A Few Observations About Policy Gradient Approximation

Mathias Winther Madsen
mathias@micropsi-industries.com

May 13, 2016

This note explains a couple of ideas used in gradient ascent methods for policy search in robotics. I have already discussed much of its content with you, but I think it will be useful to have it on paper, so we can go through it more slowly and make sure we know what's going on. I intend to produce another note on the same topic once I get further into the literature.

1 Problem and Solution Framework

Suppose we control a parameter vector θ which indexes a family of probabilities densities p_θ on a sample space Ω. A reward function R is defined on Ω, and θ thus determines the distribution of this random reward.

We would like choose θ so as to maximize the expected reward,

$$E_\theta[R] \;=\; \int_\Omega R\, dp_\theta \;=\; \int_\Omega R(\omega)\, p_\theta(\omega)\, d\omega.$$

(Throughout this note, "$E_\theta\,[X]$" means "$E\,[X \mid \theta]$," that is, "expectation under the probability density p_θ," always assuming that θ is held fixed.)

In order to solve this maximization problem, we select an initial value of θ and climb upwards along the gradient

$$\nabla E_\theta[R] \;=\; \int_\Omega R(\omega)\, \nabla p_\theta(\omega)\, d\omega$$

until we reach a local maximum.

Note that $R(\omega)$ is a real number which whose unit might be cookies, dollars, utils, or some other currency of reward. The gradient $\nabla p_\theta(\omega)$, on the other hand, is a vector of dimensionless scalars, "change in probability per change in parameter k". The product $R(\omega)\, \nabla p_\theta(\omega)$ is consequently a vector, and its entries are measured in cookies ("additional cookies per unit change in parameter k").

2 Gradient Evaluation Using Relative Growth

The idea of using gradient ascent to find a locally optimal parameter vector θ raises the practical issue of how to compute the gradient $\nabla E_\theta[R]$. In order to evaluate this function, we exploit the fact that

$$(\ln f)' \;=\; \frac{f'}{f},$$

which implies that

$$f' \;=\; (\ln f)' f$$

for any differentiable function f. This allows us to express the gradient of the expected reward as

$$\int_\Omega R\,\nabla p_\theta \;=\; \int_\Omega R\,(\nabla \ln p_\theta)\,p_\theta \;=\; E_\theta[R\,(\nabla \ln p_\theta)]\,.$$

This formulation has several advantages:

- Since the gradient is now formulated as an expected value, we can approximate it by an empirical average. The weighted integral over ω can be replaced by an unweighted average over a sample $\omega_1, \omega_2, \ldots, \omega_N \;\sim\; p_\theta$, effectively leaving it to the empirical world to do our weighting.

- Typically, logarithmic probability densities like $\ln p_\theta$ are long sums, while plain densities like p_θ are long products. This means that empirical and numerical approximations will tend to be less vulnerable to stochastic noise when we use logarithmic densities.

- Lastly, when the distribution p_θ can factored into an internal action distribution and an external world distribution,

  $$\ln p_\theta(\omega) \;=\; \ln a(\omega \mid \theta) \;+\; \ln w(\omega),$$

  then $\ln w(\omega)$ vanishes under differentiation with respect to the control variable, precisely because the world is not under our control. This effectively means that we do not need to construct any explicit world model, instead probing it only through sampling.

For these reasons, we will approximate the gradient of the expected reward using the estimation scheme

$$\nabla E_\theta[R] \;\approx\; \hat{E}_\theta[R\,(\nabla \ln p_\theta)]\,,$$

where $\hat{E}_\theta$ is an empirical average, as in

$$\hat{E}_\theta[RS] \;=\; \frac{1}{N}\,(R(\omega_1)\,\nabla \ln p_\theta(\omega_1) \;+\; \cdots \;+\; R(\omega_N)\,\nabla \ln p_\theta(\omega_N))\,.$$

Note that this average may consist of as little as one term. In that case, we are simply using the random vector $R\,\nabla \ln p_\theta$ as our gradient estimator. (Larger averages of course have lower variances and therefore yield superior estimates.)

3 Fisher Score and Information

The gradient of the logarithm of a probability density function with respect to a vector of parameters is called the **score**. The score is thus a random vector

$$S(\omega) \;=\; \nabla \ln p_\theta(\omega) \;=\; \frac{\nabla p_\theta(\omega)}{p_\theta(\omega)},$$

where the ∇ designates the gradient with respect to the parameter vector θ.

The score is a vector of unitless scalars. Its kth entry expresses a **relative rate of change**, the fraction by which $p_\theta(\omega)$ would change if we were to change the kth entry in the parameter vector θ.

As an example, consider a Gaussian random variable with mean μ and precision τ. This distribution has a logarithmic density of

$$p_{\mu,\tau}(x) \;=\; \log\sqrt{\frac{\tau}{2\pi}} \;-\; \frac{\tau}{2}(\mu \;-\; x)^2.$$

Taking the gradient of this function with respect to the vector $\theta \;=\; (\mu, \tau)^T$, we find that the score of an observation x with respect to these two parameters is the vector

$$S(x) \;=\; \nabla p_{\mu,\tau}(x) \;=\; \begin{pmatrix} -\tau(\mu \;-\; x) \\ (1/2)(1/\tau) \;-\; (1/2)(\mu \;-\; x)^2 \end{pmatrix}.$$

Both of these entries have an interpretation in terms of rate of change. Suppose for instance that we make the “too large” observation of $x \;=\; \mu \;+\; 5/\tau$. Then the first element of the vector $S(x)$ is the number 5. This suggests that the value $x \;=\; \mu \;+\; 5/\tau$ would be five times more likely if we were to increase μ by an increment of 1 (judging by the local properties of the density function).

The expected value of this Gaussian score is

$$E[S(X)] \;=\; \begin{pmatrix} -\tau(\mu \;-\; \mu) \\ (1/2)(1/\tau) \;-\; (1/2)(1/\tau) \end{pmatrix} \;=\; \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This is not a coincidence. In fact, whenever the score function is well-behaved enough for integration with respect to x to commute with differentiation with respect to θ, we get this this property:

$$\begin{aligned} E_\theta[S] \;&=\; \int_\Omega (\nabla \ln p_\theta)\, p_\theta \;=\; \int_\Omega \nabla p_\theta \\ &=\; \nabla \int_\Omega p_\theta \;=\; \nabla 1 \;=\; \mathbf{0}. \end{aligned}$$

This change of order is permissible for most common density functions (in particular, any probability density whose logarithm can be continuously differentiated with respect to the parameters in θ, and whose integral converges uniformly).

For such a well-behaved density function, the mean score is thus the zero vector. Its variance is the expectation of its squared length:

$$VAR_\theta[S] \;=\; E_\theta\left[\|S\|^2\right] \;-\; \|E_\theta[S]\|^2 \;=\; E_\theta\left[\|S\|^2\right].$$

This statistic, the variance of the score, is called the **Fisher information** (about the parameter, based on the observation). In the Gaussian example above, the Fisher information is the number

$$E_\theta\left[\|S\|^2\right] \;=\; E_\theta\left[\tau^2(\mu \,-\, x)^2 \,+\, \left(\frac{1}{2\tau} \,-\, \frac{(\mu \,-\, x)^2}{2}\right)^2\right].$$

The Fisher information is an interesting statistic because $1/E_\theta[\|S\|^2]$ is a lower bound on the variance of any unbiased estimate of θ (i.e., any random variable $\hat{\theta}$ with $E_\theta[\hat{\theta}] \;=\; \theta$). By the bias-variance decomposition,

$$E_\theta\left[(\theta \,-\, \hat{\theta})^2\right] \;=\; (\theta \,-\, E_\theta[\hat{\theta}])^2 \,+\, E_\theta\left[(\hat{\theta} \,-\, E_\theta[\hat{\theta}])^2\right],$$

so such a lower bound on $VAR_\theta[\hat{\theta}]$ is, for an unbiased estimator $\hat{\theta}$, also a lower bound on its mean squared error. In contexts where θ can be chosen so as to make the Fisher information very small, the problem of unbiased estimation is thus very difficult.

4 Consequences for Gradient Ascent

Now that we have introduced the concept of the score vector $S \;=\; \nabla \ln p_\theta$, we can reformulate our gradient estimation scheme as

$$\nabla E_\theta[R] \;\approx\; \hat{E}_\theta[RS].$$

If we take the empirical average $\hat{E}_\theta[RS]$ to have only one term, then this amounts to approximating the reward gradient by the vector

$$\nabla E_\theta[R] \;\approx\; R(\omega)\, S(\omega),$$

where ω is some observed sample point. This is equivalent to saying that we approximate the kth element of $\nabla E_\theta[R]$ by

$$\frac{\partial}{\partial \theta_k} E_\theta[R] \;\approx\; R(\omega)\, S_k(\omega) \;=\; R(\omega)\, \frac{\partial}{\partial \theta_k} \ln p_\theta(\omega).$$

Since the random vector RS is an unbiased estimate of the gradient, it follows that each of its coordinates are unbiased estimates of the corresponding partial derivatives. We will now modify these coordinate estimators slightly to improve their statistical properties.

Suppose R^* is some fixed constant (perhaps a preferred or desired number of cookies). Consider then the "overshoot estimator"

$$\frac{\partial}{\partial \theta_k} E_\theta[R] \;\approx\; (R(\omega) \;-\; R^*)\, S_k(\omega),$$

This estimator has the same mean as the absolute reward estimator $R(\omega)\, S_k(\omega)$. This follows from the fact that the mean of the score is the zero vector, and thus that $E_\theta[S_k] \;=\; 0$ for all k:

$$E_\theta[RS_k \;-\; R^* S_k] \;=\; E_\theta[RS_k] \;-\; R^* E_\theta[S_k] \;=\; E_\theta[RS_k].$$

Since RS_k is an unbiased estimator of the kth component of the reward gradient, the same is true for $(R \;-\; R^*)S_k$.

The mean squared error of an unbiased estimator is its variance. We would thus like to choose R^* so as to minimize the variance of $(R \;-\; R^*)S_k$, that is,

$$\begin{aligned} VAR_\theta[(R \;-\; R^*)S_k] \;&=\; E_\theta\left[(R \;-\; R^*)^2 S_k^2\right] \;-\; E_\theta\left[(R \;-\; R^*)S_k\right]^2 \\ &=\; E_\theta\left[(R \;-\; R^*)^2 S_k^2\right] \;-\; \left(\frac{\partial}{\partial \theta_k} E_\theta[R]\right)^2. \end{aligned}$$

The second of these terms is independent of R^* (it is, in fact, equal to the square of the constant we're trying to estimate). We can therefore minimize the whole expression by minimizing the first term.

In order to compute this minimum, we take the derivative of this term with respect to R^*:

$$\frac{\partial}{\partial R^*} E_\theta\left[(R \;-\; R^*)^2 S_k^2\right] \;=\; E_\theta\left[-2(R \;-\; R^*) S_k^2\right] \;=\; 2R^* E_\theta\left[S_k^2\right] \;-\; 2E_\theta\left[RS_k^2\right].$$

By setting this derivative equal to 0, we find that the variance of the $(R - R^*)S_k$ is minimal when

$$R^* \;=\; \frac{E_\theta\left[S_k^2 R\right]}{E_\theta[S_k^2]}.$$

This involves two expectations, but we can approximate those by empirical averages. We thus get the following empirical approximation of the optimal estimator of the kth gradient component:

$$\frac{\partial}{\partial \theta_k} E_\theta[R] \;\approx\; \hat{E}_\theta\left[\left(R \;-\; \frac{\hat{E}_\theta\left[S_k^2 R\right]}{\hat{E}_\theta[S_k^2]}\right) S_k\right].$$

This minimizes the mean squared error of the kth gradient component estimate. However, since we are free to pick a value of R^* independently for each component of $\nabla E_\theta[R]$, these approximately optimal coordinates can be stacked up into an approximately optimal vector. The optimal gradient estimator in this class thus consists of coordinates computed according to the right-hand side above, with a different value of R^* in each dimension.