# Subtleties of estimating entropy

Wei Xu

July 16, 2019

For some algorithms, we need to calculate the entropy and its derivative. If there is no analytic formula for the entropy, we can resort to sampling. Given the definition of entropy:

$$H(p) = E_{x\sim p_\theta}(-\log p_\theta(x)) \tag{1}$$

we can see that $-\log p(x)$ is an unbiased estimator of $H$ if $x$ is sampled from $p$. It is tempting to use $-\frac{\partial \log p_\theta(x)}{\partial \theta}$ as an estimator of $\frac{\partial H}{\partial \theta}$. However, it is wrong, as shown in the following:

$$\begin{aligned} & E_{x\sim p_\theta}\left(\frac{\partial \log p_\theta(x)}{\partial \theta}\right) = \int \frac{\partial \log p_\theta(x)}{\partial \theta} p_\theta(x) dx \\ = & \int \frac{\partial p_\theta(x)}{\partial \theta} dx = \frac{\partial}{\partial \theta}\int p_\theta(x) dx = \frac{\partial 1}{\partial \theta} = 0 \end{aligned}$$

We need to actually go through the process of calculating the derivative to get the unbiased estimator of $\frac{\partial H}{\partial \theta}$:

$$\begin{aligned} \frac{\partial H}{\partial \theta} &= -\frac{\partial}{\partial \theta}\int \log p_\theta(x) p_\theta(x) dx \\ &= -\int \left(\frac{\partial \log p_\theta(x)}{\partial \theta} p_\theta(x) + \log p_\theta(x) \frac{\partial p_\theta(x)}{\partial \theta}\right) dx \\ &= -\int \left(\frac{\partial \log p_\theta(x)}{\partial \theta} p_\theta(x) + \log p_\theta(x) \frac{\partial \log p_\theta(x)}{\partial \theta} p_\theta(x)\right) dx \\ &= -\int (1 + \log p_\theta(x)) \frac{\partial \log p_\theta(x)}{\partial \theta} p_\theta(x) dx \\ &= -E_{x\sim p_\theta}\left(\log p_\theta(x) \frac{\partial \log p_\theta(x)}{\partial \theta}\right) - E_{x\sim p_\theta}\left(\frac{\partial \log p_\theta(x)}{\partial \theta}\right) \\ &= -\frac{1}{2} E_{x\sim p_\theta}\left(\frac{\partial}{\partial \theta}(\log p_\theta(x))^2\right) \end{aligned}$$

This means that $-\frac{1}{2}\frac{\partial}{\partial \theta}(\log p_\theta(x))^2$ is an unbiased estimator of $\frac{\partial H}{\partial \theta}$. Actually, $-\frac{1}{2}\frac{\partial}{\partial \theta}(c + \log p_\theta(x))^2$ is an unbiased estimator for any constant $c$.

For some distributions, the sample of $p_\theta$ is generated by transforming $\epsilon \sim q$ by $f_\theta(\epsilon)$, where $q$ is a fixed distribution and $f_\theta$ is a smooth bijective mapping. $p_\theta(x)$ is implicitly defined by $q$ and $f_\theta$ as:

$$p_\theta(x) = q(f_\theta^{-1}(x)) / \left| \det \left. \frac{\partial f_\theta(\epsilon)}{\partial \epsilon} \right|_{\epsilon = f_\theta^{-1}(x)} \right|$$

Interestingly, when calculating $-\frac{\partial \log p_\theta(x)}{\partial \theta}$, if we treat $x$ as $x = f_\theta(\epsilon)$, we get an unbiased estimator of $\frac{\partial H}{\partial \theta}$:

$$\begin{aligned} & E_{x \sim p_\theta} \left( -\frac{\partial \log p_\theta(x)}{\partial \theta} \right) = E_{\epsilon \sim q} \left( -\frac{\partial \log p_\theta(f_\theta(\epsilon))}{\partial \theta} \right) \\ = \quad & -\frac{\partial}{\partial \theta} E_{\epsilon \sim q} \left( \log p_\theta(f_\theta(\epsilon)) \right) = -\frac{\partial}{\partial \theta} E_{x \sim p_\theta} \left( \log p_\theta(x) \right) = \frac{\partial}{\partial \theta} H(p) \end{aligned}$$

So we can use $-\frac{\partial \log p_\theta(x)}{\partial \theta}$ as an unbiased estimator of $\frac{\partial H(p)}{\partial \theta}$ if $x = f_\theta(\epsilon)$ and we allow gradient to propagate through $x$ to $\theta$.