Lecture 5: Advanced Bayesian Inference Methods

Advanced Topics in Machine Learning

Dr. Tom Rainforth

January 31st, 2020

rainforth@stats.ox.ac.uk

Recap: Bayesian Inference is Hard!

Image Credit: Metro

Recap: Monte Carlo

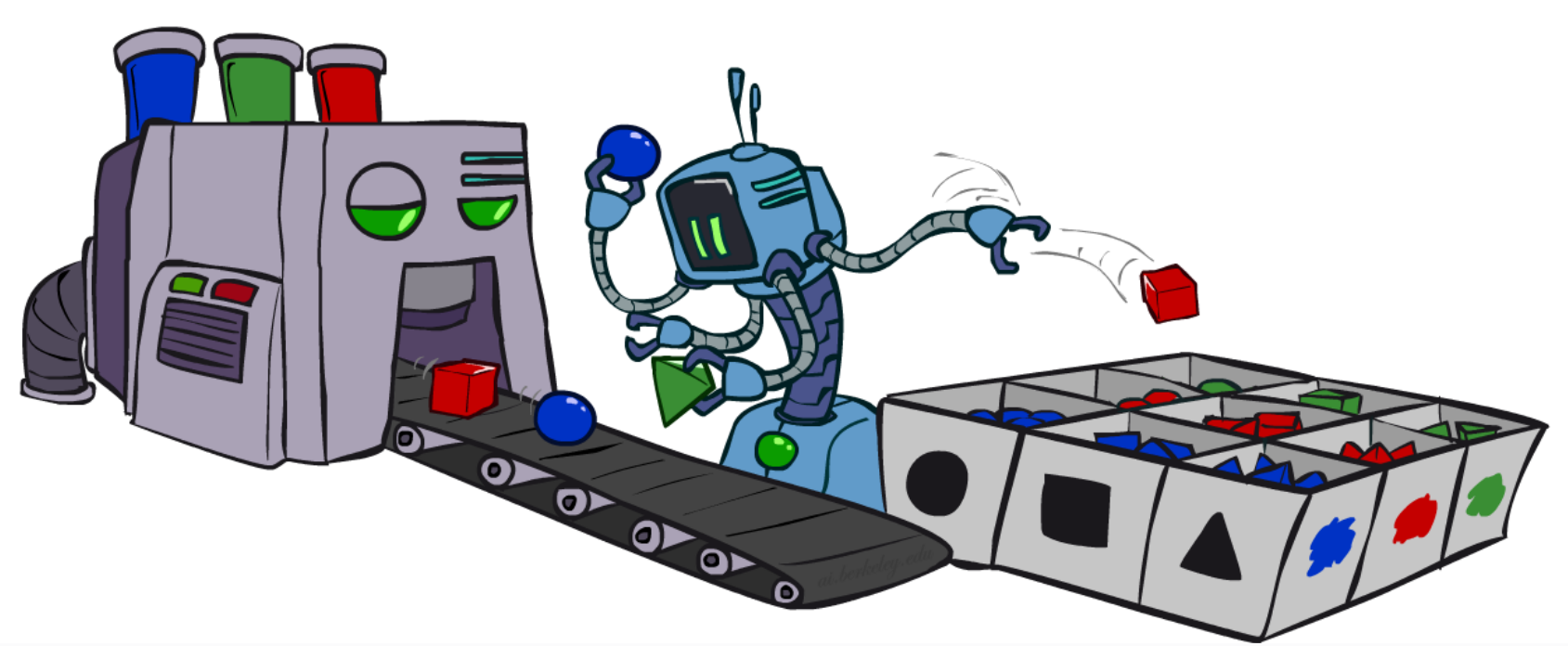
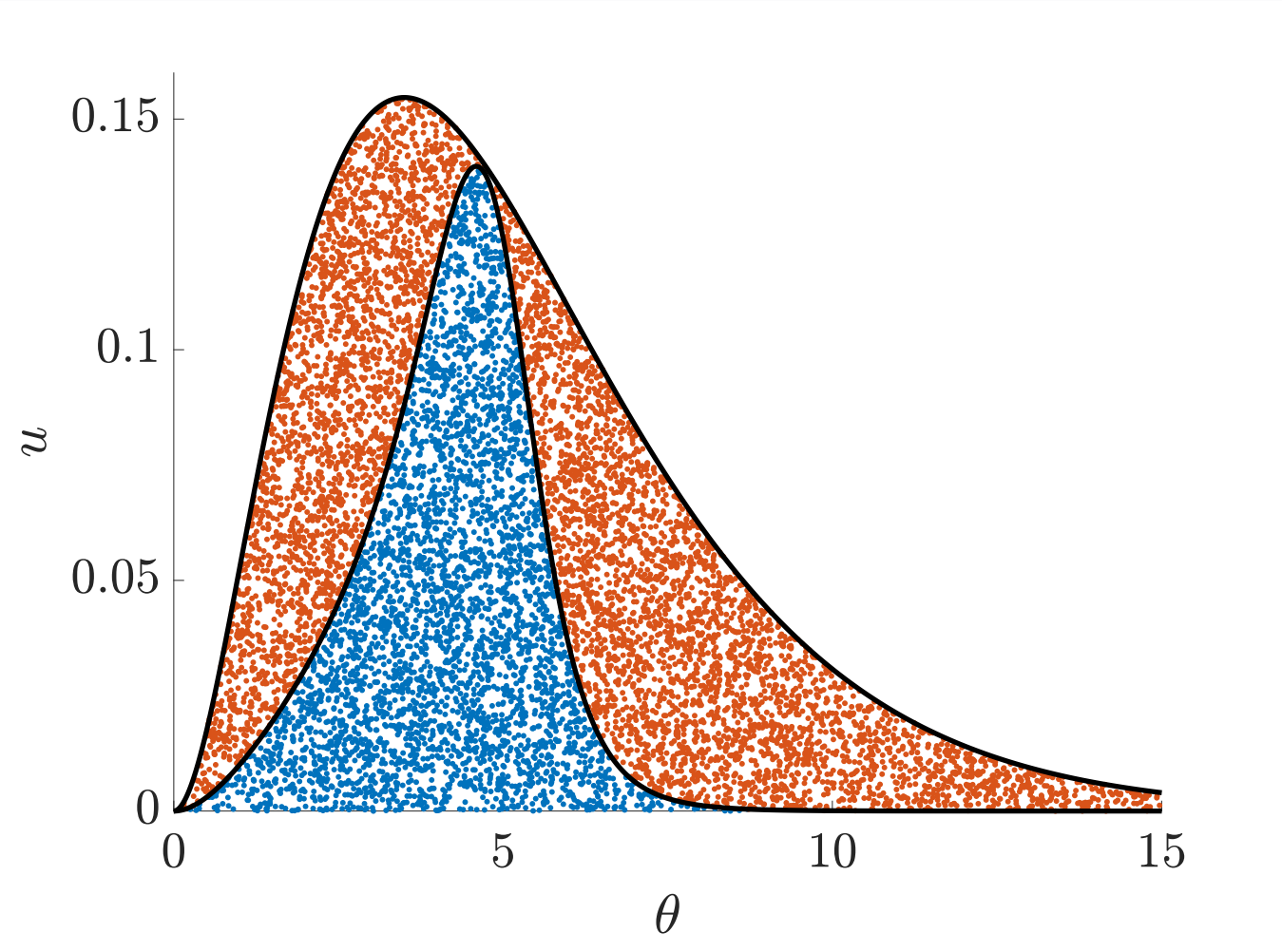
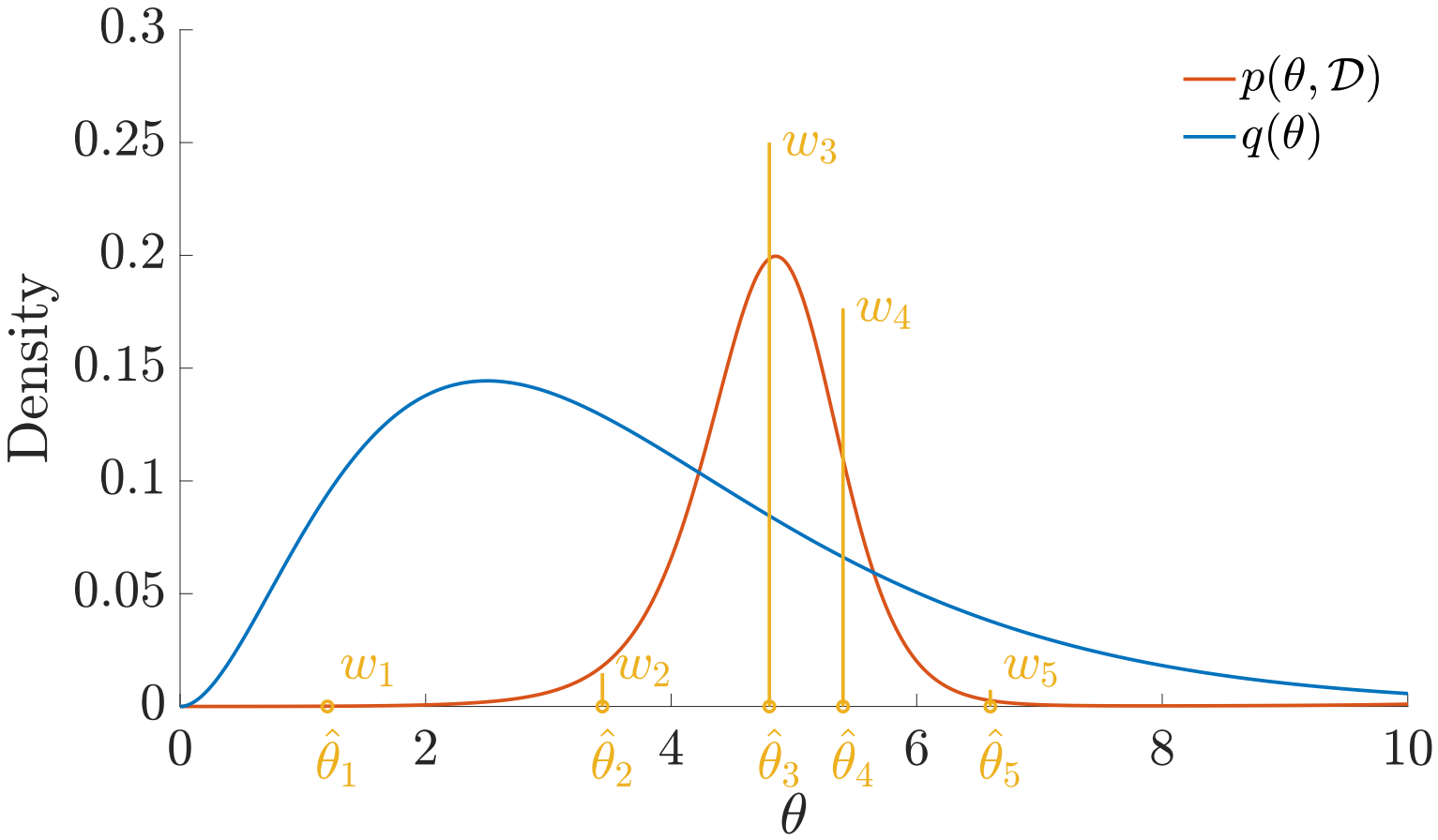


Image Credit: Pieter Abbeel

Recap: Rejection Sampling



Recap: Importance Sampling



Lecture 5 Outline

In this lecture we will we show how the foundational methods introduced in the last section are not sufficient for inference in high dimensions

Particular topics:

- The Curse of Dimensionality
- Markov Chain Monte Carlo (MCMC)
- Variational Inference

The Curse of Dimensionality

The Curse of Dimensionality

- The **curse of dimensionality** is a tendency of modeling and numerical procedures to get substantially harder as the dimensionality increases, often at an exponential rate.
- If not managed properly, it can cripple the performance of inference methods
- It is the main reason the two procedures discussed so far, rejection sampling and importance sampling, are in practice only used for very low dimensional problems
- At its core, it stems from an increase of the size (in an informal sense) of a problem as the dimensionality increases

The Curse of Dimensionality (2)

- Imagine we are calculating an expectation over a discrete distribution of dimension D, where each dimension has K possible values
- The cost of enumerating all the possible combinations scales as K^D and thus increases exponentially with D; even for modest values for K and D this will be prohibitively large
- The same problem occurs in continuous spaces: think about splitting the space into blocks, we have to reason about all the blocks to reason about the problem

The Curse of Dimensionality (3)

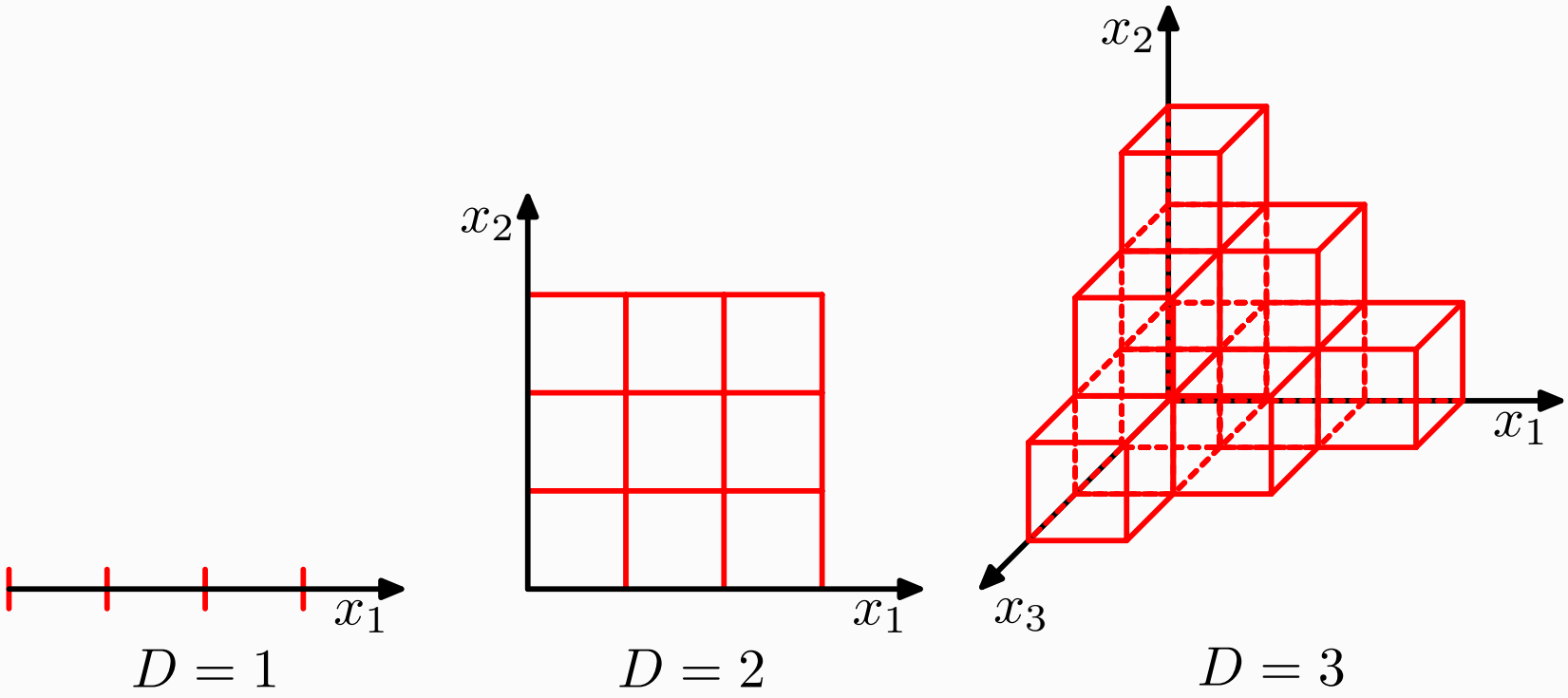


Image Credit: Bishop, Section 1.4

Example: Rejection Sampling From a Sphere

Consider rejection sampling from a D-dimensional hypersphere with radius r using the tightest possible enclosing box:

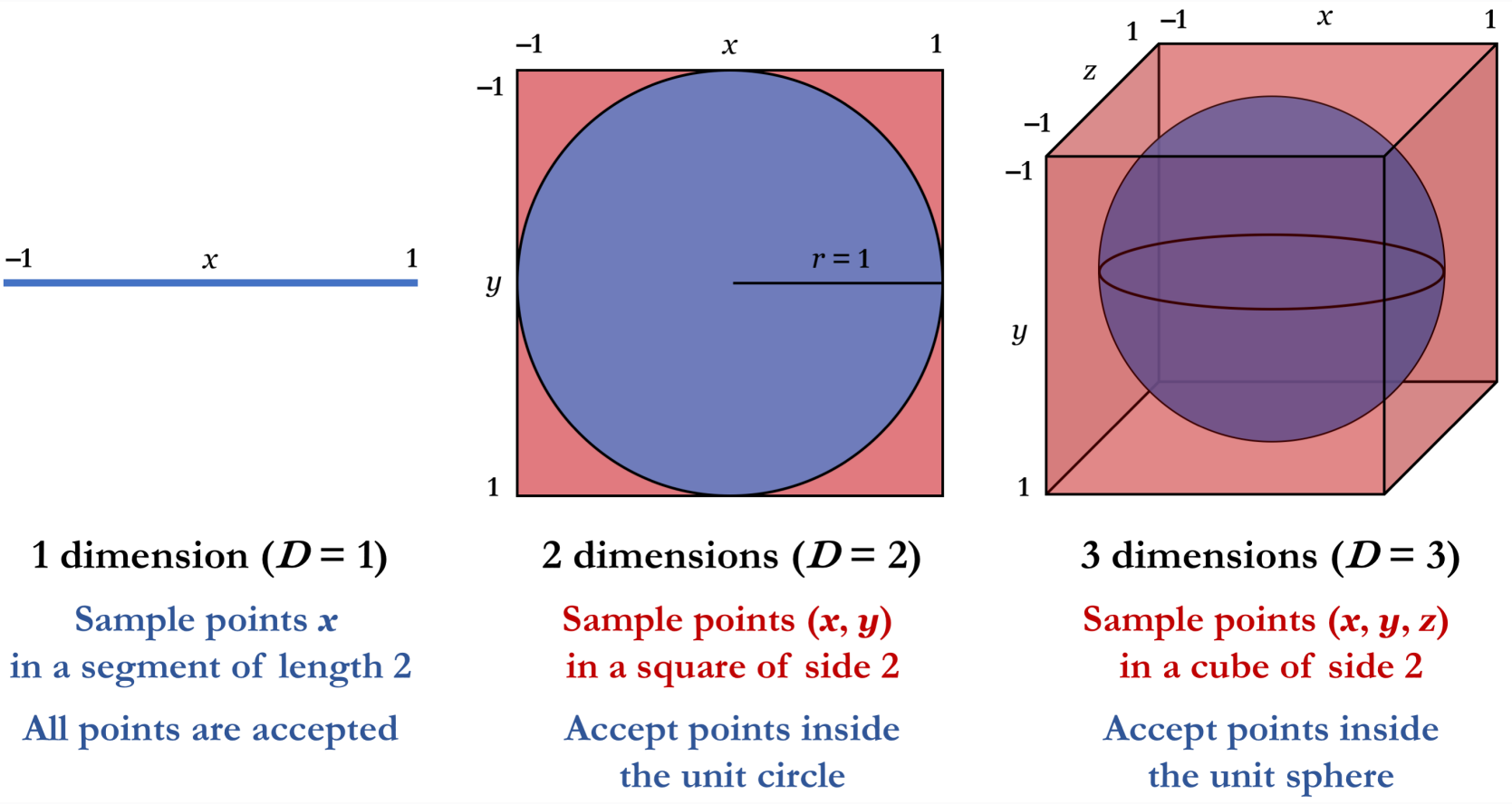


Image Credit: Adrian Baez-Ortegal

Example: Rejection Sampling From a Sphere (2)

Here the acceptance rate is equal to the ratio of the two volumes. For even values of D this is given by

$$P(\text{Accept}) = \frac{V_s}{V_c} = \frac{\pi^{D/2} r^D / (D/2)!}{(2r)^D} = \left(\frac{\sqrt{\pi}}{2}\right)^D \frac{1}{(D/2)!}$$

This now decreases super–exponentially in D (noting that $(D/2)! > (D/6)^{(D/2)}$)

$D = 2, 10, 20$, and 100 respectively gives $P(\text{Accept})$ values of 0.79, 2.5×10^{-3}, 2.5×10^{-8}, and 1.9×10^{-70}

Sampling this way was perfect in one-dimension, but quickly becomes completely infeasible in higher dimensions

Curse of Dimensionality for Importance/Rejection Sampling

- For both importance sampling and rejection sampling we use a proposal $q(\theta)$
- This proposal is an approximation of the target $p(\theta|\mathcal{D})$
- As the dimension increases, it quickly becomes much harder to find good approximations
- The performance of both methods typically diminishes exponentially as the dimension increases

Typical Sets

Another consequence of the curse of dimensionality is that most of the posterior mass becomes concentrated away from the mode.

Consider representing an isotropic Gaussian in polar coordinates. The marginal density of the radius changes with dimension:

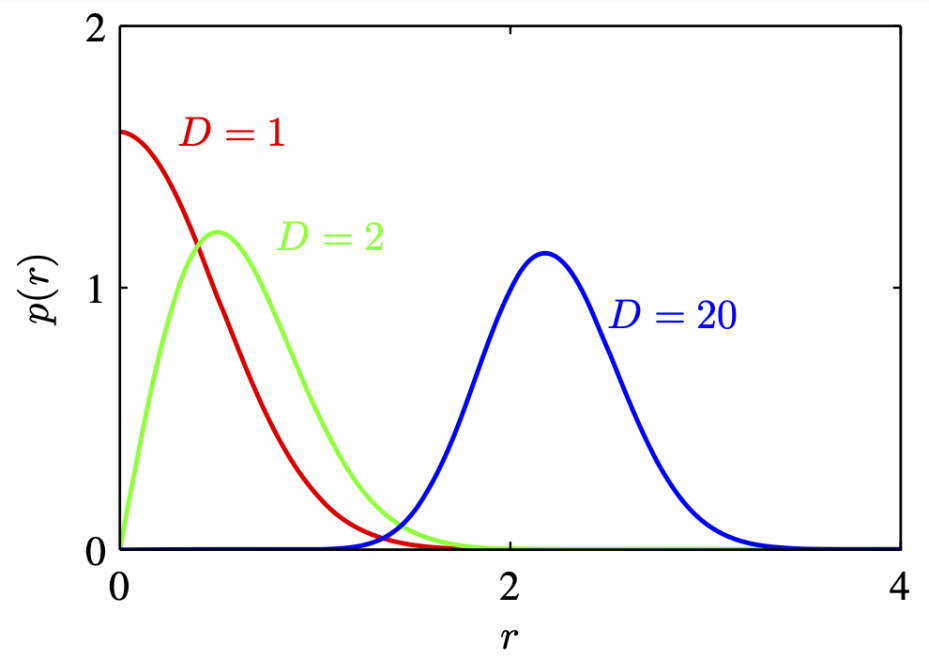
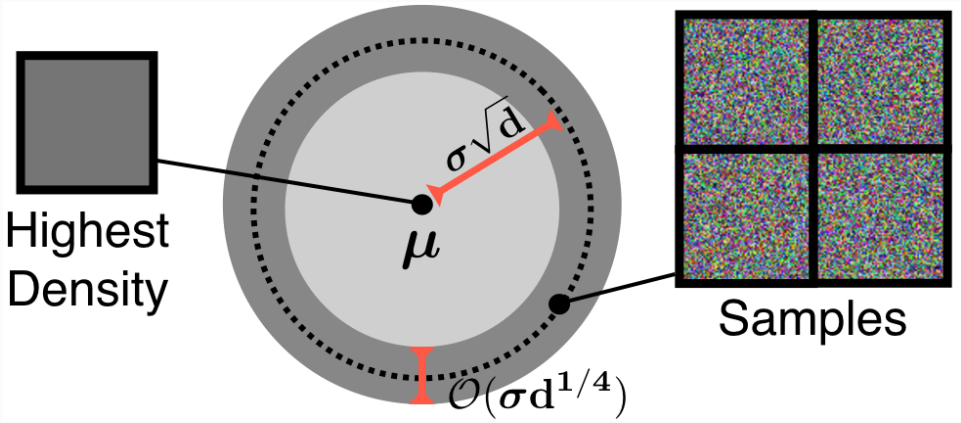
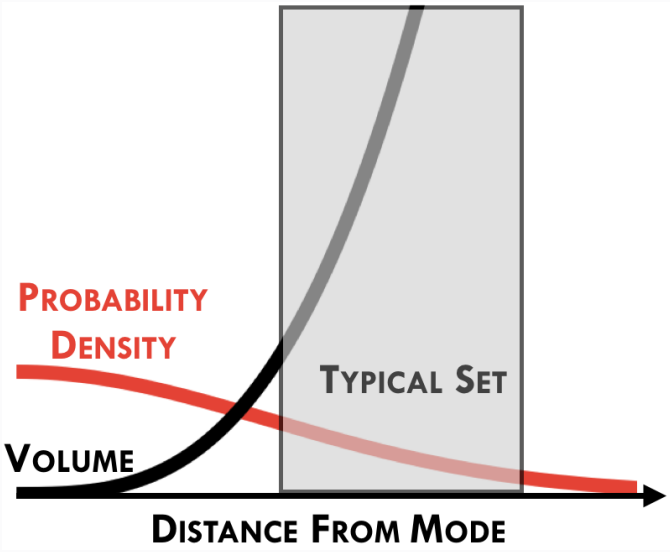


Image Credit: Bishop, Section 1.4

Typical Sets

In high dimensions, the posterior mass concentrates in an thin strip away from the mode known as the **typical set**





This means that, not only is the mass concentrated to a small proportion of the space in high dimensions, the geometry of this space can be quite complicated

[1] E Nalisnick et al. "Detecting out-of-distribution inputs to deep generative models using a test for typicality". In: *arXiv preprint arXiv:1906.02994* (2019).

How Can We Overcome The Curse of Dimensionality?

- As we showed with the typical sets, the area of significant posterior is usually only a small proportion of the overall space
- To overcome the curse, we thus need to use methods which **exploit structure** of the posterior to only search this small subset of the overall space
- All successful inference algorithms make some implicit assumptions into the structure and then try to exploit this
    - MCMC methods exploit local moves to try and stick within the typical set (thereby also implicitly assuming there are not multiple modes)
    - Variational methods assume independences between different dimensions that allow large problems to be broken into multiple smaller problems

Markov Chain Monte Carlo

Markov Chains

http://setosa.io/ev/markov-chains/

The Markov Property

In a Markovian system each state is independent of all the previous states given the last state, i.e.

$$p(\theta_n|\theta_1, \ldots, \theta_{n-1}) = p(\theta_n|\theta_{n-1})$$

The system transitions based only on its current state.

Here the series of random variables produced the system (i.e. $\Theta_1, \ldots, \Theta_n, \ldots$) is known as a Markov chain.

Defining a Markov Chain

- All the Markov chains we will deal with are **homogeneous**
- This means that each transition has the same distribution:

$$p(\Theta_{n+1} = \theta' | \Theta_n = \theta) = p(\Theta_n = \theta' | \Theta_{n-1} = \theta),$$

- In such situations, $p(\Theta_{n+1} = \theta_{n+1} | \Theta_n = \theta_n)$ is typically known as a **transition kernel** $T(\theta_{n+1} \leftarrow \theta_n)$
- The distribution of any homogeneous Markov chain is fully defined by a combination of an initial distribution $p(\theta_1)$ and the transition kernel $T(\theta_{n+1} \leftarrow \theta_n)$, e.g.

$$p(\theta_m) = \int T(\theta_m \leftarrow \theta_{m-1}) T(\theta_{m-1} \leftarrow \theta_{m-2}) \dots$$
$$T(\theta_2 \leftarrow \theta_1) p(\theta_1) d\theta_{1:m-1}$$

Random walks: Continuous Markov Chains

- Markov chains do not have to be in discrete spaces
- In continuous spaces we can informally think of them as guided random walks through the space with finite sized steps

https://youtu.be/7A83lXbs6Ik?t=114

Markov Chain Monte Carlo (MCMC)

- Markov chain Monte Carlo (MCMC) methods are one of the most ubiquitous approaches for Bayesian inference and sampling from target distributions more generally
- The key idea is to construct a valid **Markov chain** that produces sample from the target distribution
- They only require that the target distribution is only known up to a normalization constant.
- They circumvent the curse of dimensionality by exploiting local moves
    - They have a hill–climbing effect until they reach the typical set
    - They then move around the typical set using local moves
    - They tend to fail spectacularly in the presence of multi–modality

Convergence of a Markov Chain

- To use a Markov chain for consistent inference, we need it to be able to produce an infinite series of samples that converge to our posterior:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=M}^{N} \delta_{\Theta_n}(\theta) \xrightarrow{d} p(\theta|\mathcal{D}) \tag{1}$$

  where M is a number of **burn–in** samples that we discard from the start of the chain
- In most cases, a core condition for this to hold is that the distribution of individual samples converge to the target for all possible starting points:

$$\lim_{N\to\infty} p(\Theta_N = \theta|\Theta_1 = \theta_1) = p(\theta|\mathcal{D}) \quad \forall\theta_1 \tag{2}$$

Convergence of a Markov Chain

Ensuring that the chain converges to the target distribution for all possible initializations has two requirements

1. $p(\theta|\mathcal{D})$ must be the **stationary distribution** of the chain, such that if $p(\Theta_n = \theta) = p(\theta|\mathcal{D})$ then $p(\Theta_{n+1} = \theta) = p(\theta|\mathcal{D})$. This is satisfied if:

$$\int T(\theta' \leftarrow \theta) p(\theta|\mathcal{D}) d\theta = p(\theta'|\mathcal{D}) \tag{3}$$

   where we see that the target is **invariant** to the application of the transition kernel.

2. The Markov chain must be **ergodic**. This means that all possible starting points converge to this distribution.

Ergodicity

Ergodicity itself has two requirements, the chain must be:

1. **Irreducible**, i.e. all points with non-zero probability can be reached in a finite number of steps
2. **Aperiodic**, i.e. no states can only be reached at certain periods of time

These requirements for these to be satisfied are very mild for commonly used Markov chains, but are beyond the scope of the course

Stationary Distribution

Optional homework: figure out how we can get the stationary distribution from the transition kernel when θ is discrete

Hint: start by defining the transition kernel as a matrix

Detailed Balance

A sufficient condition used for constructing valid Markov chains is to ensure that the chain satisfies **detailed balance**:

$$p(\theta|\mathcal{D})\,T(\theta' \leftarrow \theta) = p(\theta'|\mathcal{D})\,T(\theta \leftarrow \theta'). \tag{4}$$

Chains that satisfy detailed balance are known as **reversible**.

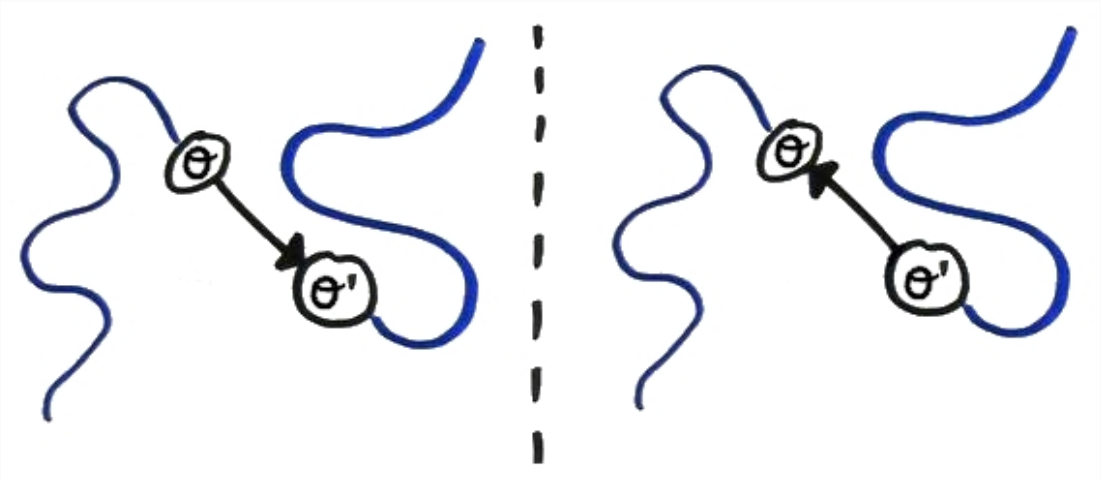


Image Credit: Iain Murray

Detailed Balance (2)

It is straightforward to see that Markov chains satisfying detailed balance will admit $p(\theta|\mathcal{D})$ as a stationary distribution by noting that

$$\int T(\theta' \leftarrow \theta)p(\theta|\mathcal{D})d\theta = \int T(\theta \leftarrow \theta')p(\theta'|\mathcal{D})d\theta = p(\theta'|\mathcal{D}). \quad (5)$$

We can thus construct valid **MCMC samplers** by using detailed balance to construct a valid transition kernel for our target, sampling a starting point, then repeatedly applying the transition kernel

Metropolis Hastings

One of the simplest and most widely used MCMC methods is **Metropolis Hastings** (MH).

Given an unnormalized target $p(\theta, \mathcal{D})$, a starting point θ_1, and a proposal $q(\theta'|\theta)$, the MH algorithm repeatedly applies the following steps ad infinitum

1. Propose a new point $\theta' \sim q(\theta'|\theta = \theta_n)$
2. Accept the new sample with probability

$$P(\text{Accept}) = \min\left(1, \frac{p(\theta', \mathcal{D})q(\theta_n|\theta')}{p(\theta_n, \mathcal{D})q(\theta'|\theta_n)}\right) \tag{6}$$

   in which case we set $\theta_{n+1} \leftarrow \theta'$
3. If the sample is rejected, set $\theta_{n+1} \leftarrow \theta_n$
4. Go back to step 1

Metropolis Hastings (2)

- This produces an infinite sequence of samples $\theta_1, \theta_2, \ldots, \theta_n, \ldots$ that converge to $p(\theta|\mathcal{D})$ and from which we can construct a Monte Carlo estimator

$$p(\theta|\mathcal{D}) \approx \frac{1}{N} \sum_{n=M}^{N} \delta_{\theta_n}(\theta) \tag{7}$$

  where we start with sample M to **burn-in** the chain
- Note that MH only requires the unnormalized target $p(\theta, \mathcal{D})$
- Unlike rejection/importance sampling, the samples are correlated and produce biased estimates for finite N
- The key though is that the proposal $q(\theta'|\theta)$ depends on the current position allowing us to make **local moves**

MCMC Methods Demo

https://chi-feng.github.io/mcmc-demo/app.html?algorithm=RandomWalkMH&target=banana

More Advanced MCMC Methods

There are loads of more advanced MCMC methods.

Two that are particularly prominent ones that you should be able to quickly pick up given what you have already learned are:

- Gibbs sampling (see the notes)
- Hamiltonian Monte Carlo: `https://arxiv.org/pdf/1206.1901.pdf?fname=cm&font=TypeI`

Hamiltonian Monte Carlo Demo

https://chi-feng.github.io/mcmc-demo/app.html?algorithm=HamiltonianMC&target=donut

Pros and Cons of MCMC Methods

Pros

- Able to work in high dimensions due to making local moves
- No requirement to have normalized target
- Consistent in the limit of running the chain for an infinitely long time
- Do not require as finely tuned proposals as importance sampling or rejection sampling
- Surprisingly effective for a huge range of problems

Pros and Cons of MCMC Methods (2)

Cons

- Produce biased estimates for finite sample sizes due to correlation between samples
- Diagnostics can be very difficult
- Typically struggle to deal with multiple modes
- Proposal still quite important: chain can **mix** very slowly if the proposal is not good
- Can be difficult to parallelize
- Deriving theoretical results is more difficult than previous approaches
- Produces no marginal likelihood estimate
- Typically far slower to converge than the variational methods we introduce next

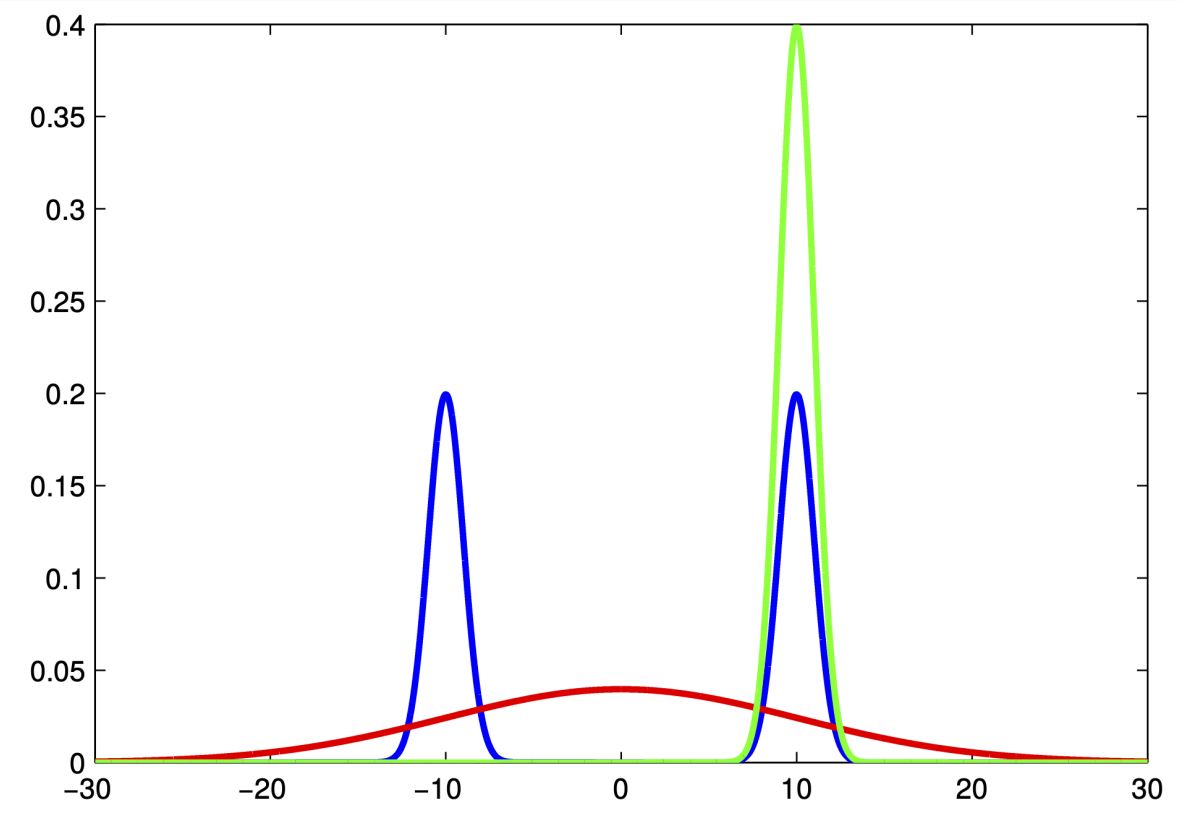
Background for Variational Inference: Divergences

Divergences

- How do we quantitatively assess how similar two distributions $p(x)$ and $q(x)$ are to one another?
- Similarity between distributions is much more subjective than you might expect, particularly for continuous variables
- Metrics for measuring the "distance" between two distributions are known as a **divergence** and typically expressed in the form $\mathbb{D}(p(x)||q(x))$
- Note that most divergences are **not** symmetric

Divergences (2)

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



Either can be the best depending how we define our divergence

Image Credit: Barber, Section 28.3.4

The Kullback–Leibler (KL) Divergence

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

$$\mathrm{KL}(q(x) \,\|\, p(x)) = \int q(x) \log \frac{q(x)}{p(x)} dx = \mathbb{E}\, q(x) \left[\log \frac{q(x)}{p(x)} \right] \quad (8)$$

Importance properties:

- $\mathrm{KL}(q(x) \,\|\, p(x)) \geq 0$ for all $q(x)$ and $p(x)$
- $\mathrm{KL}(q(x) \,\|\, p(x)) = 0$ if and only if $p(x) = q(x)\, \forall x$
- In general, $\mathrm{KL}(q(x) \,\|\, p(x)) \neq \mathrm{KL}(p(x) \,\|\, q(x))$

Asymmetry of KL Divergence

Blue: target $p(x)$

Green: Gaussian $q(x)$ that minimizes $\mathrm{KL}(q(x) \| p(x))$

Red: Gaussian $q(x)$ that minimizes $\mathrm{KL}(p(x) \| q(x))$

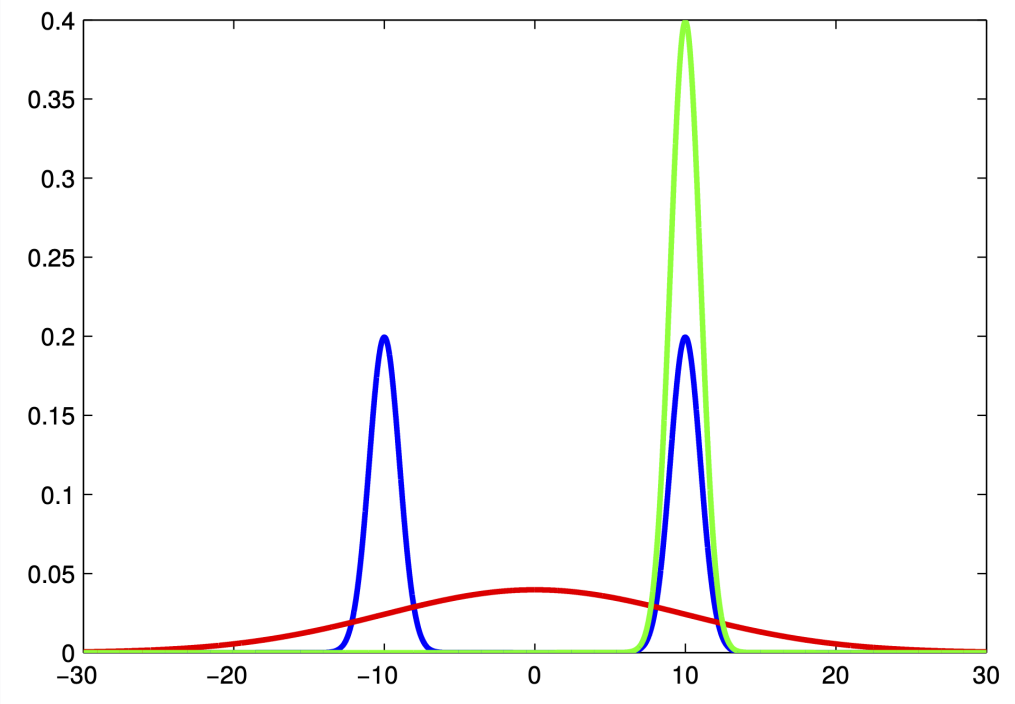


Image Credit: Barber, Section 28.3.4

Mode Covering KL

The "forward" KL, $\mathrm{KL}(p(x) \,\|\, q(x))$, is **mode covering**: $q(x)$ must place mass anywhere $p(x)$ does

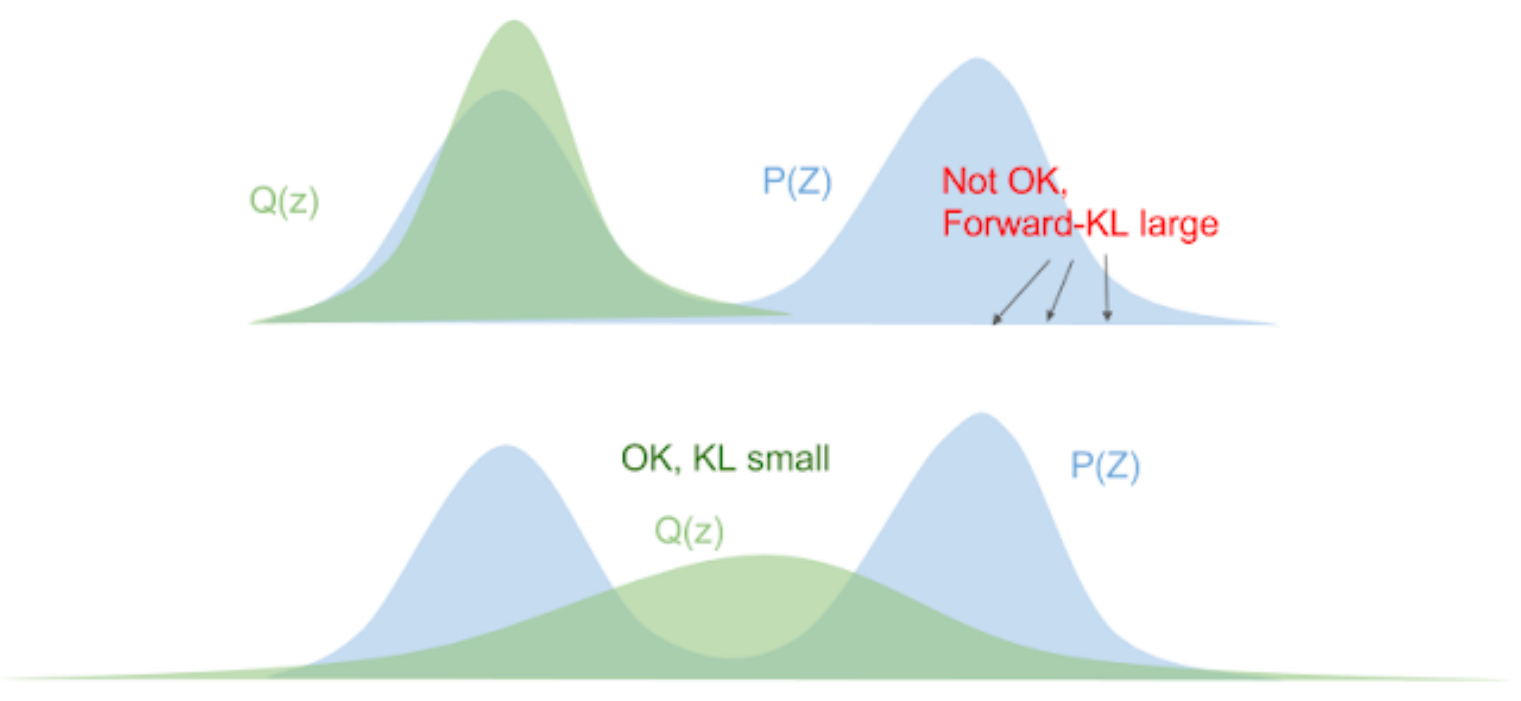


Image Credit: Eric Jang

Mode Seeking KL

The "reverse" KL, $\mathrm{KL}(q(x) \,\|\, p(x))$, is **mode seeking**: $q(x)$ must not place mass anywhere $p(x)$ does not

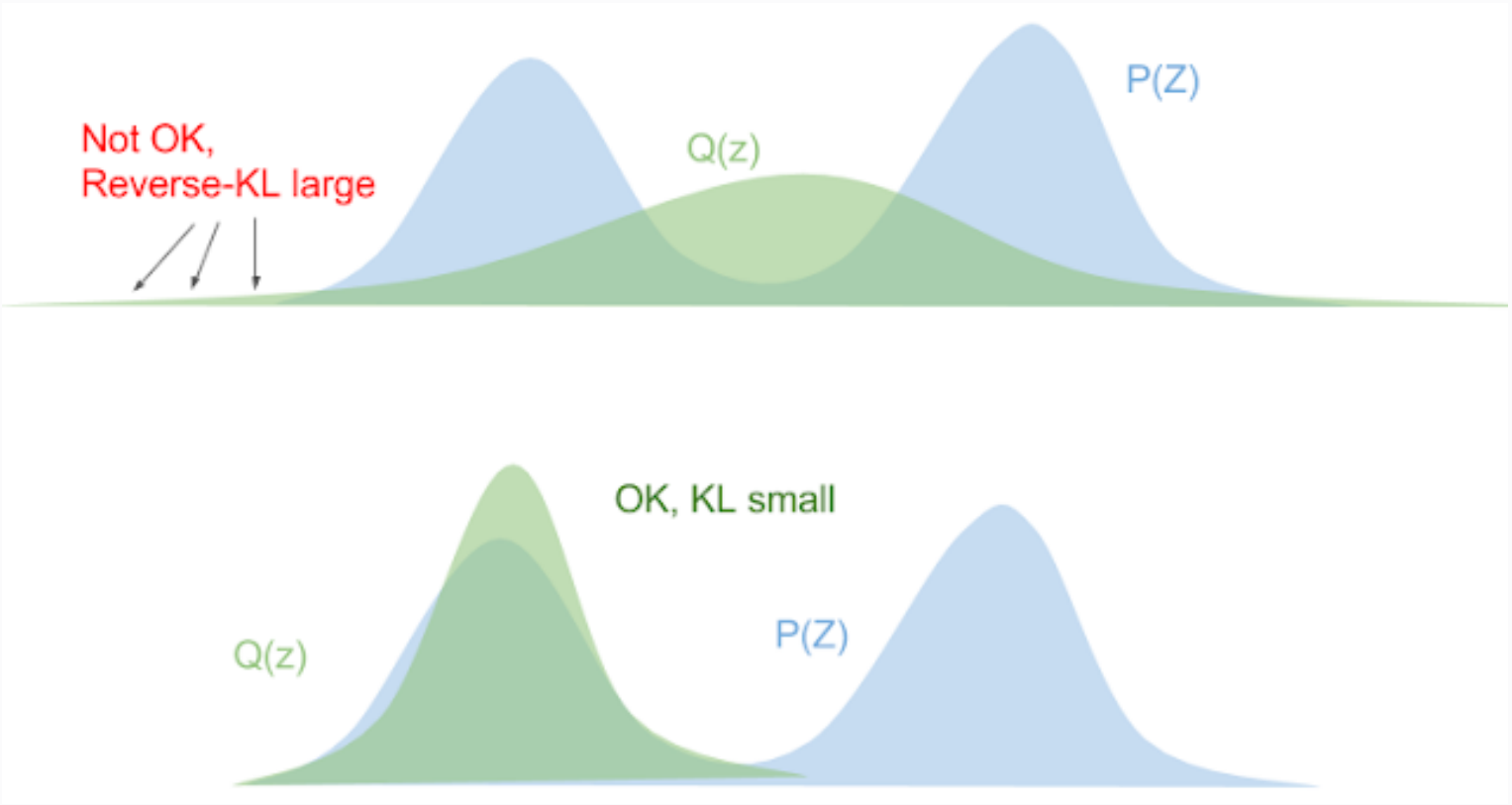


Image Credit: Eric Jang

Tail Behavior

- We can get insights into this happens by considering the cases $q(x) \to 0$ and $p(x) \to 0$, noting that $\lim_{x \to 0} x \log x = 0$
- If $q(x) = 0$ when $p(x) > 0$, then $q(x) \log(q(x)/p(x)) = 0$ and $p(x) \log(p(x)/q(x)) = \infty$
  - $\mathrm{KL}(p(x) \,\|\, q(x)) = \infty$ if $q(x) = 0$ anywhere $p(x) > 0$
  - $\mathrm{KL}(q(x) \,\|\, p(x))$ is still fine when this happens
  - By symmetry, $\mathrm{KL}(q(x) \,\|\, p(x))$ is problematic if $q(x) > 0$ anywhere $p(x) = 0$

Variational Inference

Variational Inference

- Variational inference (VI) methods are another class of ubiquitously used approaches for Bayesian inference wherein we try to learn an approximation to $p(\theta|\mathcal{D})$
- Key idea: reformulate the inference problem to an **optimization**, by learning parameters of a posterior approximation
- We do this through introducing a parameterized variational family $q_\phi(\theta),\ \phi \in \varphi$
- Then finding the $\phi^* \in \varphi$ that gives the "best" approximation
- Here "best" is based on minimizing $\mathrm{KL}(q \,\|\, p)$:

$$\phi^* = \arg\min_{\phi \in \varphi} \mathrm{KL}(q_\phi(\theta) \,\|\, p(\theta|\mathcal{D})) \tag{9}$$

The Variational Family

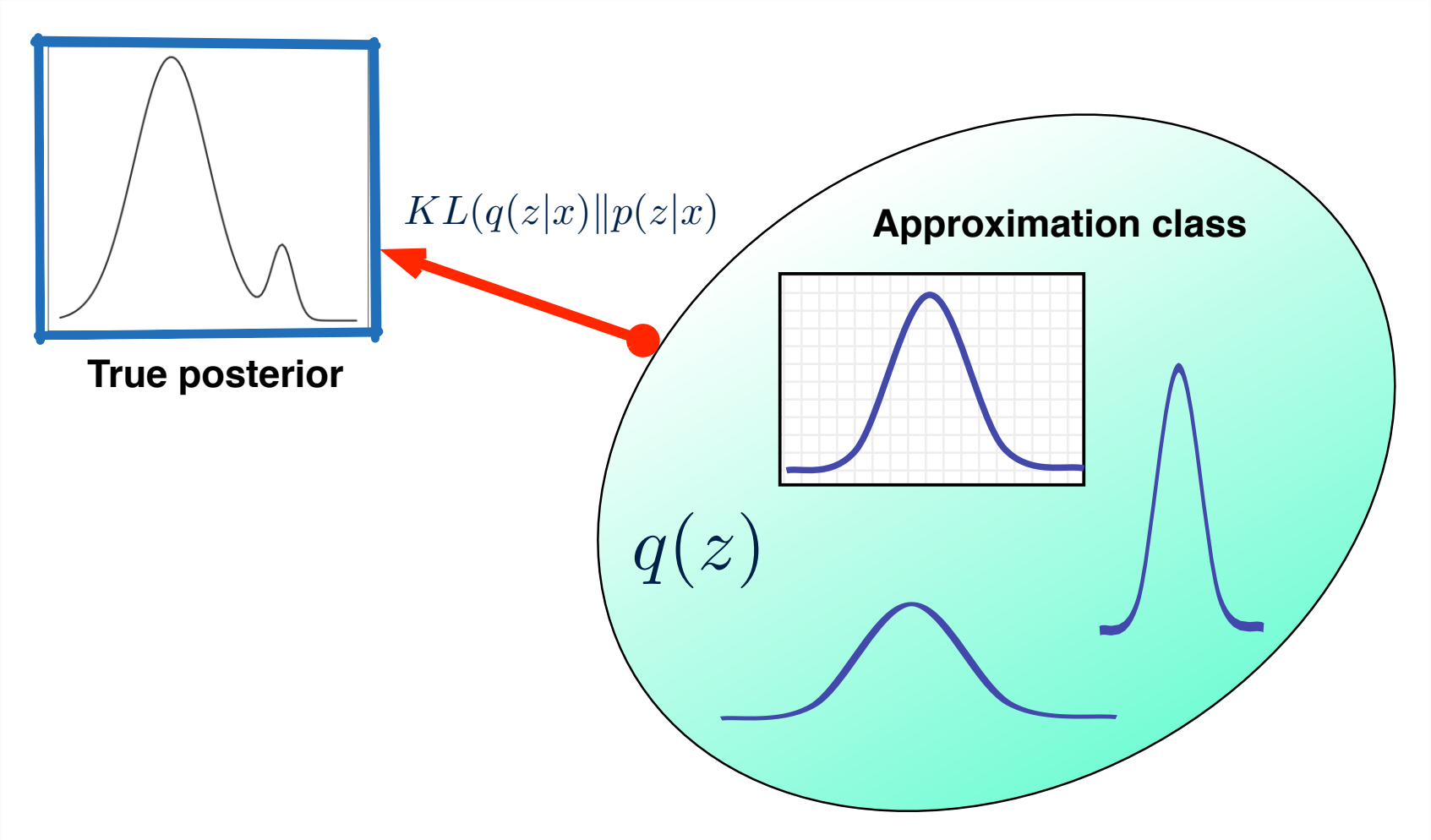
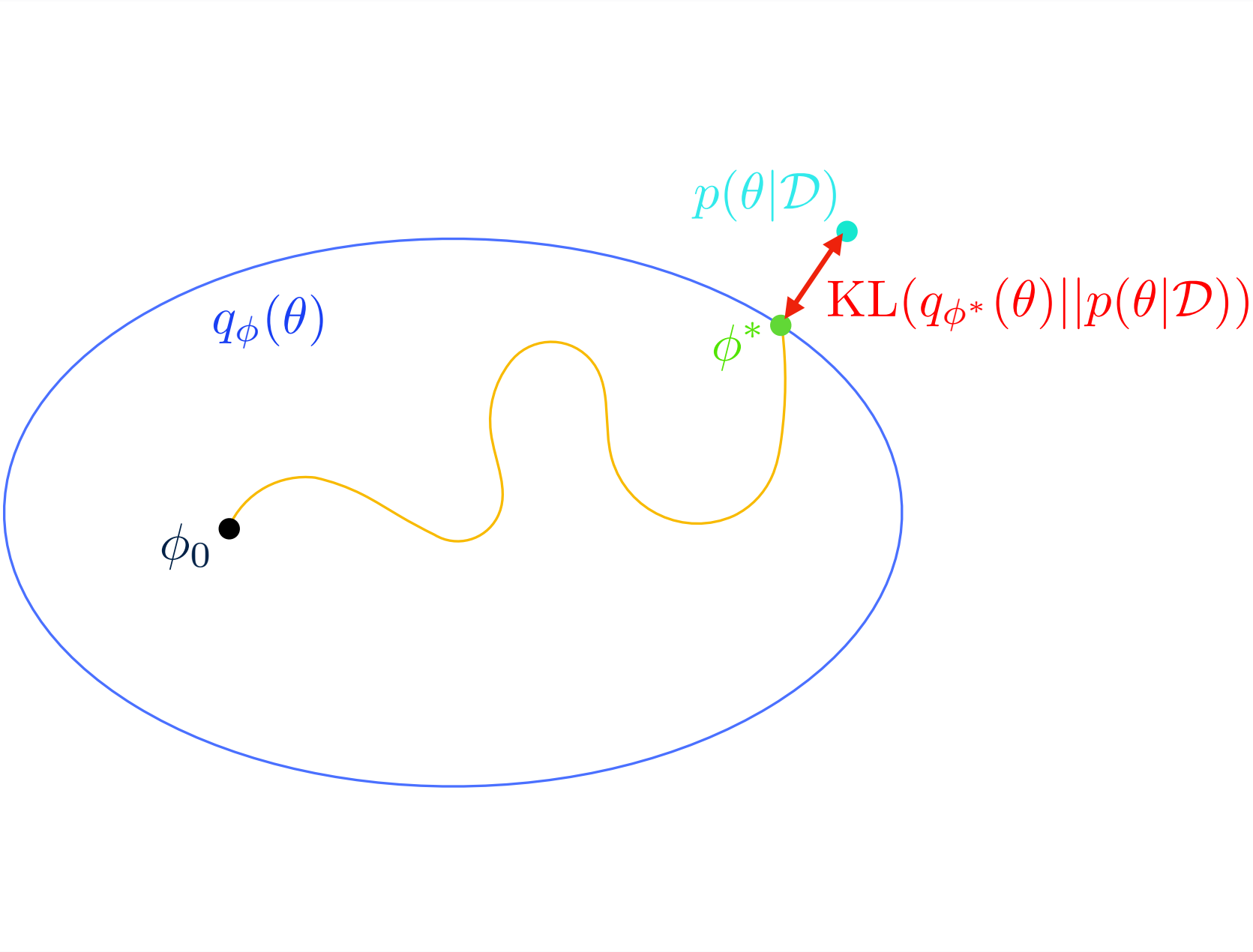


Image Credit: Shakir Mohamed

Variational Inference (2)



Variational Inference (3)

- We cannot work directly with $\mathrm{KL}(q_\phi(\theta) \,\|\, p(\theta|\mathcal{D}))$ because we don't know the posterior density
- We can note that the marginal likelihood $p(\mathcal{D})$ is independent of our variational parameters ϕ to work with the joint instead

$$\begin{aligned}
\phi^* &= \arg\min_{\phi \in \varphi} \mathrm{KL}(q_\phi(\theta) \,\|\, p(\theta|\mathcal{D})) \\
&= \arg\min_{\phi \in \varphi} \mathbb{E}_{q_\phi(\theta)} \left[\log \frac{q_\phi(\theta)}{p(\theta|\mathcal{D})} \right] \\
&= \arg\min_{\phi \in \varphi} \mathbb{E}_{q_\phi(\theta)} \left[\log \frac{q_\phi(\theta)}{p(\theta|\mathcal{D})} \right] - \log p(\mathcal{D}) \\
&= \arg\min_{\phi \in \varphi} \mathbb{E}_{q_\phi(\theta)} \left[\log \frac{q_\phi(\theta)}{p(\theta, \mathcal{D})} \right]
\end{aligned}$$

- This trick is why we work with $\mathrm{KL}(q_\phi(\theta) \,\|\, p(\theta|\mathcal{D}))$ rather than $\mathrm{KL}(p(\theta|\mathcal{D}) \,\|\, q_\phi(\theta))$: the latter is **doubly intractable**

The ELBO

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

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

$$\text{where} \quad \mathcal{L}(\phi) := \mathbb{E}_{q_\phi(\theta)} \left[\log \frac{p(\theta, \mathcal{D})}{q_\phi(\theta)} \right]$$

$$= \log p(\mathcal{D}) - \mathrm{KL}(q_\phi(\theta) \,\|\, p(\theta|\mathcal{D}))$$

is known as the **evidence lower bound** or **ELBO** for short

$\mathcal{L}(\phi)$ is also sometimes known as the **variational free energy**

The ELBO (2)

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

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(\theta)}\left[\log \frac{p(\theta, \mathcal{D})}{q_\phi(\theta)}\right] \leq \log \mathbb{E}_{q_\phi(\theta)}\left[\frac{p(\theta, \mathcal{D})}{q_\phi(\theta)}\right] = \log p(\mathcal{D})$$

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

for any $u_1, u_2 > 0$

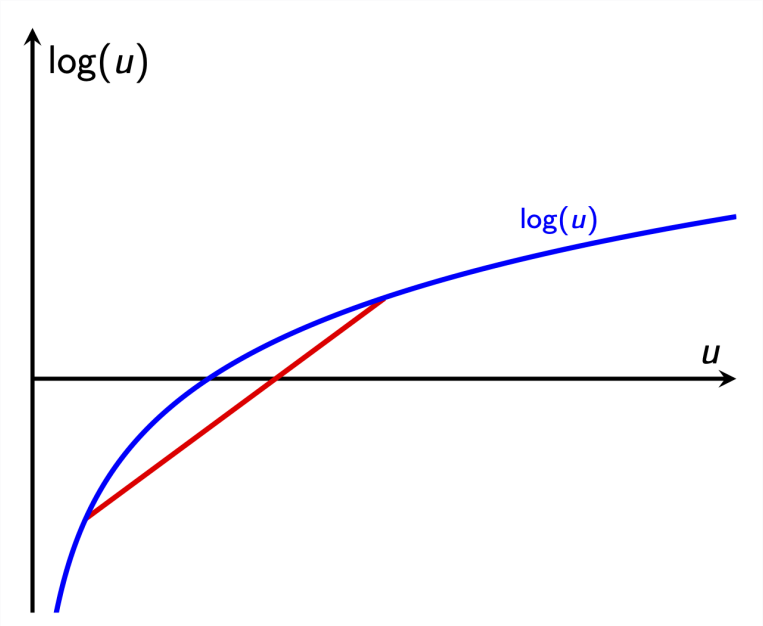
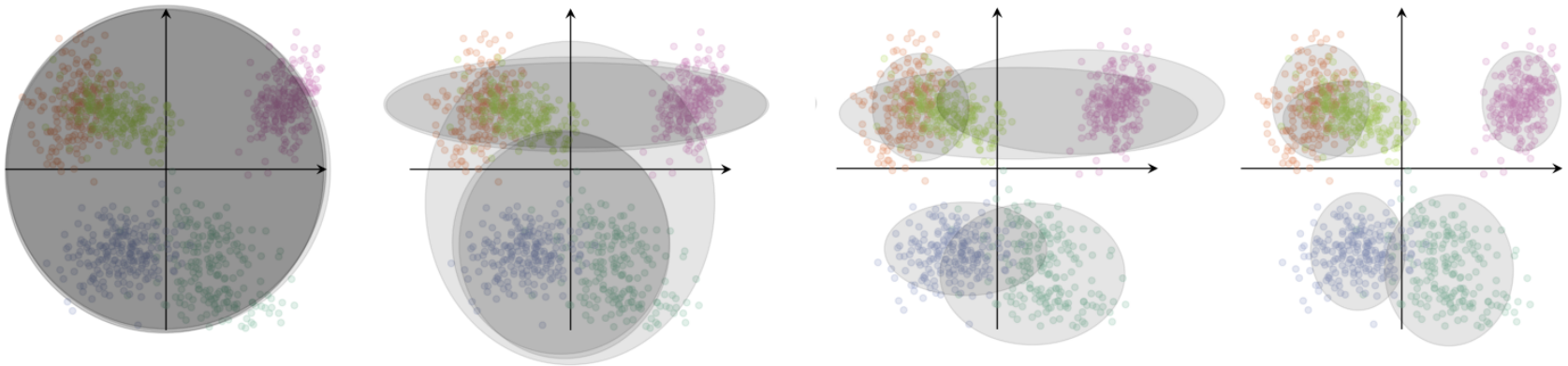
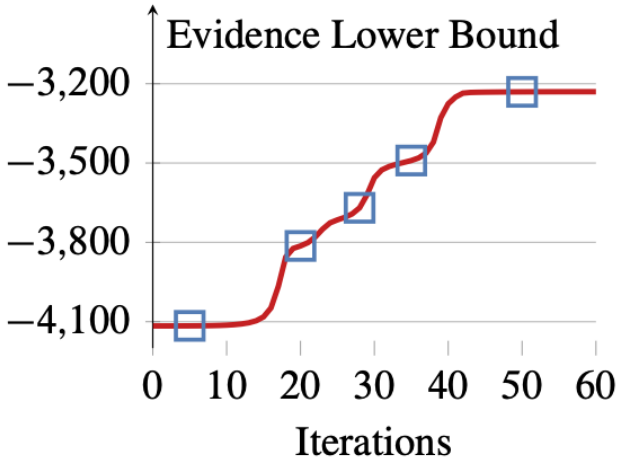


Image Credit: Michael Gutmann

Example: Mixture of Gaussians





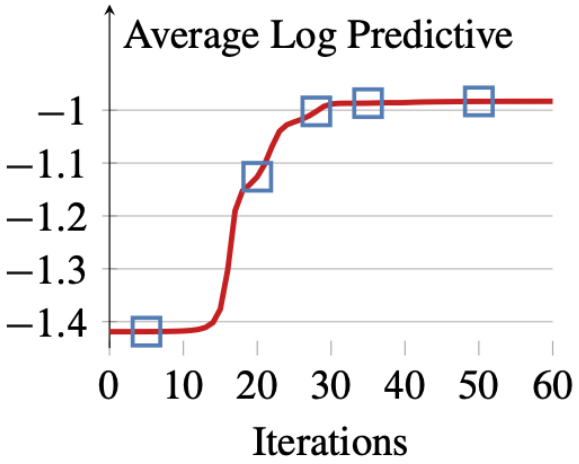


Image credit: Alp Kucukelbir

Worked Example—Gaussian with Unknown Mean and Variance

As a simple worked example (taken from Bishop 10.1.3), consider the following model where we are trying to infer to the mean μ and precision τ of a Gaussian given a set of observations $\mathcal{D} = \{x_n\}_{n=1}^N$.

Our full model is given by

$$p(\tau) = \textsc{Gamma}(\tau; \alpha, \beta)$$
$$p(\mu|\tau) = \mathcal{N}(\mu; \mu_0, (\lambda_0\tau)^{-1})$$
$$p(\mathcal{D}|\mu, \tau) = \prod_{n=1}^{N} \mathcal{N}(x_n; \mu, \tau^{-1})$$

Worked Example—Gaussian with Unknown Mean and Variance

We care about the posterior $p(\mu, \tau|\mathcal{D})$ and we are going to try and approximate this using variational inference

For our variational family we will take

$$\begin{aligned} q_\phi(\tau, \mu) &= q(\tau)q(\mu) \\ q_\phi(\tau) &= \textsc{Gamma}(\tau; \phi_a, \phi_b) \\ q_\phi(\mu) &= \mathcal{N}(\mu; \phi_c, \phi_d^{-1}) \end{aligned}$$

where we note that this factorization is an assumption: the posterior itself does not factorize

Worked Example—Gaussian with Unknown Mean and Variance

To find the best variational parameters ϕ^*, we need to optimize $\mathcal{L}(\phi)$, for which we can use gradient methods, using

$$\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi \iint q_\phi(\tau) q_\phi(\mu) \log \left(\frac{p(\mathcal{D}|\mu, \tau) p(\mu|\tau) p(\tau)}{q_\phi(\tau) q_\phi(\mu)} \right) d\tau d\mu$$

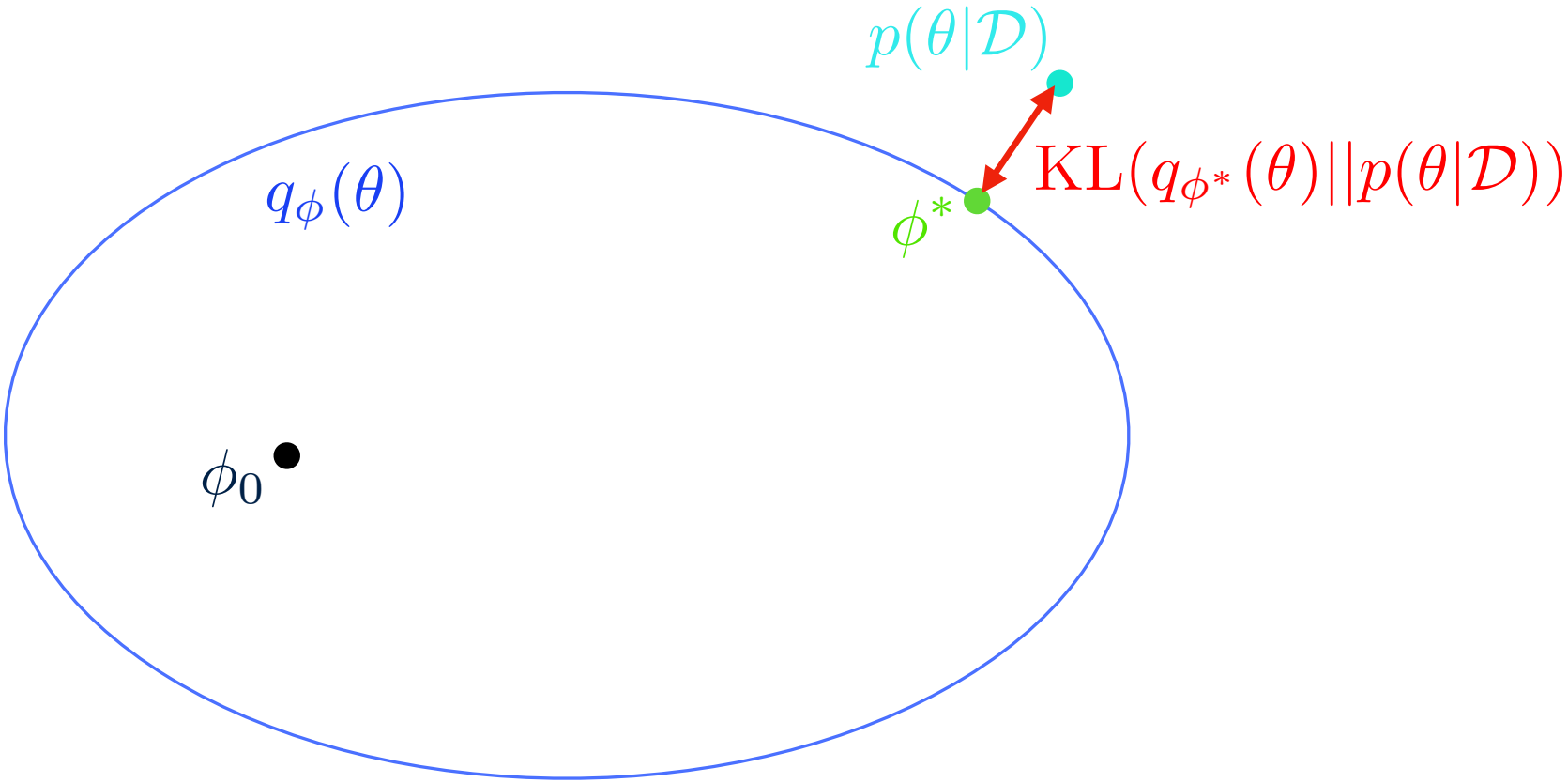
If we can calculate this gradient, this means we can optimize ϕ by performing gradient ascent.

After initializing some ϕ_0, we just repeatedly apply

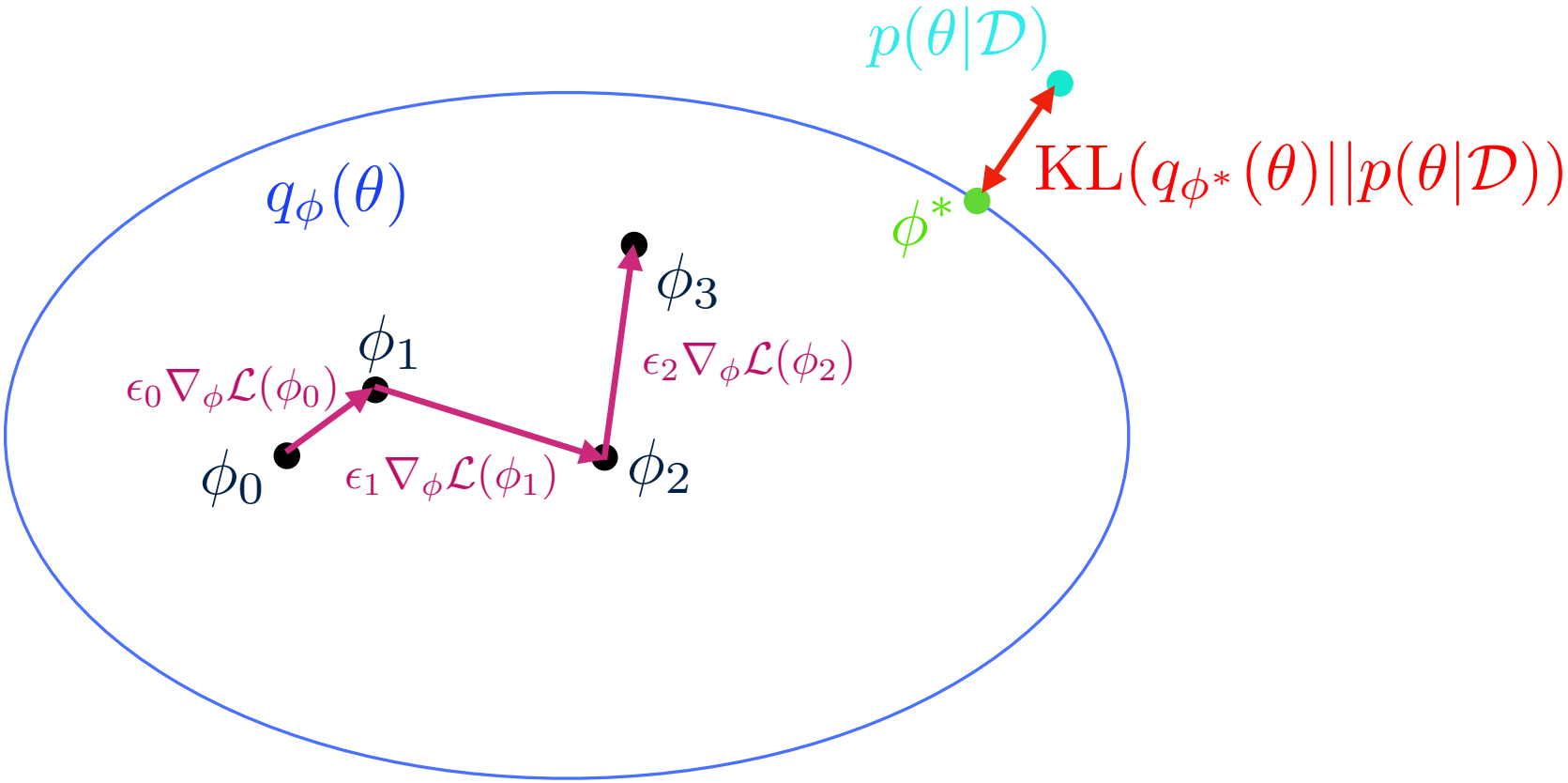
$$\phi_{n+1} \leftarrow \phi_n + \epsilon_n \nabla_\phi \mathcal{L}(\phi_n)$$

where ϵ_n are our step sizes

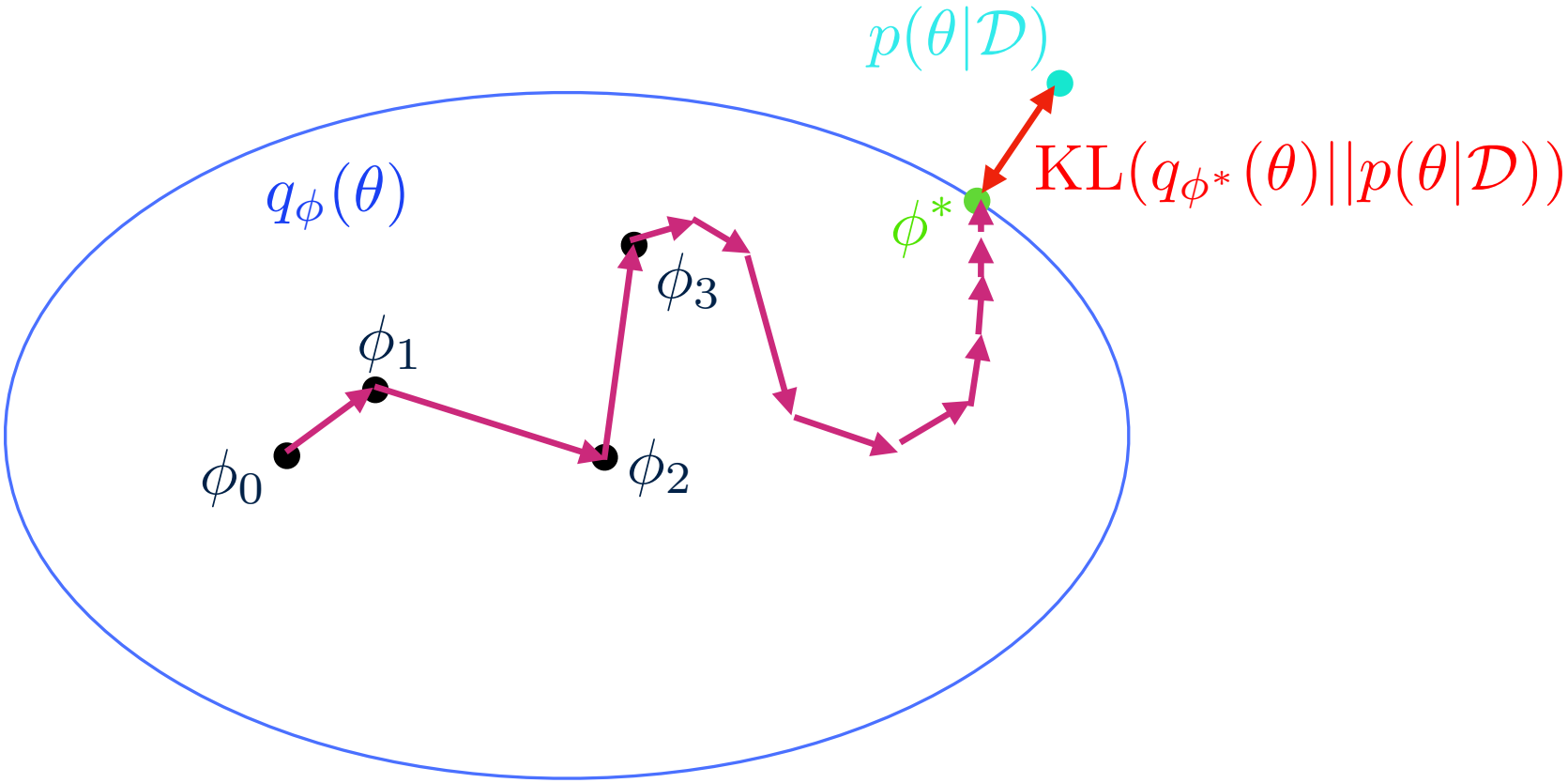
Gradient Updates of Variational Parameters



Gradient Updates of Variational Parameters



Gradient Updates of Variational Parameters



The Mean Field Assumption

- In this example we chose a factorized variational approximation: $q_\phi(\mu, \tau) = q_\phi(\mu) q_\phi(\tau)$
- This factorization assumption is actually a common assumption more generally called the **mean field** assumption
- Mathematically we can define this as

$$q_\phi(\theta) = \prod_j q_{j,\phi}(\theta_j) \tag{10}$$

- There are a number of scenarios where this can help make maximizing the ELBO more tractable
- However, it is a less necessary assumption than it used to be since the rise of AutoDiff and stochastic gradients methods

The Mean Field Assumption (2)

Perhaps the biggest upshot of the mean field assumption is that it gives a closed form solution for the optimal distribution of each $q_{j,\phi}(\theta_j)$ given the θ_i for $i \neq j$ (up to a normalization constant and assuming an otherwise constrained variational family):

$$q^*_{j,\phi}(\theta_j) \propto \exp\left(\mathbb{E}_{\prod_{i \neq j} q_{i,\phi}(\theta_i)}\left[\log p(\theta, \mathcal{D})\right]\right) \tag{11}$$

In some cases, this can be calculated analytically giving a gradientless coordinate ascent approach to optimizing the ELBO

However, it also tends to require restrictive conjugate distribution assumptions and so it is used much less often in modern approaches

Worked Example—Gaussian with Unknown Mean and Variance

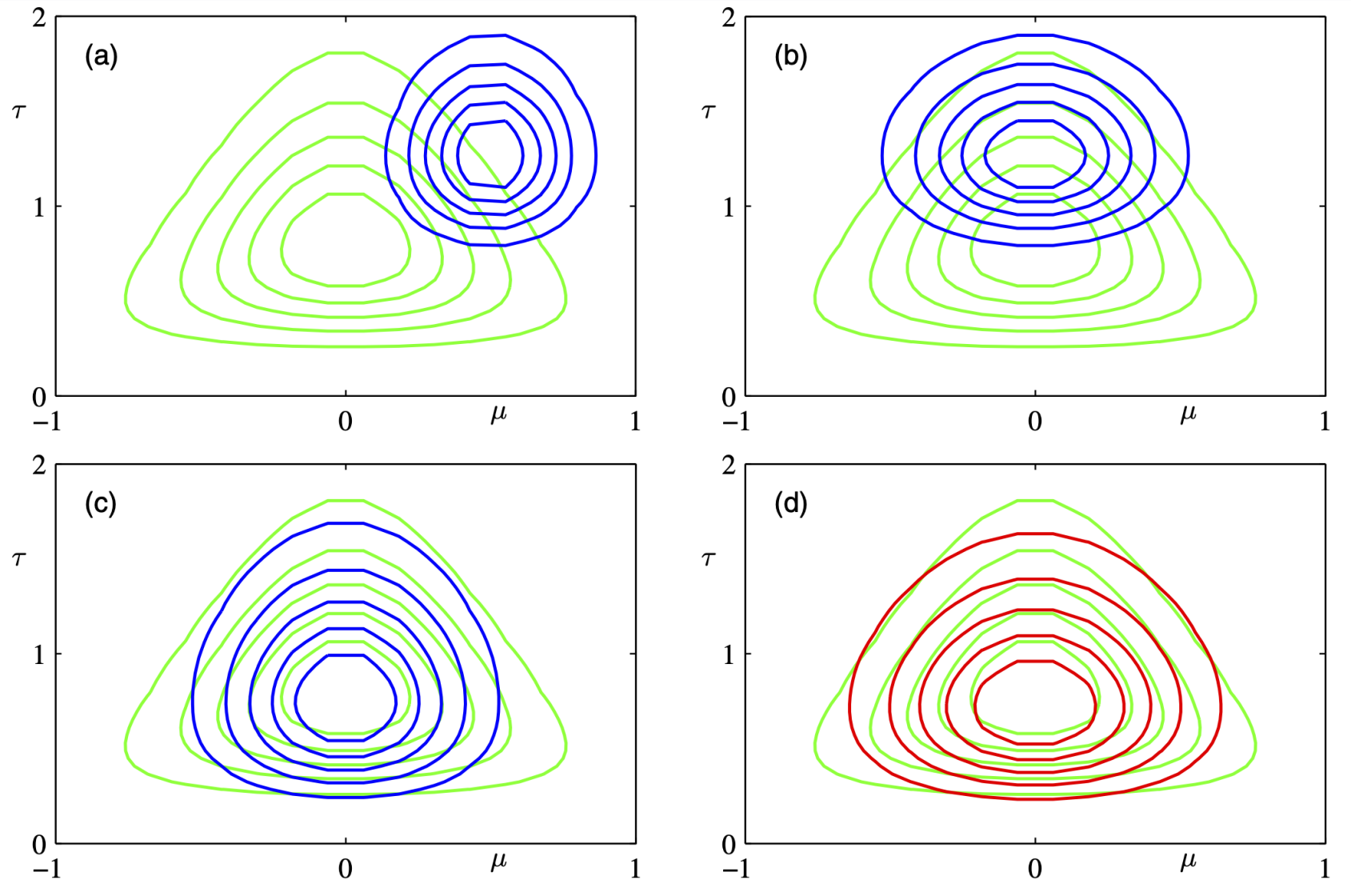


Figure 10.4 from Bishop

Pros and Cons of Variational Methods

Pros

- Typically more efficient than MCMC approaches, particularly in high dimensions once we exploit the stochastic variational approaches introduced in the next lecture
- Can often provided effective inference for models where MCMC methods have impractically slow convergence
- Though it is an approximation for the density, we can also sample directly from our variational distribution to calculate Monte Carlo estimates if needed
- Allows simultaneous optimization of model parameters as we will show in the next lecture

Pros and Cons of Variational Methods (2)

Cons

- But 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
- Often requires substantial tailoring to a particular problem
- Very difficult to estimate how much error their is in the approximation: subsequent estimates can be unreliable, particular in their uncertainty
- Tends to underestimate the variance of the posterior due to mode–seeking nature of reverse KL, particularly if using a mean field assumption

Further Reading

- The lecture notes give extra information on the curse of dimensionality and MCMC methods
- Iain Murray on MCMC
  `https://www.youtube.com/watch?v=_v4Eb09qp7Q`
- Chapters 21, 22, and 23 of K P Murphy. *Machine learning: a probabilistic perspective*. 2012
- David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* (2017)
- NeurIPS tutorial on variational inference that accompanies the previous paper:
  `https://www.youtube.com/watch?v=ogdv_6dbvVQ`