Persistent Minimum Probability Flow

Jascha Sohl-Dickstein

July 18, 2011

1 Sampling the connectivity matrix g_{ij}

As described in more detail in a separate note (available at http://redwood.berkeley.edu/jascha/), the connectivity function g_{ij} can be treated as the probability of a connection from state j to state i, rather than as a binary indicator function. In this case, g_{ij} has the constraints required of a probability distribution,

$$\sum_i g_{ij} = 1 \tag{1}$$

$$g_{ij} \geq 0, \tag{2}$$

as well as the added constraint that if $g_{ij} > 0$ then $g_{ji} > 0$. Given these constraints for g_{ij}, the following form can be chosen for the transition rates Γ_{ij},

$$\Gamma_{ij}(\theta) = \begin{cases} g_{ij}\left(\frac{g_{ji}}{g_{ij}}\right)^{\frac{1}{2}} \exp\left[\frac{1}{2}\left(E_j(\theta) - E_i(\theta)\right)\right] & i \neq j \\ -\sum_{k \neq j} \Gamma_{kj}(\theta) & i = j \end{cases}. \tag{3}$$

It can be seen by substitution that the form for Γ_{ij} in Equation 3 still satisfies detailed balance.

Using Γ_{ij} from Equation 3, the MPF objective function becomes

$$K_{MPF}(\theta; \mathbf{g}) = \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \sum_{i \notin \mathcal{D}} g_{ij} \left(\frac{g_{ji}}{g_{ij}}\right)^{\frac{1}{2}} \exp\left[\frac{1}{2}\left(E_j(\theta) - E_i(\theta)\right)\right]. \tag{4}$$

This is identical to the original MPF objective function, except for the addition of a scaling term $\left(\frac{g_{ji}}{g_{ij}}\right)^{\frac{1}{2}}$ which compensates for the differences between the forward and backward connectivity functions g_{ij} and g_{ji}.

Because g_{ij} is a probability distribution, the inner sum in Equation 4 is an expectation over g_{ij}, and can be approximated by averaging over sample states i drawn from the distribution g_{ij}.

2 Factoring K_{MPF}

If the proposed connectivity function g_{ij} depends only on the destination state, i, and not the initial state, j, then the nested sums in Equation 4 can be factored apart. For the case that g_{ij} does not depend on j, we write it simply as g_i. The MPF objective function K_{MPF} becomes

$$K_{MPF}(\theta; \mathbf{g}) = \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \sum_{i \notin \mathcal{D}} g_i \left(\frac{g_j}{g_i} \right)^{\frac{1}{2}} \exp\left[\frac{1}{2} \left(E_j(\theta) - E_i(\theta) \right) \right] \tag{5}$$

$$= \left(\frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \exp\left[\frac{1}{2} \left(E_j(\theta) + \log g_j \right) \right] \right) \cdot \left(\sum_{i \notin \mathcal{D}} g_i \exp\left[-\frac{1}{2} \left(E_i(\theta) + \log g_i \right) \right] \right). \tag{6}$$

The second sum is an expectation under g_i, and can be approximated by averaging over samples from g_i.

3 Iterative improvement of g_i

The most informative states to connect to for learning are those which are most probable under the model distribution. Therefore, it is useful for learning to make g_i as similar to $p_i^{(\infty)}(\theta)$ as possible. An effective learning procedure alternates between updating g_i to resemble the current estimate of the model distribution $p_i^{(\infty)}\left(\hat{\theta}\right)$, and updating the estimated model parameters $\hat{\theta}$ using samples from a fixed connectivity function g_i. Defining a sequence of estimated parameter vectors $\hat{\theta}^n$ and proposed connectivity distributions g_i^n, where n indicates the learning iteration, this learning procedure becomes

1. Set $\hat{\theta}^0 =$ initial parameter guess
2. For $n \in \mathcal{Z}_+$ iterate
   (a) Set $g_i^n = p_i^{(\infty)}\left(\hat{\theta}^{n-1}\right) = \frac{\exp\left[-E_i\left(\hat{\theta}^{n-1}\right)\right]}{Z\left(\hat{\theta}^{n-1}\right)}$
   (b) Find $\hat{\theta}^n$ such that $K_{MPF}^n\left(\hat{\theta}^n\right) < K_{MPF}^n\left(\hat{\theta}^{n-1}\right)$

The MPF objective function at learning step n, $K_{MPF}^n(\theta)$, is written using the proposal distribution g_i^n set in step 2a,

$$K_{MPF}^n(\theta) = \left(\frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \exp\left[\frac{1}{2} \left(E_j(\theta) - E_j\left(\hat{\theta}^{n-1}\right) \right) \right] \right) \cdot \left(\sum_{i \notin \mathcal{D}} g_i^n \exp\left[-\frac{1}{2} \left(E_i(\theta) - E_i\left(\hat{\theta}^{n-1}\right) \right) \right] \right) \tag{7}$$

(the normalization terms in $\log g_i$ cancel out between the two sums). The expectation in the second sum is still evaluated using samples from $\mathbf{g}^n$. Typically, the number of samples drawn from $\mathbf{g}^n$ will be the same as the number of observations, $|\mathcal{D}|$.

4 Persistent Minimum Probability Flow

The procedure in Section 3 will usually leave $\mathbf{g}^n$ very similar to $\mathbf{g}^{n-1}$. Additionally, the objective function in Equation 7 is typically evaluated using samples from $\mathbf{g}^n$. If sampling is done via Markov Chain Monte Carlo (MCMC), significant time can be saved when generating samples from $\mathbf{g}^n$ by initializing using the samples from $\mathbf{g}^{n-1}$.

5 Persistent Minimum Probability Flow in a continuous state space

The above description of Persistent Minimum Probability Flow (PMPF) learning applies to continuous as well as discrete state spaces. In fact, nearest neighbor schemes for setting the connectivity function $\mathbf{g}$ do not work nearly as well in continuous state spaces as in discrete state spaces, while PMPF works quite well in continuous state spaces, so PMPF is particularly applicable to the continuous state space case.

Using PMPF in an M-dimensional continuous states space $\mathcal{R}^M$, the parameter estimation procedure is as given in the steps below. $\mathcal{S}^n$ is the list of samples at learning step n. $|\mathcal{S}^n|$ is the number of samples - typically it will be the same as the number of observations $|\mathcal{D}|$.

1. Set $\hat{\theta}^0 =$ initial parameter guess
2. Initialize samples $\mathcal{S}^0$ (eg from a Gaussian)
3. For $n \in \mathcal{Z}_+$ iterate
   (a) Draw samples $\mathcal{S}^n$ from the distribution $p^{(\infty)}\left(\mathbf{x}; \hat{\theta}^{n-1}\right)$ via an MCMC sampler intialized at $\mathcal{S}^{n-1}$ (eg using Hamiltonian Monte Carlo)
   (b) Find $\hat{\theta}^n$ such that $K^n_{MPF}\left(\hat{\theta}^n\right) < K^n_{MPF}\left(\hat{\theta}^{n-1}\right)$ (eg via 10 steps of LBFGS gradient descent)

$K^n_{MPF}(\theta)$ is the MPF objective function at learning step n, and is written

$$K^n_{MPF}(\theta) = \left(\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \exp\left[\frac{1}{2}\left(E(\mathbf{x};\theta) - E\left(\mathbf{x};\hat{\theta}^{n-1}\right)\right)\right]\right) \cdot \left(\frac{1}{|\mathcal{S}^n|}\sum_{\mathbf{x}\in\mathcal{S}^n} \exp\left[-\frac{1}{2}\left(E(\mathbf{x};\theta) - E\left(\mathbf{x};\hat{\theta}^{n-1}\right)\right)\right]\right), \quad (8)$$

with derivative

$$\begin{aligned}\frac{\partial K_{MPF}^{n}(\theta)}{\partial\theta} =& \frac{1}{2}\left(\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}}\exp\left[\frac{1}{2}\left(E(\mathbf{x};\theta)-E\left(\mathbf{x};\hat{\theta}^{n-1}\right)\right)\right]\frac{\partial E(\mathbf{x};\theta)}{\partial\theta}\right)\cdot \\ & \left(\frac{1}{|\mathcal{S}^{n}|}\sum_{\mathbf{x}\in\mathcal{S}^{n}}\exp\left[-\frac{1}{2}\left(E(\mathbf{x};\theta)-E\left(\mathbf{x};\hat{\theta}^{n-1}\right)\right)\right]\right) \\ & -\frac{1}{2}\left(\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}}\exp\left[\frac{1}{2}\left(E(\mathbf{x};\theta)-E\left(\mathbf{x};\hat{\theta}^{n-1}\right)\right)\right]\right)\cdot \\ & \left(\frac{1}{|\mathcal{S}^{n}|}\sum_{\mathbf{x}\in\mathcal{S}^{n}}\exp\left[-\frac{1}{2}\left(E(\mathbf{x};\theta)-E\left(\mathbf{x};\hat{\theta}^{n-1}\right)\right)\right]\frac{\partial E(\mathbf{x};\theta)}{\partial\theta}\right).\end{aligned} \tag{9}$$