ASSIGNMENT

2457 Learning to Search
Fall 2019

University of Toronto

1. Gradient estimators, 20 points. All else being equal, it's useful for a gradient estimator to be unbiased. This is because the unbiasedness of a gradient estimator guarantees that, if we decay the step size and run stochastic gradient descent for long enough (see Robbins & Monroe), it will converge to a local optimum.

The standard REINFORCE, or score-function estimator is defined as:

$$\hat{g}_{\mathrm{SF}}[f] = f(b)\frac{\partial}{\partial\theta}\log p(b|\theta), \qquad b \sim p(b|\theta) \tag{1.1}$$

(a) **[5 points]** First, let's warm up with the score function. Prove that the score function has zero expectation, i.e. $\mathbb{E}_{p(x|\theta)}[\nabla_\theta \log p(x|\theta)] = 0$. Assume that you can swap the derivative and integral operators.

(b) **[5 points]** Show that REINFORCE is unbiased: $\mathbb{E}_{p(b|\theta)}\left[f(b)\frac{\partial}{\partial\theta}\log p(b|\theta)\right] = \frac{\partial}{\partial\theta}\mathbb{E}_{p(b|\theta)}[f(b)]$.

(c) **[5 points]** Show that REINFORCE with a fixed baseline is still unbiased, i.e. show that

$$\mathbb{E}_{p(b|\theta)}\left[[f(b) - c]\frac{\partial}{\partial\theta}\log p(b|\theta)\right] = \frac{\partial}{\partial\theta}\mathbb{E}_{p(b|\theta)}[f(b)]$$

for any fixed c.

(d) **[5 points]** If the baseline depends on b, then REINFORCE will in general give biased gradient estimates. Give a concrete for $p(b|\theta), f(), c()$, and θ example where

$$\mathbb{E}_{p(b|\theta)}\left[[f(b) - c(b)]\frac{\partial}{\partial\theta}\log p(b|\theta)\right] \neq \frac{\partial}{\partial\theta}\mathbb{E}_{p(b|\theta)}[f(b)]$$

for some function $c(b)$, and show that it is biased. That is, compute the actual expectation of both sides and write the numbers.

The takeaway is that you can use a baseline to reduce the variance of REINFORCE, but not one that depends on the current action.

2. Comparing variances of gradient estimators, 25 points. If we restrict ourselves to consider only unbiased gradient estimators, then the main property we need to worry about is the variance of our estimators. In general, optimizing with a lower-variance unbiased estimator will converge faster than a high-variance unbiased estimator. However, which estimator has the lowest variance can depend on the function being optimized. In this question, we'll look at which gradient estimators scale to large numbers of parameters, by computing their variance as a function of dimension.

For simplicity, we'll consider a toy problem. The goal will be to estimate the gradients of the expectation of a sum of D independent one-dimensional Gaussians, along with a regularization term. Each Gaussian x_i in independent, has unit variance, and its mean is given by an element of the D-dimensional parameter vector θ:

$$f(\mathbf{x}) = \sum_{d=1}^{D} x_d \tag{2.1}$$

$$L(\theta) = \mathbb{E}_{\mathbf{x} \sim p(x|\theta)}[f(\mathbf{x})] \tag{2.2}$$

(a) **[4 points]** As a warm-up, compute the variance of a single-sample simple Monte Carlo estimator of the objective $L(\theta)$:

$$\hat{L}_{MC} = \sum_{d=1}^{D} x_d, \qquad \text{where each } x_d \sim_{\text{iid}} \mathcal{N}(\theta_d, 1) \tag{2.3}$$

That is, compute $\mathbb{V}\left[\hat{L}_{MC}\right]$ as a function of D.

(b) **[5 points]** The score-function, or REINFORCE estimator with a baseline has the form:

$$\hat{g}^{\text{SF}}[f] = [f(x) - c(\theta)] \frac{\partial}{\partial \theta} \log p(x|\theta), \qquad x \sim p(x|\theta) \tag{2.4}$$

Derive a closed form for this gradient estimator for the objective defined above as a deterministic function of ϵ, a D-dimensional vector of standard normals. Set the baseline to $c(\theta) = \sum_{d=1}^{D} \theta_d$. Hint: When simplifying $\frac{\partial}{\partial \theta} \log p(x|\theta)$, you shouldn't take the derivative through x, even if it depends on θ. To help keep track of what is being differentiated, you can use the notation $\frac{\partial}{\partial \theta} g(\bar{\theta}, \theta)$ to denote taking the derivative only w.r.t. the second θ.

(c) **[8 points]** Derive the variance of the above gradient estimator. Because gradients are D-dimensional vectors, their covariance is a $D \times D$ matrix. To make things easier, we'll consider only the variance of the gradient with respect to the first element of the parameter vector, θ_1. That is, derive the scalar value $\mathbb{V}\left[\hat{g}_1^{\text{SF}}\right]$ as a function of D. Hint: The third moment of a standard normal is 0, and the fourth moment is 3. As a sanity check, consider the case where $D = 1$.

(d) **[8 points]** Next, let's look at the gold standard of gradient estimators, the reparameterization gradient estimator, where we reparameterize $x = T(\theta, \epsilon)$:

$$\hat{g}^{\text{REPARAM}}[f] = \frac{\partial f}{\partial x} \frac{\partial x}{\partial \theta}, \qquad \epsilon \sim p(\epsilon) \tag{2.5}$$

In this case, we can use the reparameterization $x = \theta + \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. Derive this gradient estimator for $\nabla_\theta L(\theta)$, and give $\mathbb{V}\left[\hat{g}_1^{\text{REPARAM}}\right]$ as a function of D.

3. Search Trees, 5 points. The expectimax computes the optimal first action in an Markov decision process, summing over all S possible states and taking a max over all A possible actions at each time step, to a horizon of T steps:

$$a_1 * = \operatorname*{argmax}_{a_1} \mathbb{E}_{p(s_1|a_1)} \left[R(s_1) + \max_{a_2} \mathbb{E}_{p(s_2|a_2, s_1)} \left[R(s_2) + \cdots + \mathbb{E}_{p(s_T|a_T, s_{T-1})} [R(s_T)] \ldots \right] \right]$$

(a) **[5 points]** What is the asymptotic time complexity of computing the exact expectimax by exhaustive enumeration, as a function of S, A, and T? You can assume R has cost $\mathcal{O}(1)$.