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 $\boldsymbol{z} = z_{1:m}$ and observations $\boldsymbol{x} = x_{1:n}$

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

  A Bayesian model draws the latent variables from a prior density $p(\boldsymbol{z})$ and then relates them back to the observations through the likelihood $p(\boldsymbol{x} \mid \boldsymbol{z})$.
- Inference amounts to conditioning on the data and computing the posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$.
- 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 $\boldsymbol{z}$ whose stationary distribution is the posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$. Then, we sample from the chain to collect samples from the stationary distribution.
  - Variational inference uses a family of tractable[1] distributions $\mathcal{Q}$ to approximate $p(\boldsymbol{z} \mid \boldsymbol{x})$. Equivalently, variational inference uses $\mathcal{Q}$ to approximate $p(\boldsymbol{x})$.[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`

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

[2] Approximating $p(\boldsymbol{z} \mid \boldsymbol{x})$ is equivalent to approximating $p(\boldsymbol{x})$ since $p(\boldsymbol{z} \mid \boldsymbol{x}) = p(\boldsymbol{z}, \boldsymbol{x})/p(\boldsymbol{x})$.

2 Variational Inference

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

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

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

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

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

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

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

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

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

$$\texttt{ELBO}[q] = \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \tag{1}$$

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

- We can also obtain ELBO by the following process.

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \int q(\boldsymbol{z}) \log p(\boldsymbol{x})\, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{z} \mid \boldsymbol{x})}\, d\boldsymbol{z} \\
&= \int q(\boldsymbol{z}) \log \frac{p(\boldsymbol{z}, \boldsymbol{x})}{q(\boldsymbol{z})} \cdot \frac{q(\boldsymbol{z})}{p(\boldsymbol{z} \mid \boldsymbol{x})}\, d\boldsymbol{z} \\
&= \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] + D_{\mathrm{KL}}(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})) \\
&= \texttt{ELBO}[q] + D_{\mathrm{KL}}(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})) \qquad (2) \\
&\geq \texttt{ELBO}[q].
\end{aligned}$$

- Hence, we can solve the optimization problem

$$\max_{q(\boldsymbol{z})\in\mathcal{Q}} \texttt{ELBO}[q]$$

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

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

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

- 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 $q(\boldsymbol{z}) \in \mathcal{Q}$ is a candidate approximation to the exact conditional $p(\boldsymbol{z} \mid \boldsymbol{x})$.
- Inference now amounts to solving the following optimization problem

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

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

$$\arg\min_{q(\boldsymbol{z})\in\mathcal{Q}} D_{\mathrm{KL}}(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})) = \arg\max_{q(\boldsymbol{z})\in\mathcal{Q}} \texttt{ELBO}[q],$$

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.

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

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

$$\boldsymbol{z}_{-j} = (z_1, \dots, z_{j-1}, z_{j+1}, \dots, z_m), \qquad q_{-j}(\boldsymbol{z}_{-j}) = \prod_{\ell \neq j} q_j(z_j).$$

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

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

Equivalently, Equation (5) is proportional to

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

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

- We now derive the CAVI update. Specifically, define

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

such that

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

and we rewrite the `ELBO` as a functional of q_j.

$$\begin{aligned}
\texttt{ELBO}[q_j] &= \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_{q(\boldsymbol{z})}[\log q(\boldsymbol{z})] \\
&= \mathbb{E}_{q(\boldsymbol{z})}[\log p(\boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_{q(\boldsymbol{z})}[\log q_j(z)] - \mathbb{E}_{q(\boldsymbol{z})}[\log q_{-j}(\boldsymbol{z}_{-j})] \\
&= \mathbb{E}_{q_j(z_j)}\left[\mathbb{E}_{q_{-j}(\boldsymbol{z}_{-j})}[\log p(z_j, \boldsymbol{z}_{-j}, \boldsymbol{x})]\right] - \mathbb{E}_{q_j(z_j)}[\log q_j(z)] - \mathbb{E}_{q_{-j}(\boldsymbol{z}_{-j})}[\log q_{-j}(\boldsymbol{z}_{-j})] \\
&= \mathbb{E}_{q_j(z_j)}\left[\log q_j^*(z_j)\right] - \mathbb{E}_{q_j(z_j)}[\log q_j(z)] - \mathbb{E}_{q_{-j}(\boldsymbol{z}_{-j})}[\log q_{-j}(\boldsymbol{z}_{-j})] + Z[q_{-j}] \\
&= -D_{\mathrm{KL}}(q_j^*(z_j) \| q_j(z_j)) - \mathbb{E}_{q_{-j}(\boldsymbol{z}_{-j})}[\log q_{-j}(\boldsymbol{z}_{-j})] + Z[q_{-j}].
\end{aligned}$$

Since the second and the third terms are constant w.r.t. q_j, $\texttt{ELBO}[q_j]$ 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 $\boldsymbol{x}$ be a set of observed variables and $\boldsymbol{y}$ be a set of latent variables with joint density
$$p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})$$
where $\boldsymbol{\theta}$ is the set of parameters of p.

4.1 EM for MLE

- The goal of MLE is to solve
$$\arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x} \mid \boldsymbol{\theta}).$$

- Since
$$\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \log \int p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})\, d\boldsymbol{z},$$
the integral appears *inside* the log. This can make optimization difficult.

- We recall that (c.f. Equation (2))
$$\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \texttt{ELBO}[q, \boldsymbol{\theta}] + D_{\mathrm{KL}}(q(\boldsymbol{y}) \| p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})) \geq \texttt{ELBO}[q, \boldsymbol{\theta}]$$
where
$$\texttt{ELBO}[q, \boldsymbol{\theta}] = \mathbb{E}_{q(\boldsymbol{y})}[\log p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})] - \mathbb{E}_{q(\boldsymbol{y})}[\log q(\boldsymbol{y})].$$
Since
$$\mathbb{E}_{q(\boldsymbol{y})}[\log p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})] = \int q(\boldsymbol{y}) \log p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})\, d\boldsymbol{z},$$
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
$$q^*(\boldsymbol{y}) = \arg\max_{q(\boldsymbol{y})} \texttt{ELBO}[q, \boldsymbol{\theta}] = p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}).$$
    We then calculate the expectation
$$\mathbb{E}_{q^*(\boldsymbol{y})}[\log p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})] = \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}[\log p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})]$$
    which is the only term in the $\texttt{ELBO}[q^*, \boldsymbol{\theta}]$ depending on $\boldsymbol{\theta}$.
  - **M (maximization) step** : hold q^* constant and solve
$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \texttt{ELBO}[q^*, \boldsymbol{\theta}] = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})}[\log p(\boldsymbol{y}, \boldsymbol{x} \mid \boldsymbol{\theta})].$$
  - Set q and $\boldsymbol{\theta}$ as q^* and $\boldsymbol{\theta}^*$ and repeat the above two steps.

- This indeed increases $\log p(\boldsymbol{x} \mid \boldsymbol{\theta})$ since
$$\begin{aligned}
\log p(\boldsymbol{x} \mid \boldsymbol{\theta}) &= \texttt{ELBO}[q, \boldsymbol{\theta}] + D_{\mathrm{KL}}(q(\boldsymbol{y}) \| p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})) \\
&= \texttt{ELBO}[q^*, \boldsymbol{\theta}] \\
&\leq \texttt{ELBO}[q^*, \boldsymbol{\theta}^*] \\
&\leq \texttt{ELBO}[q^*, \boldsymbol{\theta}^*] + D_{\mathrm{KL}}(q^*(\boldsymbol{y}) \| p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})) \\
&= \log p(\boldsymbol{x} \mid \boldsymbol{\theta}^*).
\end{aligned}$$

4.2 EM for MAP Estimation

- We introduce a prior distribution for $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}$, such that

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}).$$

- The goal of MAP is to solve

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

- Replacing $\boldsymbol{x}$ by $(\boldsymbol{x}, \boldsymbol{\theta})$ in Equation (2), we see that

$$\log p(\boldsymbol{x}, \boldsymbol{\theta}) = \texttt{LBO}[q, \boldsymbol{\theta}] + D_{\mathrm{KL}}(q(\boldsymbol{y}) \| p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})) \geq \texttt{LBO}[q, \boldsymbol{\theta}]$$

  where

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

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

- We again take a two-step approach.
  - **E (expectation) step** : hold $\boldsymbol{\theta}$ constant and solve

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

    We then calculate the expectation

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

    which is the only term in the $\texttt{LBO}[q^*, \boldsymbol{\theta}]$ depending on $\boldsymbol{\theta}$.
  - **M (maximization) step** : hold q^* constant and solve

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

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

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

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)

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

  where $\boldsymbol{y}^{(\ell)}$ are samples from $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$.
- Let $\boldsymbol{x}$ be a set of observed variables and $\boldsymbol{y}$ be a set of latent variables.
- For each model structure m, we then have the joint distribution

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

- Our goal is to calculate the evidence

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

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

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

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

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

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

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

  Equation (5) gives us the update rule

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

  We then set $q_1(\boldsymbol{y})$ and $q_2(\boldsymbol{\theta})$ as $q_1^*(\boldsymbol{y})$ and $q_2^*(\boldsymbol{\theta})$ and repeat the above two steps.
- Somewhat like EM, variational EM alternates between a $q_1(\boldsymbol{y})$ update and a $q_2(\boldsymbol{\theta})$ 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

$$\langle \nabla\mathcal{F}[u], v\rangle = \frac{d}{d\lambda}\mathcal{F}[u+\lambda v]\bigg|_{\lambda=0} \tag{7}$$

  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

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

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

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

- To this end, we form the Lagrangian

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

- 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

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

  and so to satisfy the stationarity condition, we should have

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

- Due to the optimization constraint,

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

  so we obtain

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

- It follows that

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

Reference Material.

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