Project Milestone — Literature Review: A Dual–SPMA Framework for Convex MDPs

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

Project topic (what we are building)

Goal. We study a unified way to solve *Convex Markov Decision Processes (CMDPs)* by combining a Fenchel-dual reformulation of the objective with a geometry-aware policy optimizer, *Softmax Policy Mirror Ascent (SPMA)*. The CMDP is written as minimizing a convex function of discounted occupancies, then equivalently as a saddle problem $\min_\pi \max_y \langle y, d_\pi \rangle - f^*(y)$. Fixing y turns the policy step into standard RL with a shaped reward $r_y(s,a) = -y(s,a)$ (or $-\phi(s,a)^\top y$ under features). We alternate a dual ascent step on y with an SPMA policy step, and return discounted occupancy (or feature-expectation) estimates for the next dual update.

Why now. CMDPs unify safety constraints, imitation/occupancy matching, and exploration under one convex-in-occupancy umbrella; SPMA offers a fast, mirror-descent policy learner that serves as a strong "best response" inside the saddle formulation. Our implementation plan and ablations follow directly from this literature.

Paper 1: *Reward is Enough for Convex MDPs* (NeurIPS 2021)

Core idea. Many RL goals can be posed as $\min_{d \in \mathcal{K}} f(d)$ for a convex f over the occupancy polytope $\mathcal{K}$, generalizing the linear-reward case. Using Fenchel conjugacy, this becomes the convex–concave saddle $\min_{d \in \mathcal{K}} \max_{\lambda \in \Lambda} \ \lambda \cdot d - f^*(\lambda)$, so for any fixed λ the policy subproblem reduces to vanilla RL with shaped reward $r_\lambda = -\lambda$. The paper provides a meta-algorithm that alternates a *cost player* (FTL/OMD over λ) with a *policy player* (best response or low-regret RL), yielding $O(1/\sqrt{K})$ optimization error for averaged iterates under standard OCO assumptions. It also shows how apprenticeship learning, constrained MDPs, and pure exploration emerge as concrete choices of f and of the two players (Table 1; Fig. 1). [Zahavy et al., 2021] **Why it matters for us.** This paper (i) justifies our saddle formulation; (ii) explains shaped-reward policy updates; (iii) guides our outer-loop design (dual mirror ascent + policy best response). Our implementation mirrors their Alg. 1 but swaps in a specific policy learner (SPMA).

Paper 2: *Fast Convergence of Softmax Policy Mirror Ascent* (arXiv 2025 / OPT 2024)

Core idea. SPMA performs mirror ascent *in logit space* using the log-sum-exp mirror map. In tabular MDPs, the per-state update simplifies to $\pi_{t+1}(a|s) = \pi_t(a|s)\big(1 + \eta\, A^{\pi_t}(s,a)\big)$, which avoids explicit normalization and yields *linear convergence* to the optimal value for sufficiently small constant step-size (matching NPG rates and improving over softmax PG with constant step-size). To scale, the paper projects onto function classes (log-linear or energy-based) by solving convex softmax-classification subproblems each iteration, and shows linear convergence to a neighbourhood under FA plus strong empirical results competitive with PPO/TRPO/MDPO. [Asad et al., 2024]

Why it matters for us. We need a strong policy "best response" inside the CMDP saddle; SPMA supplies both the geometry and the rates, and its FA projection matches our shaped-reward reduction for fixed y. This is the policy player in our Dual–SPMA solver.

Paper 3: *Natural Policy Gradient Primal–Dual for CMDPs* (NeurIPS 2020)

Core idea. This paper studies a policy-based primal–dual algorithm for discounted CMDPs: *natural policy gradient* (NPG) ascent for the policy and projected subgradient updates for the dual variable. Despite nonconcavity/nonconvexity under softmax parameterization, they prove *dimension-free* $O(1/\sqrt{T})$ bounds on the averaged optimality gap and constraint violation; for general FA they obtain rates up to an approximation neighbourhood, and provide sample-based variants with finite-sample guarantees. [Ding et al., 2020]

Why it matters for us. NPG–PD is a principled baseline for CMDPs with theory in both tabular and FA settings; we use it as a comparator and as a reference point for convergence/violation metrics and experimental design.

How the three fit together (and into our project)

Synthesis. Zahavy et al. provide the *formulation and outer-loop template* (Fenchel saddle; shaped-reward policy step). SPMA provides a *fast policy player* for that step (mirror ascent in logits; linear rates; FA via convex classification). NPG–PD offers a *policy-based CMDP baseline* with sublinear but dimension-free guarantees. Our project implements the Dual–SPMA solver by alternating a dual mirror-ascent step on y with an SPMA step on the r_y-shaped RL task, and evaluates against NPG–PD.

Work	Objective / saddle	Policy player	Guarantees / notes
Zahavy et al. (2021)	$\min_{d\in\mathcal{K}} f(d)$; Fenchel dual $\min_d \max_\lambda \lambda\cdot d - f^*(\lambda)$	Best response / low-regret RL under shaped reward $r_\lambda = -\lambda$	$O(1/\sqrt{K})$ via OCO; unifies AL, CMDPs, pure exploration (Fig. 1, Table 1). [Zahavy et al., 2021]
Asad et al. (2025)	Standard RL inner step (fixed y)	SPMA: mirror ascent in logits, $\pi_{t+1} = \pi_t(1+\eta A)$; FA via convex projection	Linear conv. (tabular); linear-to-neighbourhood (FA); strong empirical results. [Asad et al., 2024]
Ding et al. (2020)	Lagrangian CMDP $\max_\pi \min_{\lambda\geq 0} V_r^\pi + \lambda(V_g^\pi - b)$	NPG for π, projected subgradient for λ	Dimension-free $O(1/\sqrt{T})$ gap & violation (avg.); sample-based variants. [Ding et al., 2020]

What we will implement and measure (brief)

Method. Dual–SPMA: $y_{k+1} \leftarrow \mathrm{MA}(y_k,\, \hat{d}_{\pi_k} - \nabla f^*(y_k))$; policy step: run SPMA for K_{in} epochs on r_{y_k}; return $\hat{d}_{\pi_k}$ (or $\widehat{\mathbb{E}}[\phi]$).

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

Acknowledgements

We thank the authors of the above works and use their formulations strictly for academic purposes.

References

R. Asad, R. B. Harikandeh, I. H. Laradji, N. L. Roux, and S. Vaswani. Fast convergence of softmax policy mirror ascent for bandits & tabular MDPs. 2024. URL `https://openreview.net/forum?id=f5OjNMXIik`.

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

T. Zahavy, B. O'Donoghue, G. Desjardins, and S. Singh. Reward is enough for convex mdps. volume 34, pages 25746–25759, 2021.