A Theory of Regularized MDPs

Matthieu Geist
mfgeist@google.com
with Bruno Scherrer (Inria) and Olivier Pietquin (Google)

Motivations

$$\mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t \left(r(S_t, A_t) + \alpha \mathcal{H}(\pi(\cdot|S_t)) \right) \right]$$

- Many recent (deep) RL algorithms make use of regularization (SAC, soft Q-learning, DPP, TRPO, MPO, etc.).
- They share the use of regularization, but are derived from different principle, consider specific regularization, and have ad-hoc analysis, if any.
- This work, generalizes in two directions:
  - larger class of regularizers,
  - the general modified policy iteration scheme.
- Allows for a general theoretical analysis, suggests new algorithmic schemes.

- Bellman evaluation operator

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

For short, $T_\pi v = r_\pi + \gamma P_\pi v$. For any v, we associate

$$q(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}[v(s')].$$

We'll write $[T_\pi v](s) = \langle \pi(\cdot|s), q(s,\cdot)\rangle = \langle \pi_s, q_s\rangle$. With a slight abuse of notation, $T_\pi v = \langle \pi, q\rangle = (\langle \pi_s, q_s\rangle)_{s\in\mathcal{S}}$.

- Bellman optimality operator

$$T_* v = \max_\pi T_\pi v.$$

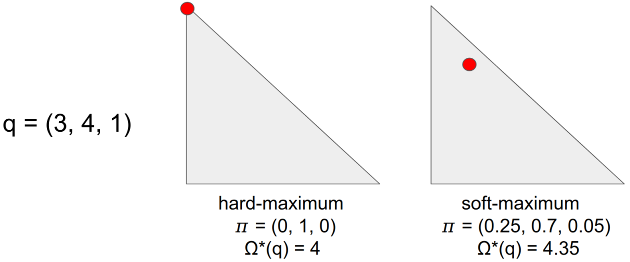
- greedy operator

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

- From T_π, get T_* and $\mathcal{G}$, and then PI, VI, MPI... RL!

Let $\Omega : \Delta_{\mathcal{A}} \to \mathbb{R}$ be a strongly convex function. The convex conjugate is (here) a smoothed maximum

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



- Negative Shannon entropy:

$$\Omega(\pi_s) = \sum_a \pi_s(a) \ln \pi_s(a), \quad \Omega^*(q_s) = \ln \sum_a \exp q_s(a)$$

$$\nabla\Omega^*(q_s) = \frac{\exp q_s(a)}{\sum_b \exp q_s(b)}$$

- Kullback-Leibler divergence

$$\Omega_\mu(\pi_s) = \sum_a \pi_s(a) \ln \frac{\pi_s(a)}{\mu_s(a)}, \quad \Omega_\mu^*(q_s) = \ln \sum_a \mu_s(a) \exp q_s(a)$$

$$\nabla\Omega_\mu^*(q_s) = \frac{\mu_s(a) \exp q_s(a)}{\sum_b \mu_s(b) \exp q_s(b)}$$

- Tsallis entropy

$$\Omega(\pi_s) = \frac{1}{2}(\|\pi_s\|_2^2 - 1).$$

Core idea

- Regularize the Bellman evaluation operator

$$[T_{\pi,\Omega} v](s) = \langle \pi_s, q_s \rangle - \Omega(\pi_s) \\ = [T_\pi v](s) - \Omega(\pi_s).$$

- From this, regularized Bellman optimality operator, regularized greediness, regularized dynamic programming, then regularized RL.

- Evaluation, optimality, greediness:

$$T_{\pi,\Omega} : v \in \mathbb{R}^{\mathcal{S}} \to T_{\pi,\Omega} v = T_\pi v - \Omega(\pi) \in \mathbb{R}^{\mathcal{S}},$$
$$T_{*,\Omega} : v \in \mathbb{R}^{\mathcal{S}} \to T_{*,\Omega} v = \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} T_{\pi,\Omega} v = \Omega^*(q) \in \mathbb{R}^{\mathcal{S}},$$
$$\pi' = \mathcal{G}_\Omega(v) = \nabla \Omega^*(q) \Leftrightarrow T_{\pi',\Omega} v = T_{*,\Omega} v.$$

- The regularized Bellman operators satisfy the same properties as the original ones:
  - $T_{\pi,\Omega}$ is affine.
  - Monotonicity, distributivity and γ-contraction of $T_{\pi,\Omega}$ and $T_{*,\Omega}$.

- Reg. value functions are fixed-points of the reg. operators,

$$q_{\pi,\Omega}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}[v_{\pi,\Omega}(s')]$$
$$\text{with } v_{\pi,\Omega}(s) = \mathbb{E}_{a\sim\pi(.|s)}[q_{\pi,\Omega}(s,a)] - \Omega(\pi(.|s)).$$
$$q_{*,\Omega}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}[v_{*,\Omega}(s')]$$
$$\text{with } v_{*,\Omega}(s) = \Omega^*(q_{*,\Omega}(s,.)).$$

- The (unique) optimal policy is greedy resp. to $v_{*,\Omega}$,

$$v_{\pi_{*,\Omega},\Omega} = v_{*,\Omega} \geq v_{\pi,\Omega} \text{ with } \pi_{*,\Omega} = \mathcal{G}_\Omega(v_{*,\Omega})$$

- However, the MDP's solution is biased by the regularizer. Assuming that $L_\Omega \leq \Omega \leq U_\Omega$,

$$v_* - \frac{U_\Omega - L_\Omega}{1-\gamma} \leq v_{\pi_{*,\Omega}} \leq v_*.$$

$$\begin{cases} \pi_{k+1} = \mathcal{G}_{\Omega}(v_k) \\ v_{k+1} = (T_{\pi_{k+1},\Omega})^m v_k \end{cases}.$$

- With $m = 1$, we get regularized VI, that can be simplified as $v_{k+1} = T_{*,\Omega} v_k$ (as π_{k+1} is greedy resp. to v_k, we have $T_{\pi_{k+1},\Omega} v_k = T_{*,\Omega} v_k$).
- With $m = \infty$, we get regularized PI, that can be simplified as $\pi_{k+1} = \mathcal{G}_\Omega(v_{\pi_k,\Omega})$ (indeed, with a slight abuse of notation, $(T_{\pi_k,\Omega})^\infty v_{k-1} = v_{\pi_k,\Omega}$).

If $m = 1$,

$$J(\theta) = \hat{\mathbb{E}}\left[\left(\hat{q}_i - q_\theta(s_i, a_i)\right)^2\right] \text{ with } \hat{q}_i = r_i + \gamma\Omega^*(q_{\bar{\theta}}(s_i', \cdot)).$$

If $m \geq 1$,

- evaluation step, $m = 1$

$$J(\theta) = \hat{\mathbb{E}}[(\hat{q}_i - q_\theta(s_i, a_i))^2] \text{ with } \hat{q}_i = r_i + \gamma(\mathbb{E}_{a\sim\pi(\cdot|s_i')}[q_{\bar{\theta}}(s_i', a)] - \Omega(\pi(\cdot, s_i')).$$

- evalution step, $m > 1$, either m-step rollouts or solve m regressions (keeping π fixed)
- greedy step

$$J(w) = \hat{\mathbb{E}}\left[\mathbb{E}_{a\sim\pi_w(\cdot|s_i)}[q_k(s_i, a)] - \Omega(\pi_w(\cdot|s_i)\right]$$
$$\text{or } J(w) = \hat{\mathbb{E}}[\mathrm{KL}(\pi_w(\cdot|s_i)||\nabla\Omega^*(q_k(s_i, .)))].$$

Soft Q-learning, SAC, DPP, MPO, TRPO are (variations of) these recipes

- Analyzed algorithmic scheme,

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

- Quantity to bound, the loss $l_{k,\Omega} = v_{*,\Omega} - v_{\pi_k,\Omega}$.
- Γ-matrix, roughly defined as $\Gamma^n = \prod_{i=1}^n (\gamma P_{\pi_i})$.

Theorem

After k iterations of reg-MPI, the loss satisfies

$$l_{k,\Omega} \leq 2 \sum_{i=1}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon_{k-i}| + \sum_{i=0}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon'_{k-i}| + h(k)$$

with $h(k) = 2 \sum_{j=k}^{\infty} \Gamma^j |d_0|$ *or* $h(k) = 2 \sum_{j=k}^{\infty} \Gamma^j |b_0|$.

- Regularizing the MDP changes the problem.
- Possible to solve the original problem with regularization?
- Idea: as DP is iterative, regularize according to the previous policy
- Bregman divergence generated by Ω:

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

  Positive, $\Omega_{\pi'}(\pi') = 0$, strongly convex in π
- Eg, KL div. generated by negative entropy

$$\mathrm{KL}(\pi_s || \pi'_s) = \sum_a \pi_s(a) \ln \frac{\pi_s(a)}{\pi'_s(a)}.$$

- greedy step, $\pi_{k+1} = \text{argmax}_\pi \langle q_k, \pi \rangle - D_\Omega(\pi || \pi_k)$.
- evaluation step, $v_{k+1} = (T_{\pi_{k+1}, \Omega_{\pi_k}})^m v_k$ or $v_{k+1} = (T_{\pi_{k+1}, \Omega_{\pi_{k+1}}})^m v_k$. As $\Omega_{\pi_{k+1}}(\pi_{k+1}) = 0$, this simplifies as $v_{k+1} = (T_{\pi_{k+1}})^m v_k$, that is a partial unregularized evaluation.
- MD-MPI type-1 and type-2

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

- TRPO: MD-MPI type 2, with $m = \infty$ and greedy step

$$J(w) = \hat{\mathbb{E}}\left[\mathbb{E}_{a\sim\pi_w(\cdot|s_i)}[q_k(s_i,a)] - \Omega(\pi_w(\cdot|s_i)\right].$$

- MPO: MD-MPI type-2, with $m = \infty$ and greedy step

$$J(w) = \hat{\mathbb{E}}[\mathrm{KL}(\pi_w(\cdot|s_i)||\nabla\Omega^*(q_k(s_i,.)))].$$

- DPP: reparameterization of MD-MPI type-1, with $m = 1$.
- etc.

Analyzed algorithmic schemes:

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

Theorem

Define $R_{\Omega_{\pi_0}} = \| \sup_\pi D_\Omega(\pi || \pi_0) \|_\infty$, *after* K *iterations of MD-MPI, for* $h = 1, 2$, *the regret* $L_K = \sum_{k=1}^K l_k$ *satisfies*

$$L_K \leq 2 \sum_{k=2}^{K} \sum_{i=1}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon_{k-i}| + \sum_{k=1}^{K} \sum_{i=0}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon'_{k-i}|$$

$$+ \sum_{k=1}^{K} h(k) + \frac{1-\gamma^K}{(1-\gamma)^2} R_{\Omega_{\pi_0}} \mathbf{1}.$$

with $h(k) = 2 \sum_{j=k}^{\infty} \Gamma^j |d_0|$ *or* $h(k) = 2 \sum_{j=k}^{\infty} \Gamma^j |b_0|$.

- Dynamic programming and optimization
- Temporal consistency equations. Eg, with entropy

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A} \quad v_{*,\Omega}(s) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}[v_{*,\Omega}(s')] - \ln \pi_{*,\Omega}(a|s).$$

- Regularized policy gradient. With $J_\Omega(\pi) = \nu v_{\pi,\Omega}$,

$$\nabla J_\Omega(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{s,a\sim d_{\nu,\pi}}\left[\left(q_{\pi,\Omega}(s,a) - \frac{\partial \Omega(\pi(.|s))}{\partial \pi(a|s)}\right)\nabla \ln \pi(a|s)\right].$$

- Regularized IRL. Uniqueness of greediness pretty useful, eg. for entropy $\hat{r}(s,a) = \ln \pi_{*,\Omega}(s,a)$ analytic solution to (regularized) IRL.
- Regularized zero-sum Markov games,

$$[T_{\mu,\nu,\Omega} v](s) = [T_{\mu,\nu} v](s) - \Omega_1(\mu(.|s)) + \Omega_2(\nu(.|s)).$$