LEARNING TO SWITCH WITH GUARANTEES: GRADIENT-BASED OPTIMIZATION OF STATE-INDEPENDENT CONTROLLER SWITCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

The design of feedback controllers plays a central role in determining how dynamical systems behave under closed-loop operation. A single fixed controller, however, may not provide the desired behavior for every performance objective. Addressing different requirements by repeatedly redesigning the controller can be difficult, especially when the design space is continuous. When a finite set of pre-designed controllers is available, the problem can be reduced to deciding how these controllers should be selected over the operating time horizon, motivating controller switching. However, the key challenge of switching is to improve closed-loop performance while maintaining stability. This is especially important in online learning, where an unstable update can destabilize the system during training. To address this, we develop `StableSwitch-PG`, a policy-gradient framework that optimizes a state-independent controller-selection distribution while preserving mean-square stability throughout learning for linear time-invariant systems. For known dynamics, we derive the exact policy gradient and establish global convergence of the objective to the optimal randomized switching policy. For unknown dynamics, we estimate the gradient from sampled trajectories and establish asymptotic and finite-time convergence guarantees. Finally, we derive a Lyapunov-based update condition that preserves mean-square stability at every learning iteration.

1 INTRODUCTION

Switching among multiple feedback laws is a well-established way of obtaining closed-loop behaviors that may be difficult to achieve with a single fixed controller. Appropriate switching can improve stabilization, accommodate different operating regimes, and exploit complementary properties of available controllers. This has motivated extensive study of switched and hybrid systems, where the behavior of the overall system depends not only on the individual closed-loop modes, but also on how transitions between those modes occur.

Different switching mechanisms arise depending on how the active mode is selected. Under *arbitrary switching*, the mode may change without restriction, whereas dwell-time and related constrained-switching rules restrict how frequently switching may occur. In *state-dependent switching*, the active mode is chosen according to the current system state, whereas under *stochastic switching* it evolves according to a probability law, including Markovian and independently randomized switching. Regardless of how the switching signal is generated, a fundamental question is whether repeated transitions among subsystems preserve stability of the overall switched system. This motivated much of the early literature, which developed common and multiple Lyapunov-function methods for stability analysis (Peleties & DeCarlo (1991); Branicky (1998)). Since requiring stability under arbitrary switching can be restrictive, Hespanha & Morse (1999) introduced the average-dwell-time framework, which guarantees stability when switching occurs sufficiently slowly. The focus subsequently expanded from analyzing a prescribed switching signal to designing switching laws that stabilize the overall system (Liberzon & Morse (1999)), while later works consolidated and extended stability and stabilizability results for switched and hybrid systems (Shorten et al. (2007); Lin & Antsaklis (2009)). For stochastic switching, the corresponding analysis led to probabilistic notions such as almost-sure and mean-square stability (Bolzern et al. (2006)). More recently,

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.

the switching mechanism itself has increasingly been treated as a design variable rather than as a prescribed process. For example, Wu et al. (2023) synthesize mode-dependent probabilistic switching policies for switched linear systems using model-based optimization while enforcing stability guarantees.

The switching mechanism considered in this paper is *controlled, stochastic, state-independent, and identically distributed*. At every time step, one controller is sampled from the same probability distribution, independently of the current state and the previously active controller.

Once the switching mechanism becomes a design variable, a natural question is whether it can be learned directly from closed-loop performance rather than computed entirely through model-based synthesis. This connects switched control with reinforcement learning over an existing collection of controllers or policies. Perkins & Barto (2002) considered safe reinforcement learning over a collection of controllers constructed using Lyapunov information. Banijamali et al. (2019) studied optimization over mixtures of given base policies, while Agarwal et al. (2020) developed a boosting framework for combining weak controllers. More closely related to our setting, Zaki et al. (2022) optimize a single state-independent distribution over a collection of given controllers using policy-gradient and actor-critic methods, with a controller sampled independently from this distribution at each time step. Our learning problem retains this controller-library viewpoint but specializes it to linear state-feedback control. The gains are fixed, and the learning variable is the probability distribution governing how often the corresponding closed-loop modes are selected.

This specialization brings the problem into direct contact with policy-gradient theory for linear control. For linear quadratic regulation, Fazel et al. (2018) established global convergence properties of direct policy optimization. Related work developed convergence theory for softmax and more general policy-gradient methods (Mei et al. (2020); Agarwal et al. (2021)). Bhandari & Russo (2024) further characterized structural conditions for global optimality of policy-gradient methods. Li et al. (2021) provided a complementary finite-time analysis of softmax policy-gradient convergence. Policy-gradient analysis has also been extended to Markovian jump linear quadratic systems; for example Jansch-Porto et al. (2023) establish global convergence while retaining mean-square stability from a stabilizing initialization. The optimization variable in these linear-control formulations is different from ours. In LQR, the feedback gain is optimized directly, while in Markovian jump linear quadratic control the mode-transition probabilities are given and the mode-dependent feedback gains are optimized. In contrast, in our setting, the feedback gains are fixed and the controller-selection probabilities themselves are the decision variables. Consequently, changing the policy changes not only how frequently each controller is selected, but also the distribution of future states and the second-moment dynamics of the switched system.

A remaining issue is that improving the policy does not by itself guarantee stability during learning. Safe reinforcement learning has addressed this problem through constrained policy updates (Achiam et al. (2017); Tessler et al. (2019)), Lyapunov-based constraints for safe policy improvement (Chow et al. (2018)), state augmentation for almost-sure safety (Sootla et al. (2022)), and control-theoretic stability certificates (Berkenkamp et al. (2017); Marvi & Kiumarsi (2022)). These works ensure safety of trajectories, states, or learned control policies. It is different in our setting where the learned variable is the controller-switching distribution itself, and we require the resulting switched linear system to remain mean-square stable after every policy update.

Existing switched-system methods provide stability analysis and model-based switching synthesis; controller-library methods provide mechanisms for selecting or combining existing policies; policy-gradient methods provide convergence theory for directly optimized control policies; and safe reinforcement learning provides tools for maintaining safety during policy improvement. What is missing is a framework that directly optimizes a single state-independent probability distribution, applied identically at every time step over a fixed library of linear feedback controllers, while simultaneously providing optimization guarantees and preserving mean-square stability as that distribution is updated.

Contributions. We develop such a framework for discrete-time linear systems. First, for known dynamics, we exploit the switched linear-quadratic structure to derive explicit matrix representations of the value function and discounted state occupancy, an exact policy-gradient expression, and a performance-difference identity. Under the stated structural and regularity conditions, these results establish global convergence of exact gradient descent, with an $O(1/k)$ rate under bounded

softmax logit spread. Second, when the dynamics are not used to compute the gradient, we derive an unbiased likelihood-ratio estimator with respect to the trajectory distribution induced by the controller-selection policy, using sampled closed-loop trajectories and establish almost-sure convergence together with a finite-time best-iterate guarantee. Finally, we distinguish finiteness of the discounted objective from mean-square stability and derive a Lyapunov-based bound on each policy update that preserves mean-square stability of the controller-selection distribution throughout learning.

2 Problem Statement and Notation

Consider the discrete-time linear time-invariant system $x_{t+1} = Ax_t + Bu_t$, where $x_t \in \mathbb{R}^n$ is the state, $u_t \in \mathbb{R}^{n_u}$ is the control input, $A \in \mathbb{R}^{n\times n}$ is the state matrix, and $B \in \mathbb{R}^{n\times n_u}$ is the input matrix. The initial state x_0 is assumed to be random with finite positive definite second-moment matrix $\Sigma_0 := \mathbb{E}[x_0 x_0^\top] \succ 0$. We assume that a finite library of pre-designed linear state-feedback controllers $\mathcal{K} := \{K_1, \dots, K_M\}$ is available. The controller gains are fixed and are not redesigned during learning.

Let $\sigma : \mathbb{Z}_{\geq 0} \to \{1, \dots, M\}$ denote the switching signal. The value $\sigma(t) = m$ indicates that controller K_m is active at physical time t, so that $u_t = -K_m x_t$. Defining the corresponding closed-loop matrix $F_m := A - BK_m$, the system evolves as

$$x_{t+1} = F_{\sigma(t)} x_t. \tag{1}$$

Thus the learning problem considered here does not optimize the feedback gains themselves; it optimizes how the available closed-loop modes are selected over time. We associate a scalar parameter $\theta_m \in \mathbb{R}$ with each controller and collect them in $\theta = [\theta_1, \dots, \theta_M]^\top \in \mathbb{R}^M$. The controller-selection probabilities are parameterized by

$$p_\theta(m) := \frac{\exp(\theta_m)}{\sum_{j=1}^M \exp(\theta_j)}, \qquad m \in \{1, \dots, M\}. \tag{2}$$

For a fixed parameter θ, the switching signal is sampled independently according to $\sigma(t) \overset{\text{i.i.d.}}{\sim} p_\theta$. Hence $\mathbb{P}_\theta(\sigma(t) = m \mid x_t) = p_\theta(m)$: the same distribution over controllers is used at every state and every physical time. We use t for physical system time and k for policy-learning iterations.

Performance objective. Let $Q \succeq 0$ and $R \succ 0$ denote the state and input weighting matrices. The one-step quadratic cost is $\ell(x_t, u_t) = x_t^\top Q x_t + u_t^\top R u_t$. When controller m is active, $u_t = -K_m x_t$, and therefore $\ell(x_t, u_t) = x_t^\top C_m x_t$, where $C_m := Q + K_m^\top R K_m$. Let $\gamma \in (0, 1)$ be a discount factor. The performance of the switching distribution p_θ is measured by

$$J(\theta) := \mathbb{E}_{x_0,\, \sigma(\cdot)\sim p_\theta} \left[\sum_{t=0}^{\infty} \gamma^t x_t^\top C_{\sigma(t)} x_t \right]. \tag{3}$$

The expectation is over both the initial state and the random switching sequence.

Discounted admissibility. Because equation 3 is infinite horizon, we first identify the switching policies for which the discounted cost is finite. For a fixed p_θ, let $\Sigma_t := \mathbb{E}_\theta[x_t x_t^\top]$. Independence of $\sigma(t)$ from x_t gives $\Sigma_{t+1} = \sum_m p_\theta(m) F_m \Sigma_t F_m^\top$. Define the second-moment transition matrix $\mathcal{M}_\theta := \sum_{m=1}^M p_\theta(m)(F_m \otimes F_m)$ which gives $\operatorname{vec}(\Sigma_{t+1}) = \mathcal{M}_\theta \operatorname{vec}(\Sigma_t)$, where $\otimes$ denotes the matrix Kronecker product (A.1). The discounted second moments therefore evolve through $(\gamma \mathcal{M}_\theta)^t$. This motivates the discounted-admissible set

$$\mathcal{D}_\gamma := \left\{ \theta \in \mathbb{R}^M : \rho(\gamma \mathcal{M}_\theta) < 1 \right\}, \tag{4}$$

where $\rho(\cdot)$ denotes spectral radius. For a general controller distribution p on the probability simplex, define $\mathcal{M}_p := \sum_m p(m)(F_m \otimes F_m)$ and let $p^\star$ denote an admissible globally optimal distribution, with $J^\star := J(p^\star)$. The condition $\rho(\gamma \mathcal{M}_p) < 1$ concerns finiteness of the *discounted* objective.

Relation to standard policy gradient. Our policy parameterization differs from the standard policy-gradient setting. In conventional policy-gradient methods, the learned policy is state dependent: given the current state x_t, the policy directly determines the control action or a distribution over actions, e.g., $u_t \sim \pi_\theta(\cdot \mid x_t)$. In our setting, the control laws are already fixed. The learner instead chooses which controller to apply according to $\mathbb{P}_\theta(\sigma(t) = m \mid x_t) = p_\theta(m)$, after which the selected controller produces the input $u_t = -K_m x_t$. Thus θ does not parameterize the state-to-control mapping; it parameterizes a single probability distribution over the given controller library that is used for all states. Nevertheless, changing $p_\theta(m)$ changes how often the closed-loop matrix $F_m = A - BK_m$ governs the system and, consequently, changes the distribution of future states. This coupling between the controller-selection probabilities and the resulting state distribution is the main feature that must be accounted for in the policy-gradient analysis.

3 Policy Optimization with Exact Gradients

We first consider the case in which the closed-loop matrices $\{F_m\}_{m=1}^M$ are known. Our goal is to determine whether gradient descent on the softmax parameters can converge to the globally optimal state-independent randomized switching policy. This is not immediate, even when the dynamics are known. The objective is generally nonconvex, and changing the probability of selecting a controller also changes the future states visited by the system. In addition, under the softmax parameterization, a controller assigned a very small probability may also have a very small gradient component. Therefore, standard convex-optimization arguments cannot be applied directly. We instead exploit the structure of the switched linear system to derive the policy gradient and establish global convergence.

3.1 Value Representation and Exact Gradient

For a fixed switching distribution, define $V_\theta(x) := \mathbb{E}_{\sigma(\cdot)\sim p_\theta}\left[\sum_{t=0}^\infty \gamma^t x_t^\top C_{\sigma(t)} x_t \mid x_0 = x\right]$.

Lemma 1 (Quadratic value representation)**.** *For every $\theta \in \mathcal{D}_\gamma$, $V_\theta(x) = x^\top P_\theta x$, where $P_\theta \succeq 0$ is the unique solution of*

$$P_\theta = \sum_{m=1}^M p_\theta(m)\left(C_m + \gamma F_m^\top P_\theta F_m\right). \tag{5}$$

Moreover, $J(\theta) = \mathrm{Tr}(P_\theta \Sigma_0)$.

Lemma 1 shows that, although the objective is defined over an infinite trajectory, its value can be represented by the finite-dimensional matrix P_θ. The matrix relation in equation 5 therefore summarizes the expected future quadratic cost under the current switching distribution.

therefore summarizes the expected future quadratic cost under the current switching distribution. The identity $J(\theta) = \mathrm{Tr}(P_\theta \Sigma_0)$ then connects this value representation directly to the optimization objective.

The discounted-admissibility condition is also what provides the regularity required by gradient-based optimization. Indeed, letting $\bar{C}_\theta := \sum_m p_\theta(m) C_m$, we get $\mathrm{vec}(P_\theta) = (I - \gamma \mathcal{M}_\theta^\top)^{-1} \mathrm{vec}(\bar{C}_\theta)$. Both $\bar{C}_\theta$ and $\mathcal{M}_\theta$ are analytic functions of θ, while $I - \gamma \mathcal{M}_\theta^\top$ is nonsingular on $\mathcal{D}_\gamma$. Hence J is real analytic on $\mathcal{D}_\gamma$. On any compact convex $\Theta \subset \mathcal{D}_\gamma$ contained inside this domain, continuity of the Hessian implies that J is L-smooth for some finite L. Detailed proofs are deferred to the appendix.

To characterize the gradient, define the discounted state-occupancy matrix (A.4),

$$X_\theta := \sum_{t=0}^\infty \gamma^t \Sigma_t, \qquad X_\theta = \Sigma_0 + \gamma \sum_{m=1}^M p_\theta(m) F_m X_\theta F_m^\top. \tag{6}$$

For controller m, define the advantage matrix

$$A_{\theta,m} := C_m + \gamma F_m^\top P_\theta F_m - P_\theta. \tag{7}$$

The quantity $x^\top A_{\theta,m} x$ compares selecting controller m at state x with following the current switching distribution.

Theorem 1 (Exact policy gradient)**.** *For every $\theta \in \mathcal{D}_\gamma$ and $m \in \{1, \dots, M\}$,*

$$\frac{\partial J(\theta)}{\partial \theta_m} = p_\theta(m) \operatorname{Tr}\left(X_\theta A_{\theta,m}\right). \tag{8}$$

Theorem 1 exposes two distinct contributions. The term $\operatorname{Tr}(X_\theta A_{\theta,m})$ measures the advantage of controller m under the current state occupancy, whereas $p_\theta(m)$ reflects how frequently the controller is selected. Thus an advantageous controller can still generate a small gradient component if its probability becomes very small.

3.2 Global Convergence

The exact gradient in Theorem 1 describes how the objective changes locally around the current switching policy. To establish global convergence, we must additionally relate this local information to the cost difference between the current policy and a globally optimal switching distribution. For any distribution q satisfying the discounted-admissibility condition in equation equation 4, i.e., $\rho(\gamma \mathcal{M}_q) < 1$, let X_q denote the corresponding discounted state-occupancy matrix.

Lemma 2 (Performance-difference identity)**.** *For any discounted-admissible switching distribution q,*

$$J(\theta) - J(q) = -\operatorname{Tr}\left[X_q \sum_{m=1}^{M} q(m) A_{\theta,m}\right]. \tag{9}$$

Lemma 2 compares the current policy with any other switching distribution using the advantage matrices defined with respect to the current policy. In particular, choosing $q = p^\star$ gives

$$J(\theta) - J^\star = \operatorname{Tr}\left(X_{p^\star} \Gamma_\theta\right), \qquad \Gamma_\theta := -\sum_{m=1}^{M} p^\star(m) A_{\theta,m}. \tag{10}$$

Thus the global optimality gap is expressed in terms of the discounted occupancy $X_{p^\star}$ of the optimal switching distribution. However, the policy gradient in Theorem 1 is determined by the current occupancy X_θ. Therefore, in order to relate the global cost gap to the policy gradient, we first need to compare $X_{p^\star}$ with X_θ.

Proposition 1 (Occupancy comparison)**.** *Suppose $\Sigma_0 \succ 0$ and $p^\star$ is discounted admissible, so that $X_{p^\star}$ is finite. Then, for every discounted-admissible θ, there exists a finite constant $\bar{\kappa} > 0$ such that*

$$X_{p^\star} \preceq \bar{\kappa} X_\theta. \tag{11}$$

Note that $\bar{\kappa} \le \frac{\lambda_{\max}(X_{p^\star})}{\lambda_{\min}(\Sigma_0)}$. Proposition 1 allows us to compare the optimal and current occupancies. However, a matrix inequality of the form $X_{p^\star} \preceq \bar{\kappa} X_\theta$ can be transferred to the optimality-gap trace only when the matrix multiplying the occupancies is positive semidefinite. We therefore make the following structural assumption.

Assumption 1 (Optimal-mixture advantage positivity)**.** *For every iterate θ under consideration,*

$$\Gamma_\theta = -\sum_{m=1}^{M} p^\star(m) A_{\theta,m} \succeq 0. \tag{12}$$

Assumption 1 has a direct cost interpretation. For any state x, the quantity $x^\top A_{\theta,m} x$ measures the change in cost-to-go obtained by selecting controller m relative to continuing with the current switching policy. Since the objective is minimized, $x^\top A_{\theta,m} x < 0$ corresponds to an improvement. Therefore $\Gamma_\theta \succeq 0$ is equivalent to $\sum_{m=1}^{M} p^\star(m) x^\top A_{\theta,m} x \le 0$. In other words, when the controllers are averaged according to the optimal switching distribution, their advantage relative to the current policy is nonpositive in every state direction. Importantly, this condition is imposed only on the weighted optimal mixture. In fact, because $\sum_{m=1}^{M} p_\theta(m) A_{\theta,m} = 0$ (A.5), the individual

advantage matrices cannot all be negative semidefinite unless $A_{\theta,m} = 0$ for every m. Thus, Assumption 1 requires only the weighted combination defined by the optimal distribution $p^\star$ to be negative semidefinite in the corresponding advantage sense. With Assumption 1, the occupancy comparison in proposition 1 can now be used inside the trace. Since $\Gamma_\theta \succeq 0$,

$$J(\theta) - J^\star \le \bar{\kappa} \operatorname{Tr}(X_\theta \Gamma_\theta) = \bar{\kappa} \left[- \sum_{m=1}^{M} p^\star(m) \operatorname{Tr}(X_\theta A_{\theta,m}) \right]. \tag{13}$$

Equation 13 expresses the global optimality gap using quantities evaluated under the current policy. To relate this bound to the policy gradient, recall from Theorem 1 that $[\nabla J(\theta)]_m = p_\theta(m) \operatorname{Tr}(X_\theta A_{\theta,m})$. Therefore, $\operatorname{Tr}(X_\theta A_{\theta,m}) = \frac{[\nabla J(\theta)]_m}{p_\theta(m)}$. Hence, substituting the policy gradient into equation 13 requires division by $p_\theta(m)$. This becomes problematic if some controller is assigned a probability arbitrarily close to zero, since the factor $1/p_\theta(m)$ can become arbitrarily large.

We therefore need to control how small the softmax probabilities can be. Define the logit spread $\Delta(\theta) := \max_m \theta_m - \min_m \theta_m$. For the softmax parameterization (A.12),

$$\min_m p_\theta(m) \ge \frac{1}{1 + (M-1)e^{\Delta(\theta)}}. \tag{14}$$

Equivalently, $\frac{1}{p_\theta(m)} \le 1 + (M-1)e^{\Delta(\theta)}$. Thus the logit spread directly controls the factor introduced when the advantage terms in equation 13 are written in terms of the policy gradient. Combining equation 13 with equation 14 yields the following global gradient-dominance relation.

Theorem 2 (Gradient dominance)**.** *Suppose Assumption 1 holds and $\Sigma_0 \succ 0$. Define*

$$D(\theta) := \bar{\kappa} \left[1 + (M-1)e^{\Delta(\theta)} \right]. \tag{15}$$

Then

$$J(\theta) - J^\star \le D(\theta) \|\nabla J(\theta)\|_2. \tag{16}$$

Theorem 2 shows that the current policy cannot remain far from the global optimum while its gradient becomes arbitrarily small, provided $D(\theta)$ remains controlled. The two factors in $D(\theta)$ have different meanings: $\bar{\kappa}$ measures the mismatch between the optimal and current state occupancies, while the exponential term captures the loss of gradient information caused by increasingly concentrated softmax probabilities. We now use this relation to establish convergence of exact policy gradient. Consider $\theta_{k+1} = \theta_k - \eta \nabla J(\theta_k)$, and define $\delta_k := J(\theta_k) - J^\star$ and $D_k := D(\theta_k)$.

Theorem 3 (Convergence of exact policy gradient)**.** *Suppose the iterates remain in a region on which J is L-smooth, Assumption 1 holds, $\Sigma_0 \succ 0$, and $0 < \eta \le 1/L$. Then*

$$\delta_{k_\mathrm{f}} \le \left[\delta_0^{-1} + \frac{\eta}{2} \sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2} \right]^{-1}. \tag{17}$$

Consequently, $J(\theta_k) \to J^\star$ whenever $\sum_{k=0}^{\infty} D_k^{-2} = \infty$.

Theorem 3 follows from two ingredients. First, L-smoothness ensures that a sufficiently small gradient step produces a decrease in the objective. Second, Theorem 2 guarantees that the gradient cannot become arbitrarily small while the current policy remains strictly suboptimal. Together, these properties force the objective value toward $J^\star$. A particularly useful case occurs when the softmax logits do not separate without bound. Suppose $\Delta(\theta_k) \le \bar{\Delta}$ for every iteration k. Then $D_k \le \bar{D} := \bar{\kappa} \left[1 + (M-1)e^{\bar{\Delta}} \right]$, so the gradient-dominance constant is uniformly bounded along the entire optimization trajectory. Substituting this bound into Theorem 3 gives

$$J(\theta_{k_\mathrm{f}}) - J^\star \le \frac{1}{\delta_0^{-1} + \eta k_\mathrm{f} / (2\bar{D}^2)} = \mathcal{O}(k_\mathrm{f}^{-1}). \tag{18}$$

Thus, when the logit spread remains bounded, the cost approaches the optimal value at rate $\mathcal{O}(1/k_\mathrm{f})$. This result concerns convergence of the objective value $J(\theta_k)$, rather than convergence of θ_k to a unique parameter vector, since multiple softmax parameter vectors can represent the same controller-selection distribution.

Remark 1. Since the smoothness constant L is generally nontrivial to compute explicitly, we use a diminishing learning-rate sequence satisfying $\sum_{k=0}^{\infty}\eta_k=\infty$ and $\sum_{k=0}^{\infty}\eta_k^2<\infty$. Because $\eta_k\to 0$, the learning rate eventually becomes sufficiently small so that $\eta_k\le 1/L$, even when L is not known.

4 Policy Learning from Sampled Trajectories

The previous section assumes that the closed-loop matrices F_m are available for evaluating P_θ, X_θ, and the exact gradient. We now remove this requirement from the *gradient computation*.

At physical time t, the learner observes x_t, the selected controller $\sigma(t)$, the applied input u_t, the realized cost $c_t:=x_t^\top Qx_t+u_t^\top Ru_t$, and the next state x_{t+1}. For a trajectory generated under p_θ, define the return $G_t:=\sum_{s=t}^{\infty}\gamma^{s-t}c_s$. For softmax, $\nabla\log p_\theta(m)=e_m-p_\theta$, where e_m is the mth canonical basis vector.

Theorem 4 (Trajectory-based policy gradient)**.** *For every policy in the discounted-admissible region,*

$$\nabla J(\theta)=\mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[\sum_{t=0}^{\infty}\gamma^t G_t\left(e_{\sigma(t)}-p_\theta\right)\right]. \tag{19}$$

Theorem 4 expresses the same policy gradient entirely through sampled trajectories; A, B, and F_m are not required to construct the estimator. At learning iteration k, let N independent trajectories begenerated using p_{θ_k}. For trajectory i, define $Z_k^{(i)}:=\sum_{t=0}^{\infty}\gamma^t G_t^{(i)}\left(e_{\sigma^{(i)}(t)}-p_{\theta_k}\right)$ and $\widehat{g}_k:=\frac{1}{N}\sum_{i=1}^{N}Z_k^{(i)}$. Conditioned on the information $\mathcal{F}_k$ available before the trajectory batch is generated, the estimator $\widehat{g}_k$ is unbiased: $\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\widehat{g}_k\mid\mathcal{F}_k\right]=\nabla J(\theta_k)$. Thus, on average, the sampled update direction agrees with the exact policy gradient. Individual trajectory estimates, however, can deviate substantially from $\nabla J(\theta_k)$. To control the effect of this sampling noise on the learning dynamics, we impose a finite-variance condition.

Assumption 2 (Finite trajectory-gradient variance)**.** *There exists a finite constant σ_{tr}^2 such that, for every learning iteration k and trajectory i,*

$$\mathbb{E}_{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}}\left[\left\|Z_k^{(i)}-\nabla J(\theta_k)\right\|_2^2\,\middle|\,\mathcal{F}_k\right]\le\sigma_{\mathrm{tr}}^2. \tag{20}$$

Assumption 2 bounds the variability of a single trajectory-gradient estimate around the exact gradient. Since $\widehat{g}_k$ averages N independent trajectory estimates, the variance is reduced by the batch size (B.4):

$$\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\left\|\widehat{g}_k-\nabla J(\theta_k)\right\|_2^2\,\middle|\,\mathcal{F}_k\right]\le\frac{\sigma_{\mathrm{tr}}^2}{N}. \tag{21}$$

Hence, increasing N produces a more accurate approximation of the exact policy-gradient direction. Using this estimator, the model-free policy update is $\theta_{k+1}=\theta_k-\eta_k\widehat{g}_k$, where $\eta_k>0$ is the learning rate. Unlike exact gradient descent, the direction $\widehat{g}_k$ contains sampling error at every iteration. Therefore, convergence requires the learning rate to decrease sufficiently slowly to continue making progress, but sufficiently fast to prevent the accumulated gradient noise from dominating the updates.

Theorem 5 (Convergence with sampled policy gradients)**.** *Suppose the iterates remain in a region where J is L-smooth, Assumptions 1 and 2 hold, and $D(\theta_k)\le\bar{D}<\infty$. If*

$$\eta_k=\frac{a}{(k+1)^\alpha},\qquad 0<a\le\frac{1}{L},\qquad \frac{1}{2}<\alpha\le 1,$$

then

$$J(\theta_k)\longrightarrow J^\star\qquad\textit{almost surely}. \tag{22}$$

Moreover, defining

$$U_{k_{\mathrm{f}}}:=\sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k,\qquad W_{k_{\mathrm{f}}}:=\sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k^2,$$

the iterates satisfy

$$\min_{0\le k<k_{\mathrm{f}}}\mathbb{E}[\delta_k^2]\le\frac{2\bar{D}^2\delta_0}{U_{k_{\mathrm{f}}}}+\frac{\bar{D}^2L\sigma_{\mathrm{tr}}^2}{N}\frac{W_{k_{\mathrm{f}}}}{U_{k_{\mathrm{f}}}}. \tag{23}$$

Theorem 5 separates the two effects governing learning. The first term in equation 23 reflects optimization progress and decreases as more gradient steps are taken. The second term reflects the error introduced by sampling and decreases as the batch size N increases. Thus, larger batches provide more accurate gradient estimates, while diminishing learning rates suppress the accumulated effect of gradient noise over time. The finite-time guarantee in equation 23 is a best-iterate bound: among the first k_{f} iterates, at least one achieves the stated expected squared optimality gap. It is not a last-iterate guarantee.

The sampled-gradient analysis therefore shows that the switching policy can be optimized without using the system matrices in the gradient computation. It does not, however, guarantee that every intermediate policy encountered during learning is mean-square stable. We address this stability requirement next.

5 Stability-Preserving Policy Learning

The discounted condition $\rho(\gamma\mathcal{M}_\theta) < 1$ guarantees that the optimization objective is finite, but it does not imply physical mean-square stability because $\gamma < 1$. It is therefore possible for a policy to have finite discounted cost while the undiscounted second moment does not converge to zero.

5.1 Mean-Square Stability

For a fixed switching distribution p_θ, the state second moment satisfies $\mathrm{vec}(\Sigma_{t+1}) = \mathcal{M}_\theta \,\mathrm{vec}(\Sigma_t)$.

Theorem 6 (Mean-square stability of a fixed policy)**.** *The randomized switched system is mean-square stable (MSS) if*

$$\rho(\mathcal{M}_\theta) = \rho\left(\sum_{m=1}^{M} p_\theta(m)(F_m \otimes F_m)\right) < 1. \tag{24}$$

In particular, $\lim_{t\to\infty} \mathbb{E}\|x_t\|_2^2 = 0$.

Theorem 6 characterizes a fixed policy. During learning, however, the parameter changes from θ_k to θ_{k+1}. Even when p_{θ_k} is Mean-Square Stable (MSS), an unrestricted cost-reducing update need not preserve this property. We therefore seek a sufficient bound on each parameter update. Let $p_k := p_{\theta_k}$ be MSS. Then there exists $S_k \succ 0$ satisfying the stochastic Lyapunov equation

$$S_k - \sum_{m=1}^{M} p_k(m) F_m^\top S_k F_m = \mathcal{Q}_k, \qquad \mathcal{Q}_k \succ 0. \tag{25}$$

The corresponding quadratic function $\mathcal{V}_k(x) := x^\top S_k x$ satisfies $\mathbb{E}[\mathcal{V}_k(x_{t+1}) \mid x_t] - \mathcal{V}_k(x_t) = -x_t^\top \mathcal{Q}_k x_t$ under the current policy. Suppose the distribution changes to p_{k+1} and define $\Delta p_{k,m} := p_{k+1}(m) - p_k(m)$. Using the same matrix S_k, we can write $S_k - \sum_m p_{k+1}(m) F_m^\top S_k F_m = \mathcal{Q}_k - E_k$, where $E_k := \sum_m \Delta p_{k,m} F_m^\top S_k F_m$. Hence the same Lyapunov matrix certifies the next policy whenever $\mathcal{Q}_k - E_k \succ 0$. A sufficient condition is $\|E_k\|_2 < \lambda_{\min}(\mathcal{Q}_k)$.

The remaining step is to relate this perturbation to the policy parameter update.

Lemma 3 (Softmax policy perturbation)**.** *For any $\theta, \vartheta \in \mathbb{R}^M$,*

$$\|p_\theta - p_\vartheta\|_1 \le \frac{1}{2}\|\theta - \vartheta\|_1. \tag{26}$$

The bound is global, not a local linear approximation. Let $J_{\mathrm{sm}}(\theta) := \nabla_\theta p_\theta$ denote the Jacobian of the softmax map. It follows from the induced 1-norm bound $\|J_{\mathrm{sm}}(\theta)\|_1 \le 1/2$ and integration of the softmax Jacobian along the segment joining the two parameter vectors. Consider an implemented update $\theta_{k+1} = \theta_k - \eta_k g_k$ and define $H_k := \max_{1\le m\le M} \|F_m^\top S_k F_m\|_2$. Lemma 3 gives $\|p_{k+1} - p_k\|_1 \le (\eta_k/2)\|g_k\|_1$ and therefore $\|E_k\|_2 \le (\eta_k H_k/2)\|g_k\|_1$.

Theorem 7 (Stability-preserving policy update)**.** *Suppose p_k is mean-square stable and S_k satisfies equation 25. The update $\theta_{k+1} = \theta_k - \eta_k g_k$ preserves mean-square stability if*

$$\eta_k < \frac{2\lambda_{\min}(\mathcal{Q}_k)}{H_k\|g_k\|_1}. \tag{27}$$

Algorithm 1 `StableSwitch-PG`

Require: Initial MSS parameter θ_0, $\eta_{\text{nom}} > 0$, $\beta \in (0, 1)$, $\mathcal{Q}_k \succ 0$

for $k = 0, 1, \ldots$ **do**
  $p_k \leftarrow \text{softmax}(\theta_k)$
  Obtain g_k
  Solve $S_k - \sum_m p_k(m) F_m^\top S_k F_m = \mathcal{Q}_k$
  $H_k \leftarrow \max_m \|F_m^\top S_k F_m\|_2$
  $\eta_k \leftarrow \min\left\{\eta_{\text{nom}}, \beta \dfrac{2\lambda_{\min}(\mathcal{Q}_k)}{H_k \|g_k\|_1}\right\}$
  $\theta_{k+1} \leftarrow \theta_k - \eta_k g_k$
end for

Theorem 7 converts the current Lyapunov margin into an admissible movement in policy-parameter space. Importantly, the guarantee depends on the update that is actually applied, rather than on how the direction g_k was obtained. Thus g_k may be the exact gradient or a sampled estimate $\widehat{g}_k$: gradient noise affects optimization progress, while stability is preserved whenever the implemented step satisfies equation 27. The roles of optimization and stability are therefore distinct. Sections 3–4 determine a direction for reducing the cost, whereas Theorem 7 limits how far the policy may move in that direction without leaving the MSS region.

Algorithm 1 combines policy improvement with a stability-preserving update. At iteration k, the current distribution and gradient direction are computed, and a Lyapunov equation is solved to obtain the admissible step-size bound.

When the system model is known, the closed-loop matrices $F_m = A - BK_m$ are available and this bound can be computed directly. When the dynamics are unknown, obtaining the same guarantee is nontrivial because S_k and H_k depend on the unknown model. A common approach is to estimate the system matrices from trajectory data and then solve the corresponding Lyapunov or Riccati equations using the estimated model (Lale et al. (2021); Mania et al. (2019)).

The experimental setup and results are described in Appendix (D).

6 Conclusion and Future Work

We developed `StableSwitch-PG`, a policy-gradient framework for optimizing a state-independent distribution over a fixed library of linear feedback controllers. The framework provides convergence guarantees for exact and trajectory-based policy gradients and a Lyapunov-based update rule that preserves mean-square stability from one policy iterate to the next.

Future work will extend the stability-preserving update to unknown system dynamics, where obtaining the Lyapunov certificate and exact learning-rate bound from data is nontrivial. Other directions include state-dependent or Markovian switching policies and extensions to nonlinear systems.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 22–31. PMLR, 2017.

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.

Naman Agarwal, Nataly Brukhim, Elad Hazan, and Zhou Lu. Boosting for Control of Dynamical Systems. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 96–103. PMLR, 2020.

Ershad Banijamali, Yasin Abbasi-Yadkori, Mohammad Ghavamzadeh, and Nikos Vlassis. Optimizing over a Restricted Policy Class in MDPs. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 3042–3050. PMLR, 2019.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe Model-based Reinforcement Learning with Stability Guarantees. In *Advances in Neural Information Processing Systems*, volume 30, pp. 908–918, 2017.

Jalaj Bhandari and Daniel Russo. Global Optimality Guarantees for Policy Gradient Methods. *Operations Research*, 72(5):1906–1927, 2024. doi: 10.1287/opre.2021.0014.

Paolo Bolzern, Patrizio Colaneri, and Giuseppe De Nicolao. Almost sure stability of stochastic linear systems with ergodic parameters: an average contractivity criterion. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 950–954, 2006. doi: 10.1109/CDC.2006.377061.

M. S. Branicky. Multiple Lyapunov Functions and Other Analysis Tools for Switched and Hybrid Systems. *IEEE Transactions on Automatic Control*, 43(4):475–482, 1998. doi: 10.1109/9.664150.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A Lyapunov-based Approach to Safe Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8103–8112, 2018.

Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global Convergence of Policy Gradient Methods for the Linear Quadratic Regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1466–1475. PMLR, 2018.

Joao P. Hespanha and A. Stephen Morse. Stability of Switched Systems with Average Dwell-Time. In *Proceedings of the 38th IEEE Conference on Decision and Control*, volume 3, pp. 2655–2660. IEEE, 1999.

Joao Paulo Jansch-Porto, Bin Hu, and Geir E. Dullerud. Policy Optimization for Markovian Jump Linear Quadratic Control: Gradient Method and Global Convergence. *IEEE Transactions on Automatic Control*, 68(4):2475–2482, 2023. doi: 10.1109/TAC.2022.3176439.

Sahin Lale, Oguzhan Teke, Babak Hassibi, and Anima Anandkumar. Stability and Identification of Random Asynchronous Linear Time-Invariant Systems. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pp. 651–663. PMLR, 07 – 08 June 2021.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax Policy Gradient Methods Can Take Exponential Time to Converge. In *Proceedings of the Thirty-Fourth Conference on Learning Theory*, volume 134, pp. 3107–3110. PMLR, 2021.

Daniel Liberzon and A. Stephen Morse. Basic Problems in Stability and Design of Switched Systems. *IEEE Control Systems Magazine*, 19(5):59–70, 1999. doi: 10.1109/37.793443.

Hai Lin and Panos J. Antsaklis. Stability and Stabilizability of Switched Linear Systems: A Survey of Recent Results. *IEEE Transactions on Automatic Control*, 54(2):308–322, 2009. doi: 10.1109/TAC.2008.2012009.

Horia Mania, Stephen Tu, and Benjamin Recht. Certainty Equivalence is Efficient for Linear Quadratic Control. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Zahra Marvi and Bahare Kiumarsi. Reinforcement Learning With Safety and Stability Guarantees During Exploration for Linear Systems. *IEEE Open Journal of Control Systems*, 1:322–334, 2022. doi: 10.1109/OJCSYS.2022.3209945.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the Global Convergence Rates of Softmax Policy Gradient Methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6820–6829. PMLR, 2020.

David Meyer. Likelihood ratio policy gradients for reinforcement learning. Technical report, University of Oregon, 2018.

Morenikeji Neri and Thomas Powell. A quantitative robbins-siegmund theorem. *Annals of Applied Probability*, 36(1):636–651, February 2026. ISSN 1050-5164. doi: 10.1214/25-AAP2242.

Matteo Papini, Matteo Pirotta, and Marcello Restelli. Smoothing policies and safe policy gradients. *Machine Learning*, 111:4081 – 4137, 2019.

Philippos Peleties and Raymond DeCarlo. Asymptotic Stability of m-Switched Systems using Lyapunov-Like Functions. In *American Control Conference*, pp. 1679–1684. IEEE, 1991.

Theodore J. Perkins and Andrew G. Barto. Lyapunov Design for Safe Reinforcement Learning. *Journal of Machine Learning Research*, 3:803–832, 2002.

Robert Shorten, Fabian Wirth, Oliver Mason, Kai Wulff, and Christopher King. Stability Criteria for Switched and Hybrid Systems. *SIAM Review*, 49(4):545–592, 2007. doi: 10.1137/05063516X.

Aivar Sootla, Alexander I. Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H. Mguni, Jun Wang, and Haitham Ammar. Saute RL: Almost Surely Safe Reinforcement Learning Using State Augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 20423–20443. PMLR, 2022.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward Constrained Policy Optimization. In *International Conference on Learning Representations*, 2019.

Bo Wu, Murat Cubuktepe, Franck Djeumou, Zhe Xu, and Ufuk Topcu. Policy Synthesis for Switched Linear Systems With Markov Decision Process Switching. *IEEE Transactions on Automatic Control*, 68(1):532–539, 2023. doi: 10.1109/TAC.2022.3145659.

Mohammadi Zaki, Avi Mohan, Aditya Gopalan, and Shie Mannor. Actor-Critic based Improper Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 25867–25919. PMLR, 2022.

A Proofs for Policy Optimization with Exact Gradients

A.1 Second-Moment Dynamics

Let $\Sigma_t := \mathbb{E}[x_t x_t^\top]$. Since $x_{t+1} = F_{\sigma(t)} x_t$ and $\sigma(t)$ is drawn independently according to p_θ,

$$\begin{aligned}
\Sigma_{t+1} &= \mathbb{E}\left[x_{t+1} x_{t+1}^\top\right] \\
&= \mathbb{E}\left[F_{\sigma(t)} x_t x_t^\top F_{\sigma(t)}^\top\right] \\
&= \sum_{m=1}^{M} p_\theta(m) \mathbb{E}\left[F_m x_t x_t^\top F_m^\top\right] \\
&= \sum_{m=1}^{M} p_\theta(m) F_m \mathbb{E}[x_t x_t^\top] F_m^\top \\
&= \sum_{m=1}^{M} p_\theta(m) F_m \Sigma_t F_m^\top.
\end{aligned}$$

Using $\mathrm{vec}(AXB) = (B^\top \otimes A)\,\mathrm{vec}(X)$,

$$\begin{aligned}\mathrm{vec}(\Sigma_{t+1}) &= \mathrm{vec}\left(\sum_{m=1}^{M} p_\theta(m) F_m \Sigma_t F_m^\top\right)\\ &= \sum_{m=1}^{M} p_\theta(m)\,\mathrm{vec}\left(F_m \Sigma_t F_m^\top\right)\\ &= \sum_{m=1}^{M} p_\theta(m)(F_m \otimes F_m)\,\mathrm{vec}(\Sigma_t)\\ &= \mathcal{M}_\theta\,\mathrm{vec}(\Sigma_t).\end{aligned}$$

where $\mathcal{M}_\theta := \sum_{m=1}^{M} p_\theta(m)(F_m \otimes F_m)$. For a fixed switching distribution p_θ, the second-moment recursion is $\mathrm{vec}(\Sigma_{t+1}) = \mathcal{M}_\theta\,\mathrm{vec}(\Sigma_t)$.

Therefore, by mathematical induction, $\mathrm{vec}(\Sigma_t) = \mathcal{M}_\theta^t\,\mathrm{vec}(\Sigma_0), \qquad \forall t \in \mathbb{Z}_{\geq 0}$.

A.2 Proof of Lemma 1

Lemma 1 (Quadratic value representation). For every $\theta \in \mathcal{D}_\gamma$, $V_\theta(x) = x^\top P_\theta x$, where $P_\theta \succeq 0$ is the unique solution of

$$P_\theta = \sum_{m=1}^{M} p_\theta(m)\left(C_m + \gamma F_m^\top P_\theta F_m\right). \tag{28}$$

Moreover, $J(\theta) = \mathrm{Tr}(P_\theta \Sigma_0)$.

Proof. Define $\bar{C}_\theta := \sum_{m=1}^{M} p_\theta(m) C_m$. For a deterministic initial state $x_0 = x$, define $\Sigma_t^x := \mathbb{E}\left[x_t x_t^\top \mid x_0 = x\right]$. Then $\Sigma_0^x = xx^\top$ and $\mathrm{vec}(\Sigma_t^x) = \mathcal{M}_\theta^t\,\mathrm{vec}(xx^\top)$. The expected stage cost satisfies

$$\begin{aligned}\mathbb{E}_\theta\left[x_t^\top C_{\sigma(t)} x_t \mid x_0 = x\right] &= \sum_{m=1}^{M} p_\theta(m)\mathbb{E}_\theta\left[x_t^\top C_m x_t \mid x_0 = x\right]\\ &= \sum_{m=1}^{M} p_\theta(m)\,\mathrm{Tr}\left(C_m \Sigma_t^x\right)\\ &= \mathrm{Tr}\left[\left(\sum_{m=1}^{M} p_\theta(m) C_m\right)\Sigma_t^x\right]\\ &= \mathrm{Tr}\left(\bar{C}_\theta \Sigma_t^x\right)\\ &= \mathrm{vec}(\bar{C}_\theta)^\top\,\mathrm{vec}(\Sigma_t^x)\\ &= \mathrm{vec}(\bar{C}_\theta)^\top \mathcal{M}_\theta^t\,\mathrm{vec}(xx^\top).\end{aligned}$$

Hence

$$\begin{aligned}V_\theta(x) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\theta\left[x_t^\top C_{\sigma(t)} x_t \mid x_0 = x\right]\\ &= \sum_{t=0}^{\infty} \gamma^t\,\mathrm{vec}(\bar{C}_\theta)^\top \mathcal{M}_\theta^t\,\mathrm{vec}(xx^\top)\\ &= \mathrm{vec}(\bar{C}_\theta)^\top\left(\sum_{t=0}^{\infty}(\gamma\mathcal{M}_\theta)^t\right)\mathrm{vec}(xx^\top).\end{aligned}$$

Since $\rho(\gamma\mathcal{M}_\theta) < 1$, the Neumann series converges and $\sum_{t=0}^{\infty}(\gamma\mathcal{M}_\theta)^t = (I - \gamma\mathcal{M}_\theta)^{-1}$. Therefore,

$$\begin{aligned}V_\theta(x) &= \mathrm{vec}(\bar{C}_\theta)^\top (I - \gamma\mathcal{M}_\theta)^{-1}\,\mathrm{vec}(xx^\top)\\ &= \mathrm{vec}(xx^\top)^\top (I - \gamma\mathcal{M}_\theta^\top)^{-1}\,\mathrm{vec}(\bar{C}_\theta).\end{aligned}$$

Define P_θ by $\mathrm{vec}(P_\theta) := (I - \gamma \mathcal{M}_\theta^\top)^{-1} \mathrm{vec}(\bar{C}_\theta)$. Then

$$\begin{aligned} V_\theta(x) &= \mathrm{vec}(xx^\top)^\top \mathrm{vec}(P_\theta) \\ &= \mathrm{Tr}\left(xx^\top P_\theta\right) \\ &= x^\top P_\theta x. \end{aligned}$$

Since every stage cost is nonnegative, $V_\theta(x) \geq 0$. Therefore, $x^\top P_\theta x \geq 0$ and hence $P_\theta \succeq 0$. From the definition of P_θ, $\mathrm{vec}(P_\theta) = (I - \gamma \mathcal{M}_\theta^\top)^{-1} \mathrm{vec}(\bar{C}_\theta)$, so $(I - \gamma \mathcal{M}_\theta^\top) \mathrm{vec}(P_\theta) = \mathrm{vec}(\bar{C}_\theta)$, and therefore $\mathrm{vec}(P_\theta) - \gamma \mathcal{M}_\theta^\top \mathrm{vec}(P_\theta) = \mathrm{vec}(\bar{C}_\theta)$. Since

$$\begin{aligned} \mathcal{M}_\theta^\top \mathrm{vec}(P_\theta) &= \sum_{m=1}^{M} p_\theta(m)(F_m^\top \otimes F_m^\top) \mathrm{vec}(P_\theta) \\ &= \sum_{m=1}^{M} p_\theta(m) \mathrm{vec}\left(F_m^\top P_\theta F_m\right), \end{aligned}$$

we obtain

$$\begin{aligned} \mathrm{vec}(P_\theta) &= \mathrm{vec}(\bar{C}_\theta) + \gamma \sum_{m=1}^{M} p_\theta(m) \mathrm{vec}\left(F_m^\top P_\theta F_m\right) \\ &= \mathrm{vec}\left[\bar{C}_\theta + \gamma \sum_{m=1}^{M} p_\theta(m) F_m^\top P_\theta F_m\right]. \end{aligned}$$

Hence

$$\begin{aligned} P_\theta &= \bar{C}_\theta + \gamma \sum_{m=1}^{M} p_\theta(m) F_m^\top P_\theta F_m \\ &= \sum_{m=1}^{M} p_\theta(m) C_m + \gamma \sum_{m=1}^{M} p_\theta(m) F_m^\top P_\theta F_m \\ &= \sum_{m=1}^{M} p_\theta(m) \left(C_m + \gamma F_m^\top P_\theta F_m\right). \end{aligned}$$

To show uniqueness, suppose P_θ and $\widetilde{P}_\theta$ both satisfy equation 5. Then $P_\theta - \widetilde{P}_\theta = \gamma \sum_{m=1}^{M} p_\theta(m) F_m^\top (P_\theta - \widetilde{P}_\theta) F_m$. Vectorizing gives $\mathrm{vec}(P_\theta - \widetilde{P}_\theta) = \gamma \mathcal{M}_\theta^\top \mathrm{vec}(P_\theta - \widetilde{P}_\theta)$, so $(I - \gamma \mathcal{M}_\theta^\top) \mathrm{vec}(P_\theta - \widetilde{P}_\theta) = 0$. Since $\rho(\gamma \mathcal{M}_\theta) < 1$, the matrix $I - \gamma \mathcal{M}_\theta^\top$ is nonsingular. Hence $\mathrm{vec}(P_\theta - \widetilde{P}_\theta) = 0$, and therefore $P_\theta = \widetilde{P}_\theta$.

Finally,

$$\begin{aligned} J(\theta) &= \mathbb{E}\left[V_\theta(x_0)\right] \\ &= \mathbb{E}\left[x_0^\top P_\theta x_0\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left(P_\theta x_0 x_0^\top\right)\right] \\ &= \mathrm{Tr}\left(P_\theta \mathbb{E}[x_0 x_0^\top]\right) \\ &= \mathrm{Tr}\left(P_\theta \Sigma_0\right). \end{aligned}$$

□

A.3 Analyticity and Smoothness of the Objective

For each controller, $p_\theta(m) = \frac{e^{\theta_m}}{\sum_{j=1}^{M} e^{\theta_j}}$ is real analytic in θ. Therefore, $\bar{C}_\theta = \sum_{m=1}^{M} p_\theta(m) C_m$ is real analytic, and $\mathcal{M}_\theta = \sum_{m=1}^{M} p_\theta(m)(F_m \otimes F_m)$ is real analytic. From Lemma 1, $\mathrm{vec}(P_\theta) =$

$(I-\gamma\mathcal{M}_\theta^\top)^{-1}\,\mathrm{vec}(\bar{C}_\theta)$. For $\theta \in \mathcal{D}_\gamma$, $\rho(\gamma\mathcal{M}_\theta) < 1$. Hence $1 \notin \mathrm{spec}(\gamma\mathcal{M}_\theta)$, so $\det(I-\gamma\mathcal{M}_\theta^\top) \neq 0$. Therefore, $(I-\gamma\mathcal{M}_\theta^\top)^{-1}$ is real analytic on $\mathcal{D}_\gamma$, and consequently P_θ is real analytic on $\mathcal{D}_\gamma$.

Since $J(\theta) = \mathrm{Tr}(P_\theta\Sigma_0)$, J is real analytic on $\mathcal{D}_\gamma$. Thus $\nabla^2 J(\theta)$ is continuous on $\mathcal{D}_\gamma$. Let $\Theta \subset \mathcal{D}_\gamma$ be compact and convex. Then $\theta \mapsto \|\nabla^2 J(\theta)\|_2$ is continuous on Θ, so $L := \max_{\theta\in\Theta} \|\nabla^2 J(\theta)\|_2 < \infty$. For $\theta, \vartheta \in \Theta$,

$$\nabla J(\theta) - \nabla J(\vartheta) = \int_0^1 \nabla^2 J\left(\vartheta + s(\theta-\vartheta)\right)(\theta-\vartheta)\,ds.$$

Therefore,

$$\begin{aligned}\|\nabla J(\theta) - \nabla J(\vartheta)\|_2 &\le \int_0^1 \left\|\nabla^2 J\left(\vartheta + s(\theta-\vartheta)\right)\right\|_2 \|\theta-\vartheta\|_2\,ds \\ &\le \int_0^1 L\|\theta-\vartheta\|_2\,ds \\ &= L\|\theta-\vartheta\|_2.\end{aligned}$$

Hence J is L-smooth on Θ.

A.4 Discounted State-Occupancy Equation

By definition, $X_\theta = \sum_{t=0}^\infty \gamma^t \Sigma_t$. Then

$$\begin{aligned}X_\theta &= \Sigma_0 + \sum_{t=1}^\infty \gamma^t \Sigma_t \\ &= \Sigma_0 + \gamma \sum_{s=0}^\infty \gamma^s \Sigma_{s+1} \\ &= \Sigma_0 + \gamma \sum_{s=0}^\infty \gamma^s \sum_{m=1}^M p_\theta(m) F_m \Sigma_s F_m^\top \\ &= \Sigma_0 + \gamma \sum_{m=1}^M p_\theta(m) F_m \left(\sum_{s=0}^\infty \gamma^s \Sigma_s\right) F_m^\top \\ &= \Sigma_0 + \gamma \sum_{m=1}^M p_\theta(m) F_m X_\theta F_m^\top.\end{aligned}$$

A.5 Zero-Average Advantage

Using $A_{\theta,m} = C_m + \gamma F_m^\top P_\theta F_m - P_\theta$, we have

$$\begin{aligned}\sum_{m=1}^M p_\theta(m) A_{\theta,m} &= \sum_{m=1}^M p_\theta(m)\left(C_m + \gamma F_m^\top P_\theta F_m - P_\theta\right) \\ &= \sum_{m=1}^M p_\theta(m)\left(C_m + \gamma F_m^\top P_\theta F_m\right) - \sum_{m=1}^M p_\theta(m) P_\theta \\ &= P_\theta - \left(\sum_{m=1}^M p_\theta(m)\right) P_\theta \\ &= P_\theta - P_\theta \\ &= 0.\end{aligned}$$

A.6 Softmax Derivative

Let $Z(\theta) := \sum_{r=1}^{M} e^{\theta_r}$. Then $p_\theta(m) = \frac{e^{\theta_m}}{Z(\theta)}$. For $j \in \{1, \dots, M\}$,

$$\begin{aligned}\frac{\partial p_\theta(m)}{\partial \theta_j} &= \frac{\mathbf{1}\{m = j\} e^{\theta_m} Z(\theta) - e^{\theta_m} e^{\theta_j}}{Z(\theta)^2} \\ &= \frac{e^{\theta_m}}{Z(\theta)} \left(\mathbf{1}\{m = j\} - \frac{e^{\theta_j}}{Z(\theta)} \right) \\ &= p_\theta(m) \left(\mathbf{1}\{m = j\} - p_\theta(j) \right).\end{aligned}$$

A.7 Proof of Theorem 1

Theorem 1 (Exact policy gradient). For every $\theta \in \mathcal{D}_\gamma$ and $m \in \{1, \dots, M\}$,

$$\frac{\partial J(\theta)}{\partial \theta_m} = p_\theta(m) \operatorname{Tr}\left(X_\theta A_{\theta,m} \right).$$

Proof. Fix $j \in \{1, \dots, M\}$ and define $\mathcal{G}_j := \frac{\partial P_\theta}{\partial \theta_j}$. Differentiate equation 5:

$$\begin{aligned}\mathcal{G}_j = &\sum_{m=1}^{M} \frac{\partial p_\theta(m)}{\partial \theta_j} \left(C_m + \gamma F_m^\top P_\theta F_m \right) \\ &+ \gamma \sum_{m=1}^{M} p_\theta(m) F_m^\top \mathcal{G}_j F_m.\end{aligned}$$

Using the softmax derivative,

$$\begin{aligned}\sum_{m=1}^{M} \frac{\partial p_\theta(m)}{\partial \theta_j} \left(C_m + \gamma F_m^\top P_\theta F_m \right) &= \sum_{m=1}^{M} p_\theta(m) \left(\mathbf{1}\{m = j\} - p_\theta(j) \right) \left(C_m + \gamma F_m^\top P_\theta F_m \right) \\ &= p_\theta(j) \left(C_j + \gamma F_j^\top P_\theta F_j \right) \\ &\quad - p_\theta(j) \sum_{m=1}^{M} p_\theta(m) \left(C_m + \gamma F_m^\top P_\theta F_m \right) \\ &= p_\theta(j) \left(C_j + \gamma F_j^\top P_\theta F_j \right) - p_\theta(j) P_\theta \\ &= p_\theta(j) \left(C_j + \gamma F_j^\top P_\theta F_j - P_\theta \right) \\ &= p_\theta(j) A_{\theta,j}.\end{aligned}$$

Hence $\mathcal{G}_j = p_\theta(j) A_{\theta,j} + \gamma \sum_{m=1}^{M} p_\theta(m) F_m^\top \mathcal{G}_j F_m$. Therefore, $\mathcal{G}_j - \gamma \sum_{m=1}^{M} p_\theta(m) F_m^\top \mathcal{G}_j F_m = p_\theta(j) A_{\theta,j}$. Since $J(\theta) = \operatorname{Tr}(P_\theta \Sigma_0)$, we have $\frac{\partial J(\theta)}{\partial \theta_j} = \operatorname{Tr}(\mathcal{G}_j \Sigma_0)$. From the occupancy equation, $\Sigma_0 = X_\theta - \gamma \sum_{m=1}^{M} p_\theta(m) F_m X_\theta F_m^\top$.

Hence

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_j} &= \operatorname{Tr}\left[\mathcal{G}_j\left(X_\theta - \gamma \sum_{m=1}^{M} p_\theta(m) F_m X_\theta F_m^\top\right)\right] \\
&= \operatorname{Tr}(\mathcal{G}_j X_\theta) - \gamma \sum_{m=1}^{M} p_\theta(m) \operatorname{Tr}\left(\mathcal{G}_j F_m X_\theta F_m^\top\right) \\
&= \operatorname{Tr}(\mathcal{G}_j X_\theta) - \gamma \sum_{m=1}^{M} p_\theta(m) \operatorname{Tr}\left(F_m^\top \mathcal{G}_j F_m X_\theta\right) \\
&= \operatorname{Tr}\left[\left(\mathcal{G}_j - \gamma \sum_{m=1}^{M} p_\theta(m) F_m^\top \mathcal{G}_j F_m\right) X_\theta\right] \\
&= \operatorname{Tr}\left[p_\theta(j) A_{\theta,j} X_\theta\right] \\
&= p_\theta(j) \operatorname{Tr}\left(X_\theta A_{\theta,j}\right).
\end{aligned}$$

Replacing j by m gives equation 8.

□

A.8 Proof of Lemma 2

Lemma 2 (Performance-difference identity). For any discounted-admissible switching distribution q,

$$J(q) - J(\theta) = \operatorname{Tr}\left[X_q \sum_{m=1}^{M} q(m) A_{\theta,m}\right].$$

Proof. For distribution q, define $\bar{C}_q := \sum_{m=1}^{M} q(m) C_m$. Then $J(q) = \operatorname{Tr}(\bar{C}_q X_q)$. The occupancy equation under q is $X_q = \Sigma_0 + \gamma \sum_{m=1}^{M} q(m) F_m X_q F_m^\top$. Hence $\Sigma_0 = X_q - \gamma \sum_{m=1}^{M} q(m) F_m X_q F_m^\top$. Since $J(\theta) = \operatorname{Tr}(P_\theta \Sigma_0)$, we obtain

$$\begin{aligned}
J(\theta) &= \operatorname{Tr}\left[P_\theta\left(X_q - \gamma \sum_{m=1}^{M} q(m) F_m X_q F_m^\top\right)\right] \\
&= \operatorname{Tr}(P_\theta X_q) - \gamma \sum_{m=1}^{M} q(m) \operatorname{Tr}\left(P_\theta F_m X_q F_m^\top\right) \\
&= \operatorname{Tr}(P_\theta X_q) - \gamma \sum_{m=1}^{M} q(m) \operatorname{Tr}\left(F_m^\top P_\theta F_m X_q\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
J(q) - J(\theta) &= \mathrm{Tr}(\bar{C}_q X_q) - \mathrm{Tr}(P_\theta X_q) \\
&\quad + \gamma \sum_{m=1}^{M} q(m)\, \mathrm{Tr}\left(F_m^\top P_\theta F_m X_q\right) \\
&= \mathrm{Tr}\left[\left(\bar{C}_q - P_\theta + \gamma \sum_{m=1}^{M} q(m) F_m^\top P_\theta F_m\right) X_q\right] \\
&= \mathrm{Tr}\left[\left(\sum_{m=1}^{M} q(m) C_m + \gamma \sum_{m=1}^{M} q(m) F_m^\top P_\theta F_m - P_\theta\right) X_q\right] \\
&= \mathrm{Tr}\left[\left(\sum_{m=1}^{M} q(m)\left(C_m + \gamma F_m^\top P_\theta F_m\right) - \left(\sum_{m=1}^{M} q(m)\right) P_\theta\right) X_q\right] \\
&= \mathrm{Tr}\left[\left(\sum_{m=1}^{M} q(m)\left(C_m + \gamma F_m^\top P_\theta F_m - P_\theta\right)\right) X_q\right] \\
&= \mathrm{Tr}\left[\left(\sum_{m=1}^{M} q(m) A_{\theta,m}\right) X_q\right] \\
&= \mathrm{Tr}\left[X_q \sum_{m=1}^{M} q(m) A_{\theta,m}\right].
\end{aligned}$$

□

A.9 Derivation of equation 10

Set $q = p^\star$ in Lemma 2. Then multiplying by -1,

$$\begin{aligned}
J(\theta) - J^\star &= \mathrm{Tr}\left[X_{p^\star}\left(-\sum_{m=1}^{M} p^\star(m) A_{\theta,m}\right)\right] \\
&= \mathrm{Tr}\left(X_{p^\star} \Gamma_\theta\right).
\end{aligned}$$

A.10 Proof of Proposition 1

Proposition 1 (Occupancy comparison). Suppose $\Sigma_0 \succ 0$ and $p^\star$ is discounted admissible, so that $X_{p^\star}$ is finite. Then, for every discounted-admissible θ, there exists a finite constant $\bar{\kappa} > 0$ such that

$$X_{p^\star} \preceq \bar{\kappa} X_\theta.$$

Proof. Since $\Sigma_t \succeq 0$, we have

$$\begin{aligned}
X_\theta &= \Sigma_0 + \sum_{t=1}^{\infty} \gamma^t \Sigma_t \\
&\succeq \Sigma_0 \\
&\succeq \lambda_{\min}(\Sigma_0) I.
\end{aligned}$$

Therefore, $I \preceq \frac{1}{\lambda_{\min}(\Sigma_0)} X_\theta$. Since $X_{p^\star} \succeq 0$, we also have $X_{p^\star} \preceq \lambda_{\max}(X_{p^\star}) I$. Hence $X_{p^\star} \preceq \lambda_{\max}(X_{p^\star}) I \preceq \frac{\lambda_{\max}(X_{p^\star})}{\lambda_{\min}(\Sigma_0)} X_\theta$. Thus equation 11 holds with

$$\bar{\kappa} = \frac{\lambda_{\max}(X_{p^\star})}{\lambda_{\min}(\Sigma_0)}.$$

□

A.11 Derivation of equation 13

From proposition 1, $X_{p^\star} \preceq \bar{\kappa} X_\theta$. Hence $\bar{\kappa} X_\theta - X_{p^\star} \succeq 0$. By Assumption 1, $\Gamma_\theta \succeq 0$. Therefore, $\operatorname{Tr}[(\bar{\kappa} X_\theta - X_{p^\star})\Gamma_\theta] = \operatorname{Tr}[\Gamma_\theta^{1/2}(\bar{\kappa} X_\theta - X_{p^\star})\Gamma_\theta^{1/2}] \geq 0$. Thus $\bar{\kappa} \operatorname{Tr}(X_\theta \Gamma_\theta) - \operatorname{Tr}(X_{p^\star}\Gamma_\theta) \geq 0$, and hence $\operatorname{Tr}(X_{p^\star}\Gamma_\theta) \leq \bar{\kappa} \operatorname{Tr}(X_\theta \Gamma_\theta)$.

Using equation 10,

$$\begin{aligned} J(\theta) - J^\star &\leq \bar{\kappa} \operatorname{Tr}(X_\theta \Gamma_\theta) \\ &= \bar{\kappa} \operatorname{Tr}\left[X_\theta \left(-\sum_{m=1}^{M} p^\star(m) A_{\theta,m} \right) \right] \\ &= \bar{\kappa} \left[-\sum_{m=1}^{M} p^\star(m) \operatorname{Tr}(X_\theta A_{\theta,m}) \right]. \end{aligned}$$

A.12 Proof of equation 14

Let $m_{\min} \in \arg\min_m \theta_m$. Then

$$\begin{aligned} \min_m p_\theta(m) &= p_\theta(m_{\min}) \\ &= \frac{e^{\theta_{m_{\min}}}}{\sum_{j=1}^{M} e^{\theta_j}} \\ &= \frac{1}{\sum_{j=1}^{M} e^{\theta_j - \theta_{m_{\min}}}} \\ &= \frac{1}{1 + \sum_{j \neq m_{\min}} e^{\theta_j - \theta_{m_{\min}}}}. \end{aligned}$$

For every j, $\theta_j - \theta_{m_{\min}} \leq \max_r \theta_r - \min_r \theta_r = \Delta(\theta)$. Therefore, $e^{\theta_j - \theta_{m_{\min}}} \leq e^{\Delta(\theta)}$. Hence, $1 + \sum_{j \neq m_{\min}} e^{\theta_j - \theta_{m_{\min}}} \leq 1 + (M-1)e^{\Delta(\theta)}$. Taking positive reciprocals,

$$\begin{aligned} & \min_m p_\theta(m) \geq \frac{1}{1 + (M-1)e^{\Delta(\theta)}} \\ \Rightarrow \quad & p_\theta(m) \geq \frac{1}{1 + (M-1)e^{\Delta(\theta)}}, \\ \Rightarrow \quad & \frac{1}{p_\theta(m)} \leq 1 + (M-1)e^{\Delta(\theta)}. \end{aligned}$$

A.13 Proof of Theorem 2

Theorem 2 (Gradient dominance). Suppose Assumption 1 holds and $\Sigma_0 \succ 0$. Define

$$D(\theta) := \bar{\kappa} \left[1 + (M-1)e^{\Delta(\theta)} \right].$$

Then

$$J(\theta) - J^\star \leq D(\theta) \|\nabla J(\theta)\|_2.$$

Proof. Define $h_m(\theta) := \operatorname{Tr}(X_\theta A_{\theta,m})$. From equation 13,

$$\begin{aligned} J(\theta) - J^\star &\leq \bar{\kappa} \left[-\sum_{m=1}^{M} p^\star(m) h_m(\theta) \right] \\ &\leq \bar{\kappa} \sum_{m=1}^{M} p^\star(m) |h_m(\theta)|. \end{aligned}$$

From Theorem 1, $[\nabla J(\theta)]_m = p_\theta(m) h_m(\theta)$. Therefore,

$$\begin{aligned} |h_m(\theta)| &= \frac{|[\nabla J(\theta)]_m|}{p_\theta(m)} \\ &\le \left[1 + (M-1)e^{\Delta(\theta)}\right] |[\nabla J(\theta)]_m|. \end{aligned}$$

Substituting,

$$\begin{aligned} J(\theta) - J^\star &\le \bar{\kappa} \left[1 + (M-1)e^{\Delta(\theta)}\right] \sum_{m=1}^{M} p^\star(m) |[\nabla J(\theta)]_m| \\ &\le \bar{\kappa} \left[1 + (M-1)e^{\Delta(\theta)}\right] \max_m |[\nabla J(\theta)]_m| \sum_{m=1}^{M} p^\star(m) \\ &= \bar{\kappa} \left[1 + (M-1)e^{\Delta(\theta)}\right] \|\nabla J(\theta)\|_\infty \\ &\le \bar{\kappa} \left[1 + (M-1)e^{\Delta(\theta)}\right] \|\nabla J(\theta)\|_2 \\ &= D(\theta) \|\nabla J(\theta)\|_2. \end{aligned}$$

□

A.14 Proof of Theorem 3

Theorem 3 (Convergence of exact policy gradient). Suppose the iterates remain in a region on which J is L-smooth, Assumption 1 holds, $\Sigma_0 \succ 0$, and $0 < \eta \le 1/L$. Then

$$\delta_{k_\mathrm{f}} \le \left[\delta_0^{-1} + \frac{\eta}{2} \sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2} \right]^{-1}.$$

Consequently, $J(\theta_k) \to J^\star$ whenever $\sum_{k=0}^{\infty} D_k^{-2} = \infty$.

Proof. If $\delta_0 = 0$, then $J(\theta_0) = J^\star$ and the claim is immediate. Assume $\delta_0 > 0$. By L-smoothness (Papini et al. (2019)),

$$J(\theta_{k+1}) \le J(\theta_k) + \nabla J(\theta_k)^\top (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|_2^2.$$

Using $\theta_{k+1} - \theta_k = -\eta \nabla J(\theta_k)$, we obtain

$$\begin{aligned} J(\theta_{k+1}) &\le J(\theta_k) - \eta \|\nabla J(\theta_k)\|_2^2 + \frac{L\eta^2}{2} \|\nabla J(\theta_k)\|_2^2 \\ &= J(\theta_k) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla J(\theta_k)\|_2^2. \end{aligned}$$

Since $0 < \eta \le \frac{1}{L}$, we have $1 - \frac{L\eta}{2} \ge \frac{1}{2}$. Therefore,

$$J(\theta_{k+1}) \le J(\theta_k) - \frac{\eta}{2} \|\nabla J(\theta_k)\|_2^2.$$

Subtracting $J^\star$,

$$\delta_{k+1} \le \delta_k - \frac{\eta}{2} \|\nabla J(\theta_k)\|_2^2.$$

From Theorem 2, $\delta_k \le D_k \|\nabla J(\theta_k)\|_2$, so $\|\nabla J(\theta_k)\|_2 \ge \frac{\delta_k}{D_k}$ and $\|\nabla J(\theta_k)\|_2^2 \ge \frac{\delta_k^2}{D_k^2}$. Hence

$$\delta_{k+1} \le \delta_k - \frac{\eta}{2D_k^2} \delta_k^2.$$

Define $c_k := \frac{\eta}{2D_k^2}$. Then $\delta_{k+1} \leq \delta_k - c_k\delta_k^2 \leq \delta_k$. If $\delta_{k+1} = 0$ for some k, then the optimum has been reached. Otherwise $\delta_{k+1} > 0$, and

$$\begin{aligned}\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k} &= \frac{\delta_k - \delta_{k+1}}{\delta_k\delta_{k+1}} \\ &\geq \frac{c_k\delta_k^2}{\delta_k\delta_{k+1}} \\ &= c_k\frac{\delta_k}{\delta_{k+1}} \\ &\geq c_k \\ &= \frac{\eta}{2D_k^2}.\end{aligned}$$

Summing from $k = 0$ to $k_\mathrm{f} - 1$,

$$\sum_{k=0}^{k_\mathrm{f}-1}\left(\frac{1}{\delta_{k+1}} - \frac{1}{\delta_k}\right) \geq \frac{\eta}{2}\sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2}.$$

The left-hand side telescopes:

$$\frac{1}{\delta_{k_\mathrm{f}}} - \frac{1}{\delta_0} \geq \frac{\eta}{2}\sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2}.$$

Therefore,

$$\frac{1}{\delta_{k_\mathrm{f}}} \geq \delta_0^{-1} + \frac{\eta}{2}\sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2}.$$

Taking positive reciprocals,

$$\delta_{k_\mathrm{f}} \leq \left[\delta_0^{-1} + \frac{\eta}{2}\sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2}\right]^{-1}.$$

If $\sum_{k=0}^{\infty} D_k^{-2} = \infty$, then $\delta_0^{-1} + \frac{\eta}{2}\sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2} \to \infty$, and consequently $\delta_{k_\mathrm{f}} \to 0$. Hence $J(\theta_{k_\mathrm{f}}) \to J^\star$.

□

A.15 Derivation of equation 18

If $\Delta(\theta_k) \leq \bar{\Delta}$, then

$$\begin{aligned}D_k &= \bar{\kappa}\left[1 + (M-1)e^{\Delta(\theta_k)}\right] \\ &\leq \bar{\kappa}\left[1 + (M-1)e^{\bar{\Delta}}\right] \\ &= \bar{D}.\end{aligned}$$

Therefore, $D_k^{-2} \geq \bar{D}^{-2}$. Hence $\sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2} \geq \sum_{k=0}^{k_\mathrm{f}-1} \bar{D}^{-2} = \frac{k_\mathrm{f}}{\bar{D}^2}$. Using Theorem 3,

$$\begin{aligned} J(\theta_{k_\mathrm{f}}) - J^\star &= \delta_{k_\mathrm{f}} \\ &\leq \left[\delta_0^{-1} + \frac{\eta}{2}\sum_{k=0}^{k_\mathrm{f}-1} D_k^{-2}\right]^{-1} \\ &\leq \left[\delta_0^{-1} + \frac{\eta}{2}\frac{k_\mathrm{f}}{\bar{D}^2}\right]^{-1} \\ &= \frac{1}{\delta_0^{-1} + \eta k_\mathrm{f}/(2\bar{D}^2)} \\ &= \mathcal{O}(k_\mathrm{f}^{-1}). \end{aligned}$$

B Proofs for Policy Learning from Sampled Trajectories

B.1 Softmax Score Function

For $p_\theta(m) = \frac{e^{\theta_m}}{\sum_{j=1}^M e^{\theta_j}}$, we have,

$$\log p_\theta(m) = \theta_m - \log\left(\sum_{j=1}^M e^{\theta_j}\right).$$

Therefore, for component r,

$$\begin{aligned} \frac{\partial}{\partial\theta_r}\log p_\theta(m) &= \mathbf{1}\{r=m\} - \frac{e^{\theta_r}}{\sum_{j=1}^M e^{\theta_j}} \\ &= \mathbf{1}\{r=m\} - p_\theta(r). \end{aligned}$$

Hence $\nabla \log p_\theta(m) = e_m - p_\theta$.

B.2 Proof of Theorem 4

Theorem 4 (Trajectory-based policy gradient). For every policy in the interior discounted-admissible region,

$$\nabla J(\theta) = \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[\sum_{t=0}^{\infty}\gamma^t G_t\left(e_{\sigma(t)} - p_\theta\right)\right].$$

Proof. Let $C(\tau) := \sum_{s=0}^{\infty}\gamma^s c_s$ denote the discounted cost of trajectory τ. The parameter-dependent part of the trajectory probability is $\prod_{t=0}^{\infty} p_\theta(\sigma(t))$. Therefore,

$$\begin{aligned} \nabla \log \mathbb{P}_\theta(\tau) &= \nabla \log\left(\prod_{t=0}^{\infty} p_\theta(\sigma(t))\right) \\ &= \nabla \sum_{t=0}^{\infty} \log p_\theta(\sigma(t)) \\ &= \sum_{t=0}^{\infty} \nabla \log p_\theta(\sigma(t)). \end{aligned}$$

Using the likelihood-ratio identity (Meyer (2018)),

$$\begin{aligned}\widehat{\nabla J}(\theta) &= \nabla \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[C(\tau)\right] \\ &= \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[C(\tau)\nabla \log \mathbb{P}_\theta(\tau)\right] \\ &= \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[\left(\sum_{s=0}^{\infty}\gamma^s c_s\right)\left(\sum_{t=0}^{\infty}\nabla \log p_\theta(\sigma(t))\right)\right] \\ &= \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[\sum_{t=0}^{\infty}\sum_{s=0}^{\infty}\gamma^s c_s \nabla \log p_\theta(\sigma(t))\right].\end{aligned}$$

For $s < t$, c_s is determined before $\sigma(t)$ is sampled. Hence

$$\mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[c_s \nabla \log p_\theta(\sigma(t))\right] = \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[c_s \mathbb{E}_{\sigma(t)\sim p_\theta}\left[\nabla \log p_\theta(\sigma(t)) \mid \mathcal{F}_t\right]\right].$$

where $\mathcal{F}_t$ denotes the trajectory history available before sampling $\sigma(t)$. Moreover,

$$\begin{aligned}\mathbb{E}_{\sigma(t)\sim p_\theta}\left[\nabla \log p_\theta(\sigma(t)) \mid \mathcal{F}_t\right] &= \sum_{m=1}^{M} p_\theta(m)\nabla \log p_\theta(m) \\ &= \sum_{m=1}^{M} \nabla p_\theta(m) \\ &= \nabla \sum_{m=1}^{M} p_\theta(m) \\ &= \nabla 1 \\ &= 0.\end{aligned}$$

Therefore, $\mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[c_s \nabla \log p_\theta(\sigma(t))\right] = 0, \qquad s < t.$

Hence

$$\widehat{\nabla J}(\theta) = \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[\sum_{t=0}^{\infty}\sum_{s=t}^{\infty}\gamma^s c_s \nabla \log p_\theta(\sigma(t))\right].$$

Now $\sum_{s=t}^{\infty}\gamma^s c_s = \gamma^t \sum_{s=t}^{\infty}\gamma^{s-t} c_s = \gamma^t G_t$. Thus

$$\begin{aligned}\widehat{\nabla J}(\theta) &= \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[\sum_{t=0}^{\infty}\gamma^t G_t \nabla \log p_\theta(\sigma(t))\right] \\ &= \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[\sum_{t=0}^{\infty}\gamma^t G_t \left(e_{\sigma(t)} - p_\theta\right)\right].\end{aligned}$$

□

B.3 Unbiasedness of the Batch Gradient Estimator

For trajectory i at iteration k, $Z_k^{(i)} := \sum_{t=0}^{\infty}\gamma^t G_t^{(i)}\left(e_{\sigma^{(i)}(t)} - p_{\theta_k}\right)$. Conditioned on $\mathcal{F}_k$, the policy parameter is fixed at θ_k. By Theorem 4, $\mathbb{E}_{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}}[Z_k^{(i)} \mid \mathcal{F}_k] = \nabla J(\theta_k)$. Since $\widehat{g}_k =$

$\frac{1}{N}\sum_{i=1}^{N} Z_k^{(i)}$, we have,

$$\begin{aligned}
\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\widehat{g}_k \mid \mathcal{F}_k\right] &= \mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\frac{1}{N}\sum_{i=1}^{N} Z_k^{(i)}\,\middle|\,\mathcal{F}_k\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}}\left[Z_k^{(i)} \mid \mathcal{F}_k\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\nabla J(\theta_k) \\
&= \frac{1}{N}N\nabla J(\theta_k) \\
&= \nabla J(\theta_k).
\end{aligned}$$

B.4 Proof of the Batch-Variance Bound equation 21

Define $\xi_k^{(i)} := Z_k^{(i)} - \nabla J(\theta_k)$. Then $\mathbb{E}_{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}}\left[\xi_k^{(i)} \mid \mathcal{F}_k\right] = 0$. Also define $e_k := \widehat{g}_k - \nabla J(\theta_k)$. Then

$$\begin{aligned}
e_k &= \frac{1}{N}\sum_{i=1}^{N} Z_k^{(i)} - \nabla J(\theta_k) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(Z_k^{(i)} - \nabla J(\theta_k)\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\xi_k^{(i)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\|e_k\|_2^2 \mid \mathcal{F}_k\right] \\
&= \frac{1}{N^2}\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\left\|\sum_{i=1}^{N}\xi_k^{(i)}\right\|_2^2\,\middle|\,\mathcal{F}_k\right] \\
&= \frac{1}{N^2}\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\left(\sum_{i=1}^{N}\xi_k^{(i)}\right)^{\top}\left(\sum_{j=1}^{N}\xi_k^{(j)}\right)\,\middle|\,\mathcal{F}_k\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}_{\{x_0^{(\ell)},\,\sigma^{(\ell)}(\cdot)\sim p_{\theta_k}\}_{\ell=1}^N}\left[(\xi_k^{(i)})^{\top}\xi_k^{(j)} \mid \mathcal{F}_k\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}}\left[\|\xi_k^{(i)}\|_2^2 \mid \mathcal{F}_k\right] \\
&\quad + \frac{1}{N^2}\sum_{i\neq j}\mathbb{E}_{\{x_0^{(\ell)},\,\sigma^{(\ell)}(\cdot)\sim p_{\theta_k}\}_{\ell=1}^N}\left[(\xi_k^{(i)})^{\top}\xi_k^{(j)} \mid \mathcal{F}_k\right].
\end{aligned}$$

For $i \neq j$, conditional independence of the trajectories and $\mathbb{E}[\xi_k^{(i)} \mid \mathcal{F}_k] = 0$ imply

$$\mathbb{E}[(\xi_k^{(i)})^{\top}\xi_k^{(j)} \mid \mathcal{F}_k] = 0.$$

Since $e_k = \widehat{g}_k - \nabla J(\theta_k)$, we obtain $\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\|\widehat{g}_k - \nabla J(\theta_k)\|_2^2\,\middle|\,\mathcal{F}_k\right] \leq \frac{\sigma_{\mathrm{tr}}^2}{N}$.

B.5 One-Step Stochastic Descent

Define $g_k := \nabla J(\theta_k)$ and $e_k := \widehat{g}_k - g_k$. Then $\widehat{g}_k = g_k + e_k$. The sampled update is $\theta_{k+1} = \theta_k - \eta_k \widehat{g}_k = \theta_k - \eta_k(g_k + e_k)$. By L-smoothness,

$$\begin{aligned} J(\theta_{k+1}) &\le J(\theta_k) + g_k^\top(\theta_{k+1} - \theta_k) + \frac{L}{2}\|\theta_{k+1} - \theta_k\|_2^2 \\ &= J(\theta_k) - \eta_k g_k^\top(g_k + e_k) + \frac{L\eta_k^2}{2}\|g_k + e_k\|_2^2 \\ &= J(\theta_k) - \eta_k\|g_k\|_2^2 - \eta_k g_k^\top e_k \\ &\quad + \frac{L\eta_k^2}{2}\left(\|g_k\|_2^2 + 2g_k^\top e_k + \|e_k\|_2^2\right). \end{aligned}$$

Taking conditional expectation,

$$\begin{aligned} &\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[J(\theta_{k+1}) \mid \mathcal{F}_k\right] \\ &\le J(\theta_k) - \eta_k\|g_k\|_2^2 - \eta_k g_k^\top \mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[e_k \mid \mathcal{F}_k\right] \\ &\quad + \frac{L\eta_k^2}{2}\Bigg(\|g_k\|_2^2 + 2g_k^\top \mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[e_k \mid \mathcal{F}_k\right] \\ &\qquad + \mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\|e_k\|_2^2 \mid \mathcal{F}_k\right]\Bigg). \end{aligned}$$

Since $\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[e_k \mid \mathcal{F}_k\right] = 0$, we obtain,

$$\begin{aligned} &\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[J(\theta_{k+1}) \mid \mathcal{F}_k\right] \\ &\le J(\theta_k) - \eta_k\|g_k\|_2^2 + \frac{L\eta_k^2}{2}\Bigg(\|g_k\|_2^2 \\ &\qquad + \mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\|e_k\|_2^2 \mid \mathcal{F}_k\right]\Bigg) \\ &\le J(\theta_k) - \eta_k\left(1 - \frac{L\eta_k}{2}\right)\|g_k\|_2^2 + \frac{L\sigma_{\mathrm{tr}}^2}{2N}\eta_k^2. \end{aligned}$$

Since $0 < \eta_k \le \frac{1}{L}$, we have $1 - \frac{L\eta_k}{2} \ge \frac{1}{2}$. Therefore,

$$\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[J(\theta_{k+1}) \mid \mathcal{F}_k\right] \le J(\theta_k) - \frac{\eta_k}{2}\|g_k\|_2^2 + \frac{L\sigma_{\mathrm{tr}}^2}{2N}\eta_k^2.$$

Since $\delta_k \le \bar{D}\|g_k\|_2$, we have $\|g_k\|_2 \ge \frac{\delta_k}{\bar{D}}$ and therefore $\|g_k\|_2^2 \ge \frac{\delta_k^2}{\bar{D}^2}$. Subtracting $J^\star$ gives

$$\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}\left[\delta_{k+1} \mid \mathcal{F}_k\right] \le \delta_k - \frac{\eta_k}{2\bar{D}^2}\delta_k^2 + \frac{L\sigma_{\mathrm{tr}}^2}{2N}\eta_k^2.$$

B.6 Proof of Theorem 5: Almost-Sure Convergence

Theorem 5 (Convergence with sampled policy gradients). Suppose the iterates remain in a region where J is L-smooth, Assumptions 1 and 2 hold, and $D(\theta_k) \le \bar{D} < \infty$. If

$$\eta_k = \frac{a}{(k+1)^\alpha}, \qquad 0 < a \le \frac{1}{L}, \qquad \frac{1}{2} < \alpha \le 1,$$

then

$$J(\theta_k) \longrightarrow J^\star \qquad \text{almost surely.}$$

Proof. Let $\eta_k = \frac{a}{(k+1)^\alpha}$ and $\frac{1}{2} < \alpha \leq 1$. Then

$$\begin{aligned}\sum_{k=0}^{\infty} \eta_k &= a \sum_{k=0}^{\infty} \frac{1}{(k+1)^\alpha} \\ &= a \sum_{r=1}^{\infty} \frac{1}{r^\alpha} \\ &= \infty.\end{aligned}$$

Also,

$$\begin{aligned}\sum_{k=0}^{\infty} \eta_k^2 &= a^2 \sum_{k=0}^{\infty} \frac{1}{(k+1)^{2\alpha}} \\ &= a^2 \sum_{r=1}^{\infty} \frac{1}{r^{2\alpha}} \\ &< \infty.\end{aligned}$$

The one-step stochastic descent inequality gives

$$\mathbb{E}_{\{x_0^{(i)},\, \sigma^{(i)}(\cdot) \sim p_{\theta_k}\}_{i=1}^N} [\delta_{k+1} \mid \mathcal{F}_k] \leq \delta_k - \frac{\eta_k}{2\bar{D}^2} \delta_k^2 + \frac{L\sigma_{\mathrm{tr}}^2}{2N} \eta_k^2.$$

Since

$$\begin{aligned}\sum_{k=0}^{\infty} \frac{L\sigma_{\mathrm{tr}}^2}{2N} \eta_k^2 &= \frac{L\sigma_{\mathrm{tr}}^2}{2N} \sum_{k=0}^{\infty} \eta_k^2 \\ &< \infty\end{aligned}$$

The Robbins–Siegmund almost-supermartingale theorem (Neri & Powell (2026)) implies that $\delta_k \longrightarrow \delta_\infty$ almost surely for some finite random variable $\delta_\infty \geq 0$, and $\sum_{k=0}^{\infty} \frac{\eta_k}{2\bar{D}^2} \delta_k^2 < \infty$ almost surely. Multiplying by $2\bar{D}^2$, $\sum_{k=0}^{\infty} \eta_k \delta_k^2 < \infty$ almost surely. Suppose $\delta_\infty > 0$ on an event of positive probability. Since $\delta_k \longrightarrow \delta_\infty$, there exists a finite random index k_0 such that

$$\delta_k \geq \frac{\delta_\infty}{2}, \qquad k \geq k_0.$$

Therefore, $\delta_k^2 \geq \frac{\delta_\infty^2}{4}$ for all $k \geq k_0$. Hence

$$\begin{aligned}\sum_{k=k_0}^{\infty} \eta_k \delta_k^2 &\geq \sum_{k=k_0}^{\infty} \eta_k \frac{\delta_\infty^2}{4} \\ &= \frac{\delta_\infty^2}{4} \sum_{k=k_0}^{\infty} \eta_k \\ &= \infty.\end{aligned}$$

This contradicts $\sum_{k=0}^{\infty} \eta_k \delta_k^2 < \infty$. Therefore, $\delta_\infty = 0$ almost surely. Hence $J(\theta_k) - J^\star \longrightarrow 0$ almost surely, and therefore $J(\theta_k) \longrightarrow J^\star$ almost surely.

□

B.7 Proof of Theorem 5: Finite-Time Bound

Theorem 5 (Convergence with sampled policy gradients). Suppose the iterates remain in a region where J is L-smooth, Assumptions 1 and 2 hold, and $D(\theta_k) \leq \bar{D} < \infty$. Define

$$U_{k_{\mathrm{f}}} := \sum_{k=0}^{k_{\mathrm{f}}-1} \eta_k, \qquad W_{k_{\mathrm{f}}} := \sum_{k=0}^{k_{\mathrm{f}}-1} \eta_k^2.$$

Then

$$\min_{0\le k<k_{\mathrm{f}}} \mathbb{E}[\delta_k^2] \le \frac{2\bar{D}^2\delta_0}{U_{k_{\mathrm{f}}}} + \frac{\bar{D}^2 L\sigma_{\mathrm{tr}}^2}{N}\frac{W_{k_{\mathrm{f}}}}{U_{k_{\mathrm{f}}}}.$$

Proof. Starting from $\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}[\delta_{k+1} \mid \mathcal{F}_k] \le \delta_k - \frac{\eta_k}{2\bar{D}^2}\delta_k^2 + \frac{L\sigma_{\mathrm{tr}}^2}{2N}\eta_k^2$. Taking total expectation:

$$\begin{aligned}\mathbb{E}[\delta_{k+1}] &= \mathbb{E}\left[\mathbb{E}_{\{x_0^{(i)},\,\sigma^{(i)}(\cdot)\sim p_{\theta_k}\}_{i=1}^N}[\delta_{k+1} \mid \mathcal{F}_k]\right] \\ &\le \mathbb{E}[\delta_k] - \frac{\eta_k}{2\bar{D}^2}\mathbb{E}[\delta_k^2] + \frac{L\sigma_{\mathrm{tr}}^2}{2N}\eta_k^2.\end{aligned}$$

Rearranging and summing from $k=0$ to $k_{\mathrm{f}}-1$,

$$\begin{aligned}\frac{1}{2\bar{D}^2}\sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k\mathbb{E}[\delta_k^2] &\le \sum_{k=0}^{k_{\mathrm{f}}-1}(\mathbb{E}[\delta_k] - \mathbb{E}[\delta_{k+1}]) + \frac{L\sigma_{\mathrm{tr}}^2}{2N}\sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k^2 \\ &= \mathbb{E}[\delta_0] - \mathbb{E}[\delta_{k_{\mathrm{f}}}] + \frac{L\sigma_{\mathrm{tr}}^2}{2N}W_{k_{\mathrm{f}}}.\end{aligned}$$

Since θ_0 is fixed, $\mathbb{E}[\delta_0] = \delta_0$. Since $\delta_{k_{\mathrm{f}}} \ge 0$, we have $-\mathbb{E}[\delta_{k_{\mathrm{f}}}] \le 0$. Therefore,

$$\frac{1}{2\bar{D}^2}\sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k\mathbb{E}[\delta_k^2] \le \delta_0 + \frac{L\sigma_{\mathrm{tr}}^2}{2N}W_{k_{\mathrm{f}}}.$$

Multiplying by $2\bar{D}^2$,

$$\sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k\mathbb{E}[\delta_k^2] \le 2\bar{D}^2\delta_0 + \frac{\bar{D}^2 L\sigma_{\mathrm{tr}}^2}{N}W_{k_{\mathrm{f}}}.$$

For every $k < k_{\mathrm{f}}$, $\mathbb{E}[\delta_k^2] \ge \min_{0\le j<k_{\mathrm{f}}}\mathbb{E}[\delta_j^2]$. Since $\eta_k > 0$, $\eta_k\mathbb{E}[\delta_k^2] \ge \eta_k \min_{0\le j<k_{\mathrm{f}}}\mathbb{E}[\delta_j^2]$. Summing,

$$\begin{aligned}\sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k\mathbb{E}[\delta_k^2] &\ge \sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k \min_{0\le j<k_{\mathrm{f}}}\mathbb{E}[\delta_j^2] \\ &= \left(\sum_{k=0}^{k_{\mathrm{f}}-1}\eta_k\right)\min_{0\le j<k_{\mathrm{f}}}\mathbb{E}[\delta_j^2] \\ &= U_{k_{\mathrm{f}}}\min_{0\le j<k_{\mathrm{f}}}\mathbb{E}[\delta_j^2].\end{aligned}$$

Thus $U_{k_{\mathrm{f}}}\min_{0\le k<k_{\mathrm{f}}}\mathbb{E}[\delta_k^2] \le 2\bar{D}^2\delta_0 + \frac{\bar{D}^2 L\sigma_{\mathrm{tr}}^2}{N}W_{k_{\mathrm{f}}}$. Dividing by $U_{k_{\mathrm{f}}} > 0$,

$$\min_{0\le k<k_{\mathrm{f}}} \mathbb{E}[\delta_k^2] \le \frac{2\bar{D}^2\delta_0}{U_{k_{\mathrm{f}}}} + \frac{\bar{D}^2 L\sigma_{\mathrm{tr}}^2}{N}\frac{W_{k_{\mathrm{f}}}}{U_{k_{\mathrm{f}}}}.$$

□

Remark 2. Here, conditional expectations given $\mathcal{F}_k$ are taken over the fresh trajectory samples generated at iteration k under p_{θ_k}, while expectations without conditioning are taken over all randomness in the learning process up to the corresponding iteration.

C Proofs for Stability-Preserving Policy Learning

C.1 Proof of Theorem 6

Theorem 6 (Mean-square stability of a fixed policy). The randomized switched system is mean-square stable if

$$\rho(\mathcal{M}_\theta) = \rho\left(\sum_{m=1}^{M} p_\theta(m)(F_m \otimes F_m)\right) < 1.$$

In particular,

$$\lim_{t\to\infty} \mathbb{E}\|x_t\|_2^2 = 0.$$

Proof. For a fixed switching distribution p_θ,

$$\begin{aligned}\Sigma_{t+1} &= \mathbb{E}\left[x_{t+1}x_{t+1}^\top\right] \\ &= \mathbb{E}\left[F_{\sigma(t)}x_t x_t^\top F_{\sigma(t)}^\top\right] \\ &= \sum_{m=1}^{M} p_\theta(m) F_m \Sigma_t F_m^\top.\end{aligned}$$

Vectorizing, $\operatorname{vec}(\Sigma_{t+1}) = \sum_{m=1}^{M} p_\theta(m)(F_m \otimes F_m)$. We get $\operatorname{vec}(\Sigma_t) = \mathcal{M}_\theta \operatorname{vec}(\Sigma_t)$. Therefore, $\operatorname{vec}(\Sigma_t) = \mathcal{M}_\theta^t \operatorname{vec}(\Sigma_0)$. If $\rho(\mathcal{M}_\theta) < 1$, then $\mathcal{M}_\theta^t \to 0$ as $t \to \infty$. Hence $\operatorname{vec}(\Sigma_t) = \mathcal{M}_\theta^t \operatorname{vec}(\Sigma_0) \to 0$, and therefore $\Sigma_t \to 0$.

Since

$$\begin{aligned}\mathbb{E}\|x_t\|_2^2 &= \mathbb{E}\left[x_t^\top x_t\right] \\ &= \mathbb{E}\left[\operatorname{Tr}(x_t x_t^\top)\right] \\ &= \operatorname{Tr}\left(\mathbb{E}[x_t x_t^\top]\right) \\ &= \operatorname{Tr}(\Sigma_t),\end{aligned}$$

we obtain $\lim_{t\to\infty} \mathbb{E}\|x_t\|_2^2 = 0$.

□

C.2 Existence of the Stochastic Lyapunov Matrix

Let $\mathcal{T}_k^*(S) := \sum_{m=1}^{M} p_k(m) F_m^\top S F_m$. Vectorizing,

$$\begin{aligned}\operatorname{vec}\left(\mathcal{T}_k^*(S)\right) &= \sum_{m=1}^{M} p_k(m)(F_m^\top \otimes F_m^\top)\operatorname{vec}(S) \\ &= \mathcal{M}_{\theta_k}^\top \operatorname{vec}(S).\end{aligned}$$

Since p_k is mean-square stable, $\rho(\mathcal{M}_{\theta_k}) < 1$. Therefore, $\rho(\mathcal{M}_{\theta_k}^\top) < 1$, and $I - \mathcal{M}_{\theta_k}^\top$ is nonsingular. For any $\mathcal{Q}_k \succ 0$, define $\operatorname{vec}(S_k) := (I - \mathcal{M}_{\theta_k}^\top)^{-1}\operatorname{vec}(\mathcal{Q}_k)$. Then $(I - \mathcal{M}_{\theta_k}^\top)\operatorname{vec}(S_k) = \operatorname{vec}(\mathcal{Q}_k)$, so $\operatorname{vec}(S_k) - \mathcal{M}_{\theta_k}^\top \operatorname{vec}(S_k) = \operatorname{vec}(\mathcal{Q}_k)$. Therefore,

$$\operatorname{vec}(S_k) - \sum_{m=1}^{M} p_k(m)\operatorname{vec}\left(F_m^\top S_k F_m\right) = \operatorname{vec}(\mathcal{Q}_k),$$

which gives

$$S_k - \sum_{m=1}^{M} p_k(m) F_m^\top S_k F_m = \mathcal{Q}_k.$$

Also,

$$\operatorname{vec}(S_k) = \sum_{r=0}^{\infty} (\mathcal{M}_{\theta_k}^{\top})^r \operatorname{vec}(\mathcal{Q}_k).$$

Equivalently,

$$S_k = \mathcal{Q}_k + \mathcal{T}_k^*(\mathcal{Q}_k) + (\mathcal{T}_k^*)^2(\mathcal{Q}_k) + \cdots .$$

Every term is positive semidefinite and $\mathcal{Q}_k \succ 0$. Hence $S_k \succeq \mathcal{Q}_k \succ 0$, so $S_k \succ 0$.

C.3 Lyapunov Drift Under the Current Policy

Let $\mathcal{V}_k(x) := x^{\top} S_k x$. Then

$$\begin{aligned}
\mathbb{E}\left[\mathcal{V}_k(x_{t+1}) \mid x_t\right] &= \sum_{m=1}^{M} p_k(m)(F_m x_t)^{\top} S_k (F_m x_t) \\
&= \sum_{m=1}^{M} p_k(m) x_t^{\top} F_m^{\top} S_k F_m x_t \\
&= x_t^{\top} \left(\sum_{m=1}^{M} p_k(m) F_m^{\top} S_k F_m \right) x_t.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{V}_k(x_{t+1}) \mid x_t\right] - \mathcal{V}_k(x_t) &= x_t^{\top} \left(\sum_{m=1}^{M} p_k(m) F_m^{\top} S_k F_m - S_k \right) x_t \\
&= -x_t^{\top} \mathcal{Q}_k x_t.
\end{aligned}$$

C.4 Lyapunov Perturbation Under a Policy Update

Define $\Delta p_{k,m} := p_{k+1}(m) - p_k(m)$. Then $p_{k+1}(m) = p_k(m) + \Delta p_{k,m}$. Hence

$$\begin{aligned}
S_k - \sum_{m=1}^{M} p_{k+1}(m) F_m^{\top} S_k F_m &= S_k - \sum_{m=1}^{M} \left(p_k(m) + \Delta p_{k,m}\right) F_m^{\top} S_k F_m \\
&= S_k - \sum_{m=1}^{M} p_k(m) F_m^{\top} S_k F_m \\
&\quad - \sum_{m=1}^{M} \Delta p_{k,m} F_m^{\top} S_k F_m \\
&= \mathcal{Q}_k - E_k,
\end{aligned}$$

where $E_k := \sum_{m=1}^{M} \Delta p_{k,m} F_m^{\top} S_k F_m$. For any $x \neq 0$,

$$\begin{aligned}
x^{\top}(\mathcal{Q}_k - E_k)x &= x^{\top}\mathcal{Q}_k x - x^{\top} E_k x \\
&\geq \lambda_{\min}(\mathcal{Q}_k)\|x\|_2^2 - |x^{\top} E_k x| \\
&\geq \lambda_{\min}(\mathcal{Q}_k)\|x\|_2^2 - \|E_k\|_2 \|x\|_2^2 \\
&= \left(\lambda_{\min}(\mathcal{Q}_k) - \|E_k\|_2\right)\|x\|_2^2.
\end{aligned}$$

Thus, if $\|E_k\|_2 < \lambda_{\min}(\mathcal{Q}_k)$, then $x^{\top}(\mathcal{Q}_k - E_k)x > 0$ and therefore $\mathcal{Q}_k - E_k \succ 0$.

C.5 Proof of Lemma 3

Lemma 3 (Softmax policy perturbation). For any $\theta, \vartheta \in \mathbb{R}^M$,

$$\|p_\theta - p_\vartheta\|_1 \leq \frac{1}{2}\|\theta - \vartheta\|_1.$$

Proof. The softmax Jacobian has entries $[J_{\mathrm{sm}}(\theta)]_{ij} = \frac{\partial p_\theta(i)}{\partial \theta_j}$. From the softmax derivative, $\frac{\partial p_\theta(i)}{\partial \theta_j} = p_\theta(i)\left(\mathbf{1}\{i=j\} - p_\theta(j)\right)$. For $i = j$, $\frac{\partial p_\theta(j)}{\partial \theta_j} = p_\theta(j)\left(1 - p_\theta(j)\right)$. For $i \neq j$, $\frac{\partial p_\theta(i)}{\partial \theta_j} = -p_\theta(i)p_\theta(j)$. The induced matrix 1-norm is

$$\|J_{\mathrm{sm}}(\theta)\|_1 = \max_j \sum_{i=1}^{M} \left|\frac{\partial p_\theta(i)}{\partial \theta_j}\right|.$$

For a fixed column j,

$$\begin{aligned}
\sum_{i=1}^{M} \left|\frac{\partial p_\theta(i)}{\partial \theta_j}\right| &= p_\theta(j)\left(1 - p_\theta(j)\right) + \sum_{i \neq j} p_\theta(i)p_\theta(j) \\
&= p_\theta(j)\left(1 - p_\theta(j)\right) + p_\theta(j)\sum_{i \neq j} p_\theta(i) \\
&= p_\theta(j)\left(1 - p_\theta(j)\right) + p_\theta(j)\left(1 - p_\theta(j)\right) \\
&= 2p_\theta(j)\left(1 - p_\theta(j)\right).
\end{aligned}$$

For $p \in [0,1]$, $2p(1-p) = 2p - 2p^2$. Differentiating gives $\frac{d}{dp}(2p - 2p^2) = 2 - 4p$. Setting $2 - 4p = 0$ gives $p = \frac{1}{2}$. Therefore, $2p(1-p) \leq 2\left(\frac{1}{2}\right)\left(1 - \frac{1}{2}\right) = \frac{1}{2}$. Hence $\|J_{\mathrm{sm}}(\theta)\|_1 \leq \frac{1}{2}$. Now define $\theta(s) := \vartheta + s(\theta - \vartheta)$, where $s \in [0,1]$. Then $\theta(0) = \vartheta, \theta(1) = \theta$, and $\frac{d\theta(s)}{ds} = \theta - \vartheta$. By the fundamental theorem of calculus,

$$\begin{aligned}
p_\theta - p_\vartheta &= p_{\theta(1)} - p_{\theta(0)} \\
&= \int_0^1 \frac{d}{ds} p_{\theta(s)}\, ds \\
&= \int_0^1 J_{\mathrm{sm}}(\theta(s))\frac{d\theta(s)}{ds}\, ds \\
&= \int_0^1 J_{\mathrm{sm}}(\theta(s))(\theta - \vartheta)\, ds.
\end{aligned}$$

Taking the 1-norm,

$$\begin{aligned}
\|p_\theta - p_\vartheta\|_1 &\leq \int_0^1 \|J_{\mathrm{sm}}(\theta(s))(\theta - \vartheta)\|_1\, ds \\
&\leq \int_0^1 \|J_{\mathrm{sm}}(\theta(s))\|_1 \|\theta - \vartheta\|_1\, ds \\
&\leq \int_0^1 \frac{1}{2}\|\theta - \vartheta\|_1\, ds \\
&= \frac{1}{2}\|\theta - \vartheta\|_1.
\end{aligned}$$

□

C.6 Bound on the Lyapunov Perturbation

Recall $E_k = \sum_{m=1}^{M} \Delta p_{k,m} F_m^\top S_k F_m$. Therefore,

$$\begin{aligned}\|E_k\|_2 &= \left\| \sum_{m=1}^{M} \Delta p_{k,m} F_m^\top S_k F_m \right\|_2 \\ &\leq \sum_{m=1}^{M} |\Delta p_{k,m}| \|F_m^\top S_k F_m\|_2 \\ &\leq H_k \sum_{m=1}^{M} |\Delta p_{k,m}| \\ &= H_k \|\Delta p_k\|_1.\end{aligned}$$

By Lemma 3,

$$\begin{aligned}\|\Delta p_k\|_1 &= \|p_{k+1} - p_k\|_1 \\ &\leq \frac{1}{2}\|\theta_{k+1} - \theta_k\|_1.\end{aligned}$$

Since $\theta_{k+1} = \theta_k - \eta_k g_k$, we have $\theta_{k+1} - \theta_k = -\eta_k g_k$ and therefore $\|\theta_{k+1} - \theta_k\|_1 = \eta_k \|g_k\|_1$. Hence $\|\Delta p_k\|_1 \leq \frac{\eta_k}{2}\|g_k\|_1$. Therefore,

$$\begin{aligned}\|E_k\|_2 &\leq H_k \|\Delta p_k\|_1 \\ &\leq \frac{\eta_k H_k}{2}\|g_k\|_1.\end{aligned}$$

C.7 Proof of Theorem 7

Theorem 7 (Stability-preserving policy update). Suppose p_k is mean-square stable and S_k satisfies equation 25. The update $\theta_{k+1} = \theta_k - \eta_k g_k$ preserves mean-square stability if

$$\eta_k < \frac{2\lambda_{\min}(\mathcal{Q}_k)}{H_k \|g_k\|_1}.$$

Proof. For the updated policy,

$$S_k - \sum_{m=1}^{M} p_{k+1}(m) F_m^\top S_k F_m = \mathcal{Q}_k - E_k.$$

A sufficient condition for mean-square stability of p_{k+1} is $\mathcal{Q}_k - E_k \succ 0$. For any $x \neq 0$,

$$\begin{aligned}x^\top(\mathcal{Q}_k - E_k)x &= x^\top \mathcal{Q}_k x - x^\top E_k x \\ &\geq \lambda_{\min}(\mathcal{Q}_k)\|x\|_2^2 - |x^\top E_k x| \\ &\geq \lambda_{\min}(\mathcal{Q}_k)\|x\|_2^2 - \|E_k\|_2\|x\|_2^2 \\ &= (\lambda_{\min}(\mathcal{Q}_k) - \|E_k\|_2)\,\|x\|_2^2.\end{aligned}$$

Therefore, $\|E_k\|_2 < \lambda_{\min}(\mathcal{Q}_k) \quad \Rightarrow \quad \mathcal{Q}_k - E_k \succ 0$. From the perturbation bound, $\|E_k\|_2 \leq \frac{\eta_k H_k}{2}\|g_k\|_1$. Hence,

$$\begin{aligned}&\frac{\eta_k H_k}{2}\|g_k\|_1 < \lambda_{\min}(\mathcal{Q}_k) \\ \Rightarrow \quad &\eta_k H_k \|g_k\|_1 < 2\lambda_{\min}(\mathcal{Q}_k) \\ \Rightarrow \quad &\eta_k < \frac{2\lambda_{\min}(\mathcal{Q}_k)}{H_k\|g_k\|_1}.\end{aligned}$$

Thus, if $\eta_k < \frac{2\lambda_{\min}(\mathcal{Q}_k)}{H_k\|g_k\|_1}$, then $\|E_k\|_2 < \lambda_{\min}(\mathcal{Q}_k)$, which implies $\mathcal{Q}_k - E_k \succ 0$. Therefore, $S_k - \sum_{m=1}^{M} p_{k+1}(m)F_m^\top S_k F_m \succ 0$, and so p_{k+1} is mean-square stable. If $g_k = 0$, then $\theta_{k+1} = \theta_k$, hence $p_{k+1} = p_k$, and mean-square stability is preserved.

□

D Experiments and Results

We evaluate the proposed StableSwitch-PG framework on a randomized switched linear CartPole system. The purpose of the experiment is to compare the learned controller-selection distributions, optimization performance, and resulting closed-loop state behavior obtained using three policy-gradient schemes: Exact PG, Model-free PG, and the proposed StableSwitch-PG. The controller library is fixed throughout training; the learning problem therefore consists only of determining the probability with which each feedback controller is selected.

D.1 CartPole Environment

The CartPole system is linearized about the upright equilibrium. The state is defined as $x_t = \begin{bmatrix} z_t & \dot{z}_t & \phi_t & \dot{\phi}_t \end{bmatrix}^\top$, where z_t is the horizontal cart position, $\dot{z}_t$ is the cart velocity, ϕ_t is the pole-angle deviation from the upright equilibrium, and $\dot{\phi}_t$ is the pole angular velocity. Thus, the state dimension is four. The continuous-time linearized dynamics are $\dot{x}(t) = A_c x(t) + B_c u(t)$, with

$$A_c = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{g(M+m)}{M\ell} & 0 \end{bmatrix}, \qquad B_c = \begin{bmatrix} 0 \\ \frac{1}{M} \\ 0 \\ \frac{1}{M\ell} \end{bmatrix}.$$

The physical parameters used in simulation are $M = 1.0$ kg, $m = 0.1$ kg, $\ell = 0.5$ m, and $g = 9.81$ m/s^2. The continuous-time model is discretized exactly using the matrix exponential with sampling period $\Delta t = 0.02$ s. Each rollout contains $H = 500$ discrete time steps, corresponding to a simulation duration of $T = H\Delta t = 10$ s. At each time instant, one controller is randomly selected according to the probability vector $p_\theta = \begin{bmatrix} p_1(\theta) & p_2(\theta) & p_3(\theta) \end{bmatrix}^\top$, where the softmax parameterization $p_m(\theta) = \frac{\exp(\theta_m)}{\sum_{j=1}^{3}\exp(\theta_j)}$ is used. The switched closed-loop system evolves according to $x_{t+1} = F_{\sigma(t)}x_t$. Controller selection is therefore performed once every $\Delta t = 0.02$ s.

All methods are initialized from $\theta_0 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^\top$, which corresponds to the uniform initial controller-selection distribution $p_{\theta_0} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}^\top$. The initial states are sampled according to $x_0 \sim \mathcal{N}(0, \Sigma_0)$, where $\Sigma_0 = \mathrm{diag}\Big(0.30^2,\, 0.15^2,\, \big(\frac{10\pi}{180}\big)^2,\, \big(\frac{8\pi}{180}\big)^2\Big)$. Thus, the initial-state standard deviations correspond to 0.30 m in cart position, 0.15 m/s in cart velocity, 10° in pole angle, and 8°/s in pole angular velocity.

D.2 Controller Library

The switching library contains three fixed state-feedback controllers. Their gains are given in Table 1. The controller library was selected to provide distinct and complementary closed-loop behaviors rather than three nearly identical feedback laws. In particular, K_1 provides comparatively strong pole damping, K_2 emphasizes cart regulation, and K_3 provides a more balanced response. The three controllers are fixed before learning; neither Exact PG, Model-free PG, nor StableSwitch-PG modifies their feedback gains. Instead, each method learns only the switching distribution p_θ over this common controller library. Each individual closed-loop matrix is Schur stable. For the gains in Table 1, the corresponding spectral radii are approximately $\rho(F_1) = 0.9986$, $\rho(F_2) = 0.9968$, and $\rho(F_3) = 0.9989$.

Table 1: Controller library used in the CartPole experiment. The state ordering is $x = [z, \dot{z}, \phi, \dot{\phi}]^\top$.

Controller	K_z	$K_{\dot{z}}$	K_ϕ	$K_{\dot{\phi}}$	Role in the library
K_1	-2.5265	-3.2413	31.4591	20.7105	Pole-damping-oriented
K_2	-8.1660	-3.2267	29.4933	5.2406	Cart-regulation-oriented
K_3	-0.1163	-2.2583	23.1730	2.3790	Balanced controller

Table 2: Mean-cost comparison of controller-selection methods.

Method	p_1	p_2	p_3	Mean Cost
K_1	1.000	0.000	0.000	111.269
K_2	0.000	1.000	0.000	127.102
K_3	0.000	0.000	1.000	99.120
Equal Mixture	0.333	0.333	0.333	93.849
Exact PG	0.103	0.115	0.782	49.784
Model-free PG	0.106	0.108	0.786	49.730
StableSwitch-PG	0.167	0.189	0.643	54.140

D.3 Performance Objective and Cost Evaluation

For the CartPole experiments, the performance cost was evaluated using the same quadratic objective defined in equation 3, with $Q = \mathrm{diag}(1,\ 0.1,\ 12,\ 0.2)$, $R = 0.05$, and discount factor $\gamma = 0.995$. Since each rollout has a finite horizon of $H = 500$ steps, the cost was approximated by the truncated discounted sum $J_H(\theta) = \mathbb{E}_{x_0,\,\sigma(\cdot)\sim p_\theta}\left[\sum_{t=0}^{H-1} \gamma^t x_t^\top C_{\sigma(t)} x_t\right]$. At every training iteration, this quantity was estimated using $N_{\mathrm{eval}} = 2000$ independent Monte Carlo trajectories, and the reported cost corresponds to the average over those evaluation trajectories. The reported expected cost is estimated independently from the gradient calculation using Monte Carlo rollouts. Specifically, at training iteration k,

$$\widehat{J}_k = \frac{1}{N_{\mathrm{eval}}} \sum_{n=1}^{N_{\mathrm{eval}}} J_k^{(n)}.$$

Here, $N_{\mathrm{eval}} = 2000$. Thus, the trajectories used to evaluate the displayed performance are fresh evaluation trajectories and are not the same trajectories used to construct the model-free gradient estimate.

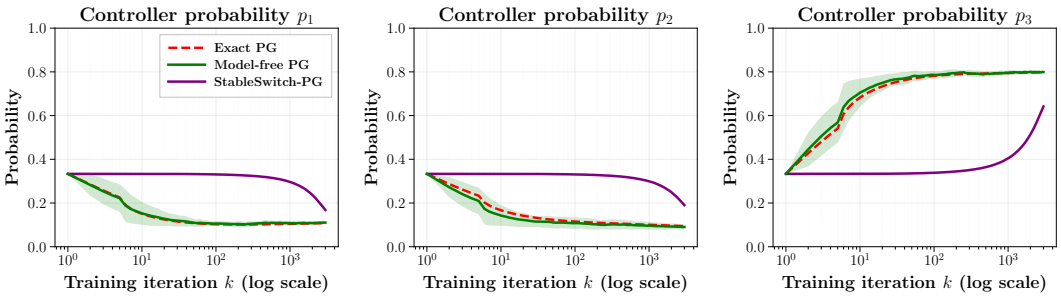
D.4 Learning Configuration

Three learning schemes are compared. Exact PG computes the finite-horizon policy gradient using the known linear system model. Model-free PG uses a trajectory-based REINFORCE gradient estimate, while StableSwitch-PG uses the same model-free gradient estimator together with the model-based stability-preserving step-size rule developed in this work. For both stochastic methods, the gradient estimator uses $N_{\mathrm{grad}} = 128$ sampled trajectories per policy update. Model-free PG and StableSwitch-PG are each repeated over 20 independent training runs in order to quantify the variability induced by trajectory sampling. For Exact PG and Model-free PG, the nominal learning-rate schedule is $\eta_k = \frac{\eta_0}{(k+1)^\alpha}$, where $\eta_0 = 0.01$ and $\alpha = 0.6$. For the stability certificate, we set $\mathcal{Q}_k = I$, so that $\lambda_{\min}(\mathcal{Q}_k) = 1$. It uses the stability-preserving learning rate $\eta_k = \frac{2\beta}{H_k \|\widehat{g}_k\|_1}$, where $\widehat{g}_k$ is the model-free policy-gradient estimate and $\beta = 0.98$. The quantity H_k is obtained from the current stochastic Lyapunov certificate and therefore adapts the allowable policy displacement according to the current switching distribution.

Table 2 summarizes the final controller-selection probabilities and corresponding mean costs. Among the individual controllers, K_3 gives the lowest mean cost, while the equal mixture further

Table 3: Learning configuration used in the CartPole experiment.

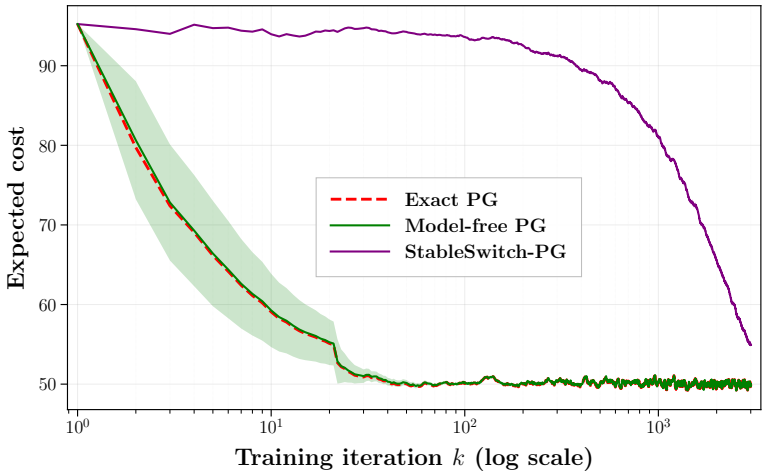
Method	Gradient	N_{grad}	Updates	Runs
Exact PG	Exact	–	3000	1
Model-free PG	REINFORCE	128	3000	20
StableSwitch-PG	REINFORCE	128	3000	20



(a) Probability p_1.

(b) Probability p_2.

(c) Probability p_3.



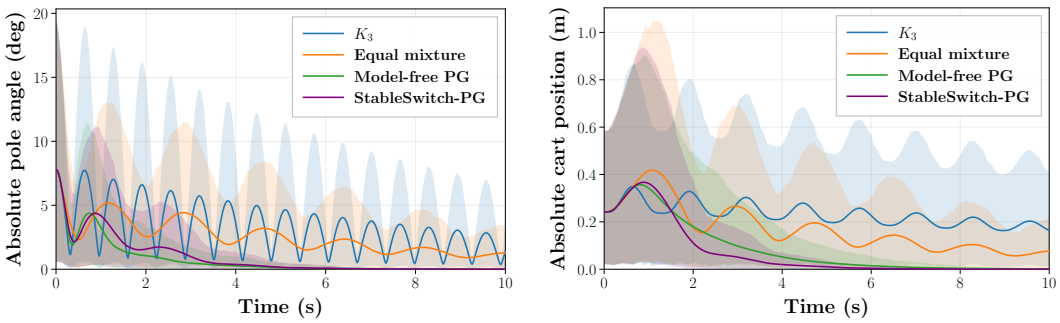
(d) Expected discounted cost during training.

Figure 1: Training behavior of Exact PG, Model-free PG, and StableSwitch-PG. (a)–(c) Evolution of the controller-selection probabilities p_1, p_2, and p_3, respectively. All methods begin from the uniform switching distribution $p_1 = p_2 = p_3 = 1/3$. Exact PG is deterministic and is therefore displayed as a single curve. For Model-free PG and StableSwitch-PG, the solid curves denote the mean over 20 independent training runs and the shaded regions denote one standard deviation across those runs. (d) Monte Carlo estimate of the expected discounted cost.

improves performance. Exact PG and Model-free PG converge to very similar controller-selection distributions and achieve nearly identical mean costs. StableSwitch-PG converges to a more conservative distribution because its update magnitude is restricted by the stability-preserving rule, while still achieving a substantially lower cost than the individual-controller and equal-mixture baselines. The main learning parameters are summarized in Table 3.

D.5 Probability Convergence and Learning Performance

Figure 1 shows the evolution of the learned switching distributions and the corresponding expected cost. Starting from the equal switching distribution, Exact PG and Model-free PG rapidly move probability mass toward controller K_3. After 3000 updates, the Exact-PG distribution is approximately $p_{\text{Exact}} \approx [0.103 \quad 0.115 \quad 0.782]^\top$, while the mean Model-free-PG distribution is approximately $p_{\text{MF}} \approx [0.106 \quad 0.108 \quad 0.786]^\top$. The close agreement between these two distributions indicates that the trajectory-based model-free gradient recovers essentially the same controller pref-



(a) Absolute pole-angle response.

(b) Absolute cart-position response.

Figure 2: Post-training closed-loop state responses for the fixed controller K_3, equal-probability switching, the final Model-free-PG switching distribution, and the final StableSwitch-PG distribution. Each case is evaluated using 1000 Monte Carlo trajectories over the 10 s simulation horizon. The same initial-state sampling seed is used across the four policies so that they are evaluated from the same set of initial conditions. The solid curve denotes the mean absolute state over the 1000 trajectories. The shaded envelope denotes the 5th–95th percentile range of the absolute state and therefore shows the central 90% of the simulated responses.

erence as the model-based exact gradient. In particular, both methods assign approximately 80% of the switching probability to K_3 while retaining smaller probabilities for K_1 and K_2. StableSwitch-PG evolves substantially more slowly because each policy update is restricted by the stability-preserving learning-rate rule. After 3000 updates, its mean switching distribution is approximately $p_{\text{StableSwitch}} \approx [0.167 \quad 0.189 \quad 0.643]^\top$. Thus, StableSwitch-PG ultimately develops the same qualitative preference for K_3, but it maintains a more distributed switching policy over the three controllers. The slower probability evolution is the expected consequence of restricting the admissible step size in order to preserve the switching-stability certificate during learning.

The same behavior is reflected in the expected-cost trajectory in Fig. 1d. All methods begin from the same uniform policy. Exact PG and Model-free PG decrease the cost rapidly and converge earlier than the StableSwitch-PG. Their cost trajectories and probability trajectories remain close throughout training, providing an empirical validation of the model-free gradient estimator with $N_{\text{grad}} = 128$.

StableSwitch-PG exhibits a slower transient because of its conservative stability-aware updates. After 3000 iterations its expected cost is approximately 54.2. Hence, the stability-preserving procedure incurs a clear convergence-rate tradeoff.

D.6 Post-Training Closed-Loop Response

Figure 2 compares the final closed-loop behavior obtained from four switching strategies: using K_3 alone, uniform random switching, the final Model-free-PG policy, and the final StableSwitch-PG policy. The comparison is performed over 1000 independent CartPole trajectories and therefore characterizes both the average behavior and the trajectory-to-trajectory dispersion.

The fixed K_3 controller stabilizes the CartPole, but its absolute cart-position and pole-angle responses exhibit persistent oscillatory behavior over the 10 s horizon. Equal-probability switching also produces a stable response, although both the mean response and its dispersion decay relatively slowly. Both learned switching distributions improve the transient response. The Model-free-PG policy substantially reduces the cart-position and pole-angle errors compared with the fixed-controller and uniform switching baselines. StableSwitch-PG also produces rapid decay of both displayed states. In particular, the mean absolute cart position under StableSwitch-PG approaches zero rapidly and its percentile envelope contracts substantially as the trajectory evolves.

It is important that these state plots and the quadratic cost measure different aspects of performance. The objective $J(\theta)$ penalizes all four state components as well as the control input through R. Consequently, a method that produces a smaller cart-position or pole-angle curve in Fig. 2 does not necessarily have the smallest total discounted cost. This explains why the state-response comparison should be interpreted together with the cost results of Fig. 1d, rather than as a separate cost ranking.

Overall, the results demonstrate three distinct properties of the learning framework. First, the model-free REINFORCE estimator closely reproduces the switching distribution obtained using the exact model-based gradient when $N_{\mathrm{grad}} = 128$. Second, learning a probabilistic mixture of the controller library produces improved closed-loop behavior relative to fixed-controller and uniform-switching baselines. Third, StableSwitch-PG reaches a useful low-cost switching distribution while employing substantially more conservative policy updates. The resulting slower convergence is the empirical price of enforcing the stability-preserving update mechanism during learning.