Lecture 6: Variational Inference

Thibault Randrianarisoa

UTSC

March 5, 2026

Intractable Posteriors

Apart from a few classical examples, the posterior $\pi[\,\cdot\,|\mathbf{X}]$ is intractable.
Popular methods to compute / sample from the posterior:

- **Sampling/Monte-Carlo methods** (Lecture 5): MCMC (Gibbs Sampler, Metropolis-Hastings), Langevin Monte-Carlo, Approximate Bayesian Computation, etc.

- **optimization methods** (Lecture 6, today): Variational inference (VI).

Variational Inference

- Variational inference (VI) methods are a class of ubiquitously used approaches for Bayesian inference wherein we try to directly learn an approximation for the posterior $\pi[\ \theta\ |\mathbf{X}]$
- **Key idea:** reformulate the inference problem to an optimization by learning parameters of a posterior approximation

optim. of a distribution $\Leftrightarrow$ optim. of some parameters

- We do this through introducing a parameterized variational family $q_\phi(\theta)$ then finding the parameter ϕ that gives the "best" approximation
- VI is especially powerful for factorized posteriors as it will allow easy and effective exploitation of the factorization (see below)

Divergences

- How do we quantitatively assess how similar two distributions P and Q are to one another?
- Similarity between distributions is much more subjective than you might expect, particularly for continuous variables
- A **divergence** $\mathbb{D}(P||Q)$ is a, typically asymmetric, way of measuring **dissimilarity** between two distributions P and Q
- We already came across an example divergence in the form of the total variation distance

Definition of VI

1. Chose a tractable family $\mathcal{F}$ of probability distributions on the parameter θ
2. Define

$$q^* = \arg\min_{q \in \mathcal{F}} \mathbb{D}(q||\pi(\theta|\mathbf{X})).$$

Examples

- parametric approximation

$$\mathcal{F} = \left\{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathcal{S}_d^+\right\}.$$

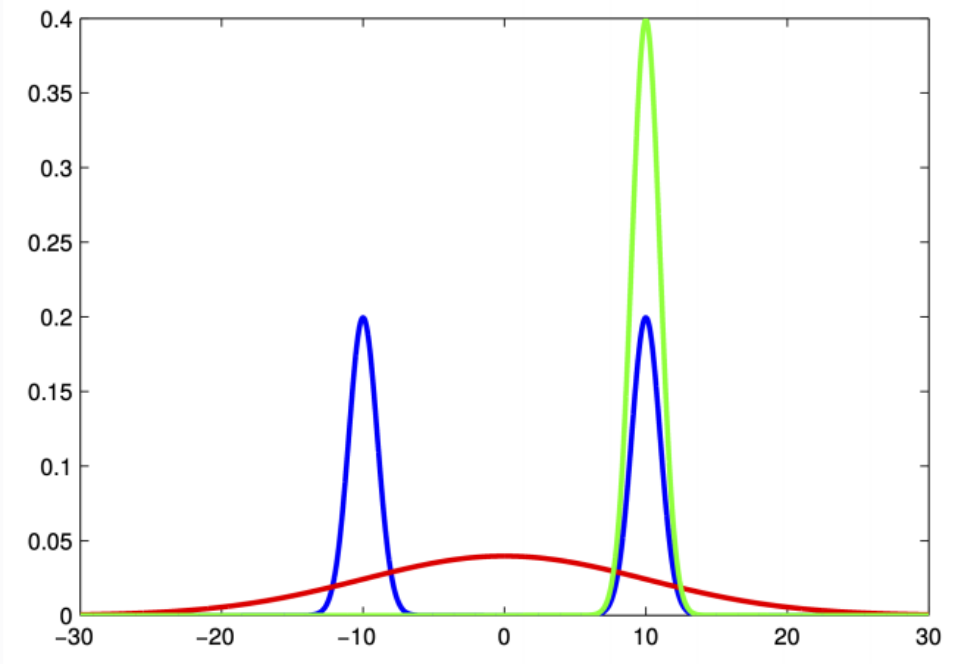
- mean-field approximation, $\theta = (\theta_1, \theta_2) \in \Theta = \Theta_1 \times \Theta_2$,

$$\mathcal{F} = \{q : q(\theta) = q_1(\theta_1) q_2(\theta_2)\}.$$

also $\theta = (\theta_1, \cdots, \theta_m)$

Subjectivity of Divergences

Which is the best fitting Gaussian to our target blue distribution?



Either can be the best depending on how we define our divergence

The Kullback–Leibler (KL) Divergence

The Kullback–Leibler (KL) divergence is one of the most commonly used due to its simplicity, useful computational properties, and the fact that it naturally arises in a number of scenarios

$$\mathbb{D}_{\mathrm{KL}}(Q \parallel P) = \mathbb{E}_{X \sim Q}\left[\log\left(\frac{q(X)}{p(X)}\right)\right] = \mathbb{E}_{q(x)}\left[\log\left(\frac{q(x)}{p(x)}\right)\right]$$

Properties

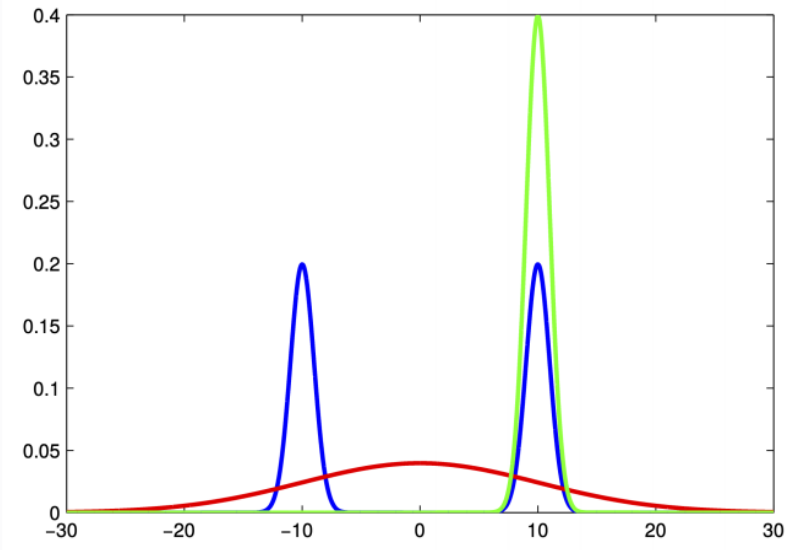
- $\mathbb{D}_{\mathrm{KL}}(Q \parallel P) \geqslant 0,\ \forall P, Q$ (**Gibbs' inequality**)
- $\mathbb{D}_{\mathrm{KL}}(Q \parallel P) = 0$ if and only if $p(x) = q(x)\ \forall x$ (or P=Q)
- In general, $\mathbb{D}_{\mathrm{KL}}(Q \parallel P) \neq \mathbb{D}_{\mathrm{KL}}(P \parallel Q)$

Asymmetry of KL Divergence

Blue: target P

Green: Gaussian Q that minimizes $\mathbb{D}_{\mathsf{KL}}(Q \parallel P)$

Red: Gaussian Q that minimizes $\mathbb{D}_{\mathsf{KL}}(P \parallel Q)$



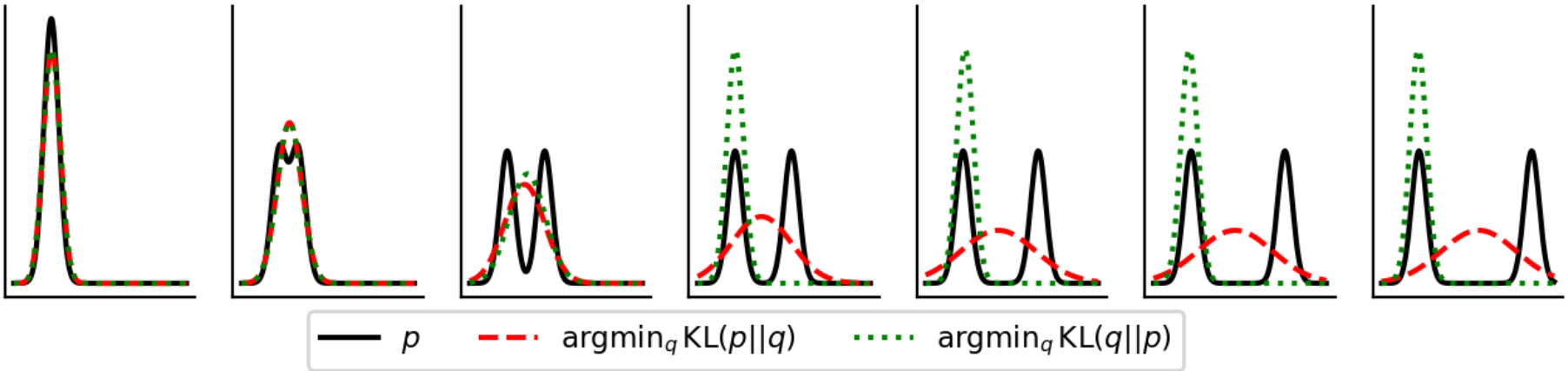
Reverse vs. Forward KL: Intuition

We are approximating a target distribution P with a tractable approximation Q.

- **Reverse KL** (Standard VI): $\mathbb{D}_{\mathrm{KL}}(Q \parallel P) = \mathbb{E}_Q\left[\log \frac{q(x)}{p(x)}\right] = \int q(x) \log \frac{q(x)}{p(x)} dx$
  - **Intuition:** To minimize the KL, $q(x)$ must be **zero** wherever $p(x)$ is zero, or the ratio blows up.
  - **Result:** This forces $q(x)$ to avoid regions where $p(x)$ has low probability. If $p(x)$ is multimodal, $q(x)$ locks onto a single peak. This is **mode seeking**.

- **Forward KL**: $\mathbb{D}_{\mathrm{KL}}(P \parallel Q) = \mathbb{E}_P\left[\log \frac{p(x)}{q(x)}\right] = \int p(x) \log \frac{p(x)}{q(x)} dx$
  - **Intuition:** To minimize the KL, $q(x)$ must be **non-zero** wherever $p(x)$ is non-zero, or the ratio blows up.
  - **Result:** This forces $q(x)$ to cover the entire support of $p(x)$. It averages across modes and bridges gaps. This is **mode covering**.

Visualizing the Difference



Black: Target distribution P

- Intractable, multimodal posterior.

Red: Forward KL Min. $\mathbb{D}_{\mathrm{KL}}(P \parallel Q)$

- **Mode Covering**.
- Overestimates variance.

Green: Reverse KL Min. $\mathbb{D}_{\mathrm{KL}}(Q \parallel P)$

- **Mode Seeking**.
- Standard behavior of VI.
- Underestimates variance.

$\mathcal{F} = \{ q_\phi , \phi \in \mathbb{R}^m \}$

Variational Inference

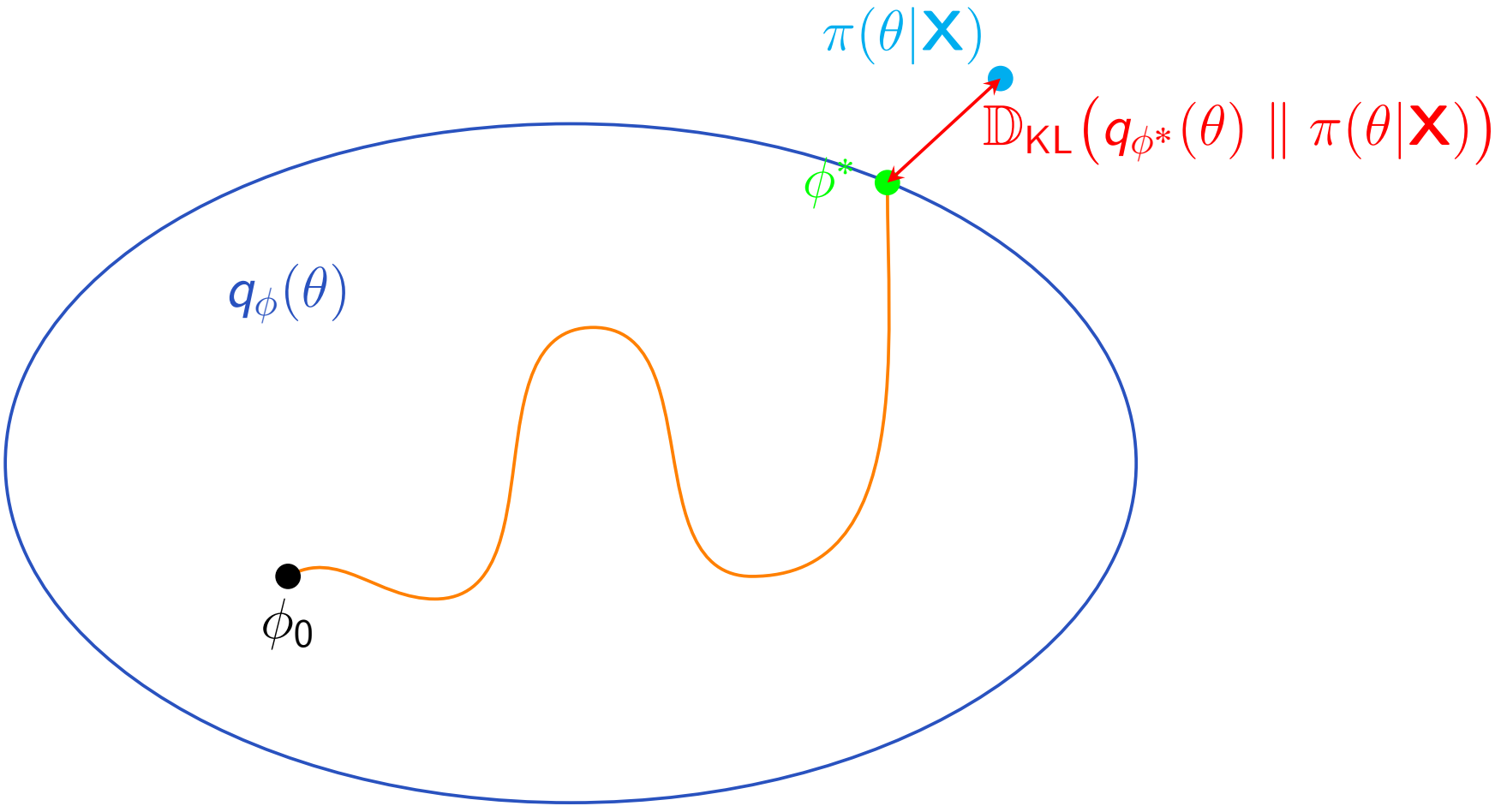
In variational inference with a **parameterized variational family** of distributions $q_\phi(\theta)$, we look for

$$\phi^* = \arg\min_\phi \mathbb{D}_{\mathsf{KL}}(q_\phi(\theta) \parallel \pi(\theta|\mathbf{X}))$$

This allows us to convert the original inference problem into an **optimization**

Critically, we will find that we only need the **unnormalized** form of this posterior, $\pi(\theta)p_\theta(\mathbf{X})$, to perform this optimization

Variational Inference Optimization



Variational Inference

We cannot work directly with $\mathbb{D}_{\mathsf{KL}}(q_\phi(\theta) \parallel \pi(\theta|\mathbf{X}))$ because we don't know the posterior density

However, by noting that the marginal likelihood $f(\mathbf{X}) = \int_\Theta p_\theta(\mathbf{X})\,\pi(\theta)\,d\nu(\theta)$ is independent of our variational parameters ϕ, we see that we can work with the joint instead

$$\begin{aligned}
\phi^* &= \arg\min_\phi \mathbb{D}_{\mathsf{KL}}(q_\phi(\theta) \parallel \pi(\theta|\mathbf{X})) \\
&= \arg\min_\phi \mathbb{E}_{q_\phi(\theta)}\left[\log\left(\frac{q_\phi(\theta)}{\pi(\theta|\mathbf{X})}\right)\right] - \log f(\mathbf{X}) \\
&= \arg\min_\phi \mathbb{E}_{q_\phi(\theta)}\left[\log\left(\frac{q_\phi(\theta)}{\pi(\theta)p_\theta(\mathbf{X})}\right)\right]
\end{aligned}$$

$\pi(\theta|x) = \frac{\pi(\theta)p_\theta(x)}{f(x)}$

This trick is a large part of why we work with $\mathbb{D}_{\mathsf{KL}}(q_\phi(\theta) \parallel \pi(\theta|\mathbf{X}))$ rather than $\mathbb{D}_{\mathsf{KL}}(\pi(\theta|\mathbf{X}) \parallel q_\phi(\theta))$

The ELBO

We can equivalently think about the optimization problem in VI as the maximization

$$\phi^* = \arg\max_{\phi} \mathcal{L}(\phi)$$

where

$$\begin{aligned}\mathcal{L}(\phi) &:= \mathbb{E}_{q_\phi(\theta)}\left[\log\left(\frac{\pi(\theta)p_\theta(\mathbf{X})}{q_\phi(\theta)}\right)\right] = \mathbb{E}_{q_\phi(\theta)}\left[\log p_\theta(\mathbf{X})\right] - \mathbb{D}_{\mathsf{KL}}(q_\phi \parallel \pi) \\ &= \log f(\mathbf{X}) - \mathbb{D}_{\mathsf{KL}}(q_\phi(\theta) \parallel \pi(\theta|\mathbf{X}))\end{aligned}$$

$\mathbb{D}_{\mathsf{KL}}(q_\phi(\theta) \parallel \pi(\theta|\mathbf{X})) \geq 0$, so $\leq \log f(\mathbf{X})$

$\mathcal{L}(\phi)$ is known as the **evidence lower bound (ELBO)** or occasionally as the **variational free energy**

Note that if our variational approximation is exact, that is $q_\phi(\theta) = \pi(\theta|\mathbf{X})$, then $\mathcal{L}(\phi) = \log f(\mathbf{X})$ such that it exactly equals the log evidence

The ELBO

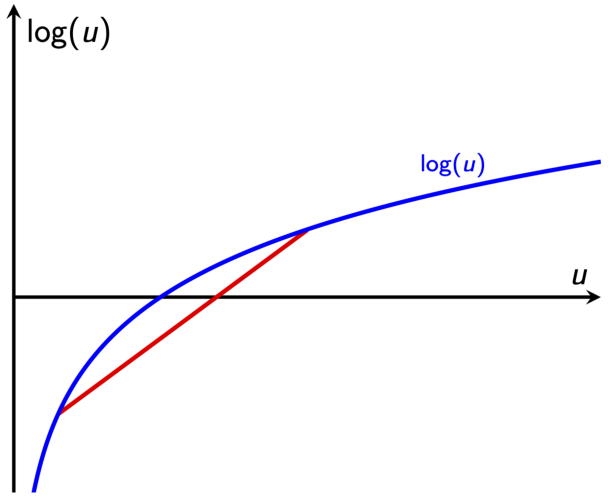
The name ELBO comes from the fact that it is a lower bound on the log evidence by Jensen's inequality using the concavity of log

$$\mathbb{E}_{q_\phi(\theta)}\left[\log\left(\frac{\pi(\theta)p_\theta(\mathbf{X})}{q_\phi(\theta)}\right)\right] \leqslant \log\left(\mathbb{E}_{q_\phi(\theta)}\left[\frac{\pi(\theta)p_\theta(\mathbf{X})}{q_\phi(\theta)}\right]\right) = \log p(\mathbf{X})$$

This bound is **tight** when $q_\phi(\theta) = \pi(\theta|\mathbf{X})$

$$\log\left(\frac{u_1+u_2}{2}\right) \geqslant \frac{\log u_1 + \log u_2}{2}$$

for any $u_1, u_2 > 0$



Optimizing the ELBO

To find the best variational parameters, we need to optimize $\mathcal{L}(\phi)$

There are different ways of doing this, but one common approach is to use a gradient method where we have

$$\nabla_\phi \mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\theta)}\left[\nabla_\phi \log\left(\frac{\pi(\theta)p_\theta(\mathbf{X})}{q_\phi(\theta)}\right)\right] + \int \log\left(\frac{\pi(\theta)p_\theta(\mathbf{X})}{q_\phi(\theta)}\right)\nabla_\phi q_\phi(\theta)d\theta$$

If we can (approximately) calculate this gradient, we can then optimize ϕ by using a (stochastic) gradient ascent approach

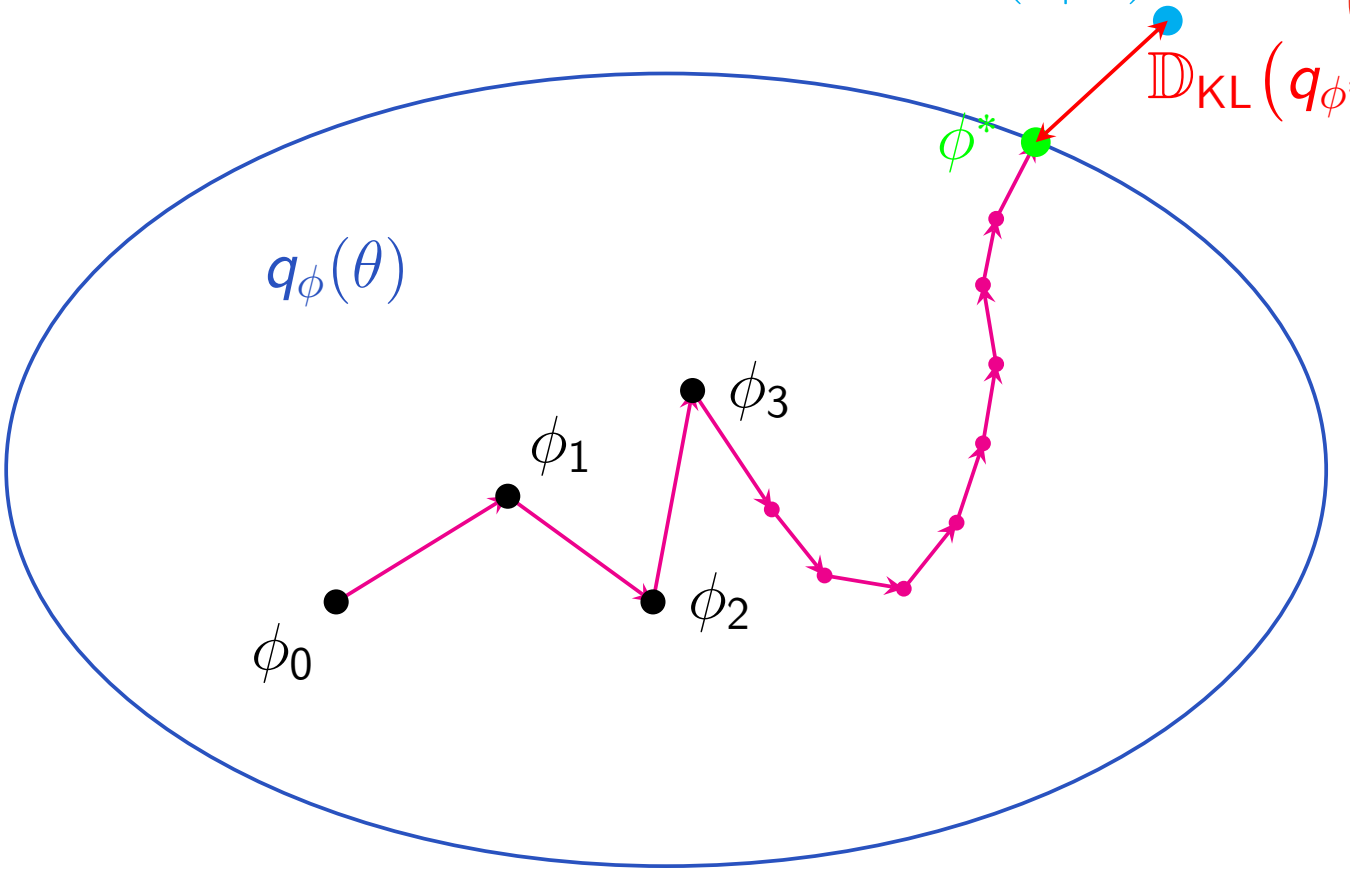
In the simplest case, we initialize at some ϕ_0 and repeatedly apply

$$\phi_{n+1} \leftarrow \phi_n + \epsilon_n \nabla_\phi \mathcal{L}(\phi_n)$$

until convergence, where ϵ_n are our step sizes

Stochastic Variational Inference

= random estimators of the gradients (Monte-Carlo)



Gradient estimator

Usually, the gradient is not available in closed-form but often it is possible to build an **unbiased estimate** of it: $\hat{\nabla}_\phi \mathcal{L}(\phi)$.

First, let's say we want to take the gradient w.r.t. ϕ of an expectation of the form,

$$\mathbb{E}_{q(\theta)}[g_\phi(\theta)]$$

where p is a density. Provided we can differentiate $g_\phi(\theta)$, we can easily compute/approximate the gradient:

$$\nabla_\phi \mathbb{E}_{q(\theta)}[g_\phi(\theta)] = \mathbb{E}_{q(\theta)}[\nabla_\phi g_\phi(\theta)] \approx \frac{1}{M} \sum_{i=1}^{M} \nabla_\phi g_\phi(\theta_i)$$

In words, the gradient of the expectation is equal to the expectation of the gradient.

Gradient estimator

But what happens if our density q is also parameterized by ϕ as in VI?

$$\begin{aligned}
\nabla_\phi \mathbb{E}_{q_\phi(\theta)}[g_\phi(\theta)] &= \nabla_\phi \left[\int q_\phi(\theta) g_\phi(\theta) d\theta \right] \\
&= \int \nabla_\phi [q_\phi(\theta) g_\phi(\theta)] d\theta \\
&= \underbrace{\int g_\phi(\theta) \nabla_\phi q_\phi(\theta) d\theta}_{\text{What about this?}} + \mathbb{E}_{q_\phi(\theta)}[\nabla_\phi g_\phi(\theta)]
\end{aligned}$$

The first term of the last equation is not guaranteed to be an expectation.

Consequence: Monte Carlo methods cannot be used to approximate the integral. This may not be a problem if we have an analytic solution to $\nabla_\phi p_\phi(\theta)$, but this is not true in general.

The Reparameterization Trick

To solve this, we decouple the randomness from the parameters using the **reparameterization trick**.

We assume our random variable $\theta \sim q_\phi$ can be expressed a deterministic, differentiable function h_ϕ of some noise $\epsilon \sim p(\epsilon)$, sampled from a parameter-free distribution:

$$\theta = h_\phi(\epsilon) \sim q_\phi$$

This allows us to rewrite our expectation so the distribution no longer depends on θ:

$$\mathbb{E}_{p_\phi(\theta)}[g_\phi(\theta)] = \mathbb{E}_{p(\epsilon)}[g_\phi(h_\phi(\epsilon))]$$

Estimating the Gradient

Because $p(\epsilon)$ does not depend on ϕ, we can again easily pass the gradient inside the expectation:

$$\begin{aligned}\nabla_\phi \mathbb{E}_{p_\phi(\theta)}[g_\phi(\theta)] &= \nabla_\phi \mathbb{E}_{p(\epsilon)}[g_\phi(h_\phi(\epsilon))] \\ &= \mathbb{E}_{p(\epsilon)}[\nabla_\phi g_\phi(h_\phi(\epsilon))] \\ &\approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\phi g_\phi(h_\phi(\epsilon^{(l)}))\end{aligned}$$

Key takeaway: We use the reparameterization trick to express a gradient of an expectation as an expectation of a gradient. Provided $g_\phi \circ h_\phi$ is differentiable, we can then use Monte Carlo methods to construct an **unbiased estimate**.

Gaussian Example: Computing the KL Divergence

Assume: $q_\phi = \mathcal{N}(\mu, \sigma^2 I)$, with $\phi = (\mu, \sigma) \in \mathbb{R}^d \times \mathbb{R}_+^*$, and prior $\pi = \mathcal{N}(0, I)$.

To compute the ELBO, we need the KL divergence $\mathbb{D}_{\mathsf{KL}}(q_\phi \parallel \pi)$. The general formula for two multivariate Gaussians $\mathcal{N}_0(\mu_0, \Sigma_0)$ and $\mathcal{N}_1(\mu_1, \Sigma_1)$ in d dimensions is:

$$\mathbb{D}_{\mathsf{KL}}(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2}\left[\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_0 - \mu_1)^T \Sigma_1^{-1}(\mu_0 - \mu_1) - d + \ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)\right]$$

Substituting $\mu_0 = \mu$, $\Sigma_0 = \sigma^2 I$, $\mu_1 = 0$, and $\Sigma_1 = I$:

$$\begin{aligned}
\mathbb{D}_{\mathsf{KL}}(q_\phi \parallel \pi) &= \frac{1}{2}\left[\mathrm{tr}(\sigma^2 I) + \mu^T I \mu - d + \ln\left(\frac{1}{|\sigma^2 I|}\right)\right] \\
&= \frac{1}{2}\left[d\sigma^2 + \|\mu\|^2 - d - \ln((\sigma^2)^d)\right] \\
&= \frac{\|\mu\|^2 + d(\sigma^2 - \log(\sigma^2) - 1)}{2}
\end{aligned}$$

Gaussian Example: SVI

Using the reparameterization trick, we write our parameter $\theta \sim q_\phi$ as $\theta = \mu + \sigma\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. $\quad h_\phi(\varepsilon) = \mu + \sigma\varepsilon$

$$\begin{aligned}\mathcal{L}(\phi) &= \mathbb{E}_{q_\phi(\theta)}\left[\log p_\theta(\mathbf{X})\right] - \mathbb{D}_{\mathsf{KL}}(q_\phi \parallel \pi) \\ &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\log p_{\mu+\sigma\epsilon}(\mathbf{X})] - \frac{\|\mu\|^2 + d(\sigma^2 - \log(\sigma^2) - 1)}{2}\end{aligned}$$

And so the true gradient is:

$$\nabla\mathcal{L}(\phi) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left[\nabla_\phi \log p_{\mu+\sigma\epsilon}(\mathbf{X})\right] - \begin{pmatrix} \mu \\ d\left(\sigma - \frac{1}{\sigma}\right)\end{pmatrix}$$

Using m Monte Carlo samples $\epsilon_1, \ldots, \epsilon_m \sim \mathcal{N}(0, I)$, we get the unbiased estimator:

$$\hat{\nabla}\mathcal{L}(\phi) = \begin{pmatrix} \frac{1}{m}\sum_{j=1}^m \nabla \log p_{\mu+\sigma\epsilon_j}(\mathbf{X}) - \mu \\ \frac{1}{m}\sum_{j=1}^m \epsilon_j^T \nabla \log p_{\mu+\sigma\epsilon_j}(\mathbf{X}) - d\left(\sigma - \frac{1}{\sigma}\right)\end{pmatrix}$$

Mean-Field Approximations

Another option, not relying on a parameter ϕ, is the mean-field variational family, where the coordinates of the parameter $\theta = (\theta_1, \ldots, \theta_m)$ are mutually independent and each governed by a distinct factor in the variational density.
A generic member of the mean-field variational family is

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

Note: we have not specified the parametric form of the individual variational factors. In principle, each can take on any parametric form

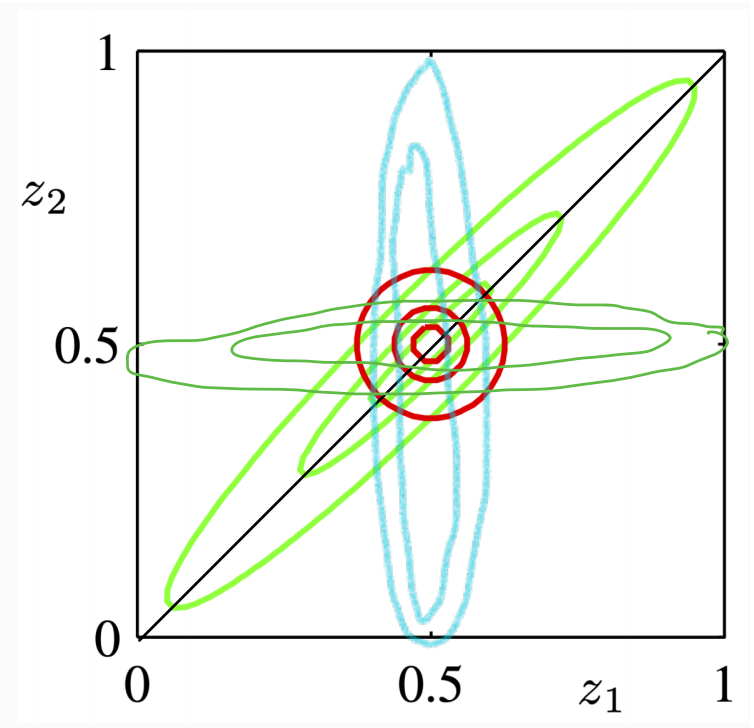
$$q_\phi(\theta) = \prod_i q_{\phi_j}(\theta_j)$$

There are a number of scenarios where this can help make maximizing the ELBO more tractable

Effect of Mean-field Approximations

The mean-field family is expressive because it can capture any marginal density of the parameter. However, it cannot capture correlation between them.

A secondary effect of mean-field approximations is that they tend to lead to underestimating the variance once coupled with the mode-seeking behavior of $\mathbb{D}_{\mathsf{KL}}(q(\theta) \parallel \pi(\theta|\mathbf{X}))$



Optimal variational approximation (red) for target in green when making a mean-field assumption of two dimensions of θ.

Optimal Mean-Field Distribution

Theorem

If $\theta = (\theta_1, \ldots, \theta_m)$, the optimal mean-field approximation q^* is such that for any $j \in [\![1, m]\!]$, if we fix the remaining distributions $q^*_{i \neq j}$,

$$q_j^* = \arg\max_{q_j} \text{ELBO}(q_j) \implies q_j^*(\theta_j) = \frac{\exp\left(\mathbb{E}_{\theta_{i\neq j} \sim q^*_{i\neq j}}[\log \pi(\theta_{i\neq j}, \theta_j) p_{\theta_{i\neq j}, \theta_j}(\mathbf{X}))]\right)}{\int \exp\left(\mathbb{E}_{\theta_{i\neq j} \sim q^*_{i\neq j}}[\log \pi(\theta_{i\neq j}, \theta_j) p_{\theta_{i\neq j}, \theta_j}(\mathbf{X})]\right) d\theta_j}$$

where

$$\text{ELBO}(q_j) = \int \left[q_j \prod_{i\neq j} q_i^*\right] \left[\log\left(\pi(\theta) p_\theta(\mathbf{X})\right) - \left(\log q_j + \sum_{i\neq j} \log q_i^*\right)\right] d\theta$$

Proof: Expanding the ELBO

Let's isolate the terms depending on θ_j:

$$\begin{aligned}
\text{ELBO}(q_j) &= \int \left[q_j \prod_{i \neq j} q_i^* \right] \left[\log\left(\pi(\theta) p_\theta(\mathbf{X})\right) - \left(\log q_j + \sum_{i \neq j} \log q_i^* \right) \right] d\theta \\
&= \int_{\theta_j} q_j(\theta_j) \int_{\theta_{i \neq j}} \left[\prod_{i \neq j} q_i^*(\theta_i) \right] \left[\log\left(\pi(\theta) p_\theta(\mathbf{X})\right) - \left(\log q_j(\theta_j) + \sum_{i \neq j} \log q_i^*(\theta_i) \right) \right] d\theta_{i \neq j} d\theta_j \\
&= \int_{\theta_j} q_j(\theta_j) \int_{\theta_{i \neq j}} \left[\prod_{i \neq j} q_i^*(\theta_i) \right] \log\left(\pi(\theta) p_\theta(\mathbf{X})\right) d\theta_{i \neq j} d\theta_j \\
&\quad - \int_{\theta_j} q_j(\theta_j) \int_{\theta_{i \neq j}} \left[\prod_{i \neq j} q_i^*(\theta_i) \right] \left(\log q_j(\theta_j) + \sum_{i \neq j} \log q_i^*(\theta_i) \right) d\theta_{i \neq j} d\theta_j
\end{aligned}$$

Proof: Evaluating the Integrals

Step 1: Let $\log \tilde{p}(\theta_j \mid \mathbf{X}) := \mathbb{E}_{z_{i\neq j}\sim q_{i\neq j}}[\log \pi(\theta) p_\theta(\mathbf{X})] + K$.

$$\int_{\theta_j} q_j(\theta_j) \int_{\theta_{i\neq j}} \left[\prod_{i\neq j} q_i^*(\theta_i)\right] \log\left(\pi(\theta) p_\theta(\mathbf{X})\right) d\theta_{i\neq j} d\theta_j = \int_{\theta_j} q_j(\theta_j)\ \mathbb{E}_{z_{i\neq j}\sim q_{i\neq j}}[\log \pi(\theta) p_\theta(\mathbf{X})] d\theta_j$$

$$= \int_{\theta_j} q_j(\theta_j)\ (\log \tilde{p}(\theta_j \mid \mathbf{X})) - K) d\theta_j$$

$$= \int_{\theta_j} q_j(\theta_j)\ \log \tilde{p}(\theta_j \mid \mathbf{X}) d\theta_j - K$$

The constant $-K$ will be absorbed into a general constant C.

Proof: Evaluating the Integrals

Step 2:

$$\int_{\theta_j} q_j(\theta_j) \int_{\theta_{i\neq j}} \left[\prod_{i\neq j} q_i^*(\theta_i)\right] \left(\log q_j(\theta_j) + \sum_{i\neq j} \log q_i^*(\theta_i)\right) d\theta_{i\neq j} d\theta_j$$

$$= \int_{z_j} q_j(\theta_j) \log q_j(\theta_j) d\theta_j \underbrace{\int_{\theta_{i\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] d\theta_{i\neq j}}_{=1} + \hat{K} \underbrace{\int_{\theta_j} q_j(\theta_j) d\theta_j}_{=1} = \int_{\theta_j} q_j \log q_j(\theta_j) d\theta_j + \hat{K}$$

where $\hat{K} := \int_{\theta_{i\neq j}} \left[\prod_{i\neq j} q_i(\theta_i)\right] \sum_{i\neq j} \log q_i(\theta_i) d\theta_{i\neq j}$ is constant w.r.t q_j and absorbed into a general constant C.

Proof: Finding the Optimal Distribution

Combining the evaluated integrals, we get:

$$\begin{aligned}\text{ELBO}(q_j) &= \int_{\theta_j} q_j(\theta_j)\ \log \tilde{p}(\theta_j \mid \mathbf{X}) d\theta_j - \int_{\theta_j} q_j \log q_j(\theta_j) d\theta_j + C \\ &= -\mathbb{D}_{\text{KL}}(q_j(\theta_j) \parallel \tilde{p}(\theta_j \mid \mathbf{X})) + C\end{aligned}$$

It is clear that $q_j^* := \tilde{p}(\cdot \mid \mathbf{X})$ maximizes $\text{ELBO}(q_j)$ because it causes the KL divergence term to be zero. Rewriting q_j^*, we have:

$$\begin{aligned}\log q_j^*(\theta_j) &= \log \tilde{p}(\theta_j \mid \mathbf{X}) \\ &= \mathbb{E}_{z_{i\neq j}\sim q_{i\neq j}}[\log \pi(\theta) p_\theta(\mathbf{X})] + K \\ \implies q_j^*(\theta_j) &= \exp\left(\mathbb{E}_{z_{i\neq j}\sim q_{i\neq j}}[\log \pi(\theta) p_\theta(\mathbf{X})]\right) \exp(K) \\ &= \frac{\exp\left(\mathbb{E}_{z_{i\neq j}\sim q_{i\neq j}}[\log \pi(\theta) p_\theta(\mathbf{X})]\right)}{\int \exp\left(\mathbb{E}_{z_{i\neq j}\sim q_{i\neq j}}[\log \pi(\theta) p_\theta(\mathbf{X})]\right) d\theta_j}\end{aligned}$$

The constant $\exp(K)$ is the normalization constant. ■

Coordinate Ascent Variational Inference (CAVI)

Algorithm: CAVI

Input: Unnormalized posterior $\pi(\theta)p_\theta(\mathbf{X})$
Output: A variational density $q(\theta) = \prod_{j=1}^{m} q_j(\theta_j)$
Initialize: Variational factors $q_j(\theta_j)$

while *the ELBO has not converged* **do**
    for $j \in \{1, \ldots, m\}$ **do**
        Set $q_j(\theta_j) \propto \exp\left(\mathbb{E}_{\theta_{i\neq j}\sim q_{i\neq j}}[\log \pi(\theta_{i\neq j}, \theta_j)p_{\theta_{i\neq j},\theta_j}(\mathbf{X}))]\right)$
    end
    Compute ELBO
end

return $q(\theta)$

CAVI and Gibbs Sampling

CAVI is closely related to **Gibbs sampling**

- **Gibbs Sampler:** Maintains a realization of the parameters and iteratively *samples* from each component's complete conditional.
- **CAVI:** Uses the same complete conditional. However, it takes the *expected log*, and uses this quantity to iteratively set each variable's variational factor.

Convergence Property

By iteratively updating these factors, CAVI goes uphill on the ELBO, eventually finding a **local optimum**.

Pros and Cons of Variational Methods

Pros

- Typically more efficient than MCMC approaches, particularly in high dimensions once we exploit the stochastic variational approach
  - Can often provided effective inference for models where MCMC methods have impractically slow convergence
- Allows simultaneous optimization of model parameters

Cons

- It produces (potentially very) biased estimates and requires strong structural assumptions to be made about the form of the posterior
  - Unlike MCMC methods, this bias stays even in the limit of large computation
- Can require substantial tailoring to a particular problem
- Very difficult to estimate how much error there is in the approximation: subsequent estimates can be unreliable, particularly in their uncertainty
- Tends to underestimate the variance of the posterior, particularly when using inexact mean-field approximations