[Paper review 24]

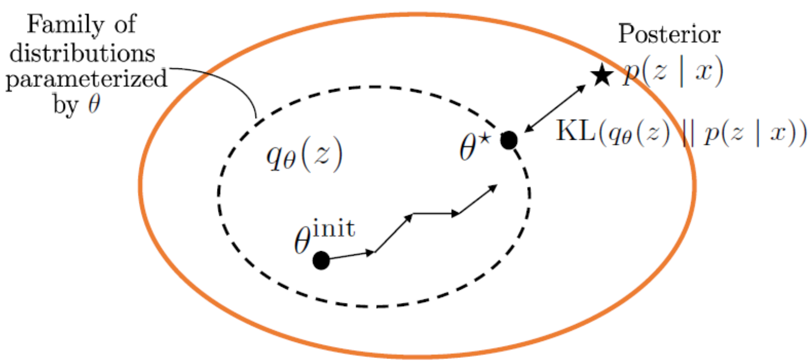
A Contrastive Divergence for Combining Variational Inference and MCMC

(Francisco J. R. Ruiz 1 2 Michalis K. Titsias, 2019)

[Contents]

1. Review : Goal

to make MORE EXPRESSIVE Variational Distributions



2. MCMC

2.1 run MCMC steps

start from "explicit" variational distribution : $q_\theta^{(0)}(z)$

- 1) know the density
- 2) can sample

Improve the distribution with t MCMC steps

- $z_0 \sim q_\theta^{(0)}(z), \quad z \sim Q^{(t)}(z \mid z_0)$
- target : posterior $p(z \mid x)$

Implicit variational distribution

- $q_\theta(z)=\int q_\theta^{(0)}\left(z_0\right) Q^{(t)}\left(z \mid z_0\right) d z_0$

2.2 Challenges of MCMC in VI

ELBO : $\mathcal{L}_{\text{improved }}(\theta)=\mathbb{E}_{q_\theta(z)}\left[\log p(x, z)-\log q_\theta(z)\right]$

- Problem #1) intractable
- Problem #2) objective depend WEAKLY on θ

  $q_\theta(z) \stackrel{t \rightarrow \infty}{\longrightarrow} p(z \mid x)$

3 Alternative Divergence : VCD

3.1 VCD

VCD : "Variational Contrastive Divergence" (= $\mathcal{L}_{\mathrm{VCD}}(\theta)$)

Desired Properties

- Non-negative for any θ
- Zero iff $q_\theta^{(0)}(z)=p(z \mid x)$

Improved distribution $q_\theta(z)$: decreases the KL :

$$\mathrm{KL}\left(q_\theta(z) \| p(z \mid x)\right) \leq \mathrm{KL}\left(q_\theta^{(0)}(z) \| p(z \mid x)\right)$$

Objective:

$$\mathcal{L}(\theta)=\mathrm{KL}\left(q_\theta^{(0)}(z) \| p(z \mid x)\right)-\mathrm{KL}\left(q_\theta(z) \| p(z \mid x)\right)$$

- have to minimize! ($q_\theta^{(0)}(z)$ should get close to $q_\theta(z)$)
- but, intractable because of $q_\theta(z)$

Add a regularizer

$$\mathcal{L}_{\mathrm{VCD}}(\theta)=\underbrace{\mathrm{KL}\left(q_\theta^{(0)}(z) \| p(z \mid x)\right)-\mathrm{KL}\left(q_\theta(z) \| p(z \mid x)\right)}_{\geq 0}+\underbrace{\mathrm{KL}\left(q_\theta(z) \| q_\theta^{(0)}(z)\right)}_{\geq 0}$$

- problem #1) (intractability)
  - solution : log $q_\theta^{(0)}(z)$ cancels out
- problem #2) (weak dependence)
  - solution : $\mathcal{L}_{\mathrm{VCD}}(\theta) \stackrel{t \rightarrow \infty}{\longrightarrow} \mathrm{KL}\left(q_\theta^{(0)}(z) \| p(z \mid x)\right)+\mathrm{KL}\left(p(z \mid x) \| q_\theta^{(0)}(z)\right)$

3.1 Gradients of VCD

$$\mathcal{L}_{\mathrm{VCD}}(\theta)=\underbrace{\mathrm{KL}\left(q_\theta^{(0)}(z)\|p(z\mid x)\right)-\mathrm{KL}\left(q_\theta(z)\|p(z\mid x)\right)}_{\geq 0}+\underbrace{\mathrm{KL}\left(q_\theta(z)\|q_\theta^{(0)}(z)\right)}_{\geq 0}$$

re express as...

$$\mathcal{L}_{\mathrm{VCD}}(\theta)=-\mathbb{E}_{q_\theta^{(0)}(z)}\left[\log p(x,z)-\log q_\theta^{(0)}(z)\right]+\mathbb{E}_{q_\theta(z)}\left[\log p(x,z)-\log q_\theta^{(0)}(z)\right]$$

First component

negative ELBO : $-\mathbb{E}_{q_\theta^{(0)}(z)}\left[\log p(x,z)-\log q_\theta^{(0)}(z)\right]$

- use reparameterization trick or score-function gradients

Second component

$\mathbb{E}_{q_\theta(z)}\left[\log p(x,z)-\log q_\theta^{(0)}(z)\right]$

- if we take derivative....

  $\nabla_\theta \mathbb{E}_{q_\theta(z)}\left[g_\theta(z)\right]=-\mathbb{E}_{q_\theta(z)}\left[\nabla_\theta \log q_\theta^{(0)}(z)\right]+\mathbb{E}_{q_\theta^{(0)}\left(z_0\right)}\left[\mathbb{E}_{Q^{(t)}\left(z \mid z_0\right)}\left[g_\theta(z)\right] \nabla_\theta \log q_\theta^{(0)}\left(z_0\right)\right]$

  (use MC approximation)

4. Algorithm to Optimize VCD

objective function :

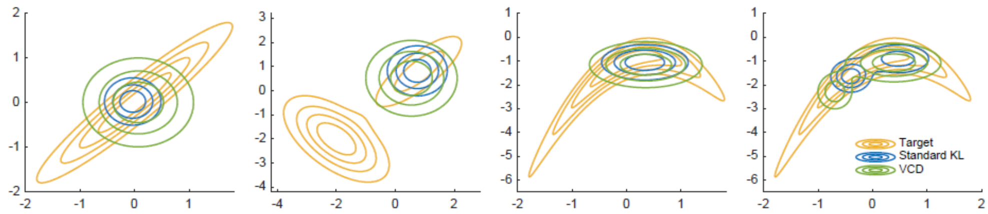
$$\mathcal{L}_{\mathrm{VCD}}(\theta)=-\mathbb{E}_{q_\theta^{(0)}(z)}\left[\log p(x,z)-\log q_\theta^{(0)}(z)\right]+\mathbb{E}_{q_\theta(z)}\left[\log p(x,z)-\log q_\theta^{(0)}(z)\right]$$

Steps :

1. Sample $z_0 \sim q_\theta^{(0)}(z)$ (reparameterization)
2. Sample $z \sim Q^{(t)}\left(z \mid z_0\right)$ (run t MCMC steps)
3. Estimate the gradient $\nabla_\theta \mathcal{L}_{\mathrm{VCD}}(\theta)$
4. Take gradient step w.r.t. θ

Leads to $q_\theta^{(0)}(z)$ with higher variances!

5. Examples


Target
Standard KL
VCD