On the connection between Bregman divergence and value in regularized Markov decision processes

Brendan O’Donoghue
DeepMind

In this short note we derive a relationship between the Bregman divergence from the current policy to the optimal policy and the suboptimality of the current value function in a regularized Markov decision process. This result has implications for multi-task reinforcement learning, offline reinforcement learning, and regret analysis under function approximation, among others.

We consider a finite state-action discounted MDP given by the tuple $M = (S, A, P, r, \rho, \gamma)$, where $S$ is the set of states and $A$ is the set of possible actions, $P : S \times A \rightarrow \Delta(S)$ denotes the state transition kernel, $r : S \times A \rightarrow \mathbb{R}$ denotes the reward function, $\rho \in \Delta(S)$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. The main result of this manuscript holds more generally, but for brevity we shall restrict ourselves to this case. A policy $\pi \in \Delta(\mathcal{A})$ is a distribution over actions for each state, and we shall denote the probability of action $a$ in state $s$ as $\pi(s, a)$. For any policy $\pi \in \Delta(\mathcal{A})$ and continuously-differentiable strictly-convex regularizer $\Omega : \Delta(\mathcal{A}) \rightarrow \mathbb{R}$ we define value functions for each $(s, a) \in S \times A$ as

$$Q^\pi_\Omega(s, a) = r(s, a) + \gamma \sum \limits_{s'} P(s' \mid s, a) V^\pi_\Omega(s'), \quad V^\pi_\Omega(s) = \max \limits_{\pi(s, \cdot) \in \Delta(\mathcal{A})} \left( \sum \limits_a \pi(s, a) Q^\pi_\Omega(s, a) - \Omega(\pi(s, \cdot)) \right).$$

The optimal value functions for each $(s, a) \in S \times A$ are given by

$$Q^*_\Omega(s, a) = r(s, a) + \gamma \sum \limits_{s'} P(s' \mid s, a) V^*_\Omega(s'), \quad V^*_\Omega(s) = \max \limits_{\pi(s, \cdot) \in \Delta(\mathcal{A})} \left( \sum \limits_a \pi(s, a) Q^*_\Omega(s, a) - \Omega(\pi(s, \cdot)) \right).$$

We denote by $\pi^*_\Omega \in \Delta(\mathcal{A})$ the policy that achieves the maximum for all $s \in S$, which is always attained due to the properties of $\Omega$. It must satisfy the first-order optimality conditions for the maximum, which is the following inclusion

$$Q^*_\Omega(s, \cdot) - \nabla \Omega(\pi^*_\Omega(s, \cdot)) \in N_{\Delta(\mathcal{A})}(\pi^*_\Omega(s, \cdot)), \quad s \in S,$$

where $N_{\Delta(\mathcal{A})}$ is the normal cone of simplex $\Delta(\mathcal{A})$ [4], i.e.,

$$N_{\Delta(\mathcal{A})}(\pi) = \{ y \mid y^\top (\pi' - \pi) \leq 0, \forall \pi' \in \Delta(\mathcal{A}) \}$$

if $\pi \in \Delta(\mathcal{A})$ otherwise $\emptyset$.

Any policy $\pi$ induces a discounted stationary distribution over states denoted $\mu^\pi \in \Delta(S)$ satisfying

$$\mu^\pi(s') = (1 - \gamma) \rho(s') + \gamma \sum \limits_{(s, a)} P(s' \mid s, a) \pi(s, a) \mu^\pi(s), \quad s' \in S.$$

Finally, the Bregman divergence generated by $\Omega$ between two points $\pi, \pi' \in \Delta(\mathcal{A})$ is defined as

$$D_\Omega(\pi, \pi') = \Omega(\pi) - \Omega(\pi') - \nabla \Omega(\pi')^\top (\pi - \pi').$$

To prove our main result we require a slight generalization of the performance difference lemma (PDL) [1] to cover the regularized MDP case.

Lemma 1. For any two policies $\pi, \pi' \in \Delta(\mathcal{A})$

$$(1 - \gamma) \mathbb{E}_{s \sim \rho} \left( V^\pi_\Omega(s) - V^{\pi'}_\Omega(s) \right) = \mathbb{E}_{s \sim \mu^\pi} \left( \sum \limits_a \pi(s, a) Q^\pi_\Omega(s, a) - V^\pi_\Omega(s) - \Omega(\pi(s, \cdot)) \right).$$

Corresponding author(s): bodonoghue@deepmind.com
The proof of this identity is included in the appendix for completeness. For our stronger result we require the following short technical lemma.

**Lemma 2.** If \( \pi^* \in \text{relint}(\Delta(\mathcal{A})) \) then \( y^\top (\pi - \pi^*) = 0 \) for all \( y \in N_{\Delta(\mathcal{A})}(\pi^*) \), \( \pi \in \Delta(\mathcal{A}) \).

**Proof.** Given \( \pi^* \in \text{relint}(\Delta(\mathcal{A})) \) assume that we can find \( \pi \in \Delta(\mathcal{A}) \) such that for some \( y \in N_{\Delta(\mathcal{A})}(\pi^*) \) we have \( y^\top (\pi - \pi^*) < 0 \). Let \( \Delta = \pi - \pi^* \). For sufficiently small \( \epsilon > 0 \) there exists policy \( \pi' = \pi^* + \epsilon \Delta \) that is in \( \Delta(\mathcal{A}) \) since \( \pi^* \) is in the relative interior. By assumption \( y^\top \Delta = y^\top (\pi - \pi^*) < 0 \). However, from the normal cone property \( 0 \leq -y^\top (\pi - \pi^*) = \epsilon y^\top \Delta \), which is a contradiction. \( \Box \)

With these we are ready to present the main result of this note.

**Theorem 1.** For any policy \( \pi \in \Delta(\mathcal{A})|_{S} \)

\[
\mathbb{E}_{s \sim \pi^*} D_{\Omega}(\pi(s, \cdot), \pi^*_\Omega(s, \cdot)) \leq (1 - \gamma) \mathbb{E}_{s \sim \rho} (V^*_\Omega(s) - V^\pi_\Omega(s)),
\]

moreover, if \( \pi^*_\Omega \in \text{relint}(\Delta(\mathcal{A})) \) then

\[
\mathbb{E}_{s \sim \pi^*} D_{\Omega}(\pi(s, \cdot), \pi^*_\Omega(s, \cdot)) = (1 - \gamma) \mathbb{E}_{s \sim \rho} (V^*_\Omega(s) - V^\pi_\Omega(s)).
\]

**Proof.** Let \( y(s) = Q^*(s, \cdot) - \nabla \Omega(\pi^*_\Omega(s, \cdot)) \) for \( s \in S \), and note that \( y(s) \in N_{\Delta(\mathcal{A})}(\pi^*_\Omega(s, \cdot)) \) from (2), then

\[
\mathbb{E}_{s \sim \pi^*} D_{\Omega}(\pi(s, \cdot), \pi^*_\Omega(s, \cdot)) = \mathbb{E}_{s \sim \pi^*} \left( \Omega(\pi(s, \cdot)) - \Omega(\pi^*_\Omega(s, \cdot)) - (Q^*(s, \cdot) - y(s))^\top (\pi(s, \cdot) - \pi^*_\Omega(s, \cdot)) \right)
\leq \mathbb{E}_{s \sim \pi^*} \left( \Omega(\pi(s, \cdot)) - \Omega(\pi^*_\Omega(s, \cdot)) - Q^*(s, \cdot)^\top (\pi(s, \cdot) - \pi^*_\Omega(s, \cdot)) \right)
= \mathbb{E}_{s \sim \pi^*} \left( V^*_\Omega(s) - \sum_a \pi(s, a) Q^*_\Omega(s, a) + \Omega(\pi(s, \cdot)) \right)
= (1 - \gamma) \mathbb{E}_{s \sim \rho} (V^*_\Omega(s) - V^\pi_\Omega(s)),
\]

where the first line replaces \( \nabla \Omega(\pi^*_\Omega(s, \cdot)) \) in the definition of the Bregman divergence (5), the second uses the normal cone property (3), the third line substitutes in the definition of \( V^*_\Omega \) from (1), and the final line uses the regularized PDL. To get the stronger statement, we use Lemma 2 which replaces the inequality in the second line with an equality. \( \Box \)

Since the KL-divergence is the Bregman divergence generated by the negative-entropy function, and since the entropy-regularized optimal policy is always in the relative interior of the simplex for bounded rewards, we have the following corollary.

**Corollary 1.** Let \( \tau > 0 \) be a regularization parameter and denote entropy by \( H \) and KL-divergence by \( KL(\cdot \mid \mid \cdot) \) and set \( \Omega = -\tau H \), then for any policy \( \pi \in \Delta(\mathcal{A})|_{S} \) we have

\[
\mathbb{E}_{s \sim \pi^*} KL(\pi(s, \cdot) \mid \mid \pi^*_\Omega(s, \cdot)) = \frac{(1 - \gamma)}{\tau} \mathbb{E}_{s \sim \rho} (V^*_\Omega(s) - V^\pi_\Omega(s)).
\]

This identity and related ones have previously been used to bound policy performance between an entropy-regularized optimistic policy and Thompson sampling [3] and to prove the linear convergence of entropy-regularized policy gradient methods [2].

**Acknowledgements.** I thank Tor Lattimore for spotting an error in a earlier draft of this note and Ted Moskovitz for useful discussions.
The result follows from (Lemma 3. For any vector \( x \in \mathbb{R}^{[S]} \))

\[
(\mu^\pi)^\top (I - \gamma P^\pi)x = (1 - \gamma) \rho^\top x.
\]

Proof. Using vector notation, we can write (4) as \( \mu^\pi = (1 - \gamma) \rho + \gamma (P^\pi)^\top \mu^\pi = (1 - \gamma) (I - \gamma (P^\pi)^\top)^{-1} \rho. \)

The result follows from \( (\mu^\pi)^\top (I - \gamma P^\pi)x = (1 - \gamma) \rho^\top (I - \gamma P^\pi)^{-1} (I - \gamma P^\pi)x = (1 - \gamma) \rho^\top x. \)

Lemma 1. For any two policies \( \pi, \pi' \in \Delta(\mathcal{A})^{[S]} \)

\[
(1 - \gamma) \mathbb{E}_{s \sim \pi} (V^\pi_\Omega(s) - V^{\pi'}_\Omega(s)) = \mathbb{E}_{s \sim \pi} \left( \sum_a \pi(s, a) Q^\pi_\Omega(s, a) - V^{\pi'}_\Omega(s) - \Omega(\pi(s, \cdot)) \right).
\]

Proof. Define vectors \( r^\pi, \Omega^\pi, q^{(\pi, \pi')} \in \mathbb{R}^{[S]} \) as

\[
r^\pi_s = \sum_a \pi(s, a)r(s, a), \quad \Omega^\pi_s = \Omega(\pi(s, \cdot)), \quad q_{(\pi, \pi')} = \sum_a \pi(s, a)Q^\pi_{\Omega}(s, a),
\]

and note that \( q^{(\pi, \pi')} \) and \( V^\pi_\Omega \) satisfy

\[
q^{(\pi, \pi')} = r^\pi + \gamma P^\pi V^{\pi'}_\Omega, \quad V^\pi_\Omega = r^\pi - \Omega^\pi + \gamma P^\pi V^\pi_\Omega.
\]

Then, using Lemma 3 we have

\[
(1 - \gamma) \rho^\top (V^\pi_\Omega - V^{\pi'}_\Omega) = (\mu^\pi)^\top \left( V^\pi_\Omega - V^{\pi'}_\Omega - \gamma P^\pi (V^\pi_\Omega - V^{\pi'}_\Omega) \right)
\]

\[
= (\mu^\pi)^\top \left( r^\pi - \Omega^\pi - V^{\pi'}_\Omega + \gamma P^\pi V^{\pi'}_\Omega \right)
\]

\[
= (\mu^\pi)^\top \left( q^{(\pi, \pi')} - \Omega^\pi - V^{\pi'}_\Omega \right).
\]