Variational inference

Last modified: 2026-11-15 19:37:53 (PST)

The posterior of an unobserved quantity is often too hard to compute. Variational inference replaces it with a simpler stand-in that is close to it, where closeness is measured by the KL divergence.

The evidence lower bound

Definition 0.1 (Evidence lower bound (ELBO))**.** Let x be observed data, and let z be an unobserved discrete random variable, with joint probability mass function $\mathbb{p}(x, z)$. Let q be a probability mass function for z. Require that q gives probability zero to every value z where the posterior $\mathbb{p}(z \mid x)$ is zero. The **evidence lower bound** (**ELBO**) of q is the average, over z drawn from q, of the difference between the logarithm of $\mathbb{p}(x, z)$ and the logarithm of $q(z)$:

$$\mathrm{ELBO}(q) = \mathbb{E}_q[\log\{\mathbb{p}(x, z)\} - \log\{q(z)\}].$$

Written as a sum over the values z with $q(z) > 0$, the average is

$$\mathrm{ELBO}(q) = \sum_z q(z)(\log\{\mathbb{p}(x, z)\} - \log\{q(z)\}).$$

Theorem 0.1 (The log evidence is the ELBO plus a divergence)**.** *Let $\mathbb{p}(x, z)$ and q be as in Definition 0.1. Let $\mathbb{p}(x) = \sum_z \mathbb{p}(x, z)$ be the evidence of the data, also called its marginal likelihood. The evidence is nonzero, because a data value x was observed. Then the logarithm of the evidence is the ELBO of q plus the Kullback–Leibler divergence of q from the posterior:*

$$\log\{\mathbb{p}(x)\} = \mathrm{ELBO}(q) + D_{\mathrm{KL}}(q \,\|\, \mathbb{p}(\cdot \mid x)).$$

Proof

Proof. Let z be a value with $q(z) > 0$. Then $\mathbb{p}(z \mid x) \neq 0$, and $\mathbb{p}(x, z) = \mathbb{p}(z \mid x)\,\mathbb{p}(x)$. Apply log to both sides and rearrange:

$$\log\{\mathbb{p}(x)\} = \log\{\mathbb{p}(x, z)\} - \log\{\mathbb{p}(z \mid x)\}.$$

Add and subtract $\log\{q(z)\}$ on the right:

$$\log\{\mathbb{p}(x)\} = (\log\{\mathbb{p}(x, z)\} - \log\{q(z)\}) + (\log\{q(z)\} - \log\{\mathbb{p}(z \mid x)\}).$$

Multiply both sides by $q(z)$, then sum over the values z with $q(z) > 0$. The left side sums to $\log\{\mathbb{p}(x)\}$, because q sums to 1:

$$\log\{\mathbb{p}(x)\} = \sum_z q(z)(\log\{\mathbb{p}(x, z)\} - \log\{q(z)\}) + \sum_z q(z)(\log\{q(z)\} - \log\{\mathbb{p}(z \mid x)\}).$$

The first sum is ELBO(q) by its definition. The second sum is $D_{\mathrm{KL}}(q \,\|\, \mathbb{p}(\cdot \mid x))$ by the definition of the divergence. □

Example 0.1 (The ELBO and the divergence for a three-valued latent variable)**.** Let z take the values $1, 2, 3$ with prior probabilities $0.5, 0.3, 0.2$. A data value x is observed, and its probabilities given z are $0.1, 0.4, 0.8$. The R chunk below finds the posterior, then computes the ELBO and the divergence for three choices of q, with natural logarithms.

```r
prior <- c(0.5, 0.3, 0.2)
likelihood <- c(0.1, 0.4, 0.8)
joint <- prior * likelihood
evidence <- sum(joint)
posterior <- joint / evidence
log_evidence <- log(evidence)

elbo_and_divergence <- function(q) {
  log_ratio <- log(joint) - log(q)
  elbo <- sum(q * log_ratio)
  log_gap <- log(q) - log(posterior)
  divergence <- sum(q * log_gap)
  c(elbo = elbo, divergence = divergence, total = elbo + divergence)
}

uniform <- rep(1 / 3, 3)
shifted <- c(0.2, 0.3, 0.5)
rows <- rbind(
  uniform = elbo_and_divergence(uniform),
  shifted = elbo_and_divergence(shifted),
  posterior = elbo_and_divergence(posterior)
)
round(rows, 4)
#>              elbo divergence   total
#> uniform   -1.2176     0.1089 -1.1087
#> shifted   -1.1219     0.0132 -1.1087
#> posterior -1.1087     0.0000 -1.1087
```

The evidence is 0.33, so its logarithm is -1.1087. In every row the ELBO plus the divergence equals this value, as Theorem 0.1 says. The ELBO is largest, and the divergence is 0, for q equal to the posterior.

Remark 0.1 (Variational inference maximizes the ELBO). The logarithm of the evidence $\log\{\mathbb{p}(x)\}$ does not depend on q. By Theorem 0.1, choosing q to maximize the ELBO is therefore the same as choosing q to minimize its Kullback–Leibler divergence from the posterior. The ELBO never exceeds the logarithm of the evidence, because the divergence is at least 0. Variational inference searches a restricted family of choices for q this way (Blei et al. 2017).

References

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. 2017. "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association* 112 (518): 859–77. https://doi.org/10.1080/01621459.2017.1285773.