CS 7545: Machine Learning Theory **Fall 2018**

Lecture 26: Variational inference

Lecturer: Andre Wibisono *Scribes: Cyrus Hettle*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

26.1 Variational inference

Inference We are given a hidden state θ, which yields an observation x. There is a prior $p(\theta)$ coming from a distribution of possible states, as well as a data likelihood probability $p(X|\theta)$. These yield a joint distribution $p(\theta, x) = p(\theta) \cdot p(x|\theta)$. We may have that x has entries $x_1, \ldots, x_n$ which are sampled iid given θ, and hence when conditioned on θ, $x_1, \ldots, x_n$ are independent. However, they may not be marginally independent, since it is possible that

$$p(x_1, \ldots, x_n) = \int_\theta p(x_1, \ldots, x_n, \theta) d\theta = \int_\theta p(\theta) \cdot \prod_{i=1}^n p(x_i|\theta)\ d\theta \neq \prod_{i=1}^n p(x_i).$$

The posterior Recall Bayes' rule

$$p(\theta|x) = \frac{p(\theta) \cdot p(x|\theta)}{p(x)}.$$

In our notation, we have

$$p(\theta|x) = \frac{p(\theta) \cdot p(x|\theta)}{\int_{\theta'} p(\theta') \cdot p(x|\theta') d\theta'}.$$

Unfortunately, this expression may be arbitrary (i.e., it has no closed form) and can be expensive to compute.

What to do? (1) The gold standard is sampling: for instance, Gibbs sampling, Langevin dynamics, or (stochastic gradient) Markov chain Monte Carlo.

(2) With variational inference, we try to find the best approximate posterior $p(\theta|x)$ from a "nice" class. Here are some possibilities:

1. VB (variational Bayes). Take

$$q^* = \arg\min_{q \in Q} KL(q||p(\cdot|x)).$$

where the KL-divergence is $KL(q||p) = \int_\theta q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta$. The KL-divergence is convex in its first argument, but in practice the set Q may be nonconvex. Note that

$$KL(q||p) \geq 0$$
$$KL(q||p) = 0 \Leftrightarrow q = p.$$

2. EP (expectation propagation). With KL as above, take

$$q^* = \arg\min_{q \in Q} KL(p(\cdot|x)||q).$$

3. Belief propagation (sum-product algorithm). Here we minimize the Bethe free energy (using the approximating graph.)

Definition 26.1 *The **evidence lower bound**, or **ELBO**, is given by*

$$ELBO(q) = \int_\theta q(\theta) \log \frac{p(\theta, x)}{q(\theta)} d\theta \leq \log p(x).$$

Hence we have that

$$\begin{aligned} KL(q||p(\cdot|x)) &= \int_\theta q(\theta) \log \frac{p(\theta, x)}{q(\theta)} d\theta \\ &= \int_\theta q(\theta) \log \frac{q(\theta) \cdot p(x)}{p(\theta, x)} d\theta \\ &= -ELBO(q) + \log p(x) \geq 0. \end{aligned}$$

26.2 Exponential family distributions

Definition 26.2 *A probability distribution $q(\theta)$ is in the **exponential family** if q is of the form*

$$q(\theta) = \exp(T(\theta)^\top \eta - A(\eta))$$

for some sufficient statistics $T(\theta) \in \mathbb{R}^m$ and natural parameter $\eta \in \mathbb{R}^m$, where

$$A(\eta) = \log \int_\theta \exp(T(\theta)^\top \eta - A(\eta)) d\theta$$

is the log-partition function.

Examples of such distributions are the Bernoulli, Poisson, geometric, exponential, Gaussian, graphical models, etc. However, uniform distributions (for instance) are not in the exponential family.

Properties of exponential distributions

1. We have that

   $$\nabla A(\eta) = \mathbb{E}q_\eta[T(x)]$$

   since

   $$\nabla A(\eta) = \frac{\int_\theta \nabla e^{\langle T(\theta), \eta \rangle} d\theta}{\int_\theta e^{\langle T(\theta), \eta \rangle} d\theta} = \frac{\int_\theta T(\theta) e^{\langle T(\theta), \eta \rangle} d\theta}{\int_\theta e^{\langle T(\theta), \eta \rangle} d\theta} = \mathbb{E}q_\eta[T(x)]$$

2. We have that

   $$\nabla^2 A(\eta) = cov_{q_\eta}(T(x)) = \mathbb{E}q_\eta[(T(\theta) - \nabla A(\eta))(T(\theta) - \nabla A(\eta))^\top] \geq 0,$$

   so A is convex.

3. $\nabla^m A(\eta) = cov(m)_{q_\eta}$.

4. $KL(q_\eta || q_\delta) = D_A(\delta, \eta) = A(\delta) - A(\eta) - \langle \nabla A(\eta), \delta - \eta \rangle$.

5. $A^*(\mu) = -A(q_{\eta(\mu)})$, where $\eta(\mu)$ is the unique parameter satisfying the **moment-matching condition** $\mathbb{E}q_{\eta(\mu)}[T(\theta)] = \mu$.

6. The exponential family distribution is the maximum entropy distribution with moment constraint

   $$q_{\eta(\mu)} \leftarrow \arg\max_q H(q) \text{ s.t. } \mathbb{E}_q[T(\theta)] = \mu$$

For VB, we usually choose

$$Q = \{\text{exp. family } q_\eta(\theta) = \exp(\langle T(\theta), \eta \rangle - A(\eta)\}.$$

Mean-field VB When $\theta = (\theta_1, \ldots, \theta_k)$, choose the approximating distribution to be independent; that is, $q(\theta) = q_1(\theta_1), \ldots, q_k(\theta_k)$. In practice, we find the optimality condition

$$\text{for } i = 1, \ldots, k: \ \frac{\partial}{\partial q_i} ELBO(q) = 0$$

and solve by iteratively setting q_i such that $\frac{\partial}{\partial q_i} ELBO(q) = 0$.

Streaming/online setting We can also approach this problem in an online setting, where we observe $x_1, \ldots, x_n$ in a stream. There are two main approaches. First, treat $ELBO$ as a sum of n terms, and use stochastic gradient descent (see Hoffman, Blei et al. 2013.) Second is the filtering method. Iteratively, use the approximate posterior q_{n-1} as the new prior p_{n-1}, then approximate again to get a new posterior q_n.