Deep Generative Models

Chapter 5: Variational Inference and VAEs

Ali Bereyhi

ali.bereyhi@utoronto.ca

Department of Electrical and Computer Engineering

University of Toronto

Summer 2025

Bayesian Setting: *Evidence*

Let's rewrite the challenge in a Bayesian setting: *we know the prior latent distribution $P(z)$ and generator $P(x|z)$ looking for the so-called evidence*

Evidence

In Bayesian formulation, the distribution of a known sample x marginalized over the latent is called evidence

$$P(x) = \int P(x|z) P(z) \, \mathrm{d}z$$

+ *Isn't this evidence simply the likelihood?!*
– *Sure! When it is computed with a learnable model $P_{\mathbf{w}}(x|z)$, it computes the likelihood of the model*

Computing Evidence: *Direct Approach*

Complexity of direct evidence computation is exponential in latent size

To see this consider a *discrete latent* $z \in \mathbb{Z}^m$ *with* $\mathbb{Z}$ *having* C *elements: the evidence in this case is computed as*

$$P(x) = \sum_z P(x|z) P(z)$$

which requires C^m *additions!*

Moral of Story

Direct computation of the evidence is not tractable even though we know both prior $P(z)$ *and generator* $P(x|z)$ $\equiv$ *direct sampling from a general distribution*

Computing Evidence: *Monte Carlo*

An alternative approach is to *estimate evidence via Monte Carlo: we note that*

$$P(x) = \int P(x|z) P(z) \, \mathrm{d}z = \mathbb{E}_{z \sim P(z)} \{ P(x|z) \}$$

We can hence estimate the evidence as

$$\hat{P}(x) = \hat{\mathbb{E}}_{z \sim P(z)} \{ P(x|z) \} = \frac{1}{n} \sum_i P(x|z_i)$$

for samples $z_i \sim P(z) \rightsquigarrow$ this is tractable as we have $P(z)$ and $P(x|z)$, but

↳ $P(x|z)$ *typically peaks close to x and is very small other points*

↳ *Only few samples $z_i \sim P(z)$ would be useful $P(x|z_i)$*

Moral of Story

MC estimate is very high variance and needs a huge set of samples, i.e., huge n

Computing of Evidence by *Importance Sampling*

There is an intuitive remedy to reduce variance of MC estimate: *assume we know $Q(z|x)$ whose samples are good for the given x; then,*

we may use samples $z_i \sim Q(z|x)$ to estimate evidence

+ *But in marginalization, we average over $P(z)$ not some $Q(z|x)$! Right?!*
– *Right! But we can do a simple modification!*

$$
\begin{aligned}
P(x) = \int P(x|z) P(z) \, \mathrm{d}z &= \int P(x|z) \frac{P(z)}{Q(z|x)} Q(z|x) \mathrm{d}z \\
&= \mathbb{E}_{z \sim Q(z|x)} \left\{ P(x|z) \frac{P(z)}{Q(z|x)} \right\}
\end{aligned}
$$

We can now use samples of $Q(z|x)$ to estimate evidence $P(x)$

Importance Sampling

Importance Sampling

We can estimate the evidence as

$$P(x) = \hat{\mathbb{E}}_{z \sim Q(z|x)} \left\{ P(x|z) \frac{P(z)}{Q(z|x)} \right\} = \frac{1}{n} \sum_i P(x|z_i) \frac{P(z_i)}{Q(z_i|x)}$$

with z_i sampled from our good distribution $z_i \sim Q(z|x)$

The core idea is that *samples $z_i \sim Q(z|x)$ are more important*

↳ *$z_i \sim Q(z|x)$ are in the region where $P(x|z_i)$ is rather large*

↳ *Samples $z_i \sim Q(z|x)$ result in useful samples $P(x|z_i)$*

Thus, the evidence estimator has less variance

Importance Sampling: *Good Distribution for Sampling*

+ *But, what is this good distribution $Q(z|x)$?!*
- Ideally, the posterior $P(z|x)$!

If we set $Q(z|x) = P(z|x)$; then, each sample reads

$$P(x|z_i)\frac{P(z_i)}{P(z_i|x)} = \frac{P(x, z_i)}{P(z_i|x)} = P(x)$$

So, only single sample is enough to estimate evidence with zero variance!

Good Distribution for Importance Sampling

A good choice of $Q(z|x)$ is the one that is close to posterior $P(z|x)$

Importance Sampling: *Learning Objective*

! Attention

Note that the posterior is given as

$$P(z|x) = \frac{P(x|z)P(z)}{\boxed{P(x)}} \xrightarrow{} \text{evidence}$$

whose computation is as complex as computing evidence

+ *How should we find a good $Q(z|x)$ at the end?!*
– We try to learn it by approximating $P(z|x)$!
+ *But how can we do this?! We don't know $P(z|x)$ in the first place!*
– We do it by an implicit approach $\rightsquigarrow$ *variational inference*

Variational Inference: *Right Distribution*

Let's say we have a class of distributions Q defined on latent space[1]: *we can learn the best choice of Q for importance sampling as*

$$Q^{\star}_{|x} = \operatorname*{argmin}_{\boxed{Q_{|x}} \longrightarrow Q(\cdot|x)} D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right)$$

The key challenge is though that we don't know $P(\cdot|x)$

Let's expand this divergence a bit

$$D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right) = \mathbb{E}_{z\sim Q_{|x}}\left\{\log \frac{Q(z|x)}{P(z|x)}\right\} = \mathbb{E}_{z\sim Q_{|x}}\left\{\log \frac{Q(z|x)P(x)}{P(z|x)\,P(x)}\right\}$$

$$= \mathbb{E}_{z\sim Q_{|x}}\left\{\log \frac{Q(z|x)P(x)}{P(x,z)}\right\}$$

[1] *We later learn Q by a computational model*

Variational Inference: *Right Distribution*

We can now use the chain rule to write

$$
\begin{aligned}
D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right) &= \mathbb{E}_{z\sim Q_{|x}}\left\{\log \frac{Q\left(z|x\right) P\left(x\right)}{P\left(x,z\right)}\right\} \\
&= \mathbb{E}_{z\sim Q_{|x}}\left\{\log \frac{Q\left(z|x\right) P\left(x\right)}{P\left(x|z\right) P\left(z\right)}\right\} \\
&= \mathbb{E}_{z\sim Q_{|x}}\left\{\log \frac{Q\left(z|x\right)}{P\left(z\right)} + \log \frac{1}{P\left(x|z\right)} + \log P\left(x\right)\right\}
\end{aligned}
$$

Let's look at each expression inside the expectation individually

Variational Inference: *Right Distribution*

The fist term is a KL divergence

$$\mathbb{E}_{z\sim Q_{|x}}\left\{\log \frac{Q\left(z|x\right)}{P\left(z\right)}\right\} = D_{\mathrm{KL}}\left(Q_{|x} \| P_z\right)$$

which we can estimate easily

---

1. *Collect samples* $z_i \sim Q\left(z|x\right)$
2. *Compute the distribution at each sample, i.e.,* $Q\left(z_i|x\right)$
3. *Compute latent distribution at each sample as well, i.e.,* $P\left(z_i\right)$

The estimate of this term is then given by

$$\hat{D}_{\mathrm{KL}}\left(Q_{|x} \| P_z\right) = \hat{\mathbb{E}}_{z\sim Q_{|x}}\left\{\log \frac{Q\left(z|x\right)}{P\left(z\right)}\right\} = \frac{1}{n}\sum_i \log \frac{Q\left(z_i|x\right)}{P\left(z_i\right)}$$

Variational Inference: *Right Distribution*

The second term is

$$\mathbb{E}_{z \sim Q_{|x}} \left\{ \log \frac{1}{P(x|z)} \right\} = -\mathbb{E}_{z \sim Q_{|x}} \left\{ \log P(x|z) \right\}$$

which we can again easily estimate

---

1. *Collect samples as* $z_i \sim Q(z|x)$
2. *Compute generator at each sample*, i.e., $P(z_i|x)$

We then estimate this term as

$$-\hat{\mathbb{E}}_{z \sim Q_{|x}} \left\{ \log P(x|z) \right\} = -\frac{1}{n} \sum_i \log P(x|z_i)$$

Variational Inference: *Right Distribution*

The third term is indeed the log-evidence

$$\mathbb{E}_{z\sim Q_{|x}}\left\{\log P\left(x\right)\right\} = \int \log P\left(x\right) Q\left(z|x\right) \mathrm{d}z$$
$$= \log P\left(x\right) \underbrace{\int Q\left(z|x\right) \mathrm{d}z}_{1} = \log P\left(x\right)$$

Putting all three terms together

$$D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right) = D_{\mathrm{KL}}\left(Q_{|x} \| P_{z}\right) - \mathbb{E}_{z\sim Q_{|x}}\left\{\log P\left(x|z\right)\right\} + \log P\left(x\right)$$

This offers a tractable way for estimating the right distribution!

ELBO: *Evidence Lower Bound*

Recall that *we were looking for*

$$Q^{\star}_{|x} = \underset{Q_{|x}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right)$$

We can use the expansion to write

$$\begin{aligned} Q^{\star}_{|x} &= \underset{Q_{|x}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\left(Q_{|x} \| P_{z}\right) - \mathbb{E}_{z\sim Q_{|x}}\left\{\log P\left(x|z\right)\right\} + \log P\left(x\right) \\ &= \underset{Q_{|x}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\left(Q_{|x} \| P_{z}\right) - \mathbb{E}_{z\sim Q_{|x}}\left\{\log P\left(x|z\right)\right\} \\ &= \underset{Q_{|x}}{\operatorname{argmax}} \boxed{\mathbb{E}_{z\sim Q_{|x}}\left\{\log P\left(x|z\right)\right\} - D_{\mathrm{KL}}\left(Q_{|x} \| P_{z}\right)} \end{aligned}$$

The objective on this maximization can be estimated $\rightsquigarrow$ $Q_{|x}$ *can be learned!*

ELBO: *Evidence Lower Bound*

ELBO: *Evidence Lower Bound*

For a given $Q_{|x}$, the ELBO is defined as

$$\mathrm{ELBO}\left(Q_{|x}\right) = \mathbb{E}_{z \sim Q_{|x}}\left\{\log P\left(x|z\right)\right\} - D_{\mathrm{KL}}\left(Q_{|x} \| P_z\right)$$

which can be estimated by sampling from $Q_{|x}$ as

$$\widehat{\mathrm{ELBO}}\left(Q_{|x}\right) = \hat{\mathbb{E}}_{z \sim Q_{|x}}\left\{\log P\left(x|z\right)\right\} - \hat{D}_{\mathrm{KL}}\left(Q_{|x} \| P_z\right)$$

The key feature of ELBO is that *it is maximized via the distribution that lies in minimum KL divergence of posterior $P\left(z|x\right)$, i.e.,*

$$\underset{Q_{|x}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right) = \underset{Q_{|x}}{\operatorname{argmax}}\, \mathrm{ELBO}\left(Q_{|x}\right)$$

ELBO Properties: *Bounding Evidence*

It is easy to see that ELBO is really a lower bound on log-evidence: *recall*

$$D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right) = \underbrace{D_{\mathrm{KL}}\left(Q_{|x} \| P_{z}\right) - \mathbb{E}_{z \sim Q_{|x}}\left\{\log P\left(x|z\right)\right\}}_{-\mathrm{ELBO}\left(Q_{|x}\right)} + \log P\left(x\right)$$

We can sort things out and write

$$\log P\left(x\right) = D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right) + \mathrm{ELBO}\left(Q_{|x}\right)$$

No matter what $Q_{|x}$ *is* $\rightsquigarrow$ *we always have* $D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right) \geqslant 0$

ELBO: *Bounding Evidence*

For any distribution $Q_{|x}$, *ELBO bounds the log-evidence from below, i.e.,*

$$\log P\left(x\right) \geqslant \mathrm{ELBO}\left(Q_{|x}\right)$$

ELBO Properties: *Optimal ELBO* $\equiv$ *Log-Likelihood*

Let's look again at the identity *and think about ideal case*

$$\log P(x) = D_{\mathrm{KL}}\left(Q_{|x} \| P_{|x}\right) + \mathrm{ELBO}\left(Q_{|x}\right)$$

if ELBO is ideally optimized $\rightsquigarrow D_{\mathrm{KL}}\left(Q^{\star}_{|x} \| P_{|x}\right) = 0$

- ↳ *This means that the optimal ELBO touches log-evidence*
- ↳ *Recall that log-evidence computes log-likelihood for us*

ELBO: *Optimal ELBO* $\equiv$ *Log-Likelihood*

Assuming true posterior $P_{|x}$ belongs to class of distributions described by $Q_{|x}$

$$\log P(x) = \max_{Q_{|x}} \mathrm{ELBO}\left(Q_{|x}\right)$$

Variational Inference: *Wrap Up*

Let us now summarize: *consider a setting in which we know*

1. *prior latent* distribution $P(z)$, *and*
2. *the conditional generator* $P(x|z)$

We want to compute the evidence $\equiv$ *likelihood* $P(x)$

Solution via *Variational Inference*

1. *Find a good estimator of posterior by maximizing the ELBO*

$$Q_{|x}^{\star} = \operatorname*{argmax}_{Q_{|x}} \mathrm{ELBO}\left(Q_{|x}\right)$$

2. *Use importance sampling to compute the evidence*

$$P(x) = \mathbb{E}_{z \sim Q_{|x}^{\star}} \left\{ P(x|z) \frac{P(z)}{Q^{\star}(z|x)} \right\}$$

Approximating Evidence Computationally

+ *How can we implement variational inference is practice?*
– We use a computational model!

We first consider a computational model for $Q_{|x} \equiv Q_{\mathbf{w}}$

```
VI(P_z:latent prior, P_{x|z}:model, x:data)
```

1: *Let $Q_{\mathbf{w}}$ be a computational model for latent distribution*
2: **for** *multiple epochs* **do**
3: *Sample a batch of latent samples* $\left\{z^j : j = 1, \dots, n\right\}$ *from* $Q_{\mathbf{w}}$
4: **for** $j = 1, \dots, n$ **do**
5: *Compute* $\mathrm{ELBO}^j = \log P\left(x | z^j\right) - \log \dfrac{Q_{\mathbf{w}}\left(z^j\right)}{P\left(z^j\right)}$
6: *Backpropagate over $Q_{\mathbf{w}}$ to compute* $\nabla_{\mathbf{w}} \mathrm{ELBO}^j$
7: **end for**
8: *Update $\mathbf{w}$ using* `Opt_avg` $\left\{\nabla_{\mathbf{w}} \mathrm{ELBO}^j\right\}$
9: **end for**
10: **return** *trained distribution $Q_{\mathbf{w}}$*

Approximating Evidence Computationally

We then use the trained model to estimate evidence by importance sampling

`VI_Evidence(`$Q_{\mathbf{w}}$`)`

1: *Sample a batch of latent samples* $\left\{z^j : j = 1, \ldots, n\right\}$ *from* $Q_{\mathbf{w}}$
2: **for** $j = 1, \ldots, n$ **do**
3:     *Compute* $\hat{P}^j = P\left(x|z^j\right) \dfrac{P\left(z^j\right)}{Q_{\mathbf{w}}\left(z^j\right)}$
4: **end for**
5: *Estimate evidence as* $\hat{P}\left(x\right) \leftarrow \mathtt{mean}\left\{\hat{P}^j\right\}$
6: **return** *Evidence estimator* $\hat{P}\left(x\right)$

Implicit MLE via Variational Inference

Variational inference provides us a tractable way to *train a*

computational probabilistic generator

Say we have a probabilistic $P_\theta(x|z)$: MLE trains this model as

$$\max_\theta \log P_\theta(x) = \max_\theta \log \int P_\theta(x|z) P(z) \, \mathrm{d}z$$

which is not tractable! Using variational inference, we can write

$$\max_\theta \log P_\theta(x) = \max_\theta \max_{\mathbf{w}} \mathrm{ELBO}(Q_{\mathbf{w}})$$

this leads to the birth of variational autoencoder (VAE) $\rightsquigarrow$ we check it next