Statistical Models & Computing Methods

Lecture 15: Advanced VI – I

Cheng Zhang

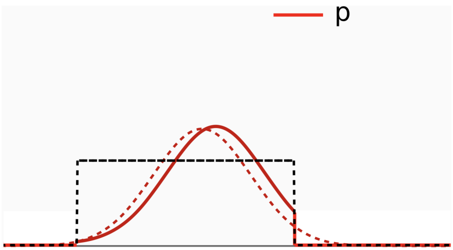
School of Mathematical Sciences, Peking University

November 25, 2024

Introduction

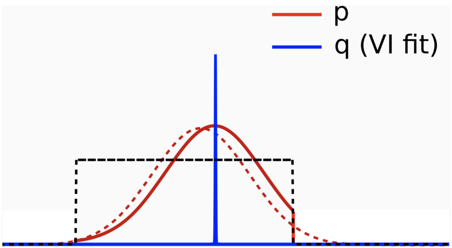
- So far, we have only used the KL divergence as a distance measure in VI.
- Other than the KL divergence, there are many alternative statistical distance measures between distributions that admit a variety of statistical properties.
- In this lecture, we will introduce several alternative divergence measures to KL, and discuss their statistical properties, with applications in VI.

Potential Problems with The KL Divergence



- VI does not work well for non-smooth potentials
- This is largely due to the zero-avoiding behaviour
  - The area where $p(\theta)$ is close to zero has very negative $\log p$, so does the variational distribution q when trained to minimize the KL.
- In this truncated normal example, VI will fit a delta function!

Potential Problems with The KL Divergence



- VI does not work well for non-smooth potentials
- This is largely due to the zero-avoiding behaviour
  - The area where $p(\theta)$ is close to zero has very negative $\log p$, so does the variational distribution q when trained to minimize the KL.
- In this truncated normal example, VI will fit a delta function!

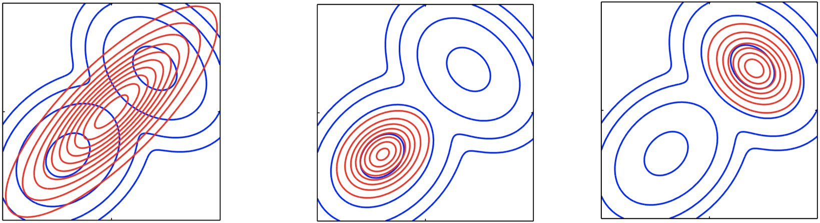
Beyond The KL Divergence

- Recall that the KL divergence from q to p is

$$D_{\mathrm{KL}}(q\|p) = \mathbb{E}_q \log \frac{q(x)}{p(x)} = \int q(x) \log \frac{q(x)}{p(x)}\, dx$$

- An alternative: the reverse KL divergence

$$D_{\mathrm{KL}}^{\mathrm{Rev}}(p\|q) = \mathbb{E}_p \log \frac{p(x)}{q(x)} = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$



Reverse KL **KL**

The f-Divergence

- The f-divergence from q to p is defined as

$$D_f(q\|p) = \int p(x) f\left(\frac{q(x)}{p(x)}\right)\, dx$$

where f is a convex function such that $f(1) = 0$.

- The f-divergence defines a family of valid divergences

$$\begin{aligned} D_f(q\|p) &= \int p(x) f\left(\frac{q(x)}{p(x)}\right)\, dx \\ &\geq f\left(\int p(x)\frac{q(x)}{p(x)}\, dx\right) = f(1) = 0 \end{aligned}$$

and

$$D_f(q\|p) = 0 \Rightarrow q(x) = p(x) \text{ a.s.}$$

The f-Divergence

Many common divergences are special cases of f-divergence, with different choices of f.

- KL divergence. $f(t) = t \log t$
- reverse KL divergence. $f(t) = -\log t$
- Hellinger distance. $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$

$$H^2(p, q) = \frac{1}{2} \int (\sqrt{q(x)} - \sqrt{p(x)})^2 dx = \frac{1}{2} \int p(x) \left(\sqrt{\frac{q(x)}{p(x)}} - 1 \right)^2 dx$$

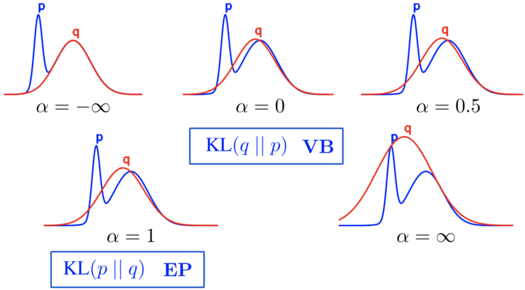
- Total variation distance. $f(t) = \frac{1}{2}|t - 1|$

$$d_{\mathrm{TV}}(p, q) = \frac{1}{2} \int |p(x) - q(x)| dx = \frac{1}{2} \int p(x) \left| \frac{q(x)}{p(x)} - 1 \right| dx$$

Amari's α-Divergence

When $f(t) = \frac{t^\alpha - t}{\alpha(\alpha-1)}$, we have the Amari's α-divergence (Amari, 1985; Zhu and Rohwer, 1995)

$$D_\alpha(p\|q) = \frac{1}{\alpha(1-\alpha)}\left(1 - \int p(\theta)^\alpha q(\theta)^{1-\alpha}\, d\theta\right)$$



$$D_{\mathrm{KL}}(q\|p) = \lim_{\alpha\to 0} D_\alpha(p\|q)$$

$$D_{\mathrm{KL}}(p\|q) = \lim_{\alpha\to 1} D_\alpha(p\|q)$$

Adapted from Hernández-Lobato et al.

Rényi's α-Divergence

$$D_\alpha(q\|p) = \frac{1}{\alpha - 1} \log \int q(\theta)^\alpha p(\theta)^{1-\alpha}\, d\theta$$

- Some special cases of Rényi's α-divergence
  - $D_1(q\|p) := \lim_{\alpha \to 1} D_\alpha(q\|p) = D_{\mathrm{KL}}(q\|p)$
  - $D_0(q\|p) = -\log \int_{q(\theta)>0} p(\theta) d\theta = 0$ *iff* $supp(p) \subset supp(q)$.
  - $D_{+\infty}(q\|p) = \log \max_\theta \frac{q(\theta)}{p(\theta)}$
  - $D_{\frac{1}{2}}(q\|p) = -2 \log \left(1 - \mathrm{Hel}^2(q\|p)\right)$
- Importance properties
  - Rényi divergence is non-decreasing in α

$$D_{\alpha_1}(q\|p) \geq D_{\alpha_2}(q\|p), \quad \text{if} \;\; \alpha_1 \geq \alpha_2$$

  - Skew symmetry: $D_{1-\alpha}(q\|p) = \frac{1-\alpha}{\alpha} D_\alpha(p\|q)$

The Rényi Lower Bound

- Consider approximating the exact posterior $p(\theta|x)$ by minimizing Rényi's α-divergence $D_\alpha(q(\theta)\|p(\theta|x))$ for some selected $\alpha > 0$
- Using $p(\theta|x) = p(\theta, x)/p(x)$, we have

$$
\begin{aligned}
D_\alpha(q(\theta)\|p(\theta|x)) &= \frac{1}{\alpha - 1} \log \int q(\theta)^\alpha p(\theta|x)^{1-\alpha}\, d\theta \\
&= \log p(x) - \frac{1}{1-\alpha} \log \int q(\theta)^\alpha p(\theta, x)^{1-\alpha}\, d\theta \\
&= \log p(x) - \frac{1}{1-\alpha} \log \mathbb{E}_q \left(\frac{p(\theta, x)}{q(\theta)} \right)^{1-\alpha}
\end{aligned}
$$

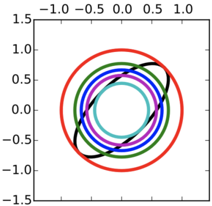
- The Rényi lower bound (Li and Turner, 2016)

$$
L_\alpha(q) \triangleq \frac{1}{1-\alpha} \log \mathbb{E}_q \left(\frac{p(\theta, x)}{q(\theta)} \right)^{1-\alpha}
$$

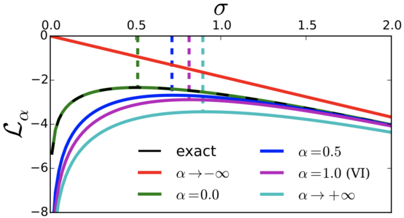
- **Theorem**(Li and Turner 2016). The Rényi lower bound is continuous and non-increasing on $\alpha \in [0,1] \cup \{|L_\alpha| < +\infty\}$. Especially for all $0 < \alpha < 1$

$$L_{\mathrm{VI}}(q) = \lim_{\alpha \to 1} L_\alpha(q) \le L_\alpha(q) \le L_0(q)$$

$L_0(q) = \log p(x)$ iff $supp(p(\theta|x)) \subset supp(q(\theta))$.



(a) Approximated posterior.



(b) Hyper-parameter optimisation.

- Monte Carlo estimation of the Rényi lower bound

$$\hat{L}_{\alpha,K}(q) = \frac{1}{1-\alpha} \log \frac{1}{K} \sum_{i=1}^{K} \left(\frac{p(\theta_i, x)}{q(\theta_i)} \right)^{1-\alpha}, \quad \theta_i \sim q(\theta)$$

- Unlike traditional VI, here the Monte Carlo estimate is **biased**. Fortunately, the bias can be characterized by the following theorem
- **Theorem**(Li and Turner, 2016). $\mathbb{E}_{\{\theta_i\}_{i=1}^K}(\hat{L}_{\alpha,K}(q))$ as a function of α and K is
  - non-decreasing in K for fixed $\alpha \leq 1$, and converges to $L_\alpha(q)$ as $K \to +\infty$ if $supp(p(\theta|x)) \subset supp(q(\theta))$.
  - continuous and non-increasing in α on $[0, 1] \cup \{|L_\alpha| < +\infty\}$

Multiple Sample ELBO

- When $\alpha = 0$, the Monte Carlo estimate reduces to the multiple sample lower bound (Burda et al., 2015)

$$\hat{L}_K(q) = \log\left(\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,\theta_i)}{q(\theta_i)}\right), \quad \theta_i \sim q(\theta)$$

- This recovers the standard ELBO when $K = 1$.
- Using more samples improves the tightness of the bound (Burda et al., 2015)

$$\log p(x) \geq \mathbb{E}(\hat{L}_{K+1}(q)) \geq \mathbb{E}(\hat{L}_K(q))$$

Moreover, if $p(x,\theta)/q(\theta)$ is bounded, then

$$\mathbb{E}(\hat{L}_K(q)) \to \log p(x), \quad \text{as } K \to +\infty$$

Using the reparameterization trick

$$\theta \sim q_\phi(\theta) \Leftrightarrow \theta = g_\phi(\epsilon),\ \epsilon \sim q_\epsilon(\epsilon)$$

$$\nabla_\phi \hat{L}_{\alpha,K}(q_\phi) = \sum_{i=1}^{K} \left(\hat{w}_{\alpha,i} \nabla_\phi \log \frac{p(g_\phi(\epsilon_i), x)}{q_\phi(g_\phi(\epsilon_i))} \right), \quad \epsilon_i \sim q_\epsilon(\epsilon)$$

where

$$\hat{w}_{\alpha,i} \propto \left(\frac{p(g_\phi(\epsilon_i), x)}{q_\phi(g_\phi(\epsilon_i))} \right)^{1-\alpha},$$

the normalized importance weight with finite samples. This is a biased estimate of $\nabla_\phi L_\alpha(q_\phi)$ (except $\alpha = 1$).

- $\alpha = 1$: Standard VI with the reparamterization trick
- $\alpha = 0$: Importance weighted VI (Burda et al., 2015)

Minibatch Training

- Full batch training for maximizing the Rényi lower bound could be very inefficient for large datasets
- Stochastic optimization is non-trivial since the Rényi lower bound can not be represented as an expectation on a datapoint-wise loss, except for $\alpha = 1$.
- Two possible methods:
  - derive the fixed point iteration on the whole dataset, then use the minibatch data to approximately compute it (Li et al., 2015)
  - approximate the bound using the minibatch data, then derive the gradient on this approximate objective (Hernández-Lobato et al., 2016)

  Remark: the two methods are equivalent when $\alpha = 1$ (standard VI).

Minibatch Training: Energy Approximation

- Suppose the true likelihood is

$$p(x|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$$

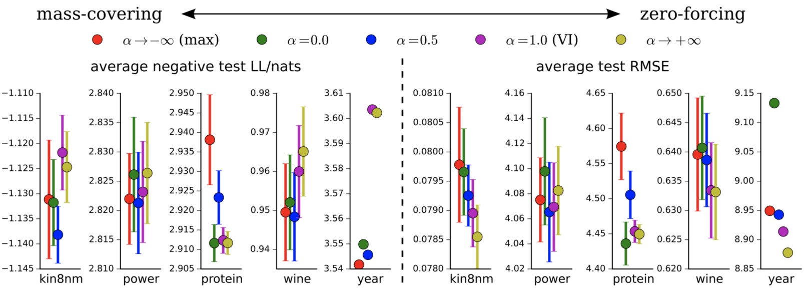
- Approximate the likelihood as

$$p(x|\theta) \approx \left(\prod_{n\in\mathcal{S}} p(x_n|\theta) \right)^{\frac{N}{|\mathcal{S}|}} \triangleq \bar{f}_{\mathcal{S}}(\theta)^N$$

- Use this approximation for the energy function

$$\tilde{L}_\alpha(q, \mathcal{S}) = \frac{1}{1-\alpha} \log \mathbb{E}_q \left(\frac{p_0(\theta)\bar{f}_{\mathcal{S}}(\theta)^N}{q(\theta)} \right)^{1-\alpha}$$

Example: Bayesian Neural Network



Adapted from Li and Turner, 2016

- The optimal α may vary for different data sets.
- Large α improves the predictive error, while small α provides better test log-likelihood.
- $\alpha = 0.5$ seems to produce overall good results for both test LL and RMSE.

- In standard VI, we often minimize $D_{\mathrm{KL}}(q\|p)$. Sometimes, we can also minimize $D_{\mathrm{KL}}(p\|q)$ (can be viewed as MLE).

$$q^* = \arg\min_q D_{\mathrm{KL}}(p\|q) = \arg\max_q \mathbb{E}_p \log q(\theta)$$

- Assume q is from the exponential family

$$q(\theta|\eta) = h(\theta)\exp\left(\eta^\top T(\theta) - A(\eta)\right)$$

- The optimal η^* satisfies

$$\begin{aligned}\eta^* &= \arg\max_\eta \mathbb{E}_p \log q(\theta|\eta) \\ &= \arg\max_\eta \left(\eta^\top \mathbb{E}_p\left(T(\theta)\right) - A(\eta)\right) + \text{Const}\end{aligned}$$

- Differentiate with respect to η

$$\mathbb{E}_p\left(T(\theta)\right) = \nabla_\eta A(\eta^*)$$

- Note that $q(\theta|\eta)$ is a valid distribution $\forall\eta$

$$\begin{aligned}
0 &= \nabla_\eta \int h(\theta)\exp\left(\eta^\top T(\theta) - A(\eta)\right)\, d\theta \\
&= \int q(\theta|\eta)\left(T(\theta) - \nabla_\eta A(\eta)\right)\, d\theta \\
&= \mathbb{E}_q\left(T(\theta)\right) - \nabla_\eta A(\eta)
\end{aligned}$$

- The KL divergence is minimized if the expected sufficient statistics are the same

$$\mathbb{E}_q\left(T(\theta)\right) = \mathbb{E}_p\left(T(\theta)\right)$$

- An approximate inference method proposed by Minka 2001.
- Suitable for approximating product forms. For example, with iid observations, the posterior takes the following form

$$p(\theta|x) \propto p(\theta) \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=0}^{n} f_i(\theta)$$

- We use an approximation

$$q(\theta) \propto \prod_{i=0}^{n} \tilde{f}_i(\theta)$$

  One common choice for $\tilde{f}_i$ is the exponential family

$$\tilde{f}_i(\theta) = h(\theta) \exp\left(\eta_i^\top T(\theta) - A(\eta_i)\right)$$

- Iteratively refinement of the terms $\tilde{f}_i(\theta)$

Iterative Updating

- Take out term approximation i

$$q^{\setminus i}(\theta) \propto \prod_{j \neq i} \tilde{f}_j(\theta)$$

- Put back in term i

$$\hat{p}(\theta) \propto f_i(\theta) \prod_{j \neq i} \tilde{f}_j(\theta)$$

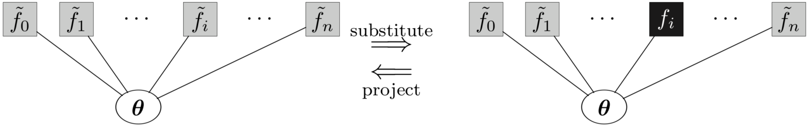
- Match moments. Find q such that

$$\mathbb{E}_q(T(\theta)) = \mathbb{E}_{\hat{p}}(T(\theta))$$

- Update the new term approximation

$$\tilde{f}_i^{\text{new}}(\theta) \propto \frac{q(\theta)}{q^{\setminus i}(\theta)}$$

How Does EP Work?



- Minimize the KL divergence from $\hat{p}$ to q

$$D_{\mathrm{KL}}(\hat{p}\|q) = \mathbb{E}_{\hat{p}} \log\left(\frac{\hat{p}(\theta)}{q(\theta)}\right)$$

- Equivalent to moment matching when q is in the exponential family.

- The approximating distributions that we discussed so far are assumed to have a parametric form, that is $q_\theta(x)$ with parameter θ.
- This parametric form often limits the power of the approximating distributions.
- In what follows, we will introduce a particle based VI introduced by Liu et al. that uses non-parameteric approximating distributions.

- A general theoretical tool for bounding differences between distributions, introduced by Charles Stein.
- The key idea is to characterize a distribution p with a Stein operator $\mathcal{A}_p$, such that

$$p = q \iff \mathbb{E}_{x \sim q}[\mathcal{A}_p f(x)] = 0, \quad \forall f \in \mathcal{F}$$

For continuous distributions with smooth density $p(x)$,

$$\mathcal{A}_p f(x) := s_p(x)^T f(x) + \nabla_x \cdot f(x)$$

where $s_p(x) = \nabla_x \log p(x)$ is the score function.

Note that $s_p(x)$ does not dependent on the normalizing constant of $p(x)$, so $p(x)$ can be unnormalized.

- When $p = q$, we have Stein's Identity

$$\mathbb{E}_{x\sim p}\left[s_p(x)^T f(x) + \nabla_x \cdot f(x)\right] = 0$$

- Stein's identity defines an infinite number of identities indexed by test function f, widely applied in learning probabilistic models, variance reduction, optimization and many more.
- When $p \neq q$, we have (also by Stein's Identity)

$$\mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)] = \mathbb{E}_{x\sim q}[(s_p(x) - s_q(x))^T f(x)] \qquad (1)$$

Easy to find test function $f(x)$ such that (1) is non-zero. For example:

$$f(x) = s_p(x) - s_q(x)$$

- We therefore, define Stein Discrepancy between p and q as follows
$$D(q\|p) := \max_{f \in \mathcal{F}} \mathbb{E}_{x \sim q}[\mathcal{A}_p f(x)] \tag{2}$$
where $\mathcal{F}$ is a rich enough set of functions.
- Traditionally, Stein's method takes $\mathcal{F}$ to be sets of functions with bounded Lipschitz norm, which is computationally difficult for practical use.
- We can use a kernel trick to construct a reproducing kernel Hilbert space (RKHS) where there is a closed form solution to (2).

- Let $k(x, x')$ be a positive definite kernel, that is
$$\int_{\mathcal{X}} g(x)k(x,x')g(x')\,dxdx' > 0, \quad \forall\, 0 < \|g\|_2^2 < \infty.$$
By Mercer's theorem,
$$k(x,x') = \sum\nolimits_i \lambda_i e_i(x) e_i(x')$$

- We can define a RKHS $\mathcal{H}$ that contains linear combinations of these eigenfunctions
$$f(x) = \sum\nolimits_i f_i e_i(x), \quad \langle f, g\rangle_{\mathcal{H}} = \sum\nolimits_i \frac{f_i g_i}{\lambda_i}$$
with $\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} = \sum_i f_i^2/\lambda_i$.

- Reproducing Property
$$f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}, \quad k(x,x') = \langle k(\cdot,x), k(\cdot,x')\rangle_{\mathcal{H}}.$$

- Given a positive definite kernel $k(x, x')$, Liu et al. define a **kernelized Stein discrepancy** (KSD) $D(q\|p)$ as follows

$$D(q\|p) = \sqrt{\mathbb{E}_{x,x'\sim q}[\delta_{p,q}(x)^T k(x,x')\delta_{p,q}(x')]}$$

where $\delta_{p,q}(x) = s_p(x) - s_q(x)$. Obviously,

$$D(q\|p) \geq 0, \quad D(q\|p) = 0 \Leftrightarrow q = p.$$

- With the spectral decomposition, we can rewrite KSD as

$$D(q\|p) = \sqrt{\sum_i \lambda_i \|\mathbb{E}_{x\sim q}[\mathcal{A}_p e_i(x)]\|_2^2}$$

- It turns out that KSD can be viewed as standard Stein discrepancy over a specific family of functions $\mathcal{F}$, i.e, the unit ball of $\mathcal{H}^d = \mathcal{H} \times \cdots \times \mathcal{H}$.
- Denote $\beta(x') = \mathbb{E}_{x\sim q}[\mathcal{A}_p k_{x'}(x)]$, then

$$D(q\|p) = \|\beta\|_{\mathcal{H}^d}$$

- Moreover, we have

$$\langle \beta, f\rangle_{\mathcal{H}^d} = \mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)], \quad \forall f \in \mathcal{H}^d$$

- Therefore,

$$D(q\|p) = \max_{f\in\mathcal{F}} \mathbb{E}_{x\sim q}[\mathcal{A}_p f(x)]$$

where $\mathcal{F} = \{f \in \mathcal{H}^d : \|f\|_{\mathcal{H}^d} \leq 1\}$. The maximum is achieved at $f^* = \beta/\|\beta\|_{\mathcal{H}^d}$.

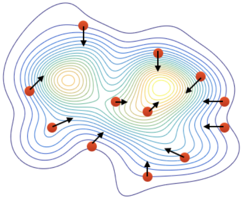
Proposed by Liu and Wang, 2016.

Idea: represent the distribution using a collection of particles $\{x_i\}_{i=1}^n$ and iteratively move these particles toward the target p by updates of form

$$x_i \leftarrow T(x_i), \quad T(x) = x + \epsilon\phi(x)$$

where ϕ is a perturbation direction chosen to maximumly decrease the KL divergence.

$$\phi = \operatorname*{arg\,max}_{\phi \in \mathcal{F}} \left\{ -\frac{\partial}{\partial \epsilon} D_{\mathrm{KL}}(q_T \| p) \Big|_{\epsilon=0} \right\}$$



where q_T is the density of $x' = T(x)$ when the current density of x is $q(x)$.

- Perturbation direction is closely related to Stein operator

$$-\frac{\partial}{\partial \epsilon} D_{\mathrm{KL}}(q_T \| p)\Big|_{\epsilon=0} = \mathbb{E}_{x\sim q}[\mathcal{A}_p \phi(x)]$$

- This gives another interpretation of Stein discrepancy

$$D(q\|p) = \max_{\phi\in\mathcal{F}} \left\{ -\frac{\partial}{\partial \epsilon} D_{\mathrm{KL}}(q_T \| p)\Big|_{\epsilon=0} \right\}$$

- Most importantly, the optimum direction has a closed form when $\mathcal{F}$ is the unit ball of RKHS $\mathcal{H}^d$:

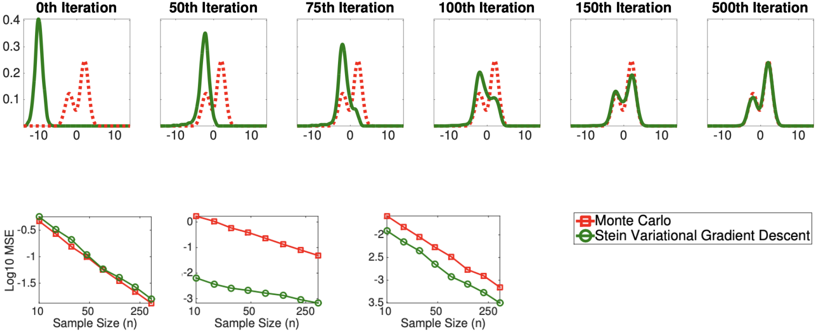
$$\begin{aligned}\phi^*(\cdot) &= \mathbb{E}_{x\sim q}[\mathcal{A}_p k(x,\cdot)] \\ &= \mathbb{E}_{x\sim q}[\nabla_x \log p(x) k(x,\cdot) + \nabla_x k(x,\cdot)]\end{aligned}$$

We can approximate the expectation $E_{x\sim q}$ with the empirical average over current particles

$$x_i \leftarrow x_i + \epsilon \frac{1}{n} \sum_{j=1}^{n} \left[\nabla_x \log p(x_j) k(x_j, x_i) + \nabla_{x_j} k(x_j, x_i) \right], \; 1 \le i \le n$$

- Deterministically transport probability mass from initial q_0 to target p.
- Reduces to standard gradient ascent for MAP when using a single particle ($n = 1$).
- $\nabla_x \log p(x_j)$: the gradient term moves the particles towards high probability domains of $p(x)$.
- $\nabla_x k(x_j, x_i)$: the repulsive force term enforces diversity in the particles and prevents them from collapsing to the modes of $p(x)$.

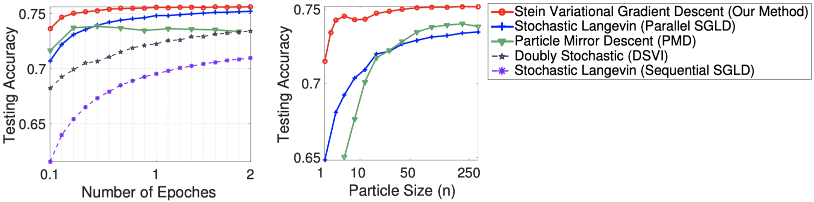
Examples: Mixture of Gaussian



(a) Estimating $\mathbb{E}(x)$ *(b) Estimating* $\mathbb{E}(x^2)$ *(c) Estimating* $\mathbb{E}(\cos(\omega x + b))$

Liu et al., 2016

Examples: Bayesian Logistic Regression



(a) Particle size $n = 100$ (b) Results at 3000 iteration ($\approx$ 0.32 epoches)

Liu et al., 2016

References

- Amari, Shun-ichi. Differential-Geometrical Methods in Statistic. Springer, New York, 1985.
- Zhu, Huaiyu and Rohwer, Richard. Information geometric measurements of generalisation. Technical report, Technical Report NCRG/4350. Aston University., 1995.
- Y. Li and R. E. Turner. Rényi Divergence Variational Inference. NIPS, pages 1073–1081, 2016.
- Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. International Conference on Learning Representations (ICLR), 2016.
- Y. Li, J. M. Hernández-Lobato, and R. E. Turner. Stochastic expectation propagation. In Advances in Neural Information Processing Systems (NIPS), 2015.

- J. M. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. E. Turner. Black-box α-divergence minimization. In Proceedings of The 33rd International Conference on Machine Learning (ICML), 2016.
- Q. Liu, J. D. Lee, and M. I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In ICML, 2016.
- Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In Advances in Neural Information Processing Systems 29, pp. 2370–2378, 2016.