1 Maximum log likelihood is minimizing KL divergence

In statistics and machine learning, we often try to maximize probability or log likelihood

$$\arg\max_w \prod_{n=1}^{N} p(x_n|w) = \arg\max_w \sum_{n=1}^{N} \log p(x_n|w) \tag{1}$$

We show that this is equivalent to $\arg\min_w KL\left(p_{data}(x) \| p(x|w)\right.$, when the samples are drawn iid from $p_{data}(x)$:

$$KL(p_{data}(x)\|p(x|w) = -H\left(p_{data}(x)\right) - E_{x\sim P_{data}(x)}\left[\log p(x|w)\right] \tag{2}$$

The entropy of $p_{data}(x)$ is unaffected by w, so it is unaffected by the minimization. Hence minimizing KL is the same as minimizing the cross entropy.

One useful property is that we can use samples to estimate the cross-entropy, without having to know the explicit probability distribution of $p_{data}(x)$. Other divergence measures may not be so direct.

We use Monte Carlo estimation to estimate the cross-entropy.

$$\frac{1}{N}\sum_{n=1}^{N} \log p(x_n|w) \to E_{x\sim p_{data}(x)}\left[\log p(x|w)\right] \tag{3}$$

This justified in statistics literature by the weak law of large numbers, and there is an almost identical proof of the Asymptotic Equipartition Property (AEP) theorem, given in Chapter 3 of Cover and Thomas.

Therefore, maximizing log likelihood is equivalent to minimizing KL divergence to the class of distributions parametrized by w

$$\arg\min_w KL\left(p_{data}(x)\|p(x|w)\right) \approx \arg\min_w -\frac{1}{N}\sum_{n=1}^{N} \log p(x_n|w) \tag{4}$$

1.1 Maximum likelihood with exponential families

From a previous problem, we showed that $\nabla_\eta \log p(x|\eta) = u(x) - E_{x\sim p(x|\eta)}\left[u(x)\right]$

Hence if we perform a gradient update on the maximum log likelihood / KL minimization objective over a minibatch, we get

$$\begin{aligned}
\nabla_\eta KL\left(p_{data}(x)\|p(x|\eta)\right) &\approx -\nabla_\eta \frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\eta) \\
&= -\frac{1}{N}\sum_{n=1}^{N}\left(u(x_n) - E_{x\sim p(x|\eta)}\left[u(x)\right]\right) \\
&= -\left(\frac{\sum_{n=1}^{N} u(x_n)}{N} - E_{x\sim p(x|\eta)}\left[u(x)\right]\right) \\
&= -\left(E_{x\sim p_{emp}(x)}\left[u(x)\right] - E_{x\sim p(x|\eta)}\left[u(x)\right]\right)
\end{aligned} \tag{5}$$

For the last step, we observe that we can compute an average over certain points as a mixture of deterministic distributions, which can be written as a delta function.

$$E\left[f(x)\right] = \int \delta(x = x')f(x)dx = f(x') \tag{6}$$

Then we define our empirical distribution as $p_{emp}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta(x = x_n)$. Then evaluating our expectation

$$\begin{aligned} E_{x\sim p_{emp}(x)}\left[u(x)\right] &= \int \frac{1}{N}\sum_{n=1}^{N}\delta(x = x_n)u(x)dx \\ &= \frac{1}{N}\sum_{n=1}^{N}\int \delta(x = x_n)u(x)dx \\ &= \frac{1}{N}\sum_{n=1}^{N}u(x_n) \end{aligned} \tag{7}$$

Hence for arbitrary exponential families, we can express both maximum gradient updates and KL minimization gradient updates as residuals.