A Dual-SPMA Framework for Convex MDPs

Shervin Khamooshian Ahmed Magd Pegah Aryadoost Danielle Nguyen
Simon Fraser University
{ska309, ams80, paa40, tdn8}@sfu.ca

1 Introduction

Goal and Approach. We aim to solve *Convex Markov Decision Processes (cMDPs)*, which generalize standard RL by allowing a convex objective $f(d_\pi)$ over the discounted state-action occupancy measure $d_\pi(s,a) = (1-\gamma)\sum_{t\geq 0}\gamma^t \Pr_\pi(s_t = s, a_t = a)$, where s_t and a_t are the state and action at time t under policy π, and $\gamma \in [0,1)$ is the discount factor. This general formulation subsumes imitation/occupancy matching, traditional *Constrained MDPs (CMDPs)*, and exploration.

A Tractable Reformulation. A convenient way to solve this problem is via its *Fenchel–dual* reformulation, $\min_{d\in\mathcal{K}} f(d) = \min_\pi \max_y \langle y, d_\pi\rangle - f^*(y)$, where $\mathcal{K}$ is the occupancy polytope and f^* is the Fenchel conjugate of f. Crucially, for any fixed dual variable y, the policy subproblem $\min_\pi \langle y, d_\pi\rangle$ becomes a standard RL problem with a shaped reward $r_y(s,a) = -y(s,a)$. (We use y for the dual, as in [**?**]; Zahavy et al. [**?**] use λ, with $r_\lambda = -\lambda$.) This suggests an alternating approach: (i) a dual ascent step on y, and (ii) a policy-improvement step on the shaped-reward RL instance.

Policy Learner Choice. For the policy step, we use *Softmax Policy Mirror Ascent (SPMA)*, which updates $\pi_{t+1}(a|s) = \pi_t(a|s)\big(1 + \eta\, A^{\pi_t}_{r_y}(s,a)\big)$, where $A^{\pi_t}_{r_y} = Q^{\pi_t}_{r_y} - V^{\pi_t}_{r_y}$ is the advantage function for the shaped reward r_y. For a small constant step size η (e.g., $\eta \leq 1-\gamma$), this yields fast tabular convergence and has a clean function-approximation (FA) variant.

Related Approaches and Baseline. Other methods tackle the *CMDP* setting directly in the policy–multiplier space, e.g., *Natural Policy Gradient Primal–Dual (NPG-PD)*. This algorithm performs natural-gradient ascent on π with projected subgradient updates on the Lagrange multiplier. We favor the more general Fenchel-dual route, as it reduces the *cMDP* problem to repeated standard RL; we will compare our approach against NPG-PD as a principled baseline for the constrained case.

Roadmap. We adopt the Fenchel–dual reduction of *cMDPs* [**?**], use SPMA as the inner-loop policy learner [**?**], and compare against NPG-PD [**?**] as a baseline. In the following sections, we summarize each of these foundational papers.

2 Literature Review: Paper Summaries

2.1 Paper 1: *Reward is Enough for Convex MDPs*

Standard RL maximizes linear reward functions over occupancy measures. This paper [**?**] extends the framework to *cMDPs*, where the objective is a convex function $f(d_\pi)$ of the discounted occupancy measure d_π (as defined in our Introduction). This generalization handles apprenticeship learning, *CMDPs*, and pure exploration within a unified framework.

A key theoretical contribution of this paper is the use of *Fenchel duality* to reformulate the cMDP objective as a saddle-point problem. The original problem, $\min_{d\in\mathcal{K}} f(d)$, is equivalently expressed, via the Fenchel–Moreau identity, as $f_{\text{OPT}} = \min_{d\in\mathcal{K}} \max_{\lambda\in\Lambda} \langle \lambda, d\rangle - f^*(\lambda)$, where $d(s,a)$ denotes the occupancy measure induced by a policy π, $\lambda(s,a)$ is a dual variable, Λ is the dual domain, and f^* is the Fenchel conjugate of f.

This formulation defines a two-player game: the policy player selects an occupancy measure d, while the cost player chooses a dual variable λ, and is regularized by the term $f^*(\lambda)$. For any fixed λ, the inner minimization with respect to d reduces to a standard reinforcement learning problem.

The authors propose a *meta-algorithm* that implements this game through alternating updates. At each iteration, the cost player (using an Online Convex Optimization method) selects a dual variable λ_k, and the policy player responds by running RL with shaped reward $r_k = -\lambda_k$ to obtain occupancy d_π^k. After K rounds, the algorithm returns the averaged occupancy $\bar{d}_\pi$. The paper proves $O(1/\sqrt{K})$ optimization error when both players use low-regret algorithms (via OCO regret bounds).

Another key contribution of the paper is the *unification* of several reinforcement-learning paradigms within the *cMDP* framework. By selecting different convex functions $f(d_\pi)$, the same Fenchel-dual formulation recovers well-known objectives: imitation learning when f is a divergence from an expert distribution, *CMDPs* when f includes penalty terms, and pure exploration when f is the entropy of the occupancy measure. This shows that diverse RL settings, supervised, constrained, and unsupervised, can all be derived from a single convex-duality perspective.

2.2 Paper 2: *Fast Convergence of Softmax Policy Mirror Ascent*

This paper [**?**] bridges the gap between theoretically grounded policy gradient (PG) algorithms, such as Natural Policy Gradient (NPG) [**?**], which enjoy strong convergence guarantees only in tabular settings, and practical deep RL algorithms like PPO [**?**], TRPO [**?**], and MDPO [**?**], which perform well empirically but lack rigorous analysis. The authors propose *Softmax Policy Mirror Ascent (SPMA)*, a normalization-free mirror ascent algorithm in the **dual (logit) space** that achieves linear convergence guarantees in tabular settings and extends naturally to function approximation.

SPMA performs mirror ascent in logit space using the log-sum-exp mirror map $\Phi(z) = \sum_s \log \sum_a \exp(z(s,a))$, which induces the softmax link $\pi(a|s) = \exp(z(s,a))/\sum_{a'} \exp(z(s,a'))$. At iteration t, the update is

$$z_{t+1} = \arg\max_z \left[\langle z - z_t, \nabla_z J(z_t) \rangle - \frac{1}{\eta} D_\Phi(z, z_t) \right],$$

where z_t is the current logit vector, $J(z)$ is the expected return as a function of the policy induced by z, $\nabla_z J(z_t)$ is the gradient of the objective with respect to logits, $\eta > 0$ is the step-size, and $D_\Phi(z, z_t) = \Phi(z) - \Phi(z_t) - \langle \nabla\Phi(z_t), z - z_t \rangle$ is the Bregman divergence induced by Φ. This update corresponds to gradient ascent in the natural geometry of the policy space and does not require explicit normalization across actions, as the softmax parameterization inherently ensures that the resulting policies remain valid probability distributions.

Weighted by the discounted state distribution $d_{\pi_t}(s) = (1-\gamma)\sum_{\tau=0}^{\infty} \gamma^\tau \Pr_{\pi_t}(s_\tau = s)$, where $\gamma \in [0, 1)$ is the discount factor, this yields a per-state policy update that is linear in both step size and advantage:

$$\pi_{t+1}(a|s) = \pi_t(a|s)\big(1 + \eta A_{\pi_t}(s,a)\big),$$

where $\pi_t(a|s)$ is the probability of taking action a in state s under the current policy at iteration t, $\eta > 0$ is the step-size, and $A_{\pi_t}(s,a)$ is the advantage function at state s and action a.

This update produces a valid probability distribution for 2 reasons. First, normalization is preserved because $\sum_a \pi_t(a|s) A^{\pi_t}(s,a) = 0$. Second, positivity requires a sufficiently small constant step size (e.g., $\eta \leq 1 - \gamma$) and a per-state margin condition, so unlike NPG, no explicit renormalization is needed after the update.

The authors prove **global linear convergence** of SPMA in both **bandit** and **tabular MDP** settings. For bandits, SPMA achieves exponential (linear) convergence to the optimal arm, and a gap-dependent variant achieves **super-linear** convergence—the first such result for policy gradient methods. For tabular MDPs, they establish:

$$\|V^{\pi^*} - V^{\pi_T}\|_\infty \leq \prod_{t=0}^{T-1} (1 - \eta C_t(1-\gamma)) \, \|V^{\pi^*} - V^{\pi_0}\|_\infty,$$

where C_t reflects the per-state advantage gap. Thus, with a constant step size $\eta < \min(1-\gamma, [C_t(1-\gamma)]^{-1})$, SPMA converges linearly to the optimal value function—matching NPG's rate but without dependence on the potentially large distribution-mismatch term.

To handle large or continuous state–action spaces, SPMA is extended using **projected mirror ascent** under a log-linear or neural policy parameterization. The logits $z_\theta(s,a)$ are constrained to lie in a

realizable function class $Z = \{f_\theta(s, a)\}$, and each update projects the ideal (tabular) step back into this feasible set via convex softmax classification:

$$\theta_{t+1} = \arg\min_\theta \sum_s d_{\pi_t}(s)\, \mathrm{KL}(\pi_{t+1/2}(\cdot|s)\,||\,\pi_\theta(\cdot|s)).$$

Unlike NPG, this does not require compatible function approximation; unlike MDPO, the per-iteration subproblem remains convex for linear FA. Under bounded rewards, realizability, and on-policy sampling, SPMA converges **linearly to a neighborhood** of the optimal value function, where the neighborhood size depends on FA bias ($\varepsilon_{\text{bias}}$) and statistical error ($\varepsilon_{\text{stat}}$).

In the Fenchel-dual cMDP framework, we use SPMA as the **policy player** to solve the shaped-reward inner RL step, leveraging its linear convergence and convex surrogate for efficient inner-loop optimization.

2.3 Paper 3: *Natural Policy Gradient Primal–Dual for CMDPs*

This paper [**?**] addresses the *CMDP* setting, which our Introduction identified as a key instance of the broader *cMDP* framework. The goal in a *CMDP* is to maximize an expected discounted reward subject to constraints on expected discounted costs. While policy gradient methods have strong theory for unconstrained problems, their guarantees for *CMDPs* have been largely asymptotic or local. This paper proposes the **Natural Policy Gradient Primal–Dual (NPG-PD)** algorithm to establish non-asymptotic, global guarantees.

The problem is formulated via the Lagrangian function: $L(\pi, \lambda) = V_r^\pi(\rho) + \lambda\big(V_g^\pi(\rho) - b\big)$, where ρ is the initial state distribution, $V_r^\pi(\rho)$ is the expected discounted reward, $V_g^\pi(\rho)$ represents the expected discounted utility (or cost), and b is the constraint threshold. The analysis assumes the *Slater condition* (strict feasibility), which ensures strong duality and the boundedness of the dual variable λ. Despite the *nonconcave* objective and *nonconvex* feasible set under policy parameterization, the paper proves convergence guarantees for **time-averaged** primal–dual iterates solving $\max_\pi \min_{\lambda \geq 0} L(\pi, \lambda)$.

Algorithm. *NPG-PD* alternates (i) a **primal** update using the natural policy gradient (Fisher-preconditioned) to adjust π, and (ii) a **dual** update using projected subgradient descent on λ, which increases when the constraint is violated and decreases when satisfied. The geometry-aware primal step stabilizes updates in policy space.

Under a *softmax policy parameterization*, the paper establishes global, dimension-free convergence for the **time-averaged** iterates. With appropriately chosen step sizes (e.g., a primal stepsize scaling with $\log|\mathcal{A}|$ and a dual stepsize on the order of $(1-\gamma)/\sqrt{T}$), both the *optimality gap* and the *constraint violation* of the averaged policy decrease as $\mathcal{O}(T^{-1/2})$. Hence, an ε-optimal and ε-feasible policy is reached in $\mathcal{O}(\varepsilon^{-2})$ iterations, independent of $|\mathcal{S}|$ and $|\mathcal{A}|$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces.

For *general policy classes* (e.g., neural network policies) where strong duality may fail, the analysis uses an *approximate* natural policy gradient. The bounds hold **up to** (i) a *compatible function approximation error* $\varepsilon_{\text{approx}}$, quantifying how well the parameterized critic matches the Lagrangian advantage, and (ii) a *distribution-mismatch factor* $\left\|\nu^*/\nu_0\right\|_\infty$, reflecting exploration from the initial state–action distribution.

Assuming policy smoothness and an *exploratory initial distribution* to mitigate state–action mismatch, the authors proves *sublinear convergence* of the time-averaged iterates: the optimality gap is $\mathcal{O}(T^{-1/2})$ and the constraint violation is $\mathcal{O}(T^{-1/4})$, up to $\varepsilon_{\text{approx}}$ and $\left\|\nu^*/\nu_0\right\|_\infty$.

Finally, the paper presents *model-free (sample-based)* variants that estimate gradients and values from trajectories. With K rollouts per iteration and standard smoothness/bounded-error assumptions, the time-averaged guarantees remain sublinear: the optimality gap scales as $\mathcal{O}(T^{-1/2})$ and the constraint violation as $\mathcal{O}(T^{-1/4})$, with constants depending on K and $(1-\gamma)^{-1}$. Under the softmax parameterization, the resulting sample complexity is tighter than in the general case.