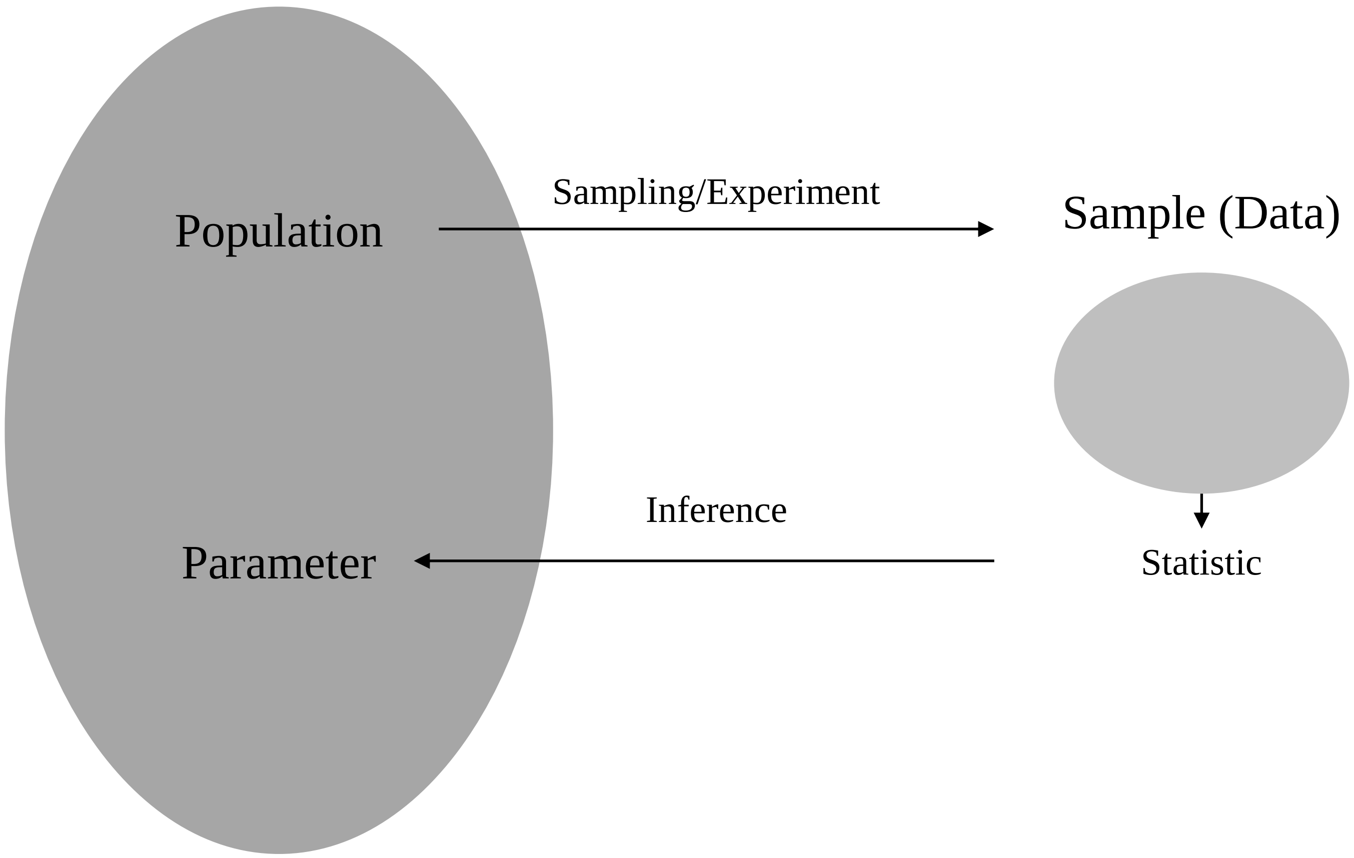
Convex Relaxation for Variational Inference

2019. 03. 08

DMQA Lab Seminar

Statistical Inference



Bayesian Inference

Bayesian Inference

Bayes' Rule

Prior (Assumption)

Likelihood (Model)

Posterior
$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}$$

Evidence

What we know: Likelihood, Prior (Model/Assumption)

What we do not know: Posterior, Evidence

What we want know: **Posterior**

Bayesian Inference

Coin Flipping Example

- We have a coin and want to infer the probability of “Head” when the coin is flipped

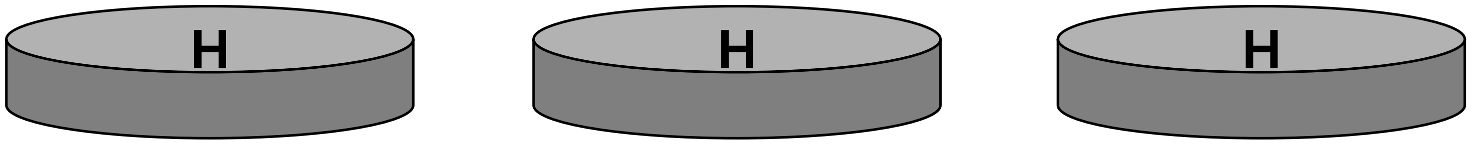
Probability of Head: $P(X = 1) = \theta$

Model

$$X \sim \mathrm{Bernoulli}(\theta)$$

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

Observation (data)



$$(x_1 = 1, x_2 = 1, x_3 = 1)$$

Bayesian Inference

Coin Flipping Example

- Frequentist Approach: Maximum Likelihood Inference

Likelihood Function

$$L(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i|\theta) \qquad (x_1 = 1, x_2 = 1, x_3 = 1)$$

$$= \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}$$

Maximum Likelihood Estimation

Likelihood function is maximized at

$$\hat{\theta}^{\mathrm{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n} = 1$$

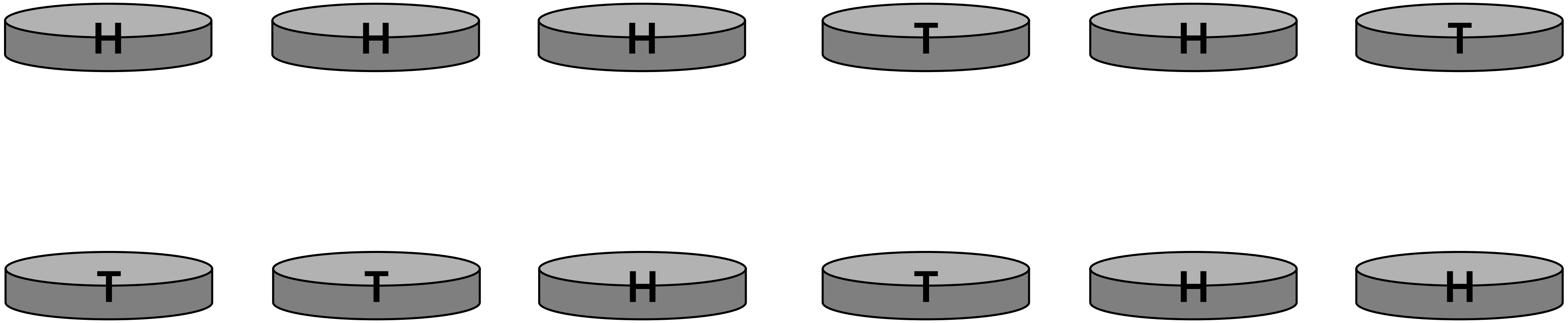
Bayesian Inference

Coin Flipping Example

- Frequentist Approach: Maximum Likelihood Inference

But we all know that the probability of Head is not 1…

Frequentists will say “We have to collect more data”



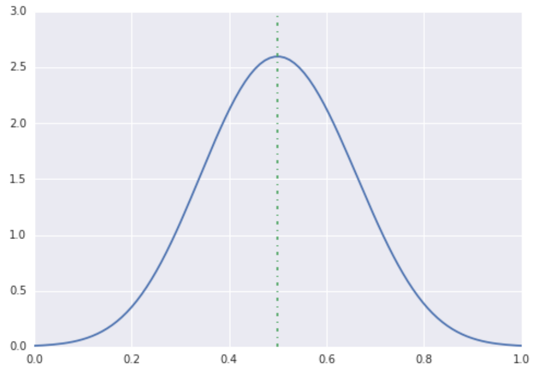
Bayesian Inference

Coin Flipping Example

- Bayesian Approach: Posterior Inference

Prior

$$\theta \sim \mathrm{Beta}(10, 10)$$



$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

We may assume the probability of "Head" is nearly 0.5

Likelihood Function

$$L(\theta; x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i|\theta)$$

$$= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

Bayesian Inference

Coin Flipping Example

- Bayesian Approach: Posterior Inference

$$p(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

$$\begin{aligned} L(\theta; x_1, \cdots, x_n) &= \prod_{i=1}^{n} p(x_i|\theta) \\ &= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \end{aligned}$$

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

$$p(\theta|x_1, \cdots, x_n) \propto p(\theta) \times L(\theta; x_1, \cdots, x_n)$$

Bayesian Inference

Coin Flipping Example

- Bayesian Approach: Posterior Inference

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

$$p(\theta|x_1, \cdots, x_n) \propto p(\theta) \times L(\theta; x_1, \cdots, x_n)$$

$$\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \times \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

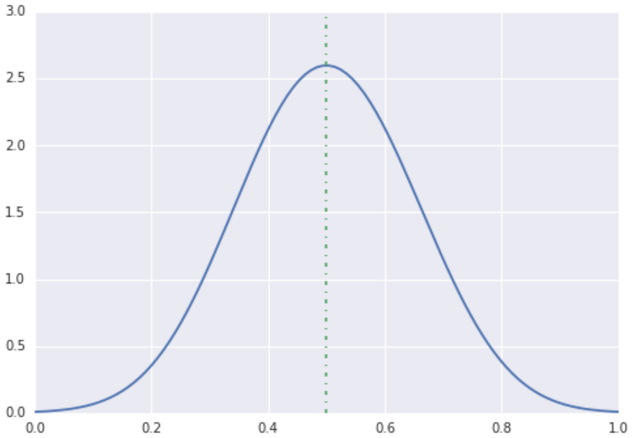
$$\propto \mathrm{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$$

Bayesian Inference

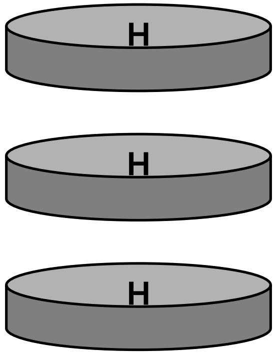
Coin Flipping Example

- Bayesian Approach: Posterior Inference

$\theta \sim \mathrm{Beta}(10, 10)$

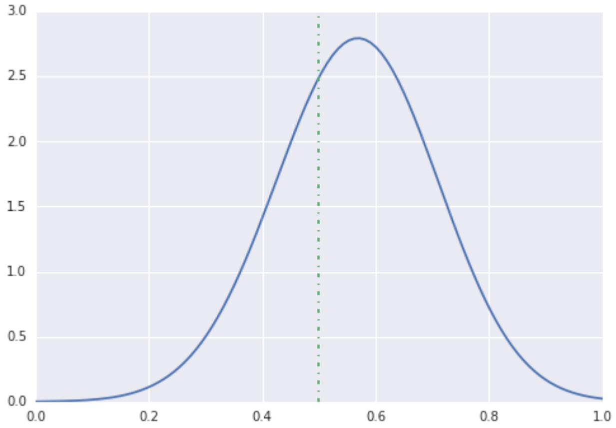


Prior



Data

$\mathrm{Beta}(13, 10)$



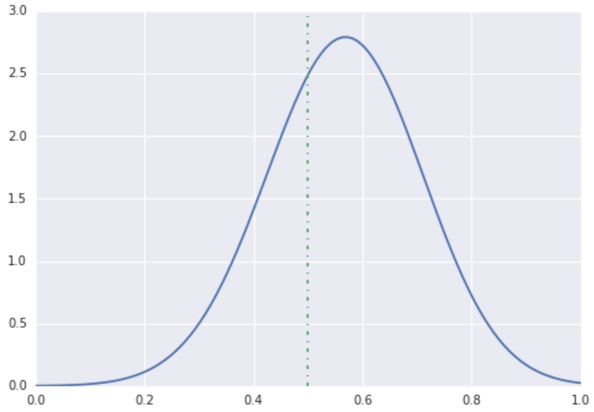
Posterior

Bayesian Inference

Coin Flipping Example

- Bayesian Approach: Posterior Inference

$\mathrm{Beta}(13, 10)$



$$p(\theta|\mathbf{x}) = \frac{x^{12}(1-x)^9}{\mathrm{B}(13, 10)}$$

Posterior Distribution

Maximum a posteriori (MAP)

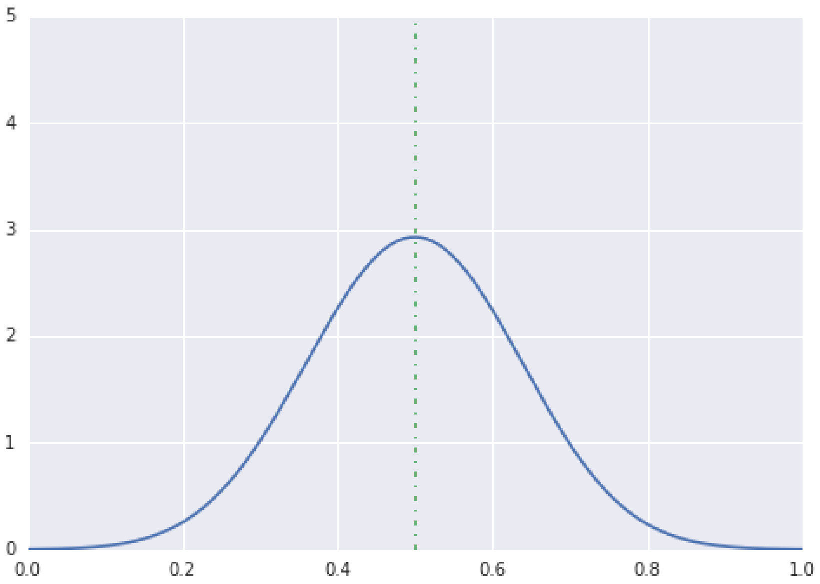
Posterior distribution is maximized at mode which is 0.5714

$$\hat{\theta}^{MAP} = 0.5714$$

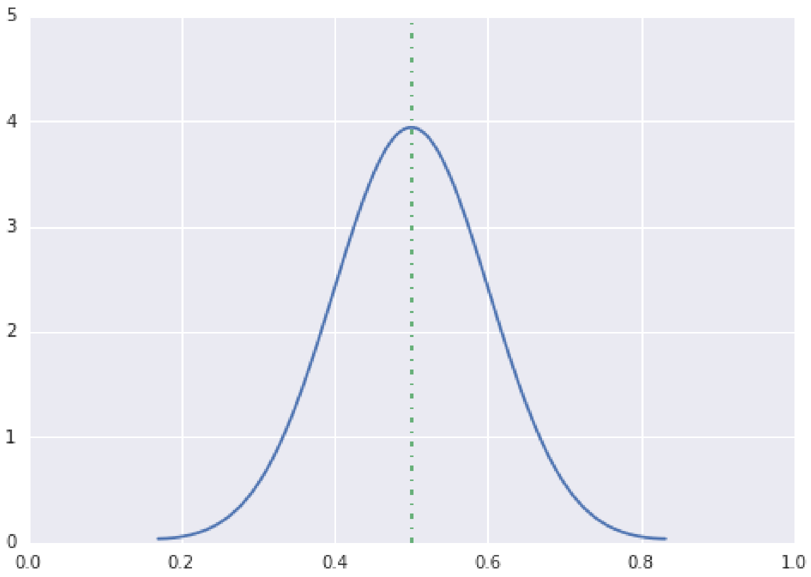
Bayesian Inference

Coin Flipping Example

- Various Priors



Larger search space

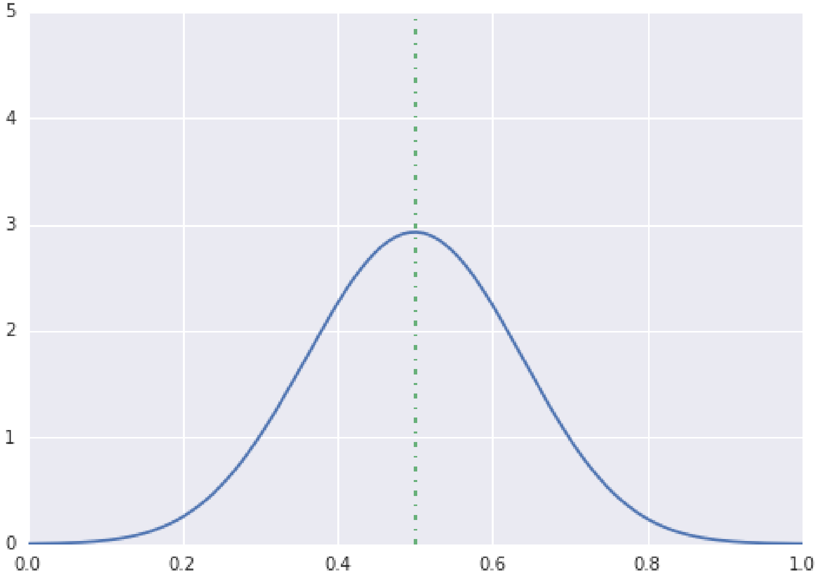


Smaller search space

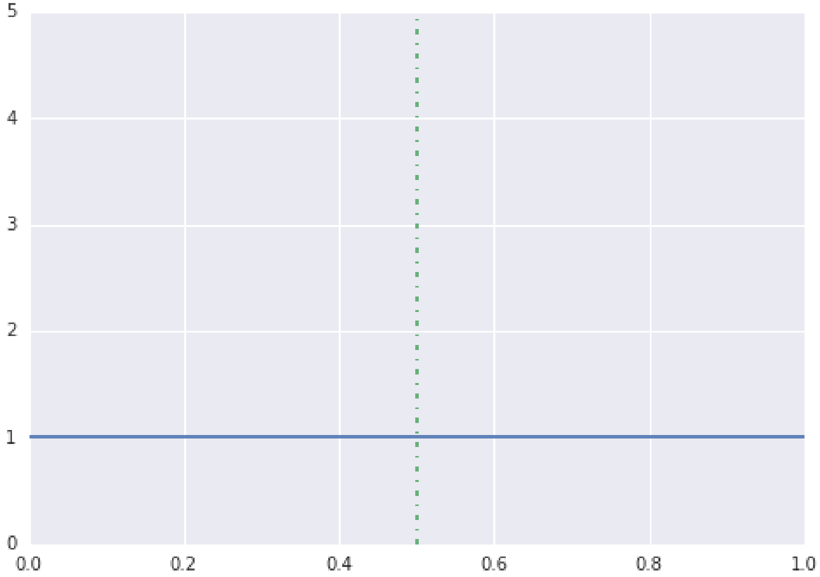
Bayesian Inference

Coin Flipping Example

- Various Priors



Informative Prior



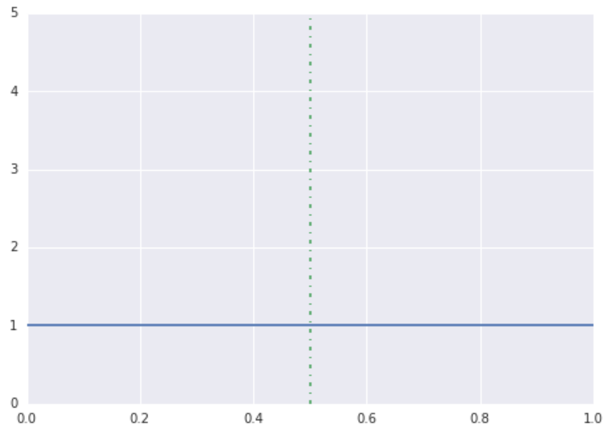
Noninformative Prior

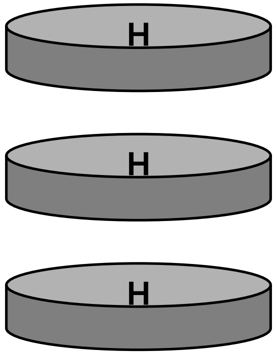
Bayesian Inference

Coin Flipping Example

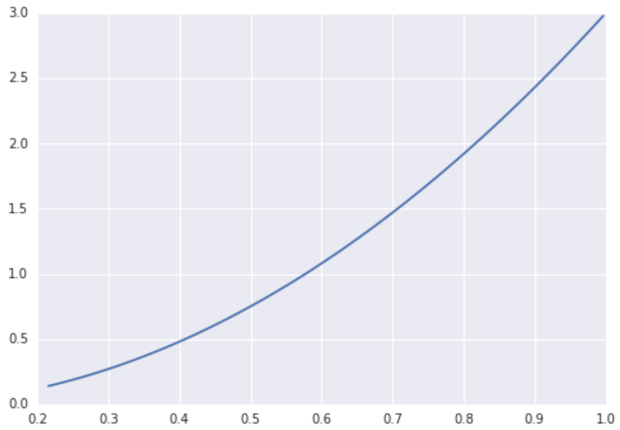
- Posterior Inference with Noninformative Prior

$\mathrm{Beta}(1, 1)$





$\mathrm{Beta}(4, 1)$



Prior

Data

Posterior
(mode at 1)

Bayesian Inference

Coin Flipping Example

- As we collect more samples…

Posterior

$$\begin{cases} \dfrac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \times \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ \propto \mathrm{Beta}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i) \end{cases}$$

MAP

$$\frac{\alpha + \sum_{i=1}^n x_i - 1}{\alpha + \sum_{i=1}^n x_i + \beta + n - \sum_{i=1}^n x_i - 2} = \frac{\alpha - 1 + \sum_{i=1}^n x_i}{\alpha + \beta - 2 + n}$$

As sample size increases
(MAP → MLE)

$$\frac{\alpha - 1 + \sum_{i=1}^n x_i}{\alpha + \beta - 2 + n} \longrightarrow \frac{\sum_{i=1}^n x_i}{n}$$

Bayesian Inference

Example: Bayesian Linear Regression

- with fixed variance of random noise

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} \in \mathbb{R}^n : \text{response variable}$$

$$\mathbf{X} \in \mathbb{R}^{n \times p} : \text{predictor variables}$$

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_p) : \text{regression coefficients}$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2) : \text{random noise}$$

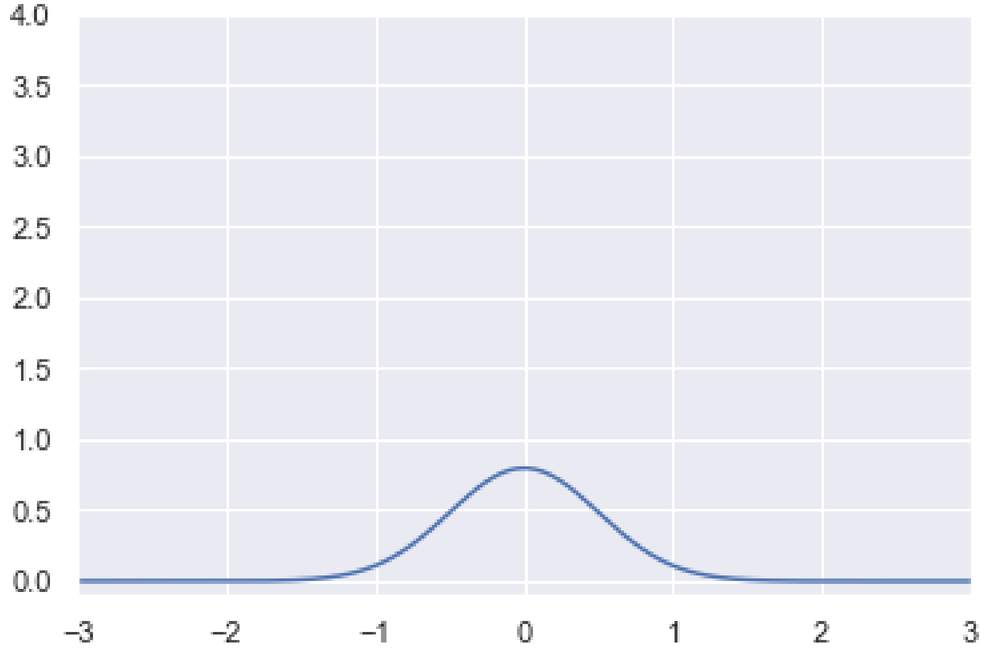
Bayesian Inference

Example: Bayesian Linear Regression
- with fixed variance of random noise

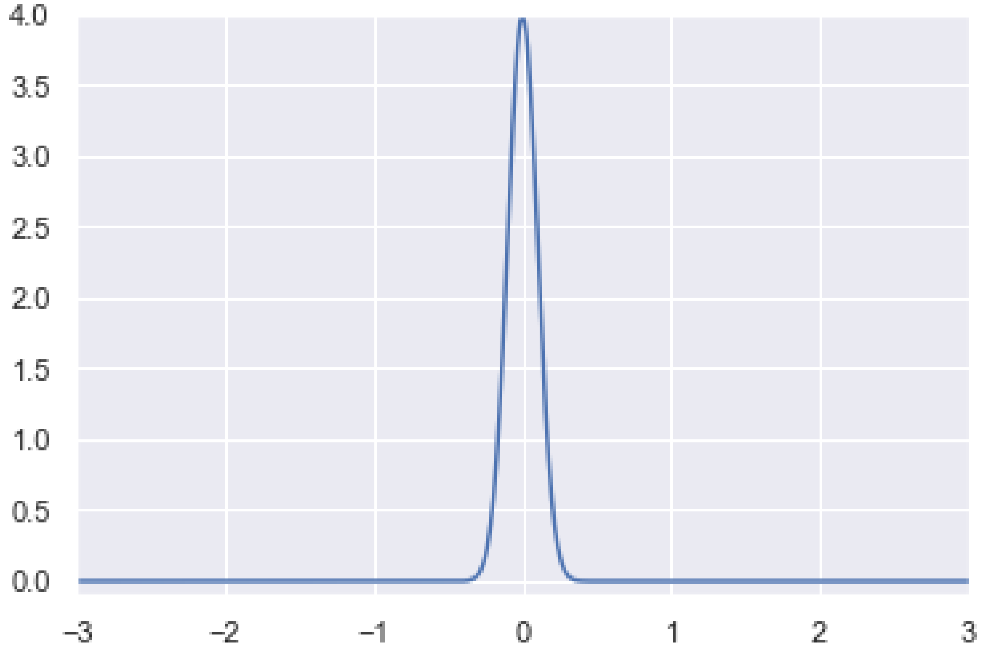
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\mathbf{y} \in \mathbb{R}^n$: response variable
$\mathbf{X} \in \mathbb{R}^{n \times p}$: predictor variables
$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}_p)$: regression coefficients
$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2)$: random noise

Small alpha



Large alpha



Bayesian Inference

Example: Bayesian Linear Regression

- with fixed variance of random noise

Prior
$$p(\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\alpha^{-p}}} e^{-\alpha\boldsymbol{\beta}^T\boldsymbol{\beta}}$$

Likelihood
$$\mathcal{L}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{2\sigma^{-2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}$$

Bayesian Inference

Example: Bayesian Linear Regression

- with fixed variance of random noise

$$p(\boldsymbol{\beta}) \times \mathcal{L}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) \propto \mathcal{N}((\mathbf{X}^T\mathbf{X} + \sigma^2\alpha\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}, \alpha\mathbf{I}_p + \sigma^{-2}\mathbf{X}^T\mathbf{X})$$

Posterior

$$\hat{\boldsymbol{\beta}}^{\text{MAP}} = (\mathbf{X}^T\mathbf{X} + \sigma^2\alpha\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\boldsymbol{\beta}}^{\text{MLE}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

Bayesian Inference

Example: Bayesian Linear Regression

- with fixed variance of random noise

$$\hat{\boldsymbol{\beta}}^{\text{MAP}} = (\mathbf{X}^T\mathbf{X} + \sigma^2\alpha\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{MAP}}$$

or.. use predictive posterior distribution

$$p(y|\boldsymbol{\beta}, \mathbf{y}) = \int p(y|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathbf{y})d\boldsymbol{\beta}$$

$$\mathcal{N}(\mathbf{y}|(\mathbf{X}^T\mathbf{X} + \sigma^2\alpha\mathbf{I}_p^{-1})\mathbf{X}^T\mathbf{y}, (\alpha\mathbf{I}_p + \sigma^{-2}\mathbf{X}^T\mathbf{X})^{-1})$$

Bayesian Inference

Example: Bayesian Linear Regression

- with fixed variance of random noise

e.g.

Pointwise Prediction: $X_{new} \rightarrow$ value of y hat — y hat is 5

Predictive Distribution: $X_{new} \rightarrow$ distribution of y hat — y hat ~ $N(4.5, 1)$

Bayesian Inference

Conjugate Prior

$$\text{Beta} \times \text{Bernoulli} \Rightarrow \text{Beta}$$

Coin Flipping

$$\text{Beta} \times \text{Binomial} \Rightarrow \text{Beta}$$

$$\text{Normal} \times \text{Normal} \Rightarrow \text{Normal}$$

Linear Regression

Approximate Bayesian Inference

Approximate Bayesian Inference

Motivation

Bayes' Rule $$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}$$

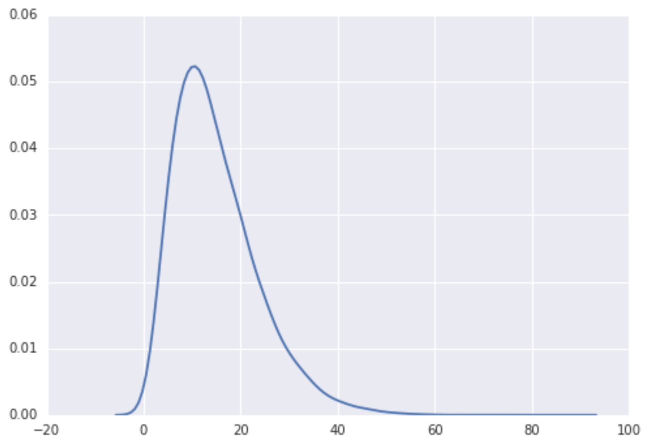
Evidence $$p(x) = \int p(\theta, x)d\theta = \int p(x|\theta)p(\theta)d\theta$$

This integration is not computable in general

Approximate Bayesian Inference

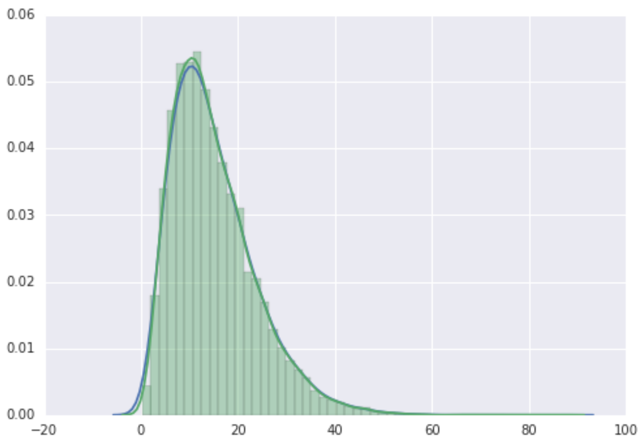
Methods for Intractable Posterior

True Posterior



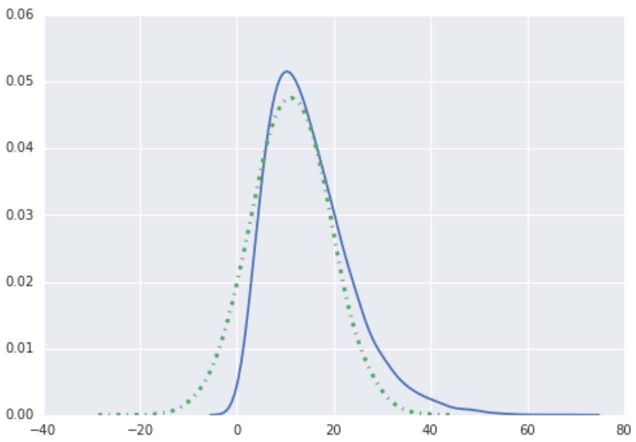
$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}$$

Sampling-based



$$\theta_1, \theta_2, \theta_3, \cdots, \theta_m \sim p(\theta|x)$$

Approximate Inference

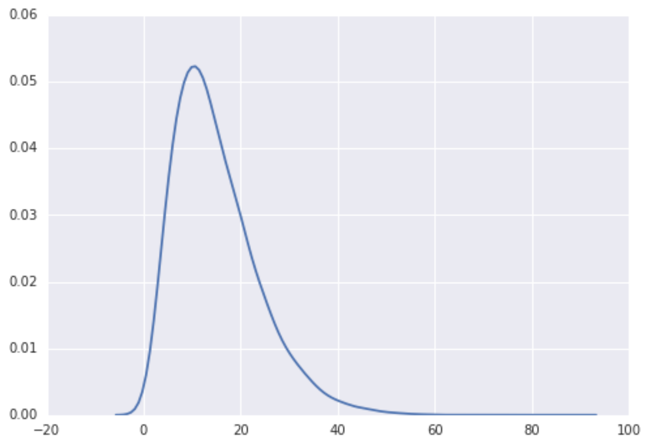


$$q(\theta) \approx p(\theta|x)$$

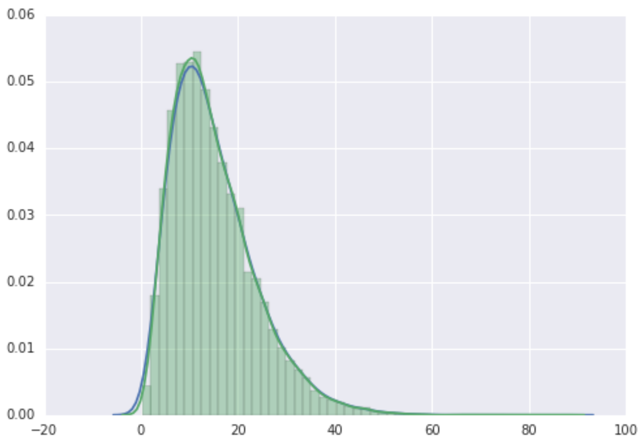
Approximate Bayesian Inference

Methods for Intractable Posterior

True Posterior

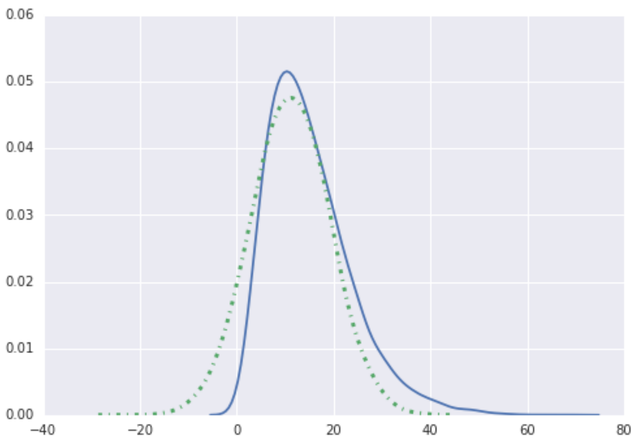


Sampling-based



- Naïve Monte Carlo
- Metropolis-Hastings
- Rejection Sampling
- Gibbs Sampling
- Importance Sampling
- Reversible-Jump MCMC

Approximate Inference



- Laplace Approximation
- Expectation Propagation
- **Variational Inference**

Approximate Bayesian Inference

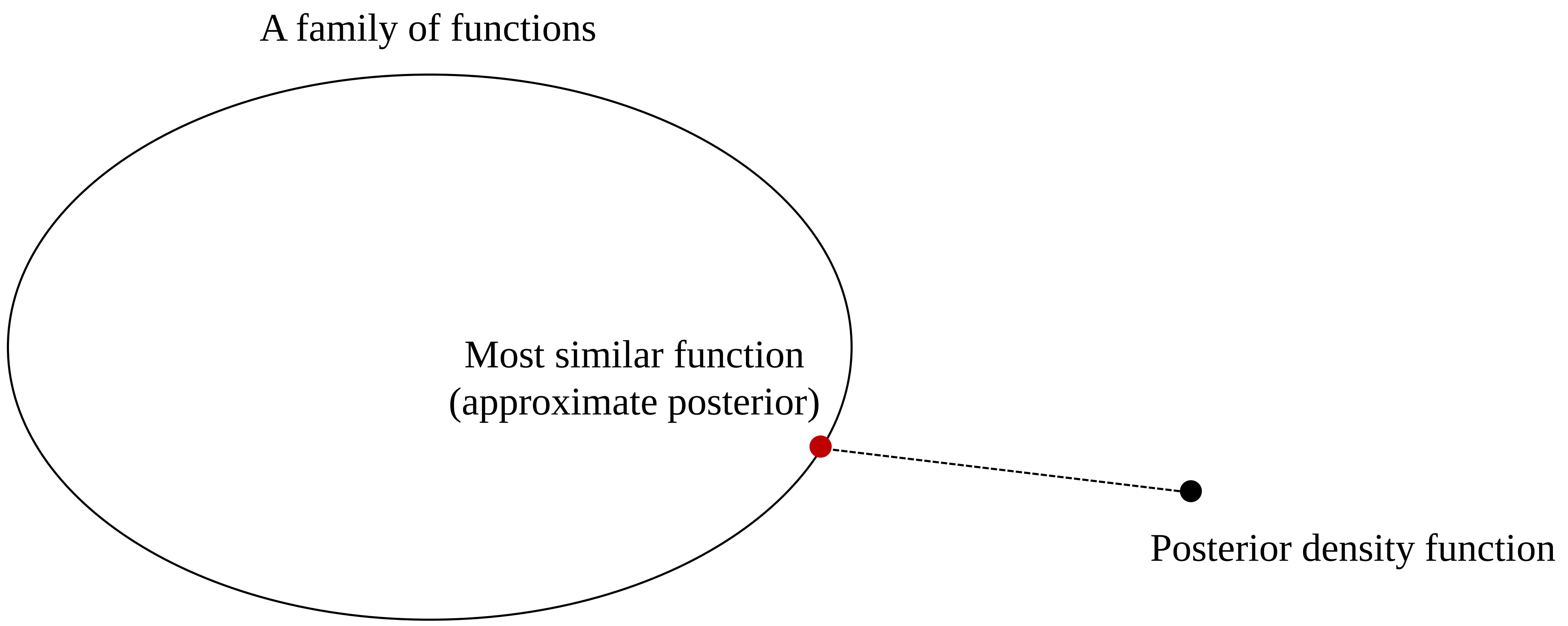
Methods for Intractable Posterior

Monte Carlo Sampling	Approximate Inference
• Monte Carlo sampling-based • Samples drawn from the true posterior (No function) • High computational cost for large data set or complex models	• Optimization-based • A function approximates the true posterior • Computationally efficient even if data set is l arge or model is complex • Accuracy of approximation is unknown • Suitable for trying many different models on l arge data set

Variational Inference

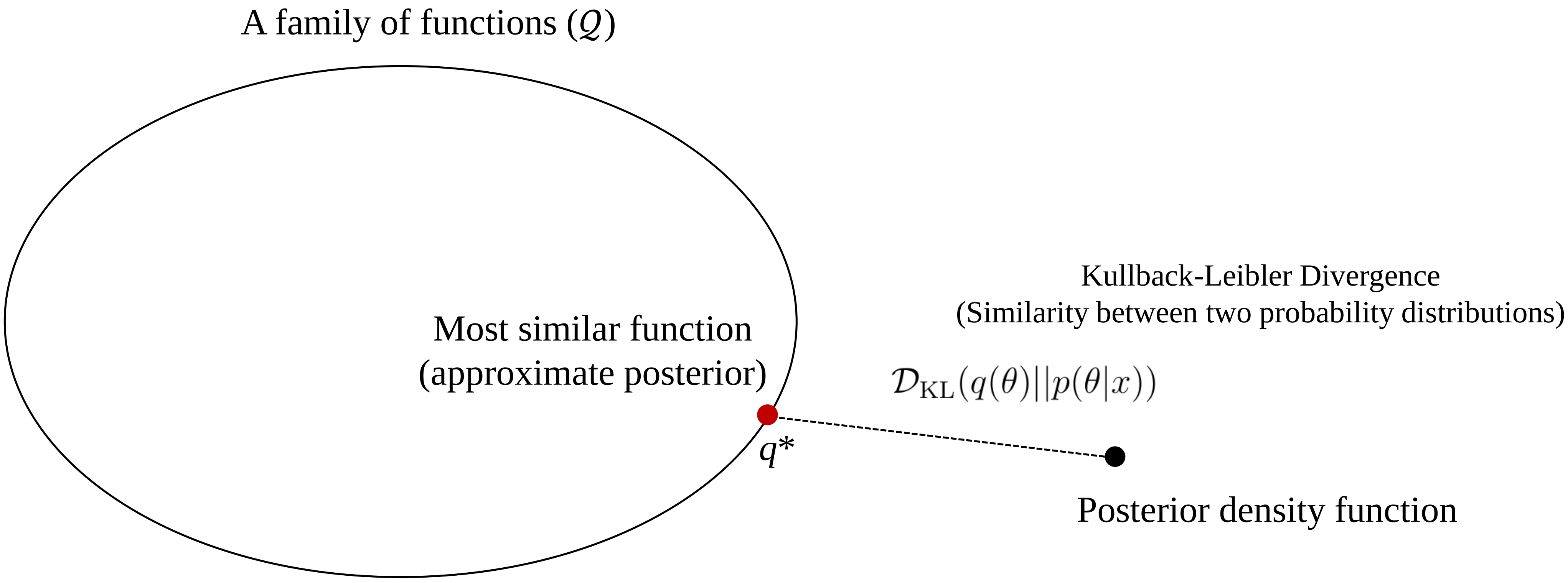
Variational Inference

Introduction



Variational Inference

Introduction



$$q^*(\theta) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \; \mathcal{D}_{\mathrm{KL}}(q(\theta)||p(\theta|x))$$

Variational Inference

Introduction

Optimal Approximation $q^*(\theta) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \ \mathcal{D}_{\mathrm{KL}}(q(\theta) || p(\theta|x))$

Class of functions

Target (Posterior)

What we do not know

→ ????

Variational Inference

Introduction

$$
\begin{aligned}
\mathcal{D}_{KL}(q(\theta)||p(\theta|x)) &= \int q(\theta) \log \frac{q(\theta)}{p(\theta|x)} d\theta \\
&= \int q(\theta) \log q(\theta) d\theta - \int q(\theta) \log p(\theta|x) d\theta \\
&= \mathbb{E}_{\theta \sim q(\theta)}[\log q(\theta)] - \mathbb{E}_{\theta \sim q(\theta)}[\log p(\theta|x)] \\
&= \mathbb{E}_{\theta \sim q(\theta)}[\log q(\theta)] - \mathbb{E}_{\theta \sim q(\theta)} \left[\log \frac{p(\theta, x)}{p(x)}\right] \\
&= \mathbb{E}_{\theta \sim q(\theta)}[\log q(\theta)] - \mathbb{E}_{\theta \sim q(\theta)}[\log p(\theta, x)] + \log p(x) \\
&= \mathbb{E}_{\theta \sim q(\theta)}[\log q(\theta)] - \mathbb{E}_{\theta \sim q(\theta)}[\log p(\theta) p(x|\theta)] + \log p(x)
\end{aligned}
$$

Variational Inference

Introduction

prior likelihood

$$\min_{q\in\mathcal{Q}} \mathcal{D}_{\mathrm{KL}}(q(\theta)||p(\theta|x)) \iff \min_{q\in\mathcal{Q}} \mathbb{E}_{\theta\sim q(\theta)}[\log q(\theta)] - \mathbb{E}_{\theta\sim q(\theta)}[\log p(\theta)p(x|\theta)]$$

Now we can solve the problem!

If q is parametrized by phi

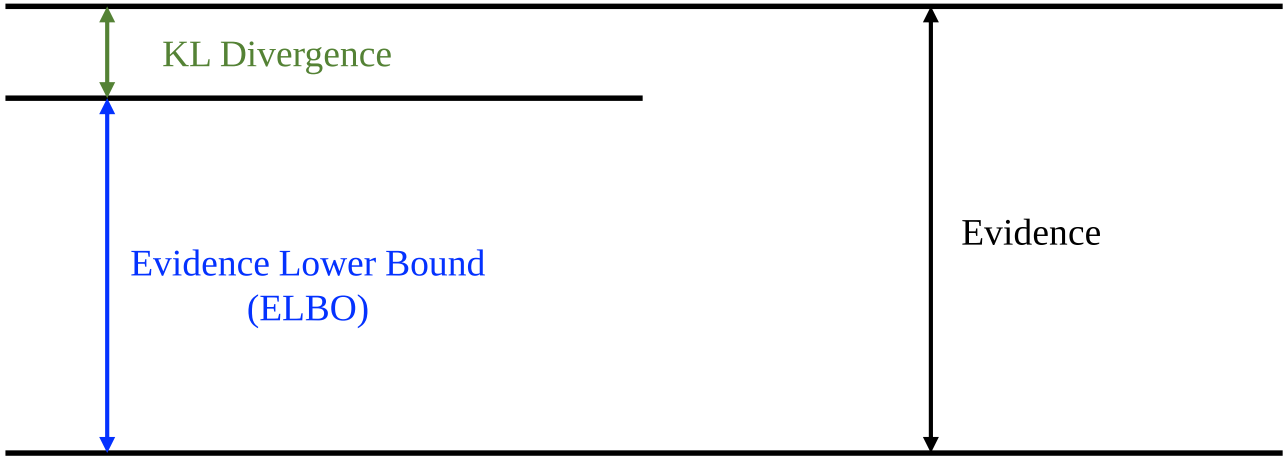
$$\min_{\phi} \mathcal{D}_{\mathrm{KL}}(q_\phi(\theta)||p(\theta|x)) \iff \min_{\phi} \mathbb{E}_{\theta\sim q_\phi(\theta)}[\log q_\phi(\theta)] - \mathbb{E}_{\theta\sim q_\phi(\theta)}[\log p(\theta)p(x|\theta)]$$

Variational Inference

Introduction

$$\mathcal{D}_{KL}(q(\theta)||p(\theta|x)) = \mathbb{E}_{\theta\sim q(\theta)}[\log q(\theta)] - \mathbb{E}_{\theta\sim q(\theta)}[\log p(\theta)p(x|\theta)] + \log p(x)$$

$$\underbrace{\log p(x)}_{\substack{\text{Evidence}\\ \text{(Constant)}}} = \underbrace{\mathcal{D}_{KL}(q(\theta)||p(\theta|x))}_{\substack{\text{KL Divergence}\\ \text{(Nonnegative)}}} \underbrace{- \mathbb{E}_{\theta\sim q(\theta)}[\log q(\theta)] + \mathbb{E}_{\theta\sim q(\theta)}[\log p(\theta)p(x|\theta)]}_{\text{Evidence Lower Bound (ELBO)}}$$



Minimizing KL Divergence = Maximizing ELBO

Variational Inference

Introduction

$$\min_{q \in \mathcal{Q}} \mathcal{D}_{\mathrm{KL}}(q(\theta)||p(\theta|x)) \iff \min_{q \in \mathcal{Q}} \mathbb{E}_{\theta \sim q(\theta)}[\log q(\theta)] - \mathbb{E}_{\theta \sim q(\theta)}[\log p(\theta)p(x|\theta)]$$

1. Determine a variational family Q (Members of Q should be tractable)

2. Solve the optimization problem

Variational Inference

Introduction

$$\min_{q\in\mathcal{Q}} \mathcal{D}_{\mathrm{KL}}(q(\theta)||p(\theta|x)) \iff \min_{q\in\mathcal{Q}} \mathbb{E}_{\theta\sim q(\theta)}[\log q(\theta)] - \mathbb{E}_{\theta\sim q(\theta)}[\log p(\theta)p(x|\theta)]$$

1. Determine a variational family Q

   → Mean-Field Variational Bayes (factorized density functions)

2. Solve the optimization problem

   → Coordinate ascent (traditional), convex relaxation (recent study)

$$q(\theta) = q(\theta_1, \cdots, \theta_k) = \prod_{i=1}^{k} q(\theta_i)$$

$$q(\theta_1, \theta_2, \theta_3, \theta_4) = q(\theta_1, \theta_2)q(\theta_3, \theta_4)$$

Factorized density functions

Variational Inference

Mean-Field Variational Bayes

- Bayesian Linear Regression

$$y_i \sim \mathcal{N}(x_i^T \boldsymbol{\beta}, \alpha^{-1})$$

Likelihood (linear model)

$$\left.\begin{aligned} \alpha &\sim \mathrm{Gamma}(a_0, b_0) \\ \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I}) \end{aligned}\right\} \text{Prior}$$

Parameter of interest: beta & alpha

= We want the posterior distribution of beta & alpha

$$p(\boldsymbol{\beta}, \alpha | \mathbf{X}, \mathbf{y})$$

Variational Inference

Mean-Field Variational Bayes

- Bayesian Linear Regression

Introduce variational distribution q, assuming factorized

$$p(\boldsymbol{\beta}, \alpha|\mathbf{X}, \mathbf{y}) \approx q(\boldsymbol{\beta}, \alpha) = q(\boldsymbol{\beta})q(\alpha)$$

$$\left.\begin{aligned} q(\boldsymbol{\beta}) &= \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\mu}, \Sigma) \\ q(\alpha) &= \mathrm{Gamma}(\alpha|a, b) \end{aligned}\right\}$$

Arbitrarily chosen
We “use” these density functions as approximators

Variational Inference

Mean-Field Variational Bayes

- Bayesian Linear Regression

Introduce variational distribution q, assuming factorized

$$p(\boldsymbol{\beta}, \alpha | \mathbf{X}, \mathbf{y}) \approx q(\boldsymbol{\beta}, \alpha) = q(\boldsymbol{\beta})q(\alpha)$$

$$q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta} | \boldsymbol{\mu}, \Sigma)$$
$$q(\alpha) = \mathrm{Gamma}(\alpha | a, b)$$

We want to solve the following optimization problem

$$\max_{\boldsymbol{\mu}, \Sigma, a, b} \mathrm{ELBO}(q(\boldsymbol{\beta}, \alpha))$$

Variational Inference

Mean-Field Variational Bayes

- Bayesian Linear Regression

$$\max_{\boldsymbol{\mu},\Sigma,a,b} \ \mathrm{ELBO}(q(\boldsymbol{\beta},\alpha)) \quad \text{s.t.} \quad a,b>0,\ \Sigma \succeq 0$$

$$\begin{aligned}
\mathrm{ELBO}(q(\boldsymbol{\beta},\alpha)) &= (a_0-1)(\psi(a)-\log b) - b_0\frac{a}{b} - \frac{\lambda}{2}(\boldsymbol{\mu}^T\boldsymbol{\mu} + \mathrm{tr}(\Sigma)) \\
&= \frac{n}{2}(\psi(a)-\log b) - \frac{a}{2b}\sum_{i=1}^{n}\left((y_i - x_i^T\boldsymbol{\mu})^2 + x_i^T\Sigma x_i\right) \\
&= a - \log b + \log\Gamma(a) + (1-a)\psi(a) + \frac{1}{2}\log|\Sigma| + \mathrm{const}
\end{aligned}$$

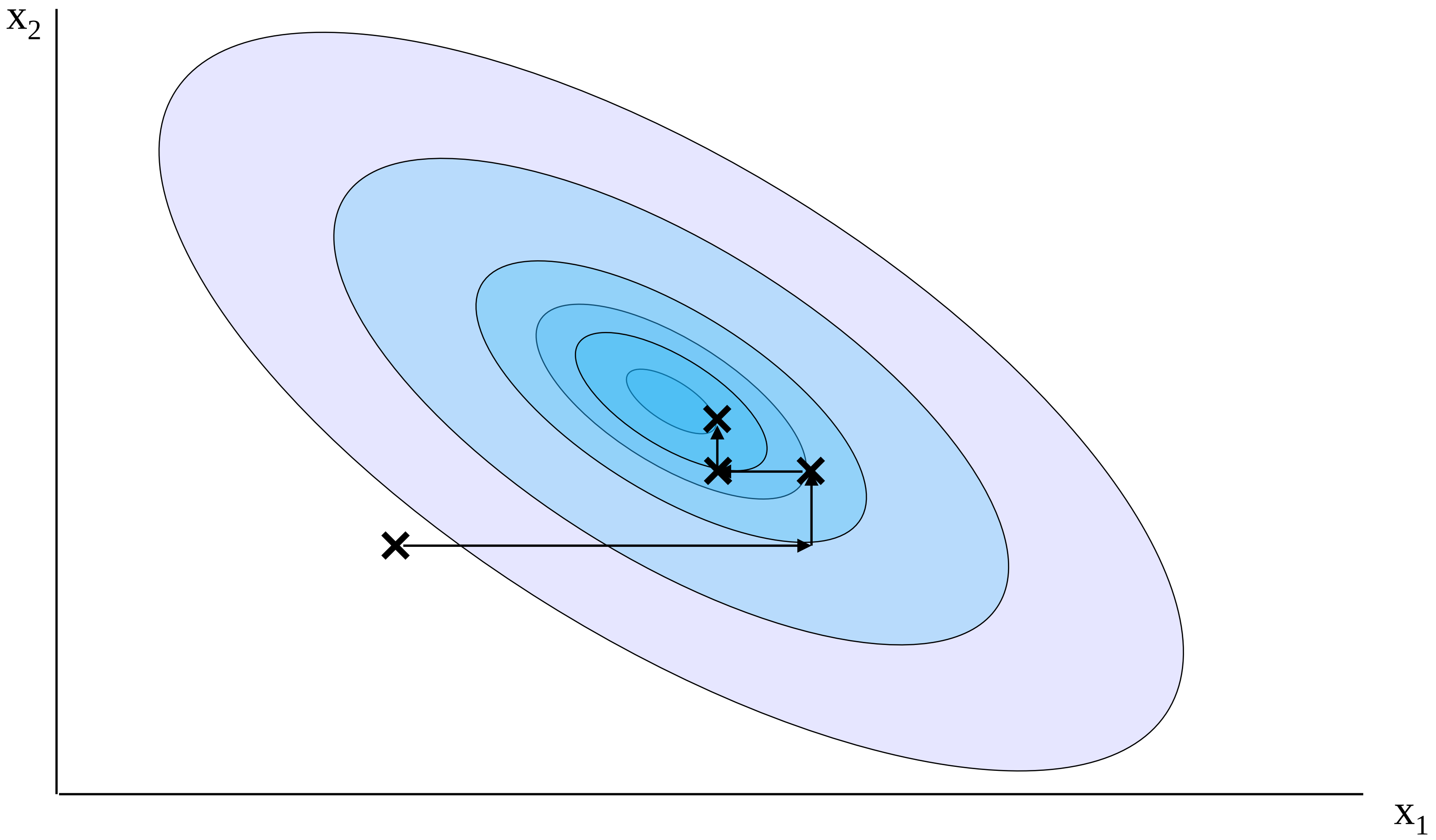
Usually, **coordinate ascent** is used to maximize this problem

Variational Inference

Mean-Field Variational Bayes

- Bayesian Linear Regression

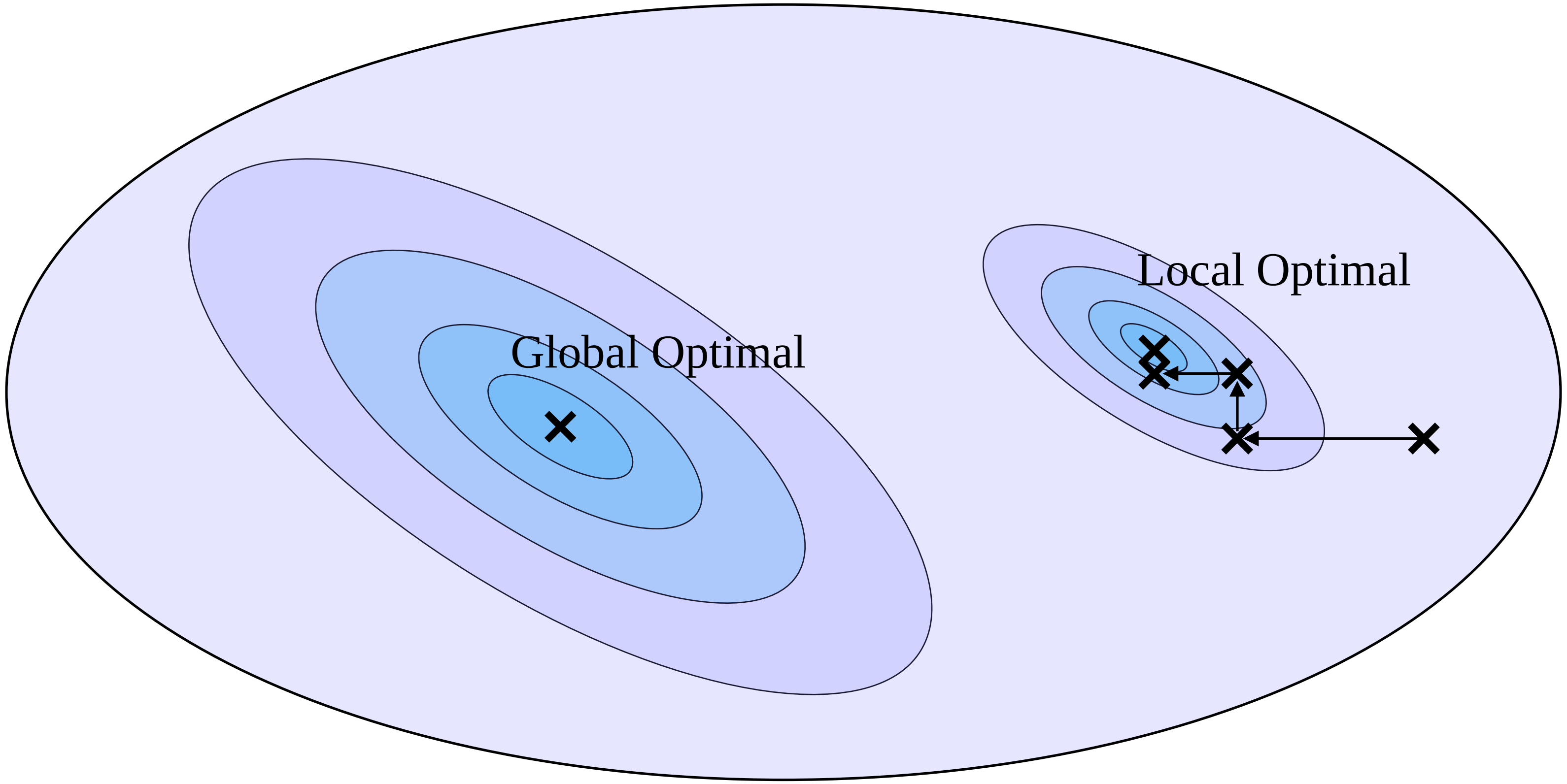
* Coordinate Ascent/Descent: Move along one axis (optimize w.r.t. one variable) at each step



Variational Inference

Mean-Field Variational Bayes

- Bayesian Linear Regression



Coordinate ascent/descent is not guaranteed to converge global optimal for nonconvex optimization problem

local optimal → may not be small enough KL divergence → not good approximation

Variational Inference

Convex Relaxation for Variational Inference (ICML 2018)

CRVI: Convex Relaxation for Variational Inference

Ghazal Fazelnia [1] **John Paisley** [1]

Abstract

We present a new technique for solving non-convex variational inference optimization problems. Variational inference is a widely used method for posterior approximation in which the inference problem is transformed into an optimization problem. For most models, this optimization is highly non-convex and so hard to solve. In this paper, we introduce a new approach to solving the variational inference optimization based on convex relaxation and semidefinite programming. Our theoretical results guarantee very tight relaxation bounds that get nearer to the global optimal solution than traditional coordinate ascent. We evaluate the performance of our approach on regression and sparse coding.

convexities in variational inference (VI) optimization for conjugate models that achieve near globally optimal solutions. Our method is based on convex relaxation and semidefinite programming (SDP). In our approach, an SDP relaxation converts a non-convex polynomial optimization of vector parameters to a convex optimization with matrix parameters via a lifting technique. We call this approach convex relaxation for variational inference (CRVI). The exactness of the relaxation can then be interpreted as the existence of a low-rank solution to this SDP. Our main contribution is to solve this variational optimization problem in an accurate way and provide theoretical guarantees for the exactness of our solution using graph theoretic tools. To the best of our knowledge, this is the first time that a relaxation for variational inference could guarantee and produce optimal solutions that are either globally optimal solution or very close to it. Our experimental results demonstrate the effectiveness of CRVI compared with coordinate ascent for sparse regression and sparse coding models.

1. Introduction

Variational Inference

Convex Relaxation for Variational Inference

- Bayesian Linear Regression

$$\max_{\boldsymbol{\mu},\Sigma,a,b} \ \mathrm{ELBO}(q(\boldsymbol{\beta},\alpha)) \quad \text{s.t.} \quad a, b > 0, \ \Sigma \succeq 0$$

$$\begin{aligned}
\mathrm{ELBO}(q(\boldsymbol{\beta},\alpha)) &= (a_0 - 1)(\psi(a) - \log b) - b_0 \frac{a}{b} - \frac{\lambda}{2}(\boldsymbol{\mu}^T \boldsymbol{\mu} + \mathrm{tr}(\Sigma)) \\
&= \frac{n}{2}(\psi(a) - \log b) - \frac{a}{2b} \sum_{i=1}^{n} \left((y_i - x_i^T \boldsymbol{\mu})^2 + x_i^T \Sigma x_i \right) \\
&= a - \log b + \log \Gamma(a) + (1 - a)\psi(a) + \frac{1}{2} \log |\Sigma| + \mathrm{const}
\end{aligned}$$

Utilize **convex relaxation** instead of coordinate ascent to find a better solution of the problem

Variational Inference

Convex Relaxation for Variational Inference

- Bayesian Linear Regression

Utilize **convex relaxation** instead of coordinate ascent to find a better solution of the problem

- A better solution (higher ELBO) than coordinate ascent in several cases
- Theoretically proved optimality gap (solution quality can be measured)
- Relatively slower but not significantly slow
- NOT as simple as coordinate ascent
- Remarkable trial to converge optimization and statistics

Variational Inference

Convex Relaxation for Variational Inference

- Bayesian Linear Regression

Utilize **convex relaxation** instead of coordinate ascent to find a better solution of the problem

Main Idea: Nonconvexities usually originated from polynomial terms

→ Polynomial terms can be represented by quadratic terms

$$\min_{x \in \mathbb{R}^d} f_0(x)$$

$$\text{subject to } f_k(x) \le 0 \text{ for } k = 1, \dots, K,$$

$$\text{where } f_k = x^\top A_k x + b_k^\top x + c_k \text{ for } k = 0, \dots, K.$$

If all the matrices A_k, $k=0, \dots, K$ are positive semidefinite, the problem is convex (but usually not)

Variational Inference

Convex Relaxation for Variational Inference

- Bayesian Linear Regression

Utilize **convex relaxation** instead of coordinate ascent to find a better solution of the problem

The problem can be reformulated as follows:

$$F_K = \begin{bmatrix} c_k & \frac{1}{2}b_k^\top \\ \frac{1}{2}b_k & A_k \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x^\top \\ x & xx^\top \end{bmatrix}$$

$$\begin{aligned} \min_{X \in \mathbb{R}^{(d+1)\times(d+1)}} \quad & \mathrm{trace}(F_0 X) \\ \text{subject to} \quad & \mathrm{trace}(F_k X) \leq 0 \ \text{ for } k = 1, .., K, \\ & X_{1,1} = 1, \ \ X \succeq 0, \\ & \boxed{\mathrm{rank}(X) = 1.} \end{aligned}$$

Nonconvex part

Variational Inference

Convex Relaxation for Variational Inference

- Bayesian Linear Regression

First, reformulate the problem as follows

$$\min_{\boldsymbol{\mu},\Sigma,a,b,c,e} \quad \sum_{i=1}^{n} \frac{1}{2}(ey_i^2 - x_i^T u + x_i^T u \boldsymbol{\mu}^T x_i) + x_i^T e \Sigma x_i)$$

$$+ \frac{\lambda}{2}(\boldsymbol{\mu}^T \boldsymbol{\mu} + \mathrm{tr}(\Sigma)) + b_0 e$$

$$- (a_0 - 1)(\psi(a) + \log c) - \frac{n}{2}(\psi(a) + \log c)$$

$$- a - \log c - \log \Gamma(a) - (1 - a)\psi(a) - \frac{1}{2}|\Sigma|$$

$$\text{s.t.} \quad a, c, e > 0, \quad \Sigma \succeq 0, \quad e = ac, \quad u = e\mu, \;\; c = b^{-1}$$

Variational Inference

Convex Relaxation for Variational Inference

- Bayesian Linear Regression

Nonconvex terms are in the first two lines

$$
\begin{aligned}
\min_{\boldsymbol{\mu},\Sigma,a,b,c,e} \quad & \left.\begin{aligned} & \sum_{i=1}^{n} \frac{1}{2}\left(ey_i^2 - x_i^T u + x_i^T u \boldsymbol{\mu}^T x_i\right) + x_i^T e \Sigma x_i) \\ & + \frac{\lambda}{2}\left(\boldsymbol{\mu}^T \boldsymbol{\mu} + \operatorname{tr}(\Sigma)\right) + b_0 e \end{aligned}\right\} \text{Nonconvex} \\
& \left.\begin{aligned} & - (a_0 - 1)(\psi(a) + \log c) - \frac{n}{2}(\psi(a) + \log c) \\ & - a - \log c - \log \Gamma(a) - (1-a)\psi(a) - \frac{1}{2}|\Sigma| \end{aligned}\right\} \text{Convex} \\
\text{s.t.} \quad & a, c, e > 0, \quad \Sigma \succeq 0, \quad e = ac, \quad u = e\mu, \;\; c = b^{-1}
\end{aligned}
$$

Variational Inference

Convex Relaxation for Variational Inference

- Bayesian Linear Regression

Utilize the following vector,

$$\boldsymbol{\nu} = [1, a, c, e, \boldsymbol{\mu}^T, \boldsymbol{u}^T, \Sigma_{1,1}, \Sigma 1, 2, \cdots, \Sigma_{p,p}]^T$$

$$\begin{aligned}
\min_{\boldsymbol{\nu}, a, c, \Sigma} \quad & f_{\text{CR}}(\boldsymbol{\nu}) + g(a, c, \Sigma) \\
\text{s.t.} & a, c, e > 0, \quad \Sigma \succeq 0, \quad e = ac, \quad u = e\boldsymbol{\mu}, \quad c = b^{-1} \\
& a = \nu_2, \quad c = \nu_3 \\
& \text{vector}(\Sigma) = [\nu_{5+2*p}, \cdots, \nu_{4+2p+p^2}]
\end{aligned}$$

$$\mathbf{A} := \nu \times \nu^\top \in \mathbb{S}^{(4+2d+d^2)\times(4+2d+d^2)}$$

Variational Inference

Convex Relaxation for Variational Inference

- CRVI vs. CAVI

Table 1. Information about the datasets, running time of the algorithms, and rank of the found solution using CRVI. We see that CRVI is slower than CAVI (coordinate ascent). However, the rank of the found CRVI solution is near 1 (and less than the theoretical upper bound of 3), indicating a solution nearer the global optimum. This is confirmed in Figure 2.

DataSet	Dim.	# of Samples	CAVI time (s)	CRVI time (s)	Rank
Birth Rate & Econ	4	30	0.281	1.115	1.11
Iris	4	150	0.231	1.807	1.20
Yacht	6	308	0.402	2.111	1.10
Pima Indian Diabetes	8	768	0.571	3.040	1.67
Bike Sharing	13	731	0.884	6.749	1.61
Parkinson	21	5875	0.962	7.309	1.98
WDBC	31	569	1.059	10.766	1.73
Online News Popularity	58	39644	9.341	15.223	1.52
Year Prediction Songs	90	515345	18.809	22.050	1.78

Variational Inference

Convex Relaxation for Variational Inference

- CRVI vs. CAVI

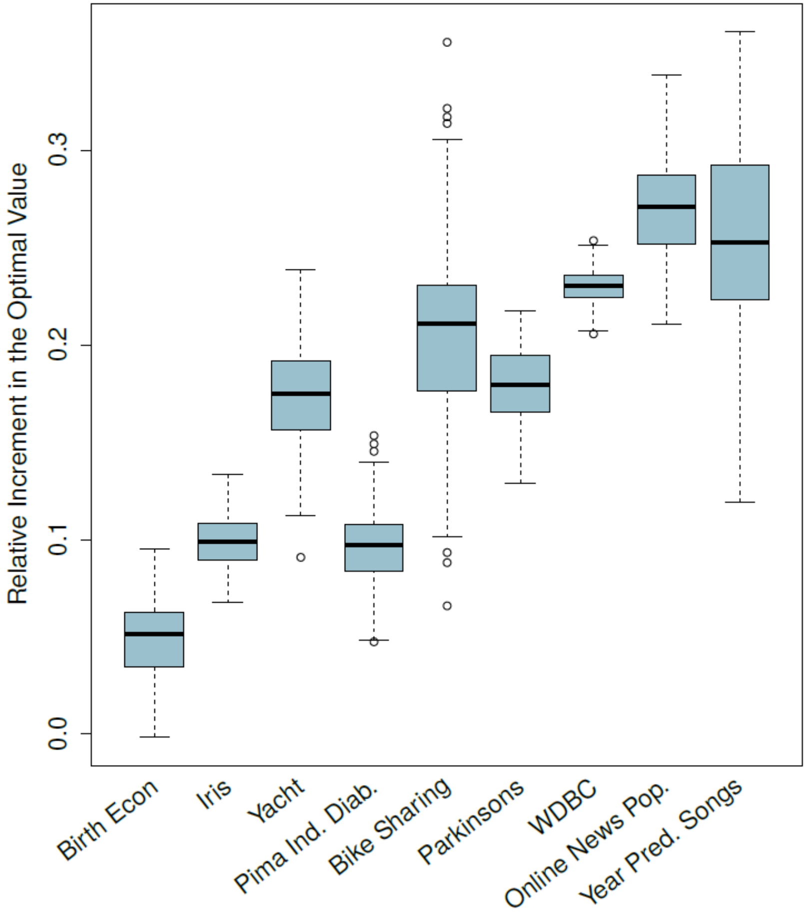


Figure 2. Boxplot of relative improvement in the calculated local optimal value of CRVI compared to CAVI. Each box represents the summary of the fractional improvement of CRVI over CAVI for 100 simulations using different prior hyper-parameters and initializations. After calculating the respective local optimal variational objective functions, the value found by CAVI is subtracted from the value from CRVI and divided by the values from CAVI to obtain the relative improvement score. As is evident, CRVI gave a significant improvement over CAVI.

Variational Inference

Convex Relaxation for Variational Inference

- Variational inference transforms an inference problem into an optimization problem
- For most models, associated variational optimization problem is highly nonconvex
- CRVI guarantee very tight relaxation bounds that get nearer to the global optimal solution than traditional coordinate ascent
- Good example of the convergence of statistics and optimization technique