Variational Inference

Advanced Statistical Inference

Simone Rossi

Refresher: Kullback-Leibler Divergence

- The Kullback-Leibler (KL) divergence is a measure of how one probability distribution diverges from a second
- Given two probability distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, the KL divergence is defined as

$$\mathrm{KL}\left(q\|p\right) = \int q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \,\mathrm{d}\boldsymbol{x} = \mathbb{E}_q \left[\log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right]$$

Properties

- $\mathrm{KL}\left(q\|p\right) \geq 0$ with equality if and only if $q(\boldsymbol{x}) = p(\boldsymbol{x})$
- $\mathrm{KL}\left(q\|p\right) \neq \mathrm{KL}\left(p\|q\right)$, i.e., it is not symmetric
- It's not a true distance measure as it's not symmetric and doesn't satisfy the triangle inequality

KL divergence for Gaussians

- The KL divergence between two Gaussians is tractable and has a closed-form solution
- For two Gaussian distributions $p(x) = \mathcal{N}(\mu_p, \sigma_p^2)$ and $q(x) = \mathcal{N}(\mu_q, \sigma_q^2)$:

$$\mathrm{KL}\left(q\|p\right) = \frac{1}{2}\left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{(\mu_q - \mu_p)^2}{\sigma_q^2} - 1 + \log \frac{\sigma_p^2}{\sigma_q^2}\right)$$

- For two multivariate Gaussians $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and $q(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$:

$$\mathrm{KL}\left(q \| p\right) = \frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}_q^{-1}\boldsymbol{\Sigma}_p\right) + (\boldsymbol{\mu}_q - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_q^{-1}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_p) - k + \log \frac{\det\{\boldsymbol{\Sigma}\}_q}{\det\{\boldsymbol{\Sigma}\}_p}\right)$$

> **Exercise**
>
> Simplify the expression when $\boldsymbol{\Sigma}_p = \sigma_p^2 \boldsymbol{I}$ and $\boldsymbol{\mu}_p = \mathbf{0}$.

Jensen's Inequality

- Another important result is Jensen's inequality, which states that for any convex function f and random variable X:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

For example, if $f(x) = \log x$, then:

$$\mathbb{E}[\log X] \leq \log(\mathbb{E}[X])$$

Variational Inference

Introduction to Variational Inference

- Given a likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ and a prior $p(\boldsymbol{\theta})$, we want to compute the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, which is intractable in most cases.
- **Variational Inference (VI)** is a method for approximating intractable posterior distributions

Intuition: Instead of trying to solve intractable integrals, we solve an optimization problem

> **Sketch of the recipe**
>
> 1. Choose a family of distributions $\mathcal{Q}$ to approximate the posterior
> 2. Define an objective function to measure the quality of the approximation
> 3. In the set of distributions $\mathcal{Q}$, find the one that minimizes the objective function

Variational Inference: Form of the Approximation

Form of the Approximation

What family of distributions $q(\boldsymbol{\theta})$ should we choose to approximate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$?

- **Mean-field approach**: each parameter θ_j is independent and has its own distribution

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q_j(\theta_j)$$

For simplicity:

- all $q_j(\theta_j)$ are Gaussian distributions, i.e., $q_j(\theta_j) = q(\theta_j) = \mathcal{N}(m_j, s_j^2)$
- each parameter θ_j has its own mean m_j and variance s_j^2

$$q(\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(\mu_j, \sigma_j^2)$$

- $\boldsymbol{\nu} = \{m_j, s_j^2\}$ are called the **variational parameters**
- The goal is to find the optimal values of $\boldsymbol{\nu}$ $\Rightarrow$ best approximation $q(\boldsymbol{\theta}; \boldsymbol{\nu})$ to the true posterior

Defining the Objective Function

Objective Function

- How to define the quality of the approximation $q(\boldsymbol{\theta}; \boldsymbol{\nu})$?

. . .

- We use the KL divergence between the approximate distribution $q(\boldsymbol{\theta}; \boldsymbol{\nu})$ and the true posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$

$$\begin{aligned} \mathrm{KL}\left(q(\boldsymbol{\theta}; \boldsymbol{\nu}) \| p(\boldsymbol{\theta} \mid \boldsymbol{y})\right) &= \int q(\boldsymbol{\theta}; \boldsymbol{\nu}) \log \frac{q(\boldsymbol{\theta}; \boldsymbol{\nu})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \, \mathrm{d}\boldsymbol{\theta} \\ &= \mathbb{E}_{q(\boldsymbol{\theta}; \boldsymbol{\nu})} \left[\log \frac{q(\boldsymbol{\theta}; \boldsymbol{\nu})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\right] \end{aligned}$$

⚠ Problem

- This expression is still intractable because the posterior $p(\theta \mid y)$ is unknown
- We need to find a way to approximate the KL divergence

Manipluating the expression:

$$\begin{aligned}
\mathrm{KL}\left(q(\theta;\nu)\|p(\theta \mid y)\right) &= \mathbb{E}_{q(\theta;\nu)} \log q(\theta;\nu) - \mathbb{E}_{q(\theta;\nu)} \log p(\theta \mid y) \\
&= \mathbb{E}_{q(\theta;\nu)} \log q(\theta;\nu) - \mathbb{E}_{q(\theta;\nu)} \log \frac{p(y \mid \theta)p(\theta)}{p(y)} \\
&= \underbrace{\mathbb{E}_{q(\theta;\nu)} \log q(\theta;\nu)}_{①} - \underbrace{\mathbb{E}_{q(\theta;\nu)} \log p(y \mid \theta)}_{②} - \underbrace{\mathbb{E}_{q(\theta;\nu)} \log p(\theta)}_{③} + \underbrace{\log p(y)}_{④}
\end{aligned}$$

Breakdown:

①: entropy of the variational distribution $q(\theta;\nu)$

②: expected log-likelihood of the data under the variational distribution

③: cross-entropy between the variational distribution and the prior

④: log marginal likelihood of the data

Rearranging the terms:

$$\begin{aligned}
\mathrm{KL}\left(q(\theta;\nu)\|p(\theta \mid y)\right) &= \mathbb{E}_{q(\theta;\nu)} \log q(\theta;\nu) - \mathbb{E}_{q(\theta;\nu)} \log p(y \mid \theta) - \mathbb{E}_{q(\theta;\nu)} \log p(\theta) + \log p(y) \\
&= -\mathbb{E}_{q(\theta;\nu)} \log p(y \mid \theta) + \mathrm{KL}\left(q(\theta;\nu)\|p(\theta)\right) + \log p(y)
\end{aligned}$$

This is an important equation in variational inference!

. . .

Note: The term $\log p(y)$ is a constant w.r.t. ν. Let's move it to the left:

$$\log p(y) - \mathrm{KL}\left(q(\theta;\nu)\|p(\theta \mid y)\right) = \mathbb{E}_{q(\theta;\nu)} \log p(y \mid \theta) - \mathrm{KL}\left(q(\theta;\nu)\|p(\theta)\right)$$

Now the right-hand side is computable: it's called **Evidence Lower Bound (ELBO)**

$$\mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}) = \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \text{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu}) \| p(\boldsymbol{\theta})\right)$$

ELBO: Evidence Lower Bound

$$\log p(\boldsymbol{y}) - \text{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu}) \| p(\boldsymbol{\theta} \mid \boldsymbol{y})\right) = \mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu})$$

- Minimizing the KL divergence is equivalent to maximizing the ELBO
- The KL divergence is non-negative:

1. $\mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}) \leq \log p(\boldsymbol{y})$
2. $\mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu})$ is a lower bound on the marginal likelihood of the data
3. If $q(\boldsymbol{\theta};\boldsymbol{\nu}) = p(\boldsymbol{\theta} \mid \boldsymbol{y})$, then $\mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}) = \log p(\boldsymbol{y})$

ELBO to be maximized w.r.t. the variational parameters $\boldsymbol{\nu}$:

$$\mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}) = \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \text{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu}) \| p(\boldsymbol{\theta})\right)$$

- The first term is a model fitting term:
  - It encourages the model to explain the data well
  - The higher, the better the parameters drawn from $q(\boldsymbol{\theta};\boldsymbol{\nu})$ are at explaining the data
- The second term is a regularization term:
  - It encourages the variational distribution to be close to the prior
  - The lower, the closer the variational distribution is to the prior

Computing the ELBO: Regularization Term

$$\mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}) = \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \text{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu}) \| p(\boldsymbol{\theta})\right)$$

- Recall our assumption that the variational distribution is a product of Gaussians $q(\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(m_j, s_j^2)$
- The second term in the ELBO is the KL divergence between the variational distribution and the prior $p(\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathcal{N}(0, \sigma^2)$
- The KL divergence between two Gaussians is tractable and has a closed-form solution

$$\mathrm{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu})\|p(\boldsymbol{\theta})\right) = \frac{1}{2}\sum_{j=1}^{J}\left(\frac{s_j^2}{\sigma^2} + \frac{m_j^2}{\sigma^2} - 1 + \log\frac{\sigma^2}{s_j^2}\right)$$

Computing the ELBO: Model Fitting Term

$$\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\nu}) = \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})}\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \mathrm{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu})\|p(\boldsymbol{\theta})\right)$$

- The first term is more complex to compute and only analytically available for simple models, but ...

. . .

- ... we can use Monte Carlo methods to estimate it

$$\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})}\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) \approx \frac{1}{S}\sum_{s=1}^{S}\log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(s)})$$

where $\boldsymbol{\theta}^{(s)} \sim q(\boldsymbol{\theta};\boldsymbol{\nu})$

Note: This estimation is unbiased and its variance decreases with $\propto 1/S$, independent of the dimensionality of $\boldsymbol{\theta}$!

ELBO Optimization

ELBO Optimization

Review:

1. We chose a family of distributions $\mathcal{Q}$ to approximate the posterior ($q(\boldsymbol{\theta};\boldsymbol{\nu}) = \prod_{j=1}^{J}\mathcal{N}(m_j, s_j^2)$)
2. We defined the ELBO as the objective function to measure the quality of the approximation

$$\mathcal{L}_{\mathrm{ELBO}}(\boldsymbol{\nu}) = \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})}\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \mathrm{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu})\|p(\boldsymbol{\theta})\right)$$

3. We discussed how to compute the regularization term and the model fitting term

. . .

4. We need to optimize the ELBO w.r.t. the variational parameters $\boldsymbol{\nu}$

$$
\begin{aligned}
\boldsymbol{\nu}^* &= \arg\max_{\boldsymbol{\nu}} \mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}) \\
&= \arg\max_{\boldsymbol{\nu}} \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \text{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu}) \| p(\boldsymbol{\theta})\right)
\end{aligned}
$$

An overview of VI optimization algorithms

VI algorithms can be divided into two categories:

1. **Coordinate Ascent Variational Inference (CAVI)**:
   - Optimize each variational parameter ν_j separately
2. **Gradient-based Variational Inference**:
   - Use gradient-based optimization methods to optimize the variational parameters simultaneously

Gradient-based methods comes in different flavors:

- **Black-box Variational Inference (BBVI)**
- **Reparameterization Gradients (RG)**
- **Stochastic Variational Inference (SVI)**
- **Automatic Differentiation Variational Inference (ADVI)**
- **Amortized Variational Inference**

Optimizing the ELBO is hard

Let's consider the optimization problem:

$$
\begin{aligned}
\boldsymbol{\nu}^* &= \arg\max_{\boldsymbol{\nu}} \mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}) \\
&= \arg\max_{\boldsymbol{\nu}} \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \text{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu}) \| p(\boldsymbol{\theta})\right)
\end{aligned}
$$

We need to compute the gradient of the ELBO w.r.t. the variational parameters $\boldsymbol{\nu}$:

$$
\nabla_{\boldsymbol{\nu}} \mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}) = \nabla_{\boldsymbol{\nu}} \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) - \nabla_{\boldsymbol{\nu}} \text{KL}\left(q(\boldsymbol{\theta};\boldsymbol{\nu}) \| p(\boldsymbol{\theta})\right)
$$

> **!** Problem
>
> We cannot move the gradient inside the expectation because the expectation is w.r.t. the variational distribution $q(\boldsymbol{\theta};\boldsymbol{\nu})$

REINFORCE: The Score Function Gradient Estimator

The **Score Function Gradient Estimator** (REINFORCE) is a general method to estimate gradients of expectations

Log-derivative trick:

$$\nabla_{\nu} q(\theta; \nu) = q(\theta; \nu) \nabla_{\nu} \log q(\theta; \nu)$$

Derivation

Derive the expression above using the chain rule

$$\nabla_z \log f(z) = \frac{\nabla_z f(z)}{f(z)}$$

Then, rearrange the terms

REINFORCE: The Score Function Gradient Estimator

Using the log-derivative trick, we can rewrite the gradient of the ELBO w.r.t. the variational parameters ν:

$$\begin{aligned}
\nabla_{\nu} \mathbb{E}_{q(\theta;\nu)} \log p(y \mid \theta) &= \int \log p(y \mid \theta) \nabla_{\nu} q(\theta; \nu) \, \mathrm{d}\theta \\
&= \mathbb{E}_{q(\theta;\nu)} \log p(y \mid \theta) \nabla_{\nu} \log q(\theta; \nu) \\
&\approx \frac{1}{S} \sum_{s=1}^{S} \log p(y \mid \theta^{(s)}) \nabla_{\nu} \log q(\theta^{(s)}; \nu)
\end{aligned}$$

where $\theta^{(s)} \sim q(\theta; \nu)$.

REINFORCE: The Score Function Gradient Estimator

Pros

- Easy to implement
- Only requires the gradient of the log-density of the variational distribution
- Can be used for any model (hence the name "black-box")

Cons

- High variance
- Slow convergence
- Needs additional variance reduction techniques
- Not popular for (modern) variational inference

Reparameterization Trick

Objective: $\nabla_{\nu} \mathbb{E}_{q(\theta;\nu)} \log p(y \mid \theta)$

. . .

Idea: Freeze the randomness in the variational distribution

1. Samples from $q(\theta;\nu)$ are generated by a deterministic transformation t of a random variable $\varepsilon \sim p(\varepsilon)$
2. The variational parameters ν are parameters of the transformation t
3. The gradient of the expectation w.r.t. the variational parameters can be computed using the chain rule

Gaussian Example

For a Gaussian variational distribution $q(\theta_i;\nu) = \mathcal{N}(m_i, s_i^2)$

1. $p(\varepsilon) = \mathcal{N}(0, 1)$
2. $t(\varepsilon;\nu) = m_i + s_i\varepsilon$

Reparameterization Trick: Derivation

Key observation

For a generic function $f(\theta)$, we have

$$\nabla_{\nu} \mathbb{E}_{q(\theta;\nu)} f(\theta) = \nabla_{\nu} \mathbb{E}_{p(\varepsilon)} f(\theta)$$

with $\theta = t(\varepsilon;\nu)$. Now the expectation is w.r.t. the random variable ε and the gradient can be moved inside the expectation

For the ELBO:

$$
\begin{aligned}
\nabla_{\nu}\mathbb{E}_{q(\theta;\nu)} \log p(\boldsymbol{y} \mid \theta) &= \nabla_{\nu}\mathbb{E}_{p(\varepsilon)} \log p(\boldsymbol{y} \mid \theta) \\
&= \mathbb{E}_{p(\varepsilon)} \nabla_{\nu} \log p(\boldsymbol{y} \mid \theta) \\
&= \mathbb{E}_{p(\varepsilon)} \nabla_{\theta} \log p(\boldsymbol{y} \mid \theta) \nabla_{\nu} \theta \\
&= \mathbb{E}_{p(\varepsilon)} \nabla_{\theta} \log p(\boldsymbol{y} \mid \theta) \nabla_{\nu} t(\varepsilon; \nu) \\
&\approx \frac{1}{S} \sum_{s=1}^{S} \nabla_{\theta} \log p(\boldsymbol{y} \mid \theta^{(s)}) \nabla_{\nu} t(\varepsilon^{(s)}; \nu)
\end{aligned}
$$

where $\varepsilon^{(s)} \sim p(\varepsilon)$ and $\theta^{(s)} = t(\varepsilon^{(s)}; \nu)$.

Reparameterization Trick: Pros and Cons

Pros

- Low variance
- Fast convergence
- No need for additional variance reduction techniques
- Popular in many models (like autoencoders)

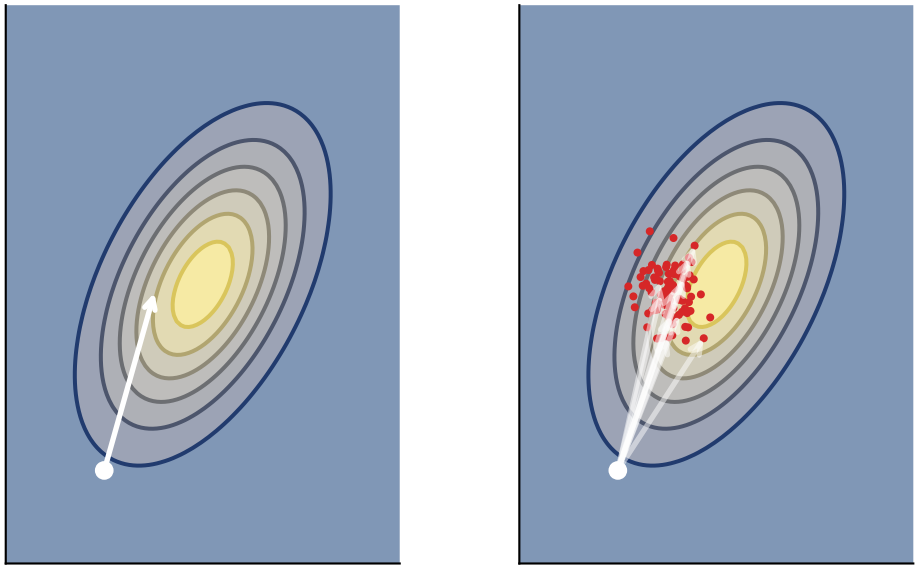
Cons

- Requires reparameterization of the variational distribution
- Needs the model/likelihood to be differentiable

Stochastic Gradient Optimization

The gradient of the ELBO w.r.t. the variational parameters ν are stochastic but unbiased

$$
\mathbb{E}_{\text{noise}} \nabla_{\nu} \widetilde{\mathcal{L}_{\text{ELBO}}}(\nu) = \nabla_{\nu} \mathcal{L}_{\text{ELBO}}(\nu)
$$



Variational inference with stochastic optimization

- **Stochastic optimization** has a long history and good theoretical properties about convergence
- Optimizing using stochastic updates reaches local optima if the learning rate α_t goes to zero with a certain rate (Robbins-Monro conditions)

$$\sum_i^T \alpha_i = \infty \quad \text{and} \quad \sum_i^T \alpha_i^2 < \infty$$

- **Price**: Covergence in $\mathcal{O}(1/\sqrt{T})$ where T is the number of iterations, vs $\mathcal{O}(1/T)$ with exact gradients (when available)

Putting It All Together

Summary

- **Variational Inference (VI)** is a method for approximating intractable posterior distributions
- The goal is to find the best approximation $q(\boldsymbol{\theta}; \boldsymbol{\nu})$ to the true posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$
- The quality of the approximation is measured using the Evidence Lower Bound (ELBO)
- Optimizing the ELBO w.r.t. the variational parameters $\boldsymbol{\nu}$ is challenging:
- The gradients of the ELBO are generally intractable
- We can use Monte Carlo methods to estimate the gradient

Extensions

Extensions of Variational Inference

1. Mini-batch optimization
2. More complex variational families

Mini-batch Optimization

- Likelihood term in the ELBO: $\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta})$
- If the likelihood factorizes over the data points (data points are independent):

$$\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$$

Problem:

- We need to compute the expectation over the variational distribution for each data point
- For large datasets, this can be computationally expensive

Solution: Use mini-batches of data points to estimate the expectation

$$\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) \approx \frac{N}{B} \sum_{b=1}^{B} \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y}_i \mid \boldsymbol{\theta}) \quad \text{with} \quad B \ll N$$

Mini-batch Optimization

$$\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) \approx \frac{N}{B} \sum_{b=1}^{B} \mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\nu})} \log p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$$

Pros:

- Faster convergence
- Scalable to large datasets

But double source of stochasticity:

- Monte Carlo estimation of the expectation
- Mini-batch optimization

More Complex Variational Families

- **Mean-field assumption**: each parameter θ_j is independent and has its own distribution
- **Problem**: The mean-field assumption can be too restrictive $\Rightarrow$ we can use more complex variational families
- If we make the variational family more complex, we get better approximation to the true posterior

Gaussian with Full Covariance

Instead of assuming that the parameters are independent, we can assume that they are correlated

- The variational distribution is a multivariate Gaussian with full covariance matrix

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Reparameterization trick is still applicable using the Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$

$$\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{\varepsilon}, \quad \text{with} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

Complex Variational Families

Important: If the *true* posterior is in the variational family, VI will recover it exactly

$$q(\boldsymbol{\theta}; \boldsymbol{\nu}^*) = p(\boldsymbol{\theta} \mid \boldsymbol{y}) \implies \mathcal{L}_{\text{ELBO}}(\boldsymbol{\nu}^*) = \log p(\boldsymbol{y}) \quad \text{and} \quad \text{KL}\left(q(\boldsymbol{\theta}; \boldsymbol{\nu}^*) \| p(\boldsymbol{\theta} \mid \boldsymbol{y})\right) = 0$$

- More complex variational families lead to better approximations of the true posterior
- But more complex variational families lead to more complex optimization problems and higher computational cost
- Trade-off between approximation quality and computational efficiency

Normalizing Flows

Refresh

Given a invertible function $f : \mathcal{X} \mapsto \mathcal{Y}$ and a simple distribution $p(\boldsymbol{x})$, we can compute the density of $\boldsymbol{y}$ as

$$p(\boldsymbol{y}) = p(\boldsymbol{x}) \left| \det \left(\frac{\partial f^{-1}}{\partial \boldsymbol{y}} \right) \right|$$

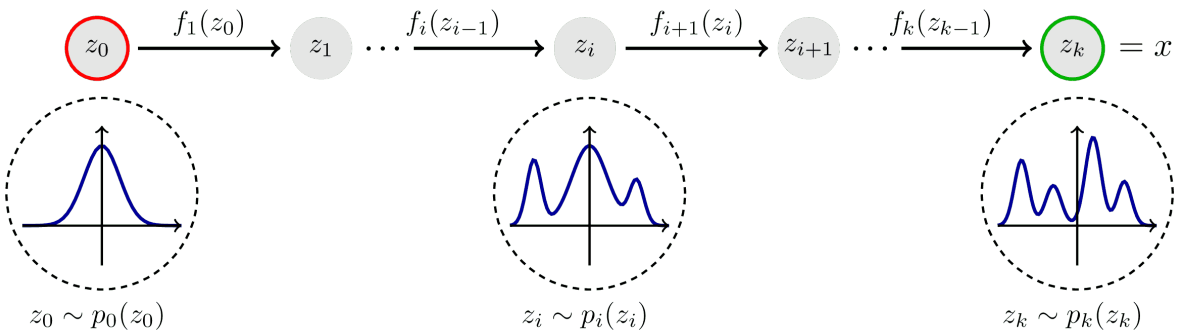
with $\boldsymbol{x} = f^{-1}(\boldsymbol{y})$

We need to build f:

- complex enough to approximate the true posterior
- simple enough to be able to compute the determinant of the Jacobian

Idea: Transform a simple distribution into a complex one using a sequence of invertible transformations

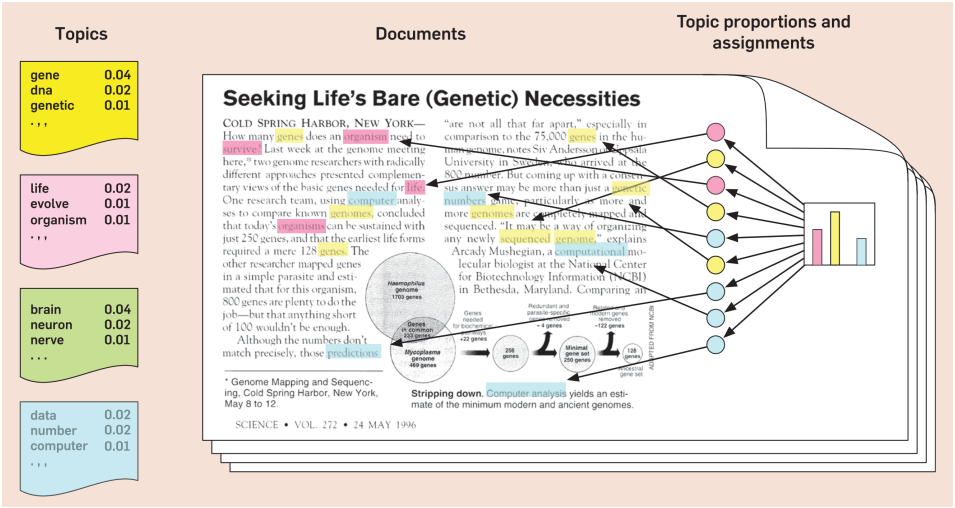
Normalizing Flows



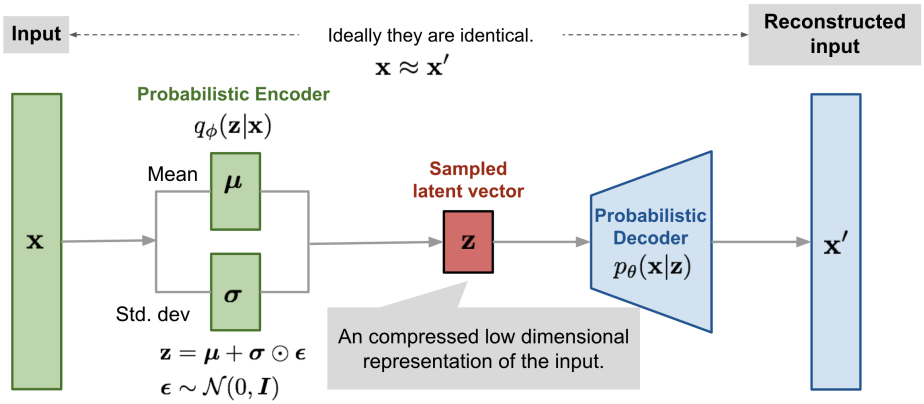
Applications of Variational Inference

Variational inference is used as inference method in many models:

1. Latent Dirichlet Allocation (LDA):
    - Topic modeling
    - Discovering topics in a collection of documents



2. Variational Autoencoders (VAE):
    - Generative models
    - Learning representations of data


Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
Probabilistic Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
Mean
$\boldsymbol{\mu}$
$\boldsymbol{\sigma}$
Std. dev
$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
$\mathbf{x}$
Sampled latent vector
$\mathbf{z}$
An compressed low dimensional representation of the input.
Probabilistic Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$