CPSC 486/586: Probabilistic Machine Learning	February 21, 2023
Lecture 11	
Lecturer: Andre Wibisono	*Scribe: Edward Frazer*

1 Exploring Variational Bayes and Connections to ELBO

1.1 Recap From Lecture 10

Suppose we have some target distribution $\nu(x)$ (e.g. $\nu(x) = p(x|y)$ is a Bayesian posterior). Our question: How can we approximate $\nu(x)$ (or $\mathbb{E}_\nu[x]$)? We've gone over some potential methods.

1. Approximate with $\hat{\nu} \in Q$, where Q is some nice family of distributions. One example of this is the Laplace approximation, where we have that $\hat{\nu} = \mathcal{N}(x^*, C^*)$. However, this example is not optimal, and so we look to variational (optimal) approximation.
2. (a) We do expectation propagation, e.g. solving this problem: $\min_{\rho \in Q} \mathbf{KL}(\nu||\rho)$.

   (b) We do variational inference, e.g. solving this problem: $\min_{\rho \in Q} \mathbf{KL}(\rho||\nu)$.

1.2 Optimal VB When $Q = \mathcal{G}$

Say our nice space of functions Q is limited to $\mathcal{G} = \{\mathcal{N}(m, c) : m \in \mathbb{R}^d, C \in \mathbb{R}^{dxd} \wedge C \succ 0\}$. Our objective function is $F(\rho) = \mathbf{KL}(\rho||\nu) = \mathbf{KL}(\mathcal{N}(m, c)||\nu)$. We claim the following.

Claim: The following ODE is the gradient flow for $\min F(m, c)$ under BW-distance.

$$\dot{m}_t = -\mathop{\mathbb{E}}_{\mathcal{N}(m_t, C_t)}[\nabla f]$$

$$\dot{C}_t = 2\left(I - C_t \mathop{\mathbb{E}}_{\mathcal{N}(m_t, C_t)}[\nabla^2 f]\right)$$

Thm [Lambert et al. '22]: If $\nu(x) \propto e^{-f(x)}$ is α-SLC ($\iff$ f is α strongly convex) then $F(\rho) = \mathbf{KL}(\rho||\nu)$ is α-strongly convex in $\mathcal{G}$ with BW-Metric $W_2(\mathcal{N}(m_1, C_1), \mathcal{N}(m_2, C_2))^2$. If $d = 1$, the metric becomes $(m_1 - M_2)^2 + (\sqrt{C_1} - \sqrt{C_2})^2$. This fact gives us exponential convergence guarantees. Specifically: $W_w(\rho_t, \hat{\nu})^2 \leq e^{-2\alpha t} W_2(\rho_0, \hat{\nu})$.

1.3 Connections to ELBO

We can view VB as maximizing ELBO (Evidence Lower Bound). Suppose we have some prior $p(x)$, a likelihood $p(y|x)$, and a posterior $p(x|y) = \nu(x)$. This implies a joint distribution $p(x, y) = p(x)p(y|x)$. We define the evidence as $p(y) = \int_X p(x, y)dx$. We also note that $p(x|y) = \frac{p(x,y)}{p(y)}$. Then we can perform VB with some approximating method ($q(x)$) of the target distribution (the posterior, $p(x|y)$).

$$
\begin{aligned}
\mathbf{KL}(q(x)||p(x|y)) &= \int_x q(x) \log\left(\frac{q(x)}{p(x|y)}\right) dx \\
&= \int_X q(x) \log\left(\frac{q(x) \cdot p(y)}{p(x,y)}\right) dx \\
&= -\int_X q(x) \log\left(\frac{p(x,y)}{q(x)}\right) dx + \int_X q(x) \log(p(y)) dx \\
&= -\mathbf{ELBO}(y, q) + \log(p(y))
\end{aligned}
$$

Where $Q(x) \in Q$. We can proceed to defining ELBO as follows:

Definition of ELBO

$$\mathbf{ELBO}(y, q) = \mathbb{E}_q\left[\log\left(\frac{p(x,y)}{q(x)}\right)\right]$$

where y is constituted of our observations and q is our approximating distribution.

Lemma:

$$\mathbf{ELBO}(y, q) = \log(p(y)) - \mathbf{KL}(q||p(\cdot|y))$$

where term 1 is our evidence and term 2 is our objective to minimize in VB. We finally get the relation:

$$\arg\min_{q \in Q} \mathbf{KL}(q||p(\cdot|y)) = \arg\max_{q \in Q} \mathbf{ELBO}(y, q)$$

2 Moving to a New Method: Sampling

2.1 Introduction to Sampling

We now look to a new method, where we seek to approximate some $\nu(x)$ on $\mathbb{X}$ by drawing *samples* (X) from ν s.t. $X \sim \nu$. We have a number of techniques to do so:

1. Markov Chain Monte Carlo (MCMC) method.
2. Random Walks
3. Metropolis-Hastings
4. Longevin Algorithm

2.2 Trying to Emulate Categorical Distributions

We first analyze an example where $\mathbb{X} = \{0,1\}$. Assume we can draw a sample from $\mathbf{Uniform}(\{0,1\}) = \mathbf{Ber}(\frac{1}{2})$ (e.g. a Bernoulli distribution where $p = \frac{1}{2}$). We claim the following: **Given our uniform sampling, we can sample from:**

1. $\mathbf{Ber}(p)$, $\forall 0 \leq p \leq 1$
2. Some categorical distribution $p_1, ...p_n$ where $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$.

Solution (Algorithm)
We note that p can be decomposed into a bitstring s.t. $p = 0.b_1 b_2 b_3 ... b_n$

1. Flip fair coin $X_1, X_2, ...X_n \sim \mathbf{Unif}(\{0,1\})$ until $X_n = 1$.
2. Return b_n

We essentially flip a fair coin until we get 1 (heads) and return the value of the bitstring at this index. This encodes a new random variable X^* over space $\{0,1\}$. We claim that $X^* \sim \mathbf{Ber}(p)$. Note the following manipulation:

$$Pr(X^* = 1) = \sum_{n=1}^{\infty} \frac{1}{2^n} \cdot \mathbb{1}\{b_n = 1\}$$
$$= \sum_{n=1}^{\infty} \frac{b_n}{2^n} = \sum_{n=1}^{\infty} b_n \cdot 2^{-n} = p$$

where the last equality holds due to the definition of a bitstring.