TOPIC Brostat 778

2007-11-30

FILE UNDER

PAGE

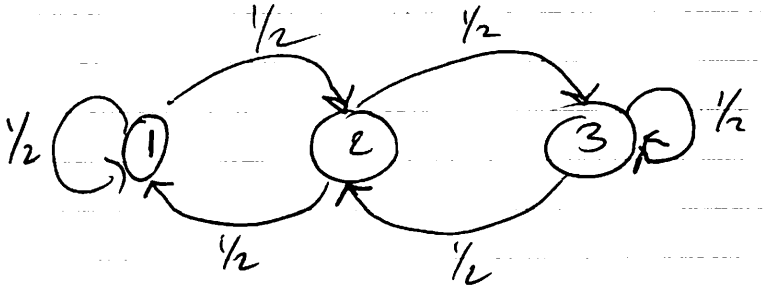
Markov Chains

⇒ A Markov chain is a stochastic process $\{X_i\}$ which satisfies

$$\mathbb{P}(X_i \mid X_{i-1}, \ldots, X_0) = \mathbb{P}(X_i \mid X_{i-1})$$

so that the current state of the chain X_i only depends on the previous value X_{i-1}.

⇒ The possible values $\{X_i\}$ can take is called the state space



$$P = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix}$$

Transition matrix (kernel)

$$\mathbb{P}(X_n = j \mid X_{n-1} = i) = P_{ij}$$

Suppose we start the chain with initial dist π_0, what is distribution after n iterations?

$(0, 0, 1)$

TOPIC

DATE

FILE UNDER

PAGE

Given transition matrix P

$$\mathbb{P}(X_n = j \mid X_0 = i) = (P^n)_{ij}$$

where $P^n = \underbrace{P \times P \times P \times \cdots \times P}_{n \text{ times}}$

$$\pi_n = \pi_0 P^n = \pi_0 \underbrace{P \times P \times P \times \cdots \times P}_{n \text{ times}}$$

Let π^* be a vector (distribution) such that

$$\pi_* P = \pi_*$$

Call π_* a stationary distribution and a MC ~~that~~ is said to be stationary if it reaches ~~has~~ this distribution.

Basic limit theorem says that under some conditions $\| \pi_* - \pi_n \| \longrightarrow 0$ as $n \to \infty$.

No matter how we start the chain (π_0) π_n approaches π_*.

TOPIC DATE

FILE UNDER PAGE

Assumptions:

(1) Irreducible: $P_{ij}^n > 0$ for some n and for every i, j

All pairs i, j are accessible from one another

(2) Aperiodic: the chain does not make deterministic visits to a subset of the state space.

ex: $X_0 = 0 \quad X_n = X_{n-1} + \varepsilon_n$

$$\varepsilon_n = \begin{cases} 1 & 1/2 \\ -1 & 1/2 \end{cases}$$

$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

$P = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$

For n even, the chain hits subset of even #s
For n odd " " " " odd #s

(3) Positive Recurrent: Every state is visited i.o.
The waiting time between visits from any 2 states is finite

(3) Stationary dist exists

TOPIC: DATE:

FILE UNDER: PAGE:

Let $X_0, X_1, \ldots$ be an irreducible, aperiodic, Markov chain w/ stationary dist. π.

Let $X_0 \sim \pi_0$, some arbitrary distribution. Then

$$\pi_n(i) \longrightarrow \pi(i) \quad \forall i \text{ as } n \to \infty$$

~~irreducible — all states communicate w/ each other~~

Time Reversibility

A Markov Chain is time reversible if

$$(X_0, X_1, \ldots X_n) \overset{D}{=} (X_n, X_{n-1}, \ldots X_0)$$

If $\{X_n\}$ is TR, then

$$(X_0, X_1) \overset{D}{=} (X_1, X_0)$$

$$\Rightarrow \quad X_0 \overset{D}{=} X_1$$

$$\Rightarrow \quad \pi_1 = \pi_0.$$

Since $\pi_1 = \pi_0 P$, the initial dist. is stationary

Let $\pi = \pi_0$. TR chains have

$$\pi_i P_{ij} = \pi_j P_{ji}$$

g. from $i \to j$ g. from $j \to i$

TOPIC

DATE

FILE UNDER:

PAGE

MCMC ⟹

Note that

$$\mathbb{P}(X_n = j \mid X_{n-1} = i) = \frac{\mathbb{P}(X_n = j, X_{n-1} = i)}{\pi_i}$$

$$TR = \frac{\mathbb{P}(X_n = i, X_{n-1} = j)}{\pi_i}$$

$$= \mathbb{P}(X_{n-1} = j \mid X_n = i)$$

$\Rightarrow P(i \to j) = P(j \to i)$ "flux" of $i \to j$ = $j \to i$

Main Ideas:

(1) We want to sample some complicated density π.

(2) We know that certain Markov chains will converge to a stationary distribution (if ergodic)

(3) How to construct a MC s.t. the sequence of values $\{X_n\}$ converges to a target distribution π?

We know that if a MC with transition matrix P or kernel $K(x,y)$ is time reversible wrt π, then π must be the stationary dist of the MC.

Given the chain we start at some point and run it until convergence

Problem: We cannot start at the stationary distribution!

TOPIC:

DATE:

FILE UNDER:

PAGE:

Metropolis-Hastings

Let $q(y|x)$ be a transition density from which we can easily simulate.

Let $X_n = x$ ——— $\frac{p(x,y)}{p(x)}$ ——— $p(y)$

① Simulate a candidate $Y \sim q(y|x)$

② Let $\alpha(y|x) = \min\left\{\frac{\pi(y)\,q(x|y)}{\pi(x)\,q(y|x)}, 1\right\}$ (not necessary)

③ Simulate $U \sim \text{unif}(0,1)$ ~~add accept~~

If $U \le \alpha(y|x)$ then set $X_{n+1} = Y$.
Otherwise set $X_{n+1} = X$.

$K(dy|x) \approx P(X_{n+1} \in dy | X_n = x)$

Why does this work? ~~We need to show~~ Let $K(y|x)$ be the transition density of the Metropolized chain.

We need to show

$$\pi(y)\,K(x|y) = \pi(x)\,K(y|x)$$

Constants of prop. cancel out

$K(y|x) = \alpha(y|x)\,q(y|x)$, so we need

$$\alpha(y|x)\,q(y|x)\,\pi(x) = \alpha(x|y)\,q(x|y)\,\pi(y)$$

$$\min\left(\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, 1\right) q(y|x)\pi(x) = \min\left(\frac{\pi(x)q(y|x)}{\pi(y)q(x|y)}, 1\right) q(x|y)\pi(y)$$

$$\min\left(\pi(y)q(x|y), q(y|x)\pi(x)\right) = \min\left(\pi(x)q(y|x), q(x|y)\pi(y)\right)$$

Obvious

TOPIC: Metropolis Hastings Time-Reversals

DATE:

FILE UNDER:

PAGE:

$$K(y|x) = \underbrace{\alpha(y|x)\, q(y|x)}_{(1)} + \mathbb{1}\{y = x\} \underbrace{\left[1 - \int \alpha(s|x)\, q(s|x)\, ds\right]}_{(2)}$$

(1) Show $K(y|x)\pi(x) = K(x|y)\pi(y)$

$$\Leftrightarrow \quad \alpha(y|x) q(y|x) \pi(x) = \alpha(x|y) q(x|y) \pi(y)$$

$$\Leftrightarrow \quad \min\left(\frac{\pi(y)}{\pi(x)} \frac{q(x|y)}{q(y|x)}, 1\right) q(y|x)\pi(x)$$

$$= \min\left(\frac{\pi(x)}{\pi(y)} \frac{q(y|x)}{q(x|y)}, 1\right) q(x|y)\pi(y)$$

$$\Leftrightarrow \quad \min(\pi(y) q(x|y), q(y|x)\pi(x)) = \min(\pi(x) q(y|x), q(x|y)\pi(y))$$

(2) $\mathbb{1}\{y = x\}\, r(x) = \mathbb{1}\{y = x\} \left[1 - \int \alpha(s|x)\, q(s|x)\, ds\right]$

$\delta_x(y)$

$\delta_x(y) r(x) \pi(x) = \delta_y(x) r(y) \pi(y)$

over set of points where $y = x$. Obvious!

TOPIC:

DATE

FILE UNDER:

PAGE:

Variations on M-H

Random Walk

Let $q(y|x)$ be defined by

$$y = x + \varepsilon$$

where $\varepsilon \sim g$ and g is symmetric about 0.

Note $q(y|x) = g(\varepsilon)$

$q(x|y) = g(-\varepsilon)$ $\Rightarrow g(\varepsilon)$

So the Metropolis acceptance prob is

$$\alpha(y|x) = \min\left\{\frac{\pi(y)\, q(x|y)}{\pi(x)\, q(y|x)}, 1\right\}$$

$$= \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}$$

Start at $X_n = x$.

(1) Simulate $\varepsilon \sim g$. Let $y = X_n + \varepsilon$

(2) Simulate $u \sim \text{unif}(0,1)$.

(3) If $u \le \alpha(y|x_n) = \min\left(\frac{\pi(y)}{\pi(x_n)}, 1\right)$

Then $X_{n+1} = y$, else $X_{n+1} = X_n$.

TOPIC: DATE:

FILE UNDER: PAGE:

Ex. Generating Normal R.V.s

Let q be $Unif[-\delta, \delta]$ dist.

(1) Simulate $\varepsilon \sim U(-\delta, \delta)$

(2) ~~Set~~ Simulate $u \sim unif(0,1)$

(3) Accept $y = x + \varepsilon$ if

$$u \leq \min\left\{ \frac{\phi(y)}{\phi(x)}, 1 \right\}$$

Independence Metropolis Algorithm (Tierney)

Define $q(y|x) = q(y)$.

So

$$\alpha(y|x) = \min\left(\frac{\pi(y) q(x)}{\pi(x) q(y)}, 1 \right)$$

$$= \min\left(\frac{w(y)}{w(x)}, 1 \right)$$

↳ importance weights.

(1) Simulate $y \sim q(y)$

(2) Simulate $u \sim unif(0,1)$

(3) Accept y if $u \leq \min\left(\frac{w(y)}{w(x)}, 1 \right)$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Independence Metropolis Sampler seems to work well in same situations as rejection sampling.

If $C = \sup_x \frac{\pi(x)}{q(x)} < \infty$, then we have

$\|\pi_n - \pi\| < K\rho^n$ for $0 < \rho < 1$

Rate of convergence ρ depends on C being small (close to 1).

Gibbs Sampler (2 variables)

Let $\pi(x,y)$ be your target density.

Let $Z = (X, Y)$ so that we want to generate a MC $\{Z_n\}$ where π is the stationary dist of the chain. The full conditionals of π are

$$p(y|x) = \frac{\pi(x,y)}{p(x)} \propto \pi(x,y)$$

$$p(x|y) = \frac{\pi(x,y)}{p(y)} \propto \pi(x,y)$$

Let $Z_n = (X_n, Y_n)$. The Gibbs sampler obtains Z_{n+1} by

(1) Simulate $X_{n+1} \sim p(x|Y_n)$

(2) Simulate $Y_{n+1} \sim p(y|X_{n+1})$

$Z_{n+1} = (X_{n+1}, Y_{n+1})$

TOPIC:

DATE:

FILE UNDER:

PAGE:

For MCs, need

- Irreducibility — positive probability of visiting any set of measure > 0
- Aperiodic — no regularly scheduled visits
- Positive Harris Recurrence
  - ↪ Any set with $\mu(A) > 0$ is visited infinitely often

true if sample space is product

$\Big\}$ Imply $X_n | X_0 = x \longrightarrow \pi$

stationary dist

Slice Sampling

Given state X_n, Generate a

auxiliary variable ←

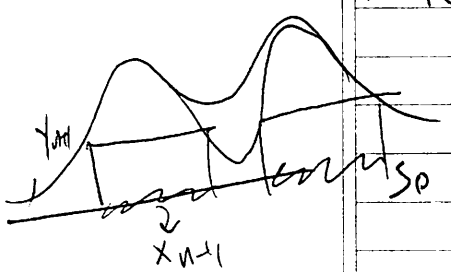
$$Y_{n+1} \sim \text{Unif}[0, \pi(X_n)]$$

Given Y_{n+1}, generate

$$X_{n+1} \sim \text{Unif}[\{x : \pi(x) \geq Y_n\}]$$

Note that

$$\pi(x) = \int \mathbb{1}\{0 \leq y \leq \pi(x)\}\, dy$$



so that

$$f(x,y) = \mathbb{1}\{0 \leq y \leq \pi(x)\}$$ is a joint density

TOPIC:

DATE

FILE UNDER:

PAGE:

Gibbs Sampling (we want to sample from some posterior or other complicated density)

For K-dim distribution, only need K full conditional dist., not K(K-1) conditionals like w/ DA.

$p(x, y, z)$ → target density

$\Downarrow$

$p(x \mid y, z)$,

$p(y \mid x, z)$

$p(z \mid x, y)$

Given the n^{th} iteration (x_n, y_n, z_n)

(1) Sample $x_{n+1} \sim p(x \mid y_n, z_n)$

(2) Sample $y_{n+1} \sim p(y \mid x_{n+1}, z_n)$

(3) Sample $z_{n+1} \sim p(z \mid x_{n+1}, y_{n+1})$

$[x_n, y_n, z_n] \xrightarrow{D} [x, y, z]$

$[x_n] \xrightarrow{D} [x]$

$[y_n] \xrightarrow{D} [y]$

$[z_n] \xrightarrow{D} [z]$

$\frac{1}{N} \sum_{n=1}^{N} f(x_n, y_n, z_n) \longrightarrow \mathbb{E}[f(x, y, z)]$

Geman + Geman

TOPIC

DATE

FILE UNDER:

PAGE:

$K(Z_{n+1}|Z_n) = p(y_{n+1}|x_{n+1})\, p(x_{n+1}|y_n)$

(no x_n)

Ex. Simulate $N(\mu, \Sigma) = N((\mu_1, \mu_2), \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix})$

Set $Z_n = (X_n, Y_n)$

① Simulate $X_{n+1} \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_n - \mu_2),\ \sigma_1^2(1-\rho)\right)$

② Simulate $Y_{n+1} \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_n - \mu_1),\ \sigma_2^2(1-\rho)\right)$

$\rho = \frac{c}{\sigma_1\sigma_2}$

Ex: Let $y_1, \dots, y_n \sim N(\mu, \tau^{-1})$ ↪ precision

Suppose we put priors (indep.)

$\mu \sim N(0, w^{-1})$

$\tau \sim \text{Gamma}(\alpha, \beta)$

No conjugacy, so we need Gibbs sampling to explore posterior $\mu, \tau | y_1, \dots, y_n$.

$p(\mu, \tau, y) \propto L(\mu, \tau | y)\, p(\mu)\, p(\tau)$

$p(\mu | \tau, y) \propto L(\mu, \tau | y)\, p(\mu)$

$p(\tau | \mu, y) \propto L(\mu, \tau | y)\, p(\tau)$

TOPIC

DATE

FILE UNDER:

PAGE

$$p(\mu, \tau, y) = \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left(-\frac{\tau}{2}\sum(y_i-\mu)^2\right)$$

$$\times \left(\frac{w}{2\pi}\right)^{1/2} \exp\left(-\frac{w}{2}\mu^2\right)$$

$$\times \frac{\tau^{\alpha-1}}{\Gamma(\alpha)}\beta^{\alpha}\exp(-\tau\beta)$$

$$p(\tau|\mu, y) \propto \tau^{n/2}\exp\left(-\frac{\tau}{2}\sum(y_i-\mu)^2\right)$$

$$\times \tau^{\alpha-1}\exp(-\tau\beta)$$

$$= \tau^{\alpha-1+n/2}\exp\left(-\tau\left[\beta+\frac{1}{2}\sum(y_i-\mu)^2\right]\right)$$

$$= \text{Gamma}\left(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum(y_i-\mu)^2\right)$$

$$p(\mu|\tau, y) \propto \exp\left(-\frac{\tau}{2}\sum(y_i-\mu)^2\right)$$

$$\times \exp\left(-\frac{w}{2}\mu^2\right)$$

$$= \exp\left(-\frac{(n\tau+w)}{2}\mu^2 - (\Sigma y_i)\tau\mu\right)$$

$$= N\left(\left(\frac{n\tau}{n\tau+w}\right)\frac{1}{n}\Sigma y_i, \frac{1}{n\tau+w}\right)$$

Gibbs sampler iterates between those densities

TOPIC:

DATE:

FILE UNDER:

PAGE:

~~$p(\alpha, \sigma^2, \mu_1 \dots y_n) = \prod_{i=1}^{n} p(y_i | \mu_i) p(\mu, \alpha, \sigma^{-2})$~~

~~$\mu^y e^{-\mu}$~~

Relationship b/w Gibbs sampling + Metropolis - Hastings

In M-H we have proposal $q(y|x)$ and acceptance prob.

$$\alpha(y|x) = \min\left(\frac{\pi(y)\, q(x|y)}{\pi(x)\, q(y|x)}, 1\right)$$

~~In Gibbs sampling~~ We can do Single-component M-H

Let $X^{(t)} = (X_1^{(t)}, X_2^{(t)}, \dots, X_K^{(t)})$ be a K-vector representing our quantities of interest (parameters) at iteration t.

Let $X_{-i}^{(t)} = (X_1^{(t)}, X_2^{(t)}, \dots, X_{i-1}^{(t)}, X_{i+1}^{(t)}, \dots, X_K^{(t)})$

SCMH, at iteration t updates the i^{th} component $X_i^{(t)}$ via:

1. Sample $Y_i \sim q_i(y | X_i^{(t)}, X_{-i})$ for component i
2. Let $\alpha(y_i | X_i^{(t)}) = \min\left(\frac{\pi(y_i | X_{-i}) q(x_i | y_i, X_{-i})}{\pi(x_i | X_{-i}) q(y_i | x_i, X_{-i})}, 1\right)$
3. Accept y_i for component i w/prob $\alpha(y_i | X_i^{(t)})$

Repeat K times.

$\pi(X_i | X_{-i})$ is a full conditional distribution

TOPIC:

DATE:

FILE UNDER:

PAGE:

Blocking

In Gibbs sampling

$$q(y_i \mid X_i, X_{-i}) = \pi(y_i \mid X_{-i})$$

This gives us

$$\alpha(y_i \mid X_i, X_{-i}) = \min\left(\frac{\pi(y_i \mid X_{-i})\,\pi(x_i \mid X_{-i})}{\pi(x_i \mid X_{-i})\,\pi(y_i \mid X_{-i})}, 1\right)$$

$$= 1$$

So Gibbs sampling is like SCMH but it always accepts

General M-H is very exploratory, maybe too much.

Hybrid Gibbs Sampler

Sometimes it is not possible to sample directly from a full conditional dist.

One can use a hybrid GS in that case. Suppose we have 2-variable problem w/ (X, Y) and we can easily sample $p(x \mid y)$ but not $p(y \mid x)$. Then for $y_{n+1} \mid x_n$

1. Simulate $y \sim q(y \mid x_n)$, draw $u \sim \text{unif}(0,1)$
2. Accept y if $u \le \min\left(\frac{p(y \mid x_n)\, q(y_n \mid x_n)}{p(y_n \mid x_n)\, q(y \mid x_n)}, 1\right)$

We just need to simulate one value and move on for things to work

TOPIC:

DATE:

FILE UNDER:

PAGE:

236hkt

Hybrid GS

$$Y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$\text{logit}\, p_{ij} = \alpha_i$$

$$\alpha_i \sim N(\mu, \sigma^2)$$

$$\mu \sim N(0, D)$$

$$\sigma^2 \sim IG(a, b)$$

$$p(Y, \alpha, \mu, \sigma^2) \propto \left[\prod_{i=1}^{n}\left[\prod_{j=1}^{G} p(Y_{ij}|\alpha_i)\right] p(\alpha_i|\mu, \sigma^2)\right] \pi(\mu)\,\pi(\sigma^2)$$

(1) $$p(\alpha_i|\cdots) \propto \left[\prod_{j=1}^{G} p(Y_{ij}|\alpha_i)\right] p(\alpha_i|\mu, \sigma^2) \qquad i = 1, \ldots, n$$

(2) $$p(\mu|\cdots) \propto \left[\prod_{i=1}^{n} p(\alpha_i|\mu, \sigma^2)\right] \pi(\mu)$$

$$= N\left(\frac{D}{D+\sigma^2}\bar{\alpha},\ \frac{\sigma^2}{\sigma^2+D} D\right)$$

(3) $$p(\sigma^2|\cdots) \propto \left[\prod_{i=1}^{n} p(\alpha_i|\mu, \sigma^2)\right] \pi(\sigma^2)$$

$$IG\left(a + \frac{n}{2},\ b + \frac{1}{2}\sum(\alpha_i - \mu)^2\right)$$

Reparametrization in MH/Gibbs

Suppose we want to sample from a full conditional $p(\lambda | w)$ where $\lambda \in (0,1)$. Random walk is tricky b/c of edge/boundary cases. We can

Let $z = \text{logit}(\lambda) = \log \frac{\lambda}{1-\lambda}$.

Then propose values of $z \in (-\infty, \infty)$, accept/reject, and transform back.

(1) propose $z^* \sim g(z | z_n)$, $z_n = \text{logit}\, \lambda_n$

(2) $$\alpha(z | z_n) = \min\left\{ \frac{p(\text{logit}^{-1}(z^*) | w)}{p(\text{logit}^{-1}(z_n) | w)} \frac{g(z_n | z^*)}{g(z^* | z_n)} \frac{|J(z^*)|}{|J(z_n)|} \right\}$$

where $|J(z)|$ is the determinant of Jacobian of transformation from $u \longmapsto z$ (i.e. logit^{-1})

$$|J(z)| = \frac{1}{|J(u)|} \quad z \longmapsto u$$

Hamiltonian MCMC

The energy in a system is defined by the Hamiltonian,

$$H(q, p) = U(q) + K(p)$$

$U(q)$: potential energy; $K(p)$: kinetic energy

The partial derivatives of $H(q, p)$ determine how q and p change over time t,

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}$$

$$\frac{dp_i}{dt} = \frac{-\partial H}{\partial q_i}$$

} Hamilton's equations

for $i = 1, \ldots d$ = # dimensions

$\Rightarrow$ Solving these differential equations gives you $q(t)$ and $p(t)$ which allow you to describe how the system changes from time t to time $t+S$.

Properties of Hamiltonian dynamics

(1) Reversible ⟶ preserve stationary distribution

(2) Volume preservation ⟶ no Jacobian

(3) Conservation of Hamiltonian ⟶ acceptance prob = 1

Basic Procedure for HMC

(1) Sample momentum

Leapfrog Method — moving around the system

Assume $H(q,p) = U(q) + K(p)$

and $K(p) = \frac{1}{2}\sum_{i=1}^{d} \frac{p_i^2}{m_i}$ (Gaussian w/ diagonal var).

For step size ε, the basic idea is

$$p_i(t+\varepsilon) = p_i(t) + \varepsilon \frac{dp_i(t)}{dt} = p_i(t) - \varepsilon \frac{\partial U}{\partial q_i}(q(t))$$

$$q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{dq_i(t)}{dt} = q_i(t) + \varepsilon \frac{p_i(t)}{m_i}$$

Basic 1st order approx doesn't work well because of error propagation

Leapfrog Method says

① $p_i(t+\varepsilon/2) = p_i(t) - \frac{\varepsilon}{2}\frac{\partial U}{\partial q_i}(q(t))$

② $q_i(t+\varepsilon) = q_i(t) + \varepsilon \frac{p_i(t+\varepsilon/2)}{m_i}$

③ $p_i(t+\varepsilon) = p_i(t+\varepsilon/2) - \frac{\varepsilon}{2}\frac{\partial U}{\partial q_i}(q(t+\varepsilon))$

~~HMC Algorithm~~
HMC Algorithm

① Simulate p from
② Run leapfrog for L steps with step ε to generate proposal
③ Accept / Reject

For statistical applications $q(t)$ is vector of parameters that "evolves" over time and $p(t)$ is arbitrary. Usually, assume

$$K(p) = \frac{1}{2} p' M^{-1} p \quad \text{where} \quad M = \begin{bmatrix} m_1 & & 0 \\ & \ddots & \\ 0 & & m_d \end{bmatrix}$$

$$U(q) = -\log \pi(q) = -\log p(q) L(q|D)$$

$$\Rightarrow H(q, p) = U(q) + K(p) = -\log \pi(q) + \frac{1}{2} p' M^{-1} p$$

$\Rightarrow$ Need $\frac{\partial U}{\partial q_i}$ the gradient of U to run the leapfrog algorithm.

Symbolic differentiation?

deriv, D.

HMC for stat : At time t, we have $q(t)$

(1) Draw $p \sim N(0, M)$

(2) Make half step

$$p \leftarrow p - \frac{\varepsilon}{2} \frac{\partial U^{k}(q(t))}{\partial q^{k}}$$

(3) Take L steps, $j = 1, \to L-1$

$$q_i(t+\varepsilon) \leftarrow q_i(t) + \varepsilon\, p_i(t+\varepsilon/2)$$

$$p_i(t+\varepsilon) = p_i(t+\varepsilon/2) - \varepsilon \frac{\partial U}{\partial q_i}(q(t))$$

(4) $$p = p - \frac{\varepsilon}{2} \frac{\partial U}{\partial q}(q(t))$$

(5) $$\alpha = \min\left\{ \frac{\exp(-H(q^*, p^*))}{\exp(-H(q, p))}, 1 \right\}$$

(proposal is symmetric)

TOPIC:

DATE:

FILE UNDER:

PAGE:

(R) Check autocorrelation of your chain (marginally) using acf. Goal: chains should have decreasing acf. Faster decrease indicates faster mixing.

$$acf(1) = \frac{\frac{1}{n}\sum (X_t - \bar{X})(X_{t-1} - \bar{X})}{\frac{1}{n}\sum (X_t - \bar{X})^2}$$

(T) Computer clusters allow easy implementation of multiple chains. So why not?

Simulated Annealing

Suppose we want to minimize a function $h(\theta)$, where θ is a vector of parameters. $\theta \in S$

The basic idea is to simulate a MC with the stationary dist.

MC for simulation

"Annealing"

$$\pi_T(\theta) \propto \exp(-h(\theta)/T)$$

The parameter T is called the "temperature".

Markov Chains for "simulation", annealing (cooling) for minimization.

Let $S^* = \{\theta^* \in S : h(\theta^*) = \min_\theta h(\theta)\}$

the set of global minimizers

Define a distribution π^* on S^* so that

$$\pi^*(\theta) \propto 1 \quad \text{for } \theta \in S^* \text{ and } 0 \text{ otherwise}$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

~~Thm: As $T \downarrow 0$ we have $\pi_T \to \pi^*$~~

~~$\pi_T(\theta) \to \pi(\theta) \quad \forall \theta \in S$~~

FACT:

As $T \downarrow 0$, we have $\pi_T \to \pi^*$

Obviously, we'd like to simulate from π^* to get the global minimum, but we can't.

But for a fixed $T > 0$, we can simulate from π_T.

Use Metropolis algorithm with a symmetric proposal density, i.e. ~~$q(x|y) = q(y|x)$~~.

$$q(\theta_n | \theta_{n-1}) = q(\theta_{n-1} | \theta_n)$$

Suppose we are at state n, θ_n,

① Simulate $\theta \sim q(\theta | \theta_n)$

② Simulate $U \sim \text{unif}(0,1)$

③ Set $\theta_{n+1} = \theta$ if $U \le \min\left(\frac{\exp(-h(\theta)/T)}{\exp(-h(\theta_n)/T)}, 1\right)$

$$= \min\left(\exp\left(-\underbrace{(h(\theta) - h(\theta_n))}_{\Delta h} / T\right), 1\right)$$

Otherwise set $\theta_{n+1} = \theta_n$

④ Decrease T.

If $\Delta h < 0$, we always go

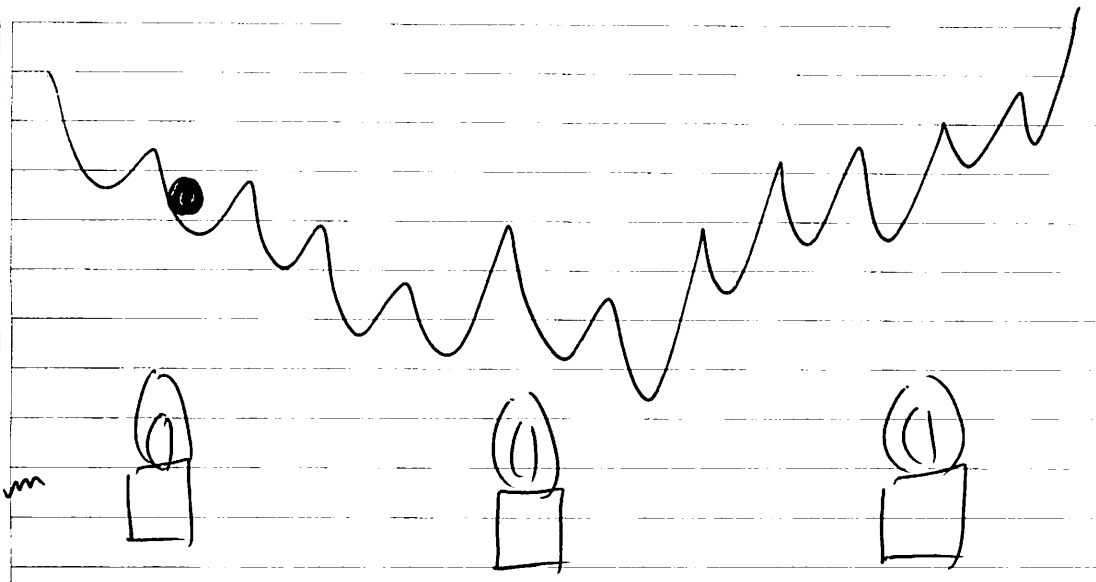
If $\Delta h > 0$, we go w/prob $\exp(-\Delta h / T)$

TOPIC:

DATE:

FILE UNDER:

PAGE:



Hi temp
↳ ball moves all over

lo temp
↳ ball stays in minimum

down hill

up hill

If $\Delta h < 0$, we always go

If $\Delta h > 0$, we go w/prob $\exp(-\Delta h/T)$

When T is closer to 0, it is less likely we will go up hill.

Can't cool too fast or will get stuck.

Need a cooling schedule like

$$T_n = \frac{a}{\log(n+b)}$$ (for global convergence)

Anything else is "too fast" and will get stuck in a local mode.

Thm: If

① $T_{n+1} < T_n \quad \forall n$

② $T_n \to 0$ as $n \to \infty$

③ $T_n = a/\log(n+b)$

Then $\| \pi_{T_n} - \pi^* \| \to 0$

TOPIC:

DATE:

FILE UNDER:

PAGE:

~~If we use~~

If we use a Metropolis step at both iterations, then we have SCMH.

Other possibilities are rejection sampling

K-Variable Gibbs sampler

let $(X_1^{(n)}, X_2^{(n)}, \dots, X_K^{(n)})$ be the state at step n. The Gibbs sampler obtains the next state $(n+1)$ by

$$X_1^{(n+1)} \sim X_1 \mid X_2^{(n)}, X_3^{(n)}, \dots, X_K^{(n)}$$

$$X_2^{(n+1)} \sim X_2 \mid X_1^{(n+1)}, X_3^{(n)}, \dots, X_K^{(n)}$$

$$\vdots$$

$$X_K^{(n+1)} \sim X_K \mid X_1^{(n+1)}, \dots, X_{K-1}^{(n+1)}$$

Sometimes, not!

Notes on Gibbs Sampling

(*) (1) It is sometimes useful to update groups of variables at a time (i.e. vector update rather than univariate). This is called "block Gibbs".

(R) (2) Starting values are a guess. Can try frequentist values (means, medians, modes) or sample from priors. Maybe multiple starts

(R) (3) Good to use a burn in period before ~~taking draws~~ storing values

TOPIC:

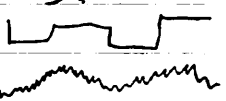
DATE:

FILE UNDER:

PAGE:

(4) Monitoring convergence is critical
- look at marginal trace plots
- try multiple chains with widely dispersed starting values
- Monitor the acceptance rate (~30% is good for > 2 dimensions, ~50% for 1,2-D.

too low acceptance → (slow convergence)
too high acceptance → (slow mixing)



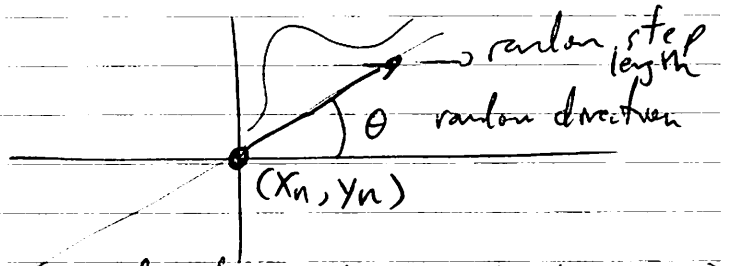
(5) Order of updating need not be sequential. One
- can do 1,2,3,4,3,2,1,2,3 ....
- or randomly choose a component ("random scan")
- randomly permute the order of updating
(1,2,3,4), (1,3,4,2), (4,3,2,1) ...

(6) Suppose X is multivariate and X_n is the current state. Choose a random direction e_n. Sample a scalar r from the density

$$p(r) \propto \pi(X_n + r e_n)$$

Set $X_{n+1} = X_n + r e_n$

In 2-D



- searches in more directions (than just along axes)
- how to sample from $p(r)$?
- using a Metropolis step results in random walk Metropolis

TOPIC: DATE:

FILE UNDER: PAGE:

(7) Gibbs sampler w/ improper priors is bad because it will run even if posterior is improper (no warning)

ex $X \sim N(\theta, \sigma^2)$

$$p(\theta, \sigma) = \frac{1}{\sigma^2}$$

$$\int\int p(x|\sigma,\theta)\, p(\theta,\sigma)\, d\theta\, d\sigma = \int\int \frac{1}{\sqrt{2\pi}\,\sigma^3} e^{-\frac{1}{2\sigma^2}(x-\theta)^2} d\theta\, d\sigma$$

$$= \int \frac{1}{\sigma^2} d\sigma = \infty$$

$\Rightarrow p(\theta, \sigma)$ is not a valid prior

But

$$p(\theta|\sigma, x) \propto e^{-\frac{1}{2\sigma^2}(x-\theta)^2} = N(x, \sigma^2)$$

$$p(\sigma|\theta, x) \propto \frac{1}{\sigma^3} e^{-(\theta-x)^2/2\sigma^2}$$

$$= \frac{1}{\sqrt{e}} \quad \text{where} \quad e \sim \text{Exp}\left(\frac{(\theta-x)^2}{2}\right)$$

Gibbs Sampler is easy to run!

↳ This chain is "null recurrent". There is a stationary/invariant function, but it is not a density.

(8) Your Gibbs Sampler may converge but that isn't always a good thing

(4)

TOPIC:

DATE:

FILE UNDER:

PAGE:

Start J chains

Markov Convergence (Gelman + Rubin)

Let $x_j^{(0)}, x_j^{(1)}, \ldots$ be the j^{th} Markov Chain

(1) Discard $X_j^{(0)}, X_j^{(1)}, \rightarrow X_j^{(b-1)}$ values (burn in)

Use values $X_j^{(b)}, X_j^{(b+1)}, \rightarrow X_j^{(b+L-1)}$

L values

$X_j^{(1)}, \rightarrow X_j^{(L)}$

(2) Calculate

$$\bar{X}_j = \frac{1}{L}\sum_{t=1}^{L} X_j^{(t)} \quad \text{(chain mean)}$$

$$\bar{X}_\cdot = \frac{1}{J}\sum_{j=1}^{J} \bar{X}_j \quad \text{(grand mean)}$$

$$B = \frac{L}{J-1}\sum_{j=1}^{J}(\bar{X}_j - \bar{X}_\cdot)^2 \quad \text{(between chain variance)}$$

$$s_j^2 = \frac{1}{L-1}\sum_{t=1}^{L}(X_j^{(t)} - \bar{X}_j)^2 \quad \text{(within chain variance)}$$

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

(3) Let $R = \dfrac{\frac{L-1}{L}W + \frac{1}{L}B}{W}$

$\sqrt{R} < 1.2$ acceptable?

TOPIC:

DATE:

FILE UNDER:

PAGE:

MCMC Dynamics

You need to babysit your Gibbs sampler. Look at output!

Try exhausting all iid sampling options

For estimation, one can use batch means. Suppose we have chain $X_1, X_2, X_3, \ldots$ and we want $\mathbb{E}h(X)$. Then we have

$$\underbrace{h(X_1)\ h(X_2)\ h(X_3) \ldots h(X_K)}_{\text{Batch 1}},\ \underbrace{h(X_{K+1}), \ldots}_{\text{Batch 2}}$$

Let $b_1 = \frac{1}{K}\sum_{i=1}^{K} h(X_i)$ $\qquad$ $b_2 = \frac{1}{K}\sum_{i=K+1}^{2K} h(X_i)$

Things for apart are roughly independent

~~Ergodic Theorem says~~

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i = \frac{1}{Km}\sum_{i=1}^{Km} h(X_i) \longrightarrow \mathbb{E}[h(X)]$$

m = # of batches

The batches should be roughly independent (by ergodic theorem) so for large m

$$\sqrt{m}\left(\frac{\bar{b} - \mathbb{E}h(x)}{s}\right) \longrightarrow N(0,1)$$

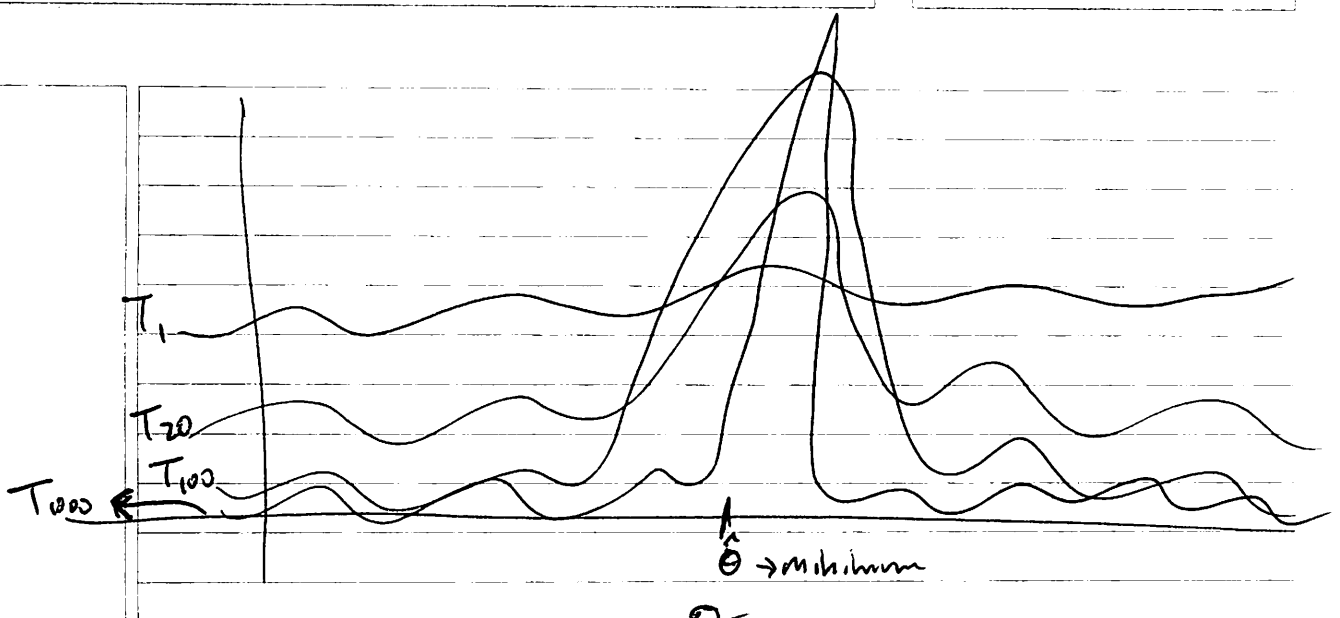
where $s^2 = \frac{K}{m}\sum (b_i - \bar{b})^2$, the sample variance of b_i's.

TOPIC:

DATE:

FILE UNDER:

PAGE:



As $T_n \to \infty$, the mass of π_{T_n} becomes increasingly concentrated about $\hat{\theta}$

Perfect Sampling (brief intro)

We know that under certain conditions, a MC will converge to a stationary distribution.
↳ at time $n = \infty$

But we don't know when!

We know a burn-in is a good idea, but how much?

When can we declare "convergence"?

Propp + Wilson (1996) discovered that perfect random samples can be obtained in finite (but stochastic) time.

⇒ Coupling from the past algorithm

TOPIC:

DATE:

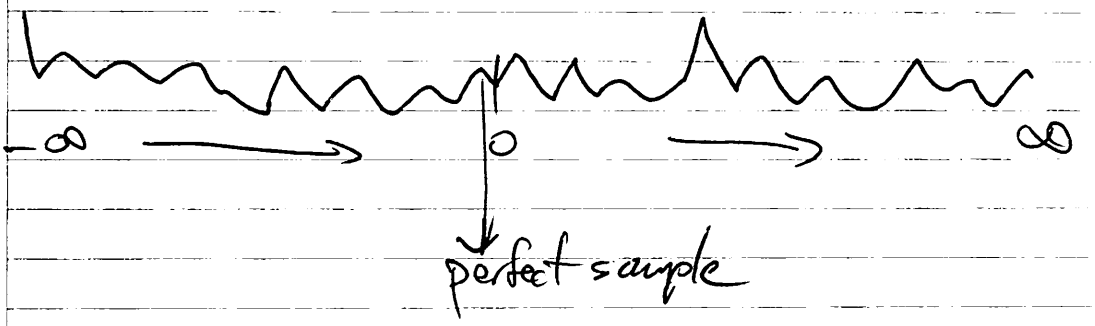
FILE UNDER:

PAGE:

Basic idea of CFTP

(1) If we start at $n=0$, then a sample from state $n=\infty$ will be a "perfect" sample from π, the stationary distribution.

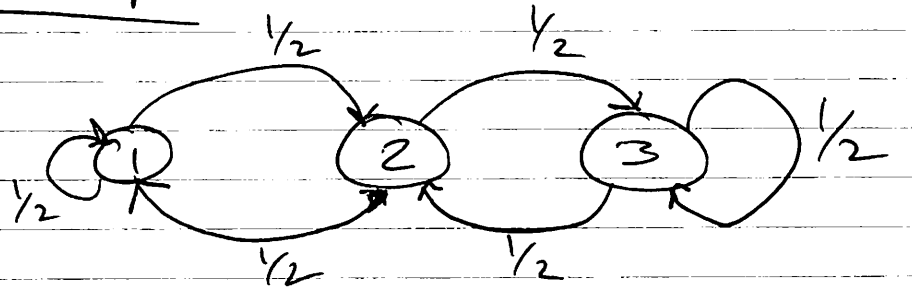
(2) So why don't we start the chain at state $n=-\infty$ and ~~take a sample from~~ use the value drawn at state $n=0$?



Both chains run for ∞ amount of time so must be stationary.

Discrete State space

Simple example



Transition matrix

$$P = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix}$$

Stationary dist: $\pi = (1/3, 1/3, 1/3)$

TOPIC:

DATE:

FILE UNDER:

PAGE:

The stationary dist. is uniform on all 3 states.

How do we simulate this Markov Chain?

Suppose X_n is the value in state n.

Draw a $U \sim \text{uniform}(0,1)$.

If $X_n = 1$ then goto $\begin{cases} X_{n+1} = 1 & U \le \frac{1}{2} \\ X_{n+1} = 2 & U > \frac{1}{2} \end{cases}$

If $X_n = 2$ then goto $\begin{cases} X_{n+1} = 1 & U \le 1/2 \\ X_{n+1} = 3 & U > 1/2 \end{cases}$

If $X_n = 3$ then goto $\begin{cases} X_{n+1} = 2 & U \le 1/2 \\ X_{n+1} = 3 & U > 1/2 \end{cases}$

Imagine running the chain in this way starting at $n = -\infty$ and ending at $n = 0$.

Can we emulate this process?

Notice:

(1) We can generate $U_n, U_{n+1}, U_{n+2}, \ldots$ because they are independent of everything.

(2) $X_{n+1} = \phi(U_n, X_n)$

TOPIC:

DATE:

FILE UNDER:

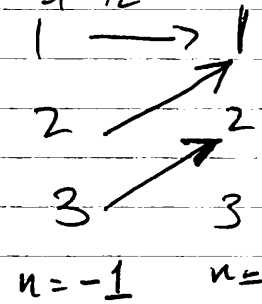
PAGE:

CFTP Algorithm for 1 perfect sample

~~Suppose we draw $U \leq 1/2$~~

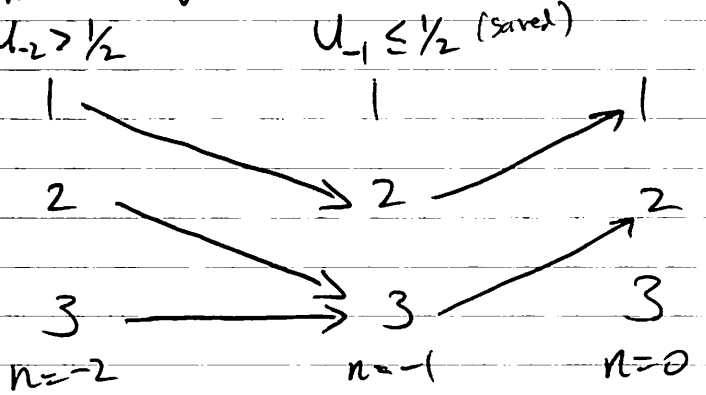
Start 3 parallel MCs each starting in state 1, 2, or 3 respectively.

Suppose we draw $U_{-1} \leq 1/2$ and start at $n=-1$



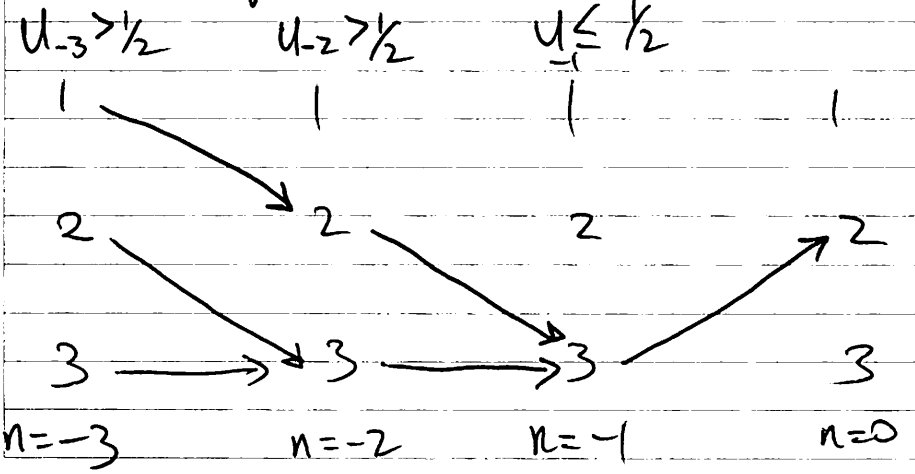
If we start the chains at $n=-1$ and draw $U \leq 1/2$, then we are at ~~state~~ state 1 or 2. (not 3)

Suppose we generate $U_{-2} > 1/2$ and start at $n=-2$.



We "lost" the second chain, or the 2nd and 3rd chains coalesced.

Now suppose generate $U_{-3} > 1/2$ and start at $n=-3$.



TOPIC:

DATE:

FILE UNDER:

PAGE:

When we start from $n=-3$, all of the chains coalesce at state $n=-1$.

$\Rightarrow$ Regardless of what happens b/w times $n=-\infty$ and $n=-3$ (we don't know), we know that at time $n=-1$ the will all coalesce at state (3).

$\Rightarrow$ Regardless of starting value, at time $n=0$, we will be at state (2).

$\Rightarrow$ (2) is a perfect sample from stationary dist.

$\Rightarrow$ Coalescence "breaks" dependence on the starting value

Note: In this chain, the arrows don't cross, so we only need to keep track of "top" chain and "bottom" chain

(1) Why can't run forward until coupling?
↳ only couples at (1) and (3)

(2) Why can't run until coupling and then a few more?
How many more?

Image Analysis (Simple)

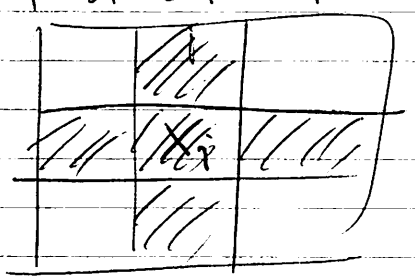
Let $X = (X_1, X_2, \ldots, X_n)$ be an image (true) where X_i is the value for pixel i.

Suppose X_i is a r.v. such that

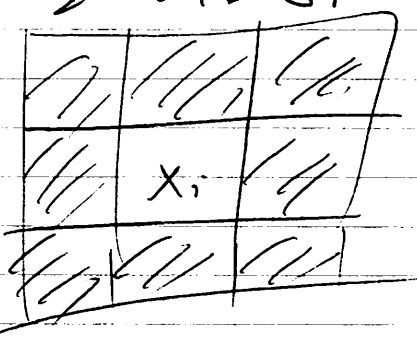
$$X_i \mid X_{-i} \overset{D}{=} X_i \mid X_{\delta_i}$$

where δ_i is a neighborhood of X_i

First order δ_i



2nd order δ_i



So $p(x_i \mid X_{-i}) = p(x_i \mid X_{\delta_i})$

Let $X_i = 0$ or 1 both with positive probability (binary image).

$\Rightarrow$ for a 40×40 image, the state space is of size 2^{1600}!

Uniquely characterize joint distribution of X (Hammersley–Clifford)

Let Y_i be a noisy version of X_i
and assume that

$$p(y_1 \to y_n \mid x_1 \to x_n) = \prod_{i=1}^{n} p(y_i \mid x_i)$$

e.g. white noise process
→ likelihood for $x_1 \to x_n$ given $y_1 \to y_n$

We want to sample $p(x|y) \propto p(y|x)\,p(x)$
One class of priors for X is

$$p(x) \propto \exp\left(\sum_{i \sim j} \phi(x_i - x_j)\right)$$

where ϕ is symmetric about 0.
Using a first order neighborhood structure for δ_i
we can Gibbs sample each pixel

$$X_i^{(t+1)} \mid X_{-i}^{(t)}, y \sim p(x_i \mid x_{\delta_i}^{(t)}, y)$$

full conditional

$$\propto p(y \mid x_{\delta_i}^{(t)}, x_i^{(t)})\, p(x_i^{(t)}, x_i^{(t)})$$

$$p(y_i \mid x_i^{(t)})\, \cancel{p(x_i^{(t)})}\, p(x_i^{(t)} \mid x_{-i}^{(t)})\, \cancel{p(x_{-i}^{(t)})}$$

$$p(x_i^{(t)} \mid x_{\delta_i}^{(t)})$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Spectral

Ex: Binary images

Given observed values $Y_1 \to Y_n \in \{0,1\}$, we want to estimate true image $X_1 \to X_n \in \{0,1\}$

$X_i = 1$ indicates presence of a species

$Y_i = 1$ observed a species

Let $p(y|x) \propto \exp\left(\alpha \sum \mathbb{1}\{Y_i = X_i\}\right)$

Use pairwise difference prior for X,

$$p(x) \propto \exp\left(\beta \sum_{i \sim j}^{n} \mathbb{1}\{X_i = X_j\}\right)$$

$$p(X_i^{(t+1)} = 1 \mid X_{-i}^{(\tau)}, Y) \propto p(Y|X_i)\, p(X_i \mid X_{-i})$$

$$\propto \exp\left(\alpha \mathbb{1}\{Y_i = X_i\}\right) \exp\left(\beta \sum_{i \sim j} \mathbb{1}\{\cdots\}\right)$$

$$\times \exp\left(\beta \sum_{i \sim j} \mathbb{1}\{X_i = X_j\}\right)$$

$$= \exp\left(\alpha \mathbb{1}\{Y_i = X_i\} + \beta \sum_{j \in \delta_{X_i}} \mathbb{1}\{X_j = X_i\}\right)$$

$$\exp\left(\alpha \mathbb{1}\{Y_i = 0\} + \beta \textstyle\sum \mathbb{1}\{X_j = 0\}\right) + \exp\left(\alpha \mathbb{1}\{Y_i = 1\} + \beta \sum_j \mathbb{1}\{X_j = 1\}\right)$$

$$= \left[1 + \exp\left(\alpha\left(\mathbb{1}\{Y_i = 0\} - \mathbb{1}\{Y_i = 1\}\right) + \beta \sum_{j \in \delta_i} \left(\mathbb{1}\{X_j = 0\} - \mathbb{1}\{X_j = 1\}\right)\right)\right]^{-1}$$

Floating Point Numbers

$n = f \times 10^{e}$ ⟶ exponent

↓

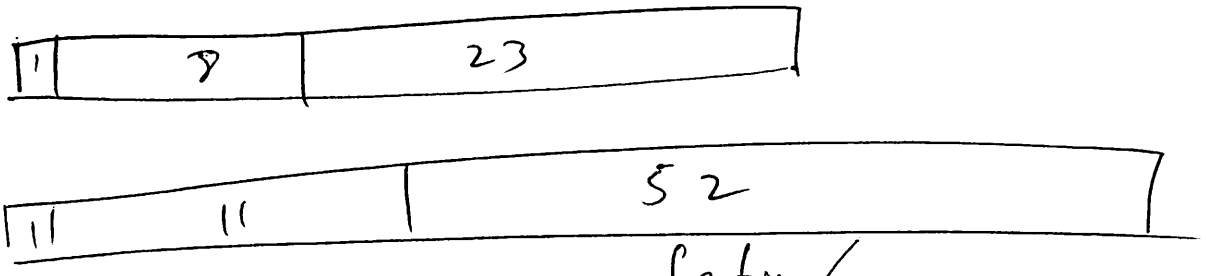
fraction / Mantissa

$3.14 = 3.14 \times 10^{0}$

$1941 = 1.941 \times 10^{3}$

Late 1970s

IEEE 754 standard



single prec.

double prec.

sign

1 = negative

0 = positive

exponent

↓

excess 127

excess 1023

$\sqrt{2} \times \sqrt{2} \stackrel{?}{=} 2$

a ← sqrt(2)

a * a == 2 FALSE

$\times 10^{-16}$

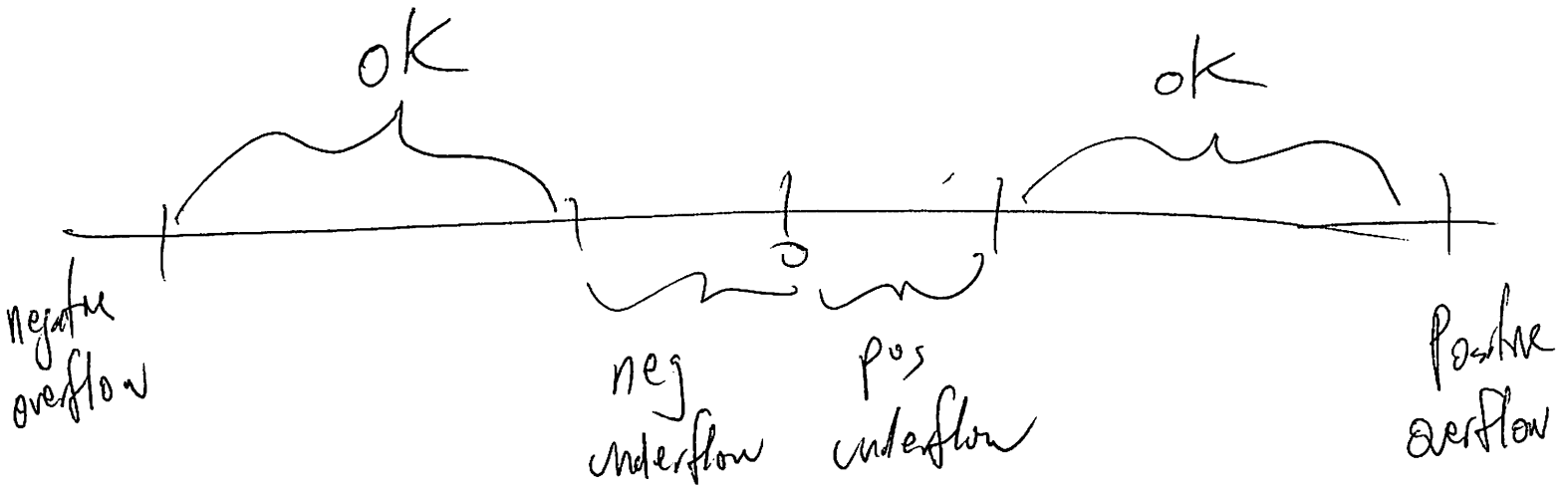
↓

rounding

Inf = exp ⇒ all 1s, fraction = 0

NaN = exp ⇒ all 1s, fraction = any non-zero pattern

	single	double
smallest	2^{-126}	2^{-1022}
largest	2^{128}	2^{1024}



Linear Models:

$$Y = f(X) = \sum_{j=1}^{p} \beta_j X_j = \beta' X$$

Additive Model

$$f(X) = f(X_1, \ldots X_p) = \sum_{j=1}^{p} s_j(X_j)$$

More General:

$$f(X) = \sum_{k=1}^{M} g_m(\underbrace{w_m' X}_{V_m = \text{derived feature}})$$

Projection Pursuit

Linear Regression: $M = 1$, $g_m(X) = \beta X$, w_m unknown

Additive model: $M = p$, g_m smooth, $w_m = (0, 0, \underset{k}{1}, 0 \ldots)$

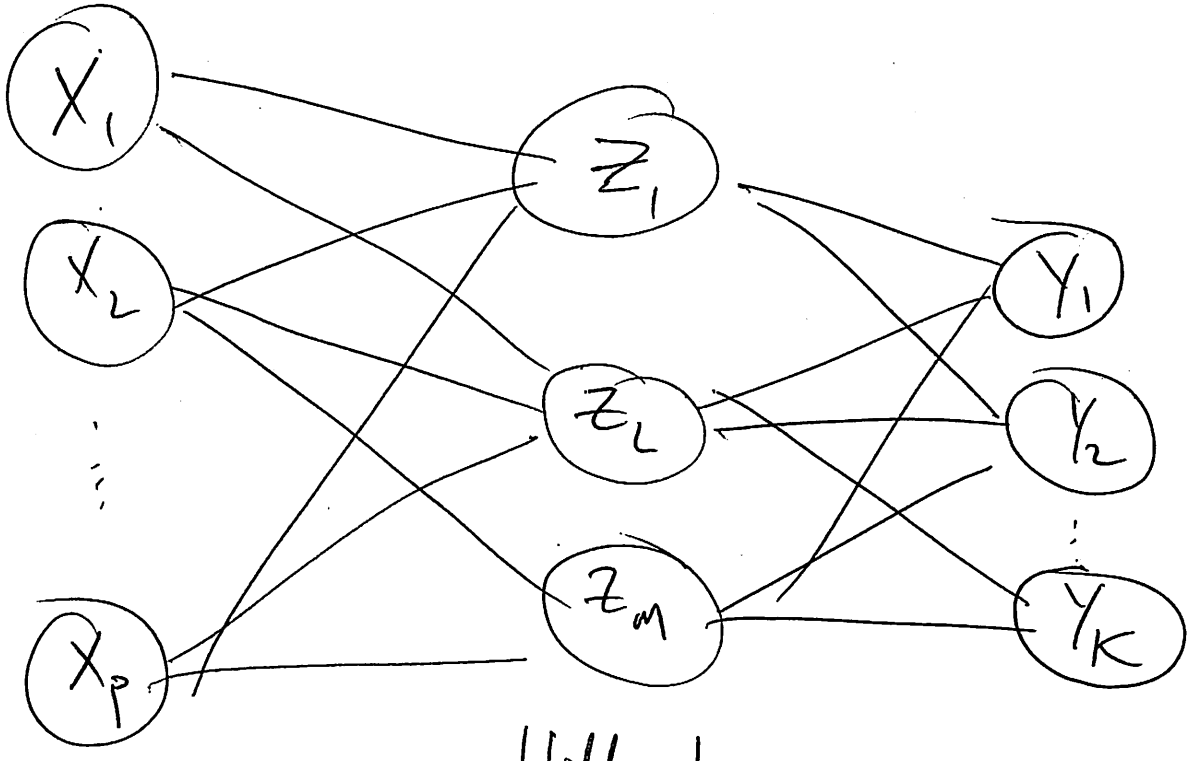
PPR = Additive model + non-coordinate axis directions

PCR: $M < p$, $g_m(X) = \beta_m X$, w_m = PC directions only dep on X

Neural Network: M = large, w_m unknown

w/ 1 hidden layer $\quad g_m(X) = \beta_m \sigma(\alpha_{0m} + \|\alpha_m\| X)$

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad \text{(logistic)}$$



Hidden layer

Fitting General projection pursuit: estimate w_m and g_m

~~$g_m(w_m'X) \approx g(w_0'X) + g'(w_0'X)(w - w_0)X$~~

$$f(X_i) = \sum_{m=1}^{M} g_m(w_m' X_i)$$

Minimize $$\sum_{i=1}^{N} \left(Y_i - f(X_i)\right)^2 = \sum_{i=1}^{N} \left(Y_i - \sum_{m=1}^{M} g_m(w_m' X_i)\right)$$

For fitting g_m, w_m

(1) Let $r_i = Y_i - \sum_{k \neq m} g_k(w_k' X_i)$

(2) Approximate $g_m(w' X_i) \doteq g(w_0' X_i) + g'(w_0' X_i)(w - w_0)' X_i$

(3) $$\sum_{i=1}^{N} \left(r_i - g_m(w' X_i)\right)^2 \doteq \sum_{i=1}^{N} \left(r_i - g(w_0' X_i) - g'(w_0' X_i)(w - w_0)' X_i\right)^2$$

$$= \sum_{i=1}^{N} g'(w_0' X_i)^2 \left[\left(w_0' X_i + \frac{r_i - g(w_0' X_i)}{g'(w_0' X_i)}\right) - w' X_i\right]^2$$

Linear regression

④ Given W_m, estimate g_m with smooth of r_i on $g_m(w_m' x_i)$ or simple linear regression.

⑤ Go back to ① until convergence

Strategy:

① Start with $M = 1$

② Build in new terms in forward stage-wise manner (Fix previous terms)

③ Use CV to estimate OOS error

④ Increase M until error stops decreasing

Neural Networks

$$Z_m = \sigma(\alpha_{0m} + \alpha_m' X) \quad m = 1, \dots M$$

Regression $f(X) = \beta_0 + \beta_1' Z$, $Z = (Z_1, \dots Z_M)$

Classification: $f(x) =$

Neural Networks $Y_1 \to Y_K$

$$Z_m = \sigma(\alpha_{0m} + \alpha_m' X), \quad m = 1, \to M$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

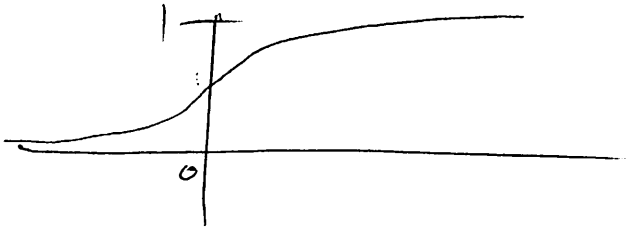
$$T_k = \beta_{0k} + \beta_k' Z, \quad Z = (Z_1, \to Z_M)$$

$$T = (T_1, \to T_K)$$

Regression: $f_k(x) = g_k(T_k) = \beta_{0k} + \beta_k' Z = \beta_{0k} + \sum_{m=1}^{M} \beta_{km} \sigma(\quad)$

Classification: $f_k(x) = \dfrac{e^{T_k}}{\sum_{l=1}^{K} e^{T_k}} =$ Softmax function

$$f_1(x) = \frac{e^{T_1}}{e^{T_1} + e^{T_0}} =$$



Error function:

$$R(\beta, \alpha) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2$$

$$R(\beta, \alpha) = \sum_{i=1}^{N} \sum_{k=1}^{K} (y_{ik} - f_k(x_i))^2$$

~~Recall~~ *

Fitting NN via back propagation:

Recall X_i = i^{th} input, $i = 1 \to N$

$$z_{im} = \sigma(\alpha_{0m} + \alpha_m' X_i) \qquad z = \begin{bmatrix} 1 & & M \\ \vdots & \cdots & z_{im} \\ N & & \end{bmatrix}$$

$z_i = (z_{i1}, \to z_{iM})$

The error function $R(\beta, \alpha)$, ~~this gradient~~

$$R(\beta,\alpha) = \sum_{i=1}^{N} R_i = \sum_{i=1}^{N}\left[\sum_{k=1}^{K} \underbrace{(y_{ik} - f_k(x_i))}_{\beta_m}^2\right]$$

$$f_k(x_i) = g_k(\beta_k' z_i)$$

$$\frac{\partial R_i}{\partial \beta_{km}} = \underbrace{-2(y_{ik} - f_k(x_i))\, g_k'(\beta_k' z_i)}_{\delta_{ki}} z_{im}$$

$k = 1 \to K$, $m = 1 \to M$

$$\frac{\partial R_i}{\partial \alpha_{ml}} = \sum_{k=1}^{K} \underbrace{-2(y_{ik} - f_k(x_i))\, g_k'(\beta_k' z_i)\, \beta_{km}\, \sigma'(\alpha_m' x_i)}\, x_{il}$$

$l = 1 \to p$

s_{mi}

Gradient descent update 13

$$\beta_{km}^{(t+1)} = \beta_{km}^{(t)} - \gamma_t \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}^{(t)}}$$

$$\alpha_{m\ell}^{(t+1)} = \alpha_{m\ell}^{(t)} - \gamma_t \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{m\ell}^{(t)}}$$

γ_t is learning rate (step length)

~~$\gamma_r = \frac{1}{r}$ for online learning~~

~~$\log \frac{P(Y=1)}{1-P}$~~

If we write

$$z_{im} = \sigma(\alpha_{0m} + \alpha_m' x_i)$$

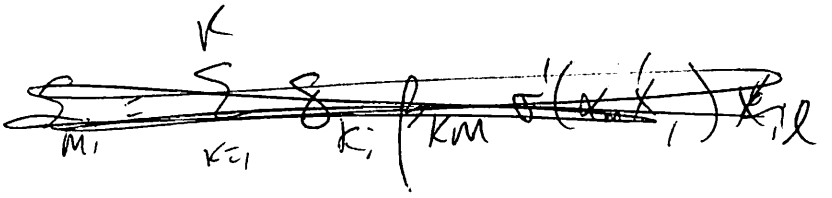
$$\frac{\partial R_i}{\partial \beta_{km}} = \delta_{ki} z_{im}$$

$$\frac{\partial R_i}{\partial \alpha_{m\ell}} = s_{mi} x_{i\ell}$$

$$s_{mi} = \sum_{k=1}^{K} \delta_{ki} \beta_{km} \sigma'(\alpha_m' x_i)$$

$$= \sigma'(\alpha_m' x_i) \sum_{k=1}^{K} \delta_{ki} \beta_{km}$$

Back propagation equations



~~Forward pass~~:

(1) Fix weights, Compute $\hat{f}_k(x_i)$

(2) Compute errors δ_{ki} and then s_{mi}

(3) Compute gradients $\frac{\partial R_i}{\partial \beta_{km}}$, $\frac{\partial R_i}{\partial \alpha_{m\ell}}$

(4) Update β_{km}, $\alpha_{m\ell}$

(5) Goto (1)

Learning rate γ_t can be fixed

Online learning

$$\beta_{km}^{(t+1)} = \beta_{km}^{(t)} - \gamma_t \frac{\partial R_i}{\partial \beta_{km}^{(t)}} \quad \text{for each } i = 1, \to N$$

$$\alpha_{m\ell}^{(t+1)} = \alpha_{m\ell}^{(t)} - \gamma_t \frac{\partial R_i}{\partial \alpha_{m\ell}^{(t)}}$$

For online learning γ_t must satisfy

$\gamma_t \to 0 \quad t \to \infty$

$\sum \gamma_t = \infty$

$\sum \gamma_t^2 < \infty$

$\gamma_t = 1/t$ works.

Regularization

① Stop early

② $J(\beta, \alpha) = \sum_{km} \beta_{km}^2 + \sum_{ml} \alpha_{ml}^2$

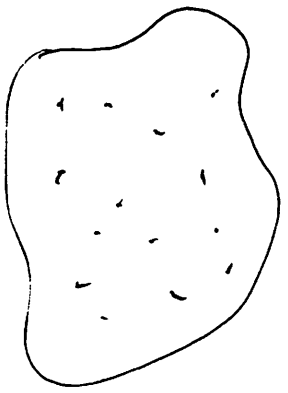
Let $R^*(\beta, \alpha) = R(\beta, \alpha) + \lambda J(\beta, \alpha)$ Ridge penalty

TOPIC: Biostat 778

DATE:

FILE UNDER:

PAGE:



Spatial Modelling

$Y(s) = \mu(s) + W(s) + \varepsilon(s)$

- $\mu(s)$: fixed effect
- $W(s)$: latent Gaussian spatial process
- $\varepsilon(s)$: measurement error? microscale noise?

$\mu(s) = x^T(s)\beta$

~~$W(s) \sim N(0,$~~ meaning?

$W = (W(s_1), W(s_2), \ldots, W(s_n)) \sim N(0, \sigma^2 H(\phi))$

- σ^2: sill
- ϕ: range

$\varepsilon(s) \sim N(0, \tau^2)$

↳ nugget effect

What is meaning of $W(s)$?

$H_{ij} = Cov(W(s_i), W(s_j))$

$= \rho(s_i - s_j ; \phi)$ anisotropic

$= \rho(\|s_i - s_j\| ; \phi)$ isotropic

e.g. $\rho(s_i - s_j, \phi) = \exp(-\phi \|s_i - s_j\|)$

Let $Y = (Y(s_1), Y(s_2), \ldots, Y(s_n))$

~~θ~~ $\theta = (\beta, \sigma^2, \phi, \tau^2)$

$Y|\theta, W \sim N(X\beta + W, \tau^2 I)$

$W|\sigma^2, \phi \sim N(0, \sigma^2 H(\phi))$

σ^2

ϕ $\sim$ Priors

β

τ^2

TOPIC:

DATE:

FILE UNDER:

PAGE:

Get posterios

$$p(\gamma, \theta, w) \propto p(y|\theta, w)\, p(w|\theta^{\beta,\tau^2,\sigma^2,\phi})\, p(\theta)$$

$$p(\theta) = p(\sigma^2)\,p(\tau^2)\,p(\phi)\,p(\beta)$$ (uninformative)

$$p(\beta|w) \propto p(y|\beta,\tau^2,w)\,p(\beta)$$

$$N(X\beta+w, \tau^2 I)\, N(0, D)$$

$$= N\left(S[X(X'X)^{-1}X'(y-w)],\ (I-S)D\right) \qquad S = D[\tau^2 I + D]^{-1}$$

$$p(\tau^2|w) \propto p(y|\beta,\tau^2,w)\,p(\tau^2)$$

$$= IG_1\left(a_2+1,\ b_2 + \tfrac{1}{2}(y-X\beta-w)'(y-X\beta-w)\right)$$

$$p(\sigma^2|w) \propto p(w|\sigma^2,\phi)\,p(\sigma^2)$$

$$IG_2\left(a_1+1,\ b_1 + \tfrac{1}{2}w'w\right)$$

$$p(\phi|w) \propto p(w|\sigma^2,\phi)\,p(\phi)$$

$$N(0, \sigma^2 H(\phi))\ \phi^{e-1} e^{-\phi/f}$$

$$|\sigma^2 H(\phi)|^{-1/2} \exp\left(-\frac{1}{2\sigma^2} w' H(\phi)^{-1} w\right)$$

$$H(\phi) = U(\phi) D(\phi) U(\phi)' \qquad D(\phi) = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \end{pmatrix}$$

$$|H(\phi)^{-1}| = |U(\phi) D(\phi)^{-1} U(\phi)'| = \prod \lambda_i^{-1}$$

$$|H(\phi)| = \prod \lambda_i$$

$$\left|\prod_{i=1}^{n} \sigma^2 \lambda_i\right|^{-1/2} \exp\left(-\frac{1}{2\sigma^2} w' U(\phi) D(\phi)^{-1} U(\phi)' w\right)$$

$$p(w|w) \propto p(y|\theta,w)\,p(w|\theta) \qquad \begin{pmatrix} 1/\lambda_1 & & \\ & \ddots & \\ & & 1/\lambda_n \end{pmatrix}$$

need Metropolis step →

$y - w$

z

$x(x'x)^{-1}x'(y-w)$

$\phi^{-(\alpha+1)} e^{-\beta/\phi}$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Bayesian way

$$p(\theta|y) \propto p(y|\theta, w)\, p(w|\theta)\, p(\theta)$$

$\sigma^2 \sim IG(a_1, b_1)$

$\tau^2 \sim IG(a_2, b_2)$ $\quad \phi \sim Gamma(a_3, b_3)$

$\beta \sim N(0, D)$ (vague)

~~$\phi \sim IG(c, f)$~~

} informative

$y|X \sim N(X, \tau^2)$

$X \sim N(0, \sigma^2$

$X|y \sim N(By, (1-B)\sigma^2)$

$\hat{\beta}|\beta \sim N(\beta, \Sigma^2)$

$\beta \sim N(\mu, \tau^2)$

$\beta|\hat{\beta} \sim N(c\hat{\beta}, (1-c)\tau^2)$

Compute MLEs

$$CDL = p(y|\theta, w)\, p(w|\theta)$$

$$N(X\beta + W, \tau^2 I)\, N(0, \sigma^2 H(\phi))$$

Need

$$p(w|\theta, y) \propto p(y|\theta, w)\, p(w|\theta)$$

$$CDLL = -\frac{1}{2}\log|\tau^2 I| - \frac{1}{2}(Y - X\beta - W)^T (Y - X\beta - W)/\tau^2$$

$$- \frac{1}{2}\log|\sigma^2 H(\phi)| - \frac{1}{2} W^T H(\phi)^{-1} W/\sigma^2$$

$1\times n \quad n\times n \quad n\times 1$

$$p(w|\theta, y) = N(BY, (I-B)\sigma^2 H(\phi))$$

$$B = \sigma^2 H(\phi)\left[\tau^2 I + \sigma^2 H(\phi)\right]^{-1}$$

Can iterate EM algorithm directly by filling in W and max Q. Estimate β, τ^2 okay but σ^2 and ϕ are harder (ϕ nonlinear)

TOPIC

DATE

FILE UNDER

PAGE

Prediction at new location.

Suppose we have $x(s_0)$ and we want $Y(s_0)$

↳ (Bayesian) Kriging, interpolation

We want predictive distribution;

$$p(y(s_0) \mid y, X, x(s_0))$$

↓ obs y ↘ obs X

$$= \int \underbrace{p(y_0 \mid y, X, \theta, x_0)}_{\text{Normal}} \underbrace{p(\theta \mid y, X)}_{\text{posterior}} d\theta$$

$$= \int \left[\int \underbrace{p(y_0 \mid y, X, \theta, x_0, w) p(w \mid \theta) dw}_{\text{If both normal, just use } p(y_0 \mid y, X, \theta, x_0)} \right] p(\theta \mid y, X) d\theta$$

$$= N(X\beta, \tau^2 I + \sigma^2 H(\phi))$$

But if $p(y \mid \theta, w)$ is not normal (maybe Poisson, Bernoulli) then we are stuck, need to sample.

$Y \mid \theta \sim N(\cdot, \cdot)$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ Y_0 \end{pmatrix} \sim N\left(\begin{pmatrix} x_1 - \\ x_2 - \\ \\ x_n - \\ x_0 - \end{pmatrix} \beta \right.$$

$p(w_0, w \mid \theta)$

$= p(w_0 \mid w, \theta) p(w \mid \theta)$

TOPIC:

DATE:

FILE UNDER:

PAGE:

y|

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ w_0 \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma^2 & \cdots & \text{cov} \\ & \ddots & \\ \text{cov} & \cdots & \sigma^2 \end{pmatrix}\right)$$

$$w_0 \mid w_1, \ldots, w_n \sim N(0 + B w, (1 - B))$$

$$\begin{pmatrix} w_0 \\ \hline w_1 \\ \vdots \\ w_n \end{pmatrix} \sim N\left(0, \left(\begin{array}{c|ccc} \sigma^2 & \text{cov} & \cdots & R \\ \hline \text{cov} & \sigma^2 & & \text{cov} \\ \vdots & & \ddots & \\ & \text{cov} & & \sigma^2 \end{array}\right)\right)$$

$$w_0 \mid w_1, \ldots, w_n \sim N\left(\underset{1\times n}{R^T} \underset{n\times n}{H(\phi)^{-1}} \underset{n\times 1}{W}, \sigma^2 - R^T H(\phi)^{-1} R\right)$$

Forward Stagewise Additive Modeling

For the regression problem

$$Y_i = f(X_i) + \varepsilon_i$$

Y_i = continuous or $\{-1, 1\}$

X_i = p-dim vector

We want to minimize

$$\sum_{i=1}^{N} L(y_i, f(x_i))$$

over all f in some class. For modeling f, often useful to use an additive model

$$f(x) = \sum_{m=1}^{M} \beta_m b(x | \gamma_m)$$

Regression: $b(x|\gamma_m) = x_m$

Neural Net: $b(x|\gamma_m) = \sigma(\gamma_{0m} + \gamma_{1m}' x)$

Proj Pursuit: $b(x|\gamma_m) = g_m(\gamma' x)$ (g_m smooth)

Trees: ~~2 parameters~~ $b(x|\gamma_m) = \mathbb{I}\{x \in R_m(\gamma_m)\}$

Nonlinear:

$$\sum_{i=1}^{N} L\left(y_i, \sum_{m=1}^{M} \beta_m b(x_i|\gamma_m)\right) \qquad (*)$$

Loss function	$L(y, f(x))$
Least squares:	$(y - f(x))^2$
Robust regression:	$\lvert y - f(x)\rvert$
Deviance:	$\log(1 + \exp(-y f(x)))$
Exponential:	$\exp(-y f(x))$
Support vector:	$(1 - y f(x))_+$

$\Rightarrow$ Solving $(*)$ Globally can be hard

$\Rightarrow$ What if Solving $\sum_{i=1}^{N} L(y_i, \beta\, b(x_i|\gamma_m))$

is easy?

Forward Stagewise Additive Modeling

1. Initialize $f_0(x) = 0$
2. For $m = 1, \ldots, M$
   a. $(\beta_m, \gamma_m) = \underset{\beta, \gamma}{\arg\min} \sum_{i=1}^{N} L(y_i, f_{m-1}(x_i) + \beta\, b(x_i \mid \gamma))$
   b. Set $f_m(x) = f_{m-1}(x) + \beta_m b(x \mid \gamma_m)$

$\Rightarrow$ An approximation (greedy) to the global problem.

$\Rightarrow$ Not backfitting b/c previously added terms not modified

Linear regression w/ least squares

$$L(y_i, f_{m-1}(x_i) + \beta b(x_i|\gamma))$$

$$= (\underbrace{y_i - f_{m-1}(x_i)}_{r_i} - \beta\, b(x_i|\gamma))^2$$

$$(r_i - \beta\, b(x_i|\gamma))^2$$

~~$y_i - \beta_{m-1} x_{i,m-1}$~~

$$(\underbrace{y_i - \beta'_{m-1} x_{i,m-1}}_{r_i} - \beta x_{i,m})^2$$

$$(r_i - \beta x_{i,m})^2$$

$$\hat{\beta} = \frac{\mathrm{Cov}(x_m, r)}{\mathrm{Var}(x_m)}$$

Boosting (AdaBoost)

$$L(y, f(x)) = \exp(-\underbrace{y f(x)}_{\text{margin}})$$

$$f(x) = \sum_{m=1}^{M} \alpha_m G_m(x)$$

individual trees / classifiers

$\Rightarrow$ We must solve at each $m = 1, \dots, M$

$$(\beta_m, G_m) = \underset{\beta, G}{\operatorname{argmin}} \sum_{i=1}^{N} \exp\left[-y_i \left(f_{m-1}(x_i) + \beta G(x_i)\right)\right]$$

$$= \underset{\beta, G}{\operatorname{argmin}} \sum_{i=1}^{N} w_i^{(m)} \exp\left(-y \beta G(x_i)\right)$$

G_m minimizes $\sum_{i=1}^{N} w_i^{(m)} \mathbb{1}\{y_i \neq G(x_i)\}$ for $\beta > 0$

$$\beta_m = \frac{1}{2} \log \frac{1 - err_m}{err_m}, \quad err_m = \frac{\sum w_i^{(m)} \mathbb{1}\{y_i \neq G_m(x_i)\}}{\sum w_i^{(m)}}$$

$$f_m(x) = f_{m-1}(x) + \beta_m G_m(x)$$

Gradient Boosting

① $f_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_i, \gamma)$

② For $m = 1, \ldots, M$

(a) Compute $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f = f_{m-1}}$

(b) regress r_{im} on predictors to produce J_m regions at leaf nodes

$\Rightarrow \hat{f}_m(x) \quad \forall x$

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

(d) $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbb{1}\{x \in R_{jm}\}$

③ $\hat{f}(x) = f_M(x)$

Bagging

(1) For $b = 1, \ldots, B$

(a) Draw Z^* w/ replace

(b) Fit tree to Z^* to get T_b

(2) Output $T_b, b = 1, \ldots, B$

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

not iid $\Rightarrow$ id

$$\mathrm{Var}\left(\frac{1}{B}(X_1 + \cdots + X_B)\right) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

If iid $\mathrm{Var}\left(\frac{1}{B}(X_1 + \cdots + X_B)\right) = \frac{1}{B}\sigma^2$

Random Forest

Given (Y_i, X_i) $i = 1 \to N$

(1) For $b = 1 \to B$

(a) Draw bootstrap sample Z^* of size N from training data

(b) RF tree T_b

(1) Select m vars from p vars

(2) Pick best split variable

(3) Split node into 2 nodes

(4) etc.

(2) Output T_b for $1 \to B$

$$\hat{f}(x) = \frac{1}{B} \sum T_b(x)$$ Regression

$$\hat{f}(x) = \text{Majority vote} \left\{ T_b(x) \right\}_1^B$$ Sample w/o replace!

Use OOB samples to estimate error rate