Learning deep energy models by NICE

August 10, 2018

For a system with dimensionless potential energy $U(x)$ in the space $\mathbb{R}^d$, the equilibrium distribution $\mu(x)$ can be represented as

$$\mu(x) = \frac{1}{C}\exp(-U(x)), \tag{1}$$

where C is the partition function. We consider here how to approximate the equilibrium distribution by the NICE network for given u.

Assume that x is modeled as a nonlinear transformation of a latent variable $z \in \mathbb{R}^d$ as

$$x = S \cdot F(z), \tag{2}$$

where z follows the standard Gaussian distribution $\mathcal{N}(z|0, I)$, F is a NICE network with

$$\left|\frac{\partial F}{\partial z}\right| \equiv 1, \tag{3}$$

and $S \in \mathbb{R}^{d\times d}$ is an invertible matrix. Then for a given z, the probability density of x is

$$\mathbb{P}(x) = \mathcal{N}(z|0, I) \cdot |\det(S)|^{-1}, \tag{4}$$

and the KL divergence between $\mathbb{P}(x)$ and $\mu(x)$ is

$$\begin{aligned} J &= \mathrm{KL}(\mathbb{P}(x)||\mu(x)) \\ &= \mathbb{E}_{z\sim\mathcal{N}(0,I)}[\log \mathbb{P}(x(z)) + \log U(x(z)) + \log C] \qquad (5) \\ &= \mathbb{E}_{z\sim\mathcal{N}(0,I)}[\log U(x(z))] - \log|\det(S)| \\ &\quad -\mathbb{H}(z) + \log C, \qquad (6) \end{aligned}$$

where $\mathbb{H}(z)$ denotes the entropy of z.

Remark 1. The information entropy of x is

$$\mathbb{H}(x) = \log|\det(S)| + \mathbb{H}(z). \tag{7}$$

Since $\mathbb{H}(z)$ is a constant, it is necessary to obtain a suitable S via learning.

Based on the above analysis, we can get the following stochastic gradient algorithm for learning deep energy models:

1. Draw B latent variables $z_1, \ldots, z_B$ from the standard Gaussian distribution.

2. Calculate
$$\hat{J} = -\log|\det(S)| + \frac{1}{B}\sum_b \log U(x_b)$$
with $x_b = S \cdot F(z_b)$.

3. Update all parameters w in S and F as $w \leftarrow w - \eta\frac{\partial \hat{J}}{\partial w}$, where η is the step size.

4. Repeat Steps 1-3 until convergence of w.

Learning with redundant latent variables

We now consider the case where $z \in \mathbb{R}^D$ and $S \in \mathbb{R}^{D\times D}$ with $D > d$, i.e., the dimension of z is larger than x. In this case, the NICE model becomes

$$\begin{pmatrix} x \\ y \end{pmatrix} = S \cdot F(z), \tag{8}$$

where $y \in \mathbb{R}^{D-d}$. If the conditional distribution $\mathbb{P}(y|x)$ in the NICE model is known, we can also calculate the KL divergence between $\mathbb{P}(x)$ and $\mu(x)$ according to

$$\begin{aligned} \mathrm{KL}(\mathbb{P}(x)||\mu(x)) &= \mathrm{KL}(\mathbb{P}(x)\cdot\mathbb{P}(y|x)||\mu(x)\cdot\mathbb{P}(y|x)) \\ &= \mathrm{KL}(\mathbb{P}(x,y)||\mu(x)\cdot\mathbb{P}(y|x)). \end{aligned} \tag{9}$$

However, the computation of $\mathbb{P}(y|x)$ is intractable in practice. So we approximate the conditional distribution by a Gaussian distribution

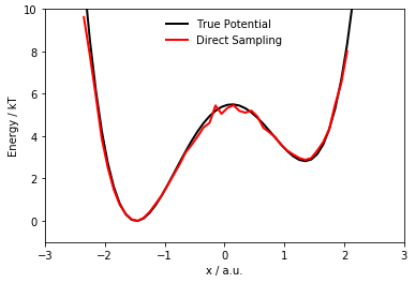
$$\mathbb{P}(y|x) \approx \hat{\mathbb{P}}(y|x) = \mathcal{N}(y|M(x), \Sigma(x)), \tag{10}$$

where $M(x), \Sigma(x)$ are also expressed as deep networks, and all parameters of S, F, M, Σ are optimized by minimizing

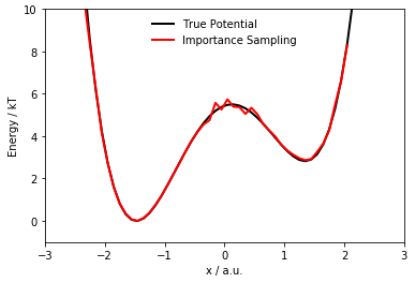
$$\begin{aligned} J_R &= \mathrm{KL}(\mathbb{P}(x,y)||\mu(x)\cdot\hat{\mathbb{P}}(y|x)) \\ &= \mathbb{E}_{z\sim\mathcal{N}(0,I)}[\log U(x)] - \log|\det(S)| \\ &\quad + \frac{1}{2}\mathbb{E}_{z\sim\mathcal{N}(0,I)}[\log\det(\Sigma(x))] \\ &\quad + \frac{1}{2}\mathbb{E}_{z\sim\mathcal{N}(0,I)}[(y-M(x))^\top\Sigma(x)^{-1}(y-M(x))] \\ &\quad - \mathbb{H}(z) + \log C. \end{aligned} \tag{11}$$

It can be shown that $J \leq J_R$, and the minimal value of J_R is achieved if $\mathbb{P}(x) = \mu(x)$ and $\mathbb{P}(y|x) = \hat{\mathbb{P}}(y|x)$.

Therefore, the learning algorithm with redundant latent variables is:



(a) Estimated by direct sampling from $\mathbb{P}(x)$



(b) Estimated by the importance sampling with the proposal $\mathbb{P}(x)$

Figure 1: Potential energy U, where the black line represents the true value and the red line is the estimate.

1. Draw B latent variables $z_1, \ldots, z_B$ from the standard Gaussian distribution.
2. Calculate

$$\hat{J}_R = -\log|\det(S)| + \frac{1}{B}\sum_b \left(\log U(x_b) + \frac{1}{2}\log\det(\Sigma(x_b)) + \frac{1}{2}(y_b - M(x_b))^\top \Sigma(x_b)^{-1}(y_b - M(x_b))\right)$$

   with $(x_b, y_b) = S \cdot F(z_b)$.
3. Update all parameters w as $w \leftarrow w - \eta\frac{\partial \hat{J}}{\partial w}$, where η is the step size.
4. Repeat Steps 1-3 until convergence of w.

Remark 2. It is unclear what is a good choice of the dimension D of z. Considering J_R provides an upper bound of J, we can choose D with the minimal J_R in practice.

Example

We consider a one-dimensional example, where

$$U(x) = x^4 - 4x^2 + x, \tag{12}$$

and $z \in \mathbb{R}^2$. The learning results are shown in Fig. 1.