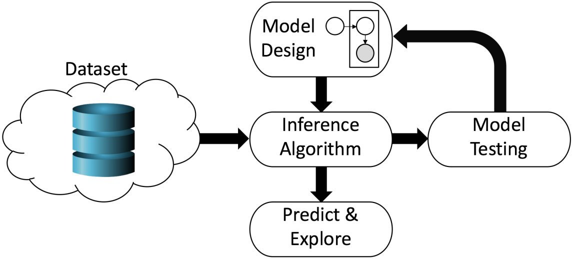
Beyond the Mean-Field Family: Variational Inference with Implicit Distributions

Francisco J. R. Ruiz
Marie Skłodowska-Curie Fellow

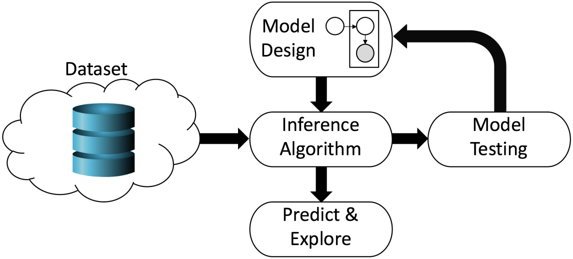
Linköping University
May 8th, 2019

Probabilistic Modeling Pipeline



- Posit generative process with hidden and observed variables
- Given the data, reverse the process to infer hidden variables
- Use hidden structure to make predictions, explore the dataset, etc.

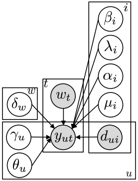
Probabilistic Modeling Pipeline

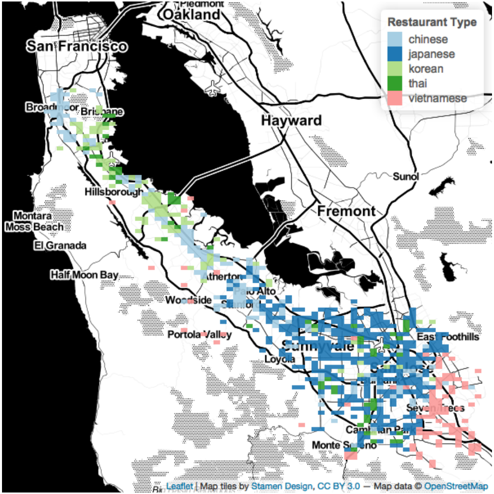


- Incorporate domain knowledge
- Separate assumptions from computation
- Facilitate collaboration with domain experts

Applications: Consumer Preferences

Can we use mobile location data to find
the most promising location for a new restaurant?

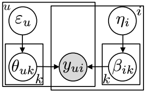


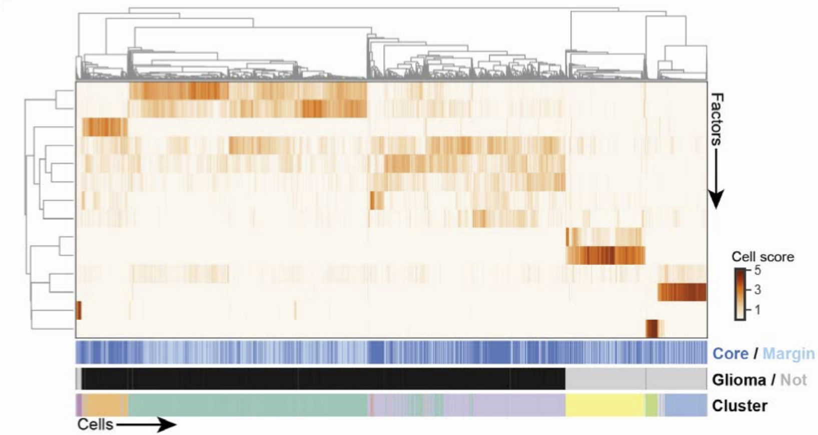


Restaurants in the Bay Area

Applications: Gene Signature Discovery

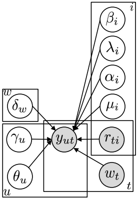
Can we identify *de novo* gene expression patterns in scRNA-seq?

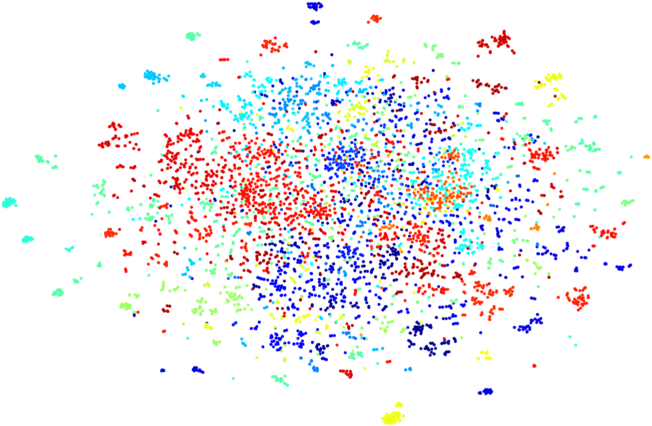




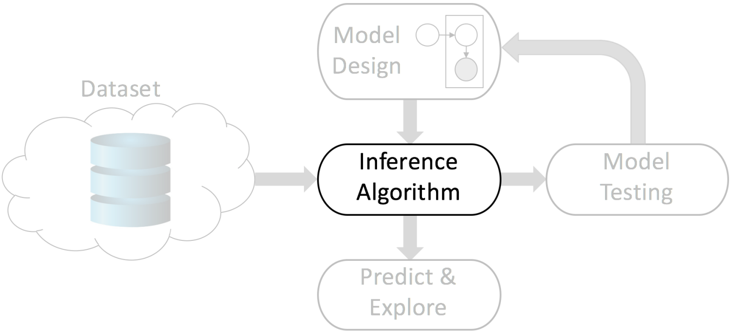
Applications: Shopping Behavior

Can we use past shopping transactions to learn customer preferences and predict demand under price interventions?





Inference


Dataset
Model Design
Inference Algorithm
Model Testing
Predict & Explore

Notation

- Model: Joint distribution $p(x, z)$
- Latent variables z
- Observations x

The Posterior Distribution

$$p(z \mid x) = \frac{p(x, z)}{\int p(x, z) dz}$$

- The posterior allows us to explore the data and make predictions
- Intractable in general
- Approximate the posterior: Bayesian inference

Variational Inference

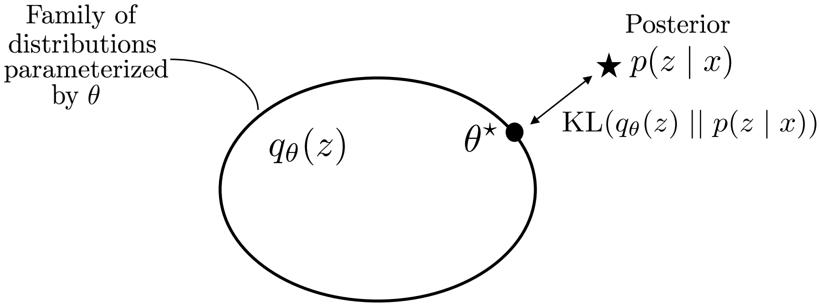
$$p(z \mid x) = \frac{p(x, z)}{\int p(x, z) dz}$$

- Define a simple family of distributions $q_\theta(z)$ with parameters θ
- Fit θ by minimizing the KL divergence to the posterior,

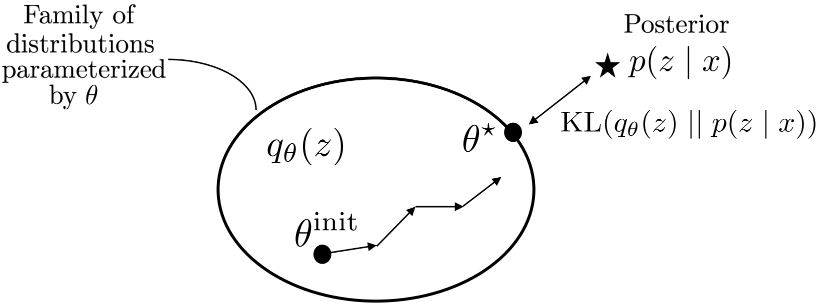
$$\theta^\star = \arg\min_\theta \mathrm{KL}\big(q_\theta(z) \,||\, p(z \mid x)\big)$$

- Variational inference solves an optimization problem

Variational Inference



Variational Inference



Variational Inference

- Minimizing the KL $\equiv$ Maximizing the ELBO

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

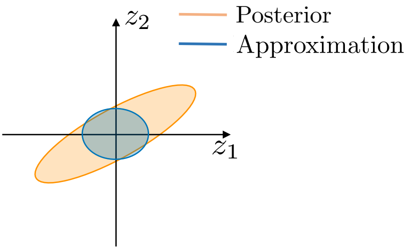
- Variational inference finds θ to maximize $\mathcal{L}(\theta)$

Mean-Field Variational Inference

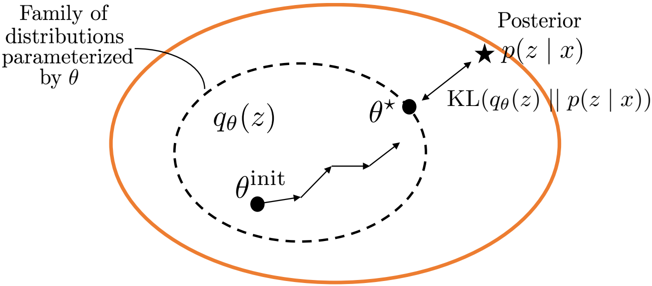
- Classical VI: Mean-field variational distribution:

$$q_\theta(z) = \prod_n q_{\theta_n}(z_n)$$

- Useful and simple, but might not be accurate



This Talk



Family of distributions parameterized by θ
$q_\theta(z)$
$\theta^\star$
θ^{init}
Posterior
$p(z \mid x)$
$\text{KL}(q_\theta(z) \mid\mid p(z \mid x))$

This Talk

- Expand the variational family $q_\theta(z)$
- Key idea: Use *implicit distributions*
  - Easy to sample from, $z \sim q_\theta(z)$
  - Intractable density, $q_\theta(z)$
- Challenge: Solve the optimization problem with intractable $q_\theta(z)$

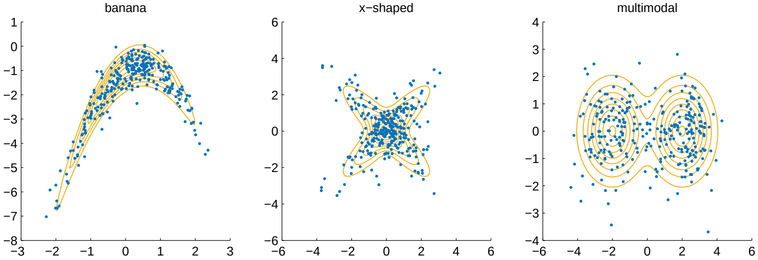
Related Work

- Structured variational inference
- Mixtures
- Hierarchical variational models
- Normalizing flows
- Other methods for implicit distributions

Part I:

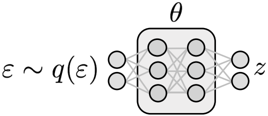
Semi-Implicit Construction

Our Goal: More Expressive Variational Distributions



Blue dots: samples from $q_\theta(z)$

Variational Inference with Implicit Distributions



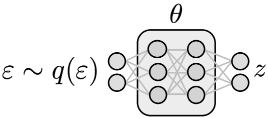
- Easy to draw samples from $q_\theta(z)$:

$$\text{sample } \varepsilon \sim q(\varepsilon); \qquad \text{set } z = f_\theta(\varepsilon)$$

- Cannot evaluate the density $q_\theta(z)$

- Flexible distribution due to the NN

VI with Implicit Distributions is Hard



- The VI objective is the ELBO (equivalent to minimizing KL),

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(z)}\Big[\underbrace{\log p(x,z)}_{\text{model}} - \underbrace{\log q_\theta(z)}_{\text{entropy}}\Big]$$

- Gradient of the objective $\nabla_\theta \mathcal{L}(\theta)$ (reparameterization)

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)}\Big[\nabla_\theta\Big(\log p(x, f_\theta(\varepsilon)) - \log q_\theta\left(f_\theta(\varepsilon)\right)\Big)\Big]$$

- Monte Carlo estimates require $\nabla_z \log q_\theta(z)$ (not available)

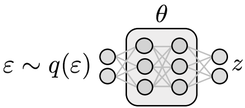
Unbiased Implicit Variational Inference

- UIVI obtains an unbiased Monte Carlo estimator of $\nabla_z \log q_\theta(z)$
- It avoids density ratio estimation
- Key ideas:
  1. Semi-implicit construction of $q_\theta(z)$
  2. Gradient of the entropy component as an expectation,

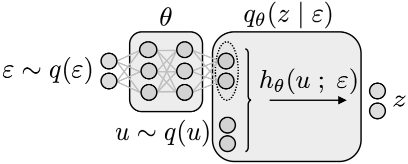
$$\nabla_z \log q_\theta(z) = \mathbb{E}_{\mathrm{distrib}(\cdot)}\left[\mathrm{function}(z, \cdot)\right]$$

UIVI Step 1: Semi-Implicit Distribution

- Implicit distribution:

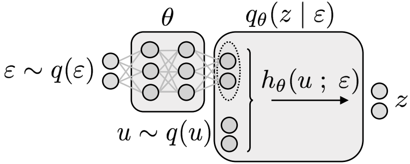


- (Semi-)implicit distribution:



UIVI Step 1: Semi-Implicit Distribution

- (Semi-)implicit distribution

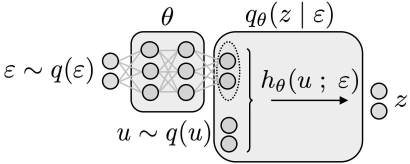


- **Example:** The conditional $q_\theta(z \,|\, \varepsilon)$ is a Gaussian,

$$q_\theta(z \,|\, \varepsilon) = \mathcal{N}\left(z \,|\, \mu_\theta(\varepsilon), \Sigma_\theta(\varepsilon)\right)$$

UIVI Step 1: Semi-Implicit Distribution

- (Semi-)implicit distribution



- The distribution $q_\theta(z)$ is still **implicit**,
  - Easy to sample,

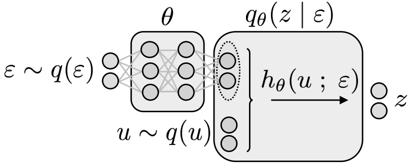
$$\text{sample } \varepsilon \sim q(\varepsilon),$$
$$\text{obtain } \mu_\theta(\varepsilon) \text{ and } \Sigma_\theta(\varepsilon)$$
$$\text{sample } z \sim \mathcal{N}(z \mid \mu_\theta(\varepsilon), \Sigma_\theta(\varepsilon))$$

  - The variational distribution $q_\theta(z)$ is not tractable,

$$q_\theta(z) = \int q(\varepsilon) q_\theta(z \mid \varepsilon) d\varepsilon$$

UIVI Step 1: Semi-Implicit Distribution

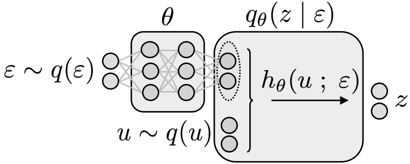
- (Semi-)implicit distribution



- **Assumptions** on the conditional $q_\theta(z \mid \varepsilon)$:
  - Reparameterizable
  - Tractable gradient $\nabla_z \log q_\theta(z \mid \varepsilon)$
    Note: this is different from $\nabla_z \log q_\theta(z)$ (still intractable)

UIVI Step 1: Semi-Implicit Distribution

- (Semi-)implicit distribution



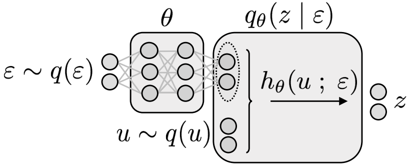
- The Gaussian meets both assumptions:
  - Reparameterizable,

$$u \sim \mathcal{N}(u \mid 0, I), \qquad z = h_\theta(u\,;\,\varepsilon) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2} u$$

  - Tractable gradient,

$$\nabla_z \log q_\theta(z \mid \varepsilon) = -\Sigma_\theta(\varepsilon)^{-1}(z - \mu_\theta(\varepsilon))$$

UIVI Step 2: Gradient as Expectation



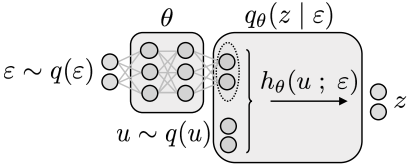
- Goal: Estimate the gradient of the entropy component, $\nabla_z \log q_\theta(z)$
- Rewrite as an expectation,

$$\nabla_z \log q_\theta(z) = \mathbb{E}_{q_\theta(\varepsilon' \mid z)}\left[\nabla_z \log q_\theta(z \mid \varepsilon')\right]$$

- Form Monte Carlo estimate,

$$\nabla_z \log q_\theta(z) \approx \nabla_z \log q_\theta(z \mid \varepsilon'), \qquad \varepsilon' \sim q_\theta(\varepsilon' \mid z)$$

UIVI: Full Algorithm

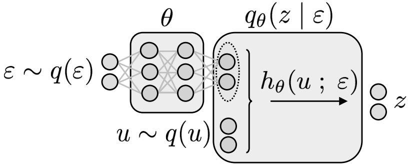


- The gradient of the ELBO is

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)q(u)}\left[\nabla_z \left(\log p(x,z) - \log q_\theta(z)\right)\Big|_{z=h_\theta(u\,;\,\varepsilon)} \times \nabla_\theta h_\theta(u\,;\,\varepsilon)\right]$$

- Estimate the gradient based on samples:
  1. Sample $\varepsilon \sim q(\varepsilon)$, $u \sim q(u)$ (standard Gaussians)
  2. Set $z = h_\theta(\varepsilon\,;\,u) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2}u$
  3. Evaluate $\nabla_z \log p(x,z)$ and $\nabla_\theta h_\theta(u\,;\,\varepsilon)$
  4. Sample $\varepsilon' \sim q_\theta(\varepsilon' \mid z)$
  5. Approximate $\nabla_z \log q_\theta(z) \approx \nabla_z \log q_\theta(z \mid \varepsilon')$

UIVI: The Reverse Conditional

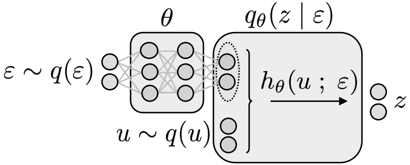


- The distribution $q_\theta(\varepsilon' \mid z)$ is the **reverse conditional**
  The conditional is $q_\theta(z \mid \varepsilon)$
- Sample from $q_\theta(\varepsilon' \mid z)$ using HMC, targeting

$$q(\varepsilon' \mid z) \propto q(\varepsilon') q_\theta(z \mid \varepsilon')$$

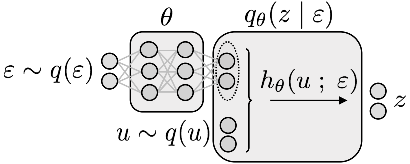
- Problem: HMC is slow... How to accelerate this?

UIVI: The Reverse Conditional



- Recall the UIVI algorithm,
  1. Sample $\varepsilon \sim q(\varepsilon)$, $u \sim q(u)$ (standard Gaussians)
  2. Set $z = h_\theta(\varepsilon\,;\,u) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2}u$
  3. Evaluate $\nabla_z \log p(x, z)$ and $\nabla_\theta h_\theta(u\,;\,\varepsilon)$
  4. Sample $\varepsilon' \sim q_\theta(\varepsilon' \mid z)$
  5. Approximate $\nabla_z \log q_\theta(z) \approx \nabla_z \log q_\theta(z \mid \varepsilon')$
- We have that $(\varepsilon, z) \sim q_\theta(\varepsilon, z) = q(\varepsilon)q_\theta(z \mid \varepsilon) = q_\theta(z)q_\theta(\varepsilon \mid z)$
- Thus, ε is a sample from $q_\theta(\varepsilon \mid z)$
- To accelerate sampling $\varepsilon' \sim q(\varepsilon' \mid z)$, initialize HMC at ε

UIVI: The Reverse Conditional

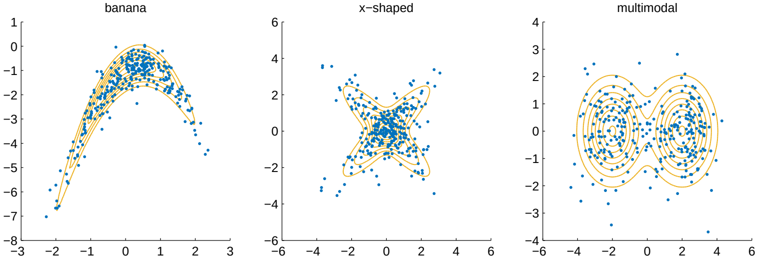


- Sample from $q_\theta(\varepsilon' \mid z)$ using HMC targeting

$$q(\varepsilon' \mid z) \propto q(\varepsilon') q_\theta(z \mid \varepsilon')$$

- Initialize HMC at stationarity (using ε)
- A few HMC iterations to reduce correlation between ε and ε'

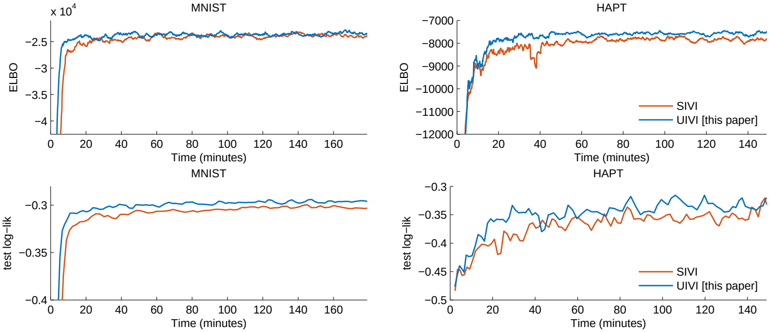
Toy Experiments



Blue dots: samples from $q_\theta(z)$

Experiments: Multinomial Logistic Regression

$$p(x,z) = p(z)\prod_{n=1}^{N}\frac{\exp\{x_n^\top z_{y_n} + z_{0y_n}\}}{\sum_k \exp\{x_n^\top z_k + z_{0k}\}}$$



UIVI provides better ELBO and predictive performance than SIVI

Experiments: VAE

- Model is $p_\phi(x, z) = \prod_n p(z_n) p_\phi(x_n \mid z_n)$
- Amortized variational distrib. $q_\theta(z_n \mid x_n) = \int q(\varepsilon_n) q_\theta(z_n \mid \varepsilon_n, x_n) d\varepsilon_n$
- Goal: Find model parameters ϕ and variational parameters θ

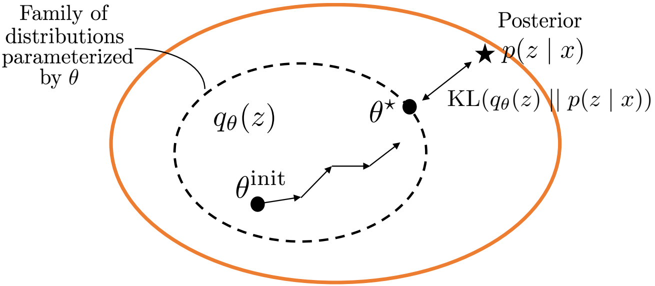
	average test log-likelihood	
method	MNIST	Fashion-MNIST
Explicit (standard VAE)	−98.29	−126.73
SIVI	−97.77	−121.53
UIVI (this talk)	**−94.09**	**−110.72**

UIVI provides better predictive performance

Part II:

MCMC-Improved Approximation

Our Goal: More Expressive Variational Distributions



Main Idea: Use 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 variational 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

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

(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,

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

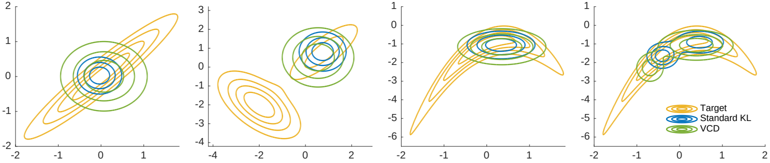
(can be approximated via Monte Carlo)

Algorithm to Optimize the VCD

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

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 $\nabla_\theta \mathcal{L}_{\text{VCD}}(\theta)$
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

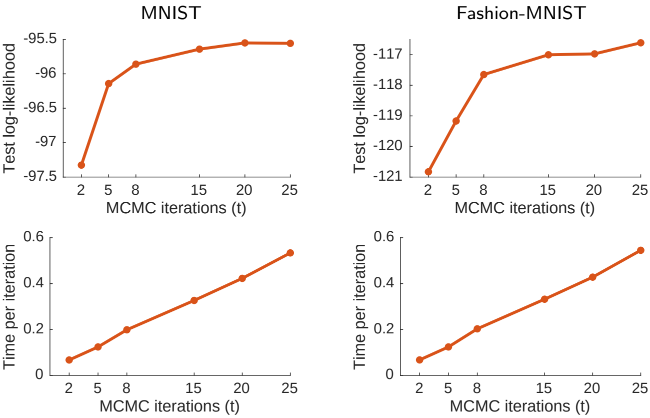
method	average test log-likelihood MNIST	average test log-likelihood Fashion-MNIST
Explicit + KL	−111.20	−127.43
Implicit + KL (Hoffman, 2017)	−103.61	−121.86
VCD (this talk)	**−101.26**	**−121.11**

(a) Logistic matrix factorization

method	average test log-likelihood MNIST	average test log-likelihood 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

Experiments: Impact of Number of MCMC Steps



Summary

- Use *implicit distributions* to form expressive variational posteriors
- UIVI: Hierarchy of tractable distributions
- VCD: Refine the variational approximation with MCMC
  - Optimize a novel divergence (VCD)
  - Leverage the advantages of both VI and MCMC
- Stable training
- Good empirical results on (deep) probabilistic models

EU H2020 (MSCA Actions)

Proof of the Key Equation in UIVI

- Goal: Prove that

$$\nabla_z \log q_\theta(z) = \mathbb{E}_{q_\theta(\varepsilon \mid z)} \left[\nabla_z \log q_\theta(z \mid \varepsilon)\right]$$

- Start with log-derivative identity,

$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \nabla_z q_\theta(z)$$

- Apply the definition of $q_\theta(z)$ through a mixture,

$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \int \nabla_z q_\theta(z \mid \varepsilon) q(\varepsilon) d\varepsilon$$

- Apply the log-derivative identity on $q_\theta(z \mid \varepsilon)$,

$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \int q_\theta(z \mid \varepsilon) q(\varepsilon) \nabla_z \log q_\theta(z \mid \varepsilon) d\varepsilon.$$

- Apply Bayes' theorem

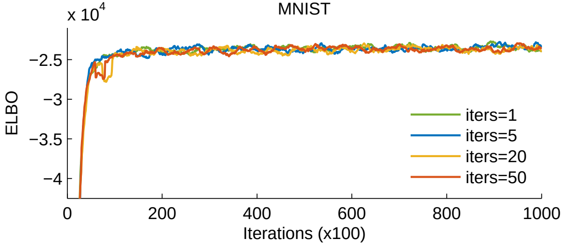
SIVI

- SIVI optimizes a lower bound of the ELBO,

$$\mathcal{L}_{\mathrm{SIVI}}^{(L)}(\theta) = \mathbb{E}_{\varepsilon \sim q(\varepsilon)}\left[\mathbb{E}_{z \sim q_\theta(z \,|\, \varepsilon)}\left[\mathbb{E}_{\varepsilon^{(1)},\ldots,\varepsilon^{(L)} \sim q(\varepsilon)}\left[\log p(x, z) - \log\left(\frac{1}{L+1}\left(q_\theta(z \,|\, \varepsilon) + \sum_{\ell=1}^{L} q_\theta(z \,|\, \varepsilon^{(\ell)})\right)\right)\right]\right]\right]$$

UIVI Experiments: Multinomial Logistic Regression

$$p(x, z) = p(z) \prod_{n=1}^{N} \frac{\exp\{x_n^\top z_{y_n} + z_{0y_n}\}}{\sum_k \exp\{x_n^\top z_k + z_{0k}\}}$$



Number of HMC iterations does not significantly impact results

Generalized VCD

- VCD

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

- α-generalized VCD

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