Project Title

Author 1
Email

1 Reward is Enough

While maximizing a cumulative reward function that is Markov and stationary is sufficient to capture many kinds of goals in a Markov Decision Process (MDP), not all objectives can be represented in this way. To address this limitation, the paper considers *convex MDPs*, where goals are expressed as convex functions of the stationary distribution rather than as the linear functions used in the standard RL formulation. This generalization enables the framework to handle a broader range of problems in both supervised and unsupervised RL settings, including apprenticeship learning, constrained MDPs, and pure exploration.

A key theoretical contribution is the use of *Fenchel duality* to reformulate this convex minimization problem. This powerful tool transforms the original objective $\min_{d_\pi \in K} f(d_\pi)$ into an equivalent convex–concave (zero-sum) game between the agent and an adversary. Applying the Fenchel–Moreau identity to the objective yields the equivalent min–max saddle-point formulation:

$$f_{\text{OPT}} = \min_{d_\pi \in K} \max_{\lambda \in \Lambda} (\lambda \cdot d_\pi - f^*(\lambda)).$$

The resulting formulation defines a two-player zero-sum game between a **policy player**, who seeks to maximize a reward, and a **cost player**, who adjusts the dual variable λ while being regularized by $f^*(\lambda)$. For any fixed value of λ, minimizing the Lagrangian with respect to the policy's occupancy measure d_π is equivalent to solving a standard reinforcement learning problem with a reward function defined as $r = -\lambda$. This demonstrates that the complex convex goal can be achieved by simply maximizing a reward, effectively showing that "reward is enough" for this class of problems.

Inspired by this result, the authors propose a *meta-algorithm* that solves convex MDPs by implementing the game through an alternating learning process. At each iteration, the cost player (using an Online Convex Optimization method) selects a dual variable λ_k, and the policy player responds by running a standard RL algorithm to maximize the instantaneous reward $r_k = -\lambda_k$, yielding an occupancy measure d_π^k. After K rounds, the algorithm outputs the averaged occupancy $\bar{d}_\pi = \frac{1}{K}\sum_{k=1}^{K} d_\pi^k$. The paper proves that if both subroutines achieve sublinear regret, this average policy converges to the optimal solution of the convex MDP.

Another key contribution of the paper is the *unification* of several reinforcement-learning paradigms within the convex MDP framework. By selecting different convex functions $f(d_\pi)$, the same Fenchel-dual formulation recovers well-known objectives: imitation learning when f is a divergence from an expert distribution, constrained RL when f includes penalty terms, and pure exploration when f is the entropy of the occupancy measure. This shows that diverse RL settings—supervised, constrained, and unsupervised—can all be derived from a single convex-duality perspective.

2 Natural Policy Gradient Primal–Dual Method for Constrained Markov Decision Processes

This paper studies problems modeled as *Constrained Markov Decision Processes (CMDPs)*, where an agent seeks to maximize the expected discounted reward while satisfying a constraint on the expected discounted utility (or cost).

While policy gradient methods have achieved great success for unconstrained MDPs, their theoretical guarantees for CMDPs have been mostly asymptotic or local. To address the lack of non-asymptotic global convergence guarantees, the authors propose a new algorithm called the *Natural Policy Gradient Primal–Dual (NPG-PD)* method.

The problem is formulated via the Lagrangian function:

$$L(\pi, \lambda) = V_r^\pi(\rho) + \lambda\big(V_g^\pi(\rho) - b\big),$$

where V_r^π denotes the expected reward, V_g^π represents the expected utility (or safety metric), and b is the constraint threshold.

The analysis assumes the *Slater condition* (strict feasibility), which ensures strong duality and the boundedness of the dual variable λ.

This leads to a *primal–dual* optimization problem:

$$\max_{\pi} \min_{\lambda \geq 0} L(\pi, \lambda),$$

where λ is a nonnegative Lagrange multiplier that enforces the constraint.

The NPG-PD algorithm alternates between improving the policy (the *primal variable*) and adjusting the Lagrange multiplier (the *dual variable*) that enforces the constraint. Each iteration consists of two coordinated updates:

- **Primal step:** The policy parameters are updated using the *natural policy gradient*, which rescales the gradient by the inverse Fisher information matrix. This geometry-aware update enables smoother and more efficient adjustments to the policy's probability distribution, improving stability and accelerating convergence.
- **Dual step:** The Lagrange multiplier λ is updated via *projected subgradient descent*, increasing when the constraint is violated and decreasing when it is satisfied.

The authors first analyze the NPG-PD algorithm under the *softmax policy parametrization*, where the policy assigns action probabilities through a softmax function. Using this setting, they establish a **global convergence guarantee** (Theorem 1), showing that, with appropriate step sizes, both the *optimality gap* (the difference between the achieved and optimal reward) and the *constraint violation* decrease at a rate of $\mathcal{O}(1/\sqrt{T})$. Thus, to achieve an ε-optimal and ε-feasible policy, the algorithm requires $\mathcal{O}(1/\varepsilon^2)$ iterations, independent of the sizes of the state and action spaces. This result represents a significant improvement over prior CMDP methods that exhibited slower convergence.

The analysis is then extended to *general policy classes*, such as neural network policies, where *strong duality may not hold*. The method employs an approximate natural policy gradient characterized by the *compatible function approximation error*, which measures how well the parameterized policy represents the true Lagrangian advantage function.

Assuming policy smoothness and using an *exploratory initial distribution* to mitigate state–action mismatch, **Theorem 2** establishes *sublinear convergence*: the optimality gap decreases as $\mathcal{O}(1/\sqrt{T})$ and the constraint violation as $\mathcal{O}(T^{-1/4})$. Despite approximation and exploration errors, the NPG-PD algorithm achieves stable convergence for general differentiable policy classes.

Finally, the paper presents *model-free (sample-based)* versions of the NPG-PD algorithm, which estimate gradients and value functions from sampled trajectories. Under standard smoothness and bounded-error assumptions, these algorithms retain sublinear convergence: the optimality gap scales as $\mathcal{O}(1/\sqrt{T})$ and the constraint violation as $\mathcal{O}(T^{-1/4})$.

For *softmax policies*, convergence is faster and dimension-free. Empirical results show that the sample-based NPG-PD performs comparably to *dualDescent* (comparison in Appendix) while being simpler and computationally more efficient.

3 Softmax Policy Mirror Ascent (SPMA): Tabular and Function Approximation Settings

This paper bridges the gap between theoretically grounded policy gradient (PG) algorithms—such as Natural Policy Gradient (NPG), which enjoy strong convergence guarantees only in tabular settings—and practical deep RL algorithms like PPO, TRPO, and MDPO, which perform well empirically but lack rigorous analysis. The authors propose *Softmax Policy Mirror Ascent (SPMA)*, a refined, normalization-free mirror ascent algorithm that operates in the **dual (logit) space**. SPMA retains the linear convergence guarantees of NPG while extending naturally to large-scale problems with function approximation.

SPMA represents the policy as a softmax over logits,

$$\pi(\cdot|s) = \mathrm{softmax}(z(s,\cdot)),$$

and performs **mirror ascent** in the logit space using the log-sum-exp mirror map:

$$z_{t+1} = \arg\max_{z} \left[\langle z - z_t, \nabla_z J(z_t)\rangle - \tfrac{1}{\eta} D_\Phi(z, z_t)\right],$$

where D_Φ is the Bregman divergence. This corresponds to gradient ascent in the policy's natural geometry *without the need for explicit normalization across actions.* In the MDP case, weighting by the discounted state distribution $d_{\pi_t}(s)$ yields the per-state update:

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

which is linear in both step size and advantage. Unlike NPG's exponential normalization or SPG's direct logit updates, this formulation improves numerical stability and convergence speed.

The authors prove **global linear convergence** of SPMA in both **bandit** and **tabular MDP** settings. For bandits, SPMA achieves exponential (linear) convergence to the optimal arm, and a gap-dependent variant achieves **super-linear** convergence—the first such result for policy gradient methods. For tabular MDPs, they establish:

$$\|V^{\pi^*} - V^{\pi_T}\|_\infty \le \prod_{t=0}^{T-1}(1 - \eta C_t(1-\gamma))\, \|V^{\pi^*} - V^{\pi_0}\|_\infty,$$

where C_t reflects the per-state advantage gap. Thus, with a constant step size $\eta < \min(1-\gamma, [C_t(1-\gamma)]^{-1})$, SPMA converges linearly to the optimal value function—matching NPG's rate but without dependence on the potentially large distribution-mismatch term.

To handle large or continuous state–action spaces, SPMA is extended using **projected mirror ascent** under a log-linear or neural policy parameterization. The logits $z_\theta(s,a)$ are constrained to lie in a realizable function class $Z = \{f_\theta(s,a)\}$, and each update projects the ideal (tabular) step back into this feasible set via convex softmax classification:

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

Unlike NPG, this does not require compatible funcStion approximation; unlike MDPO, the per-iteration subproblem remains convex for linear FA. Under bounded rewards, expressive parameterization, and adequate sampling, **Algorithm 1** achieves **linear convergence to a neighborhood** of the optimal value function. The residual error depends on the function-approximation bias and statistical error terms ($\varepsilon_{\text{bias}}, \varepsilon_{\text{stat}}$), which vanish as data or inner-loop optimization improves.

SPMA is evaluated on (i) tabular MDPs, (ii) continuous-state discrete-action MDPs (Atari), and (iii) continuous control tasks (MuJoCo).

- In **tabular settings**, SPMA matches NPG and consistently outperforms SPG.
- With **linear FA**, it surpasses MDPO in challenging benchmarks like *CliffWorld*.
- In **deep RL benchmarks**, SPMA achieves performance comparable to or exceeding PPO, TRPO, and MDPO—especially on Atari, where it surpasses both TRPO variants.

Unlike TRPO-constrained, SPMA is computationally efficient and free of heavy hyperparameter tuning, demonstrating that strong theoretical guarantees can coexist with practical robustness.

A Appendix