$

where the hidden constants depend only on the signal derivative bounds.

Proof. For $T \sim \mathrm{Geo}(1-\gamma)$,

$$\mathbb{E}[T+1] = \frac{1}{1-\gamma}.$$

The constraint-return trajectory generates one constraint encoding at every sampled state-action pair, hence contributes $(1-\gamma)^{-1}$ constraint encodings in expectation.

The Q-trajectory generates both one reward and one constraint encoding at each sampled state-action pair, and the V-trajectory does the same. Therefore a complete oracle call produces

$$\frac{2}{1-\gamma}$$

reward encodings and

$$\frac{3}{1-\gamma}$$

constraint encodings in expectation.

There are N critic calls and one independent constraint/critic call per primal-dual iteration. Multiplying by $K(N+1)$ proves Eq. 78–Eq. 79. Substitution into Eq. 71–Eq. 72 gives Eq. 80. □

E.3 Upper-Gradient Estimation Using Only Gradient Oracles

We next construct an unbiased estimator of Eq. 53 without assuming access to exact values of G.

We choose the reference policy parameter to be the lower initialization θ_0. Let

$$d(Z) = \theta(Z) - \theta_0.$$

Lemma 15 (Mean-square displacement of the approximate lower-level response)**.** *Suppose Assumptions 5 and 6 hold and let β be the inner vanilla-SGD critic stepsize used in Algorithm 2, chosen as required by Lemma 5. Let $Z \sim P_v$ be the complete lower trace and let $\theta_R = \theta(Z)$ be the lower-level response returned by the same algorithm. Then*

$$\mathbb{E}\left[\|\theta_R - \theta_0\|^2\right] \le \bar{D}^2, \tag{81}$$

where

$$\bar{D}^2 := \eta^2 K^2 W_\omega^2. \tag{82}$$

Here W_ω^2 is the constant defined in Lemma 5.

In particular, if

$$\eta = \frac{\eta_0}{\sqrt{K}}, \tag{83}$$

then

$$\bar{D}^2 \le \eta_0^2 K W_\omega^2. \tag{84}$$

Proof. By Lemma 5,

$$\mathbb{E}_k\|\widehat{\omega}_k\|^2 \le W_\omega^2.$$

Taking expectation over the lower-level history gives

$$\mathbb{E}\|\widehat{\omega}_k\|^2 \le W_\omega^2 \tag{85}$$

uniformly in k.

The policy update in Algorithm 2 gives, for any $r \in \{0, \dots, K\}$,

$$\theta_r - \theta_0 = \eta \sum_{k=0}^{r-1} \widehat{\omega}_k. \tag{86}$$

Therefore,

$$\|\theta_r - \theta_0\|^2 = \eta^2 \left\| \sum_{k=0}^{r-1} \widehat{\omega}_k \right\|^2. \tag{87}$$

By the Cauchy–Schwarz inequality,

$$\left\|\sum_{k=0}^{r-1}\widehat{\omega}_k\right\|^2 \le r\sum_{k=0}^{r-1}\|\widehat{\omega}_k\|^2. \tag{88}$$

Substituting Eq. 88 into Eq. 87 gives

$$\|\theta_r-\theta_0\|^2 \le \eta^2 r\sum_{k=0}^{r-1}\|\widehat{\omega}_k\|^2. \tag{89}$$

Taking expectation of Eq. 89,

$$\mathbb{E}\left[\|\theta_r-\theta_0\|^2\right] \le \eta^2 r\sum_{k=0}^{r-1}\mathbb{E}\left[\|\widehat{\omega}_k\|^2\right]. \tag{90}$$

Applying Eq. 85,

$$\begin{aligned}\mathbb{E}\left[\|\theta_r-\theta_0\|^2\right] &\le \eta^2 r\sum_{k=0}^{r-1}W_\omega^2\\ &= \eta^2 r^2 W_\omega^2.\end{aligned} \tag{91}$$

Since $r \le K$,

$$\mathbb{E}\left[\|\theta_r-\theta_0\|^2\right] \le \eta^2 K^2 W_\omega^2. \tag{92}$$

We finally average over the sampled output index

$$R \sim \mathrm{Unif}\{0,\dots,K-1\}.$$

Because R is independent of the lower-level optimization randomness,

$$\begin{aligned}\mathbb{E}\left[\|\theta_R-\theta_0\|^2\right] &= \frac{1}{K}\sum_{r=0}^{K-1}\mathbb{E}\left[\|\theta_r-\theta_0\|^2\right]\\ &\le \frac{1}{K}\sum_{r=0}^{K-1}\eta^2K^2W_\omega^2\\ &= \eta^2K^2W_\omega^2\\ &= \bar{D}^2.\end{aligned} \tag{93}$$

This proves Eq. 81.

Finally, if

$$\eta = \frac{\eta_0}{\sqrt{K}},$$

then

$$\begin{aligned}\bar{D}^2 &= \eta^2K^2W_\omega^2\\ &= \frac{\eta_0^2}{K}K^2W_\omega^2\\ &= \eta_0^2KW_\omega^2,\end{aligned} \tag{94}$$

which proves Eq. 84. □

Lemma 16 (Unbiased value-difference estimator)**.** *Under Assumption 3, for a complete lower trace Z with output $\theta(Z)$, define*

$$d := \theta(Z)-\theta_0, \qquad V_\theta := M_G^2+\sigma_\theta^2,$$

and, conditionally on Z, set

$$B_G(Z) := \max\{1,\lceil\|d\|^2\rceil\}. \tag{95}$$

Draw conditionally independent $U_m \sim \mathrm{Unif}[0,1]$ *and fresh upper-gradient oracle samples, and define*

$$Y_m := \left\langle \widehat{\nabla}_\theta G(v, \theta_0 + U_m d), d \right\rangle, \qquad m = 1, \dots, B_G(Z), \tag{96}$$

$$\widehat{\delta}_G(Z; v) := \frac{1}{B_G(Z)} \sum_{m=1}^{B_G(Z)} Y_m. \tag{97}$$

Then, conditionally on Z*,*

$$\mathbb{E}[\widehat{\delta}_G(Z; v) \mid Z] = G(v, \theta(Z)) - G(v, \theta_0), \tag{98}$$

$$\mathrm{Var}(\widehat{\delta}_G(Z; v) \mid Z) \le V_\theta. \tag{99}$$

Proof. Fix v and let $Z \sim P_v$ be a complete lower training trace. Conditionally on Z, both d and $B_G(Z)$ are fixed. Conditional unbiasedness of each fresh oracle sample and averaging over U_m give

$$\begin{aligned}\mathbb{E}[Y_m \mid Z] &= \int_0^1 \langle \nabla_\theta G(v, \theta_0 + ud), d \rangle \, du \\ &= G(v, \theta_0 + d) - G(v, \theta_0) = G(v, \theta(Z)) - G(v, \theta_0),\end{aligned}$$

where the second equality is the fundamental theorem of calculus along the segment. Averaging the $B_G(Z)$ terms proves Eq. 98.

Conditioning additionally on U_m fixes the oracle's evaluation point. The unbiasedness and variance bound in Assumption 3 imply

$$\begin{aligned}&\mathbb{E}\left[\|\widehat{\nabla}_\theta G(v, \theta_0 + U_m d)\|^2 \mid Z, U_m \right] \\ &\quad = \|\nabla_\theta G(v, \theta_0 + U_m d)\|^2 + \mathbb{E}\left[\|\widehat{\nabla}_\theta G(v, \theta_0 + U_m d) - \nabla_\theta G(v, \theta_0 + U_m d)\|^2 \mid Z, U_m \right] \\ &\quad \le M_G^2 + \sigma_\theta^2 = V_\theta.\end{aligned}$$

Cauchy–Schwarz and the tower property therefore yield

$$\begin{aligned}\mathbb{E}[Y_m^2 \mid Z] &\le \|d\|^2 \mathbb{E}\left[\mathbb{E}\left[\|\widehat{\nabla}_\theta G(v, \theta_0 + U_m d)\|^2 \mid Z, U_m \right] \mid Z \right] \\ &\le V_\theta \|d\|^2.\end{aligned}$$

Since the Y_m are independent conditionally on Z,

$$\mathrm{Var}(\widehat{\delta}_G \mid Z) = \frac{1}{B_G(Z)^2} \sum_{m=1}^{B_G(Z)} \mathrm{Var}(Y_m \mid Z) \le \frac{V_\theta \|d\|^2}{B_G(Z)} \le V_\theta,$$

proving Eq. 99. The bound $|G(v, \theta(Z)) - G(v, \theta_0)| \le \Delta_G$ then gives

$$\mathbb{E}[\widehat{\delta}_G^2 \mid Z] = \mathrm{Var}(\widehat{\delta}_G \mid Z) + \left(\mathbb{E}[\widehat{\delta}_G \mid Z] \right)^2 \le V_\theta + \Delta_G^2. \tag{100}$$

□

Lemma 17 (Smoothness of the approximate-response objective)**.** *Suppose Assumptions 1 and 3 hold. Let*

$$\Phi_{\epsilon_\ell}(v) = \mathbb{E}_{Z \sim P_v}\left[G(v, \theta(Z)) \right]$$

be the approximate-response objective defined in Eq. 60.

Let

$$S(Z; v) = \sum_{j=1}^{N_r(Z)} s_{r,j}(Z; v) + \sum_{j=1}^{N_c(Z)} s_{c,j}(Z; v)$$

denote the complete training score, and define

$$H_{\log}(Z; v) := \nabla_v^2 \log P_v(Z \mid \mathcal{T}),$$

where $\mathcal{T}$ denotes the collection of geometric horizons generated during the complete lower-level training run.

By Lemma 13, the score moments satisfy

$$\mathbb{E}_v\left[\|S(Z;v)\|^2\right] \le \bar{I}, \qquad \mathbb{E}_v\left[\|H_{\log}(Z;v)\|\right] \le \bar{J},$$

where $\bar{I}$ and $\bar{J}$ are defined in Eq. 71 and Eq. 72, respectively.

Then Φ_{ϵ_ℓ} has globally Lipschitz gradient. In particular, for every $v, v' \in \mathbb{R}^{d_v}$,

$$\|\nabla\Phi_{\epsilon_\ell}(v) - \nabla\Phi_{\epsilon_\ell}(v')\| \le L_\Phi\|v - v'\|, \tag{101}$$

where a valid Lipschitz constant is

$$L_\Phi = L_G + 2L_{G,0}\sqrt{\bar{I}} + \frac{\Delta_G}{2}\left(\bar{I} + \bar{J}\right). \tag{102}$$

Moreover, for Algorithm 2,

$$L_\Phi = O\left(1 + \frac{K(N+1)}{1-\gamma}\right). \tag{103}$$

Proof. We first prove the result under the additional assumption that $G(\cdot,\theta)$ is twice continuously differentiable for every θ. We subsequently explain how to extend the result to the $C^{1,1}$ regularity in Assumption 3.

Recall that the complete training score is

$$S(Z;v) = \sum_{j=1}^{N_r(Z)} s_{r,j}(Z;v) + \sum_{j=1}^{N_c(Z)} s_{c,j}(Z;v), \tag{104}$$

where $s_{r,j}$ and $s_{c,j}$ are, respectively, the score increments corresponding to the j-th encoded reward and constraint signals.

We also define the complete log-likelihood Hessian by

$$H_{\log}(Z;v) := \nabla_v^2 \log P_v(Z \mid \mathcal{T}), \tag{105}$$

where

$$\mathcal{T} = (T_1, \dots, T_{3K(N+1)})$$

collects all geometric horizons appearing in one complete lower-level training run.

The geometric horizons are independent of the upper-level parameter v. Conditional on a realization $\mathcal{T} = t$, the lower training trace contains finitely many encoding events. Therefore, likelihood differentiation can first be performed conditional on $\mathcal{T} = t$.

Step 1: Second derivative of the conditional trace law. Fix an arbitrary realization $\mathcal{T} = t$. Let

$$p_v(z \mid t)$$

denote the conditional probability density or probability mass function of the complete lower-level trace $Z = z$ given $\mathcal{T} = t$, with respect to a common reference measure.

By construction of the encoded lower-level algorithm, once the complete trace z is fixed, the decoded reward and constraint signals are fixed. Consequently, all policy, critic, and dual iterates, as well as the final output $\theta(z)$, are fixed functions of z and have no additional explicit dependence on v.

The score of the conditional trace distribution is

$$S(z;v) = \nabla_v \log p_v(z \mid t). \tag{106}$$

Therefore,

$$\nabla_v p_v(z \mid t) = p_v(z \mid t) S(z;v). \tag{107}$$

Differentiating Eq. 107 once more gives

$$\begin{aligned}\nabla_v^2 p_v(z \mid t) &= \nabla_v \left(p_v(z \mid t) S(z; v)\right) \\ &= \nabla_v p_v(z \mid t) S(z; v)^\top + p_v(z \mid t) \nabla_v S(z; v) \\ &= p_v(z \mid t) S(z; v) S(z; v)^\top + p_v(z \mid t) H_{\log}(z; v), \end{aligned} \tag{108}$$

where we used

$$\nabla_v S(z; v) = \nabla_v^2 \log p_v(z \mid t) = H_{\log}(z; v).$$

Hence,

$$\nabla_v^2 p_v(z \mid t) = p_v(z \mid t) \left(S(z; v) S(z; v)^\top + H_{\log}(z; v)\right). \tag{109}$$

Step 2: Conditional Hessian of the approximate-response objective. For a fixed horizon realization t, define

$$\Phi_{\epsilon_\ell, t}(v) := \mathbb{E}_v \left[G(v, \theta(Z)) \mid \mathcal{T} = t\right]. \tag{110}$$

Equivalently,

$$\Phi_{\epsilon_\ell, t}(v) = \int G(v, \theta(z)) p_v(z \mid t)\, dz. \tag{111}$$

Differentiating Eq. 111 once yields

$$\begin{aligned}\nabla_v \Phi_{\epsilon_\ell, t}(v) = \int \Big[&\nabla_v G(v, \theta(z)) p_v(z \mid t) \\ &+ G(v, \theta(z)) \nabla_v p_v(z \mid t)\Big]\, dz. \end{aligned} \tag{112}$$

Using Eq. 107,

$$\nabla_v \Phi_{\epsilon_\ell, t}(v) = \mathbb{E}_v \left[\nabla_v G(v, \theta(Z)) + G(v, \theta(Z)) S(Z; v) \mid \mathcal{T} = t\right]. \tag{113}$$

We now differentiate once more. Since $\theta(z)$ is fixed when z is fixed, the derivative of $\nabla_v G(v, \theta(z))$ with respect to v is simply $\nabla_{vv}^2 G(v, \theta(z))$. Applying the product rule gives

$$\begin{aligned}\nabla_v^2 \Phi_{\epsilon_\ell, t}(v) = \mathbb{E}_v \Big[&\nabla_{vv}^2 G(v, \theta(Z)) \\ &+ \nabla_v G(v, \theta(Z)) S(Z; v)^\top \\ &+ S(Z; v) \nabla_v G(v, \theta(Z))^\top \\ &+ G(v, \theta(Z)) \big(S(Z; v) S(Z; v)^\top + H_{\log}(Z; v)\big) \mid \mathcal{T} = t\Big]. \end{aligned} \tag{114}$$

Taking expectation with respect to $\mathcal{T}$ gives

$$\begin{aligned}\nabla_v^2 \Phi_{\epsilon_\ell}(v) = \mathbb{E}_v \Big[&\nabla_{vv}^2 G(v, \theta(Z)) \\ &+ \nabla_v G(v, \theta(Z)) S(Z; v)^\top \\ &+ S(Z; v) \nabla_v G(v, \theta(Z))^\top \\ &+ G(v, \theta(Z)) \big(S(Z; v) S(Z; v)^\top + H_{\log}(Z; v)\big)\Big]. \end{aligned} \tag{115}$$

We now justify the interchange of differentiation and expectation over the geometric horizons. Each horizon $T_i \sim \mathrm{Geo}(1 - \gamma)$ has finite moments of every order. Since the complete lower-level run contains only the finite collection $\mathcal{T} = (T_1, \ldots, T_{3K(N+1)})$ and its total number of encoding events satisfies

$$M(\mathcal{T}) \le 2 \sum_{i=1}^{3K(N+1)} (T_i + 1),$$

we have

$$\mathbb{E}[M(\mathcal{T})^2] < \infty.$$

Let

$$\kappa := \max\left\{\frac{R_1}{R}, \frac{C_1}{C}\right\}.$$

The uniform bound $\|s_j\| \le \kappa$ on each encoding-score increment gives

$$\|S(Z;v)\| \le \kappa M(\mathcal{T}), \qquad \|S(Z;v)S(Z;v)^\top\| \le \kappa^2 M(\mathcal{T})^2.$$

Assumption 1 also gives the uniform eventwise log-likelihood Hessian bound, where H_j denotes the reward or constraint log-likelihood Hessian increment at event j,

$$\|H_j\| \le \kappa_H := \max\left\{\frac{R_2}{R} + \left(\frac{R_1}{R}\right)^2, \frac{C_2}{C} + \left(\frac{C_1}{C}\right)^2\right\},$$

and hence $\|H_{\log}(Z;v)\| \le \kappa_H M(\mathcal{T})$. Writing $C_G := \max\{|G_{\min}|, |G_{\max}|\}$, the norm of the integrand in Eq. 114 is therefore bounded by

$$L_G + 2L_{G,0}\kappa M(\mathcal{T}) + C_G\left(\kappa^2 M(\mathcal{T})^2 + \kappa_H M(\mathcal{T})\right).$$

This dominating function is integrable with respect to $\mathcal{T}$ because $\mathbb{E}[M(\mathcal{T})^2] < \infty$. Dominated convergence therefore justifies interchanging ∇_v^2 and the expectation over the geometric horizons in Eq. 115.

Step 3: A normalization identity for the score Hessian term. For every fixed realization $\mathcal{T} = t$,

$$\int p_v(z \mid t)\, dz = 1. \tag{116}$$

Differentiating Eq. 116 twice and using Eq. 109 yields

$$\mathbb{E}_v\left[S(Z;v)S(Z;v)^\top + H_{\log}(Z;v) \mid \mathcal{T} = t\right] = 0. \tag{117}$$

Taking expectation over $\mathcal{T}$ gives

$$\mathbb{E}_v\left[S(Z;v)S(Z;v)^\top + H_{\log}(Z;v)\right] = 0. \tag{118}$$

Define the midpoint of the range of the upper objective by

$$G_{\text{mid}} := \frac{G_{\min} + G_{\max}}{2}. \tag{119}$$

Since Eq. 118 is zero,

$$\begin{aligned}&\mathbb{E}_v\left[G(v,\theta(Z))\left(S(Z;v)S(Z;v)^\top + H_{\log}(Z;v)\right)\right]\\ &= \mathbb{E}_v\left[\left(G(v,\theta(Z)) - G_{\text{mid}}\right)\left(S(Z;v)S(Z;v)^\top + H_{\log}(Z;v)\right)\right].\end{aligned} \tag{120}$$

Furthermore,

$$|G(v,\theta(Z)) - G_{\text{mid}}| \le \frac{\Delta_G}{2}. \tag{121}$$

Substituting Eq. 120 into Eq. 115, we obtain

$$\begin{aligned}\nabla_v^2 \Phi_{\epsilon_\ell}(v) = \mathbb{E}_v\Big[&\nabla_{vv}^2 G(v,\theta(Z))\\ &+ \nabla_v G(v,\theta(Z))S(Z;v)^\top\\ &+ S(Z;v)\nabla_v G(v,\theta(Z))^\top\\ &+ \left(G(v,\theta(Z)) - G_{\text{mid}}\right)\left(S(Z;v)S(Z;v)^\top + H_{\log}(Z;v)\right)\Big].\end{aligned} \tag{122}$$

Step 4: Bound each term in the Hessian. Taking the operator norm on both sides of Eq. 122 and using the triangle inequality gives

$$\begin{aligned}\left\|\nabla_v^2 \Phi_{\epsilon_\ell}(v)\right\| \le\ & \mathbb{E}_v\left[\left\|\nabla_{vv}^2 G(v,\theta(Z))\right\|\right] \\ & + \mathbb{E}_v\left[\left\|\nabla_v G(v,\theta(Z)) S(Z;v)^\top\right\|\right] \\ & + \mathbb{E}_v\left[\left\|S(Z;v)\nabla_v G(v,\theta(Z))^\top\right\|\right] \\ & + \mathbb{E}_v\Big[\,|G(v,\theta(Z)) - G_{\mathrm{mid}}| \\ & \qquad \times \left\|S(Z;v)S(Z;v)^\top + H_{\log}(Z;v)\right\|\Big]. \end{aligned} \tag{123}$$

Because $\nabla_v G(\cdot,\theta)$ is L_G-Lipschitz in v, wherever $\nabla_{vv}^2 G$ exists,

$$\left\|\nabla_{vv}^2 G(v,\theta(Z))\right\| \le L_G. \tag{124}$$

Therefore,

$$\mathbb{E}_v\left[\left\|\nabla_{vv}^2 G(v,\theta(Z))\right\|\right] \le L_G. \tag{125}$$

For any two vectors a and b,

$$\|ab^\top\| = \|a\|\|b\|.$$

Using

$$\|\nabla_v G(v,\theta(Z))\| \le L_{G,0},$$

we obtain

$$\mathbb{E}_v\left[\left\|\nabla_v G(v,\theta(Z)) S(Z;v)^\top\right\|\right] \le L_{G,0}\mathbb{E}_v\left[\|S(Z;v)\|\right], \tag{126}$$

and similarly,

$$\mathbb{E}_v\left[\left\|S(Z;v)\nabla_v G(v,\theta(Z))^\top\right\|\right] \le L_{G,0}\mathbb{E}_v\left[\|S(Z;v)\|\right]. \tag{127}$$

For the fourth term, using Eq. 121 and the triangle inequality gives

$$\begin{aligned}&\mathbb{E}_v\Big[\,|G(v,\theta(Z)) - G_{\mathrm{mid}}|\left\|S(Z;v)S(Z;v)^\top + H_{\log}(Z;v)\right\|\Big] \\ &\le \frac{\Delta_G}{2}\mathbb{E}_v\left[\left\|S(Z;v)S(Z;v)^\top\right\| + \|H_{\log}(Z;v)\|\right]. \end{aligned} \tag{128}$$

Since

$$\left\|S(Z;v)S(Z;v)^\top\right\| = \|S(Z;v)\|^2, \tag{129}$$

we have

$$\begin{aligned}&\mathbb{E}_v\Big[\,|G(v,\theta(Z)) - G_{\mathrm{mid}}|\left\|S(Z;v)S(Z;v)^\top + H_{\log}(Z;v)\right\|\Big] \\ &\le \frac{\Delta_G}{2}\left(\mathbb{E}_v\left[\|S(Z;v)\|^2\right] + \mathbb{E}_v\left[\|H_{\log}(Z;v)\|\right]\right). \end{aligned} \tag{130}$$

Combining Eq. 125, Eq. 126, Eq. 127, and Eq. 130 in Eq. 123 yields

$$\begin{aligned}\left\|\nabla_v^2 \Phi_{\epsilon_\ell}(v)\right\| \le\ & L_G + 2L_{G,0}\mathbb{E}_v\left[\|S(Z;v)\|\right] \\ & + \frac{\Delta_G}{2}\left(\mathbb{E}_v\left[\|S(Z;v)\|^2\right] + \mathbb{E}_v\left[\|H_{\log}(Z;v)\|\right]\right). \end{aligned} \tag{131}$$

Step 5: Substitute the score-moment bounds. By Jensen's inequality,

$$\mathbb{E}_v\left[\|S(Z;v)\|\right] \le \sqrt{\mathbb{E}_v\left[\|S(Z;v)\|^2\right]}. \tag{132}$$

Lemma 13 gives

$$\mathbb{E}_v\left[\|S(Z;v)\|^2\right] \le \bar{I}, \tag{133}$$

and

$$\mathbb{E}_v\left[\|H_{\log}(Z;v)\|\right] \le \bar{J}. \tag{134}$$

Combining Eq. 132 and Eq. 133,

$$\mathbb{E}_v\left[\|S(Z;v)\|\right] \le \sqrt{\bar{I}}. \tag{135}$$

Substituting Eq. 133, Eq. 134, and Eq. 135 into Eq. 131 yields

$$\left\|\nabla_v^2 \Phi_{\epsilon_\ell}(v)\right\| \le L_G + 2L_{G,0}\sqrt{\bar{I}} + \frac{\Delta_G}{2}\left(\bar{I} + \bar{J}\right) = L_\Phi. \tag{136}$$

The bound in Eq. 136 holds uniformly in v. Hence, for any $v, v' \in \mathbb{R}^{d_v}$,

$$\nabla \Phi_{\epsilon_\ell}(v) - \nabla \Phi_{\epsilon_\ell}(v') = \int_0^1 \nabla_v^2 \Phi_{\epsilon_\ell}\left(v' + \tau(v - v')\right)(v - v')\, d\tau. \tag{137}$$

Taking norms and applying Eq. 136 gives

$$\|\nabla \Phi_{\epsilon_\ell}(v) - \nabla \Phi_{\epsilon_\ell}(v')\| \le L_\Phi \|v - v'\|. \tag{138}$$

Thus $\nabla \Phi_{\epsilon_\ell}$ is globally L_Φ-Lipschitz.

Step 6: Extension to $C^{1,1}$ regularity. Assumption 3 requires only that $\nabla_v G(\cdot, \theta)$ be globally L_G-Lipschitz. Such a function is twice differentiable almost everywhere, with

$$\left\|\nabla_{vv}^2 G(v, \theta)\right\| \le L_G$$

at every point at which the Hessian exists.

To obtain the global result without requiring pointwise twice differentiability, let G_δ be a standard smooth mollification of G in the upper-level variable v. The mollified function satisfies the same uniform gradient-Lipschitz bound,

$$\left\|\nabla_{vv}^2 G_\delta(v, \theta)\right\| \le L_G,$$

and the same uniform bounds on G and $\nabla_v G$, up to an arbitrarily small approximation error. Applying the preceding argument to G_δ gives

$$\|\nabla \Phi_{\epsilon_\ell, \delta}(v) - \nabla \Phi_{\epsilon_\ell, \delta}(v')\| \le L_\Phi \|v - v'\|$$

with a constant L_Φ independent of δ. Passing to the limit $\delta \downarrow 0$ yields Eq. 138 for Φ_{ϵ_ℓ}.

Step 7: Dependence on the lower-level training length. By Lemma 14,

$$\bar{I} = O\left(\frac{K(N+1)}{1-\gamma}\right), \tag{139}$$

and

$$\bar{J} = O\left(\frac{K(N+1)}{1-\gamma}\right). \tag{140}$$

Consequently,

$$\sqrt{\bar{I}} = O\left(\sqrt{\frac{K(N+1)}{1-\gamma}}\right). \tag{141}$$

Substituting Eq. 139–Eq. 141 into Eq. 102 gives

$$\begin{aligned} L_\Phi &= L_G + O\left(\sqrt{\frac{K(N+1)}{1-\gamma}}\right) + O\left(\frac{K(N+1)}{1-\gamma}\right) \\ &= O\left(1 + \frac{K(N+1)}{1-\gamma}\right), \end{aligned} \tag{142}$$

which proves Eq. 103. □

Lemma 18 (Unbiased upper-gradient estimator)**.** *Under Assumption 3 and the hypotheses of Lemmas 10, 13, and 14, let $\widehat{\delta}_G(Z; v)$ be the value-difference estimator defined in Lemma 16, and let $S(Z; v)$ be the complete score defined in Lemma 13. Using an independent fresh v-gradient oracle call, define*

$$\widehat{g}(v) := \widehat{\nabla}_v G(v, \theta(Z)) + \widehat{\delta}_G(Z; v) S(Z; v). \tag{143}$$

Let $\bar{I}$ be the uniform score second-moment bound from Lemma 13, let $V_\theta := M_G^2 + \sigma_\theta^2$, and define

$$V_\Phi := 2(L_{G,0}^2 + \sigma_v^2) + 2(\Delta_G^2 + V_\theta)\bar{I}. \tag{144}$$

Then

$$\mathbb{E}[\widehat{g}(v)] = \nabla\Phi_{\epsilon_\ell}(v), \tag{145}$$

$$\mathbb{E}\|\widehat{g}(v)\|^2 \le V_\Phi, \tag{146}$$

and

$$V_\Phi = O\left(1 + \frac{K(N+1)}{1-\gamma}\right), \tag{147}$$

where $\Phi_{\epsilon_\ell}(v) = \mathbb{E}_{Z\sim P_v}[G(v,\theta(Z))]$. Expectations include the lower trace and the fresh upper-oracle randomness.

Proof. Fix v and let $Z \sim P_v$ be a complete lower training trace. For the vector estimator, $S(Z;v)$ is fixed conditionally on Z. The fresh v-gradient oracle and Lemma 16 give

$$\mathbb{E}[\widehat{g}(v) \mid Z] = \nabla_v G(v,\theta(Z)) + \big(G(v,\theta(Z)) - G(v,\theta_0)\big)S(Z;v).$$

Taking expectation over Z and applying Lemma 10 with baseline $c(v) = G(v,\theta_0)$ proves Eq. 145.

The inequality $\|u+w\|^2 \le 2\|u\|^2 + 2\|w\|^2$ gives

$$\begin{aligned}\mathbb{E}[\|\widehat{g}(v)\|^2 \mid Z] &\le 2\mathbb{E}[\|\widehat{\nabla}_v G(v,\theta(Z))\|^2 \mid Z] + 2\mathbb{E}[\widehat{\delta}_G^2 \mid Z]\|S(Z;v)\|^2 \\ &\le 2(L_{G,0}^2 + \sigma_v^2) + 2(\Delta_G^2 + V_\theta)\|S(Z;v)\|^2.\end{aligned}$$

Here the first oracle second moment is bounded by its squared mean plus its variance. The scalar second-moment bound follows from Eq. 100. Taking expectation and using $\mathbb{E}\|S(Z;v)\|^2 \le \bar{I}$ proves the second-moment bound in Eq. 146. Unbiasedness then implies

$$\mathbb{E}\|\widehat{g}(v) - \nabla\Phi_{\epsilon_\ell}(v)\|^2 = \mathbb{E}\|\widehat{g}(v)\|^2 - \|\nabla\Phi_{\epsilon_\ell}(v)\|^2 \le V_\Phi.$$

Finally, Lemma 14 gives $\bar{I} = O(K(N+1)/(1-\gamma))$. Substitution into the definition of V_Φ proves Eq. 147. □

E.4 Upper-Level Stationarity and End-to-End Guarantee

Upper-Level Convergence

Lemma 19 (Convergence to upper-level stationarity)**.** *Suppose the hypotheses of Lemmas 17 and 18 hold uniformly in v. Let L_Φ and V_Φ be the smoothness and estimator second-moment constants from these lemmas, respectively, and define*

$$\bar{L}_\Phi := \max\{1, L_\Phi\}, \qquad \bar{V}_\Phi := \max\{1, V_\Phi\}.$$

Starting from v_0, generate

$$v_{t+1} := v_t - \alpha\widehat{g}_t,$$

where, conditionally on v_t, $\widehat{g}_t$ is an independent realization of the upper-gradient estimator defined in Lemma 18. For a desired upper accuracy $\epsilon_o \in (0,1]$, choose

$$\alpha = \frac{\epsilon_o}{\bar{L}_\Phi \bar{V}_\Phi}, \tag{148}$$

and

$$T_o = \max\left\{1, \left\lceil \frac{2\Delta_G \bar{L}_\Phi \bar{V}_\Phi}{\epsilon_o^2} \right\rceil\right\}. \tag{149}$$

Let

$$R_o \sim \mathrm{Unif}\{0, \dots, T_o - 1\}$$

be selected independently of all previous randomness. Then

$$\mathbb{E}\left[\|\nabla\Phi_{\epsilon_\ell}(v_{R_o})\|^2\right] \le \epsilon_o. \tag{150}$$

Proof. By Lemma 17, Φ_{ϵ_ℓ} is L_Φ-smooth. Therefore,

$$\Phi_{\epsilon_\ell}(v_{t+1}) \le \Phi_{\epsilon_\ell}(v_t) + \langle \nabla\Phi_{\epsilon_\ell}(v_t), v_{t+1} - v_t\rangle + \frac{L_\Phi}{2}\|v_{t+1} - v_t\|^2.$$

Using

$$v_{t+1} - v_t = -\alpha \widehat{g}_t,$$

we obtain

$$\Phi_{\epsilon_\ell}(v_{t+1}) \le \Phi_{\epsilon_\ell}(v_t) - \alpha \langle \nabla\Phi_{\epsilon_\ell}(v_t), \widehat{g}_t\rangle + \frac{L_\Phi \alpha^2}{2}\|\widehat{g}_t\|^2.$$

Taking conditional expectation given v_t and applying Lemma 18,

$$\mathbb{E}_t[\Phi_{\epsilon_\ell}(v_{t+1})] \le \Phi_{\epsilon_\ell}(v_t) - \alpha\|\nabla\Phi_{\epsilon_\ell}(v_t)\|^2 + \frac{L_\Phi \alpha^2}{2} V_\Phi.$$

Taking total expectation and summing over $t = 0, \dots, T_o - 1$ gives

$$\alpha \sum_{t=0}^{T_o-1} \mathbb{E}\|\nabla\Phi_{\epsilon_\ell}(v_t)\|^2 \le \mathbb{E}[\Phi_{\epsilon_\ell}(v_0) - \Phi_{\epsilon_\ell}(v_{T_o})] + \frac{L_\Phi \alpha^2 V_\Phi T_o}{2}.$$

Since

$$G_{\min} \le \Phi_{\epsilon_\ell}(v) \le G_{\max},$$

we have

$$\Phi_{\epsilon_\ell}(v_0) - \Phi_{\epsilon_\ell}(v_{T_o}) \le \Delta_G.$$

Consequently,

$$\frac{1}{T_o}\sum_{t=0}^{T_o-1} \mathbb{E}\|\nabla\Phi_{\epsilon_\ell}(v_t)\|^2 \le \frac{\Delta_G}{\alpha T_o} + \frac{L_\Phi \alpha V_\Phi}{2}. \tag{151}$$

By Eq. 149,

$$\frac{\Delta_G}{\alpha T_o} \le \frac{\epsilon_o}{2}.$$

Furthermore, using Eq. 148,

$$\frac{L_\Phi \alpha V_\Phi}{2} \le \frac{\epsilon_o}{2}.$$

Thus

$$\frac{1}{T_o}\sum_{t=0}^{T_o-1} \mathbb{E}\|\nabla\Phi_{\epsilon_\ell}(v_t)\|^2 \le \epsilon_o.$$

Uniform independent selection of R_o yields

$$\mathbb{E}\|\nabla\Phi_{\epsilon_\ell}(v_{R_o})\|^2 = \frac{1}{T_o}\sum_{t=0}^{T_o-1} \mathbb{E}\|\nabla\Phi_{\epsilon_\ell}(v_t)\|^2,$$

which proves Eq. 150. □

End-to-End Guarantee and Sample Complexity

We combine the upper stationarity guarantee, the approximate lower-level response guarantee, and the resulting sample requirements in one statement.

Theorem 5 (End-to-end constrained bilevel guarantee and complexity)**.** *Suppose the hypotheses of Theorem 4 and Lemma 19 hold uniformly in v. For $\delta_\ell, \epsilon_o \in (0, 1]$, set $K = \lceil \delta_\ell^{-2} \rceil$ and $\epsilon_\ell := \max\{\sqrt{\epsilon_{\text{bias}}}, \delta_\ell\}$. Take N to be the smallest positive integer satisfying the critic-iteration condition in Theorem 4, and use the remaining lower and upper parameter choices stated in Theorem 4 and Lemma 19. Let $(v_{R_o}, \widehat{\theta}_{\text{out}})$ be the output of Algorithm 1, including its final independent lower run. Conditional on v_{R_o}, denote the trace of this fresh run by $Z_{\text{out}} \sim P_{v_{R_o}}$, so that $\widehat{\theta}_{\text{out}} = \theta(Z_{\text{out}})$. Then*

$$\mathbb{E}\|\nabla\Phi_{\epsilon_\ell}(v_{R_o})\|^2 \le \epsilon_o, \tag{152}$$

$$\mathbb{E}\left[J_r^\star(v_{R_o}) - J_r(\widehat{\theta}_{\mathrm{out}}, x_{R_o})\right] \le \overline{C}_r \epsilon_\ell, \tag{153}$$

and

$$\mathbb{E}\left[-J_c(\widehat{\theta}_{\mathrm{out}}, y_{R_o})\right] \le \overline{C}_c \epsilon_\ell. \tag{154}$$

Here $\overline{C}_r := 1 + C_r$ and $\overline{C}_c := C_{\mathrm{sl}} + C_c$, where C_r, C_c, and C_{sl} are the constants in Theorem 4, and ϵ_{bias} is defined in Assumption 7.

Suppose additionally that each upper-gradient oracle call uses at most $2H$ environment transitions. Let N_{env} denote the expected total number of environment transitions used by Algorithm 1, including both lower training and upper-gradient sampling. Then, for fixed primitive problem constants,

$$N_{\mathrm{env}} = \widetilde{\mathcal{O}}(\delta_\ell^{-6} \epsilon_o^{-2}). \tag{155}$$

Proof. The choices of η, ζ, N, and $\lambda_{\max}$ satisfy the conditions of Theorem 4, while the choices of α and T_o satisfy the conditions of Lemma 19. Therefore Lemma 19 gives Eq. 152.

Condition on v_{R_o}. The deployment parameter $\widehat{\theta}_{\mathrm{out}} = \theta(Z_{\mathrm{out}})$ is generated by an independent fresh trace $Z_{\mathrm{out}} \sim P_{v_{R_o}}$ at this fixed upper-level parameter. Uniform application of Theorem 4 gives

$$\mathbb{E}\left[J_r^\star(v_{R_o}) - J_r(\widehat{\theta}_{\mathrm{out}}, x_{R_o}) \mid v_{R_o}\right] \le \sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_r}{\sqrt{K}},$$

and

$$\mathbb{E}\left[-J_c(\widehat{\theta}_{\mathrm{out}}, y_{R_o}) \mid v_{R_o}\right] \le C_{\mathrm{sl}}\sqrt{\epsilon_{\mathrm{bias}}} + \frac{C_c}{\sqrt{K}}.$$

Because $K^{-1/2} \le \delta_\ell \le \epsilon_\ell$ and $\sqrt{\epsilon_{\mathrm{bias}}} \le \epsilon_\ell$, taking expectation over v_{R_o} proves Eq. 153–Eq. 154.

For the environment-interaction count, define

$$Q_\ell := \frac{K(N+1)}{1-\gamma}.$$

Let M_ℓ be a uniform upper bound on the expected number of environment transitions in one complete lower run. Also let $N_{\mathrm{lower-env}}$ denote the expected transitions used by all lower runs and let N_{upper} denote the expected number of upper-gradient oracle calls. Each complete lower run makes $K(N+1)$ encoded critic calls. Each call uses a constant number of geometric-horizon trajectories whose expected lengths are $O((1-\gamma)^{-1})$, so $M_\ell = O(Q_\ell)$. Lemmas 17, 18, and 14 give

$$L_\Phi = O(1 + Q_\ell), \qquad V_\Phi = O(1 + Q_\ell).$$

Substitution into the definition of T_o in Lemma 19 gives

$$T_o = O\left((1+Q_\ell)^2 \epsilon_o^{-2}\right).$$

The algorithm performs one lower run at each of the T_o upper-level iterations and one fresh deployment run after selecting R_o. Hence

$$N_{\mathrm{lower-env}} \le M_\ell(T_o + 1),$$

and therefore

$$N_{\mathrm{lower-env}} = O\left(Q_\ell + Q_\ell(1+Q_\ell)^2 \epsilon_o^{-2}\right).$$

For $a \ge 0$, $\max\{1, \lceil a \rceil\} \le 1 + a$. The stated adaptive batch size and Lemma 15 therefore give

$$\mathbb{E}[B_G(Z)] \le 1 + \mathbb{E}\|\theta(Z) - \theta_0\|^2 \le 1 + \bar{D}^2.$$

Each upper-level iteration uses $B_G(Z)$ fresh θ-gradient calls and one fresh v-gradient call. Therefore the expected number of calls per upper-level iteration is at most $2 + \bar{D}^2$, and hence

$$N_{\mathrm{upper}} \le (2 + \bar{D}^2) T_o.$$

By the assumed upper-oracle cost bound, each such call uses at most $2H$ environment transitions. Lemma 12 (in Appendix) provides one concrete realization through policy-generated Bradley–Terry samples formed from two trajectories of length at most H. Consequently,

$$N_{\mathrm{env}} \le N_{\mathrm{lower-env}} + 2H N_{\mathrm{upper}}. \tag{156}$$

Finally, the stated choices give

$$K = \Theta(\delta_\ell^{-2}), \qquad N = \Theta(\log(\delta_\ell^{-1})), \qquad Q_\ell = \widetilde{\mathcal{O}}(\delta_\ell^{-2}).$$

Lemma 15 gives $\bar{D}^2 \le \eta_0^2 K W_\omega^2 = O(\delta_\ell^{-2})$. Thus both $N_{\mathrm{lower-env}}$ and $2HN_{\mathrm{upper}}$ are $\widetilde{\mathcal{O}}(\delta_\ell^{-6}\epsilon_o^{-2})$. Combining these bounds in Eq. 156 proves Eq. 155. □

Remark 3 (No differentiation through the lower-level algorithm)**.** *None of the bounds above requires differentiation through the full sequence of lower-level updates, including the projected multiplier update, or an inverse Hessian of a lower-level saddle system. The dependence on the length of lower training appears only through the score moments $\bar{I}$ and $\bar{J}$, which grow linearly with the expected number of encoded signals.*