

Paper Review: Variational Inference

Beomsu Kim*

August 1, 2021

Paper Information.

- David M. Blei et. al. Variational Inference: A Review for Statisticians. [arXiv preprint arXiv:1601.00670](#), 2016.

1 Introduction

- A core problem of modern statistics is to approximate difficult-to-compute probability densities.
- Consider a joint density of latent variables $\mathbf{z} = z_{1:m}$ and observations $\mathbf{x} = x_{1:n}$

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z})p(\mathbf{x} | \mathbf{z}).$$

A Bayesian model draws the latent variables from a prior density $p(\mathbf{z})$ and then relates them back to the observations through the likelihood $p(\mathbf{x} | \mathbf{z})$.

- Inference amounts to conditioning on the data and computing the posterior $p(\mathbf{z} | \mathbf{x})$.
- In complex or high-dimensional Bayesian models, this computation is often intractable.
- There are two approaches to approximate inference: MCMC and variational inference.
 - MCMC first constructs an ergodic Markov chain on \mathbf{z} whose stationary distribution is the posterior $p(\mathbf{z} | \mathbf{x})$. Then, we sample from the chain to collect samples from the stationary distribution.
 - Variational inference uses a family of tractable¹ distributions \mathcal{Q} to approximate $p(\mathbf{z} | \mathbf{x})$. Equivalently, variational inference uses \mathcal{Q} to approximate $p(\mathbf{x})$.²
- Comparing variational inference and MCMC.
 - MCMC methods tend to be more computationally expensive than variational inference, but they also provide guarantees of producing (asymptotically) exact samples from the target density.
 - Variational inference does not enjoy such guarantees—it can only find a density close to the target—but tends to be faster than MCMC. Because it rests on optimization, variational inference easily takes advantage of methods like stochastic optimization and distributed optimization.
- In the following sections, I omit examples for clarity of exposition. Please read the reference materials for detailed examples.

*Department of Mathematical Sciences, KAIST. Email beomsu.kim@kaist.ac.kr

¹A distribution is *tractable* if it has a closed form density function or we can easily sample from it.

²Approximating $p(\mathbf{z} | \mathbf{x})$ is equivalent to approximating $p(\mathbf{x})$ since $p(\mathbf{z} | \mathbf{x}) = p(\mathbf{z}, \mathbf{x})/p(\mathbf{x})$.

2 Variational Inference

- Let $\mathbf{x} = x_{1:n}$ be a set of observed variables and $\mathbf{z} = z_{1:m}$ be a set of latent variables with joint density

$$p(\mathbf{z}, \mathbf{x}).$$

- The *inference problem* is to compute the conditional density of \mathbf{z} given \mathbf{x}

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}.$$

The denominator is called the *evidence*. We calculate it by marginalizing out the latent variables

$$p(\mathbf{x}) = \int p(\mathbf{z}, \mathbf{x}) d\mathbf{z}.$$

For many models, the evidence integral is unavailable in closed form or requires exponential (w.r.t. the dimension n) time to compute. This is why inference in such models is hard.

- Hence, we resort to approximate inference. There are two equivalent approaches.

2.1 Approach 1: Evidence Lower Bound (ELBO)

- Assuming we have $p(\mathbf{z}, \mathbf{x})$, calculating $p(\mathbf{z} \mid \mathbf{x})$ is equivalent to calculating $p(\mathbf{x})$.
- Instead of directly calculating $p(\mathbf{x})$, we maximize a lower bound of $p(\mathbf{x})$.
- Specifically, we first define a *variational family* \mathcal{Q} of tractable densities over the latent variables.
- Then, for any $q(\mathbf{z}) \in \mathcal{Q}$,

$$\begin{aligned} \log p(\mathbf{x}) &= \log \int p(\mathbf{z}, \mathbf{x}) d\mathbf{z} \\ &= \log \mathbb{E}_{q(\mathbf{z})} \left[\frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})} \right] \\ &\geq \mathbb{E}_{q(\mathbf{z})} \left[\log \frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})} \right] \\ &= \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}_{q(\mathbf{z})} [\log q(\mathbf{z})] \\ &= \text{ELBO}[q] \end{aligned}$$

where we have used Jensen's inequality at the fourth line. **ELBO** is defined as

$$\text{ELBO}[q] = \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}_{q(\mathbf{z})} [\log q(\mathbf{z})] \quad (1)$$

and it lower bounds the (log) evidence $p(\mathbf{x})$. From here comes its name “evidence lower bound”.

- We can also obtain **ELBO** by the following process.

$$\begin{aligned} \log p(\mathbf{x}) &= \int q(\mathbf{z}) \log p(\mathbf{x}) d\mathbf{z} \\ &= \int q(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\ &= \int q(\mathbf{z}) \log \frac{p(\mathbf{z}, \mathbf{x})}{q(\mathbf{z})} \cdot \frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})} d\mathbf{z} \\ &= \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}_{q(\mathbf{z})} [\log q(\mathbf{z})] + D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) \\ &= \text{ELBO}[q] + D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) \\ &\geq \text{ELBO}[q]. \end{aligned} \quad (2)$$

- Hence, we can solve the optimization problem

$$\max_{q(\mathbf{z}) \in \mathcal{Q}} \text{ELBO}[q]$$

to obtain the best approximation to $\log p(\mathbf{x})$.

- Equation (2) shows that ELBO is maximized when $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x})$.
- Calculus of variations can also be used to prove that $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x})$ maximizes the ELBO.
- For a detailed proof, see Appendix A.2.
- We also observe that

$$\begin{aligned} \text{ELBO}[q] &= \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] \\ &= \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})] + \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] \\ &= \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{x} \mid \mathbf{z})] - D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z})). \end{aligned} \tag{3}$$

- The first term is an expected likelihood, and it encourages densities that place their mass on configurations of the latent variables that explain the observed data.
- The second term is the negative divergence between the variational density and the prior; it encourages densities close to the prior.

2.2 Approach 2: Posterior Approximation

- We specify a family \mathcal{Q} of densities over the latent variables.
- Each candidate $q(\mathbf{z}) \in \mathcal{Q}$ is a candidate approximation to the exact conditional $p(\mathbf{z} \mid \mathbf{x})$.
- Inference now amounts to solving the following optimization problem

$$\begin{aligned} q^*(\mathbf{z}) &= \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) \\ &= \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z} \mid \mathbf{x})] \\ &= \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z}, \mathbf{x})] + \log p(\mathbf{x}) \\ &= \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} \mathbb{E}_{q(\mathbf{z})}[\log q(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z})}[\log p(\mathbf{z}, \mathbf{x})] \\ &= \arg \min_{q(\mathbf{z}) \in \mathcal{Q}} -\text{ELBO}[q] \\ &= \arg \max_{q(\mathbf{z}) \in \mathcal{Q}} \text{ELBO}[q]. \end{aligned}$$

- Once found, $q^*(\mathbf{z})$ is the best approximation of the conditional, within the family \mathcal{Q} .
- The complexity of the family determines the complexity of this optimization.
- Since

$$\arg \min_{q(\mathbf{z}) \in \mathcal{Q}} D_{\text{KL}}(q(\mathbf{z}) \parallel p(\mathbf{z} \mid \mathbf{x})) = \arg \max_{q(\mathbf{z}) \in \mathcal{Q}} \text{ELBO}[q],$$

approaches 1 and 2 are equivalent.

3 Mean-Field Variational Inference

- We now know we can do approximate inference by maximizing the ELBO w.r.t. a variational family \mathcal{Q} .
- We give an example of a variational family \mathcal{Q} that is often used in the literature.
- We focus on the *mean-field variational family*, where the latent variables are mutual independent and each governed by a distinct factor in the variational density.

$$q(\mathbf{z}) = \prod_{j=1}^m q_j(z_j). \quad (4)$$

Each latent variable z_j is governed by its own variational factor, the density $q_j(z_j)$.

3.1 Coordinate Ascent Mean-Field Variational Inference (CAVI)

- CAVI optimizes each factor of the mean-field variational density, while holding the others fixed.
- It climbs the ELBO to a local maximum.
- Define

$$\mathbf{z}_{-j} = (z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m), \quad q_{-j}(\mathbf{z}_{-j}) = \prod_{\ell \neq j} q_\ell(z_\ell).$$

- The *complete conditional* of z_j is its conditional density given all of the other latent variables in the model and the observations $p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})$.
- The CAVI update is given by

$$q_j^*(z_j) \propto \exp \left\{ \mathbb{E}_{q_{-j}(\mathbf{z}_{-j})} [\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})] \right\}. \quad (5)$$

Equivalently, Equation (5) is proportional to

$$q_j^*(z_j) \propto \exp \left\{ \mathbb{E}_{q_{-j}(\mathbf{z}_{-j})} [\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})] \right\}. \quad (6)$$

Because of the mean-field family assumption, the expectations of on the RHS do not involve the j th variational factor. Thus this is a valid coordinate update.

- We now derive the CAVI update. Specifically, define

$$Z[q_{-j}] = \int \exp \left\{ \mathbb{E}_{q_{-j}(\mathbf{z}_{-j})} [\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})] \right\} dz_j$$

such that

$$q_j^*(z_j) = \frac{1}{Z[q_{-j}]} \exp \left\{ \mathbb{E}_{q_{-j}(\mathbf{z}_{-j})} [\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})] \right\}$$

and we rewrite the ELBO as a functional of q_j .

$$\begin{aligned} \text{ELBO}[q_j] &= \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}_{q(\mathbf{z})} [\log q(\mathbf{z})] \\ &= \mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}_{q(\mathbf{z})} [\log q_j(z_j)] - \mathbb{E}_{q(\mathbf{z})} [\log q_{-j}(\mathbf{z}_{-j})] \\ &= \mathbb{E}_{q_j(z_j)} [\mathbb{E}_{q_{-j}(\mathbf{z}_{-j})} [\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})]] - \mathbb{E}_{q_j(z_j)} [\log q_j(z_j)] - \mathbb{E}_{q_{-j}(\mathbf{z}_{-j})} [\log q_{-j}(\mathbf{z}_{-j})] \\ &= \mathbb{E}_{q_j(z_j)} [\log q_j^*(z_j)] - \mathbb{E}_{q_j(z_j)} [\log q_j(z_j)] - \mathbb{E}_{q_{-j}(\mathbf{z}_{-j})} [\log q_{-j}(\mathbf{z}_{-j})] + Z[q_{-j}] \\ &= -D_{\text{KL}}(q_j^*(z_j) \parallel q_j(z_j)) - \mathbb{E}_{q_{-j}(\mathbf{z}_{-j})} [\log q_{-j}(\mathbf{z}_{-j})] + Z[q_{-j}]. \end{aligned}$$

Since the second and the third terms are constant w.r.t. q_j , $\text{ELBO}[q_j]$ is maximized when $q_j = q_j^*$.

- We can also use Calculus of Variations to derive the CAVI update.
- See <http://www2.imm.dtu.dk/pubdb/edoc/imm3314.pdf> for an example.

4 Expectation Maximization (EM)

- Let \mathbf{x} be a set of observed variables and \mathbf{y} be a set of latent variables with joint density

$$p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is the set of parameters of p .

4.1 EM for MLE

- The goal of MLE is to solve

$$\arg \max_{\boldsymbol{\theta}} \log p(\mathbf{x} \mid \boldsymbol{\theta}).$$

- Since

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}) = \log \int p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta}) d\mathbf{z},$$

the integral appears *inside* the log. This can make optimization difficult.

- We recall that (c.f. Equation (2))

$$\log p(\mathbf{x} \mid \boldsymbol{\theta}) = \text{ELBO}[q, \boldsymbol{\theta}] + D_{\text{KL}}(q(\mathbf{y}) \parallel p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})) \geq \text{ELBO}[q, \boldsymbol{\theta}]$$

where

$$\text{ELBO}[q, \boldsymbol{\theta}] = \mathbb{E}_{q(\mathbf{y})}[\log p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})] - \mathbb{E}_{q(\mathbf{y})}[\log q(\mathbf{y})].$$

Since

$$\mathbb{E}_{q(\mathbf{y})}[\log p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})] = \int q(\mathbf{y}) \log p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta}) d\mathbf{z},$$

the integral appears *outside* the log so the maximization of the ELBO w.r.t. q or $\boldsymbol{\theta}$ can be easier.

- Motivated by this observation, we take a two-step approach.
 - **E (expectation) step** : hold $\boldsymbol{\theta}$ constant and solve

$$q^*(\mathbf{y}) = \arg \max_{q(\mathbf{y})} \text{ELBO}[q, \boldsymbol{\theta}] = p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}).$$

We then calculate the expectation

$$\mathbb{E}_{q^*(\mathbf{y})}[\log p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})] = \mathbb{E}_{p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})}[\log p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})]$$

which is the only term in the $\text{ELBO}[q^*, \boldsymbol{\theta}]$ depending on $\boldsymbol{\theta}$.

- **M (maximization) step** : hold q^* constant and solve

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \text{ELBO}[q^*, \boldsymbol{\theta}] = \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})}[\log p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})].$$

- Set q and $\boldsymbol{\theta}$ as q^* and $\boldsymbol{\theta}^*$ and repeat the above two steps.
- This indeed increases $\log p(\mathbf{x} \mid \boldsymbol{\theta})$ since

$$\begin{aligned} \log p(\mathbf{x} \mid \boldsymbol{\theta}) &= \text{ELBO}[q, \boldsymbol{\theta}] + D_{\text{KL}}(q(\mathbf{y}) \parallel p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})) \\ &= \text{ELBO}[q^*, \boldsymbol{\theta}] \\ &\leq \text{ELBO}[q^*, \boldsymbol{\theta}^*] \\ &\leq \text{ELBO}[q^*, \boldsymbol{\theta}^*] + D_{\text{KL}}(q^*(\mathbf{y}) \parallel p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})) \\ &= \log p(\mathbf{x} \mid \boldsymbol{\theta}^*). \end{aligned}$$

4.2 EM for MAP Estimation

- We introduce a prior distribution for θ , denoted $p(\theta)$, such that

$$p(\mathbf{x}, \theta) = p(\mathbf{x} | \theta)p(\theta).$$

- The goal of MAP is to solve

$$\arg \max_{\theta} \log p(\theta | \mathbf{x}) = \arg \max_{\theta} \log p(\mathbf{x}, \theta).$$

- Replacing \mathbf{x} by (\mathbf{x}, θ) in Equation (2), we see that

$$\log p(\mathbf{x}, \theta) = \text{LBO}[q, \theta] + D_{\text{KL}}(q(\mathbf{y}) \| p(\mathbf{y} | \mathbf{x}, \theta)) \geq \text{LBO}[q, \theta]$$

where

$$\text{LBO}[q, \theta] = \mathbb{E}_{q(\mathbf{y})}[\log p(\mathbf{y}, \mathbf{x}, \theta)] - \mathbb{E}_{q(\mathbf{y})}[\log q(\mathbf{y})].$$

I will call the above term just “lower bound” not ELBO, since $p(\mathbf{x}, \theta)$ is not evidence.

- We again take a two-step approach.
 - **E (expectation) step** : hold θ constant and solve

$$q^*(\mathbf{y}) = \arg \max_{q(\mathbf{y})} \text{LBO}[q, \theta] = p(\mathbf{y} | \mathbf{x}, \theta).$$

We then calculate the expectation

$$\mathbb{E}_{q^*(\mathbf{y})}[\log p(\mathbf{y}, \mathbf{x}, \theta)] = \mathbb{E}_{p(\mathbf{y} | \mathbf{x}, \theta)}[\log p(\mathbf{y}, \mathbf{x}, \theta)]$$

which is the only term in the $\text{LBO}[q^*, \theta]$ depending on θ .

- **M (maximization) step** : hold q^* constant and solve

$$\theta^* = \arg \max_{\theta} \text{LBO}[q^*, \theta] = \arg \max_{\theta} \mathbb{E}_{p(\mathbf{y} | \mathbf{x}, \theta)}[\log p(\mathbf{y}, \mathbf{x}, \theta)].$$

- Set q and θ as q^* and θ^* and repeat the above two steps.
- This indeed increases $\log p(\mathbf{x}, \theta)$ since

$$\begin{aligned} \log p(\mathbf{x}, \theta) &= \text{LBO}[q, \theta] + D_{\text{KL}}(q(\mathbf{y}) \| p(\mathbf{y} | \mathbf{x}, \theta)) \\ &= \text{LBO}[q^*, \theta] \\ &\leq \text{LBO}[q^*, \theta^*] \\ &\leq \text{LBO}[q^*, \theta^*] + D_{\text{KL}}(q^*(\mathbf{y}) \| p(\mathbf{y} | \mathbf{x}, \theta)) \\ &= \log p(\mathbf{x}, \theta^*). \end{aligned}$$

Remark.

- EM maximizes the ELBO exactly while variational inference maximizes the ELBO approximately.
- EM which uses Monte Carlo approximation in the E step (in this case, E step of EM for MLE)

$$\mathbb{E}_{p(\mathbf{y} | \mathbf{x}, \theta)}[\log p(\mathbf{y}, \mathbf{x} | \theta)] \approx \frac{1}{L} \sum_{\ell=1}^L \log p(\mathbf{y}^{(\ell)}, \mathbf{x} | \theta)$$

where $\mathbf{y}^{(\ell)}$ are samples from $p(\mathbf{y} | \mathbf{x}, \theta)$ is called *Monte Carlo EM*.

Reference Material.

- *Pattern Recognition and Machine Learning* by Christopher M. Bishop.

5 Variational EM

- Let \mathcal{M} be a collection of model structures.
- Each model structure $m \in \mathcal{M}$ has a set of parameters θ .
- Let \mathbf{x} be a set of observed variables and \mathbf{y} be a set of latent variables.
- For each model structure m , we then have the joint distribution

$$p(\mathbf{x}, \mathbf{y}, \theta \mid m).$$

- Our goal is to calculate the evidence

$$\log p(\mathbf{x} \mid m) = \log \iint p(\mathbf{x}, \mathbf{y}, \theta \mid m) d\mathbf{y} d\theta.$$

Then, we can either perform MLE w.r.t. m

$$\arg \max_{m \in \mathcal{M}} \log p(\mathbf{x} \mid m)$$

or given a prior distribution over model structures $p(m)$, we can perform MAP w.r.t. m

$$\arg \max_{m \in \mathcal{M}} \log p(m \mid \mathbf{x}) = \arg \max_{m \in \mathcal{M}} \log p(\mathbf{x}, m) = \arg \max_{m \in \mathcal{M}} \log p(\mathbf{x} \mid m)p(m).$$

- Recall that maximizing the ELBO leads to approximating the log evidence (Section 2.1).
- Hence, setting $z_1 = \mathbf{y}$ and $z_2 = \theta$ such that $\mathbf{z} = (\mathbf{y}, \theta)$, we may use CAVI (Section 3.1).
- Specifically, define the mean-field variational distribution (c.f. Equation (4))

$$q(\mathbf{z}) = q(\mathbf{y}, \theta) = q_1(\mathbf{y})q_2(\theta).$$

Equation (5) gives us the update rule

$$\begin{aligned} q_1^*(\mathbf{y}) &\propto \exp \left\{ \mathbb{E}_{q_2(\theta)} [\log p(\mathbf{y} \mid \theta, \mathbf{x}, m)] \right\}, \\ q_2^*(\theta) &\propto \exp \left\{ \mathbb{E}_{q_1^*(\mathbf{y})} [\log p(\theta \mid \mathbf{y}, \mathbf{x}, m)] \right\}. \end{aligned}$$

We then set $q_1(\mathbf{y})$ and $q_2(\theta)$ as $q_1^*(\mathbf{y})$ and $q_2^*(\theta)$ and repeat the above two steps.

- Somewhat like EM, variational EM alternates between a $q_1(\mathbf{y})$ update and a $q_2(\theta)$ update.
- This is why it is called variational “EM”.

Reference Material.

- Matthew J. Beal and Zoubin Ghahramani. *The Variational Bayesian EM Algorithm for Incomplete Data: with Applications to Scoring Graphical Model Structures*. In Bayesian Statistics 7, 2003.

A Calculus of Variations

A.1 Preliminaries

- A *functional* is a scalar-valued function defined on the space of functions.
- Formally, a functional \mathcal{F} , when given a function u , returns a scalar $\mathcal{F}[u]$.
- The *calculus of variations* is a field of mathematics that uses variations, which are small changes in functions and functionals, to find maxima and minima of functionals.
- We use the functional gradient, denoted $\nabla\mathcal{F}[u]$, to find maxima or minima of functionals.
- We treat the functional gradient as a directional derivative

$$\langle \nabla\mathcal{F}[u], v \rangle = \left. \frac{d}{d\lambda} \mathcal{F}[u + \lambda v] \right|_{\lambda=0} \quad (7)$$

where $\lambda \in \mathbb{R}$.

- The function v representing the direction of the derivative is called the *variation* of the function u .
- The inner product is the standard L^2 inner product for real functions

$$\langle f, g \rangle = \int f(x)g(x) dx.$$

Proposition 1. *For a differentiable function f , if*

$$\mathcal{F}[u] = \int f(u(x)) dx,$$

then the functional gradient is given by

$$\nabla\mathcal{F}[u] = \frac{\partial}{\partial u} f(u) = f'(u).$$

Proof. Observe that

$$\begin{aligned} \frac{d}{d\lambda} \mathcal{F}[u + \lambda v] &= \frac{d}{d\lambda} \int f(u(x) + \lambda v(x)) dx \\ &= \int \frac{d}{d\lambda} f(u(x) + \lambda v(x)) dx \\ &= \int f'(u(x) + \lambda v(x)) v(x) dx \end{aligned}$$

and so

$$\left. \frac{d}{d\lambda} \mathcal{F}[u + \lambda v] \right|_{\lambda=0} = \int f'(u(x)) v(x) dx = \langle f'(u), v \rangle.$$

Since (7) must hold for any choice of v , we see that the claim is true. □

Reference Materials.

- https://en.wikipedia.org/wiki/Calculus_of_variations
- <https://www2.math.uconn.edu/~gordina/NelsonAaronHonorsThesis2012.pdf>

A.2 Finding the Maximizer of ELBO

- We solve the constrained optimization problem

$$\max_q \text{ELBO}[q] \quad \text{s.t.} \quad \int q(\mathbf{z}) d\mathbf{z} = 1.$$

- To this end, we form the Lagrangian

$$L(q, \lambda) = \text{ELBO}[q] + \lambda \left(\int q(\mathbf{z}) d\mathbf{z} - 1 \right).$$

- The maximizer of ELBO should satisfy the Lagrangian stationarity condition

$$\nabla L(q, \lambda) = 0.$$

Using the definition of ELBO and Proposition 1, we see that

$$\begin{aligned} \nabla L(q, \lambda) &= \nabla \left[\text{ELBO}[q] + \lambda \left(\int q(\mathbf{z}) d\mathbf{z} - 1 \right) \right] \\ &= \nabla \left[\mathbb{E}_{q(\mathbf{z})} [\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}_{q(\mathbf{z})} [\log q(\mathbf{z})] + \lambda \left(\int q(\mathbf{z}) d\mathbf{z} - 1 \right) \right] \\ &= \nabla \left[\int q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}) d\mathbf{z} - \int q(\mathbf{z}) \log q(\mathbf{z}) d\mathbf{z} + \lambda \left(\int q(\mathbf{z}) d\mathbf{z} - 1 \right) \right] \\ &= \nabla \int q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}) d\mathbf{z} - \nabla \int q(\mathbf{z}) \log q(\mathbf{z}) d\mathbf{z} + \lambda \nabla \int q(\mathbf{z}) d\mathbf{z} \\ &= \log p(\mathbf{z}, \mathbf{x}) - \log q(\mathbf{z}) - 1 + \lambda \end{aligned}$$

and so to satisfy the stationarity condition, we should have

$$q(\mathbf{z}) = p(\mathbf{z}, \mathbf{x}) e^{\lambda-1}.$$

- Due to the optimization constraint,

$$\begin{aligned} 1 &= \int q(\mathbf{z}) d\mathbf{z} = \int p(\mathbf{z}, \mathbf{x}) e^{\lambda-1} d\mathbf{z} \\ &= e^{\lambda-1} \int p(\mathbf{z}, \mathbf{x}) d\mathbf{z} \\ &= e^{\lambda-1} p(\mathbf{x}) \end{aligned}$$

so we obtain

$$e^{\lambda-1} = \frac{1}{p(\mathbf{x})}.$$

- It follows that

$$q(\mathbf{z}) = p(\mathbf{z}, \mathbf{x}) e^{\lambda-1} = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})} = p(\mathbf{z} \mid \mathbf{x}).$$

Reference Material.

- <https://wiki.inf.ed.ac.uk/twiki/pub/MLforNLP/WebHome/bkj-VBwalkthrough.pdf>