Variational Inference

Notes for Reading Group

Hongwei Jin

1 Problem

One of the core problems of modern statistics is to approximate difficult-to-compute probability densities. This problem is especially important in Bayesian statistics, which frames all inference about unknown quantities as a calculation involving the posterior density. This note is for the review of variational inference, a method used to approximate posterior densities for Bayesian models.

Let's set up the the general problem. Consider a joint density of latent variables $\mathbf{z} = z_{1:m}$ and observations $\mathbf{x} = x_{1:n}$, the task is to calculate the posterior:

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

1.1 Motivation

MCMC. We can get the probability density by MCMC sampling.

1. construct an ergodic Markov chain on $\mathbf{z}$ whose stationary distribution is the posterior $p(\mathbf{z}|\mathbf{x})$;
2. sample from the chain to collect samples from the stationary distribution, such as Metropolis-Hastings algorithm and Gibbs sampler;
3. approximate the posterior with an empirical estimate constructed from the collected samples.

Variational Inference. VI measures the posterior probability density by optimizing a family of densities, instead of MCMC sampling.

1. posit a family of approximate densities $\mathcal{Q}$, a set of densities over the latent variables;
2. try to find the member of that family which minimizing the Kullback-Leibler (KL) divergence to the exact posterior:

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z})\in\mathcal{Q}}{\operatorname{argmin}} D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})). \tag{1}$$

3. approximate the posterior with the optimized member of the family $q^*(\cdot)$.

Comparison Both variational inference and MCMC solve the same problem to estimate the posterior probability density, while they have different approaches and properties. The relative accuracy of variational inference and MCMC is still known. In real applications, choice between variational inference and MCMC still needs more cares.

2 KL Divergence

In the classical VI, the model adopts the KL divergence to measure the distance between two distributions. However, we will take a closer look why and how the KL divergence works for the distance measure.

	VI	MCMC
approach	optimize on a set of densities	sample from Markov chain
guarantee	find a density close to the target	can find the exact density
computation	faster	slower
data size	suit for large	suit for small
distribution	adopt stochastic optimization methods easily	hard to be parallel
model complexity	mixture models (not suit for Gibbs sampling)	mixture model (Gibbs sampling)

Table 1: VI v.s. MCMC

2.1 Entropy

In the information theory, the information is defined as $\mathbf{I}(x) = -\log p(x)$, and the entropy is defined as the measurement of the expected information you can get if one of the events happens.

$$H(P) = \mathbb{E}\left[I(\mathbf{x})\right] = \mathbb{E}\left[-\log p(\mathbf{x})\right],$$

and explicitly, it is written as $\mathbf{H}(P) = -\sum_i p(x_i)\log p(x_i)$.

Example 1. *The entropy of exponential distribution $p(x) = \lambda e^{-\lambda x}$ is $1 - \log\lambda$.*

Another definition is the *cross entropy*, which is defined as $H(P, Q) = -\sum_i p(x_i)\log q(x_i)$.

2.2 KL divergence

KL divergence is also known as relative entropy, which is more descriptive name. The entropy measures the average information, and the relative entropy measures the "distance" from one distribution to another.

Definition 1 (KL divergence)**.** *For two probability distributions P and Q, the KL divergence with respect to P is defined as*

$$D_{KL}(P||Q) = \sum\nolimits_i p(x_i)\log\frac{p(x_i)}{q(x_i)}.$$

From the definition, we can see if P and Q are close "almost everywhere", then the divergence goes to 0. KL divergence can also be written as the difference of expectation with respect to P, that is

$$D_{KL}(P||Q) = \mathbb{E}\left[\log p\right] - \mathbb{E}\left[\log q\right].$$

2.3 Properties of KL divergence

- Non-negative: $D_{KL}(P||Q) \geq 0$ (to show it).

  Lemma 1 (Gibbs inequality)**.** *Suppose $P = \{p_1, \cdots, p_n\}$ is a probability distribution, then for any other probability distribution Q, then following inequality holds*

  $$-\sum_i^n p_i \log p_i \leq -\sum_i^n p_i \log q_i.$$

  Note that the difference of two sides is exactly the KL divergence with respect to P.

- None symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. It is obvious based on the definition, but the geometric meaning is more intuitive. Thus KL divergence is not a proper divergence measure between two distributions.
  To enrich the content, a large class of different divergences are the so called f-divergences. And a list of common divergences is listed in the table.

  Definition 2 (f-divergence)**.** *Given two distributions P and Q that possess, respectively, an absolutely continuos density function p and q with respect to base measure dx defined on the domain χ, we define the f-divergence as*

  $$D_f(P||Q) = \int_\chi q(x) f\left(\frac{p(x)}{q(x)}\right) dx,$$

  where f is the generator function: $\mathbb{R}_+ \to \mathbb{R}$ is a convex, lower-semicontinous function satisfying $f(1) = 0$.

Name	$D_f(P\|\|Q)$	$f(\cdot)$
KL	$\int p(x) \log \frac{p(x)}{q(x)} dx$	$u \log u$
Reverse KL	$\int q(x) \log \frac{p(x)}{q(x)} dx$	$-\log u$
Person χ^2	$\int \frac{(p(x)-q(x))^2}{p(x)} dx$	$(u-1)^2$
Squared Hellinger	$\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$	$(\sqrt{u}-1)^2$
Jensen-Shannon	$\frac{1}{2} \int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} dx$	$u \log u - (u+1) \log \frac{u+1}{2}$
GAN	$\int p(x) \log \frac{2p(x)}{p(x)+q(x)} + q(x) \log \frac{2q(x)}{p(x)+q(x)} dx - \log(4)$	$u \log u - (u+1) \log(u+1)$

- Additive for independent distributions: if P_1, P_2 are independent distributions and similar like Q_1, Q_2, then the KL divergence for the joint distribution has

$$D_{KL}(P||Q) = D_{KL}(P_1||Q_1) + D_{KL}(P_2||Q_2).$$

- Joint convexity: for any $0 \leq \lambda \leq 1$

$$D_{KL}(\lambda P_1 + (1-\lambda)P_2||\lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D_{KL}(P_1||Q_1) + (1-\lambda) D_{KL}(P_2||Q_2).$$

  This follows from the convexity of the mapping $(p, q) \mapsto qf(p/q)$ on $\mathbb{R}_+^2$,

3 Variational Inference

Bring our problem of estimating the posterior $p(\mathbf{z}|\mathbf{x})$ together with KL divergence, our main task is to find a best approximation density $q(\mathbf{z})$ from the family $\mathcal{Q}$ such that it has the minimum KL divergence. In other words, the approximated density is the closest one among the density family.

However, this problem itself is still hard to solve. We can write the conditional density as $p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})}$. The denominator contains the marginal density of the observations, also called the *evidence*. We calculate it by marginalizing out the latent variables from the joint density:

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

The integral itself is unavailable or requires exponential time to compute. That's why the inference is hard.

3.1 ELBO

Consider the KL divergence between the candidate distribution and the posterior

$$\begin{aligned} D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) &= \sum_i q(z_i) \log \frac{q(z_i)}{p(z_i|x_i)} = \sum_i q(z_i) \log \frac{q(z_i)p(x_i)}{p(z_i, x_i)} \\ &= \sum_i q(z_i) \log \frac{q(z_i)}{p(z_i, x_i)} p(x_i) = \sum_i q(z_i) \log \frac{q(z_i)}{p(z_i, x_i)} + \sum_i q(z_i) \log p(x_i) \\ &= D_{KL}(q(\mathbf{z})||p(\mathbf{z}, \mathbf{x})) + \log p(\mathbf{x}) \end{aligned}$$

Note that the LHS is our objective, and the first term of RHS is the KL divergence between the $q(\mathbf{z})$ and $p(\mathbf{z}, \mathbf{x})$, plus a fixed term $\log p(\mathbf{x})$. Rewrite the equation as:

$$D_{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})) + ELBO(q) = \log p(\mathbf{x}), \tag{2}$$

where $ELBO(q)$ is called the evidence lower bound (EBLO). Specifically, it is

$$\begin{aligned} ELBO(q) &= -D_{KL}(q(\mathbf{z})||p(\mathbf{z}, \mathbf{x})) \\ &= \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] - \mathbb{E}\left[\log q(\mathbf{z})\right] \\ &= \mathbb{E}\left[\log p(\mathbf{x}|\mathbf{z})\right] + \mathbb{E}\left[\log p(\mathbf{z})\right] - \mathbb{E}\left[\log q(\mathbf{z})\right] \\ &= \mathbb{E}\left[\log p(\mathbf{x}|\mathbf{z})\right] - D_{KL}(q(\mathbf{z})||p(\mathbf{z})). \end{aligned}$$

Observations:

- $\log p(x) \in [-\infty, 0]$,
- KL divergence is always positive, the smaller the better,
- ELBO is the negation of KL divergence between $q(\mathbf{z})$ and $p(\mathbf{z}, \mathbf{x})$,
- minimizing the KL divergence is equivalent to maximizing the ELBO,
- the first term of ELBO is the log-likelihood, which is optimized by the EM algorithm,
- maximizing the ELBO is still hard.

3.2 Mean-field variational family

Minimizing the KL divergence is the same as maximizing the ELBO. Therefore, to find $q(z)$ is equivalent to maximize the ELBO

$$ELBO(q) = \sum q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}, \mathbf{x})}$$

This is still hard to solve, so why not make assumptions on z, in order to restrict the family of distribution.

Assumption 1. $\mathbf{z}$ *are independent with each other, i.e.,*

$$q(\mathbf{z}) = \prod_{i=1}^{m} q(z_i)$$

There is no further assumptions about the distribution. In particular, we place no restriction on the function forms of the individual factor $q(z_i)$. This factorized form is called *mean field variational inference*, which originally comes from physics.

3.3 Coordinate Ascent

Using the ELBO and the mean-field family, we can approximate conditional inference as an optimization problem. Here we will introduce a commonly used algorithm for solving this optimization problem, *coordinate ascent variational inference*.

Observations.

- Instead of optimizing $q(\mathbf{z})$, under the independence assumption, we can try to alternatively optimize one by one;
- The complete conditional of latent variable z_j is its conditional density given all of the other latent variables and the observations, $p(z_j|\mathbf{z}_{-j}, \mathbf{x})$;
- The optimal $q(z_j)$ is proportional to the expected log of the complete condition:

$$q^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j|\mathbf{z}_{-j}, \mathbf{x})\right]\right\}$$

- $q(z_j)$ is also proportional to

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

- coordinate ascent is closely related to Gibbs sampling. The Gibbs sampler maintains a realization of latent variables and iteratively samples from each variable's complete conditional, while coordinate ascent takes the expected log and uses this quantity to set each variable's variational factor.
- The ELBO is a non-convex problem, but the coordinate ascent will eventually converge to local optimum?

Algorithm 1: Coordinate ascent variational inference (CAVI)

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$

Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

Initialize: Variational factors $q_j(z_j)$

while *the* ELBO *has not converged* **do**

  for $j \in \{1, \ldots, m\}$ **do**

    Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$

  end

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

end

return $q(\mathbf{z})$

3.4 Exponential Family Conditionals

One remaining question is whether there is a general form for models in which the coordinate updates in mean field variational inference are easy to compute and lead to closed-form updates? The answer is yes, and the form is called exponential family.

Models with conditional densities that are in an exponential family has the form in short

$$p(\mathbf{x}) = h(\mathbf{x}) \exp\left\{\eta^\top t(\mathbf{x}) - a(\eta)\right\} \tag{3}$$

And explicitly, in our problem it is

$$p(z_j|\mathbf{z}_{-j}, \mathbf{x}) = h(z_j) \exp\left\{\eta(\mathbf{z}_{-j}, \mathbf{x})^\top t(z_j) - a(\eta(\mathbf{z}_{-j}, \mathbf{x}))\right\}, \tag{4}$$

where h, η, t, a are functions that parameterize the exponential family.

Facts.

- it is called exponential family conditional models, a special case is conditional conjugate models with local and global variables;
- the log normalizer $a(\eta) = \log \int \exp\left\{\eta^\top t(\mathbf{x})\right\} d\mathbf{x}$ ensures the density integrates to one;
- the gradient calculates the expected sufficient statistics: $\mathbf{x} = \nabla_\eta a(\eta)$
- different choices of parameters lead to many popular densities, including normal, gamma, exponential, etc.

Derivation.

- the log of the conditional:

$$\log p(z_j|\mathbf{z}_{-j}, \mathbf{x}) = \log h(z_j) + \eta(\mathbf{z}_{-j}, \mathbf{x})^\top t(z_j) - a(\eta(\mathbf{z}_{-j}, \mathbf{x})).$$

- expectation of this with respect to $q(\mathbf{z}_{-j})$:

$$\mathbb{E}_{q_{-j}}\left[\log p(z_j|\mathbf{z}_{-j}, \mathbf{x})\right] = \log h(z_j) + \mathbb{E}_{q_{-j}}\left[\eta(\mathbf{z}_{-j}, \mathbf{x})^\top t(z_j)\right] - \mathbb{E}_{q_{-j}}\left[a(\eta(\mathbf{z}_{-j}, \mathbf{x}))\right].$$

- last term doesn't depend on $q(\mathbf{z}_{-j})$, then:

$$q^*(z_j) \propto h(z_j) \exp\left\{\mathbb{E}_{q_{-j}}\left[\eta(\mathbf{z}_{-j}, \mathbf{x})\right]^\top t(z_j)\right\}$$

  It is in the same exponential family as the conditional.

- Given each latent variable a variational parameter ν_j, the full approximation is

$$q(\mathbf{z}|\boldsymbol{\nu}) = \prod_j q(z_j|\nu_j).$$

  then the coordinate ascent algorithm updates each variational parameter as

$$\nu_j^* = \mathbb{E}_{q_{-j}}\left[\eta(\mathbf{z}_{-j}, \mathbf{x})\right].$$

- extend to stochastic variational inference (SVI).

3.5 Open problems

- optimize over other measures, instead of KL divergence;
- get rid of independence assumptions with mean field family;
- explore the interface between VI and MCMC;
- understand statistical profile of VI.

References

[Bis06] Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[BKM17] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

[JGJS99] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[KFB09] Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[WJ+08] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.