Policy Exploration in a Cold Universe

Mathias Winther Madsen
mathias@micropsi-industries.com

May 25, 2016

This note defines a game which is particularly hard to learn by means of local optimization methods such as gradient ascent. The point of the note is to focus our intuitions about exploration and exploitation by means of a particularly transparent example.

1 Policies and Rewards

Consider a game in which the agent can turn T light switches on or off, resulting in a binary vector $a \in \{0, 1\}^T$.

We will assume that

- the agent uses an i.d.d. strategy, that is, selects the T switch settings by means of T independent coin flips with a (shared) bias of θ;
- the world rewards the agent by +1 point for every light switch that is on, but pays a special bounty of $R(a) = +2T$ for the zero vector $a = (0, 0, 0, \dots, 0)$.

For a light-flicking policy with parameter θ, the agent's expected reward is

$$E_\theta[R] \;=\; T\theta + 2T(1-\theta)^T.$$

The gradient of this reward (whicnh is here just a derivative) is

$$\frac{\partial}{\partial\theta} E_\theta[R] \;=\; T - 2T^2(1-\theta)^{T-1}.$$

These two functions are sketched in Figure 1.

This derivative is negative below a certain tipping point

$$\theta^* \;=\; 1 - \frac{(1/T)^{1/(T-1)}}{2^{1/(T-1)}}$$

which is decreasing in T. A gradient ascent agent initialized with $\theta_0 \in [0, \theta^*)$ will therefore be attracted to the globally optimal solution of $\theta = 0$. An agent initialized at $\theta_0 \in (\theta^*, 1]$ will conversely be attracted to the globally worse maximum $\theta = 1$.

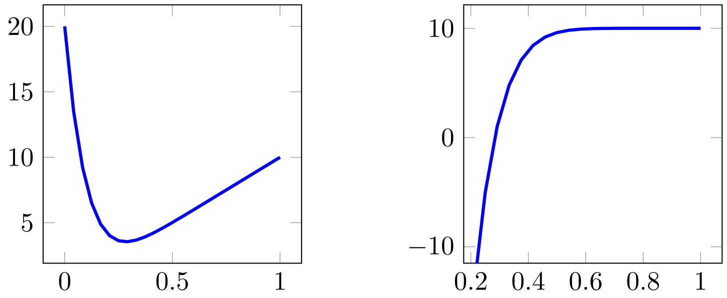

Figure 1: The expected reward and its derivative with respect to the policy parameter θ. Both plots assume that $T = 10$.

2 Gradient Estimation

Let's try to get a better sense of how we would estimate the reward gradient of this game if we couldn't compute it explicitly. For this purpose it will be convenient to define the binomial random variable

$$K = K(a) = \sum_{t=1}^{T} a_t,$$

which counts the number of 1s in the random vector a.

We can then express the logarithmic probability of a binary vector a as

$$\ln p_\theta = K \ln \theta + (T - K) \ln(1 - \theta).$$

The derivative of this logarithmic density is the Fisher score,

$$S = \frac{\partial}{\partial \theta} \ln p_\theta = \frac{K}{\theta} - \frac{T - K}{1 - \theta} = \frac{K}{\theta(1 - \theta)} - \frac{T}{1 - \theta}.$$

According to the log-likelihood trick, the random variable

$$SR = \frac{KR}{\theta(1 - \theta)} - \frac{TR}{1 - \theta}$$

is thus an unbiased estimate of the derivative of the expected reward. We will now verify this by means of a direct computation.

3 The Mean Estimate

We split the expectation $E_\theta[SR]$ up into two conditional expectations according to the value of K.

When $K = 0$, we find

$$E_\theta[RS \mid K = 0] = 2T\left(-\frac{T}{1 - \theta}\right) = -\frac{2T^2}{1 - \theta}.$$

Since $P(K = 0) = (1 - \theta)^{-T}$, this entails that

$$P(K = 0)\, E_\theta[RS \mid K = 0] = -2T^2(1 - \theta)^{T-1}.$$

This accounts for the second term in the derivative, and it quantifies the incentive to make θ smaller in order to make the reward $2T$ occur more frequently.

When $K > 0$, we note that $R = K$, so

$$P(K > 0)\, E_\theta[RS \mid K > 0] \quad = \quad E_\theta[KS].$$

This expected value, in turn, is equal to

$$\begin{aligned} E_\theta[KS] &= \frac{E_\theta[K^2]}{\theta(1-\theta)} - \frac{T\, E_\theta[K]}{1-\theta} \\ &= \frac{T\theta(1-\theta) + T^2\theta^2}{\theta(1-\theta)} - \frac{T^2\theta}{1-\theta} \\ &= T + \frac{T^2\theta}{1-\theta} - \frac{T^2\theta}{1-\theta} \\ &= T. \end{aligned}$$

Putting these two terms together, we have thus confirmed directly that

$$\underbrace{E_\theta\left[\frac{KR}{\theta(1-\theta)} - \frac{TR}{1-\theta}\right]}_{\text{estimator}} = \underbrace{T - 2T^2(1-\theta)^{T-1}}_{\text{gradient}}.$$

4 Estimation Problems

This is all well and good. However, the issue here is that the gradient is very likely to be estimated wrong, causing the agent to converge on the suboptimal equilibrium, even when it starts in the right basin of attraction.

Specifically, the random variable RS takes a small positive value with probability $1 - (1 - \theta)^T$ and a large negative value with probability $(1 - \theta)^T$. When T is large, the variable thus spends most of its time being positive, even though its mean is negative.

For instance, for $T = 5$ and $\theta = 2/5$, we get the following sample space:

X	x_0	x_1	x_2	x_3	x_4	x_5	$E[X]$
K	0	1	2	3	4	5	2.0
R	10	1	2	3	4	5	2.8
S	-8.3	-4.2	0.0	4.2	8.3	12.5	0.0
RS	-83.3	-4.2	0.0	12.5	33.3	62.5	-1.5
$P(X = x_k)$	.08	.26	.35	.23	.08	.01	.25

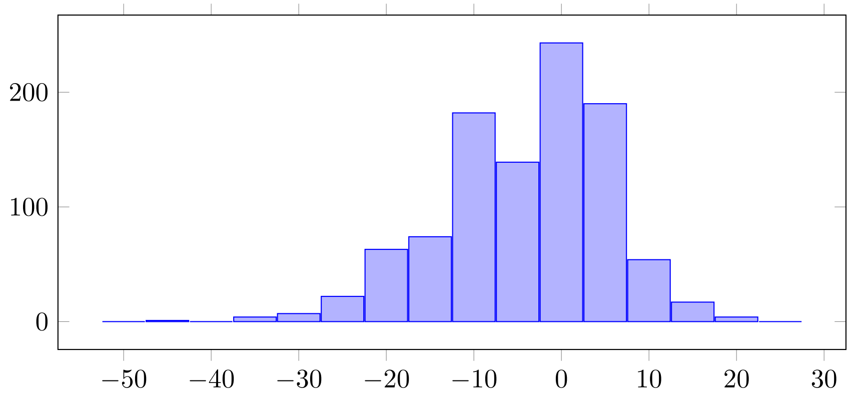

Figure 2: A histogram of 1000 averages $\hat{E}_\theta[RS]$, with $N = 8$ and $T = 5$.

In this case, the gradient estimate is negative with about 68% probability, so there is a fair chance of observing a positive value of RS even though the true gradient is -1. If we estimate the gradient from a single sample of RS, we are thus very likely to get an estimate with the wrong sign.

This effect remains even if we estimate the gradient from a batch of N trajectories, that is, as an empirical average of the form

$$\frac{\partial}{\partial\theta} E_\theta[R] \;\approx\; \sum_{n=1}^{N} R(a_n)S(a_n).$$

Figure 2 shows the distribution of this average for $N = 8$. As the figure indicates, the estimate has a very large variance, and it has a substantial probability of being positive.

In fact, since $P(K(a_n) = 0) \approx 8\%$ for a single trajectory a_n, there is just over 50% probability that $K(a_n) > 0$ for all 8 trajectories, in spite of the fairly low parameter setting of $\theta = 2/5$. Conditional on the event that $K(a_n)$ for all n, the mean estimate is

$$E_\theta[RS \mid K > 0] \;=\; \frac{T}{P(K > 0)} \;=\; \frac{5}{.92} \;=\; 5.4,$$

even though the true value is -1.5. (When $K = 0$, the gradient estimate is -83.3).

We thus have a substantial probability of walking in the wrong direction, even if we estimate the gradient from $N = 8$ trajectories of length $T = 5$. This reflects the fact that when T is large enough, even a near-optimal policy with $\theta \approx 0$ is unlikely to hit upon the huge treasure buried under the zero vector.