Introduction to Probabilistic Programming
Maria Han Veiga
AI in Science and Engineering Summer Academy 2023

About me

Fall 2023: (Incoming) Assistant Professor, Dpt. of Mathematics, OSU

2020 - now: Postdoctoral Fellow at MIDAS, UofM

2021 - 2023: Assistant Professor, Dpt. of Mathematics, UofM

2015 - 2019: PhD in Mathematics, University of Zurich

Interests:

Numerical analysis for PDEs/ODEs

Scientific Machine Learning

Reinforcement Learning

Session structure

Part 1: Theoretical concepts for Bayesian inference

1. Introduction to Bayesian inference
2. Exact inference and sampling
3. Approximate inference with variational inference

Part 2: Deep dive into existing programming frameworks

1. Revisiting examples
2. Pyro framework

Posterior inference

Data generating process
with unknown parameter(s) θ

Tossing a coin
Probability of 'head'

Observed data $\{d_i\}_{i=1}^{N}$

Outcomes of coin toss

Posterior inference

Data generating process with unknown parameter(s) θ	Tossing a coin Probability of ‘head’	Spread and prevalence of X virus Infection rate, recovery rate
Observed data $\{d_i\}_{i=1}^{N}$	Outcomes of coin toss	Number of infected patients

Posterior inference

Data generating process with unknown parameter(s) θ	Tossing a coin Probability of 'head'	Spread and prevalence of X virus Infection rate, recovery rate	Neural network Weights and biases
Observed data $\{d_i\}_{i=1}^{N}$	Outcomes of coin toss	Number of infected patients	Observed labels

Posterior inference

Data generating process
with unknown parameter(s) θ

Observed data $\{d_i\}_{i=1}^N$

	Tossing a coin	Spread and prevalence of X virus	Neural network
θ	Probability of 'head'	Infection rate, recovery rate	Weights and biases
$\{d_i\}_{i=1}^N$	Outcomes of coin toss	Number of infected patients	Observed labels

Question of interest:

Given a (model of a) data generating process and observed data, what are the parameters θ?

- We can perform point estimates of the parameters θ (e.g. Maximum Likelihood estimation)
  - Disadvantage: hard to come up with confidence intervals for the parameters
- Let the parameter be a random variable (RV) and describe the distribution of that RV

Intro to Bayesian Inference

What is Bayesian statistics?

Thomas Bayes (1701-1761)
What a BAyE!

- Bayesian statistics gives a way to integrate **prior information** with **data** to draw inferences

- Probabilities are **subjective measures of uncertainty**

- Data and parameters are represented by **random variables**

Basic set-up

- Data and parameters are represented by **random variables**. The data is observed, whereas the parameters are not.

Basic set-up

- Data and parameters are represented by **random variables**. The data is observed, whereas the parameters are not.

- A **model** $p(d \mid \theta)$ for the data generating process (also called **likelihood**) is specified. This process depends on some **unknown parameters** θ

Basic set-up

- Data and parameters are represented by **random variables**. The data is observed, whereas the parameters are not.
- A **model** $p(d \mid \theta)$ for the data generating process (also called **likelihood**) is specified. This process depends on some **unknown parameters** θ
- Information that we might have about the **unknown parameters** θ is represented by a **prior probability distribution** $p(\theta)$

Basic set-up

- Data and parameters are represented by **random variables**. The data is observed, whereas the parameters are not.

- A **model** $p(d \mid \theta)$ for the data generating process (also called **likelihood**) is specified. This process depends on some **unknown parameters** θ

- Information that we might have about the **unknown parameters** θ is represented by a **prior probability distribution** $p(\theta)$

- Bayesian inference uses *Bayes theorem* to combine the **prior** with the **observed data** to obtain a **posterior probability distribution for the parameters** $p(\theta \mid d)$.

Bayes' theorem:

Let $A, B \in \mathscr{F}$ such that $p(A),\ p(B) > 0$. The Bayes' theorem states

$$p(B \mid A) = \frac{p(B)p(A \mid B)}{p(A)}.$$

Bayes' theorem:

Let $A, B \in \mathscr{F}$ such that $p(A),\ p(B) > 0$. The Bayes' theorem states

$$p(B \mid A) = \frac{p(B)p(A \mid B)}{p(A)}.$$

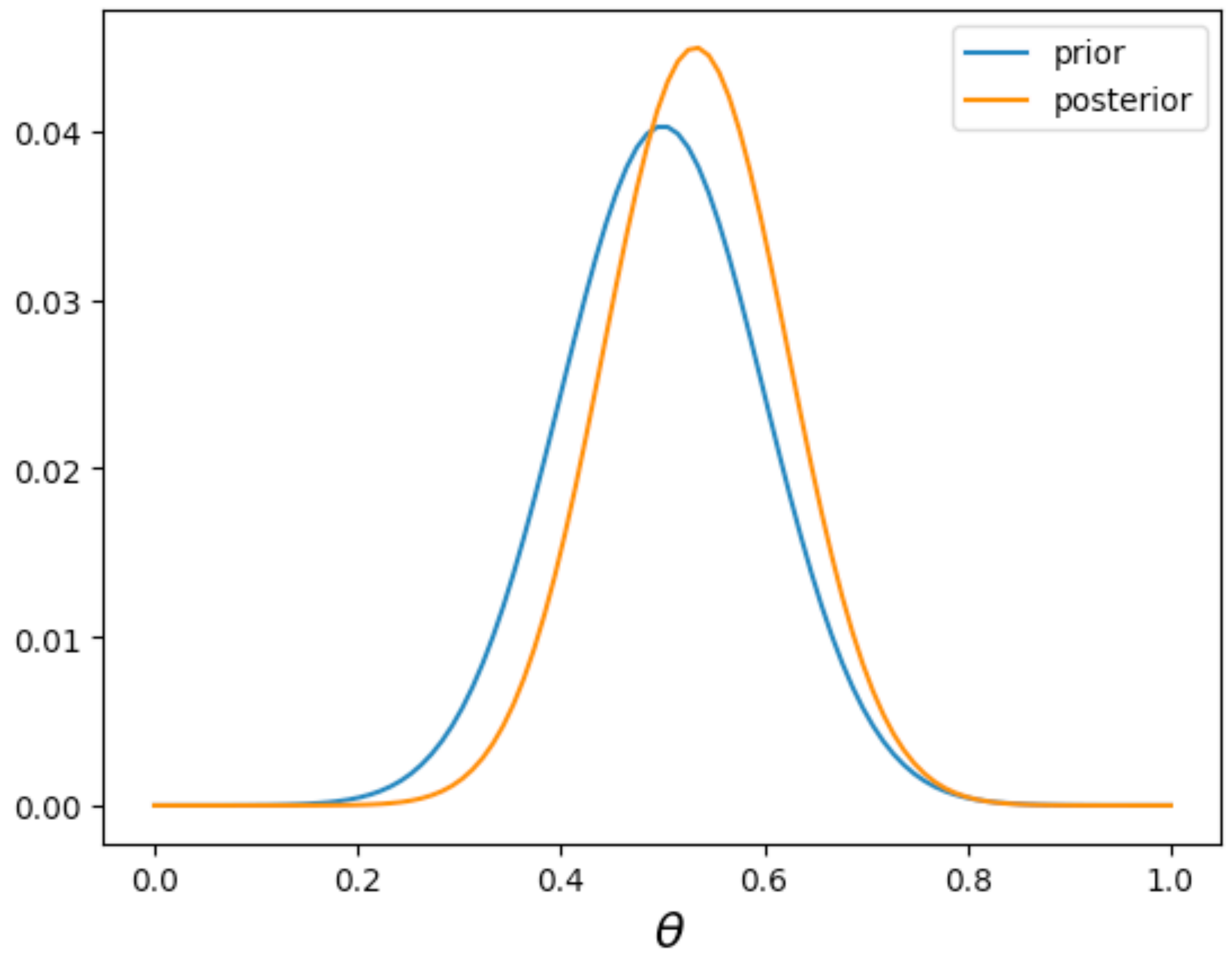
- In the context of Bayesian inference:
  - B represents your *a priori* beliefs of the world.
  - A is some observation related to that belief.
  - This tells us how to update our beliefs about B, given A (*a posteriori*)

- **Example:**
  - I want to estimate whether a coin is fair or not (probability of getting "Head" is my parameter θ)
    - My prior belief is that my coin is fair, e.g. $\theta \sim \mathcal{N}(0.5, 0.1)$
    - I observe the data d, which is the number of heads after 6 tosses.
      - The true data generating process is $d \sim Bin(6, \theta^*)$
      - The likelihood computes $p(d \mid \theta = 0.5)$

$$p(\theta \mid d) = \frac{p(\theta)p(d \mid \theta)}{p(d)}.$$

- **Example:**
  - I want to estimate whether a coin is fair or not (probability of getting "Head" is my parameter θ)
    - My prior belief is that my coin is fair, e.g. $\theta \sim \mathcal{N}(0.5, 0.1)$
    - I observe the data d, which is the number of heads after 6 tosses.
      - The true data generating process is $d \sim Bin(6, \theta^*)$
      - The likelihood computes $p(d \mid \theta = 0.5)$

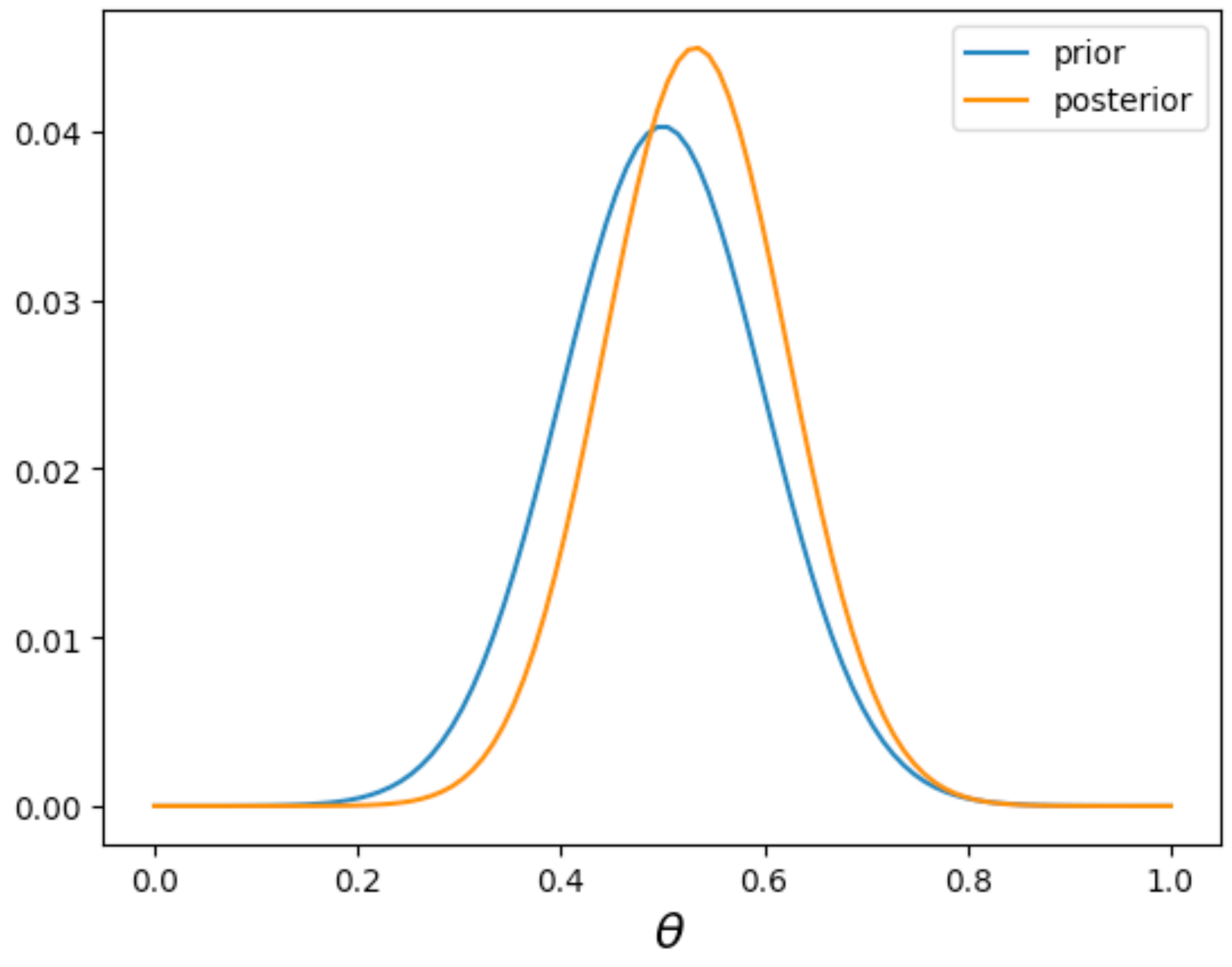
$$p(\theta \mid d) = \frac{p(\theta)p(d \mid \theta)}{p(d)}.$$

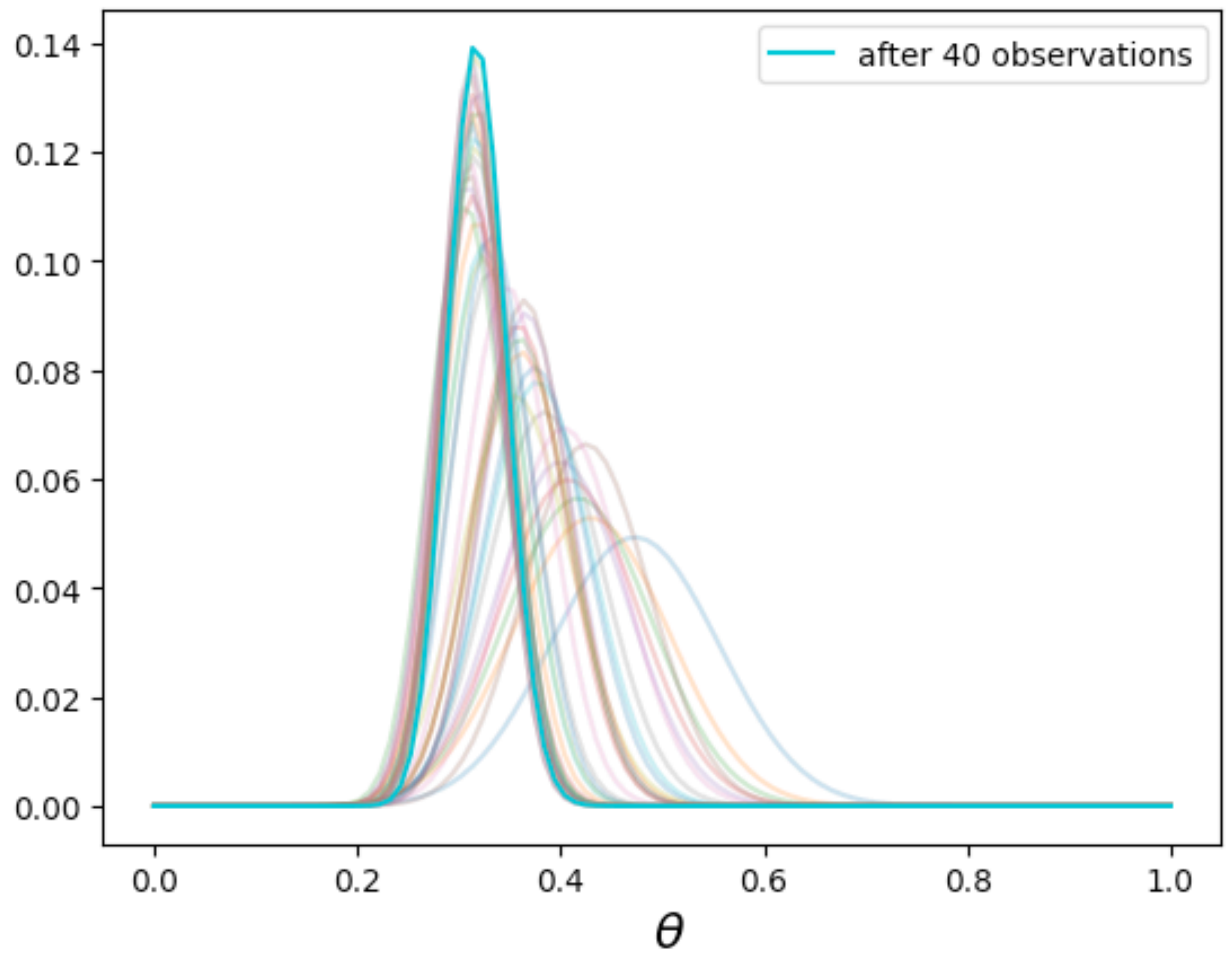


d = 4

- **Example:**
  - I want to estimate whether a coin is fair or not (probability of getting "Head" is my parameter θ)
    - My prior belief is that my coin is fair, e.g. $\theta \sim \mathcal{N}(0.5, 0.1)$
    - I observe the data d, which is the number of heads after 6 tosses.
      - The true data generating process is $d \sim Bin(6, \theta^*)$
      - The likelihood computes $p(d \mid \theta = 0.5)$

$$p(\theta \mid d) = \frac{p(\theta)p(d \mid \theta)}{p(d)}.$$

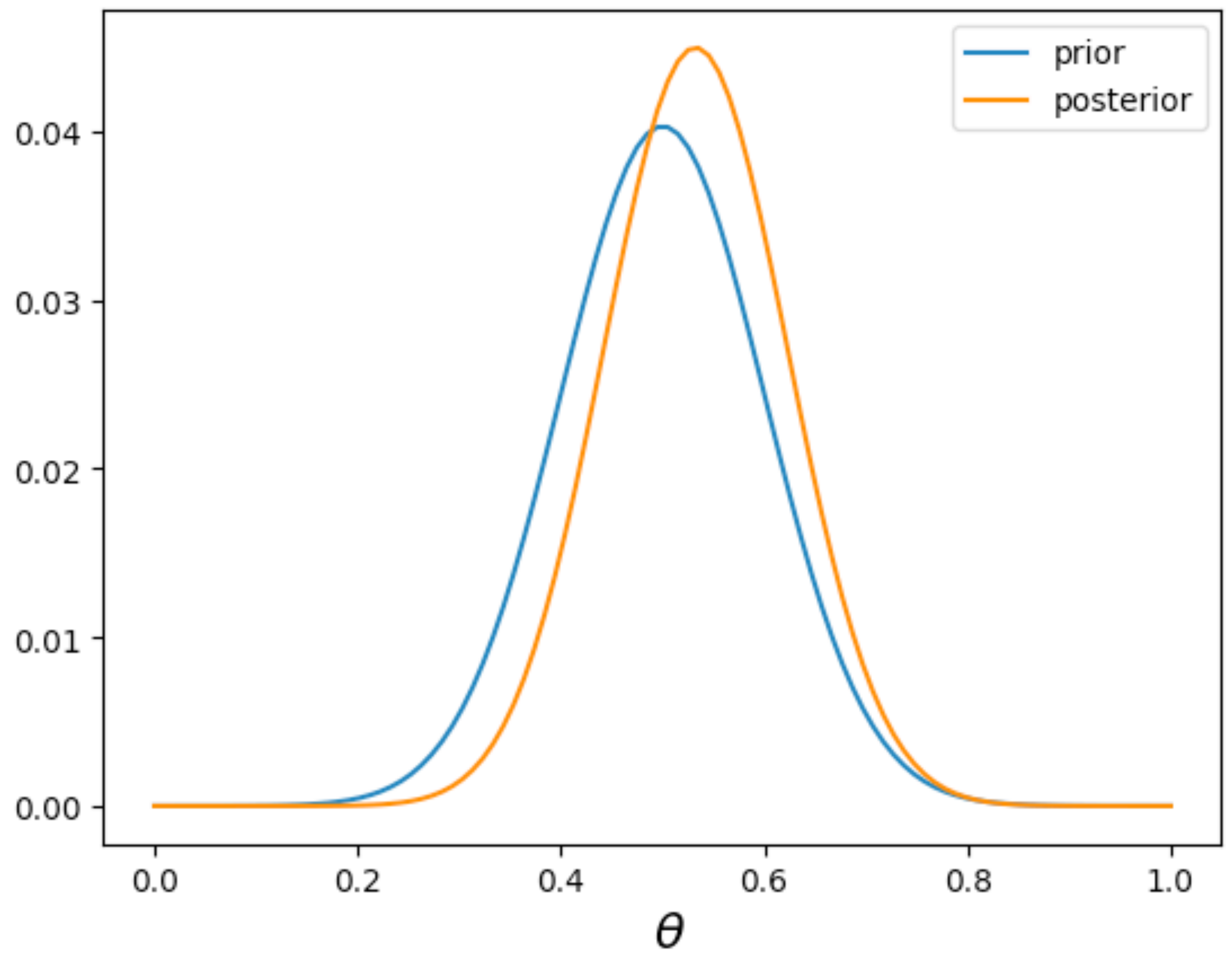




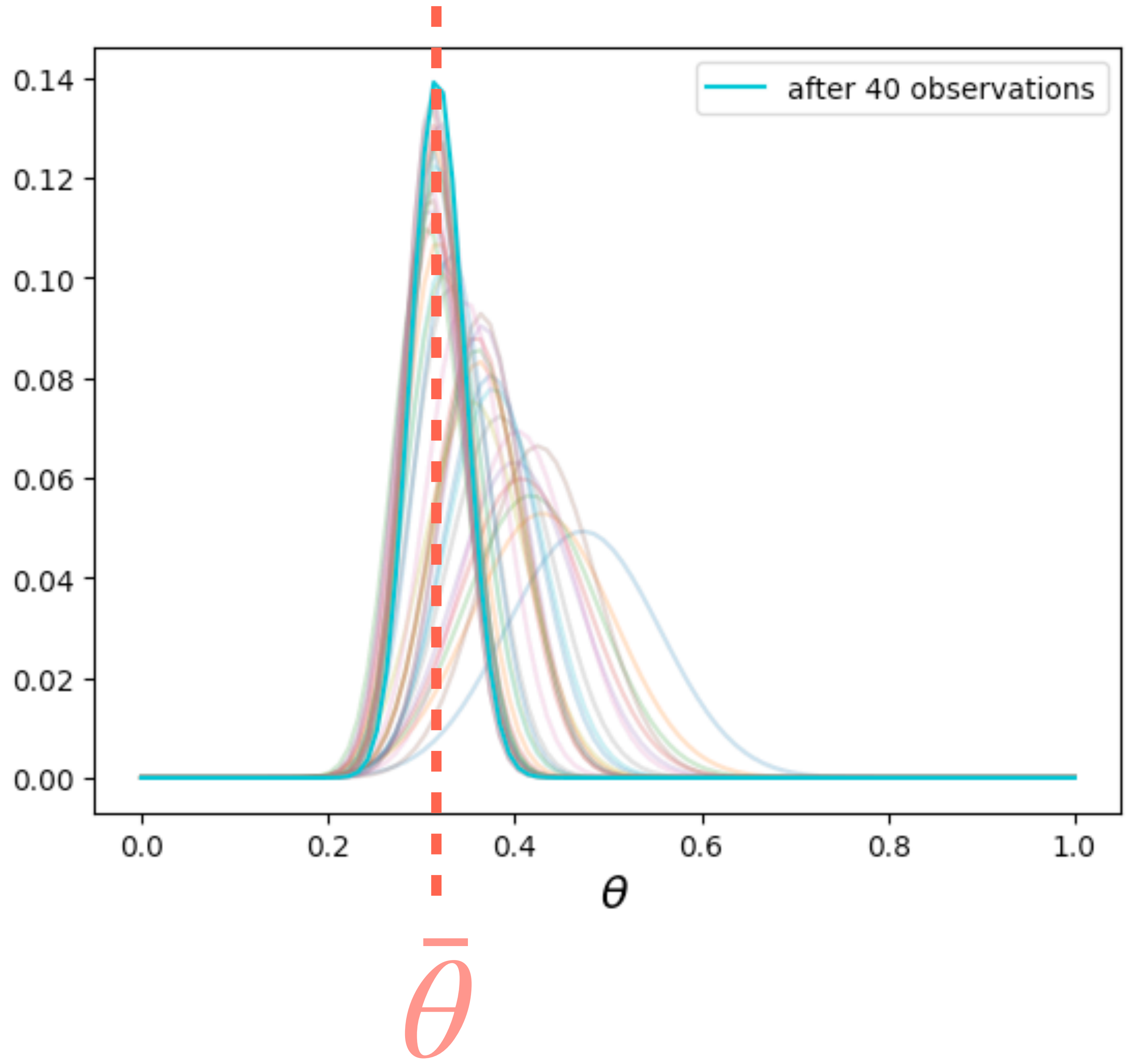
d = 4

- **Example:**
  - I want to estimate whether a coin is fair or not (probability of getting "Head" is my parameter θ)
    - My prior belief is that my coin is fair, e.g. $\theta \sim \mathcal{N}(0.5, 0.1)$
    - I observe the data d, which is the number of heads after 6 tosses.
      - The true data generating process is $d \sim Bin(6, \theta^*)$
      - The likelihood computes $p(d \mid \theta = 0.5)$

$$p(\theta \mid d) = \frac{p(\theta)p(d \mid \theta)}{p(d)}.$$



d = 4



$$\theta^* = 0.3 \approx \bar{\theta}$$

Wait a minute…

- What about the denominator $p(d)$?

Wait a minute…

- What about the denominator $p(d)$?
  - Assume θ is a discrete RV, then we can decompose it:
    - $p(d) = p(d\,|\,\theta)p(\theta) + p(d\,|\,\theta^c)p(\theta^c)$
    - We can compute $p(d)$ according to whether our beliefs are true or not, and the prior probability we assign to our beliefs.
    - If θ continuous, we must integrate over all possible θ. We will see this in general is a quantity that is intractable to compute in full generality…

Notation

- **Data** $d = (d_1, \ldots, d_n)$
- **True generating process** $f(\theta^*)$
- **Parameters** $\theta = (\theta_1, \cdots, \theta_m)$
- **Prior distribution** $p(\theta) = p(\theta_1, \cdots, \theta_m)$
- **Model** or **likelihood function** $p(d \,|\, \theta)$
- **Posterior distribution** $p(\theta \,|\, d)$

Notation

- **Data** $d = (d_1, \ldots, d_n)$ Observable
- **True generating process** $f(\theta^*)$
- **Parameters** $\theta = (\theta_1, \cdots, \theta_m)$
- **Prior distribution** $p(\theta) = p(\theta_1, \cdots, \theta_m)$
- **Model** or **likelihood function** $p(d \mid \theta)$
- **Posterior distribution** $p(\theta \mid d)$

Notation

- **Data** $d = (d_1, \ldots, d_n)$ Observable
- **True generating process** $f(\theta^*)$
- **Parameters** $\theta = (\theta_1, \cdots, \theta_m)$
- **Prior distribution** $p(\theta) = p(\theta_1, \cdots, \theta_m)$ Modelling choices
- **Model** or **likelihood function** $p(d \mid \theta)$
- **Posterior distribution** $p(\theta \mid d)$

Notation

- **Data** $d = (d_1, \ldots, d_n)$ Observable
- **True generating process** $f(\theta^*)$
- **Parameters** $\theta = (\theta_1, \cdots, \theta_m)$
- **Prior distribution** $p(\theta) = p(\theta_1, \cdots, \theta_m)$ Modelling choices
- **Model** or **likelihood function** $p(d \mid \theta)$
- **Posterior distribution** $p(\theta \mid d)$ Computed quantity of interest

Notation

- **Data** $d = (d_1, \ldots, d_n)$
- **True generating process** $f(\theta^*)$
- **Parameters** $\theta = (\theta_1, \cdots, \theta_m)$
- **Prior distribution** $p(\theta) = p(\theta_1, \cdots, \theta_m)$
- **Model** or **likelihood function** $p(d \mid \theta)$
- **Posterior distribution** $p(\theta \mid d)$

Observable

Modelling choices

Computed quantity of interest

Notation

- **Data** $d = (d_1, \ldots, d_n)$ Observable
- **True generating process** $f(\theta^*)$
- **Parameters** $\theta = (\theta_1, \cdots, \theta_m)$
- **Prior distribution** $p(\theta) = p(\theta_1, \cdots, \theta_m)$ Modelling choices
- **Model** or **likelihood function** $p(d \mid \theta)$
- **Posterior distribution** $p(\theta \mid d)$ Computed quantity of interest

Remark: We assumed the likelihood function and the true generating process are the same distribution, up to the parameter θ. In reality, we might don't know the function form of the true generating process, it might not even depend on parameters θ. This is called **model misspecification**.

Beyond parameter inference: posterior predictive

- Consider a new data sample $\tilde{d}$
- Find $p(\tilde{d} \mid d)$, the probability of the new data given our current data d:

$$p(\tilde{d} \mid d) = \int_{\Theta} p(\tilde{d} \mid \theta, d) p(\theta \mid d) \mathrm{d}\theta$$

$$= \int_{\Theta} p(\tilde{d} \mid \theta) p(\theta \mid d) \mathrm{d}\theta$$

(By independence of $\tilde{d}$ and d)

Beyond parameter inference: posterior predictive

- Consider a new data sample $\tilde{d}$
- Find $p(\tilde{d} \,|\, d)$, the probability of the new data given our current data d:

$$p(\tilde{d} \,|\, d) = \int_{\Theta} p(\tilde{d} \,|\, \theta, d) p(\theta \,|\, d) \mathrm{d}\theta$$

$$= \int_{\Theta} p(\tilde{d} \,|\, \theta) p(\theta \,|\, d) \mathrm{d}\theta$$

(By independence of $\tilde{d}$ and d)

Beyond parameter inference: posterior predictive

- Consider a new data sample $\tilde{d}$
- Find $p(\tilde{d} \mid d)$, the probability of the new data given our current data d:

$$p(\tilde{d} \mid d) = \int_{\Theta} p(\tilde{d} \mid \theta, d) p(\theta \mid d) \mathrm{d}\theta$$

$$= \int_{\Theta} p(\tilde{d} \mid \theta) p(\theta \mid d) \mathrm{d}\theta$$

(By independence of $\tilde{d}$ and d)

- $p(\tilde{d} \mid d)$ is the **posterior predictive distribution** and it can be used to:
  - Forecast
  - Check model (likelihood function) correctness: if the data we did observe follows this pattern closely, it indicates we chose our model / likelihood and prior well.

How to solve Bayesian inference problems?

- Exactly
- Through sampling
- Approximately

Exact inference & Sampling

Exact inference

Recall Bayes' theorem: $p(\theta \,|\, X = d) = \dfrac{p(X = d \,|\, \theta) \times p(\theta)}{p(X = d)}$

Exact inference

Recall Bayes' theorem: $p(\theta \,|\, X = d) = \dfrac{p(X = d \,|\, \theta) \times p(\theta)}{p(X = d)}$

Computing the denominator:

$$p(X = d) = \int_{\Theta} p(X = d \,|\, \theta) \times p(\theta) \mathrm{d}\theta$$

is not always straightforward:

- Generally solve integral approximately
- If $\vec{\theta} = (\theta_1, \ldots, \theta_n)$, integrate over $n-$dimensional parameter space

$\implies$ computationally intractable

Exact inference

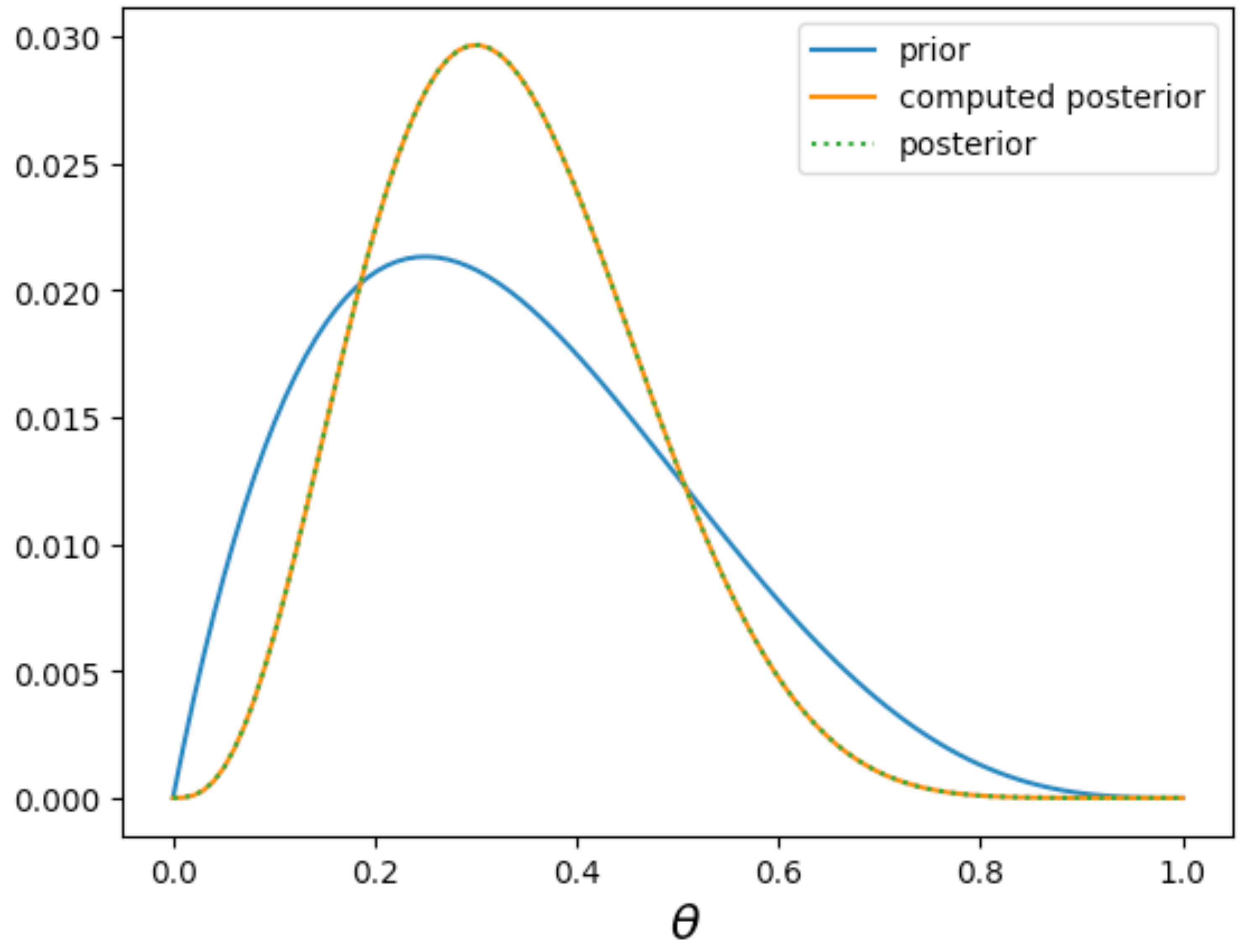
- In some case, we can write a closed-form expression for the posterior using conjugate priors
- For some likelihood functions, there exists a prior such that the posterior is the same as the prior (up to parameters)

Example:

Likelihood function $p(\mathbf{x}\|\theta)$	Model parameters θ	Conjugate Prior $p(\theta)$	Posteriori $p(\theta\|\mathbf{x})$
Gaussian	μ (mean)	Gaussian	Gaussian
Gaussian	σ^2 (variance)	Inverse Gamma	Inverse Gamma
Exponential	λ (rate)	Gamma	Gamma
Binomial	p (success prob.)	Beta	Beta
Geometric	p (success prob.)	Beta	Beta
Poisson	λ (mean)	Gamma	Gamma

Coin example

- Let the prior $p(\theta)$ be given by a Beta distribution Beta(α_0, β_0)
- The likelihood is again $d \sim Bin(6,\theta^*)$
- Let observed data be: d = 2 (2 heads out of 6 tosses)
- Posterior is also a Beta distribution Beta$(\alpha_0 + d, \beta_0 + 6 - d)$

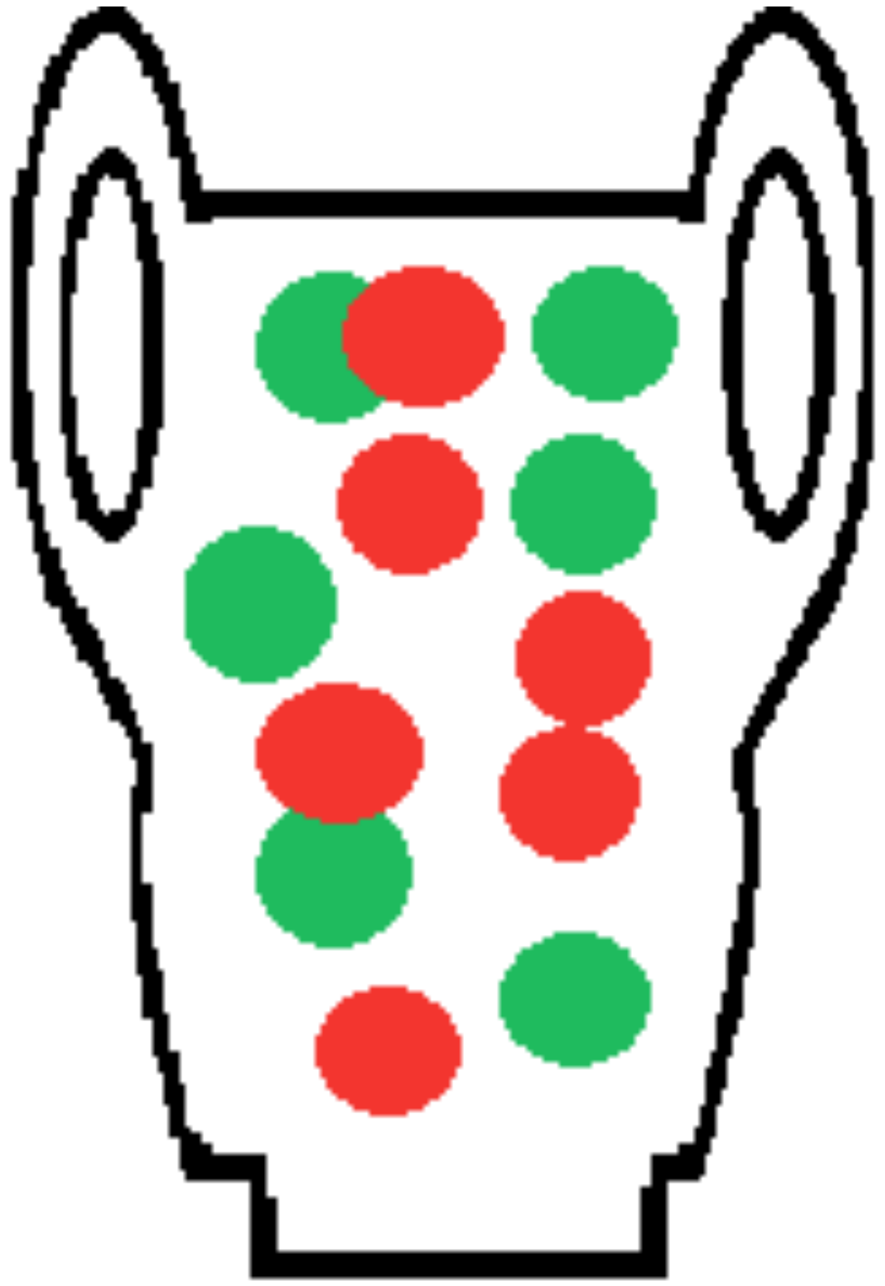


Exact inference

- **Disadvantage:**
  - At most 1-dimensional or 2-dimensional
  - Rigid form for the prior and likelihood
  - Not useful for general prior/likelihood choices and high-dimensional problems

Ice breaker: What problems in your research you could use these ideas?

Sampling



Idea:

- Draw independent samples from this urn
- By sampling we can characterise the distribution of the ball distribution

Question:

- If we can't compute $p(\theta \mid d)$ explicitly, can we *sample* from it, to then characterise the posterior? How?

Characterising the posterior through sampling

- Sampling from $p(\theta\,|\,d)$ is difficult. What if all we can do is evaluate something related to $p(\theta\,|\,d)$? Namely:

$$p(\theta\,|\,d) \propto p(d\,|\,\theta) \times p(\theta)$$

- (Handwavy) Let $p(\theta\,|\,d)$ be our target distribution, we can use a candidate distribution $w(\theta)$ that is easy to handle to help with the sampling

Characterising the posterior through sampling

- **Markov Chain Monte Carlo** methods are a class of algorithms to sample from a probability distribution.
- We need a few key concepts to generally understand the algorithm.

Markov Chain

- A stochastic process $X = \{X_n : n \geq 0\}$ is a Markov chain if for any state j:

$$P(X_{n+1} = j \,|\, X_n, \cdots, X_0) = P(X_{n+1} = j \,|\, X_n)$$

- $P(X_{n+1} = j \,|\, X_n = i) = p_{ij}$ denotes the transition probability of passing from state i to state j.
- Let P denote the transition probabilities matrix
- π_n denotes the state distribution in the n step

Stationary distribution

- The probability distribution of states evolves as $\pi_1 = P\pi_0$, and so on…
- Let $P\pi* = \pi*$. Then $\pi*$ is the stationary distribution of the Markov Chain.

Stationary distribution

- The probability distribution of states evolves as $\pi_1 = P\pi_0$, and so on…
- Let $P\pi^* = \pi^*$. Then π^* is the stationary distribution of the Markov Chain.

The basic limit theorem for Markov chains, under some assumptions, gives:

$$||\pi^* - \pi_n|| \to 0, \quad n \to \infty$$

Stationary distribution

- The probability distribution of states evolves as $\pi_1 = P\pi_0$, and so on…
- Let $P\pi^* = \pi^*$. Then π^* is the stationary distribution of the Markov Chain.

The basic limit theorem for Markov chains, under some assumptions, gives:

$$||\pi^* - \pi_n|| \to 0, \quad n \to \infty$$

No matter where we start the Markov Chain, π_n will eventually approach the stationary distribution π^*.

Stationary distribution

- The probability distribution of states evolves as $\pi_1 = P\pi_0$, and so on…
- Let $P\pi^* = \pi^*$. Then π^* is the stationary distribution of the Markov Chain.

The basic limit theorem for Markov chains, under some assumptions, gives:

$$||\pi^* - \pi_n|| \to 0, \quad n \to \infty$$

No matter where we start the Markov Chain, π_n will eventually approach the stationary distribution π^*.

Key idea: Let this stationary distribution π^* the target distribution

Markov Chain Monte Carlo

Metropolis-Hastings algorithm (1953):

- Let $w(\theta\,|\,\theta')$ be the transition density and $p(\theta\,|\,d)$ the target density
- Given state θ, sample a candidate value $\theta' \sim w(\theta'\,|\,\theta)$
- Compute the acceptance ratio:

$$\alpha(\theta'\,|\,\theta) = \min\left\{\frac{p(\theta'\,|\,d)w(\theta\,|\,\theta')}{p(\theta\,|\,d)w(\theta'\,|\,\theta)},1\right\}$$

- Sample $u \sim U(0,1)$. If $u \leq \alpha(\theta'\,|\,\theta)$, then the next state is equal to $\theta_{n+1} = \theta'$. Otherwise, the next state remains θ_n.

Markov Chain Monte Carlo

Metropolis-Hastings algorithm (1953):

- Let $w(\theta \mid \theta')$ be the transition density and $p(\theta \mid d)$ the target density
- Given state θ, sample a candidate value $\theta' \sim w(\theta' \mid \theta)$
- Compute the acceptance ratio:

$$\alpha(\theta' \mid \theta) = \min \left\{ \frac{p(\theta' \mid d) w(\theta \mid \theta')}{p(\theta \mid d) w(\theta' \mid \theta)}, 1 \right\}$$

- Sample $u \sim U(0,1)$. If $u \leq \alpha(\theta' \mid \theta)$, then the next state is equal to $\theta_{n+1} = \theta'$. Otherwise, the next state remains θ_n.

If $\alpha(\theta' \mid \theta)$ is symmetric, and plugging in the definition of the posterior, we have:

$$\alpha(\theta' \mid \theta) = \min \left\{ \frac{p(d \mid \theta') p(\theta')}{p(d \mid \theta) p(\theta)}, 1 \right\}$$

Markov Chain Monte Carlo

Metropolis-Hastings algorithm (1953):

- Let $w(\theta\,|\,\theta')$ be the transition density and $p(\theta\,|\,d)$ the target density
- Given state θ, sample a candidate value $\theta' \sim w(\theta'\,|\,\theta)$
- Compute the acceptance ratio:

$$\alpha(\theta'\,|\,\theta) = \min\left\{\frac{p(\theta'\,|\,d)w(\theta\,|\,\theta')}{p(\theta\,|\,d)w(\theta'\,|\,\theta)}, 1\right\}$$

- Sample $u \sim U(0,1)$. If $u \leq \alpha(\theta'\,|\,\theta)$, then the next state is equal to $\theta_{n+1} = \theta'$. Otherwise, the next state remains θ_n.

If $\alpha(\theta'\,|\,\theta)$ is symmetric, and plugging in the definition of the posterior, we have:

$$\alpha(\theta'\,|\,\theta) = \min\left\{\frac{p(d\,|\,\theta')p(\theta')}{p(d\,|\,\theta)p(\theta)}, 1\right\}$$

The $p(d)$ term cancels out!

Markov Chain Monte Carlo

Metropolis-Hastings algorithm (1953):

- Let $w(\theta\,|\,\theta')$ be the transition density and $p(\theta\,|\,d)$ the target density
- Given state θ, sample a candidate value $\theta' \sim w(\theta'\,|\,\theta)$
- Compute the acceptance ratio:

$$\alpha(\theta'\,|\,\theta) = \min\left\{\frac{p(\theta'\,|\,d)w(\theta\,|\,\theta')}{p(\theta\,|\,d)w(\theta'\,|\,\theta)},1\right\}$$

- Sample $u \sim U(0,1)$. If $u \leq \alpha(\theta'\,|\,\theta)$, then the next state is equal to $\theta_{n+1} = \theta'$. Otherwise, the next state remains θ_n.

If $\alpha(\theta'\,|\,\theta)$ is symmetric, and plugging in the definition of the posterior, we have:

$$\alpha(\theta'\,|\,\theta) = \min\left\{\frac{p(d\,|\,\theta')p(\theta')}{p(d\,|\,\theta)p(\theta)},1\right\}$$

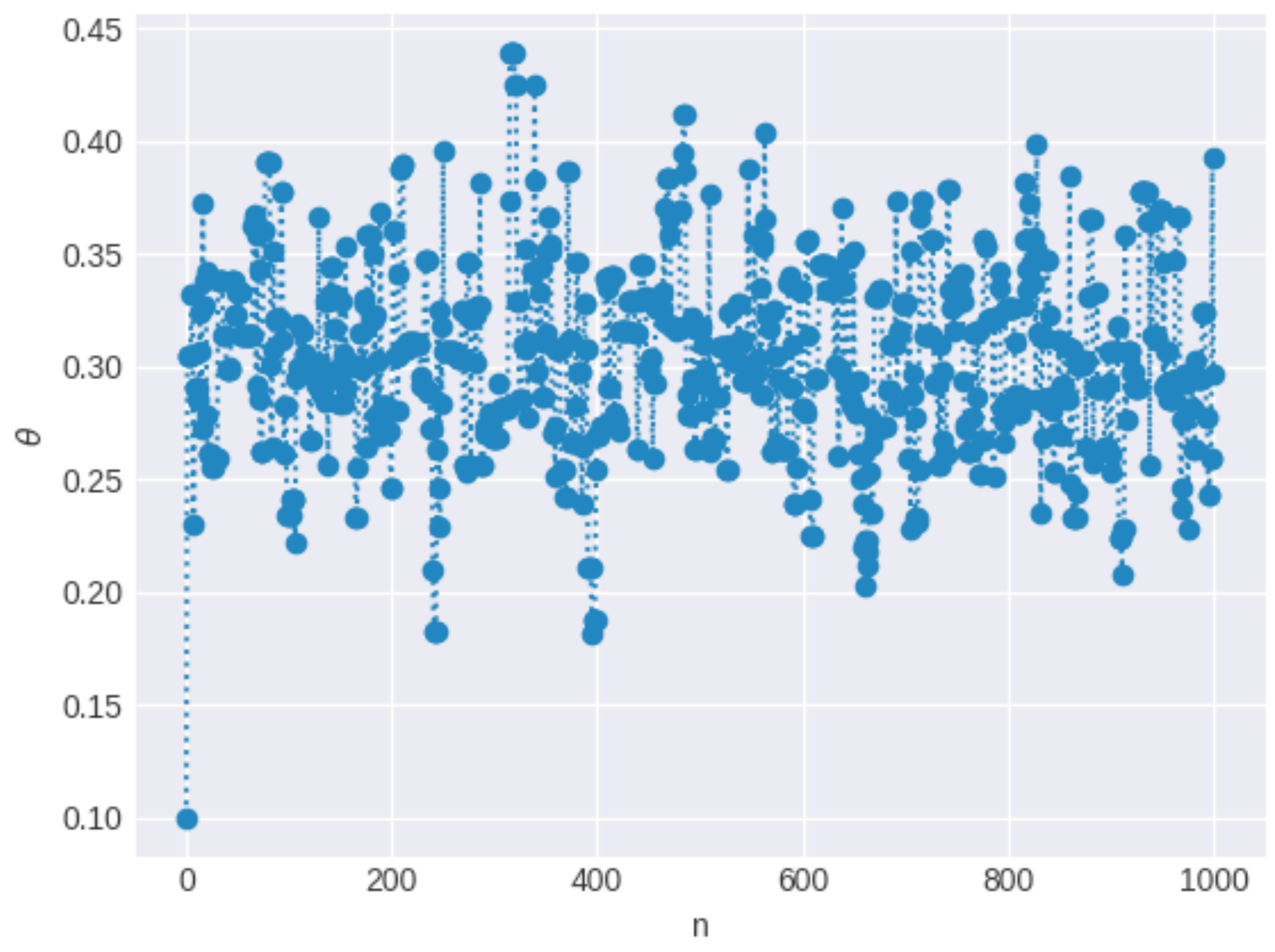
The $p(d)$ term cancels out!

We sample from likelihood x prior, the unnormalised posterior

Markov Chain Monte Carlo

- **The Metropolis-Hastings algorithm:** a way to obtain a sequence of random samples from a probability distribution with some density $p(x)$ while knowing only some function proportional to it: we only know $f(x) \propto p(x)$
- In the context of posterior estimation, allows us to sample from the **unnormalised posterior:** $p(d \mid \theta) \times p(\theta)$

Example



Again, let's look at the coin flip:

- Prior $p(\theta) \sim Beta(10,10)$
- Let $\theta' = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,0.1)$
- Then, $w(\theta' \mid \theta)$ is given by the distribution of ε
- Acceptance ratio:

$$\alpha(\theta' \mid \theta) = \min\left\{\frac{p(d \mid \theta')p(\theta')}{p(d \mid \theta)p(\theta)}, 1\right\}$$

  (symmetry of ε)
- $u \sim U(0,1)$
- If $u < \alpha$, $\theta_{n+1} = \theta'$, else $\theta_{n+1} = \theta_n$

Example

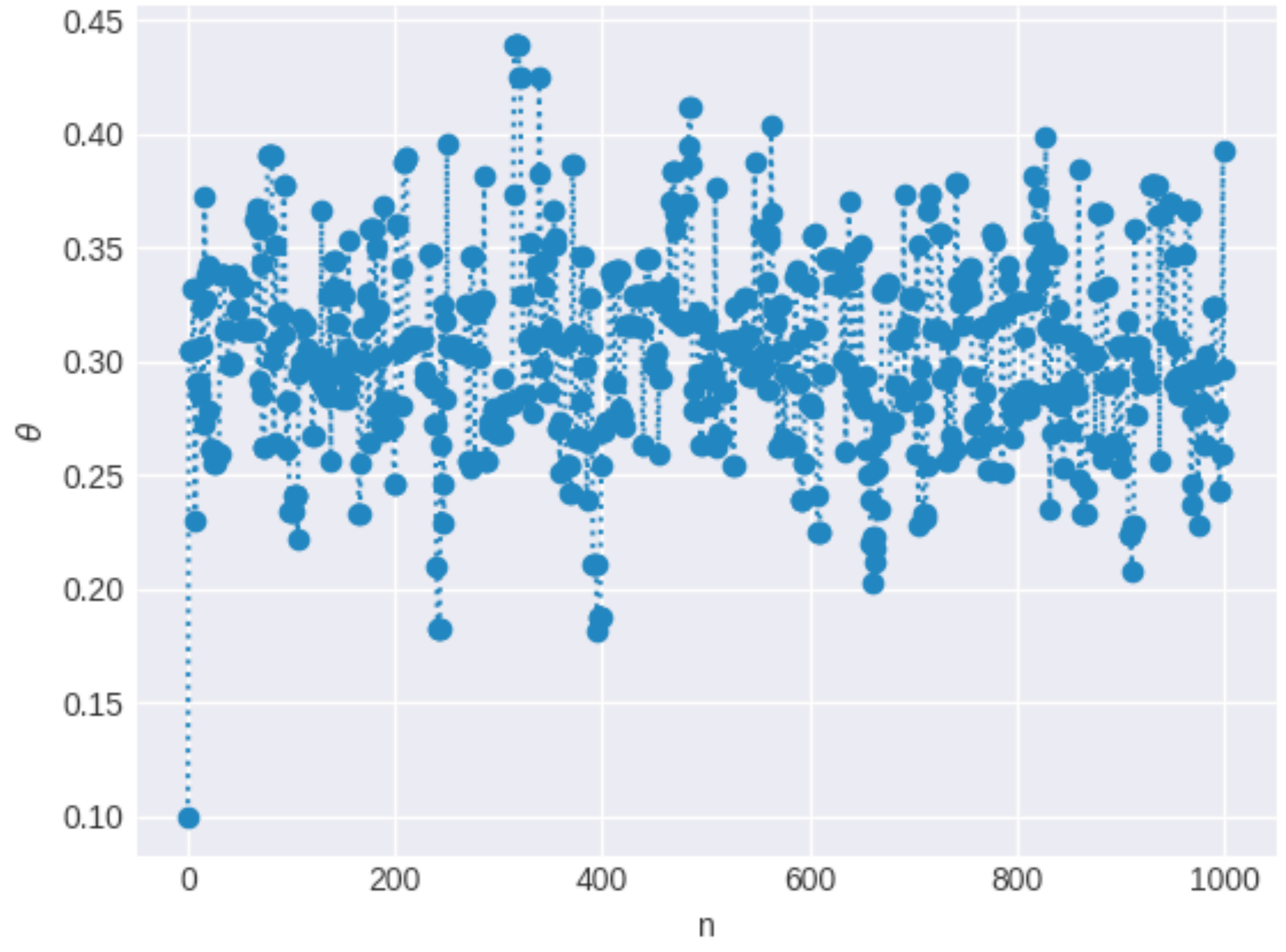
Again, let's look at the coin flip:

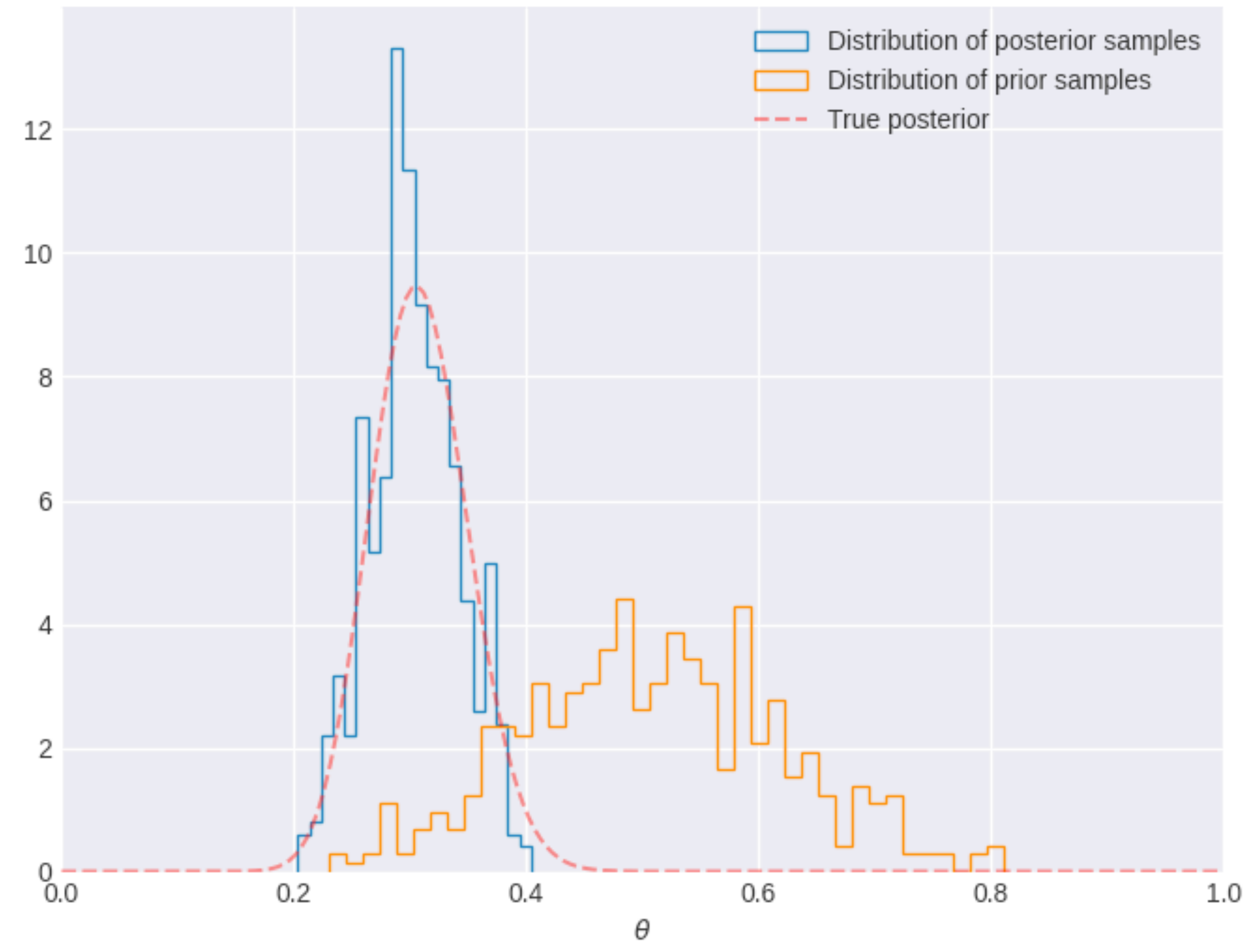
- Prior $p(\theta) \sim Beta(10,10)$
- Let $\theta' = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,0.1)$
- Then, $w(\theta' \mid \theta)$ is given by the distribution of ε
- Acceptance ratio:

$$\alpha(\theta' \mid \theta) = \min\left\{ \frac{p(d \mid \theta')p(\theta')}{p(d \mid \theta)p(\theta)}, 1 \right\}$$

(symmetry of ε)

- $u \sim U(0,1)$
- If $u < \alpha$, $\theta_{n+1} = \theta'$, else $\theta_{n+1} = \theta_n$





Example

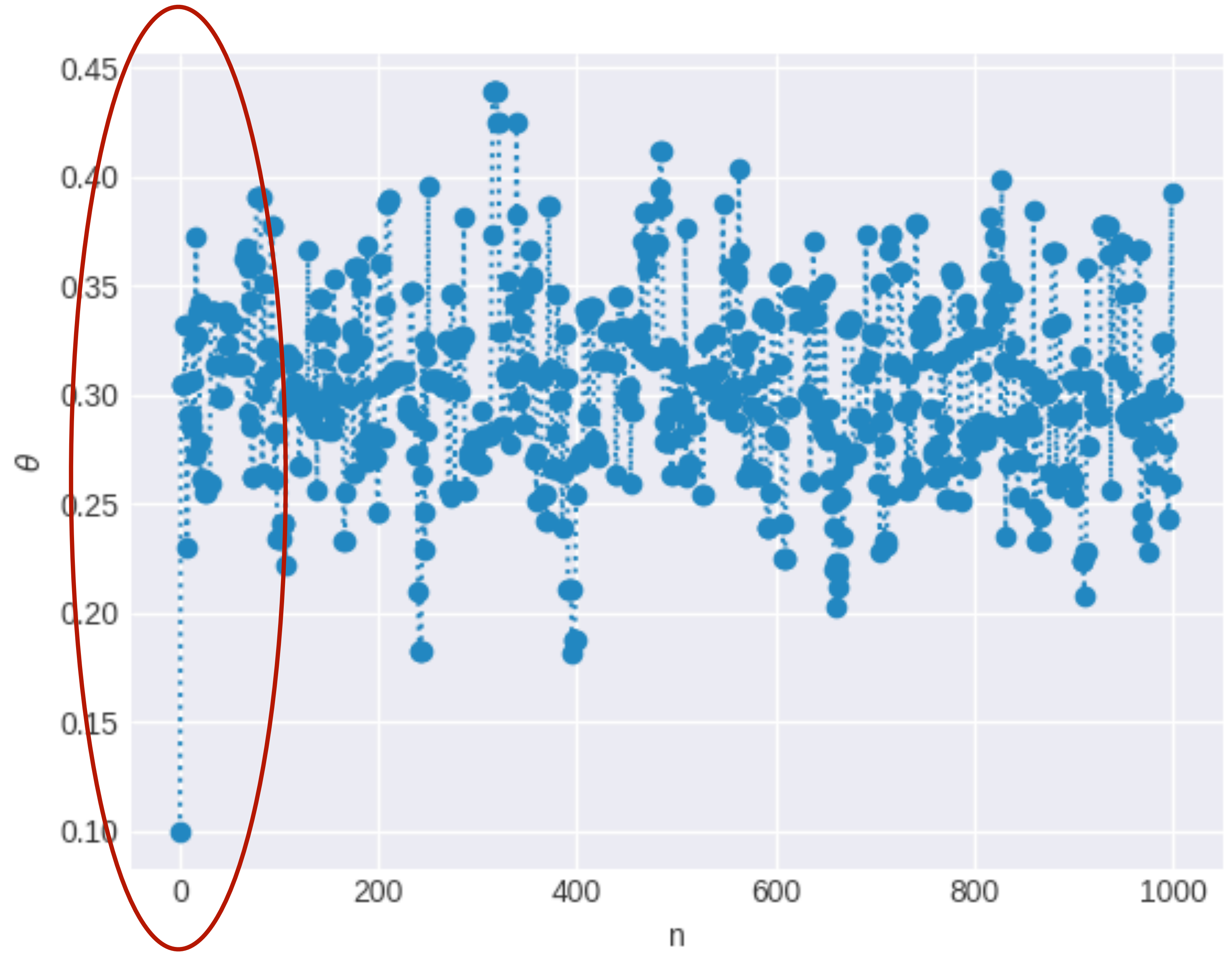
Again, let's look at the coin flip:

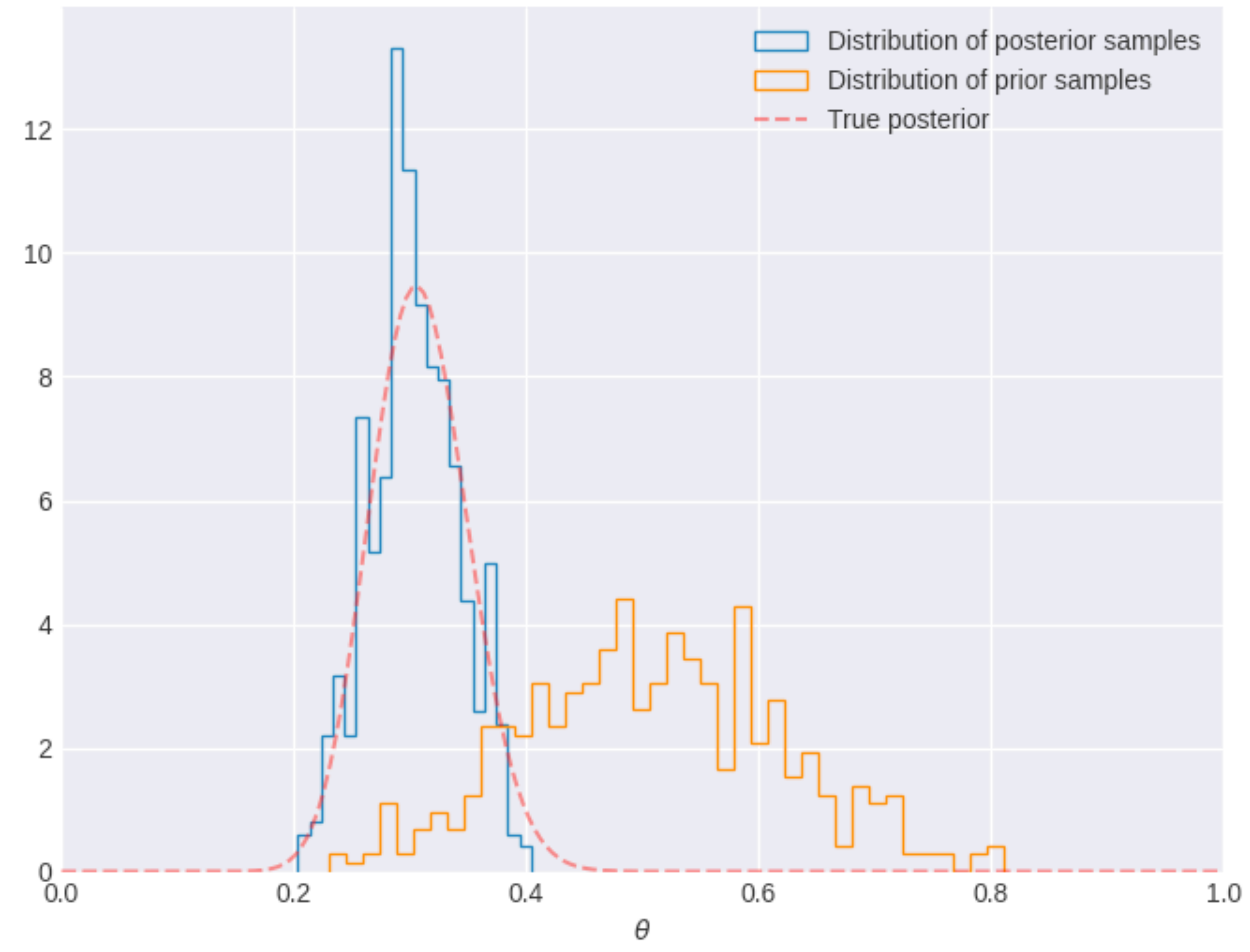
- Prior $p(\theta) \sim Beta(10{,}10)$
- Let $\theta' = \theta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0{,}0.1)$
- Then, $w(\theta' \,|\, \theta)$ is given by the distribution of ε
- Acceptance ratio:

$$\alpha(\theta' \,|\, \theta) = \min\left\{ \frac{p(d\,|\,\theta')p(\theta')}{p(d\,|\,\theta)p(\theta)}, 1 \right\}$$

(symmetry of ε)

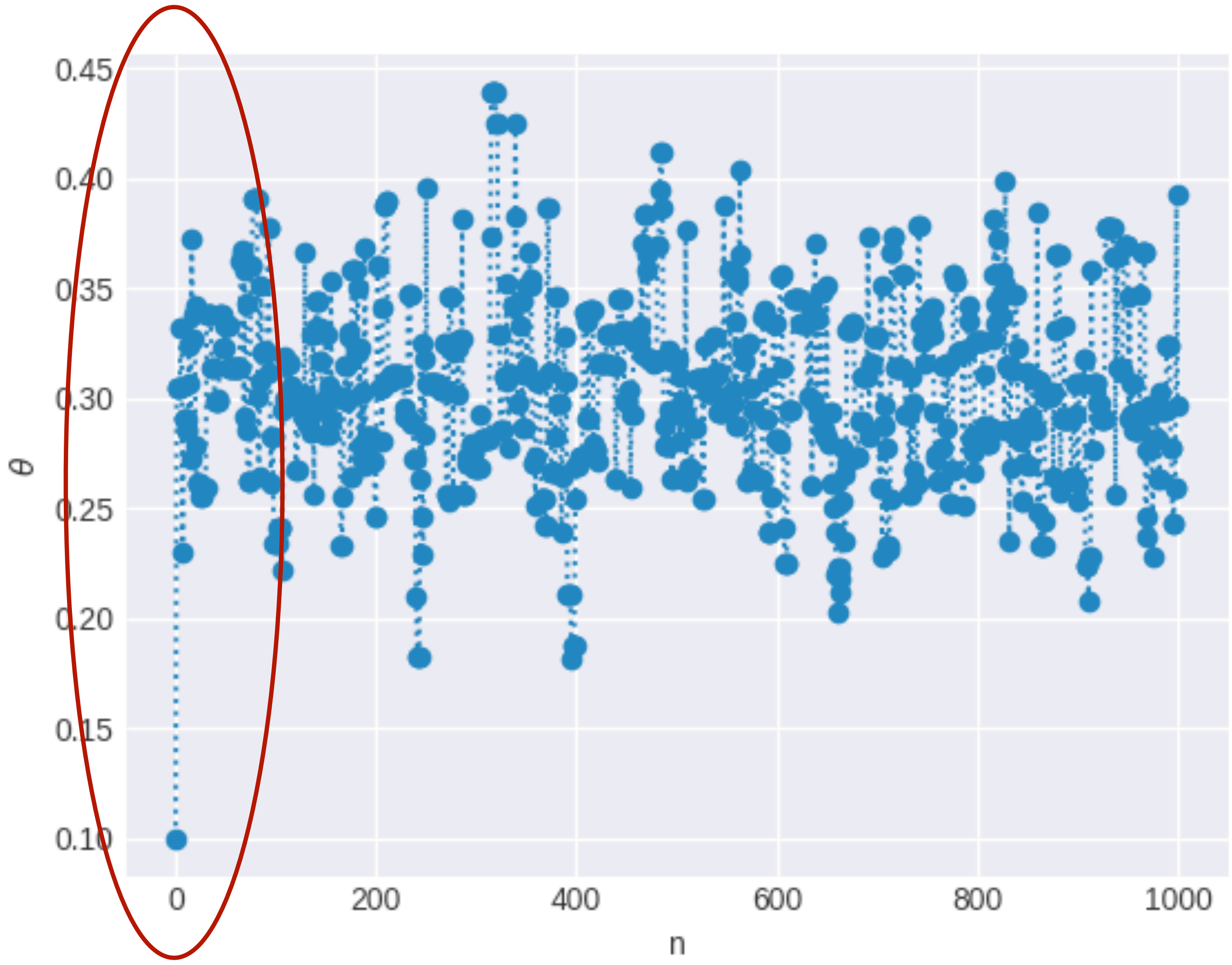
- $u \sim U(0{,}1)$
- If $u < \alpha$, $\theta_{n+1} = \theta'$, else $\theta_{n+1} = \theta_n$





Markov Chain Monte Carlo

- We have an assumption that at some point we reach the stationary distribution.
- In the beginning of the chain, this is not the case — *burn-in* period.



Markov Chain Monte Carlo Convergence

- Analytical upper bound for number of iterations to distance to stationarity (Rosenthal 2002). I.e. How long is the burn-in phase?
- Analytical bounds on the MCMC mean/variance and true parameter mean (Jones and Hobert, 2001)
- Eventually, we sample from the true posterior distribution.

Markov Chain Monte Carlo

- **Advantages:**
  - Easy to implement
  - Better at handling high-dimensional parameter spaces
  - Produces samples from the target distribution (asymptotically)
- **Disadvantages:**
  - Can be computationally costly to go to very high-dimensional problems/large datasets
  - Requires careful fine-tuning of parameters: step-size, proposal distribution, etc…

Questions?

Approximate inference through Variational inference

Variational inference

- When computing $p(\theta \mid d)$ is intractable
  - E.g. many parameters θ

- **Idea:** Replace the exact, but intractable posterior $p(\theta \mid d)$ with a tractable approximate posterior $q(\theta \mid d)$

Variational inference

- Let $q(\theta \mid d)$ belong to a family of probability distributions $\mathcal{Q}$

- Solve the optimisation problem:

$$q^*(\theta) := \arg\min_{q \in \mathcal{Q}} KL(q \mid p)$$

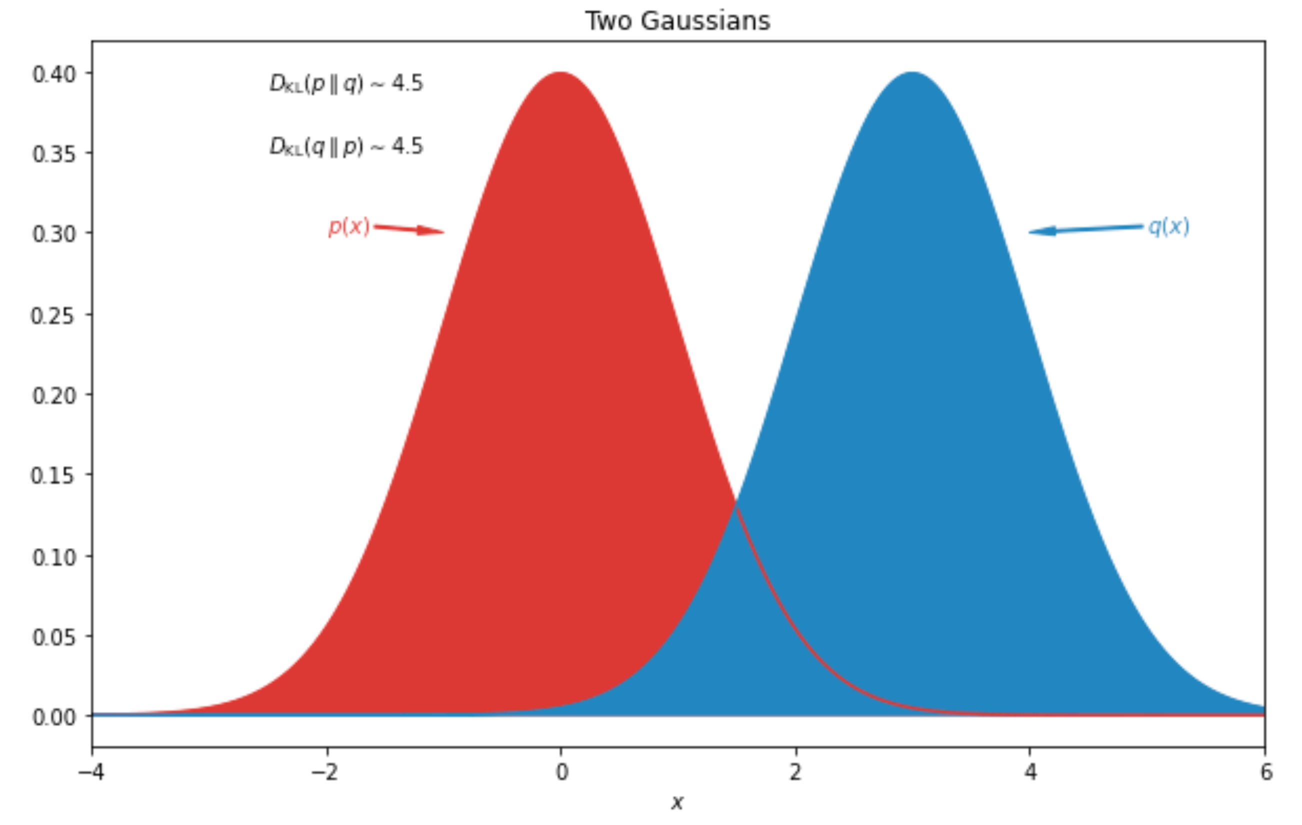
- We seek $q(\theta \mid d)$ that *approximates* the posterior $p(\theta \mid d)$.

Quick detour: KL divergence

- Kullback-Leibler (KL) divergence is a measure of dissimilarity between two probability distributions.

Let X and Y be two random variables with support R_X and R_Y and probability density functions $p_X(x)$ and $p_Y(y)$. Let $R_X \subseteq R_Y$. Then, the KL divergence of $p_Y(y)$ from $p_X(x)$ is

$$KL(p_X | p_Y) = \mathbb{E}_{x \sim X}\left[\ln\left(\frac{p_X(x)}{p_Y(y)}\right)\right].$$

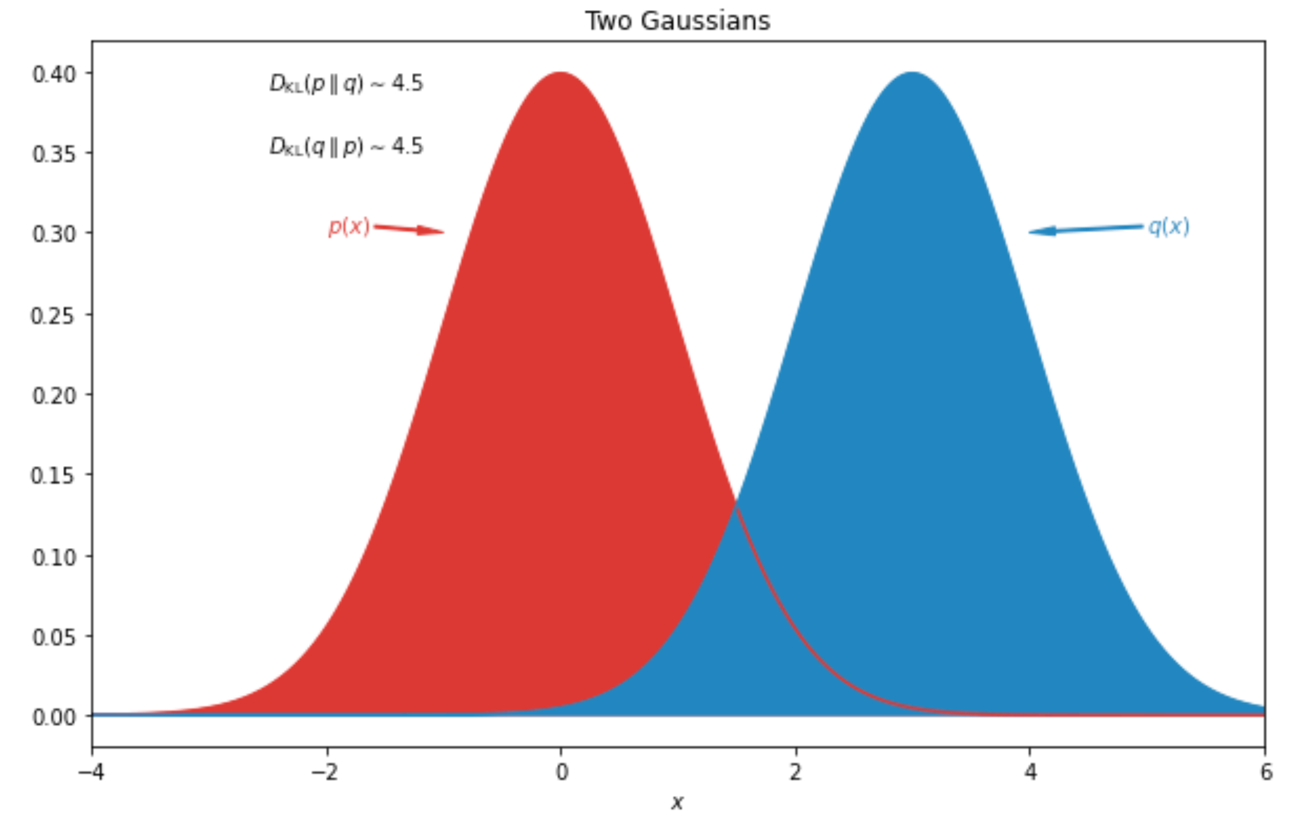


Quick detour: KL divergence

- Kullback-Leibler (KL) divergence is a measure of dissimilarity between two probability distributions.

Let X and Y be two random variables with support R_X and R_Y and probability density functions $p_X(x)$ and $p_Y(y)$. Let $R_X \subseteq R_Y$. Then, the KL divergence of $p_Y(y)$ from $p_X(x)$ is

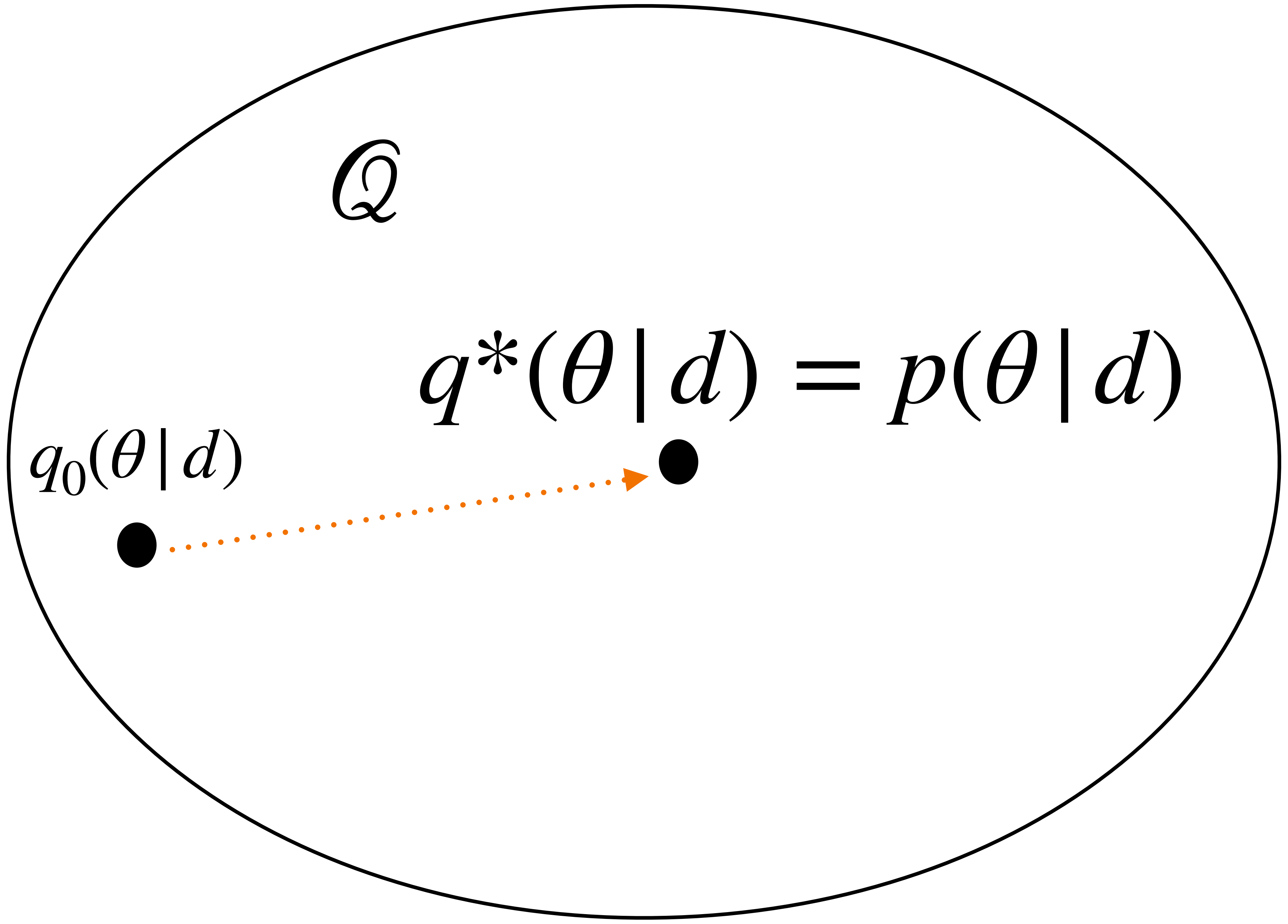
$$KL(p_X | p_Y) = \mathbb{E}_{x \sim X}\left[\ln\left(\frac{p_X(x)}{p_Y(y)}\right)\right].$$



- KL divergence is non-negative
- If $KL(p | q) = 0 \implies p = q$

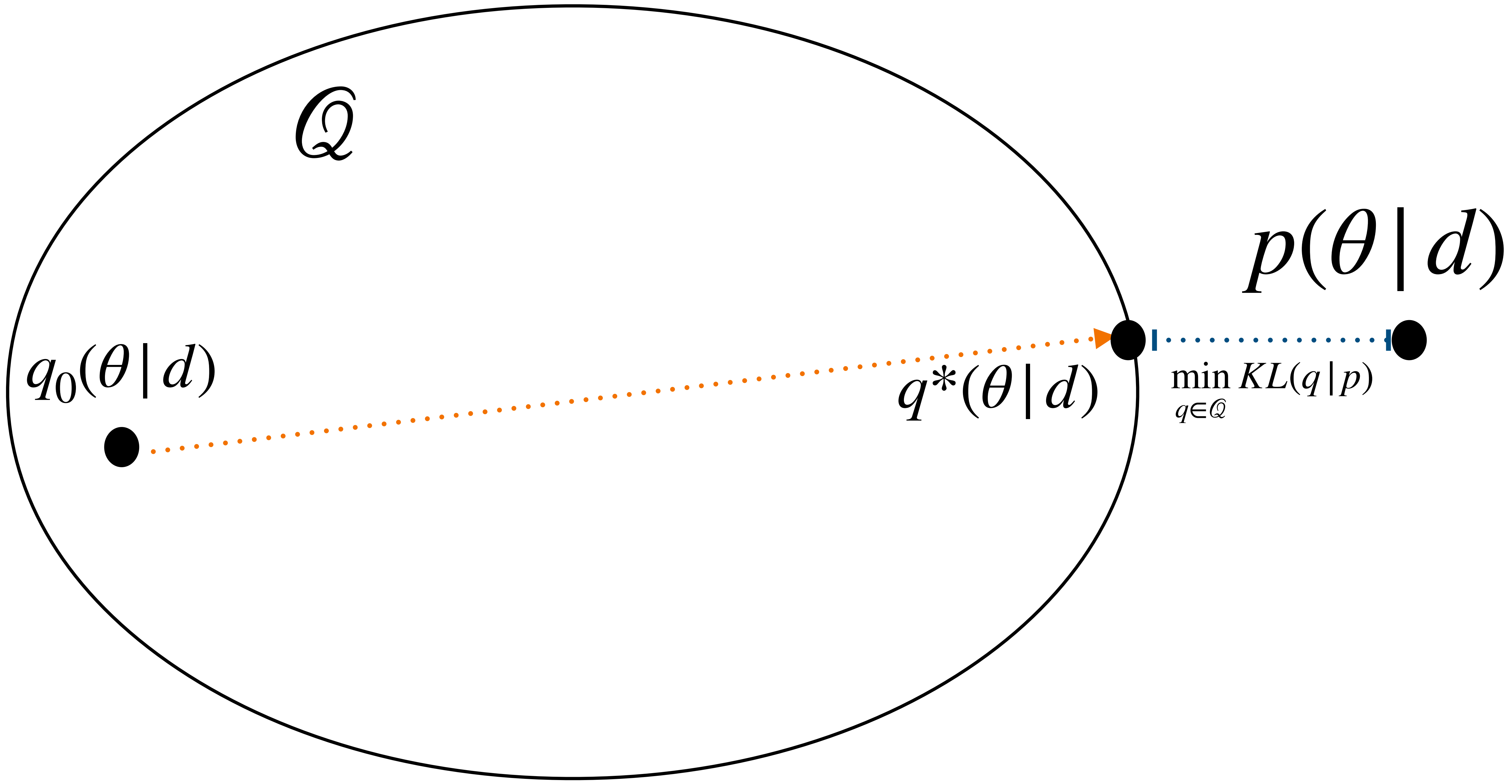
Variational inference

- If $p(\theta \,|\, d) \in \mathcal{Q}$, then $q^*(\theta \,|\, d) = p(\theta \,|\, d)$ (under some assumptions).



Variational inference

If $p(\theta | d) \notin \mathcal{Q}$, then $q^*(\theta | d)$ minimises the Kullback-Leibler divergence between the two distributions.



How to solve the minimisation?

$$q_\lambda(\theta) := \arg\min_{q \in \mathcal{Q}} KL(q \,|\, p) \iff \arg\max_{q \in \mathcal{Q}} ELBO(q, \theta)$$

How to solve the minimisation?

$$q_\lambda(\theta) := \underset{q \in \mathcal{Q}}{\arg\min}\, KL(q\,|\,p) \iff \underset{q \in \mathcal{Q}}{\arg\max}\, ELBO(q, \theta)$$

$$KL(q\,|\,p) = E_{\theta\sim q}\left[\ln\left(\frac{q(\theta\,|\,d)}{p(\theta\,|\,d)}\right)\right]$$

$$= E_{\theta\sim q}\left[\ln\left(q(\theta\,|\,d)\right)\right] - E_{\theta\sim q}\left[\ln\left(p(\theta\,|\,d)\right)\right]$$ Log properties

$$= E_{\theta\sim q}\left[\ln\left(q(\theta\,|\,d)\right)\right] - E_{\theta\sim q}\left[\ln\left(\frac{p(\theta, d)}{p(d)}\right)\right]$$ Definition of posterior

$$= E_{\theta\sim q}\left[\ln\left(q(\theta\,|\,d)\right)\right] - E_{\theta\sim q}\left[\ln\left(p(\theta, d)\right)\right] + E_{\theta\sim q}\left[\ln(p(d))\right]$$ Log properties

$$= -\left(E_{\theta\sim q}\left[\ln\left(p(\theta, d)\right)\right] - E_{\theta\sim q}\left[\ln\left(q(\theta\,|\,d)\right)\right]\right) + \ln(p(d))$$ Independence of θ and d

How to solve the minimisation?

$$q_\lambda(\theta) := \arg\min_{q \in \mathcal{Q}} KL(q\,|\,p) \iff \arg\max_{q \in \mathcal{Q}} \boxed{ELBO(q, \theta)}$$

$$\begin{aligned}
KL(q\,|\,p) &= E_{\theta \sim q}\left[\ln\left(\frac{q(\theta\,|\,d)}{p(\theta\,|\,d)}\right)\right] \\
&= E_{\theta \sim q}\left[\ln\left(q(\theta\,|\,d)\right)\right] - E_{\theta \sim q}\left[\ln\left(p(\theta\,|\,d)\right)\right] && \text{Log properties} \\
&= E_{\theta \sim q}\left[\ln\left(q(\theta\,|\,d)\right)\right] - E_{\theta \sim q}\left[\ln\left(\frac{p(\theta, d)}{p(d)}\right)\right] && \text{Definition of posterior} \\
&= E_{\theta \sim q}\left[\ln\left(q(\theta\,|\,d)\right)\right] - E_{\theta \sim q}\left[\ln\left(p(\theta, d)\right)\right] + E_{\theta \sim q}\left[\ln(p(d))\right] && \text{Log properties} \\
&= -\underbrace{\boxed{\left(E_{\theta \sim q}\left[\ln\left(p(\theta, d)\right)\right] - E_{\theta \sim q}\left[\ln\left(q(\theta\,|\,d)\right)\right]\right)}}_{\text{Evidence Lower Bound (ELBO)}} + \ln(p(d)) && \text{Independence of } \theta \text{ and } d
\end{aligned}$$

Variational inference

- Formulate the approximate Bayesian inference problem as an optimisation problem $\Longrightarrow$ use optimisation tools to solve the inference problem
  - e.g. Use gradient descent-like method

What can be said of $\mathcal{Q}$?

- **Mean field approximation:**
  - Assume the variational distribution over the parameters θ factorizes as:

$$q(\theta_1, \cdots, \theta_m) = \prod_{j=1}^{m} q(\theta_j)$$

    - Assumes the parameters are independent from each other
    - Usually $p(\theta \mid d) \notin \mathcal{Q}$

What can be said of $\mathcal{Q}$?

- **Mean field approximation:**
  - Assume the variational distribution over the parameters θ factorizes as:

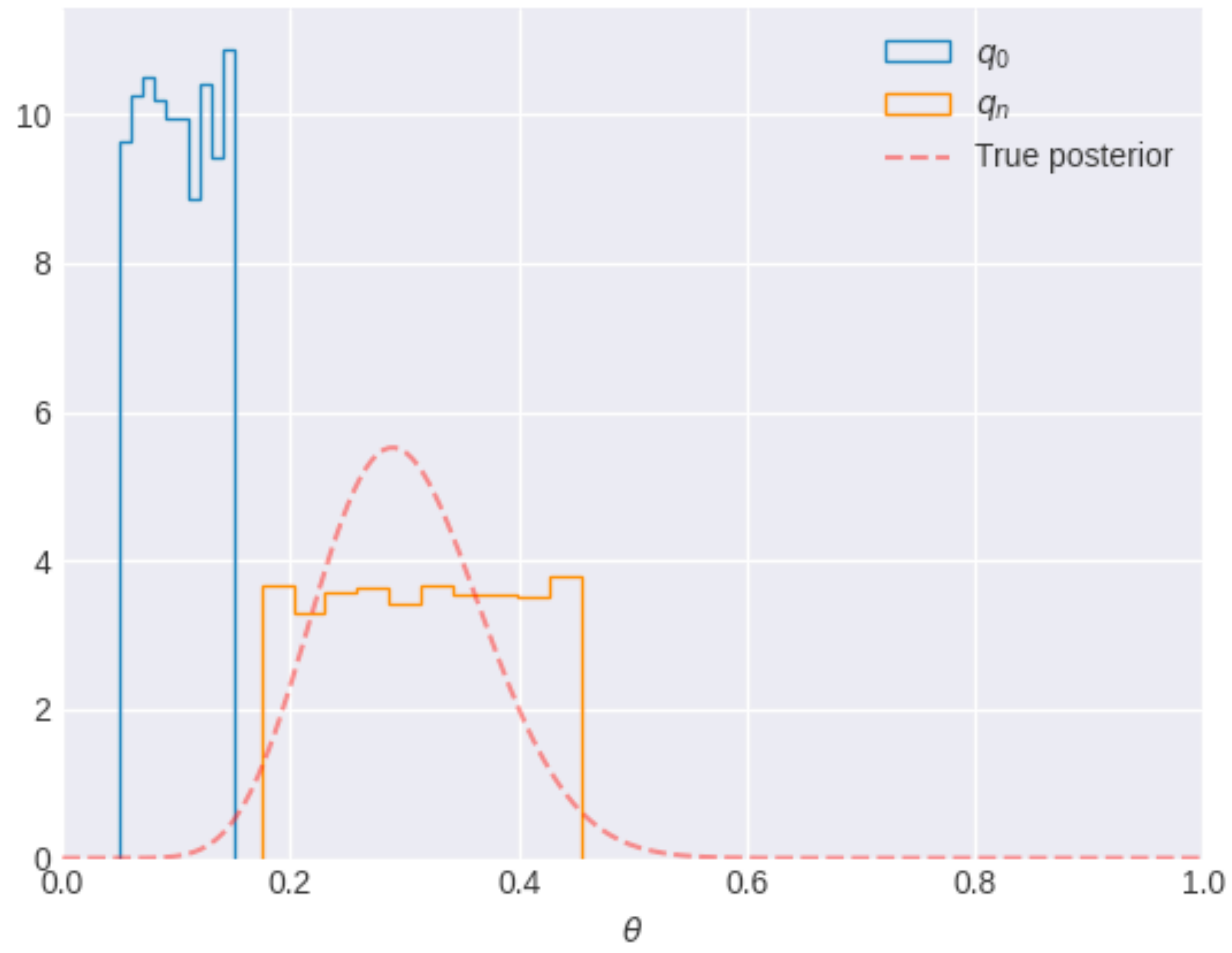
$$q(\theta_1, \cdots, \theta_m) = \prod_{j=1}^{m} q(\theta_j)$$

    - Assumes the parameters are independent from each other
    - Usually $p(\theta | d) \notin \mathcal{Q}$
- **Fixed form approximation:**
  - Assume the variational distribution $q \in \mathcal{Q}$, some class of distributions indexed by a vector λ (*variational parameter*)

Example 1: $\mathcal{Q} :=$ family of n-dimensional Gaussian distributions, variational parameters $\lambda :=$ vector of means $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$
Example 2: $\mathcal{Q} :=$ d-deep neural network, variational parameters $\lambda :=$ weights and biases

Example

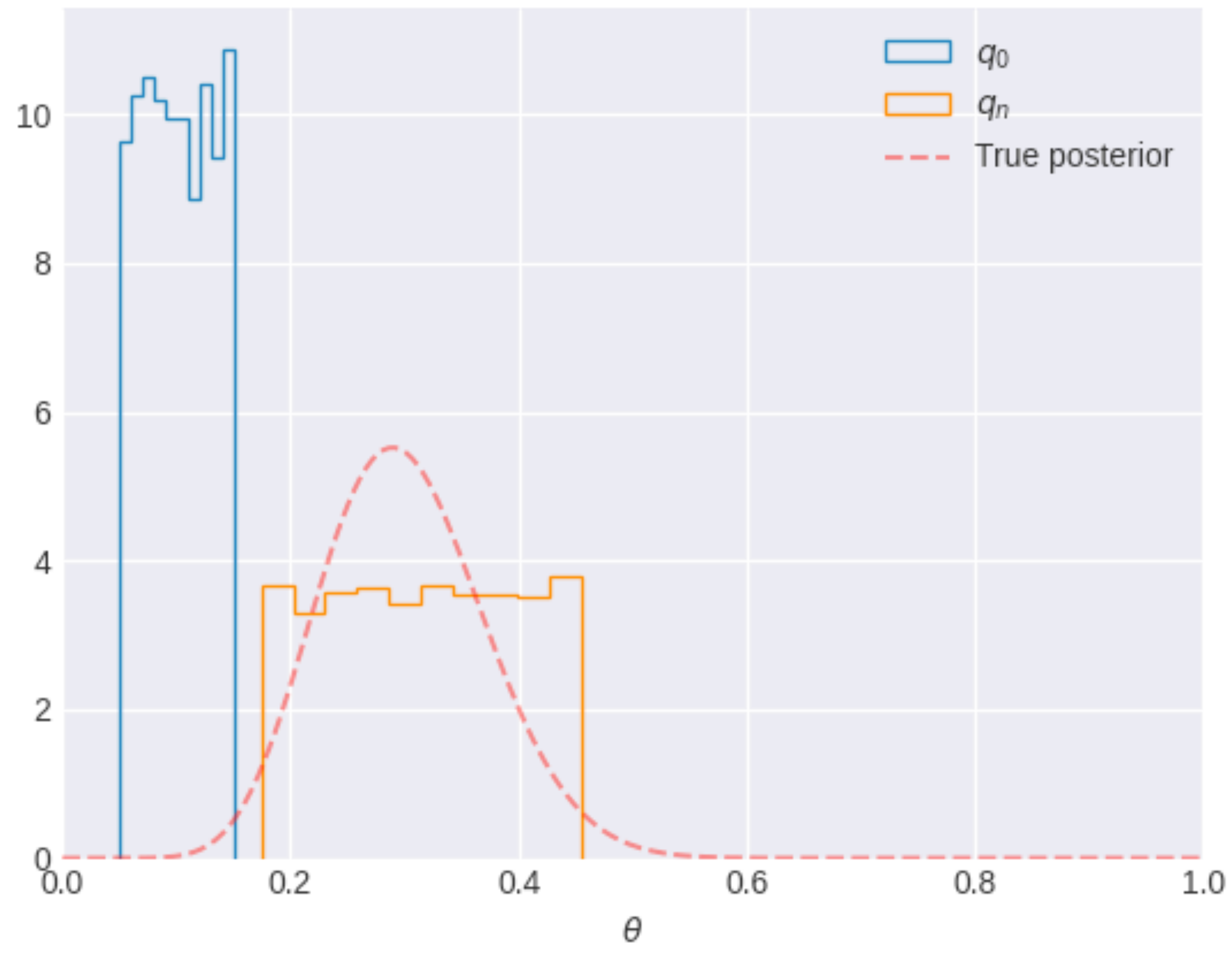
Again, let's look at the coin flip.
Let us consider $\mathcal{Q} := U(a, b)$, then, $p(\theta \mid d) \notin \mathcal{Q}$.



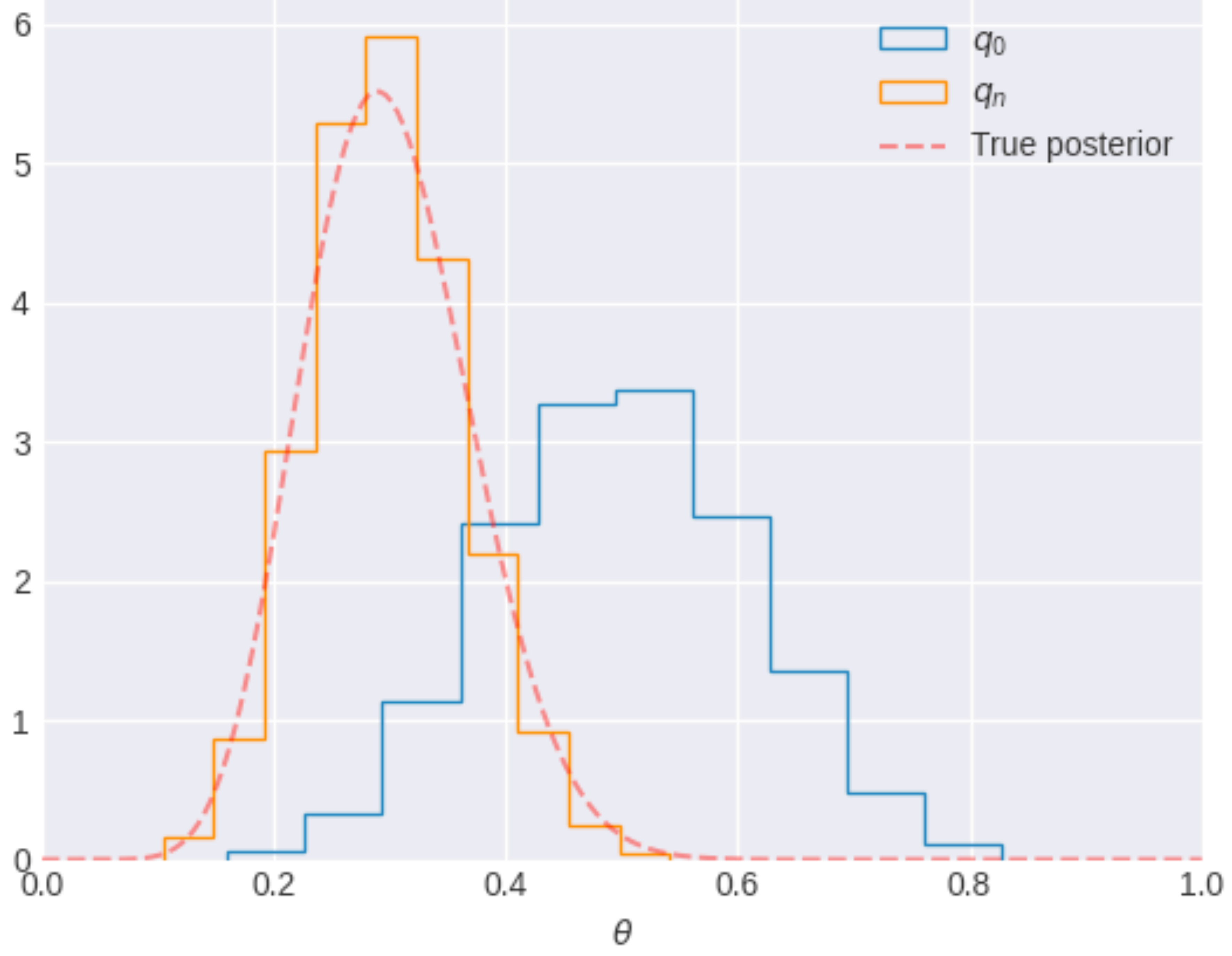
Example

Again, let's look at the coin flip.

Let us consider $\mathcal{Q} := U(a, b)$, then, $p(\theta \mid d) \notin \mathcal{Q}$.



Let us consider $\mathcal{Q} :=$ Beta(a, b), then, $p(\theta \mid d) \in \mathcal{Q}$.



What can be said about convergence?

- Not much.
  - On the convergence of the mean of the variational posterior to the true mean of the posterior: (Wang and Blei, 2021)
  - On the convergence of the variational posterior to true posterior distribution moments: (Zhang and Gao, 2020)

- We might never be close to the true posterior distribution.

Variational inference

- **Advantages:**
  - Scalable
  - Fast
- **Disadvantages:**
  - Little theory on convergence
  - Computationally complex

Summary

	Dimension	Expressivity	Efficiency	Computational Complexity
Conjugate priors	Low	Low	High	Low
Sampling	Low	High	Low	Low
Variational inference	High	Varying	High	High

Break time

Hands-on session: http://bit.ly/43OLjUh