Lecture 5: Sparse GLMs

STATS305B: Applied Statistics II

Scott Linderman

January 27, 2025

Announcements

- Great work on HW1!
- **Midterm next Wednesday in MCCULL 115**
- HW2 to be posted tonight or early tomorrow. Due Mon, Feb 10

Course Schedule

- Weeks 1-3: Classics: Exponential family distributions and GLMs
- **Weeks 4-5: Bayesian Inference algorithms: MCMC and variational inference**
- Weeks 6-7: Latent variable models: mixture models, HMMs, etc.
- Weeks 8-9: Deep generative models: VAEs, Transformers, Deep SSMs, Denoising diffusion models
- Week 10: Bonus: Point processes, survival analysis, etc.

Learning Objectives

- **Models:** Exponential Family Distributions, (Sparse) GLMs
- **Algorithms:** Gradient Descent, Newton's Method, IRLS, Proximal methods
- **Code:** Logistic regression from scratch in Python/PyTorch

Outline

Today...

- Bayesian Inference
- Conjugate Priors
- Laplace Approximation
- Monte Carlo
- Start on MCMC

Bayesian Inference

So far we've focused on *frequentist* inference techniques: asymptotically normal approximations, Wald confidence intervals, etc.

Today we'll discuss an alternative approach based on Bayesian inference.

While the two approaches are philosophically quite different, we'll see that the statistical inferences they lead to can be quite similar in many cases.

Introduction

It is tempting to interpret the confidence interval as saying that θ is in a confidence interval with probability $1 - \alpha$ given the observed data, but **that is not justified!**

In the setting above, the parameter θ is **not** a random variable. This fallacy is a common misinterpretation of frequentist confidence intervals.

To make such a claim, we need to adopt a Bayesian perspective and reason about the *posterior* distribution of the parameters, θ, given the data, x.

Introduction

To obtain a posterior, we first need to specify a *prior* distribution on parameters, $p(\theta)$. Given a prior and likelihood, the posterior follows from Bayes' rule,

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)},$$

where

- $p(\theta \mid x)$ is the **posterior**,
- $p(x \mid \theta)$ is the **likelihood**,
- $p(\theta)$ is the **prior**, and
- $p(x) = \int p(x \mid \theta)\, p(\theta)\, \mathrm{d}\theta$ is the **marginal likelihood**

What do we want from the Posterior?

Often, we are particularly interested in **posterior expectations**, like:

- $\mathbb{E}_{p(\theta|x)}[\theta]$, the posterior mean,
- $\mathbb{E}_{p(\theta|x)}[\mathbb{I}[\theta \in \mathscr{A}]]$, the probability of the parameters being in set $\mathscr{A}$,
- $\mathbb{E}_{p(\theta|x)}[p(x' \mid \theta)]$, the posterior predictive density of new data x'.

All of these can be written as $\mathbb{E}_{p(\theta|x)}[f(\theta)]$ for some function f.

For point estimation, we may choose the mode, $\hat{\theta}_{\text{MAP}} = \arg\max p(\theta \mid x)$ a.k.a., the ***maximum a posteriori*** **(MAP)** estimate.

We can also obtain an analogue of frequentist confidence intervals by summarizing the posterior in terms of a **Bayesian credible interval**: a set of parameters that captures $1-\alpha$ probability under the posterior.

Conjugate Priors

The posterior distribution depends on the choice of prior. Indeed, the subjective choice of prior distributions is the source of much of the criticism of Bayesian approaches.

- **Uninformative priors:** We can often specify “weak” or “uninformative” prior distributions. Then we’ll find that Bayesian and frequentist approaches can yield similar estimates.
- **Tractability** The hard part of Bayesian inference is typically integration: to normalize the posterior we need to compute the marginal likelihood, which is an integral over the parameter space. To compute posterior expectations, we need to do the same.

Conjugate priors are distributions on θ that often render these integrals tractable and can vary in informativeness.

Example: Bernoulli Likelihood with a Beta Prior

The beta distribution is a conjugate prior for a Bernoulli likelihood,

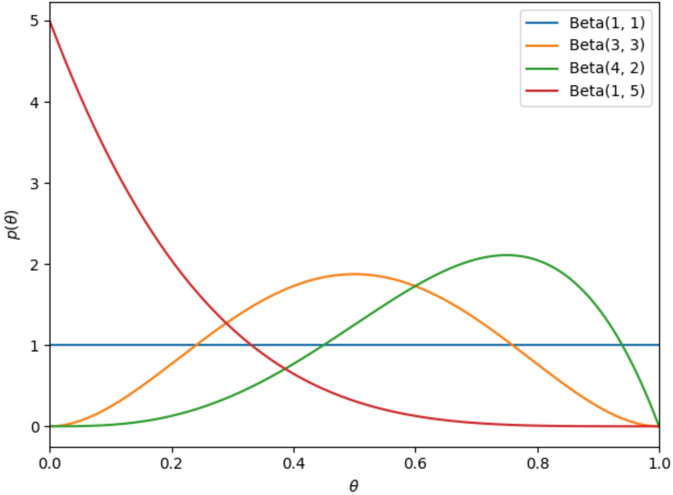
$$\theta \sim \mathrm{Beta}(\alpha, \beta)$$

with support on $\theta \in [0, 1]$. Its probability density function (pdf) is,

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

where $\mathrm{B}(\alpha, \beta)$ is the beta function and the hyperparameters $\alpha, \beta \in \mathbb{R}_+$ determine the shape of the prior. When $\alpha = \beta = 1$, the prior reduces to a uniform distribution on $[0, 1]$.

Example: Bernoulli Likelihood with a Beta Prior



Example: Bernoulli Likelihood with a Beta Prior

Under the beta prior, the posterior distribution over θ is,

$$\begin{aligned} p(\theta \mid \{x_i\}_{i=1}^n) &\propto \mathrm{Beta}(\theta; \alpha, \beta) \prod_{i=1}^n \mathrm{Bern}(x_i \mid \theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \\ &\propto \mathrm{Beta}(\theta; \alpha', \beta') \end{aligned}$$

where $x = \sum_{i=1}^n x_i$ is the number of coins that came up heads and

$$\begin{aligned} \alpha' &= x + \alpha \\ \beta' &= n - x + \beta \end{aligned}$$

Example: Bernoulli Likelihood with a Beta Prior

The posterior mode – i.e., the maximum a posteriori (MAP) estimate – is

$$\hat{\theta}_{\mathrm{MAP}} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{x + \alpha - 1}{n + \alpha + \beta - 2}.$$

Under an uninformative prior with $\alpha = \beta = 1$, it is equivalent to the MLE, $\hat{\theta}_{\mathrm{MLE}} = x/n$.

Bayesian credible intervals can be derived using the cumulative distribution function (cdf) of the beta distribution, which is given by the incomplete beta function.

In the large sample limit, the beta posterior is approximately Gaussian. The variance of the posterior beta distribution is,

$$\mathrm{Var}[\theta \mid X] = \frac{\alpha'\beta'}{(\alpha' + \beta')^2(\alpha' + \beta' + 1)} = \frac{(x + \alpha)(n - x + \beta)}{(n + \alpha + \beta)^2(n + \alpha + \beta + 1)}$$

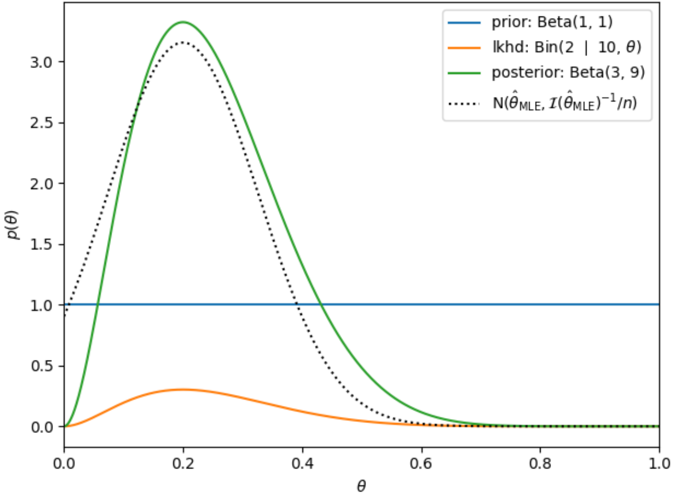
Example: Bernoulli Likelihood with a Beta Prior

In this limit, α and β are much smaller than n and x. Thus, the posterior variance is approximately

$$\mathrm{Var}[\theta \mid X] \approx \frac{x(n-x)}{n^3} = \frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{n} = \mathscr{I}(\hat{\theta}_{\mathrm{MLE}})^{-1}/n,$$

and the Bayesian credible intervals match the Wald confidence interval.

Example: Bernoulli Likelihood with a Beta Prior



Exponential Family Likelihoods

Consider a general **exponential family** likelihood with natural parameter θ,

$$p(x \mid \theta) = h(x) \exp\left\{\langle t(x), \theta\rangle - A(\theta)\right\}.$$

Exponential family distributions have conjugate priors,

$$p(\theta; \chi, \nu) \propto g(\theta) \exp\left\{\langle \chi, \theta\rangle - \nu A(\theta)\right\} = g(\theta) \exp\left\{\langle \chi, \theta\rangle + \langle \nu, -A(\theta)\rangle - B(\chi, \nu)\right\}.$$

We recognize the conjugate prior as another exponential family distribution in which,

- the natural parameter χ are **pseudo-observations** of the sufficient statistics,
- the natural parameter ν is a **pseudo-count** (like the number of fake data points),
- the prior sufficient statistics are $(\theta, -A(\theta))$,
- the prior log normalizer is $B(\chi, \nu)$

Exponential Family Likelihoods

With a conjugate prior, the posterior distribution belongs to the same family as the prior,

$$
\begin{aligned}
p(\theta \mid \{x_i\}_{i=1}^{n}; \chi, \nu) &\propto p(\theta; \chi, \nu) \prod_{i=1}^{n} p(x_i \mid \theta) \\
&\propto \exp\left\{ \chi + \sum_{i=1}^{n} t(x_i), \theta \rangle + \langle \nu + n, -A(\theta) \rangle \right\} \\
&= p(\theta \mid \chi', \nu')
\end{aligned}
$$

where

$$
\begin{aligned}
\chi' &= \chi + \sum_{i=1}^{n} t(x_i) \\
\nu' &= \nu + n.
\end{aligned}
$$

Question: With a conjugate prior, the posterior is just a function of χ' and ν', regardless of how many data points are observed.. Does that make it computationally tractable?

Laplace Approximation

Conjugate priors are a common choice for simple exponential family models, but we need more general approaches for more complex models.

Suppose you wanted to perform Bayesian inference of the weights in a logistic regression model,

$$p(y \mid x, \boldsymbol{\beta}) = \prod_{i=1}^{n} \mathrm{Bern}(y_i \mid \sigma(x_i^\top \boldsymbol{\beta})).$$

Assume a Gaussian prior,

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \gamma^{-1} \boldsymbol{I}).$$

Laplace Approximation

Unfortunately, the posterior does not have a closed formation solution. Instead, a common form of approximate posterior inference is the **Laplace approximation**,

$$p(\boldsymbol{\beta} \mid x, y) \approx \mathrm{N}(\hat{\boldsymbol{\beta}}_{\mathsf{MAP}}, \widehat{\boldsymbol{\Sigma}})$$

where

$$\hat{\boldsymbol{\beta}}_{\mathsf{MAP}} = \arg\max_{\boldsymbol{\beta}} \mathscr{L}(\boldsymbol{\beta})$$

is the *maximum a posteriori (MAP)* estimate,

$$\widehat{\boldsymbol{\Sigma}} = -[\nabla^2 \mathsf{L}(\hat{\boldsymbol{\beta}}_{\mathsf{MAP}})]^{-1}$$

is an approximation of the posterior covariance,

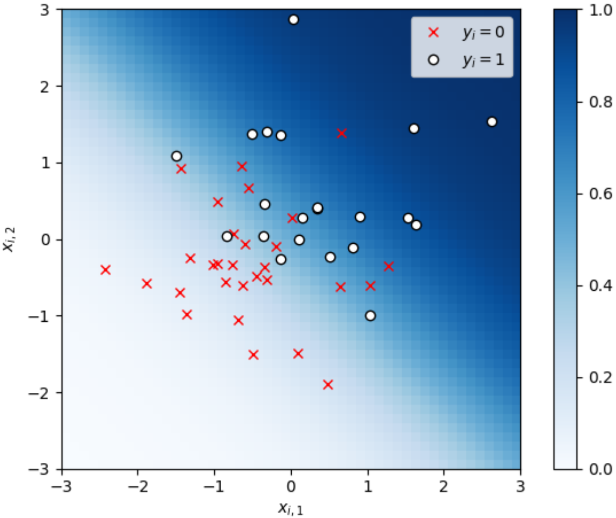
Laplace Approximation

and

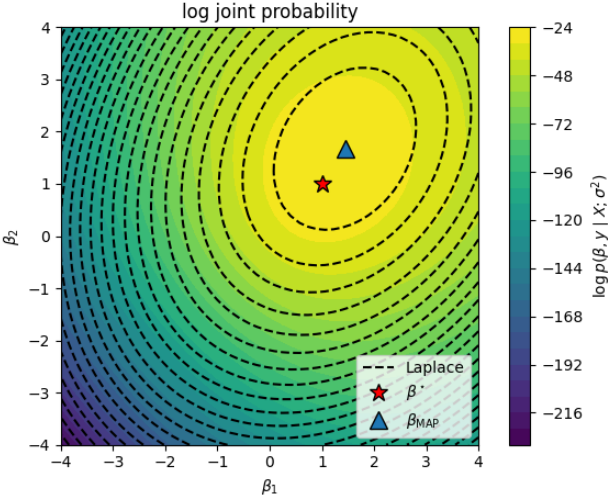
$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta}) &= \log p(\boldsymbol{\beta}) + \sum_{i=1}^{n} \log p(y_i \mid x_i, \boldsymbol{\beta}) \\ &= \log \mathrm{N}(\boldsymbol{\beta}; \mathbf{0}, \gamma^{-1}\boldsymbol{I}) + \sum_{i=1}^{n} \log \mathrm{Bern}(y_i \mid \sigma(x_i^\top \boldsymbol{\beta}))\end{aligned}$$

is the log joint probability, *not the loss function from previous chapters!*

Synthetic Demo



Synthetic Demo



Bernstein-von Mises Theorem

In the large data limit (as $n \to \infty$), the posterior is asymptotically normal, justifying the Laplace approximation in this regime.

Consider a simpler setting in which we have data $\{x_i\}_{i=1}^{n} \overset{\text{iid}}{\sim} p(x \mid \theta^\star)$.

Under some conditions (e.g. $\theta^\star$ not on the boundary of Θ and $\theta^\star$ has nonzero prior probability), then the MAP estimate is consistent. As $n \to \infty$, $\theta_{\text{MAP}} \to \theta^\star$.

Likewise,

$$p(\theta \mid \{x_i\}_{i=1}^{n}) \to \mathrm{N}\big(\theta \mid \theta^\star, \tfrac{1}{n}\mathscr{I}(\theta^\star)^{-1}\big)$$

where $\mathscr{I}(\theta)$ is the Fisher information matrix.

Approximating Posterior Expectations

Generally, we can't analytically compute posterior expectations. In these cases, we need to resort to approximations. For example, we could use *quadrature methods* like Simpson's rule or the trapezoid rule to numerically approximate the integral over Θ.

Roughly,

$$\mathbb{E}_{p(\theta|x)}[f(\theta)] \approx \sum_{m=1}^{M} p(\theta_m \mid x) f(\theta_m)\, \Delta_m$$

where $\theta_m \subset \Theta$ is a grid of points and Δ_m is a volume around that point.

This works for low-dimensional problems (say, up to 5 dimensions), but the number of points (M) needed to get a good estimate grows exponentially with the parameter dimension.

Monte Carlo Approximations

Idea: approximate the expectation via sampling,

$$\mathbb{E}_{p(\theta|x)}[f(\theta)] \approx \frac{1}{M}\sum_{m=1}^{M} f(\theta_m) \quad \text{where} \quad \theta_m \sim p(\theta \mid x).$$

Let $\hat{f} = \frac{1}{M}\sum_{m=1}^{M} f(\theta_m)$ denote the Monte Carlo estimate. It is a random variable, since it's a function of random samples θ_m. As such, we can reason about its mean and variance.

Unbiasedness

Clearly,

$$\mathbb{E}[\hat{f}] = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{p(\theta|x)}[f(\theta)] = \mathbb{E}_{p(\theta|x)}[f(\theta)].$$

Thus, $\hat{f}$ is an *unbiased* estimate of the desired expectation.

Monte Carlo Variance

What about its variance?

$$\mathrm{Var}[\hat{f}] = \mathrm{Var}\left(\frac{1}{M} \sum_{m=1}^{M} f(\theta_m) \right) = \frac{1}{M^2} \left(\sum_{m=1}^{M} \mathrm{Var}[f(\theta)] + 2 \sum_{1 \leq m < m' \leq M} \mathrm{Cov}[f(\theta_m), f(\theta_{m'})] \right)$$

Comparison to Numerical Quadrature

- If the samples are not only identically distributed but also *uncorrelated*, then $\mathrm{Var}[\hat{f}] = \frac{1}{M}\mathrm{Var}[f(\theta)]$.
- In this case, the *root mean squared error* (RMSE) of the estimate is $\sqrt{\mathrm{Var}[\hat{f}]} = O(M^{-\frac{1}{2}})$.
- Compare this to Simpson's rule, which for smooth 1D problems has an error rate of $O(M^{-4})$. That's roughly 8 times better than Monte Carlo!
- However, for multidimensional problems, Simpson's rule is $O(M^{-\frac{4}{D}})$, whereas the **error rate of Monte Carlo does not depend on the dimensionality!**

The Catch

So far so good: we'll just draw a lot of samples to drive down our Monte Carlo error.

Here's the catch! How do you draw samples from the posterior $p(\theta \mid x)$?

We're interested in Monte Carlo for cases where the posterior does not admit a simple closed form!

In general, sampling the posterior is as hard as computing the marginal likelihood.

Markov Chains

A *Markov chain* is a joint distribution of a sequence of variables, $\pi(\theta_1, \theta_2, \ldots, \theta_M)$. (To avoid confusion with the model *p*, we denote the densities associated with the Markov chain by π.) The Markov chain factorizes so that each variable is drawn conditional on the previous variable,

$$\pi(\theta_1, \theta_2, \ldots, \theta_M) = \pi_1(\theta_1) \prod_{m=2}^{M} \pi(\theta_m \mid \theta_{m-1}).$$

This is called the *Markov property*.

- The distribution $\pi_1(\theta_1)$ is called the *initial distribution*.
- The distribution $\pi(\theta_m \mid \theta_{m-1})$ is called the *transition distribution*. If the transition distribution is the same for each *m*, the Markov chain is *homogeneous*.

Stationary distributions

Let $\pi_m(\theta_m)$ denote the marginal distribution of sample θ_m. It can be obtained recursively as,

$$\pi_m(\theta_m) = \int \pi_{m-1}(\theta_{m-1})\,\pi(\theta_m \mid \theta_{m-1})\,\mathrm{d}\theta_{m-1}.$$

We are interested in the asymptotic behavior of the marginal distributions as $m \to \infty$.

A distribution $\pi^\star(\theta)$ is a **stationary distribution** if,

$$\pi^\star(\theta) = \int \pi^\star(\theta')\,\pi(\theta \mid \theta')\,\mathrm{d}\theta'.$$

That is, suppose the marginal of sample θ' is $\pi^\star(\theta)$. Then the marginal of the next time point is also $\pi^\star(\theta)$.

Detailed balance

How can we relate transition distributions and stationary distributions? A sufficient (but not necessary) condition for $\pi^\star(\theta)$ to be a stationary distribution is that it satisfies *detailed balance*,

$$\pi^\star(\theta')\pi(\theta \mid \theta') = \pi^\star(\theta)\pi(\theta' \mid \theta).$$

In words, the probability of starting at θ' and moving to θ is the same as that of starting at θ and moving to θ', if you draw the starting point from the stationary distribution.

To see that detailed balance is sufficient, integrate both sides to get,

$$\int \pi^\star(\theta')\pi(\theta \mid \theta')\,\mathrm{d}\theta' = \int \pi^\star(\theta)\pi(\theta' \mid \theta)\,\mathrm{d}\theta' = \pi^\star(\theta).$$

Thus, $\pi^\star(\theta)$ is a stationary distribution of the Markov chain with transitions $\pi(\theta \mid \theta')$.

Ergodicity

Detailed balance can be used to show that $\pi^\star(\theta)$ is *a* stationary distribution, but not that it is *the unique* one. This is where *ergodicity* comes in. A Markov chain is ergodic if $\pi_m(\theta_m) \to \pi^\star(\theta)$ regardless of $\pi_1(\theta_1)$. An ergodic chain has only one stationary distribution, $\pi^\star(\theta)$.

The easiest way to prove ergodicity is to show that it is possible to reach any θ' from any other θ. E.g. this is trivially so if $\pi(\theta' \mid \theta) > 0$.

Note: A more technical definition is that all pairs of sets *communicate*, in which case the chain is *irreducible*, and that each state is *aperiodic*. The definitions can be a bit overwhelming.

Markov Chain Monte Carlo (MCMC)

Finally, we come to our **main objective**: designing a Markov chain for which *the posterior is the unique stationary distribution*. That is, we want $\pi^\star(\theta) = p(\theta \mid x)$.

Recall our **constraint**: we can only compute the joint probability (the numerator in Bayes' rule), not the marginal likelihood (the denominator). Fortunately, that still allows us to compute ratios of posterior densities! We have,

$$\frac{p(\theta \mid x)}{p(\theta' \mid x)} = \frac{p(\theta, x)}{p(x)} \frac{p(x)}{p(\theta', x)} = \frac{p(\theta, x)}{p(\theta', x)}.$$

Now rearrange the detailed balance condition to relate ratios of transition probabilities to ratios of joint probabilities,

$$\frac{\pi(\theta \mid \theta')}{\pi(\theta' \mid \theta)} = \frac{\pi^\star(\theta)}{\pi^\star(\theta')} = \frac{p(\theta \mid x)}{p(\theta' \mid x)} = \frac{p(\theta, x)}{p(\theta', x)}$$

The Metropolis-Hastings algorithm

To construct such a transition distribution $\pi(\theta \mid \theta')$, break it down into two steps.

1. Sample a proposal θ from a *proposal distribution* $q(\theta \mid \theta')$,
2. Accept the proposal with *acceptance probability* $a(\theta' \to \theta)$. (Otherwise, set $\theta = \theta'$.)

Thus,

$$\pi(\theta \mid \theta') = \begin{cases} q(\theta \mid \theta')\, a(\theta' \to \theta) & \text{if } \theta' \neq \theta \\ \int q(\theta'' \mid \theta')\,(1 - a(\theta' \to \theta''))\,\mathrm{d}\theta'' & \text{if } \theta' = \theta \end{cases}$$

Detailed balance is trivially satisfied when $\theta = \theta'$. When $\theta \neq \theta'$, we need

$$\frac{\pi(\theta \mid \theta')}{\pi(\theta' \mid \theta)} = \frac{q(\theta \mid \theta')\, a(\theta' \to \theta)}{q(\theta' \mid \theta)\, a(\theta \to \theta')} = \frac{p(\theta, x)}{p(\theta', x)} \Rightarrow \frac{a(\theta' \to \theta)}{a(\theta \to \theta')} = \underbrace{\frac{p(\theta, x)\, q(\theta' \mid \theta)}{p(\theta', x)\, q(\theta \mid \theta')}}_{\triangleq A(\theta' \to \theta)}$$

The Metropolis-Hastings algorithm

WLOG, assume $A(\theta' \to \theta) \leq 1$. (If it's not, its inverse $A(\theta \to \theta')$ must be.) A simple way to ensure detailed balance is to set $a(\theta' \to \theta) = A(\theta' \to \theta)$ and $a(\theta \to \theta') = 1$.

We can succinctly capture both cases with,

$$a(\theta' \to \theta) = \min\left\{1, A(\theta' \to \theta)\right\} = \min\left\{1, \frac{p(\theta, x)\, q(\theta' \mid \theta)}{p(\theta', x)\, q(\theta \mid \theta')}\right\}.$$

The Metropolis algorithm

Now consider the special case in which the proposal distribution is symmetric; i.e. $q(\theta \mid \theta') = q(\theta' \mid \theta)$. Then the proposal densities cancel in the acceptance probability and,

$$a(\theta' \to \theta) = \min\left\{1, \frac{p(\theta, x)}{p(\theta', x)}\right\}.$$

In other words, you accept any proposal that moves “uphill,” and only accept “downhill” moves with some probability.

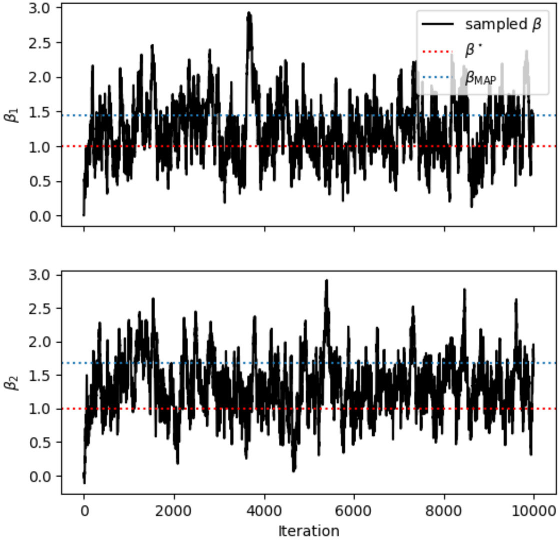
This is called the *Metropolis algorithm* and it has close connections to *simulated annealing*.

Synthetic Demo

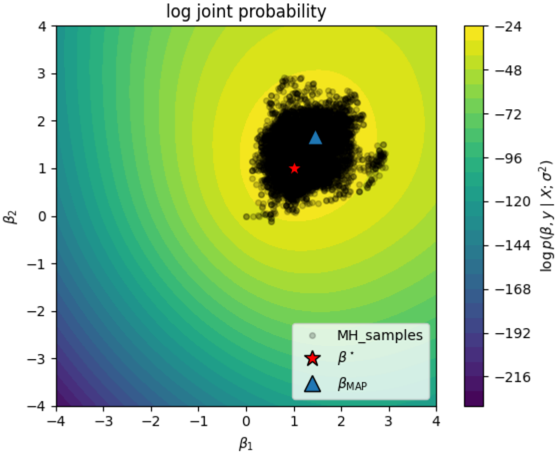
Let’s implement a simple Metropolis-Hastings sampler for the logistic regression model above.

We’ll use a spherical Gaussian proposal and play with the proposal variance.

Synthetic Demo



Synthetic Demo



Gibbs Sampling

Gibbs is a special case of MH with proposals that always accept. Gibbs sampling updates one "coordinate" of $\theta \in \mathbb{R}^D$ at a time by sampling from its conditional distribution. The algorithm is:

Algorithm: Gibbs Sampling

Input: Initial parameters $\theta^{(0)}$, observations x

```
► For $t = 1, \dots, T$
    ► For $d = 1, \dots, D$
        ► Sample $\theta_d^{(t)} \sim p\big(\theta_d \mid \theta_1^{(t)}, \dots, \theta_{d-}^{(t)}, \theta_{d+1}^{(t-1)}, \dots, \theta_D^{(t-1)}, x\big)$
► Return samples $\{\theta^{(t)}\}_{t=1}^{T}$
```

Gibbs Sampling

You can think of Gibbs as cycling through D Metropolis-Hastings proposals, one for each coordinate $d \in 1, \ldots, D$,

$$q_d(\theta \mid \theta') = p(\theta_d \mid \theta'_{\neg d}, x) \, \delta_{\theta'_{\neg d}}(\theta_{\neg d}),$$

where $\theta_{\neg d} = (\theta_1, \ldots, \theta_{d-1}, \theta_{d+1}, \ldots, \theta_D)$ denotes all parameters except θ_d.

In other words, the proposal distribution q_d samples θ_d from its conditional distribution and leaves all the other parameters unchanged.

Gibbs Sampling

What is the probability of accepting this proposal?

$$\begin{aligned}
a_d(\theta' \to \theta) &= \min\left\{1, \frac{p(\theta, x) q_d(\theta' \mid \theta)}{p(\theta', x) q_d(\theta \mid \theta')}\right\} \\
&= \min\left\{1, \frac{p(\theta, x) p(\theta'_d \mid \theta_{\neg d}, x) \delta_{\theta_{\neg d}}(\theta'_{\neg d})}{p(\theta', x) p(\theta_d \mid \theta'_{\neg d}, x) \delta_{\theta'_{\neg d}}(\theta_{\neg d})}\right\} \\
&= \min\left\{1, \frac{p(\theta_{\neg d}, x) p(\theta_d \mid \theta_{\neg d}, x) p(\theta'_d \mid \theta_{\neg d}, x) \delta_{\theta_{\neg d}}(\theta'_{\neg d})}{p(\theta'_{\neg d}, x) p(\theta'_d \mid \theta'_{\neg d}, x) p(\theta_d \mid \theta'_{\neg d}, x) \delta_{\theta'_{\neg d}}(\theta_{\neg d})}\right\} \\
&= \min\{1, 1\} = 1
\end{aligned}$$

for all θ, θ' that differ only in their d-th coordinate.

The Godfather: The Gibbs proposal is an offer you cannot refuse.

Gibbs Sampling

Of course, if we only update one coordinate, the chain can’t be ergodic. However, if we cycle through coordinates it generally will be.

Question: If Gibbs sampling always accepts, is it strictly better than other Metropolis-Hastings algorithms?

Conclusion

There's plenty more to be said about Bayesian statistics — choosing a prior, subjective vs objective vs empirical Bayesian approaches, the role of the marginal likelihood in Bayesian model comparison, varieties of MCMC, and other approaches to approximate Bayesian inference.

We'll dig into some of these topics as the course goes on, but for now, we have some valuable tools for developing Bayesian modeling and inference with discrete data!