Lectures 1–2: Bayesian inference and MCMC foundations

Youssef Marzouk

Department of Aeronautics and Astronautics
Center for Computational Engineering
Statistics and Data Science Center

Massachusetts Institute of Technology
http://uqgroup.mit.edu
ymarz@mit.edu

19–22 March 2018

Agenda

Plan for the lectures:

1. Lectures 1–2: Bayesian inference and MCMC foundations
   - Bayesian modeling
   - Computational approaches/demos
2. Lecture 3: Bayesian approach to *inverse problems*
   - What distinguishes inverse problems? Elements of a Bayesian inverse problem formulation
   - Linear–Gaussian problems in detail
   - Computational issues: surrogate modeling/likelihood approximations, parameter dimension reduction, ...
3. Lecture 4: Bayesian optimal experimental design *or some other topic TBD*

Statistical inference

Why is a statistical perspective useful in data assimilation?

- To characterize *uncertainty* in the *parameters* and/or *state* of a system
  - To understand how this uncertainty depends on the number and quality of observations, features of the model, prior information, etc.
- To make probabilistic *predictions*
- To choose useful observations or experiments
- To address questions of model error and model validity; to perform model selection

Bayesian inference

Bayes' rule

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

Key idea: model parameters θ are treated as random variables
(For simplicity, we let our random variables have densities)

Notation

- θ are model parameters; y are the data; assume both to be *finite-dimensional* unless otherwise indicated
- $p(\theta)$ is the *prior* probability density
- $L(\theta) := p(y|\theta)$ is the likelihood function
- $p(\theta|y)$ is the *posterior* probability density
- $p(y)$ is the *evidence*, or equivalently, the *marginal likelihood*

Bayesian inference example

Infer the bias $\theta \in [0,1]$ of a coin, given flip outcomes $(y_i)_{i=1}^n \in \{0,1\}^n$.
Convention: outcome $y_i = 1$ is "heads" and $y_i = 0$ is "tails."

Elements of the Bayesian formulation:

- **Likelihood function**
  - Single observation: $Y_i \sim \text{Ber}(\theta)$, where $\theta := \mathbb{P}[Y_i = 1]$. Hence:

$$P(y_i|\theta) = \theta^{y_i}(1-\theta)^{(1-y_i)}$$

  - Multiple observations are conditionally independent given θ:

$$P(y_1, \ldots, y_n|\theta) = \prod_{i=1}^n P(y_i|\theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}$$

  - Rewrite more compactly, with $k := \sum_{i=1}^n y_i$ heads in n trials:

$$P(k|\theta, n) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad \text{i.e., } K \sim \text{Binomial}(\theta, n)$$

Bayesian inference example

Infer the bias $\theta \in [0,1]$ of a coin, given flip outcomes $(y_i)_{i=1}^n \in \{0,1\}^n$.
Convention: outcome $y_i = 1$ is "heads" and $y_i = 0$ is "tails."

Elements of the Bayesian formulation:

- **Likelihood function**
  - Single observation: $Y_i \sim \text{Ber}(\theta)$, where $\theta := \mathbb{P}[Y_i = 1]$. Hence:

$$P(y_i|\theta) = \theta^{y_i}(1-\theta)^{(1-y_i)}$$

  - Multiple observations are conditionally independent given θ:

$$P(y_1, \ldots, y_n|\theta) = \prod_{i=1}^n P(y_i|\theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}$$

  - Rewrite more compactly, with $k := \sum_{i=1}^n y_i$ heads in n trials:

$$P(k|\theta, n) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad \text{i.e., } K \sim \text{Binomial}(\theta, n)$$

Bayesian inference example

Infer the bias $\theta \in [0,1]$ of a coin, given flip outcomes $(y_i)_{i=1}^n \in \{0,1\}^n$.
Convention: outcome $y_i = 1$ is "heads" and $y_i = 0$ is "tails."

Elements of the Bayesian formulation:

- **Likelihood function**
  - Single observation: $Y_i \sim \text{Ber}(\theta)$, where $\theta := \mathbb{P}[Y_i = 1]$. Hence:

$$P(y_i|\theta) = \theta^{y_i}(1-\theta)^{(1-y_i)}$$

  - Multiple observations are conditionally independent given θ:

$$P(y_1, \ldots, y_n|\theta) = \prod_{i=1}^n P(y_i|\theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i}$$

  - Rewrite more compactly, with $k := \sum_{i=1}^n y_i$ heads in n trials:

$$P(k|\theta, n) = \binom{n}{k}\theta^k(1-\theta)^{n-k}, \quad \text{i.e., } K \sim \text{Binomial}(\theta, n)$$

Bayesian inference example

Infer the bias $\theta \in [0,1]$ of a coin, given flip outcomes $(y_i)_{i=1}^n \in \{0,1\}^n$.
Convention: outcome $y_i = 1$ is "heads" and $y_i = 0$ is "tails."

Elements of the Bayesian formulation:

- **Prior distribution**
  - Use $\Theta \sim \text{Beta}(\beta_1, \beta_2)$:

$$p(\theta \,|\, \beta_1, \beta_2) \propto \theta^{\beta_1 - 1}(1-\theta)^{\beta_2 - 1}$$

  - Uniform distribution is a special case: $\text{Beta}(1,1) = \mathcal{U}(0,1)$
- **Posterior distribution**
  - Posterior density follows from simple algebra:

$$p(\theta|k,n,\beta_1,\beta_2) \propto p(k|\theta,n)p(\theta \,|\, \beta_1,\beta_2) \propto \theta^{k+\beta_1-1}(1-\theta)^{n-k+\beta_2-1}$$

    i.e., $\Theta|k,n \sim \text{Beta}(\beta_1 + k, \beta_2 + n - k)$
  - This happens to be a *conjugate* Bayesian model!

Bayesian inference example

Infer the bias $\theta \in [0,1]$ of a coin, given flip outcomes $(y_i)_{i=1}^n \in \{0,1\}^n$.
Convention: outcome $y_i = 1$ is "heads" and $y_i = 0$ is "tails."

Elements of the Bayesian formulation:

- **Prior distribution**
  - Use $\Theta \sim \text{Beta}(\beta_1, \beta_2)$:

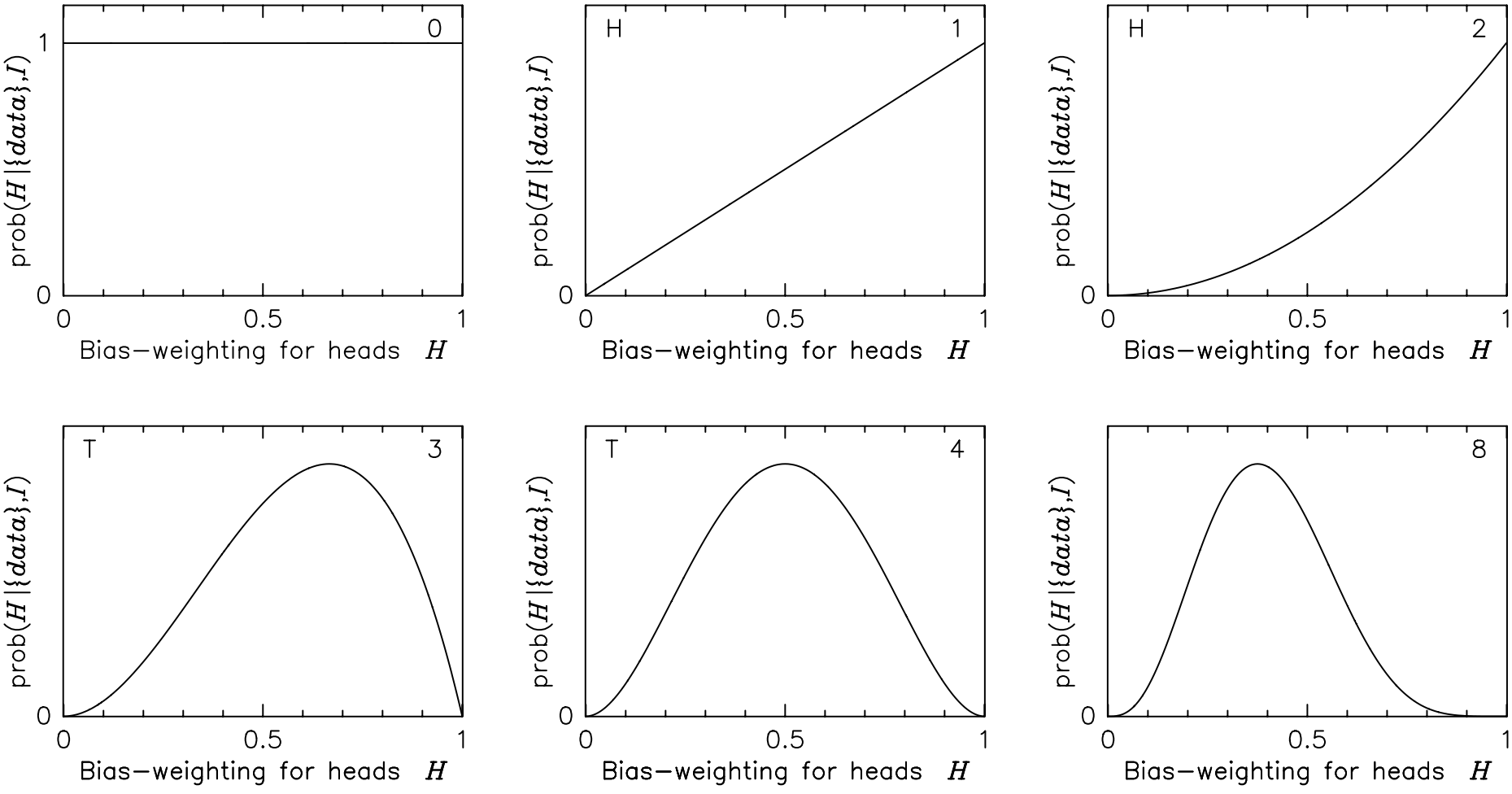
$$p(\theta \,|\, \beta_1, \beta_2) \propto \theta^{\beta_1 - 1}(1-\theta)^{\beta_2 - 1}$$

  - Uniform distribution is a special case: $\text{Beta}(1,1) = \mathcal{U}(0,1)$
- **Posterior distribution**
  - Posterior density follows from simple algebra:

$$p(\theta | k, n, \beta_1, \beta_2) \propto p(k|\theta, n) p(\theta \,|\, \beta_1, \beta_2) \propto \theta^{k+\beta_1 - 1}(1-\theta)^{n-k+\beta_2 - 1}$$

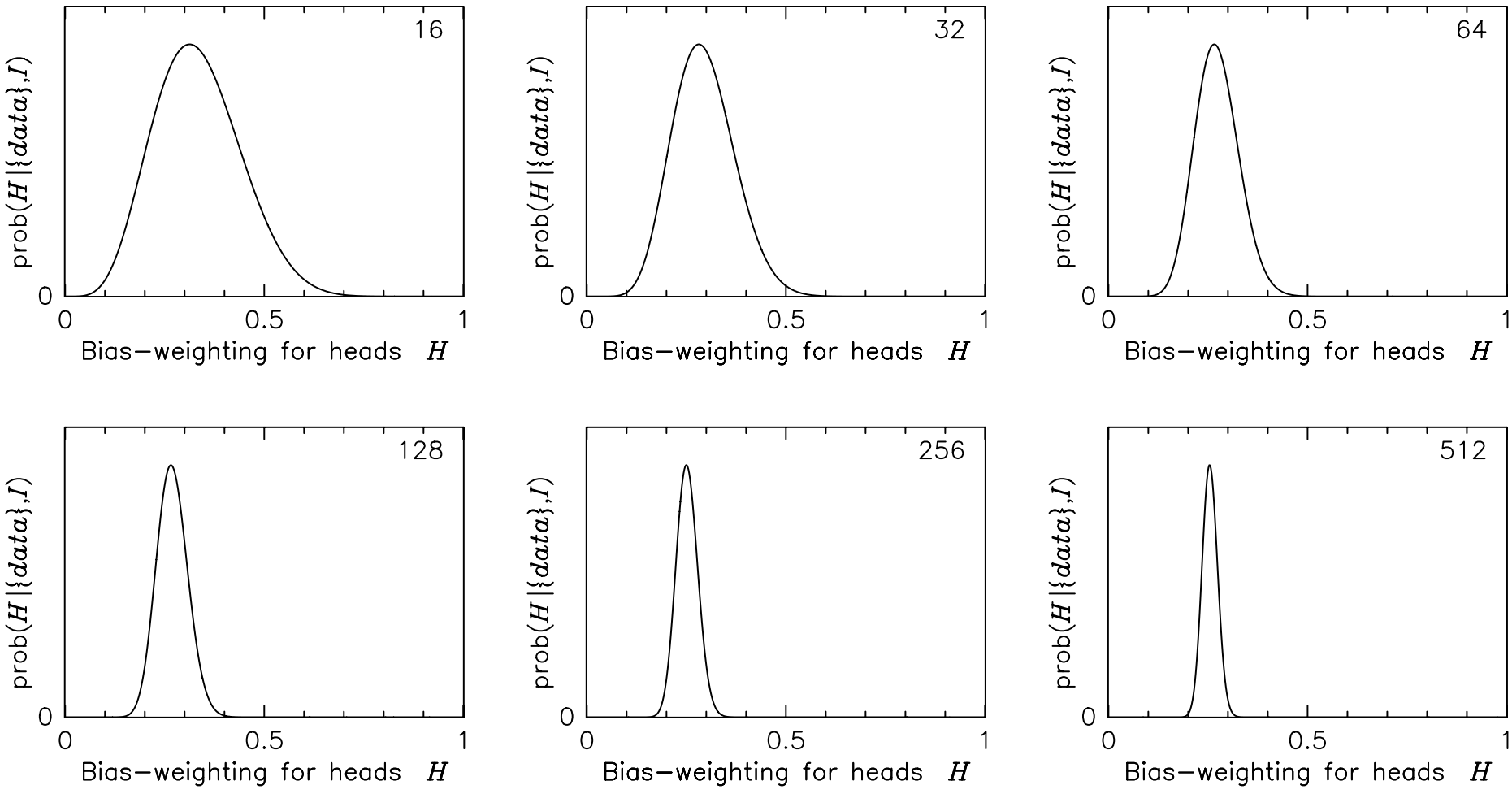
    i.e., $\Theta | k, n \sim \text{Beta}(\beta_1 + k, \beta_2 + n - k)$
  - This happens to be a *conjugate* Bayesian model!

Coin flip example 1/2 [Sivia 2006]



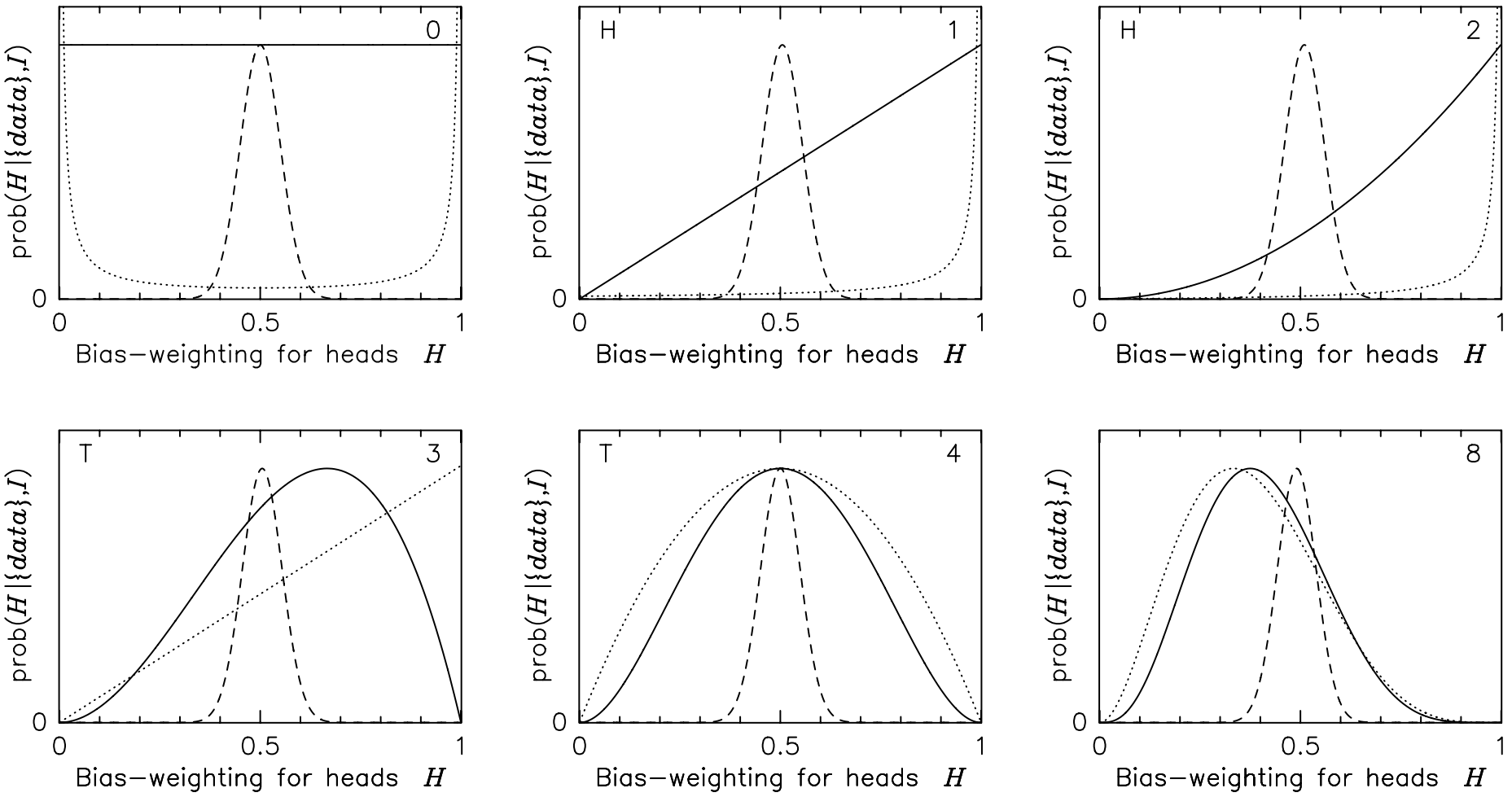
from Sivia & Skilling, *Data Analysis: a Bayesian Tutorial*, Oxford UP (2006).

Coin flip example 1/2 [Sivia 2006]



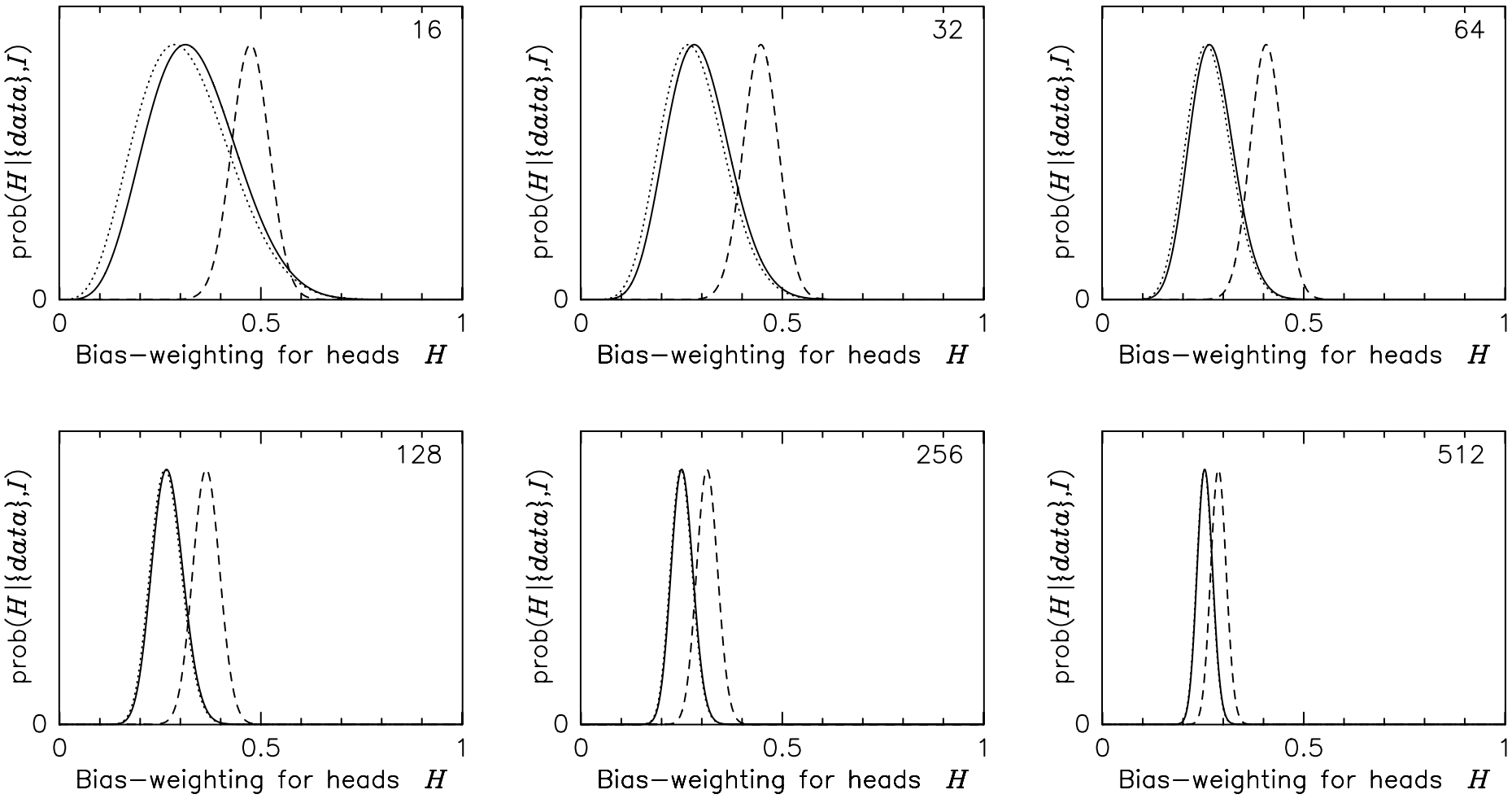
from Sivia & Skilling, *Data Analysis: a Bayesian Tutorial*, Oxford UP (2006).

Coin flip example 2/2 [Sivia 2006]



from Sivia & Skilling, *Data Analysis: a Bayesian Tutorial*, Oxford UP (2006).

Coin flip example 2/2 [Sivia 2006]



from Sivia & Skilling, *Data Analysis: a Bayesian Tutorial*, Oxford UP (2006).

Linear regression example

Bayesian linear regression

$$y_i = \theta_0 + \boldsymbol{\theta}_1^\top \mathbf{x}_i + \xi_i, \quad y_i, \theta_0 \in \mathbb{R};\ \mathbf{x}_i, \boldsymbol{\theta}_1 \in \mathbb{R}^d;\ \xi_i \sim N(0, \sigma^2)$$

- Let the data consist of n observations $\mathcal{D}_n \equiv \{(y_i, \mathbf{x}_i)\}_{i=1}^n$
- Define the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with rows $\mathbf{x}_i$, $\mathbf{y} \in \mathbb{R}^n$ as a column vector of $y_1 \dots y_n$, $\bar{\boldsymbol{\theta}} = [\theta_0; \boldsymbol{\theta}_1] \in \mathbb{R}^{d+1}$, $\mathbf{1} \in \mathbb{R}^n$ as a vector of ones, and $\mathbf{I}_n$ as the n-dimensional identity matrix

- Bayesian model: likelihood and prior:

$$\begin{aligned} \mathbf{y} \,|\, \theta_0, \boldsymbol{\theta}_1 &\sim N\left(\mathbf{1}\theta_0 + \mathbf{X}\boldsymbol{\theta}_1, \sigma^2 \mathbf{I}_n\right) \\ \bar{\boldsymbol{\theta}} &\sim N\left(\boldsymbol{\mu}_{pr}, \boldsymbol{\Sigma}_{pr}\right) \end{aligned}$$

- Yields the joint ($d+1$–dimensional) posterior distribution of constant term θ_0 and "slopes" $\boldsymbol{\theta}_1$

Summaries of the posterior distribution

What information to extract?

- Posterior mean of θ; maximum *a posteriori* (MAP) estimate of θ
- Posterior covariance or higher moments of θ
- Quantiles
- Credibile intervals: $C(y)$ such that $\mathbb{P}\left[\theta \in C(y) \,|\, y\right] = 1 - \alpha$.
  - Credible intervals are not uniquely defined above; thus consider, for example, the HPD (highest posterior density) region.
- Posterior realizations: for direct assessment, or to estimate posterior expectations

Bayesian and frequentist paradigms

Understanding both perspectives is useful and important...

Key differences between these two statistical paradigms

- Frequentists do **not** assign probabilities to unknown parameters θ. One can write likelihoods $p_\theta(y) \equiv p(y|\theta)$ but not priors $p(\theta)$ or posteriors. θ is *not* a random variable.
- In the frequentist viewpoint, there is no single preferred methodology for inverting the relationship between parameters and data. Instead, consider various **estimators** $\hat{\theta}(y)$ of θ.
- The estimator $\hat{\theta}$ *is* a random variable. Why? Frequentist paradigm considers y to result from a random and repeatable experiment.

Bayesian and frequentist paradigms

Key differences (continued)

- Evaluate quality of $\hat{\theta}$ through various criteria: bias, variance, mean-square error, consistency, efficiency, ...
- One common frequentist approach is *maximum likelihood* estimation: $\hat{\theta}_{ML} = \text{argmax}_\theta \, p(y|\theta)$. (View $p(y|\theta)$ as a family of distributions indexed by θ.)
- Link to Bayesian approach: MAP estimate maximizes a "penalized likelihood."
- What about Bayesian versus frequentist prediction of $y_{new} \perp\!\!\!\perp y \,|\, \theta$?
  - Frequentist: use "plug-in" estimate of θ
  - Bayesian: posterior prediction via integration

Statistical problems

Canonical statistical problems

- **Density estimation:** observe realizations $\{y^{(i)}\}$ of a random variable Y and use them to learn the probability distribution (density) of Y. Parametric (e.g., $p_\theta(y)$) and nonparametric approaches.
- **Regression:** observe dependence of a *response* or *output* variable Y on a *covariate* or *input* variable X. Consider a model $p(y|x,\theta)$; learn θ and predict future $y|x$.
- **Classification:** like regression, but response variable ranges over a finite set.

Not all statistical problems fall cleanly into one of these three categories. But core aspects of these problems are worth studying!

Bayesian inference

Likelihood functions (initial summary)

- In general, $p(y|\theta) = p_\theta(y)$ is a probabilistic model for the data
- *Preview:* in inverse problems, the likelihood function is where the *forward model* appears, along with a noise model and (if applicable) an expression for model discrepancy
- *Preview:* in filtering, the likelihood function might be simpler (e.g., direct noisy observations of the state)

Prior distributions (initial summary)

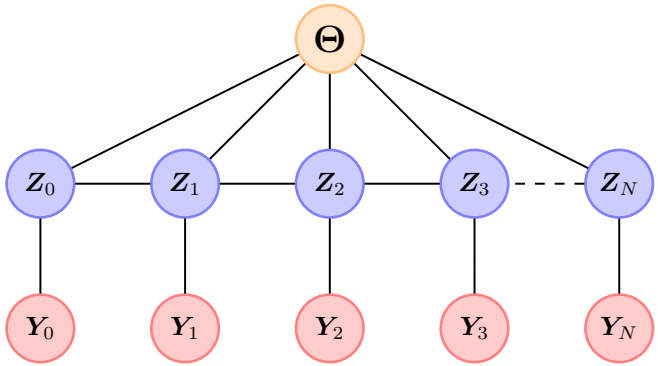
- **Much** can be written about choosing priors.
- Intuitive idea: assign lower probability to neighborhoods of θ that you don't expect to see, higher probability to neighborhoods of θ that you *do* expect to see.
- *Preview:* in ill-posed parameter estimation problems, e.g., inverse problems, prior information plays a key role!
- *Preview:* in filtering problems, the prior is often the result of "applying" the dynamics to an earlier distribution on the state

Hierarchical modeling

- One of the key flexibilities of the Bayesian construction!
- Hierarchical modeling has important implications for the design of efficient MCMC samplers (later in the lecture)
- Examples:
  1. Unknown noise variance
  2. Unknown scale of the prior (cf. choosing the **regularization** parameter in an inverse problem)
  3. Many more, as dictated by the physical and statistical models at hand

Hierarchical modeling example

- State-space model with static parameters

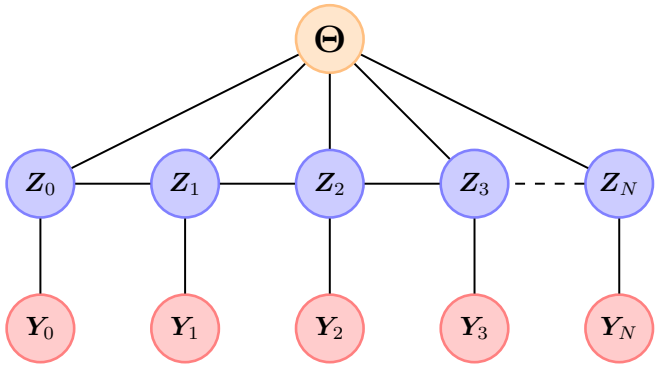


- Ingredients of the Bayesian model:
  - Transition density $\pi_{Z_k|Z_{k-1},\Theta}$
  - Observation density (likelihood) $\pi_{Y_k|Z_k}$
  - Prior on static parameters π_Θ
  - Prior on initial condition π_{Z_0}
- Posterior density:

$$\pi_{Z_{0:N},\Theta\,|\,y_{0:N}} \propto \pi_\Theta \pi_{Z_0} \left(\prod_{k=1}^{N} \pi_{Z_k|Z_{k-1},\Theta} \right) \left(\prod_{j=1}^{N} \pi_{y_j|Z_j} \right)$$

Hierarchical modeling example

- State-space model with static parameters



- Ingredients of the Bayesian model:
  - Transition density $\pi_{Z_k|Z_{k-1},\Theta}$
  - Observation density (likelihood) $\pi_{Y_k|Z_k}$
  - Prior on static parameters π_Θ
  - Prior on initial condition π_{Z_0}
- Posterior density:

$$\pi_{Z_{0:N},\Theta\,|\,y_{0:N}} \propto \pi_\Theta \pi_{Z_0} \left(\prod_{k=1}^{N} \pi_{Z_k|Z_{k-1},\Theta} \right) \left(\prod_{j=1}^{N} \pi_{y_j|Z_j} \right)$$

Questions yet to answer

- How to simulate from or explore general *non-Gaussian* posterior distributions? (This lecture)
- How to make Bayesian inference computationally tractable when the forward model is expensive (e.g., a PDE) and the parameters are high- or infinite-dimensional? (Lecture #3)

MCMC overview

Markov chain Monte Carlo (MCMC)

- Metropolis-Hastings algorithm, transition kernels, ergodicity
- Mixture and cycles of kernels
- Gibbs sampling
- Gradient-exploiting MCMC, adaptive MCMC, other practicalities
- Using *approximations* (e.g., approximate likelihoods) within MCMC

Why Markov chain Monte Carlo (MCMC)?

In general, MCMC provides a means of sampling ("simulating") from an arbitrary distribution.

- The density $\pi(x)$ need be known only up to a normalizing constant
- Utility in *inference* and *prediction*: write both as posterior expectations, $\mathbb{E}_\pi f$.

Then

$$\mathbb{E}_\pi f \approx \frac{1}{n} \sum_i^n f\left(x^{(i)}\right)$$

- $x^{(i)}$ will be asymptotically distributed according to π
- $x^{(i)}$ will **not** be i.i.d. In other words, we must pay a price!

Why Markov chain Monte Carlo (MCMC)?

In general, MCMC provides a means of sampling ("simulating") from an arbitrary distribution.

- The density $\pi(x)$ need be known only up to a normalizing constant
- Utility in *inference* and *prediction*: write both as posterior expectations, $\mathbb{E}_\pi f$.

Then

$$\mathbb{E}_\pi f \approx \frac{1}{n} \sum_i^n f\left(x^{(i)}\right)$$

- $x^{(i)}$ will be asymptotically distributed according to π
- $x^{(i)}$ will **not** be i.i.d. In other words, we must pay a price!

Construction of an MCMC sampler

Define a **Markov chain** (i.e., discrete time). For real-valued random variables, the chain has a continuous-valued state space (e.g., $\mathbb{R}^d$). Ingredients of the definition:

- Initial distribution, $x_0 \sim \pi_0$
- Transition kernel $K(x_n, x_{n+1})$.

$$\mathbb{P}(X_{n+1} \in A | X_n = x) = \int_A K(x, x')\, dx'$$

(Analogy: consider matrix of transition probabilities for a finite state space.)

Markov property: X_{n+1} depends only on X_n.

Goal: design transition kernel K such that chain converges asymptotically to the *target distribution* π independently of the initial distribution (starting point).

Construction of an MCMC sampler (cont.)

Goal: choose transition kernel K such that chain converges asymptotically to the *target distribution* π independently of the starting point.

- Use realizations of $X_n, X_{n-1}, \ldots$ in a Monte Carlo estimator of posterior expectations (an ergodic average)
- Would like to converge to the target distribution *quickly* and to have samples as close to independent as possible
- Price for non-i.i.d. samples: greater variance in MC estimates of posterior expectations

Metropolis-Hastings algorithm

A simple recipe! From x_n to x_{n+1}:

1. Draw a proposal y from $q(y|x_n)$
2. Calculate acceptance ratio
$$\alpha(x_n, y) = \min\left\{1, \frac{\pi(y)q(x_n|y)}{\pi(x_n)q(y|x_n)}\right\}$$
3. Put
$$x_{n+1} = \begin{cases} y, & \text{with probability } \alpha(x_n, y) \\ x_n, & \text{with probability } 1 - \alpha(x_n, y) \end{cases}$$

Metropolis-Hastings algorithm

Notes on the algorithm:

- If $q(y|x_n) \propto \pi(y)$ then $\alpha = 1$. Thus we "correct" for sampling from q, rather than from π, via the Metropolis acceptance step.
- q does not have to be symmetric. If the proposal is symmetric, the acceptance probability simplifies (a "Hastings" proposal).
- π need be evaluated only up to a multiplicative constant

Metropolis-Hastings algorithm

What is the **transition kernel** of the Markov chain we have just defined?

- *Hint:* it is not q!
- Informally, it is

$$K(x_n, x_{n+1}) = p(x_{n+1}|\text{accept})\,\mathbb{P}[\text{accept}] + p(x_{n+1}|\text{reject})\,\mathbb{P}[\text{reject}]$$

- More precisely, we have:

$$\begin{aligned} K(x_n, x_{n+1}) &= p(x_{n+1}|x_n) \\ &= q(x_{n+1}|x_n)\alpha(x_n, x_{n+1}) + \delta_{x_n}(x_{n+1})\, r(x_n), \\ \text{where } r(x_n) &\equiv \int q(y|x_n)\,(1-\alpha(x_n, y))\, dy \end{aligned}$$

Metropolis-Hastings algorithm

What is the **transition kernel** of the Markov chain we have just defined?

- *Hint:* it is not q!
- Informally, it is

$$K\left(x_n, x_{n+1}\right) = p\left(x_{n+1}|\text{accept}\right)\mathbb{P}[\text{accept}] + p\left(x_{n+1}|\text{reject}\right)\mathbb{P}[\text{reject}]$$

- More precisely, we have:

$$\begin{aligned} K\left(x_n, x_{n+1}\right) &= p\left(x_{n+1}|x_n\right) \\ &= q(x_{n+1}|x_n)\alpha(x_n, x_{n+1}) + \delta_{x_n}\left(x_{n+1}\right) r(x_n), \\ \text{where } r(x_n) &\equiv \int q(y|x_n)\left(1-\alpha(x_n, y)\right) dy \end{aligned}$$

Metropolis-Hastings algorithm

What is the **transition kernel** of the Markov chain we have just defined?

- *Hint:* it is not q!
- Informally, it is

$$K\left(x_n, x_{n+1}\right) = p\left(x_{n+1}|\text{accept}\right)\mathbb{P}[\text{accept}] + p\left(x_{n+1}|\text{reject}\right)\mathbb{P}[\text{reject}]$$

- More precisely, we have:

$$\begin{aligned} K\left(x_n, x_{n+1}\right) &= p\left(x_{n+1}|x_n\right) \\ &= q(x_{n+1}|x_n)\alpha(x_n, x_{n+1}) + \delta_{x_n}\left(x_{n+1}\right) r(x_n), \\ \text{where } r(x_n) &\equiv \int q(y|x_n)\left(1-\alpha(x_n, y)\right) dy \end{aligned}$$

Metropolis-Hastings algorithm

Now, some theory. What are the key questions?

1. Is π a stationary distribution of the chain? (Is the chain π-invariant?)
   - Stationarity: π is such that $X_n \sim \pi \Rightarrow X_{n+1} \sim \pi$
2. Does the chain converge to stationarity? In other words, as $n \to \infty$, does $\mathcal{L}(X_n)$ converge to π?
3. Can we use paths of the chain in Monte Carlo estimates?

A *sufficient* (but not necessary) condition for (1) is **detailed balance** (also called 'reversibility'):

$$\pi(x_n)K(x_n, x_{n+1}) = \pi(x_{n+1})K(x_{n+1}, x_n)$$

- This expresses an equilibrium in the flow of the chain
- Hence $\int \pi(x_n)K(x_n, x_{n+1})\, dx_n = \int \pi(x_{n+1})K(x_{n+1}, x_n)\, dx_n = \pi(x_{n+1}) \int K(x_{n+1}, x_n)\, dx_n = \pi(x_{n+1})$.
- As an exercise, verify detailed balance for the M-H kernel defined on the previous slide.

Metropolis-Hastings algorithm

Now, some theory. What are the key questions?

1. Is π a stationary distribution of the chain? (Is the chain π-invariant?)
   - Stationarity: π is such that $X_n \sim \pi \Rightarrow X_{n+1} \sim \pi$
2. Does the chain converge to stationarity? In other words, as $n \to \infty$, does $\mathcal{L}(X_n)$ converge to π?
3. Can we use paths of the chain in Monte Carlo estimates?

A *sufficient* (but not necessary) condition for (1) is **detailed balance** (also called 'reversibility'):

$$\pi(x_n)K(x_n, x_{n+1}) = \pi(x_{n+1})K(x_{n+1}, x_n)$$

- This expresses an equilibrium in the flow of the chain
- Hence $\int \pi(x_n)K(x_n, x_{n+1})\,dx_n = \int \pi(x_{n+1})K(x_{n+1}, x_n)\,dx_n = \pi(x_{n+1})\int K(x_{n+1}, x_n)\,dx_n = \pi(x_{n+1})$.
- As an exercise, verify detailed balance for the M-H kernel defined on the previous slide.

Metropolis-Hastings algorithm

Now, some theory. What are the key questions?

1. Is π a stationary distribution of the chain? (Is the chain π-invariant?)
   - Stationarity: π is such that $X_n \sim \pi \Rightarrow X_{n+1} \sim \pi$
2. Does the chain converge to stationarity? In other words, as $n \to \infty$, does $\mathcal{L}(X_n)$ converge to π?
3. Can we use paths of the chain in Monte Carlo estimates?

A *sufficient* (but not necessary) condition for (1) is **detailed balance** (also called 'reversibility'):

$$\pi(x_n)K(x_n, x_{n+1}) = \pi(x_{n+1})K(x_{n+1}, x_n)$$

- This expresses an equilibrium in the flow of the chain
- Hence $\int \pi(x_n)K(x_n, x_{n+1})\,dx_n = \int \pi(x_{n+1})K(x_{n+1}, x_n)\,dx_n = \pi(x_{n+1})\int K(x_{n+1}, x_n)\,dx_n = \pi(x_{n+1})$.
- As an exercise, verify detailed balance for the M-H kernel defined on the previous slide.

Metropolis-Hastings algorithm

Beyond π-invariance, we also need to establish (2) and (3) from the previous slide. This leads to additional technical requirements:

- π-irreducibility: for every set A with $\pi(A) > 0$, there exists n such that $K^n(x, A) > 0 \ \forall x$.
  - *Intuition:* chain visits any measurable subset with nonzero probability in a finite number of steps. Helps you "forget" the initial condition. Sufficient to have $q(y|x) > 0$ for every $(x, y) \in \chi \times \chi$.
- Aperiodicity: "don't get trapped in cycles"

Metropolis-Hastings algorithm

When these requirements are satisfied (i.e., chain is *irreducible* and *aperiodic*, with *stationary* distribution π) we have

1. $$\lim_{n\to\infty} \left\| \int K^n(x,\cdot)\,\mu(dx) - \pi(\cdot) \right\|_{TV} = 0$$

   for every initial distribution μ.

   - K^n is the kernel for n transitions
   - This yields the law of X_n: $\int K^n(x,\cdot)\,\mu(dx) = \mathcal{L}(X_n)$
   - The total variation distance $\|\mu_1 - \mu_2\|_{TV} = \sup_A |\mu_1(A) - \mu_2(A)|$ is the largest possible difference between the probabilities that the two measures can assign to the same event.

Metropolis-Hastings algorithm

When these requirements are satisfied (i.e., chain is *irreducible* and *aperiodic*, with *stationary* distribution π) we have

2. For $h \in L^1_\pi$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i}^{n} h\left(x^{(i)}\right) = \mathbb{E}_\pi[h] \text{ w.p. } 1$$

This is a *strong law of large numbers* that allows computation of posterior expectations.

Obtaining a central limit theorem, or more generally saying anything about the *rate* of convergence to stationarity, requires additional conditions (e.g., geometric ergodicity).

See [Roberts & Rosenthal 2004] for an excellent survey of MCMC convergence results.

Metropolis-Hastings diagnostics

What about the **quality** of MCMC estimates?

What is the price one pays for correlated samples?

Compare Monte Carlo (iid) and MCMC estimates of $\mathbb{E}_\pi h$ (and for the latter, assume we have a CLT):

Monte Carlo

$$\mathbb{V}\text{ar}\left[\bar{h}_n\right] = \frac{\mathbb{V}\text{ar}_\pi\left[h(X)\right]}{n}$$

MCMC

$$\mathbb{V}\text{ar}\left[\bar{h}_n\right] = \theta\frac{\mathbb{V}\text{ar}_\pi\left[h(X)\right]}{n}$$

where

$$\theta = 1 + 2\sum_{s>0}^{\infty}\text{corr}\left(h(X_i), h(X_{i+s})\right)$$

is the **integrated autocorrelation time**.

Effective sample size (ESS) is then n/θ

Metropolis-Hastings diagnostics

What about the **quality** of MCMC estimates?
What is the price one pays for correlated samples?

Compare Monte Carlo (iid) and MCMC estimates of $\mathbb{E}_\pi h$ (and for the latter, assume we have a CLT):

Monte Carlo

$$\mathbb{V}\text{ar}\left[\bar{h}_n\right] = \frac{\mathbb{V}\text{ar}_\pi\left[h(X)\right]}{n}$$

MCMC

$$\mathbb{V}\text{ar}\left[\bar{h}_n\right] = \theta\frac{\mathbb{V}\text{ar}_\pi\left[h(X)\right]}{n}$$

where

$$\theta = 1 + 2\sum_{s>0}^{\infty} \text{corr}\left(h(X_i), h(X_{i+s})\right)$$

is the **integrated autocorrelation time**.

Effective sample size (ESS) is then n/θ

Metropolis-Hastings diagnostics

What about the **quality** of MCMC estimates?
What is the price one pays for correlated samples?

Compare Monte Carlo (iid) and MCMC estimates of $\mathbb{E}_\pi h$ (and for the latter, assume we have a CLT):

Monte Carlo

$$\mathbb{V}\text{ar}\left[\bar{h}_n\right] = \frac{\mathbb{V}\text{ar}_\pi\left[h(X)\right]}{n}$$

MCMC

$$\mathbb{V}\text{ar}\left[\bar{h}_n\right] = \theta\frac{\mathbb{V}\text{ar}_\pi\left[h(X)\right]}{n}$$

where

$$\theta = 1 + 2\sum_{s>0}^{\infty}\text{corr}\left(h(X_i), h(X_{i+s})\right)$$

is the **integrated autocorrelation time**.

Effective sample size (ESS) is then n/θ

Metropolis-Hastings algorithm

Now try a very simple computational demonstration (in MATLAB): MCMC sampling from a univariate distribution

Look at autocorrelation and visual diagnostics (e.g., trace of chain)

Other Metropolis-Hastings diagnostics

Example: *multivariate potential scale reduction factor* (MPSRF) [Brooks & Gelman 1998]

- Run multiple "replicate" chains from over-dispersed starting points.
- Compute:
  - Pooled-sample covariance estimate (across all chains) $\hat{\mathbf{V}}$ *(tends to over-estimate)*
  - Average of individual-chain sample covariance estimates $\mathbf{W}$ *(tends to under-estimate)*
- Let $\hat{R}^{1/2}$ be the largest generalized eigenvalue of the pencil $(\hat{\mathbf{V}}, \mathbf{W})$.
- Diagnostic: value of statistic $\hat{R}^{1/2}$ approaches 1 (from above) as the chains become similar

More sophisticated Metropolis-Hastings

- M-H construction was extremely general.
- Achieving efficient sampling (good "mixing") requires more exploitation of problem structure.
  1. Mixtures of kernels
  2. Cycles of kernels; Gibbs sampling
  3. Adaptive MCMC
  4. Gradient- and Hessian-exploiting MCMC
  5. MCMC in infinite dimensions

Mixtures and cycles

Mixtures of kernels

- Let K_i all have π as limiting distribution
- Use a convex combination: $K^* = \sum_i \nu_i K_i$
- ν_i is the probability of picking transition kernel K_i at a given step of the chain
- Kernels can correspond to transitions that each have desirable properties, e.g., local versus global proposals

Cycles of kernels

- Split multivariate state vector into *blocks* that are updated separately; each update is accomplished by transition kernel K_j
- Need to combine kernels. **Cycle** = a systematic scan, $K^* = \prod_j K_j$

Componentwise Metropolis-Hastings

This is an example of using a cycle of kernels

- Let $\mathbf{x} = (x^1, \ldots, x^d) \in \mathbb{R}^d$
- Proposal $q_i(y|\mathbf{x})$ updates only component i
- Walk through components of the state sequentially, $i = 1 \ldots d$:
  - Propose a new value for component i using

$$q_i\left(y^i | x_{n+1}^1, \ldots, x_{n+1}^{i-1}, x_n^i, x_n^{i+1}, \ldots, x_n^d\right)$$

  - Accept $(x_{n+1}^i = y^i)$ or reject $(x_{n+1}^i = x_n^i)$ this component update with acceptance probability

$$\alpha_i(\mathbf{x}_i, \mathbf{y}_i) = \min\left\{1, \frac{\pi(\mathbf{y}_i) q_i(x_n^i | \mathbf{y}_i)}{\pi(\mathbf{x}_i) q_i(y^i | \mathbf{x}_i)}\right\}$$

    where $\mathbf{x}_i$ and $\mathbf{y}_i$ differ only in component i

$$\begin{aligned} \mathbf{y}_i &\equiv \left(x_{n+1}^1, \ldots, x_{n+1}^{i-1}, y, x_n^{i+1}, \ldots, x_n^d\right) \text{ and} \\ \mathbf{x}_i &\equiv \left(x_{n+1}^1, \ldots, x_{n+1}^{i-1}, x_n^i, x_n^{i+1}, \ldots, x_n^d\right) \end{aligned}$$

Gibbs sampling

- One very useful *cycle* is the Gibbs sampler.
- Requires the ability to sample directly from the *full conditional distribution* $\pi(x_i|\mathbf{x}_{\sim i})$.
  - $\mathbf{x}_{\sim i}$ denotes all components of $\mathbf{x}$ other than x_i
  - In problems with appropriate *structure*, generating independent samples from the full conditional may be feasible while sampling from π is not.
  - x_i can represent a block of the state vector, rather than just an individual component
- A Gibbs update is a proposal from the full conditional; the acceptance probability is **identically one**!

$$\begin{aligned}\alpha_i(\mathbf{x}_i, \mathbf{y}_i) &= \min\left\{1, \frac{\pi(\mathbf{y}_i)\, q_i(x_n^i|\mathbf{y}_i)}{\pi(\mathbf{x}_i)\, q_i(y^i|\mathbf{x}_i)}\right\} \\ &= \min\left\{1, \frac{\pi(y_i|\mathbf{x}_{\sim i})\pi(\mathbf{x}_{\sim i})\, \pi(x_n^i|\mathbf{x}_{\sim i})}{\pi(x_n^i|\mathbf{x}_{\sim i})\pi(\mathbf{x}_{\sim i})\, \pi(y^i|\mathbf{x}_{\sim i})}\right\} = 1.\end{aligned}$$

Gibbs sampling example #1

Correlated bivariate normal

$$x \sim N\left(\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}\right)$$

Full conditionals are:

$$\begin{aligned} x_1|x_2 &\sim N\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2-\mu_2), (1-\rho^2)\sigma_1^2\right)\\ x_2|x_1 &\sim \ldots \end{aligned}$$

See computational demo

Gibbs sampling example #2

Bayesian linear regression with a variance hyperparameter

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \sigma z_i, \quad y_i \in \mathbb{R};\ \boldsymbol{\beta}, \mathbf{x}_i \in \mathbb{R}^d;\ z_i \sim N(0,1)$$

- This problem has a non-Gaussian posterior but is amenable to block Gibbs sampling
- Let the data consist of n observations $\mathcal{D}_n \equiv \{(y_i, \mathbf{x}_i)\}_{i=1}^n$
- Bayesian hierarchical model, likelihood and priors:

$$\begin{aligned} \mathbf{y} \,|\, \boldsymbol{\beta}, \sigma^2 &\sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right) \\ \boldsymbol{\beta} \,|\, \sigma^2 &\sim N\left(0, \tau^2 \sigma^2 \mathbf{I}_d\right) \\ 1/\sigma^2 &\sim \Gamma\left(\alpha, \gamma\right) \end{aligned}$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ has rows $\mathbf{x}_i$ and $\mathbf{y} \in \mathbb{R}^n$ is a vector of $y_1 \ldots y_n$.

Gibbs sampling example #2 (cont.)

- Posterior density:

$$\begin{aligned}
\pi\left(\boldsymbol{\beta}, \sigma^2\right) &\equiv p\left(\boldsymbol{\beta}, \sigma^2 | \mathcal{D}_n\right) \\
&\propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right) \\
&\quad \frac{1}{(\tau\sigma)^d} \exp\left(-\frac{1}{2\tau^2\sigma^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \\
&\quad \left(\frac{1}{\sigma^2}\right)^{\alpha-1} \exp\left(-\gamma/\sigma^2\right)
\end{aligned}$$

- Full conditionals $\boldsymbol{\beta}|\sigma^2, \mathcal{D}_n$ and $\sigma^2|\boldsymbol{\beta}, \mathcal{D}_n$ have a closed form! Try to obtain by inspecting the joint density above. (See next page for answer.)

Gibbs sampling example #2 (cont.)

- Full conditional for $\boldsymbol{\beta}$ is Gaussian:

$$\boldsymbol{\beta} \,|\, \sigma^2, \mathcal{D}_n \sim N\left(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}\right)$$

where

$$\boldsymbol{\Sigma}^{-1} = \left(\frac{1}{\tau^2}\mathbf{I}_d + \mathbf{X}^T\mathbf{X}\right) \text{ and } \boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{X}^T\mathbf{y}.$$

- Full conditional for $1/\sigma^2$ is Gamma:

$$1/\sigma^2 \,|\boldsymbol{\beta}, \mathcal{D}_n \sim \Gamma\left(\hat{\alpha}, \hat{\gamma}\right)$$

where

$\hat{a} = a + n/2 + d/2$ and $\hat{\gamma} = \gamma + \dfrac{1}{2\tau^2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \dfrac{1}{2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right).$

- Alternately sample from these FCs in order to simulate the joint posterior.
- Also, this is an example of the use of **conjugate priors**.

Metropolis-within-Gibbs

What if we cannot sample from the full conditionals?

- Solution: "Metropolis-within-Gibbs"
- This is just componentwise Metropolis-Hastings (which is where we started)

Metropolis-within-Gibbs

What if we cannot sample from the full conditionals?

- Solution: "Metropolis-within-Gibbs"
- This is just componentwise Metropolis-Hastings (which is where we started)

Adaptive Metropolis

- Intuitive idea: learn a better proposal $q(y|x)$ from past samples.
  - Learn an appropriate proposal **scale**.
  - Learn an appropriate proposal **orientation** and anisotropy; this is *essential* in problems with strong correlation in π
- Adaptive Metropolis scheme of [Haario *et al.* 2001]:
  - Covariance matrix at step n

$$C_n^* = s_d \, \mathrm{Cov}\,(x_0, \ldots, x_n) + s_d \epsilon I_d$$

    where $\epsilon > 0$, d is the dimension of the state, and $s_d = 2.4^2/d$ (scaling rule-of-thumb).
  - Proposals are Gaussians centered at x_n. Use a fixed covariance C_0 for the first n_0 steps, then use C_n^*.
  - Chain is not Markov, and previous convergence proofs do not apply. Nonetheless, one can prove that the chain converges to π. See paper in references.
- Many other adaptive MCMC ideas have been developed in recent years

Adaptive Metropolized independence samplers

- Independence proposal: does not depend on current state
- Consider a proposal $q(x;\psi)$ with parameter ψ.
- Key idea: minimize Kullback-Leibler divergence between this proposal and the target distribution:

$$\min_{\psi} D_{KL}\left(\pi(x) \| q(x;\psi)\right)$$

- Equivalently, maximize $\int \pi(x) \log q(x;\psi) dx$

- Solve this optimization problem with successive steps of stochastic approximation (e.g., Robbins-Monro), while approximating the integral via MCMC samples
- Common choice: let q be a mixture of Gaussians or other exponential-family distributions

Online MCMC demo

Very cool demo, thanks to Chi Feng (MIT):
`https://chi-feng.github.io/mcmc-demo`

Let's look at RWM and AM on various targets

Langevin MCMC

- Intuitive idea: use gradient of the posterior to steer samples towards higher density regions
- Consider the SDE

$$dX_t = \frac{1}{2}\nabla \log \pi(X_t)dt + dW_t$$

  This SDE has π as its stationary distribution
- Discretize the SDE (e.g., Euler-Maruyama)

$$X^{t+1} = X^t + \frac{\sigma^2}{2}\nabla \log \pi(X^t) + \sigma\epsilon^t,\ \epsilon^t \sim N(0, I)$$

  - Discretized process X^t no longer has π as its stationary distribution! But we can use X^{t+1} as a **proposal** in the regular Metropolis-Hastings framework, and accept or reject it accordingly.
  - σ^2 (discretization time step) is an adjustable free parameter.
- Langevin schemes require access to the gradient of the posterior.

Preconditioned Langevin

- Introduce a positive definite matrix $\mathbf{A}$ to the Langevin SDE:

$$dX_t = \frac{1}{2}\mathbf{A}\nabla \log \pi(X_t)dt + \mathbf{A}^{1/2}dW_t$$

- Let $\mathbf{A}$ reflect covariance structure of target
- For example: let $\mathbf{A}$ be the local inverse Hessian of the log-posterior, or the inverse Hessian at the posterior mode, or posterior-averaged Hessian information, or some other estimate of the posterior covariance

- Key idea for *inverse problems:* use **low-rank approximations** of the posterior covariance/precision developed for the linear-Gaussian case

Preconditioned Langevin

- Introduce a positive definite matrix $\mathbf{A}$ to the Langevin SDE:

$$dX_t = \frac{1}{2}\mathbf{A}\nabla \log \pi(X_t)dt + \mathbf{A}^{1/2}dW_t$$

- Let $\mathbf{A}$ reflect covariance structure of target
- For example: let $\mathbf{A}$ be the local inverse Hessian of the log-posterior, or the inverse Hessian at the posterior mode, or posterior-averaged Hessian information, or some other estimate of the posterior covariance

- Key idea for *inverse problems:* use **low-rank approximations** of the posterior covariance/precision developed for the linear-Gaussian case

Hamiltonian MCMC

- Let x be "position" variables; introduce auxiliary "momentum" variables w
- Consider a separable Hamiltonian, $H(x, w) = U(x) + w^T M^{-1} w/2$. Put $U(x) = -\log \pi(x)$.
- Hamiltonian dynamics are *reversible* and conserve H. Use them to propose new states x!
- In particular, sample from $p(x, w) = \frac{1}{Z} \exp\left(-H(x, w)\right)$:
  - First, sample the momentum variables w from their Gaussian distribution
  - Second, integrate Hamilton's equations to propose a new state (x, w); then apply Metropolis accept/reject
- **Features:**
  - Enables faraway moves in x-space while leaving the value of the density (essentially) unchanged. Good mixing!
  - Requires good symplectic integrators and access to derivatives
  - Recent extension: Riemannian manifold HMC [Girolami & Calderhead JRSSB 2011]

Online MCMC demo

Back to the demo:
`https://chi-feng.github.io/mcmc-demo`

Now look at MALA and HMC/NUTS on various targets

MCMC in infinite dimensions

- Would like to construct a well-defined MCMC sampler for **functions** $u \in \mathcal{H}$.
- First, the posterior measure μ_y should be a well-defined probability measure on $\mathcal{H}$ (see Stuart *Acta Numerica* 2010). For simplicity, let the prior μ_0 be $\mathcal{N}(0, C)$.
- Now let q be the proposal distribution, and consider pair of measures

$$\nu(du, du') = q(u, du')\mu_y(du), \;\; \nu^{\perp}(du, du') = q(u', du)\mu_y(du');$$

- Then the MCMC acceptance probability is

$$\alpha(u_k, u') = \min\left\{1, \frac{d\nu^{\perp}}{d\nu}(u_k, u')\right\}$$

- To define a **valid** transition kernel, we need absolute continuity $\nu^{\perp} \ll \nu$; in turn, this places requirements on the proposal q

MCMC in infinite dimensions

- Would like to construct a well-defined MCMC sampler for **functions** $u \in \mathcal{H}$.
- First, the posterior measure μ_y should be a well-defined probability measure on $\mathcal{H}$ (see Stuart *Acta Numerica* 2010). For simplicity, let the prior μ_0 be $\mathcal{N}(0, C)$.

- Now let q be the proposal distribution, and consider pair of measures
$$\nu(du, du') = q(u, du')\mu_y(du), \;\; \nu^{\perp}(du, du') = q(u', du)\mu_y(du');$$
- Then the MCMC acceptance probability is
$$\alpha(u_k, u') = \min\left\{1, \frac{d\nu^{\perp}}{d\nu}(u_k, u')\right\}$$
- To define a **valid** transition kernel, we need absolute continuity $\nu^{\perp} \ll \nu$; in turn, this places requirements on the proposal q

MCMC in infinite dimensions (cont.)

- One way to produce a valid transition kernel is the preconditioned Crank-Nicolson (pCN) proposal (Cotter *et al.* 2013):

$$u' = (1-\beta^2)^{1/2} u_k + \beta \xi_k, \ \xi_k \sim \mathcal{N}(0, C), \ \beta \in (0,1).$$

- Practical impact: sampling efficiency does not degenerate as discretization of u is refined
- More sophisticated versions: combine pCN with Hessian/geometry information, e.g., DILI (dimension-independent likelihood-informed) proposals (Cui, Law, M 2016)

MCMC practicalities

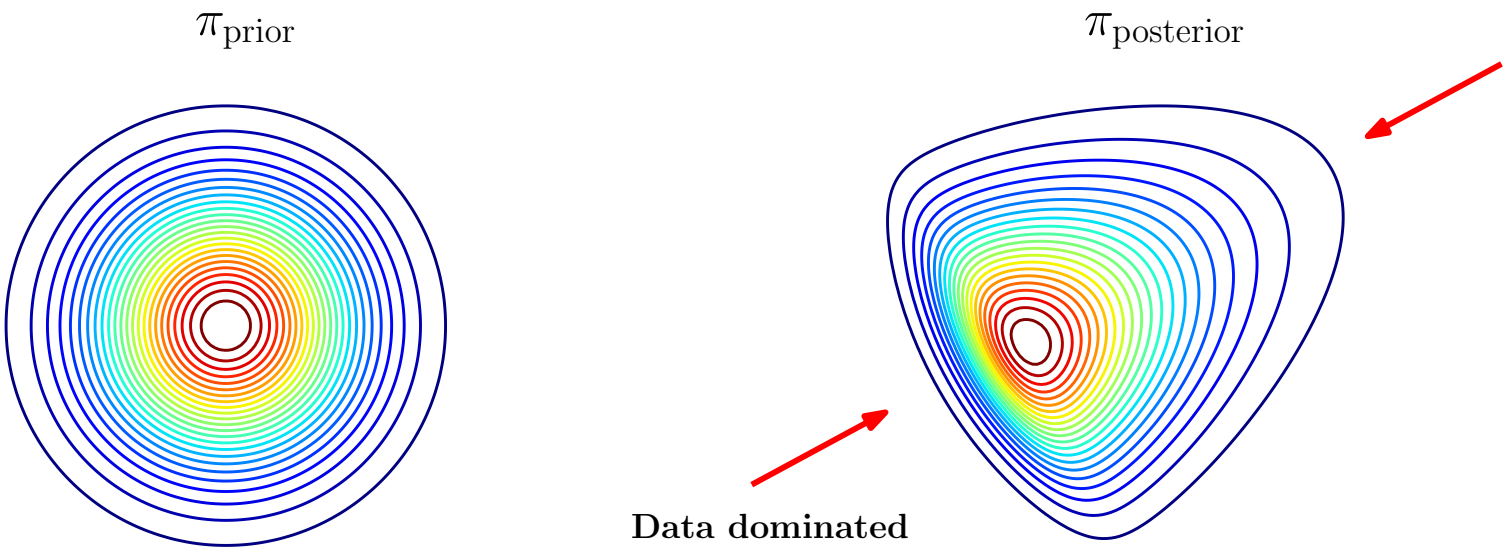
Effective use of MCMC still requires some (problem-specific) experience. Some useful rules of thumb:

- Adaptive schemes are not a panacea.
- Whenever possible, (re-)parameterize the problem in order to minimize posterior correlations.
- What to do, if anything, about "burn-in?"
- Visual inspection of chain components is often the first and best convergence diagnostic.
- Also look at autocorrelation plots. Run multiple chains from different starting points. Evaluate MPSRF or other diagnostics.

MCMC practicalities

Additional advice:

- "The best Monte Carlo is a dead Monte Carlo": If you can tackle any part of the problem analytically, do it!
  - Example: **Rao-Blackwellization** in Cui *et al.*, "Likelihood-informed dimension reduction for nonlinear inverse problems," *Inverse Problems* 30: 114015 (2014).



MCMC references

A small selection of useful "general" MCMC references.

- C. Andrieu, N. de Freitas, A. Doucet, M. I. Jordan, "An introduction to MCMC for machine learning," *Machine Learning* 50 (2003) 5–43.
- S. Brooks, A. Gelman, G. Jones and X. Meng, editors. *Handbook of MCMC*. Chapman & Hall/CRC, 2011.
- A. Gelman, J. B. Carlin, H. S. Stern, D. Dunson, A. Vehtari, D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall CRC, 3rd edition, 2013.
- P. J. Green. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732, 1995.
- H. Haario, M. Laine, A. Mira, and E. Saksman. "DRAM: Efficient adaptive MCMC." *Statistics and Computing*, 16(4): 339–354, 2006.
- C. P. Robert, G. Casella, *Monte Carlo Statistical Methods*, 2nd Edition, Springer, 2004.
- G. Roberts, J. Rosenthal. "General state space Markov chains and MCMC algorithms." *Probability Surveys*, 1: 20–71, 2004.