Outline of the lecture

This is about Monte Carlo methods.

- We will revise importance sampling.
- Revise how Google works (Markov chains).
- Introduce Markov chain Monte Carlo (MCMC)

Bayesian logistic regression

The logistic regression model specifies the probability of a binary output $y_i \in \{0,1\}$ given the input $\mathbf{x}_i$ as follows:

$$\pi(\theta) = -\log P(\theta|X,y) = \text{const} - \log P(y|x,\theta) - \log P(\theta)$$

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) &= \prod_{i=1}^{n} \text{Ber}(y_i|\text{sigm}(\mathbf{x}_i\boldsymbol{\theta})) \quad = \text{Const} - \sum_{i=1}^{n} y_i \log \pi_i + (1-y_i)\log(1-\pi_i) \\ &= \prod_{i=1}^{n} \Bigg[\underbrace{\frac{1}{1+e^{-\mathbf{x}_i\boldsymbol{\theta}}}}_{\pi_i}\Bigg]^{y_i} \left[1 - \frac{1}{1+e^{-\mathbf{x}_i\boldsymbol{\theta}}}\right]^{1-y_i} \quad - \frac{1}{2\sigma^2}\|\theta\|_2^2 \end{aligned}$$

We also assume a Gaussian prior

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\theta - \overset{0}{\cancel{\mu}})'(\theta - \overset{\theta}{\cancel{\mu}})\right)$$

Posterior: $\quad P(\theta|X,y) = \frac{1}{Z} P(y|X,\theta)\, P(\theta)$

$\quad Z = \int P(y|X,\theta)\, P(\theta)\, d\theta$ is unknown / hard

Importance sampling

$$Z = \int P(Y|\theta)\,P(\theta)\,d\theta$$

$$Z = \int \boxed{\frac{P(Y|\theta)\,P(\theta)}{q(\theta)}}\, q(\theta)\, d\theta \qquad q(\theta) = \mathcal{N}(0, 1000)$$

$$Z = \int w(\theta)\, q(\theta)\, d\theta \qquad \theta^{(i)} \sim q(\theta), \quad i = 1:N$$

SLLN

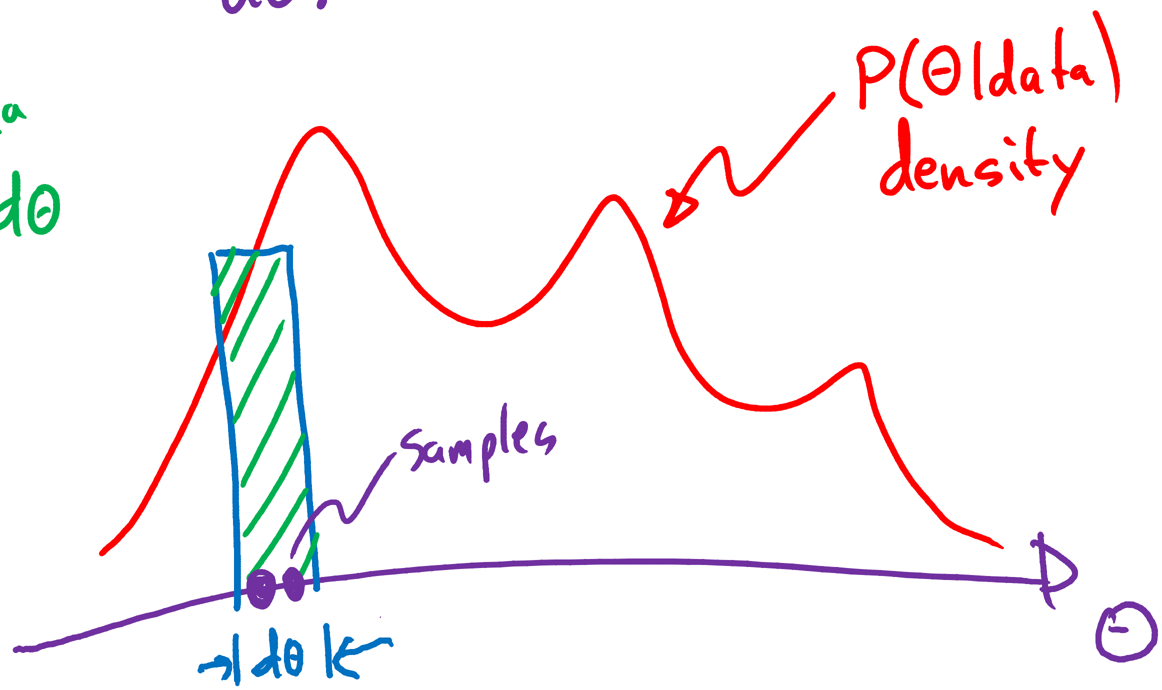
$$Z \approx \frac{1}{N} \sum_{i=1}^{N} w(\theta^{(i)})$$

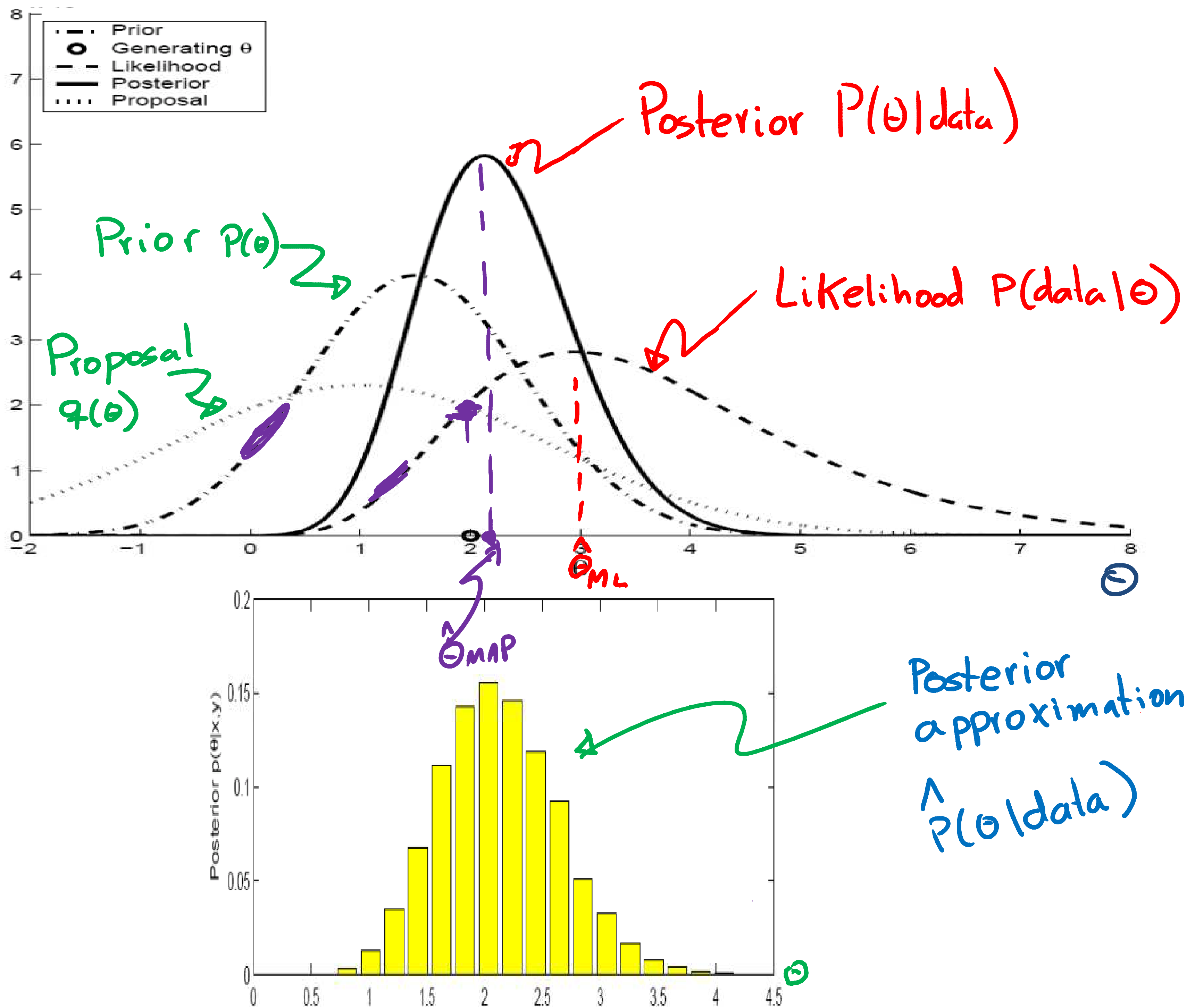
Importance sampling

$$\boxed{p(d\theta \mid \text{data}) = \frac{1}{N} \sum_{i=1}^{N} w(\theta^{(i)})\, \delta_{\theta^{(i)}}(d\theta)}$$

$\delta_{\theta^{(i)}}(d\theta) =$ Number of samples $\theta^{(i)}$ in the interval $d\theta$.

$$P(d\theta \mid \text{data}) = P(\theta \mid \text{data})\, d\theta$$





Prior
Generating θ
Likelihood
Posterior
Proposal
Posterior P(θ|data)
Prior P(θ)
Likelihood P(data|θ)
Proposal q(θ)
θ_ML
θ_MAP
θ
Posterior p(θ|x,y)
Posterior approximation
p̂(θ|data)

Importance sampling

$$P(y_{t+1} \mid x_{t+1}, y_{1:t}, x_{1:t}) = \int P(y_{t+1} \mid x_{t+1}, \theta)\, P(d\theta \mid x_{1:t}, y_{1:t})$$

$$= \int P(y_{t+1} \mid x_{t+1}, \theta)\, P(\theta \mid x_{1:t}, y_{1:t})\, d\theta$$

$$\approx \int P(y_{t+1} \mid x_{t+1}, \theta)\, \frac{1}{N} \sum_{i=1}^{N} \omega(\theta^{(i)})\, \delta_{\theta^{(i)}}(d\theta)$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \int P(y_{t+1} \mid x_{t+1}, \theta)\, \omega(\theta^{(i)})\, \delta_{\theta^{(i)}}(\underline{d\theta})$$

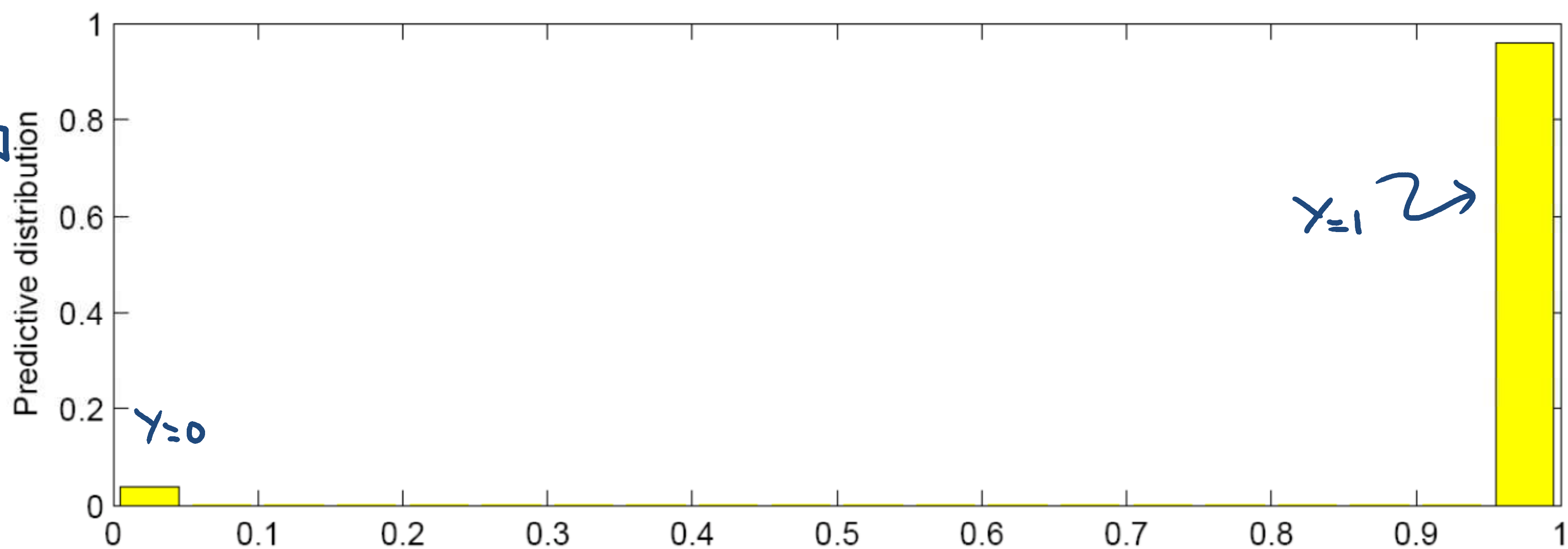
$$\approx \frac{1}{N} \sum_{i=1}^{N} \underbrace{P(y_{t+1} \mid x_{t+1}, \theta^{(i)})}_{\text{Likelihood}}\, \omega(\theta^{(i)})$$

Example: Logistic Regression

$$p(y_{T+1}|x_{1:T+1}) = \int_{\Theta} p(y_{T+1}|x_{T+1},\theta)p(\theta|x_{1:T},y_{1:T})d\theta$$

$$p(y_{T+1}|x_{1:T+1}) = \frac{1}{N}\sum_{i=1}^{N} p(y_{T+1}|x_{T+1},\theta^{(i)})$$

$\Theta^{(i)} \sim P(\Theta|X_{1:T}, Y_{1:T})$



Un-normalized importance sampling

D = data

$$P(\theta|D) = \frac{1}{Z} P(D|\theta)P(\theta) = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)\,d\theta}$$

$$P(y_{t+1}|x_{t+1}, D)$$

$$= P(y_{t+1}|x_{1:t+1}, y_{1:t}) = \frac{1}{Z}\int P(y_{t+1}|x_{t+1},\theta)\, P(D|\theta)P(\theta)\,d\theta$$

$Q(\mu, \sigma^2 I)$

$$= \frac{\int P(y_{t+1}|x_{t+1},\theta)\, P(D|\theta)P(\theta)\frac{q(\theta)}{q(\theta)}\,d\theta}{\int P[D|\theta]P(\theta)\frac{q(\theta)}{q(\theta)}\,d\theta}$$

$$= \frac{\int P(y_{t+1}|x_{t+1},\theta)\, w(\theta)\, q(\theta)\,d\theta}{\int w(\theta)\, q(\theta)\,d\theta} = \frac{\cancel{\frac{1}{N}}\sum_{i=1}^{N} w(\theta^{(i)})P(y_{t+1}|x_{t+1},\theta^{(i)})}{\cancel{\frac{1}{N}}\sum_{j=1}^{N} w(\theta^{(j)})}$$

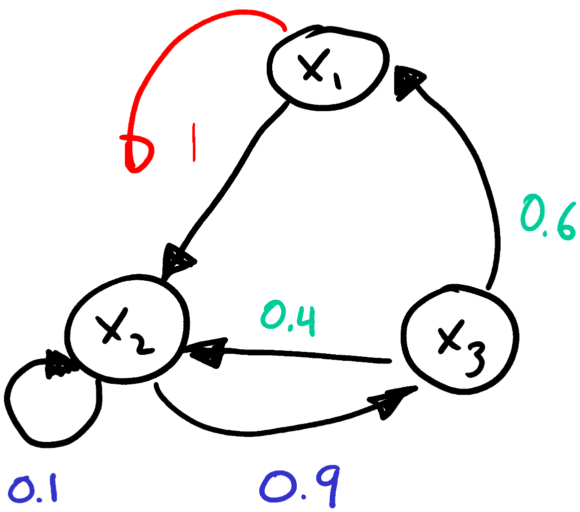
$$\tilde{w}^i = \frac{w^i}{\sum_j w^j}$$

$$= \sum_{i=1}^{N} \tilde{w}(\theta^{(i)})\, P(y_{t+1}|x_{t+1},\theta^{(i)})$$

Markov Chain Monte Carlo

For simplicity, let's consider only 3 states:

$x_t \in \mathcal{X} = \{x_1, x_2, x_3\}$



$$T = P(x_t \mid x_{t-1}) = \begin{array}{c|ccc} & x_1 & x_2 & x_3 \\ \hline x_1 & 0 & 1 & 0 \\ x_2 & 0 & 0.1 & 0.9 \\ x_3 & 0.6 & 0.4 & 0 \end{array} \quad 0.9 \to 1$$

Think of this as a webgraph. Our goal is to crawl it to find the "relevance" of each node.

Markov Chain Monte Carlo

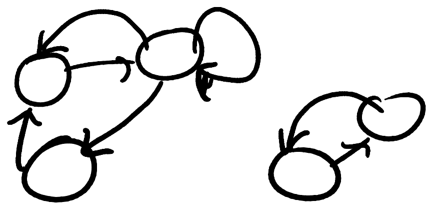
T is a stochastic matrix. As long as the graph (state space) is aperiodic and irreducible, we have that for any initial vector of probabilities ν:

$$\nu' T^{(t)} \longrightarrow \pi' \quad \text{as} \quad t \to \infty$$

where π is the invariant or Stationary distribution of the chain. It is unique.

Markov Chain Monte Carlo

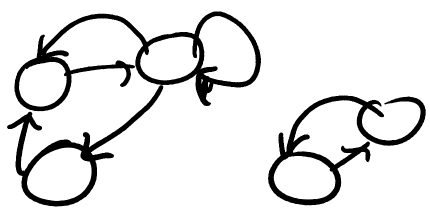
Need for irreducibility:



One cluster might never be visited!

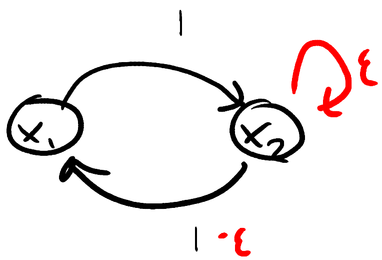
Markov Chain Monte Carlo

Need for irreducibility:



One cluster might never be visited!

Need for aperiodicity:



$$T = \begin{bmatrix} \overset{\varepsilon}{0} & 1-\varepsilon \\ 1-\varepsilon & 0^{\varepsilon} \end{bmatrix}$$

Let $\pi = \begin{bmatrix} 1/3 & 2/3 \end{bmatrix}$

$$\pi' T = \begin{bmatrix} 2/3 & 1/3 \end{bmatrix}$$

$$\pi' T^2 = \begin{bmatrix} 1/3 & 2/3 \end{bmatrix}$$

$$\vdots$$

Oscillation!

Markov Chain Monte Carlo

In the limit:

$$\pi' T = \pi'$$

π is the left eigenvector of T with corresponding eigenvalue 1.

Markov Chain Monte Carlo

In the limit:

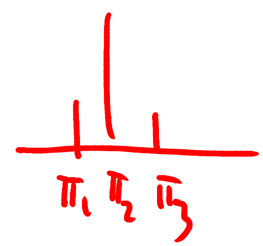
$\sum \pi_i = 1$

$\pi' = [\pi_1 \;\; \pi_2 \;\; \pi_3]$

$$\underline{\pi' T = 1 \pi'}$$

π is the left eigenvector of T with corresponding eigenvalue 1. Componentwise, we have:

$$\sum_{i=1}^{3} \pi_i T_{ij} = \pi_j$$



Markov Chain Monte Carlo

In the limit:

$$\underline{\pi}' \underline{T} = \underline{\pi}'$$

π is the left eigenvector of T with corresponding eigenvalue 1. Componentwise, we have:

$$\sum_{i=1}^{3} \pi_i T_{ij} = \pi_j$$

As the state space grows:

$$\boxed{\int \pi(x) \underbrace{p(y|x)}_{\text{Markov Chain Kernel}} dx = \pi(y)}$$

Markov Chain Monte Carlo

Detailed Balance:

If $\int_{x_t} \underbrace{\pi(x_t)\, P(x_{t+1} \mid x_t)} = \int_{x_t} \pi(x_{t+1})\, P(\cancel{x_t} \mid x_{t+1})$

$= \pi(x_{t+1}) \cancel{\int_{x_t} P(x_t \mid x_{t+1})}^{\,1}$

Integrating over x_t yields

$$\boxed{\int_{x_t} \pi(x_t)\, P(x_{t+1} \mid x_t) = 1\, \pi(x_{t+1})}$$

Which is the ergodic behaviour we want.
Now we have a sufficient condition for designing $P(x_{t+1} \mid x_t)$ so as to get samples from π

Metropolis-Hastings for logistic regression

$$P(\theta \mid x_{1:t}, y_{1:t}) = \frac{1}{Z} \prod_{i=1}^{t} \left[\pi_i^{y_i} (1-\pi_i)^{1-y_i} \right] e^{-\frac{1}{2\delta^2} \theta^T \theta}$$

want $\theta^{(i)} \sim P(\theta \mid y_{1:t}, x_{1:t})$

$$\boxed{\begin{array}{l} \pi(x) \\ x^{i} \sim \pi(x) \end{array}}$$

$$\tilde{\pi}(x) = \frac{1}{Z} \pi(x)$$

$$Z = \int \pi(x)\, dx$$

MCMC: Metropolis-Hastings

$\pi(x) = \frac{P(x)}{Z}$

e.g.

$x^{*} = x^{(i)} + N(0, \sigma^2)$

- Initialise $x^{(0)}$. $x^{(i)}$
- For $i = 0$ to $N - 1$
  - Sample $\underline{u} \sim U_{[0,1]}$.
  - Sample $x^{\star} \sim q(x^{\star}|x^{(i)})$.
  - If $u < \underbrace{A(x^{(i)}, x^{\star})}_{0.8} = \min\left\{1, \frac{\cancel{Z}\, p(x^{\star})q(x^{(i)}|x^{\star})}{\cancel{Z}\, p(x^{(i)})q(x^{\star}|x^{(i)})}\right\}$

    posterior

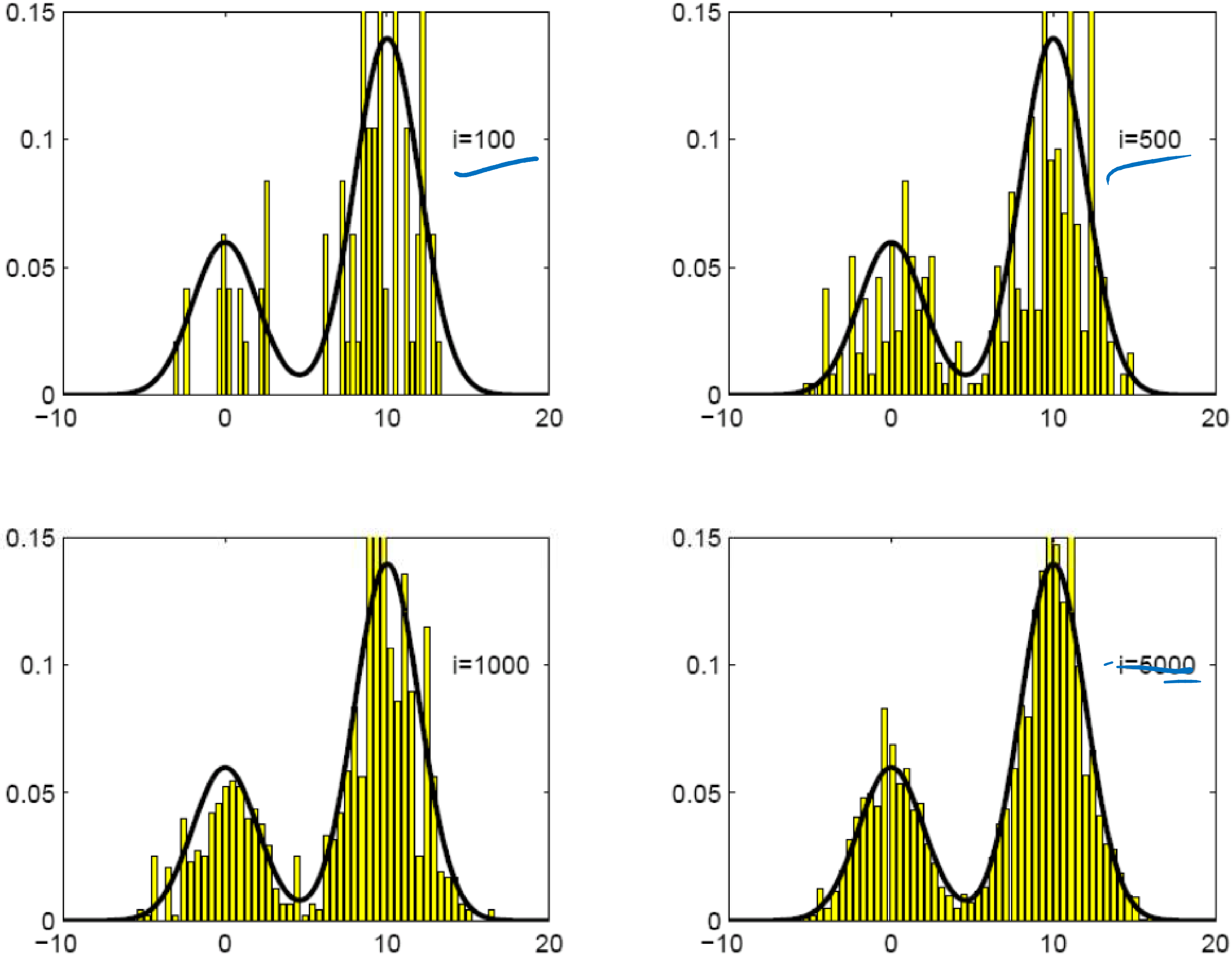
    $$\underline{x^{(i+1)} = x^{\star}}$$

    else

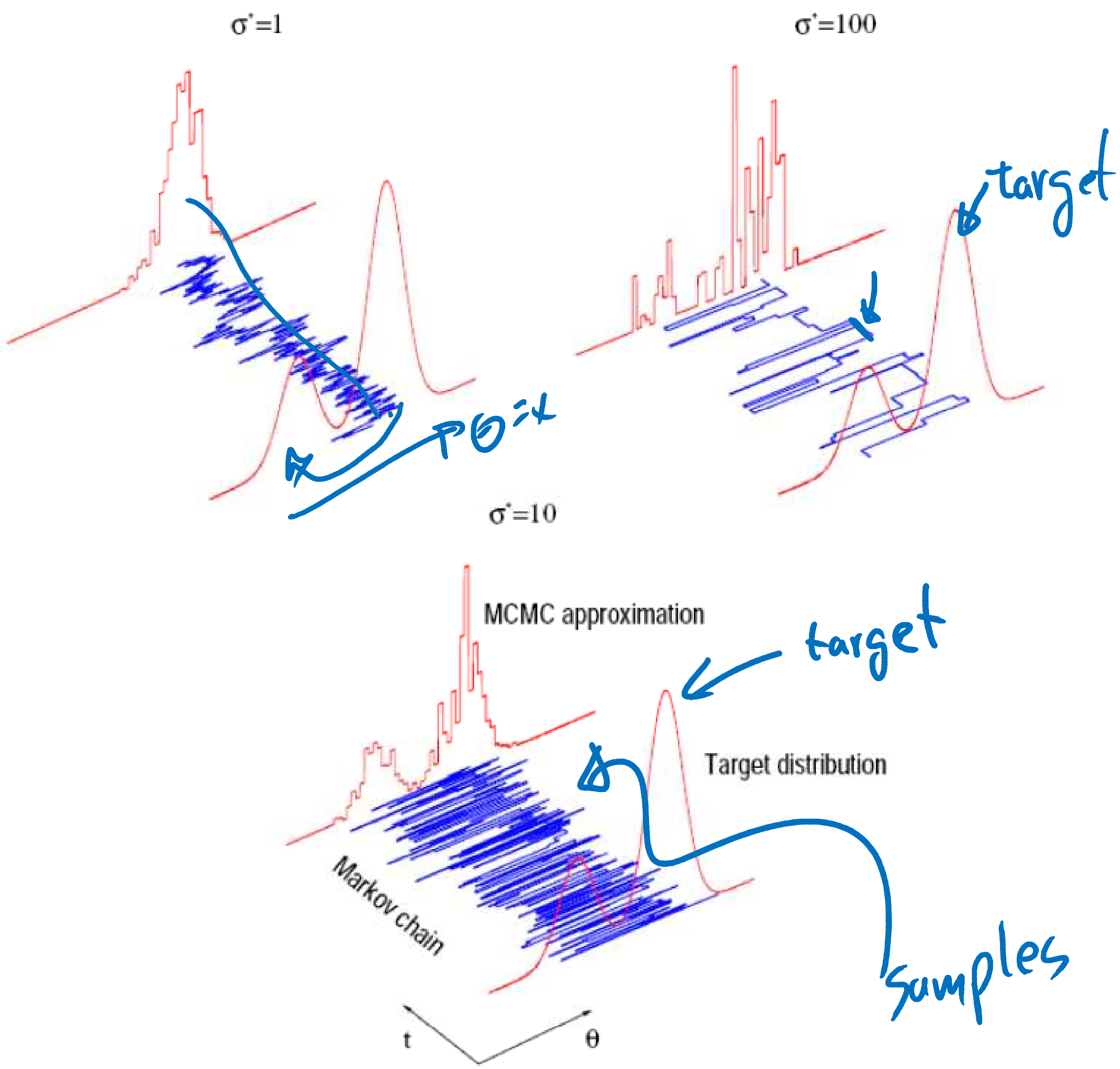
    $$\underline{x^{(i+1)} = x^{(i)}}$$

$x^{(i+1)}$

MCMC: Metropolis-Hastings



MCMC: Choosing the Right Proposal



MCMC: Theory

Kernel:

$$T = K(x, B) = \begin{cases} \underline{q(B|x)}\ \underline{A(x,B)} & \underline{x \notin B} \\ 1 - \int_{x' \in \{\mathcal{X} \setminus B\}} q(x'|x)\, A(x, x') & \underline{x \in B} \end{cases}$$

Prob of going from x to interval B.

all space $\mathcal{X}$ minus B

$$\therefore\ K(x,B) = q(B|x)\,A(x,B) + \mathbb{I}_{x\in B}\left\{ 1 - q(B|x)\,A(x,B) - \int_{x' \in \{\mathcal{X}\setminus B\}} q(x'|x)\,A(x,x') \right\}$$

$$K(x,B) = q(B|x)\,A(x,B) + \mathbb{I}_{x \in B}\left\{ 1 - \int_{x' \in \mathcal{X}} q(x'|x)\,A(x,x') \right\}$$

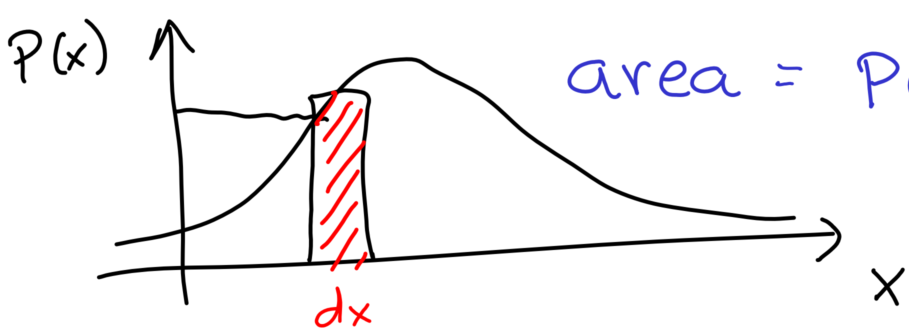
MCMC: Theory

Detailed balance:

$$\underline{\underline{\pi(A)}}\,\underline{\underline{K(A,B)}} = \pi(B)\,K(B,A)$$

$$\int_{x\in A} \pi(dx)\,K(x,B) = \int_{y\in B} \pi(dy)\,K(y,A)$$

Note: $\int f(x)\,p(x)\,dx \equiv \int f(x)\,p(dx)$



area $= P(dx) = P(x)\,dx$

Variations of Metropolis-Hastings

$$\min\left\{1, \frac{P(x^*)}{P(x^{(i)})} \frac{\cancel{q(x^{(i)}|x^*)}}{\cancel{q(x^*|x^{(i)})}}\right\}$$

$$x^* = x^{(i)} + N(0, \sigma^2)$$

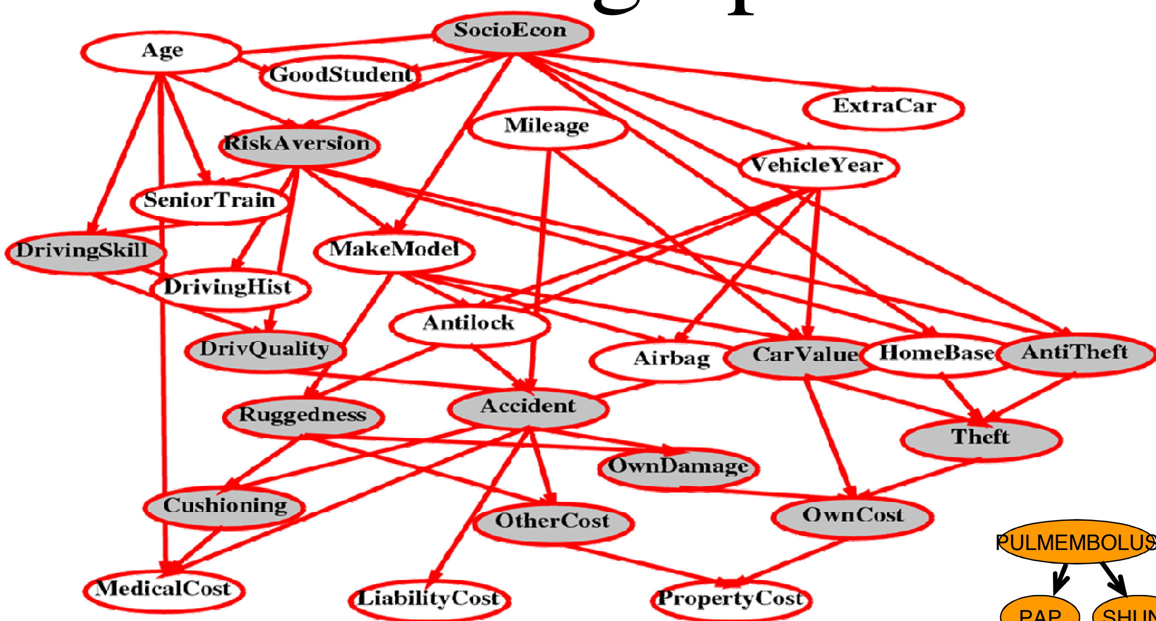
$$q(x^{(i)}|x^*) \propto e^{-\|x^* - x^{(i)}\|^2/\sigma^2}$$

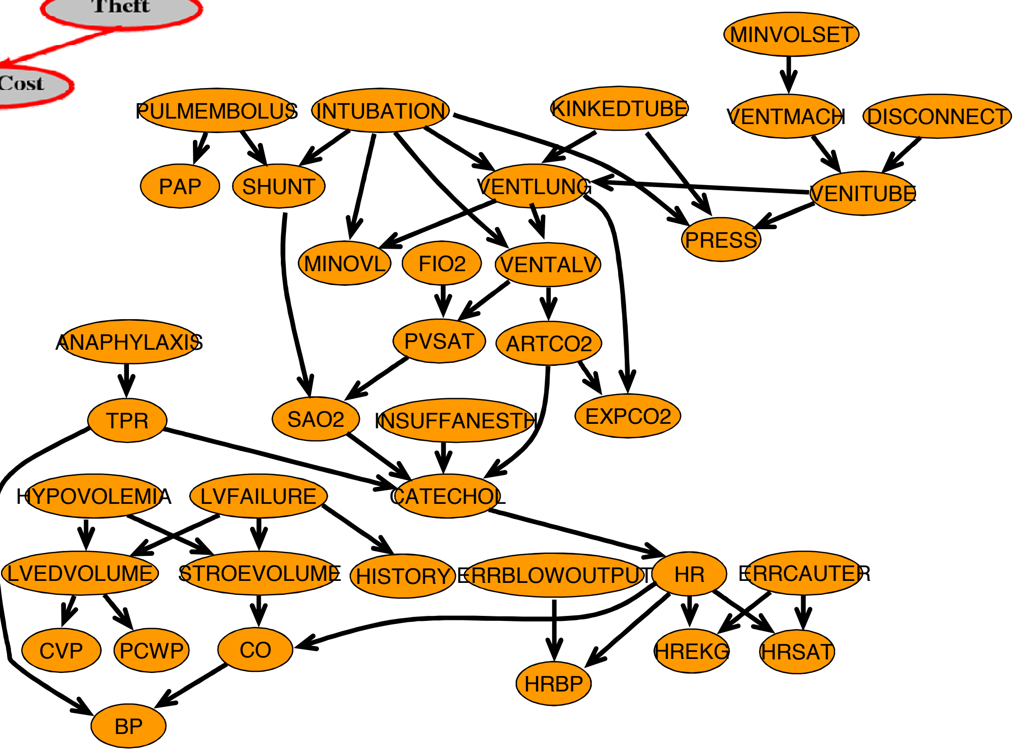
$$q(x^*|x^{(i)}) \propto e^{-\|x^{(i)} - x^*\|^2/\sigma^2}$$

$$\min\left\{1, \frac{P^{1/T}(x^*)}{P^{1/T}(x^{(i)})}\right\}$$

If anneal with T, concentrate on peaks of $P(x)$

Extending MH to directed probabilistic graphical models



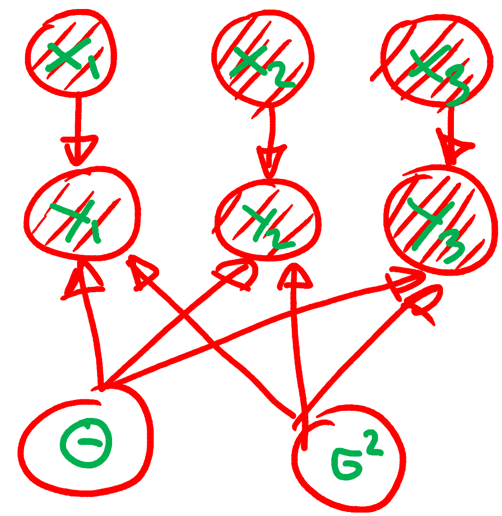


Bayesian graphical models and Gibbs

$$P(\sigma^2) = IG(a, b)$$

$$P(\theta) = \mathcal{N}(0, \delta^2 I) \quad \equiv$$

$$P(y_i \mid x_i, \theta) = \mathcal{N}(x\theta, \sigma^2)$$



GIBBS:

FOR $i = 1$ to Nsamples

$$\theta^{(i+1)} \sim P(\theta \mid \sigma^{2(i)}, x, y)$$

$$\sigma^{2(i+1)} \sim P(\sigma^2 \mid \theta^{(i)}, x, y)$$

END

Gibbs Sampling

Choose the following proposal:

$$q(x^\star|x^{(i)}) = \begin{cases} p(x_j^\star|x_{-j}^{(i)}) & \text{If } x_{-j}^\star = x_{-j}^{(i)} \\ 0 & \text{Otherwise.} \end{cases}$$

where $x_{-j} = \{x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n\}$.

Then the acceptance is:

$$A(x^{(i)}, x^\star) \quad = \quad \min\left\{1, \frac{p(x^\star)q(x^{(i)}|x^\star)}{p(x^{(i)})q(x^\star|x^{(i)})}\right\} = 1.$$

Gibbs Sampling

- Initialise $x_{1:n}^{(0)}$.
- For $i = 0$ to $N - 1$
  - Sample $x_1^{(i+1)} \sim p(x_1 | x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
  - Sample $x_2^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})$.

$$\vdots$$

  - Sample
    $x_j^{(i+1)} \sim p(x_j | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.

$$\vdots$$

  - Sample $x_n^{(i+1)} \sim p(x_n | x_1^{(i+1)}, x_2^{(i+1)}, \ldots x_{n-1}^{(i+1)})$.

Gibbs Sampling For Graphical models

A large-dimensional joint distribution is factored into a directed graph that encodes the conditional independencies in the model. In particular, if $x_{pa(j)}$ denotes the parent nodes of node x_j, we have

$$p(x) = \prod_j p(x_j | x_{pa(j)}).$$

It follows that the full conditionals simplify as follows

$$p(x_j | x_{-j}) = p(x_j | x_{pa(j)}) \prod_{k \in ch(j)} p(x_k | x_{pa(k)})$$

where $ch(j)$ denotes the children nodes of x_j.

Auxiliary Variable Samplers

- It is often easier to sample from an augmented distribution $p(x, u)$, where u is an auxiliary variable, than from $p(x)$.
- It is possible to obtain marginal samples $x^{(i)}$ by sampling $(x^{(i)}, u^{(i)})$ according to $p(x, u)$ and, then, ignoring the samples $u^{(i)}$.
- This very useful idea was proposed in the physics literature (Swendsen and Wang, 1987).

Hybrid (Hamiltonian) Monte Carlo

- The idea is to exploit gradient information.
- Define the extended target distribution:

$$p(x, u) = p(x)N(u; 0, I_{n_x}).$$

- Introduce the gradient vector: $\Delta(x) = \partial \log p(x)/\partial x$
- Introduce the parameters ρ and L.
- Next we “leapfrog”.

Hybrid Monte Carlo

- Sample $v \sim U_{[0,1]}$ and $u^\star \sim N(0, I_{n_x})$.
- Let $x_0 = x^{(i)}$ and $u_0 = u^\star + \rho\Delta(x_0)/2$.
- For $l = 1, \ldots, L$, take steps

$$x_l = x_{l-1} + \rho u_{l-1}$$

$$u_l = u_{l-1} + \rho_l \Delta(x_l)$$

  where $\rho_l = \rho$ for $l < L$ and $\rho_L = \rho/2$.

- If $v < A = \min\left\{1, \frac{p(x_L)}{p(x^{(i)})} \exp\left(-\frac{1}{2}(u_L^{\mathrm{T}} u_L - u^{\star\mathrm{T}} u^\star)\right)\right\}$

$$(x^{(i+1)}, u^{(i+1)}) = (x_L, u_L)$$

  else $$(x^{(i+1)}, u^{(i+1)}) = (x^{(i)}, u^\star)$$