02477 – Bayesian Machine Learning: Lecture 9

Michael Riis Andersen

Technical University of Denmark,
DTU Compute, Department of Applied Math and Computer Science

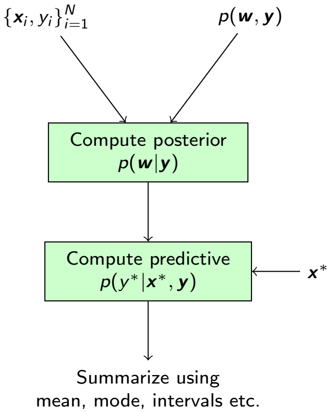
Outline

1. Markov chain Monte Carlo
2. Gibbs sampling
3. Convergence diagnostics
4. Hierarchical models

Markov chain Monte Carlo

Bayesian inference and probabilistic modelling

- Bayesian supervised learning in general:
  1. Joint model of data $\boldsymbol{y}$ and parameters $\boldsymbol{w}$.
  2. Summarize knowledge of $\boldsymbol{w}$ given data $\boldsymbol{y}$.
  3. Compute posterior predictive distribution.
- Goal: separate modelling from inference:
  1. Build models reflecting domain knowledge.
  2. Push "inference button" and get results.
- Inference methods:
  1. Laplace approximations.
  2. Markov chain Monte Carlo.
  3. Variational approximations.



Monte Carlo: Posterior inference using samples

- Many posterior summaries can be phrased as expectations $\mathbb{E}_p[f(\boldsymbol{z})]$ for some function f

- We can compute expectations wrt. p using the *Monte Carlo estimator*

$$\bar{f} = \mathbb{E}_p[f(\boldsymbol{z})] = \int f(\boldsymbol{z})p(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \approx \frac{1}{S}\sum_{i=1}^{S} f(\boldsymbol{z}^i) \equiv \hat{f},$$

  where $\boldsymbol{z}^i \sim p(\boldsymbol{z})$ for $i = 1, \ldots, S$

- We showed last time that ...
  1. the Monte Carlo estimator $\hat{f}$ is *unbiased*
  2. the variance of $\hat{f}$ decreases with $1/S$ when the samples are *i.i.d.*

A zoo of sampling-based methods

- Simple sampling methods
  1. Rejection sampling
  2. Ancestral sampling
  3. Importance sampling
  4. Transformation methods
  5. Inverse transform sampling
  6. ...
- MCMC methods
  1. Metropolis-Hastings
  2. Gibbs Sampling
  3. Slice sampling
  4. Hamiltonian Monte Carlo
  5. ...

MCMC using the Metropolis-Hastings algorithm

- We can use the MH to generate samples from a distribution of interest $p(\boldsymbol{z})$.

The Metropolis-Hastings algorithm

- Start from some initial value $\boldsymbol{z}^1$ (e.g., a sample from the prior).
- Repeat for $k = 1$ to K:
  1. Given last value $\boldsymbol{z}^{k-1}$, generate *candidate sample* using proposal distribution
  $$\boldsymbol{z}^\star \sim q(\boldsymbol{z}^\star | \boldsymbol{z}^{k-1}).$$
  2. Compute *acceptance probability* A_k as follows
  $$A_k = \min\left(1, \frac{p(\boldsymbol{z}^\star)q(\boldsymbol{z}^{k-1}|\boldsymbol{z}^\star)}{p(\boldsymbol{z}^{k-1})q(\boldsymbol{z}^\star|\boldsymbol{z}^{k-1})}\right).$$
  3. Simulate $u_k \sim \mathcal{U}(0,1)$ and define $\boldsymbol{z}^k$ as
  $$\boldsymbol{z}^{k+1} = \begin{cases} \boldsymbol{z}^\star & \text{if } u_k < A_k \\ \boldsymbol{z}^{k-1} & \text{otherwise} \end{cases}$$

- What do we need in order to implement MH for a given model?

Markov chain Monte Carlo theory I

- Metropolis-Hastings defines a chain of samples $\boldsymbol{z}^0, \boldsymbol{z}^1, \boldsymbol{z}^2, \ldots$ with a *Markov property*

$$p(\boldsymbol{z}^{k+1}|\boldsymbol{z}^1, \boldsymbol{z}^2, .., \boldsymbol{z}^k) = p(\boldsymbol{z}^{k+1}|\boldsymbol{z}^k)$$

- The *transition kernel* tells us how to iterate the chain

$$T(\boldsymbol{z}^{k+1}|\boldsymbol{z}^k) \equiv p(\boldsymbol{z}^{k+1}|\boldsymbol{z}^k)$$

- The distribution of $\boldsymbol{z}^{k+1}$ is given by *sum rule*

$$p(\boldsymbol{z}^{k+1}) = \int T(\boldsymbol{z}^{k+1}|\boldsymbol{z}^k) p(\boldsymbol{z}^k) \mathrm{d}\boldsymbol{z}^k$$

- A distribution $p^*(\boldsymbol{z})$ is said to be *invariant* or *stationary* wrt. the Markov chain if each step does not change the distribution

$$p^*(\boldsymbol{z}) = \int T(\boldsymbol{z}|\boldsymbol{z}') p^*(\boldsymbol{z}') \mathrm{d}\boldsymbol{z}'$$

- We require $p^*(\boldsymbol{z})$ to be a limiting distribution of the chain *(independent of the initial distribution)*.

$$p(\boldsymbol{z}^k) \quad \rightarrow \quad p^*(\boldsymbol{z}) \qquad \text{for} \qquad k \rightarrow \infty$$

Transistion kernel for Metropolis-Hastings

- Recall the acceptance probability for Metropolis-Hastings

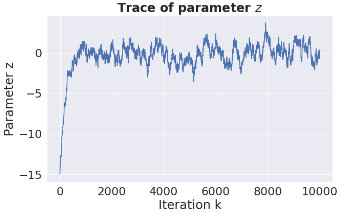
$$A(\boldsymbol{z}^\star|\boldsymbol{z}^k) = \min\left[1, \frac{p(\boldsymbol{z}^\star)q(\boldsymbol{z}^k|\boldsymbol{z}^\star)}{p(\boldsymbol{z}^k)q(\boldsymbol{z}^\star|\boldsymbol{z}^k)}\right]$$

- The transition kernel for Metropolis-Hasting

$$T(\boldsymbol{z}'|\boldsymbol{z}) = \begin{cases} q(\boldsymbol{z}'|\boldsymbol{z})A(\boldsymbol{z}'|\boldsymbol{z}) & \text{if} \quad \boldsymbol{z}' \neq \boldsymbol{z} \\ q(\boldsymbol{z}|\boldsymbol{z})A(\boldsymbol{z}|\boldsymbol{z}) + \int q(\boldsymbol{z}''|\boldsymbol{z}) \left[1 - A(\boldsymbol{z}''|\boldsymbol{z})\right] \mathrm{d}\boldsymbol{z}'' & \text{if} \quad \boldsymbol{z}' = \boldsymbol{z} \end{cases}$$

- The big picture

  1. We initialize $\boldsymbol{z}^1$.
  2. We iterate $\boldsymbol{z}^{k+1}|\boldsymbol{z}^k \sim T(\boldsymbol{z}^{k+1}|\boldsymbol{z}^k)$ (*warm-up phase*).
  3. Eventually the distribution of $\boldsymbol{z}^k$ will converge to the target distribution p^* (*sampling phase*).



Markov chain Monte Carlo theory II

$$p^*(\boldsymbol{z}) = \int T(\boldsymbol{z}|\boldsymbol{z}')p^*(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}'$$

- We require $p^*(\boldsymbol{z})$ to be a limiting distribution of the chain *(independent of the initial distribution)*.

$$p(\boldsymbol{z}^k) \quad \to \quad p^*(\boldsymbol{z}) \qquad \text{for} \qquad k \to \infty$$

- Conditions required for a Markov chain to have a limiting distribution equal to the unique stationary distribution:

(A1) *Irreducible*: all states $\boldsymbol{z}$ can be reached.

$$\text{Counter example:} \quad z^{k+1} = z^k + |e_k|, \quad \text{where} \quad e_k \sim \mathcal{N}(0,1)$$

(A2) *Aperiodic*: no deterministic cycles.

$$\text{Counter example:} \quad z^1 = 1, z^2 = 2, z^3 = 3, z^4 = 1, z^5 = 2, \ldots$$

(A3) *Positive recurrent*: the chain has positive probability of returning to any given state (+ finite expected return-time).

$$P(\boldsymbol{z}^k \in A|\boldsymbol{z}^0) > 0 \quad \text{for all sets } A \quad \text{where} \quad P(A) > 0$$

- A chain that satisfies *(A1)–(A3)* is said to be *ergodic* and ensures $p(\boldsymbol{z}^k) \to p^*(\boldsymbol{z})$.

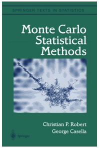
Markov chain Monte Carlo theory III

- These conditions are generally hard to check in practice.
- Simpler condition: if a chain satisfies the *detailed balance condition*, then p^* is its stationary distribution

$$T(z'|z)p^*(z) = T(z|z')p^*(z').$$

- MH with reasonable proposal distributions satisfies detailed balance (see slide 33).
- It's often not a question of convergence or not, but rather how fast we converge for a given proposal distribution.
- For example, all (non-degenerate) Gaussian proposals lead to detailed balance, but convergence time may vary drastically depending on the proposal variance.

- More theory and details: Monte Carlo Statistical Methods by Robert and Casella.



Pros and cons of Metropolis-Hastings

Pros

- Strong mathematical guarantees: If we sample long enough, the iterates $\boldsymbol{z}^k$ will converge to the exact target distribution

$$p(\boldsymbol{z}^k) \quad \rightarrow \quad p^*(\boldsymbol{z}) \qquad \text{for} \qquad k \rightarrow \infty$$

- Easy to implement.
- Easy to prototype and evaluate different models.

Cons

- May have to sample "infinitely" long for difficult distributions.
- Acceptance ratio can be low.
- Slow for large datasets.
- Proposal distribution may require tuning.

Questions: True or false?

Quiz via DTU Learn:

Lecture 9: Metropolis-Hastings (12 questions)

Check you knowledge

Gibbs sampling

Gibbs sampling

- When using Metropolis-Hastings,
  1. we have to choose a proposal distribution (and sometimes tune it), and
  2. it may suffer from low acceptance rates.

- *Gibbs sampling* works by iteratively updating each coordinate of $\boldsymbol{z}$ by sampling from the posterior conditionals $p(z_i|\boldsymbol{z}_{-i})$ ($\boldsymbol{z}_{-i}$ means the entire vector except index i).

The Gibbs Sampler

- Initialize all parameter values $\left\{z_i^0\right\}_{i=1}^D$
- Repeat for $k = 1$ to K:
  - Sample $z_1^k \sim p(z_1|z_2^{k-1}, z_3^{k-1}, \ldots, z_D^{k-1})$.
  - Sample $z_2^k \sim p(z_2|z_1^{k}, z_3^{k-1}, \ldots, z_D^{k-1})$.
  - Sample $z_3^k \sim p(z_3|z_1^{k}, z_2^{k}, z_4^{k-1}, \ldots, z_D^{k-1})$
  - Sample . . .
  - Sample $z_D^k \sim p(z_D|z_1^{k}, z_2^{k}, z_3^{k}, \ldots, z_{D-1}^{k})$.

Example: Gaussian linear model I

- Suppose we want to derive a Gibbs sampler for the following target distribution

$$
\begin{aligned}
y|w &\sim \mathcal{N}(y|w_1x_1 + w_2x_2, \sigma^2)\\
w_1 &\sim \mathcal{N}(w_1|0, \kappa^2)\\
w_2 &\sim \mathcal{N}(w_2|0, \kappa^2)
\end{aligned}
$$

- The posterior distribution is proportional to the joint density $p(y, w_1, w_2)$

$$
\begin{aligned}
p(w_1, w_2|y) &= \frac{p(y|w_1, w_2)p(w_1)p(w_2)}{p(y)}\\
&\propto p(y|w_1, w_2)p(w_1)p(w_2)\\
&= \mathcal{N}(y|w_1x_1 + w_2x_2, \sigma^2)\mathcal{N}(w_1|0, \kappa^2)\mathcal{N}(w_2|0, \kappa^2)
\end{aligned}
$$

- *Gibbs sampling* requires us to derive the *posterior conditionals* $p(w_1|y, w_2)$ and $p(w_2|y, w_1)$

- *General technique* for identifying $p(w_i|y, \boldsymbol{w}_{-i})$:
  1. Write up the log joint density.
  2. Identify all quantities that depends on w_i and ignore the rest.
  3. Identify the distribution $p(w_i|y, \boldsymbol{w}_{-i})$ from its *functional form*.

Example: Gaussian linear model II

- Write out the logarithm of the joint density

$$p(w_1, w_2|y) \propto \mathcal{N}(t|w_1x_1 + w_2x_2, \sigma^2)\mathcal{N}(w_1|0, \kappa^2)\mathcal{N}(w_2|0, \kappa^2)$$

- Recall the expression for a Gaussian density

$$\mathcal{N}(x|m, v) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(x-m)^2}{2v}\right)$$

- Let's write it out the log density and identify all terms that depend on w_1 or w_2

$$\begin{aligned}
\log p(w_1, w_2|y) = & -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - w_1x_1 - w_2x_2)^2 + \\
& -\frac{1}{2}\log(2\pi\kappa^2) - \frac{1}{2\kappa^2}w_1^2 - \frac{1}{2}\log(2\pi\kappa^2) - \frac{1}{2\kappa^2}w_2^2 + K \\
= & -\frac{1}{2\sigma^2}(y - w_1x_1 - w_2x_2)^2 - \frac{1}{2\kappa^2}w_1^2 - \frac{1}{2\kappa^2}w_2^2 + K' \\
= & -\frac{1}{2\sigma^2}(y^2 + (w_1x_1 + w_2x_2)^2 - 2y(w_1x_1 - w_2x_2)) - \frac{1}{2\kappa^2}w_1^2 - \frac{1}{2\kappa^2}w_2^2 + K' \\
= & -\frac{1}{2\sigma^2}(w_1x_1 + w_2x_2)^2 + \frac{1}{\sigma^2}y(w_1x_1 - w_2x_2) - \frac{1}{2\kappa^2}w_1^2 - \frac{1}{2\kappa^2}w_2^2 + K'' \\
= & -\frac{1}{2\sigma^2}(w_1^2x_1^2 + w_2^2x_2^2 + 2w_1x_1w_2x_2) + \frac{1}{\sigma^2}y(w_1x_1 - w_2x_2) - \frac{1}{2\kappa^2}w_1^2 - \frac{1}{2\kappa^2}w_2^2 + K''
\end{aligned}$$

Example: Gaussian linear model III

- We just arrived at

$$\log p(w_1, w_2|y) = -\frac{1}{2\sigma^2}(w_1^2 x_1^2 + w_2^2 x_2^2 + 2w_1 x_1 w_2 x_2) + \frac{1}{\sigma^2} y(w_1 x_1 - w_2 x_2) - \frac{1}{2\kappa^2} w_1^2 - \frac{1}{2\kappa^2} w_2^2 + K''$$

- Let's compare that to a generic Gaussian distribution:

$$\begin{aligned}\log \mathcal{N}(w_1|m, v) &= -\frac{1}{2}\log(2\pi v) - \frac{1}{2v}(w_1 - m)^2 \\ &= -\frac{1}{2}\log(2\pi v) - \frac{1}{2v}(w_1^2 + m^2 - 2w_1 m) = -\frac{1}{2v} w_1^2 + \frac{1}{v} m w_1 + C\end{aligned}$$

- Recall: *The functional form* of the logarithm of a Gaussian density is a quadratic function.

- Identify the distribution $p(w_1|y, w_2)$ based on the functional dependence on w_1:

$$\begin{aligned}\log p(w_1|y, w_2) &= -\frac{1}{2\sigma^2}(w_1^2 x_1^2 + w_2^2 x_2^2 + 2w_1 x_1 w_2 x_2) + \frac{1}{\sigma^2} y(w_1 x_1 - w_2 x_2) - \frac{1}{2\kappa^2} w_1^2 - \frac{1}{2\kappa^2} w_2^2 + K'' \\ &= -\frac{1}{2\sigma^2}(w_1^2 x_1^2 + 2w_1 x_1 w_2 x_2) + \frac{1}{\sigma^2} y w_1 x_1 - \frac{1}{2\kappa^2} w_1^2 + K''' \\ &= -\frac{1}{2}\left(\frac{1}{\sigma^2} x_1^2 + \frac{1}{\kappa^2}\right) w_1^2 + \left(\frac{1}{\sigma^2} y x_1 - \frac{1}{\sigma^2} x_1 w_2 x_2\right) w_1 + K'''\end{aligned}$$

- We conclude that the distribution $\log p(w_1|y, w_2)$ must be a Gaussian, because *its functional form is quadratic wrt. w_1*.

Example: Gaussian linear model IV

- A generic Gaussian distribution

$$\ln \mathcal{N}(w_1|m, v) = -\frac{1}{2v} w_1^2 + \frac{1}{v} m w_1 + C$$

- We know $p(w_1|y, w_2) = \mathcal{N}(w_1|m_1, v_1)$ is Gaussian, so all we need is a mean and variance

$$\log p(w_1|y, w_2) = -\frac{1}{2}\left(\frac{1}{\sigma^2} x_1^2 + \frac{1}{\kappa^2}\right) w_1^2 + \left(\frac{1}{\sigma^2} y x_1 - \frac{1}{\sigma^2} x_1 w_2 x_2\right) w_1 + K'''$$

- *Comparing the coefficients* for the *second order term* w_1^2, we get the variance

$$v_1 = \left(\frac{1}{\sigma^2} x_1^2 + \frac{1}{\kappa^2}\right)^{-1}$$

- and by comparing coefficients for the *first order* term w_1, we get the mean

$$\frac{m_1}{v_1} = \left(\frac{1}{\sigma^2} y x_1 - \frac{1}{\sigma^2} x_1 w_2 x_2\right) \quad \Longleftrightarrow \quad m_1 = \frac{v_1}{\sigma^2}\left(y x_1 - x_1 w_2 x_2\right)$$

- *By symmetry*, we get $p(w_2|y, w_1) = \mathcal{N}(w_2|m_2, v_2)$

$$v_2 = \left(\frac{1}{\sigma^2} x_1^1 + \frac{1}{\kappa^2}\right)^{-1} \quad \Longleftrightarrow \quad m_2 = \frac{v_2}{\sigma^2}\left(y x_2 - x_2 w_1 x_1\right)$$

Example: Gaussian linear model V

- Initialize w_1 and w_2.

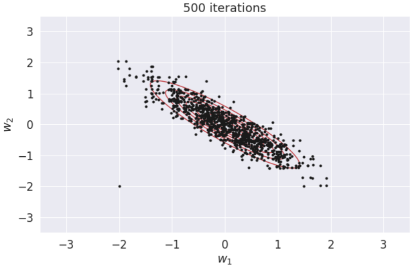
  1. Sample w_1 conditioned w_2,

  $$w_1 \sim p(w_1|y, w_2) = \mathcal{N}(w_1|m_1, v_1).$$

  2. Sample w_2 conditioned w_1,

  $$w_2 \sim p(w_2|y, w_1) = \mathcal{N}(w_2|m_2, v_2).$$

  3. Repeat.



- Example shows typical *staircase behavior* of Gibbs samplers due sampling from the *posterior conditionals*

Why does Gibbs sampling work?

- The acceptance probability in Metropolis-Hastings algorithm:

$$A_k = \min\left[1, \frac{p(\mathbf{z}^\star)q(\mathbf{z}^k|\mathbf{z}^\star)}{p(\mathbf{z}^k)q(\mathbf{z}^\star|\mathbf{z}^k)}\right]$$

- Writing $\mathbf{z}^k = \left\{z_i^k, \mathbf{z}_{-i}^k\right\}$, the proposal distribution for the Gibbs sampler is

$$q(z_i^\star|\mathbf{z}^k) = p(z_i^\star|\mathbf{z}_{-i}^k). \tag{1}$$

- We need the following fact

$$p(\mathbf{z}) = p(z_i, \mathbf{z}_{-i}) = p(z_i|\mathbf{z}_{-i})p(\mathbf{z}_{-i}). \tag{2}$$

- Plugging in the proposal

$$\begin{aligned}
A_k &= \min\left[1, \frac{p(\mathbf{z}^\star)p(z_i^k|\mathbf{z}_{-i}^\star)}{p(\mathbf{z}^k)p(z_i^\star|\mathbf{z}_{-i}^k)}\right] && \text{(Plugging in the proposal in eq. (1))} \\
&= \min\left[1, \frac{p(z_i^\star|\mathbf{z}_{-i}^\star)p(\mathbf{z}_{-i}^k)p(z_i^k|\mathbf{z}_{-i}^\star)}{p(z_i^k|\mathbf{z}_{-i}^k)p(\mathbf{z}_{-i}^k)p(z_i^\star|\mathbf{z}_{-i}^k)}\right] && \text{(Using eq. (2))} \\
&= \min\left[1, \frac{p(z_i^\star|\mathbf{z}_{-i}^k)p(\mathbf{z}_{-i}^k)p(z_i^k|\mathbf{z}_{-i}^k)}{p(z_i^k|\mathbf{z}_{-i}^k)p(\mathbf{z}_{-i}^k)p(z_i^\star|\mathbf{z}_{-i}^k)}\right] && \text{(Using } \mathbf{z}_{-i}^\star = \mathbf{z}_{-i}^k\text{)} \\
&= 1
\end{aligned}$$

- A Gibbs sampler is a special case of MH, where the proposed candidate is always accepted.

Questions: True or false?

Quiz via DTU Learn:

Lecture 9: Gibbs (5 questions)

Check you knowledge

Convergence diagnostics

Are we there yet?

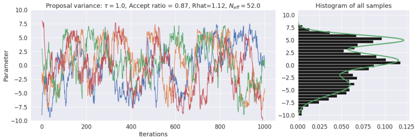
- MCMC theory states that samplers converge to the true target distribution as $S \to \infty$.

- *Intuitive heuristic for assessing stationarity*: Run *multiple* chains from *different initial conditions*. After K iterations, we compare the distributions for each chain. If they are different, the chains have not yet reached the stationary distribution.

- Let B denote the *between-chain variance* and let W denote the *within-chain variance*, then *$\hat{R}$-statistic* (or the *potential scale reduction factor*) for chains of length N is defined by

$$\hat{R}^2 = \frac{S-1}{S} + \frac{1}{S}\frac{B}{W}$$

- If $B = W$, then $\hat{R} = 1$. If $B > W$, then $\hat{R} > 1$. In practice, we say that the *chains have mixed* if $\hat{R} < 1.1$



For more details and motivation, see p. 284 in Bayesian Data Analysis (`http://www.stat.columbia.edu/~gelman/book/BDA3.pdf`)

How accurate is MCMC? Quantifying the error

- Recall for i.i.d. samples, the error of the MC estimator decreases with rate $1/\sqrt{S}$:

$$\mathbb{V}\left[\hat{f}\right] = \frac{1}{S}\mathbb{V}\left[f(\theta)\right]$$

- We can *estimate the variance* based on the samples and use this to *quantify the Monte Carlo error*. Let $\widehat{\mathrm{sd}}(f(\theta))$ be the standard deviation of the MCMC samples,

$$\mathrm{MCSE} = \frac{1}{\sqrt{S}}\widehat{\mathrm{sd}}(f(\theta))$$

- MCMC methods produces *highly correlated* samples. The *autocorrelation function* ρ_t measures the correlation between two samples θ^i and θ^{i+t}

$$\rho_t = \frac{1}{\sigma^2}\int(\theta^i - \mu)(\theta^{i+t} - \mu)p(\theta)\mathrm{d}\theta$$

- The *effective sample size (ESS)* S_{eff} takes this correlation into account

$$S_{\mathrm{eff}} = \frac{S}{\sum_{t=-\infty}^{\infty}\rho_t} = \frac{S}{1 + 2\sum_{t=1}^{\infty}\rho_t}$$

- and then

$$\mathrm{MCSE} = \frac{1}{\sqrt{S_{\mathrm{eff}}}}\widehat{\mathrm{sd}}(f(\theta))$$

Example: MCMC diagnostics

- The $\hat{R} - statistic$

$$\hat{R}^2 = \frac{N-1}{N} + \frac{1}{N}\frac{B}{W}$$

- The effective sample size

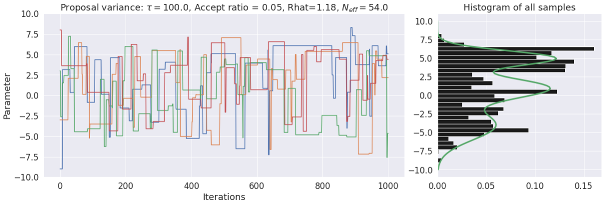
$$S_{\text{eff}} = \frac{S}{1 + 2\sum_{t=1}^{\infty} \rho_t}$$

- The Monte Carlo Standard Error

$$\text{MCSE} = \frac{1}{\sqrt{S_{\text{eff}}}} \widehat{\text{sd}}(f(\theta))$$

- Next week

1. A few words on Hamiltonian Monte Carlo
2. We will do a short discussion on pros and cons of MCMC in practice
3. Start discussing variational inference.



Hierarchical models

Revisiting the Poisson regression

- "*Being Bayesian*" usually refers to treating quantities of interest as *random variable* and reason using the rules of *probability theory*

- When we talked about "*fully Bayesian*" inference, we refer the setting, where we have *prior distributions on all parameters, including hyperparameters*

- Before the holidays, we worked with a fully Bayesian Poisson regression model via MCMC

$$\begin{aligned} y_n|\mu_n &\sim \text{Poisson}(\mu_n), \\ \mu_n &= \exp(f_n) \\ f_n &= \mathbf{w}^T \mathbf{x}_n \\ \mathbf{w}|\kappa &= \mathcal{N}(0, \kappa^2 \mathbf{I}) \\ \kappa &\sim \mathcal{N}_+(0, 1) \end{aligned}$$

  with the following joint distribution

$$p(\mathbf{y}, \mathbf{w}, \kappa) = \prod_{n=1}^{N} p(y_n|\mathbf{w}) p(\mathbf{w}|\kappa) p(\kappa),$$

- This is an example of a *hierarchical model*

Hierarchical modelling

- *Hierarchical* or *multi-level* models are one of the key strength of the Bayesian framework

- Suppose we have a model $p(\mathcal{D}|\boldsymbol{\theta})$ with data $\mathcal{D}$, parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\xi}$

$$\xi \longrightarrow \theta \longrightarrow \mathcal{D}$$

  with the following joint distribution

$$p(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{\xi}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})p(\boldsymbol{\xi})$$

- Useful when you want to
  1. make inference more robust
  2. reason probabilistically about data, parameters and hyperparameters
  3. model hierarchical structure in data (random effects models)
  4. squeeze out every bit of predictive performance of a model/dataset

- Degrees of "Bayesianity" according to Murphy1

Method	Definition
Maximum likelihood	$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})$
MAP	$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})$
ML-II	$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})\mathrm{d}\boldsymbol{\theta}$
MAP-II	$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\xi})p(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\theta}$
Full Bayes	$p(\boldsymbol{\theta}, \boldsymbol{\xi}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{\xi})$

Example

- Recall the model for Bayesian linear regression

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I}) \qquad (\textit{prior})$$

$$p(\boldsymbol{y}|\boldsymbol{w}, \beta) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \beta^{-1}\boldsymbol{I}) \qquad (\textit{likelihood})$$

$$p(\boldsymbol{w}|\boldsymbol{y}, \alpha, \beta) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}, \boldsymbol{S}) \qquad (\textit{posterior})$$

$$p(\boldsymbol{y}|\alpha, \beta) = \mathcal{N}(\boldsymbol{y}|\mathbf{0}, \beta^{-1}\boldsymbol{I} + \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^T) \qquad (\textit{marginal likelihood})$$

- Fully Bayesian inference on the full joint distribution

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

- Fully Bayesian inference on the marginalized joint distribution

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

- Making predictions via MCMC

$$p(y^*|\mathbf{y}) = \mathbb{E}_{p(\alpha,\beta|\mathbf{y})}\left[p(y^*|\boldsymbol{y}, \alpha, \beta)\right] \approx \frac{1}{S}\sum_{i=1}^{S} p(y^*|\boldsymbol{y}, \alpha^{(i)}, \beta^{(i)})$$

for $\alpha^{(i)}, \beta^{(i)} \sim p(\alpha, \beta|\mathbf{y})$

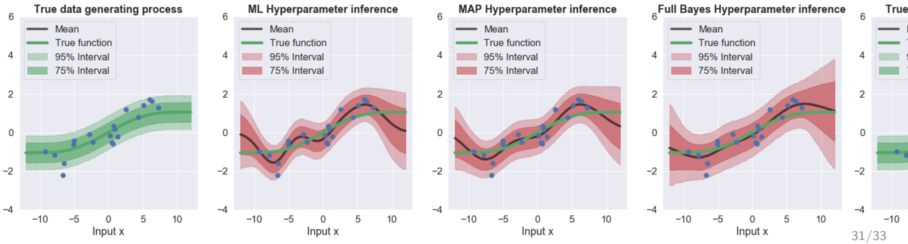
Fully Bayesian Gaussian process regression: Example

- Example with $N = 20$ data points and additive Gaussian noise
- Gaussian process regression with squared exponential kernel
- We impose a *weakly informative* prior on the lengthscale to
  1. rule out really short length scales (smaller than the grid size)
  2. rule out really long length scales (larger than span data of data)
- Joint distribution

$$p(\mathbf{y}, \mathbf{f}, \ell, \sigma, \kappa) = p(\mathbf{y}|\mathbf{f}, \sigma)p(\boldsymbol{f}|\ell, \kappa)p(\kappa)p(\ell)$$

- Marginalized joint distribution

$$p(\mathbf{y}, \ell, \sigma, \kappa) = p(\mathbf{y}|\sigma, \ell, \kappa)p(\kappa)p(\ell)$$



Medical example

- Suppose you work for a medical company and are asked to analyze data from a drug evaluation on rats prior to human trials

- Suppose the drug was administered to N rats, where y rats ended up developing tumors.

- We could use a Beta-Binomial model to estimate the probability of developing tumors θ

$$p(\theta|y) \propto \text{Bin}(y|N,\theta)\text{Beta}(\theta|\alpha_0,\beta_0)$$

- What if the company tested the drugs on J different types of rats, where (y_j, N_j) denote the data for the j'th group. How to analyze the data?

  1. We could fit J *individual* models

  $$p(\theta_j|y_j) \propto \text{Bin}(y_j|N_j,\theta_j)\text{Beta}(\theta_j|\alpha_0,\beta_0)$$

  2. We could *pool* all the data such that $N_p = \sum_j N_j$ and $y_p = \sum_j y_j$

  $$p(\theta_p|y_p) \propto \text{Bin}(y_p|N_p,\theta)\text{Beta}(\theta_p|\alpha_0,\beta_0)$$

- *Individual* models may perform poorly if N_j is small, but the *pooled* model might perform poorly if the different groups exhibit different behaviors

- *Bayesian hierarchical models* allows us to borrow statistical strength from groups with lots of data to help groups with less data

Medical example cont.

- Let $\boldsymbol{y} = \{y_j\}_{j=1}^J$ and $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^J$, then

$$p(\boldsymbol{y}, \boldsymbol{\theta}, \alpha, \beta) = \prod_{j=1}^{J} \text{Bin}(y_j|N_j, \theta_j) \prod_{j=1}^{J} \text{Beta}(\theta_j|\alpha, \beta) p(\alpha, \beta)$$

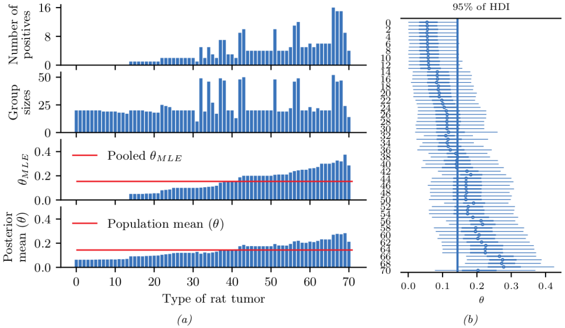


Figure 3.15: Data and inferences for the hierarchical binomial model fit using HMC. Generated by hierarchical_binom_rats.ipynb.

The detailed balance condition I

- If a chain satisfies the *detailed balance condition*, then p^* is its stationary distribution

$$T(\boldsymbol{z}'|\boldsymbol{z})p^*(\boldsymbol{z}) = T(\boldsymbol{z}|\boldsymbol{z}')p^*(\boldsymbol{z}')$$

- Integrating both sides wrt. $\boldsymbol{z}'$ yields the stationary distribution p^*

$$\int T(\boldsymbol{z}'|\boldsymbol{z})p^*(\boldsymbol{z})\mathrm{d}\boldsymbol{z}' = p^*(\boldsymbol{z}) = \int T(\boldsymbol{z}|\boldsymbol{z}')p^*(\boldsymbol{z}')\mathrm{d}\boldsymbol{z}'$$

- Does the Metropolis-Hastings satisfy this condition? Let's check

- Recall the acceptance probability for Metropolis-Hastings

$$A(\boldsymbol{z}^\star|\boldsymbol{z}^k) = \min\left[1, \frac{p(\boldsymbol{z}^\star)q(\boldsymbol{z}^k|\boldsymbol{z}^\star)}{p(\boldsymbol{z}^k)q(\boldsymbol{z}^\star|\boldsymbol{z}^k)}\right]$$

- The transition kernel for Metropolis-Hasting

$$T(\boldsymbol{z}'|\boldsymbol{z}) = \begin{cases} q(\boldsymbol{z}'|\boldsymbol{z})A(\boldsymbol{z}'|\boldsymbol{z}) & \text{if} \quad \boldsymbol{z}' \neq \boldsymbol{z} \\ q(\boldsymbol{z}|\boldsymbol{z})A(\boldsymbol{z}|\boldsymbol{z}) + \int q(\boldsymbol{z}''|\boldsymbol{z})\left[1 - A(\boldsymbol{z}''|\boldsymbol{z})\right]\mathrm{d}\boldsymbol{z}'' & \text{if} \quad \boldsymbol{z}' = \boldsymbol{z} \end{cases}$$

The detailed balance condition II

- The acceptance probability

$$A(\boldsymbol{z}^\star, \boldsymbol{z}^k) = \min\left[1, \frac{p(\boldsymbol{z}^\star)q(\boldsymbol{z}^k|\boldsymbol{z}^\star)}{p(\boldsymbol{z}^k)q(\boldsymbol{z}^\star|\boldsymbol{z}^k)}\right]$$

- The transition kernel

$$T(\boldsymbol{z}'|\boldsymbol{z}) = \begin{cases} q(\boldsymbol{z}'|\boldsymbol{z})A(\boldsymbol{z}'|\boldsymbol{z}) & \text{if} \quad \boldsymbol{z}' \neq \boldsymbol{z} \\ q(\boldsymbol{z}|\boldsymbol{z})A(\boldsymbol{z}|\boldsymbol{z}) + \int q(\boldsymbol{z}''|\boldsymbol{z})\left[1 - A(\boldsymbol{z}''|\boldsymbol{z})\right] \mathrm{d}\boldsymbol{z}'' & \text{if} \quad \boldsymbol{z}' = \boldsymbol{z} \end{cases}$$

- If $p(\boldsymbol{z}^k)q(\boldsymbol{z}^\star|\boldsymbol{z}^k) > p(\boldsymbol{z}^\star)q(\boldsymbol{z}^k|\boldsymbol{z}^\star)$, then $A(\boldsymbol{z}^\star|\boldsymbol{z}^k) < 1$ and $A(\boldsymbol{z}^k|\boldsymbol{z}^\star) = 1$

- To jump from $\boldsymbol{z}^k$ to $\boldsymbol{z}^\star$, we first need to propose it and then accept it

$$T(\boldsymbol{z}^\star|\boldsymbol{z}^k) = q(\boldsymbol{z}^\star|\boldsymbol{z}^k)A(\boldsymbol{z}^\star|\boldsymbol{z}^k) = q(\boldsymbol{z}^\star|\boldsymbol{z}^k)\frac{p(\boldsymbol{z}^\star)q(\boldsymbol{z}^k|\boldsymbol{z}^\star)}{p(\boldsymbol{z}^k)q(\boldsymbol{z}^\star|\boldsymbol{z}^k)} = \frac{p(\boldsymbol{z}^\star)q(\boldsymbol{z}^k|\boldsymbol{z}^\star)}{p(\boldsymbol{z}^k)}$$

- It follows

$$p(\boldsymbol{z}^k)T(\boldsymbol{z}^\star|\boldsymbol{z}) = p(\boldsymbol{z}^\star)q(\boldsymbol{z}^k|\boldsymbol{z}^\star)$$

The detailed balance condition III

- We just arrived at

$$p(\boldsymbol{z}^k)T(\boldsymbol{z}^\star|\boldsymbol{z}^k) = p(\boldsymbol{z}^\star)q(\boldsymbol{z}^k|\boldsymbol{z}^\star)$$

- What about the opposite direction?

$$T(\boldsymbol{z}^k|\boldsymbol{z}^\star) = q(\boldsymbol{z}^k|\boldsymbol{z}^\star)A(\boldsymbol{z}^k|\boldsymbol{z}^\star) = q(\boldsymbol{z}^k|\boldsymbol{z}^\star)$$

- Combining the two and we are done

$$p(\boldsymbol{z}^k)T(\boldsymbol{z}^\star|\boldsymbol{z}^k) = p(\boldsymbol{z}^\star)T(\boldsymbol{z}^k|\boldsymbol{z}^\star)$$