A Dual–SPMA Framework for Convex MDPs

Fenchel Duality + Softmax Policy Mirror Ascent

Shervin Khamooshian Ahmed Magd Pegah Aryadoost Danielle Nguyen

School of Computing Science, Simon Fraser University

Project Presentation

Main Goal

Implement Fenchel duality + a fast policy optimizer (SPMA) to solve convex MDPs;

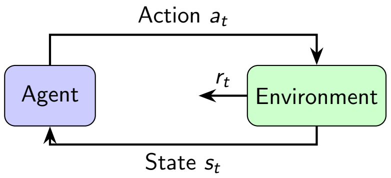
Outline

What is Reinforcement Learning?

Reinforcement Learning (RL) is a learning framework where an agent learns to make decisions by interacting with an environment.

At each time step t, the agent:

1. Observes a **state** s_t
2. Chooses an **action** a_t (based on a policy)
3. Receives a **reward** r_t
4. Transitions to a new **state** s_{t+1}



Markov Decision Process (MDP): Formal Definition

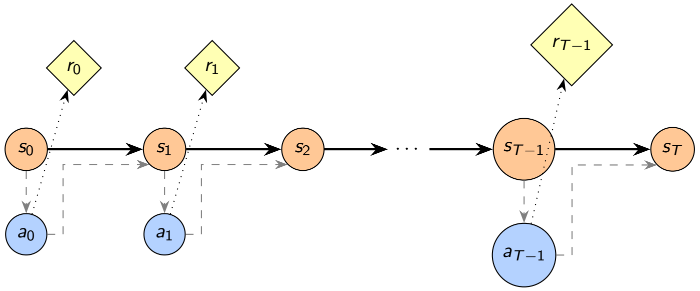
An **MDP** is defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho)$:

Symbol	**Meaning**
$\mathcal{S}$	State space (set of all possible states)
$\mathcal{A}$	Action space (set of all possible actions)
$P(s' \mid s, a)$	Transition probability: probability of reaching s' from (s, a)
$r(s, a)$	Reward function: immediate reward for taking action a in state s
$\gamma \in [0, 1)$	Discount factor: how much to value future vs. immediate rewards
$\rho(s)$	Initial state distribution

Policy $\pi(a \mid s)$: probability of taking action a in state s.

Trajectory and Return

A **trajectory** τ is a sequence of states, actions, and rewards:



Discounted Return

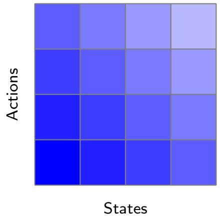
The **expected discounted return** under policy π is: $J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$

Goal of RL: Find $\pi^\star = \arg\max_\pi J(\pi)$

Occupancy Measure: Where the Policy Spends Time

Discounted Occupancy Measure

$$d_\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \Pr_\pi(s_t = s, a_t = a)$$



Think of d_π as a **heatmap**: bright = often visited

Notation:

- $d_\pi(s,a)$: probability of being in state s and taking action a under policy π
- $(1-\gamma)$: normalization factor
- $\Pr_\pi$: probability under policy π

Key property: d_π is a probability distribution: $\sum_{s,a} d_\pi(s,a) = 1$

Key Identity: RL is Linear in Occupancy

Fundamental Identity

$$\langle r, d_\pi \rangle = \sum_{s,a} r(s,a)\, d_\pi(s,a) = (1-\gamma)\, J(\pi)$$

What does this mean?

- Standard RL $\Rightarrow$ maximize a **linear** function of d_π

But many interesting goals are NOT linear in d_π:

Goal	**Objective**
Safety constraints	$\max J_r(\pi)$ subject to $\langle c, d_\pi \rangle \leq \tau$
Imitation learning	$\min \|d_\pi - d_{\text{expert}}\|$
Exploration	$\max J_r(\pi) + \alpha H(d_\pi)$

Why Convex MDPs?

Problem: Linear RL is insufficient for:

- Safety constraints
- Matching expert behavior
- Encouraging exploration
- Risk-sensitive objectives

Solution: Convex MDPs

$$\min_{\pi} f(d_\pi)$$

where f is a **convex function**

Challenge:

- Solving CMDPs directly involves optimization over a high-dimensional constrained space of stationary distributions

Our approach: Use Fenchel duality to transform the convex MDP into a minimax (saddle-point) problem.

- Apply RL algorithms to optimize over occupancy measures via policies.
- Use existing optimization techniques on the dual variables.

Roadmap

✓ Background: Reinforcement Learning

✓ Motivation: Convex MDPs

$\rightarrow$ **Problem Formulation: Fenchel Duality**

Related Work: “Reward Is Enough” & SPMA

Our Method: Dual–SPMA

Experiments

Conclusion & Future Work

Fenchel Conjugate: Definition

Given a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$

Fenchel Conjugate (Convex Conjugate)

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle y, x \rangle - f(x)\}$$

Notation:

- $f^*(y)$: the conjugate function evaluated at dual variable y
- $\langle y, x \rangle = \sum_i y_i x_i$: inner product (dot product)
- sup: supremum (least upper bound)

Intuition: $f^*(y)$ measures “how much $\langle y, x \rangle$ can exceed $f(x)$”

Fenchel–Moreau Theorem

Fenchel–Moreau Identity

For any proper, closed, convex function f:

$$f(d) = \sup_y \{\langle y, d\rangle - f^*(y)\}$$

What does this say?

- We can **recover** f from its conjugate f^*
- f is the conjugate of its conjugate: $f = (f^*)^*$
- This is called **biconjugation**

Why is this useful?

- Transforms a minimization problem into a **min-max** problem
- Introduces a **dual variable** y that we can optimize over

Applying Fenchel Duality to Convex MDPs

Step 1: Start with the convex MDP problem

$$\min_{d\in\mathcal{D}} f(d)$$

Step 2: Apply Fenchel–Moreau identity

$$\min_{d\in\mathcal{D}} f(d) = \min_{d\in\mathcal{D}} \sup_{y} \{\langle y, d\rangle - f^*(y)\}$$

Step 3: This is a **convex-concave saddle-point problem**

$$= \min_{d\in\mathcal{D}} \max_{y} \{\langle y, d\rangle - f^*(y)\}$$

(Under standard conditions, solving this saddle-point is equivalent to the original problem.)

Step 4: Replace d with d_π (occupancy induced by policy)

Saddle-Point Formulation

$$\min_{\pi} \max_{y} \underbrace{\langle y, d_\pi\rangle - f^*(y)}_{L(\pi,y)}$$

Two-Player Game Interpretation

Saddle-Point Problem

$$\min_{\pi} \max_{y} L(\pi, y), \quad \text{where } L(\pi, y) = \langle y, d_\pi \rangle - f^*(y)$$

This is a **min-max game** between two players:

Policy Player (min)

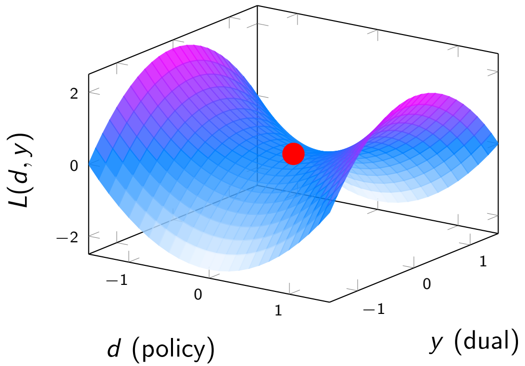
- Chooses policy π
- Wants to minimize L
- Controls occupancy d_π

Dual Player (max)

- Chooses dual variable y
- Wants to maximize L
- Shapes the reward signal

At equilibrium: policy player finds optimal π^*,
dual player finds optimal y^*

Visualizing the Saddle Point



Min-max solution:

$$\min_d \max_y L(d, y)$$

At the **red point** (saddle point):

- Along d-direction: **minimal**
- Along y-direction: **maximal**

Policy player (min) and dual player (max) reach **equilibrium** here.

From Saddle Point to Shaped Reward

For **fixed** dual variable y, the policy player solves: $\min_{\pi} \langle y, d_\pi \rangle$

Expand using the occupancy definition:

$$\langle y, d_\pi \rangle = \sum_{s,a} y(s,a) \cdot d_\pi(s,a) = (1-\gamma)\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t y(s_t, a_t) \right]$$

Key Insight: Shaped Reward

$$\min_{\pi} \langle y, d_\pi \rangle = \min_{\pi} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t y(s_t, a_t) \right] = \max_{\pi} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \underbrace{(-y(s_t, a_t))}_{r_y(s_t, a_t)} \right]$$

Conclusion: Policy player just does **standard RL** with shaped reward: $\boxed{r_y(s,a) = -y(s,a)}$

Summary: The Fenchel Dual Reduction

$$\begin{tikzcd}
\boxed{\begin{array}{c}\text{Convex MDP}\\ \min_\pi f(d_\pi)\end{array}} \arrow[r, "\text{Fenchel}"] & \boxed{\begin{array}{c}\text{Saddle-Point}\\ \min_\pi \max_y L(\pi, y)\end{array}} \arrow[r, "\text{Fix } y"] & \boxed{\begin{array}{c}\text{Shaped-}\\ \text{Reward RL}\\ r_y = -y\end{array}}
\end{tikzcd}$$

Algorithm structure:

1. **Dual step:** Update y using gradient of L w.r.t. y

$$y_{k+1} = y_k + \alpha \left(d_{\pi_k} - \nabla f^*(y_k)\right)$$

2. **Policy step:** Run RL algorithm with reward $r_{y_k} = -y_k$

$$\pi_{k+1} = \text{RL-Update}(\pi_k, r_{y_k})$$

Roadmap

✓ Background: Reinforcement Learning

✓ Motivation: Convex MDPs

✓ Problem Formulation: Fenchel Duality

$\rightarrow$ **Related Work: “Reward Is Enough” & SPMA**

Our Method: Dual–SPMA

Experiments

Conclusion & Future Work

Where We Are

We’ve established the Fenchel dual reduction. Now we review the key papers that inform our method: the theoretical foundation and the policy optimizer we’ll use.

Related Work: “Reward Is Enough” (Zahavy et al., 2021)

Main contributions of this foundational paper:

1. **Fenchel dual reduction:**
   - Reformulate convex MDP as saddle-point problem
   - The theoretical foundation we just presented
2. **Meta-algorithm:**
   - Alternating updates between policy and dual players
   - Any RL algorithm can be the policy player
   - Any online convex optimization (OCO) can be the dual player
3. **Unification:**
   - Shows many RL paradigms are special cases of convex MDPs
   - Imitation learning, constrained RL, entropy-regularized RL

Our contribution: Implement this framework with SPMA as the policy player

Related Work: Softmax Policy Mirror Ascent (Asad et al., 2024)

Softmax Policy Mirror Ascent (SPMA):

- Mirror ascent in *logit space* using log-sum-exp mirror map
- Achieves **linear convergence** in tabular MDPs

Tabular SPMA Update Rule

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

where $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ is the advantage function[1]

Why use SPMA as our policy player?

- **Geometry-aware:** Updates respect the simplex structure
- **No normalization:** Unlike vanilla PG, no per-state renormalization
- **Fast:** Linear convergence vs. sublinear for vanilla PG

[1] In tabular theory, this uses $[1+\eta A]_+$ and implicit renormalization; simplified form shown here.

Roadmap

- ✓ Background: Reinforcement Learning
- ✓ Motivation: Convex MDPs
- ✓ Problem Formulation: Fenchel Duality
- ✓ Related Work: “Reward Is Enough” & SPMA
- → **Our Method: Dual–SPMA**
- Experiments
- Conclusion & Future Work

Dual–SPMA Loop: High-Level View

Saddle-Point Problem

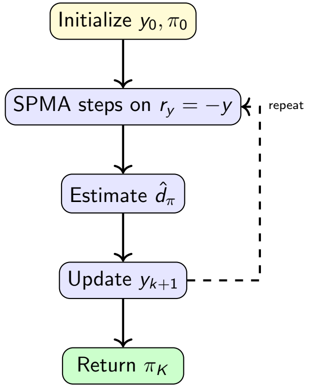
$$\min_{\pi} \max_{y} \underbrace{\langle y, d_\pi \rangle - f^*(y)}_{L(\pi,y)}$$

Outer loop (dual):

$$y_{k+1} = \arg\max_{y} L(\pi, y)$$

Inner loop (policy):

- Run K_{in} SPMA steps
- Shaped reward: $r_{y_k} = -y_k$



Occupancy Estimation: MC vs MLE

Monte Carlo (our default)

Tabular Estimator

$$\hat{d}_\pi(s,a) = \frac{1-\gamma}{N}\sum_{i=1}^{N}\sum_{t=0}^{T}\gamma^t \mathbf{1}\{s_t^{(i)} = s, a_t^{(i)} = a\}$$

- Simple: count discounted visits

MLE (Barakat et al., 2024)

Log-Linear Model

$$\lambda_\omega(s,a) \propto \exp(\omega^\top \phi(s,a))$$

- Fit ω by max-likelihood
- Independent of $|S||A|$!

Example: Constrained Safety CMDP

Problem: Maximize reward subject to safety constraint

$$\max_{\pi} J_r(\pi) \quad \text{s.t.} \quad J_c(\pi) \le \tau$$

where $J_r(\pi) = \mathbb{E}_\pi[\sum_t \gamma^t r(s_t, a_t)]$ and $J_c(\pi) = \mathbb{E}_\pi[\sum_t \gamma^t c(s_t, a_t)]$.

Dual–SPMA approach:

1. Build dual variable: $y_\lambda(s, a) = \lambda c(s, a) - r(s, a)$
2. Policy sees shaped reward: $r_y = -y = r - \lambda c$
3. Run SPMA inner loop on r_y
4. Update dual: $\lambda_{k+1} = [\lambda_k + \beta(J_c(\pi_k) - \tau)]_+$

Roadmap

- ✓ Background: Reinforcement Learning
- ✓ Motivation: Convex MDPs
- ✓ Problem Formulation: Fenchel Duality
- ✓ Related Work: “Reward Is Enough” & SPMA
- ✓ Our Method: Dual–SPMA
- $\rightarrow$ **Experiments**
- Conclusion & Future Work

Dual Game as Online Convex Optimization

Objective: find a saddle point of

$$\min_{d_\pi \in \mathcal{D}} \max_{\lambda \in \Lambda} L(d_\pi, \lambda) := \langle \lambda, d_\pi \rangle - f^*(\lambda).$$

Algorithm sketch:

1. Initialize dual variable λ_1 (e.g. $\lambda_1 = 0$) and a policy π_1.
2. For $k = 1, 2, \ldots, K$:
    1. **Policy player (min):** chooses π_k and induces occupancy $d_k := d_{\pi_k}$.
    2. **Cost player (max):** chooses λ_k.
    3. Both players interact through the *same* payoff

       $$L_k(d, \lambda) := L(d, \lambda).$$

       The policy player wants to make L_k *small* in d, the cost player wants to make L_k *large* in λ.

OCO/game perspective:

- Each round defines a convex–concave function $L_k(d, \lambda)$.
- Policy player runs an online **minimization** algorithm on $d \mapsto L_k(d, \lambda_k)$.
- Cost player runs an online **maximization** algorithm on $\lambda \mapsto L_k(d_k, \lambda)$.

Cost Player via FTL

Cost player uses Follow-The-Leader (FTL):

$$\lambda_k = \arg\max_{\lambda} \sum_{j=1}^{k-1} L(d^j, \lambda) = \arg\max_{\lambda} \sum_{j=1}^{k-1} \left(d^j \cdot \lambda - f^*(\lambda)\right).$$

Factor out $(k-1)$ and define the average occupancy $\bar{d}_{k-1} := \frac{1}{k-1} \sum_{j=1}^{k-1} d^j$:

$$\lambda_k = \arg\max_{\lambda} (k-1)\left(\bar{d}_{k-1} \cdot \lambda - f^*(\lambda)\right) = \arg\max_{\lambda} \left(\bar{d}_{k-1} \cdot \lambda - f^*(\lambda)\right).$$

For a smooth convex f the maximizer satisfies

$$\nabla_{\lambda}\left(\bar{d}_{k-1} \cdot \lambda - f^*(\lambda)\right) = 0 \quad \Longrightarrow \quad \bar{d}_{k-1} = \nabla f^*(\lambda_k^\star).$$

Using the convex-analysis identity

$$(\nabla f^*)^{-1} = \nabla f,$$

we obtain the clean update

$$\boxed{\lambda_k^\star = \nabla f(\bar{d}_{k-1})}.$$

Policy Player and Entropy Objective

Policy player with fixed λ:

$$\min_{\pi} L(d_\pi, \lambda) = \min_{\pi} \big(\langle \lambda, d_\pi \rangle - f^*(\lambda)\big) = \min_{\pi} \langle \lambda, d_\pi \rangle$$

since $f^*(\lambda)$ is constant w.r.t. π.

$$\Rightarrow \quad \text{RL problem with reward } r_\lambda(s, a) = -\lambda(s, a).$$

Entropy-based objective we test:

$$f(d_\pi) = -\beta \langle r, d_\pi \rangle - (1-\beta) H(d_\pi), \qquad H(d_\pi) = -\sum_{s,a} d_\pi(s, a) \log d_\pi(s, a).$$

Expanded form:

$$f(d_\pi) = -\beta \langle r, d_\pi \rangle + (1-\beta) \sum_{s,a} d_\pi(s, a) \log d_\pi(s, a).$$

This combines **reward maximization** (weighted by β) and **entropy regularization** (weighted by $1 - \beta$), encouraging diverse and high-reward state–action occupancies.

Algorithm

Gradient for cost player (FTL update):

$$\nabla f(d_\pi)(s,a) = -\beta\, r(s,a) + (1-\beta)\big(\log d_\pi(s,a) + 1\big).$$

Dual–SPMA algorithm (entropy objective):

1. Initialize policy π_1, estimate d_1, set $\bar{d}_1 = d_1$.
2. For $k = 1, \dots, K$:
    1. **Cost player (FTL):** $\lambda_k = \nabla f(\bar{d}_k) = -\beta r + (1-\beta)\big(\log \bar{d}_k + 1\big)$.
    2. **Policy player (RL):** run RL on reward $r_{\lambda_k}(s,a) = -\lambda_k(s,a)$, get new policy π_{k+1} and occupancy d_{k+1}.
    3. **Average occupancies:** $\bar{d}_{k+1} = \frac{k\,\bar{d}_k + d_{k+1}}{k+1}$.
3. Return $\bar{d}_{K+1}$.

(Zahavy et al., 2021) proved that the average occupancy $\bar{d}_K$ satisfies the regret bound $f(\bar{d}_K) - f_{\mathrm{OPT}} \leq O\Big(\frac{1}{\sqrt{K}}\Big)$, provided that both the cost and policy players have sub-linear regret.

Experimental Setup

Environments:

- FrozenLake 8×8 (tabular)
- Deterministic transitions
- Unsafe states = holes (cost $c = 1$)

Methods Compared:

- Dual–SPMA (ours)
- NPG–PD baseline

Hyperparameters:

- Discount $\gamma = 0.99$
- 100 outer iterations
- 1000 RL iterations
- 128 env steps/iteration

FrozenLake Environments (8×8)

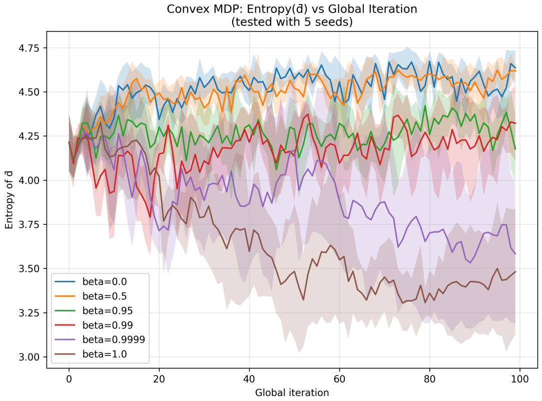
Agent: starts at S and tries to reach G

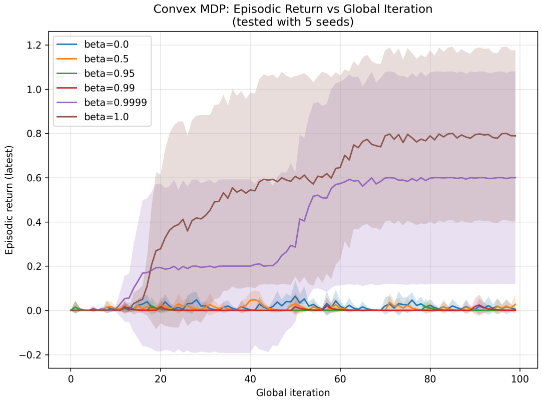
Holes $\Rightarrow$ episode ends, cost $c(s, a) = 1$

Reward: $+1$ at G, 0 elsewhere

Actions: up, down, left, right

Results Quantitative: Entropy-Regularized RL





Entropy of average occupancy $H(\bar{d}_\pi)$ vs iterations

Episodic return $J(\pi_k)$ vs iterations

Results Qualitative: Entropy-Regularized RL

$\beta = 1.0$ vs iterations

$\beta = 0.0$ vs iterations

Takeaways & Limitations

The Recipe:

$$\text{Convex MDP} \xrightarrow{\text{Fenchel}} \text{Saddle-Point Game} \xrightarrow{\text{FTL+RL}} \text{Shaped-Reward RL}$$

What we built:

- Dual–SPMA loops for entropy-regularized RL
- Three occupancy estimators (tabular MC)

Limitations:

- Experiments only in low-dimensional (tabular) environments
- Need to tune β
- Limitations of keeping track/saving d_π

Future Work

1. **Scale up:**
   - Test on larger CMDPs (continuous states/actions)
   - MuJoCo safety tasks
2. **Better estimation of d:**
   - Evaluate MLE estimator on high-dimensional tasks
   - Compare variance of different estimators
3. **More objectives:**
   - Risk-sensitive RL
   - Imitation learning via convex MDP
4. **Theoretical analysis:**
   - Convergence rates for Dual–SPMA
   - Sample complexity comparison with NPG–PD

Q&A

Questions?

References

1. **Zahavy, T., et al.** (2021). "Reward is Enough for Convex MDPs." *NeurIPS 2021*. arXiv:2108.06389
2. **Asad, A., et al.** (2024). "Fast Convergence of Softmax Policy Mirror Ascent." *arXiv:2405.09781*
3. **Ding, D., et al.** (2020). "Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes." *NeurIPS 2020*
4. **Barakat, A., et al.** (2024). "Reinforcement Learning with General Utilities: Simpler Variance Reduction and Large State-Action Space." *ICML 2024*
5. **Schulman, J., et al.** (2015). "Trust Region Policy Optimization." *ICML 2015*
6. **Sutton, R. & Barto, A.** (2018). "Reinforcement Learning: An Introduction." *MIT Press*

Backup: Flow Constraints (Occupancy Polytope)

The set $\mathcal{D}$ of valid occupancy measures satisfies **Bellman flow constraints**:
For all states s:

$$\sum_{a} d(s,a) = (1-\gamma)\rho(s) + \gamma \sum_{s',a'} P(s|s',a')\, d(s',a')$$

Also: $d(s,a) \geq 0$ for all (s,a)

Interpretation:

- Flow into state s = initial distribution + discounted flow from other states
- This is a **convex polytope** in $\mathbb{R}^{|S|\times|A|}$

Backup: Entropy-Regularized Conjugate

For entropy-regularized objective:

$$f(d) = -\langle r, d \rangle + \alpha \sum_{s,a} d(s,a) \log d(s,a)$$

$$f^*(y) = \alpha \log \sum_{s,a} \exp\left(\frac{y(s,a) + r(s,a)}{\alpha}\right)$$

$$\nabla f^*(y) = \mathrm{softmax}\left(\frac{y + r}{\alpha}\right)$$

The gradient is a softmax distribution—very convenient for computation!

Backup: SPMA with Function Approximation

In function approximation, the SPMA projection step becomes:

$$\theta_{t+1} = \arg\min_{\theta} \sum_{s} d^{\pi_t}(s)\, \mathrm{KL}\big(\pi_{t+1/2}(\cdot|s) \,\|\, \pi_\theta(\cdot|s)\big)$$

This is a **convex optimization problem** (weighted KL minimization).

Equivalent to **softmax classification**:

- Labels: actions from $\pi_{t+1/2}$
- Weights: state occupancies $d^{\pi_t}(s)$
- Features: state representations

Backup: Algorithm Pseudocode

Dual–SPMA Algorithm:

```
1. Initialize dual variable $y_1 = 0$, policy $\pi_1$ randomly
2. For $k = 1, 2, \ldots, K_{\text{outer}}$:
    1. **Policy step:** Run $K_{\text{inner}}$ SPMA iterations on shaped reward $r_{y_k} = -y_k$
           $$\pi_{k+1} = \text{SPMA}(\pi_k, r_{y_k}, K_{\text{inner}})$$
    2. **Estimate occupancy:** Collect trajectories, compute $\hat{d}^{\pi_{k+1}}$
    3. **Dual step:** Update dual variable
           $$y_{k+1} = y_k + \alpha \left( \hat{d}^{\pi_{k+1}} - \nabla f^*(y_k) \right)$$
3. Return final policy $\pi_K$
```

Backup: Notation Summary

Symbol	**Meaning**
$\mathcal{S}, \mathcal{A}$	State and action spaces
$\pi(a\|s)$	Policy (probability of action a in state s)
$r(s, a)$	Reward function
$c(s, a)$	Cost function (for CMDPs)
γ	Discount factor
$d_\pi(s, a)$	Occupancy measure under policy π
$J(\pi)$	Expected return
$f(d)$	Convex objective over occupancies
$f^*(y)$	Fenchel conjugate of f
y	Dual variable
λ	Lagrange multiplier (for CMDPs)
τ	Safety threshold
η	SPMA step size
α	Dual step size