Approximating KL Divergence

annxhe

June 2021

Introduction

- These notes are from John Schulman's blog (http://joschu.net/blog/kl-approx.html) and Augstinus Kristiadi's Blog (https://wiseodd.github.io/techblog/2018/03/11/fisher-information/)
- I'm rewriting them to understand how to implement approximating the KL divergence of two policies in TRPO

Statement

We want to Monte-Carlo approximate the KL divergence

$$KL[q,p] = \sum_x q(x) \log \frac{q(x)}{p(x)} = \mathbb{E}_{x\sim q}[\log \frac{q(x)}{p(x)}] \tag{1}$$

as

$$\frac{1}{2}(\log p(x) - \log q(x))^2 \tag{2}$$

We want to un-bias the above estimator while preserving its low variance.

We assume that we can compute the probabilities $p(x)$ and $q(x)$ for any x, but we can't calculate the sum over x analytically

Estimators

- We know that one unbiased estimator (under samples from q) is $\log \frac{q(x)}{p(x)}$
- Call it $k_1 = \log \frac{q(x)}{p(x)} = -\log r$ where $r = \frac{p(x)}{q(x)}$
- an alternate lower variance estimator that is biased is $\frac{1}{2}(\log \frac{p(x)}{q(x)})^2 = \frac{1}{2}(\log r)^2 = k_2$
- Empirically, k_2 has lower variance than k_1, and also has low bias

Reason for Low Bias

- f-divergence
- An f-divergence is defined as $D_f(p,q) = \mathbb{E}_{x\sim q}[f(\frac{p(x)}{q(x)})]$ for a convex function f
- KL divergence and and other well-known probability distances are f-divergences
- Key insight: all f-divergences with differentiable f look like KL divergence up to second order when q and p are close, i.e.

$$D_f(p_0, p_\theta) = \frac{f''(1)}{2}\theta^T F\theta + O(\theta^3) \tag{3}$$

where F is the Fisher information matrix for p_θ evaluated at $p_\theta = p_0$

- $\mathbb{E}_q[k_2] = \mathbb{E}_q[\frac{1}{2}(\log r)]$ is the f-divergence where $f(x) = \frac{1}{2}(\log x)^2$
- $KL[q,p]$ has $f(x) = -\log x$
- We can check that $f''(1) = 1$, so both look like the same quadratic distance fn for $p \approx q$

Q: Did we ever use the fact $p \approx q$?

Variance

- The general way to lower variance is with a control variate
- I.e., take k_1 and add something that has expectation zero but is negatively correlated with k_1
- The only interesting quantity that's guaranteed to have zero expectation is $\frac{p(x)}{q(x)} - 1 = r - 1$
- $\log r + \lambda(r-1)$ is an unbiased estimator of $KL[q,p]$
- Since log is concave, $\log(x) \leq x - 1$
- So letting $\lambda = 1$, the expression $\log r + \lambda(r-1)$ is guaranteed positive
- It measures the vertical distance between $\log(x)$ and its tangent
- Now our estimator is $k_3 = (r-1) - \log r$
- The idea of mesauring distance by looking at the difference between a convex function and its tangent plan appears in many places. It's called a Bregman divergence

Fisher Information Matrix

- S'pose we have a model paramterized by θ modeling $p(x|\theta)$
- In frequentist stats, the way we learn θ is by maximizing the likelihood $p(x|\theta)$ wrt θ
- To assess the goodness of our estimate of θ we define a score function

$$s(\theta) = \nabla_\theta \log p(x|\theta) \tag{4}$$

- The score function is the gradient of the log likelihood function

Lemma 0.1. *The expected value of score wrt our model is zero.*

Proof. Below, the gradient is wrt to θ

$$\begin{aligned}
\mathbb{E}_{p(x|\theta)}[s(\theta)] &= \mathbb{E}_{p(x|\theta)}[s(\theta)] \\
&= \int \nabla \log p(x|\theta) p(x|\theta) dx \\
&= \int \frac{\nabla \log p(x|\theta)}{p(x|\theta)} p(x|\theta) dx \\
&= \int \nabla p(x|\theta) dx \\
&= \nabla \int p(x|\theta) dx \\
&= 1 \\
&= 0
\end{aligned}$$

We can also define an uncertainty measure around the expected estimate. That is, we look at the covariance of the score of our model.

$$\mathbb{E}_{p(x|\theta)}[(s(\theta) - 0)(s(\theta) - 0)^T] \tag{5}$$

The covariance of the score function above is the definition of Fisher Information.

$$F = \mathbb{E}_{p(x|\theta)}[\nabla \log p(x|\theta) \nabla \log p(x|\theta)^T] \tag{6}$$

As usual, computing the expectation is intractable. We can approximated it using hte empirical distribution, and then F is the Empirical Fisher.

$$F = \frac{1}{N} \sum_{i=1}^{N} \nabla \log p(x_i|\theta) \nabla \log p(x_i|\theta)^T \tag{7}$$

Fisher and Hessian

A super not obvious property of F is that it is the negative expected Hessian ov the model's log likelihood.

Claim: The Hessian of the log likelihood is given by the Jacobian of its gradient.

$$\begin{aligned}
H_{\log p(x|\theta)} &= J(\frac{\nabla p(x|\theta)}{p(x|\theta)}) \\
&= \frac{H_{p(x|\theta)} p(x|\theta) - \nabla p(x|\theta) \nabla p(x|\theta)^T}{p(x|\theta) p(x|\theta)} \\
&= \frac{H_{p(x|\theta)} p(x|\theta)}{p(x|\theta) p(x|\theta)} - \frac{\nabla p(x|\theta) \nabla p(x|\theta)^T}{p(x|\theta) p(x|\theta)} \\
&= \frac{H_{p(x|\theta)}}{p(x|\theta)} - (\frac{\nabla p(x|\theta)}{p(x|\theta)})(\frac{\nabla p(x|\theta)}{p(x|\theta)})^T
\end{aligned}$$

Where we apply the quotient rule to the first line. Taking expectation wrt the model,

$$\begin{aligned}
\mathbb{E}_{p(x|\theta)}[H_{\log p(x|\theta)}] &= \mathbb{E}_{p(x|\theta)}\big[\frac{H_{p(x|\theta)}}{p(x|\theta)} - (\frac{\nabla p(x|\theta)}{p(x|\theta)})(\frac{\nabla p(x|\theta)}{p(x|\theta)})^T\big] \\
&= \mathbb{E}_{p(x|\theta)}\big[\frac{H_{p(x|\theta)}}{p(x|\theta)}\big] - \mathbb{E}_{p(x|\theta)}\big[(\frac{\nabla p(x|\theta)}{p(x|\theta)})(\frac{\nabla p(x|\theta)}{p(x|\theta)})^T\big] \\
&= \int \frac{H_{\log p(x|\theta)}}{p(x|\theta)} p(x|\theta) dx - \mathbb{E}_{p(x|\theta)}\big[\log p(x|\theta) \nabla \log p(x|\theta)^T\big] \\
&= H_{\int p(x|\theta) dx} - F \\
&= H_1 - F \\
&= -F
\end{aligned}$$

F, covariance of the model's score, is a measure of the curvature of the log likelihood function.

Restated, Fisher Information Matrix is defined as the covariance of the score function. It has interpretation as the negative expected Hessian of the log likelihood function. This suggests immediate application of F as the drop-in replacmenet of H in second order optimization methods.

Connections to KL Divergence

TRPO plan

- We need to compute $\hat{H}_k$, the Hessian of the sample average KL divergence
- The Fisher matrix is the Hessian of the KL-divergence
- So, naively as a first step, we can use the sample Fisher matrix as the Hessian of the sample KL-divergence

- Then we can replace this estimator with a Hessian of the sample KL-divergence where KL-divergence is approximate using John's method; I think by deriving a sampling form of it

Proof that F is the Hessian of the KL-divergence between $p(x|\theta)$ and $p(x|\theta')$ with respect to θ', evaluated at $\theta' = theta$

Note: "evaluated at y = sampled from y"

KL-divergence can be decomposed into entropy and cross-entropy term:

$$KL\big[p(x|\theta)||p(x|\theta')\big] = \mathbb{E}_{p(x|\theta)}\big[\log p(x|\theta)\big] - \mathbb{E}_{p(x|\theta)}\big[\log p(x|\theta')\big] \tag{8}$$

The first derivaitve wrt θ' is:
To-do