Evidence lower bound

In variational Bayesian methods, the **evidence lower bound** (often abbreviated **ELBO**, also sometimes called the **variational lower bound**[1] or **negative variational free energy**) is a useful lower bound on the log-likelihood of some observed data.

The ELBO is useful because it provides a guarantee on the worst-case for the log-likelihood of some distribution (e.g. $p(X)$) which models a set of data. The actual log-likelihood may be higher (indicating an even better fit to the distribution) because the ELBO includes a Kullback-Leibler divergence (KL divergence) term which decreases the ELBO due to an internal part of the model being inaccurate despite good fit of the model overall. Thus improving the ELBO score indicates either improving the likelihood of the model $p(X)$ or the fit of a component internal to the model, or both, and the ELBO score makes a good loss function, e.g., for training a deep neural network to improve both the model overall and the internal component. (The internal component is $q_\phi(\cdot|x)$, defined in detail later in this article.)

Definition

Let X and Z be random variables, jointly distributed with distribution p_θ. For example, $p_\theta(X)$ is the marginal distribution of X, and $p_\theta(Z \mid X)$ is the conditional distribution of Z given X. Then, for a sample $x \sim p_{\text{data}}$, and any distribution q_ϕ, the ELBO is defined as

$$L(\phi, \theta; x) := \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\ln \frac{p_\theta(x,z)}{q_\phi(z|x)}\right].$$

The ELBO can equivalently be written as[2]

$$\begin{aligned} L(\phi, \theta; x) =& \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\ln p_\theta(x,z)\right] + H[q_\phi(z|x)] \\ =& \ln p_\theta(x) - D_{KL}(q_\phi(z|x)||p_\theta(z|x)). \end{aligned}$$

In the first line, $H[q_\phi(z|x)]$ is the entropy of q_ϕ, which relates the ELBO to the Helmholtz free energy.[3] In the second line, $\ln p_\theta(x)$ is called the *evidence* for x, and $D_{KL}(q_\phi(z|x)||p_\theta(z|x))$ is the Kullback-Leibler divergence between q_ϕ and p_θ. Since the Kullback-Leibler divergence is non-

negative, $L(\phi, \theta; x)$ forms a lower bound on the evidence (*ELBO inequality*)

$$\ln p_\theta(x) \geq \mathbb{E}_{z \sim q_\phi(\cdot|x)} \left[\ln \frac{p_\theta(x, z)}{q_\phi(z|x)} \right].$$

Motivation

Variational Bayesian inference

Suppose we have an observable random variable X, and we want to find its true distribution p^*. This would allow us to generate data by sampling, and estimate probabilities of future events. In general, it is impossible to find p^* exactly, forcing us to search for a good **approximation.**

That is, we define a sufficiently large parametric family $\{p_\theta\}_{\theta \in \Theta}$ of distributions, then solve for $\min_\theta L(p_\theta, p^*)$ for some loss function L. One possible way to solve this is by considering small variation from p_θ to $p_{\theta + \delta\theta}$, and solve for $L(p_\theta, p^*) - L(p_{\theta+\delta\theta}, p^*) = 0$. This is a problem in the calculus of variations, thus it is called the **variational method**.

Since there are not many explicitly parametrized distribution families (all the classical distribution families, such as the normal distribution, the Gumbel distribution, etc, are far too simplistic to model the true distribution), we consider *implicitly parametrized* probability distributions:

- First, define a simple distribution $p(z)$ over a latent random variable Z. Usually a normal distribution or a uniform distribution suffices.
- Next, define a family of complicated functions f_θ (such as a deep neural network) parametrized by θ.
- Finally, define a way to convert any $f_\theta(z)$ into a distribution (in general simple too, but unrelated to $p(z)$) over the observable random variable X. For example, let $f_\theta(z) = (f_1(z), f_2(z))$ have two outputs, then we can define the corresponding distribution over X to be the normal distribution $\mathcal{N}(f_1(z), e^{f_2(z)})$.

This defines a family of joint distributions p_θ over (X, Z). It is very easy to sample $(x, z) \sim p_\theta$: simply sample $z \sim p$, then compute $f_\theta(z)$, and finally sample $x \sim p_\theta(\cdot|z)$ using $f_\theta(z)$.

In other words, we have a **generative model** for both the observable and the latent. Now, we consider a distribution p_θ good, if it is a close approximation of p^*:

$$p_\theta(X) \approx p^*(X)$$

since the distribution on the right side is over X only, the distribution on the left side must marginalize the latent variable Z away.

In general, it's impossible to perform the integral $p_\theta(x) = \int p_\theta(x|z)p(z)dz$, forcing us to perform another approximation.

Since $p_\theta(x) = \frac{p_\theta(x|z)p(z)}{p_\theta(z|x)}$ (Bayes' Rule), it suffices to find a good approximation of $p_\theta(z|x)$. So define another distribution family $q_\phi(z|x)$ and use it to approximate $p_\theta(z|x)$. This is a **discriminative model** for the latent.

The entire situation is summarized in the following table:

X: observable	X, Z	Z: latent
$p^*(x) \approx p_\theta(x) \approx \frac{p_\theta(x\|z)p(z)}{q_\phi(z\|x)}$ approximable		$p(z)$, easy
	$p_\theta(x\|z)p(z)$, easy	
$p_\theta(z\|x) \approx q_\phi(z\|x)$ approximable		$p_\theta(x\|z)$, easy

In **Bayesian** language, X is the observed evidence, and Z is the latent/unobserved. The distribution p over Z is the *prior distribution* over Z, $p_\theta(x|z)$ is the likelihood function, and $p_\theta(z|x)$ is the *posterior distribution* over Z.

Given an observation x, we can *infer* what z likely gave rise to x by computing $p_\theta(z|x)$. The usual Bayesian method is to estimate the integral $p_\theta(x) = \int p_\theta(x|z)p(z)dz$, then compute by Bayes' rule $p_\theta(z|x) = \frac{p_\theta(x|z)p(z)}{p_\theta(x)}$. This is expensive to perform in general, but if we can simply find a good approximation $q_\phi(z|x) \approx p_\theta(z|x)$ for most x, z, then we can infer z from x cheaply. Thus, the search for a good q_ϕ is also called **amortized inference**.

All in all, we have found a problem of **variational Bayesian inference**.

Deriving the ELBO

A basic result in variational inference is that minimizing the Kullback–Leibler divergence (KL-divergence) is equivalent to maximizing the log-likelihood:

$$\mathbb{E}_{x\sim p^*(x)}[\ln p_\theta(x)] = -H(p^*) - D_{KL}(p^*(x)\|p_\theta(x))$$

where $H(p^*) = -\mathbb{E}_{x\sim p^*}[\ln p^*(x)]$ is the entropy of the true distribution. So if we can maximize $\mathbb{E}_{x\sim p^*(x)}[\ln p_\theta(x)]$, we can minimize $D_{KL}(p^*(x)\|p_\theta(x))$, and consequently find an accurate approximation $p_\theta \approx p^*$.

To maximize $\mathbb{E}_{x\sim p^*(x)}[\ln p_\theta(x)]$, we simply sample many $x_i \sim p^*(x)$, i.e. use importance sampling

$$N \max_\theta \mathbb{E}_{x\sim p^*(x)}[\ln p_\theta(x)] \approx \max_\theta \sum_i \ln p_\theta(x_i)$$

where N is the number of samples drawn from the true distribution. This approximation can be seen as overfitting.[note 1]

In order to maximize $\sum_i \ln p_\theta(x_i)$, it's necessary to find $\ln p_\theta(x)$:

$$\ln p_\theta(x) = \ln \int p_\theta(x|z)p(z)dz$$

This usually has no closed form and must be estimated. The usual way to estimate integrals is

Monte Carlo integration with importance sampling:

$$\int p_\theta(x|z)p(z)dz = \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

where $q_\phi(z|x)$ is a sampling distribution over z that we use to perform the Monte Carlo integration.

So we see that if we sample $z \sim q_\phi(\cdot|x)$, then $\frac{p_\theta(x,z)}{q_\phi(z|x)}$ is an unbiased estimator of $p_\theta(x)$. Unfortunately, this does not give us an unbiased estimator of $\ln p_\theta(x)$, because $\ln$ is nonlinear. Indeed, we have by Jensen's inequality,

$$\ln p_\theta(x) = \ln \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \geq \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\ln \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

In fact, all the obvious estimators of $\ln p_\theta(x)$ are biased downwards, because no matter how many samples of $z_i \sim q_\phi(\cdot|x)$ we take, we have by Jensen's inequality:

$$\mathbb{E}_{z_i\sim q_\phi(\cdot|x)}\left[\ln\left(\frac{1}{N}\sum_i \frac{p_\theta(x,z_i)}{q_\phi(z_i|x)}\right)\right] \leq \ln \mathbb{E}_{z_i\sim q_\phi(\cdot|x)}\left[\frac{1}{N}\sum_i \frac{p_\theta(x,z_i)}{q_\phi(z_i|x)}\right] = \ln p_\theta(x)$$

Subtracting the right side, we see that the problem comes down to a biased estimator of zero:

$$\mathbb{E}_{z_i\sim q_\phi(\cdot|x)}\left[\ln\left(\frac{1}{N}\sum_i \frac{p_\theta(z_i|x)}{q_\phi(z_i|x)}\right)\right] \leq 0$$

At this point, we could branch off towards the development of an importance-weighted autoencoder[note 2], but we will instead continue with the simplest case with $N = 1$:

$$\ln p_\theta(x) = \ln \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \geq \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\ln \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

The tightness of the inequality has a closed form:

$$\ln p_\theta(x) - \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\ln \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = D_{KL}(q_\phi(\cdot|x)\|p_\theta(\cdot|x)) \geq 0$$

We have thus obtained the ELBO function:

$$L(\phi,\theta;x) := \ln p_\theta(x) - D_{KL}(q_\phi(\cdot|x)\|p_\theta(\cdot|x))$$

Maximizing the ELBO

For fixed x, the optimization $\max_{\theta,\phi} L(\phi,\theta;x)$ simultaneously attempts to maximize $\ln p_\theta(x)$ and minimize $D_{KL}(q_\phi(\cdot|x)\|p_\theta(\cdot|x))$. If the parametrization for p_θ and q_ϕ are flexible enough, we would obtain some $\hat{\phi}, \hat{\theta}$, such that we have simultaneously

$$\ln p_{\hat{\theta}}(x) \approx \max_\theta \ln p_\theta(x); \quad q_{\hat{\phi}}(\cdot|x) \approx p_{\hat{\theta}}(\cdot|x)$$

Since

$$\mathbb{E}_{x\sim p^*(x)}[\ln p_\theta(x)] = -H(p^*) - D_{KL}(p^*(x)\|p_\theta(x))$$

we have

$$\ln p_{\hat{\theta}}(x) \approx \max_\theta -H(p^*) - D_{KL}(p^*(x)\|p_\theta(x))$$

and so

$$\hat{\theta} \approx \arg\min D_{KL}(p^*(x)\|p_\theta(x))$$

In other words, maximizing the ELBO would simultaneously allow us to obtain an accurate generative model $p_{\hat{\theta}} \approx p^*$ and an accurate discriminative model $q_{\hat{\phi}}(\cdot|x) \approx p_{\hat{\theta}}(\cdot|x)$.[5]

Main forms

The ELBO has many possible expressions, each with some different emphasis.

$$\mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\ln \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \int q_\phi(z|x) \ln \frac{p_\theta(x,z)}{q_\phi(z|x)} dz$$

This form shows that if we sample $z \sim q_\phi(\cdot|x)$, then $\ln \frac{p_\theta(x,z)}{q_\phi(z|x)}$ is an unbiased estimator of the ELBO.

$$\ln p_\theta(x) - D_{KL}(q_\phi(\cdot|x) \parallel p_\theta(\cdot|x))$$

This form shows that the ELBO is a lower bound on the evidence $\ln p_\theta(x)$, and that maximizing the ELBO with respect to ϕ is equivalent to minimizing the KL-divergence from $p_\theta(\cdot|x)$ to $q_\phi(\cdot|x)$.

$$\mathbb{E}_{z\sim q_\phi(\cdot|x)}[\ln p_\theta(x|z)] - D_{KL}(q_\phi(\cdot|x) \parallel p)$$

This form shows that maximizing the ELBO simultaneously attempts to keep $q_\phi(\cdot|x)$ close to p and concentrate $q_\phi(\cdot|x)$ on those z that maximizes $\ln p_\theta(x|z)$. That is, the approximate posterior $q_\phi(\cdot|x)$ balances between staying close to the prior p and moving towards the maximum likelihood $\arg\max_z \ln p_\theta(x|z)$.

Data-processing inequality

Suppose we take N independent samples from p^*, and collect them in the dataset $D = \{x_1, \dots, x_N\}$, then we have empirical distribution $q_D(x) = \frac{1}{N}\sum_i \delta_{x_i}$.

Fitting $p_\theta(x)$ to $q_D(x)$ can be done, as usual, by maximizing the loglikelihood $\ln p_\theta(D)$:

$$D_{KL}(q_D(x)\|p_\theta(x)) = -\frac{1}{N}\sum_i \ln p_\theta(x_i) - H(q_D) = -\frac{1}{N}\ln p_\theta(D) - H(q_D)$$

Now, by the ELBO inequality, we can bound $\ln p_\theta(D)$, and thus

$$D_{KL}(q_D(x)\|p_\theta(x)) \leq -\frac{1}{N}L(\phi,\theta;D) - H(q_D)$$

The right-hand-side simplifies to a KL-divergence, and so we get:

$$D_{KL}(q_D(x)\|p_\theta(x)) \leq -\frac{1}{N}\sum_i L(\phi,\theta;x_i) - H(q_D) = D_{KL}(q_{D,\phi}(x,z);p_\theta(x,z))$$

This result can be interpreted as a special case of the data processing inequality.

In this interpretation, maximizing $L(\phi,\theta;D) = \sum_i L(\phi,\theta;x_i)$ is minimizing $D_{KL}(q_{D,\phi}(x,z);p_\theta(x,z))$, which upper-bounds the real quantity of interest $D_{KL}(q_D(x);p_\theta(x))$ via the data-processing inequality. That is, we append a latent space to the observable space, paying the price of a weaker inequality for the sake of more computationally efficient minimization of the KL-divergence.[6]

References

1. Kingma, Diederik P.; Welling, Max (2014-05-01). "Auto-Encoding Variational Bayes". arXiv:1312.6114 (https://arxiv.org/abs/1312.6114) [stat.ML (https://arxiv.org/archive/stat.ML)].
2. Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron (2016). "Chapter 19". *Deep learning*. Adaptive computation and machine learning. Cambridge, Mass: The MIT press. ISBN 978-0-262-03561-3.
3. Hinton, Geoffrey E; Zemel, Richard (1993). "Autoencoders, Minimum Description Length and Helmholtz Free Energy" (https://proceedings.neurips.cc/paper/1993/hash/9e3cfc48eccf81a0d57663e129aef3cb-Abstract.html). *Advances in Neural Information Processing Systems*. **6**. Morgan-Kaufmann.
4. Burda, Yuri; Grosse, Roger; Salakhutdinov, Ruslan (2015-09-01). "Importance Weighted Autoencoders". arXiv:1509.00519 (https://arxiv.org/abs/1509.00519) [stat.ML (https://arxiv.org/archive/stat.ML)].
5. Neal, Radford M.; Hinton, Geoffrey E. (1998), "A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants" (https://dx.doi.org/10.1007/978-94-011-5014-9_12), *Learning in Graphical Models*, Dordrecht: Springer Netherlands, pp. 355–368, doi:10.1007/978-94-011-5014-9_12 (https://doi.org/10.1007%2F978-94-011-5014-9_12), ISBN 978-94-010-6104-9, S2CID 17947141 (https://api.semanticscholar.org/CorpusID:17947141)
6. Kingma, Diederik P.; Welling, Max (2019-11-27). "An Introduction to Variational Autoencoders" (https://www.nowpublishers.com/article/Details/MAL-056). *Foundations and Trends in Machine Learning*. **12** (4). Section 2.7. arXiv:1906.02691 (https://arxiv.org/abs/1906.02691). doi:10.1561/2200000056 (https://doi.org/10.1561%2F2200000056). ISSN 1935-8237 (https://search.worldcat.org/issn/1935-8237). S2CID 174802445 (https://api.semanticscholar.org/CorpusID:174802445).

Notes

1. In fact, by Jensen's inequality,

$$\mathbb{E}_{x\sim p^*(x)}\left[\max_\theta \sum_i \ln p_\theta(x_i)\right] \geq \max_\theta \mathbb{E}_{x\sim p^*(x)}\left[\sum_i \ln p_\theta(x_i)\right] = N\max_\theta \mathbb{E}_{x\sim p^*(x)}[\ln p_\theta(x)]$$

The estimator is biased upwards. This can be seen as overfitting: for some finite set of sampled data x_i , there is usually some θ that fits them better than the entire p^* distribution.

2. By the delta method, we have

$$\mathbb{E}_{z_i \sim q_\phi(\cdot|x)}\left[\ln\left(\frac{1}{N}\sum_i \frac{p_\theta(z_i|x)}{q_\phi(z_i|x)}\right)\right] \approx -\frac{1}{2N}\mathbb{V}_{z\sim q_\phi(\cdot|x)}\left[\frac{p_\theta(z|x)}{q_\phi(z|x)}\right] = O(N^{-1})$$

If we continue with this, we would obtain the importance-weighted autoencoder.[4]
