10-708: Probabilistic Graphical Models, Spring 2020

8: Variational Inference 2

Lecturer: Xun Zheng *Scribe: Zhaoqi Cheng, Eric Dexheimer, Yifei Li, Chiyu Wu, Junwoong Yoon*

1 Practice of Variational Inference

1.1 Stochastic Variational Inference

1.1.1 Drawback of Coordinate Ascent

In the previous lecture on variational inference with coordinate ascent, there were two types of parameters to update: the local variables, which were documents, and the global parameters λ. The algorithm is summarized in a high level in Algorithm 1:

Algorithm 1: Coordinate Ascent

```
Initialize global parameters λ;
repeat
    for each document d ∈ {1, 2, ..., D} do
        Update document-specific variational distributions;
    end
    Update global parameters λ;
until Lower bound L(λ) converges;
```

As in the algorithm, all documents are looped through first before updating the global parameters. This becomes a problem when there are too many documents, so the updating of the model will be infrequent.

To combat this, the lower bound can be separated into a per-data point term parameterized by local parameters ϕ_i, and a global term parameterized by the global parameter λ:

$$\mathcal{L}(\lambda, \phi_{1:n}) = \underbrace{\mathbb{E}_q\left[\log p(\beta) - \log q(\beta|\lambda)\right]}_{\text{global contribution}} + \sum_{i=1}^{n} \underbrace{\left\{\mathbb{E}_q\left[\log p(w_i, z_i|\beta] - \log q(z_i|\phi_i)\right\}\right.}_{\text{per-data point contribution}}. \tag{1}$$

Let

$$f(\lambda) \coloneqq \mathbb{E}_q\left[\log p(\beta) - \log q(\beta|\lambda)\right] \tag{2}$$

be the global contribution, and

$$g_i(\lambda, \phi_i) \coloneqq \mathbb{E}_q\left[\log p(w_i, z_i|\beta] - \log q(z_i|\phi_i)\right. \tag{3}$$

be the per-data point contribution of the i-th data point. Then the lower bound can be simplified as

$$\mathcal{L}(\lambda, \phi_{1:n}) = f(\lambda) + \sum_{i=1}^{n} g_i(\lambda, \phi_i). \tag{4}$$

To optimize our objective, we can maximize it w.r.t. parameters $\phi_{1:n}$ first, which results in a one-parameter lower bound

$$\mathcal{L}(\lambda) = f(\lambda) + \sum_{i=1}^{n} \max_{\phi_i} g_i(\lambda, \phi_i). \tag{5}$$

Let the optimizer for each data point be

$$\phi_i^* = \arg\max_{\phi} g_i(\lambda, \phi_i), \tag{6}$$

then the gradient of the one-parameter lower bound $\mathcal{L}(\lambda)$ has the following form,

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = \frac{\partial f(\lambda)}{\partial \lambda} + \sum_{i=1}^{n} \frac{\partial g_i(\lambda, \phi_i)}{\partial \lambda}. \tag{7}$$

Since the per-data point term is a sum of each data point contribution, the derivative of each data-point can be summed to compute the derivative of the per-data point term as well, as shown in the second term of RHS of equation 7. This allows us to use stochastic gradient algorithms and update the model more frequently for better convergence. And once the global parameter λ is estimated, each ϕ_i can be estimated online if needed.

1.1.2 Stochastic Variational Inference using Natural Inference

However, the gradient of the lower bound with respect to the parameters will not be the steepest direction, since the parameters represent a distribution. For example, a pair of Gaussian distributions with the same two means will have the same "distance" in the mean parameter space, but the KL divergence is very different, as seen by the overlap in Figure 1. In addition, parameterizing the spread using variance or the inverse variance would lead to different gradients due to the difference in representation.

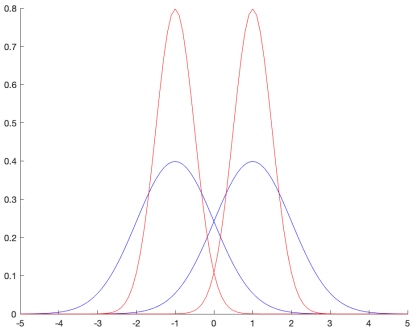

Figure 1: Two pairs of Gaussian distributions with the same two means

The reason is, using gradient algorithms, the objective function $f(\theta)$ at θ_0 is approximated by a local quadratic second-order approximation

$$f(\theta) \approx f(\theta_0) + \nabla f(\theta_0)^\top (\theta - \theta_0) + \frac{1}{2t}(\theta - \theta_0)^\top (\theta - \theta_0), \tag{8}$$

where $\nabla^2 f(\theta_0)$ is replaced by $\frac{1}{t} I$. But this approximation will not work for distributions. Instead, we use the Fisher Information Matrix to for the gradient since the variational distribution is in the exponential family.

Now, we can perform stochastic variational inference by sampling data points and updating the parameters using the natural gradient as in Algorithm 2.

Algorithm 2: Stochastic Variational Inference using Natural Inference

Initialize global parameters λ_0, $t = 0$;
Set step-size schedule ρ_t;
for $t = 1, \ldots, \infty$ **do**
  Sample a data point $i \sim \text{Uniform}(1, \ldots, n)$;
  Compute the optimal local parameter $\phi_i^*(\lambda_t)$;
  Perform natural gradient ascent on global parameter λ,
    $\lambda_{t+1} \leftarrow \lambda_t + \rho_t g(\lambda_t) = (1 - \rho_t)\lambda_t + \rho_t \left(\eta + nt_{\phi_i^*}(x_i)\right)$;
end

1.2 Black-box Variational Inference (BBVI)

We have derived variational inference specific for Latent Dirichlet allocation (LDA) above. But there are innumerable conjugate/non-conjugate models. We want to have a general solution which does not entail model-specific work. This solution will serve like a black box, which outputs a variational distribution when input any model and massive data. It is called Black-box Variational Inference (BBVI).

There are generally two types of BBVI: BBVI with the score gradient, and BBVI with the reparameterization gradient. The latter is the foundation of Variational AutoEncoder (VAE), which will be discussed more in lecture 12.

1.2.1 BBVI with the Score Gradient

Consider a probabilistic model where x is the observed variable and z is the latent variable. The corresponding variational distribution is $q(z|\lambda)$. The evidence lower bound (ELBO) is

$$\mathcal{L}(\lambda) \coloneqq \mathbb{E}_{q_\lambda(z)} \left[\log p(x, z) - \log q(z|\lambda)\right]. \tag{9}$$

Its gradient w.r.t. λ is

$$\nabla_\lambda \mathcal{L} = \mathbb{E}_q \left[\nabla_\lambda \log q(z|\lambda)(\log p(x, z) - \log q(z|\lambda))\right], \tag{10}$$

where $\nabla_\lambda \log q(z|\lambda)$ is called the *score function*.

To derive equation 10, we need two essential facts:

1. $\nabla_\lambda q_\lambda(z) = \frac{1}{q_\lambda(z)} \nabla_\lambda q_\lambda(z) = q_\lambda(z) \nabla_\lambda q_\lambda(z)$;
2. $\mathbb{E}_q \left[\nabla_\lambda \log q_\lambda(z)\right] = 0$, i.e. the expectation of the gradient of log-likelihood (score function) is zero.

Based on these two facts, we can derive the score gradient of our ELBO,

$$
\begin{aligned}
\nabla_\lambda \mathcal{L} &= \nabla_\lambda \int_z \left[q_\lambda(z) \log p(x,z) - q_\lambda(z) \log q_\lambda(z)\right] \mathrm{d}z \\
&= \int_z \left\{ \log p(x,z) \nabla_\lambda q_\lambda(z) - \left[\nabla_\lambda q_\lambda(z) \log q_\lambda(z) + q_\lambda(z) \frac{1}{q_\lambda(z)} \nabla_\lambda q_\lambda(z) \right] \right\} \mathrm{d}z \\
&= \int_z \nabla_\lambda q_\lambda(z) \left[\log p(x,z) - \log q_\lambda(z) - 1\right] \mathrm{d}z \\
&= \int_z q_\lambda(z) \nabla_\lambda q_\lambda(z) \left[\log p(x,z) - \log q_\lambda(z) - 1\right] \mathrm{d}z \\
&= \mathbb{E}_{q_\lambda} \left[\nabla_\lambda q_\lambda(z) \left(\log p(x,z) - \log q_\lambda(z)\right)\right] - \mathbb{E}_{q_\lambda} \left[\nabla_\lambda q_\lambda(z)\right] \\
&= \mathbb{E}_{q_\lambda} \left[\nabla_\lambda q_\lambda(z) \left(\log p(x,z) - \log q_\lambda(z)\right)\right].
\end{aligned}
$$

Using the score gradient in equation 10, We can take advantage of Monte Carlo to compute the noisy unbiased gradient of the ELBO with samples from the vaiational distribution,

$$
\nabla_\lambda \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z_s|\lambda)(\log p(x,z_s) - \log q(z_s|\lambda)), \tag{11}
$$

where $z_s \sim q(z|\lambda)$.

2 Theory of Variational Inference

We have seen how variational inference can be viewed as an optimization problem in which we approximate the true solution by relaxing/approximating the intractable optimization problem. Now, a unified view based on the variational principle is presented where the mean-field approximation is an inner approximation of the set of canonical parameters, while loopy belief propagation is an outer approximation.

Thus far, common inference problems include marginal distributions and normalization constants, but now we want to represent these quantities in a variational form via exponential families and convex analysis.

2.1 Graphical Models as Exponential Families

The exponential family described previously is:

$$
p_\theta(x_1, ..., x_n) = \exp\left(\theta^T \phi(x) - A(\theta)\right), \tag{12}
$$

where θ is the set of canonical parameters, $\phi(x)$ are the sufficient statistics, and $A(\theta)$ is the log partition function for normalization,

$$
A(\theta) = \log \int \exp\left\{\theta^\top \phi(x)\right\} \mathrm{d}x.
$$

Note that whatever distribution p_θ, the log partition function $A(\theta)$ is a convex function, which can be proved via its Hessian. There is also a constraint for domain of the effective canonical parameters that the value of the log partition function $A(\theta)$ must be finite:

$$
\Omega \coloneqq \left\{\theta \in \mathbb{R}^d | A(\theta) < +\infty\right\}.
$$

Both Gaussian MRFs and discrete MRFs can be represented in this way. The Gaussian MRF has a joint distribution

$$p(\mathbf{x}) = \exp\left\{\frac{1}{2}\langle\Theta, \mathbf{x}\mathbf{x}^T\rangle - A(\Theta)\right\}, \text{ where } \Theta = -\Lambda, \tag{13}$$

with sufficient statistics

$$\{x_s^2, s \in V; x_s x_t, (s,t) \in E\}. \tag{14}$$

The discrete MRF joint distribution can be parameterized as

$$p(\mathbf{x};\theta) = \exp\left\{\underbrace{\sum_{s\in V}\sum_{j}\theta_{s;j}\mathbb{I}_j(x_s)}_{\text{marginal potential}} + \underbrace{\sum_{(s,t)\in E}\theta_{st;jk}\mathbb{I}_j(x_s)\mathbb{I}_k(x_t)}_{\text{pair-wise potential}}\right\}, \tag{15}$$

where $\mathbb{I}$ denotes the binary indicator function. The indicators can also be regarded as one-hot variables. One example of such distribution is the Ising model.

The exponential family is useful because computing the expectation over sufficient statistics given canonical parameters yields the marginals, which is called the marginal/pair-wise inference:

$$\mu_{s;j} = \mathbb{E}_p[\mathbb{I}_j(X_s)] = \mathbb{P}[X_s = j] \quad \forall j \in \mathcal{X}_s, \tag{16}$$

$$\mu_{st;jk} = \mathbb{E}_p[\mathbb{I}_{st;jk}(X_s, X_t)] = \mathbb{P}[X_s = j, X_t = k] \quad \forall (j,k) \in \mathcal{X}_s \in \mathcal{X}_t. \tag{17}$$

The sufficient statistics and canonical parameters are duals, in that one can be computed from the other. In addition, the normalizer yields the log partition function:

$$\log Z(\theta) = A(\theta). \tag{18}$$

The above duality is presented for the Bernoulli distribution whose joint distribution is

$$p(x;\theta) = \exp\left(\theta x - A(\theta)\right), x \in \{0,1\}, A(\theta) = \log(1 + e^\theta), \tag{19}$$

and the inference (computing the mean parameter) is

$$\mu(\theta) = \mathbb{E}_\theta[X] = 1 \cdot p(X = 1;\theta) + 0 \cdot p(X = 0;\theta) = \frac{e^\theta}{1 + e^\theta}. \tag{20}$$

Now we want to formulate the inference in a variational manner, i.e. as an optimization problem.

2.2 Conjugate Dual

First, we define the conjugate dual function for a function $f(\theta)$:

$$f^*(\mu) := \sup_\theta\{\langle\theta, \mu\rangle - f(\theta)\} \tag{21}$$

A visual representation of the conjugate dual can be seen in Figure 2. The conjugate dual is always a convex function since it is a point-wise supremum of a class of linear functions. Note that the dual is the dual of itself if f is convex and lower semi-continuous, i.e.

$$f(\theta) = \sup_\mu\{\langle\theta, \mu\rangle - f^*(\mu)\}. \tag{22}$$

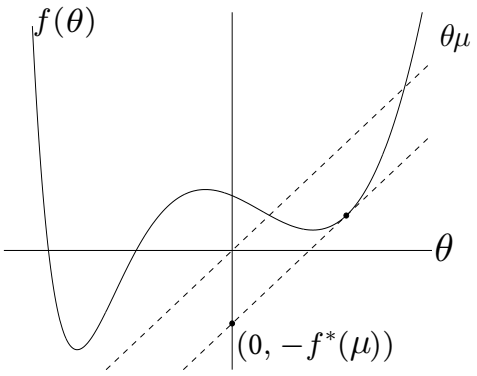

Figure 2: Example of conjugate dual function

In this sense the log partition function can be written this way under its domain:

$$A(\theta) = \sup_{\mu}\{\langle\theta, \mu\rangle - f^*(\mu)\}, \quad \theta \in \Omega. \tag{23}$$

This way, the dual variable μ has a natural interpretation as the mean parameter.

2.2.1 Example: Computing Mean Parameter of Bernoulli Distribution

Firstly consider a simple example of the Bernoulli distribution. Using its partition function in equation 19, its conjugate dual is

$$A^*(\mu) \coloneqq \sup_{\theta\in\mathbb{R}} \{\mu\theta - \log\left[1 + \exp(\theta)\right]\}. \tag{24}$$

It's kind of cheating because we have already known the form of $A(\theta)$. Solving the equation we write down the stationary condition which the optimal θ will satisfy

$$\mu = \frac{e^{\theta^*}}{1 + e^{\theta^*}}. \tag{25}$$

Inversing this condition we know the θ^* is

$$\theta^* = \log\frac{\mu}{1-\mu}, \ \mu \in (0, 1). \tag{26}$$

Plug this optimal θ^* into equation 24,

$$A^*(\mu) = \begin{cases} \mu\log\mu + (1-\mu)\log(1-\mu), & \text{if } \mu \in (0,1), \\ +\infty, & \text{otherwise.} \end{cases} \tag{27}$$

As we see here, $A^*(\mu)$ can only be defined on a restricted domain of μ. Using the dual of dual, we can solve the original partition function $A(\theta)$,

$$A(\theta) = \sup_{\mu\in(0,1)} \{\mu\cdot\theta - A^*(\mu)\}. \tag{28}$$

The optimum is achieved at

$$\mu(\theta) = \frac{e^{\theta}}{1 + e^{\theta}}, \tag{29}$$

which is exactly the mean.

2.2.2 Computation of Conjugate Dual

The example of Bernoulli is by no means practical. It's just for purpose of understanding. But in general, the process of the variational representation is similar. Given a distribution of exponential family

$$p(x;\theta) = \exp\left\{\sum_{i=1}^{d}\theta_i\phi_i(x) - A(\theta)\right\}, \tag{30}$$

we firstly write down its dual function

$$A^*(\theta) \coloneqq \sup_{\theta\in\Omega}\left\{\langle\mu,\theta\rangle - A(\theta)\right\}. \tag{31}$$

In order to represent this dual function, the optimal θ^* should satisfy the stationary condition

$$\mu - \nabla A(\theta^*) = 0. \tag{32}$$

Note that the derivatives of A essentially yields the mean parameters, i.e.

$$\frac{\partial A}{\partial\theta_i}(\theta) = \mathbb{E}_\theta\left[\phi_i(x)\right] = \int\phi_i(x)p(x;\theta)\mathrm{d}x. \tag{33}$$

Thus the stationary condition becomes

$$\mu = \mathbb{E}_{\theta^*}\left[\phi(x)\right]. \tag{34}$$

However, we should be careful that μ is restricted to a domain Ω. Now assume there is a solution $\theta(\mu)$ s.t. $\mu = \mathbb{E}_\theta\left[\phi(X)\right]$, then the dual takes the form

$$A^*(\mu) = \langle\theta(\mu),\mu\rangle - A(\theta(\mu)) = \mathbb{E}_{\theta(\mu)}\left[\langle\theta(\mu),\phi(x)\rangle - A(\theta(\mu)\right] \tag{35}$$

$$= \mathbb{E}_{\theta(\mu)}\left[\log p(x;\theta(\mu))\right], \tag{36}$$

which is derived by plugging in the optimal $\theta(\mu)$ and the definition of the joint probability of this exponential distribution.

Recall that the entropy is defined as,

$$H(p(x)) = -\int p(x)\log p(x)\mathrm{d}x, \tag{37}$$

so the dual is

$$A^*(\mu) = -H(p(x;\theta(\mu))), \tag{38}$$

when there is such a solution $\theta(\mu)$.

2.3 Marginal Polytope

Though we provide a framework for general exponential families/graphical models above, there are still several problems. Computing the entropy is generally intractable, and the constrain set Ω of mean parameter is hard to characterize. For the latter one, the answer is the marginal polytope.

For any distribution $p(x)$ (which does not necessarily belong to exponential family) and a set of sufficient statistics $\phi(x)$, define a vector of mean parameters as

$$\mu_i = \mathbb{E}_p\left[\phi_i(X)\right] = \int\phi_i(x)p(x)dx. \tag{39}$$

The set of all realizable mean parameters $\mathcal{M} := \left\{\mu \in \mathbb{R}^d \mid \exists p \text{s.t.} \mathbb{E}_p[\phi(X)] = \mu\right\}$ is a convex set.

When $p(x)$ belongs to exponential families, $\mathcal{M}$ is called marginal polytope and has the following convex hull representation:

$$\mathcal{M} = \left\{\mu \in \mathbb{R}^d \mid \sum_{x \in \mathcal{X}^m} \phi(x)p(x) = \mu, \text{ for some } p(x) \geq 0, \sum_{x \in \mathcal{X}^m} p(x) = 1\right\} = \text{conv}\left\{\phi(x), x \in \mathcal{X}^m\right\} \tag{40}$$

By Minkowski-Weyl Theorem, any non-empty convex polytope can be characterized by a finite collection of linear inequality constraints

$$\mathcal{M} = \left\{\mu \in \mathbb{R}^d | a_j^\top \mu \geq b_j, \forall j \in \mathcal{J}\right\} \tag{41}$$

where $|\mathcal{J}|$ is finite.

2.4 Variational Principle and Approximation

The marginal polytope is constrained by a set of half-planes. The number of such half-planes is called *facet complexity*. For a tree graphical model, the facet complexity grows only linearly in the graph size. However, for a general graph it grows so fast that it's extremely hard to characterize the marginal polytope. Thus we want to introduce some variational method.

As shown before, the dual function takes the form:

$$A^*(\mu) = \begin{cases} -H\left(p_{\theta(\mu)}\right) & \text{if } \mu \in \mathcal{M}^\circ \\ +\infty & \text{if } \mu \notin \overline{\mathcal{M}} \end{cases}, \tag{42}$$

where $\theta(\mu)$ satisfies $\mu = \mathbb{E}_{\theta(u)}[\phi(X)]$.

The exact variational formulation is

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \left\{\theta^T \mu - A^*(\mu)\right\}, \tag{43}$$

where $\mathcal{M}$ is the marginal polytope and A^* is the conjugate dual (negative entropy function).

As discussed, the marginal polytope $\mathcal{M}$ is difficult to characterize, while the conjugate dual has no explicit form. To approximate, there are 2 common methods:

1. Mean field approximation method: non-convex inner bound and exact form of entropy;
2. Bethe approximation and loopy belief propagation: polyhedral outer bound and non-convex Bethe approximation.

Here we will focus on the former one.

2.4.1 Mean Field Approximation

For an exponential family with sufficient statistics ϕ defined on an arbitrary graph G, the set of realizable mean parameter set is:

$$\mathcal{M}(G; \phi) = \left\{\mu \in \mathbb{R}^d \mid \exists p \text{ s.t. } E_p[\phi(X)] = \mu\right\} \tag{44}$$

The idea is to restrict p to a subset of distributions associated with a tractable subgraph. For instance, we can transform a general graph with mean parameter set $\Omega = \left\{\theta \in \mathbb{R}^d \mid A(\theta) < +\infty\right\}$ to subgraph F_0 with $\Omega\left(F_0\right) = \left\{\theta \in \Omega \mid \theta_{(s,t)} = 0 \forall (s,t) \in E\right\}$ or T with $\Omega(T) = \left\{\theta \in \Omega \mid \theta_{(s,t)} = 0 \forall (s,t) \in E(T)\right\}$, as illustrated in Figure 3.

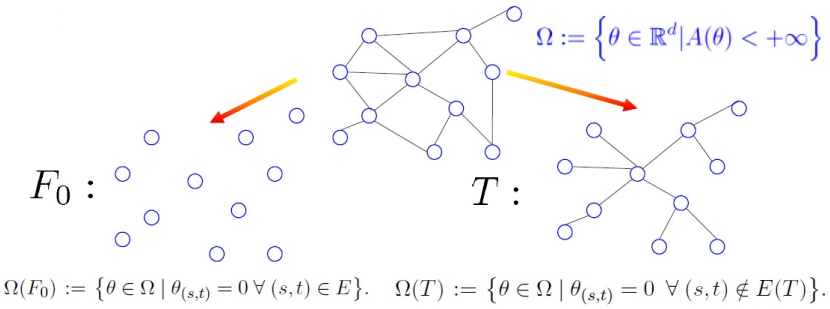

Figure 3: Transform a general graph to tractable subgraph

2.4.2 Geometry of Mean Field

Mean field optimization is always **non-convex** for any exponential family in which the state space $\mathcal{X}^m$ is finite. Recall the marginal polytope $\mathcal{M}(G)$ is a convex hull,

$$\mathcal{M}(G) = \text{conv}\left\{\phi(e); e \in \mathcal{X}^m\right\} \tag{45}$$

and $\mathcal{M}_F(G)$ contains all the extreme points of this polytope. This implies that the convex hull $\mathcal{M}(G)$ must be non-convex if $\mathcal{M}_F(G)$ is a strict subset of the convex hull.
For example, let's consider a two-node Ising model:

$$\mathcal{M}_F(G) = \{\tau_1, \tau_2 \in [0,1] s.t. \tau_{12} = \tau_1\tau_2\} \tag{46}$$

This model has a parabolic cross section along $\tau_1 = \tau_2$, hence it is non-convex.