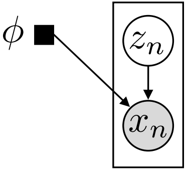
A Contrastive Divergence for Combining Variational Inference and MCMC

Francisco J. R. Ruiz* Michalis K. Titsias

EMS 2019

Inference for Latent Variable Models

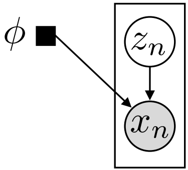
- Inference and learning in latent variable models
  - Probabilistic PCA
  - Matrix factorization
  - Variational autoencoders
  - ...

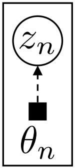


$$p_\phi(x, z) = \prod_n p_\phi(x_n \mid z_n) p(z_n)$$

Variational Inference

- Variational inference: Joint inference and learning
- Approximate the posterior $p_\phi(z \mid x) \approx q_\theta(z)$
- Factorization $q_\theta(z) = \prod_n q_\theta(z_n)$



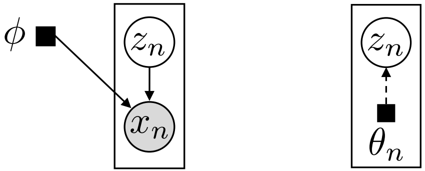


Variational Inference

- Maximize the ELBO w.r.t. model and variational parameters

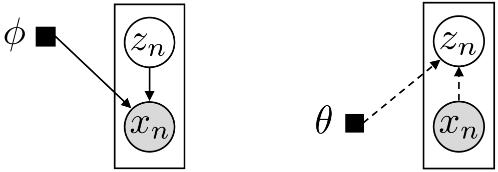
$$\mathcal{L}_{\text{standard}} = \sum_n \mathbb{E}_{q_\theta(z_n)} \left[\log p_\phi(z_n, x_n) - \log q_\theta(z_n)\right]$$

- Equivalent to minimizing $\mathrm{KL}(q_\theta(z) \,||\, p_\phi(z \,|\, x))$



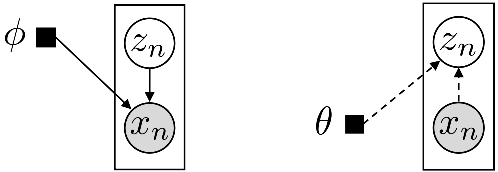
Advantages of Variational Inference

- Amortization quickly forms an approximation of the posterior $p_\phi(z_n \mid x_n) \approx q_\theta(z_n \mid x_n)$
    - Reduces number of parameters
    - Improves scalability



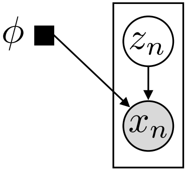
Limitations of Variational Inference

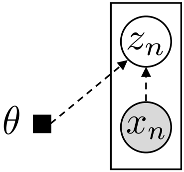
- Approximation gap: $q_\theta(z_n \,|\, x_n)$ has parametric form (Gaussian)
- Amortization gap: the parameters of $q_\theta(z_n \,|\, x_n)$ are not optimal (they are a function of x_n)



This Work: Improve VI using MCMC

- VI: Scalable but might be inaccurate
- MCMC: Asymptotically unbiased but typically slower
- This work: Combine the advantages of both





Main Idea: Refine the Approximation with MCMC

- Goals:
  - Increase the expressiveness of the variational family
  - Improve a variational distribution $q_\theta(z)$
- Draw samples from $q_\theta(z)$ and refine them with MCMC
- Optimize $q_\theta(z)$ to provide a good initialization for MCMC
- For tractable inference: Replace the KL with the **VCD divergence**

Refine the Variational Distribution with MCMC

- Start from an *explicit* variational distribution, $q_\theta^{(0)}(z)$

- Improve the distribution with t MCMC steps,

$$z_0 \sim q_\theta^{(0)}(z), \qquad z \sim Q^{(t)}(z \mid z_0)$$

  The MCMC sampler targets the posterior $p(z \mid x)$

- Implicit distribution

$$q_\theta(z) = \int q_\theta^{(0)}(z_0) Q^{(t)}(z \mid z_0) dz_0$$

Challenges of Using MCMC in VI

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

- Challenge #1: The variational objective becomes intractable
- Challenge #2: The variational objective may depend *weakly* on θ

$$q_\theta(z) \xrightarrow{t \to \infty} p(z \mid x)$$

Alternative Divergence: VCD

- We would like an objective that avoids these challenges
- We call the objective *Variational Contrastive Divergence*, $\mathcal{L}_{\mathrm{VCD}}(\theta)$
- Desired properties:
  - Non-negative for any θ
  - Zero only if $q_\theta^{(0)}(z) = p(z \mid x)$

Variational Contrastive Divergence

- Key idea: The improved distribution $q_\theta(z)$ decreases the KL

$$\mathrm{KL}(q_\theta^{(0)}(z) \,||\, p(z \,|\, x)) \geq \mathrm{KL}(q_\theta(z) \,||\, p(z \,|\, x))$$

(equality only if $q_\theta^{(0)}(z) = p(z \,|\, x)$)

- A first objective:

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

(it is a proper divergence)

Variational Contrastive Divergence

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

- Still intractable: $\log q_\theta(z)$ in the second term
- Add regularizer,

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

(still a proper divergence)

Variational Contrastive Divergence

$$\mathcal{L}_{\mathrm{VCD}}(\theta) = \mathrm{KL}(q_\theta^{(0)}(z) \,||\, p(z \,|\, x)) - \mathrm{KL}(q_\theta(z) \,||\, p(z \,|\, x)) + \mathrm{KL}(q_\theta(z) \,||\, q_\theta^{(0)}(z))$$

- Addresses Challenge #1 (intractability):
  - The intractable term $\log q_\theta(z)$ cancels out
- Addresses Challenge #2 (weak dependence):
  - $\mathcal{L}_{\mathrm{VCD}}(\theta) \xrightarrow{t \to \infty} \mathrm{KL}(q_\theta^{(0)}(z) \,||\, p(z \,|\, x)) + \mathrm{KL}(p(z \,|\, x) \,||\, q_\theta^{(0)}(z))$

Taking Gradients of the VCD

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

- The first component is the (negative) standard ELBO
  - Use reparameterization or score-function gradients
- The second component is the new part,

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

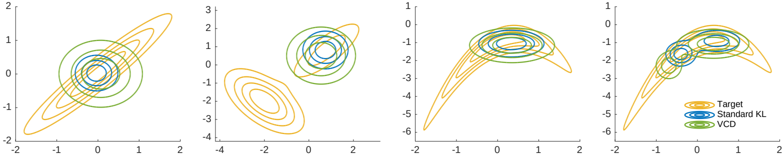
(can be approximated via Monte Carlo)

Algorithm to Optimize the VCD

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

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

Toy Experiments



Optimizing the VCD leads to a distribution $q_\theta^{(0)}(z)$ with higher variance

$$\mathcal{L}_{\mathrm{VCD}}(\theta) \xrightarrow{t\to\infty} \mathrm{KL}_{\mathrm{sym}}(q_\theta^{(0)}(z)\ ,\ p(z\,|\,x))$$

Experiments: Latent Variable Models

- Model is $p_\phi(x, z) = \prod_n p(z_n) p_\phi(x_n \mid z_n)$
- Amortized distribution $q_\theta(z_n \mid x_n) = \int Q^{(t)}(z_n \mid z_0) q_\theta^{(0)}(z_0 \mid x_n) dz_0$
- Goal: Find model parameters ϕ and variational parameters θ

method	average test log-likelihood MNIST	Fashion-MNIST
Explicit + KL	−111.20	−127.43
Implicit + KL (Hoffman, 2017)	−103.61	−121.86
VCD (this talk)	**−101.26**	**−121.11**

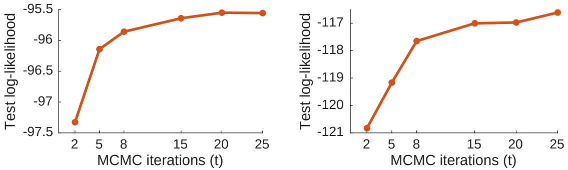
(a) Logistic matrix factorization

method	average test log-likelihood MNIST	Fashion-MNIST
Explicit + KL	−98.46	−124.63
Implicit + KL (Hoffman, 2017)	−96.23	−117.74
VCD (this talk)	**−95.86**	**−117.65**

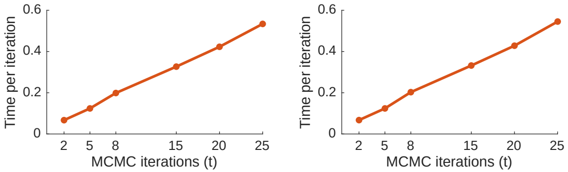
(b) VAE

Impact of Number of MCMC Steps

- More MCMC steps: Models with better predictive performance



- More MCMC steps: Higher computational cost



Conclusion

- Expand the variational family $q_\theta(z)$
- Key ideas: Define an *implicit* distribution
  - Improve the variational approximation with a few MCMC steps
  - Tractable inference by optimizing the *VCD divergence*
- Better predictive performance in latent variable models

This project is funded by the European Commission's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement 706760