Learning Deep Generative Models with Annealed Importance Sampling

Xinqiang Ding[1] and David J. Freedman[1,2]

[1] Department of Neurobiology, The University of Chicago, [2]The Grossman Institute for Neuroscience, Quantitative Biology and Human Behavior, The University of Chicago

Learning Deep Generative Models via Maximizing Likelihood Requires Computing the Gradient of the Log-Likelihood

Given a generative model defined by a joint distribution of observed data variables ***x*** and continuous latent variable ***z***:

$$p_\theta(x, z) = p_\theta(z)p_\theta(x|z)$$

With training data x, we are interested in learning θ by maximizing its marginal likelihood:

$$\log p_\theta(x) = \log \int p_\theta(x, z)\mathrm{d}z$$

Computing log $p_\theta(x)$ is expensive because it involves a high dimensional integral. However, we note that maximizing log $p_\theta(x)$ does not necessarily require computing the value of log $p_\theta(x)$. What is necessarily required is the gradient of log $p_\theta(x)$ with respect to θ:

$$\nabla_\theta \log p_\theta(x) = \mathop{\mathbb{E}}_{z\sim p_\theta(z|x)}\left[\nabla_\theta \log p_\theta(x, z)\right] \qquad (1)$$

which can be written as an expectation.

Several Previous Methods Used for Training Generative Models Can Be Viewed as Devising Ways of Approximating the Expectation in (1)

- **Variational Auto-Encoder (VAE)**
  In maximizing the ELBO:
  $$\mathcal{L}(\theta, \phi) = \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x, z) - \log q_\phi(z|x)\right]$$
  θ follows the gradient:
  $$\nabla_\theta\mathcal{L}(\theta, \phi) = \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)}\left[\nabla_\theta \log p_\theta(x, z)\right]$$
  which is an approximation to the expectation in (1).

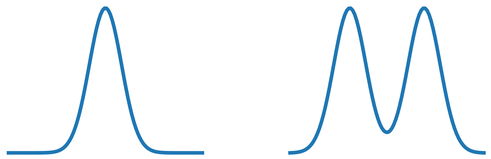
- **Importance Weighted Auto-Encoder (IWAE)**
  $$\mathcal{L}_K(\theta, \phi) = \mathop{\mathbb{E}}_{z_1,\ldots,z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x, z_k)}{q_\phi(z_k|x)}\right]$$
  $$\nabla_\theta\mathcal{L}_K(\theta, \phi) = \sum_{k=1}^{K}\widetilde{w_k}\nabla_\theta \log p_\theta(x, z_k) \qquad (2)$$
  $$\widetilde{w_k} = w_k/\sum_{i=1}^{K} w_i \qquad w_k = w(x, z_k, \theta, \phi) = \frac{p_\theta(x, z_k)}{q_\phi(z_k|x)}$$

IWAE uses (2) as an approximation of (1) via importance sampling.

- **Markov Chain Monte Carlo (MCMC) Methods**
  Directly estimate the expectation in (1) by drawing samples from the posterior distribution $p_\theta(z|x)$ with MCMC. It takes a large number of steps for a Markov chain to converge, so it is much slower than variational methods.

Estimate the Gradient $\nabla_\theta \log p_\theta(x)$, which is an Expectation in (1), Using Annealed Importance Sampling/Jarzynski Equality

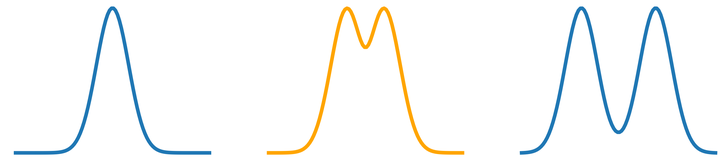
Importance Sampling



$q(z|x)$ $\quad$ $p_\theta(z|x)$

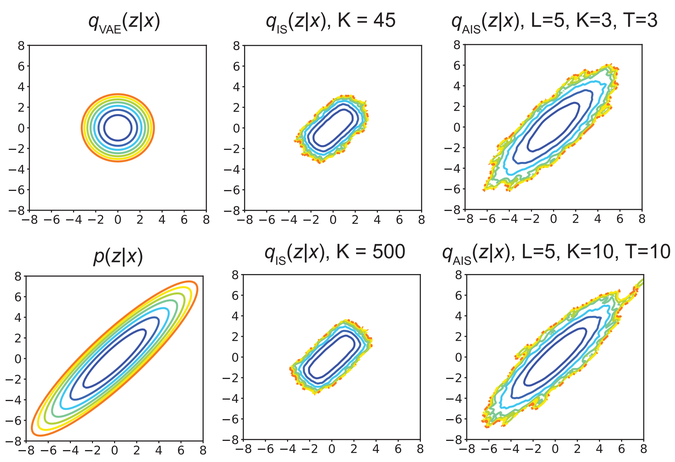
sample $z_1, \ldots, z_K \sim q_\phi(z|x)$
set $w_k = \frac{p_\theta(x,z_k)}{q_\phi(z_k|x)}$ and $\widetilde{w_k} = w_k/\sum_{k=1}^{K} w_k$

Annealed Importance Sampling



$q(z|x)$ $\quad$ $q(z|x)^{1-\beta}p_\theta(z|x)^\beta$ $\quad$ $p_\theta(z|x)$

sample $z_1^0, \ldots, z_K^0 \sim q_\phi(z|x)$
sample $z_1^1, \ldots, z_K^1$ from $q(z|x)^{1-\beta}p_\theta(z|x)^\beta$ by running L steps of Hamiltonian Monte Carlo (HMC) with the potential energy of $U^\beta(z) = -\log q(z|x)^{1-\beta}p_\theta(z|x)^\beta$
set $w_k = \frac{q(z_k^0|x)^{1-\beta}p_\theta(z_k^0|x)^\beta}{q(z_k^0|x)} \cdot \frac{p_\theta(z_k^1|x)}{q(z_k^1|x)^{1-\beta}p_\theta(z_k^1|x)^\beta}$



Algorithm : Sampling $q_{\mathrm{IS}}(z|x)$ or $q_{\mathrm{AIS}}(z|x)$

K: number of importance samples
L: number of integration steps in HMC
T: num of intermediate distributions
case 1: *when importance sampling is used*
  sample $z_1, \ldots, z_K \sim q_\phi(z|x)$
  set $w_k = \frac{p_\theta(x,z_k)}{q_\phi(z_k|x)}$ and $\widetilde{w_k} = w_k/\sum_{k=1}^{K} w_k$
case 2: *when AIS is used*
  sample $z_1, \ldots, z_K$ and compute $\widetilde{w_1}, \ldots, \widetilde{w_K}$ with Algorithm 1
sample $j \sim$ *Categorical*$(\widetilde{w_1}, \ldots, \widetilde{w_K})$
return z_j

Learn Deep Generative Models with Annealed Importance Sampling

Results of IWAE-DReG and our approach on the Omniglot and the MNIST dataset

$\approx \log p(x)$	Omniglot			MNIST		
	K = 1	K=5	K=50	K = 1	K=5	K=50
IWAE-DReG	-109.41	-106.11	-103.91	-86.90	-85.52	-84.38
Ours (T = 5)	-103.22	-102.47	-102.03	-84.56	-84.25	-83.93
Ours (T = 11)	-102.45	-101.94	-101.64	-84.14	-83.78	-83.63

Results of IWAE-DReG, MH-HMC and our approach on the Omniglot dataset with same computational cost.

	$\approx \log p(x)$
MH-HMC (L = 5, K = 5, T = 5)	-102.57
Ours (L = 5, K = 5, T = 5)	-102.47
MH-HMC (L = 5, K = 5, T = 11)	-101.32
Ours (L = 5, K = 5, T = 11)	-101.94
MH-HMC (L = 5, K = 50, T = 5)	-102.32
Ours (L = 5, K = 50, T = 5)	-102.03
MH-HMC (L = 5, K = 50, T = 11)	-101.25
Ours (L = 5, K = 50, T = 11)	-101.64

	$\approx \log p(x)$
IWAE-DReG (K = 55)	-103.85
Ours (L = 5, K = 1, T = 11)	-102.45
IWAE-DReG (K = 275)	-103.13
Ours (L = 5, K = 5, T = 11)	-101.94
IWAE-DReG (K = 2750)	-102.40
Ours (L = 5, K = 50, T = 11)	-101.64