A very quick introduction to

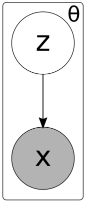
Auto-encoding variational Bayes and the reparameterization trick

J. Bill

February 12, 2019

The problem with inference in and learning of generative models

Consider a generative model:



$p^*(x)$

We want to perform

- Inference: $p_\theta(z \,|\, x) \propto p_\theta(x \,|\, z) \cdot p_\theta(z)$
- Learning: $\Delta\theta = \eta \cdot \nabla_\theta \, \mathbb{E}_{p^*(x)} \left[\log p_\theta(x) \right]$

Problem:

For expressive model classes $p_\theta(x, z)$, typically either

- inference is intractable (e.g., neural nets), or
- learning becomes computationally expensive (e.g., Boltzmann machines).

Idea: Separate inference model and learning model

Define a variational posterior,

$$q_\phi(z \mid x) \approx p_\theta(z \mid x) \ ,$$

that makes inference tractable while keep learning the model,

$$\nabla_\theta \log p_\theta(x) \ ,$$

easy.

We have to keep $q_\phi(z \mid x)$ and $p_\theta(z \mid x)$ "consistent" with each other while learning their separate parameters ϕ and θ.

This will be formalized in the objective function.

The evidence lower bound (ELBO) of the data log-likelihood

Three identical ways of writing the ELBO that grant different insights:

$$\mathcal{L}(\theta, \phi; x) := \log p_\theta(x) - \mathrm{D}_{\mathrm{KL}}\left[\, q_\phi(z \,|\, x) || p_\theta(z \,|\, x) \,\right] \tag{1}$$

$\rightarrow \nabla_\phi$: Keep the variational posterior close the correct one.

$$= \mathbb{E}_{q_\phi(z \,|\, x)}\left[\, \log p_\theta(x, z) \,\right] + \mathbb{H}\left[\, q_\phi(z \,|\, x) \,\right] \tag{2}$$

$\rightarrow \nabla_\theta$: Improve the generative model (alternate with (1) to $\uparrow \log p_\theta(x)$)

$$= \mathbb{E}_{q_\phi(z \,|\, x)}\left[\, \log p_\theta(x \,|\, z) \,\right] - \mathrm{D}_{\mathrm{KL}}\left[\, q_\phi(z \,|\, x) || p_\theta(z) \,\right] \tag{3}$$

$\rightarrow$ Auto-encode the input while keeping the avg. posterior close to the prior.

The decomposition holds for any choice of $q_\phi(z \,|\, x)$!

We can choose $p_\theta(x, z)$ and $q_\phi(z \,|\, x)$ such that the gradients $\nabla_\theta\, p_\theta(x, z)$ and $\nabla_\phi\, q_\phi(z \,|\, x)$ are easy to compute while evading the complex posterior of $p_\theta(x, z)$.

A technical problem: High variance gradient estimates

Gradient ascent on the ELBO always involves terms

$$\nabla_\phi \, \mathbb{E}_{q_\phi(z \mid x)} \left[f(z) \right] \quad .$$

Yet, we can bring the gradient into the form,

$$\mathbb{E}_{q_\phi(z \mid x)} \left[f(z) \cdot \nabla_\phi \log q_\phi(z \mid x) \right] \quad .$$

to make it accessible to sample-based estimation. Hooray!

Hooray..?

It turns out that low probability samples from $q_\phi(z \mid x)$ can have a strong contribution via $\nabla_\phi \log q_\phi(z \mid x)$. This makes the total gradient estimate very "noisy"!

To noisy, actually, for practical purposes ☹

Solution: The reparameterization trick

Choose $q_\phi(z \mid x)$ such that

$$z = g_\phi(x, \epsilon) \ \text{ with } \ \epsilon \sim p(\epsilon) \quad \leftarrow \text{simple distribution.}$$

Then:

$$\nabla_\phi \, \mathbb{E}_{q_\phi(z \mid x)} \left[f(z) \right] = \nabla_\phi \, \mathbb{E}_{p(\epsilon)} \left[f(g_\phi(x, \epsilon)) \right] = \mathbb{E}_{p(\epsilon)} \left[\nabla_\phi \, f(g_\phi(x, \epsilon)) \right].$$

Now, $\mathbb{E}_{p(\epsilon)} [\, \cdot \,]$ is easy to sample from while covering the $q_\phi(z \mid x)$ space.
And $\nabla_\phi \, f(g_\phi(x, \epsilon))$ is a "backprop-style" chain rule.

It is found, that this estimator has much lower variance, i.e., is less "noisy".

AEVB in practice

In practice, we consider the form (3) of the ELBO for learning:

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z \,|\, x)} [\log p_\theta(x \,|\, z)] - \mathrm{D}_{\mathrm{KL}} [q_\phi(z \,|\, x) || p_\theta(z)]$$

We choose to fix the prior, $p_\theta(z) \equiv p(z)$, thereby turning it into a target distribution for $q_\phi(z \,|\, x)$. Further, we can often choose $p(z)$ and $q_\phi(z \,|\, x)$ such that

$$\nabla_\phi \, \mathrm{D}_{\mathrm{KL}} [q_\phi(z \,|\, x) || p(z)]$$

can be calculated analytically ($\rightarrow$ no sampling error).
Then, we only apply the reparameterization trick to $\mathbb{E}_{q_\phi(z \,|\, x)} [\log p_\theta(x \,|\, z)]$:

Sample-based part of the gradient estimation

$$\mathbb{E}_{p(\epsilon)} \left[\nabla_{\theta/\phi} \, p_\theta(x \,|\, g_\phi(x, \epsilon)) \right]$$