**CPSC 486/586: Probabilistic Machine Learning** February 15, 2023

Lecture 10

*Lecturer: Andre Wibisono* *Scribe: Nhi Nguyen*

# 1 Problem: Gaussian Variational Inference

Let $\mathcal{P}_2\left(\mathbb{R}^d\right)$ denote the set of all probability distributions on $\mathbb{R}^d$ with finite second moment, i.e., $\mathbb{E}\|x\|^2 < \infty$, and let $\mathcal{G}$ denote the set of all non-degenerate Gaussian distributions on $\mathbb{R}^d$.

$$\mathcal{P}_2\left(\mathbb{R}^d\right) = \left\{\rho : \rho \text{ is a probability distribution on } \mathbb{R}^d \text{ with } \mathop{\mathbb{E}}_{x\sim\rho}\left[\|x\|^2\right] < \infty\right\}$$
$$\mathcal{G} = \mathsf{BW}\left(\mathbb{R}^d\right) = \left\{\rho : \rho = \mathcal{N}(m, C) \text{ for some } m \in \mathbb{R}^d, C \succ 0 \in \mathbb{R}^{d\times d}\right\}$$

Given a target distribution $\nu(x) \propto e^{-f(x)} \in \mathcal{P}_2\left(\mathbb{R}^d\right)$, the goal of Gaussian variational inference is to approximate the distribution $\nu$ by a more tractable distribution $\hat{\nu} \in \mathcal{G}$ such that

$$\hat{\nu} = \mathop{\arg\min}_{\rho\in\mathcal{G}} \mathsf{KL}\left(\rho \| \nu\right)$$

# 2 Algorithm: Gradient Flow

In order to find $\hat{\nu}$, we want to to use Gradient Flow to minimize $F(\rho) = \mathsf{KL}\left(\rho \mid \nu\right)$ over $\rho \in \mathcal{G}$.

Recall from last time that we can computed the gradients of $F(\rho) = F(m, C)$ with respect to $m$ and $C$ as follows:

$$\frac{\partial F(m,C)}{\partial m} = \mathop{\mathbb{E}}_{\mathcal{N}(m,C)}\left[\nabla f\right]$$
$$\frac{\partial F(m,C)}{\partial C} = \frac{1}{2}\left(\mathop{\mathbb{E}}_{\mathcal{N}(m,C)}\left[\nabla^2 f\right] - C^{-1}\right)$$

where $\rho = \mathcal{N}(m, C)$ and $\nu \propto e^{-f}$.

We claim that the following ODE is good for minimizing $F(m, C)$

$$\begin{cases} \dot{m}_t &= -\mathop{\mathbb{E}}_{\mathcal{N}(m_t,C_t)}\left[\nabla f\right] \\ \dot{C}_t &= 2\left(I - C_t \mathop{\mathbb{E}}_{\mathcal{N}(m_t,C_t)}\left[\nabla^2 f\right]\right) \end{cases} \tag{1}$$

We call this the Bures–Wasserstein Gradient Flow ($\mathsf{BW-GF}$). This ODE yields a well-defined evolution of Gaussian distributions $\left(\rho_t = \mathcal{N}\left(m_t, C_t\right)\right)_{t\geq 0}$, which we may optimistically believe to be a good approximation of $\nu$.

**Theorem 1.** *(Lambert et al. 2022) The ODE* $\mathsf{BW}-\mathsf{GF}$ *define* $\rho_t = \mathcal{N}(m_t, C_t)$*, which follow the gradient flow*

$$\frac{\partial \rho_t}{\partial t} = \dot{\rho}_t = -\mathsf{grad} F(\rho_t)$$

*where* $F(\rho) = \mathsf{KL}(\rho \| \nu)$ *and the gradient is with respect to the Bures-Wasserstein (*$\mathsf{BW}$*) metric*

$$W_2^2(\rho_1, \rho_2) = \|m_1 - m_2\|^2 + \mathsf{Tr}\left(C_1 + C_2 - 2\left(C_1^{\frac{1}{2}} C_2 C_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)$$

Note that the gradient with respect to the $\mathsf{BW}$ metric defines a gradient flow on a manifold. For instance, when $d = 1$, the $\mathsf{BW}$ metric becomes

$$W_2^2(\rho_1, \rho_2) = (m_1 - m_2)^2 + \left(\sqrt{C_1} - \sqrt{C_2}\right)^2$$

whereas the Euclidean distance between $\rho_1 = \mathcal{N}(m_1, C_1)$ and $\rho_2 = \mathcal{N}(m_2, C_2)$ is

$$\mathsf{Euc}(\rho_1, \rho_2) = (m_1 - m_2)^2 + (C_1 - C_2)^2$$

## 3 Convergence results

Now, we are interested in the convergence rate for $\mathsf{BW}-\mathsf{GF}$, i.e. the gradient flow for $F(\rho) = \mathsf{KL}(\rho\|\nu)$ on $(\mathcal{G}, W_2)$

**Theorem 2.** *(Lambert et al. 2022) Assume that* $\nu(x) \propto e^{-f(x)}$ *for some* $f : \mathbb{R}^d \to \mathbb{R}$*, and assume that* $f$ *is* $\alpha$*-strongly convex for some* $\alpha \geq 0$*. In other words,* $\nabla^2 f(x) \succeq \alpha I$ *for all* $x \in \mathbb{R}^d$*. This also means* $\nu \propto e^{-f}$ *is* $\alpha$*-strongly log-concave (*$\alpha - \mathsf{SLC}$*). Also recall that* $\hat{\nu} = \underset{\rho \in \mathcal{G}}{\arg\min}\, \mathsf{KL}(\rho\|\nu)$*. Then, along the* $\mathsf{BW}-\mathsf{GF}$ *that defines* $\rho_t = \mathcal{N}(m_t, C_t)$*:*

*(1) For all* $t \geq 0$*, we have*

$$W_2^2(\rho_t, \hat{\nu}) \leq e^{-2\alpha t} W_2^2(\rho_0, \hat{\nu}).$$

*That means* $\rho_t$ *converges to the minimizer exponentially fast when* $\alpha > 0$*.*

*(2) For all* $t \geq 0$*, we have*

$$\underbrace{\mathsf{KL}(\rho_t\|\nu)}_{F(\rho_t)} - \underbrace{\mathsf{KL}(\hat{\nu}\|\nu)}_{\min_{\rho\in\mathcal{G}} F(\rho)} \leq e^{-2\alpha t}\left(\underbrace{\mathsf{KL}(\rho_0\|\nu)}_{F(\rho_0)} - \underbrace{\mathsf{KL}(\hat{\nu}\|\nu)}_{\min_{\rho\in\mathcal{G}} F(\rho)}\right)$$

*That means* $F(\rho_t)$ *converges to the minima exponentially fast when* $\alpha > 0$*.*

*(3) If* $\alpha = 0$*, then for all* $t \geq 0$*, we have*

$$\mathsf{KL}(\rho_t\|\nu) - \mathsf{KL}(\hat{\nu}\|\nu) \leq \frac{1}{2t} W_2^2(\rho_0, \hat{\nu})$$

# 4 Approximation results

Exponential contraction is great! But we are also interested in how accurate the approximation $\hat{\nu} \in \mathcal{G}$ could be to the target distribution $\nu$.

Recall that for a posterior distribution

$$\begin{aligned}\nu(x) &= p_n(X \mid y_1, \cdots, y_n) \\ &\propto p_0(x) \prod_{i=1}^{n} p(y_i \mid x) \quad (\text{as } Y_i \mid X = x \text{ are i.i.d.}) \\ &\propto \exp\left(-f_0(x) - \sum_{i=1}^{n} \ell(x, y_i)\right)\end{aligned}$$

where $p_0 \propto e^{-f_0(x)}$ is the prior distribution and $\ell(x, y) = -\log p(y \mid x)$.

That means

$$v = p_n \propto e^{-f_n}$$

where

$$f_n(x) = f_0(x) + \sum_{i=1}^{n} \ell(x, y_i)$$

This implies

$$\lim_{n\to\infty} \frac{1}{n} f_n(x) = \lim_{n\to\infty} \frac{1}{n} f_0(x) + \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ell(x, y_i) = 0 + \mathbb{E}_Y\left[\ell(x, Y)\right] = g(x)$$

where $g(x) = \mathbb{E}_Y\left[\ell(x, Y)\right]$. then

$$p_n(x) \propto e^{-f_n(x)} \propto \exp\left(-n\left(\frac{1}{n} f(x)\right)\right) \overset{n\gg 1}{\approx} \exp\left(-n \cdot g(x)\right)$$

So if we assume that the target distribution is $\nu(x) \propto \exp(-n \cdot g(x))$, and if $n \gg 1$, then under certain assumptions (HW), we would obtain

$$\mathrm{Var}_\nu(X) = O\left(\frac{1}{n}\right),$$

i.e. $\nu_n$ looks roughly unimodel. We would then expect that Gaussian variational inference would yields a good approximation $\hat{\nu} \in \mathcal{G}$ for $\nu_n$.

**Theorem 3.** *(Katsevich and Rigollet 2023) Assume a target distribution $\nu(x) \propto \exp(-n \cdot g(x))$, $g : \mathbb{R}^d \to \mathbb{R}$ such that*

- *$g$ has a unique global minimizer $m^*$*

- $n \gtrsim d^3$, *where* $d$ *is the number of dimensions and* $n$ *is the number of observations*

*Let* $m_\nu = m_{\mathsf{EP}} = \mathbb{E}_\nu[x]$ *and* $C_\nu = C_{\mathsf{EP}} = \mathsf{Cov}_\nu(x)$*, and let* $m_{\mathsf{VI}}$ *and* $C_{\mathsf{VI}}$ *be the mean and covariance of the variational Gaussian approximation to* $\nu$*. Then*

$$\sqrt{n}\,\|m_{\mathsf{VI}} - m_\nu\|_{\ell_2} \lesssim \left(\frac{d^3}{n}\right)^{\frac{3}{2}}$$
$$n\,\|C_{\mathsf{VI}} - C_\nu\|_{op} \lesssim \frac{d^3}{n}$$

Notice that if $m_{\mathsf{Lap}}$ and $C_{\mathsf{Lap}}$ are the mean and covariance of the Laplace approximation to $\nu$, then

$$\sqrt{n}\,\|m_{\mathsf{Lap}} - m_\nu\|_{\ell_2} \lesssim \left(\frac{d^3}{n}\right)^{\frac{1}{2}}$$
$$n\,\|C_{\mathsf{Lap}} - C_\nu\|_{\mathrm{op}} \lesssim \frac{\mathrm{Const}(d)}{n}$$

The following are plots from the paper by Lambert et al. 2022 demonstrating the effectiveness of their $\mathsf{BW} - \mathsf{GF}$ algorithm on a Bayesian logistic regression problem.

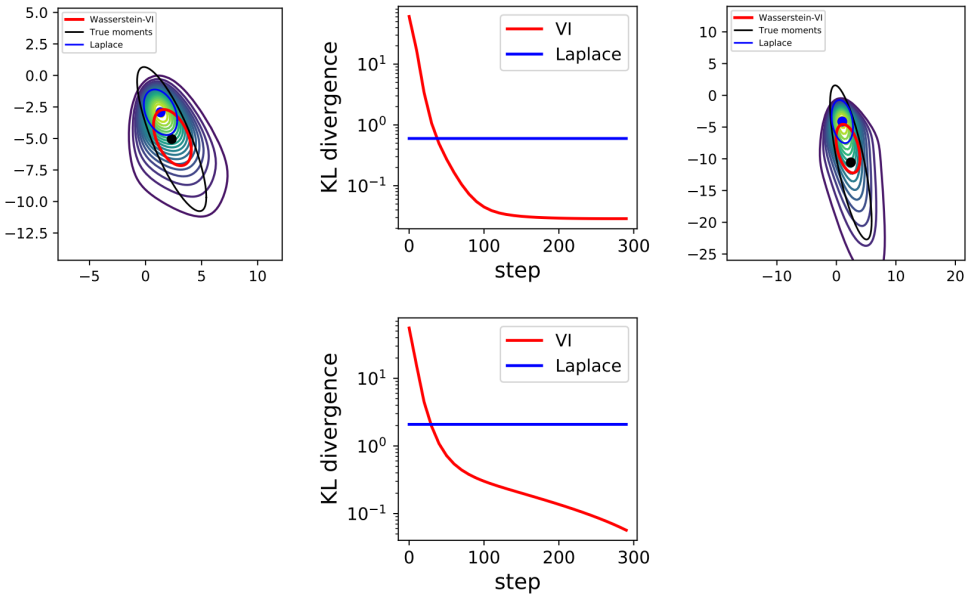


Figure 7: Results in dimension $d = 2$, $N = 10$ for a separation factor $s = 1.5$ (upper row) and $s = 2$ (lower row). The left column shows the true density via contour lines, the true mean (black dot) and covariance (black ellipsoid), and the results of the Laplace and Wasserstein VI approximations as blue and red ellipsoids respectively. The right column shows the evolution of the left KL divergence for Gaussian VI on a logarithmic scale. The corresponding KL divergence obtained with Laplace approximation is shown as a blue straight line.

## References

Katsevich, Anya and Philippe Rigollet (2023). "On the Approximation Accuracy of Gaussian Variational Inference". In: *arXiv preprint arXiv:2301.02168*.

Lambert, Marc et al. (2022). "Variational inference via Wasserstein gradient flows". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=K2PTuvVTF1L.