MARKOVIAN NONCONVEX ADMM FOR REINFORCEMENT LEARNING: BELLMAN-RESOLVENT STABILITY BEYOND SMOOTH BLOCKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify and study a structural mechanism for Markovian nonconvex ADMM in reinforcement learning. Using finite discounted MDPs as a canonical proving ground, we show that the discounted Bellman resolvent $(I - \gamma P_\pi)^{-1}$ can provide the multiplier stability that classical nonconvex ADMM analyses often obtain from a designated smooth block. Starting from this mechanism, we establish convergence under controlled Markov sampling and then under stochastic observations using an empirical Bellman surrogate that jointly represents the random residual and its Jacobian. Markov mixing, initialization drift, observation noise, and decaying bias enter as one operator perturbation, avoiding unbiased product and double sampling requirements. When the perturbations are square summable, the true KKT residual converges almost surely to zero. Under a finite conditional fourth moment condition, a companion iterate satisfies $\mathbb{E}[\widetilde{G}_{K+1}] \le A/T + (B/T)\sum_{k<T} m_k^{-1}$, which becomes $O(T^{-1} + T/N)$ for total Markov sample budget N, giving $O(\epsilon^{-1})$ iteration complexity and $O(\epsilon^{-2})$ sample complexity for squared KKT accuracy ϵ. Beyond stationarity, discounted occupancy coverage yields $J^\star - J(\pi) = O(\sqrt{G})$ for direct tabular policies, so covered exact KKT points are globally optimal, while a statewise quadratic Bellman-improvement condition sharpens the relation to $O(G)$. Finally, nonlinear policy, projected Bellman, and explicit occupancy formulations exhibit the same chain of operator invertibility, dual representation, and multiplier stability. This supports discounted operator invertibility as a reusable structural principle for primal-dual reinforcement learning.

1 INTRODUCTION

The alternating direction method of multipliers is a natural tool for constrained nonconvex optimization because it separates difficult variables while preserving the interaction between primal and dual updates. Its analysis is delicate because the multiplier step increases the augmented Lagrangian. What the convergence proof ultimately needs is a structural estimate that controls $\|\Delta\lambda_{k+1}\|$ by quantities dissipated in the primal steps. Existing nonconvex analyses commonly obtain such multiplier stability from a designated smooth block and a suitable linear operator (Hong et al., 2016; Zeng et al., 2022; Yuan, 2025). This observation shifts the question from smoothness itself to the source of dual stability: when regularity is carried by the constraint geometry, can that geometry provide the mechanism for controlling the multiplier directly? We use the finite discounted MDP as a canonical instantiation in which this mechanism can be exposed and verified end to end, from multiplier stability and Markovian perturbations to complexity bounds at a finite time horizon and policy performance.

Discounted reinforcement learning provides such a constraint. For a policy π, the Bellman equation is $M_\pi V - r_\pi = 0$ with $M_\pi = I - \gamma P_\pi$. Since $0 < \gamma < 1$ and P_π is stochastic, M_π is invertible and $M_\pi^{-1} = \sum_{t\ge 0} \gamma^t P_\pi^t$. The central question is whether this resolvent can replace the classical mechanism based on a smooth block in a nonconvex ADMM proof. The answer is affirmative. The

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.

optimality condition for the value update and the dual update give $\lambda^{k+1} = -\widehat{M}_{k+1}^{-\mathsf{T}}(c+\eta_V^{-1}\Delta V_{k+1})$. Stability of the inverse resolvent then converts changes in the policy and value variables into a bound on the multiplier increment.

The theory is organized in three layers. The first is a structural layer. Bellman resolvent stability closes Lyapunov arguments for deterministic and controlled Markov settings without assuming that ϕ is smooth. The second is a stochastic layer that preserves the structure of the Bellman operator. Here the difficulty is not variance alone: the augmented policy gradient couples a random Bellman Jacobian with a random Bellman residual, so separate unbiasedness does not remove their covariance. We therefore generate the residual, its Jacobians, and the dual update from one empirical Bellman operator, turning Markov dependence, initialization drift, observation noise, and decaying bias into a single operator perturbation. The third is an RL geometry layer. Occupancy measures, advantage functions, and the performance difference identity turn the optimization residual into a global policy guarantee. This layer further characterizes two performance regimes: occupancy coverage yields the $O(\sqrt{G})$ guarantee, while a curvature condition on the Bellman improvement at each state sharpens the conversion to $O(G)$.

Our contributions are as follows. First, we identify the discounted Bellman resolvent as a source of multiplier stability specific to RL and prove the corresponding bounds on dual increments and the Lyapunov function. Second, we show that the stochastic extension must preserve the Bellman operator coupling. Under Markov sampling controlled by resets and uniform exploration, one shared empirical Bellman surrogate jointly generates the residual, its Jacobians, and the dual update, converting dependence between the residual and its Jacobian, observation noise, bias, and Markov dependence into a controlled operator perturbation and yielding convergence of the KKT residual to zero almost surely. Third, assuming a finite conditional fourth moment, we introduce a companion iterate aligned with the policy subproblem and prove $\mathbb{E}[\widetilde{G}_{K+1}] = O(T^{-1}+T/N)$, yielding $O(\varepsilon^{-1})$ outer iterations and $O(\varepsilon^{-2})$ Markov samples for accuracy ε measured by the squared KKT residual. Fourth, for direct tabular policies, occupancy coverage gives $J^\star - J(\pi) = O(\sqrt{G})$ and exact covered KKT points are globally optimal. Statewise quadratic Bellman improvement strengthens this to $O(G)$. Matching constructions characterize the roles of occupancy coverage and curvature of statewise Bellman improvement in the two performance regimes. Fifth, we use formulations based on nonlinear policies, projected Bellman equations, and occupancy measures to test which parts of the mechanism are specific to the tabular Bellman model. These extensions show that the reusable ingredient is the existence of a stable invertible discounted operator yielding an explicit dual representation.

2 Related Work

Nonconvex ADMM. Classical ADMM is well understood in convex settings (Gabay and Mercier, 1976; Boyd et al., 2011; He and Yuan, 2012), while nonconvex analyses require additional structure to control multiplier motion and recover descent (Hong et al., 2016; Li and Pong, 2016; Zeng et al., 2022). Recent approaches weaken objective regularity through restricted strong convexity, smoothing, increasing penalization, or inexact block updates (Barber and Sidky, 2024; Yuan, 2024; 2025; Bai et al., 2025). The multiaffine ADMM theory of (Gao et al., 2020) controls multiplier increments through a fixed linear subblock whose image contains the image of the remaining constraint map. In our Bellman formulation, the value-block Jacobian $M_\pi = I - \gamma P_\pi$ varies with the policy, so this fixed-operator structure does not directly cover the policy–value splitting. We exploit uniform invertibility and inverse-map stability of M_π to control multiplier motion, and preserve the resulting value–dual identity under Markov sampling through a shared empirical Bellman operator.

Stochastic ADMM and Markovian optimization. Recent stochastic ADMM methods mainly study finite-sum or expectation objectives using variance reduction, momentum, Bregman geometry, adaptive batches, or hybrid estimators (Zeng et al., 2024; Liu et al., 2025; Jin et al., 2026; Zeng et al., 2026). Complementary Markovian optimization theory develops finite-time stochastic approximation, unbounded-noise analysis, lower bounds, and single-trajectory variance reduction (Adibi et al., 2024; Haque and Maguluri, 2025; Sun and Wei, 2025; Peng, 2026; Moulos, 2019). Our stochasticity enters the equality-constraint operator itself: the Bellman residual and its Jacobian are generated from the same controlled trajectory. We therefore use one shared empirical Bellman operator across

Table 1: Structural positioning across neighboring theories. Rates are reported in each paper's native stationarity or performance metric.

Ref.	Method	Data	Structural assumption	Stability / analysis device	Guarantee	RL meaning
Hong et al. (2016)	ADMM	Det.	Nonconvex consensus / sharing; large penalty	Consensus / sharing structure + AL descent	Stationary-set convergence	–
Zeng et al. (2022)	MEAL	Det.	Weakly convex nonsmooth, linear constraints	Moreau envelope of AL	$o(\varepsilon^{-2})$ FOSP complexity	–
Barber and Sidky (2024)	ADMM	Det.	Nonsmooth nonconvex under RSC	Restricted strong convexity	Convergence without differentiability	–
Yuan (2025)	Prox-ADMM	Det.	Multi-block; one continuous block; BI / SU operator	Increasing penalty + decreasing smoothing	$O(\varepsilon^{-3})$ critical-point complexity	–
Zeng et al. (2024)	Stoch. ADMM	Finite sum	Nonconvex nonsmooth linear constraint	SVRG + momentum + potential	$O(1/T)$ stationarity; linear under KL	–
Sun and Wei (2025)	1st-order opt.	Markov	Smooth nonconvex gradient oracle	Markov-oracle lower-bound construction	$\Omega(\varepsilon^{-4})$ countable; $\Omega(\varepsilon^{-2})$ finite chain	RL-relevant oracle
Gaur et al. (2024)	Actor–critic	Markov	Neural actor / critic; weak gradient domination	Critic-error control + gradient domination	$\widetilde{O}(\varepsilon^{-3})$ global sample complexity	Global last iterate
Ours	**Bellman-ADMM**	**Ctrl. Markov**	**Proper l.s.c. ϕ; exact policy subproblem; invertible discounted RL operator**	**Resolvent + shared empirical operator + companion iterate**	$\mathbb{E}\widetilde{G} = O(T^{-1} + T/N)$; $O(\sqrt{G})$; $N = O(\varepsilon^{-2})$ **for squared KKT**	$O(G)$ **with curvature**

the primal and dual updates, preserving row stochasticity, resolvent invertibility, and the value–dual identity while treating residual–Jacobian dependence as operator perturbation.

Policy optimization and occupancy geometry. Performance-difference and occupancy identities connect local policy information to global return (Kakade and Langford, 2002; Agarwal et al., 2021). Recent actor–critic and policy-optimization analyses establish global or nonasymptotic guarantees under Markov sampling, neural or general function approximation, average-reward objectives, constraints, and occupancy-induced geometry (Gaur et al., 2024; Wang et al., 2024; Patel et al., 2024; Ganesh et al., 2025a; Chen and Zhao, 2025; Ganesh et al., 2025b; Montenegro et al., 2024; Satheesh et al., 2026; Sherman et al., 2024; 2025; Asad et al., 2025). Occupancy has also been used directly as an optimization representation (Huang and Jiang, 2024; Neu and Okolo, 2025; Barakat et al., 2025). We first establish convergence of the constrained primal–dual system and then use occupancy geometry to convert its KKT residual into global policy performance and the same discounted occupancy structure also yields a second invertible operator in Section 7.

3 Problem Formulation and Algorithm

We develop the mechanism through a finite discounted MDP, where the discounted operator, its inverse, the induced Markov sampling process, and the policy-performance geometry can all be characterized explicitly. This model serves as a concrete proving ground for the operator-stability principle rather than an abstract definition of its scope. Section 7 then tests the same mechanism under nonlinear policies, projected Bellman equations, and an independent occupancy formulation.

In detail, we consider a finite discounted MDP $(\mathcal{S}, \mathcal{A}, P, r, \rho, \gamma)$ with $0 < \gamma < 1$. The direct policy space is $\Pi = \prod_{s\in\mathcal{S}} \Delta(\mathcal{A})$. For $\pi \in \Pi$, define $P_\pi(s, s') = \sum_{a\in\mathcal{A}} \pi(a \mid s)P(s' \mid s, a), \quad r_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s)r(s, a)$, and $M_\pi = I - \gamma P_\pi, \quad R(\pi, V) = M_\pi V - r_\pi$. We study

$$\min_{\pi\in\Pi,\, V\in\mathbb{R}^{|\mathcal{S}|}} \phi(\pi) + c^{\mathsf{T}}V \quad \text{subject to} \quad R(\pi, V) = 0. \tag{1}$$

The function ϕ may be nonsmooth and nonconvex. It is proper, lower semicontinuous, and bounded below on Π, with $\operatorname{dom}\phi \cap \Pi \neq \varnothing$. Define $\psi = \phi + \delta_\Pi$, where δ_Π is the indicator of Π. We use the limiting subdifferential $\partial\psi$ for policy stationarity. Standard policy optimization is recovered with $\phi \equiv 0$ and $c = -\mu$, where μ is a training initial-state distribution.

The augmented Lagrangian is $\mathcal{L}_\beta(\pi, V, \lambda) = \phi(\pi) + c^{\mathsf{T}}V + \langle\lambda, R(\pi, V)\rangle + \frac{\beta}{2}\|R(\pi, V)\|^2$.

Definition 3.1 (Squared KKT residual)**.** *For $z = (\pi, V, \lambda)$, let*

$$G(z) = \operatorname{dist}^2\left(0, \partial\psi(\pi) + J_\pi(\pi, V)^{\mathsf{T}}\lambda\right) + \|c + M_\pi^{\mathsf{T}}\lambda\|^2 + \|R(\pi, V)\|^2, \tag{2}$$

where $J_\pi(\pi, V) = D_\pi R(\pi, V)$. For $\phi \equiv 0$, $\partial\psi = N_\Pi$.

Algorithm 1 Markovian Proximal Bellman-ADMM

1: Initialize (π^0, V^0, λ^0) and choose $\beta, \eta_\pi, \eta_V > 0$
2: **for** $k = 0, 1, \ldots, T-1$ **do**
3: Freeze $\bar{\pi}^k$ and collect a reset-controlled batch of length m_k
4: Construct $(\widehat{P}_k, \widehat{r}_k)$ from non-reset transitions
5: $\pi^{k+1} \in \arg\min_{\pi \in \Pi} \left\{ \widehat{\mathcal{L}}_{\beta,k}(\pi, V^k, \lambda^k) + \frac{1}{2\eta_\pi} \|\pi - \pi^k\|^2 \right\}$
6: $V^{k+1} = \arg\min_{V} \left\{ \widehat{\mathcal{L}}_{\beta,k}(\pi^{k+1}, V, \lambda^k) + \frac{1}{2\eta_V} \|V - V^k\|^2 \right\}$
7: $\lambda^{k+1} = \lambda^k + \beta \widehat{R}_k(\pi^{k+1}, V^{k+1})$
8: **end for**

3.1 Controlled Markov sampling

At iteration k, the algorithm freezes an exploratory behavior policy $\bar{\pi}^k$ and simulates the reset kernel $K_k(s, s') = (1-\gamma)\rho(s') + \gamma P_{\bar{\pi}^k}(s, s')$. A reset indicator records whether a transition came from the environment or from ρ. Only environment transitions are used to estimate P. This distinction prevents the estimator from converging to K_k.

Assumption 3.2 (Coverage and stochastic observations)**.** *There are constants $\rho_{\min} > 0$, $\alpha > 0$, and $R_{\max} < \infty$ such that $\rho(s) \geq \rho_{\min}, \quad \bar{\pi}^k(a \mid s) \geq \alpha, \quad |r(s,a)| \leq R_{\max}$. A reward observation has the form $Y_{k,t} = r(S_{k,t}, A_{k,t}) + \zeta_{k,t}$, where $\mathbb{E}[\zeta_{k,t} \mid \mathcal{G}_{k,t}] = 0, \quad \mathbb{E}[|\zeta_{k,t}|^4 \mid \mathcal{G}_{k,t}] \leq M_4$. Here $\mathcal{G}_{k,t}$ contains all past observations and the current state, action, and reset indicator, before the current reward and next state are drawn. Fresh reset coins, reset states, and action randomization are independent of the preceding history. On non-reset steps, the conditional next-state law is $P(\cdot \mid S_{k,t}, A_{k,t})$. The current reward and next state may be dependent.*

A batch of length m_k produces action-conditional empirical rows $\widehat{P}_k(\cdot \mid s, a)$ and rewards $\widehat{r}_k(s, a)$. Every unvisited transition row is replaced by a fixed probability vector, and every unvisited reward row uses a fixed bounded value. Hence all empirical rows remain stochastic. For each candidate policy, define $\widehat{P}_{\pi,k} = \sum_a \pi(a \mid s)\widehat{P}_k(\cdot \mid s, a), \quad \widehat{r}_{\pi,k} = \sum_a \pi(a \mid s)\widehat{r}_k(s, a), \ \widehat{M}_{\pi,k} = I - \gamma \widehat{P}_{\pi,k}, \quad \widehat{R}_k(\pi, V) = \widehat{M}_{\pi,k} V - \widehat{r}_{\pi,k}$. Row stochasticity gives $\|M_\pi^{-1}\|_\infty \leq \frac{1}{1-\gamma}, \quad \|\widehat{M}_{\pi,k}^{-1}\|_\infty \leq \frac{1}{1-\gamma}$. Finite-dimensional norm equivalence therefore yields a common spectral bound $\bar{\kappa} < \infty$.

3.2 Markovian proximal Bellman-ADMM

Let $\widehat{\mathcal{L}}_{\beta,k}$ denote (3) with R replaced by $\widehat{R}_k$. Algorithm 1 uses the same empirical model in the policy, value, and dual updates.

Assumption 3.3 (Algorithmic margins)**.** *The policy subproblem admits a global minimizer and is solved exactly. The value subproblem is solved exactly. The initial policy is deterministic and belongs to $\operatorname{dom}\psi$. There is $L_M < \infty$ such that $\|M_{\pi'} - M_\pi\| \leq L_M \|\pi' - \pi\|$. The reduced objective $b(\pi) = \phi(\pi) + c^\mathsf{T} M_\pi^{-1} r_\pi$ is bounded below on Π. The parameters satisfy*

$$\beta > \max\left\{ \frac{60\bar{\kappa}^2}{\eta_V}, 30\eta_\pi \bar{\kappa}^4 L_M^2 \|c\|^2 \right\}, \qquad \tau = \frac{1}{8\eta_V}. \tag{3}$$

The exact policy solve and a structured nonsmooth instance are described in Appendix C. The initialization satisfies $\mathbb{E}\|V^0\|^4 + \mathbb{E}\|\lambda^0\|^4 < \infty$.

4 Resolvent Stability and Lyapunov Descent

The value-step optimality condition is $c + \widehat{M}_{k+1}^\mathsf{T} \lambda^k + \beta \widehat{M}_{k+1}^\mathsf{T} \widehat{R}_k(\pi^{k+1}, V^{k+1}) + \eta_V^{-1} \Delta V_{k+1} = 0$, where $\widehat{M}_{k+1} = \widehat{M}_{\pi^{k+1},k}$ and $\Delta V_{k+1} = V^{k+1} - V^k$. Combining this relation with the dual update gives $c + \widehat{M}_{k+1}^\mathsf{T} \lambda^{k+1} + \eta_V^{-1} \Delta V_{k+1} = 0$. Thus $\lambda^{k+1} = -\widehat{M}_{k+1}^{-\mathsf{T}} \left(c + \eta_V^{-1} \Delta V_{k+1} \right)$.

Define the empirical-model error $\varepsilon_k = \sup_{\pi \in \Pi} \left(\|\widehat{P}_{\pi,k} - P_\pi\| + \|\widehat{r}_{\pi,k} - r_\pi\| \right)$.

Lemma 4.1 (Empirical Bellman moments)**.** *There are deterministic constants $C_2, C_4 < \infty$ such that*

$$\mathbb{E}[\varepsilon_k^2 \mid \mathcal{F}_k] \le \frac{C_2}{m_k}, \qquad \mathbb{E}[\varepsilon_k^4 \mid \mathcal{F}_k] \le \frac{C_4}{m_k^2}. \tag{4}$$

Lemma 4.2 (A priori fourth-moment stability)**.** *Under Assumptions 3.2 and 3.3, there are constants $C_V^{(4)}, C_\lambda^{(4)} < \infty$ such that $\sup_{k\ge 0} \mathbb{E}\|V^k\|^4 \le C_V^{(4)}$, $\sup_{k\ge 0} \mathbb{E}\|\lambda^k\|^4 \le C_\lambda^{(4)}$.*

This argument establishes iterate stability directly from the empirical resolvent and the coupled value–dual recursion, providing the independent boundedness needed by the subsequent Lyapunov analysis.

Lemma 4.3 (Bellman-resolvent multiplier control)**.** *For $k \ge 1$, there are deterministic constants $C_\pi, C_V, C_\varepsilon > 0$ such that*

$$\mathbb{E}\|\Delta\lambda_{k+1}\|^2 \le C_\pi \mathbb{E}\|\Delta\pi_{k+1}\|^2 + C_V \mathbb{E}\|\Delta V_{k+1}\|^2 + C_V \mathbb{E}\|\Delta V_k\|^2 + C_\varepsilon \left(\frac{1}{m_k} + \frac{1}{m_{k-1}}\right). \tag{5}$$

One valid choice for the deterministic primal coefficients is $C_\pi = 5\bar{\kappa}^4 L_M^2 \|c\|^2$, $C_V = \frac{5\bar{\kappa}^2}{\eta_V^2}$.

The lemma is the main structural result. It follows from the value-dual identity, inverse-map stability, and the empirical-model moment bounds.

Structure-preserving stochasticization. Sharing an empirical Bellman operator preserves the multiplier identity under transition estimation. For $\widehat{R} = R + e$ and $\widehat{J} = J + G$, separate zero-mean errors do not eliminate the product term $G^\mathsf{T} e$ in the augmented policy gradient. We therefore construct one complete empirical Bellman operator per batch and use it consistently in the residual, Jacobians, and dual update. Lemma 4.4 shows that the resulting stochastic effects enter the true primal–dual system through the single operator error ε_k.

Lemma 4.4 (Structure-preserving surrogate closure)**.** *Let $e_k(\pi, V) = \widehat{R}_k(\pi, V) - R(\pi, V)$. On every path for which $\sum_k \varepsilon_k^2 < \infty$ and the a priori iterate bounds hold, there are finite constants such that four true-problem relations hold simultaneously. First, the primal steps satisfy $\mathcal{L}_\beta(\pi^k, V^k, \lambda^k) - \mathcal{L}_\beta(\pi^{k+1}, V^{k+1}, \lambda^k) \ge a_\pi \|\Delta\pi_{k+1}\|^2 + a_V \|\Delta V_{k+1}\|^2 - C\varepsilon_k^2$. Second, the empirical value-dual identity becomes $c + M_{k+1}^\mathsf{T} \lambda^{k+1} + \eta_V^{-1} \Delta V_{k+1} = r_{k+1}$ with $\|r_{k+1}\| \le C\varepsilon_k$. Third, the multiplier update becomes $\Delta\lambda_{k+1} = \beta R_{k+1} + d_{k+1}$ with $\|d_{k+1}\| \le C\varepsilon_k$. Fourth, the true KKT residual obeys $\mathrm{KKT}_{k+1} \le C(\|\Delta\pi_{k+1}\| + \|\Delta V_{k+1}\| + \|\Delta\lambda_{k+1}\| + \varepsilon_k)$.*

Remark 4.5 (Shared empirical operators control product bias). Separate unbiasedness of a random residual $\widehat{R} = R + e$ and a random Jacobian $\widehat{J} = J + G$ does not imply an unbiased augmented gradient. Indeed, $\widehat{J}^\mathsf{T}(\lambda + \beta\widehat{R}) - J^\mathsf{T}(\lambda + \beta R) = G^\mathsf{T}(\lambda + \beta R) + \beta J^\mathsf{T} e + \beta G^\mathsf{T} e$, and generally $\mathbb{E}[G^\mathsf{T} e] \ne 0$. Coupling $\widehat{R}$ and $\widehat{J}$ through one empirical Bellman surrogate turns $G^\mathsf{T} e$ into a pathwise $O(\varepsilon_k^2)$ perturbation. The stochastic analysis therefore proceeds through operator error without an unbiased-product or double-sampling condition.

Define the corrected potential $\Phi_k = \mathcal{L}_\beta(\pi^k, V^k, \lambda^k) + \tau\|\Delta V_k\|^2$, $\Delta V_0 = 0$.

Lemma 4.6 (Expected Lyapunov descent)**.** *For $k \ge 1$, there are constants $c_\pi, c_V, c_V', C_E > 0$ such that*

$$\mathbb{E}[\Phi_k] - \mathbb{E}[\Phi_{k+1}] \ge c_\pi \mathbb{E}\|\Delta\pi_{k+1}\|^2 + c_V \mathbb{E}\|\Delta V_{k+1}\|^2 + c_V' \mathbb{E}\|\Delta V_k\|^2 - C_E \left(\frac{1}{m_k} + \frac{1}{m_{k-1}}\right). \tag{6}$$

There are $\underline{\Phi} \in \mathbb{R}$ and $C_L > 0$ such that $\mathbb{E}[\Phi_k] \ge \underline{\Phi} - \frac{C_L}{m_{k-1}}$.

The empirical subproblems produce sufficient decrease for $\widehat{\mathcal{L}}_{\beta,k}$. We transfer this decrease to $\mathcal{L}_\beta$ by retaining the full empirical Bellman operator. Conditional Cauchy–Schwarz, Lemma 4.1, and Lemma 4.2 reduce the statistical contribution to $O(m_k^{-1})$.

Theorem 4.7 (Controlled-Markov Bellman-ADMM convergence)**.** *Assume the finite-MDP, reset-access, uniform-exploration, resolvent, and parameter conditions above, with no additional observation noise. If $\sum_k m_k^{-1} < \infty$, then almost surely $\Delta\pi_k \to 0$, $\Delta V_k \to 0$, $\Delta\lambda_k \to 0$, and $R(\pi^k, V^k) \to 0$. Every accumulation point of (π^k, V^k, λ^k) satisfies the KKT conditions of* (1)*.*

Theorem 4.8 (Structure-preserving stochastic KKT convergence)**.** *Assume the model, sampling, and algorithmic conditions above, with the fourth-moment noise condition replaced by conditional variance at most σ_k^2. Allow the average bias in each visited state-action reward row to have magnitude at most b_k. If $\sum_{k=0}^{\infty} \frac{1+\sigma_k^2}{m_k} < \infty, \quad \sum_{k=0}^{\infty} b_k^2 < \infty$, then, almost surely, $\Delta\pi_k \to 0, \quad \Delta V_k \to 0, \quad \Delta\lambda_k \to 0, \quad R(\pi^k, V^k) \to 0$. The true KKT residual converges to zero. Every accumulation point satisfies the KKT conditions of* (1)*.*

5 KKT Convergence and Complexity

The policy subproblem is solved at (π^{k+1}, V^k). We evaluate a companion iterate that preserves this first-order alignment. Define $\widetilde{\lambda}^{k+1} = \lambda^k + \beta \widehat{R}_k(\pi^{k+1}, V^k), \quad \widetilde{z}^{k+1} = (\pi^{k+1}, V^k, \widetilde{\lambda}^{k+1})$, and $\widetilde{G}_{k+1} = G(\widetilde{z}^{k+1})$.

Lemma 5.1 (Companion residual)**.** *For $k \geq 1$, there are deterministic constants $C_G, C'_G > 0$ such that*

$$\mathbb{E}[\widetilde{G}_{k+1}] \leq C_G \mathbb{E}\left[\|\Delta\pi_{k+1}\|^2 + \|\Delta V_{k+1}\|^2 + \|\Delta V_k\|^2\right] + C'_G \left(\frac{1}{m_k} + \frac{1}{m_{k-1}}\right). \tag{7}$$

Theorem 5.2 (Finite-time squared-KKT complexity)**.** *Under Assumptions 3.2 and 3.3, let K be uniform on $\{0, \ldots, T-1\}$ and independent of the algorithmic randomness. There are constants $A, B > 0$ that do not depend on T or the batch schedule such that $\mathbb{E}[\widetilde{G}_{K+1}] \leq \frac{A}{T} + \frac{B}{T} \sum_{k=0}^{T-1} \frac{1}{m_k}$. If $N = \sum_{k=0}^{T-1} m_k$ and $m_k = N/T$, then $\mathbb{E}[\widetilde{G}_{K+1}] \leq \frac{A}{T} + \frac{BT}{N}$. Consequently, $T(\varepsilon) = O(\varepsilon^{-1}), \quad N(\varepsilon) = O(\varepsilon^{-2})$ are sufficient for $\mathbb{E}[\widetilde{G}_{K+1}] \leq \varepsilon$.*

The theorem measures a squared residual. For the unsquared residual $\widetilde{R} = \sqrt{\widetilde{G}}$, Jensen's inequality gives $T = O(\varepsilon^{-2})$ and $N = O(\varepsilon^{-4})$.

Corollary 5.3 (Iteration and Markov sample complexity)**.** *If $(1/T)\sum_{k<T} m_k^{-1} = O(T^{-1})$, then $\mathbb{E}\widetilde{G}_{K+1} = O(T^{-1})$. For fixed T and total budget N, equal batches minimize $\sum_k m_k^{-1}$ and give $\mathbb{E}\widetilde{G}_{K+1} \leq A/T + BT/N$. Thus $T = O(\varepsilon^{-1})$ and $N = O(\varepsilon^{-2})$ suffice for squared residual ε.*

6 From KKT Residual to Policy Performance

We now set $\phi \equiv 0$ and $c = -\mu$. Let $V^\pi = M_\pi^{-1} r_\pi, \quad J_\nu(\pi) = \nu^\mathsf{T} V^\pi, \quad F_\mu(\pi) = -J_\mu(\pi)$. Define the policy-stationarity residual, following the occupancy/gradient-domination viewpoint in (Agarwal et al., 2021; Huang and Jiang, 2024; Sherman et al., 2025; Barakat et al., 2025), $R_{\mathrm{pol}}(\pi) = \mathrm{dist}\left(0, \nabla F_\mu(\pi) + N_\Pi(\pi)\right)$.

Residual-transfer chain. The policy-performance result rests on five explicit intermediate statements. Lemma D.1 gives the Bellman adjoint representation $\nabla F_\mu(\pi) = J_\pi(\pi, V^\pi)^\mathsf{T} \lambda_\mu^\pi$ with $\lambda_\mu^\pi = M_\pi^{-\mathsf{T}} \mu$. Lemma D.2 gives the tabular Jacobian $\partial R_s(\pi, V)/\partial\pi(a \mid s) = -Q_V(s, a)$ and therefore $[J_\pi(\pi, V)^\mathsf{T}\lambda]_{s,a} = -\lambda_s Q_V(s, a)$. Lemma D.3 gives $\|V - V^\pi\| \leq \bar{\kappa} e_R$ from Bellman feasibility. Lemma D.4 gives $\|\lambda - \lambda_\mu^\pi\| \leq \bar{\kappa} e_V$ from value-stationarity error. Lemma D.5 combines them into $\|J_\pi(\pi, V)^\mathsf{T}\lambda - \nabla F_\mu(\pi)\| \leq C_V e_V + C_R e_R + C_{VR} e_V e_R$.

Lemma 6.1 (KKT residual controls policy stationarity)**.** *Let $G = G(\pi, V, \lambda)$. If $G \leq 1$, then $R_{\mathrm{pol}}(\pi) \leq C_{\mathrm{stat}} \sqrt{G}$, where $C_{\mathrm{stat}} < \infty$ depends only on the discounted resolvent and bounded model quantities.*

Let $d_\mu^\pi = (1-\gamma) M_\pi^{-\mathsf{T}} \mu$ be the normalized discounted state occupancy. For an optimal policy $\pi^\star$ under evaluation distribution ν, define $\mathcal{C}_2(\pi^\star, \pi, \nu, \mu) = \left\| \frac{d_\nu^{\pi^\star}}{d_\mu^\pi} \right\|_2$, with $0/0 = 0$ and value $+\infty$ when the numerator is positive and the denominator is zero.

Theorem 6.2 (Occupancy-mismatch performance bound)**.** *If $\mathrm{supp}\, d_\nu^{\pi^\star} \subseteq \mathrm{supp}\, d_\mu^\pi$, then $J_\nu(\pi^\star) - J_\nu(\pi) \leq \sqrt{2}\, \mathcal{C}_2(\pi^\star, \pi, \nu, \mu) R_{\mathrm{pol}}(\pi)$. If $G \leq 1$, then $J_\nu(\pi^\star) - J_\nu(\pi) \leq \sqrt{2}\, \mathcal{C}_2 C_{\mathrm{stat}} \sqrt{G}$. Every covered exact KKT point is globally optimal under ν.*

Corollary 6.3 (Full-support global optimality)**.** *If the training distribution satisfies $\mu(s) \geq \mu_{\min} > 0$, then $J_\mu^\star - J_\mu(\pi) \leq \sqrt{2}R_{\mathrm{pol}}(\pi)/((1-\gamma)\mu_{\min})$. Hence every exact policy-stationary point is globally optimal.*

Corollary 6.4 (KKT global optimality and finite-time policy bound)**.** *Every exact KKT point satisfying the occupancy-cover condition in Theorem 6.2 is globally optimal under ν. If along the random output iterate K the mismatch coefficient is bounded by $\bar{C}_{\mathrm{occ}}$ and $\mathbb{E}G_K \leq A/T + (B/T)\sum_{k<T} m_k^{-1}$, then $\mathbb{E}[J_\nu^\star - J_\nu(\pi_K)] \leq C\sqrt{A/T + (B/T)\sum_{k<T} m_k^{-1}}$.*

Proposition 6.5 (Coverage cannot be removed)**.** *There exists a three-state discounted MDP and a training distribution concentrated at the initial state for which a policy is first-order stationary but strictly globally suboptimal. In the construction, the current policy never visits a downstream state. Changing only the upstream action or only the downstream action is non-improving, while changing both raises return from 1 to 9. Hence occupancy coverage is a first-order identifiability condition.*

Proposition 6.6 (Separation between performance gap and squared policy stationarity)**.** *There is a three-stage chain and policies π_ε such that $J^\star - J(\pi_\varepsilon) = \gamma^2\varepsilon^3$ while $R_{\mathrm{pol}}(\pi_\varepsilon)^2 = (3/2)\gamma^4\varepsilon^4$. Therefore no finite constant C can satisfy $J^\star - J(\pi) \leq CR_{\mathrm{pol}}(\pi)^2$ for all direct tabular policies.*

For state s, define $\bar{r}_s(\pi) = \operatorname{dist}\left(0, -Q^\pi(s,\cdot) + N_{\Delta(\mathcal{A})}(\pi_s)\right)$, $\quad \delta_s^\pi = TV^\pi(s) - V^\pi(s)$.

Assumption 6.7 (Quadratic Bellman improvement)**.** *There is $\mu_B > 0$ such that $\delta_s^\pi \leq \frac{1}{2\mu_B}\bar{r}_s(\pi)^2$ for every relevant state and policy.*

Theorem 6.8 (Quadratic performance conversion)**.** *Suppose Assumption 6.7 holds and $\mathcal{C}_{\mathrm{PL}} = \left\| \frac{d_\nu^{\pi^\star}}{(d_\mu^\pi)^2} \right\|_\infty < \infty$. Then $J_\nu(\pi^\star) - J_\nu(\pi) \leq \frac{1-\gamma}{2\mu_B}\mathcal{C}_{\mathrm{PL}}R_{\mathrm{pol}}(\pi)^2$. If $G \leq 1$, the right-hand side is at most $\frac{1-\gamma}{2\mu_B}\mathcal{C}_{\mathrm{PL}}C_{\mathrm{stat}}^2 G$.*

A uniform statewise action gap provides a concrete sufficient condition. If every suboptimal action satisfies $\max_b Q^\pi(s,b) - Q^\pi(s,a) \geq \Delta > 0$ on the policy set of interest, then Assumption 6.7 holds with $\mu_B = \Delta/4$. Appendix D proves this implication.

Theorem 6.9 (RL-specific stronger conclusion)**.** *Suppose a random companion iterate satisfies the finite-time squared-KKT bound in Theorem 5.2. Under a uniform occupancy-mismatch bound, its expected policy gap is $O(\sqrt{\mathbb{E}G})$. Hence $G \to 0$ implies convergence to globally optimal policy performance and every covered exact KKT point is globally optimal. If the statewise quadratic Bellman-improvement condition and the stronger occupancy ratio bound hold, the expected policy gap is $O(\mathbb{E}G)$.*

7 Extensions to Other RL Formulations

The preceding theory raises a structural question: is multiplier stability specific to the tabular Bellman resolvent, or does it follow from a broader property of discounted RL operators? We show that the same mechanism persists across nonlinear policy parameterizations, projected Bellman equations, and explicit occupancy formulations whenever the induced operator is uniformly invertible and stable. In each case, the proof follows the same chain: *operator invertibility → explicit dual representation → multiplier control → Lyapunov descent*. The stochastic occupancy formulation further shows that this principle survives empirical operator perturbations when the same structure-preserving operator is shared across primal and dual updates.

7.1 Smooth Nonlinear Policies

The operator-stability mechanism does not depend on direct tabular policy coordinates. Let π_θ be smooth, define $R(\theta, V) = M_\theta V - r_\theta$, and suppose $\|M_\theta - M_{\theta'}\| \leq L_M\|\theta - \theta'\|$ and $b_A(\theta) = \phi(\theta) + c^{\mathsf{T}}M_\theta^{-1}r_\theta$ is lower bounded. **E.1** gives $\|M_\theta^{-1}\|_2 \leq \sqrt{|\mathcal{S}|}/(1-\gamma)$. **E.2** gives $\|M_\theta^{-1} - M_{\theta'}^{-1}\| \leq \kappa_M^2 L_M\|\theta - \theta'\|$. **E.3** gives exact parameter-step descent $\mathcal{L}_\beta(\theta^k, V^k, \lambda^k) - \mathcal{L}_\beta(\theta^{k+1}, V^k, \lambda^k) \geq (2\eta_\theta)^{-1}\|\Delta\theta_{k+1}\|^2$. **E.4** gives $\lambda^{k+1} = -M_{k+1}^{-\mathsf{T}}(c + \eta_V^{-1}\Delta V_{k+1})$. **E.5** gives $\|\Delta\lambda_{k+1}\|^2 \leq A_A\|\Delta\theta_{k+1}\|^2 + B_A\|\Delta V_{k+1}\|^2 + B_A\|\Delta V_k\|^2$. **E.6** gives $\Phi_k^A \geq b_A^{\inf} + (\beta/4)\|R_k\|^2 + (\tau - \kappa_M^2/(\beta\eta_V^2))\|\Delta V_k\|^2$. **E.7** shows that the corresponding positive coefficient conditions imply $\Delta\theta_k, \Delta V_k, \Delta\lambda_k, R_k \to 0$ and KKT convergence.

For a computable proximal-gradient parameter update, define the smooth augmented term $h_k(\theta)$. **E.8** gives $\mathcal{L}_\beta(\theta^k, V^k, \lambda^k) - \mathcal{L}_\beta(\theta^{k+1}, V^k, \lambda^k) \geq (1/(2\eta_{\theta,k}) - L_{\theta,k}/2)\|\Delta\theta_{k+1}\|^2$. **E.9** shows that the margin $1/(2\eta_{\theta,k}) - L_{\theta,k}/2 - A_A/\beta \geq \alpha_\theta > 0$ yields Lyapunov descent and $\min_{k\leq T} G_k^A = O(T^{-1})$. The step sizes also satisfy $\eta_{\theta,k} \geq \underline{\eta} > 0$, with uniformly bounded $L_{\theta,k}$, $D_\theta M_\theta$, and $D_\theta r_\theta$. A uniform smoothness bound and the restarted backtracking rule in Appendix E ensure these conditions.

7.2 Projected Bellman Equations

The same mechanism also survives value-function approximation. What is required for primal–dual stability is an invertible projected Bellman operator, rather than an exact state-value representation. Let $V_w = \Xi w$, fix positive diagonal D, and write $A_\theta = \Xi^\mathsf{T} D M_\theta \Xi$, $b_\theta = \Xi^\mathsf{T} D r_\theta$. Assume A_θ is uniformly invertible and Lipschitz, with the transformed regularity and step-size conditions of Theorem F.2. **F.1** gives the projected value-dual identity $\nu^{k+1} = -A_{k+1}^{-\mathsf{T}}(c_w + \eta_w^{-1}\Delta w_{k+1})$. Theorem **F.2** gives projected-KKT convergence and an $O(T^{-1})$ minimum squared residual. If the projected Bellman map contracts with modulus $q < 1$, **F.3** gives $\|\bar{V}^\theta - V^\theta\| \leq \|(I - \Pi_D)V^\theta\|/(1-q)$. Proposition **F.4** shows that an $O(\varepsilon)$ value error can coexist with an $O(\sqrt{\varepsilon})$ parameter-derivative error, so value approximation alone cannot control the policy gradient. **F.5** instead gives $\|D_\theta R(\theta, V^\theta)^\mathsf{T}\lambda^\theta - D_\theta R(\theta, \bar{V}^\theta)^\mathsf{T}\widehat{\lambda}^\theta\| \leq C_V\|V^\theta - \bar{V}^\theta\| + C_\lambda\|\lambda^\theta - \widehat{\lambda}^\theta\|$, while **F.6** gives $\|\lambda^\theta - \widehat{\lambda}^\theta\| \leq C_{\mathrm{dual}} \inf_{z\in\mathrm{range}(D\Xi)} \|\lambda^\theta - z\|$.

7.3 Explicit Occupancy Measures

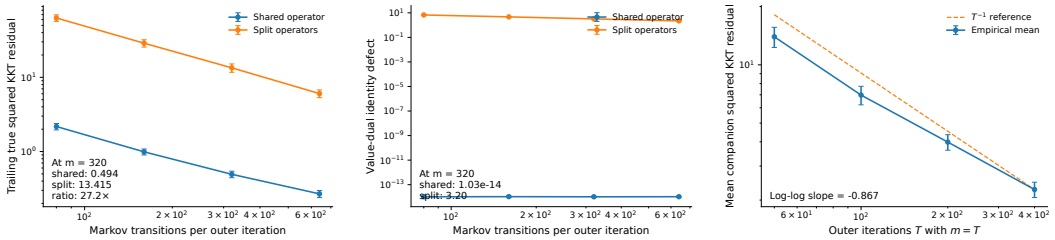
More importantly, the stability principle is not tied to the Bellman value formulation itself. Discounted occupancy flow provides a second RL operator with the same algebraic role. Let $x = (q, d)$, $b = ((1-\gamma)\rho, 0)$, and $K_\theta = \begin{bmatrix} -\gamma P^\mathsf{T} & I \\ I & -\Pi_\theta \end{bmatrix}$. **G.1** proves K_θ invertible and $\|K_\theta^{-1}\|_{1,\oplus} \leq 2/(1-\gamma)$. Proposition **G.2** shows that $K_\theta x = b$ has the unique nonnegative solution $x^\theta = (q^\theta, d^\theta)$. Theorem **G.3** proves KKT equivalence between the constrained occupancy problem and the reduced policy objective. **G.4** gives the occupancy-dual identity $y^{k+1} = -K_{k+1}^{-\mathsf{T}}(c_x + \eta_x^{-1}\Delta x_{k+1})$. **G.5** gives deterministic occupancy-KKT convergence under the regularity, positive step-size lower bound, and strict margins in Theorem G.5. Lemma **G.6** gives the error bound $\|x - x^\theta\| \leq \kappa_K\|K_\theta x - b\|$.

7.4 Stochastic occupancy operator

The stochastic occupancy formulation provides a second test of the structure-preserving principle from Section 4. The occupancy gradient contains $\widehat{K}^\mathsf{T}\widehat{K}$, and sharing the empirical operator preserves the value–dual identity. We therefore construct one empirical transition matrix per iteration and share the induced $\widehat{K}_k$ across the policy, occupancy, and dual steps. **H.1** proves a regenerative estimator with $\mathbb{E}_k\|\widehat{P}_k - P\|_F^2 \leq C_P/m_k$ and $\Pr_k(\|\widehat{P}_k - P\|_F \geq t) \leq 2|\mathcal{S}|^2|\mathcal{A}|e^{-c_P m_k t^2}$. **H.2** shows every empirical occupancy operator is automatically invertible with a deterministic uniform inverse bound, and $\|\widehat{K}_k^{-1} - K_\theta^{-1}\| \leq \bar{\kappa}_K\kappa_K\delta_k$. **H.3** gives $y^{k+1} = -\widehat{K}_k(\theta^{k+1})^{-\mathsf{T}}(c_x + \eta_x^{-1}\Delta x_{k+1})$. **H.4** bounds Δy_{k+1} by $\Delta\theta_{k+1}$, Δx_{k+1}, Δx_k, δ_k, and δ_{k-1}.

H.5 gives the perturbed true Lyapunov lower bound $\Phi_k \geq B_C^{\inf} + (\beta/4)\|F_k\|^2 + q_x\|\Delta x_k\|^2 - C_0\delta_{k-1}^2$. Defining $\Psi_k = \Phi_k - B_C^{\inf} + 1 + C_0\delta_{k-1}^2$, **H.6** proves that the direct policy-gradient perturbation vanishes by block-row orthogonality, and hence bounds the policy-gradient perturbation by $|\langle \widehat{g}_{\theta,k} - g_{\theta,k}, \Delta\theta_{k+1}\rangle| \leq \epsilon\|\Delta\theta_{k+1}\|^2 + C_\epsilon\delta_k^2\Psi_k$. Under the uniform bounds and parameter conditions in Appendix H, **H.7** gives, for $k \geq 1$, the one-step recursion $\mathbb{E}_k\Psi_{k+1} \leq (1+a/m_k)\Psi_k - c_\theta\mathbb{E}_k\|\Delta\theta_{k+1}\|^2 - c_x\mathbb{E}_k\|\Delta x_{k+1}\|^2 - c_x'\|\Delta x_k\|^2 + b/m_k$. **H.8** controls the true squared KKT residual by the increment energy plus $m_k^{-1}\Psi_k + m_k^{-1} + \delta_{k-1}^2$.

H.9 Let $A_T = \sum_{k<T} m_k^{-1}$. Then $\sup_{1\leq k\leq T}\mathbb{E}\Psi_k \leq e^{C_1 A_T}(C_{\mathrm{init}} + C_2 A_T)$ and, for an independent uniform output $K \in \{1, \ldots, T\}$, $\mathbb{E}G_K^{\mathrm{true}} \leq C_3 e^{C_1 A_T}(1 + A_T)(T^{-1} + A_T/T)$. Corollary **H.10** gives $\mathbb{E}G_K^{\mathrm{true}} = O(T^{-1} + T/N)$ for equal batches when $N \gtrsim T^2$. **H.11** If $\sum_k m_k^{-1} < \infty$, then Ψ_k converges, $\sum_k(\|\Delta\theta_{k+1}\|^2 + \|\Delta x_{k+1}\|^2 + \|\Delta x_k\|^2) < \infty$, $\delta_k \to 0$, and $G_k^{\mathrm{true}} \to 0$ almost



(a) KKT residual **(b)** Value–dual identity **(c)** Finite-time scaling

Figure 1: **Numerical validation of the operator-stability mechanism.** **(a)** A shared empirical Bellman operator yields substantially smaller true squared KKT residuals than an equal-budget split-operator construction. **(b)** Sharing the operator preserves the empirical value–dual identity to numerical precision, whereas inconsistent blockwise operators break this identity. **(c)** With $m_k = T$ and $N = T^2$, the companion squared KKT residual exhibits slope -0.87, close to the T^{-1} scaling predicted by Theorem 5.2.

surely. Corollary **H.12** gives the corresponding nonlinear-policy performance conversion under an additional parameterization-specific domination condition.

Unified principle and rate hierarchy. The Bellman resolvent M_π^{-1}, the projected Bellman operator A_θ^{-1}, and the occupancy operator K_θ^{-1} are three realizations of the same stability mechanism. In each case, uniform invertibility converts primal optimality into an explicit dual representation, whose perturbation stability controls multiplier motion. Under stochastic data, sharing one empirical operator across the coupled updates preserves this representation pathwise and converts Markov dependence, observation noise, and bias into operator perturbations. For the Bellman formulation, this yields the squared-KKT rate $\mathbb{E}[\widetilde{G}_{K+1}] = O\big(T^{-1} + T^{-1}\sum_{k<T} m_k^{-1}\big)$, while the policy-performance conversion gives $J^\star - J(\pi) = O(\sqrt{G})$ under occupancy coverage and $O(G)$ under the stronger statewise quadratic improvement condition. The stochastic occupancy formulation exhibits the same operator-stability mechanism with its corresponding sampling rate. Taken together, these results identify discounted-operator invertibility as a reusable structural source of primal–dual stability in reinforcement learning.

8 Numerical Mechanism Validation

We validate the structural predictions on eight independently generated finite discounted MDPs, with five trajectory seeds per instance. Figure 1 summarizes shared-operator stability and finite-time scaling. Additional three-way ablations in Appendix I.5 show that, at 320 steps per empirical model, sharing reduces the trailing true squared KKT residual by a factor of 8.27 relative to separate operators, while using one third of the sampling budget. Sharing between the value and dual updates preserves the value–dual identity even with a separate policy model, identifying the consistency underlying multiplier control. Under $m_k = T$, hence $N = T^2$, the mean companion squared KKT residual exhibits slope -0.87, consistent with the $O(T^{-1})$ dependence in Theorem 5.2. Further statistical controls and sensitivity results appear in Appendix I.

9 Conclusion

We established a Markovian nonconvex ADMM theory in which discounted RL operator structure supplies primal-dual stability. The Bellman resolvent replaces the classical smooth-block mechanism, the structure-preserving empirical Bellman model lifts the argument to controlled Markov and stochastic data, the companion iterate yields finite-time KKT complexity, and occupancy geometry converts stationarity into global policy quality. The nonlinear-policy, projected-Bellman, deterministic occupancy, and stochastic occupancy results show that this mechanism extends beyond the tabular Bellman formulation. Taken together, these results identify uniformly invertible discounted RL operators as a reusable structural source of primal-dual stability.

AI USE STATEMENT

Generative AI tools assisted with language editing and debugging selected parts of the Latex code.

REFERENCES

Arman Adibi, Nicolò Dal Fabbro, Luca Schenato, Sanjeev Kulkarni, H. Vincent Poor, George J. Pappas, Hamed Hassani, and Aritra Mitra. Stochastic approximation with delayed updates: Finite-time rates under markovian sampling. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2746–2754. PMLR, 2024.

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

Reza Asad, Reza Babanezhad Harikandeh, Issam H. Laradji, Nicolas Le Roux, and Sharan Vaswani. Fast convergence of softmax policy mirror ascent. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3943–3951. PMLR, 2025.

Jianchao Bai, Miao Zhang, and Hongchao Zhang. An inexact admm for separable nonconvex and nonsmooth optimization. *Computational Optimization and Applications*, 90:445–479, 2025. doi: 10.1007/s10589-024-00643-y.

Anas Barakat, Souradip Chakraborty, Peihong Yu, Pratap Tokekar, and Amrit Singh Bedi. On the global optimality of policy gradient methods in general utility reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Rina Foygel Barber and Emil Y. Sidky. Convergence for nonconvex admm, with applications to ct imaging. *Journal of Machine Learning Research*, 25(38):1–46, 2024.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

Xuyang Chen and Lin Zhao. On the convergence of continuous single-timescale actor-critic. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 9829–9859. PMLR, 2025.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1): 17–40, 1976.

Swetha Ganesh, Jiayu Chen, Washim Uddin Mondal, and Vaneet Aggarwal. Order-optimal global convergence for actor-critic with general policy and neural critic parametrization. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 1358–1380. PMLR, 2025a.

Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. A sharper global convergence analysis for average reward reinforcement learning via an actor-critic approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18206–18227. PMLR, 2025b.

Wenbo Gao, Donald Goldfarb, and Frank E. Curtis. ADMM for multiaffine constrained optimization. *Optimization Methods and Software*, 35(2):257–303, 2020. doi: 10.1080/10556788.2019.1683553.

Mudit Gaur, Amrit Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: Achieving global convergence (last iterate) of actor-critic under markovian sampling with neural network parametrization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 15153–15179. PMLR, 2024.

Shaan Ul Haque and Siva Theja Maguluri. Stochastic approximation with unbounded markovian noise: A general-purpose theorem. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3718–3726. PMLR, 2025.

Bingsheng He and Xiaoming Yuan. On the o(1/n) convergence rate of the douglas–rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.

Audrey Huang and Nan Jiang. Occupancy-based policy gradient: Estimation, convergence, and optimality. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jiachen Jin, Kangkang Deng, Boyu Wang, and Hongxia Wang. Stochastic admm with batch size adaptation for nonconvex nonsmooth optimization. *Applied Numerical Mathematics*, 222:87–107, 2026. doi: 10.1016/j.apnum.2025.12.007.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, 2002.

Guoyin Li and Ting Kei Pong. Douglas–rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical Programming*, 159:371–401, 2016.

Longhui Liu, Congying Han, Tiande Guo, and Shichen Liao. An inertial stochastic bregman generalized alternating direction method of multipliers for nonconvex and nonsmooth optimization. *Expert Systems with Applications*, 276:126939, 2025. doi: 10.1016/j.eswa.2025.126939.

Alessandro Montenegro, Marco Mussi, Matteo Papini, and Alberto Maria Metelli. Last-iterate global convergence of policy gradients for constrained reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Vrettos Moulos. Optimal best markovian arm identification with fixed confidence. *Advances in Neural Information Processing Systems*, 32, 2019.

Gergely Neu and Nneka Okolo. Offline rl via feature-occupancy gradient ascent. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3637–3645. PMLR, 2025.

Bhrij Patel, Wesley A. Suttle, Alec Koppel, Vaneet Aggarwal, Brian M. Sadler, Dinesh Manocha, and Amrit Bedi. Towards global optimality for practical average reward reinforcement learning without mixing time oracles. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39889–39907. PMLR, 2024.

Zhaojun Peng. Variance-reduced conditional gradient methods under markovian sampling for nonconvex composite optimization, 2026.

Anirudh Satheesh, Pankaj Kumar Barman, Washim Uddin Mondal, and Vaneet Aggarwal. Global convergence of average reward constrained mdps with neural critic and general policy parameterization. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pages 5997–6025. PMLR, 2026.

Uri Sherman, Alon Cohen, Tomer Koren, and Yishay Mansour. Rate-optimal policy optimization for linear markov decision processes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44815–44837. PMLR, 2024.

Uri Sherman, Tomer Koren, and Yishay Mansour. Convergence of policy mirror descent beyond compatible function approximation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 54825–54863. PMLR, 2025.

Zhenyu Sun and Ermin Wei. Improved lower bounds for first-order stochastic non-convex optimization under markov sampling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 57754–57772. PMLR, 2025.

Yudan Wang, Yue Wang, Yi Zhou, and Shaofeng Zou. Non-asymptotic analysis for single-loop (natural) actor-critic with compatible function approximation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 51771–51824. PMLR, 2024.

Ganzhao Yuan. Smoothing proximal gradient methods for nonsmooth sparsity constrained optimization: Optimality conditions and global convergence. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57842–57870. PMLR, 2024.

Ganzhao Yuan. Admm for nonconvex optimization under minimal continuity assumption. In *International Conference on Learning Representations*, 2025.

Jinshan Zeng, Wotao Yin, and Ding-Xuan Zhou. Moreau envelope augmented lagrangian method for nonconvex optimization with linear constraints. *Journal of Scientific Computing*, 91(2):61, 2022.

Yuxuan Zeng, Zhiguo Wang, Jianchao Bai, and Xiaojing Shen. An accelerated stochastic admm for nonconvex and nonsmooth finite-sum optimization. *Automatica*, 163:111554, 2024. doi: 10.1016/j.automatica.2024.111554.

Yuxuan Zeng, Jianchao Bai, Shengjia Wang, Zhiguo Wang, and Xiaojing Shen. A hybrid stochastic alternating direction method of multipliers for nonconvex and nonsmooth composite optimization. *European Journal of Operational Research*, 329(1):63–78, 2026. doi: 10.1016/j.ejor.2025.10.024.

A Proof Map, Conventions, and a Reusable Resolvent Lemma

The appendix follows the same proof organization used in the reference ADMM analysis: the main text states the results and points to an appendix chain, while the appendix proves the lemmas in the order in which they are consumed by the convergence argument. In particular, the logical flow is

$$\text{empirical-model control} \Longrightarrow \text{iterate stability} \Longrightarrow \text{dual control}$$
$$\Longrightarrow \text{Lyapunov descent} \Longrightarrow \text{KKT convergence and complexity} \Longrightarrow \text{policy performance}.$$

The extensions repeat the same architecture after replacing the Bellman operator by a nonlinear-policy Bellman operator, a projected Bellman matrix, or an occupancy operator.

We use $\Delta\pi_{k+1} = \pi^{k+1} - \pi^k$, $\Delta V_{k+1} = V^{k+1} - V^k$, and $\Delta\lambda_{k+1} = \lambda^{k+1} - \lambda^k$. We set $\varepsilon_{-1} = 0$ and $m_{-1}^{-1} = 0$. The initialization satisfies $\pi^0 \in \operatorname{dom}\psi$ and has the moments in Assumption 3.3. Estimates obtained by subtracting consecutive value-dual identities, and their Lyapunov and residual consequences, are stated for $k \geq 1$. The first update contributes a finite initialization constant to finite-time sums. The deterministic extensions use the same convention. All vector norms are Euclidean unless stated otherwise, and matrix norms are induced spectral norms. Constants denoted by C may change from line to line but never depend on k, T, or the batch schedule.

Dependency map. The statistical Lemmas B.1–B.3 prove Lemma 4.1 and supply the stochastic error used by Theorems 4.7 and 4.8. Lemma A.1 and Lemma 4.2 are used in Lemma 4.3. Lemmas 4.4 and 4.3 are the two inputs to Lemma 4.6. Lemmas 4.6 and 5.1 imply Theorem 5.2. The policy-performance chain uses Lemmas D.1–D.6 before Theorems 6.2 and 6.8. Every extension result is used either to prove the next result in its subsection or to justify the operator-stability extension stated in Section 7.

Lemma A.1 (Inverse-map stability)**.** *Let M and M' be invertible with $\|M^{-1}\| \leq \kappa$ and $\|M'^{-1}\| \leq \kappa$. Then*

$$\|M'^{-1} - M^{-1}\| \leq \kappa^2 \|M' - M\|.$$

Consequently, if $\|M_{\pi'} - M_\pi\| \le L_M \|\pi' - \pi\|$*, then*

$$\|M_{\pi'}^{-1} - M_\pi^{-1}\| \le \kappa^2 L_M \|\pi' - \pi\|.$$

Proof. The resolvent identity gives

$$M'^{-1} - M^{-1} = M'^{-1}(M - M')M^{-1}.$$

Taking operator norms yields

$$\|M'^{-1} - M^{-1}\| \le \|M'^{-1}\| \, \|M' - M\| \, \|M^{-1}\| \le \kappa^2 \|M' - M\|.$$

The policy-dependent bound follows by substituting the Lipschitz estimate for M_π. □

Role. Lemma A.1 is the algebraic step that converts a primal policy displacement into a change of the inverse Bellman operator. It is used in Lemma 4.3, in the nonlinear-policy extension, and in the empirical occupancy inverse perturbation bound.

B Controlled-Markov Empirical-Model Bounds

Fix an outer iteration k and condition on $\mathcal{F}_k$. During the batch the behavior policy is frozen, so the reset chain is time homogeneous. Let $B_t = 0$ denote a reset and $B_t = 1$ a genuine environment transition. For a state-action pair (s, a) define

$$Y_t^{s,a} = \mathbf{1}\{S_t = s, A_t = a, B_t = 1\}, \qquad N_{s,a}^{(k)} = \sum_{t=1}^{m_k} Y_t^{s,a}.$$

The quantity $N_{s,a}^{(k)}$ is the denominator of the empirical conditional transition and reward estimators.

Lemma B.1 (Uniform effective-visitation lower tail)**.** *Under Assumption 3.2, there are constants* $p_0, c_0, C_0 > 0$*, independent of* k *and* m_k*, such that*

$$\mathbb{P}\Big(N_{s,a}^{(k)} < p_0 m_k \mid \mathcal{F}_k\Big) \le C_0 e^{-c_0 m_k}$$

for every state-action pair (s, a)*.*

Proof. Split the batch into $J = \lfloor m_k/2 \rfloor$ nonoverlapping two-step blocks and let

$$Q_j = \mathbf{1}\{B_{2j-1} = 0,\ S_{2j} = s,\ A_{2j} = a,\ B_{2j} = 1\}.$$

If $Q_j = 1$, then the second step of the block contributes one usable (s, a) transition. Hence $N_{s,a}^{(k)} \ge \sum_{j=1}^{J} Q_j$. Given the history before block j, the reset draws a fresh state from ρ, the frozen behavior policy chooses action a, and the next reset coin equals one. Therefore

$$\mathbb{E}[Q_j \mid \mathcal{H}_{j-1}] = (1-\gamma)\rho(s)\bar{\pi}^k(a \mid s)\gamma \ge q_\star := \gamma(1-\gamma)\rho_{\min}\alpha > 0.$$

Set $D_j = Q_j - \mathbb{E}[Q_j \mid \mathcal{H}_{j-1}]$. Then (D_j) is a martingale-difference sequence and $|D_j| \le 1$. If $\sum_j Q_j < q_\star J/2$, then $\sum_j D_j \le -q_\star J/2$. Azuma–Hoeffding gives

$$\mathbb{P}\left(\sum_{j=1}^{J} Q_j < \frac{q_\star J}{2} \,\middle|\, \mathcal{F}_k\right) \le \exp\left(-\frac{q_\star^2 J}{8}\right).$$

Since $J \ge m_k/3$ for all sufficiently large m_k, the desired inequality follows with $p_0 = q_\star/6$. The finitely many smaller batch sizes are absorbed by increasing C_0. □

Role. Lemma B.1 prevents the random denominators of all state-action conditional estimators from becoming too small. It is used in Lemma B.2 and the proof of Lemma 4.1.

Lemma B.2 (Inverse moments of random visit counts)**.** *For every fixed $q \geq 1$ there exists $C_q < \infty$ such that*

$$\mathbb{E}\Big[(N_{s,a}^{(k)})^{-q}\mathbf{1}\{N_{s,a}^{(k)} > 0\} \mid \mathcal{F}_k\Big] \leq \frac{C_q}{m_k^q}.$$

Proof. Write $N = N_{s,a}^{(k)}$ and split the expectation over $\{N \geq p_0 m_k\}$ and $\{1 \leq N < p_0 m_k\}$. On the first event, $N^{-q} \leq (p_0 m_k)^{-q}$. On the second event, $N^{-q} \leq 1$ and Lemma B.1 gives probability at most $C_0 e^{-c_0 m_k}$. Therefore

$$\mathbb{E}[N^{-q}\mathbf{1}\{N > 0\} \mid \mathcal{F}_k] \leq (p_0 m_k)^{-q} + C_0 e^{-c_0 m_k}.$$

For fixed q, $e^{-c_0 m} \leq C_q' m^{-q}$ for all integers $m \geq 1$. Combining the two terms proves the claim. □

Role. Lemma B.2 is used to convert martingale moment bounds with a random sample count into deterministic orders m_k^{-1} and m_k^{-2}.

Proof of Lemma 4.1. We prove the transition and reward parts separately.

Part (a): transition rows. For fixed (s, a) define

$$Z_t^{s,a} = Y_t^{s,a}\big(e_{S_{t+1}} - P(\cdot \mid s, a)\big).$$

Because only non-reset transitions are used, $Y_t^{s,a}$ is measurable before S_{t+1} is generated and, on $Y_t^{s,a} = 1$, the next state is drawn from the true environment row. Hence $\mathbb{E}[Z_t^{s,a} \mid \mathcal{H}_t] = 0$. The increments are bounded. The standard Burkholder–Rosenthal fourth-moment inequality for finite-dimensional martingales yields

$$\mathbb{E}\left[\left\|\sum_{t=1}^{m_k} Z_t^{s,a}\right\|^4 \,\middle|\, \mathcal{F}_k\right] \leq C_Z m_k^2.$$

On the good-count event $\mathcal{E}_{s,a}^{(k)} = \{N_{s,a}^{(k)} \geq p_0 m_k\}$,

$$\widehat{P}_k(\cdot \mid s, a) - P(\cdot \mid s, a) = \frac{1}{N_{s,a}^{(k)}} \sum_{t=1}^{m_k} Z_t^{s,a},$$

so

$$\mathbb{E}\Big[\|\widehat{P}_k(\cdot \mid s, a) - P(\cdot \mid s, a)\|^4 \mathbf{1}_{\mathcal{E}_{s,a}^{(k)}} \,\Big|\, \mathcal{F}_k\Big] \leq \frac{C}{m_k^2}.$$

On the bad-count event two probability rows are uniformly bounded in norm, while Lemma B.1 gives an exponentially small probability. Thus the bad-event contribution is also $O(m_k^{-2})$. Consequently

$$\mathbb{E}\Big[\|\widehat{P}_k(\cdot \mid s, a) - P(\cdot \mid s, a)\|^4 \mid \mathcal{F}_k\Big] \leq \frac{C_P}{m_k^2}.$$

Part (b): reward rows. Write $m = m_k$, $I_t = Y_t^{s,a}$, $N = \sum_t I_t$, and $S_m = \sum_{t=1}^m I_t \zeta_{k,t}$. The variables I_t are predictable for the pre-observation filtration $\mathcal{G}_{k,t}$. Thus $I_t \zeta_{k,t}$ is a martingale difference. Work throughout conditionally on $\mathcal{F}_k$. The martingale fourth-moment inequality gives

$$\mathbb{E}_k |S_m|^4 \leq C M_4 m^2.$$

Let $p_\star$ be the visitation constant from Lemma B.1 and set $p = p_\star/4$. On $\{N \geq pm\}$,

$$\mathbb{E}_k\left[|S_m/N|^4 \mathbf{1}\{N \geq pm\}\right] \leq (pm)^{-4} \mathbb{E}_k |S_m|^4 \leq C M_4 m^{-2}.$$

To control small counts without conditioning on the full trajectory, split the batch into two contiguous halves of lengths $\lfloor m/2 \rfloor$ and $\lceil m/2 \rceil$. Let N_1, N_2 be their usable visit counts and let $\mathcal{H}$ contain all observations in the first half. For $m \geq 4$, each half has length at least $m/3$, so Lemma B.1, applied from any starting history, gives

$$\mathbb{P}_k(N_1 < pm) \leq C e^{-cm}, \qquad \mathbb{P}(N_2 < pm \mid \mathcal{H}) \leq C e^{-cm}.$$

On $\mathcal{B} = \{1 \le N < pm\}$, Jensen's inequality gives

$$|S_m/N|^4 \le N^{-1} \sum_t I_t |\zeta_{k,t}|^4 \le \sum_t I_t |\zeta_{k,t}|^4.$$

For a first-half index t, $I_t|\zeta_{k,t}|^4$ is $\mathcal{H}$-measurable and $\mathcal{B} \subset \{N_2 < pm\}$. Hence

$$\mathbb{E}_k[I_t|\zeta_{k,t}|^4 \mathbf{1}_{\mathcal{B}}] \le Ce^{-cm}\mathbb{E}_k[I_t|\zeta_{k,t}|^4] \le CM_4 e^{-cm}.$$

For a second-half index t, the event $\{N_1 < pm\}$ belongs to $\mathcal{G}_{k,t}$. Since $\mathcal{B} \subset \{N_1 < pm\}$, the conditional moment assumption gives the same bound. Summing over t yields

$$\mathbb{E}_k[|S_m/N|^4 \mathbf{1}_{\mathcal{B}}] \le CM_4 m e^{-cm} \le CM_4 m^{-2}.$$

For $N = 0$, the bounded fallback reward contributes Ce^{-cm}. The finitely many $m < 4$ are absorbed in the constant. Therefore $\mathbb{E}_k|\widehat{r}_k(s,a) - r(s,a)|^4 \le C/m_k^2$. This argument allows the reward and next state at a given transition to be dependent.

Part (c): policy-uniform model error. Because the state-action space is finite and P_π, r_π are convex combinations of primitive rows,

$$\varepsilon_k \le C_\Theta \left(\max_{s,a} \|\widehat{P}_k(\cdot \mid s,a) - P(\cdot \mid s,a)\| + \max_{s,a} |\widehat{r}_k(s,a) - r(s,a)| \right).$$

Finite-dimensional norm equivalence and the bounds above give $\mathbb{E}[\varepsilon_k^4 \mid \mathcal{F}_k] \le C_4/m_k^2$. Conditional Jensen then yields $\mathbb{E}[\varepsilon_k^2 \mid \mathcal{F}_k] \le (\mathbb{E}[\varepsilon_k^4 \mid \mathcal{F}_k])^{1/2} \le C_2/m_k$. □

Role. Lemma 4.1 supplies the deterministic $O(m_k^{-1})$ error terms in the expected Lyapunov and finite-time complexity arguments. Its fourth-moment part is exactly what permits Cauchy–Schwarz with the random iterates.

Lemma B.3 (General stochastic empirical-model error)**.** *Suppose an additional observation perturbation is decomposed as $\xi_{k,t} = b_{k,t} + \zeta_{k,t}$ with $\mathbb{E}[\zeta_{k,t} \mid \mathcal{G}_{k,t}] = 0$, $\mathbb{E}[\|\zeta_{k,t}\|^2 \mid \mathcal{G}_{k,t}] \le \sigma_k^2$, and the average bias in every visited state-action reward row bounded in magnitude by b_k. Then there are deterministic constants C_0, C_σ, C_b such that*

$$\mathbb{E}[\varepsilon_k^2 \mid \mathcal{F}_k] \le \frac{C_0 + C_\sigma \sigma_k^2}{m_k} + C_b b_k^2.$$

Consequently,

$$\sum_k \frac{1+\sigma_k^2}{m_k} < \infty, \qquad \sum_k b_k^2 < \infty \quad \Longrightarrow \quad \sum_k \varepsilon_k^2 < \infty \text{ almost surely.}$$

Proof. The transition part contributes C/m_k. For each reward row, use the predictable martingale sum $S_m = \sum_{t=1}^m I_t \zeta_{k,t}$ from Part (b) above. Orthogonality of martingale differences gives $\mathbb{E}_k|S_m|^2 \le m\sigma_k^2$. On $N \ge pm$ the squared average therefore contributes at most $C\sigma_k^2/m$. On $1 \le N < pm$, repeat the two-half argument with exponent two. The first-half terms are controlled by the conditional lower-tail bound for the second half. The second-half terms are controlled by $\mathbb{E}[|\zeta_{k,t}|^2 \mid \mathcal{G}_{k,t}] \le \sigma_k^2$ and the first-half lower-tail event. Their total is at most $C\sigma_k^2 m e^{-cm} \le C\sigma_k^2/m$. The bounded zero-count fallback contributes C/m. Adding the squared row-average bias and summing over the finite row set proves the bound.

Taking expectations and summing under the stated schedule gives $\sum_k \mathbb{E}\varepsilon_k^2 < \infty$. Since the summands are nonnegative, Tonelli's theorem implies $\sum_k \varepsilon_k^2 < \infty$ almost surely. □

Role. Lemma B.3 is the statistical input to Theorem 4.8. The summability condition controls both observation noise and row-average bias.

C Bellman-Resolvent Stability, Lyapunov Descent, and KKT Complexity

C.1 A priori iterate stability

Proof of Lemma 4.2. The empirical value first-order condition and the dual update give the exact identity

$$\lambda^{k+1} = -\widehat{M}_{k+1}^{-\mathsf{T}} \left(c + \eta_V^{-1} \Delta V_{k+1} \right). \tag{8}$$

Hence

$$\|\lambda^{k+1}\| \le \bar{\kappa}\|c\| + \frac{\bar{\kappa}}{\eta_V} \left(\|V^{k+1}\| + \|V^k\| \right). \tag{9}$$

The dual update also gives

$$\widehat{M}_{k+1} V^{k+1} = \widehat{r}_{\pi^{k+1},k} + \beta^{-1} \Delta\lambda_{k+1},$$

so

$$\|V^{k+1}\| \le \bar{\kappa}\|\widehat{r}_{\pi^{k+1},k}\| + \frac{\bar{\kappa}}{\beta} \left(\|\lambda^{k+1}\| + \|\lambda^k\| \right). \tag{10}$$

Lemma 4.1, bounded true rewards, and the conditional fourth-moment assumption imply $R_4 := \sup_k (\mathbb{E}\|\widehat{r}_{\pi^{k+1},k}\|^4)^{1/4} < \infty$. Set

$$X_k = (\mathbb{E}\|V^k\|^4)^{1/4}, \quad Y_k = (\mathbb{E}\|\lambda^k\|^4)^{1/4}, \quad a = \frac{\bar{\kappa}}{\eta_V}, \quad b = \frac{\bar{\kappa}}{\beta}.$$

Minkowski's inequality applied to (9)–(10) gives

$$Y_{k+1} \le \bar{\kappa}\|c\| + a(X_{k+1} + X_k), \qquad X_{k+1} \le \bar{\kappa} R_4 + b(Y_{k+1} + Y_k).$$

Eliminating the next-iterate variables yields

$$\begin{bmatrix} X_{k+1} \\ Y_{k+1} \end{bmatrix} \le \mathsf{A} \begin{bmatrix} X_k \\ Y_k \end{bmatrix} + d, \qquad \mathsf{A} = \frac{1}{1-ab} \begin{bmatrix} ab & b \\ a & ab \end{bmatrix},$$

where d is a fixed nonnegative vector. Let $z = \sqrt{ab} = \bar{\kappa}/\sqrt{\beta\eta_V}$. The largest eigenvalue of A is

$$\rho(\mathsf{A}) = \frac{ab + \sqrt{ab}}{1-ab} = \frac{z}{1-z}.$$

Assumption 3.3 implies $\beta\eta_V > 4\bar{\kappa}^2$, hence $z < 1/2$ and $\rho(\mathsf{A}) < 1$. Iterating the recursion gives

$$\begin{bmatrix} X_k \\ Y_k \end{bmatrix} \le \mathsf{A}^k \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} + \sum_{j=0}^{k-1} \mathsf{A}^j d.$$

The matrix series converges. Thus $\sup_k X_k < \infty$ and $\sup_k Y_k < \infty$, which is the claimed fourth-moment stability. □

Role. This lemma is proved before any Lyapunov decrease. It removes the potential circularity that would arise if bounded multipliers were assumed inside the empirical-to-true perturbation estimate.

C.2 Multiplier control

Proof of Lemma 4.3. Let $M_{k+1} = M_{\pi^{k+1}}$. Replacing $\widehat{M}_{k+1}$ in the value-dual identity by the true matrix gives

$$c + M_{k+1}^{\mathsf{T}} \lambda^{k+1} + \eta_V^{-1} \Delta V_{k+1} = r_{k+1}, \qquad r_{k+1} = (M_{k+1} - \widehat{M}_{k+1})^{\mathsf{T}} \lambda^{k+1}. \tag{11}$$

Since $\|M_{k+1} - \widehat{M}_{k+1}\| \le C\varepsilon_k$, conditional Cauchy–Schwarz, Lemma 4.1, and Lemma 4.2 give

$$\mathbb{E}\|r_{k+1}\|^2 \le C\mathbb{E}[\varepsilon_k^2 \|\lambda^{k+1}\|^2] \tag{12}$$

$$\le C(\mathbb{E}\varepsilon_k^4)^{1/2} (\mathbb{E}\|\lambda^{k+1}\|^4)^{1/2} \le \frac{C_r}{m_k}. \tag{13}$$

Solve (11) for λ^{k+1} and subtract the corresponding identity at iteration k:

$$\begin{aligned}\Delta\lambda_{k+1} = &-(M_{k+1}^{-\mathsf{T}} - M_k^{-\mathsf{T}})c - \eta_V^{-1} M_{k+1}^{-\mathsf{T}} \Delta V_{k+1} + \eta_V^{-1} M_k^{-\mathsf{T}} \Delta V_k \\ &+ M_{k+1}^{-\mathsf{T}} r_{k+1} - M_k^{-\mathsf{T}} r_k. \end{aligned} \tag{14}$$

Lemma A.1 gives $\|(M_{k+1}^{-\mathsf{T}} - M_k^{-\mathsf{T}})c\| \le \bar{\kappa}^2 L_M \|c\| \|\Delta\pi_{k+1}\|$. Using $\|\sum_{i=1}^5 x_i\|^2 \le 5\sum_{i=1}^5 \|x_i\|^2$ in (14),

$$\begin{aligned}\|\Delta\lambda_{k+1}\|^2 \le 5\bar{\kappa}^4 L_M^2 \|c\|^2 \|\Delta\pi_{k+1}\|^2 + \frac{5\bar{\kappa}^2}{\eta_V^2}\left(\|\Delta V_{k+1}\|^2 + \|\Delta V_k\|^2\right) \\ + 5\bar{\kappa}^2(\|r_{k+1}\|^2 + \|r_k\|^2). \end{aligned} \tag{15}$$

Taking expectations and using (13) proves (5) with the stated C_π and C_V and a suitable C_ε. □

Role. Lemma 4.3 is the replacement for the classical smooth-block multiplier estimate. It is the term that allows the positive primal decrease to dominate the dual ascent.

C.3 Structure-preserving surrogate closure

Proof of Lemma 4.4. Fix a sample path for which $\sum_k \varepsilon_k^2 < \infty$ and the a priori iterate bounds hold. The empirical and true residuals satisfy

$$e_k(\pi, V) = \widehat{R}_k(\pi, V) - R(\pi, V) = (\widehat{M}_{\pi,k} - M_\pi)V - (\widehat{r}_{\pi,k} - r_\pi).$$

On the bounded iterate set there is $C_e < \infty$ such that

$$\|e_k(\pi, V)\| + \|D_{\pi,V} e_k(\pi, V)\| \le C_e \varepsilon_k. \tag{16}$$

We verify the four relations in the statement.

Part (a): true primal decrease. Let $\Psi_k = \widehat{\mathcal{L}}_{\beta,k} - \mathcal{L}_\beta$. Expanding,

$$\Psi_k = \langle \lambda, e_k \rangle + \beta \langle R, e_k \rangle + \frac{\beta}{2}\|e_k\|^2.$$

Equation (16) and bounded iterates imply $\|D_{\pi,V}\Psi_k\| \le C\varepsilon_k$ along the two primal line segments. Optimality of the empirical policy subproblem gives

$$\widehat{\mathcal{L}}_{\beta,k}(\pi^k, V^k, \lambda^k) - \widehat{\mathcal{L}}_{\beta,k}(\pi^{k+1}, V^k, \lambda^k) \ge \frac{1}{2\eta_\pi}\|\Delta\pi_{k+1}\|^2.$$

The perturbation difference is at most $C\varepsilon_k\|\Delta\pi_{k+1}\|$, which by Young's inequality is bounded by $(4\eta_\pi)^{-1}\|\Delta\pi_{k+1}\|^2 + C\varepsilon_k^2$. Thus the true policy step decreases by at least $(4\eta_\pi)^{-1}\|\Delta\pi_{k+1}\|^2 - C\varepsilon_k^2$. The same argument for the value step gives $(4\eta_V)^{-1}\|\Delta V_{k+1}\|^2 - C\varepsilon_k^2$. Adding the two proves the first relation with $a_\pi = 1/(4\eta_\pi)$ and $a_V = 1/(4\eta_V)$.

Part (b): perturbed value-dual identity. The exact empirical identity is the value-dual identity. Hence

$$c + M_{k+1}^{\mathsf{T}}\lambda^{k+1} + \eta_V^{-1}\Delta V_{k+1} = (M_{k+1} - \widehat{M}_{k+1})^{\mathsf{T}}\lambda^{k+1} =: r_{k+1}.$$

The pathwise iterate bound and (16) give $\|r_{k+1}\| \le C\varepsilon_k$.

Part (c): inexact true-residual multiplier update. Since $\Delta\lambda_{k+1} = \beta\widehat{R}_k(\pi^{k+1}, V^{k+1})$,

$$\Delta\lambda_{k+1} = \beta R_{k+1} + d_{k+1}, \qquad d_{k+1} = \beta e_k(\pi^{k+1}, V^{k+1}),$$

and $\|d_{k+1}\| \le C\varepsilon_k$ by (16).

Part (d): true KKT residual. The empirical policy Fermat condition is

$$0 \in \partial\psi(\pi^{k+1}) + \widehat{J}_k(\pi^{k+1}, V^k)^{\mathsf{T}}\left[\lambda^k + \beta\widehat{R}_k(\pi^{k+1}, V^k)\right] + \eta_\pi^{-1}\Delta\pi_{k+1}.$$

The bracket differs from λ^{k+1} by $\beta\widehat{M}_{k+1}(V^k - V^{k+1})$, and the true and empirical Jacobians differ by $O(\varepsilon_k)$. The tabular Bellman Jacobian is affine in V, so replacing V^k by V^{k+1} costs $O(\|\Delta V_{k+1}\|)$. Hence the true policy residual is at most $C(\|\Delta\pi_{k+1}\| + \|\Delta V_{k+1}\| + \varepsilon_k)$. The value residual follows from Part (b), and feasibility follows from Part (c). Combining the three components gives the stated true KKT bound. □

Role. This lemma is the interface between stochastic RL data and deterministic primal-dual analysis. It packages all sampling effects into one scalar operator error ε_k.

C.4 Expected Lyapunov descent and lower bound

Proof of Lemma 4.6. We prove the descent and lower bound separately.

Part (a): expected primal decrease. The empirical policy and value subproblems give the same comparison inequalities as above. To justify the moment transfer, define

$$U_k = 1 + \|V^k\| + \|\lambda^k\|, \quad p_k = \max_{s,a} \|\widehat{P}_k(\cdot \mid s,a) - P(\cdot \mid s,a)\|, \quad u_k = \max_{s,a} |\widehat{r}_k(s,a) - r(s,a)|.$$

Then U_k is $\mathcal{F}_k$-measurable, p_k is uniformly bounded, and Lemma 4.1 gives conditional second and fourth moments of orders m_k^{-1} and m_k^{-2} for both p_k and u_k. Expanding $\Psi_k = \widehat{\mathcal{L}}_{\beta,k} - \mathcal{L}_\beta$ on the policy comparison segment, where $V = V^k$ and $\lambda = \lambda^k$, gives

$$\sup_{\pi \in \Pi} \|D_\pi \Psi_k(\pi, V^k, \lambda^k)\| \le C\left(p_k U_k^2 + u_k U_k + u_k^2\right).$$

Indeed, the Jacobian error is at most $C(p_k\|V^k\| + u_k)$ and the residual error has the same bound. Boundedness of p_k absorbs its square. Conditioning before the batch therefore yields

$$\mathbb{E}_k \sup_{\pi \in \Pi} \|D_\pi \Psi_k\|^2 \le \frac{C(1 + U_k^4)}{m_k}.$$

For the value segment, the explicit value solve gives

$$V^{k+1} = (\beta \widehat{M}_{k+1}^\top \widehat{M}_{k+1} + \eta_V^{-1} I)^{-1} \left(\beta \widehat{M}_{k+1}^\top \widehat{r}_{\pi^{k+1},k} - \widehat{M}_{k+1}^\top \lambda^k - c + \eta_V^{-1} V^k\right),$$

so $\|V^{k+1}\| \le C(U_k + u_k)$. On the segment from V^k to V^{k+1},

$$\|D_V \Psi_k\| \le C\{p_k(U_k + u_k) + u_k\}, \qquad \mathbb{E}_k \sup_{V \in [V^k, V^{k+1}]} \|D_V \Psi_k\|^2 \le C(1 + U_k^2)/m_k.$$

Taking expectations and applying Lemma 4.2 controls both comparison-segment errors by C/m_k. Young's inequality therefore yields

$$\mathbb{E}[\mathcal{L}_\beta(\pi^k, V^k, \lambda^k) - \mathcal{L}_\beta(\pi^{k+1}, V^k, \lambda^k)] \ge \frac{1}{4\eta_\pi} \mathbb{E}\|\Delta\pi_{k+1}\|^2 - \frac{C}{m_k}, \tag{17}$$

$$\mathbb{E}[\mathcal{L}_\beta(\pi^{k+1}, V^k, \lambda^k) - \mathcal{L}_\beta(\pi^{k+1}, V^{k+1}, \lambda^k)] \ge \frac{1}{4\eta_V} \mathbb{E}\|\Delta V_{k+1}\|^2 - \frac{C}{m_k}. \tag{18}$$

Let $a_\pi = 1/(4\eta_\pi)$ and $a_V = 1/(4\eta_V)$.

Part (b): dual ascent. By Lemma 4.4, $\Delta\lambda_{k+1} = \beta R_{k+1} + d_{k+1}$. Therefore

$$\mathcal{L}_\beta(\pi^{k+1}, V^{k+1}, \lambda^{k+1}) - \mathcal{L}_\beta(\pi^{k+1}, V^{k+1}, \lambda^k) \tag{19}$$

$$= \langle \Delta\lambda_{k+1}, R_{k+1} \rangle = \frac{1}{\beta}\|\Delta\lambda_{k+1}\|^2 - \frac{1}{\beta}\langle \Delta\lambda_{k+1}, d_{k+1} \rangle \tag{20}$$

$$\le \frac{3}{2\beta}\|\Delta\lambda_{k+1}\|^2 + \frac{1}{2\beta}\|d_{k+1}\|^2. \tag{21}$$

The second moment of d_{k+1} is $O(m_k^{-1})$. Substituting Lemma 4.3 into (21) and combining with (17)–(18) gives

$$\mathbb{E}[\mathcal{L}_\beta^k] - \mathbb{E}[\mathcal{L}_\beta^{k+1}] \ge \left(a_\pi - \frac{3C_\pi}{2\beta}\right) \mathbb{E}\|\Delta\pi_{k+1}\|^2 + (a_V - D_V)\, \mathbb{E}\|\Delta V_{k+1}\|^2 - D_V \mathbb{E}\|\Delta V_k\|^2 - C_E\left(m_k^{-1} + m_{k-1}^{-1}\right),$$

where $D_V = 3C_V/(2\beta)$. Adding $\tau\|\Delta V_k\|^2$ to the potential gives coefficients

$$c_\pi = a_\pi - \frac{3C_\pi}{2\beta}, \quad c_V = a_V - D_V - \tau, \quad c_V' = \tau - D_V.$$

Assumption 3.3 makes all three positive. This proves (6).

Part (c): lower bound. From the true perturbed value-dual identity, $\lambda^k = -M_k^{-\mathsf{T}}(c + \eta_V^{-1}\Delta V_k - r_k)$, while $V^k = M_k^{-1}(r_{\pi^k} + R_k)$. Substituting both into $\mathcal{L}_\beta^k$ gives

$$\mathcal{L}_\beta^k = b(\pi^k) - \eta_V^{-1}\langle \Delta V_k, M_k^{-1}R_k\rangle + \langle r_k, M_k^{-1}R_k\rangle + \frac{\beta}{2}\|R_k\|^2.$$

Two applications of Young's inequality yield

$$\Phi_k \geq b_{\min} + \frac{\beta}{4}\|R_k\|^2 + \left(\tau - \frac{2\bar{\kappa}^2}{\beta\eta_V^2}\right)\|\Delta V_k\|^2 - \frac{2\bar{\kappa}^2}{\beta}\|r_k\|^2.$$

The parameter condition makes the coefficient of $\|\Delta V_k\|^2$ positive. Taking expectations and using (13) gives $\mathbb{E}\Phi_k \geq \underline{\Phi} - C_L m_{k-1}^{-1}$. □

Role. Lemma 4.6 is the summability engine for both the asymptotic and finite-time results.

C.5 Controlled-Markov convergence

Proof of Theorem 4.7. With no additional observation noise, Lemma 4.1 gives $\mathbb{E}\varepsilon_k^2 \leq C/m_k$. Hence $\sum_k m_k^{-1} < \infty$ implies $\sum_k \mathbb{E}\varepsilon_k^2 < \infty$, and Tonelli gives $\sum_k \varepsilon_k^2 < \infty$ almost surely.

Fix a path in this probability-one event. The empirical rewards are pathwise bounded because the true rewards are bounded and $\varepsilon_k \to 0$. The same two-dimensional recursion used in Lemma 4.2, now pathwise, gives bounded V^k and λ^k. Lemma 4.4 and the pathwise version of the Lyapunov calculation then yield

$$\Phi_k - \Phi_{k+1} \geq c_\pi\|\Delta\pi_{k+1}\|^2 + c_V\|\Delta V_{k+1}\|^2 + c_V'\|\Delta V_k\|^2 - C(\varepsilon_k^2 + \varepsilon_{k-1}^2).$$

The lower-bound calculation in Lemma 4.6 remains valid pathwise up to a summable $O(\varepsilon_{k-1}^2)$ term. Summing over k shows

$$\sum_k \|\Delta\pi_{k+1}\|^2 < \infty, \qquad \sum_k \|\Delta V_{k+1}\|^2 < \infty.$$

The pathwise version of Lemma 4.3 and $\sum_k \varepsilon_k^2 < \infty$ give $\sum_k \|\Delta\lambda_{k+1}\|^2 < \infty$. Thus all increments converge to zero. Lemma 4.4, Part (c), gives $R_{k+1} = \beta^{-1}(\Delta\lambda_{k+1} - d_{k+1}) \to 0$. Part (d) then gives KKT-residual convergence.

The policy space is compact and the value-dual iterates are bounded, so accumulation points exist. Fix a convergent subsequence $z^{k_j} \to z^\star$. Vanishing increments imply $z^{k_j-1} \to z^\star$. The descent inequality bounds Φ_k above, so $\phi(\pi^k)$ is bounded above on this path. Lower semicontinuity gives $\phi(\pi^\star) < \infty$.

Write $H_j(\pi)$ for the smooth part of $\widehat{\mathcal{L}}_{\beta,k_j-1}(\pi, V^{k_j-1}, \lambda^{k_j-1})$. Model-error convergence and bounded iterates imply that H_j converges uniformly on Π to a continuous function. Exact policy minimization, compared with $\pi^\star$, gives

$$\phi(\pi^{k_j}) + H_j(\pi^{k_j}) + \frac{\|\pi^{k_j} - \pi^{k_j-1}\|^2}{2\eta_\pi} \leq \phi(\pi^\star) + H_j(\pi^\star) + \frac{\|\pi^\star - \pi^{k_j-1}\|^2}{2\eta_\pi}.$$

Taking the upper limit and using lower semicontinuity proves $\psi(\pi^{k_j}) \to \psi(\pi^\star)$. The policy Fermat condition supplies $w_j \in \partial\psi(\pi^{k_j})$ with $w_j \to -J_\pi(\pi^\star, V^\star)^{\mathsf{T}}\lambda^\star$. The limiting subdifferential is closed along this function-value-convergent sequence. Hence $0 \in \partial\psi(\pi^\star) + J_\pi(\pi^\star, V^\star)^{\mathsf{T}}\lambda^\star$. The other two KKT conditions follow by continuity. This proves the accumulation-point claim without a subdifferential sum rule. □

Role. Theorem 4.7 is the first complete Markovian nonconvex ADMM convergence theorem in the paper. It validates that Bellman-resolvent stability survives controlled Markov sampling.

C.6 COMPANION RESIDUAL

Proof of Lemma 5.1. The policy-subproblem Fermat condition is

$$0 \in \partial\psi(\pi^{k+1}) + \widehat{J}_k(\pi^{k+1}, V^k)^\mathsf{T} \left[\lambda^k + \beta \widehat{R}_k(\pi^{k+1}, V^k)\right] + \eta_\pi^{-1}\Delta\pi_{k+1}. \tag{22}$$

The bracket is $\widetilde{\lambda}^{k+1}$. Hence the true policy residual at $\widetilde{z}^{k+1}$ is bounded by

$$e_{\pi,k+1} \le \eta_\pi^{-1}\|\Delta\pi_{k+1}\| + \|J_\pi(\pi^{k+1}, V^k) - \widehat{J}_k(\pi^{k+1}, V^k)\|\|\widetilde{\lambda}^{k+1}\|.$$

Use the conditional variables U_k, p_k, u_k from the preceding descent proof. The empirical companion multiplier satisfies $\|\widetilde{\lambda}^{k+1}\| \le C(U_k + u_k)$. Consequently

$$\|(J_\pi - \widehat{J}_k)^\mathsf{T}\widetilde{\lambda}^{k+1}\| \le C(p_k U_k^2 + u_k U_k + u_k^2).$$

Conditioning on $\mathcal{F}_k$ and then using the fourth-moment stability yields

$$\mathbb{E}e_{\pi,k+1}^2 \le C\mathbb{E}\|\Delta\pi_{k+1}\|^2 + \frac{C}{m_k}. \tag{23}$$

The companion and final multipliers satisfy

$$\lambda^{k+1} - \widetilde{\lambda}^{k+1} = \beta\widehat{M}_{k+1}\Delta V_{k+1}.$$

Substitution into the empirical value-dual identity gives

$$c + \widehat{M}_{k+1}^\mathsf{T}\widetilde{\lambda}^{k+1} = -\left(\eta_V^{-1}I + \beta\widehat{M}_{k+1}^\mathsf{T}\widehat{M}_{k+1}\right)\Delta V_{k+1}.$$

Replacing $\widehat{M}_{k+1}$ by M_{k+1} and applying the fourth-moment bounds yields

$$\mathbb{E}\|c + M_{k+1}^\mathsf{T}\widetilde{\lambda}^{k+1}\|^2 \le C\mathbb{E}\|\Delta V_{k+1}\|^2 + \frac{C}{m_k}. \tag{24}$$

Finally,

$$\widehat{R}_k(\pi^{k+1}, V^k) = \beta^{-1}\Delta\lambda_{k+1} - \widehat{M}_{k+1}\Delta V_{k+1}.$$

Thus

$$\|R(\pi^{k+1}, V^k)\| \le \beta^{-1}\|\Delta\lambda_{k+1}\| + \|\widehat{M}_{k+1}\|\|\Delta V_{k+1}\| + \|e_k(\pi^{k+1}, V^k)\|.$$

Squaring, taking expectations, and applying Lemma 4.3 controls feasibility by the three increment energies plus $m_k^{-1} + m_{k-1}^{-1}$. Adding this to (23) and (24) proves (7). □

Role. The companion iterate removes endpoint products between random multipliers and the latest value increment. This is the technical step that makes a deterministic-constant fourth-moment complexity bound possible.

C.7 FINITE-TIME COMPLEXITY

Proof of Theorem 5.2. Sum (6) from $k = 1$ to $T-1$ and let $c_\star = \min\{c_\pi, c_V, c_V'\}$. The first update has uniformly bounded fourth moments by the explicit value solve and the empirical-model bounds. Its policy comparison with the deterministic $\pi^0 \in \operatorname{dom}\psi$ also gives $\sup_{m_0}\mathbb{E}\Phi_1 < \infty$, and the companion Fermat condition gives $\sup_{m_0}\mathbb{E}\widetilde{G}_1 < \infty$. Absorb these first-step quantities into the initial constant. The lower bound in Lemma 4.6 absorbs the terminal potential and gives constants A_0, A_1 such that

$$\sum_{k=0}^{T-1}\mathbb{E}\left[\|\Delta\pi_{k+1}\|^2 + \|\Delta V_{k+1}\|^2 + \|\Delta V_k\|^2\right] \le A_0 + A_1\sum_{k=0}^{T-1} m_k^{-1}.$$

Apply Lemma 5.1, sum over k, and divide by T:

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\widetilde{G}_{k+1} \le \frac{A}{T} + \frac{B}{T}\sum_{k=0}^{T-1} m_k^{-1}.$$

Since K is uniform and independent, the left side equals $\mathbb{E}\widetilde{G}_{K+1}$, proving Theorem 5.2.

For a fixed total budget $N = \sum_{k<T} m_k$, convexity of $x \mapsto 1/x$ gives

$$\sum_{k=0}^{T-1} m_k^{-1} \geq \frac{T^2}{N},$$

with equality for equal batches $m_k = N/T$. Substitution gives the equal-batch bound in Theorem 5.2. Taking $T \asymp \sqrt{N}$ yields $\mathbb{E}\widetilde{G}_{K+1} = O(N^{-1/2})$. To make the squared residual at most ε, it is sufficient to choose $T = O(\varepsilon^{-1})$ and $N = O(\varepsilon^{-2})$. □

Role. Theorem 5.2 is the finite-time optimization guarantee that the policy-performance analysis later converts into a return-gap rate.

Proof of Corollary 5.3. If $(1/T)\sum_{k<T} m_k^{-1} = O(T^{-1})$, Theorem 5.2 gives $\mathbb{E}\widetilde{G}_{K+1} = O(T^{-1})$. For fixed N the equal-batch conclusion follows from Jensen's inequality as in the theorem proof. Setting $A/T \leq \varepsilon/2$ and $BT/N \leq \varepsilon/2$ gives $T = O(\varepsilon^{-1})$ and $N = O(\varepsilon^{-2})$. □

Role. The corollary translates the theorem into the conventional iteration and sample-complexity language used in the abstract and introduction.

C.8 General stochastic observations

Proof of Theorem 4.8. Lemma B.3 and the summability conditions in Theorem 4.8 imply $\sum_k \varepsilon_k^2 < \infty$ almost surely. Fix a path in this event. Since $\varepsilon_k \to 0$ and the true reward is bounded, the empirical rewards are pathwise bounded. The pathwise two-dimensional recursion in the proof of Lemma 4.2 therefore gives bounded V^k and λ^k.

Lemma 4.4 and the pathwise Lyapunov calculation give

$$\Phi_k - \Phi_{k+1} \geq c_\pi \|\Delta\pi_{k+1}\|^2 + c_V \|\Delta V_{k+1}\|^2 + c'_V \|\Delta V_k\|^2 - C(\varepsilon_k^2 + \varepsilon_{k-1}^2).$$

The lower bound differs from a fixed constant by at most $C\varepsilon_{k-1}^2$. Summation yields square summability of the policy and value increments. The pathwise five-term dual estimate then gives square summability of $\Delta\lambda_{k+1}$. Hence all increments vanish. The inexact dual update gives feasibility, and the true KKT bound in Lemma 4.4 gives stationarity. Accumulation-point KKT optimality follows from the exact policy comparison and function-value convergence argument in the proof of Theorem 4.7. □

Role. Theorem 4.8 extends the controlled-Markov result to centered observation noise and decaying bias without imposing pathwise bounded noise.

C.9 Exact policy subproblems

For fixed (V^k, λ^k), the empirical Bellman residual is affine in π. Its squared norm plus the proximal term is a strongly convex quadratic. When $\phi \equiv 0$, the policy step is therefore a simplex-constrained quadratic program, separable across states. More generally, consider the nonsmooth, nonconvex regularizer $\phi(\pi) = a\sum_s \|\pi_s\|_0$ with $a > 0$. For each state, enumerate its nonempty action supports, minimize the quadratic on each corresponding simplex face, and select the candidate with smallest original objective. The support of a global minimizer occurs in this finite enumeration, which proves exactness. The enumeration uses at most $2^{|\mathcal{A}|} - 1$ face problems per state. This example supplies an exact nonsmooth policy oracle for small action spaces. The stated iteration and sample bounds count outer iterations and transitions. They do not count internal policy-subproblem operations.

D From Bellman KKT Residuals to Global Policy Performance

Throughout this section $\phi \equiv 0$ and $c = -\mu$. Let $V^\pi = M_\pi^{-1} r_\pi$, $F_\mu(\pi) = -\mu^\mathsf{T} V^\pi$, and $\lambda_\mu^\pi = M_\pi^{-\mathsf{T}}\mu$. For a point (π, V, λ) write

$$e_\pi = \operatorname{dist}(0, J_\pi(\pi, V)^\mathsf{T}\lambda + N_\Pi(\pi)), \quad e_V = \|M_\pi^\mathsf{T}\lambda - \mu\|, \quad e_R = \|M_\pi V - r_\pi\|,$$

so $G = e_\pi^2 + e_V^2 + e_R^2$.

Lemma D.1 (Bellman adjoint representation of the true policy gradient)**.** *The reduced policy objective satisfies*

$$\nabla F_\mu(\pi) = J_\pi(\pi, V^\pi)^\mathsf{T} \lambda_\mu^\pi.$$

Proof. Differentiate the Bellman equation $R(\pi, V^\pi) = 0$ in a feasible policy direction $d\pi$:

$$J_\pi(\pi, V^\pi) d\pi + M_\pi dV^\pi = 0, \qquad dV^\pi = -M_\pi^{-1} J_\pi(\pi, V^\pi) d\pi.$$

Since $F_\mu(\pi) = -\mu^\mathsf{T} V^\pi$,

$$dF_\mu = \mu^\mathsf{T} M_\pi^{-1} J_\pi(\pi, V^\pi) d\pi = (\lambda_\mu^\pi)^\mathsf{T} J_\pi(\pi, V^\pi) d\pi.$$

The coefficient of $d\pi$ is the claimed gradient. □

Role. This identity is the first bridge from the constrained Bellman KKT system to stationarity of the reduced policy objective.

Lemma D.2 (Tabular Bellman policy Jacobian)**.** *For direct tabular policies and* $Q_V(s,a) = r(s,a) + \gamma P(\cdot \mid s,a)^\mathsf{T} V$*,*

$$\frac{\partial R_s(\pi, V)}{\partial \pi(a \mid s)} = -Q_V(s,a), \qquad [J_\pi(\pi, V)^\mathsf{T} \lambda]_{s,a} = -\lambda_s Q_V(s,a).$$

In particular $[\nabla F_\mu(\pi)]_{s,a} = -\lambda_{\mu,s}^\pi Q^\pi(s,a)$.

Proof. For each state,

$$R_s(\pi, V) = V_s - \sum_a \pi(a \mid s) \left[r(s,a) + \gamma P(\cdot \mid s,a)^\mathsf{T} V \right].$$

Differentiating with respect to $\pi(a \mid s)$ gives the first identity. Multiplication by λ gives the second. The last statement follows from Lemma D.1 with $V = V^\pi$. □

Role. The explicit Jacobian isolates the only two discrepancies between the ADMM stationarity vector and the true policy gradient: $V - V^\pi$ and $\lambda - \lambda_\mu^\pi$.

Lemma D.3 (Bellman feasibility controls the value error)**.** *For every* (π, V)*,*

$$\|V - V^\pi\| \le \bar{\kappa} e_R.$$

Proof. Since $M_\pi V - r_\pi = R(\pi, V)$ and $M_\pi V^\pi - r_\pi = 0$,

$$M_\pi (V - V^\pi) = R(\pi, V).$$

Multiplication by M_π^{-1} and the resolvent bound give the claim. □

Role. This lemma converts primal feasibility error into the value mismatch needed in Lemma D.5.

Lemma D.4 (Value stationarity controls the adjoint error)**.** *For every* (π, λ)*,*

$$\|\lambda - \lambda_\mu^\pi\| \le \bar{\kappa} e_V.$$

Proof. By definition $M_\pi^\mathsf{T} \lambda_\mu^\pi = \mu$. Hence

$$M_\pi^\mathsf{T} (\lambda - \lambda_\mu^\pi) = M_\pi^\mathsf{T} \lambda - \mu.$$

Multiplication by $M_\pi^{-\mathsf{T}}$ proves the result. □

Role. This is the dual counterpart of Lemma D.3 and controls the occupancy-weight error in the policy gradient.

Lemma D.5 (Adjoint-gradient perturbation)**.** *Assume* $|r(s,a)| \le R_{\max}$ *and let* $Q_{\max} = R_{\max}/(1-\gamma)$*. There are finite constants* C_V, C_R, C_{VR} *such that*

$$\|J_\pi(\pi, V)^\mathsf{T}\lambda - \nabla F_\mu(\pi)\| \le C_V e_V + C_R e_R + C_{VR} e_V e_R.$$

One may choose constants proportional to $\sqrt{|\mathcal{A}|}Q_{\max}\bar{\kappa}$*,* $\gamma\sqrt{|\mathcal{A}|}\bar{\kappa}^2\|\mu\|$*, and* $\gamma\sqrt{|\mathcal{A}|}\bar{\kappa}^2$*, respectively.*

Proof. Using Lemma D.1, add and subtract the two mixed terms:

$$\begin{aligned} J_\pi(\pi, V)^\mathsf{T}\lambda - \nabla F_\mu(\pi) = {} & J_\pi(\pi, V^\pi)^\mathsf{T}(\lambda - \lambda_\mu^\pi) \\ & + [J_\pi(\pi, V) - J_\pi(\pi, V^\pi)]^\mathsf{T}\lambda_\mu^\pi \\ & + [J_\pi(\pi, V) - J_\pi(\pi, V^\pi)]^\mathsf{T}(\lambda - \lambda_\mu^\pi). \end{aligned}$$

For the first term, Lemma D.2 and $|Q^\pi(s,a)| \le Q_{\max}$ give an operator bound proportional to $\sqrt{|\mathcal{A}|}Q_{\max}\|\lambda - \lambda_\mu^\pi\|$. Lemma D.4 yields the e_V term.

For the second and third terms, the tabular Jacobian satisfies

$$Q_V(s,a) - Q^\pi(s,a) = \gamma P(\cdot \mid s,a)^\mathsf{T}(V - V^\pi),$$

so its change is bounded by $\gamma\sqrt{|\mathcal{A}|}\|V - V^\pi\|$. Also $\|\lambda_\mu^\pi\| \le \bar{\kappa}\|\mu\|$. Lemmas D.3 and D.4 then give the e_R and $e_V e_R$ terms. □

Role. This is the final bridge needed to prove that a small constrained KKT residual implies a small stationarity residual for the true reduced policy objective.

Proof of Lemma 6.1. Distance to a translated closed set is 1-Lipschitz. Therefore

$$\begin{aligned} R_{\mathrm{pol}}(\pi) &= \operatorname{dist}(0, \nabla F_\mu(\pi) + N_\Pi(\pi)) \\ &\le \operatorname{dist}(0, J_\pi(\pi, V)^\mathsf{T}\lambda + N_\Pi(\pi)) + \|J_\pi(\pi, V)^\mathsf{T}\lambda - \nabla F_\mu(\pi)\| \\ &\le e_\pi + C_V e_V + C_R e_R + C_{VR} e_V e_R. \end{aligned}$$

If $G \le 1$, then $e_\pi, e_V, e_R \le \sqrt{G}$ and $e_V e_R \le (e_V^2 + e_R^2)/2 \le G/2 \le \sqrt{G}/2$. Hence

$$R_{\mathrm{pol}}(\pi) \le \left(1 + C_V + C_R + \frac{1}{2}C_{VR}\right)\sqrt{G} =: C_{\mathrm{stat}}\sqrt{G}.$$

□

Role. Lemma 6.1 composes the optimization theory with policy geometry. Every global performance theorem below uses it.

Lemma D.6 (Simplex normal-cone geometry)**.** *Let* $p \in \Delta_m$*,* $q \in \mathbb{R}^m$*, and* $\lambda > 0$*. If*

$$r = \operatorname{dist}(0, -\lambda q + N_{\Delta_m}(p)),$$

then

$$\max_i q_i - p^\mathsf{T} q \le \frac{\sqrt{2}}{\lambda} r.$$

Proof. Choose $n^\star \in N_{\Delta_m}(p)$ attaining the distance and set $h = -\lambda q + n^\star$, so $\|h\| = r$. Let $p^\star \in \arg\max_{u \in \Delta_m} u^\mathsf{T} q$. By the normal-cone inequality, $\langle n^\star, p^\star - p\rangle \le 0$. Therefore

$$\begin{aligned} \lambda(p^{\star\mathsf{T}} q - p^\mathsf{T} q) &= -\langle -\lambda q, p^\star - p\rangle \\ &= -\langle h - n^\star, p^\star - p\rangle \\ &\le -\langle h, p^\star - p\rangle \le \|h\|\|p^\star - p\|. \end{aligned}$$

The Euclidean diameter of the probability simplex is $\sqrt{2}$, giving the claim. □

Role. This lemma converts statewise policy stationarity into a bound on the maximum advantage and is the key geometric input to Theorem 6.2.

Proof of Theorem 6.2. Let $r_s(\pi)$ be the statewise component of $R_{\mathrm{pol}}(\pi)$. By Lemma D.2, the statewise gradient equals $-\lambda^{\pi}_{\mu,s} Q^{\pi}(s,\cdot)$, where

$$\lambda^{\pi}_{\mu,s} = \frac{d^{\pi}_{\mu}(s)}{1-\gamma}.$$

Applying Lemma D.6 statewise gives

$$\max_a A^{\pi}(s,a) \le \sqrt{2}(1-\gamma)\frac{r_s(\pi)}{d^{\pi}_{\mu}(s)}$$

whenever $d^{\pi}_{\mu}(s) > 0$. Under the support condition this is sufficient on every state weighted by $d^{\pi^\star}_{\nu}$. The performance-difference identity yields

$$\begin{aligned} J_\nu(\pi^\star) - J_\nu(\pi) &\le \frac{1}{1-\gamma}\sum_s d^{\pi^\star}_{\nu}(s)\max_a A^{\pi}(s,a) \\ &\le \sqrt{2}\sum_s \frac{d^{\pi^\star}_{\nu}(s)}{d^{\pi}_{\mu}(s)} r_s(\pi) \\ &\le \sqrt{2}\left\| \frac{d^{\pi^\star}_{\nu}}{d^{\pi}_{\mu}} \right\|_2 \left(\sum_s r_s(\pi)^2 \right)^{1/2}. \end{aligned}$$

The last factor is $R_{\mathrm{pol}}(\pi)$, proving Theorem 6.2. Equation the KKT bound in Theorem 6.2 follows from Lemma 6.1. If $G = 0$, the right side vanishes, proving covered KKT global optimality. □

Role. This theorem is the principal RL-specific strengthening of the ADMM stationarity result.

Proof of Corollary 6.3. If $\mu(s) \ge \mu_{\min}$, then the unnormalized discounted occupancy satisfies $\lambda^{\pi}_{\mu,s} = \sum_{t\ge 0}\gamma^t \mathbb{P}_{\mu,\pi}(S_t = s) \ge \mu(s) \ge \mu_{\min}$. Lemma D.6 therefore gives $\max_a A^{\pi}(s,a) \le \sqrt{2} r_s(\pi)/\mu_{\min}$. Substitution into the performance-difference identity under the same evaluation distribution gives

$$J^\star_\mu - J_\mu(\pi) \le \frac{\sqrt{2}}{(1-\gamma)\mu_{\min}} R_{\mathrm{pol}}(\pi).$$

Setting $R_{\mathrm{pol}}(\pi) = 0$ proves global optimality. □

Role. The corollary gives a simple assumption under which the coverage coefficient in Theorem 6.2 is automatically finite.

Proof of Corollary 6.4. The exact-KKT statement is the zero-residual case of Theorem 6.2. For the finite-time bound, let $D_{\max} = 2R_{\max}/(1-\gamma)$. The return gap is bounded above by $D_{\max}$. On $G_K \le 1$, Theorem 6.2 gives the bound $C_0\sqrt{G_K}$ under the uniform occupancy-mismatch assumption. On $G_K > 1$, the gap is at most $D_{\max}\sqrt{G_K}$. Hence the pointwise bound holds for every $G_K \ge 0$ with $C = \max\{C_0, D_{\max}\}$. Taking expectations and applying Jensen gives

$$\mathbb{E}[J^\star_\nu - J_\nu(\pi_K)] \le C\sqrt{\mathbb{E}G_K}.$$

Substitution of the companion-residual bound proves the claim, with the policy component of the companion output. □

Role. This corollary directly composes the finite-time KKT bound with the policy-performance conversion.

Proof of Proposition 6.5. Consider states s_0, s_1, s_T and training distribution $\mu = \delta_{s_0}$. At s_0, action a terminates with reward 1, while action b moves to s_1 with reward 0. At s_1, action c terminates with reward 0, while action d terminates with reward 10. Take $\gamma = 0.9$ and the policy choosing a at s_0 and c at s_1.

Then $V^\pi(s_0) = 1$ and $V^\pi(s_1) = 0$. At s_0, $Q^\pi(s_0, a) = 1$ and $Q^\pi(s_0, b) = 0$, so the current action is strictly greedy. The policy never visits s_1, hence $d^\pi_\mu(s_1) = 0$ and the policy gradient has zero weight there. Thus the policy is first-order stationary. However the coordinated policy choosing b at s_0 and d at s_1 has return $0.9 \times 10 = 9 > 1$. The failure is exactly the missing occupancy support. □

Role. The proposition shows that the coverage assumption in Theorem 6.2 is necessary for first-order information to identify all globally relevant states.

Proof of Proposition 6.6. Consider the three-stage chain $s_0 \to s_1 \to s_2 \to s_T$ with actions stop and continue. All rewards are zero except that continuing at s_2 gives reward -1. Let π_ε continue with probability ε at every state. The optimal value is zero, while a loss occurs only if all three continue actions are chosen, so

$$J^\star - J(\pi_\varepsilon) = \gamma^2 \varepsilon^3.$$

The unnormalized occupancies are $1, \gamma\varepsilon, \gamma^2\varepsilon^2$. The corresponding action-value gaps are $\gamma^2\varepsilon^2, \gamma\varepsilon, 1$. For an interior two-action simplex, the statewise normal-cone residual is the weighted action gap divided by $\sqrt{2}$, so all three state residuals equal $\gamma^2\varepsilon^2/\sqrt{2}$. Therefore

$$R_{\mathrm{pol}}(\pi_\varepsilon)^2 = 3\frac{\gamma^4\varepsilon^4}{2}.$$

Consequently

$$\frac{J^\star - J(\pi_\varepsilon)}{R_{\mathrm{pol}}(\pi_\varepsilon)^2} = \frac{2}{3\gamma^2\varepsilon} \to \infty.$$

No uniform constant can satisfy a quadratic PL-type inequality for all direct tabular policies. □

Role. This negative result justifies why the generic policy-performance theorem has an $O(\sqrt{G})$ conversion and why Theorem 6.8 requires additional curvature.

Proof of Theorem 6.8. Let $\bar{r}_s(\pi)$ be the unweighted statewise stationarity residual. Because the true statewise policy gradient is scaled by $\lambda^\pi_{\mu,s} = d^\pi_\mu(s)/(1-\gamma)$,

$$r_s(\pi) = \frac{d^\pi_\mu(s)}{1-\gamma}\bar{r}_s(\pi), \qquad \bar{r}_s(\pi) = \frac{1-\gamma}{d^\pi_\mu(s)} r_s(\pi).$$

Assumption 6.7 therefore gives

$$\delta^\pi_s \le \frac{(1-\gamma)^2}{2\mu_B [d^\pi_\mu(s)]^2} r_s(\pi)^2.$$

The Bellman performance bound is

$$J_\nu(\pi^\star) - J_\nu(\pi) \le \frac{1}{1-\gamma}\sum_s d^{\pi^\star}_\nu(s)\delta^\pi_s.$$

Substituting the previous inequality gives

$$\begin{aligned} J_\nu(\pi^\star) - J_\nu(\pi) &\le \frac{1-\gamma}{2\mu_B}\sum_s \frac{d^{\pi^\star}_\nu(s)}{[d^\pi_\mu(s)]^2} r_s(\pi)^2 \\ &\le \frac{1-\gamma}{2\mu_B}\mathcal{C}_{\mathrm{PL}}\sum_s r_s(\pi)^2, \end{aligned}$$

which proves (6.8). Lemma 6.1 then gives the $O(G)$ form. □

Role. The theorem identifies the additional one-step curvature needed for policy performance to inherit the squared-KKT rate without the square-root loss.

Proof of Theorem 6.9. The global truncation argument in Corollary 6.4 gives $\mathbb{E}[J^\star - J(\pi_K)] \le C\sqrt{\mathbb{E}G_K}$. Under the quadratic improvement and stronger occupancy-ratio assumptions, Theorem 6.8 gives $J^\star - J(\pi_K) \le C_0 G_K$ on $G_K \le 1$. On $G_K > 1$, bounded returns give $J^\star - J(\pi_K) \le D_{\max} G_K$. Thus $\mathbb{E}[J^\star - J(\pi_K)] \le \max\{C_0, D_{\max}\}\mathbb{E}G_K$. The pointwise bounds also imply policy-performance convergence whenever $G_K \to 0$, and global optimality at covered exact KKT points. □

Role. This theorem summarizes the second layer of value supplied by RL structure: it interprets the optimization limit point rather than merely proving that the ADMM iterates become stationary.

A sufficient action-gap condition. Suppose that on the policy set of interest every suboptimal action has gap $\max_b Q^\pi(s,b) - Q^\pi(s,a) \ge \Delta > 0$. Fix a state and write $p = \pi_s$, $q = Q^\pi(s,\cdot)$, and $r = \operatorname{dist}(0, -q + N_\Delta(p))$. If p is supported on maximizing actions, $\delta_s^\pi = 0$. Otherwise choose a supported suboptimal action a and a maximizing action b. For every $n \in N_\Delta(p)$, the feasible direction $e_b - e_a$ gives $n_b - n_a \le 0$. Thus

$$\langle -q + n, e_b - e_a \rangle \le -\Delta, \qquad r \ge \Delta/\sqrt{2}.$$

Lemma D.6 gives $\delta_s^\pi \le \sqrt{2} r \le 2r^2/\Delta$. Therefore Assumption 6.7 holds with $\mu_B = \Delta/4$. For example, when transitions are action independent and rewards have a positive gap between optimal and suboptimal actions at each nontrivial state, the same Δ works for every policy.

E Smooth Nonlinear Policy Parameterization

Let $\Theta \subset \mathbb{R}^p$ be closed and convex and let π_θ be a smooth policy on a neighborhood of Θ. Set $\psi_\Theta = \phi + \delta_\Theta$ and assume it is proper, lower semicontinuous, and bounded below. Policy stationarity below uses $\partial\psi_\Theta$. Initialize deterministically with $\theta^0 \in \operatorname{dom}\psi_\Theta$ and finite V^0, λ^0. Define

$$P_\theta(s' \mid s) = \sum_a \pi_\theta(a \mid s) P(s' \mid s, a), \quad r_\theta(s) = \sum_a \pi_\theta(a \mid s) r(s,a), \quad M_\theta = I - \gamma P_\theta,$$

and $R(\theta, V) = M_\theta V - r_\theta$. Consider

$$\min_{\theta, V} \ \phi(\theta) + c^\mathsf{T} V \quad \text{s.t.} \quad R(\theta, V) = 0.$$

Assume $\|M_\theta - M_{\theta'}\| \le L_M \|\theta - \theta'\|$ and $b_A(\theta) = \phi(\theta) + c^\mathsf{T} M_\theta^{-1} r_\theta$ is bounded below.

Lemma E.1 (Euclidean Bellman-resolvent bound)**.** *For every row-stochastic P_θ,*

$$\|(I - \gamma P_\theta)^{-1}\|_2 \le \frac{\sqrt{|\mathcal{S}|}}{1-\gamma}.$$

Proof. The Neumann series gives $M_\theta^{-1} = \sum_{t \ge 0} \gamma^t P_\theta^t$. Since each P_θ^t is row stochastic,

$$\|M_\theta^{-1}\|_\infty \le \sum_{t \ge 0} \gamma^t = \frac{1}{1-\gamma}.$$

For every vector x, $\|M_\theta^{-1} x\|_2 \le \sqrt{|\mathcal{S}|}\|M_\theta^{-1} x\|_\infty \le \sqrt{|\mathcal{S}|}(1-\gamma)^{-1}\|x\|_\infty \le \sqrt{|\mathcal{S}|}(1-\gamma)^{-1}\|x\|_2$. □

Role. This lemma supplies the uniform inverse constant κ_M used in every nonlinear-policy stability bound.

Lemma E.2 (Nonlinear inverse-map stability)**.** *If $\kappa_M = \sup_\theta \|M_\theta^{-1}\| < \infty$, then*

$$\|M_\theta^{-1} - M_{\theta'}^{-1}\| \le \kappa_M^2 L_M \|\theta - \theta'\|.$$

Proof. Apply Lemma A.1 with $M = M_{\theta'}$ and $M' = M_\theta$ and then use the assumed Lipschitz bound for M_θ. □

Role. This is the nonlinear-policy version of the resolvent sensitivity used to control multiplier increments.

The exact proximal updates are

$$\theta^{k+1} \in \arg\min_{\theta\in\Theta}\left\{\mathcal{L}_\beta(\theta, V^k, \lambda^k) + \frac{1}{2\eta_\theta}\|\theta-\theta^k\|^2\right\},$$

$$V^{k+1} = \arg\min_{V}\left\{\mathcal{L}_\beta(\theta^{k+1}, V, \lambda^k) + \frac{1}{2\eta_V}\|V-V^k\|^2\right\}, \qquad \lambda^{k+1} = \lambda^k + \beta R(\theta^{k+1}, V^{k+1}).$$

Lemma E.3 (Exact policy-step decrease)**.** *The exact proximal policy update satisfies*

$$\mathcal{L}_\beta(\theta^k, V^k, \lambda^k) - \mathcal{L}_\beta(\theta^{k+1}, V^k, \lambda^k) \geq \frac{1}{2\eta_\theta}\|\Delta\theta_{k+1}\|^2.$$

Proof. Use θ^k as a feasible comparison point in the definition of θ^{k+1}. The proximal term vanishes at θ^k, while at θ^{k+1} it contributes $(2\eta_\theta)^{-1}\|\Delta\theta_{k+1}\|^2$. □

Role. This is the primal decrease that will dominate the policy-dependent part of the dual ascent.

Lemma E.4 (Nonlinear value-dual identity)**.** *For every k,*

$$\lambda^{k+1} = -M_{k+1}^{-\mathsf{T}}\left(c + \eta_V^{-1}\Delta V_{k+1}\right), \qquad M_{k+1} = M_{\theta^{k+1}}.$$

Proof. The value-subproblem first-order condition is

$$0 = c + M_{k+1}^{\mathsf{T}}\lambda^k + \beta M_{k+1}^{\mathsf{T}} R_{k+1} + \eta_V^{-1}\Delta V_{k+1}.$$

The dual update gives $\lambda^{k+1} = \lambda^k + \beta R_{k+1}$. Substitution gives $M_{k+1}^{\mathsf{T}}\lambda^{k+1} = -(c + \eta_V^{-1}\Delta V_{k+1})$, and Lemma E.1 guarantees invertibility. □

Role. This identity is the exact nonlinear analogue of the tabular Bellman multiplier representation.

Lemma E.5 (Nonlinear multiplier increment)**.** *There are constants*

$$A_A = 3\kappa_M^4 L_M^2\|c\|^2, \qquad B_A = \frac{3\kappa_M^2}{\eta_V^2},$$

such that

$$\|\Delta\lambda_{k+1}\|^2 \leq A_A\|\Delta\theta_{k+1}\|^2 + B_A\|\Delta V_{k+1}\|^2 + B_A\|\Delta V_k\|^2.$$

Proof. Subtract Lemma E.4 at consecutive iterations:

$$\Delta\lambda_{k+1} = -(M_{k+1}^{-\mathsf{T}} - M_k^{-\mathsf{T}})c - \eta_V^{-1}M_{k+1}^{-\mathsf{T}}\Delta V_{k+1} + \eta_V^{-1}M_k^{-\mathsf{T}}\Delta V_k.$$

Lemma E.2 bounds the first term by $\kappa_M^2 L_M\|c\|\|\Delta\theta_{k+1}\|$. Bound the other two by $\kappa_M\eta_V^{-1}$ times their value increments and apply $\|x+y+z\|^2 \leq 3(\|x\|^2 + \|y\|^2 + \|z\|^2)$. □

Role. This lemma is the deterministic nonlinear-policy multiplier-control result used in Theorems E.7 and E.9.

Lemma E.6 (Nonlinear Lyapunov lower bound)**.** *Let*

$$\Phi_k^A = \mathcal{L}_\beta(\theta^k, V^k, \lambda^k) + \tau\|\Delta V_k\|^2.$$

Then

$$\Phi_k^A \geq b_A^{\inf} + \frac{\beta}{4}\|R_k\|^2 + \left(\tau - \frac{\kappa_M^2}{\beta\eta_V^2}\right)\|\Delta V_k\|^2.$$

Proof. Since $M_k V^k = r_k + R_k$, $V^k = M_k^{-1}(r_k + R_k)$. Lemma E.4 at the previous update gives $\lambda^k = -M_k^{-\mathsf{T}}(c + \eta_V^{-1}\Delta V_k)$. Substitute both identities into the augmented Lagrangian:

$$\mathcal{L}_\beta^k = b_A(\theta^k) - \eta_V^{-1}\langle\Delta V_k, M_k^{-1}R_k\rangle + \frac{\beta}{2}\|R_k\|^2.$$

Young's inequality gives

$$\eta_V^{-1}\kappa_M\|\Delta V_k\|\|R_k\| \leq \frac{\beta}{4}\|R_k\|^2 + \frac{\kappa_M^2}{\beta\eta_V^2}\|\Delta V_k\|^2.$$

Use $b_A(\theta^k) \geq b_A^{\inf}$ and add $\tau\|\Delta V_k\|^2$. □

Role. This lemma prevents the deterministic nonlinear-policy Lyapunov from diverging to $-\infty$ and permits summation of the descent inequality.

Theorem E.7 (Exact nonlinear-policy convergence)**.** *If*

$$\frac{B_A}{\beta} < \tau < \frac{1}{2\eta_V} - \frac{B_A}{\beta}, \qquad \frac{1}{2\eta_\theta} > \frac{A_A}{\beta},$$

then there are $c_\theta, c_V, c_V' > 0$ such that

$$\Phi_k^A - \Phi_{k+1}^A \geq c_\theta \|\Delta\theta_{k+1}\|^2 + c_V \|\Delta V_{k+1}\|^2 + c_V' \|\Delta V_k\|^2.$$

Consequently $\Delta\theta_k, \Delta V_k, \Delta\lambda_k, R_k \to 0$. If $D_\theta M_\theta$ and $D_\theta r_\theta$ are bounded on the visited region, the KKT residual converges to zero.

Proof. Lemma E.3 gives policy decrease. The value comparison point V^k similarly gives

$$\mathcal{L}_\beta(\theta^{k+1}, V^k, \lambda^k) - \mathcal{L}_\beta(\theta^{k+1}, V^{k+1}, \lambda^k) \geq \frac{1}{2\eta_V} \|\Delta V_{k+1}\|^2.$$

Because the deterministic dual update is exact, the multiplier step increases the augmented Lagrangian by

$$\langle \Delta\lambda_{k+1}, R_{k+1} \rangle = \beta^{-1} \|\Delta\lambda_{k+1}\|^2.$$

Insert Lemma E.5. After adding $\tau\|\Delta V_k\|^2 - \tau\|\Delta V_{k+1}\|^2$, the coefficients are

$$c_\theta = \frac{1}{2\eta_\theta} - \frac{A_A}{\beta}, \quad c_V = \frac{1}{2\eta_V} - \frac{B_A}{\beta} - \tau, \quad c_V' = \tau - \frac{B_A}{\beta},$$

which are positive by assumption. Lemma E.6 bounds Φ_k^A from below, so summation gives square summability of the policy and value increments. Lemma E.5 then gives $\Delta\lambda_k \to 0$, while $\Delta\lambda_{k+1} = \beta R_{k+1}$ gives feasibility.

For stationarity, the policy-subproblem Fermat condition is

$$0 \in \partial\psi_\Theta(\theta^{k+1}) + D_\theta R(\theta^{k+1}, V^k)^\mathsf{T} [\lambda^k + \beta R(\theta^{k+1}, V^k)] + \eta_\theta^{-1} \Delta\theta_{k+1}.$$

The bracket differs from λ^{k+1} by $O(\|\Delta V_{k+1}\|)$, and bounded $D_\theta M_\theta, D_\theta r_\theta$ make the change from V^k to V^{k+1} linear in $\|\Delta V_{k+1}\|$. Thus the policy KKT residual tends to zero. Lemma E.4 gives the value residual, and feasibility has already been proved. □

Role. This theorem shows that the Bellman-resolvent mechanism is not tied to tabular direct policies.

For an implementable parameter step define

$$h_k(\theta) = \langle \lambda^k, R(\theta, V^k) \rangle + \frac{\beta}{2} \|R(\theta, V^k)\|^2,$$

and use

$$\theta^{k+1} \in \arg\min_{\theta \in \Theta} \left\{ \phi(\theta) + \langle \nabla h_k(\theta^k), \theta - \theta^k \rangle + \frac{1}{2\eta_{\theta,k}} \|\theta - \theta^k\|^2 \right\}.$$

Lemma E.8 (Proximal-gradient policy decrease)**.** *If ∇h_k is $L_{\theta,k}$-Lipschitz on the visited level set, then*

$$\mathcal{L}_\beta(\theta^k, V^k, \lambda^k) - \mathcal{L}_\beta(\theta^{k+1}, V^k, \lambda^k) \geq \left(\frac{1}{2\eta_{\theta,k}} - \frac{L_{\theta,k}}{2} \right) \|\Delta\theta_{k+1}\|^2.$$

Proof. Optimality of the proximal-gradient subproblem, compared with θ^k, gives

$$\phi(\theta^{k+1}) - \phi(\theta^k) + \langle \nabla h_k(\theta^k), \Delta\theta_{k+1} \rangle + \frac{1}{2\eta_{\theta,k}} \|\Delta\theta_{k+1}\|^2 \leq 0.$$

The descent lemma gives $h_k(\theta^{k+1}) - h_k(\theta^k) \leq \langle \nabla h_k(\theta^k), \Delta\theta_{k+1} \rangle + (L_{\theta,k}/2)\|\Delta\theta_{k+1}\|^2$. Adding the two inequalities proves the claim. □

Role. This lemma replaces the exact policy-subproblem decrease in Theorem E.7 and makes the nonlinear-policy algorithm implementable by one first-order step.

Theorem E.9 (Implementable nonlinear-policy convergence and complexity)**.** *Suppose the value-step parameter conditions of Theorem E.7 hold and the policy step is the proximal-gradient update above. Assume $D_\theta M_\theta$, $D_\theta r_\theta$ are uniformly bounded on the visited region and ∇h_k is $L_{\theta,k}$-Lipschitz on every policy comparison segment, with*

$$0 < \underline{\eta} \le \eta_{\theta,k}, \qquad L_{\theta,k} \le \overline{L} < \infty.$$

If

$$\frac{1}{2\eta_{\theta,k}} - \frac{L_{\theta,k}}{2} - \frac{A_A}{\beta} \ge \alpha_\theta > 0$$

for all k, then the corrected Lyapunov decreases and the KKT residual converges to zero. Moreover there is $C_G < \infty$ such that

$$G^A_{k+1} \le C_G \left(\|\Delta\theta_{k+1}\|^2 + \|\Delta V_{k+1}\|^2 + \|\Delta V_k\|^2 \right),$$

so $\min_{1\le k\le T} G^A_k \le C/T$.

Proof. Replace Lemma E.3 by Lemma E.8 in the proof of Theorem E.7. The policy coefficient after dual ascent becomes $1/(2\eta_{\theta,k}) - L_{\theta,k}/2 - A_A/\beta$, which is uniformly at least α_θ. The value coefficients and Lyapunov lower bound are unchanged. Hence the three increment-square sequences are summable. The value and multiplier iterates are bounded independently of this descent argument: the recursion in Lemma 4.2 applies with bounded deterministic rewards, and the value margins imply $\beta\eta_V > 12\kappa_M^2 > 4\kappa_M^2$.

The proximal-gradient optimality condition is

$$0 \in \partial\psi_\Theta(\theta^{k+1}) + \nabla h_k(\theta^k) + \eta_{\theta,k}^{-1}\Delta\theta_{k+1}.$$

Comparison with true policy stationarity at $(\theta^{k+1}, V^{k+1}, \lambda^{k+1})$ gives

$$e_{\theta,k+1} \le (\underline{\eta}^{-1} + \overline{L})\|\Delta\theta_{k+1}\| + C\|\Delta V_{k+1}\|.$$

The second term follows from bounded value and multiplier iterates, bounded policy derivatives, and $\lambda^{k+1} - [\lambda^k + \beta R(\theta^{k+1}, V^k)] = \beta M_{k+1}\Delta V_{k+1}$. The value residual is $\eta_V^{-1}\|\Delta V_{k+1}\|$. Feasibility is controlled by Lemma E.5. Thus all coefficients in the residual bound are uniform in k and the residual converges to zero. Squaring gives the displayed G^A_{k+1} bound. Summing the Lyapunov decrease from 1 to T gives a constant upper bound on the sum of the increment energies, and therefore

$$\min_{1\le k\le T} G^A_k \le \frac{1}{T}\sum_{k=1}^{T} G^A_k \le \frac{C}{T}.$$

□

Role. This theorem is the practical nonlinear-policy extension used in the paper's general operator-stability claim.

Uniform constants and backtracking. A compact convex parameter set and a twice continuously differentiable policy map on its neighborhood give bounded first and second derivatives. Together with the value-dual bound above, they yield a uniform $\overline{L}$. Choose a fixed trial step $\eta_{\mathrm{trial}} > 0$ and a shrink factor $q \in (0,1)$. Restart backtracking at η_{trial} on every iteration and shrink until the descent condition with margin $A_A/\beta + \alpha_\theta$ holds. Every step at most $\eta_\star = (\overline{L} + 2A_A/\beta + 2\alpha_\theta)^{-1}$ is acceptable. Therefore the accepted steps satisfy $\eta_{\theta,k} \ge \min\{\eta_{\mathrm{trial}}, q\eta_\star\} > 0$.

F Projected Bellman Equations with Function Approximation

Let $V_w = \Xi w$ with a full-column-rank $\Xi \in \mathbb{R}^{|\mathcal{S}|\times m}$ and $m < |\mathcal{S}|$. Fix a positive diagonal matrix D and define the D-orthogonal projection

$$\Pi_D = \Xi(\Xi^\mathsf{T} D\Xi)^{-1}\Xi^\mathsf{T} D.$$

The projected Bellman equation $\Xi w = \Pi_D T_\theta(\Xi w)$ is equivalent to

$$G(\theta, w) = A_\theta w - b_\theta = 0, \qquad A_\theta = \Xi^\mathsf{T} D M_\theta \Xi, \qquad b_\theta = \Xi^\mathsf{T} D r_\theta.$$

Assume $\sup_\theta \|A_\theta^{-1}\| \le \kappa_A$ and $\|A_\theta - A_{\theta'}\| \le L_A\|\theta - \theta'\|$. Let $c_w = \Xi^\mathsf{T} c$.

Lemma F.1 (Projected value-dual identity)**.** *If the w-subproblem uses a proximal term $(2\eta_w)^{-1}\|w - w^k\|^2$ and the projected multiplier update is*

$$\nu^{k+1} = \nu^k + \beta(A_{k+1}w^{k+1} - b_{k+1}),$$

then

$$\nu^{k+1} = -A_{k+1}^{-\mathsf{T}}\left(c_w + \eta_w^{-1}\Delta w_{k+1}\right).$$

Proof. The w-subproblem first-order condition is

$$0 = c_w + A_{k+1}^\mathsf{T}\nu^k + \beta A_{k+1}^\mathsf{T}(A_{k+1}w^{k+1} - b_{k+1}) + \eta_w^{-1}\Delta w_{k+1}.$$

The bracket $\nu^k + \beta(A_{k+1}w^{k+1} - b_{k+1})$ equals ν^{k+1}. Thus $A_{k+1}^\mathsf{T}\nu^{k+1} = -(c_w + \eta_w^{-1}\Delta w_{k+1})$, and uniform invertibility gives the formula. □

Role. This identity restores the same multiplier representation that drives the Bellman-resolvent proof, now in the lower-dimensional projected space.

Theorem F.2 (Projected Bellman-ADMM convergence and complexity)**.** *Assume A_θ is uniformly invertible and Lipschitz, $D_\theta A_\theta, D_\theta b_\theta$ are bounded, and the projected reduced objective is bounded below. Use the composite policy regularizer ψ_Θ and a deterministic feasible-domain initialization as in Appendix E. Require all conditions of Theorem E.9 after the replacements below, including the value-step margins, $\inf_k \eta_{\theta,k} > 0$, and a uniform Lipschitz constant for the projected smooth augmented gradient. Then a corrected projected Lyapunov function decreases, the projected KKT residual converges to zero, and the minimum squared projected KKT residual satisfies $O(T^{-1})$.*

Proof. We verify that every ingredient of Theorem E.9 transfers under the replacements

$$M_\theta \mapsto A_\theta, \qquad V \mapsto w, \qquad \lambda \mapsto \nu, \qquad r_\theta \mapsto b_\theta, \qquad c \mapsto c_w.$$

First, the inverse identity and the assumptions give

$$\|A_\theta^{-1} - A_{\theta'}^{-1}\| \le \kappa_A^2 L_A\|\theta - \theta'\|.$$

Second, Lemma F.1 is exactly the value-dual identity needed to compare consecutive multipliers. Subtraction therefore gives a dual-increment estimate

$$\|\Delta\nu_{k+1}\|^2 \le A_B\|\Delta\theta_{k+1}\|^2 + B_B\|\Delta w_{k+1}\|^2 + B_B\|\Delta w_k\|^2$$

for finite constants A_B, B_B. Third, the projected w comparison step yields $(2\eta_w)^{-1}\|\Delta w_{k+1}\|^2$ decrease, and the policy step gives the same proximal-gradient decrease as Lemma E.8. Fourth, substituting $w^k = A_k^{-1}(b_k + G_k)$ and $\nu^k = -A_k^{-\mathsf{T}}(c_w + \eta_w^{-1}\Delta w_k)$ into the projected augmented Lagrangian gives the same lower-bound calculation as Lemma E.6, with κ_M, η_V replaced by κ_A, η_w.

Under the strict parameter margins, the resulting corrected Lyapunov therefore has positive coefficients on $\|\Delta\theta_{k+1}\|^2$, $\|\Delta w_{k+1}\|^2$, and $\|\Delta w_k\|^2$. Summation gives square summability of the increments, the multiplier identity gives projected feasibility, and the policy/w optimality conditions give convergence of the projected KKT residual. The same residual-to-increment argument as in Theorem E.9 yields $G_{k+1}^B \le CD_{k+1}^B$, where the sum of D_{k+1}^B is bounded. Hence $\min_{k\le T} G_k^B \le C/T$. □

Role. The theorem shows that exact state-value representation is not required for the primal-dual stability mechanism. What is required is an invertible projected Bellman matrix.

Assume now that the projected Bellman map is contractive in a chosen norm: for some $q < 1$,

$$\|\Pi_D T_\theta U - \Pi_D T_\theta V\| \le q\|U - V\|.$$

Let $\bar{V}^\theta = \Xi\bar{w}^\theta$ be its fixed point.

Lemma F.3 (Projected value approximation). *With* $\varepsilon_{\mathrm{val}}(\theta) = \|(I - \Pi_D)V^\theta\|$,

$$\|\bar{V}^\theta - V^\theta\| \le \frac{\varepsilon_{\mathrm{val}}(\theta)}{1-q}.$$

Proof. Use $\bar{V}^\theta = \Pi_D T_\theta \bar{V}^\theta$ and $V^\theta = T_\theta V^\theta$:

$$\begin{aligned}\|\bar{V}^\theta - V^\theta\| &\le \|\Pi_D T_\theta \bar{V}^\theta - \Pi_D T_\theta V^\theta\| + \|\Pi_D V^\theta - V^\theta\| \\ &\le q\|\bar{V}^\theta - V^\theta\| + \varepsilon_{\mathrm{val}}(\theta).\end{aligned}$$

Move the first term to the left. □

Role. This lemma quantifies the value-representation error, but the next proposition shows why this alone is insufficient for policy-gradient accuracy.

Proposition F.4 (Value error does not control parameter-derivative error). *There exists a smooth policy family for which*

$$\sup_\theta \|\bar{V}^\theta - V^\theta\| = \varepsilon, \qquad \sup_\theta \|D_\theta \bar{V}^\theta - D_\theta V^\theta\| = \sqrt{\varepsilon}.$$

Thus no constant independent of ε *can bound the derivative error linearly by the value error.*

Proof. Consider a one-step terminating MDP with two states. Let the approximation class represent the first state's value exactly but force the second coordinate to zero. At the second state there are two actions with rewards $+1$ and -1. Define

$$\pi_\theta(+ \mid s_2) = \frac{1}{2} + \frac{\varepsilon}{2}\sin(\theta/\sqrt{\varepsilon}).$$

Then the true second-state value is $V_2^\theta = \varepsilon \sin(\theta/\sqrt{\varepsilon})$, while the projected value is $\bar{V}_2^\theta = 0$. Hence the uniform value error is ε. But

$$\frac{d}{d\theta} V_2^\theta = \sqrt{\varepsilon}\cos(\theta/\sqrt{\varepsilon}),$$

whose supremum is $\sqrt{\varepsilon}$. Therefore a uniform inequality $\|D_\theta \bar{V} - D_\theta V\| \le C\|\bar{V} - V\|$ would require $\sqrt{\varepsilon} \le C\varepsilon$ for arbitrarily small ε, which is impossible. □

Role. This negative result motivates the adjoint-based error decomposition in Lemmas F.5 and F.6.

Let the true adjoint satisfy $M_\theta^\mathsf{T}\lambda^\theta = -c$. Let ν^θ solve the projected adjoint equation and define the lifted projected dual $\widehat{\lambda}^\theta = D\Xi\nu^\theta$.

Lemma F.5 (Adjoint-gradient approximation). *Suppose* $D_\theta M_\theta$ *and* $D_\theta r_\theta$ *are bounded on the relevant set. Then there are constants* $C_V, C_\lambda < \infty$ *such that*

$$\|D_\theta R(\theta, V^\theta)^\mathsf{T}\lambda^\theta - D_\theta R(\theta, \bar{V}^\theta)^\mathsf{T}\widehat{\lambda}^\theta\| \le C_V\|V^\theta - \bar{V}^\theta\| + C_\lambda\|\lambda^\theta - \widehat{\lambda}^\theta\|.$$

Proof. Add and subtract $D_\theta R(\theta, \bar{V}^\theta)^\mathsf{T}\lambda^\theta$:

$$\begin{aligned}&D_\theta R(\theta, V^\theta)^\mathsf{T}\lambda^\theta - D_\theta R(\theta, \bar{V}^\theta)^\mathsf{T}\widehat{\lambda}^\theta \\ &= \left[D_\theta R(\theta, V^\theta) - D_\theta R(\theta, \bar{V}^\theta)\right]^\mathsf{T}\lambda^\theta + D_\theta R(\theta, \bar{V}^\theta)^\mathsf{T}(\lambda^\theta - \widehat{\lambda}^\theta).\end{aligned}$$

Because $D_\theta R(\theta, V)[h] = D_\theta M_\theta[h]V - D_\theta r_\theta[h]$, the first bracket is linear in $V^\theta - \bar{V}^\theta$ and is bounded by $C_V\|V^\theta - \bar{V}^\theta\|$ after using bounded λ^θ. Boundedness of $D_\theta R$ gives the second term with constant C_λ. □

Role. This lemma replaces the false derivative-from-value estimate by a correct value-plus-adjoint approximation bound.

Lemma F.6 (Lifted-dual quasi-optimality)**.** *There is $C_{\mathrm{dual}} < \infty$ such that*

$$\|\lambda^\theta - \widehat{\lambda}^\theta\| \le C_{\mathrm{dual}} \inf_{z \in \mathrm{range}(D\Xi)} \|\lambda^\theta - z\|.$$

Proof. Take any $z = D\Xi a$ in the approximation space. The true adjoint equation and the projected adjoint equation imply

$$A_\theta^\mathsf{T} \nu^\theta = \Xi^\mathsf{T} M_\theta^\mathsf{T} \lambda^\theta, \qquad A_\theta^\mathsf{T} a = \Xi^\mathsf{T} M_\theta^\mathsf{T} z.$$

Subtracting gives

$$\nu^\theta - a = A_\theta^{-\mathsf{T}} \Xi^\mathsf{T} M_\theta^\mathsf{T} (\lambda^\theta - z).$$

Therefore

$$\|\widehat{\lambda}^\theta - z\| \le \|D\Xi\| \, \|A_\theta^{-\mathsf{T}} \Xi^\mathsf{T} M_\theta^\mathsf{T}\| \|\lambda^\theta - z\|.$$

By the triangle inequality,

$$\|\lambda^\theta - \widehat{\lambda}^\theta\| \le \left(1 + \|D\Xi\| \sup_\theta \|A_\theta^{-\mathsf{T}} \Xi^\mathsf{T} M_\theta^\mathsf{T}\|\right) \|\lambda^\theta - z\|.$$

Minimize over $z \in \mathrm{range}(D\Xi)$. □

Role. Together with Lemmas F.3 and F.5, this lemma gives a controlled route from projected-Bellman stationarity back to the true policy gradient.

G Explicit Discounted Occupancy Formulation

Use the policy domain, composite regularizer ψ_Θ, and initialization conventions of Appendix E. Let

$$d_\theta(s) = (1-\gamma) \sum_{t \ge 0} \gamma^t \mathbb{P}_{\rho, \pi_\theta}(S_t = s), \qquad q_\theta(s,a) = d_\theta(s) \pi_\theta(a \mid s).$$

Let $P^\mathsf{T} q$ denote the state inflow and let $\Pi_\theta d$ be the state-action vector with entries $\pi_\theta(a \mid s) d(s)$. The occupancy constraints are

$$d - \gamma P^\mathsf{T} q = (1-\gamma)\rho, \qquad q - \Pi_\theta d = 0.$$

Define

$$x = \begin{bmatrix} q \\ d \end{bmatrix}, \qquad b = \begin{bmatrix} (1-\gamma)\rho \\ 0 \end{bmatrix}, \qquad K_\theta = \begin{bmatrix} -\gamma P^\mathsf{T} & I \\ I & -\Pi_\theta \end{bmatrix}.$$

Lemma G.1 (Occupancy-operator invertibility)**.** *For every policy π_θ, K_θ is invertible. If*

$$K_\theta \begin{bmatrix} q \\ d \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix},$$

then

$$d = (I - \gamma P_\theta^\mathsf{T})^{-1}(u + \gamma P^\mathsf{T} v), \qquad q = v + \Pi_\theta d.$$

In the block ℓ_1 norm, $\|K_\theta^{-1}\|_{1,\oplus} \le 2/(1-\gamma)$. Hence $\sup_\theta \|K_\theta^{-1}\|_2 < \infty$.

Proof. The second block row gives $q = v + \Pi_\theta d$. Substitute this into the first row:

$$-\gamma P^\mathsf{T}(v + \Pi_\theta d) + d = u.$$

Since $P^\mathsf{T} \Pi_\theta = P_\theta^\mathsf{T}$,

$$(I - \gamma P_\theta^\mathsf{T}) d = u + \gamma P^\mathsf{T} v,$$

which yields the stated formulas because $I - \gamma P_\theta^\mathsf{T}$ is invertible.

For the norm bound, $\|(I - \gamma P_\theta^\mathsf{T})^{-1}\|_1 \le (1-\gamma)^{-1}$, $\|P^\mathsf{T}\|_1 = 1$, and $\|\Pi_\theta\|_1 = 1$. Hence

$$\|d\|_1 \le \frac{\|u\|_1 + \gamma \|v\|_1}{1-\gamma}, \qquad \|q\|_1 \le \|v\|_1 + \|d\|_1.$$

Adding and simplifying gives a valid bound no larger than $2(1-\gamma)^{-1}(\|u\|_1 + \|v\|_1)$ after absorbing the fixed coefficients. Finite-dimensional norm equivalence gives the Euclidean bound. □

Role. This lemma shows that discounted occupancy flow provides a second RL-specific invertible operator, independent of the Bellman value representation.

Proposition G.2 (Occupancy equivalence)**.** *The equation $K_\theta x = b$ has the unique solution $x^\theta = (q^\theta, d^\theta)$ equal to the discounted state-action and state occupancies. Moreover $q^\theta \geq 0$ and $d^\theta \geq 0$.*

Proof. Apply Lemma G.1 with $u = (1-\gamma)\rho$ and $v = 0$:

$$d = (1-\gamma)(I - \gamma P_\theta^\mathsf{T})^{-1}\rho, \qquad q = \Pi_\theta d.$$

The first expression is exactly the normalized discounted state occupancy and the second is the corresponding state-action occupancy. Uniqueness follows from invertibility of K_θ. The Neumann series for $(I - \gamma P_\theta^\mathsf{T})^{-1}$ is entrywise nonnegative, so $d \geq 0$, and then $q = \Pi_\theta d \geq 0$. □

Role. The proposition establishes that the square operator formulation is exactly equivalent to the usual discounted occupancy definition, rather than a relaxation.

Let

$$c_x = \begin{bmatrix} -r/(1-\gamma) \\ 0 \end{bmatrix}$$

and consider

$$\min_{\theta, x} \ \phi(\theta) + c_x^\mathsf{T} x \quad \text{s.t.} \quad K_\theta x = b.$$

Theorem G.3 (KKT equivalence)**.** *Define the reduced objective $f(\theta) = \phi(\theta) + c_x^\mathsf{T} K_\theta^{-1} b$. A triple (θ, x, y) satisfies the KKT conditions of the occupancy-constrained problem if and only if*

$$x = K_\theta^{-1} b, \qquad y = -K_\theta^{-\mathsf{T}} c_x, \qquad 0 \in \partial(f + \delta_\Theta)(\theta).$$

Proof. By Proposition G.2, feasibility identifies x with the true discounted occupancy. The feasibility and x-stationarity conditions of the constrained problem are

$$K_\theta x = b, \qquad c_x + K_\theta^\mathsf{T} y = 0.$$

Lemma G.1 gives $x = K_\theta^{-1} b$ and $y = -K_\theta^{-\mathsf{T}} c_x$. Differentiate the identity $K_\theta x^\theta = b$ in direction h:

$$D_\theta x^\theta[h] = -K_\theta^{-1} D_\theta K_\theta[h] x^\theta.$$

For the smooth reduced part $\psi(\theta) = c_x^\mathsf{T} x^\theta$,

$$\begin{aligned} D\psi(\theta)[h] &= -c_x^\mathsf{T} K_\theta^{-1} D_\theta K_\theta[h] x^\theta \\ &= \langle y^\theta, D_\theta K_\theta[h] x^\theta \rangle. \end{aligned}$$

Thus the reduced first-order condition $0 \in \partial\psi_\Theta(\theta) + \nabla\psi(\theta)$ is exactly the θ-stationarity condition of the constrained problem. This proves both directions. □

Role. This theorem verifies that the occupancy KKT residual has the same policy meaning as the reduced parameterized objective.

The deterministic occupancy-ADMM performs a proximal-gradient θ step, then

$$x^{k+1} = \arg\min_x \left\{ \mathcal{L}_\beta(\theta^{k+1}, x, y^k) + \frac{1}{2\eta_x}\|x - x^k\|^2 \right\}, \qquad y^{k+1} = y^k + \beta(K_{k+1} x^{k+1} - b).$$

Lemma G.4 (Occupancy-dual identity)**.** *The occupancy step and dual update satisfy*

$$y^{k+1} = -K_{k+1}^{-\mathsf{T}} \left(c_x + \eta_x^{-1} \Delta x_{k+1} \right).$$

Proof. The x-subproblem first-order condition is

$$0 = c_x + K_{k+1}^\mathsf{T} y^k + \beta K_{k+1}^\mathsf{T}(K_{k+1} x^{k+1} - b) + \eta_x^{-1} \Delta x_{k+1}.$$

The bracket $y^k + \beta(K_{k+1} x^{k+1} - b)$ equals y^{k+1}. Thus $K_{k+1}^\mathsf{T} y^{k+1} = -(c_x + \eta_x^{-1}\Delta x_{k+1})$, and Lemma G.1 gives the formula. □

Role. This is the occupancy analogue of the Bellman value-dual identity and is the key input to the next theorem.

Theorem G.5 (Deterministic occupancy-ADMM convergence)**.** *Assume $\theta \mapsto K_\theta$ is Lipschitz, $D_\theta K_\theta$ is bounded on the visited set, and the policy proximal-gradient and (x, y) parameters satisfy the strict margins analogous to Theorem E.9. Require the occupancy reduced objective to be bounded below, a positive lower bound on the policy steps, and a uniform Lipschitz bound for the smooth augmented policy gradients, with all constants and value-step margins taken for K_θ. Then there exists a corrected Lyapunov function*

$$\Phi_k^C = \mathcal{L}_\beta(\theta^k, x^k, y^k) + \tau\|\Delta x_k\|^2$$

with positive one-step decrease in $\|\Delta\theta_{k+1}\|^2$, $\|\Delta x_{k+1}\|^2$, and $\|\Delta x_k\|^2$. Consequently the true occupancy KKT residual converges to zero.

Proof. The proof instantiates the nonlinear-policy argument with the operator K_θ. Lemma G.1 supplies a uniform inverse bound and the inverse-map identity gives

$$\|K_\theta^{-1} - K_{\theta'}^{-1}\| \le \kappa_K^2 L_K \|\theta - \theta'\|.$$

Lemma G.4 therefore implies a dual-increment bound

$$\|\Delta y_{k+1}\|^2 \le A_C\|\Delta\theta_{k+1}\|^2 + B_C\|\Delta x_{k+1}\|^2 + B_C\|\Delta x_k\|^2.$$

The proximal-gradient policy step gives the same descent form as Lemma E.8, while the exact x step gives $(2\eta_x)^{-1}\|\Delta x_{k+1}\|^2$ decrease. The exact dual update contributes $\beta^{-1}\|\Delta y_{k+1}\|^2$ ascent. Under the stated strict margins these terms leave positive coefficients after adding $\tau\|\Delta x_k\|^2 - \tau\|\Delta x_{k+1}\|^2$.

For the lower bound, write $x^k = K_k^{-1}(b + F_k)$ with $F_k = K_k x^k - b$ and use Lemma G.4 at the preceding step. The same Young-inequality calculation as Lemma E.6 bounds Φ_k^C below by the reduced occupancy objective plus a positive multiple of $\|F_k\|^2$ and $\|\Delta x_k\|^2$. Summation forces all increments to zero. Since $\Delta y_{k+1} = \beta F_{k+1}$, feasibility follows. The policy and occupancy first-order residuals are linear in the adjacent increments because $D_\theta K_\theta$ is bounded. Hence the true occupancy KKT residual converges to zero. □

Role. The theorem establishes a second complete deterministic ADMM theory supported by discounted RL operator invertibility.

Lemma G.6 (Occupancy residual controls primal error)**.** *Let $x^\theta = K_\theta^{-1} b$. Then*

$$\|x - x^\theta\| \le \kappa_K \|K_\theta x - b\|, \qquad \kappa_K = \sup_\theta \|K_\theta^{-1}\|.$$

Proof. Subtract $K_\theta x^\theta = b$ from $K_\theta x$:

$$x - x^\theta = K_\theta^{-1}(K_\theta x - b).$$

Take norms and use the uniform inverse bound. □

Role. This error bound is used in the stochastic occupancy analysis to control the size of the occupancy variable by the true feasibility residual.

H Stochastic Occupancy-ADMM with a Shared Empirical Operator

The occupancy gradient contains $K^\mathsf{T}(Kx - b)$, and its empirical quadratic term generally satisfies $\mathbb{E}[\widehat{K}^\mathsf{T}\widehat{K}] \ne K^\mathsf{T}K$. Iteration k constructs one empirical environment matrix $\widehat{P}_k$ and uses its induced occupancy operator consistently in the coupled updates.

Assume the reset distribution and all policies used for sampling satisfy $\rho(s) \ge \rho_{\min} > 0$ and $\pi_\theta(a \mid s) \ge \pi_{\min} > 0$. Assume $m_k \ge 2$ and fresh sampling innovations as in Assumption 3.2. Use ψ_Θ from Appendix E, a lower-bounded reduced objective, known bounded rewards, and deterministic initial variables with $\theta^0 \in \operatorname{dom}\psi_\Theta$. Split a length-$m_k$ trajectory into $L_k = \lfloor m_k/2 \rfloor$ nonoverlapping two-step blocks. An anchor block is one in which the first step resets and the second step performs an environment transition.

Lemma H.1 (Regenerative transition estimator). *Define $p_0 = \gamma(1-\gamma)$ and, for each block j,*

$$Y_j(s,a,s') = \frac{\mathbf{1}\{\textit{anchor},\ S=s,\ A=a,\ S'=s'\}}{p_0\rho(s)\pi_{\theta^k}(a \mid s)}.$$

Conditioned on $\mathcal{F}_k$, the random tensors Y_j are iid and $\mathbb{E}_k Y_j(s,a,s') = P(s' \mid s,a)$. Let $\bar{P}_k = L_k^{-1}\sum_j Y_j$ and project each row of $\bar{P}_k$ onto the probability simplex to obtain $\widehat{P}_k$. Then there are constants $C_P, c_P > 0$ such that

$$\mathbb{E}_k\|\widehat{P}_k - P\|_F^2 \le \frac{C_P}{m_k}, \qquad \mathbb{P}_k(\|\widehat{P}_k - P\|_F \ge t) \le 2|\mathcal{S}|^2|\mathcal{A}|e^{-c_P m_k t^2}.$$

Proof. Different blocks use disjoint reset coins, reset states, action randomization, and environment transition noise. If the first step of a block is not a reset, the corresponding Y_j is zero, so the inherited chain state does not enter the nonzero sample. Hence the Y_j are conditionally iid.

An anchor has probability p_0. Given an anchor, $S \sim \rho$, $A \sim \pi_{\theta^k}(\cdot \mid S)$, and $S' \sim P(\cdot \mid S, A)$. Therefore

$$\mathbb{E}_k Y_j(s,a,s') = \frac{p_0\rho(s)\pi_{\theta^k}(a \mid s)P(s' \mid s,a)}{p_0\rho(s)\pi_{\theta^k}(a \mid s)} = P(s' \mid s,a).$$

Moreover $0 \le Y_j(s,a,s') \le B = [p_0\rho_{\min}\pi_{\min}]^{-1}$. Thus, for every coordinate, Hoeffding's inequality gives

$$\mathbb{P}_k(|\bar{P}_k(s' \mid s,a) - P(s' \mid s,a)| \ge u) \le 2e^{-2L_k u^2/B^2}.$$

A union bound over at most $|\mathcal{S}|^2|\mathcal{A}|$ coordinates and the implication $\|A\|_F \ge t \Rightarrow \max_i |A_i| \ge t/\sqrt{|\mathcal{S}|^2|\mathcal{A}|}$ gives the stated Frobenius tail with a modified constant c_P. Independence and bounded variance also give $\mathbb{E}_k\|\bar{P}_k - P\|_F^2 \le C/L_k \le C_P/m_k$.

Each true transition row lies in the probability simplex. Euclidean projection onto a closed convex set is nonexpansive, so rowwise $\|\Pi_\Delta(\bar{p}) - p\|_2 \le \|\bar{p} - p\|_2$. Summing over rows preserves both the mean-square and tail orders for $\widehat{P}_k$. □

Role. This estimator gives an empirical transition matrix that is both statistically controlled and exactly row stochastic on every sample path. The latter property is essential for deterministic invertibility of the empirical occupancy operator.

Let $\widehat{K}_k(\theta)$ be obtained from K_θ by replacing P with $\widehat{P}_k$. Define

$$E_k(\theta) = \widehat{K}_k(\theta) - K_\theta, \qquad \delta_k = \sup_\theta \|E_k(\theta)\|.$$

Since the environment-transition block does not depend on θ, $D_\theta E_k(\theta) = 0$. Lemma H.1 implies $\mathbb{E}_k\delta_k^2 \le C_E/m_k$.

Lemma H.2 (Empirical occupancy-operator invertibility). *For every row-stochastic $\widehat{P}_k$ and every policy, $\widehat{K}_k(\theta)$ is invertible with a deterministic bound $\sup_{k,\theta}\|\widehat{K}_k(\theta)^{-1}\| \le \bar{\kappa}_K < \infty$. Moreover*

$$\|\widehat{K}_k(\theta)^{-1} - K_\theta^{-1}\| \le \kappa_K\bar{\kappa}_K\delta_k.$$

Proof. The proof of Lemma G.1 uses only row stochasticity of the environment transition matrix. Since every row of $\widehat{P}_k$ is projected onto the simplex, the same Schur-complement argument gives invertibility of $\widehat{K}_k(\theta)$ and the same block-ℓ_1 inverse bound, hence a deterministic Euclidean bound $\bar{\kappa}_K$.

For the perturbation estimate, use

$$\widehat{K}_k^{-1} - K_\theta^{-1} = K_\theta^{-1}(K_\theta - \widehat{K}_k)\widehat{K}_k^{-1}.$$

Taking norms and using the two uniform inverse bounds gives the result. □

Role. This lemma removes any small-error invertibility event. Empirical primal-dual updates are well defined on every sample path.

Let $\widehat{F}_k(\theta, x) = \widehat{K}_k(\theta)x - b$. The stochastic occupancy-ADMM uses one shared $\widehat{K}_k$ in all updates.

Lemma H.3 (Empirical occupancy-dual identity)**.** *If the x update minimizes*

$$c_x^\mathsf{T} x + \langle y^k, \widehat{F}_k(\theta^{k+1}, x)\rangle + \frac{\beta}{2}\|\widehat{F}_k(\theta^{k+1}, x)\|^2 + \frac{1}{2\eta_x}\|x - x^k\|^2$$

and $y^{k+1} = y^k + \beta\widehat{F}_k(\theta^{k+1}, x^{k+1})$, then

$$y^{k+1} = -\widehat{K}_k(\theta^{k+1})^{-\mathsf{T}}\left(c_x + \eta_x^{-1}\Delta x_{k+1}\right).$$

Proof. The first-order condition of the x subproblem is

$$0 = c_x + \widehat{K}_k(\theta^{k+1})^\mathsf{T}\left[y^k + \beta\widehat{F}_k(\theta^{k+1}, x^{k+1})\right] + \eta_x^{-1}\Delta x_{k+1}.$$

The bracket equals y^{k+1} by the shared empirical dual update. Lemma H.2 allows inversion. □

Role. Sharing the same empirical operator preserves an exact empirical multiplier representation, which is the stochastic counterpart of the Bellman value-dual identity.

Uniform bounds and parameter conditions. Let $\bar{\kappa}_K$ bound both true and empirical inverse operators. Assume $D_\theta K_\theta$ is bounded and the smooth augmented policy gradients have a common Lipschitz constant $\overline{L}$ on all policy comparison segments. Require

$$\beta > 36\bar{\kappa}_K^2/\eta_x, \qquad \tau = 1/(8\eta_x), \qquad 0 < \underline{\eta} \le \eta_{\theta,k},$$

$$\frac{1}{2\eta_{\theta,k}} - \frac{\overline{L}}{2} - \frac{3A_C}{2\beta} \ge a_\theta > 0, \qquad A_C = 9\bar{\kappa}_K^4 L_K^2\|c_x\|^2.$$

These conditions can be met with a fixed sufficiently small policy step. A compact convex parameter set and a twice continuously differentiable policy map supply the required derivative bounds. In particular, the empirical dual identity and the dual update imply

$$\|y^{k+1}\| \le \bar{\kappa}_K\|c_x\| + \frac{\bar{\kappa}_K}{\eta_x}(\|x^{k+1}\| + \|x^k\|), \qquad \|x^{k+1}\| \le \bar{\kappa}_K\|b\| + \frac{\bar{\kappa}_K}{\beta}(\|y^{k+1}\| + \|y^k\|).$$

The two-dimensional recursion from Lemma 4.2, applied pathwise, has spectral radius less than one because $\beta\eta_x > 4\bar{\kappa}_K^2$. Thus x^k, y^k have deterministic uniform bounds for every sampling path. This also yields a uniform smoothness bound under the compact-set sufficient condition above. All subsequent two-iterate estimates apply for $k \ge 1$.

Lemma H.4 (Empirical multiplier increment)**.** *Assume $\|K_\theta - K_{\theta'}\| \le L_K\|\theta - \theta'\|$. One may take $A_C = 9\bar{\kappa}_K^4 L_K^2\|c_x\|^2$, $B_C = 3\bar{\kappa}_K^2/\eta_x^2$, and $D_C = 9\bar{\kappa}_K^4\|c_x\|^2$, so that*

$$\|\Delta y_{k+1}\|^2 \le A_C\|\Delta\theta_{k+1}\|^2 + B_C\|\Delta x_{k+1}\|^2 + B_C\|\Delta x_k\|^2 + D_C(\delta_k^2 + \delta_{k-1}^2).$$

Proof. Subtract Lemma H.3 at iterations k and $k-1$:

$$\begin{aligned}\Delta y_{k+1} = &-\left(\widehat{K}_k(\theta^{k+1})^{-\mathsf{T}} - \widehat{K}_{k-1}(\theta^k)^{-\mathsf{T}}\right)c_x \\ &- \eta_x^{-1}\widehat{K}_k(\theta^{k+1})^{-\mathsf{T}}\Delta x_{k+1} + \eta_x^{-1}\widehat{K}_{k-1}(\theta^k)^{-\mathsf{T}}\Delta x_k.\end{aligned}$$

The operator difference satisfies

$$\begin{aligned}\|\widehat{K}_k(\theta^{k+1}) - \widehat{K}_{k-1}(\theta^k)\| &\le \|K_{\theta^{k+1}} - K_{\theta^k}\| + \|E_k(\theta^{k+1})\| + \|E_{k-1}(\theta^k)\| \\ &\le L_K\|\Delta\theta_{k+1}\| + \delta_k + \delta_{k-1}.\end{aligned}$$

Apply the inverse identity with the uniform empirical inverse bound to the first term. Then use $\|a+b+c\|^2 \le 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$ and $(u+v+w)^2 \le 3(u^2+v^2+w^2)$. All fixed coefficients are absorbed into A_C, B_C, D_C. □

Role. This is the stochastic occupancy analogue of multiplier control. It enters the dual-ascent term of Lemma H.7.

Let the true residual be $F_k = K_{\theta^k} x^k - b$ and the true augmented Lagrangian be

$$\mathcal{L}_\beta^C(\theta, x, y) = \phi(\theta) + c_x^\mathsf{T} x + \langle y, K_\theta x - b\rangle + \frac{\beta}{2}\|K_\theta x - b\|^2.$$

Define

$$\Phi_k = \mathcal{L}_\beta^C(\theta^k, x^k, y^k) + \tau\|\Delta x_k\|^2, \qquad B_C(\theta) = \phi(\theta) + c_x^\mathsf{T} K_\theta^{-1} b.$$

Lemma H.5 (Perturbed true Lyapunov lower bound)**.** *For suitable τ there are $q_x > 0$ and $C_0 < \infty$ such that*

$$\Phi_k \geq B_C^{\inf} + \frac{\beta}{4}\|F_k\|^2 + q_x\|\Delta x_k\|^2 - C_0\delta_{k-1}^2.$$

Therefore

$$\Psi_k = \Phi_k - B_C^{\inf} + 1 + C_0\delta_{k-1}^2$$

is nonnegative and satisfies $\Psi_k \geq 1 + (\beta/4)\|F_k\|^2 + q_x\|\Delta x_k\|^2$.

Proof. Write $x^k = K_k^{-1}(b + F_k)$. Lemma H.3 at the previous iteration gives

$$y^k = -\widehat{K}_{k-1}(\theta^k)^{-\mathsf{T}}(c_x + \eta_x^{-1}\Delta x_k).$$

Substitute these into the true augmented Lagrangian. The reduced objective contributes $B_C(\theta^k) \geq B_C^{\inf}$. The terms involving Δx_k and F_k have the same form as in Lemma E.6. The additional mismatch is produced by $\widehat{K}_{k-1}(\theta^k)^{-1} - K_{\theta^k}^{-1}$, whose norm is at most $\kappa_K \bar{\kappa}_K \delta_{k-1}$ by Lemma H.2. Thus, for fixed constants,

$$\Phi_k \geq B_C^{\inf} + \frac{\beta}{2}\|F_k\|^2 - C_1\|\Delta x_k\|\|F_k\| - C_2\delta_{k-1}\|F_k\| + \tau\|\Delta x_k\|^2.$$

Apply Young's inequality separately to the two cross terms, allocating at most $\beta\|F_k\|^2/8$ to each. This leaves $\beta\|F_k\|^2/4$, a coefficient $q_x = \tau - C_1'$ on $\|\Delta x_k\|^2$, and $-C_0\delta_{k-1}^2$. Choosing τ above C_1' gives $q_x > 0$. Adding the correcting terms in the definition of Ψ_k proves nonnegativity. □

Role. This lemma supplies a nonnegative stochastic Lyapunov variable despite empirical-operator perturbations and provides the $\sqrt{\Psi_k}$ bound used in the next lemma.

Lemma H.6 (Empirical policy-gradient perturbation)**.** *Let*

$$g_{\theta,k} = D_\theta F(\theta^k, x^k)^\mathsf{T}(y^k + \beta F_k)$$

be the true smooth augmented policy gradient and let $\widehat{g}_{\theta,k}$ be the corresponding empirical gradient. If $\sup\|D_\theta K_\theta\| \leq G_K$, then

$$\|\widehat{g}_{\theta,k} - g_{\theta,k}\| \leq C\delta_k\sqrt{\Psi_k}.$$

Consequently, for every $\epsilon > 0$,

$$|\langle \widehat{g}_{\theta,k} - g_{\theta,k}, \Delta\theta_{k+1}\rangle| \leq \epsilon\|\Delta\theta_{k+1}\|^2 + C_\epsilon\delta_k^2\Psi_k.$$

Proof. The error and policy derivative have the block forms

$$E_k x = \begin{bmatrix} -\gamma(\widehat{P}_k - P)^\mathsf{T} q \\ 0 \end{bmatrix}, \qquad D_\theta F(\theta, x)[h] = \begin{bmatrix} 0 \\ -D_\theta\Pi_\theta[h]d \end{bmatrix}.$$

Also $D_\theta E_k = 0$. Hence their inner product is zero for every direction h, and

$$\widehat{g}_{\theta,k} - g_{\theta,k} = \beta D_\theta F(\theta^k, x^k)^\mathsf{T} E_k x^k = 0.$$

Both displayed bounds follow. This cancellation is specific to the block structure of the occupancy constraint. □

Role. This lemma is the key benefit of sharing one empirical operator. It converts stochastic policy-gradient error into a multiplicative Lyapunov perturbation without any unbiased-product assumption.

Lemma H.7 (True one-step stochastic descent)**.** *Under the uniform bounds and parameter conditions stated above, for $k \geq 1$ there are constants $c_\theta, c_x, c'_x, a, b > 0$ such that*

$$\mathbb{E}_k \Psi_{k+1} \leq \left(1 + \frac{a}{m_k}\right) \Psi_k - c_\theta \mathbb{E}_k \|\Delta\theta_{k+1}\|^2 - c_x \mathbb{E}_k \|\Delta x_{k+1}\|^2 - c'_x \|\Delta x_k\|^2 + \frac{b}{m_k}.$$

Proof. The policy gradient is unchanged by the transition perturbation, by Lemma H.6. The true policy step therefore decreases the augmented Lagrangian by at least $(1/(2\eta_{\theta,k}) - \overline{L}/2)\|\Delta\theta_{k+1}\|^2$.

On the value comparison segment, deterministic boundedness of x, y and uniform boundedness of $K, \widehat{K}$ imply

$$\|\nabla_x \widehat{\mathcal{L}}_\beta - \nabla_x \mathcal{L}_\beta\| = \|E_k^\mathsf{T}(y + \beta F) + \beta K^\mathsf{T} E_k x + \beta E_k^\mathsf{T} E_k x\| \leq C\delta_k.$$

Transferring the empirical value-step decrease by Young's inequality gives $(4\eta_x)^{-1}\|\Delta x_{k+1}\|^2 - C\delta_k^2$. The true dual ascent satisfies

$$\langle \Delta y_{k+1}, F_{k+1} \rangle \leq \frac{3}{2\beta}\|\Delta y_{k+1}\|^2 + C\delta_k^2,$$

since x^{k+1} is uniformly bounded. Substituting Lemma H.4 and adding the potential correction yields

$$\Phi_k - \Phi_{k+1} \geq c_\theta \|\Delta\theta_{k+1}\|^2 + c_x \|\Delta x_{k+1}\|^2 + c'_x \|\Delta x_k\|^2 - C_1\delta_k^2 - C_2\delta_{k-1}^2,$$

where

$$c_\theta \geq a_\theta, \quad c_x = \frac{1}{4\eta_x} - \frac{3B_C}{2\beta} - \tau > 0, \quad c'_x = \tau - \frac{3B_C}{2\beta} > 0.$$

In Lemma H.5, increase C_0 to satisfy $C_0 \geq C_2$. With $c = \min\{c_\theta, c_x, c'_x\}$, the corrected nonnegative potential then obeys the pathwise bound

$$\Psi_{k+1} \leq \Psi_k - cD_{k+1} + C\delta_k^2. \tag{25}$$

Conditioning gives $\mathbb{E}_k \Psi_{k+1} \leq \Psi_k - c\mathbb{E}_k D_{k+1} + C/m_k$. Since $\Psi_k \geq 0$, this also implies the stated inequality with any fixed $a > 0$ after enlarging b. □

Role. This is the stochastic occupancy Lyapunov recursion. Theorem H.9 and Theorem H.11 are direct consequences.

Define the true squared KKT residual

$$G_k^{\mathrm{true}} = e_{\theta,k}^2 + e_{x,k}^2 + \|F_k\|^2,$$

and let $D_{k+1} = \|\Delta\theta_{k+1}\|^2 + \|\Delta x_{k+1}\|^2 + \|\Delta x_k\|^2$.

Lemma H.8 (True KKT residual bound)**.** *For $k \geq 1$, there is $C_G < \infty$ such that*

$$\mathbb{E}_k G_{k+1}^{\mathrm{true}} \leq C_G \mathbb{E}_k D_{k+1} + \frac{C_G}{m_k}\Psi_k + \frac{C_G}{m_k} + C_G\delta_{k-1}^2.$$

Proof. Use the composite policy Fermat condition, the equality of true and empirical policy gradients, the uniform policy smoothness bound, and $\eta_{\theta,k}^{-1} \leq \underline{\eta}^{-1}$. They give

$$e_{\theta,k+1} \leq C(\|\Delta\theta_{k+1}\| + \|\Delta x_{k+1}\| + \|\Delta y_{k+1}\|).$$

The empirical dual identity and update give

$$c_x + K_{k+1}^\mathsf{T} y^{k+1} = -\eta_x^{-1}\Delta x_{k+1} - E_k^\mathsf{T} y^{k+1}, \qquad F_{k+1} = \beta^{-1}\Delta y_{k+1} - E_k x^{k+1}.$$

Uniform boundedness of x^{k+1}, y^{k+1} and Lemma H.4 imply the pathwise estimate

$$G_{k+1}^{\mathrm{true}} \leq C\{D_{k+1} + \delta_k^2 + \delta_{k-1}^2\}.$$

Taking $\mathcal{F}_k$-conditional expectations leaves the previous error δ_{k-1}^2 unchanged and bounds the current error by C/m_k. This is stronger than the displayed estimate because $\Psi_k \geq 0$. □

Role. This lemma translates Lyapunov descent into the true optimization residual and is the final input to the stochastic complexity theorem.

Let $A_T = \sum_{k=0}^{T-1} m_k^{-1}$.

Theorem H.9 (Finite-time stochastic occupancy theory)**.** *There are constants $C_1, C_2, C_3, C_{\mathrm{init}}$ independent of the batch schedule such that*

$$\sup_{1\le k\le T} \mathbb{E}\Psi_k \le e^{C_1 A_T}(C_{\mathrm{init}} + C_2 A_T).$$

If K is uniform on $\{1,\dots,T\}$ and independent of the algorithmic randomness, then

$$\mathbb{E}G_K^{\mathrm{true}} \le C_3 e^{C_1 A_T}(1+A_T)\left(\frac{1}{T} + \frac{A_T}{T}\right).$$

Proof. The first-step comparison with θ^0 and the deterministic iterate bounds give uniform upper bounds on $\mathbb{E}\Psi_1$ and $\mathbb{E}G_1^{\mathrm{true}}$. Absorb these into C_{init}. Summing the additive bound (25) in expectation and using $\Psi_k \ge 0$ yields

$$\sup_{1\le k\le T} \mathbb{E}\Psi_k \le C_{\mathrm{init}} + CA_T, \qquad \sum_{k=1}^{T-1} \mathbb{E}D_{k+1} \le C(1+A_T).$$

The pathwise residual estimate in Lemma H.8 has two consecutive operator errors. Their expected sum is at most $2CA_T$. Including the first residual gives

$$\frac{1}{T}\sum_{j=1}^{T} \mathbb{E}G_j^{\mathrm{true}} \le C(1+A_T)/T.$$

This proves the stated bounds after increasing constants, since the exponential and additional $(1 + A_T)$ factors are at least one. Uniform randomization identifies the left side with $\mathbb{E}G_K^{\mathrm{true}}$. □

Role. This theorem gives the finite-time guarantee for the stochastic occupancy operator and verifies that the operator-stability mechanism survives a second, explicit RL formulation.

Corollary H.10 (Equal-batch complexity)**.** *If $m_k = m = N/T$, then $A_T = T^2/N$. If $N \ge c_0 T^2$ for a fixed $c_0 > 0$, then*

$$\mathbb{E}G_K^{\mathrm{true}} = O\left(\frac{1}{T} + \frac{T}{N}\right).$$

For an unsquared KKT target ϵ, it is sufficient to take $T = O(\epsilon^{-2})$ and $N = O(\epsilon^{-4})$.

Proof. Equal batches give $A_T = T/m = T^2/N$. If $N \ge c_0 T^2$, then $A_T \le 1/c_0$, so the exponential and $(1 + A_T)$ factors in Theorem H.9 are bounded by constants. Moreover $A_T/T = T/N$, proving the rate. Jensen gives $\mathbb{E}\sqrt{G_K^{\mathrm{true}}} \le \sqrt{\mathbb{E}G_K^{\mathrm{true}}}$. Requiring the squared residual to be $O(\epsilon^2)$ yields $T = O(\epsilon^{-2})$ and $N = O(\epsilon^{-4})$. □

Role. The corollary puts the stochastic occupancy theorem on the same squared-versus-unsquared residual scale as the main Bellman complexity result.

Theorem H.11 (Almost-sure stochastic occupancy convergence)**.** *If $\sum_k m_k^{-1} < \infty$, then Ψ_k converges almost surely,*

$$\sum_k D_{k+1} < \infty \quad \textit{almost surely},$$

$\delta_k \to 0$, and $G_k^{\mathrm{true}} \to 0$ almost surely.

Proof. Lemma H.1 and Tonelli's theorem give $\sum_k \delta_k^2 < \infty$ almost surely. Fix such a path. From (25),

$$\Psi_{k+1} + C\sum_{j\ge k+1} \delta_j^2 \le \Psi_k + C\sum_{j\ge k} \delta_j^2 - cD_{k+1}.$$

The corrected sequence is nonnegative and decreasing. It converges and $\sum_k D_{k+1} < \infty$. Since the error tails vanish, Ψ_k converges as well. The pathwise residual estimate in Lemma H.8 then gives $G_k^{\mathrm{true}} \to 0$. □

Role. This theorem is the asymptotic stochastic counterpart of Theorem G.5.

Corollary H.12 (Conditional nonlinear-policy performance)**.** *Assume in addition a parameter-space domination inequality*

$$J^\star - J(\pi_\theta) \le C_{\mathrm{par}} R_\theta(\theta), \qquad R_\theta(\theta)^2 \le C_G G^{\mathrm{true}}.$$

Then

$$\mathbb{E}[J^\star - J(\pi_{\theta_K})] \le C\sqrt{\mathbb{E}G_K^{\mathrm{true}}}.$$

If instead $J^\star - J(\pi_\theta) \le C R_\theta(\theta)^2$, then the performance gap is $O(\mathbb{E}G_K^{\mathrm{true}})$.

Proof. The first two inequalities imply $J^\star - J(\pi_{\theta_K}) \le C\sqrt{G_K^{\mathrm{true}}}$. Taking expectations and applying Jensen proves the first claim. Under quadratic parameter-space domination, $J^\star - J(\pi_{\theta_K}) \le CC_G G_K^{\mathrm{true}}$, and expectation preserves the squared-residual rate. □

Role. This corollary states the domination conditions for performance guarantees with nonlinear policies.

I Additional Experimental Results

This section provides additional numerical evidence for the structural predictions studied in Sections 8. All experiments use finite discounted MDPs with direct tabular policies and the same reset-controlled Markov sampling mechanism as in the theoretical analysis. Unless stated otherwise, results are averaged over multiple independently generated MDP instances and trajectory seeds, and error bars denote 95% confidence intervals.

The experiments are designed to isolate four components of the theory: (i) the statistical contribution of the Markov batch size, (ii) the correlated residual–Jacobian product created by shared sampling, (iii) the conversion from KKT residual to policy performance, and (iv) the numerical sensitivity to the ADMM penalty parameter. The corresponding main structural tests for shared-operator stability and finite-time scaling are reported in Section 8.

I.1 Batch-Size Sensitivity

The finite-time result predicts

$$\mathbb{E}[\widetilde{G}_{K+1}] \le \frac{A}{T} + \frac{B}{T}\sum_{k<T}\frac{1}{m_k}.$$

For a constant batch size $m_k = m$, this reduces to

$$O(T^{-1} + m^{-1}).$$

We therefore fix the outer iteration budget at $T = 240$ and vary $m \in \{60, 120, 240, 480, 960\}$.

Figure 2 shows that the mean companion squared KKT residual decreases monotonically with the batch size. A log–log fit over the tested range gives slope approximately -0.87, close to the m^{-1} statistical contribution predicted by the finite-time bound. The deviation from an exact -1 slope is expected because the bound contains the additional A/T term, which produces a finite-iteration floor as m increases.

I.2 Residual–Jacobian Product under Shared Sampling

A central difficulty in the stochastic analysis is that the Bellman residual and its Jacobian are constructed from the same empirical operator. Writing

$$e = \widehat{R} - R, \qquad G = \widehat{J} - J,$$

the stochastic policy-gradient perturbation contains the correlated term

$$\beta G^\top e.$$

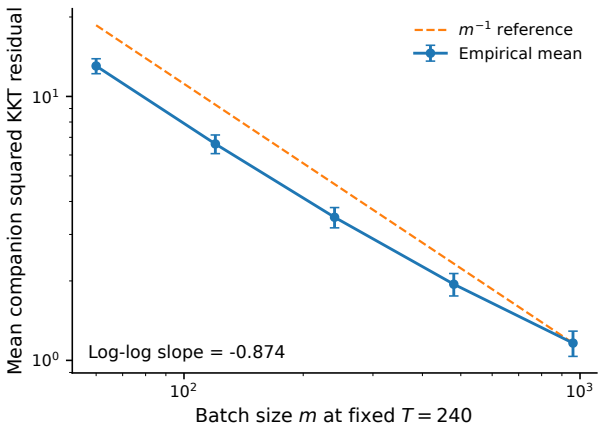


Figure 2: **Batch-size sensitivity.** Mean companion squared KKT residual at fixed $T = 240$ as the Markov batch size increases. The dashed line shows an m^{-1} reference slope.

Separate unbiasedness of G and e does not imply that this product has zero mean. Remark 4.5 instead controls the product pathwise through the shared empirical operator error:

$$\|G^\top e\| = O(\epsilon_k^2).$$

Since the empirical operator satisfies $\mathbb{E}[\epsilon_k^2] = O(m^{-1})$, the resulting perturbation is predicted to occur at the same m^{-1} scale.

To isolate this effect, we fix a primal–dual point and repeatedly construct empirical Bellman operators with different batch sizes. We compare two estimators. The *same-batch* estimator uses one empirical operator for both G and e, matching the algorithm. The *independent-batch* estimator constructs them from separate samples, thereby suppressing the correlation at the cost of double sampling.

As shown in Figure 3, the same-batch product decays with empirical log–log slope approximately -1.13, consistent with the $O(m^{-1})$ prediction. Independent batches produce a substantially smaller mean product term, as expected from decorrelation. The latter, however, no longer represents the single shared empirical operator used by the algorithm. Together with the exact value–dual identity experiment in the main text, this result illustrates the role of structure-preserving stochasticization: the algorithm retains the coupled empirical operator and controls the induced product perturbation rather than removing it through double sampling.

I.3 From KKT Residual to Policy Performance

Section 6 establishes that, under discounted occupancy coverage, the optimization residual has a direct RL interpretation. In particular, the policy stationarity residual is controlled by the KKT residual, and the policy-performance gap consequently satisfies

$$J^\star - J(\pi) = O(\sqrt{G}).$$

We examine this relationship along the iterates generated by the Bellman-ADMM dynamics.

For every sampled iterate, we compute the true KKT residual G using the underlying MDP model and evaluate the corresponding policy return exactly. Figure 4 plots $J^\star - J(\pi)$ against $\sqrt{G}$. The points exhibit the predicted first-order relation throughout the observed regime. The empirical 95th percentile of

$$\frac{J^\star - J(\pi)}{\sqrt{G}}$$

is 2.72, while the ratio remains bounded across the collected iterates. This experiment complements the optimization-based residual results by showing that decreasing KKT error is accompanied by decreasing policy suboptimality.

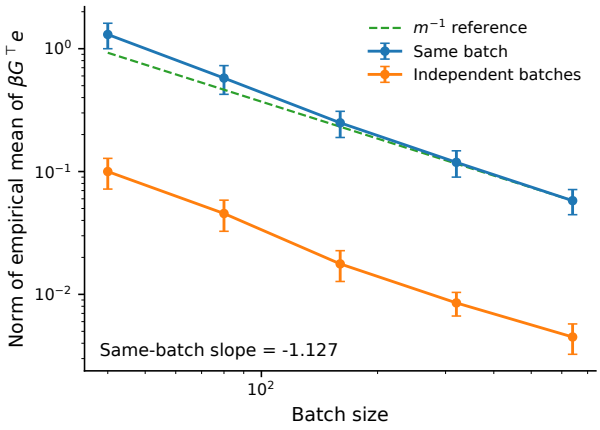


Figure 3: **Residual–Jacobian product scaling.** Magnitude of the empirical product term $\beta G^\top e$ as the batch size increases. The same-batch construction follows approximately the predicted m^{-1} scale. Independent batches reduce the correlation but require separate samples.

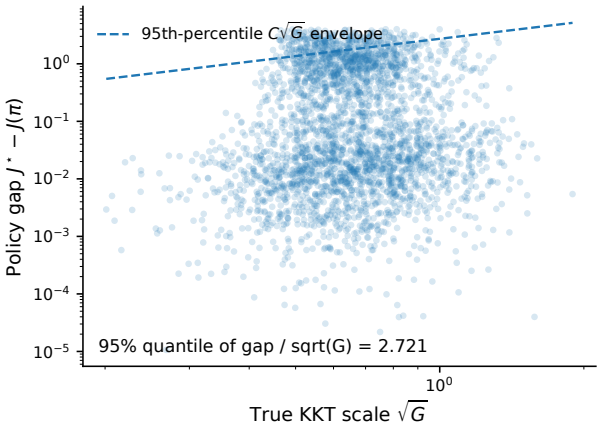


Figure 4: **KKT residual and policy performance.** Policy suboptimality versus the true KKT scale $\sqrt{G}$. The dashed line denotes the empirical 95th-percentile $C\sqrt{G}$ envelope.

I.4 Sensitivity to the Penalty Parameter

The convergence analysis imposes a sufficient lower bound on the ADMM penalty parameter β to guarantee positivity of the coefficients in the Lyapunov descent argument. Such analytical margins are designed to hold uniformly over the admissible problem class and need not coincide with empirically optimal parameter choices.

We therefore vary

$$\beta \in \{2.5, 5, 10, 20, 40\}$$

while keeping the remaining algorithmic parameters fixed. As shown in Figure 5, the method remains numerically stable over the entire tested range. The trailing KKT residual changes smoothly with β, with smaller penalties yielding lower residuals on these instances. This behavior is consistent with the role of the theoretical condition as a uniform sufficient stability margin rather than a tuning prescription for individual MDPs.

I.5 Operator Consistency: Experimental Design

The following ablations separate the role of empirical-operator consistency from the precision of each empirical model and the total sampling cost. We compare three constructions: *all shared*,

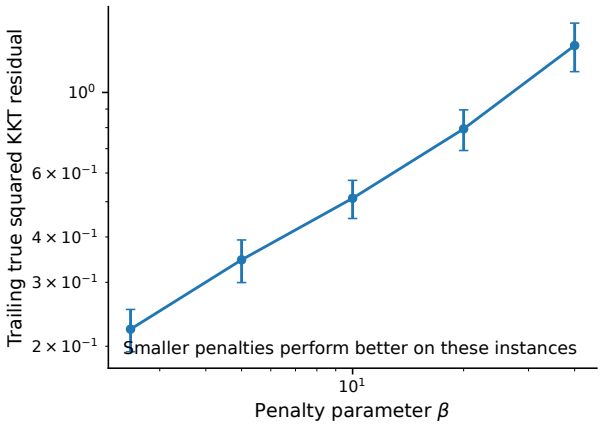


Figure 5: **Penalty sensitivity.** Trailing true squared KKT residual as a function of the ADMM penalty parameter β.

value–dual shared, and *all separate*. Table 2 specifies the model used by each update. The value–dual-shared construction isolates the consistency needed by the value–dual identity, while retaining an independently estimated model for the policy update.

Table 2: Empirical-model assignments. Distinct indices denote separately sampled models. Under an equal per-model batch size m, the last column gives the total number of reset-controlled sampling steps per iteration.

Construction	Policy	Value	Dual	Total steps
All shared	0	0	0	m
Value–dual shared	0	1	1	$2m$
All separate	0	1	2	$3m$

Instances and parameters. Each MDP has five states and two actions. Every transition row is sampled independently from $\mathrm{Dirichlet}(0.7\mathbf{1}_5)$. Rewards are independent $\mathcal{N}(0, 0.45^2)$ draws plus statewise action trends. For states $s = 0, \ldots, 4$, these trends are $0.30 - 0.125s$ for action 0 and $-0.35 + 0.20s$ for action 1. The generated transition kernel and mean rewards remain fixed during each run. The reset, training, and evaluation distributions are uniform, $\rho = \mu = \nu = \mathbf{1}_5/5$, with $c = -\mu$ and $\phi \equiv 0$. Unless specified otherwise, $\gamma = 0.85$, $\beta = 10$, and $\eta_\pi = \eta_V = 0.15$. We initialize π^0 uniformly and set $V^0 = \lambda^0 = 0$. Each run uses 180 outer iterations. The MDP seeds are $3, 7, 11, 19, 23, 31, 43, 59$, and each instance uses trajectory seeds $0, 1, 2, 3, 4$.

Markov sampling and budget accounting. At iteration k, the behavior policy is frozen at $\bar{\pi}^k = 0.7\pi^k + 0.3\mathbf{1}_2/2$. Each step resets to ρ with probability $1-\gamma$ and otherwise samples an environment transition. Environment rewards are observed with independent $\mathcal{N}(0, 0.3^2)$ noise. Only non-reset transitions contribute to the empirical transition and reward rows. Unvisited rows use the uniform transition distribution and reward zero. Each distinct model has a separate random stream and a persistent chain state, initialized at state 0. Corresponding streams use matched random seeds across constructions. With adaptive behavior, subsequent observations can differ as the policies evolve. A fixed-behavior control below removes this dependence.

We vary $m \in \{80, 160, 320, 640\}$ under two budget conventions. For *equal total budget*, m reset-controlled steps are divided among one, two, or three models, with remainders assigned to the earliest model indices. For *equal per-model batch*, every distinct model receives m steps, giving total costs m, $2m$, and $3m$. Reset steps are included in these costs. Counts of environment transitions and resets are also recorded separately. All three constructions in this study use the same stream protocol. The original two-way experiment uses serial blocks of one trajectory and assigns $\lfloor m/3 \rfloor$ steps to each separate model. The new controls use exactly m total steps in the equal-budget comparison.

Exact subproblem solutions. The two-action policy subproblem separates over states. Write $x_s^k = \pi^k(1 \mid s)$, $q_{sa} = \widehat{r}_k^p(s,a) + \gamma \sum_{s'} \widehat{P}_k^p(s' \mid s,a) V^k(s')$, $d_s = q_{s1} - q_{s0}$, and $b_s = V^k(s) - q_{s0}$, where the superscript p denotes the policy model. Its exact solution is

$$x_s^{k+1} = \mathrm{proj}_{[0,1]} \frac{d_s(\lambda_s^k + \beta b_s) + 2x_s^k/\eta_\pi}{\beta d_s^2 + 2/\eta_\pi}. \tag{26}$$

Let $\widehat{M}_k^v$ and $\widehat{r}_k^v$ denote the value model evaluated at π^{k+1}. The value update solves

$$\left[\beta(\widehat{M}_k^v)^\top \widehat{M}_k^v + \eta_V^{-1} I\right] V^{k+1} = -c - (\widehat{M}_k^v)^\top \lambda^k + \beta(\widehat{M}_k^v)^\top \widehat{r}_k^v + \eta_V^{-1} V^k. \tag{27}$$

The coefficient matrix is positive definite. These exact solves ensure that subproblem approximation does not enter the comparison.

Metrics and statistical aggregation. We evaluate the true squared KKT residual

$$G(\pi, V, \lambda) = \mathrm{dist}^2\big(0, N_\Pi(\pi) + J_\pi(\pi, V)^\top \lambda\big) + \|c + M_\pi^\top \lambda\|^2 + \|R(\pi, V)\|^2 \tag{28}$$

using the underlying MDP. Policy return is evaluated by an exact Bellman solve, and $J^\star$ is obtained by enumerating the 2^5 deterministic policies. We also record $\|\lambda^{k+1} - \lambda^k\|^2$, the return gap $J^\star - J(\pi^{k+1})$, and the identity defect defined below. Unless stated otherwise, each reported outcome averages iterations 151–180. We first average the five trajectory seeds within each MDP, then average the eight MDP means. Figure intervals are 95% percentile bootstrap intervals from 10,000 resamples of these eight means. Paired comparisons use the eight within-MDP differences and a 95% Student-t interval with seven degrees of freedom. These are individual, unadjusted intervals.

I.6 Identifying the Value–Dual Consistency Mechanism

Let $\widehat{R}_k^v$ and $\widehat{R}_k^d$ denote the empirical residuals used by the value and dual updates. In this subsection, both are evaluated at (π^{k+1}, V^{k+1}), and $\widehat{M}_k^v$ is evaluated at π^{k+1}. The exact value solve and the multiplier update give

$$0 = c + (\widehat{M}_k^v)^\top \lambda^k + \beta(\widehat{M}_k^v)^\top \widehat{R}_k^v + \eta_V^{-1} \Delta V_{k+1}, \tag{29}$$

$$\lambda^{k+1} = \lambda^k + \beta \widehat{R}_k^d. \tag{30}$$

Substitution yields the pathwise relation

$$\underbrace{c + (\widehat{M}_k^v)^\top \lambda^{k+1} + \eta_V^{-1} \Delta V_{k+1}}_{h_{k+1}} = \beta(\widehat{M}_k^v)^\top (\widehat{R}_k^d - \widehat{R}_k^v). \tag{31}$$

Thus, sharing the value and dual models sets the right-hand side to zero, regardless of the model used in the policy update. This identity explains why both all-shared and value–dual-shared updates retain an explicit multiplier representation. All sharing also reuses the same empirical model for the policy update, reducing the number of sampled models.

We measure the identity defect $D_{k+1} = \|h_{k+1}\|_2$ and verify Equation (31) at every update. At equal per-model batch size 320, the mean trailing defects are 1.05×10^{-14} for both sharing constructions and 1.793 for all separate updates. Figure 6 shows the same distinction across batch sizes. Across the full experiment suite, including the large-penalty check below, the maximum norm of the difference between the two sides of Equation (31) is 2.02×10^{-10}. The value–dual-shared construction serves as a mechanism ablation. The convergence results in the main text apply to the specified all-shared algorithm.

I.7 Separating Operator Consistency from Sampling Cost

Figure 7 compares true squared KKT residuals, squared multiplier increments, and return gaps under both budget conventions. Table 3 reports the batch-320 slice. At equal per-model batch size, the mean residual is 0.503 for all shared and 4.159 for all separate, a ratio of 8.27. All sharing uses one third of the total sampling steps in this comparison. Value–dual sharing attains residual 0.639, placing its optimization behavior closer to all sharing while preserving the same multiplier identity.

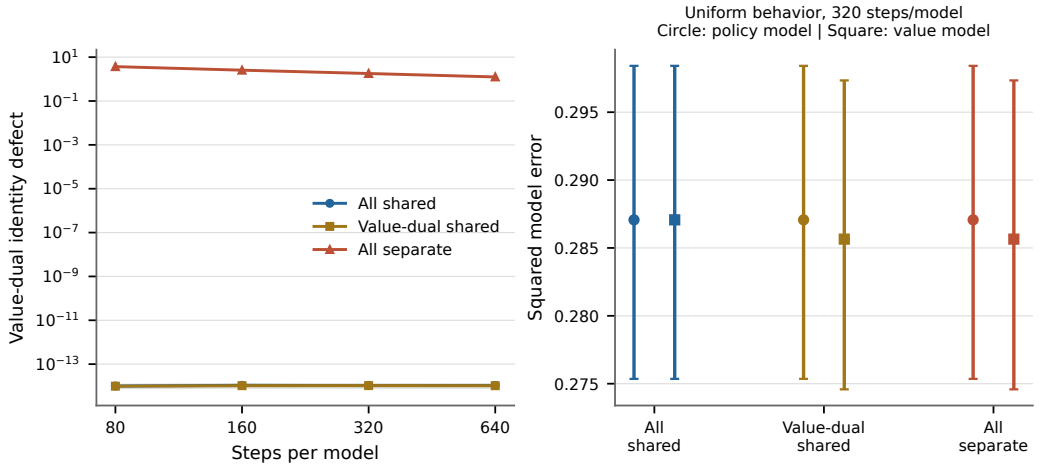


Figure 6: **Identity preservation and estimator precision.** Left: trailing identity defect under equal per-model batches and adaptive behavior. All-shared and value–dual-shared curves overlap at numerical precision. Right: empirical-model squared error under fixed uniform behavior and 320 steps per model. Circles denote policy models and squares denote value models. Model error is measured as the sum of squared Frobenius transition error and squared Euclidean reward error. Error intervals resample MDP means.

Table 3: **Batch-320 comparison.** Entries average the last 30 of 180 iterations and then the eight MDP means. G is the true squared KKT residual, Δ_λ is the squared multiplier increment, and $\Delta_J = J^\star - J(\pi)$ is the return gap. Total cost includes reset steps.

Budget	Construction	Steps	G	Δ_λ	Δ_J
Equal total	All shared	320	0.5028	1.0452	0.01425
	Value–dual shared	320	1.3813	2.2538	0.05144
	All separate	320	13.1920	12.8006	0.17656
Equal per model	All shared	320	0.5028	1.0452	0.01425
	Value–dual shared	640	0.6390	1.1283	0.02419
	All separate	960	4.1585	4.0575	0.04168

Under equal total budget, the corresponding residuals are 0.503, 1.381, and 13.192. These complementary comparisons demonstrate the contribution of consistency at matched model batch sizes and the sampling benefit of reusing one model.

The paired differences in Table 4 quantify variation across instances. At equal per-model batch size 320, all separate exceeds all shared in squared KKT residual by 3.656, with interval $[2.617, 4.694]$. Its squared multiplier increment exceeds all shared by 3.012, with interval $[2.213, 3.812]$. The all-shared versus value–dual-shared distinction is supported for the residual and multiplier increment. Their return-gap difference has an interval containing zero. We therefore use this comparison to locate the operator-consistency mechanism and quantify its sampling cost.

I.8 Fixed-Behavior Control

To separate operator consistency from changes in data collection, we repeat the comparisons with $\bar{\pi}^k(a \mid s) = 1/2$ at every iteration. At equal per-model batch size 320, the policy model is identical across all three constructions under the matched streams. The value model is also identical between value–dual shared and all separate. All shared reuses its policy model for the value and dual updates. Each distinct model follows the same estimation procedure and sampling distribution. This control removes the feedback from the learned policy to the sampling distribution.

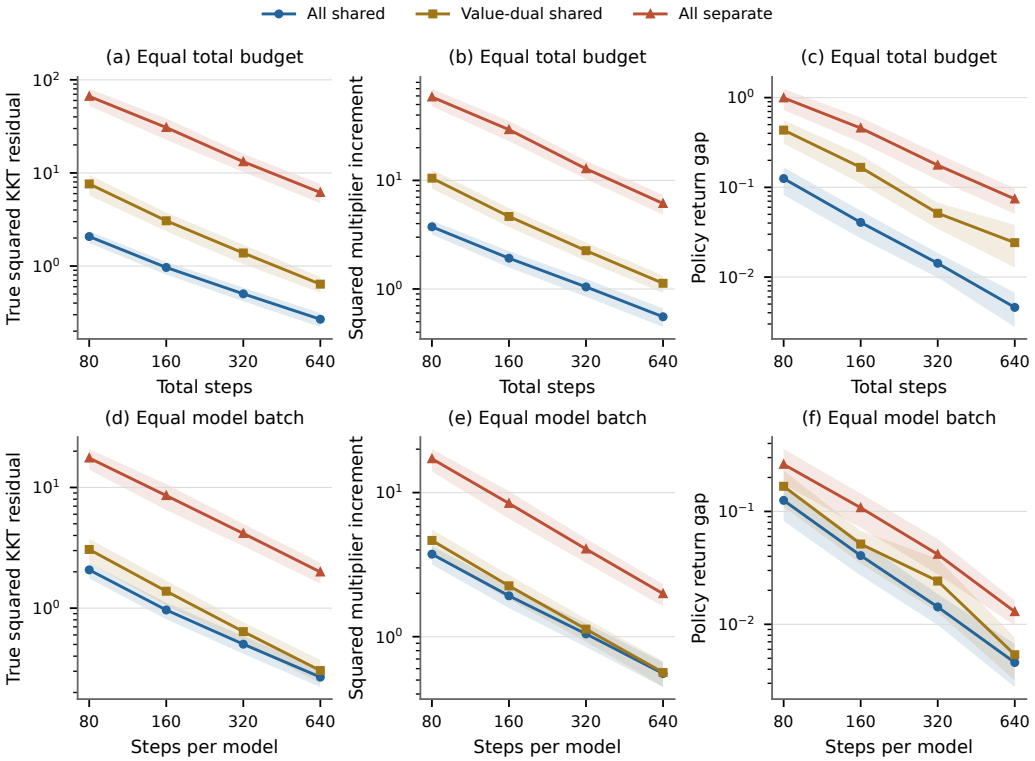


Figure 7: **Consistency and sampling-budget controls.** Top: equal total steps per iteration. Bottom: equal steps per empirical model, with total costs m, $2m$, and $3m$ for all shared, value–dual shared, and all separate. Columns show the true squared KKT residual, squared multiplier increment, and return gap. Curves average the last 30 of 180 iterations. Shading denotes 95% bootstrap intervals over MDP means.

Table 4: **Paired contrasts at equal per-model batch size 320.** Each contrast subtracts the second construction from the first. Intervals are based on eight paired MDP means.

Metric	Contrast	Mean difference	95% interval
G	Value–dual shared − all shared	0.1361	$[0.0463, 0.2260]$
	All separate − all shared	3.6557	$[2.6170, 4.6944]$
	All separate − value–dual shared	3.5196	$[2.5340, 4.5052]$
Δ_λ	Value–dual shared − all shared	0.0831	$[0.0205, 0.1457]$
	All separate − all shared	3.0123	$[2.2128, 3.8118]$
	All separate − value–dual shared	2.9293	$[2.1526, 3.7060]$
Δ_J	Value–dual shared − all shared	0.00994	$[-0.00384, 0.02372]$
	All separate − all shared	0.02742	$[0.01143, 0.04342]$
	All separate − value–dual shared	0.01748	$[0.01011, 0.02486]$

Table 5 retains the same ordering of the three residuals. The paired residual difference between all separate and all shared is 6.151, with 95% interval $[4.013, 8.288]$. Between all separate and value–dual shared it is 5.854, with interval $[3.882, 7.825]$. Thus, the residual separation persists when the behavior policy and per-model sampling distributions are fixed.

I.9 Discounted-Resolvent Sensitivity

The inverse-map argument predicts that the sensitivity of the dual representation depends on the discounted resolvent. To examine this dependence with fixed estimation error, we collect one 320-

Table 5: **Uniform-behavior control at 320 steps per model.** Total costs are 320, 640, and 960 steps, respectively. The notation and aggregation match Table 3.

Construction	G	Δ_λ	Δ_J
All shared	0.8505	1.7776	0.02818
Value–dual shared	1.1477	1.8908	0.04018
All separate	7.0014	6.7681	0.07335

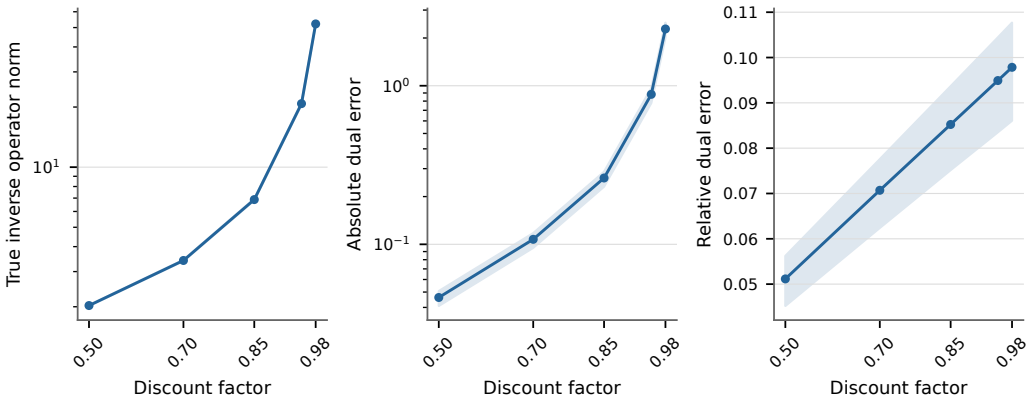


Figure 8: **Discount sensitivity with fixed policy and transition error.** One empirical transition model per run is reused across all discounts. Panels show the true resolvent norm and the absolute and relative errors in the stationary dual representation. Shading denotes 95% bootstrap intervals over MDP means.

step batch using uniform behavior and sampling discount 0.85 for each MDP and trajectory seed. We then hold both the empirical transition matrix and the uniform evaluation policy fixed while varying $\gamma \in \{0.50, 0.70, 0.85, 0.95, 0.98\}$. Write $M_\gamma = I - \gamma P_\pi$ and $\widehat{M}_\gamma = I - \gamma \widehat{P}_\pi$. The resolvent identity gives

$$\|\widehat{M}_\gamma^{-1} - M_\gamma^{-1}\|_2 \le \gamma \|\widehat{M}_\gamma^{-1}\|_2 \|\widehat{P}_\pi - P_\pi\|_2 \|M_\gamma^{-1}\|_2. \tag{32}$$

We evaluate the stationary dual representations $\lambda_\gamma = M_\gamma^{-\top}\mu$ and $\widehat{\lambda}_\gamma = \widehat{M}_\gamma^{-\top}\mu$. Figure 8 reports the true inverse norm, the absolute dual error, and the relative error $\|\widehat{\lambda}_\gamma - \lambda_\gamma\|_2 / \|\lambda_\gamma\|_2$. As γ increases from 0.50 to 0.98, the mean inverse norm rises from 2.03 to 52.15, the absolute dual error from 0.0462 to 2.2808, and the relative error from 0.0512 to 0.0978. The relative metric accounts for the change in discounted occupancy mass, $\mathbf{1}^\top \lambda_\gamma = (1-\gamma)^{-1}$. These fixed-perturbation measurements support the sensitivity mechanism used in multiplier control without estimating a worst-case asymptotic rate.

We also run the three algorithms at the same discount grid using uniform behavior and 320 steps per model, holding the remaining parameters fixed. Figure 9 reports these dynamic comparisons. Here, the reset probability is $1-\gamma$, so the experiment includes both operator sensitivity and changes in the sampling process. All-shared updates have lower mean KKT residuals than all-separate updates throughout this grid. The fixed-perturbation experiment above isolates the inverse-operator component of this dependence.

I.10 Parameter-Margin Check and Numerical Verification

The preceding finite-budget comparisons use $\beta = 10$. A separate parameter-margin check uses the uniform bounds $\bar{\kappa} = \sqrt{5}/(1-\gamma)$ and $L_M = \gamma\sqrt{2}$. At $\gamma = 0.85$ and $\eta_\pi = \eta_V = 0.15$, the sufficient lower bound in the main text is

$$\beta > \max\left\{\frac{60\bar{\kappa}^2}{\eta_V}, 30\eta_\pi \bar{\kappa}^4 L_M^2 \|c\|^2\right\} = 88888.889. \tag{33}$$

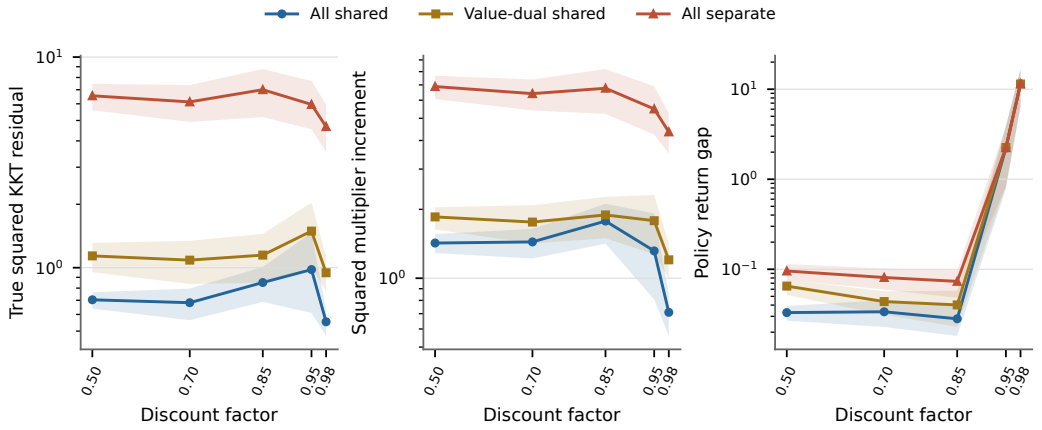


Figure 9: **Dynamic discount sweep.** Uniform behavior and 320 steps per model, with all other algorithmic parameters fixed. The reset probability changes with the discount. Outcomes average the last 30 of 180 iterations, and shaded intervals resample MDP means.

Table 6: **Large-penalty check.** Uniform behavior, 320 steps per model, and $\beta = 97777.778$. All quantities use the same trailing aggregation as the other controls. D denotes the identity defect.

Construction	G	Δ_λ	Δ_J	D
All shared	6.735	76.779	2.034	2.22×10^{-11}
Value–dual shared	266.470	1813.805	2.177	3.13×10^{-11}
All separate	5.12×10^{8}	8.33×10^{8}	2.210	1.82×10^{4}

We set $\beta = 97777.778$, which is 1.1 times this threshold, and run all three constructions with uniform behavior and 320 steps per model. The three coefficients in the all-shared Lyapunov descent bound are 0.57197, 0.07576, and 0.07576, respectively. Table 6 reports the corresponding finite-time outcomes. This check concerns the penalty margin of the all-shared theory. The main comparisons assess finite-budget behavior, and the fixed-batch runs do not invoke the square-summability condition for almost-sure asymptotic convergence.

The suite contains 960 budget-comparison runs, 240 uniform-behavior runs, 600 discount-sweep runs, and 120 parameter-margin runs, totaling 1920 algorithm runs. The fixed-perturbation study adds 200 evaluation points. Some settings coincide across studies and are not treated as additional independent replicates in the statistical analysis. Numerical verification reproduces the 80 original batch-320 runs to a maximum absolute discrepancy of 3.55×10^{-15} across the recorded summary metrics. Independent scalar optimization verifies the policy subproblem objective to 1.07×10^{-14}, and the checked value solves have first-order residual at most 6.44×10^{-15}. Configurations, per-run results, selected per-iteration traces, and the analysis scripts are retained with the experiment package.