Project Report — A Dual–SPMA Framework for Convex MDPs

Shervin Khamooshian **Ahmed Magd** **Pegah Aryadoost**

Danielle Nguyen
School of Computing Science, Simon Fraser University
{ska309,ams80,paa40,tdn8}@sfu.ca

Abstract

We study a practical solver for Convex MDPs that combines a Fenchel–dual reformulation with a fast policy optimizer, Softmax Policy Mirror Ascent (SPMA). Fixing the dual variable yields a shaped-reward RL subproblem; SPMA serves as the policy player. We compare against a principled CMDP baseline, Natural Policy Gradient Primal–Dual (NPG–PD). We outline our method, experimental plan, and ablations, and provide early results placeholders.

1 Introduction

Convex MDPs (cMDPs) extend standard RL to objectives that are convex in the discounted occupancy measure d^{π}. Many problems (constrained RL, imitation, exploration) fit this form but are hard to optimize directly over d^{π}. Following Zahavy et al. [2021], we use Fenchel duality to obtain a *policy–cost* saddle formulation and then solve the policy step with SPMA [Asad et al., 2024]. We compare to NPG–PD, a policy-based CMDP method with non-asymptotic guarantees [Ding et al., 2020].

Contributions (draft). (i) Implement an outer–inner Dual–SPMA loop that reduces cMDPs to shaped-reward RL; (ii) instantiate SPMA as the policy player (tabular and linear FA) and dual updates via OMD/FTL; (iii) evaluate against NPG–PD on constrained tasks with convergence, violation, and efficiency metrics.

2 Related Work

Convex MDPs via Fenchel duality. Zahavy et al. [2021] reformulate $\min_{d \in K} f(d)$ as a saddle $\min_{\pi} \max_{y} \{\langle y, d^{\pi} \rangle - f^{*}(y)\}$ and propose a meta-algorithm alternating a cost player (FTL/OMD) with a policy player (standard RL under $r_y = -y$). **Policy optimization geometry.** SPMA performs mirror ascent in logit space and achieves linear convergence in tabular MDPs; with FA it uses a convex softmax classification projection [Asad et al., 2024]. **CMDP primal–dual.** NPG–PD updates policy by natural PG and multipliers by projected subgradient, with $O(1/\sqrt{T})$ averaged gap/violation guarantees [Ding et al., 2020].

3 Preliminaries

MDP and occupancies. For discounted MDP $(\mathcal{S}, \mathcal{A}, P, \rho, \gamma)$,

$$d^{\pi}(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^{t} \Pr_{\pi}(s_t = s, a_t = a), \quad K = \{d \geq 0 : \text{flow constraints hold}\}. \tag{1}$$

Convex MDP. We seek $\min_{d\in K} f(d)$ where $f : K \to \mathbb{R}$ is convex. Using the Fenchel conjugate f^*, the problem equals

$$\min_{\pi} \max_{y} \; L(\pi, y) := \langle y, d^\pi \rangle - f^*(y), \tag{2}$$

and for fixed y the policy subproblem is standard RL with shaped reward $r_y = -y$ (or $r_y(s,a) = -\phi(s,a)^\top y$ under features) [Zahavy et al., 2021].

4 Method: Dual–SPMA

Outer (dual) update. Given estimate $\hat{d}^{\pi_t}$ or $\widehat{\mathbb{E}}[\phi]$, update y via OMD or FTL:

$$\text{OMD: } y_{t+1} = \arg\max_{y} \; \big\langle y - y_t, \; \hat{d}^{\pi_t} - \nabla f^*(y_t) \big\rangle - \tfrac{1}{\eta_2} B_r(y, y_t), \tag{3}$$

$$\text{FTL: } y_{t+1} = \nabla f \left(\tfrac{1}{t} \sum_{k=1}^{t} \hat{d}^{\pi_k} \right). \tag{4}$$

Inner (policy) update via SPMA. For tabular policies and step-size $\eta \leq 1 - \gamma$,

$$\pi_{t+1}(a|s) = \pi_t(a|s) \big(1 + \eta \, A_{\pi_t}^{(r_y)}(s,a) \big), \tag{5}$$

and under log-linear FA we project the ideal tabular step back into the class via convex softmax classification [Asad et al., 2024].

Estimators. In tabular settings we estimate d^π from finite-horizon rollouts; under FA we estimate feature expectations $\mathbb{E}_{d^\pi}[\phi]$ (Appendix **??**).

Algorithm (sketch). Alternate K outer dual steps with m inner SPMA steps per outer iteration; return averaged iterates $(\bar{\pi}, \bar{y})$.

5 Baseline: NPG–PD for CMDPs

We implement NPG–PD with softmax policies: natural PG ascent on π, projected subgradient on λ in the Lagrangian $V_r^\pi(\rho) + \lambda(V_g^\pi(\rho) - b)$; report averaged optimality gap and constraint violation as in Ding et al. [2020].

6 Experimental Setup

Environments. Small tabular MDPs (grid/chain) with constraints or imitation; linear-feature variants for FA.

Metrics. (i) Saddle $L(\pi, y)$ (when f^* known); (ii) constraint value/violation; (iii) return under r_y; (iv) convergence of $\|d^\pi\|_1$ (tabular) or $\|\mathbb{E}[\phi]\|$ (FA); (v) wall-clock/sample efficiency.

Baselines. NPG–PD; (optional) PPO/TRPO/MDPO for context.

Hyperparameters. Dual step η_2, SPMA step η, inner steps m, rollout lengths, seeds (Appendix **??**).

7 Results (placeholders)

Main comparisons. Dual–SPMA vs NPG–PD: convergence of $L(\pi, y)$, constraint violation, sample efficiency.

Ablations. OMD vs FTL; impact of m and η; tabular vs FA.

Analysis. Stability, estimator variance, effect of dual regularization.

8 Discussion

When does Dual–SPMA help (simple inner loops, convex FA projection) and when does NPG–PD remain preferable (tight constraint control, established guarantees)?

9 Limitations and Future Work

Occupancy/feature estimation variance; coupling of inner/outer steps; FA bias. Next steps: variance reduction, neural FA, robust costs.

10 Conclusion

Dual–SPMA is a simple outer–inner approach to cMDPs that leverages shaped-reward RL and a fast policy optimizer. Early experiments (to be inserted) indicate competitive performance with NPG–PD on constrained tasks.

Reproducibility. Code, configs, and scripts to regenerate figures will be released with the final report.

References

Reza Asad, Reza Babanezhad Harikandeh, Issam H. Laradji, Nicolas Le Roux, and Sharan Vaswani. Fast convergence of softmax policy mirror ascent for bandits & tabular MDPs. 2024. URL `https://openreview.net/forum?id=f5OjNMXIik`.

Dongsheng Ding, Kaiqing Zhang, Tamer Başar, and Mihailo R. Jovanović. Natural policy gradient primal-dual method for constrained markov decision processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tom Zahavy, Brendan O'Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. volume 34, pages 25746–25759, 2021.