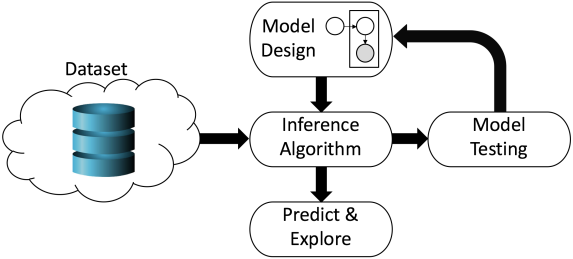
Variational Inference with Implicit and Semi-Implicit Distributions

Francisco J. R. Ruiz
Research Scientist - DeepMind, London

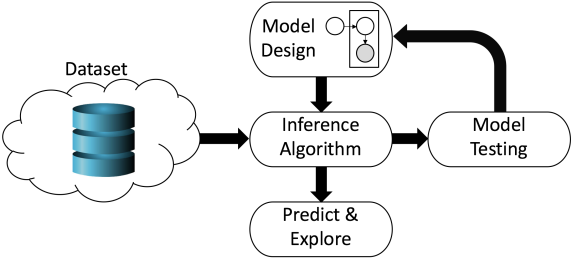
ProbAI Summer School
June 17, 2021

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.

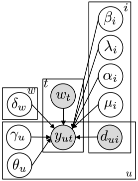
Probabilistic Modeling Pipeline

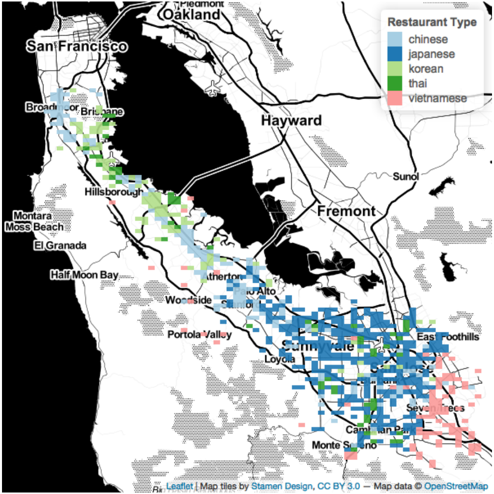


- Incorporate domain knowledge
- Separate assumptions from computation
- Facilitate collaboration with domain experts

Applications: Consumer Preferences

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

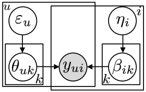


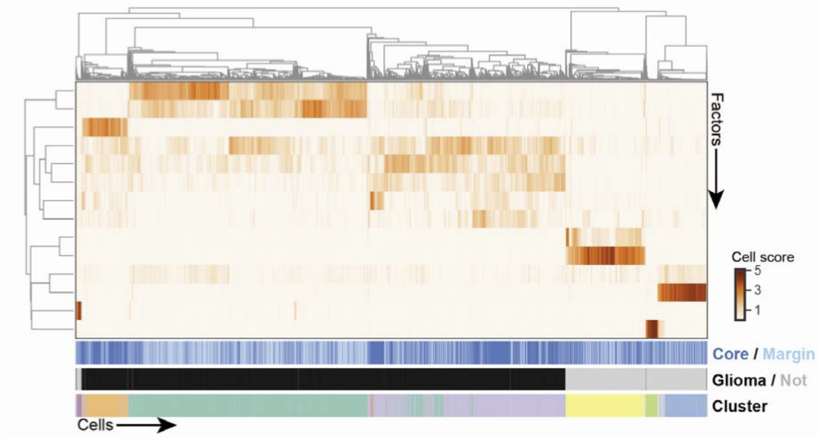


Restaurants in the Bay Area

Applications: Gene Signature Discovery

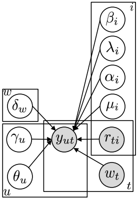
Can we identify *de novo* gene expression patterns in scRNA-seq?

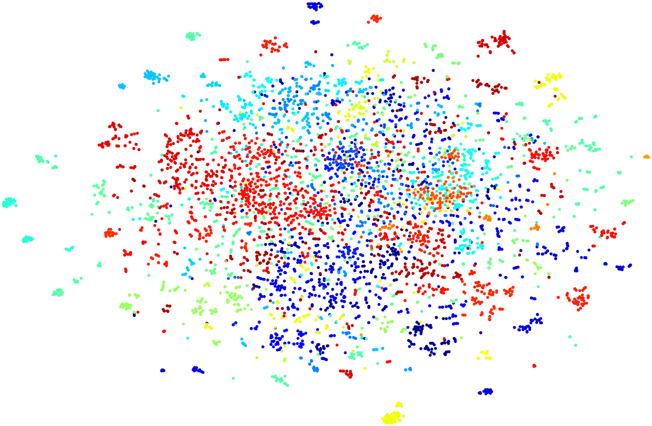




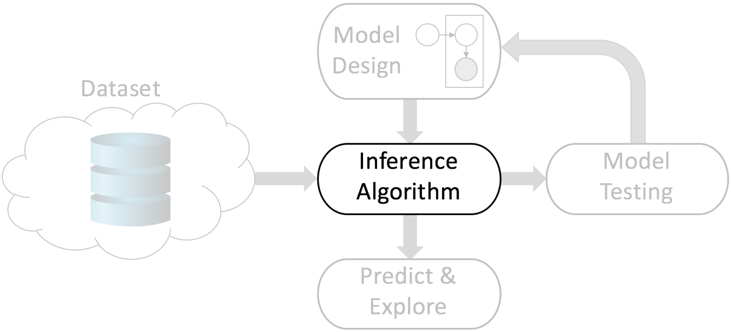
Applications: Shopping Behavior

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



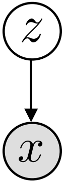


Inference


Model
Design
Dataset
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 (Quick Review)

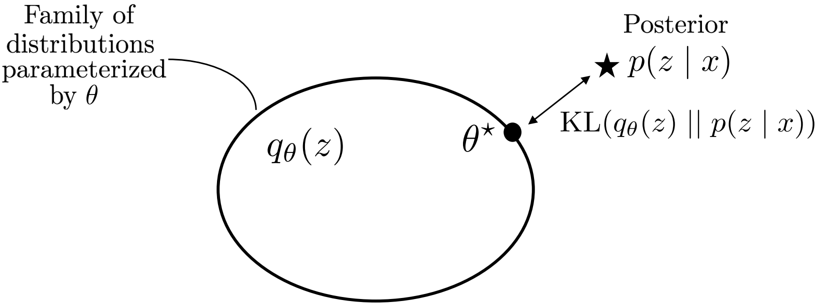
$$p(z \mid x) = \frac{p(x, z)}{\int p(x, z) dz}$$

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

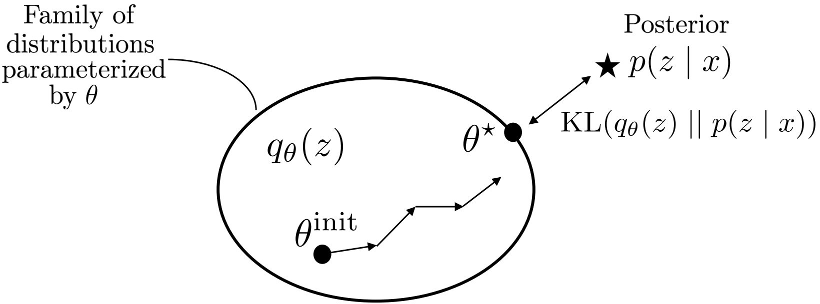
$$\theta^\star = \arg\min_\theta \mathrm{KL}\big(q_\theta(z) \,||\, p(z \mid x)\big)$$

- Variational inference solves an optimization problem

Variational Inference (Quick Review)



Variational Inference (Quick Review)



Variational Inference (Quick Review)

- Minimizing the KL $\equiv$ Maximizing the ELBO

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

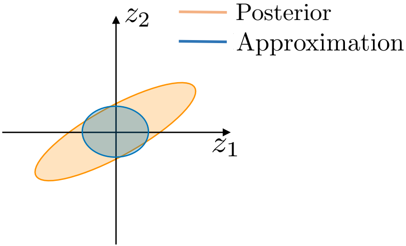
- Variational inference finds θ to maximize $\mathcal{L}(\theta)$

Mean-Field Variational Inference

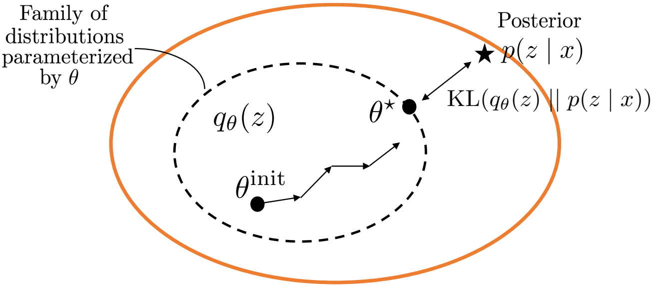
- Classical VI: Mean-field variational distribution:

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

- Useful, simple, and fast, but might not be accurate



This Lecture: Expand the Variational Family



Beyond the Mean-Field Family

- **Structured VI** [Saul+, 1996; Ghahramani+, 1997; Titsias+, 2011]
- **Mixtures** [Bishop+, 1998; Gershman+, 2012; Salimans+, 2013; Guo+, 2016; Miller+, 2017]
- **Sampling mechanisms** [Salimans+, 2015; Hoffman, 2017; Maddison+, 2017; Naesseth+, 2017; Li+, 2017; Titsias, 2017; Naesseth+, 2018; Le+, 2018; Grover+, 2018; Zhang+, 2018; Habib+, 2019; Neklyudov+, 2019; Ruiz+, 2019]
- **Spectral methods** [Shi+, 2018]
- **Linear response estimates** [Giordano+, 2015; Giordano+, 2017]
- **Copulas** [Tran+, 2015; Han+, 2016]
- **Invertible transformations** [Rezende+, 2014; Kingma+, 2014; Titsias+, 2014; Kucukelbir+, 2015] **& Normalizing flows** [Rezende+, 2015; Kingma+, 2016; Papamakarios+, 2017; Tomczak+, 2016; Tomczak+, 2017; Dinh+, 2017]
- **Hierarchical models** [Ranganath+, 2016; Tran+, 2016; Maaløe+, 2016; Sobolev+, 2019]
- **Implicit distributions** [Mescheder+, 2017; Huszár, 2017; Tran+, 2017; Shi+, 2018] **& Semi-implicit distributions** [Yin+, 2018; Titsias+, 2019; Molchanov+, 2019]

Beyond the Mean-Field Family

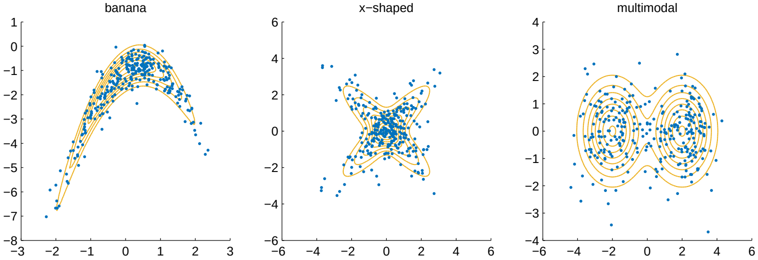
- **Structured VI** [Saul+, 1996; Ghahramani+, 1997; Titsias+, 2011]
- **Mixtures** [Bishop+, 1998; Gershman+, 2012; Salimans+, 2013; Guo+, 2016; Miller+, 2017]
- **Sampling mechanisms** [Salimans+, 2015; Hoffman, 2017; Maddison+, 2017; Naesseth+, 2017; Li+, 2017; Titsias, 2017; Naesseth+, 2018; Le+, 2018; Grover+, 2018; Zhang+, 2018; Habib+, 2019; Neklyudov+, 2019; Ruiz+, 2019]
- **Spectral methods** [Shi+, 2018]
- **Linear response estimates** [Giordano+, 2015; Giordano+, 2017]
- **Copulas** [Tran+, 2015; Han+, 2016]
- **Invertible transformations** [Rezende+, 2014; Kingma+, 2014; Titsias+, 2014; Kucukelbir+, 2015] **& Normalizing flows** [Rezende+, 2015; Kingma+, 2016; Papamakarios+, 2017; Tomczak+, 2016; Tomczak+, 2017; Dinh+, 2017]
- **Hierarchical models** [Ranganath+, 2016; Tran+, 2016; Maaløe+, 2016; Sobolev+, 2019]
- **Implicit distributions** [Mescheder+, 2017; Huszár, 2017; Tran+, 2017; Shi+, 2018] **& Semi-implicit distributions** [Yin+, 2018; Titsias+, 2019; Molchanov+, 2019]

This Lecture

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

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

Goal: More Expressive Variational Distributions

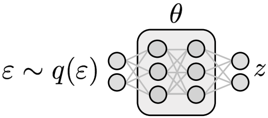


Blue dots: samples from $q_\theta(z)$

Part I:

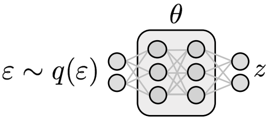
Implicit Distributions and Adversarial Training

How to Form an Expressive Implicit Distribution



- Generate random noise $\varepsilon \sim q(\varepsilon)$
- Pass the noise through a NN with parameters θ
- Let z be the output of the NN

How to Form an Expressive Implicit Distribution



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

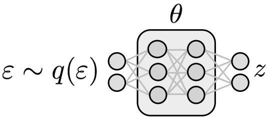
$$\text{sample } \varepsilon \sim q(\varepsilon); \qquad \text{set } z = f_\theta(\varepsilon)$$

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

- Flexible distribution $q_\theta(z)$ due to the NN

- Goal: Tune θ so that $q_\theta(z)$ approximates the posterior $p(z \mid x)$

Why VI with Implicit Distributions is Hard



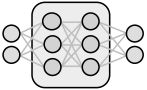
- The VI objective is the ELBO (equivalent to minimizing KL),

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

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

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

Why VI with Implicit Distributions is Hard



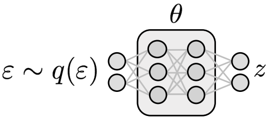
- Gradient of the objective $\nabla_\theta \mathcal{L}(\theta)$ (reparameterization)

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

- For the model term:

$$\mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \log p(x, f_\theta(\varepsilon))\right] \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\theta \log p(x, f_\theta(\varepsilon^{(s)})), \quad \varepsilon^{(s)} \sim q(\varepsilon)$$

Why VI with Implicit Distributions is Hard



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

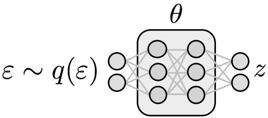
$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \Big(\log p(x, f_\theta(\varepsilon)) - \log q_\theta \left(f_\theta(\varepsilon)\right)\Big)\right]$$

- For the entropy term:

$$\nabla_\theta \log q_\theta \left(f_\theta(\varepsilon)\right) = \nabla_z \log q_\theta(z) \times \nabla_\theta f_\theta(\varepsilon) + \underbrace{\nabla_\theta \log q_\theta(z)\Big|_{z=f_\theta(\varepsilon)}}_{=0 \text{ (in expectation)}}$$

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

How Density Ratio Estimation Can Help



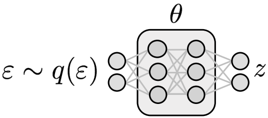
- The VI objective is the ELBO (equivalent to minimizing KL),

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

- Rewrite the ELBO as "log-likelihood minus KL to prior,"

$$\begin{aligned}\mathcal{L}(\theta) &= \mathbb{E}_{q_\theta(z)}\left[\log p(x \mid z)\right] - \mathrm{KL}\left(q_\theta(z) \mid\mid p(z)\right) \\ &= \mathbb{E}_{q_\theta(z)}\left[\log p(x \mid z)\right] - \mathbb{E}_{q_\theta(z)}\left[\log \frac{q_\theta(z)}{p(z)}\right]\end{aligned}$$

How Density Ratio Estimation Can Help



- ELBO objective:

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

- Key idea: Approximate the density ratio $\log \frac{q_\theta(z)}{p(z)}$

Density Ratio Estimation

- ELBO objective:

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

- Imagine that we had labelled samples from $q_\theta(z)$ and $p(z)$:
  - Class $y = 1$: The sample z comes from $q_\theta(z)$
  - Class $y = 0$: The sample z comes from $p(z)$

- If you observe z, what is the class? (under equal class prior)
  - Optimal classifier is $D^\star(z) = \frac{q_\theta(z)}{q_\theta(z)+p(z)}$

- The density ratio can be expressed as a function of the classifier:

$$\log \frac{q_\theta(z)}{p(z)} = \log D^\star(z) - \log(1 - D^\star(z))$$

Density Ratio Estimation

- The density ratio can be expressed as a function of the classifier:

$$\log \frac{q_\theta(z)}{p(z)} = \log D^\star(z) - \log(1 - D^\star(z))$$

- Train a (flexible) classifier $D(z)$ that distinguishes samples:

$$D^\star(z) = \max_D \mathbb{E}_{q_\theta(z)}[D(z)] + \mathbb{E}_{p(z)}[\log(1 - D(z))]$$

- Rewrite the ELBO using $D(z)$,

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(z)}[\log p(x \mid z)] - \mathbb{E}_{q_\theta(z)}[\log D(z) - \log(1 - D(z))]$$

Density Ratio Estimation: Optimization

- ELBO objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(z)}[\log p(x \mid z)] - \mathbb{E}_{q_\theta(z)}[\log D(z) - \log(1 - D(z))]$$

- Algorithm:

  1. Follow gradient estimates of the ELBO w.r.t. θ (reparameterization)
  2. For each θ, fit a flexible classifier $D(z)$ so that $D(z) \approx D^\star(z)$

Limitations of Density Ratio Estimation

- ELBO objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(z)}[\log p(x \mid z)] - \mathbb{E}_{q_\theta(z)}[\log D(z) - \log(1 - D(z))]$$

- Limitations:
  - The discriminator $D(z)$ needs to be trained to optimum after each update of θ (in practice, optimization is truncated to a few iterations)
  - Unstable training when discriminator does not catch up quickly
  - In **high dimensions**, the discriminator overfits easily, giving values close to 0 or 1

Alternatives

- Kernel-based density ratio estimation (KIVI) [Shi+, 2018]
- Semi-implicit distributions [Yin+, 2018; Titsias+, 2019; Molchanov+, 2019]

Alternatives

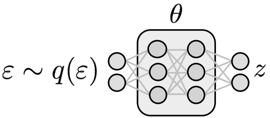
- Kernel-based density ratio estimation (KIVI) [Shi+, 2018]
- Semi-implicit distributions [Yin+, 2018; Titsias+, 2019; Molchanov+, 2019]

Part II:

Semi-Implicit Distributions

[Yin+, 2018; Titsias+, 2019; Molchanov+, 2019]

Recap: VI with Implicit Distributions



- ELBO objective:

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

- 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)

Semi-Implicit Distributions

- ELBO objective:

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

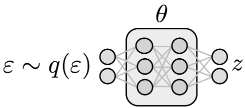
- Goal: Tractable inference avoiding density ratio estimation

- Two methods:
  - Lower-bound the ELBO (SIVI) [Yin+, 2018; Molchanov+, 2019]
  - Estimate gradients with sampling (UIVI) [Titsias+, 2019]

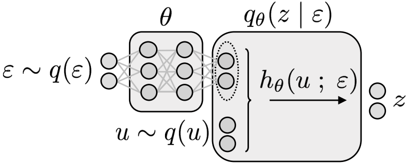
- First step: use a semi-implicit construction of $q_\theta(z)$

Semi-Implicit Distributions

- Implicit distribution:

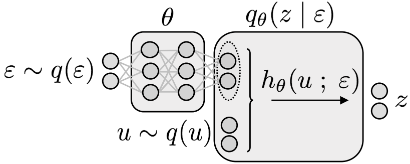


- (Semi-)implicit distribution:



Semi-Implicit Distributions

- (Semi-)implicit distribution

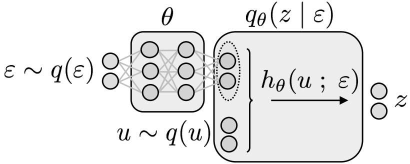


- **Example:** The conditional $q_\theta(z \,|\, \varepsilon)$ is a Gaussian,

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

Semi-Implicit Distributions

- (Semi-)implicit distribution



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

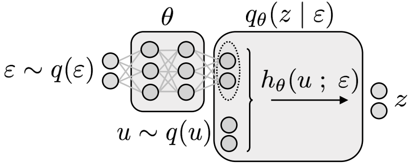
$$
\begin{aligned}
&\text{sample } \varepsilon \sim q(\varepsilon),\\
&\text{obtain } \mu_\theta(\varepsilon) \text{ and } \Sigma_\theta(\varepsilon)\\
&\text{sample } z \sim \mathcal{N}\left(z \mid \mu_\theta(\varepsilon), \Sigma_\theta(\varepsilon)\right)
\end{aligned}
$$

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

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

Semi-Implicit Distributions

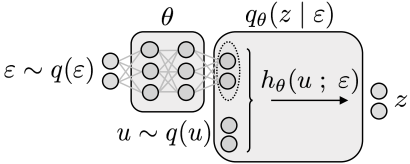
- (Semi-)implicit distribution



- **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)

Semi-Implicit Distributions

- (Semi-)implicit distribution



- The Gaussian meets both assumptions:
  - Reparameterizable,

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

  - Tractable gradient,

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

Method 1: SIVI

- Define a lower bound of the ELBO,

$$\mathcal{L}(\theta) \geq \overline{\mathcal{L}}(\theta), \quad \text{where}$$

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

- Optimize the lower bound instead of the ELBO

- The lower bound does not depend on the intractable $q_\theta(z)$

Method 1: SIVI

- SIVI bound:

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

- Free parameter L controls the tightness of the bound
  - As $L \to \infty$, $\overline{\mathcal{L}}(\theta) \to \mathcal{L}(\theta)$
  - Computational complexity increases with L

- SIVI allows for semi-implicit construction of prior in VAEs [Molchanov+, 2019]

Method 2: UIVI

- Recall the reparameterization gradient of the ELBO,

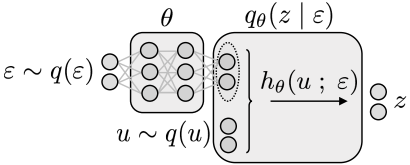
$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \Big(\log p(x, f_\theta(\varepsilon)) - \log q_\theta \left(f_\theta(\varepsilon)\right)\Big)\right]$$

- UIVI obtains an unbiased Monte Carlo estimator of $\nabla_z \log q_\theta(z)$
  - Avoid density ratio estimation
  - Directly optimize the ELBO (instead of a bound)

- Key idea: Gradient of the entropy component as an expectation,

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

Method 2: UIVI



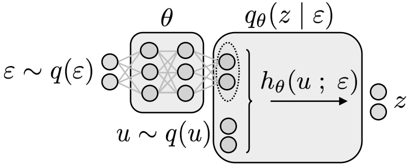
- Rewrite as an expectation,

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

- Form Monte Carlo estimate,

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

Method 2: UIVI (Full Algorithm)

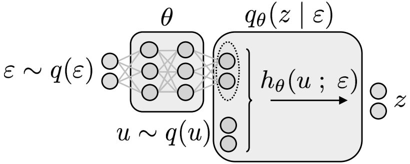


- The gradient of the ELBO is

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

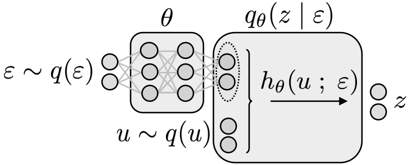
- Estimate the gradient based on samples:
  1. Sample $\varepsilon \sim q(\varepsilon)$, $u \sim q(u)$ (standard Gaussians)
  2. Set $z = h_\theta(\varepsilon\,;\,u) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2} u$
  3. Evaluate $\nabla_z \log p(x,z)$ and $\nabla_\theta h_\theta(u\,;\,\varepsilon)$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

Method 2: UIVI (The Reverse Conditional)



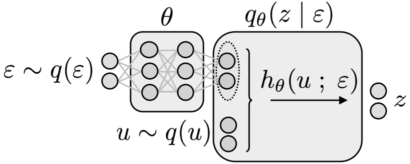
- The distribution $q_\theta(\varepsilon' \mid z)$ is the **reverse conditional**
  The conditional is $q_\theta(z \mid \varepsilon)$
- How to sample from $q_\theta(\varepsilon' \mid z)$ in the UIVI algorithm?

Method 2: UIVI (The Reverse Conditional)



- It turns out we already have a sample $\varepsilon \sim q_\theta(\varepsilon' \mid z)$
- 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 $q_\theta(\varepsilon \mid z)$

Method 2: UIVI (The Reverse Conditional)

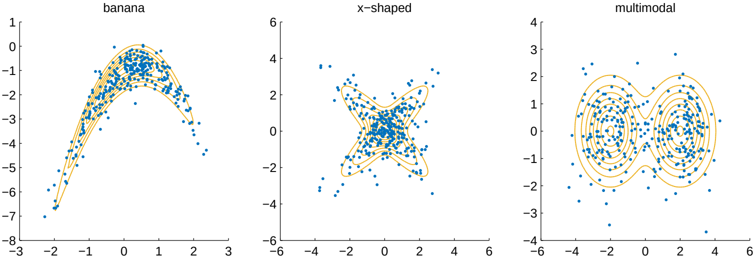


- But setting $\varepsilon' = \varepsilon$ is not correct, because both must be independent
- We use ε to initialize a HMC sampler targeting

$$q(\varepsilon' \,|\, z) \propto q(\varepsilon') q_\theta(z \,|\, \varepsilon')$$

A few HMC iterations reduce the correlation between ε' and ε

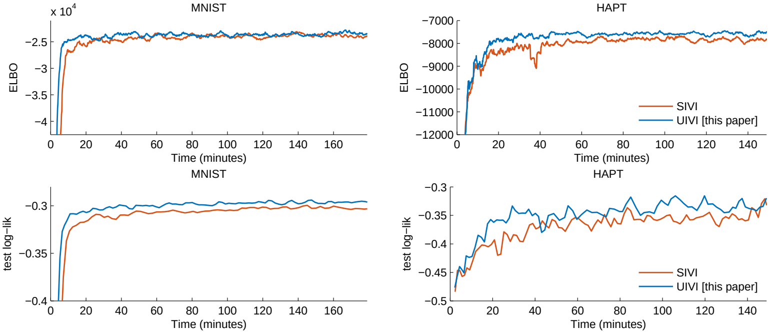
UIVI: Toy Experiments



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

UIVI: Multinomial Logistic Regression Experiments

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



UIVI provides better ELBO and predictive performance than SIVI

UIVI: VAE Experiments

- 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

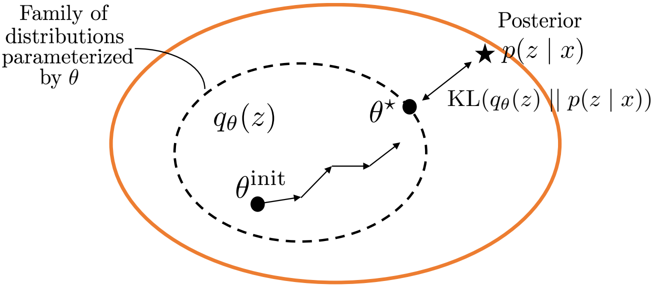
	average test log-likelihood	
method	MNIST	Fashion-MNIST
Explicit (standard VAE)	-98.29	-126.73
SIVI	-97.77	-121.53
UIVI	$-\mathbf{94.09}$	$-\mathbf{110.72}$

UIVI provides better predictive performance

Part III:

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,

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

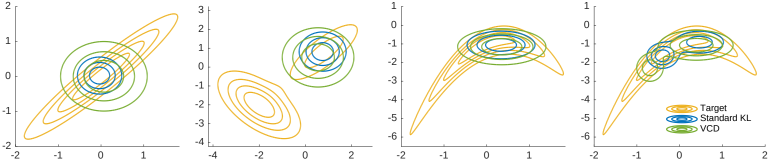
(can be approximated via Monte Carlo)

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

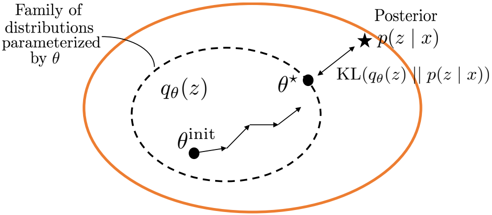
method	average test log-likelihood MNIST	Fashion-MNIST
Explicit + KL	−111.20	−127.43
Implicit + KL [Hoffman, 2017]	−103.61	−121.86
VCD	**−101.26**	**−121.11**

(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	**−95.86**	**−117.65**

(b) VAE

Summary



- Use *implicit distributions* to form expressive variational posteriors
  - Density ratio estimation
  - Semi-implicit distributions (SIVI, UIVI)
  - Refine the variational distribution with MCMC (VCD)
- Stable training
- Good empirical results on (deep) probabilistic models

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,

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

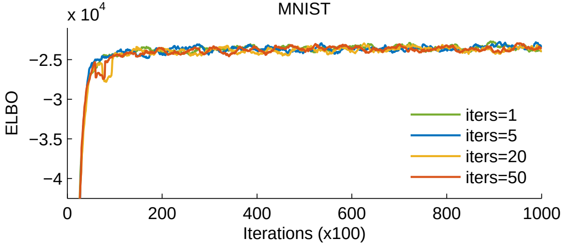
- Apply the log-derivative identity on $q_\theta(z \mid \varepsilon)$,

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

- Apply Bayes' theorem

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 ($0 < \alpha \leq 1$)

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

VCD Experiments: Impact of Number of MCMC Steps