Deep learning of Boltzmann distributions with parallel tempering

September 30, 2018

1 Problem

For a Boltzmann distribution

$$\mathbb{P}(x) = \frac{1}{Z} \exp\left(-U(x)\right), \tag{1}$$

we can model it by a NICE network as

$$x = SF(z) \text{ or } F(Sz) \tag{2}$$

where

$$z \sim \mathcal{N}(z|0, I) \tag{3}$$

is the latent variable distributed according to the standard Gaussian distribution, S is a diagonal matrix, and F is a NICE network with $\frac{\partial F(z)}{\partial z} \equiv I$.

Then the KL divergence between the distribution $\hat{\mathbb{P}}$ defined by (2) and $\mathbb{P}$ can be expressed as

$$\mathrm{KL}(\hat{\mathbb{P}}||\mathbb{P}) = \mathbb{E}_{z\sim N(z|0,I)}\left[U(x)\right] - \log|\det(S)| - \mathbb{E}_{z\sim N(z|0,I)}\left[\log \mathcal{N}(z|0,I)\right] + \log Z, \tag{4}$$

and F, S can be optimized by minimizing $\mathrm{KL}(\hat{\mathbb{P}}||\mathbb{P})$. However, according to our experience, this method possibly ignores shallow potential wells in the landscape of U.

2 Idea

Let

$$\mathbb{P}_T(x) = \frac{1}{Z} \exp\left(-\frac{U(x)}{T}\right) \tag{5}$$

be the Boltzmann distribution with temperature T and $\hat{\mathbb{P}}_T$ be the distribution of model (2) with

$$z \sim \mathcal{N}(z|0, T \cdot I). \tag{6}$$

We can conclude that

$$\hat{\mathbb{P}} = \mathbb{P} \Rightarrow \hat{\mathbb{P}}_T = \mathbb{P}_T \tag{7}$$

Therefore, if $\hat{\mathbb{P}}$ is an accurate approximation of $\mathbb{P}$, we can also get the approximate distributions at different termperatures by simply changing the variance of the latent variable. When the temperature T is high, the energy landscape is more flat and can be more efficiently explored with a large variance of z.

Based on the above analysis, we can obtain such a method: Learn the model (2) by minimizing the KL divergence at different termperatures, i.e.,

$$\min_{F,S} J = \int \rho(T)\mathrm{KL}(\hat{\mathbb{P}}_T||\mathbb{P}_T)\mathrm{d}T, \tag{8}$$

where $\rho(T)$ represents the weight of temperature T and is assumed to be a probability density function.

This method is close to parallel tempering methods in MCMC and MD, where the approximation performance can be improved by combining simulations at both low and high temperatures. Unlike REMD and related methods, here T can be continuous-valued in $(0, +\infty)$ and independently sampled.

3 Algorithm

After a few steps, we have

$$\begin{aligned} J &= -\log|\det(S)| + \iint T^{-1}\rho(T)\mathcal{N}(z|0,T)U(x(z))\mathrm{d}z\mathrm{d}T + \mathrm{const} \\ &= -\log|\det(S)| + C \cdot \mathbb{E}_{T\sim\rho'(T), z|T\sim\mathcal{N}(z|0,T)}\left[U(x(z))\right] + \mathrm{const} \end{aligned} \tag{9}$$

where

$$C = \int T^{-1}\rho(T)\mathrm{d}T \tag{10}$$

and

$$\rho'(T) = C^{-1} \cdot T^{-1}\rho(T) \tag{11}$$

is a new distribution of the temperature.

The choice of $\rho(T)$ requires further investigations. An example is

$$\rho(T) \propto T^{-1} \cdot 1_{T_L \le T \le T_U}, \tag{12}$$

where T_L, T_U are the bounds.

We can then get the following algorithm:

1. Sample $\{(T_i, z_i)\}_{i=1}^B$ according to

$$\begin{aligned} T &\sim \rho'(T), \\ z|T &\sim \mathcal{N}(z|0,T), \end{aligned}$$

   where

$$\rho'(T) \propto T^{-1}\rho(T).$$

2. Let

$$J = -\log|\det(S)| + \frac{C}{B}\sum_i U(x(z_i))$$

3. Update all parameters W as

$$W \leftarrow W + \eta\frac{\partial J}{\partial W}$$