Modern Computational Statistics

Lecture 15: Training Objectives in VI

Cheng Zhang

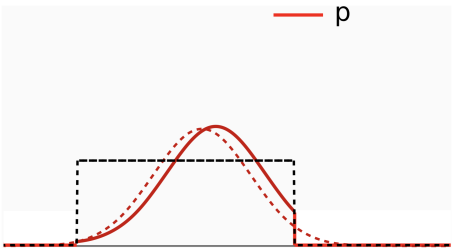
School of Mathematical Sciences, Peking University

November 18, 2019

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.

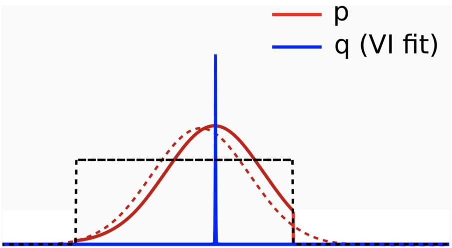
Potential Problems with The KL Divergence



- 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 distribution when trained to minimize the KL.
- In this truncated normal example, VI will fit a delta function!

PEKING UNIVERSITY

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 distribution when trained to minimize the KL.
- In this truncated normal example, VI will fit a delta function!

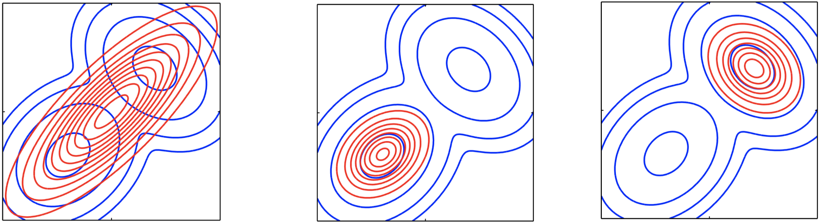
Beyond The KL Divergence

- 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

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

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. $f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$

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

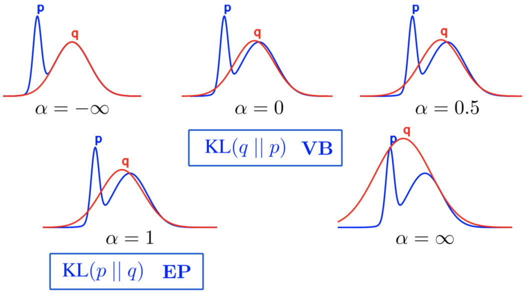
- Total variation distance. $f(t) = \frac{1}{2}|t - 1|$

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

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)

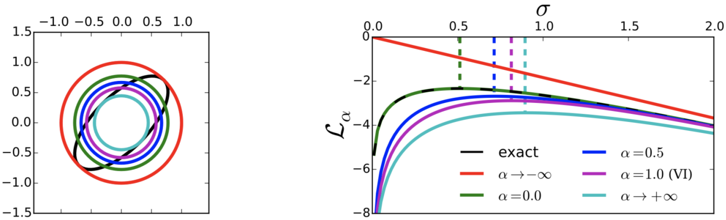
$$L_\alpha(q) \triangleq \frac{1}{1-\alpha} \log \mathbb{E}_q \left(\frac{p(\theta, x)}{q(\theta)}\right)^{1-\alpha}$$

The Rényi Lower Bound

- **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 $0 < \alpha < 1$

$$L_{\text{VI}}(q) = \lim_{\alpha \to 1} L_\alpha(q) \leq L_\alpha(q) \leq L_0(q)$$

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



(a) Approximated posterior.

(b) Hyper-parameter optimisation.

Monte Carlo Estimation

- 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

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

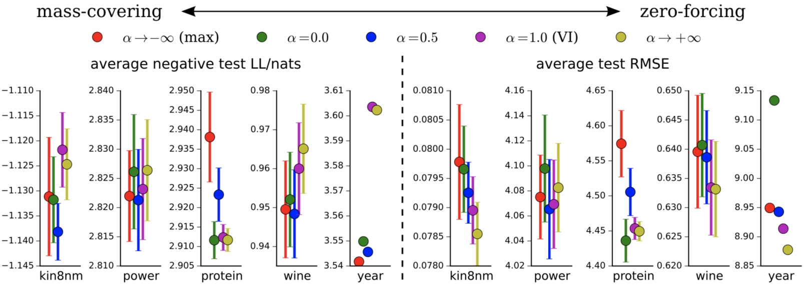
- Approximate the likelihood as

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

- 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.

Expectation Propagation

- 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

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

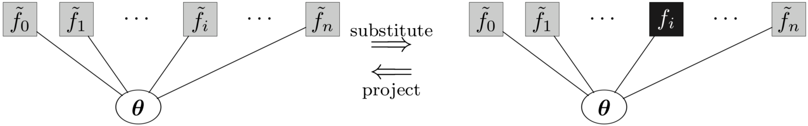
- Match moments. Find q such that

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

- Update the new term approximation

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

How Does EP Work?

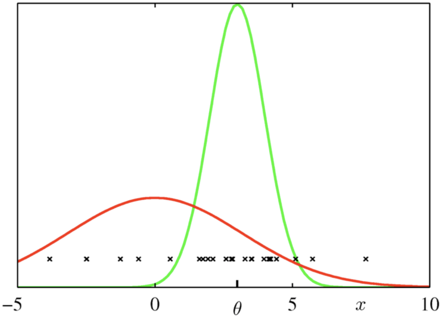


- Minimize the KL divergence from $\hat{p}$ to q

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

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

Example: The Clutter Problem



- **Goal**: fit a multivariate Gaussian into data in the presence of background clutter (also Gaussian)

$$p(x|\theta) = (1-w)\mathcal{N}(x|\theta, I) + w\mathcal{N}(x|0, aI)$$

- The prior is Gaussian: $p(\theta) = \mathcal{N}(\theta|0, bI)$.

- The joint distribution

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

  is a mixture of 2^n Gaussians, intractable for large n.
- We approximate it using a spherical Gaussian

$$q(\theta) = \mathcal{N}(\theta|m, vI)$$

- This is an exponential family with
  - sufficient statistics $T(\theta) = (\theta, \theta^\top \theta)$
  - natural parameters $\eta = (v^{-1}m, -\frac{1}{2}v^{-1})$
  - normalizing constant $Z(\eta) = (2\pi v)^{d/2} \exp\left(\frac{m^\top m}{2v}\right)$

Initialization

- For the clutter problem, we have

$$
\begin{aligned}
f_0(\theta) &= p(\theta) \\
f_i(\theta) &= p(x_i|\theta), \; i = 1, \ldots, n
\end{aligned}
$$

- The approxmation is of the form

$$
\begin{aligned}
\tilde{f}_0(\theta) &= f_0(\theta) = p(\theta) \\
\tilde{f}_i(\theta) &= s_i \exp(\eta_i^\top T(\theta)), \; i = 1, \ldots, n \\
q(\theta) &\propto \prod_{i=0}^{n} \tilde{f}_i(\theta) = s\mathcal{N}(\theta; \eta)
\end{aligned}
$$

- Initialize $\eta_i = (0, 0)$ for $i = 1, \ldots, n$

- With natural parameters, taking out term approximation i is trivial.

$$q^{\backslash i}(\theta) \propto \frac{q(\theta)}{\tilde{f}_i(\theta)} \propto \mathcal{N}(\theta; \eta^{\backslash i})$$

where

$$\eta^{\backslash i} = \eta - \eta_i$$

- Now we put back in term i

$$\begin{aligned}
\hat{p}(\theta) &\propto ((1-w)\mathcal{N}(x_i|\theta, I) + w\mathcal{N}(x_i|0, aI))\,\mathcal{N}(\theta; \eta^{\backslash i}) \\
&= (1-w)\frac{Z(\eta^{+})}{Z(\eta^{x_i})Z(\eta^{\backslash i})}\mathcal{N}(\theta; \eta^{+}) + w\mathcal{N}(x_i|0, aI)\mathcal{N}(\theta; \eta^{\backslash i}) \\
&\propto r\mathcal{N}(\theta; \eta^{+}) + (1-r)\mathcal{N}(\theta; \eta^{\backslash i})
\end{aligned}$$

where $\eta^{+} = \eta^{\backslash i} + \eta^{x_i}, \quad \eta^{x_i} = (x_i, -\frac{1}{2})$.

Match Moments and Update

- Now we match the sufficient statistics of the Gaussian mixture

$$\hat{p}(\theta) = r\mathcal{N}(\theta; \eta^{+}) + (1-r)\mathcal{N}(\theta; \eta^{\setminus i})$$

From $\mathbb{E}_q(T(\theta)) = \mathbb{E}_{\hat{p}}(T(\theta))$, we have

$$m = rm^{+} + (1-r)m^{\setminus i}$$

$$v + m^{\top}m = r\left(v^{+} + (m^{+})^{\top}m^{+}\right) + (1-r)\left(v^{\setminus i} + (m^{\setminus i})^{\top}m^{\setminus i}\right)$$

- Similarly, the update of $\tilde{f}_i$ is trivial

$$\tilde{f}_i(\theta) \propto \frac{q(\theta)}{q^{\setminus i}(\theta)} \propto \mathcal{N}(\theta; \eta_i)$$

where

$$\eta_i = \eta - \eta^{\setminus i}$$

Marginal Likelihood by EP

- We can use EP to evaluate the marginal likelihood $p(x)$
- To do this, we include a scale on $\tilde{f}_i(\theta)$

$$\tilde{f}_i(\theta) = Z_i \frac{q^*(\theta)}{q^{\setminus i}(\theta)}$$

where $q^*(\theta)$ is a normalized version of $q(\theta)$ and

$$Z_i = \int q^{\setminus i}(\theta) f_i(\theta) \, d\theta$$

- Use the normalizing constant of $q(x)$ to approximate $p(x)$

$$p(x) \approx \int \prod_{i=0}^{n} \tilde{f}_i(\theta) \, d\theta$$

- For the clutter problem

$$s_i \exp(\eta_i^\top T(\theta)) = \tilde{f}_i(\theta) = Z_i \frac{q^*(\theta)}{q^{\setminus i}(\theta)}$$

implies

$$s_i = Z_i \frac{Z(\eta^{\setminus i})}{Z(\eta)}$$

$$Z_i = (1-w) \frac{Z(\eta^+)}{Z(\eta^{x_i}) Z(\eta^{\setminus i})} + w\mathcal{N}(x_i | 0, aI)$$

- The marginal likelihood estimate is

$$p(x) \approx \int \prod_{i=0}^{n} \tilde{f}_i(\theta)\, d\theta = \frac{Z(\eta)}{Z(\eta_0)} \prod_{i=1}^{n} s_i$$

An Energy Function For EP

- The EP iterations can be shown to always have a fixed point when the approximations are in an exponential family.
- With an exact prior, the final approximation is

$$q(\theta) \propto p(\theta) \exp\left(\nu^\top T(\theta)\right)$$

- The leave-one-out approximations

$$q^{\setminus i}(\theta) \propto p(\theta) \exp\left(\lambda_i^\top T(\theta)\right)$$

An Energy Function For EP

- EP fixed points correspond to stationary points of the objective

$$\min_{\nu} \max_{\lambda} (n-1) \log \int p(\theta) \exp(\nu^\top T(\theta)) \, d\theta - \sum_{i=1}^{n} \log \int f_i(\theta) p(\theta) \exp(\lambda_i^\top T(\theta)) \, d\theta$$

such that $(n-1)\nu_j = \sum_i \lambda_{ij}$

- Taking derivatives we get the stationary conditions

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

- Note that this is a non-convex optimization problem.

Summary

- Other than the standard KL divergence, there are many alternative distance measures for VI (e.g., f-divergence, Rényi α-divergence).
- The Rényi α-divergences allow tractable lower bound and promote different learning behaviors through the choice of α (from mode-covering to model-seeking as α goes from $-\infty$ to ∞), which can be adapted to specific learning tasks.
- We also introduced another approximate inference method, expectation propagation (EP), that uses the reversed KL. More recent development on EP (Li et al., 2015, Hernández-Lobato et al., 2016).

- 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.