1 KL Divergence Gradient

1.1 Formulation

This is what I have been able to piece together on how to (approximately) compute the gradient of the KL divergence for a Restricted Boltzmann Machine (RBM) based on Hinton[1] and Torlai and Melko[2]. [1] is the source of this scheme, and (while not immediately clear on the first reading) contains all the necessary information. [2], on the other hand, makes reference to [1], but on its own it *does not* have all the necessary information and I would even say that it is misleading; it was still mildly helpful in pieceing the puzzle together, though.

The (approximate) formulae for the KL divergence gradients may be written as

$$\nabla_{W_{ij}} \mathbb{KL}(P^0||P^\infty_\lambda) \approx \left\langle \sigma_j \langle h_i \rangle_{P^\infty_\lambda(\mathbf{h}|\boldsymbol{\sigma})} \right\rangle_{P^k_\lambda(\boldsymbol{\sigma})} - \left\langle \sigma_j \langle h_i \rangle_{P^\infty_\lambda(\mathbf{h}|\boldsymbol{\sigma})} \right\rangle_{P^0(\boldsymbol{\sigma})} \tag{1a}$$

$$\nabla_{b_j} \mathbb{KL}(P^0||P^\infty_\lambda) \approx \langle \sigma_j \rangle_{P^k_\lambda(\boldsymbol{\sigma})} - \langle \sigma_j \rangle_{P^0(\boldsymbol{\sigma})} \tag{1b}$$

$$\nabla_{c_i} \mathbb{KL}(P^0||P^\infty_\lambda) \approx \left\langle \langle h_i \rangle_{P^\infty_\lambda(\mathbf{h}|\boldsymbol{\sigma})} \right\rangle_{P^k_\lambda(\boldsymbol{\sigma},\mathbf{h})} - \left\langle \langle h_i \rangle_{P^\infty_\lambda(\mathbf{h}|\boldsymbol{\sigma})} \right\rangle_{P^0(\boldsymbol{\sigma})} \tag{1c}$$

This is the *contrastive-divergence* or CD-k approximation. $\lambda = (\mathbf{W}, \mathbf{b}, \mathbf{c})$ represents the parameters in the energy

$$E_\lambda(\boldsymbol{\sigma}, \mathbf{h}) = \mathbf{b}^{\mathrm{T}}\boldsymbol{\sigma} + \mathbf{h}^{\mathrm{T}}\mathbf{c} + \mathbf{h}^{\mathrm{T}}\mathbf{W}\boldsymbol{\sigma}. \tag{2}$$

$P^0(\boldsymbol{\sigma})$ is the distribution of input data. $P^\infty_\lambda(\boldsymbol{\sigma}, \mathbf{h})$ is the joint distribution over input and hidden nodes of the RBM, with the marginal distribution

$$P^\infty_\lambda(\boldsymbol{\sigma}) = \sum_{\mathbf{h}} P^\infty_\lambda(\boldsymbol{\sigma}, \mathbf{h}) \tag{3}$$

being the RBM's approximation to the exact input data distribution, and

$$P^\infty_\lambda(\mathbf{h}|\boldsymbol{\sigma}) = \frac{P^\infty_\lambda(\boldsymbol{\sigma}, \mathbf{h})}{P^\infty_\lambda(\boldsymbol{\sigma})} \tag{4}$$

is the conditional probability of $\mathbf{h}$.

$P^k_\lambda(\boldsymbol{\sigma})$ for some k is the distribution acquired from k Gibbs sampling steps on P^0 via P^∞_λ. More precisely, if we have an input data sequence $(\boldsymbol{\sigma}_n)_{n=1}^N$ so that

$$P^0(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{n=1}^{N} [\boldsymbol{\sigma} = \boldsymbol{\sigma}_n] \tag{5}$$

where $[\cdot]$ is the Iverson bracket, then we form the Markov chain

$$\boldsymbol{\sigma}^{(0)}_n = \boldsymbol{\sigma}_n \to \mathbf{h}^{(0)}_n \to \boldsymbol{\sigma}^{(1)}_n \to \mathbf{h}^{(1)}_n \to \cdots \to \boldsymbol{\sigma}^{(k)}_n \to \mathbf{h}^{(k)}_n \tag{6}$$

by sampling $\mathbf{h}^{(l)}_n$ from $P^\infty_\lambda(\mathbf{h}|\boldsymbol{\sigma}^{(l)}_n)$ and sampling $\boldsymbol{\sigma}^{(l+1)}_n$ from $P^\infty_\lambda(\boldsymbol{\sigma}|\mathbf{h}^{(l)}_n)$, with $l = 0, \ldots, k$. The distribution P^k_λ is then

$$P^k_\lambda(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{n=1}^{N} [\boldsymbol{\sigma} = \sigma^{(k)}_n]. \tag{7}$$

Forming the Markov chain is tractable, since the conditional probabilities factor and are easily computed:

$$P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}) = \prod_i P_\lambda^\infty(h_i|\boldsymbol{\sigma}), \quad P_\lambda^\infty(\boldsymbol{\sigma}|\mathbf{h}) = \prod_j P_\lambda^\infty(\sigma_j|\mathbf{h}), \tag{8a}$$

$$P_\lambda^\infty(h_i = 1|\boldsymbol{\sigma}) = \sum_j \text{sigm}(\mathbf{W}_j\boldsymbol{\sigma} + b_j), \tag{8b}$$

$$P_\lambda^\infty(\sigma_j = 1|\mathbf{h}) = \sum_i \text{sigm}(\mathbf{h}^\text{T}\mathbf{W}^i + c_i). \tag{8c}$$

The function sigm is the sigmoid function

$$\text{sigm}(x) = \frac{1}{1+e^{-x}}, \tag{9}$$

and $\mathbf{W}_j$ is the j^th row of $\mathbf{W}$, and $\mathbf{W}^i$ the i^th column of $\mathbf{W}$. Equations (8b) and (8c) are sufficient since h_i, σ_j are binary random variables.

1.2 Implementation

Since each $\boldsymbol{\sigma}_n$ corresponds one-to-one with a $\boldsymbol{\sigma}_n^{(k)}$, we see that we can perform both averages in Equations (1) in the same loop, and so we do not have to precompute $(\boldsymbol{\sigma}_n^{(k)})_{n=1}^N$. For example,

$$\begin{aligned}
\nabla_{W_{ij}}\mathbb{KL}(P^0||P_\lambda^\infty) &\approx \left\langle \sigma_j \langle h_i \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma})} \right\rangle_{P_\lambda^k(\boldsymbol{\sigma})} - \left\langle \sigma_j \langle h_i \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma})} \right\rangle_{P^0(\boldsymbol{\sigma})} \\
&= \frac{1}{N}\sum_{n=1}^N \sigma_{n,j}^{(k)} \langle h_i \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}_n^{(k)})} - \frac{1}{N}\sum \sigma_{n,j} \langle h_i \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}_n)} \\
&= \frac{1}{N}\sum_{n=1}^N \left[\sigma_{n,j}^{(k)} \langle h_i \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}_n^{(k)})} - \sigma_{n,j} \langle h_i \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}_n)} \right].
\end{aligned}$$

At this point, all we have left is to compute the averages over h_i. I have identified two options:

1. Since we have Equations (8), we can compute the average over $\mathbf{h}$ exactly:

$$\begin{aligned}
\langle h_i \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma})} &= 1 \cdot P_\lambda^\infty(h_i = 1|\boldsymbol{\sigma}) + 0 \cdot P_\lambda^\infty(h_i = 0|\boldsymbol{\sigma}) \\
&= \sum_j \text{sigm}(\mathbf{W}_j\sigma + b_j).
\end{aligned} \tag{10}$$

2. It appears to be more common in the implementation of RBM's to instead make the approximations

$$\langle \mathbf{h} \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}_n)} = \mathbf{h}_n^{(0)} P_\lambda^\infty(\mathbf{h}_n^{(0)}|\boldsymbol{\sigma}_n) \quad + \sum_{\mathbf{h}} \mathbf{h} P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}_n)[\mathbf{h} \neq \mathbf{h}_n^{(0)}] \quad \approx \mathbf{h}_n^{(0)}, \tag{11a}$$

$$\langle \mathbf{h} \rangle_{P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}_n^{(k)})} = \mathbf{h}_n^{(k)} P_\lambda^\infty(\mathbf{h}_n^{(k)}|\boldsymbol{\sigma}_n^{(k)}) + \sum_{\mathbf{h}} \mathbf{h} P_\lambda^\infty(\mathbf{h}|\boldsymbol{\sigma}_n^{(k)})[\mathbf{h} \neq \mathbf{h}_n^{(k)}] \approx \mathbf{h}_n^{(k)}. \tag{11b}$$

By the definition of the Markov chain (6), the probabilities $P_\lambda^\infty(\mathbf{h}_n^{(0)}|\boldsymbol{\sigma}_n)$ and $P_\lambda^\infty(\mathbf{h}_n^{(k)}|\boldsymbol{\sigma}_n^{(k)})$ will usually be high or approximately the same as that of each of the terms thrown away, hence why this approximation could be considered reasonable. At the very least, this is more computationally efficient than the exact method in Equation (10). It is worth noting that in this approximate scheme $(\boldsymbol{\sigma}_n^{(0)}, \mathbf{h}_n^{(0)})_{n=1}^N$ are aften called the *positive* units, and $(\boldsymbol{\sigma}_n^{(k)}, \mathbf{h}_n^{(k)})_{n=1}^N$ are called the *negative* units.

References

[1] G. Hinton. "Training Products of Experts by Minimizing Contrastive Divergence". In: *Neural Comput.* 14 (2002), pp. 1771–1800.

[2] G. Torlai and R. G. Melko. "Learning Thermodynamics with Boltzmann Machines". In: `arXiv:1606.02718v1 [cond-mat.stat-mech]` (2016). URL: `https://arxiv.org/abs/1606.02718`.