CS 329M: Topics in Artificial Intelligence **Spring 2017**

Lecture 5: May 1

Lecturer: James Zou *Scribe: Ruishan Liu, Feng Ruan, Can Wang*

5.1 Variational Inference 1

In this note, we introduce the basic idea of variational inference, show typical algorithms and illustrate with an example.

5.1.1 Motivation

In Bayesian inference and machine learning, we are interested in estimating the distribution

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} \tag{5.1}$$

Here x and z can be general variables. When x is the data and z is the parameter, $p(z|x)$ corresponds to the posterior distribution.

In Eq. (5.1), the conditional probability $p(x|z)$ and the prior $p(z)$ are relatively easy to calculate, but $p(x)$ can be hard to estimate. In the last lecture, we introduced the Hamiltonian Monte Carlo method to sample from data and do inference. This week, we discuss another approach, that is, variational inference.

5.1.2 Variational Inference

The basic idea of variational inference is to find a simpler distribution $q(z|\lambda)$ to approximate than the original distribution $p(z|x)$. Here $q(z|\lambda)$ should be simple for estimation and is parametrized by $\lambda \in \Lambda$.

In order to do a good approximation, we first need to measure the distance between the estimated distribution $q(z|\lambda)$ and the real one $p(z|x)$. There are a lot of potential metrics to use, which is still a hot research topic today. One wildly-used and computationally efficient measure is *KL divergence*, defined as

$$\begin{aligned} \mathrm{KL}[q(z|\lambda)||p(z|x)] &= \int q(z|\lambda)\log\frac{q(z|\lambda)}{p(z|x)}\mathrm{d}z \\ &= \mathbb{E}_{z\sim q(z|\lambda)}\left[\log\frac{q(z|\lambda)}{p(z|x)}\right] \end{aligned}$$

After deriving the "distance" between $q(z|\lambda)$ and $p(z|x)$, our goal is then to minimize the divergence and get an optimal $\lambda \in \Lambda$, i.e., find the minimum of the optimization problem below,

$$\underset{\lambda\in\Lambda}{\text{minimize}}\, \mathbb{E}_{z\sim q(\cdot|\lambda)}\left[\log\frac{q(z|\lambda)}{p(z|x)}\right] = \int q(z|\lambda)\log\frac{q(z|\lambda)}{p(z|x)}dz \tag{5.2}$$

5.2 Numerical Algorithms

5.2.1 SGD Method

A natural idea to solve the optimization problem (5.2) is to use gradient descent method. To do so, we first compute the derivative of the objective (5.2) w.r.t λ

$$\begin{aligned}\nabla_\lambda \left[\int q(z|\lambda)\log\frac{q(z|\lambda)}{p(z|x)}dz\right] &= \int [\nabla_\lambda q(z|\lambda)]\log\frac{q(z|\lambda)}{p(z|x)}dz - \int \nabla_\lambda q(z|\lambda)dz \\ &= \int q(z|\lambda)\nabla_\lambda[\log(q(z|\lambda)]\log\frac{q(z|\lambda)}{p(z|x)}dz - \underbrace{\nabla_\lambda\left[\int q(z|\lambda)dz\right]}_{=0} \\ &= \mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\nabla_\lambda[\log(q(Z|\lambda)]\log\frac{q(Z|\lambda)}{p(Z|x)}\right]\end{aligned}$$

Clearly, exact computation of the gradient shown above is challenging, as it involves a nontrivial evaluation of certain expectation (or equivalently, evaluation of a nontrivial integral). Nevertheless, since it takes the form of certain expectation, it has some natural unbiased estimate. For instance, if one simulates some sample $Z \sim q(\cdot|\lambda)$ (which we assumed was easy), then

$$\sum_{i=1}^{M}\nabla_\lambda[\log(q(Z_i|\lambda)]\log\frac{q(Z_i|\lambda)}{p(Z_i|x)}$$

is an unbiased estimator of the gradient. This observation motivates us to consider the following stochastic gradient descent (SGD) method.

Algorithm 1: Minimize objective (5.2) via SGD

Data: Initialize $\lambda^0 \in \Lambda$, stepsize $\{\alpha_k\}_{k=1}^{\infty}$, sample size $M \in \mathbb{N}$.

while λ *not converged* **do**

  Sample $z^{(1)}, z^{(2)}, \ldots, z^{(M)} \sim q(\cdot|\lambda)$.

  Update λ^{k+1} via a stochastic gradient step from λ^k

$$\lambda^{k+1} = \lambda^k - \frac{\alpha_k}{M}\sum_{i=1}^{M}\nabla_\lambda[\log(q(z^{(i)}|\lambda)]\log\frac{q(z^{(i)}|\lambda)}{p(z^{(i)}|x)} \tag{5.3}$$

end

Remark

1. Note here, in the above updating step, we draw M samples i.i.d $\{Z^{(i)}\}_{i=1}^{M}$ and since each of

$$\nabla_\lambda[\log(q(Z^{(i)}|\lambda)]\log\frac{q(Z^{(i)}|\lambda)}{p(Z^{(i)}|x)}$$

is an unbiased estimator of the gradient, its average is also unbiased. A choice of large M certainly leads to a more accurate estimate of true gradient, yet it could also bring huge computational burden that one does not wish to have. Such tradeoff might suggests that one would choose a moderately large M in practice.

2. Observe that the updating rule (5.3) can sometimes cause huge computational challenges as evaluation of $p(z|x)$ can be computationally intensive. In this case, notice the fact that, the optimization problem (5.2) is related to the following optimization problem:

$$\underset{\lambda\in\Lambda}{\text{minimize}}\, \mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\log\frac{q(Z|\lambda)}{p(Z,x)}\right] = \int q(z|\lambda)\log\frac{q(z|\lambda)}{p(z,x)}dz = \int q(z|\lambda)\log\frac{q(z|\lambda)}{p(z|x)}dz - \log p(x) \tag{5.4}$$

Hence, a similar derivation of the gradient step shows that one can in fact substitute $p(z^{(i)}|x)$ with $p(z^{(i)},x)$ in Eq (5.3) and this leads to an SGD solver for problem (5.4), which is often computationally efficient as $p(z,x)=p(z)p(x|z)$ is often easy to evaluate numerically.

5.2.2 Coordinate Descent Method

Another idea to solve optimization problem (5.2) is to assume some nice structures on the candidate set $\{q(\cdot|\lambda)\}_{\lambda\in\Lambda}$. To be more concrete, suppose the hidden variable z is an m dimensional vector. Consider the following set Q as the set of candidates to approximate the target density $p(z|x)$:

$$Q = \left\{q(\cdot|\lambda), \lambda=(\lambda_1,\lambda_2,\ldots,\lambda_m)^T\in\mathbb{R}^m : q(z|\lambda)=\Pi_{i=1}^m q(z_i|\lambda_i) \text{ for some density function } \{q(\cdot|\lambda_i)\}_{i=1}^m\right\} \tag{5.5}$$

Similar to the discussion in section 5.1.2, one seeks to find some $q(\cdot|\lambda)\in Q$ that minimizes the KL divergence between $q(\cdot|\lambda)$ and $p(z|x)$. Hence, let us consider the optimization problem below

$$\underset{q(\cdot|\lambda)\in Q}{\text{minimize}}\ \mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\log\frac{q(Z|\lambda)}{p(Z|x)}\right] = \mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\sum_{i=1}^m \log q(Z_i|\lambda_i) - \log p(Z|x)\right] \tag{5.6}$$

Now, notice that (5.6) has a simple "decomposable structure" (see the RHS of the above equation). Thus, this motivates us to consider the following coordinate descent method to solve the above optimization problem (5.6).

Algorithm 2: Minimize objective (5.6) via CD

Data: Initialize $\lambda^0=(\lambda_1^0,\lambda_2^0,\ldots\lambda_m^0)$.

while $\{\lambda^k\}$ *not converged* **do**

  for $i_0\in\{1,2,\ldots,m\}$ **do**

$$\lambda_{i_0}^{k+1} = \underset{\lambda_{i_0}\text{ with }\lambda_{-i_0}\text{ fixed}}{\operatorname{argmin}} \mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\sum_{i=1}^m \log q(Z_i|\lambda_i) - \log p(Z|x)\right] \tag{5.7}$$

  end

end

In coordinate descent algorithm, using the "decomposable structure" of the objective in (5.6), one can find a closed form solution for the update (5.7). To start with the analysis, we need the following lemma.

Lemma 1. *For any function $f(x)\geq 0$, consider the optimization problem below*

$$\underset{g(\cdot)\in G}{\text{minimize}}\ \mathbb{E}_{Z\sim g(\cdot)}\left[\log g(Z) - \log f(Z)\right] \tag{5.8}$$

where we use G to denote the set of all possible probability densities, i.e.,

$$G = \left\{g(\cdot) : g(\cdot)\geq 0, \int g = 1\right\}$$

Then, the minimizer of the optimization problem (5.8)*, denoted by* $g^*(\cdot)$*, satisfies* $g^*(\cdot) \propto f(\cdot)$*.*

Proof: Note that, if we define $F = \int f \geq 0$, and $\tilde{f}(\cdot) = f(\cdot)/F$, then we have,

$$\begin{aligned}\mathbb{E}_{Z\sim g(\cdot)}\left[\log g(Z) - \log f(Z)\right] &= \mathbb{E}_{Z\sim g(\cdot)}\left[\log g(Z) - \log \tilde{f}(Z)\right] - \log(F) \\ &= \mathrm{KL}\left(g(\cdot)||\tilde{f}(\cdot)\right) - \log(F)\end{aligned}$$

Hence, by the nonnegativity of KL divergence, we get that, the objective function has minimum $-\log(F)$, and attains $-\log(F)$ if and only if $g(\cdot) = \tilde{f}(\cdot)$. In particular, this gives that the minimum $g^*(\cdot) \propto f(\cdot)$, as desired. ■

We are now ready to derive an explicit form for the update (3). To do so, we first notice the fact that when $Z \sim q(Z|\lambda) = \Pi_{i=1}^m q(Z_i|\lambda_i)$, the coordinates $\{Z_i\}_{i=1}^m$ are mutually independent and its marginal distribution is $Z_i \sim q(Z_i|\lambda_i)$. Hence, we have the algebraic results below:

$$\begin{aligned}\mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\sum_{i=1}^m \log q(Z_i|\lambda_i)\right] &= \sum_{i\neq i_0} \mathbb{E}_{Z\sim q(\cdot|\lambda)} \log q(Z_i|\lambda_i) + \mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\log q(Z_{i_0}|\lambda_{i_0})\right] \\ &= \sum_{i\neq i_0} \mathbb{E}_{Z_i\sim q_i(\cdot|\lambda_i)} \log q(Z_i|\lambda_i) + \mathbb{E}_{Z_{i_0}\sim q(\cdot|\lambda_{i_0})}\left[\log q(Z_{i_0}|\lambda_{i_0})\right]\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{Z\sim q(\cdot|\lambda)}[\log p(Z|x)] &= \mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\log p(Z_{i_0}|Z_{-i_0}, x) + \log p(Z_{-i_0}|x)\right] \\ &= \mathbb{E}_{Z_{i_0}\sim q(Z_{i_0}|\lambda_{i_0})}\left[\mathbb{E}_{Z_{-i_0}\sim\Pi_{i\neq i_0}q(z_i|\lambda_i)} \log p(Z_{i_0}|Z_{-i_0}, x)\right] + \mathbb{E}_{Z_{-i_0}\sim\Pi_{i\neq i_0}q(z_i|\lambda_i)}[\log p(Z_{-i_0}|x)]\end{aligned}$$

Therefore, one can rewrite the objective in the updating rule (5.7) as

$$\begin{aligned}&\mathbb{E}_{Z\sim q(\cdot|\lambda)}\left[\sum_{i=1}^m \log q(Z_i|\lambda_i) - \log p(Z|x)\right] \\ &= \left(\mathbb{E}_{Z_{i_0}\sim q(\cdot|\lambda_{i_0})}\left[\log q(Z_{i_0}|\lambda_{i_0})\right] - \mathbb{E}_{Z_{i_0}\sim q(Z_{i_0}|\lambda_{i_0})}\left[\mathbb{E}_{Z_{-i_0}\sim\Pi_{i\neq i_0}q(z_i|\lambda_i)} \log p(Z_{i_0}|Z_{-i_0}, x)\right]\right) \\ &\qquad + \left(\sum_{i\neq i_0} \mathbb{E}_{Z_i\sim q_i(\cdot|\lambda_i)} \log q(Z_i|\lambda_i) - \mathbb{E}_{Z_{-i_0}\sim\Pi_{i\neq i_0}q(z_i|\lambda_i)} \log p(Z_{-i_0}|x)\right) \\ &= \mathbb{E}_{Z_{i_0}\sim q(\cdot|\lambda_{i_0})}\left[\log q(Z_{i_0}|\lambda_{i_0})\right] - \mathbb{E}_{Z_{i_0}\sim q(\cdot|\lambda_{i_0})}\left[\mathbb{E}_{Z_{-i_0}\sim\Pi_{i\neq i_0}q(z_i|\lambda_i)} \log p(Z_{i_0}|Z_{-i_0}, x)\right] + \Phi(\lambda_{-i_0}) \qquad (5.9)\end{aligned}$$

for some function $\Phi : \mathbb{R}^{m-1} \to \mathbb{R}$. Since in update step (5.7), λ_{-i_0} is fixed and the step requires optimizing over λ_{i_0}. With expression (5.9), update (5.7) is equivalent to solving the following minimization problem:

$$\underset{\lambda_{i_0}}{\text{minimize}}\ \mathbb{E}_{Z_{i_0}\sim q(\cdot|\lambda_{i_0})}\left[\log q(Z_{i_0}|\lambda_{i_0})\right] - \mathbb{E}_{Z_{i_0}\sim q(\cdot|\lambda_{i_0})}\left[\mathbb{E}_{Z_{-i_0}\sim\Pi_{i\neq i_0}q(z_i|\lambda_i)} \log p(Z_{i_0}|Z_{-i_0}, x)\right] \qquad (5.10)$$

Therefore, if one use $h_{i_0}(\cdot)$ to denote

$$h_{i_0}(\cdot) := \exp\left(\mathbb{E}_{Z_{-i_0}\sim\Pi_{i\neq i_0}q(z_i|\lambda_i)} \log p(\cdot|Z_{-i_0}, x)\right) \qquad (5.11)$$

then we get by lemma 1 that the minimizer for problem (5.10) is $q(\cdot|\lambda_{i_0}^*) = h(\cdot)/H$, where $H := \int h \geq 0$ is the normalizing constant for the function $h(\cdot)$. To summarize, one can replace the update step (5.7) with the following formula:

$$q(\cdot|\lambda_{i_0}^{k+1}) = \frac{1}{\int h_{i_0}} h_{i_0}(\cdot) \qquad (5.12)$$

where h_{i_0} is defined in Eq (5.11).

Remark

1. Although in general, formula (5.12) shows that an exact update for λ_i can sometimes be computationally intractable whenever the normalizing constant $h(\cdot)$ is hard to compute (which is often the case), it also brings some simplications under some situations, as we will see in example 1.

2. Suppose $p(\cdot|x)$ satisfies the structure that for all $1 \leq i \leq m$, $p(z_i|z_{-i}, x)$ belongs to exponential family, in the sense that,
$$p(z_i|z_{-i}, x) = h_i(z_i) \exp\left(\eta_i(z_{-i}, x)z_i - \phi_i(z_{-i}, x)\right)$$
for some function $\{\eta_i\}_{i=1}^m$ and $\{\phi_i\}_{i=1}^m$. In this case, the explicit updating rule (5.12) shows that, we have
$$q(z|\lambda_{i_0}^*) \propto \exp\left(\mathbb{E}_{Z_{-i_0}} \log p(z|z_{-i}, x)\right) \propto h_{i_0}(z_{i_0}) \exp\left(z\mathbb{E}_{Z_{-i_0}}[\eta_i(z_{-i}, x)]\right)$$

Example 1: In this example, we consider the hierachical Gaussian mixture model. Let $\{\mu_k\}_{k=1}^K$ be i.i.d following $\mathsf{N}(0, \sigma^2)$ for some known σ^2 and let $\{C_i\}_{i=1}^N$ be i.i.d following multinomial distribution with parameter $(\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K}) = \frac{1}{K}\mathbf{1}_K^T$. Conditioning on the hidden variables $Z = \left(\{\mu_k\}_{k=1}^K, \{C_i\}_{i=1}^N\right)$, the observed samples $\{X_i\}_{i=1}^N$ has conditional distribution $\mathsf{N}(\mu_{C_i}, 1)$. Overall, the above data generating process can be viewed as a hierachical gaussian mixture model, i.e.,

$$\mu_k \sim \mathsf{N}\left(0, \sigma^2\right), \quad C_i \sim \text{Mult}\left(\frac{1}{K}\mathbf{1}_K^T\right), \quad X_i|C_i, \mu \sim \mathsf{N}\left(\mu_{C_i}, 1\right)$$

The goal is to use the observed data X to infer the posterior distributions of $\{\mu_k\}_{k=1}^N$ and $\{C_i\}_{i=1}^K$. To do so, consider using the following class of candidates to approximate the posterior distribution of $\{\mu_k, C_i\}$,

$$q(\mu, c|x) = \Pi_{k=1}^K q_k(\mu_k)\Pi_{i=1}^N q^i(C_i)$$

As is discussed in previous section, one might wish to apply the coordinate descent technique to find the optimal density function $\{q_k\}_{k=1}^K$ and $\{q^i\}_{i=1}^N$ that best approximate the true posterior distribution. In order to implement the CD algorithm, one crucial step is to compute the update rule via Eq (5.7), or equivalently, via Eq (5.12). For instance, to compute the update rule for $\{q_i\}_{i=1}^N$, using Eq (5.12), the updating rule follows that,

$$q^i(c) \propto \exp\left(\mathbb{E}_{C_{-i},\mu}\left[\log p(c|C_{-i}, \mu, X)\right]\right) \propto \exp\left(\mathbb{E}_{C_{-i},\mu}\left[\log p(c, C_{-i}, \mu, X)\right]\right) \tag{5.13}$$

Now notice that, according to our model assumption, we have,

$$p(C, \mu, X) = \Pi_{i=1}^K p_i(\mu_i)\Pi_{i=1}^N p^i(C_i)\Pi_{i=1}^N \phi(X_i|\mu_i, C_i)$$

where in above, $p_i(\cdot)$ denotes the marginal distribution of μ_i, $p^i(\cdot)$ denotes the marginal distribution of C_i and $\phi(\cdot|\mu, c)$ denotes the density function for $\mathsf{N}(\mu_c, 1)$. Therefore, we have,

$$\log p(C_i, C_{-i}, \mu, X) = \log p^i(c_i) + \log \phi(X_i|C_i, \mu) + \Psi(C_{-i}, \mu, X) \tag{5.14}$$

for some function Ψ. As $p^i(C_i) = 1/K$ by our model assumption, by plugging Eq (5.14) into Eq (5.13), one gets that,

$$q^i(c) \propto \exp\left(\mathbb{E}_{C_{-i},\mu}\left[\log p(c, C_{-i}, \mu, X)\right]\right) \propto \exp\left(\mathbb{E}_{C_{-i},\mu} \log \phi(X_i|c, \mu)\right) \propto \exp\left(\mathbb{E}_{C_{-i},\mu}\left[\mu_c - \frac{1}{2}\mu_c^2\right]\right)$$

So if one restricts $\{q_k(\cdot)\}_{k=1}^K$ to take the form of normal density with mean m_k and variance s_k^2, the update rule for q^i becomes

$$q^i(c) \propto \exp\left(m_c - \frac{1}{2}\left(m_c^2 + s_c^2\right)\right)$$

As $\{C_i\}_{i=1}^k$ is supported on $\{1, 2, \ldots, K\}$, it is natural to let $q^i(\cdot)$ take the form of multinomial density with parameter $\{\tau_{i,k}\}_{k=1}^K$. In this sense, one gets the update rule for $\{\tau_{i,k}\}_{k=1}^K$. The update rule for (m_k, s_k^2) can be similarly computed (we leave this actual derivation as an exercise for the reader, and its computation does not require more than a few lines)

$$m_k = \frac{\sum_{i=1}^N \tau_{i,k} X_i}{\frac{1}{\sigma^2} + \sum_{i=1}^N \tau_{i,k}} \quad \text{and} \quad s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_{i=1}^N \tau_{i,k}}$$

An interesting fact here is that, if one starts to initialize $q^i(\cdot)$ with some multinomial distribution and $q_i(\cdot)$ with some normal distribution, then the update will maintain those $q^i(\cdot)$ to be multinomial and $q_i(\cdot)$ to be normal, though the parameters for the distributions keep updating! ♣