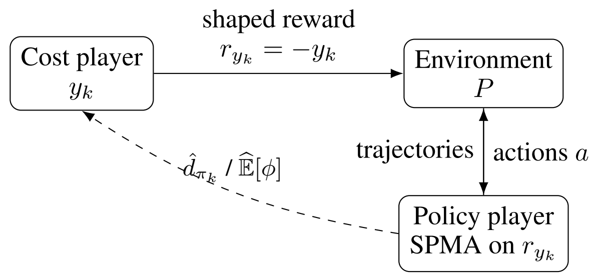

Figure 1: **Dual–SPMA loop.** Dual ascent chooses y_k, which induces a shaped reward for the SPMA policy step; discounted occupancies feed the next dual update.

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, tdSn8}@sfu.ca`

Project topic (what we are building)

Goal. We study a unified way to solve *Convex MDPs (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 $\min_\pi \max_y \langle y, d_\pi \rangle - f^*(y)$. Fixing y turns the policy step into standard RL with shaped reward $r_y(s,a) = -y(s,a)$ (or $-\phi(s,a)^\top y$ 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 (Fig. 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 $r_\lambda = -\lambda$.

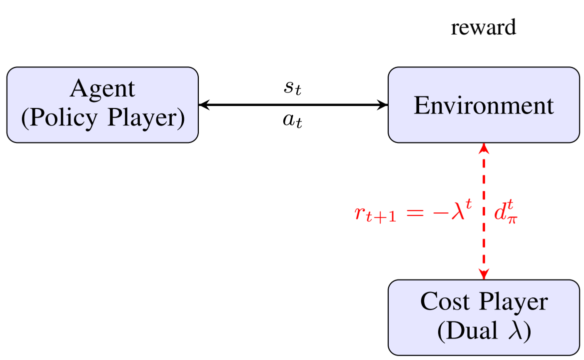

Figure 2: Convex MDP as a two-player game (adapted from Zahavy et al. [2021], Fig. 1). The cost player provides non-stationary shaped rewards $r_t = -\lambda^t$ to the agent, observing the resulting occupancy measures d_π^t. From the agent's perspective, this reduces to standard RL with time-varying rewards.

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, which reduces to "just RL" under the shaped reward), 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 $\bar{d}_\pi^K$ rather than single iterates. It unifies apprenticeship learning, CMDPs and pure exploration (Table 2). [Zahavy et al., 2021]

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

Input: Convex-concave payoff $\mathcal{L} : \mathcal{K} \times \Lambda \to \mathbb{R}$, algorithms Alg_λ, Alg_π, $K \in \mathbb{N}$

for $k = 1, \ldots, K$ **do**

  $\lambda^k \leftarrow \mathrm{Alg}_\lambda(d_\pi^1, \ldots, d_\pi^{k-1}; \mathcal{L})$; `// Cost player update`

  $d_\pi^k \leftarrow \mathrm{Alg}_\pi(-\lambda^k)$; `// Policy: solve RL with` $r = -\lambda^k$

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

Relevance. This work justifies the saddle and the shaped-reward reduction we implement and informs our outer-loop design (dual MA + policy best response).

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 tabular MDPs the per-state update $\pi_{t+1}(a|s) = \pi_t(a|s)\big(1 + \eta\, A^{\pi_t}(s,a)\big)$ avoids explicit normalization and achieves *linear convergence* for sufficiently small constant step-size, improving over softmax PG. For large problems, SPMA projects onto function classes via convex *softmax classification* subproblems and proves linear convergence to a neighbourhood under FA; empirically it competes with PPO/TRPO/MDPO. [Asad et al., 2024]

Relevance. We need a strong policy "best response" in Zahavy's saddle; SPMA provides the geometry and rates, and its FA projection matches our shaped-reward reduction.

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

Core idea. A policy-based primal–dual method: *natural policy gradient* (NPG) ascent for the policy and projected subgradient updates for the multiplier. Despite nonconcavity/nonconvexity under softmax parameterization, it proves *dimension-free* $O(1/\sqrt{T})$ bounds on averaged optimality gap and constraint violation; with FA, rates hold up to an approximation neighbourhood; sample-based variants have finite-sample guarantees. [Ding et al., 2020]

Relevance. NPG–PD is our principled CMDP baseline for both guarantees and practice; we compare Dual–SPMA against it in convergence/violation/sample-efficiency.

How the three fit together (and into our project)

Zahavy et al. provide the *formulation and outer-loop template* (Fenchel saddle; shaped-reward RL). SPMA supplies a *fast policy player* for that RL 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 dual mirror-ascent on y with an SPMA step on the r_y-shaped RL task and evaluates against NPG–PD.

What we will implement and measure (brief)

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

Work	Objective / Saddle	Policy Player	Guarantees / Notes
Zahavy et al. (2021) [Zahavy et al., 2021]	$\min_d f(d)$; Fenchel dual $\min_d \max_\lambda \lambda \cdot d - f^*(\lambda)$	Best response / low-regret RL under $r_\lambda = -\lambda$	$O(1/\sqrt{K})$ via OCO; unifies AL, CMDPs, exploration
Asad et al. (2025) [Asad et al., 2024]	RL inner step (fixed y)	SPMA: $\pi_{t+1} = \pi_t(1 + \eta A)$; FA via convex projection	Linear (tabular); neighbourhood (FA); strong empirical results
Ding et al. (2020) [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.)

Table 1: Three perspectives that our project unifies or compares against.

Table 2: Instantiations of the convex MDP framework (adapted from Zahavy et al. [2021], Table 1). Different choices of objective f and player algorithms recover well-known RL problems.

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

Metrics. (i) Saddle value $L(\pi, y)$ (when f^* known); (ii) constraint value/violation; (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.

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.