Policy Gradient Methods

Jan Peters, Max Planck Institute for Biological Cybernetics
J. Andrew Bagnell, Carnegie Mellon University

Definition

A policy gradient method is a **reinforcement learning** approach that directly optimizes a parametrized control policy by gradient descent. It belongs to the class of **policy search** techniques that maximize the expected return of a policy in a fixed policy class while traditional **value function approximation** approaches derive policies from a value function. Policy gradient approaches have various advantages: they allow the straightforward incorporation of domain knowledge in the policy parametrization and often significantly fewer parameters are needed for representing the optimal policy than the corresponding value function. They are guaranteed to converge to at least a locally optimal policy and can handle continuous states and action, and often even imperfect state information. Their major drawbags are the difficult use in off-policy settings, their slow convergence in discrete problems and that global optima are not attained.

Structure of the Learning System

Policy gradient methods are centered around a parametrized policy π_θ with parameters θ that allows the selection of actions a given the state s, also known as a **direct controller.** Such a policy may either be deterministic $a = \pi_\theta(s)$ or stochastic $a \sim \pi_\theta(a|s)$. This choice also affects the policy gradient approach (e.g., a deterministic policy requires a model-based formulation when used for likelihood ratio policy gradients), chooses how the exploration-exploitation dilemma is addressed (e.g., a stochastic policy tries new actions while a deterministic policy requires the perturbation of policy parameters or sufficient stochasticity in the system), and may affect the optimal solution (e.g., for a time-invariant or stationary policy, the optimal policy can be stochastic [7]). Frequently used policy are Gibbs policies $\pi_\theta(a|s) = \exp(\phi(s,a)^T\theta)/\sum_b \exp(\phi(s,b)^T\theta)$ for discrete problems [7, 1] and, for continuous problems, Gaussian policies $\pi_\theta(a|s) = \mathcal{N}(\phi(s,a)^T\theta_1, \theta_2)$ with an exploration parameter θ_2, see [8, 5].

Expected Return

The goal of policy gradient methods is to optimize the expected return of a policy π_θ with respect to the expected return

$$J(\theta) \quad = \quad Z_\gamma E\left\{\sum_{k=0}^{H} \gamma^k r_k\right\},$$

where $\gamma \in [0, 1]$ denotes a discount factor, a normalization Z_γ and H the planning horizon. For finite H, we have an episodic reinforcement learning scenario where the truly optimal policy is non-stationary and the normalization does not matter. For an infinite horizon $H = \infty$, we choose the normalization to be $Z_\gamma \equiv (1 - \gamma)$ for $\gamma < 1$ and $Z_1 \equiv \lim_{\gamma \to 1}(1 - \gamma) = 1/H$ for **average reward reinforcement learning** problem where $\gamma = 1$.

Gradient Descent in Policy Space

Policy gradient methods follow the gradient of the expected return

$$\theta_{k+1} = \theta_k + \alpha_k \left. \nabla_\theta J(\pi_\theta) \right|_{\theta=\theta_k},$$

where θ_k denotes the parameters after update k with initial policy θ_0 and α_k denotes a learning rate. If the gradient estimator is unbiased, $\sum_{k=0}^{\infty} \alpha_k \to \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 = \text{const}$, the convergence to a local minimum can be guaranteed. In optimal control, model-based gradient methods have been used for optimizing policies since the late 1960s. While these are used machine learning community (e.g., differential dynamic programming with learned models), they are numerically very brittle and rely on accurate, deterministic models. Hence, they may suffer significantly from optimization biases and are not generally applicable, especially not in a model-free case. Several model-free alternatives can be found in the simulation optimization literature [2], i.e., finite-difference gradients, likelihood ratio approaches, response-surface methods, and mean-valued, "weak" derivatives. The advantages and disadvantages of these different approaches are still a fiercely debated topic [2]. In machine learning, the first two approaches have been dominating the field.

Finite Difference Gradients

The simplest policy gradient approaches with the most practical applications (see [5] for a list of robotics application of this method) estimate the gradient by perturbing the policy parameters. For a current policy is θ_k with expected return $J(\theta_k)$, this approach will create explorative policies $\hat{\theta}_i = \theta_k + \delta\theta_i$ with the approximated expected returns given by $J(\hat{\theta}_i) \approx J(\theta_k) + \delta\theta_i^T g$ where $g = \left. \nabla_\theta J(\pi_\theta) \right|_{\theta=\theta_k}$. In this case, it fully sufficies to determine the gradient by linear regression, i.e., we obtain

$$g = (\Delta\Theta^T \Delta\Theta)^{-1} \Delta\Theta^T \Delta J,$$

with parameter perturbations $\Delta\Theta = [\delta\theta_1, \ldots, \delta\theta_n]$ and the mean-subtracted rollout returns $\delta J_n = J(\hat{\theta}_i) - \overline{J(\theta_k)}$ form $\Delta J = [\delta J_1, \ldots, \delta J_n]$. The choice of the parameter perturbation determines the performance of the approach [6]. Problems of this approach the sensitivity of the system with respect to each parameter differs by orders of magnitude, that a small changes in a single parameter may render a system unstable and that it cannot cope well with stochasticity unless used in simulation with a deterministically re-used random numbers [3, 6].

Likelihood-Ratio Gradients

Likelihood ratio gradients rely upon the stochasticity of either the policy for model-free approaches, or the system in the model-based case, and, hence, they may cope better with noise and the sensitivity problems. Assume that you have a path distribution $p_\theta(\tau)$ and rewards $R(\tau) = Z_\gamma \sum_{k=0}^{H} \gamma^k r_k$ along a path τ. Thus, you can write the gradient of the expected return as

$$\nabla_\theta J(\theta) \quad = \quad \nabla_\theta \int p_\theta(\tau) R(\tau) d\tau = \int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) R(\tau) d\tau = E\{\nabla_\theta \log p_\theta(\tau) R(\tau)\}.$$

If our system $p(s'|s,a)$ is Markovian, we can use $p_\theta(\tau) = p(s_0) \prod_{h=0}^{H} p(s_{k+1}|s_k, a_k) \pi_\theta(a_k|s_k)$ for a stochastic policy $a \sim \pi_\theta(a|s)$ to obtain the model-free policy gradient estimator known as Episodic REINFORCE [8]

$$\nabla_\theta J(\theta) = Z_\gamma E \left\{ \sum_{h=0}^{H} \gamma^k \nabla_\theta \log \pi_\theta(a_k|s_k) \sum_{k=h}^{H} \gamma^{k-h} r_k \right\},$$

and for the deterministic policy $a = \pi_\theta(s)$, the model-based policy gradient

$$\nabla_\theta J(\theta) \quad = \quad Z_\gamma E \left\{ \sum_{h=0}^{H} \gamma^k \left(\nabla_a \log p(s_{k+1}|s_k, a_k)^T \nabla_\theta \pi_\theta(s) \right) \sum_{k=h}^{H} \gamma^{k-h} r_k \right\},$$

follows from $p_\theta(\tau) = p(s_0) \prod_{h=0}^{H} p(s_{k+1}|s_k, \pi_\theta(s_k))$. Note that all rewards preceeding an action may be omitted as the cancel out in expectation. Using a state-action value function $Q^{\pi_\theta}(s,a,h) = E\left\{ \sum_{k=h}^{H} \gamma^{k-h} r_k \middle| s, a, \pi_\theta \right\}$ (see **value function approximation**), we can rewrite REINFORCE in its modern form

$$\nabla_\theta J(\theta) = Z_\gamma E \left\{ \sum_{h=0}^{H} \gamma^k \nabla_\theta \log \pi_\theta(a_k|s_k) \left(Q^{\pi_\theta}(s,a,h) - b(s,h) \right) \right\},$$

known as the policy gradient theorem where the baseline $b(s,h)$ is an arbitrary function that may be used to reduce the variance.

While likelihood ratio gradients have been known since the late 1980s, they have recently experienced an upsurge of interest due to progress towards a reduction variance using optimal baselines [4] a compatible function approximation [7], policy gradients in reproducing kernel Hilbert space [1] as well as faster, more robust convergence using natural policy gradients, see [1, 5] for these developments.

See Also

reinforcement learning, **policy search**, **value function approximation**

References and Recommended Reading

[1] James Andrew Bagnell. *Learning Decisions: Robustness, Uncertainty, and Approximation.* Doctoral dissertation, Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, August 2004.

[2] Michael C. Fu. *Handbook on Operations Research and Management Science: Simulation*, chapter Stochastic Gradient Estimation, pages 575–616. Number 19. Elsevier, 2006.

[3] Peter Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, October 1990.

[4] Greg Lawrence, Noah Cowan, and Stuart Russell. Efficient gradient estimation for motor control learning. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, Acapulco, Mexico, 2003.

[5] J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–97, 2008.

[6] J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control.* Wiley, Hoboken, NJ, 2003.

[7] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, and K.-R. Mueller, editors, *Advances in Neural Information Processing Systems (NIPS)*, Denver, CO, 2000. MIT Press.

[8] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.