Mean squared advantage minimization as a consequence of entropic policy improvement regularization

Boris Belousov BELOUSOV@IAS.TU-DARMSTADT.DE
Department of Computer Science
Technische Universität Darmstadt
FG IAS, Hochschulstr. 10, 64289 Darmstadt, Germany

Jan Peters[*,†] PETERS@IAS.TU-DARMSTADT.DE
[*] *Department of Computer Science, TU Darmstadt*
[†] *Max Planck Institute for Intelligent Systems*
Max-Planck-Ring 4, 72076 Tübingen, Germany

Abstract

Policy improvement regularization with entropy-like f-divergence penalties provides a unifying perspective on actor-critic algorithms, rendering policy improvement and policy evaluation steps as primal and dual subproblems of the same optimization problem. For small policy improvement steps, we show that all f-divergences with twice differentiable generator function f yield a mean squared advantage minimization objective for the policy evaluation step and an advantage-weighted maximum log-likelihood objective for the policy improvement step. The mean squared advantage objective fits in-between the well-known mean squared Bellman error and the mean squared temporal difference error objectives, requiring only the expectation of the temporal difference error with respect to the next state and not the policy, in contrast to the Bellman error, which requires both, and the temporal difference error, which requires none. The advantage-weighted maximum log-likelihood policy improvement rule emerges as a linear approximation to a more general weighting scheme where weights are a monotone function of the advantage. Thus, the entropic policy regularization framework provides a rigorous justification for the common practice of least squares value function fitting accompanied by advantage-weighted maximum log-likelihood policy parameters estimation, at the same time pointing at the direction in which this classical actor-critic approach can be extended.

Keywords: policy optimization, entropic proximal mappings, actor-critic algorithms

1. Introduction

Recent progress in reinforcement learning on challenging continuous control tasks showed that combining benefits of policy-based and value-based methods in actor-critic architectures yields highly competitive algorithms that achieve state of the art results on a variety of benchmark control problems (Mnih et al., 2016; Wu et al., 2017; Schulman et al., 2017).

These algorithms follow the generalized policy iteration scheme (Sutton and Barto, 1998), consisting of a policy evaluation and a policy improvement steps that have a specific form. Namely, parameters of the value function are found by minimizing the average squared error over a batch of samples from a current policy

$$w = \underset{\tilde{w}}{\text{minimize}}\ \hat{E}_t \left[\| V^{\tilde{w}}(s_t) - \hat{V}_t \|^2 \right]. \tag{1}$$

The sample average over data points (s_t, a_t, s'_t, r_t) is denoted by $\hat{E}_t$ and the target $\hat{V}_t$ is given in the simplest case by the Monte Carlo estimate of the value function $\hat{V}_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k}$.

Parameters of the policy, on the other hand, are updated by maximizing the advantage-weighted log-likelihood

$$\theta = \underset{\tilde{\theta}}{\text{maximize}}\ \hat{E}_t \left[\log \pi_{\tilde{\theta}} \hat{A}^w_t \right] \tag{2}$$

A popular way of computing the advantage estimate $\hat{A}^w_t$ is by performing exponential averaging over Bellman residuals $\hat{A}^w_t = \sum_{k=0}^{\infty} (\gamma\lambda)^k \delta^w_{t+k}$ with a decay factor $\lambda \in [0, 1]$. Here, the Bellman residual δ^w_t, also known as the temporal difference (TD) error, is defined as $\delta^w_t = R_t + \gamma V^w(s_{t+1}) - V^w(s_t)$. Such technique can be interpreted as an application of the TD(λ) algorithm to advantage function estimation (Schulman et al., 2016).

It is important to point out that the advantage estimate $\hat{A}^w_t$ in (2) is assumed to be independent of the policy parameters $\tilde{\theta}$. Ideally, one would recompute $\hat{A}^w_t$ after every gradient descent step in $\tilde{\theta}$; however, this is rather sample-inefficient. More aggressive policy parameter updates are possible if one postulates a trust region within which the advantage estimate $\hat{A}^w_t$ is deemed to not vary much when $\tilde{\theta}$ is varied. Such approach is taken by the trust-region policy optimization (TRPO) algorithm (Schulman et al., 2015) and its derivatives. Another strategy is to add an entropy bonus to (2) and perform just a few gradient descent update steps (Mnih et al., 2016). The curvature information coming from the entropy bonus helps to dampen the change in $\tilde{\theta}$, however the resulting algorithm may diverge (Neu et al., 2017).

Expressions (1) and (2) constitute a pair of optimization problems: the policy evaluation step (1) and the policy improvement step (2). Later we will see that this pair is just one representative from a family of such pairs that arise for different choices of an f-divergence penalty within our entropic proximal policy optimization framework. What is very special about this particular choice is that it is, in a certain sense, a linear-quadratic approximation arising in the limit of small policy update steps for any other choice of the twice differentiable generator function f.

2. Entropic proximal policy optimization

The relative entropy policy search (REPS) framework (Peters et al., 2010) views the pair of policy improvement and policy evaluation steps as a primal-dual pair of a single optimization problem. Such interpretation results from imposing a Kullback-Leibler (KL) constraint on policy improvement. We show that this formulation can be straightforwardly generalized to any f-divergence (Csiszár, 1963) with twice differentiable generator function f. For simplicity of exposition, we employ an f-divergence penalty instead of a hard constraint. Although the mathematical treatment of the constrained case is analogous, in practice, the constrained formulation might be advantageous because it simplifies hyper-parameter tuning (Schulman et al., 2017).

2.1 Variational derivation

Following the intuition of REPS, we introduce an f-divergence penalized optimization problem that the learning agent has to solve at every policy iteration step in the average-reward

reinforcement learning setting (Sutton and Barto, 1998)

$$\begin{aligned}
\underset{\pi}{\text{maximize}} \quad & J_\eta(\pi) = \int_{S\times A} \rho_\pi(s,a) R(s,a) ds da - \eta \int_{S\times A} \rho_{\pi_0}(s,a) f\left(\frac{\rho_\pi(s,a)}{\rho_{\pi_0}(s,a)}\right) ds da \\
\text{subject to} \quad & \int_A \rho_\pi(s',a') da' = \int_{S\times A} \rho_\pi(s,a) p(s'|s,a) ds da, \quad \forall s' \in S, \\
& \int_{S\times A} \rho_\pi(s,a) ds da = 1, \\
& \rho_\pi(s,a) \geq 0, \quad \forall (s,a) \in S \times A.
\end{aligned} \tag{3}$$

Here, $\rho_\pi(s,a)$ is the state-action distribution $\rho_\pi(s,a) = \mu_\pi(s)\pi(a|s)$ induced by policy $\pi(a|s)$, and $\rho_{\pi_0}(s,a)$ is the state-action distribution induced by the current policy $\pi_0(a|s)$. Let us denote the dual variables corresponding to the constraints by $\{V(s), \lambda, \kappa(s,a)\}$, respectively. Then, the policy update can be expressed through the derivative of the convex conjugate function $f_*(y)$ as

$$\rho_\pi(s,a) = \rho_{\pi_0}(s,a) f_*'\left(\frac{R(s,a) + \int_S V(s') p(s'|s,a) ds' - V(s) - \lambda + \kappa(s,a)}{\eta}\right). \tag{4}$$

This solution can be viewed as an application of the general entropic proximal mappings optimization framework (Teboulle, 1992; Neu et al., 2017) to Markov decision processes. The resulting dual optimization problem has a straightforward form

$$\begin{aligned}
\underset{V,\lambda,\kappa}{\text{minimize}} \quad & g(V,\lambda,\kappa) = \eta \int_{S\times A} \rho_{\pi_0}(s,a) f_*\left(\frac{A^V(s,a) - \lambda + \kappa(s,a)}{\eta}\right) ds da + \lambda \\
\text{subject to} \quad & \kappa(s,a) \geq 0, \quad \forall (s,a) \in S \times A, \\
& \arg f_* \in \text{range}_{x\geq 0} f'(x), \quad \forall (s,a) \in S \times A.
\end{aligned} \tag{5}$$

Here, the advantage function $A^V(s,a) = R(s,a) + \int_S V(s') p(s'|s,a) ds' - V(s)$ was introduced for notational convenience. Its dependence on the value function is symbolized by the superscript V.

Policy evaluation/policy improvement pair (5)-(4) in principle provides a way to perform policy iteration to find an optimal policy for a given Markov decision process. However, in a reinforcement learning scenario, neither the system dynamics $p(s'|s,a)$ nor the reward function $R(s,a)$ are known. Moreover, we are mainly interested in continuous state-action spaces, which makes the problem even harder (Bellman, 1957). To address all these issues, we have to resort to function approximation and sample-based estimation.

2.2 Value function and policy approximation, model-free learning

Assume that the value function $V^w(s)$ is parameterized by a vector w, and the policy $\pi_\theta(a|s)$ is parameterized by a vector θ. Supposing that a batch of samples was collected under the current policy, the integral in the dual objective (5) can be replaced by a sample average

$$\hat{g}(w,\lambda,\kappa) = \eta \hat{E}_t \left[f_*\left(\frac{\hat{A}^w(s_t,a_t) - \lambda + \kappa(s_t,a_t)}{\eta}\right)\right] + \lambda. \tag{6}$$

Constrained maximization of the objective function (6) constitutes the policy evaluation step, which results in an optimal value for the dual parameters w. The corresponding policy improvement step (4) again cannot be computed in closed form because we do not have the model $\{p(s'|s,a), R(s,a)\}$; therefore, one commonly employs the weighted maximum likelihood estimation approach to fit policy parameters (Deisenroth et al., 2013), which in our f-divergence penalized setting gives

$$\hat{L}(\theta) = \hat{E}_t \left[\log \pi_\theta(a_t|s_t) f'_* \left(\frac{\hat{A}^w(s_t,a_t) - \lambda + \kappa(s_t,a_t)}{\eta} \right) \right]. \tag{7}$$

The pair of optimization objectives (6)-(7) is strikingly similar to the commonly used pair (1)-(2), and they are actually equivalent when f is a quadratic function $f(x) = \frac{1}{2}(x-1)^2$, corresponding to the Pearson χ^2-divergence (Cichocki and Amari, 2010). Furthermore, it can be shown by Taylor approximation that in the limit $\eta \gg 1$, corresponding to small policy update steps, the pair (6)-(7) tends towards (1)-(2) for any choice of twice differentiable divergence generating function f.

Observer that no extra trust region constraint needs to be added to the optimization problem (7) to ensure closeness of the new policy to the old one. In contrast, most state-of-the-art algorithms heuristically add the trust region constraint exactly at this stage. The reason why no additional constraint is required in our formulation is that the closeness to the previous policy is ensured by the f-divergence term in the primal optimization objective (3). The closed-form solution (4) gives the exact expression for the new state-action distribution expressed as a function of the dual variables. The dual optimization problem (5) in its turn provides the optimal values of the dual variables that guarantee that ρ_π remains sufficiently close to ρ_{π_0} according to (4).

3. Mean squared advantage objective

Instantiating Equations (6)-(7) with the conjugate of the Pearson χ^2-divergence generator function $f_*(y) = \frac{1}{2}(y+1)^2 - \frac{1}{2}$, which corresponds to the high-temperature $\eta \gg 1$ limit, or to small policy update steps, one obtains the following critic-actor pair

$$\hat{g}(w) \propto \hat{E}_t \left[\left(\hat{A}^w(s_t,a_t) \right)^2 \right] \tag{8}$$

$$\hat{L}_2(\theta) \propto \hat{E}_t \left[\log \pi_\theta(a_t|s_t) \hat{A}^w(s_t,a_t) \right]. \tag{9}$$

These objectives are written for the original discounted infinite horizon problem formulation (1)-(2), which differs from the average reward setting (3) only in the absence of the average return baseline λ and the presence of the discount factor γ in the definition of the advantage $A^w(s,a) = R(s,a) + \gamma \int_S V^w(s') p(s'|s,a) ds' - V^w(s)$ (Sutton and Barto, 1998).

3.1 Relation to common linear-quadratic actor-critics

The policy improvement objective (9) emerging in the limit of small policy update steps is exactly the same as the customary objective (2) used in the actor-critic algorithms mentioned in this paper, such as TRPO and related. However, note that TRPO adds a trust-region constraint to (9) because it fits the advantage parameters w in an independent,

unrelated to policy optimization way. In contrast, since we fit the advantage parameters w by optimizing the dual problem (5), no additional trust region constraint is required.

The policy evaluation objective (8), on the other hand, given by the mean squared advantage (MSA) minimization, is not exactly the same as the mean squared error minimization (1). The discrepancy stems from the fact that (1) introduces a fixed target $\hat{V}_t$, although in theory it should also depend on the value function parameters w (Sutton and Barto, 1998). Thus, if one either uses the same target $\hat{V}_t$ in both (8) and (1) or if the target is assumed to vary with w, then the two objectives are equivalent. For algorithm stability, it is advisable to use a fixed target and employ TD(λ)-like bootstrapping (Schulman et al., 2016).

3.2 Relation between different error objectives

It is instructive to systematically write down the objective functions for the mean squared temporal difference error (MSTDE), the mean squared advantage (MSA), and the mean squared Bellman error (MSBE) to compare them against each other. Let

$$\delta^w(s,a,s') = R(s,a) + \gamma V^w(s') - V^w(s) \tag{10}$$

denote the TD error parameterized by w through the value function. The advantage is defined as the expectation of the TD error with respect to the next state

$$A^w(s,a) = R(s,a) + \gamma E_{s'\sim p(s'|s,a)}[V^w(s')] - V^w(s). \tag{11}$$

The Bellman error goes one level deeper and requires the expectation of the advantage with respect to the action

$$\epsilon^w(s) = E_{a\sim\pi(a|s)}\left[R(s,a) + \gamma E_{s'\sim p(s'|s,a)}[V^w(s')]\right] - V^w(s). \tag{12}$$

From each of these errors, one can construct a mean squared (MS) objective:

$$\mathrm{MSTDE}(w) = E_{s\sim\mu(s),a\sim\pi(a|s),s'\sim p(s'|s,a)}\left[(\delta^w(s,a,s'))^2\right], \tag{13}$$

$$\mathrm{MSA}(w) = E_{s\sim\mu(s),a\sim\pi(a|s)}\left[\left(E_{s'\sim p(s'|s,a)}[\delta^w(s,a,s')]\right)^2\right], \tag{14}$$

$$\mathrm{MSBE}(w) = E_{s\sim\mu(s)}\left[\left(E_{a\sim\pi(a|s),s'\sim p(s'|s,a)}[\delta^w(s,a,s')]\right)^2\right]. \tag{15}$$

Thus, the MSA objective that results from our derivation fits precisely in-between MSTDE and MSBE. Stochastic gradient descent (SGD) on these objectives leads to various flavors of the residual gradient algorithm (Baird, 1995; Dann et al., 2014). Curiously, all three objectives result in the naive residual gradient algorithm (Sutton and Barto, 1998) when expectations are 'naively' replaced by one-sample estimates.

Interestingly, since the dual objective (5) tends towards the MSA objective in the limit of high temperatures $\eta \to \infty$ independent of the divergence function f, dual minimization (5) can be viewed as a continuous generalization of the MSA minimization (8) to finite temperatures $\eta > 0$; therefore, SGD on the dual objective (5) can be seen as a continuous generalization of the residual gradient algorithm to finite temperatures.

4. Conclusion

Regularized policy improvement framework yields actor-critic update pairs as pairs of primal-dual optimization problems. Whereas the Kullback-Leibler divergence is normally employed for regularization, we showed that one can actually use any f-divergence penalty. However, for small policy update steps, which is the usual and desired regime in reinforcement learning, all f-divergences with twice differentiable generator function f are equivalent and act as the Pearson χ^2-divergence, which, in turn, is equivalent to the natural policy gradient (Kakade, 2001). Thus, a promising direction to explore in terms of finding novel actor-critic pairs that even for small update steps behave differently is to use continuous but possibly not everywhere differentiable functions f, such as the absolute value $f(x) = \frac{1}{2}|x-1|$, corresponding to the total variation distance. Another quite straightforward extension is to incorporate Bregman divergences, which also generalize the KL but in a different direction, and which can be treated similarly to f-divergences (Teboulle, 1992), having in addition rather attractive analytical properties.

It is important to point out that in our entropic policy optimization framework, the trust region constraint between the current and the new policy appears in the original complete optimization problem (3). This has the advantage that value function estimation becomes linked to the way in which the policy parameters are update. In contrast, other trust region policy optimization algorithms fit the V-function without any relation to the policy update step, being forced to subsequently introduce a trust region constraint heruristically at the policy update step in order to avoid large policy changes.

The objective for policy evaluation resulting from the Pearson χ^2-divergence penalty, called the mean squared advantage (MSA) minimization, was shown to be closely related to the well-known MSTDE and MSBE objectives. In deterministic environments, MSA is equivalent to MSTDE, so all intuitions about MSTDE carry over to MSA when there is no uncertainty in system dynamics. However, in stochastic environments all three objectives are different. Clarifying how exactly MSA is different from MSTDE and MSBE and characterizing the range of approximations available for minimizing the MSA from sampled data is a subject of ongoing investigation.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 640554.

References

Leemon Baird. Residual Algorithms: Reinforcement Learning with Function Approximation. *Proceedings of the 12th International Conference on Machine Learning*, (July): 30–37, 1995. ISSN 00043702. doi: 10.1.1.48.3256.

Richard Bellman. *Dynamic Programming*, volume 70. 1957. ISBN 978-0-691-07951-6. doi: 10.1108/eb059970.

Andrzej Cichocki and Shun'ichi Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of Similarities. *Entropy*, 12(6):1532–1568, 2010. ISSN 10994300. doi: 10.3390/e12061532.

Imre Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.*, 8:85–108, 1963.

Christoph Dann, Gerhard Neumann, and Jan Peters. Policy Evaluation with Temporal Differences: A Survey and Comparison. *Journal of Machine Learning Research*, 15:809–883, 2014. ISSN 15337928.

Marc Peter Deisenroth, Gerhard Neumann, and Jan Peters. A survey on policy search for robotics. *Foundations and Trends®in Robotics*, 2(1–2):1–142, 2013.

Sham Machandranath Kakade. A Natural Policy Gradient. In *NIPS*, pages 1531–1538, 2001. ISBN 9780874216561. doi: 10.1.1.19.8165.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pages 1928–1937, 2016.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv:1705.07798*, 2017.

Jan Peters, Katharina Mülling, and Yasemin Altun. Relative Entropy Policy Search. In *AAAI*, pages 1607–1612, 2010.

John Schulman, Sergey Levine, Michael Jordan, and Pieter Abbeel. Trust Region Policy Optimization. In *ICML*, 2015. ISBN 0375-9687. doi: 10.1063/1.4927398.

John Schulman, Philipp Moritz, Sergey Levine, Michael I Jordan, and Pieter Abbeel. High Dimensional Continuous Control Using Generalized Advantage Estimation. In *ICLR*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

Marc Teboulle. Entropic Proximal Mappings with Applications to Nonlinear Programming. *Mathematics of Operations Research*, 17(3):670–690, 1992. ISSN 0364-765X. doi: 10.1287/moor.17.3.670.

Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. In *NIPS*, pages 5279–5288. Curran Associates, Inc., 2017.