Machine Learning 2

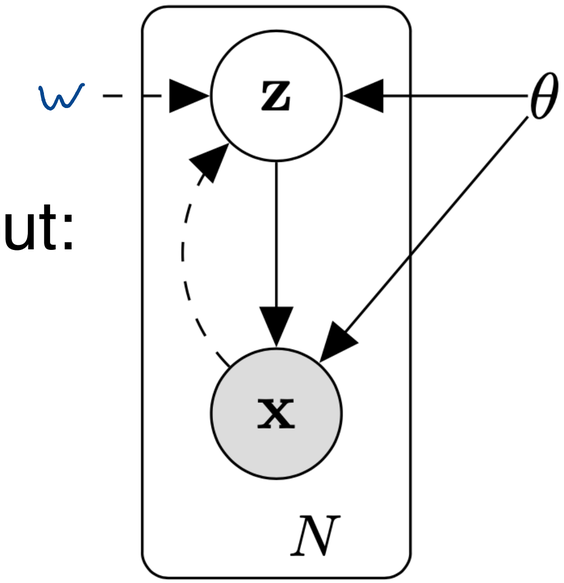
Variational Inference
- Variational AutoEncoders (VAE)

Patrick Forré

Variational AutoEncoders - Setup

- Let $q(x)$ be an underlying 'true' distribution in a high dimensional space $\mathcal{X}$ and $\hat{X}$ data sampled from it, e.g. images.
- We consider the 'support' of $q(x)$ to lie close to a lower dimensional 'manifold', which we consider given by a latent variable.
- So we assume a lower dimensional latent space $\mathcal{Z}$ that 'parameterizes' the 'data manifold'.
- Since deep neural networks are universal approximators we hope that they can learn the 'data manifold'.
- We also introduce an inference network (the 'encoder') such that we can use Variational Inference for training: maximizing ELBO.

Variational AutoEncoders



- To maximize ELBO: $L_{\hat{X}}(w,\theta) = \mathbb{E}_{\hat{q}(X)}\mathbb{E}_{q_w(Z|X)}\left[\log \frac{p_\theta(Z,X)}{q_w(Z|X)}\right]$ we put:
  - $p(z) = \mathcal{N}(z|0,I)$
  - $p_\theta(x|z) = \mathcal{N}(x|\mu_{\theta_1}(z), \sigma^2_{\theta_2}(z))$ generator / decoder
  - $q_w(z|x) = \mathcal{N}(z|\mu_{w_1}(x), \sigma^2_{w_2}(x))$ encoder
- where $\mu_{\theta_1}(z)$, $\sigma_{\theta_2}(z)$, $\mu_{w_1}(x)$, $\sigma_{w_2}(x)$ are parameterized by neural networks.
- The ELBO is then maximized via backpropagation on mini-batches.

Reparameterization Trick

- To evaluate the expectation $\mathbb{E}_{q_w(Z|X)}[f(Z)]$ in the ELBO one approximates it with an empirical version based on samples:
- For each $x^{(n)}$ sample M i.i.d. $\epsilon^{(n,1)}, \ldots, \epsilon^{(n,M)} \sim \mathcal{N}(0,I)$ and put $z^{(n,m)} := \mu_{w_1}(x^{(n)}) + \sigma_{w_2}(x^{(n)}) \cdot \epsilon^{(n,m)}$, which are then i.i.d.
  $\sim \mathcal{N}(z \,|\, \mu_{w_1}(x), \sigma^2_{w_2}(x))$,
  separating parameters from random noise for back-propagation. We get:
- $$\mathbb{E}_{q_w(Z|x^{(n)})}[f(Z)] \approx \frac{1}{M}\sum_{m=1}^{M} f\left(z^{(n,m)}\right) = \frac{1}{M}\sum_{m=1}^{M} f\left(\mu_{w_1}(x^{(n)}) + \sigma_{w_2}(x^{(n)}) \cdot \epsilon^{(n,m)}\right)$$

Training the VAE

- To maximize ELBO: $L_{\hat{X}}(w, \theta) = \mathbb{E}_{q(X)} \mathbb{E}_{q_w(Z|X)} \left[\log \frac{p_\theta(Z, X)}{q_w(Z \mid X)} \right]$ do until convergence:
  - Sample mini-batch from data: $x^{(1)}, \ldots, x^{(N)} \sim q(X)$
  - For each n sample mini-batch:
    $z^{(n,1)}, \ldots, z^{(n,M)} \sim \mathcal{N}(z \mid \mu_{w_1}(x^{(n)}), \sigma^2_{w_2}(x^{(n)}))$
    using the reparameterization trick.
  - Forward pass: $\hat{L}(\theta, w) = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \left[\log \frac{p_\theta(z^{(n,m)}, x^{(n)})}{q_w(z^{(n,m)} \mid x^{(n)})} \right]$
  - Update parameters via Back-propagation (α learning rate)
    - $[\theta, w]^\top := [\theta, w]^\top + \alpha \cdot \nabla_{\theta,w} \hat{L}(\theta, w)$

Interpreting the ELBO

- The ELBO can be written as:

$$L_{\hat{X}}(w, \theta) = \mathbb{E}_{q(x)} \mathbb{E}_{q_w(z|x)} [\log p_\theta(x|z)] - \mathbb{E}_{q(x)} [KL(q_w(z|x) \| p(z))]$$

$x \rightsquigarrow z \rightsquigarrow \hat{x}$

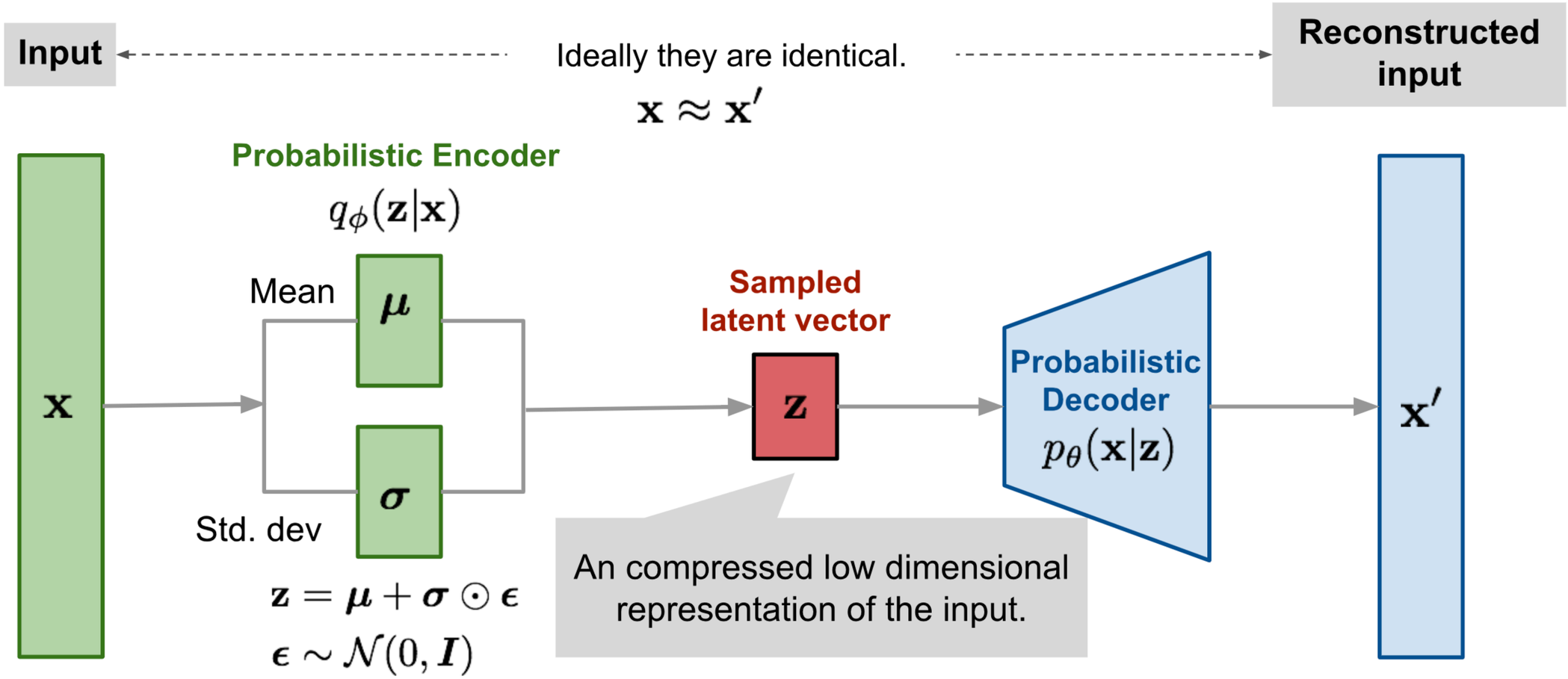
$\rightsquigarrow -\|x - \hat{x}\|_2^2$

reconstruction

regularization.

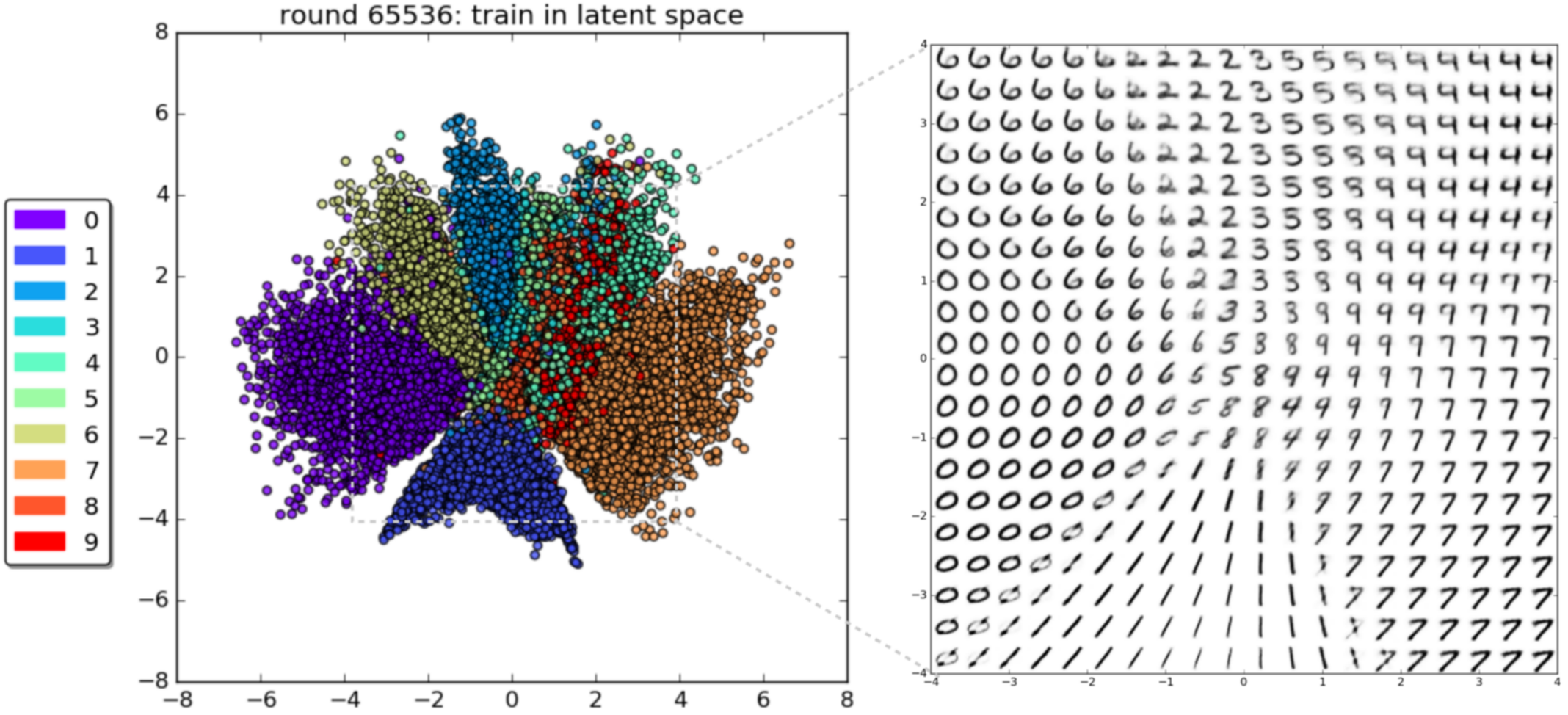
- splitting in reconstruction term, and
- regularization term

Training the VAE



- Source: https://lilianweng.github.io/lil-log/2018/08/12/from-autoencoder-to-beta-vae.html

VAE - Latent Space



- https://blog.fastforwardlabs.com/2016/08/12/introducing-variational-autoencoders-in-prose-and-code.html

Machine Learning 2

Variational Inference
- Variational Bayes /
/ Mean Field Approximation

Patrick Forré

Recap: The Framework of Variational Inference (VI)

- Let $q(x)$ be an underlying 'true' distribution of interest and $\hat{X}$ data sampled from it. Let $\{p_\theta(x,z) \mid \theta \in \Theta\}$ be a latent variable model for $q(x)$ (or a Bayesian setting).
- Instead of maximizing the evidence/log-likelihood $\log p_\theta(\hat{X})$ w.r.t. θ,
- in Variational Inference (VI) one introduces a variational family of inference distributions $\{q_w(z \mid x) \mid w \in \mathscr{W}\}$ and Maximizes the Evidence Lower BOund (ELBO):

$$L_{\hat{X}}(w,\theta) := \mathbb{E}_{q_w(Z|\hat{X})}\left[\log \frac{p_\theta(Z,\hat{X})}{q_w(Z \mid \hat{X})}\right]$$

w.r.t. both distributions / parameters w, θ.

Variational Bayes - Mean Field Approximation

- Goal: Maximize ELBO wrt w, θ: $L_{\hat{X}}(w, \theta) = \mathbb{E}_{q_w(Z|\hat{X})}\left[\log \frac{p_\theta(Z, \hat{X})}{q_w(Z \mid \hat{X})}\right]$

- Latent variables $Z = (Z_1, \ldots, Z_K)$ with several components or component blocks.
- Variational Bayes / Mean Field Approximation: Family of inference distribution to be taken as product distributions:

$$q_w(z_1, \ldots, z_K \mid \hat{X}) = \prod_{k=1}^{K} q_{w_k}(z_k \mid \hat{X}) \quad \approx \quad p_\theta(z \mid \hat{X})$$

- This is said to be an "approximation", as $q_w(z \mid \hat{X})$ is considered an approximation for the 'true' posterior: $p_\theta(z \mid \hat{X})$ in the Bayesian setting and $L_{\hat{X}}(w, \theta)$ after fitting for evidence: $\log p_\theta(\hat{X})$.
- Note that this is a 'high' entropy approximation as: $H(Z_1, \ldots, Z_K) \leq H(Z_1) + \cdots + H(Z_K)$.

Optimizing w.r.t. q

dependent on w and $\hat{x}$

- Abbrev. inference distributions with: $q(z) = q_1(z_1) \cdots q_K(z_K)$
- Langrangian (ELBO plus constraints, θ fixed):

$$\mathcal{L}(q) = \mathbb{E}_{q(z)}\left[\log \frac{p(z,\hat{x})}{q(z)}\right]$$

$$= \mathbb{E}_{q(z)}[\log p(z,\hat{x})] - \mathbb{E}_{q(z)}[\log q(z)]$$

$$= \int q_1(z_1) \cdots q_K(z_K) \cdot \log p(z,\hat{x})\, dz - \sum_{k=1}^{K} \int q_k(z_k) \cdot \log q_k(z_k)\, dz_k$$

$$+ \sum_{k=1}^{K} \lambda_k \cdot \left(\int q_k(z_k)\, dz_k - 1 \right)$$

$$+ \sum_{k=1}^{K} \lambda_k \cdot \left(\int q_k(z_k)\, dz_k - 1 \right)$$

$$+ \sum_{k=1}^{K} \lambda_k \left(\int q_k(z_k) - 1 \right)$$

Functional derivatives w.r.t. q

$$0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial q_k(z_k)} = \int \prod_{j \neq k} q_j(z_j) \cdot \log p(z, \hat{x}) \, dz_{\neg k} - \log q_k(z_k) - 1 + \lambda_k$$

$$\mathcal{L}(q) = \int q_1(z_1) \cdots q_K(z_K) \cdot \log p(z, \hat{x}) \, dz - \sum_{k=1}^{K} \int q_k(z_k) \cdot \log q_k(z_k) \, dz_k + \sum_{k=1}^{K} \lambda_k \left(\int q_k(z_k) - 1 \right)$$

- $$q_k(z_k) = \exp(\lambda_k - 1) \cdot \exp\left(\int \prod_{j \neq k} q_j(z_j) \log p(z, \hat{x}) \, dz_{\neg k} \right)$$

$$\propto \exp\left(\mathbb{E}_{q_{\neg k}(z_{\neg k})} \left[\log p(z_{\neg k}, z_k, \hat{x}) \right] \right)$$

VB-MF Fixpoint Equations

- We arrive at the following fixpoint equations for $k = 1, \ldots, K$:

$$q_k(z_k) \propto \exp\left(\mathbb{E}_{q(Z_{\neg k})}\left[\log p(\hat{X}, z_k, Z_{\neg k})\right]\right)$$

- Note that the rhs is dependent on the all other $q_j(z_j)$ for $j \neq k$.
- We are left to solve this system of fixpoint equation:
  - Often from this an iterative update rule for corresponding parameters can be derived that converges.

Machine Learning 2

Variational Inference
- Variational Bayes /
/ Mean Field Approximation
- Example

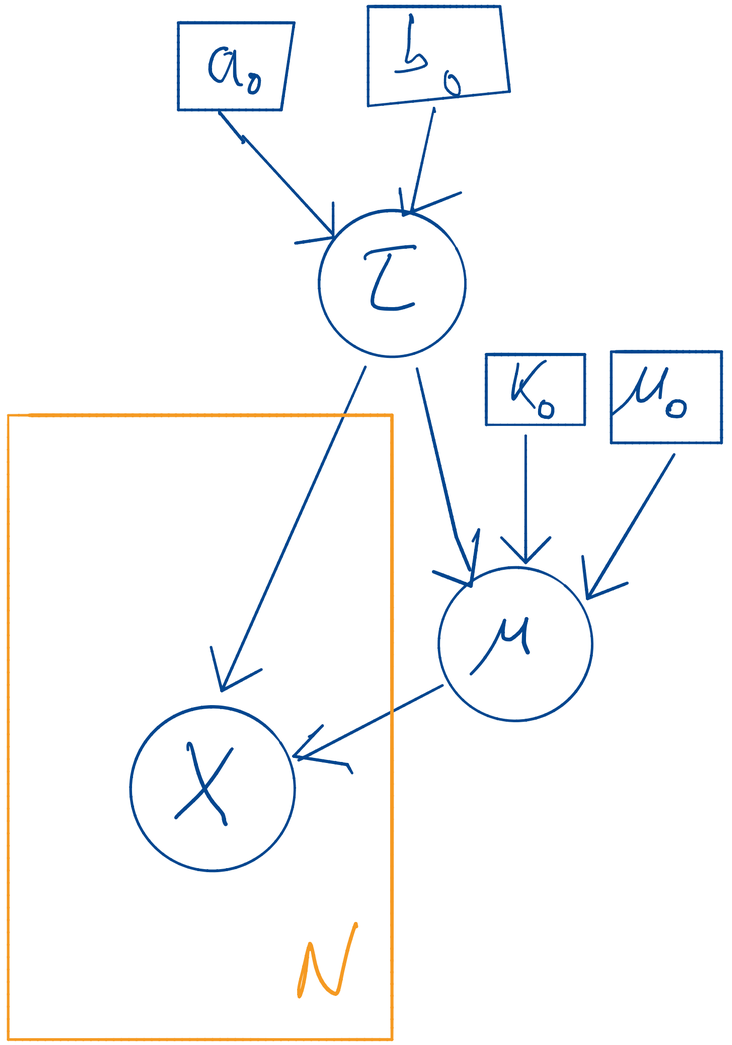
Patrick Forré

$x_1, \dots, x_N \overset{\text{iid}}{\sim}$ $\hat{x}$

$$p(x|\mu,\tau) = \mathcal{N}(x|\mu,\tau^{-1})$$

$$p(\mu|\mu_0,\kappa_0,\tau) = \mathcal{N}(\mu|\mu_0,(\kappa_0\cdot\tau)^{-1})$$

$$p(\tau|a_0,b_0) = \Gamma(\tau|a_0,b_0)$$



$$\log p(\hat{x},\mu,\tau) = \log p(\hat{x}|\mu,\tau) + \log p(\mu|\tau) + \log p(\tau)$$

$$= \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2$$

$$+ \frac{1}{2}\log(\kappa_0\tau) - \frac{\kappa_0\tau}{2}(\mu-\mu_0)^2$$

$$+ (a_0-1)\cdot\log\tau - b_0\tau + \text{const.}$$

Goal: $p(\mu,\tau|\hat{x}) \approx q_1(\mu)\cdot q_2(\tau)$

VB-MF Fixpoint equations:

$$\log q_1(\mu) = \mathbb{E}_{q_2(\tau)}\left[\log p(\hat{x}, \mu, \tau)\right] + \text{const.}$$

$$\log q_2(\tau) = \mathbb{E}_{q_1(\mu)}\left[\log p(\hat{x}, \mu, \tau)\right] + \text{const.}$$

$$\log q_1(\mu) = \mathbb{E}_{q_2(\tau)}\left[\log p(\hat{x}, \mu, \tau)\right] + \text{const.}$$

$$\log p(\hat{x}, \mu, \tau) = \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{1}{2}\log(\kappa_0\tau) - \frac{\kappa_0\tau}{2}(\mu-\mu_0)^2$$

$$+ (a_0-1)\cdot\log\tau - b_0\tau \qquad + \text{const.}$$

$$\log q_1(\mu) = \mathbb{E}_{q_2(\tau)}\left[-\frac{\tau}{2}\cdot\left(\sum_{n=1}^{N}(x_n-\mu)^2 + \kappa_0(\mu-\mu_0)^2\right)\right] + \text{const}$$

$$= -\frac{\mathbb{E}[\tau]}{2}\cdot\left(\kappa_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right) + \text{const.}$$

$$= -\frac{\tau_N}{2}(\mu-\mu_N)^2 + \text{const.}$$

$$\mu_N = \frac{\kappa_0\cdot\mu_0 + \sum_{n=1}^{N} x_n}{\kappa_0+N} \quad \text{known}$$

$$\tau_N = (\kappa_0+N)\cdot\mathbb{E}_{q(\tau)}[\tau] \quad \text{unknown}$$

$$\log q_1(\mu) = -\frac{\tau_N}{2}(\mu - \mu_N)^2 + \text{const.}$$

$$\Rightarrow q_1(\mu) = \mathcal{N}(\mu \mid \mu_N, \tau_N^{-1})$$

where:

$$\mu_N = \frac{\kappa_0 \mu_0 + \sum_{n=1}^{N} x_n}{\kappa_0 + N}$$

$$\tau_N = (\kappa_0 + N) \cdot \mathbb{E}_{q_2(\tau)}[\tau]$$

$$\log q_2(\tau) = \mathbb{E}_{q_1(\mu)}\left[\log p(\hat{x}, \mu, \tau)\right] + \text{const.}$$

$$\log p(\hat{x}, \mu, \tau) = \frac{N}{2}\log\tau - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2 + \frac{1}{2}\log(\kappa_0\tau) - \frac{\kappa_0\tau}{2}(\mu-\mu_0)^2$$

$$+ (a_0-1)\cdot\log\tau - b_0\tau + \text{const.}$$

$$\log q_2(\tau) = \frac{N}{2}\log\tau + \frac{1}{2}\log\tau + (a_0-1)\log\tau - b_0\cdot\tau$$

$$- \frac{\tau}{2}\mathbb{E}_{q_1(\mu)}\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \kappa_0(\mu-\mu_0)^2\right] + \text{const.}$$

$$= \left(a_0 + \frac{N+1}{2} - 1\right)\log\tau - \left[b_0 + \frac{1}{2}\mathbb{E}_{q_1(\mu)}\left[\sum_{n}(x_n-\mu)^2 + \kappa_0(\mu-\mu_0)^2\right]\right]\cdot\tau + \text{const.}$$

$$= (a_N-1)\cdot\log\tau - b_N\cdot\tau + \text{const.}$$

$$\log q_2(\tau) = (a_N - 1)\log\tau - b_N \cdot \tau + \text{const.}$$

$$\Rightarrow \quad q_2(\tau) = \Gamma(\tau \mid a_N, b_N) \qquad \rightarrow \mathbb{E}_\tau[\tau] = \frac{a_N}{b_N}$$

where

$$a_N = a_0 + \frac{N+1}{2} \qquad \leftarrow \text{known}$$

$$b_N = b_0 + \frac{1}{2}\mathbb{E}_{q_1(\mu)}\left[K_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right] \qquad \leftarrow \text{unknown}$$

$$= b_0 + \frac{K_0}{2}\left(\mathbb{E}_\mu[\mu^2] + \mu_0^2 - 2\,\mathbb{E}_\mu[\mu]\cdot\mu_0\right)$$

$$+ \frac{1}{2}\sum_{n=1}^{N}\left(x_i^2 + \mathbb{E}_\mu[\mu^2] - 2\,\mathbb{E}_\mu[\mu]\cdot x_i\right)$$

$$q_1(\mu) = \mathcal{N}(\mu \mid \mu_N, \tau_N^{-1}), \quad \mathbb{E}_\mu[\mu] = \mu_N, \quad \mathbb{E}_\mu[\mu^2] = \frac{1}{\tau_N} + \mu_N^2$$

Known

μ_N

a_N

to be determined

$$\tau_N = (\kappa_0 + N) \cdot \frac{a_N}{b_N} \quad \Rightarrow \quad \tau_N^{-1} = \frac{b_N}{(\kappa_0 + N) \cdot a_N}$$

$$b_N = b_0 + \frac{\kappa_0}{2}\left(\frac{1}{\tau_N} + \mu_N^2 + \mu_0^2 - 2 \cdot \mu_N \cdot \mu_0\right) + \frac{1}{2}\sum_{n=1}^{N}\left(x_n^2 + \frac{1}{\tau_N} + \mu_N^2 - 2\,\mu_N \cdot x_n\right)$$

$$= \tau_N^{-1}\left(\frac{\kappa_0 + N}{2}\right) + \text{rest}$$

$$\Rightarrow \quad b_N = \frac{1}{2}\frac{b_N}{a_N} + b_0 + \frac{\kappa_0}{2}(\mu_N - \mu_0)^2 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu_N)^2$$

$$\Leftrightarrow \quad b_N = \left(1 - \frac{1}{2a_N}\right)^{-1}\left[b_0 + \frac{\kappa_0}{2}(\mu_N - \mu_0)^2 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu_N)^2\right]$$

$$\Rightarrow \quad \tau_N = (\kappa_0 + N)\frac{a_N}{b_N}$$

$$q_1(\mu) = \mathcal{N}(\mu \mid \mu_N, \tau_N^{-1}) \qquad \text{and} \qquad q_2(\tau) = \Gamma(\tau \mid a_N, b_N)$$

where

$$a_N = a_0 + \frac{N+1}{2}$$

$$\mu_N = \frac{\kappa_0 \mu_0 + \sum_{n=1}^{N} x_n}{\kappa_0 + N}$$

$$b_N = \left(1 - \frac{1}{2a_N}\right)^{-1} \left[b_0 + \frac{\kappa_0}{2} (\mu_N - \mu_0)^2 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu_N)^2 \right]$$

$$\tau_N = (\kappa_0 + N) \frac{a_N}{b_N}$$

$$p(\mu, \tau \mid \hat{X}) \approx N(\mu \mid \mu_N, \tau_N^{-1}) \cdot \Gamma(\tau \mid a_N, b_N)$$

$$\log p(\hat{X}) \gtrapprox \mathrm{ELBO}\left(N(\mu \mid \mu_N, \tau_N^{-1}) \cdot \Gamma(\tau \mid a_N, b_N) \right)$$

evidence

VB-MF approximation.

Machine Learning 2

Variational Inference
- Universal Variational Inference
as its own learning framework

Patrick Forré

Universal Variational Inference (UVI) - Setting

- Let $q(x)$ be an underlying 'true' distribution of interest and $\hat{X} = [x^{(1)}, \ldots, x^{(N)}]^\top$ data sampled from it.
- We aim at:
  - 'learning the data distribution',
  - prediction,
  - quantifying model uncertainty,
  - incorporating prior knowledge
  - regularization
  - computational efficiency, etc.

Universal Variational Inference - Ingredients

- In the general setting of UVI we need to specify:

1. Statistical model $\mathscr{P} = \{p_\theta(x) = p(x \,|\, \theta) \,|\, \theta \in \Theta\}$ parameterized by θ.
2. Loss function: $\mathscr{L}(\hat{X} \,|\, p_\theta)$ measuring the goodness of fit between (discrete) data points $\hat{X}$ and (implicit/explicit) model p_θ.
3. Inference model: $\mathscr{Q} = \{q_w(\theta) = q(\theta \,|\, w) \,|\, w \in \mathscr{W}\}$, which is usually given by comp. restrictions, hard constraints or simplifying assumptions.
4. Regularizing functional: $\Omega(q)$ for $q \in \mathscr{Q}$ addressing uncertainty, prior knowledge, outcome expectations, soft constraints, complexity, etc.
5. Optimizer $\mathscr{O}$ and evalution scheme $\mathscr{E}$ for the following:

UVI - Learning Objective + Predictive Distribution

- After p_θ, $\mathscr{L}$, $\mathbb{Q}$, Ω are specified we use an optimizer $\mathcal{O}$ to solve:

$$q_{\hat{X}} \in \arg\min_{q \in \mathbb{Q}} \left(\mathbb{E}_{q(\theta)} \left[\mathscr{L}(\hat{X} \,|\, p_\theta) \right] + \Omega(q) \right)$$

- Finally, for predictions do model averaging (via $\mathscr{E}$):

$$p(x \,|\, \hat{X}) := \int p(x \,|\, \theta) \cdot q_{\hat{X}}(\theta)\, d\theta$$

goal: $\approx q(x|$

Remark: Loss functions

- The loss function often comes in form: $\mathscr{L}(\hat{X}\,|\,p_\theta) = \sum_{n=1}^{N} \ell(x^{(n)}\,|\,p_\theta)$ with:
  - $\ell(\hat{x}\,|\,p_\theta) = -\log p_\theta(\hat{x})$ (log-loss)
  - $\ell(\hat{x}\,|\,p_\theta) = \frac{1}{2}\left(\hat{x} - \mathbb{E}_{p_\theta(X)}[X]\right)^2$ (square-loss)
  - $\ell(\hat{x}\,|\,p_\theta) = \alpha \cdot [\hat{x} - q_\alpha^{p_\theta}]_+ + (1-\alpha)\cdot[q_\alpha^{p_\theta} - \hat{x}]_+$ (α-quantile loss)
  - $\ell(\hat{x}\,|\,p_\theta) = 1 - p_\theta(\hat{x})$ (misclassification rate)
  - $\ell(\hat{x}\,|\,p_\theta) = -\left(1 - p_\theta(\hat{x})\right)^\gamma \cdot \log p_\theta(\hat{x})$ (focal loss)
  - Maximum-Mean-Discrepancies, Optimal Transport divergencies, Stein divergencies, etc.

Remark: Inference models

- Possible Inference models: $\mathcal{Q} = \{q_w(\theta) = q(\theta \mid w) \mid w \in \mathcal{W}\}$ are
  - Point estimators: $q_w(\theta) = \delta_w(\theta)$ for $w \in \Theta = \mathcal{W}$
  - Neural networks
  - Product distributions: $q_w(\theta) = q_{w_1}(\theta_1) \cdots q_{w_K}(\theta_K)$
  - All distributions

Remark: Regularizing functional

- Possible Regularizing functionals:
  - $\Omega(q) = 0$
  - $\Omega(q) = C(q\|\pi)$ with a prior $\pi(\theta)$
  - $\Omega(q) = \frac{1}{\beta}\mathrm{KL}(q\|\pi)$ with a prior $\pi(\theta)$ and ‘inverse temperature’ β
  - $\Omega(q) = \pm H(q)$
  - $\Omega(q) = \mathbb{E}_{q(\theta)}\left[d(p_\theta\|p_0)\right]$ with reference model $p_0(x)$ and divergence d.

Remark: UVI for the supervised setting

- Assumption of true distribution $q(y, x)$ or at least $q(y \mid x)$
- Statistical model $p_\theta(y, x) = p_\theta(y \mid x) \cdot q(x)$,

  where $q(x)$ is either assumed to be known, given or ignored or coincides with empirical version $\hat{q}(x)$.
- Note we are here interested in $q(y \mid x)$ and not so much in $q(x)$.
- Then everything goes similar in the UVI famework.

Machine Learning 2

Variational Inference
- Universal Variational Inference
- recovering special cases

Patrick Forré

Standard Variational Inference via UVI

1. $\mathscr{P} = \{p_\theta(x) \mid \theta \in \Theta\}$
2. $\mathscr{L}(\hat{X} \mid p_\theta) =$ $-\log p(\hat{x} \mid \theta)$
3. $\mathcal{Q} =$ {any/all}
4. $\Omega(q) =$ $KL(q \| \pi)$, $\pi(\theta)$ prior
5. $\arg\min_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} \left[\mathscr{L}(\hat{X} \mid p_\theta) \right] + \Omega(q) \right) = \mathbb{E}_{q(\theta)} \left[-\log p(\hat{x} \mid \theta) \right] + \mathbb{E}_{q(\theta)} \left[\log \frac{q(\theta)}{\pi(\theta)} \right]$
6. $p(x \mid \hat{X}) = \int p(x \mid \theta) \cdot q_{\hat{X}}(\theta)\, d\theta$

$$
\begin{aligned}
&= -ELBO(q) \\
&= -\mathbb{E}_{q(\theta)} \left[\log \frac{p(\hat{x}, \theta)}{q(\theta)} \right] \\
&= \mathbb{E}_{q(\theta)} \log \frac{q(\theta)}{p(\hat{x} \mid \theta)\, \pi(\theta)}
\end{aligned}
$$

Full Bayesian approach via UVI

1. $\mathscr{P} = \{p_\theta(x) \mid \theta \in \Theta\}$

2. $\mathscr{L}(\hat{X} \mid p_\theta) =$ $-\log p(\hat{x} \mid \theta)$

3. $\mathcal{Q} =$ all ?

4. $\Omega(q) =$ $KL(q \| \pi)$, prior $\pi(\theta)$

5. $\arg\min_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} \left[\mathscr{L}(\hat{X} \mid p_\theta) \right] + \Omega(q) \right) =$ $= p(\theta \mid \hat{x})$

$= \arg\min_{q \in \text{all}} KL(q(\theta) \| p(\theta \mid \hat{x}))$

$= \arg\min_{q \in \text{all}} \mathbb{E}_{q(\theta)} \left[\log \frac{q(\theta)}{p(\theta \mid \hat{x})} \right] - \log p(\hat{x})$

$= \arg\min_{q \in \text{all}} \mathbb{E}_{q(\theta)} \log \frac{q(\theta)}{p(\theta, \hat{x})}$

6. $p(x \mid \hat{X}) = \int p(x \mid \theta) \cdot q_{\hat{X}}(\theta)\, d\theta =$ $\int p(x \mid \theta)\, p(\theta \mid \hat{x})\, d\theta$

Maximum Likelihood Estimation via UVI

1. $\mathscr{P} = \{p_\theta(x) \mid \theta \in \Theta\}$
2. $\mathscr{L}(\hat{X} \mid p_\theta) =$ $-\log p_\theta(\hat{x})$
3. $\mathcal{Q} =$ $\{\delta_w(\theta) \mid w \in \Theta\}$
4. $\Omega(q) =$ 0

$= \delta_{\theta_{MLE}}(\theta)$

$= \arg\max_{\delta_w} \log p_w(\hat{x})$

5. $\arg\min_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} \left[\mathscr{L}(\hat{X} \mid p_\theta) \right] + \Omega(q) \right) = \arg\min_{\delta_w} -\log p_w(\hat{x})$

6. $p(x \mid \hat{X}) = \int p(x \mid \theta) \cdot q_{\hat{X}}(\theta)\, d\theta =$ $p(x \mid \theta_{MLE})$

Maximum A Posteriori Estimation via UVI

1. $\mathscr{P} = \{p_\theta(x) \mid \theta \in \Theta\}$
2. $\mathscr{L}(\hat{X} \mid p_\theta) =$ $-\log p_\theta(\hat{x})$
3. $\mathcal{Q} =$ $\{\delta_w(\theta) \mid w \in \Theta\}$
4. $\Omega(q) =$ $C(q \| \pi)$, prior $\pi(\theta)$

$= \delta_{\theta_{MAP}}(\theta)$

$= \underset{\delta_w}{\operatorname{argmax}} \log p(w|\hat{x})$

$= \underset{\delta_w}{\operatorname{argmax}} \log p(w, \hat{x})$

5. $\underset{q \in \mathcal{Q}}{\arg\min} \left(\mathbb{E}_{q(\theta)} \left[\mathscr{L}(\hat{X} \mid p_\theta) \right] + \Omega(q) \right) =$ $\underset{\delta_w}{\operatorname{argmin}} -\log p_w(\hat{x}) - \log \pi(w)$

6. $p(x \mid \hat{X}) = \int p(x \mid \theta) \cdot q_{\hat{X}}(\theta)\, d\theta =$ $p(x \mid \theta_{MAP})$

Empirical Risk Minimization via UVI

1. $\mathscr{P} = \{p_\theta(x) \mid \theta \in \Theta\}$
2. $\mathscr{L}(\hat{X} \mid p_\theta) =$ any
3. $\mathcal{Q} =$ $\{\delta_w(\theta) \mid w \in \Theta\}$
4. $\Omega(q) =$ 0 $= \delta_{\theta_{ERM}}(\theta)$
5. $\arg\min_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} \left[\mathscr{L}(\hat{X} \mid p_\theta) \right] + \Omega(q) \right) =$ $\underset{\delta_w}{\operatorname{argmin}} \; \mathcal{L}(\hat{X} \mid p_w)$
6. $p(x \mid \hat{X}) = \int p(x \mid \theta) \cdot q_{\hat{X}}(\theta) \, d\theta =$ $p(x \mid \theta_{ERM})$

Variational Bayes - Mean Field Approximation via UVI

1. $\mathscr{P} = \{p_\theta(x) \mid \theta \in \Theta\}$
2. $\mathscr{L}(\hat{X} \mid p_\theta) =$ $-\log p_\theta(\hat{x})$
3. $\mathcal{Q} =$ $\{ q_\nu(\theta) = q_{\nu_1}(\theta_1) \cdots q_{\nu_n}(\theta_n) \}$
4. $\Omega(q) =$ $KL(q \| \pi)$, $\pi(\theta)$ prior.
5. $\arg\min_{q \in \mathcal{Q}} \left(\mathbb{E}_{q(\theta)} \left[\mathscr{L}(\hat{X} \mid p_\theta) \right] + \Omega(q) \right) =$ $\underset{q \in \mathcal{Q}}{\arg\max}\ ELBO(q)$
6. $p(x \mid \hat{X}) = \int p(x \mid \theta) \cdot q_{\hat{X}}(\theta)\, d\theta$

Machine Learning 2

Variational Inference
- Summary

Patrick Forré

Variational Inference

- True distribution $q(x)$, data $\hat{X}$, latent variable model or Bayesian setting:
$$\{p_\theta(z, x) \mid \theta \in \Theta\}.$$
- Introduce inference model $\{q_w(z \mid x) \mid w \in \mathscr{W}\}$ meant to approximate $p_\theta(z \mid \hat{X})$
- Key objective in VI is Maximizing the Evidence Lower Bound (ELBO):

$$L_{\hat{X}}(w, \theta) := \mathbb{E}_{q_w(Z|\hat{X})}\left[\log \frac{p_\theta(Z, \hat{X})}{q_w(Z \mid \hat{X})}\right] \leq \log p_\theta(\hat{X})$$

w.r.t. parameters w and θ (if present).

"=" iff $q_w(z|\hat{x}) = p_\theta(z|\hat{x})$

EM-Algorithm and VAEs

- Expectation-Maximization (EM) Algorithm: Alternate:
  - E-step: Evaluate: $Q(\theta, \theta^{(t)}) := \mathbb{E}_{p_{\theta^{(t)}}(Z|\hat{X})}\left[\log p_\theta(Z, \hat{X})\right]$
  - M-step: Compute: $\theta^{(t+1)} := \arg\max_\theta Q(\theta, \theta^{(t)})$ + constraints.
- Variational AutoEncoders (VAE):
  - Parameterize everything with deep neural networks.
  - Maximize ELBO with SGD and reparameterization trick.

Variational Bayes Mean Field Approximation

- Variational Bayes Mean Field Approximation:

$$p(z_1, \ldots, z_K \,|\, \hat{X}) \approx \prod_{k=1}^{K} q_{w_k}(z_k \,|\, \hat{X}) =: q_w(z \,|\, \hat{X})$$

  - lead to VB-MF Fixpoint Equations: for all $k = 1, \ldots, K$:

$$q_k(z_k) \propto \exp\left(\mathbb{E}_{q(Z_{\neg k})} \left[\log p(\hat{X}, z_k, Z_{\neg k}) \right] \right)$$

  - solve system for (the parameters of) the $q_k(z_k)$'s explicitly or via iterative updating.

Universal Variational Inference Learning Framework

- In the general setting of UVI we need to specify:
  1. Statistical model: $\mathscr{P} = \{p_\theta(x) = p(x \mid \theta) \mid \theta \in \Theta\}$
  2. Loss function: $\mathscr{L}(\hat{X} \mid p_\theta)$
  3. Inference model: $\mathscr{Q} = \{q_w(\theta) = q(\theta \mid w) \mid w \in \mathscr{W}\}$
  4. Regularizing functional: $\Omega(q)$ for $q \in \mathscr{Q}$.
  5. Solve optimization problem: $$q_{\hat{X}} \in \arg\min_{q \in \mathscr{Q}} \left(\mathbb{E}_{q(\theta)} \left[\mathscr{L}(\hat{X} \mid p_\theta) \right] + \Omega(q) \right)$$
  6. Predictive distribution via model averaging: $$p(x \mid \hat{X}) := \int p(x \mid \theta) \cdot q_{\hat{X}}(\theta)\, d\theta$$