

Figure 1: **Dual-SPMA algorithm loop.** Dual ascent chooses y_k , which induces a shaped reward for the SPMA policy step; discounted occupancies feed the next dual update.

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

1 Project Topic

Goal. We study a unified way to solve *Constrained Markov Decision Processes (CMDPs)* by combining a Fenchel-dual saddle formulation with a geometry-aware policy optimizer, *Softmax Policy Mirror Ascent (SPMA)*. CMDPs minimize a convex function of discounted occupancies and are equivalent to the saddle problem:

$$\min_{\pi} \max_y \langle y, d_{\pi} \rangle - f^*(y).$$

Fixing y turns the policy step into standard RL with shaped reward $r_y(s, a) = -y(s, a)$ (or $-\phi(s, a)^T y$ under features). We alternate a mirror-ascent step on y with an SPMA policy step and return discounted occupancy (or feature-expectation) estimates for the next dual update.

Implementation. Building on Zahavy’s formulation, we alternate a dual step

$$y_{k+1} \leftarrow \text{MA}\left(y_k, \hat{d}_{\pi_k} - \nabla f^*(y_k)\right)$$

with a policy step that runs SPMA on the shaped reward $r_{y_k} = -y_k$, returning discounted occupancies \hat{d}_{π_k} (or feature expectations) for the next dual update (Fig. 1). We will benchmark this Dual-SPMA against NPG-PD, focusing on convergence speed, constraint satisfaction, and sample efficiency.

2 Literature Review: Paper Summaries

We review three foundational papers that inform our Dual-SPMA framework: (1) the Fenchel duality approach for convex MDPs, (2) the SPMA policy optimizer with fast convergence, and (3) primal-dual methods for constrained MDPs.

2.1 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 convex f over the occupancy polytope \mathcal{K} . Using Fenchel conjugacy,

$$\min_{d \in \mathcal{K}} f(d) = \min_{d \in \mathcal{K}} \max_{\lambda \in \Lambda} \lambda \cdot d - f^*(\lambda),$$

so for fixed λ the policy subproblem is vanilla RL with shaped reward $r_{\lambda} = -\lambda$.

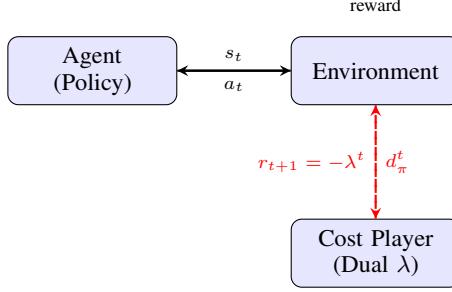


Figure 2: Convex MDP as a two-player game (adapted from Zahavy et al. [2021], Fig. 1). The cost player provides shaped rewards $r_t = -\lambda^t$ to the agent, observing occupancy measures d_π^t .

Algorithm 1: Meta-algorithm for Convex MDPs [Zahavy et al., 2021]

Input: Convex-concave payoff $\mathcal{L} : \mathcal{K} \times \Lambda \rightarrow \mathbb{R}$, algorithms $\text{Alg}_\lambda, \text{Alg}_\pi, K \in \mathbb{N}$

- 1 **for** $k = 1, \dots, K$ **do**
- 2 $\lambda^k \leftarrow \text{Alg}_\lambda(d_\pi^1, \dots, d_\pi^{k-1}; \mathcal{L})$; *// Cost player*
- 3 $a_\pi^k \leftarrow \text{Alg}_\pi(-\lambda^k)$; *// Policy: RL with $r = -\lambda^k$*

Output: $\bar{d}_\pi^K = \frac{1}{K} \sum_{k=1}^K d_\pi^k, \bar{\lambda}^K = \frac{1}{K} \sum_{k=1}^K \lambda^k$

Figure 2 illustrates this as a two-player game where the agent sees non-stationary rewards from the cost player. A meta-algorithm (Algorithm 1) alternates a *cost player* (FTL/OMD in λ , a convex ascent step) with a *policy player* (best response or low-regret RL), yielding $O(1/\sqrt{K})$ optimization error for averaged iterates under standard OCO assumptions. The paper shows best-response is ideal but often intractable in deep RL, so low-regret learners (e.g., UCRL2, MDPO) suffice; the guarantees hold for averaged occupancies \bar{d}_π^K rather than single iterates. It unifies apprenticeship learning, CMDPs and pure exploration (Table 1). [Zahavy et al., 2021]

Relevance to our project. This work justifies the saddle formulation and the shaped-reward reduction we implement. It informs our outer-loop design (dual mirror ascent + policy best response).

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

Core idea. SPMA performs mirror ascent in *logit* space using the log-sum-exp mirror map, in contrast to NPG which uses exponential reweighting in probability space (Sec. 3.2). In tabular MDPs the per-state update (Eq. 3)

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

avoids explicit normalization and achieves *linear convergence* for sufficiently small constant step-size: bandit linear convergence (Thm. 1) and tabular MDP linear convergence to the optimal value (Thm. 2), improving over softmax PG (Sec. 3.3).

For large problems, SPMA projects onto function classes via convex *softmax classification* subproblems (Eq. 4–5, Algorithm 1):

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

It proves linear convergence to a neighbourhood under FA (Sec. 4); the theory assumes exact or low-noise gradients, and performance depends on advantage-estimation quality (Sec. 4.2). Empirically it competes with PPO/TRPO/MDPO (Sec. 4.1). [Asad et al., 2024]

Relevance to our project. We need a strong policy “best response” in Zahavy’s saddle; SPMA provides the geometry and fast rates in tabular settings, plus a practical FA implementation (convex surrogates that fit deep RL) matching our shaped-reward reduction.

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

Core idea. A policy-based primal–dual method for the Lagrangian CMDP formulation $V_r^\pi(\rho) + \lambda(V_g^\pi(\rho) - b)$: *natural policy gradient* (NPG) ascent for the policy and projected subgradient updates

Table 1: Instantiations of the convex MDP framework (adapted from Zahavy et al. [2021], Table 1).

Application	Objective $f(d_\pi)$	Cost Player	Policy Player
Standard RL	$-\lambda \cdot d_\pi$	FTL	RL
Apprenticeship Learning	$\ d_\pi - d_E\ _2^2$	FTL	Best Response
Pure Exploration	$d_\pi \cdot \log(d_\pi)$	FTL	Best Response
Constrained MDPs	$\lambda_1 \cdot d_\pi$ s.t. $\lambda_2 \cdot d_\pi \leq c$	OMD	RL
GAIL / State Matching	$\text{KL}(d_\pi \ d_E)$	FTL	RL

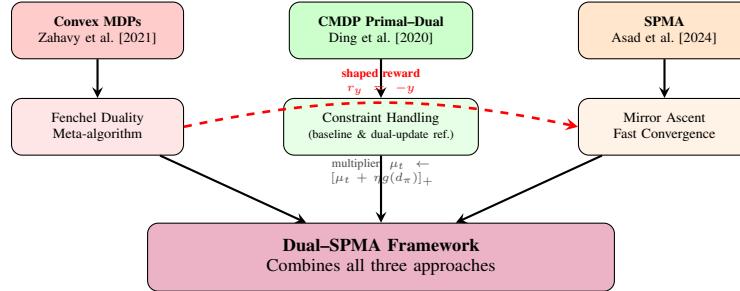


Figure 3: Overview of how the three papers inform our Dual-SPMA framework. The key insight is that Zahavy’s shaped reward reduction ($r_y = -y$) enables us to use SPMA as the policy player, while Ding et al. inform our constraint handling and serve as a baseline.

for the multiplier (Eq. 7–8, Sec. 3–4), showing a multiplicative update for softmax policies and projection for λ .

Despite nonconcavity/nonconvexity under softmax parameterization, it proves *dimension-free* $O(1/\sqrt{T})$ bounds on averaged optimality gap and constraint violation (Thm. 1, Eq. 9a–9b) under Slater’s condition; with FA, rates hold up to an approximation neighbourhood (Sec. 5, Thm. 3); sample-based variants have finite-sample guarantees (Thm. 4). [Ding et al., 2020]

Relevance to our project. NPG–PD is our principled CMDP baseline for both guarantees and practice: we compare Dual–SPMA (Fenchel saddle + SPMA policy player) against NPG–PD (Lagrangian saddle + NPG) in terms of geometry (logit-space vs. probability-space), convergence rates, constraint violation, and sample-efficiency.

3 How the Papers Relate to Each Other

The three papers provide complementary perspectives on our Dual–SPMA framework (Fig. 3):

Formulation. Zahavy et al. provide the *outer-loop template*: the Fenchel-dual saddle formulation that reduces constrained/structured RL to alternating between a cost player (dual ascent) and a policy player (shaped-reward RL). This formulation is key because fixing the dual variable y yields standard RL with shaped reward $r_y = -y$.

Policy player. SPMA (Asad et al.) supplies the *fast policy player* for the RL step in Zahavy’s framework. It performs mirror ascent in logit space, achieving linear convergence rates in tabular settings and practical function approximation via convex softmax classification. This makes SPMA an ideal candidate for the policy subproblem.

Baseline and dual updates. NPG–PD (Ding et al.) offers a *policy-based CMDP baseline* with dimension-free sublinear guarantees. While we use SPMA (not NPG) as our policy player, Ding et al. inform how we handle constraint multipliers via projected subgradient updates and serve as our principled comparison baseline.

Table 2 summarizes how these three approaches differ in their formulation, policy optimization method, and convergence guarantees.

4 Project Implementation and Evaluation

Method. Our Dual–SPMA algorithm alternates:

Table 2: Comparison of the three approaches that inform our project.

Work	Formulation	Policy Method	Guarantees
Zahavy et al. (2021) [Zahavy et al., 2021]	Fenchel dual: $\min_d \max_\lambda \lambda \cdot d - f^*(\lambda)$	Best response / low-regret RL under $r = -\lambda$	$O(1/\sqrt{K})$ via OCO
Asad et al. (2024) [Asad et al., 2024]	RL inner step (fixed y)	SPMA: $\pi_{t+1} = \pi_t(1 + \eta A)$	Linear (tabular); neighbourhood (FA)
Ding et al. (2020) [Ding et al., 2020]	Lagrangian: $\max_\pi \min_{\lambda \geq 0} V_r^\pi - \lambda(V_g^\pi - b)$	NPG + projected subgradient	Dimension-free $O(1/\sqrt{T})$

1. **Dual update:** $y_{k+1} \leftarrow \text{MA}\left(y_k, \hat{d}_{\pi_k} - \nabla f^*(y_k)\right)$
2. **Policy step:** Run SPMA for K_{in} epochs on shaped reward $r_{y_k} = -y_k$
3. **Return:** Discounted occupancies \hat{d}_{π_k} (or $\hat{\mathbb{E}}[\phi]$ under FA)

Evaluation metrics. We will measure:

- (i) Saddle value $L(\pi, y)$ (when f^* is known)
- (ii) Constraint value and constraint violation
- (iii) Policy return under shaped reward r_y
- (iv) Convergence of $\|\hat{d}_\pi\|_1$ (tabular) or $\|\hat{\mathbb{E}}[\phi]\|$ (FA)
- (v) Wall-clock time and sample efficiency

Baselines. We will compare against NPG–PD (Ding et al., 2020) as our principled CMDP baseline, evaluating differences in convergence speed, constraint satisfaction, and sample efficiency.

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=f50jNMXIik>.
- D. Ding, K. Zhang, T. Bařáš, 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.