Divergence Triangle for Joint Training of Generator Model, Energy-based Model, and Inferential Model

Tian Han*[1], Erik Nijkamp*[1], Xiaolin Fang[2], Mitch Hill[1], Song-Chun Zhu[1] and Ying Nian Wu[1]

[1]University of California, Los Angeles, [2]Zhejiang University

Objectives

Learning two deep probabilistic models in a unified framework.

- The Generator Model
- The Energy-based Model

Generator Model

Top-down mapping
hidden vector z
$\Downarrow$
signal $x \approx g_\theta(z)$
(a) Generator model

$$z \sim \mathrm{N}(0, I_d),\ x = g_\theta(z) + \epsilon,$$

- Latent variable model. $p(z)$, $p_\theta(x|z)$
- g_θ: top-down ConvNet.
- $p_\theta(x) = \int_z p(z)p_\theta(x|z)dz$.

Maximum likelihood learning of generator model:

$$-\frac{\partial}{\partial\theta}\mathrm{KL}(q_{\text{data}}(x)\|p_\theta(x)) = \mathrm{E}_{q_{\text{data}}(x)p_\theta(z|x)}\left[\frac{\partial}{\partial\theta}\log p_\theta(z,x)\right].$$

- Intractable: posterior distribution $p_\theta(z|x)$ intractable $\rightarrow$ MCMC approximation to infer z.
- EM-type learning: iteratively impute the missing latent z, then use imputed z to update the generator.

Energy-based Model

Bottom-up mapping
energy $-f_\alpha(x)$
$\Uparrow$
signal x
(b) Energy-based model

$$\pi_\alpha(x) = \frac{1}{Z(\alpha)}\exp\left[f_\alpha(x)\right],$$

- $Z(\alpha)$: normalizing constant.
- f_α: bottom-up ConvNet.
- Gibbs distribution, FRAME model.

Maximum likelihood learning of energy-based model:

$$-\frac{\partial}{\partial\alpha}\mathrm{KL}(q_{\text{data}}(x)\|\pi_\alpha(x)) = \mathrm{E}_{q_{\text{data}}}\left[\frac{\partial}{\partial\alpha}f_\alpha(x)\right] - \mathrm{E}_{\pi_\alpha}\left[\frac{\partial}{\partial\alpha}f_\alpha(x)\right].$$

- Intractable: model distribution $\pi_\alpha(x)$ intractable $\rightarrow$ MCMC approximation to sample x.
- Self-critic learning: model π_α gets samples from its current version, then criticize such samples, i.e., the model is its own adversary or its own critic.

Divergence Triangle

Define three joint distributions on (z, x):

- Q-distribution: $Q(z, x) = q_{\text{data}}(x)q_\phi(z|x)$.
- P-distribution: $P(z, x) = p(z)p_\theta(x|z)$.
- Π-distribution: $\Pi(z, x) = \pi_\alpha(x)q_\phi(z|x)$.

Divergence Triangle Functional:

$$\max_\alpha \min_\theta \min_\phi \mathcal{D}(\alpha, \theta, \phi),$$
$$\mathcal{D} = \mathrm{KL}(Q\|P) + \mathrm{KL}(P\|\Pi) - \mathrm{KL}(Q\|\Pi).$$

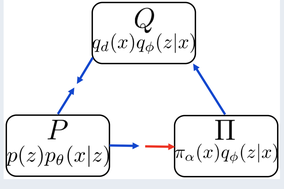

Figure 1:Divergence triangle is based on the Kullback-Leibler divergences between three joint distributions of (z, x). The blue arrow indicates the "running toward" behavior and the the red arrow indicates the "running away" behavior.

Unpacking Loss

$\max_\alpha \min_\theta \min_\phi \mathrm{KL}(Q\|P) + \mathrm{KL}(P\|\Pi) - \mathrm{KL}(Q\|\Pi)$

1. Variational Learning:

$$\mathrm{KL}(Q\|P) = \underbrace{\mathrm{KL}(q_{\text{data}}(x)\|p_\theta(x))}_{\text{MLE for Generator}} + \mathrm{KL}(q_\phi(z|x)\|p_\theta(z|x)),$$

$q_\phi(z|x) \rightarrow$ (Intractable) $p_\theta(z|x) \Longrightarrow$ No MCMC for sampling $p_\theta(z|x)$
Variational Auto-encoder.

2. Adversarial Learning:

$$S = \underbrace{\mathrm{KL}(q_{\text{data}}(x)\|\pi_\alpha(x))}_{\text{MLE for energy-based model}} - \mathrm{KL}(p_\theta(x)\|\pi_\alpha(x))$$

$p_\theta(x) \rightarrow$ (Intractable) $\pi_\alpha(x) \Longrightarrow$ No MCMC for sampling $\pi_\alpha(x) \Longrightarrow \pi_\alpha(x)$ is learned tractable.
Further, (Intractable) $p_\theta(x) \rightarrow q_\phi(z|x)$ variational approximation:

$$\mathrm{KL}(P\|\Pi) = \mathrm{KL}(p_\theta(x)\|\pi_\alpha(x)) + \mathrm{KL}(p_\theta(z|x)\|q_\phi(z|x)).$$

$\min_\alpha \max_\theta S \Longleftrightarrow \max_\alpha \min_\theta \min_\phi \mathrm{KL}(P\|\Pi) - \mathrm{KL}(Q\|\Pi)$

Advantages:

Jointly and tractably learn both models without MCMC steps.
Integrates maximum likelihood learning, variational learning, adversarial learning etc.

Generation

Figure 2:Generated samples. Clockwise: 32×32 ImageNet, CIFAR10, 64×64 LSUN (bedroom), CelebA

High Resolution Image

Progressively training for $1,024 \times 1,024$ image.

Figure 3:Top two: generated samples. Bottom two: linear interpolation.

Reconstruction

Figure 4:Test image reconstruction for CIFAR-10. Left: test images. Right: reconstructed images.

Energy-Landscape Mapping

Evaluate the learned energy-based model by mapping the structure of the energy landscape.

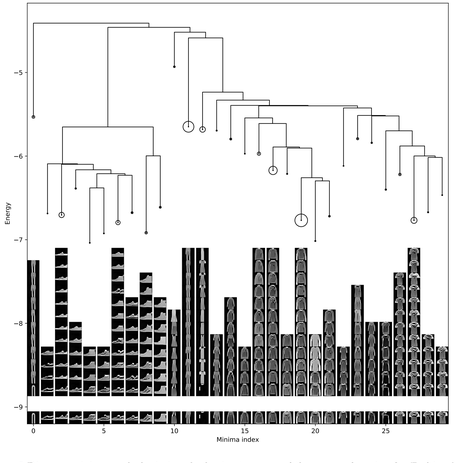

Figure 5:Disconnectivity-graph depicting the basin structure of the energy function for Fashion-MNIST.