CPSC 486/586: Probabilistic Machine Learning	February 13, 2023
Lecture 9	
Lecturer: Andre Wibisono	*Scribe: Shivesh Mehrotra*

1 Recap of Expectation Propagation

Recall the problem of Expectation Propagation:

$$\min_{\rho \in Q} \mathrm{KL}(\nu || \rho) \iff \max_{\rho \in Q} \mathbb{E}_\nu[\log \rho]$$

where ν is our target distribution and Q is a family of distributions we choose to estimate ν. In particular, we have shown that if $Q = \{q_\theta(x) = \exp(\langle \theta, T(x) \rangle - A(\theta)) : \theta \in \Theta\}$, that is Q is an exponential family, then expectation propagation is equivalent to

$$\min_{\theta \in \Theta} (A(\theta) - \langle \theta, \mathbb{E}_\nu[T(x)] \rangle)$$

and the minimizer θ^* satisfies moment matching:

$$\mathbb{E}_{q^*(\theta)}[T(x)] = \mathbb{E}_\nu[T(x)].$$

We now move on to the problem of variational inference.

2 Variational Inference

We now consider the problem:

$$\min_{\rho \in Q} \mathrm{KL}(\rho || \nu).$$

For today's discussion we consider the Gaussian family, $Q = \{\mathcal{N}(m, C) : m \in \mathbb{R}^d, C \in \mathbb{R}^{d \times d}\}$, and that $\nu(x) = \frac{e^{-f(x)}}{z}$ where $z = \int_X e^{-f(x)} dx$ is the normalizing constant. In this regime, where $\rho \sim \mathcal{N}(m, C)$ we have that

$$\begin{aligned}
\mathrm{KL}(\rho || \nu) &= \int \rho \log \frac{\rho}{\nu} \\
&= \int \rho \log \rho - \int \rho \log \nu \\
&= -H(\rho) + \mathbb{E}_\rho[f] + \log(z) \\
&= -\frac{d}{2} \log(2\pi) - \frac{1}{2} \log(\det C) + \mathbb{E}_\rho[f] + \log(z).
\end{aligned}$$

Thus we have objective function

$$F(m, C) \coloneqq -\frac{1}{2}\log(\det C) + \mathbb{E}_{\mathcal{N}(m,C)}[f].$$

For simplicity, we consider the case of $d = 1$ and note that the multivariate case follows analogously. We note the following relationships:

$$\begin{aligned}
\rho(x) &= e^{\frac{-(x-m)^2}{2C}}/\sqrt{2\pi C} \\
\frac{\partial \rho}{\partial m} &= \frac{-(m-x)}{C}\rho(x) \\
\frac{\partial \rho}{\partial x} &= \frac{-(x-m)}{C}\rho(x) \\
\frac{\partial \rho}{\partial x} &= -\frac{\partial \rho}{\partial m}.
\end{aligned}$$

Now returning to our objective function we have that

$$\begin{aligned}
\frac{\partial F}{\partial m} &= \frac{\partial}{\partial m}\int_{\mathbb{R}} \rho(x) f(x) dx \\
&= \int_{\mathbb{R}} \frac{\partial \rho}{\partial m} f(x) dx && \text{(Dominated Convergence Theorem)} \\
&= -\int_{\mathbb{R}} \frac{\partial \rho}{\partial x} f(x) dx \\
&= -\left(\rho(x) f(x)|_{-\infty}^{\infty} - \int_{\mathbb{R}} \rho(x) f'(x) dx\right) \\
&= \int_{\mathbb{R}} \rho(x) f'(x) \\
&= \mathbb{E}_\rho[f'(x)].
\end{aligned}$$

So in general at the minimizer m^* we have that

$$\mathbb{E}_\rho[\nabla f(x)] = 0.$$

Now to find C^*:

$$\begin{aligned}
\frac{\partial F}{\partial C} &= \frac{\partial}{\partial C}\left(-\frac{1}{2}\log(C) + \int_{\mathbb{R}} \rho(x) f(x) dx\right) \\
&= \int_{\mathbb{R}} \frac{\partial \rho}{\partial C}(x) f(x) dx - \frac{1}{2C} \\
&= \int_{R} \rho(x)\left(\frac{(x-m)^2}{2C^2} - \frac{1}{2C}\right) f(x) dx - \frac{1}{2C} \\
&= \frac{1}{2C^2}\underbrace{\mathbb{E}_\rho[f(x)(x-m)^2]} - \frac{1}{2C}(1 + \mathbb{E}_\rho[f(x)]).
\end{aligned}$$

Now manipulating the underlined term by using integration by parts twice we have:

$$\mathbb{E}_\rho[f(x)(x-m)^2] = C\mathbb{E}_\rho[f(x)] + C^2\mathbb{E}_\rho[f''(x)].$$

Thus

$$\begin{aligned}\frac{\partial F}{\partial C} &= \frac{1}{2C^2}\mathbb{E}_\rho[f(x)(x-m)^2] - \frac{1}{2C}(1+\mathbb{E}_\rho[f(x)]) \\ &= \frac{1}{2C^2}(C\mathbb{E}_\rho[f(x)] + C^2\mathbb{E}_\rho[f''(x)]) - \frac{1}{2C}(1+\mathbb{E}_\rho[f(x)]) \\ &= \frac{1}{2}(\mathbb{E}_\rho[f''(x)] - \frac{1}{C}).\end{aligned}$$

In general,

$$\frac{\partial F}{\partial C} = \frac{1}{2}(\mathbb{E}_\rho[\nabla^2 f(x)] - C^{-1}).$$

In summary for variational inference (VI):

$$\begin{aligned}\frac{\partial F}{\partial m} &= \mathbb{E}_\rho[\nabla f(x)] & \frac{\partial F}{\partial C} &= (\mathbb{E}_\rho[\nabla^2 f(x)] - C^{-1}) \\ m^* &\implies \mathbb{E}_\rho[\nabla f(x)] = 0 & C^* &\implies \mathbb{E}_\rho[\nabla^2 f(x)] = (C^*)^{-1}.\end{aligned}$$

For expectation propagation (EP):

$$m^* = \mathbb{E}_\nu[x] \qquad\qquad C^* = \mathrm{Cov}_\nu[x].$$

For Laplace:

$$m^* = x^* = \mathrm{argmin}_x\, f(x) \qquad\qquad C^* = (\nabla^2 f(x^*))^{-1}.$$

The gradient equations which pop out of variational inference naturally lend themselves to Gradient Flow dynamics in order to find the minimizer. In the next lecture, we discuss this and a better ODE to follow.