SAMPLE-EFFICIENT REINFORCEMENT LEARNING WITH GENERAL UTILITIES VIA NEURAL DUAL CRITICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with general utilities optimizes nonlinear functionals of policy-induced occupancy measures. We develop a natural actor-critic method for concave RLGU objectives admitting an exact variational dual representation, with explicit constructions for a broad class of separable occupancy utilities. Rather than estimating the occupancy measure and then forming the policy-gradient pseudo-reward, our method learns this pseudo-reward directly as a function of the state-action pair using a neural critic. The critic is trained by projected stochastic updates, while multilevel Monte Carlo combines outputs across training horizons to control its bias at logarithmic expected cost. Our analysis separates conditional actor bias from estimator second moments and accounts explicitly for dual optimization, critic approximation, and neural linearization errors. Under policy-transfer, dual-stability, and local critic-regularity conditions, the method achieves $\widetilde{\mathcal{O}}(T^{-1/2})$ global convergence above policy, critic-class, and finite-width approximation errors using $\widetilde{\mathcal{O}}(T)$ expected environment transitions and reference samples. For a fixed critic representation, this yields $\widetilde{\mathcal{O}}(\epsilon^{-2})$ sampling complexity above the approximation floor, covering infinite state-action spaces and occupancy utilities beyond fixed finite vectors of returns.

1 INTRODUCTION

Standard reinforcement learning (RL) maximizes an expected discounted sum of rewards, an objective linear in the policy-induced occupancy measure. Many applications, however, depend on the occupancy distribution itself rather than on a fixed linear reward. For example, imitation learning seeks to match the learner's occupancy measure to an expert's target distribution, while exploration may encourage high visitation entropy (Ho & Ermon, 2016; Hazan et al., 2019). Reinforcement learning with general utilities (RLGU) provides a unified formulation:

$$J(\theta) = F(\lambda_\theta),$$

where λ_θ is the normalized discounted state-action occupancy measure and F is a utility functional. We focus on concave utilities and generally parameterized policies. Although J is generally nonconcave in θ, concavity of F in the occupancy measure provides the global comparison structure used in our analysis (Zhang et al., 2020; Barakat et al., 2025).

The main statistical challenge is the policy-dependent pseudo-reward $r_\theta = \nabla_\lambda F(\lambda_\theta)$. Holding this reward fixed, the policy gradient takes the form

$$\nabla_\theta J(\theta) = (1-\gamma)\, \nabla_\vartheta V^{\pi_\vartheta}(r_\theta)|_{\vartheta=\theta}\,. \tag{1}$$

For small finite spaces, one can estimate the occupancy measure directly and then evaluate r_θ. In larger spaces, however, occupancy approximation becomes an additional learning problem. Existing linear and parametric occupancy estimators yield $\widetilde{\mathcal{O}}(\epsilon^{-4})$ guarantees for stationarity (Barakat et al., 2023) and global optimality (Barakat et al., 2026), respectively.

In this work, we study concave RLGU objectives admitting an exact variational dual representation, and give explicit constructions for a broad class of separable occupancy utilities. Instead of first

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.

Table 1: Comparison of representations, theoretical support for infinite state–action spaces, and sample complexity in RLGU. Here, $\widetilde{\mathcal{O}}$ suppresses logarithmic factors. Our bound holds above the approximation floor for a fixed critic representation under the stated policy-transfer, dual-stability, and local neural-regularity assumptions. † Assumes an oracle cost player whose implementation cost is excluded. ‡ Establishes first-order stationarity; all other entries concern global optimality.

Reference	Representation	Infinite Spaces	Sample Complexity
Zhang et al. (2021)	Tabular	✗	$\widetilde{\mathcal{O}}(\epsilon^{-2})$
Zahavy et al. (2021)	Tabular (UCRL2)	✗	$\widetilde{\mathcal{O}}(\epsilon^{-2})^{\dagger}$
Barakat et al. (2023)	Linear occupancy	✗	$\widetilde{\mathcal{O}}(\epsilon^{-4})^{\ddagger}$
Barakat et al. (2026)	Parametric occupancy	✗	$\widetilde{\mathcal{O}}(\epsilon^{-4})$
This work	**Neural dual critic**	✓	$\widetilde{\mathcal{O}}(\epsilon^{-2})$

estimating λ_θ and then forming r_θ, we directly learn the dual pseudo-reward as a function of (s, a) using a neural critic. This places the approximation requirement on the reward function actually needed by the actor, rather than on the full occupancy density, and allows the framework to extend naturally to infinite state-action spaces.

Our method and analysis hinge on two technical ideas.

Technical novelty: learning the dual reward. Building on Fenchel duality and variational policy gradients (Zhang et al., 2020; Zahavy et al., 2021), we learn the pseudo-reward through stochastic neural dual optimization (Section 2.2). Our finite-time analysis uses dual curvature and feature conditioning to establish a stable recursion toward the constrained dual minimizer, separating dual-class approximation from neural linearization error. Table 2 in Appendix E gives reference-sampling estimators for the required dual gradients.

Technical novelty: controlling the bias of a trained critic. The critic's $\mathcal{O}(H^{-1/2})$ root-mean-square error above its approximation floor would require long training if used directly as actor bias. Adapting MLMC and neural actor-critic analyses (Blanchet & Glynn, 2015; Ganesh et al., 2025a; Naskar & Aggarwal, 2026), we apply randomized corrections to critic output functions at dyadic horizons, matching the longest run's mean at logarithmic expected training cost (Lemma 4). Here MLMC controls a learned dual reward rather than finite scalarization coefficients. Separating actor bias from second moments in the KL-potential bound (Lemma 1) and adjusting the actor stepsize preserves the $\widetilde{\mathcal{O}}(T^{-1/2})$ statistical rate with explicit finite-width error.

Contributions: Our key contributions are as follows:

- We develop a neural dual approach to concave RLGU that learns the utility-induced pseudo-reward directly, avoiding explicit occupancy estimation. For separable utilities, we derive explicit dual objectives and sampling formulas (Table 2, and Proposition 1 in Appendix E).
- We introduce a function-valued MLMC correction across critic-training horizons that inherits the bias of a long critic run at logarithmic expected training cost. We establish critic bounds and a natural-gradient analysis separating conditional actor bias from estimator second moments (Lemmas 3, 4, and 2).
- Combining these ingredients, we obtain $\widetilde{\mathcal{O}}(\epsilon^{-2})$ expected sampling complexity above an explicit approximation floor for a fixed critic representation, including infinite state–action spaces (Theorem 1). Table 1 contrasts this guarantee with existing RLGU results in representation, scope, and complexity.

1.1 Related Works:

General-utility and convex reinforcement learning: Hazan et al. (2019) study concave visitation objectives through planning, and Zhang et al. (2020) develop the variational policy-gradient and Fenchel-dual formulation. Zahavy et al. (2021) express convex RL as a policy-cost game, while

Geist et al. (2022) connect it to mean-field games and Frank-Wolfe updates. We build on these objective and duality foundations, but analyze a trained neural pseudo-reward.

Large-state RLGU and occupancy approximation: Barakat et al. (2023) provide variance-reduced tabular methods and a linear occupancy approximation method. Barakat et al. (2025) obtain global guarantees using maximum-likelihood occupancy estimation. Table 1 distinguishes their representation assumptions and guarantees. Our dual-reward approach instead imposes approximation and stability conditions on the utility gradient rather than the occupancy density.

Concave utility RL. Finite-objective concave utility RL (CURL) optimizes a concave function of finitely many expected returns (Agarwal & Aggarwal, 2023; Agarwal et al., 2022; Bai et al., 2022). For finite state-action spaces, it is equivalent to concave RLGU: $M = |\mathcal{S}||\mathcal{A}|$ indicator rewards recover the occupancy vector through normalized returns (Ganesh & Aggarwal, 2026). In infinite spaces, general RLGU is strictly more expressive, since finitely many occupancy integrals cannot represent every occupancy utility. Proposition 2 in Appendix F gives the finite reduction and a continuous-state counterexample.

Variational and Fenchel-dual approaches: Our stochastic dual learning also relates to variational divergence estimation (Nowozin et al., 2016). The actor analysis adapts approximate NPG arguments from Agarwal et al. (2021); Ganesh et al. (2025a;b), while the critic-output correction adapts MLMC ideas from Blanchet & Glynn (2015); Naskar & Aggarwal (2026). Ganesh & Aggarwal (2026) obtain the ϵ^{-2} exponent for finite-objective concave scalarization; our extension treats a function-valued dual reward and accounts for its optimization, approximation, and sampling errors.

2 Problem Setup

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \rho, \gamma)$ be a discounted MDP on compact metric state and action spaces with their Borel sigma-algebras. Here $\mathcal{P}$ is a transition kernel, ρ an initial-state distribution, and $\gamma \in (0, 1)$. Finite spaces are included. A stationary policy π induces $s_0 \sim \rho$, $a_t \sim \pi(\cdot \mid s_t)$, and $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$. Its normalized occupancy measure is

$$\lambda_\pi(B) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi\big((s_t, a_t) \in B\big) \tag{2}$$

for measurable $B \subseteq \mathcal{S} \times \mathcal{A}$, where $\mathbb{P}_\pi$ denotes the trajectory law induced by the policy π.

We consider differentiable policy densities $\pi_\theta(a \mid s)$ with respect to a fixed action reference measure and parameter-independent support, with $\theta \in \mathbb{R}^d$. The policy assumptions stated below hold along all update segments and at the comparison policy. Write $\lambda_\theta = \lambda_{\pi_\theta}$ and d_θ for its state marginal. For a convex domain $\mathcal{D}$ of probability measures containing all policy occupancies, our objective is

$$F^* = F(\lambda_{\theta^*}) = \max_{\theta \in \mathbb{R}^d} F(\lambda_\theta). \tag{3}$$

We assume the maximum is attained; the same analysis applies to any fixed comparator with finite initial KL divergence. We now state the regularity required of the utility functional.

Assumption 1 (Concavity and Differentiability) *$F : \mathcal{D} \to \mathbb{R}$ is concave and differentiable at each policy occupancy. Its functional derivative $r_\theta = \nabla_\lambda F(\lambda_\theta)$ represents directional derivatives along segments in $\mathcal{D}$, and the chain rule holds along differentiable policy paths. Moreover, $|r_\theta|_\infty \leq R_{\max}$ uniformly over the policies under consideration.*

Concavity provides the global comparison used in the convergence analysis, while bounded pseudo-rewards control the moments of the policy-gradient estimator.

2.1 Natural Policy Gradient

We next express the utility gradient in a form suitable for natural policy optimization. Let $u_\theta(s, a) = \nabla_\theta \log \pi_\theta(a \mid s)$ denote the score. For a fixed bounded reward r, define

$$V^{\pi_\theta}(r) = \mathbb{E}_{\pi_\theta}\left[\sum_{t \geq 0} \gamma^t r(s_t, a_t)\right] = \frac{1}{1-\gamma} \int r d\lambda_\theta. \tag{4}$$

Applying the functional chain rule to $J(\theta) = F(\lambda_\theta)$ and then the likelihood-ratio identity gives

$$\nabla_\theta J(\theta) = (1-\gamma)\, \nabla_\vartheta V^{\pi_\vartheta}(r_\theta)|_{\vartheta=\theta} = (1-\gamma)\mathbb{E}_{\pi_\theta}\left[\sum_{t\geq 0}\gamma^t r_\theta(s_t,a_t)\sum_{j=0}^{t} u_\theta(s_j,a_j)\right]. \quad (5)$$

Here r_θ is held fixed when differentiating with respect to ϑ. This is the general-utility policy-gradient identity (Zhang et al., 2020; Kumar et al., 2024); bounded rewards and scores justify the interchange with the discounted sum.

Natural policy gradient (NPG) preconditions this gradient using the Fisher matrix (Kakade, 2001):

$$\mathcal{I}(\theta) = \mathbb{E}_{\lambda_\theta}[u_\theta u_\theta^\top], \qquad w^*(\theta) = \mathcal{I}(\theta)^{-1}\nabla_\theta J(\theta), \quad (6)$$

where to find the NPG vector $w^*(\theta)$, we minimize the quadratic objective

$$\ell(\theta, w) = \frac{1}{2} w^\top \mathcal{I}(\theta) w - w^\top \nabla_\theta J(\theta). \quad (7)$$

The policy-gradient theorem (Sutton et al., 1999) gives $\nabla_\theta J(\theta) = \mathbb{E}_{\lambda_\theta}[u_\theta A_\theta^{(r_\theta)}]$, where $A_\theta^{(r)}$ is the advantage function associated with a fixed reward r. Specifically, define

$$\begin{aligned} Q_\theta^{(r)}(s,a) &= \mathbb{E}_{\pi_\theta}\left[\sum_{t\geq 0}\gamma^t r(s_t,a_t) \mid s_0 = s, a_0 = a\right], \\ A_\theta^{(r)}(s,a) &= Q_\theta^{(r)}(s,a) - \int Q_\theta^{(r)}(s,a')\pi_\theta(da' \mid s). \end{aligned} \quad (8)$$

Combining this identity with (6), $w^*(\theta)$ is the least-squares coefficient obtained by regressing $A_\theta^{(r_\theta)}$ onto the score u_θ.

2.2 Pseudo-reward via Exact Dual Representation

We next characterize the pseudo-reward through a dual representation that avoids explicit occupancy estimation. Let $\mathcal{R}$ denote the bounded measurable functions on $\mathcal{S} \times \mathcal{A}$.

Assumption 2 (Exact Dual Representation) *There exists a concave functional $\Psi : \mathcal{R} \to \mathbb{R} \cup -\infty$ such that, for every policy under consideration,*

$$F(\lambda_\theta) = \min_{r\in\mathcal{R}} \left\{\mathbb{E}_{\lambda_\theta}[r] - \Psi(r)\right\}, \qquad r_\theta \in \arg\min_{r\in\mathcal{R}} \left\{\mathbb{E}_{\lambda_\theta}[r] - \Psi(r)\right\}. \quad (9)$$

This is the variational RLGU representation of Zhang et al. (2020). Its key advantage is that the unknown occupancy enters only through an expectation. Since the dual objective touches F at r_θ, differentiability gives

$$\nabla_\theta J(\theta) = (1-\gamma)\, \nabla_\vartheta V^{\pi_\vartheta}(r_\theta)|_{\vartheta=\theta}. \quad (10)$$

Thus, learning an exact dual minimizer is sufficient to recover the pseudo-reward required by the actor.

We now give a concrete class of utilities satisfying Assumption 2. Let ν be a finite reference measure with $\nu_0 = \nu(\mathcal{S}\times\mathcal{A}) > 0$, assume $\lambda_\theta \ll \nu$, and write $p_\theta = d\lambda_\theta/d\nu$.

Assumption 3 (Separable Concave Utility) *The utility has the form*

$$F(\lambda_\theta) = C - \int f(p_\theta)\, d\nu - \int c p_\theta\, d\nu, \quad (11)$$

where f is proper, closed, and convex, and c is measurable. At attained density values in the interior of its effective domain, f is differentiable; at an attained endpoint, f' denotes a finite one-sided derivative in ∂f. All integrals are finite, and $|f'(p_\theta) + c|$ is bounded uniformly in θ, ν-almost everywhere,

Remark 1 *Assumption 3 covers visitation-entropy and relative-entropy objectives used for exploration (Hazan et al., 2019; Zhang et al., 2020), KL-based occupancy matching for imitation (Zhang et al., 2020), more general f-divergence objectives (Nowozin et al., 2016), and quadratic density matching. The bounded-derivative condition holds under the corresponding density-range conditions summarized in Table 2 in the appendix; details are given in Appendix E.*

2.3 NEURAL CRITIC ESTIMATION

We represent the pseudo-reward by an L-layer neural network. For bounded features $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^p$, let $x^{(0)} = \phi(s, a)$ and

$$x^{(\ell)} = m^{-1/2}\sigma(W^{(\ell)}x^{(\ell-1)}), \quad \ell = 1, \dots, L, \qquad r(s, a, \zeta) = m^{-1/2}b^\top x^{(L)}. \tag{12}$$

Here $W^{(1)} \in \mathbb{R}^{m\times p}$, $W^{(\ell)} \in \mathbb{R}^{m\times m}$ for $\ell \geq 2$, and b is fixed. The trainable weights form $\zeta \in \mathbb{R}^{D_m}$, where $D_m = mp + (L-1)m^2$.

To control the network locally around its initialization, we impose the following common assumption:

Assumption 4 (Smooth Activation) *The activation σ is Lipschitz with a Lipschitz derivative.*

Assumption 4 is satisfied by commonly used smooth approximations of ReLU such as Sigmoid, ELU, and GeLU Hendrycks & Gimpel (2016); Clevert et al. (2015). Moreover, this ensures that the network admits a controlled first-order linearization around its initialization, as used in NTK analyses (Jacot et al., 2018; Ganesh et al., 2025a; Satheesh et al., 2026).

To control the critic using NTK analysis around its initialization, fix ζ_0 and restrict the parameters to a ball Ω of radius R centered at ζ_0, and define

$$\begin{aligned} g_0(s, a) &:= \nabla_\zeta r(s, a, \zeta_0), \qquad \mathcal{F}_{\Omega,m} := r_{\text{lin}}(\cdot, \zeta) : \zeta \in \Omega, \\ r_{\text{lin}}(s, a, \zeta) &:= r(s, a, \zeta_0) + g_0(s, a)^\top(\zeta - \zeta_0). \end{aligned} \tag{13}$$

We then require the initialization, tangent features, and linearization error to satisfy

$$\begin{aligned} &\sup_{(s,a)} |r(s, a, \zeta_0)| \leq C_0, \\ &\sup_{(s,a),\zeta\in\Omega} |r(s, a, \zeta) - r_{\text{lin}}(s, a, \zeta)| \leq \eta_m. \\ &\sup_{(s,a)} \|g_0(s, a)\|_2 \leq C_\phi, \end{aligned} \tag{14}$$

The term η_m captures the finite-width error. For example, a uniform parameter-Hessian bound $C_H/\sqrt{m}$ gives $\eta_m \leq C_H R^2/(2\sqrt{m})$ by Lemma 6.

3 ALGORITHM

At outer iteration k, we hold π_{θ_k} fixed, learn a dual reward, estimate a natural-gradient direction, and update

$$\theta_{k+1} = \theta_k + \alpha w_k, \qquad k = 0, \dots, K-1. \tag{15}$$

Policy trajectories are independently restarted from ρ. States within each trajectory follow the MDP and may be dependent. Conditional on the current policy, critic trajectories, NPG trajectories, and reference samples use independent randomness. One environment transition counts as one unit of trajectory cost.

Algorithm 1: Neural Dual-Functional NAC

Require: $\theta_0, \zeta_0, K, B, H_{\max} = 2^J, \alpha, \{\beta_{\zeta,h}\}, \beta_w, H_w$

for $k = 0, \dots, K-1$ **do**
  Draw $N_k \sim \text{Geom}(1/2)$ and set $H_k = 2^{N_k}$ if $N_k \leq J$, otherwise $H_k = 1$
  Initialize $\zeta_0^k = \zeta_0$
  for $h = 0, \dots, H_k - 1$ **do**
    Sample a fresh length-B trajectory under π_{θ_k} and an independent dual-oracle draw
    Update ζ_{h+1}^k by (18)
  end for
  If $N_k \leq J$, retain the shared-run outputs $R_{k,1}, R_{k,2^{N_k-1}}, R_{k,2^{N_k}}$; otherwise retain $R_{k,1}$
  Form $r_{k,\text{MLMC}}$ by (16)
  Initialize $w_0^k = 0$
  for $h = 0, \dots, H_w - 1$ **do**
    Sample a fresh length-B trajectory under π_{θ_k}
    Compute $\widehat{\mathcal{I}}^{k,h}$ and $\widehat{g}^{k,h}$ by (19)
    Update w_{h+1}^k by (20)
  end for
  Set $w_k = w_{H_w}^k$ and $\theta_{k+1} = \theta_k + \alpha w_k$
end for

3.1 CRITIC ESTIMATOR

For fixed θ_k, the critic minimizes the dual objective

$$\mathcal{E}(\theta_k, \zeta) = \mathbb{E}_{\lambda_{\theta_k}}[r(\cdot, \zeta)] - \Psi(r(\cdot, \zeta)).$$

We use the frozen tangent feature g_0 from (13) in the critic update, while evaluating the nonlinear reward inside the dual derivative. The resulting error is controlled by the local linearization bound (14).

MLMC Critic Function. Choose $H_{\max} = 2^J$ and draw $N_k \in 1, 2, \ldots$ with $\mathbb{P}(N_k = j) = 2^{-j}$. If $N_k \le J$, run the critic for 2^{N_k} updates and retain the outputs at horizons 1, 2^{N_k-1}, and 2^{N_k}; otherwise, run one update. Writing $R_{k,H} = r(\cdot, \zeta_H^k)$, define

$$r_{k,\mathrm{MLMC}} = R_{k,1} + \mathbf{1}_{N_k \le J} 2^{N_k} \left(R_{k,2^{N_k}} - R_{k,2^{N_k-1}} \right). \tag{16}$$

All retained outputs are generated from the same critic run, so the dyadic levels share the required prefixes. The correction acts on critic functions and may lie outside the neural class; its bias and second moment are controlled in the analysis.

Vanilla Critic Subroutine. To estimate the dual-gradient term, assume access at each $\zeta \in \Omega$ to a random vector $Z_\Psi(\zeta)$ satisfying

$$\mathbb{E}[Z_\Psi(\zeta) \mid \zeta] = \int g_0(s,a)\, \nabla_r \Psi(r(\cdot,\zeta))(s,a)\, d\nu(s,a), \qquad \mathbb{E}[\|Z_\Psi(\zeta)\|_2^2 \mid \zeta] \le S_\Psi^2. \tag{17}$$

For the utilities in Table 2, one or two independent reference draws provide such an estimator; explicit formulas are given in Proposition 1. Exact integration is a special case, and reference draws are counted separately from environment transitions.

Initialize $\zeta_0^k = \zeta_0$. For $h = 0, 1, \ldots$, generate a fresh length-B trajectory under π_{θ_k} and an independent dual-oracle draw, and update

$$\zeta_{h+1}^k = \Pi_\Omega \left(\zeta_h^k - \beta_{\zeta,h} \widehat{v}^{k,h} \right), \qquad \widehat{v}^{k,h} = (1-\gamma) \sum_{t=0}^{B-1} \gamma^t g_0(s_t^{k,h}, a_t^{k,h}) - Z_\Psi(\zeta_h^k). \tag{18}$$

Here $\beta_{\zeta,h} = \frac{2}{\mu(h+2)}$ with $\mu = \mu_c \kappa$, where μ_c and κ are defined in Section 4. The decreasing stepsize gives uniform control across the dyadic horizons used by the MLMC estimator, while projection keeps the critic within the region covered by (14).

3.2 NPG Estimator

Hold the sampled function $r_{k,\mathrm{MLMC}}$ fixed throughout. Start at $w_0^k = 0$. At each of H_w iterations, generate a fresh length-B trajectory, independent of critic training, and compute

$$\widehat{\mathcal{I}}^{k,h} = (1-\gamma) \sum_{t=0}^{B-1} \gamma^t u_{\theta_k}(s_t^{k,h}, a_t^{k,h}) u_{\theta_k}(s_t^{k,h}, a_t^{k,h})^\top, \tag{19}$$

$$\widehat{g}^{k,h} = (1-\gamma) \sum_{t=0}^{B-1} \gamma^t r_{k,\mathrm{MLMC}}(s_t^{k,h}, a_t^{k,h}) \sum_{j=0}^{t} u_{\theta_k}(s_j^{k,h}, a_j^{k,h}).$$

Apply the stochastic quadratic update

$$w_{h+1}^k = w_h^k - \beta_w \big(\widehat{\mathcal{I}}^{k,h} w_h^k - \widehat{g}^{k,h} \big), \qquad w_k = w_{H_w}^k. \tag{20}$$

The Fisher and gradient estimates may use the same trajectory. Conditional independence across iterations yields a linear mean recursion, so the actor bias depends on the *mean* critic error while the second-moment analysis captures the full MLMC randomness.

4 Assumptions and Main Result

Before we state our main result in Theorem 1, we present our key assumptions.

Assumption 5 (KL Potential) *The comparison policy π_{θ^*} satisfies*

$$\Phi_0 := \mathbb{E}_{s \sim d_{\theta^*}} \mathrm{KL}\big(\pi_{\theta^*}(\cdot \mid s) \| \pi_{\theta_0}(\cdot \mid s) \big) < \infty.$$

Assumption 6 (Score Function) *For all relevant parameters and state-action pairs,*

$$\|u_\theta(s,a)\|_2 \le G_1, \qquad \|u_\theta(s,a) - u_{\theta'}(s,a)\|_2 \le G_2\|\theta - \theta'\|_2.$$

Assumption 7 (Fisher Non-Degeneracy) *There exists $\mu_F > 0$ such that $\mathcal{I}(\theta) \succeq \mu_F I_d$ for all policies under consideration.*

Assumption 8 (Transferred Compatibility) *Let $w^*(\theta)$ be the on-policy natural-gradient coefficient in* (6)*. For some $\varepsilon_{\text{bias}} \ge 0$,*

$$\mathcal{L}_{\lambda_{\theta^*}}(\theta, w^*(\theta)) := \mathbb{E}_{\lambda_{\theta^*}}\left[\left(u_\theta(s,a)^\top w^*(\theta) - A_\theta^{(r_\theta)}(s,a)\right)^2\right] \le \varepsilon_{\text{bias}}. \tag{21}$$

We next state the critic conditions. All functional norms below are in $L_2(\nu)$.

Assumption 9 (Uniform Dual Stability) *The functional Ψ is differentiable, μ_c-strongly concave, and L_c-smooth on a convex set containing $\{r(\cdot,\zeta) : \zeta \in \Omega\}$ and $\mathcal{F}_{\Omega,m}$. The oracle condition* (17) *holds uniformly on Ω.*

Assumption 10 (Gram Matrix Non-Degeneracy) *The tangent-feature Gram matrix*

$$G = \int g_0(s,a) g_0(s,a)^\top \, \nu(ds,da) \tag{22}$$

satisfies $G \succeq \kappa I_{D_m}$ for some $\kappa > 0$.

Define the linearized-dual objective and its minimizer by

$$\mathcal{E}^{\text{lin}}(\theta,\zeta) := \mathbb{E}_{\lambda_\theta}[r_{\text{lin}}(\cdot,\zeta)] - \Psi(r_{\text{lin}}(\cdot,\zeta)), \quad \zeta_\theta^* := \arg\min_{\zeta\in\Omega} \mathcal{E}^{\text{lin}}(\theta,\zeta), \quad \bar{r}_\theta := r_{\text{lin}}(\cdot,\zeta_\theta^*). \tag{23}$$

Assumption 11 (Critic Approximation Error) *For some $\varepsilon_{\text{app}} \ge 0$, uniformly over the policies considered,*

$$\|r_\theta - \bar{r}_\theta\|^2_{L_2(\lambda_\theta)} \le \varepsilon_{\text{app}}. \tag{24}$$

Discussion of assumptions. Assumptions 5–8 are standard in global analyses of approximate natural policy gradient (Agarwal et al., 2021; Ganesh et al., 2025a;b). The KL potential supports comparison with π_{θ^*}; bounded and Lipschitz scores control likelihood-ratio moments and the KL remainder; and Fisher non-degeneracy ensures a well-conditioned natural-gradient direction. Transferred compatibility captures approximation error and distribution shift to the comparator occupancy; an on-policy residual bound e together with $\lambda_{\theta^*} \le C\lambda_\theta$ gives $\varepsilon_{\text{bias}} \le Ce$ (Agarwal et al., 2021; Ganesh et al., 2025a). Assumptions 9–11 govern the dual critic. Strong concavity and smoothness stabilize the variational reward-learning problem (Zhang et al., 2020; Nowozin et al., 2016), while Gram non-degeneracy provides the required feature conditioning (Ganesh et al., 2025a; Satheesh et al., 2026). Together, they give the critic strong-convexity constant $\mu = \mu_c\kappa$. Moreover,

$$D_m\kappa \le \operatorname{tr} G \le \nu_0 C_\phi^2, \tag{25}$$

so the conditioning may deteriorate as the critic dimension grows. For finite spaces, $\operatorname{rank} G \le |\mathcal{S}||\mathcal{A}|$, and hence Assumption 10 cannot hold with $\kappa > 0$ once $D_m > |\mathcal{S}||\mathcal{A}|$. Finally, ε_{app} isolates pseudo-reward approximation error, paralleling occupancy-model realizability assumptions (Barakat et al., 2025); Proposition 3 relates it to $L_2(\nu)$ approximation. These conditions separate policy approximation, critic approximation, conditioning, and finite-width effects in the final bound.

Under Assumptions 1–11, we have the following main result:

Theorem 1 (Global convergence) *Fix the critic representation (m, R, ζ_0) and the constants in* (14)*. Suppose Assumptions 1, 2, 4, and 5–11 hold under the sampling model of Section 3. For $T \ge 2$, choose $K = T$, $H_{\max} = 2^{\lceil \log_2 T\rceil}$, and*

$$B = \left\lceil \frac{\max\{2\log T, \log(2G_1^2/\mu_F)\}}{\log(1/\gamma)} \right\rceil, \qquad 0 < \beta_w \le \min\left\{\frac{1}{G_1^2}, \frac{\mu_F}{8G_1^4}\right\},$$
$$H_w = \left\lceil \frac{4\log T}{\mu_F \beta_w} \right\rceil, \qquad V_T = 1 + \log(2H_{\max}) + H_{\max}(\eta_m^2 + \gamma^{2B}),$$
$$\beta_{\zeta,h} = \frac{2}{\mu(h+2)}, \qquad \alpha = \frac{1-\gamma}{\sqrt{KV_T}}. \tag{26}$$

Then there exists a constant C, independent of T, such that

$$\mathcal{G}_K := F^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}F(\lambda_{\theta_k}) \le \sqrt{\varepsilon_{\text{bias}}} + \frac{C}{1-\gamma}\left(\sqrt{\varepsilon_{\text{app}}} + \eta_m + \sqrt{\frac{1+\log T}{T}}\right). \tag{27}$$

Remark 2 (Expected Sample Complexity) *The expected environment-transition cost is $\mathcal{O}(KB[H_w + 1 + \log H_{\max}])$. For fixed representation and problem constants, Theorem 1 therefore gives $\widetilde{\mathcal{O}}(\epsilon^{-2})$ expected sample complexity above the approximation floor $\sqrt{\varepsilon_{\text{bias}}} + C(\sqrt{\varepsilon_{\text{app}}} + \eta_m)/(1-\gamma)$. An $\mathcal{O}(m^{-1/2})$ Hessian bound gives $\eta_m = \mathcal{O}(m^{-1/2})$ for fixed R. If m grows with T, the width dependence of C, particularly through κ in* (25)*, must also be tracked. Reference samples are counted separately, while representation-dependent neural and matrix computations are excluded.*

5 Proof Outline

In this section, we outline the proof of Theorem 1, which proceeds in four key steps. Before presenting them, let $\mathcal{F}_k$ denote the σ-field generated by all randomness prior to outer iteration k, and write $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$. We also fix a critic initialization satisfying (14).

Step 1. General Framework for NPG: Combining the KL-potential argument (Agarwal et al., 2021; Ganesh et al., 2025a;b) with concavity and transferred compatibility yields the following bound, proved in Appendix A.

Lemma 1 (Optimality Gap) *Under Assumptions 1 and 5–8,*

$$\mathcal{G}_K \le \sqrt{\varepsilon_{\text{bias}}} + \frac{\Phi_0}{\alpha K} + \frac{G_1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\mathbb{E}_k w_k - w^*(\theta_k)\|_2 + \frac{G_2\alpha}{2K}\sum_{k=0}^{K-1}\mathbb{E}\|w_k\|_2^2. \tag{28}$$

Actor bias enters linearly, while the second moment is weighted by α, allowing a noisy MLMC estimator with an accurate mean, at the cost of a logarithmically increasing second moment.

Step 2. NPG Analysis: For the trained pseudo-reward, define

$$\delta_k^2 = \|\mathbb{E}_k r_{k,\text{MLMC}} - \bar{r}_{\theta_k}\|_{L_2(\lambda_{\theta_k})}^2, \qquad \sigma_k^2 = \mathbb{E}_k\|r_{k,\text{MLMC}} - \bar{r}_{\theta_k}\|_{L_2(\lambda_{\theta_k})}^2. \tag{29}$$

The conditional expectation in σ_k^2 includes critic training and the random level. The NPG trajectories are sampled afterward. Given the frozen critic, their mean update is a linear recursion with matrix $\mathcal{I}_B(\theta_k) = \mathbb{E}\widehat{\mathcal{I}}^{k,h}$. Its difference from $\mathcal{I}(\theta_k)$ is at most $G_1^2\gamma^B$. Crucially, the gradient estimate is linear in the supplied reward. This transfers δ_k to the actor bias rather than replacing it by σ_k, giving us the following result:

Lemma 2 (NPG Bounds) *Suppose the bounded reward, score, Fisher, and critic-approximation conditions hold, $G_1^2\gamma^B \le \mu_F/2$, and β_w satisfies* (26)*. Let $\tau_B = \gamma^B(B + (1-\gamma)^{-1})$. For a constant C depending only on the fixed regularity bounds,*

$$\begin{aligned}\|\mathbb{E}_k w_k - w^*(\theta_k)\|_2^2 &\le C\left[\tau_B^2 + \frac{\varepsilon_{\text{app}} + \delta_k^2 + e^{-\mu_F\beta_w H_w}}{(1-\gamma)^2}\right],\\ \mathbb{E}_k\|w_k\|_2^2 + \mathbb{E}_k\|w_k - w^*(\theta_k)\|_2^2 &\le \frac{C}{(1-\gamma)^2}(1+\sigma_k^2).\end{aligned} \tag{30}$$

Appendix B proves these bounds using a reset-trajectory specialization of the general linear stochastic-recursion result in Lemma 7. The required second moment is bounded rather than vanishing with H_w; logarithmically many iterations are used to reduce the mean initialization error.

Step 3. Dual-Critic Analysis: We next control δ_k and σ_k. Write $\zeta_k^* = \zeta_{\theta_k}^*$ and $q_m = \eta_m^2 + \gamma^{2B}$. Trajectory truncation and neural linearization induce the conditional gradient bias

$$d_v = C_\phi\gamma^B + \nu_0 C_\phi L_c\eta_m,$$

while the reference-oracle noise remains conditionally unbiased with bounded second moment. Dual curvature, the Gram condition $G \succeq \kappa I$, and the constrained minimizer's variational inequality yield a projected-SGD contraction with effective curvature $\mu_c \kappa$. Transferring the resulting parameter error through the tangent features, we prove the following result in Appendix C:

Lemma 3 (Vanilla-Critic Bounds) *Under* (14)*,* (17)*, and Assumptions 9–10, the updates* (18) *satisfy, for all* $H \geq 1$*,*

$$\max\left\{\mathbb{E}_k\|\zeta_H^k - \zeta_k^*\|_2^2,\, \mathbb{E}_k\|R_{k,H} - \bar{r}_{\theta_k}\|_{L_2(\lambda_{\theta_k})}^2\right\} \leq \frac{C}{H+1} + Cq_m. \tag{31}$$

Thus, a vanilla critic trained for H steps has squared error of order $\mathcal{O}(H^{-1})$ above the approximation floor. Consequently, obtaining the $\widetilde{\mathcal{O}}(T^{-1})$ squared critic bias required by the actor would necessitate $H = \Theta(T)$ critic updates at every outer iteration. Since the actor itself performs $K = \Theta(T)$ outer iterations, using such a critic directly would lead to a quadratic-order critic-training cost and would therefore preclude the desired $\widetilde{\mathcal{O}}(T)$ overall sampling complexity. The purpose of the MLMC construction is to retain the mean accuracy of a critic trained for $H_{\max}$ steps while reducing its expected optimization cost to logarithmic order.

Rather than correcting a scalar statistic, we exploit the MLMC construction in (16). Recall that $H_{\max} = 2^J$ and $N_k \sim \text{Geom}(1/2)$. This gives an expected training horizon of

$$\mathbb{E}[H_k] = \sum_{j=1}^{J} 2^{-j} 2^j + \mathbb{P}(N_k > J) = J + 2^{-J} \leq 1 + \log_2 H_{\max}.$$

At the same time, the dyadic correction in (16) telescopes in function space, so its conditional expectation matches that of the critic trained for the maximum horizon $H_{\max}$. Consequently, the long-horizon bound in Lemma 3 transfers directly to the conditional bias quantity δ_k. In contrast, the second moment σ_k^2 depends on the random MLMC correction itself; controlling it requires bounds on the differences between critic outputs at adjacent dyadic horizons. Lemma 3 provides precisely these level-wise estimates. Combining the telescoping identity, the dyadic-difference bounds, and the logarithmic expected horizon established above yields the following MLMC result in Appendix C:

Lemma 4 (MLMC-Critic Bounds) *Under Lemma 3, the estimator* (16) *satisfies*

$$\delta_k^2 \leq C(H_{\max}^{-1} + q_m), \qquad \sigma_k^2 \leq C[1 + \log H_{\max} + H_{\max} q_m], \tag{32}$$

using at most $1 + \log_2 H_{\max}$ *expected critic updates.*

Hence the MLMC construction exchanges the $\mathcal{O}(H_{\max})$ cost of a long vanilla critic for $\mathcal{O}(\log H_{\max})$ expected cost, while preserving its conditional mean and incurring only a controlled increase in the second moment.

Step 4. Matching the Bias and Variance Scales: Finally, choose $H_{\max} \asymp K = T$ and $\alpha = (1-\gamma)/\sqrt{KV_T}$. Lemma 4 gives the long-horizon critic bias at logarithmic expected inner cost, while Lemma 2 converts this conditional critic bias into actor bias and confines the larger MLMC second moment to the α-weighted term in (28). Substitution yields the claimed rate because

$$\sqrt{V_T/T} = \mathcal{O}\left(\sqrt{\frac{1+\log T}{T}} + \eta_m + \gamma^B\right).$$

Thus the statistical term retains the $\widetilde{\mathcal{O}}(T^{-1/2})$ rate, while critic approximation and finite-width errors appear only through the final approximation floor. Details are in Appendix D.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, and verifying the proofs. We have not used generative AI tools to generate experimental results or datasets; experimental implementation and data analysis are not applicable to this theoretical work. We have reviewed all AI-assisted work, independently checked the mathematical arguments, and verified the cited sources and comparisons with prior work. We take responsibility for the accuracy, originality, and integrity of the final manuscript.

Ethics Statement

This work studies the theoretical foundations of reinforcement learning with general utilities. It involves no human-subject studies, personal data, or real-world deployment. We identify no specific ethical concerns arising directly from the research presented. Downstream applications should assess the potential effects of their objectives and learned policies on safety, privacy, and fairness in the intended deployment context.

Reproducibility Statement

This paper presents theoretical results and does not involve empirical experiments or datasets. To facilitate verification, the main text specifies the proposed algorithm, sampling model, assumptions, and parameter choices. Complete proofs and supporting derivations are provided in the appendices, together with explicit formulas for the example applications. These materials provide the mathematical details necessary to reproduce the analysis and assess the scope of the stated guarantees.

References

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

M. Agarwal and V. Aggarwal. Reinforcement learning for joint optimization of multiple rewards. *Journal of Machine Learning Research*, 24(49):1–41, 2023.

M. Agarwal, Q. Bai, and V. Aggarwal. Concave utility reinforcement learning with zero-constraint violations. *Transactions on Machine Learning Research*, 2022.

Q. Bai, M. Agarwal, and V. Aggarwal. Joint optimization of concave scalarized multi-objective reinforcement learning with policy gradient based algorithm. *Journal of Artificial Intelligence Research*, 74:1565–1597, 2022.

A. Barakat, I. Fatkhullin, and N. He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, 2023.

A. Barakat, S. Chakraborty, P. Yu, P. Tokekar, and A. S. Bedi. On the global optimality of policy gradient methods in general utility reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

A. Barakat, S. Chakraborty, P. Yu, P. Tokekar, and A. S. Bedi. On the global optimality of policy gradient methods in general utility reinforcement learning. *Advances in Neural Information Processing Systems*, 38:79601–79633, 2026.

A. Beznosikov, S. Samsonov, M. Sheshukova, A. Gasnikov, A. Naumov, and E. Moulines. First order methods with markovian noise: from acceleration to variational inequalities. *Advances in Neural Information Processing Systems*, 36:44820–44835, 2023.

J. H. Blanchet and P. W. Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *Winter Simulation Conference*, 2015.

D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 4(5):11, 2015.

S. Ganesh and V. Aggarwal. Breaking the bias barrier in concave multi-objective reinforcement learning. *arXiv preprint arXiv:2603.08518*, 2026.

S. Ganesh, J. Chen, W. U. Mondal, and V. Aggarwal. Order-optimal global convergence for actor-critic with general policy and neural critic parametrization. In *Conference on Uncertainty in Artificial Intelligence*, 2025a.

S. Ganesh, W. U. Mondal, and V. Aggarwal. A sharper global convergence analysis for average reward reinforcement learning via an actor-critic approach. In *International Conference on Machine Learning*, 2025b.

M. Geist, J. Pérolat, M. Laurière, R. Elie, S. Perrin, O. Bachem, R. Munos, and O. Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. In *International Conference on Autonomous Agents and Multiagent Systems*, 2022.

E. Hazan, S. M. Kakade, K. Singh, and A. Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.

D. Hendrycks and K. Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

S. M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001.

N. Kumar, K. Wang, U. Pratiush, K. Y. Levy, and S. Mannor. Policy gradient for reinforcement learning with general utilities. In *Tiny Papers at the International Conference on Learning Representations*, 2024.

A. Naskar and V. Aggarwal. Hierarchical multilevel Monte Carlo for order-optimal neural actor-critic in average-reward CMDPs. *arXiv preprint arXiv:2607.28390*, 2026.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

A. Satheesh, P. K. Barman, W. U. Mondal, and V. Aggarwal. Global convergence of average reward constrained MDPs with neural critic and general policy parameterization. In *Conference on Uncertainty in Artificial Intelligence*, 2026.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

T. Zahavy, B. O'Donoghue, G. Desjardins, and S. Singh. Reward is enough for convex MDPs. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

J. Zhang, A. Koppel, A. S. Bedi, C. Szepesvári, and M. Wang. Variational policy gradient method for reinforcement learning with general utilities. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

J. Zhang, C. Ni, C. Szepesvari, M. Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

A Optimality Gap: Proof of Lemma 1

The proof adapts the KL-potential argument for approximate NPG (Agarwal et al., 2021; Ganesh et al., 2025a). We give it in full because the conditional mean, rather than the mean norm of the random critic error, is needed here. Define

$$\Phi_k = \mathbb{E}_{s\sim d_{\theta^*}} \mathrm{KL}(\pi_{\theta^*}(\cdot \mid s) \| \pi_{\theta_k}(\cdot \mid s)), \qquad b_k = \mathbb{E}_{\lambda_{\theta^*}} u_{\theta_k}(s,a).$$

Assumption 6 and the actor update imply, pathwise, 5

$$\begin{aligned}\Phi_k - \Phi_{k+1} &= \mathbb{E}_{\lambda_{\theta^*}}\left[\log \pi_{\theta_{k+1}}(a \mid s) - \log \pi_{\theta_k}(a \mid s)\right] \\ &\geq \alpha b_k^\top w_k - \frac{G_2 \alpha^2}{2} \|w_k\|_2^2. \end{aligned} \tag{33}$$

The finite initial KL and bounded score ensure the displayed difference is integrable whenever the update has a finite first moment; the NPG second-moment bound supplies this property for the algorithm.

By Lemma 5 and transferred compatibility,

$$\begin{aligned} F^* - F(\lambda_{\theta_k}) &\leq \mathbb{E}_{\lambda_{\theta^*}} A_{\theta_k}^{(r_{\theta_k})} \\ &\leq b_k^\top w^*(\theta_k) + \left(\mathbb{E}_{\lambda_{\theta^*}}[A_{\theta_k}^{(r_{\theta_k})} - u_{\theta_k}^\top w^*(\theta_k)]^2\right)^{1/2} \\ &\leq b_k^\top w^*(\theta_k) + \sqrt{\varepsilon_{\mathrm{bias}}}. \end{aligned} \tag{34}$$

Both b_k and $w^*(\theta_k)$ are $\mathcal{F}_k$-measurable, and $\|b_k\|_2 \leq G_1$. Taking conditional expectations in (33), inserting (34), and rearranging gives

$$F^* - F(\lambda_{\theta_k}) \leq \sqrt{\varepsilon_{\mathrm{bias}}} + \frac{\Phi_k - \mathbb{E}_k \Phi_{k+1}}{\alpha} + G_1 \|\mathbb{E}_k w_k - w^*(\theta_k)\|_2 + \frac{G_2 \alpha}{2} \mathbb{E}_k \|w_k\|_2^2.$$

Take expectations and sum from $k = 0$ to $K - 1$. The potentials telescope, and $\Phi_K \geq 0$. Division by K yields (28). This proves Lemma 1. □

B NPG Analysis: Proof of Lemma 2

This is a reset-trajectory adaptation of the conditional-bias approach in Ganesh et al. (2025b). We prove the needed reward-to-gradient bounds and the stochastic solve explicitly. Fix k and condition on $\mathcal{F}_k$ throughout. Write $\theta = \theta_k$, $\lambda = \lambda_\theta$, $r^\circ = r_\theta$, $\bar{r} = \bar{r}_\theta$, and $\widehat{r} = r_{k,\mathrm{MLMC}}$. The random function $\widehat{r}$ is sampled before the NPG trajectories.

Let $A = \mathcal{I}_B(\theta) = \mathbb{E}\widehat{\mathcal{I}}^{k,h}$ and $I_\theta = \mathcal{I}(\theta)$. The missing tail is positive semidefinite, and

$$0 \preceq I_\theta - A \preceq G_1^2 \gamma^B I_d, \qquad \|I_\theta - A\|_{\mathrm{op}} \leq G_1^2 \gamma^B. \tag{35}$$

Consequently, $A \succeq a I_d$ with $a = \mu_F/2$ under the theorem's truncation condition. Every sampled matrix obeys

$$0 \preceq \widehat{\mathcal{I}}^{k,h} \preceq G_1^2 I_d. \tag{36}$$

Both statements follow directly from the bounded score and the normalized discounted weights.

For a deterministic square-integrable reward v, let

$$\mathcal{Q}_B(v) = (1-\gamma)\mathbb{E}_{\pi_\theta} \sum_{t=0}^{B-1} \gamma^t v(s_t, a_t) \sum_{j=0}^{t} u_\theta(s_j, a_j). \tag{37}$$

Put $C_\gamma = G_1\sqrt{1+\gamma}/(1-\gamma)$. The identity

$$(1-\gamma)\sum_{t\geq 0} \gamma^t (t+1)^2 = \frac{1+\gamma}{(1-\gamma)^2}$$

and Cauchy-Schwarz imply

$$\|\mathcal{Q}_B(v)\|_2 \leq C_\gamma \|v\|_{L_2(\lambda)}. \tag{38}$$

Indeed, the norm of the summand's score sum is at most $G_1(t+1)$; apply Cauchy-Schwarz to the finite measure assigning mass $(1-\gamma)\gamma^t$ to time t and then enlarge the nonnegative second sum to all times. The same argument pathwise, followed by expectation, proves the second-moment inequality

$$\mathbb{E}\left\|(1-\gamma)\sum_{t=0}^{B-1}\gamma^t v(s_t,a_t)\sum_{j=0}^{t}u_\theta(s_j,a_j)\right\|_2^2 \le C_\gamma^2\|v\|_{L_2(\lambda)}^2. \tag{39}$$

It assumes no independence between times within a single trajectory.

The true reward is bounded, so the gradient tail satisfies

$$\|\mathcal{Q}_B(r^\circ)-\nabla_\theta J(\theta)\|_2 \le G_1 R_{\max}\gamma^B\left(B+\frac{1}{1-\gamma}\right). \tag{40}$$

This follows from $(1-\gamma)\sum_{t=B}^{\infty}\gamma^t(t+1)=\gamma^B[B+(1-\gamma)^{-1}]$.

Fresh NPG trajectories are independent of $\widehat{r}$. Therefore, the tower property and linearity give $\mathbb{E}_k\widehat{g}^{k,h}=\mathcal{Q}_B(\mathbb{E}_k\widehat{r})$ for each inner iteration after averaging over its past. Decompose

$$\mathbb{E}_k\widehat{r}-r^\circ=(\mathbb{E}_k\widehat{r}-\bar{r})+(\bar{r}-r^\circ).$$

By (38), (40), and Assumption 11,

$$\|\mathcal{Q}_B(\mathbb{E}_k\widehat{r})-\nabla_\theta J(\theta)\|_2 \le G_1 R_{\max}\tau_B + C_\gamma(\delta_k+\sqrt{\varepsilon_{\text{app}}}). \tag{41}$$

For the second moment, (14) gives $\|\bar{r}\|_\infty \le C_r := C_0 + C_\phi R$. Then

$$\mathbb{E}_k\|r_{k,\text{MLMC}}\|_{L_2(\lambda)}^2 \le 2\sigma_k^2+2C_r^2, \quad \mathbb{E}_k\|\hat{g}^{k,h}\|_2^2 \le 2C_\gamma^2(\sigma_k^2+C_r^2). \tag{42}$$

The second inequality also holds conditional on a fixed critic with $\|\widehat{r}\|_{L_2(\lambda)}^2$ in place of the averaged bound.

Condition further on the sampled critic. The hypotheses of Lemma 7 hold with $L=G_1^2$, $a=\mu_F/2$, $q=\mathcal{Q}_B(\widehat{r})$, and $V=C_\gamma^2\|\widehat{r}\|_{L_2(\lambda)}^2$. Hence

$$\mathbb{E}[w_{H_w}^k \mid \mathcal{F}_k,\widehat{r}] = \left[I_d-(I_d-\beta_w A)^{H_w}\right]A^{-1}\mathcal{Q}_B(\widehat{r}).$$

The matrix in this expression is independent of $\widehat{r}$. Averaging over the critic thus replaces $\widehat{r}$ by $\mathbb{E}_k\widehat{r}$ exactly. Let $M=I_d-\beta_w A$, $w^*=I_\theta^{-1}\nabla_\theta J(\theta)$, and $\bar{q}=\mathcal{Q}_B(\mathbb{E}_k\widehat{r})$. Then

$$\mathbb{E}_k w_k - w^* = -M^{H_w}w^* + (I_d-M^{H_w})A^{-1}(\bar{q}-Aw^*). \tag{43}$$

Since $\|M^{H_w}\|_{\text{op}} \le e^{-a\beta_w H_w}$, $\|(I_d-M^{H_w})A^{-1}\|_{\text{op}} \le a^{-1}$, and

$$\|w^*\|_2 \le W_* := \frac{G_1 R_{\max}}{\mu_F(1-\gamma)},$$

we obtain

$$\|\mathbb{E}_k w_k - w^*\|_2 \le e^{-a\beta_w H_w}W_* + a^{-1}\left[G_1R_{\max}\tau_B + C_\gamma(\delta_k+\sqrt{\varepsilon_{\text{app}}}) + G_1^2\gamma^B W_*\right]. \tag{44}$$

Squaring and using $2a=\mu_F$ proves the first line of (30); the extra $\gamma^B/(1-\gamma)$ term is bounded by a constant times τ_B.

Finally, the second-moment part of Lemma 7, averaged over the critic, yields

$$\mathbb{E}_k\|w_k\|_2^2 \le \left(\frac{2}{a^2}+\frac{4\beta_w}{a}\right)C_\gamma^2(2\sigma_k^2+2C_r^2).$$

The inequality $\|w_k-w^*\|^2 \le 2\|w_k\|^2+2W_*^2$ establishes the second line of (30). All constants depend only on the fixed representation and policy bounds. This proves Lemma 2. □

C Dual Critic Analysis: Proofs of Lemmas 3 and 4

Proof of Lemma 3: Fix k and condition on $\mathcal{F}_k$. Write $f_k(\zeta) = \mathcal{E}^{\mathrm{lin}}(\theta_k, \zeta)$ and $\zeta^* = \zeta_k^*$. The linear map J_0 satisfies

$$\|J_0\|_{\mathbb{R}^{D_m} \to L_2(\nu)} = \|J_0^*\|_{L_2(\nu) \to \mathbb{R}^{D_m}} \le \sqrt{\nu_0}\, C_\phi. \tag{45}$$

The composition of the strongly convex functional $-\Psi$ with this affine map is $\mu = \mu_c \kappa$ strongly convex because $\|J_0 v\|_{L_2(\nu)}^2 = v^\top G v \ge \kappa \|v\|_2^2$. The linear occupancy term preserves that curvature. Consequently, f_k has a unique minimizer on the compact convex set Ω.

Let $\mathcal{F}_{k,h}$ contain the critic history before update h. Direct differentiation of the linearized objective and the oracle mean gives

$$\begin{aligned} e_h &:= \mathbb{E}[\widehat{v}^{k,h} \mid \mathcal{F}_{k,h}] - \nabla f_k(\zeta_h^k) \\ &= -(1-\gamma) \sum_{t=B}^{\infty} \gamma^t \mathbb{E}_{\pi_{\theta_k}} g_0(s_t, a_t) + J_0^* \big[\nabla \Psi(r_{\mathrm{lin}}(\cdot, \zeta_h^k)) - \nabla \Psi(r(\cdot, \zeta_h^k)) \big]. \end{aligned} \tag{46}$$

Using (45), dual smoothness on the set containing both arguments, and (14),

$$\|e_h\|_2 \le C_\phi \gamma^B + \nu_0 C_\phi L_c \eta_m =: d_v. \tag{47}$$

The reference estimator contributes no term to this bias because its conditional mean is the required dual derivative. The trajectory part of (18) has norm at most C_ϕ, so

$$\mathbb{E}[\|\widehat{v}^{k,h}\|_2^2 \mid \mathcal{F}_{k,h}] \le 2C_\phi^2 + 2S_\Psi^2 =: V_c. \tag{48}$$

These bounds make the dependence on the dual oracle explicit.

The variational inequality at the constrained minimizer gives $\langle \nabla f_k(\zeta^*), \zeta - \zeta^* \rangle \ge 0$ for $\zeta \in \Omega$. Strong convexity then implies

$$\langle \nabla f_k(\zeta), \zeta - \zeta^* \rangle \ge \mu \|\zeta - \zeta^*\|_2^2.$$

Projection nonexpansiveness, expansion of the square, and Young's inequality yield

$$\begin{aligned} \mathbb{E}[\|\zeta_{h+1}^k - \zeta^*\|_2^2 \mid \mathcal{F}_{k,h}] &\le \|\zeta_h^k - \zeta^*\|_2^2 - 2\beta_{\zeta,h} \langle \zeta_h^k - \zeta^*, \nabla f_k(\zeta_h^k) + e_h \rangle + \beta_{\zeta,h}^2 V_c \\ &\le (1 - \mu\beta_{\zeta,h}) \|\zeta_h^k - \zeta^*\|_2^2 + \beta_{\zeta,h} \frac{d_v^2}{\mu} + \beta_{\zeta,h}^2 V_c. \end{aligned} \tag{49}$$

This proof works at boundary minimizers and uses only bounded second moments of the stochastic update.

To solve the recursion, set $D_h = \mathbb{E}_k \|\zeta_h^k - \zeta^*\|_2^2$ and $E_0 = d_v^2/\mu^2$. With $\beta_{\zeta,h} = 2/[\mu(h+2)]$,

$$D_{h+1} - E_0 \le \frac{h}{h+2}(D_h - E_0) + \frac{4V_c}{\mu^2 (h+2)^2}.$$

Let $A_c = \max\{R^2, 4V_c/\mu^2\}$. We claim $D_h \le E_0 + A_c/(h+1)$. It holds at $h = 0$ since ζ_0 is the center of Ω. If it holds at h, then

$$D_{h+1} - E_0 \le \frac{A_c h}{(h+2)(h+1)} + \frac{A_c}{(h+2)^2} \le \frac{A_c}{h+2}.$$

This proves

$$\mathbb{E}_k \|\zeta_H^k - \zeta^*\|_2^2 \le \frac{A_c}{H+1} + \frac{2C_\phi^2}{\mu^2} \gamma^{2B} + \frac{2\nu_0^2 C_\phi^2 L_c^2}{\mu^2} \eta_m^2. \tag{50}$$

In particular, the constants include $(\mu_c \kappa)^{-2}$.

For every $x = (s, a)$,

$$|r(x, \zeta_H^k) - r_{\mathrm{lin}}(x, \zeta^*)| \le \eta_m + C_\phi \|\zeta_H^k - \zeta^*\|_2.$$

Square, integrate against the probability measure λ_{θ_k}, and use (50). This proves both inequalities in (31). □

Proof of Lemma 4 The randomized-level construction follows the MLMC telescoping principle of Beznosikov et al. (2023); we verify both its output law and cost for the coupled critic runs used here. Write $H_{\max} = 2^J$, $R_H = R_{k,H}$, and $r_*^{\mathrm{lin}} = \bar{r}_{\theta_k}$. The common stepsize schedule ensures that, conditional on a chosen level, the prefix of length H has the same law as a standalone run of length H.

Since $\mathbb{P}(N_k = j) = 2^{-j}$, we have

$$\mathbb{E}_k r_{k,\mathrm{MLMC}} = \mathbb{E}_k R_1 + \sum_{j=1}^{J} 2^{-j} 2^j \mathbb{E}_k (R_{2^j} - R_{2^{j-1}}) = \mathbb{E}_k R_{H_{\max}}. \tag{51}$$

All sums are finite. Jensen's inequality in $L_2(\lambda_{\theta_k})$ and Lemma 3 give

$$\delta_k^2 \le \mathbb{E}_k \| R_{H_{\max}} - r_*^{\mathrm{lin}} \|_{L_2(\lambda_{\theta_k})}^2 \le C(H_{\max}^{-1} + q_m).$$

Let $\Delta_j = R_{2^j} - R_{2^{j-1}}$. No additional cancellation assumption on coupled runs is needed for the following bound: the two marginals and the triangle-square inequality give

$$\mathbb{E}_k \|\Delta_j\|_{L_2(\lambda_{\theta_k})}^2 \le 2\mathbb{E}_k \|R_{2^j} - r_*^{\mathrm{lin}}\|^2 + 2\mathbb{E}_k \|R_{2^{j-1}} - r_*^{\mathrm{lin}}\|^2 \le C(2^{-j} + q_m). \tag{52}$$

Each norm on the middle line is in $L_2(\lambda_{\theta_k})$. Therefore

$$\begin{aligned} \sigma_k^2 &\le 2\mathbb{E}_k \|R_1 - r_*^{\mathrm{lin}}\|^2 + 2\sum_{j=1}^{J} 2^{-j} 4^j \mathbb{E}_k \|\Delta_j\|^2 \\ &\le C(1 + q_m) + C\sum_{j=1}^{J} (1 + 2^j q_m) \le C[1 + J + H_{\max} q_m]. \end{aligned} \tag{53}$$

This accounts for both the base estimator and the statistical contribution at each level.

Finally, the number of updates is 2^{N_k} on $\{N_k \le J\}$ and one otherwise. Its expectation is

$$\sum_{j=1}^{J} 2^{-j} 2^j + \mathbb{P}(N_k > J) = J + 2^{-J} \le J + 1.$$

Shared prefixes produce all three needed outputs within this budget. Each update uses one length-B trajectory and one dual-oracle call. This proves the Lemma 4. □

D Proof of Theorem 1

The proof combines the previous results while keeping the fixed finite-width error explicit. Let $T \ge 2$ and use (26). The choice of B implies

$$\gamma^B \le T^{-2}, \qquad G_1^2 \gamma^B \le \mu_F/2.$$

Also $T \le H_{\max} < 2T$ and $e^{-\mu_F \beta_w H_w} \le T^{-4}$. By Lemma 4,

$$\delta_k^2 \le C(T^{-1} + \eta_m^2), \qquad 1 + \sigma_k^2 \le C V_T, \quad V_T = 1 + \log(2H_{\max}) + H_{\max}(\eta_m^2 + \gamma^{2B}). \tag{54}$$

The constants are uniform over k by the assumptions.

The tail τ_B is harmless at this horizon. To see this with discount dependence displayed, use $\log(1/\gamma) \ge 1 - \gamma$ and the definition of B to obtain

$$B \le 1 + \frac{2\log T + \log(2G_1^2/\mu_F)}{1 - \gamma}.$$

Since $\mu_F \le G_1^2$, the logarithm on the right is positive. Hence $\tau_B \le C(1 + \log T)T^{-2}/(1-\gamma)$, with a fixed constant C. Lemma 2 now gives

$$\begin{aligned} \|\mathbb{E}_k w_k - w^*(\theta_k)\|_2 &\le \frac{C}{1-\gamma}\big(\sqrt{\varepsilon_{\mathrm{app}}} + \eta_m + T^{-1/2}\big), \\ \mathbb{E}_k \|w_k\|_2^2 &\le \frac{C V_T}{(1-\gamma)^2}. \end{aligned} \tag{55}$$

The small $T^{-2}\log T$ term is absorbed into $T^{-1/2}$.

Insert these bounds into Lemma 1:

$$\mathcal{G}_K \le \sqrt{\varepsilon_{\text{bias}}} + \frac{\Phi_0}{\alpha K} + \frac{C}{1-\gamma}(\sqrt{\varepsilon_{\text{app}}} + \eta_m + T^{-1/2}) + \frac{C\alpha V_T}{(1-\gamma)^2}.$$

For $\alpha = (1-\gamma)/\sqrt{KV_T}$ and $K = T$, the potential and second-moment terms are bounded by

$$\frac{C}{1-\gamma}\sqrt{\frac{V_T}{T}} \le \frac{C}{1-\gamma}\left(\sqrt{\frac{1+\log T}{T}} + \eta_m + T^{-2}\right).$$

This proves (27). The stepsize adjustment is what allows a fixed representation: a nonzero η_m appears in the approximation floor rather than as an increasing term from the critic second moment.

The NPG subroutine uses BH_w environment transitions per outer iteration. Lemma 4 bounds expected critic cost by $B(1+\log_2 H_{\max})$ transitions and $1+\log_2 H_{\max}$ dual-oracle calls. Thus

$$\mathbb{E}N_{\text{env}} \le KB(H_w + 1 + \log_2 H_{\max}), \qquad \mathbb{E}N_{\text{ref}} \le c_\Psi K(1+\log_2 H_{\max}), \tag{56}$$

where c_Ψ is the reference-draw count per oracle call. For fixed discount and regularity constants, $B, H_w = \mathcal{O}(\log T)$, and the total is $\widetilde{\mathcal{O}}(T)$. Generic duals can instead report the number of oracle calls in (56). With fixed m, R, ζ_0 and their conditioning constants, taking $T = \widetilde{\mathcal{O}}(\epsilon^{-2})$ makes the statistical term at most ϵ above the floor. A uniformly random index from $\{0, \dots, K-1\}$ therefore returns a policy with this expected gap. This proves Theorem 1. □

E Separable Utility Duals and Examples

This appendix develops the separable-utility specialization introduced following Assumption 3. We first derive the general Fenchel-dual representation and associated pseudo-reward, and then specialize the construction to entropy, smoothed entropy, KL imitation, and quadratic matching. For each example, we give the explicit dual functional, pseudo-reward, unbiased dual-gradient estimator, and the conditions ensuring bounded rewards and the dual regularity required by Assumption 9. The derivations use the Fenchel-Young identity (Rockafellar, 1997), as in variational RLGU (Zhang et al., 2020) and variational divergence estimation (Nowozin et al., 2016).

Proposition 1 (Separable dual construction) *Under Assumption 3, the variational dual and the corresponding pseudo-reward admit the explicit forms*

$$\Psi(r) = -C - \int f^*(-r-c)\,d\nu, \qquad r_\theta = -f'(p_\theta) - c, \tag{57}$$

where $f^*(y) = \sup_{u\in\mathbb{R}}\{uy - f(u)\}$. Furthermore, whenever differentiation under the reference integral is justified,

$$J_0^* \nabla_r \Psi(r) = \int (f^*)'(-r-c) g_0\,d\nu. \tag{58}$$

Proof. Fenchel's inequality at $p = p_\theta(x)$ gives $f(p) + f^*(-r-c) \ge (-r-c)p$. Substituting (57), rearranging, and integrating yields

$$\mathbb{E}_{\lambda_\theta} r - \Psi(r) = C + \int \{pr + f^*(-r-c)\}\,d\nu \ge C - \int f(p)\,d\nu - \int cp\,d\nu = F(\lambda_\theta).$$

Equality holds at $-r-c = f'(p)$ by Fenchel-Young equality. Hence $r_\theta = -f'(p_\theta) - c$ attains the minimum. The negative integral of a convex function is concave. Differentiating it gives (58). □

We next specialize Proposition 1 to the utilities summarized in Table 2. The table records the resulting dual functionals, exact pseudo-rewards, and boundedness conditions; below, we derive these expressions, construct unbiased estimators of the dual-gradient term in (58), and verify the curvature and moment conditions required by Assumption 9.

For entropy, $f(u) = u\log u$ on $u \ge 0$ (with $f(0) = 0$) and $f(u) = +\infty$ for $u < 0$. Maximizing $uy - u\log u$ gives $u = e^{y-1}$ and $f^*(y) = e^{y-1}$. Thus $\Psi(r) = -\int e^{-r-1}d\nu$ and $r_\theta = -1 - \log p_\theta$.

Positive lower and finite upper density bounds make this reward bounded. If $X \sim \nu/\nu_0$, an unbiased dual-gradient sample is

$$Z_\Psi(\zeta) = \nu_0 e^{-r(X,\zeta)-1} g_0(X).$$

The bounded critic range ensures that its second moment is finite. The second variation of $-\Psi$ is multiplication by e^{-r-1}, which is bounded above and below by positive constants on a bounded reward range.

Table 2: Explicit duals. Each row uses its stated reference measure; KL imitation uses $\nu = Q$. In the smoothed row, $f_\tau(u) = u\log(u+\tau)$ for $u \geq 0$ and $+\infty$ otherwise. The rightmost column ensures bounded pseudo-rewards; dual stability imposes the additional range conditions explained after the table. Full conjugate and gradient calculations appear in Proposition 1.

Utility	Dual functional $\Psi(r)$	Exact pseudo-reward	Boundedness conditions
Entropy $-\int p\log p\, d\nu$	$-\int e^{-r-1}\, d\nu$	$-1-\log p$	$0 < a \leq p \leq b < \infty$
Smoothed entropy $-\int f_\tau(p)\, d\nu$	$-\int f_\tau^*(-r)\, d\nu$	$-\log(p+\tau) - p/(p+\tau)$	$\tau > 0$, $0 \leq p \leq b$
KL imitation $-\mathrm{KL}(\lambda\|Q)$	$-\mathbb{E}_Q e^{-r-1}$	$-1-\log(d\lambda/dQ)$	$0 < a \leq d\lambda/dQ \leq b$
Quadratic matching $-\frac{1}{2}\int (p-q)^2 d\nu$	$\mathbb{E}_Q r - \frac{1}{2}\int r^2 d\nu$	$q-p$	p, q bounded; $dQ = q\, d\nu$

KL imitation applies the same calculation with reference measure Q and density $d\lambda_\theta/dQ$. It gives $-\mathrm{KL}(\lambda_\theta\|Q)$, the stated pseudo-reward, and the estimator $e^{-r(X,\zeta)-1}g_0(X)$ for $X \sim Q$.

For smoothed entropy, $f_\tau(u) = u\log(u+\tau)$ on $[0,\infty)$ and $+\infty$ elsewhere. Its right derivative at zero is $\log\tau$, and

$$f_\tau'(u) = \log(u+\tau) + \frac{u}{u+\tau}, \qquad f_\tau''(u) = \frac{1}{u+\tau} + \frac{\tau}{(u+\tau)^2} > 0.$$

The conjugate maximizer is zero when $y \leq \log\tau$, and otherwise is the unique positive solution of $f_\tau'(u) = y$. Hence $(f_\tau^*)'(y)$ equals that maximizer and

$$Z_\Psi(\zeta) = \nu_0 (f_\tau^*)'(-r(X,\zeta)) g_0(X), \qquad X \sim \nu/\nu_0,$$

is unbiased. On compact reward ranges its magnitude is bounded. Strong concavity holds when $-r$ stays in a compact interval strictly above $\log\tau$, where the conjugate curvature equals $1/f_\tau''((f_\tau^*)'(-r))$ and has positive finite bounds.

For quadratic matching, take $f(u) = u^2/2$ on all of $\mathbb{R}$, $c = -q$, and $C = -\frac{1}{2}\int q^2 d\nu$. Since $f^*(y) = y^2/2$, (57) gives

$$\Psi(r) = \frac{1}{2}\int q^2 d\nu - \frac{1}{2}\int (q-r)^2 d\nu = \mathbb{E}_Q r - \frac{1}{2}\int r^2 d\nu, \qquad r_\theta = q - p_\theta.$$

The gradient is $\mathbb{E}_Q g_0 - \int r g_0 d\nu$, and an unbiased estimator is

$$Z_\Psi(\zeta) = g_0(Y) - \nu_0 r(X,\zeta) g_0(X), \qquad Y \sim Q, \quad X \sim \nu/\nu_0,$$

with independent draws. Its second moment is bounded by the local critic and feature bounds. The dual is one strongly concave and one smooth in $L_2(\nu)$. The unrestricted quadratic extension is what yields this quadratic conjugate; a nonnegative-domain restriction would instead give a positive-part conjugate.

F Connection to Concave Utility RL

Proposition 2 *For finite $\mathcal{S}$ and $\mathcal{A}$, every concave occupancy utility can be written as a concave utility of $M = |\mathcal{S}||\mathcal{A}|$ normalized discounted returns. Conversely, any concave utility of finitely many expected returns is a concave occupancy utility in finite or infinite spaces. There is a continuous-state, two-action MDP and a concave occupancy utility on it that cannot be represented by any fixed finite collection of bounded-reward returns.*

Proof. Index the finite state-action pairs by $x_1, \dots, x_M$, and choose $r_i(x) = \mathbf{1}_{\{x=x_i\}}$. Their normalized returns satisfy

$$J_i^\pi = (1-\gamma)\mathbb{E}_\pi \sum_{t\geq 0} \gamma^t r_i(s_t, a_t) = \lambda_\pi(x_i).$$

Thus $F(\lambda_\pi) = f(J_1^\pi, \dots, J_M^\pi)$ with $f = F$ on the occupancy vectors. When CURL uses unnormalized returns $\widetilde{J}_i^\pi$, take $f(z) = F((1-\gamma)z)$. Concavity is preserved by this linear scaling. Conversely, fixed rewards $r_1, \dots, r_M$ define the linear map $\lambda \mapsto (\int r_1\, d\lambda, \dots, \int r_M\, d\lambda)$; composing it with a concave scalarization gives a concave occupancy functional. An unnormalized-return convention adds the corresponding factor $(1-\gamma)^{-1}$.

For the infinite-state statement, take $\mathcal{S} = [0,1]$, $\mathcal{A} = \{0,1\}$, ρ uniform, and transitions that draw the next state uniformly, independently of the current state and action. Let ν be Lebesgue measure times counting measure on the actions. Every stationary policy has an occupancy density

$$p_\pi(s,1) = \pi(1 \mid s), \qquad p_\pi(s,0) = 1 - \pi(1 \mid s).$$

Consider the concave quadratic utility

$$F(\lambda_\pi) = -\frac{1}{2}\sum_{a=0}^{1}\int_0^1 [p_\pi(s,a) - 1/2]^2\, ds.$$

Fix any finite collection of bounded rewards $r_1, \dots, r_M$. Choose $M+1$ disjoint intervals of positive length. Their vectors of integrals of $r_i(s,1) - r_i(s,0)$ in $\mathbb{R}^M$ are linearly dependent. Therefore a nonzero step function h, constant on these intervals, satisfies

$$\int_0^1 h(s)[r_i(s,1) - r_i(s,0)]\, ds = 0, \qquad i = 1, \dots, M.$$

Scale h so $\|h\|_\infty \leq 1$. The policies $\pi_0(1 \mid s) = 1/2$ and $\pi_1(1 \mid s) = 1/2 + h(s)/4$ have identical normalized and unnormalized returns for all M rewards. However,

$$F(\lambda_{\pi_0}) = 0, \qquad F(\lambda_{\pi_1}) = -\frac{1}{16}\int_0^1 h(s)^2\, ds < 0.$$

Thus, no scalarization of that finite return vector represents this utility over all stationary policies. Since the finite collection was arbitrary, the statement follows. This is an objective-class separation; the convergence theorem separately imposes its policy and critic assumptions. □

G Technical Lemmas

The following result is adapted from the standard discounted performance-difference lemma used in policy-gradient analysis (Agarwal et al., 2021). Concavity supplies the extension to a utility functional, as in Barakat et al. (2025).

Lemma 5 (Performance Difference Lemma) *For policies $\pi_\theta, \pi_{\theta'}$ satisfying Assumption 1,*

$$F(\lambda_\theta) - F(\lambda_{\theta'}) \leq \mathbb{E}_{\lambda_\theta} A_{\theta'}^{(r_{\theta'})}.$$

Proof. Concavity of F gives us

$$F(\lambda_\theta) - F(\lambda_{\theta'}) \leq \int r_{\theta'}\, d(\lambda_\theta - \lambda_{\theta'}).$$

For any fixed bounded reward r, write $V_{\theta'}^{(r)}(s) = \int Q_{\theta'}^{(r)}(s,a)\pi_{\theta'}(da \mid s)$. The Bellman equation gives

$$A_{\theta'}^{(r)}(s,a) = r(s,a) + \gamma\mathbb{E}[V_{\theta'}^{(r)}(s') \mid s,a] - V_{\theta'}^{(r)}(s).$$

Integrating along a trajectory of π_θ, multiplying by $(1-\gamma)\gamma^t$, and summing makes the value terms telescope. Their tail vanishes since $V_{\theta'}^{(r)}$ is bounded. Thus

$$\mathbb{E}_{\lambda_\theta} A_{\theta'}^{(r)} = (1-\gamma)\big(V^{\pi_\theta}(r) - V^{\pi_{\theta'}}(r)\big) = \int r\, d(\lambda_\theta - \lambda_{\theta'}).$$

Setting $r = r_{\theta'}$ then proves the claim. □

The following deterministic Taylor statement is the local ingredient needed from the NTK viewpoint. It is consistent with the local-linearization approach of Jacot et al. (2018); Ganesh et al. (2025a); Satheesh et al. (2026). Its hypotheses state the required uniform control explicitly.

Lemma 6 (NTK bounds) *Suppose $\|g_0(x)\|_2 \le C_\phi$ and $|r(x,\zeta_0)| \le C_0$ uniformly in x. If*

$$\sup_{x,\zeta\in\Omega} \|\nabla_\zeta^2 r(x,\zeta)\|_{\mathrm{op}} \le C_H/\sqrt{m},$$

then, uniformly over x and $\zeta \in \Omega$,

$$|r - r_{\mathrm{lin}}| \le \frac{C_H R^2}{2\sqrt{m}}, \qquad \|\nabla_\zeta r - g_0\|_2 \le \frac{C_H R}{\sqrt{m}}, \qquad |r(x,\zeta)| \le C_0 + C_\phi R + \frac{C_H R^2}{2\sqrt{m}}.$$

Proof. For $v = \zeta - \zeta_0$, Taylor's formula along the segment inside Ω gives

$$r(x,\zeta) - r(x,\zeta_0) - g_0(x)^\top v = \int_0^1 (1-t) v^\top \nabla_\zeta^2 r(x,\zeta_0 + tv) v \, dt.$$

Its absolute value is at most $C_H\|v\|^2/(2\sqrt{m})$. Applying the fundamental theorem of calculus to the gradient gives the second bound. The final bound follows from the triangle inequality. □

The Hessian condition is a quantitative hypothesis on the chosen architecture, normalization, initialization, and radius. The main theorem can instead use any certified bound η_m in (14). All expectations in that theorem are conditional on this fixed representation; a random-initialization statement additionally requires a probability bound for the event on which these uniform inequalities hold.

The next lemma is the independent-trajectory specialization of the mean-versus-second-moment method used in Ganesh et al. (2025b). We give the complete proof, allowing the sampled matrix and right-hand side within an update to be dependent.

Lemma 7 (Stochastic Linear Recursion) *Let*

$$x_{h+1} = x_h - \beta(\widehat{A}_h x_h - \widehat{b}_h), \qquad x_0 = 0.$$

For the history $\mathcal{H}_h$ before update h, suppose

$$0 \preceq \widehat{A}_h \preceq L I_d, \quad \mathbb{E}[\widehat{A}_h \mid \mathcal{H}_h] = A \succeq a I_d, \quad \mathbb{E}[\widehat{b}_h \mid \mathcal{H}_h] = q, \quad \mathbb{E}[\|\widehat{b}_h\|_2^2 \mid \mathcal{H}_h] \le V.$$

The constants may be conditional on a fixed auxiliary random variable, with A independent of that variable. If $0 < \beta \le \min\{1/L, a/(4L^2)\}$, then

$$\mathbb{E}x_H = [I_d - (I_d - \beta A)^H]A^{-1}q, \qquad \mathbb{E}\|x_H\|_2^2 \le \left(\frac{2}{a^2} + \frac{4\beta}{a}\right) V.$$

Proof. Conditional expectation gives the exact mean recursion $\mathbb{E}x_{h+1} = (I_d - \beta A)\mathbb{E}x_h + \beta q$. Summing its geometric series proves the mean identity. Since $aI_d \preceq A \preceq LI_d$ and $\beta \le 1/L$, its contraction matrix has norm at most $1 - \beta a$.

For the second moment, expansion and the inequality $\|v - z\|^2 \le 2\|v\|^2 + 2\|z\|^2$ give

$$\begin{aligned}
\mathbb{E}[\|x_{h+1}\|_2^2 \mid \mathcal{H}_h] &\le \|x_h\|_2^2 - 2\beta x_h^\top A x_h + 2\beta x_h^\top q + 2\beta^2 L^2 \|x_h\|_2^2 + 2\beta^2 V \\
&\le (1 - \beta a + 2\beta^2 L^2)\|x_h\|_2^2 + (\beta/a)\|q\|_2^2 + 2\beta^2 V \\
&\le (1 - \beta a/2)\|x_h\|_2^2 + (\beta/a + 2\beta^2)V.
\end{aligned}$$

The middle line uses $2x_h^\top q \le a\|x_h\|^2 + \|q\|^2/a$; the final line uses $\|q\|^2 \le V$ and the stepsize condition. The contraction coefficient is nonnegative because $a \le L$ and $\beta \le a/(4L^2)$. Sum the scalar recursion from $x_0 = 0$ and use $\sum_{h\ge 0}(1 - \beta a/2)^h = 2/(\beta a)$. This proves the claim. □

The following elementary consequence of strong convexity explains the target used in Assumption 11. Its proof is included because the dual minimizer and the best L_2 projection coincide only in special cases.

Proposition 3 (Approximation at the dual minimizer) *Fix a policy θ and suppose $p_\theta = d\lambda_\theta/d\nu \le D$ almost everywhere. Let $\mathcal{E}_\theta(r) = \mathbb{E}_{\lambda_\theta} r - \Psi(r)$ be μ_c-strongly convex and L_c-smooth on a convex set containing r_θ and $\mathcal{F}_{\Omega,m}$, with $\nabla\mathcal{E}_\theta(r_\theta) = 0$. Then its minimizer $\bar{r}_\theta$ over $\mathcal{F}_{\Omega,m}$ satisfies*

$$\|\bar{r}_\theta - r_\theta\|^2_{L_2(\lambda_\theta)} \le D\frac{L_c}{\mu_c} \inf_{v\in\mathcal{F}_{\Omega,m}} \|v - r_\theta\|^2_{L_2(\nu)}.$$

Proof. Strong convexity at the unrestricted stationary minimizer and smoothness give, for every v in the critic class,

$$\frac{\mu_c}{2}\|\bar{r}_\theta - r_\theta\|^2_{L_2(\nu)} \le \mathcal{E}_\theta(\bar{r}_\theta) - \mathcal{E}_\theta(r_\theta) \le \mathcal{E}_\theta(v) - \mathcal{E}_\theta(r_\theta) \le \frac{L_c}{2}\|v - r_\theta\|^2_{L_2(\nu)}.$$

The density bound transfers the left-hand norm to $L_2(\lambda_\theta)$. Take the infimum over v. □

H Additional Related Works

Variational duality and policy-gradient analysis. Our analysis builds on the general-utility policy-gradient identity and concave-dual representation of Zhang et al. (2020); Section 2.2 fixes the reward convention. Lemma 5 combines the classical performance-difference identity with the supporting hyperplane inequality for concave utilities. Appendix A adapts approximate NPG analysis with transferred errors (Agarwal et al., 2021; Ganesh et al., 2025a), isolating the conditional actor bias needed for the multilevel estimator.

Occupancy approximation in large-state RLGU. Barakat et al. (2023) study linear occupancy fitting, while Barakat et al. (2026) use maximum-likelihood occupancy estimation. Their representation assumptions concern distributions; ours concern the population minimizer of a dual-reward problem, with sufficient approximation conditions given in Proposition 3. The improved sampling exponent in Table 1 depends on dual stability, reward approximation, and the stated sampling access; it does not imply that either representation class uniformly dominates the other.

Finite-objective CURL. Agarwal & Aggarwal (2023); Agarwal et al. (2022); Bai et al. (2022) develop finite-objective CURL through model-based and policy-gradient analyses. Since CURL also denotes occupancy-functional objectives (Geist et al., 2022), our comparison specifically concerns the fixed-finite-return formulation. Ganesh & Aggarwal (2026) apply MLMC to a finite vector of scalarization derivatives, achieving the ϵ^{-2} sampling exponent. Proposition 2 connects this result to finite RLGU, with objective-count and regularity dependence inherited from the reduction. In infinite spaces, our function-valued formulation accommodates utilities beyond finite moment summaries.

MLMC estimation and neural actor-critic analysis. MLMC is an established debiasing technique (Blanchet & Glynn, 2015). Neural actor-critic analyses use local linearization and bias-sensitive NPG bounds (Ganesh et al., 2025a; Satheesh et al., 2026), while Naskar & Aggarwal (2026) combines multilevel estimation with neural critics for average-reward constrained problems. We specialize these tools to stochastic concave-dual optimization with reset discounted trajectories, explicitly counting dual-reference samples and retaining neural linearization error in the MLMC second moment and actor stepsize. Our fixed-representation guarantee also tracks feature conditioning through (50) and (25).