Black Box Variational Inference

Complete conditionals → Gibbs, CAVI

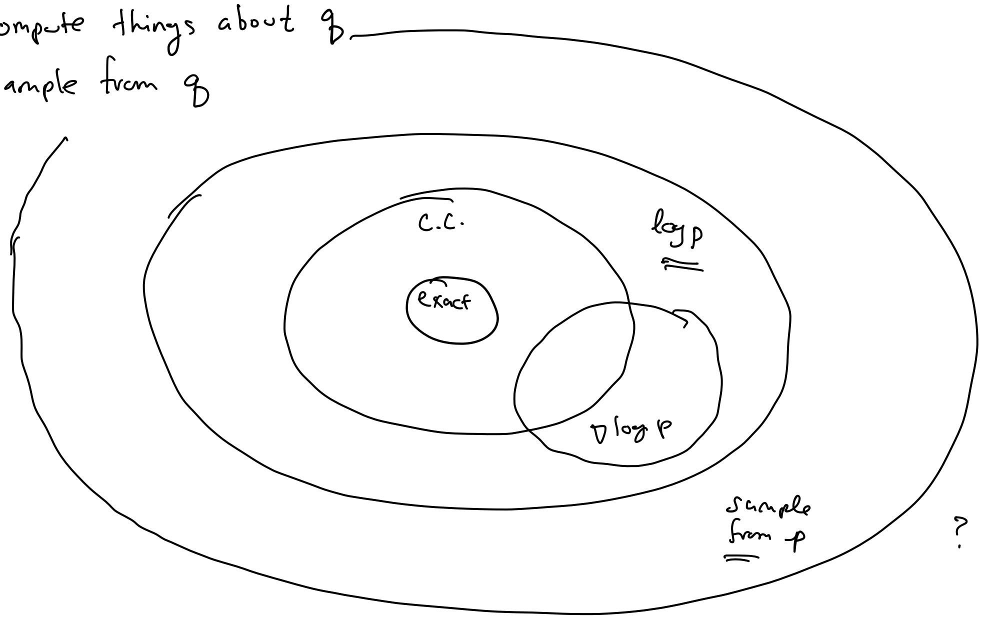
$p(\theta, z_{1:n}, x_{1:n})$ global θ local $z_{1:n}$

$$= p(\theta) \prod_{i=1}^{n} p(z_i \mid \theta)\, p(x_i \mid z_i, \theta)$$

$p(\theta, z_{1:n} \mid x_{1:n})$ - posterior. Goal: approximate it with $q(\theta, z_{1:n}; \nu)$

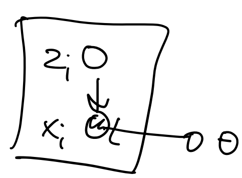
Black box criteria:

- compute the log joint (compute $\nabla_{\theta, z} \log p(z, \theta, x)$)
- compute things about q
- sample from q

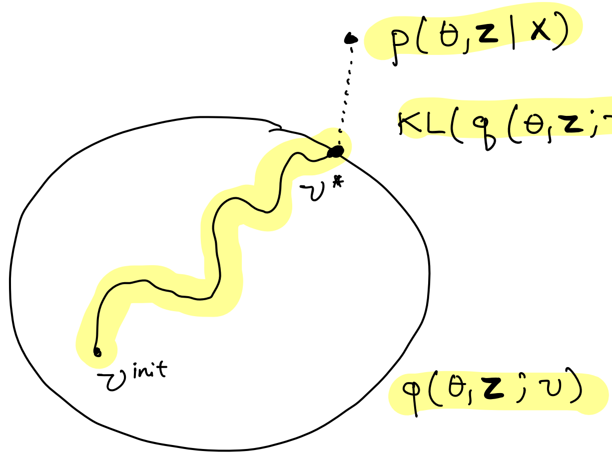


Recall: Deep generative model.

$$z_i \sim N_p(0, 1)$$
$$x_i \mid z_i, \theta \sim N(\Omega_\mu(z_i), \Omega^2_\sigma(z_i))$$



$$p(z_i \mid x_i, \theta) = \frac{p(z_i, x_i \mid \theta)}{p(x_i \mid \theta)} \qquad p(z_i, x_i \mid \theta) = p(z_i)\, p(x_i \mid z_i, \theta)$$



Objective function

EM: maximum likelihood estimation alg. when there are latent variables in the model.

ELBO

$$\mathcal{L}(\nu) = \mathbb{E}_{\nu}\left[\log p(\theta, Z, x) - \log q(\theta, Z; \nu)\right]$$

Strategy:

1. Write $\nabla_{\nu}\mathcal{L}$ as an expectation $\mathbb{E}_q[\sim]$
2. Take a Monte Carlo approximation of $\nabla_{\nu}\mathcal{L}$ $\approx \frac{1}{B}\sum_b \sim_b$
3. Use stochastic optimization

Score gradient

$$\nabla_{\nu}\mathcal{L} = \mathbb{E}\left[\underbrace{\nabla_{\nu}\log q(\theta, Z; \nu)}_{\text{score function}}\underbrace{\left(\log p(\theta, \mathbf{Z}, x) - \log q(\theta, \mathbf{Z}; \nu)\right)}_{\text{instantaneous ELBO}}\right]$$

$$\left.\begin{array}{r} \nabla_{\nu}\log q(\theta, Z; \nu) \\ \nabla_{\theta, z}\log q(\theta, Z; \nu) \\ \nabla \log q(\theta, Z; \nu) \end{array}\right] \text{"score function"}$$

Score VI

For iteration t:

$$\theta_b, \mathbf{z}_b \sim q(\theta, \mathbf{z}; v_t) \qquad b = 1..B$$

$$g_t = \frac{1}{B} \sum_{b=1}^{B} \nabla_v \log q(\theta_b, \mathbf{z}_b; v_t)\left(\log p(\theta_b, \mathbf{z}_b, \mathbf{x}) - \log q(\theta_b, \mathbf{z}_b; v_t)\right)$$

$$v_{t+1} = v_t + \rho_t g_t$$

Example: DGM

$$\log p(\theta, \mathbf{z}, \mathbf{x}) = \log p(\theta) + \sum_{i=1}^{n} \log p(z_i) + \log p(x_i \mid z_i, \theta)$$

$$q(\theta, \mathbf{z}; v) = \underset{\uparrow \atop \mathcal{N}}{q(\theta; v_\theta)} \prod_{i=1}^{n} \underset{\uparrow \atop \mathcal{N}}{q(z_i; v_i)}$$

In practice:

Variance is high. — Andy Miller, Justin Domke, ...

control variates.

Let $g(\theta, \mathbf{z}; v) \triangleq$ stochastic score gradient, $\mathbb{E}[g(\theta, \mathbf{z}; v)] = \nabla_v \mathcal{L}(v)$

Choose $h(\theta, \mathbf{z}; v)$ s/t $\mathbb{E}[h(\theta, \mathbf{z}; v)]$

Define $\tilde{g}(\theta, \mathbf{z}; v) = g(\theta, \mathbf{z}; v) - \underset{\uparrow \atop \text{scalar}}{\xi}\left(h(\theta, \mathbf{z}; v) - \mathbb{E}[h]\right)$

$$\operatorname{Var}(\tilde{g}) = \operatorname{Var}(g) + \xi^2 \operatorname{Var}(h) - 2\xi \operatorname{Cov}(g, h)$$

$$\xi^* = \frac{\operatorname{Cov}(g, h)}{\operatorname{Var}(h)}$$

$$\mathbb{E}\left[\nabla_\nu \log q(\mathbf{z};\nu)\right] = \int q(z;\nu)\, \nabla_\nu \log q(\mathbf{z};\nu)\, d\mathbf{z}$$

$$= \int \nabla_\nu q(z;\nu)\, dz$$

$$= \nabla_\nu \int q(z;\nu)\, dz$$

$$= \nabla_\nu 1 = \emptyset$$

FACT:

$$\nabla_\nu q(z;\nu) = q(z;\nu) \nabla_\nu \log q(z;\nu)$$

Reparameterization gradient

transform the variables in q:

$$\varepsilon \sim s(\varepsilon) \qquad \rightarrow \quad z \sim q(z;\nu)$$
$$z = t(\varepsilon, \nu)$$

e.g. for a normal:

$$\varepsilon \sim \mathcal{N}(0,1) \qquad \rightarrow \quad z \sim \mathcal{N}(\nu_\mu, \nu_\sigma^2)$$
$$z = \varepsilon \nu_\sigma + \nu_\mu$$

Reparameterization gradient

$$\nabla \mathcal{L}(\nu) =$$

$$\mathcal{L}(\nu) = \mathbb{E}_{s(\varepsilon)}\left[\log p(\theta, z, x) - \log q(\theta, z; \nu)\right]$$

where: $\theta = t(\varepsilon_\theta, \nu_\theta)$, $z = t(\varepsilon_z, \nu_z)$

$$\nabla_\nu \mathcal{L}(\nu) = \mathbb{E}_{s(\varepsilon)}\Big[\nabla_{\theta,z}\left[\log p(\theta,z,x) - \log q(\theta,z;\nu)\right]\nabla_\nu t(\varepsilon,\nu)$$
$$- \nabla_\nu \log q(\theta, z; \nu)\Big]$$

$$= \mathbb{E}_{s(\varepsilon)}\left[\nabla_{\theta,z}\left[\log p(\theta,z,x) - \log q(\theta,z;\nu)\right]\nabla_\nu t(\varepsilon,\nu)\right]$$

Score gradient vs. Reparameterization gradient

Score gradient	Reparameterization gradient
discrete + cont LV's	continuous LV's
larger variance	smaller variance
cheaper	more expensive

ADVI

Stan HMC

ELBO surgery

$$\mathcal{L} = \mathbb{E}\left[\log p(x \mid z,\theta)\right] - \underbrace{KL\left(q(z,\theta) \,\|\, p(z,\theta)\right)}$$

Sticking the Landing

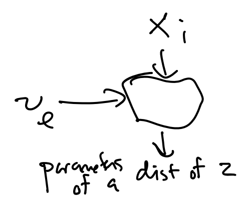
Amortized inference and the variational autoencoder

DGM: $\theta \sim p(\theta)$

for each data point i:

$z_i \sim \mathcal{N}_p(0,1)$

$x_i \sim \mathcal{N}\left(\Omega_\mu(z_i;\theta), \Omega_\sigma(z_i;\theta)^2\right)$



observe: $\{x_i\}_{i=1}^n$ goal: $p(\theta, z_{1:n} \mid x_{1:n})$

$q(z_i; v_e, x_i) = \Phi(x_i; v_e)$

inference network: $\Phi(x_i; v) \to q(z_i)$

$$q(\theta, z_{1:n}; v) = q(\theta; v_\theta)\prod_{i=1}^n q(z_i; v_e, x_i)$$ Amortized mean field

$$q(\theta; v_\theta)\prod_{i=1}^n q(z_i; v_i)$$ Classical mean field

$$q(z_i; v_e, x_i) = \mathcal{N}\left(\Phi_\mu(x_i; v_e), \Phi_\sigma(x_i; v_e)^2\right)$$

$$\mathcal{L}(v) = \mathbb{E}\left[\log p(\theta) - \log q(\theta; v_\theta) + \sum_{i=1}^n \left(\log p(z_i, x_i \mid \theta) - \log q(z_i; v_e, x_i)\right)\right]$$

Transformation of local latent variables.

$\varepsilon_i \sim \mathcal{N}_p(0,1)$

$t(\varepsilon_i, x_i, v_e) = \varepsilon_i \cdot \Phi_\sigma(x_i; v_e) + \Phi_\mu(x_i; v_e)$ $\to z_i \sim q(z_i; x_i; v_e)$

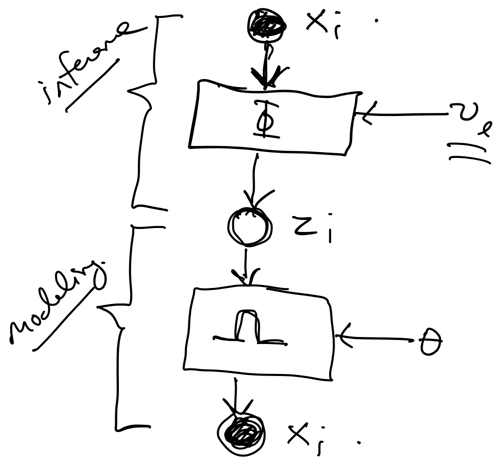
$$\nabla_{v_\theta}\mathcal{L} = \mathbb{E}\left[\nabla_\theta\left(\log p(\theta) + \sum_{i=1}^n \log p(z_i, x_i \mid \theta) - \log q(\theta; v_\theta)\right)\nabla_{v_\theta} t(\varepsilon_\theta, v_\theta)\right]$$

$$\nabla_{\nu_e} \mathcal{L} = \sum_{i=1}^{n} \mathbb{E}\left[\nabla_{z_i}\left(\log p(z_i, x_i \mid \theta) - \log q(z_i ; x_i, \nu_e) \right) \nabla_{\nu_e} t(\varepsilon_i, x_i, \nu_e) \right]$$

$\Omega(z_i ; \theta)$ = "decoder"

$\Phi(x_i ; \nu_e)$ = "encoder"

$\frac{1}{B}\sum$



Zemel.

Helmholtz Machine

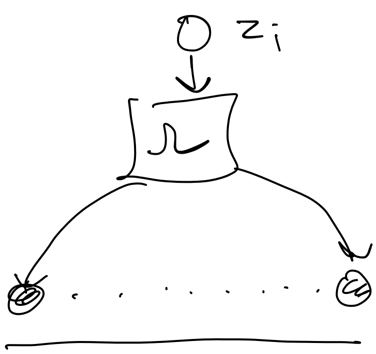
Autoencoder

ScVI

$z_i \sim \mathcal{N}_p(0, 1)$

$x_i \sim NB(\Omega(z_i ; \theta))$

$q(z_i ; x_i, \nu_e) = \mathcal{N}(\Phi_\mu(z_i ; \nu_e), \Phi_\sigma(z_i ; \nu_e)^2)$

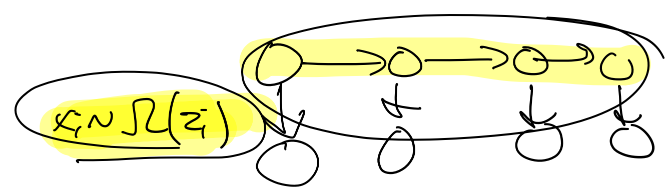


counts, sparse.

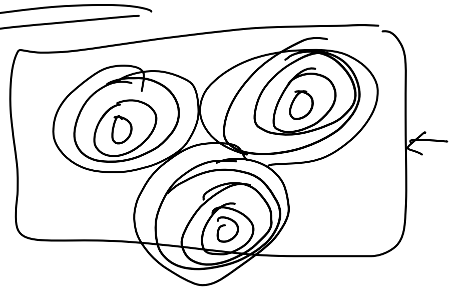
$$\nabla \mathcal{L}(\nu) = \mathbb{E}_q \left[\nabla_\nu \log q(\ ) \left(\log p(-) - \log q(\ ) \right) \right]$$

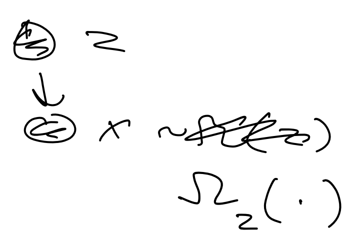
Galaxy properties (from sim) z_i's.

low-dim.



Structured VAEs — "old" (2016)





EDM.

Conditional models

linear + logistic regression

mixtures – cluster the data.

mixed-membership – grouped data

matrix factorization – interaction data.

exponential family + conjugacy

Generalized linear models

deep learning – regression

generalized neural models

deep generative models

= Deep exp. families.

exact inference.

MAP inference.

Gibbs sampler.

CAVI

generic inference w/ BBVI.