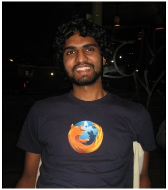
The χ-Divergence for Approximate Inference

Adji Bousso Dieng
Columbia University

Collaborators

Bayesian inference

data - $\mathbf{x}$

latent variables - $\mathbf{z}$

probabilistic model - $p(\mathbf{x}, \mathbf{z})$

Goal: Compute $p(\mathbf{z}|\mathbf{x})$

Bayesian inference

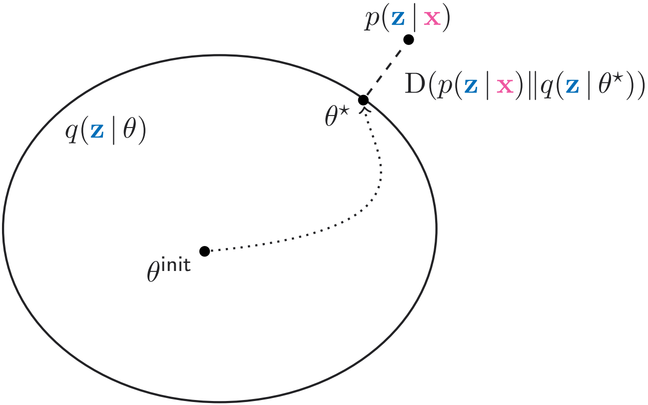
How to compute $p(\mathbf{z}|\mathbf{x})$

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})}$$

$\rightarrow$ easy for conjugate models

$\rightarrow$ intractable for most cases

Variational inference



$\mathrm{D}(p(\mathbf{z} \mid \mathbf{x}) \| q(\mathbf{z} \mid \theta)) \geq 0$

$\mathrm{D}(p(\mathbf{z} \mid \mathbf{x}) \| q(\mathbf{z} \mid \theta)) = 0 \iff p(\mathbf{z} \mid \mathbf{x}) = q(\mathbf{z} \mid \theta)$ a.e.

Variational inference

Pick a tractable family of approximating distribution $q(\mathbf{z} \mid \theta)$

Pick a divergence measure $\mathrm{D}(p(\mathbf{z} \mid \mathbf{x}) \| q(\mathbf{z} \mid \theta))$

Optimize this divergence over θ

$$q^*(\mathbf{z} \mid \theta^*) = \operatorname*{arg\,min}_{q(\mathbf{z} \mid \theta)} \mathrm{D}(p(\mathbf{z} \mid \mathbf{x}) \| q(\mathbf{z} \mid \theta))$$

Use $q^*(\mathbf{z} \mid \theta^*)$ as a surrogate to $p(\mathbf{z} \mid \mathbf{x})$

Which divergence measure to choose?

Divergence measures

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

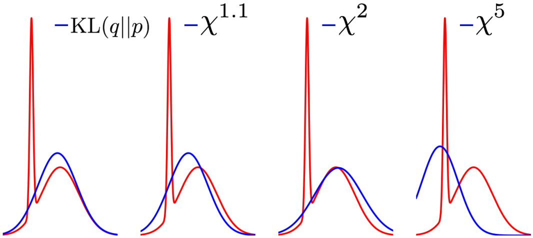
$\rightarrow$ f convex and $f(1) = 0$

$\rightarrow$ Special cases:

$KL(q||p)$ (BBVI) and $KL(p||q)$ ($\approx$ EP)

$D_\alpha(q||p)$(VR_α) and $D_\alpha(p||q)$ (CHIVI)

Divergence measures



Three types of behaviors:

zero-forcing, zero-avoiding, neither

Variational inference with $KL(q||p)$

$$\mathrm{KL}(q(\mathbf{z} \mid \theta) \| p(\mathbf{z} \mid \mathbf{x})) = E_{q(\mathbf{z} \mid \theta)} \left[\log \left(\frac{q(\mathbf{z} \mid \theta)}{p(\mathbf{z} \mid \mathbf{x})} \right) \right]$$

$\rightarrow$ Maximizes the evidence lower bound

$$\mathsf{ELBO}(\theta) = E_{q(\mathbf{z} \mid \theta)} [\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z} \mid \theta)]$$

$\rightarrow$ Perform Stochastic optimization using Monte Carlo estimates of $\nabla_\theta ELBO(\theta)$

Variational inference with $KL(q||p)$

$$\mathsf{ELBO}(\theta) = E_{q(\mathbf{z}\,|\,\theta)}\Big[\log p(\mathbf{x}|\mathbf{z})\Big] - \mathrm{KL}(q(\mathbf{z}\,|\,\theta)\|p(\mathbf{z}))$$

$$\nabla_\theta ELBO(\theta) = E_{q(\mathbf{z}\,|\,\theta)}\left[\nabla_\theta\Big(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}\,|\,\theta)\Big)\right]$$

$\rightarrow$ uses unbiased gradients

$\rightarrow$ very efficient and works well in many settings

$\rightarrow$ underdispersion ... too confident

$\rightarrow$ model overpruning in VAEs

Variational inference with $D_\alpha(q||p)$

$$\mathrm{D}_\alpha(q(\mathbf{z}\,|\,\theta)\|p(\mathbf{z}\,|\,\mathbf{x})) = \tfrac{1}{\alpha-1}\log\int p(\mathbf{z}\,|\,\mathbf{x})^{1-\alpha}q(\mathbf{z}\,|\,\theta)^{\alpha}d\mathbf{z}$$

$\rightarrow$ Maximizes the variational Renyi lower bound

$$\mathcal{L}(\theta) = \frac{1}{1-\alpha}\log E_{q(\mathbf{z}\,|\,\theta)}\left[\left(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}\,|\,\theta)}\right)^{1-\alpha}\right]$$

$$\hat{\mathcal{L}}(\theta) = \frac{1}{1-\alpha}\log\frac{1}{K}\sum_{k=1}^{K}\left[\left(\frac{p(\mathbf{x},\mathbf{z}_k)}{q(\mathbf{z}_k\,|\,\theta)}\right)^{1-\alpha}\right] \text{ where } \mathbf{z}_k \sim q(\mathbf{z}\,|\,\theta)$$

$\rightarrow$ Perform Stochastic optimization using noisy estimates of $\nabla_\theta\mathcal{L}(\theta)$

Variational inference with $D_\alpha(q||p)$

$$\nabla_\theta \mathcal{L}(\theta) = E_{q(\mathbf{z}\,|\,\theta)}\left[\frac{w^{1-\alpha}}{E_{q(\mathbf{z}\,|\,\theta)}(w^{1-\alpha})}\nabla_\theta \log w\right]$$

$$\log w = \log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}\,|\,\theta)$$

$\rightarrow$ uses biased gradients

$\rightarrow$ works well in many settings

$\rightarrow$ underdispersion by the nature of the divergence

$\rightarrow$ cannot handle upper bounds

Variational inference with $D_\alpha(p||q)$

$\rightarrow$ Equivalent to minimizing:

$$\mathrm{D}_\chi(p(\mathbf{z}\,|\,\mathbf{x})\|q(\mathbf{z}\,|\,\theta)) = \frac{1}{n} E_{q(\mathbf{z}\,|\,\theta)}\Big[\Big(\frac{p(\mathbf{z}\,|\,\mathbf{x})}{q(\mathbf{z}\,|\,\theta)}\Big)^n\Big],\ n = 1+\alpha$$

$\rightarrow$ Equivalent to minimizing the χ upper bound:

$$\mathsf{CUBO}_n(\theta) = \frac{1}{n}\log E_{q(\mathbf{z}\,|\,\theta)}\Big[\Big(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}\,|\,\theta)}\Big)^n\Big]$$

$\rightarrow$ Equivalent to minimizing:

$$\mathcal{L}(\theta) = \exp(n * \mathsf{CUBO}(\theta))$$

$\rightarrow$ Perform Stochastic optimization using noisy estimates of $\nabla_\theta \mathcal{L}(\theta)$

Variational inference with $D_\alpha(p||q)$

$$\nabla_\theta \mathcal{L}(\theta) = E_{q(\mathbf{z}\,|\,\theta)}\Big[w^n\ \nabla_\theta \log w\Big]$$

$\rightarrow$ uses unbiased gradients

$\rightarrow$ overdispersed posterior approximations

$\rightarrow$ performs upper bound minimization

$\rightarrow$ enables sandwich estimation of the evidence

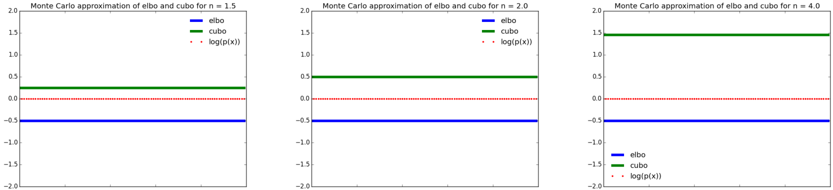
$\rightarrow$ black box alternative to EP

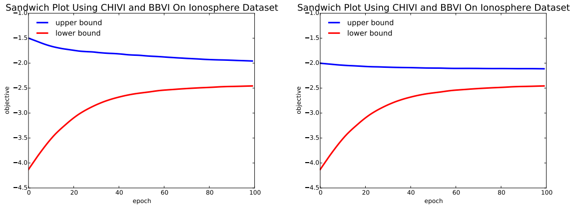
Sandwich estimation

$$\mathsf{ELBO}(\theta) \leq \log p(\mathbf{x}) \leq \mathsf{CUBO}_n(\theta)$$

$$\lim_{n \to 0} \mathsf{CUBO}_n(\theta) = \mathsf{ELBO}(\theta)$$

Sandwich estimation





CHIVI for probit regression

Table: Test error for Bayesian probit regression.

Dataset	BBVI	EP	CHIVI
Pima	0.235 ± 0.006	0.234 ± 0.006	0.222 ± 0.048
Ionos	0.123 ± 0.008	0.124 ± 0.008	0.116 ± 0.05
Madelon	0.457 ± 0.005	0.445 ± 0.005	0.453 ± 0.029
Covertype	0.157 ± 0.01	0.155 ± 0.018	0.154 ± 0.014

CHIVI for Cox processes

Table: Average L_1 error for posterior uncertainty estimates (ground truth from HMC).

-	Curry	Demarcus	Lebron	Duncan
CHIVI	0.060	0.073	0.0825	0.0849
BBVI	0.066	0.082	0.0812	0.0871

Take-away message

CHIVI

- $\rightarrow$ uses unbiased gradients
- $\rightarrow$ favors overdispersed posterior approximations
- $\rightarrow$ performs upper bound minimization
- $\rightarrow$ enables sandwich estimation of the evidence
- $\rightarrow$ is a black box alternative to EP
- $\rightarrow$ needs variance reduction techniques