A Theory of Regularized Markov Decision Processes

Matthieu Geist, Bruno Scherrer, Olivier Pietquin
ICML 2019

STA4273 Paper Presentation

Presenter: Weizheng Zhang, University of Toronto

March 2021

This paper introduces a framework of dynamic programming (DP) algorithms to study the propagation of errors of regularized RL algorithms, where the greediness is softened by convex regularizers.

Main ideas

- incorporate a larger class of regularizers;
- penalize a divergence between consecutive policies;
- consider the general modified policy iteration.

Key contributions

- propose a general theory of regularized MDPs;
- allow for error propagation analyses of general algorithmic schemes of which classical algorithms are special cases;
- use tools in constrained convex optimization to analyze trust-region algorithms in RL.

Why use regularization in RL algorithms?

- To encourage exploration targeted at high-value actions, and prevent agents from overfitting to certain actions.
- To increase robustness to stochastic noises and environment perturbations.
- To improve convergence properties.
- To prevent earlier convergence to sub-optimal policies.
- Some regularizers can be used for sparse and non-deterministic greedy policies, e.g., Tsallis entropy.
- ....

Notations

Δ_X	the set of probability distributions over a set X
Y^X	the set of applications from X to a set Y
$\mathbb{E}$	expectation
$\langle \cdot, \cdot \rangle$	dot product
$\| \cdot \|_p$	ℓ_p-norm
$\| \cdot \|_\infty$	supremum norm

Markov Decision Processes (MDPs)

Consider an infinite-horizon discounted MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$.

- state space $\mathcal{S}$ and action space $\mathcal{A}$
- Markovian transition kernel $P \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$
  $P(s'|s,a)$ is the prob. of transiting to s' when action a is taken in state s.
- reward function $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$
  $r(s,a)$ is the reward when action a is taken in state s.
- discount factor $\gamma \in (0,1)$

A **policy** $\pi \in \Delta_{\mathcal{S}}^{\mathcal{A}}$ associates each state to a distribution over $\mathcal{A}$. $\pi(a|s)$ is the prob. of taking action a in state s.
The **value function** of state s under policy π, is the expected cumulative discounted reward of starting in s and following π:

$$v_\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, a_k) \,\middle|\, s_0 = s\right], \ \forall s \in \mathcal{S}.$$

Bellman operator

The **Bellman operator** of π applied to a value function $v \in \mathbb{R}^{\mathcal{S}}$ is

$$[T_\pi v](s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[r(s,a) + \gamma \mathbb{E}_{s'|s,a}[v(s')] \right], \ \forall s \in \mathcal{S}.$$

Equivalently,

$$T_\pi v = r_\pi + \gamma P_\pi v,$$

where $r_\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[r(s,a)]$ and $P_\pi(s'|s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[P(s'|s,a)]$.

- T_π is a γ-contraction in supremum norm, i.e.,

$$\|T_\pi v_1 - T_\pi v_2\|_\infty \leqslant \gamma \|v_1 - v_2\|_\infty, \ \forall v_1, v_2 \in \mathbb{R}^{\mathcal{S}}.$$

- The unique fixed point of T_π is the value function v_π.

Bellman optimallity operator

The **Bellman optimality operator** is,

$$T_* v = \max_{\pi} T_\pi v, \; \forall v \in \mathbb{R}^{\mathcal{S}}.$$

- T_* is also a γ-contraction in supremum norm.
- The fixed point of T_* is the optimal value function v_*.

Greedy policy

π' is called a **greedy policy** w.r.t. v if

$$T_* v = T_{\pi'} v,$$

and denoted as

$$\pi' \in \mathcal{G}(v) = \operatorname*{argmax}_{\pi} T_\pi v.$$

Legendre-Fenchel transform

The **Legendre-Fenchel transform** (or **convex conjugate**) of a strongly convex function $\Omega : \Delta_{\mathcal{A}} \to \mathbb{R}$, is $\Omega^* : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$ with

$$\Omega^*(q_s) = \max_{\pi_s \in \Delta_{\mathcal{A}}} \left\{ \langle \pi_s, q_s \rangle - \Omega(\pi_s) \right\}, \ \forall q_s \in \mathbb{R}^{\mathcal{A}}.$$

Example

- $\Omega(\pi_s) = \sum_{a \in \mathcal{A}} \pi_s(a) \ln \pi_s(a)$ is the negative entropy, and its convex conjugate is

$$\Omega^*(q_s) = \ln \sum_{a \in \mathcal{A}} \exp q_s(a),$$

with the maximizing argument

$$\pi_s = \left(\frac{\exp q_s(a)}{\sum_{a' \in \mathcal{A}} \exp q_s(a')} \right)_{a \in \mathcal{A}}.$$

Regularized Bellman operator

The core idea is to regularize the Bellman operator by a strongly convex and differentiable function Ω of policy π.

The **regularized Bellman operator** is [1]

$$T_{\pi,\Omega}v := T_\pi v - \Omega(\pi),\ \forall v \in \mathbb{R}^{\mathcal{S}}.$$

- $T_{\pi,\Omega}$ is also a γ-contraction in supremum norm.

[1] $\Omega(\pi) := (\Omega(\pi_s))_{s\in\mathcal{S}} \in \mathbb{R}^{\mathcal{S}}$ with a slight abuse of notation.

Regularized Bellman optimallity operator

To get the optimality operator, perform state-wise maximization over $\pi_s \in \Delta_{\mathcal{A}}$ which gives the convex conjugate of $[T_{\pi,\Omega}v](s)$.

- The **regularized Bellman optimality operator** is

$$T_{*,\Omega}v = \max_{\pi} T_{\pi,\Omega}v = \Omega^*(q),\ \forall v \in \mathbb{R}^{\mathcal{S}}.$$

- $T_{*,\Omega}$ is also a γ-contraction in supremum norm.

The related maximizing argument defines the greediness.

- For any $v \in \mathbb{R}^{\mathcal{S}}$, the associated **unique** greedy policy is

$$\pi' = \mathcal{G}_{\Omega}(v) = \nabla\Omega^*(q) \iff T_{\pi',\Omega}v = T_{*,\Omega}v.$$

Regularized value functions

The regularized operators being contractions, we can define regularized value functions as their unique fixed points.

- The **regularized value function** of policy π, $v_{\pi,\Omega}$, is the unique fixed point of $T_{\pi,\Omega}$, i.e.,

$$v_{\pi,\Omega} = T_{\pi,\Omega} v_{\pi,\Omega}.$$

- Alternatively, the regularized value is just the unregularized value of π for the reward $r_\pi - \Omega(\pi)$.

$$v_{\pi,\Omega} = (I - \gamma P_\pi)^{-1} (r_\pi - \Omega(\pi)) .$$

Regularized optimal value functions

Regularized optimality operators are contractions, so we can define regularized optimal value functions as their unique fixed points.

- The **regularized optimal value function** $v_{*,\Omega}$ is the unique fixed point of $T_{*,\Omega}$, i.e.,

$$v_{*,\Omega} = T_{*,\Omega} v_{*,\Omega}.$$

- $v_{*,\Omega}$ is indeed the optimal value function. (see the next theorem)

Optimal regularized policy

Theorem (Optimal regularized policy)

$\pi_{*,\Omega} = \mathcal{G}_\Omega(v_{*,\Omega})$ is the unique optimal regularized policy, i.e.,

$$v_{\pi_{*,\Omega},\Omega} = v_{*,\Omega} \geqslant v_{\pi,\Omega},\ \forall \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}.$$

- This theorem shows that in a regularized MDP, the policy greedy w.r.t. the optimal value function is indeed the optimal policy.
- The optimal regularized policy is unique, because of the strong convexity of Ω. In contrast, there may exist multiple optimal policies in an unregularized MDP.

When regularizing the MDP, the optimal policy changes. The following results relate value functions in (un)regularized MDPs.

Theorem

For any policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ *and any state* $s \in \mathcal{S}$,

$$\frac{L_\Omega}{1-\gamma} \leqslant v_\pi(s) - v_{\pi,\Omega}(s) \leqslant \frac{U_\Omega}{1-\gamma},$$

$$\frac{L_\Omega}{1-\gamma} \leqslant v_*(s) - v_{*,\Omega}(s) \leqslant \frac{U_\Omega}{1-\gamma},$$

and

$$0 \leqslant v_*(s) - v_{\pi_*,\Omega}(s) \leqslant \frac{U_\Omega - L_\Omega}{1-\gamma},$$

where $U_\Omega = \sup \Omega$ *and* $L_\Omega = \inf \Omega$.

Modified Policy Iteration (MPI)

MPI is a classical DP algorithm that alternates between policy improvement and policy evaluation.

Given initial value v_0 and $m \in \mathbb{Z}_+ \cup \{\infty\}$, at $(k+1)$-th iteration,

$$\begin{cases} \pi_{k+1} = \mathcal{G}(v_k) & \text{greedy step,} \\ v_{k+1} = \left(T_{\pi_{k+1}}\right)^m v_k & \text{evaluation step.} \end{cases}$$

- Value Iteration (VI, $m = 1$): Since π_{k+1} is greedy w.r.t. v_k,

$$v_{k+1} = T_{\pi_{k+1}} v_k = T_* v_k.$$

- Policy Iteration (PI, $m = \infty$): Since $v_k = \left(T_{\pi_k}\right)^\infty v_{k-1} = v_{\pi_k}$,

$$\pi_{k+1} = \mathcal{G}(v_{\pi_k}).$$

Modified Policy Iteration (MPI)

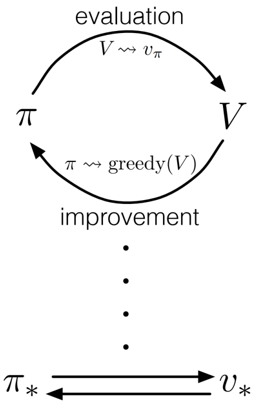
evaluation
$V \rightsquigarrow v_\pi$
π
V
$\pi \rightsquigarrow \text{greedy}(V)$
improvement
π_*
v_*

Figure: evaluation and improvement processes interact[1]

[1] Sutton and Barto (2018). Reinforcement Learning: An Introduction.

Regularized Modified Policy Iteration (Reg-MPI)

Reg-MPI:

$$\begin{cases} \pi_{k+1} = \mathcal{G}_\Omega(v_k) & \text{regularized greedy step,} \\ v_{k+1} = \left(T_{\pi_{k+1},\Omega}\right)^m v_k & \text{regularized evaluation step.} \end{cases}$$

Extreme cases

- If $m = 1$, then $v_{k+1} = T_{*,\Omega} v_k$. $\longrightarrow$ Regularized VI.
  (e.g. Soft Q-Learning)
- If $m = \infty$, then $\pi_{k+1} = \mathcal{G}_\Omega(v_{\pi_k,\Omega})$. $\longrightarrow$ Regularized PI.
  (e.g. Soft Actor Critic)

Bregman divergence

The **Bregman divergence** generated by a strongly convex regularizer Ω, between π and a reference policy π', is

$$\Omega_{\pi'}(\pi) := D_{\Omega}(\pi || \pi') \equiv \Omega(\pi) - \Omega(\pi') - \langle \nabla\Omega(\pi'), \pi - \pi' \rangle.$$

Examples:

- The KL divergence is generated by the negative entropy.
- The Euclidean distance is generated by the ℓ_2-norm.

Properties:

- $\Omega_{\pi'}(\pi)$ is strongly convex in π;
- $\Omega_{\pi'}(\pi) \geqslant 0$ and $\Omega_{\pi'}(\pi') = 0$;
- $\Omega_{\pi'}(\pi) < \infty \iff \mathrm{supp}(\pi) \subset \mathrm{supp}(\pi')$.

Mirror Descent (MD)

Mirror Descent (MD) is a first-order optimization method for solving constrained convex problems. It solves a maximization problem by iterating as follows[1]:

$$\begin{cases} g_k = \nabla f(x_k), \\ x_{k+1} = \underset{x}{\operatorname{argmax}} \left\{ \eta \langle x, g_k \rangle - D_\Omega(x || x_k) \right\}. \end{cases}$$

MD consists at each step in maximizing a linearization of the function of interest, with the constraint of not moving to far from the previous iterate, this constraint being quantified by the Bregman divergence.

[1] Beck and Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 2003.

Mirror Descent MPI (MD-MPI)

The idea of **MD-MPI** is to regularize the greediness by a Bregman divergence between π and the previous policy.

At the $(k+1)$-th greedy step,

$$\pi_{k+1} = \mathcal{G}_{\Omega_{\pi_k}}(v_k) \equiv \underset{\pi}{\operatorname{argmax}} \left\{ T_\pi v_k - \Omega_{\pi_k}(\pi) \right\} .$$

- In Reg-MPI, the regularization is fixed across all iterations, while in MD-MPI it changes at different iterations.
- In each policy improvement step, the output policy won't be very far from the previous one.

Mirror Descent MPI (MD-MPI)

Similarly, in the evaluation step, employ regularization according to the previous policy[1]:

$$v_{k+1} = \left(T_{\pi_{k+1},\Omega_{\pi_k}}\right)^m v_k.$$

In summary, **MD-MPI** is a general algorithmic scheme based on Bregman divergence:

$$\begin{cases} \pi_{k+1} = \mathcal{G}_{\Omega_{\pi_k}}(v_k) & \text{regularized greedy step,} \\ v_{k+1} = \left(T_{\pi_{k+1},\Omega_{\pi_k}}\right)^m v_k & \text{regularized evaluation step.} \end{cases}$$

[1] We can also use the current policy π_{k+1}, leading to an unregularized evaluation.

Error propagation of MD-MPI

Approximation arises from practical settings with large spaces. Consider evaluation error ϵ_k and greedy error ϵ'_k in the k-th step.

$J_k(\pi) = T_{\pi,\Omega_{\pi_k}} v_k$ is the optimization problem corresponding to the Bregman divergence regularized greediness.

We write $\pi_{k+1} \in \mathcal{G}^{\epsilon'_{k+1}}_{\Omega_{\pi_k}}(v_k)$ if

$$\langle \nabla J_k(\pi_{k+1}), \pi - \pi_{k+1} \rangle \leqslant \epsilon'_{k+1}, \ \forall \pi.$$

This condition leads to

$$T_{\pi,\Omega_{\pi_k}} v_k - T_{\pi_{k+1},\Omega_{\pi_k}} v_k \leqslant \epsilon'_{k+1}, \ \forall \pi,$$

which implies that π_{k+1} is ϵ'_{k+1}-close to the regularized greedy policy.

Error propagation of MD-MPI

Consider MD-MPI with errors in greedy and evaluation steps:

$$\begin{cases} \pi_{k+1} = \mathcal{G}^{\epsilon'_{k+1}}_{\Omega_{\pi_k}}(v_k), \\ v_{k+1} = \left(T_{\pi_{k+1},\Omega_{\pi_k}}\right)^m v_k + \epsilon_{k+1}. \end{cases}$$

Regularization generally leads to convergence to a policy different from the optimal greedy policy of the unregularized problem.

It's important to control the sub-optimality of regularized optimal policy, and we're interested in the **loss** $l_k = v_* - v_{\pi_k}$ and the **regret** $L_K = \sum_{k=1}^K l_k$, measuring the sub-optimality in unregularized MDPs.

At the k-th iteration, the distance between the optimal value function and the value before approximation is

$$d_k = v_* - \left(T_{\pi_k, \Omega_{\pi_{k-1}}}\right)^m v_{k-1} = v_* - (v_k - \epsilon_k),$$

and the Bellman residual is

$$b_k = v_k - T_{\pi_{k+1}, \Omega_{\pi_k}} v_k.$$

Let ρ and μ be distributions and $p, q, q' > 0$ such that $\frac{1}{q} + \frac{1}{q'} = 1$. Define the **concentrability coefficient**

$$C_q^i := \frac{1-\gamma}{\gamma^i} \sum_{j=i}^{\infty} \gamma^j \max_{\pi_1, \dots, \pi_j} \left\| \frac{\rho P_{\pi_1} P_{\pi_2} \cdots P_{\pi_j}}{\mu} \right\|_{q,\mu}.$$

Let $R_{\Omega_{\pi_0}} := \| \sup_\pi \Omega_{\pi_0}(\pi) \|_\infty$.

Theorem (weighted ℓ_p-bound for the regret)

$$
\begin{aligned}
\|L_K\|_{p,\rho} \leqslant{}& \sum_{k=1}^{K}\sum_{i=0}^{k-1}\frac{\gamma^i}{1-\gamma}\left(C_q^i\right)^{\frac{1}{p}}\left\{2\|\epsilon_{k-i}\|_{pq',\mu}+\|\epsilon'_{k-i}\|_{pq',\mu}\right\}\\
&+\sum_{k=1}^{K}\frac{2\gamma^k}{1-\gamma}\left(C_q^k\right)^{\frac{1}{p}}\min\left\{\|d_0\|_{pq',\mu},\|b_0\|_{pq',\mu}\right\}\\
&+\frac{1-\gamma^K}{(1-\gamma)^2}R_{\Omega_{\pi_0}}.
\end{aligned}
$$

Regularized optimal policies are more stochastic than their unregularized counterparts in classical DP, so they can improve exploration. Stochastic policies induce lower concentrability coefficients, and hence lower regret bounds.

Theorem (weighted ℓ_1-bound for the loss)

For $p \geqslant 1$ and a distribution ρ, we have

$$\min_{1 \leqslant k \leqslant K} \|v_* - v_{\pi_k}\|_{1,\rho} \leqslant \frac{\|L_K\|_{p,\rho}}{K}.$$

This result implies that if we can control the average regret, then we can control the loss of the best policy computed so far. Therefore, practically we should not use the last policy, but this best policy.

Theorem (Rate of convergence in the exact case)

If no approximation is done, i.e., $\epsilon_k = \epsilon'_k = 0$*, then*

$$\|L_K\|_\infty \leqslant \frac{1-\gamma^K}{(1-\gamma)^2}\left(2\gamma\|v_* - v_0\|_\infty + R_{\Omega_{\pi_0}}\right).$$

In this exact case, we only have a logarithmic convergence rate, while in classical DP, there is a linear convergence rate ($\frac{2\gamma^K}{1-\gamma}\|v_* - v_0\|_\infty$).

We also pay an horizon factor with a quadratic dependency in $\frac{1}{1-\gamma}$ instead of linear.

Experiment

MDP setting:

- state space $\mathcal{S} = \{s_1, s_2\}$ and action space $\mathcal{A} = \{a_1, a_2\}$
- transition kernel $P(s'|a, s)$:

$P(s'\|a, s = s_1)$		next state s'	
		s_1	s_2
action a	a_1	0.3	0.7
	a_2	0.8	0.2

$P(s'\|a, s = s_2)$		next state s'	
		s_1	s_2
action a	a_1	0.6	0.4
	a_2	0.1	0.9

- reward function $r(a, s)$:

$r(a, s)$		state s	
		s_1	s_2
action a	a_1	0.1	0.3
	a_2	0.2	0.1

- discount factor $\gamma = 0.9$

Experiment

Algorithm setting:

- MD-MPI with $m = 3$
- regularizers: KL divergence, Euclidean and Itakura-Saito (IS) distance
- approximation errors: $\epsilon_k = \epsilon'_k = 0$ (the exact case)
- initial value $v_0 = (0, 0)$

Results: red points for regrets in $\|\cdot\|_\infty$ and blue curves for upper bounds

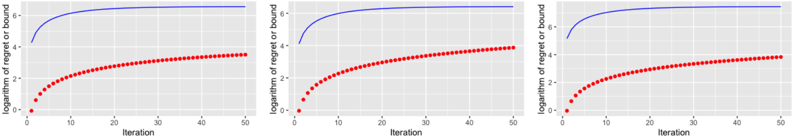

Figure: Regularizers are KL divergence, Euclidean distance and IS distance, resp.

Why is this framework general?

- negative entropy & KL divergence ⟶ Bregman divergence
- Value Iteration & Policy Iteration ⟶ Modified Policy Iteration

RL algorithms using regularization

Algorithm	Regularizer	Scheme
DPP[1]	KL divergence	VI
TRPO[2]	KL divergence	PI
SQL[3]	negative entropy	VI
SAC[4]	negative entropy	PI
CSTE[5]	Tsallis entropy	VI
This Paper	Bregman divergence	MPI

[1] Dynamic Policy Programming (Azar et al., 2012)

[2] Trust Region Policy Optimization (Schulman et al., 2015)

[3] Soft Q-Learning (e.g. Fox et al., 2016; Schulman et al., 2017)

[4] Soft Actor Critic (Haarnoja et al., 2018a)

[5] Causal Sparse Tsallis Entropy Regularization (Lee et al., 2018)

Limitations

1. As the concentrability coefficients C_q^i's depend on the worst case of all policies, they may be infinite and lead to meaningless bounds.

2. Require a closed-form relation between policy and optimal value function, and the knowledge on model dynamics, which may be intractable or unavailable.

3. Assume access to an oracle that returns the gradient of value functions for any policy. When parameters are partially known, the gradients have to be estimated from observations or simulations.

4. Experiment plots show that the regret bounds are not very sharp, perhaps sacrificing tightness of bounds for generality of analysis.

Future research

1. Combine the propagation of errors with a finite sample analysis.

2. What specific regularizer should be chosen for what context.

3. Generalize the Bregman divergence regularized MDPs to multi-agent settings like Markov games and mean-field games.

4. Connect links between approximate DP and proximal convex optimization, going beyond mirror descent.

5. Study Bregman divergence regularized policy search methods.

6. Use the proposed scheme to analyze inverse RL approaches.

......

Summary

This paper proposes a general theory of regularized MDPs, where the Bellman operator is modified by fixed convex functions or Bregman divergence between consecutive policies.

For both cases, it proposes a general algorithmic scheme based on MPI, which generalizes both value and policy-based regularized methods. It shows this algorithmic scheme incorporates many variations of existing algorithms for approximate planning under regularized notions of optimality.

This paper also analyzes the propagation of errors, and provides provable average case guarantees for the regret.

Thanks!