Background: Two key problems: Estimation of Integrals, Sampling

Reason: Want to learn $I = \int_{\mathbb{X}} f(x)\,dx$, often rely on samples to approximate

Basic MC gets $\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} \varphi(\theta_i).$, when $I = \int_{\Theta} \varphi(\theta)\,\pi(\theta)\,d\theta,$ pi pdf and psi 'test' function

Proposition 2.1 (LLN): $If\ \mathbb{E}(|\varphi(X)|) < \infty\ then\ \hat{I}_n\ is\ a\ strongly\ consistent\ estimator$ (of I)

Proposition 2.2 (CLT): If $\sigma^2 = \mathbb{V}(\varphi(X)) = \int [\varphi(x) - I]^2\,\pi(x)\,dx < \infty$ then

$\mathbb{E}\left(\left(\hat{I}_n - I\right)^2\right) = \mathbb{V}\left(\hat{I}_n\right) = \frac{\sigma^2}{n}$ and $\frac{\sqrt{n}}{\sigma}\left(\hat{I}_n - I\right) \xrightarrow{D} \mathcal{N}(0,1)$

Proposition 2.3: If conditions for 2.2 is satisfied then $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\varphi(X_i) - \hat{I}_n\right)^2$ is unbiased for sigma^2

Proof: $let\ Y_i = \varphi(X_i)$ write E[Sn^2] in terms of this, get E[Ybar^2) and sub in

---

Error Estimates: Chebyshev's Inequality $\mathbb{P}\left(\left|\hat{I}_n - I\right| > c\frac{\sigma}{\sqrt{n}}\right) \le \frac{\mathbb{V}\left(\hat{I}_n\right)}{c^2\sigma^2/n} = \frac{1}{c^2}.$

CLT $\mathbb{P}\left(\left|\hat{I}_n - I\right| > c\frac{\sigma}{\sqrt{n}}\right) \approx 2(1 - \Phi(c)) = \mathcal{O}\left(\frac{e^{-c^2/2}}{c}\right).$, CI $\left(\hat{I}_n \pm c_\alpha\frac{\sigma}{\sqrt{n}}\right) \approx \left(\hat{I}_n \pm c_\alpha\frac{S_n}{\sqrt{n}}\right)$

Toy Example: draw points randomly in a 2x2 square, the proportion of points in the circle circumscribed by the square is pi/4, can use this to get estimate of pi

---

Drawing Random Numbers: How do we do it? Needs to be *pseudo-random*, can't do fully

One approach: $x_{n+1} = (ax_n + c) \bmod M$, maximum period M (then cycles). Looks quite uniform and random

Assuming we can sample from uniform[0,1]: Galtons machine to draw normal samples (the physical machine

, or **Inversion Method;** define **generalised inverse**

$F^-(u) = \inf\{x \in \mathbb{R}; F(x) \ge u\}$ **,** F is cdf, F- is inverse (can do as F is right cont)

Proposition 4.1: $Let\ F\ be\ a\ cdf\ and\ U \sim \mathcal{U}_{[0,1]}.\ Then\ X = F^-(U)\ has\ cdf\ F.$

Fact: $F^-(u) \le x \Leftrightarrow u \le F(x)$. Thus for $U \sim \mathcal{U}_{[0,1]}$, we have $\mathbb{P}(F^-(U) \le x) = \mathbb{P}(U \le F(x)) = F(x).$

Lectures 3&4 are always assessed

Transformation method:

We can simulate Y ~ q, want to simulate X ~ f(q), pdf = pi

Change of variables formula gets us $\pi(x) = q \circ \varphi^{-1}(x) \left|\det(D\varphi^{-1})(x)\right|$, φ is a bijection,
Note: Sometimes will be easier to do the ^-1 outside the determinant
E.g. getting Gamma(a,b) from exp(1)'s. Proof by checking MGFs, *checking MGFs is a common strategy that works*

Box-Muller Algorithm - For simulating Gaussian Distribution

U1 U2 uniform, $R = \sqrt{-2\log(U_1)},\ \vartheta = 2\pi U_2$, $R^2 \sim \text{Exp}(1/2),\ \vartheta \sim \mathcal{U}_{[0,2\pi]}$,
$X = R\cos(\vartheta), Y = R\sin(\vartheta)$ a bijection, Get f(X,Y) in terms of f(R^2,theta) using change of variables, find (X,Y) are independent standard normal
Also a method for **multivariate normal** in slides 3, similar only difference is this time we get the inverse on the inside

---

Sampling via composition: Assume pi marginal, i.e. $\pi(x) = \int \overline{\pi}_{X,Y}(x,y)\,dy$, decompose $\overline{\pi}_{X,Y}(x,y) = \overline{\pi}_Y(y)\,\overline{\pi}_{X|Y}(x|y)$. Sometimes it's easy to compute pi hat but hard/impossible to get pi, so can sample $Y \sim \overline{\pi}_Y$ then $X|Y \sim \overline{\pi}_{X|Y}(\cdot|Y)$ so $(X,Y) \sim \overline{\pi}_{X,Y}$ and hence $X \sim \pi$.

Rejection Sampling 1. Draw $X \sim q$, draw $U \sim \mathcal{U}_{[0,1]}$. $\pi(x) \le M q(x)$
2. Accept $X = x$ as a sample from π if $U \le \frac{\pi(x)}{M q(x)}$, constraint need such a pi to exist
Proposition: The distribution of the samples accepted by rejection sampling is π. Proof in slides, uses bayes on $\mathbb{P}(X \in A | X \text{ accepted})$, writes being in A and constrain on U as indicator functions to get that probability under q(x)

$$\mathbb{P}(X \in A | X \text{ accepted}) = \frac{\mathbb{P}(X \in A, X \text{ accepted})}{\mathbb{P}(X \text{ accepted})} \quad \mathbb{P}(X \in A, X \text{ accepted})$$

$$= \int_{\mathbb{X}}\int_0^1 \mathbb{I}_A(x)\,\mathbb{I}\left(u \le \frac{\pi(x)}{M q(x)}\right) q(x)\,du\,dx = \int_{\mathbb{X}} \mathbb{I}_A(x)\frac{\pi(x)}{M q(x)} q(x)\,dx = \int_{\mathbb{X}} \mathbb{I}_A(x)\frac{\pi(x)}{M}dx = \frac{\pi(A)}{M}.$$

Often we only know pi and q up to some unknown normalising constant $\pi = \tilde{\pi}/Z_\pi \quad \text{and} \quad q = \tilde{q}/Z_q$

If you can upper bound $\tilde{\pi}(x)/\tilde{q}(x) \le \tilde{M}$, then using $\tilde{\pi}$, $\tilde{q}$ and $\tilde{M}$ in the algorithm is correct.

$$\frac{\tilde{\pi}(x)}{\tilde{q}(x)} \le \tilde{M} \Leftrightarrow \frac{\pi(x)}{q(x)} \le \tilde{M}\frac{Z_q}{Z_\pi} = M.$$

Lemma 1: Let T denote the number of pairs (X,U) that have to be generated until X is accepted for the first time. T is geometrically distributed with parameter 1/M and in particular E (T) = M. *Gives a way to estimating ratio of normalising constants* as $\mathbb{E}(T) = M = \frac{Z_q \tilde{M}}{Z_\pi}$ Example: Uniform from bounded subset B of R^p, set $\tilde{M} = 1$ (If the bounded subset filled the uniform), $\tilde{\pi}(x)/\left(\tilde{M}'\tilde{q}(x)\right) = \mathbb{I}_B(x)$, prob(accept) = Z_π/Z_q.
In summary: Requires evaluation of Pi pointwise, an upper bound of pi/q (or same on tildas). Upper Bound not always feasible

Importance Sampling: Want to compute $I = \mathbb{E}_\pi(\varphi(X)) = \int_{\mathbb{X}} \varphi(x)\,\pi(x)\,dx.$, don't know pi but have a proposal distribution q

Require the support of pi to be in q $\pi(x) > 0 \Rightarrow q(x) > 0;$ q called the **proposal/importance distribution**

Identity $I = \mathbb{E}_\pi(\varphi(X)) = \mathbb{E}_q(\varphi(X)w(X))$ w is the **importance weight function** pi/q ⇒ for X iid q,

$\widehat{I}_n^{\mathrm{IS}} = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i)w(X_i).$ **properties of importance sampling:** Unbiased, strongly consistent, CLT

$\mathbb{V}_q(\widehat{I}_n^{IS}) = \sigma_{IS}^2/n \quad \sigma_{IS}^2 := \mathbb{V}_q(\varphi(X)w(X)) \quad If\ \sigma_{IS}^2 < \infty \lim_{n\to\infty} \sqrt{n}\left(\widehat{I}_n^{IS} - I\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_{IS}^2\right)$

Same proofs as for rejection sampling

Question: Is there a best proposal that minimises the variance sigma (IS)? **Proposition:** optimisal proposal $q_{opt}(x) = \frac{|\varphi(x)|\,\pi(x)}{\int_{\mathbb{X}} |\varphi(x)|\,\pi(x)\,dx}$ minimising V_q(I^nIS) is (modulus to make it a true density)]

We have indeed

$$\sigma_{\mathrm{IS}}^2 = \mathbb{V}_q(\varphi(X)w(X)) = \mathbb{E}_q\left(\varphi^2(X)w^2(X)\right) - I^2.$$

We also have by Jensen's inequality for any q

$$\mathbb{E}_q\left(\varphi^2(X)w^2(X)\right) \geq \left(\int_{\mathbb{X}} |\varphi(x)|\,\pi(x)\,dx\right)^2.$$

For $q = q_{\mathrm{opt}}$, we have

$$\mathbb{E}_{q_{\mathrm{opt}}}\left(\varphi^2(X)w^2(X)\right) = \int_{\mathbb{X}} \frac{\varphi^2(x)\pi^2(x)}{|\varphi(x)|\,\pi(x)}dx \times \int_{\mathbb{X}} |\varphi(x)|\,\pi(x)\,dx = \left(\int_{\mathbb{X}} |\varphi(x)|\,\pi(x)\,dx\right)^2. \quad \square$$

Proof Variance equation gets an equation with a part I we can't minimise on
Jensens gives us a lower bound, subbing in qopt gives us equality to the lower bound

.

Problem: qopt(x) can never be used in practice. Requires us to know I Can use as a guideline to choose q though

Not always best to sample from the actual distribution pi, e.g. another distribution that is most concentrated at the most likely points of pi can be better (toy example in slide 4, t distribution)

Normalised Importance Sampling: Assume $\pi(x) = \widetilde{\pi}(x)/Z_\pi$ and $q(x) = \widetilde{q}(x)/Z_q$, $\pi(x) > 0 \Rightarrow q(x) > 0$ and and define $\widetilde{w}(x) = \frac{\widetilde{\pi}(x)}{\widetilde{q}(x)}$ Alternative Identity is

$$I = \mathbb{E}_\pi(\varphi(X)) = \frac{\int_{\mathbb{X}} \varphi(x)\,\widetilde{w}(x)\,q(x)dx}{\int_{\mathbb{X}} \widetilde{w}(x)q(x)dx}.$$

Main complication is this is the ratio of expectations (pi ~)

Propositions 2.1 (SLLN for NIS) $Let\ X_1, \ldots, X_n \overset{i.i.d.}{\sim} q\ and\ assume\ that\ \mathbb{E}_q(|\varphi(X)|\,w(X)) < \infty.$ then is strongly consistent. **Proof:** divide top and bottom by n, both converge a.s by SLLN. BUT

for finite n $\widehat{I}_n^{\mathrm{NIS}}$ is **biased**

Proposition 2.2 $If\ \mathbb{V}_q(\varphi(X)w(X)) < \infty\ and\ \mathbb{V}_q(w(X)) < \infty\ then$

$$\sqrt{n}(\widehat{I}_n^{NIS} - I) \Rightarrow \mathcal{N}(0, \sigma_{NIS}^2),\ \sigma_{NIS}^2 := \mathbb{V}_q\left(\left[w(X)\left(\varphi(X) - I\right)\right]\right) = \int \frac{\pi(x)^2\left(\varphi(x) - I\right)^2}{q(x)}\mathrm{d}x.$$

Proof

First notice that with $X_1, \ldots, X_n$ i.i.d. $\sim q$

$$\sqrt{n}(\widehat{I}_n^{\text{NIS}} - I) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \widetilde{w}(X_i)[\varphi(X_i) - I]}{\frac{1}{n}\sum_{i=1}^n \widetilde{w}(X_i)}$$

where since $\widetilde{w}(x) = \widetilde{\pi}/\widetilde{q}$

$$\mathbb{E}_q\left[\widetilde{w}(X_n)(\varphi(X_i) - I)\right] = 0.$$

Since $\mathbb{V}_q(\varphi(X)w(X)) < \infty$ by standard CLT

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \widetilde{w}(X_i)[\varphi(X_i) - I] \Rightarrow \mathcal{N}\left(0, \mathbb{V}_q\left(\widetilde{w}(X_1)[\varphi(X_1) - I]\right)\right).$$

The strong law of large numbers applied to the denominator

$$\frac{1}{n}\sum_{i=1}^n \widetilde{w}(X_i) \to \mathbb{E}_q[\widetilde{w}(X_1)] = Z_\pi/Z_q, \quad \text{a.s.}$$

By Slutsky's theorem, combining the two

$$\sqrt{n}(\widehat{I}_n^{\text{NIS}} - I) \Rightarrow \mathcal{N}\left(0, \mathbb{V}_q(\widetilde{w}(X_1)[\varphi(X_1) - I])\frac{Z_q^2}{Z_\pi^2}\right)$$
$$\sim \mathcal{N}\left(0, \sigma_{\text{NIS}}^2\right).$$

Writes out LHS., take expectation, and then use CLT
Use SLLN, then **Slutskys Theorem** to combine
(Alternatively use Delta method instead of Slutskys)

SLLN, CLT, and Slutskys are the three main tools we will use'

Variance of importance Sampling estimators

- Standard Importance Sampling: $X_1, \ldots, X_n \overset{iid}{\sim} q$,

$$\widehat{I}_n^{\text{IS}} = \frac{1}{n}\sum_{i=1}^n \varphi(X_i)w(X_i).$$

- Asymptotic Variance:

$$\mathbb{V}_{as}\left(\widehat{I}_n^{\text{IS}}\right) = \mathbb{E}_q\left[\left(\varphi(X)w(X) - \mathbb{E}_q\left(\varphi(X)w(X)\right)\right)^2\right]$$
$$\approx \frac{1}{n}\sum_{i=1}^n \left(\varphi(X_i)w(X_i) - \widehat{I}_n^{\text{IS}}\right)^2.$$

- Thus the asymptotic variance can be estimated consistently with

$$\frac{1}{n}\sum_{i=1}^n \left(\varphi(X_i)w(X_i) - \widehat{I}_n^{\text{IS}}\right)^2.$$

- Normalised Importance Sampling: $X_1, \ldots, X_n \overset{iid}{\sim} q$,

$$\widehat{I}_n^{\text{NIS}} = \frac{\sum_{i=1}^n \varphi(X_i)\widetilde{w}(X_i)}{\sum_{i=1}^n \widetilde{w}(X_i)}.$$

- Asymptotic Variance:

$$\mathbb{V}_{as}\left(\widehat{I}_n^{\text{NIS}}\right) = \frac{\mathbb{E}_q\left[\left(\widetilde{w}(x)\left(\varphi(X) - I\right)\right)^2\right]}{\mathbb{E}_q\left[\widetilde{w}(X)\right]^2}.$$

- Thus the asymptotic variance can be estimated consistently with

$$\frac{\frac{1}{n}\sum_{i=1}^N \widetilde{w}(X_i)^2\left(\varphi(X_i) - \widehat{I}_n^{\text{NIS}}\right)^2}{\left(\frac{1}{n}\sum_{i=1}^N \widetilde{w}(X_i)\right)^2}.$$

MC vs IS vs NIS: NIS is sometimes better than IS, despite seeming to have extra variance from the ratio of distributions and biassed (while IS is unbiased)

- **Toy example:** Let $\mathbb{X} = \mathbb{R}^d$ and

$$\pi(x) = \frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{\sum_{i=1}^d x_i^2}{2}\right)$$

and

$$q(x) = \frac{1}{(2\pi\sigma^2)^{d/2}}\exp\left(-\frac{\sum_{i=1}^d x_i^2}{2\sigma^2}\right).$$

Important toy example

$$\mathbb{P}(X \text{ accepted}) = \frac{1}{\sigma^d} \to 0 \text{ as } d \to \infty,$$

very inefficient as d gets big and sig>1

$$\mathbb{V}_q[w(X)] = \left(\frac{\sigma^4}{2\sigma^2 - 1}\right)^{d/2} - 1$$

Importance sampling, exponentially increasing variance
Weird as we wanted the rate to be independent of d, and here this is true but the 'constant' (in n) explodes exponentially in d

Markov Chain Monte Carlo:

Pseudo Marginal is *not examinable*
Markov Chain Theory is *not examinable*, application is (won't get question like what is irreducible)
Hamiltonian Monte Carlo is *not examinable kinda*, too hard to make a question on
1 question from MCMC, Gibbs Sampling, Metropolis Hasting, Parallel Tempering Capacity and Diagnostics are examinable (PTC and Diagnostics are part of Metropolis Hasting kinda)

MC methods give convergence to pi of rate 1/Root(n), independent of dimension d of x. But error still depends on d, through constant in front of rate, and it usually explodes in d. *If we could get something dependence of pi(x)/pi(y) this constant would disappear*

Want a distribution X such that $\frac{1}{n}\sum_{t=1}^{n}\phi(X_t) \to \int \phi(x)\pi(x)\,dx$ works for X correctly distributed by Ergodic theorem, CLT for rate but we have a Markov Chain, so don't have independence, so need a different thing to get the rate.

- The state space $\mathbb{X}$ is now continuous, e.g. $\mathbb{R}^d$.
- $(X_t)_{t\geq 1}$ is a Markov chain if for any (measurable) set A,

$$\begin{aligned}\mathbb{P}(X_t \in A | X_1 = x_1, X_2 = x_2, ..., X_{t-1} = x_{t-1}) \\ = \mathbb{P}(X_t \in A | X_{t-1} = x_{t-1}).\end{aligned}$$

- We have

$$\mathbb{P}(X_t \in A | X_{t-1} = x) = \int_A K(x, y)\,dy = K(x, A),$$

that is conditional on $X_{t-1} = x$, X_t is a random variable which admits a probability density function $K(x, \cdot)$.

- $K : \mathbb{X}^2 \to \mathbb{R}$ is the **kernel** of the Markov chain.

- Consider the autoregressive (AR) model

$$X_t = \rho X_{t-1} + V_t$$

where $V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2)$. This defines a Markov chain such that

$$K(x, y) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(y - \rho x)^2\right).$$

Autoregressive model

Continuous markov chain definitions

rreducibility and aperiodicity

Definition

Given a probability measure μ over $\mathbb{X}$, a Markov chain is μ-irreducible if

$$\forall x \in \mathbb{X} \quad \forall A : \mu(A) > 0 \quad \exists t \in \mathbb{N} \quad K^t(x, A) > 0.$$

A μ-irreducible Markov chain of transition kernel K is periodic if there exists some partition of the state space $\mathbb{X}_1, ..., \mathbb{X}_d$ for $d \geq 2$, such that

$$\forall i, j, t, s : \ \mathbb{P}(X_{t+s} \in \mathbb{X}_j | X_t \in \mathbb{X}_i) = \begin{cases} 1 & j = i + s \bmod d \\ 0 & \text{otherwise.} \end{cases}.$$

Otherwise the chain is aperiodic.

Recurrence and Harris Recurrence

For any measurable set A of $\mathbb{X}$, let

$$\eta_A = \sum_{k=1}^{\infty} \mathbb{1}_A(X_k),$$

be *the number of visits to the set A*.

Definition

A μ-irreducible Markov chain is recurrent if for any measurable set $A \subset \mathbb{X} : \mu(A) > 0$, then

$$\forall x \in A \quad \mathbb{E}_x(\eta_A) = \infty.$$

A μ-irreducible Markov chain is Harris recurrent if for any measurable set $A \subset \mathbb{X} : \mu(A) > 0$, then

$$\forall x \in \mathbb{X} \quad \mathbb{P}_x(\eta_A = \infty) = 1.$$

Note: Harris recurrence is stronger than recurrence.

Lecture 3: Markov Chains

Invariant Distribution and Reversibility

Definition

A distribution of density π is invariant or stationary for a Markov kernel K, if

$$\int_{\mathbb{X}} \pi(x) K(x, y) \, dx = \pi(y).$$

A Markov kernel K is π-reversible if

$$\forall f \quad \iint f(x, y) \pi(x) K(x, y) \, dxdy = \iint f(x, y) \pi(y) K(y, x) \, dxdy$$

where f is a bounded measurable function.

Gibbs Sampling

$$\pi(x) = \pi(x_1, x_2, ..., x_d), \quad x \in \mathbb{R}^d.$$

Interested in sampling

Notation: $x_{-i} := (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d)$.

Systematic scan Gibbs sampler. Let $\left(X_1^{(1)}, ..., X_d^{(1)}\right)$ be the initial state then iterate for $t = 2, 3, ...$

1. Sample $X_1^{(t)} \sim \pi_{X_1|X_{-1}}\left(\cdot | X_2^{(t-1)}, ..., X_d^{(t-1)}\right)$.

   ⋮

j. Sample $X_j^{(t)} \sim \pi_{X_j|X_{-j}}\left(\cdot | X_1^{(t)}, ..., X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)}\right)$

   ⋮

d. Sample $X_d^{(t)} \sim \pi_{X_d|X_{-d}}\left(\cdot | X_1^{(t)}, ..., X_{d-1}^{(t)}\right)$.

This is a markov chain:New state only depends on the current state. If instead of using the updated Xj(t)s in updating future ones and just used Xj(t-1) we'd still have a markov chain, but a different one
A few questions about this algorithm:

1. Is the joint distribution π uniquely specified by the conditional distributions πXi |X−i ?
2. Does the Gibbs sampler provide a Markov chain with the correct stationary distribution π?
3. If yes, does the Markov chain converge towards this invariant distribution?

It will turn out to be the case under some mild conditions

Hammersley–Clifford Theorem

Consider a distribution with continuous density $\pi(x_1, x_2, ..., x_d)$ *such that*

$$supp(\pi) = supp\left(\bigotimes_{i=1}^{d} \pi_{X_i}\right).$$

Then for any $(z_1, ..., z_d) \in supp(\pi)$, *we have*

$$\pi(x_1, x_2, ..., x_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j|X_{-j}}(x_j | x_{1:j-1}, z_{j+1:d})}{\pi_{X_j|X_{-j}}(z_j | x_{1:j-1}, z_{j+1:d})}.$$

Note: the condition above is known as the positivity condition.

Equivalently, if $\pi_{X_i}(x_i) > 0$ for $i = 1, \ldots, d$, then

$$\pi(x_1, \ldots, x_d) > 0.$$

Proof.
Similarly, we have

$$\begin{aligned}\pi(x_{1:d-1}, z_d) &= \pi(x_{1:d-2}, z_{d-1}, z_d)\frac{\pi(x_{1:d-1}, z_d)}{\pi(x_{1:d-2}, z_{d-1}, z_d)} \\ &= \pi(x_{1:d-2}, z_{d-1}, z_d)\frac{\pi(x_{1:d-1}, z_d)/\pi(x_{1:d-2}, z_d)}{\pi(x_{1:d-2}, z_{d-1}, z_d)/\pi(x_{1:d-2}, z_d)} \\ &= \pi(x_{1:d-2}, z_{d-1}, z_d)\frac{\pi_{X_{d-1}|X_{-(d-1)}}(x_{d-1} \mid x_{1:d-2}, z_d)}{\pi_{X_{d-1}|X_{-(d-1)}}(z_{d-1} \mid x_{1:d-2}, z_d)}\end{aligned}$$

hence

$$\begin{aligned}\pi(x_{1:d}) = \pi(x_{1:d-2}, z_{d-1}, z_d)&\frac{\pi_{X_{d-1}|X_{-(d-1)}}(x_{d-1}|x_{1:d-2}, z_d)}{\pi_{X_{d-1}|X_{-(d-1)}}(z_{d-1}|x_{1:d-2}, z_d)} \\ &\times \frac{\pi_{X_d|X_{-d}}(x_d|x_{1:d-1})}{\pi_{X_d|X_{-d}}(z_d|x_{1:d-1})}\end{aligned}$$

Proof.
We have

$$\begin{aligned}\pi(x_{1:d-1}, x_d) &= \pi_{X_d|X_{-d}}(x_d|x_{1:d-1})\pi(x_{1:d-1}), \\ \pi(x_{1:d-1}, z_d) &= \pi_{X_d|X_{-d}}(z_d|x_{1:d-1})\pi(x_{1:d-1}).\end{aligned}$$

Therefore

$$\begin{aligned}\pi(x_{1:d}) &= \pi(x_{1:d-1}, z_d)\frac{\pi(x_{1:d-1}, x_d)}{\pi(x_{1:d-1}, z_d)} \\ &= \pi(x_{1:d-1}, z_d)\frac{\pi(x_{1:d-1}, x_d)/\pi(x_{1:d-1})}{\pi(x_{1:d-1}, z_d)/\pi(x_{1:d-1})} \\ &= \pi(x_{1:d-1}, z_d)\frac{\pi_{X_d|X_{1:d-1}}(x_d \mid x_{1:d-1})}{\pi_{X_d|X_{1:d-1}}(z_d \mid x_{1:d-1})}.\end{aligned}$$

Proof.
By $z \in \text{supp}(\pi)$ we have that $\pi_{X_i}(z_i) > 0$ for all i.

Also, we are allowed to suppose that $\pi_{X_i}(x_i) > 0$ for all i.

Thus all the conditional probabilities we introduce are positive since

$$\begin{aligned}&\pi_{X_j|X_{-j}}(x_j \mid x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_d) \\ &= \frac{\pi(x_1, \ldots, x_{j-1}, x_j, z_{j+1}, \ldots, z_d)}{\pi(x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_d)} > 0.\end{aligned}$$

By iterating we have the theorem. □

Example: Non-Integrable target (in slide 6)

Invariance of the Gibbs sampler

The systematic scan Gibbs sampler kernel admits π as invariant distribution.

- The kernel of the Gibbs sampler (case $d = 2$) is

$$K(x^{(t-1)}, x^{(t)}) = \pi_{X_1|X_2}(x_1^{(t)} \mid x_2^{(t-1)})\pi_{X_2|X_1}(x_2^{(t)} \mid x_1^{(t)})$$

- Case $d > 2$:

$$K(x^{(t-1)}, x^{(t)}) = \prod_{\cdot}^{d} \pi_{X_j|X_{-j}}(x_j^{(t)} \mid x_{1:j-1}^{(t)}, x_{j+1:d}^{(t-1)})$$

Proof for $d = 2$.
Let $x = (x_1, x_2)$ and $y = (y_1, y_2)$. Then we have

$$\int K(x, y)\pi(x)dx = \int \pi(y_2 \mid y_1)\pi(y_1 \mid x_2)\pi(x_1, x_2)dx_1 dx_2$$
$$= \pi(y_2 \mid y_1) \int \pi(y_1 \mid x_2)\pi(x_2)dx_2$$
$$= \pi(y_2 \mid y_1)\pi(y_1) = \pi(y_1, y_2) = \pi(y).$$

Can use this for bigger d by grouping them (e.g X1~ = X1, X2~ = (X2,X3,...))

Proposition

*Assume π satisfies the positivity condition, then the Gibbs sampler yields a π-irreducible and recurrent * Markov chain.*

Proof.

Recurrence follows from irreducibility and the fact that π is invariant (see Meyn and Tweedie, Prop'n 10.1.1.).
Irreducibility. Let $\mathbb{X} \subset \mathbb{R}^d$, such that $\pi(\mathbb{X}) = 1$. Write K for the kernel and let $A \subset \mathbb{X}$ such that $\pi(A) > 0$. Then for any $x \in \mathbb{X}$

$$\begin{aligned} K(x, A) &= \int_A K(x, y)\mathrm{d}y \\ &= \int_A \pi_{X_1|X_{-1}}(y_1 \mid x_2, \ldots, x_d) \times \cdots \\ &\quad \times \pi_{X_d|X_{-d}}(y_d \mid y_1, \ldots, y_{d-1})\mathrm{d}y. \end{aligned}$$

□

Irreducibility and Recurrence (cont.)

Proof.

Thus if for some $x \in \mathbb{X}$ and A with $\pi(A) > 0$ we have $K(x, A) = 0$, we must have that

$$\pi_{X_1|X_{-1}}(y_1 \mid x_2, \ldots, x_d) \times \cdots \times \pi_{X_d|X_{-d}}(y_d \mid y_1, \ldots, y_{d-1}) = 0,$$

for almost all $y = (y_1, \ldots, y_d) \in A$.

Therefore, by the Hammersley–Clifford theorem, we must also have that

$$\pi(y_1, y_2, \ldots, y_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j|X_{-j}}(y_j \mid y_{1:j-1}, x_{j+1:d})}{\pi_{X_j|X_{-j}}(x_j \mid y_{1:j-1}, x_{j+1:d})} = 0,$$

for almost all $y = (y_1, \ldots, y_d) \in A$ and thus $\pi(A) = 0$ obtaining a contradiction. □

Theorem

If the positivity condition is satisfied then for any π-integrable function $\phi : \mathbb{X} \to \mathbb{R}$:

$$\lim \frac{1}{t} \sum_{i=1}^{t} \phi\left(X^{(i)}\right) = \int_{\mathbb{X}} \phi(x)\, \pi(x)\, \mathrm{d}x$$

for π-almost all starting values $X^{(1)}$.

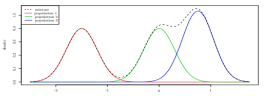
LLN for Gibbs Sampler

Gibbs sampling and auxiliary variables

- Gibbs sampling requires sampling from $\pi_{X_j|X_{-j}}$.
- In many scenarios, we can include a set of auxiliary variables $Z_1, ..., Z_p$ and have an "extended" distribution of joint density $\overline{\pi}(x_1, ..., x_d, z_1, ..., z_p)$ such that

$$\int \overline{\pi}(x_1, ..., x_d, z_1, ..., z_p)\, dz_1 ... dz_d = \pi(x_1, ..., x_d).$$

Mixture models, Capture-recapture,tobit,probit etc.
Sometimes adding extra 'auxillary' variables (Z1,....Zp) actually makes the distribution easier
Can be hard to decide what auxiliary variables to get. E.g. didn't have a nice method for sampling Bayesian Logistic Regression until 2013

Example: Mixture of Normals. K = 2 for simplicity, data is a mixture of gaussians with different mean but same variance 1, want to estimate the means  Not obvious how to do it, because the conditional distribution (below) is pretty messy

- Independent data $y_1, ..., y_n$ $\quad Y_i|\theta \sim \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k, \sigma_k^2)$ where $\theta = (p_1, ..., p_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2)$
- Likelihood function

$$p(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \frac{p_k}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right) \right)$$

- For example, let's fix $K = 2$, $\sigma_k^2 = 1$ and $p_k = 1/K$ for all k.

Suppose prior $p(\theta) = p(\mu_1)\, p(\mu_2)$

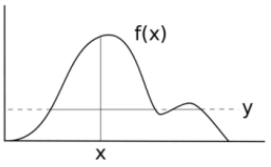
Goal: introduce a variable for each data point that tells us which gaussian the datapoint comes from, in a way that preserves original marginal density

Associate to each Yi an auxiliary Zi where $\mathbb{P}(Z_i = k|\theta) = p_k$ and $Y_i|Z_i = k, \theta \sim \mathcal{N}(\mu_k, \sigma_k^2)$.

$\Rightarrow$ $p(y_i|\theta) = \sum_{k=1}^{K} \mathbb{P}(Z_i = k)\mathcal{N}(y_i; \mu_k, \sigma_k^2)$ extended posterior $p(\theta, z_1, ..., z_n|y_1, ..., y_n) \propto p(\theta) \prod_{i=1}^{n} \mathbb{P}(z_i|\theta)\, p(y_i|z_i, \theta)$.

Integrating out the Zi in the new likelihood function gives the original Likelihood

GIbbs Samples alternately $\mathbb{P}(z_{1:n}|y_{1:n}, \theta)$ $p(\theta|y_{1:n}, z_{1:n})$ (saw alternate sampling last time)
Notes has an example for set variance 1 and pk uniform.



Slice Sampling: A form of Gibbs Sampling, y is the auxillary.
Assumes you can compute and sample the distribution along each line.
Start from a random x coordinate sample along that sample from random y = f(x) proportional to pi(x) coordinate sample along that.

We Consider the joint density over $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ uniform over the region $\{(x, y) : 0 < y < f(x)\}$ $\quad p(x, y) = (1/Z)\mathbb{1}_{\{0<y<f(x)\}}, \quad Z = \int f(x)dx$

The marginal density for x is then $\int_0^{f(x)} (1/Z)dy = f(x)/Z = \pi(x)$

Exam question: want to show this is a gibbs sampler for a given distribution. Two steps pi-bar(x|y) and pi-bar(y|x), show they're uniform

Gibbs Sampling Summary

- Given a target $\pi(x) = \pi(x_1, x_2, ..., x_d)$, Gibbs sampling works by sampling from $\pi_{X_j|X_{-j}}(x_j|x_{-j})$ for $j = 1, ..., d$.
- Sampling exactly from one of these full conditionals might be a hard problem itself.
- Even if it is possible, the Gibbs sampler might converge slowly if components are highly correlated.
- If the components are not highly correlated then Gibbs sampling performs well, even when $d \to \infty$ (e.g. mixing times of some Gibbs samplers can even be dimension-free, see [Yang and Rosenthal, 2017]).
- Metropolis–Hastings algorithm (1953, 1970) is a more general algorithm that can bypass these problems.

Hard to ask a question on Gibbs because the only thing you can really do is ask us to define an auxiliary variable for something

Metropolis Hastings Algorithm: Frameworks:

1. Target distribution on $\mathbb{X} = \mathbb{R}^d$ of density $\pi(x)$
2. Proposal Distribution for any $x, x' \in \mathbb{X}$, we have $q(x'|x) \geq 0$ and $\int_{\mathbb{X}} q(x'|x)\,dx' = 1$.
3. Starting with $X^{(1)}$, for $t = 2, 3, ...$
   a. 1. Sample $X^\star \sim q(\cdot|X^{(t-1)})$
   b. Compute $$\alpha\left(X^\star|X^{(t-1)}\right) = \min\left(1, \frac{\pi(X^\star)\,q\left(X^{(t-1)}|X^\star\right)}{\pi\left(X^{(t-1)}\right)\,q\left(X^\star|X^{(t-1)}\right)}\right)$$
   c. Sample $U \sim \mathcal{U}_{[0,1]}$. If $U \leq \alpha\left(X^\star|X^{(t-1)}\right)$, set $X^{(t)} = X^\star$, otherwise set $X^{(t)} = X^{(t-1)}$.
4. Q is something we can sample from. 3c, is if we decide to accept a change to a new state, moving around the distribution
5. Is clearly a markov chain: movement only dependent on proposal and current position

Average Acceptance probability from current state $a\left(x^{(t-1)}\right) := \int_{\mathbb{X}} \alpha\left(x|x^{(t-1)}\right) q\left(x|x^{(t-1)}\right) dx$

in which case $X^{(t)} = X$, otherwise $X^{(t)} = X^{(t-1)}$.

Only requires point-wise evaluations of pi(x) up to a normalising constant. if $\widetilde{\pi}(x) \propto \pi(x)$

$$\frac{\pi(x^\star)\,q\left(x^{(t-1)}|x^\star\right)}{\pi\left(x^{(t-1)}\right)\,q\left(x^\star|x^{(t-1)}\right)} = \frac{\widetilde{\pi}(x^\star)\,q\left(x^{(t-1)}|x^\star\right)}{\widetilde{\pi}\left(x^{(t-1)}\right)\,q\left(x^\star|x^{(t-1)}\right)}.$$

Lemma: kernel of MH algorithm is $K(x, y) = \alpha(y \mid x)q(y \mid x) + (1 - a(x))\delta_x(y).$

Proof: $K(x, y) = \int q(x^\star \mid x)\{\alpha(x^\star \mid x)\delta_{x^\star}(y) + (1 - \alpha(x^\star \mid x))\delta_x(y)\}dx^\star$

$= q(y \mid x)\alpha(y \mid x) + \left\{\int q(x^\star \mid x)(1 - \alpha(x^\star \mid x))dx^\star\right\}\delta_x(y) = q(y \mid x)\alpha(y \mid x) + \left\{1 - \int q(x^\star \mid x)\alpha(x^\star \mid x)dx^\star\right\}\delta_x(y)$

$= q(y \mid x)\alpha(y \mid x) + \{1 - a(x)\}\,\delta_x(y).$

Need to be careful as distribution itself is continuous, but for K can assume x is given (as for if accepted (y=/=x) or rejected (y=x)

Advice for above: In exam consider y=x and y=/=x separately

Reversibility: The Metropolis–Hastings kernel K is π-reversible and thus admits π as invariant distribution.

Proof: For any $x, y \in \mathbb{X}$, with $x \neq y$ $\pi(x)K(x,y) = \pi(x)q(y \mid x)\alpha(y \mid x) = \pi(x)q(y \mid x)\left(1 \wedge \frac{\pi(y)q(x \mid y)}{\pi(x)q(y \mid x)}\right)$

$$= \left(\pi(x)q(y \mid x) \wedge \pi(y)q(x \mid y)\right) = \pi(y)q(x \mid y)\left(\frac{\pi(x)q(y \mid x)}{\pi(y)q(x \mid y)} \wedge 1\right) = \pi(y)K(y,x).$$

If $x = y$, then obviously $\pi(x)K(x,y) = \pi(y)K(y,x)$

Consider target $\pi(x) = \left(\mathcal{U}_{[0,1]}(x) + \mathcal{U}_{[2,3]}(x)\right)/2$ and proposal $q(x^\star \mid x) = \mathcal{U}_{(x-\delta, x+\delta)}(x^\star)$

The MH chain is reducible if $\delta \leq 1$: the chain stays either in [0,1] or [2,3]

Proposition: *If $q(x^\star \mid x) > 0$ for any $x, x^\star \in supp(\pi)$ then the Metropolis-Hastings chain is irreducible, in fact every state can be reached in a single step (strongly irreducible).*

If the MH chain is *irreducible* then it is also *Harris recurrent*

LLN for MH: If the Markov chain generated by the Metropolis–Hastings sampler is π-irreducible, then we have for any integrable function $\phi : \mathbb{X} \to \mathbb{R}$:

$$\lim_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \phi\left(X^{(i)}\right) = \int_{\mathbb{X}} \phi(x)\,\pi(x)\,dx \quad \textit{for every starting value } X^{(1)}.$$

Independent proposal: a proposal distribution $q(x^\star \mid x)$ not dependent on x

Acceptance probability then $\alpha(x^\star \mid x) = \min\left(1, \frac{\pi(x^\star)q(x)}{\pi(x)q(x^\star)}\right)$

Uniformly Ergodic: If $\pi(x)/q(x) < M$ for all x and some $M < \infty$, (hard to satisfy)

If it is satisfied, then the expected acceptance probability is at least 1/M (was an exam question to prove). If such M doesn't exist, not even geometrically ergodic.

Choosing a good proposal distribution:

Goal: design a Markov chain with small correlation $\rho\left(X^{(t-1)}, X^{(t)}\right)$ between subsequent values (why?).

Two sources of correlation:
- between the current state $X^{(t-1)}$ and proposed value $X \sim q\left(\cdot \mid X^{(t-1)}\right)$,
- correlation induced if $X^{(t)} = X^{(t-1)}$, if proposal is rejected.

Trade-off: there is a compromise between
- proposing large moves,
- obtaining a decent acceptance probability.

Don't want to keep rejecting (so may want to be close to current place where we know we were accepted), but don't want to explore only one area of the state space

Random walk Metropolis (RWM)

In the Metropolis–Hastings, pick $q(x^\star \mid x) = g(x^\star - x)$ with g being a *symmetric* distribution, thus $X^\star = X + \varepsilon, \quad \varepsilon \sim g;$

So acceptance probability is $\alpha(x^\star \mid x) = \min\left(1, \frac{\pi(x^\star)}{\pi(x)}\right)$

Accept:
- a move to a more probable state with probability 1;
- a move to a less probable state with probability $\pi(x^\star)/\pi(x) \leq 1.$

Example in notes

One choice to consider is how much variance we want epsilon to have (more for explore)

Knowing how big an acceptance rate to have is 'Optimal Scaling Theory': some say 23.4%

Some maximise expected squared jumping distance $\mathbb{E}\left[||X_{t+1} - X_t||^2\right]$
Steps: Test variance for T iterations, measure criterion, change to get better performance on criterion. Called **Adaptive MCMC (**change Kernel Kt)
Using this method actually breaks markov property, as we're using the previous data to make a decision about the new distribution. So need to prove convergence again, though fine if the adapting eventually stops
Can also have Adaptive Gibbs Samplers

Langevin Diffusion $dX_t = \frac{1}{2}\nabla \log \pi(X_t)dt + dB_t$ has a stationary distribution π(if π is smooth)

⇒ ULA Unadjusted Langevin algorithm $X^{(t)} = X^{(t-1)} + \frac{\sigma}{2}\nabla \log \pi\left(X^{(t-1)}\right) + \sigma W$

MALA Metropolis adjusted Langevin Algorithm $X^\star = X^{(t-1)} + \frac{\sigma}{2}\nabla \log \pi\left(X^{(t-1)}\right) + \sigma W$

so the Metropolis–Hastings acceptance ratio is

$$\frac{\pi(X^\star)}{\pi(X^{(t-1)})}\frac{\mathcal{N}(X^{(t-1)}; X^\star + \frac{\sigma}{2}.\nabla \log \pi\left(X^\star\right); \sigma^2)}{\mathcal{N}(X^\star; X^{(t-1)} + \frac{\sigma}{2}.\nabla \log \pi\left(X^{(t-1)}\right); \sigma^2)}.$$ The sigmaW pulls more to middle

Sampling and Optimisation ▶ π is log-concave if $U(x) := -\log \pi(x)$ is convex.

$q(X^\star|X^{(t-1)}) = g(X^\star; \varphi(X^{(t-1)}))$ psi is a deterministic mapping

$$\frac{\pi(X^\star)q(X^{(t-1)}|X^\star)}{\pi(X^{(t-1)})q(X^\star|X^{(t-1)})} = \frac{\pi(X^\star)g(X^{(t-1)}; \varphi(X^\star))}{\pi(X^{(t-1)})g(X^\star; \varphi(X^{(t-1)}))}.$$ Can use higher derivatives

Transformations:

Want to sample $\text{supp}(\pi) \subset \mathbb{R}^+$, the posterior distribution of a variance/scale parameter
Any proposed move to R- is a waste of time, so can transform proposals in R to only ones in R+: Given $X^{(t-1)}$, propose $X^\star = \exp(\log X^{(t-1)} + \epsilon)$. Then acceptance =

$$\alpha(X^\star \mid X^{(t-1)}) = \min\left(1, \frac{\pi(X^\star)}{\pi(X^{(t-1)})}\frac{q(X^{(t-1)} \mid X^\star)}{q(X^\star \mid X^{(t-1)})}\right) = \min\left(1, \frac{\pi(X^\star)}{\pi(X^{(t-1)})}\frac{X^\star}{X^{(t-1)}}\right)$$

Because $$\frac{q(y \mid x)}{q(x \mid y)} = \frac{\frac{1}{y\sigma\sqrt{2\pi}}\exp\left[-\frac{(\log y - \log x)^2}{2\sigma^2}\right]}{\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{(\log x - \log y)^2}{2\sigma^2}\right]} = \frac{x}{y}.$$

Random Proposals: Acceptance not considering integral

Assume you want to use $q_{\sigma^2}(X^\star|X^{(t-1)}) = \mathcal{N}(X; X^{(t-1)}, \sigma^2)$ but you don't know how to pick σ^2. You decide to pick a random $\sigma^{2,\star}$ from a distribution $f(\sigma^2)$:

$$\sigma^{2,\star} \sim f(\sigma^{2,\star}),\ X^\star|\sigma^{2,\star} \sim q_{\sigma^{2,\star}}(\cdot|X^{(t-1)})$$

$$q(X^\star|X^{(t-1)}) = \int q_{\sigma^{2,\star}}(X^\star|X^{(t-1)})f(\sigma^{2,\star})d\sigma^{2,\star}.$$

Perhaps $q(X^\star|X^{(t-1)})$ cannot be evaluated, e.g. the above integral is intractable. Hence the acceptance probability

$$\min\left\{1, \frac{\pi(X^\star)q(X^{(t-1)}|X^\star)}{\pi(X^{(t-1)})q(X^\star|X^{(t-1)})}\right\}$$

cannot be computed.

- Instead you decide to accept your proposal with probability

$$\alpha_t = \min\left\{1, \frac{\pi\left(X^\star\right) q_{\sigma^{2,(t-1)}}\left(X^{(t-1)}|X^\star\right)}{\pi\left(X^{(t-1)}\right) q_{\sigma^{2,\star}}\left(X^\star|X^{(t-1)}\right)}\right\}$$

where $\sigma^{2,(t-1)}$ corresponds to parameter of the last accepted proposal.

With probability α_t, set $\sigma^{2,(t)} = \sigma^{2,\star}$, $X^{(t)} = X^\star$, otherwise $\sigma^{2,(t)} = \sigma^{2,(t-1)}$, $X^{(t)} = X^{(t-1)}$.

Makes MC (X x Sigma^2)

$$\widetilde{\pi}\left(x, \sigma^2\right) := \pi\left(x\right) f\left(\sigma^2\right) \quad q(y, \tau^2|x, \sigma^2) = f(\tau^2) q_{\tau^2}(y|x)$$

$$\frac{\widetilde{\pi}(y, \tau^2)}{\widetilde{\pi}(x, \sigma^2)} \frac{q(x, \sigma^2|y, \tau^2)}{q(y, \tau^2|x, \sigma^2)} = \frac{\pi(y) f(\tau^2)}{\pi(x) f(\sigma^2)} \frac{f(\sigma^2) q_{\sigma^2}(x|y)}{f(\tau^2) q_{\tau^2}(y|x)} = \frac{\pi(y)}{\pi(x)} \frac{q_{\sigma^2}(x|y)}{q_{\tau^2}(y|x)}$$

Convergence Diagnostics

In general, impossible to see whether MCMC has converged, *but can sometimes know there it hasn't*

Autocorrelogram: Visual diagnostics, shows autocorrelation (correlation of a signal with a delayed copy of itself), almost all MCMC produce positively correlated samples

Gewekes Diagnostics: Uses partial mean of the samples. Determines the burn-in period: the smallest early portion of the chain that passes the diagnostic (i.e. mean of first x%). Example if the target mean is -2, you ask for the burn in period to be within 0.5 of -2

Gelman Rubin Diagnostics: In order to check for getting stuck at one section**.** start M chains from various starting points, should be similar

Look at classical sum of squares within group and without group

$$\sum_{m=1}^{M}\sum_{t=1}^{T}(X_{m,t} - \bar{X}_{\cdot,\cdot})^2 = \sum_{m=1}^{M}\sum_{t=1}^{T}(\bar{X}_{m,\cdot} - \bar{X}_{\cdot,\cdot})^2 \quad \text{inter-group} \quad + \sum_{m=1}^{M}\sum_{t=1}^{T}(X_{m,t} - \bar{X}_{m,\cdot})^2 \quad \text{intra-group}$$

Expect inter average inter and intra group value to be the same

$$W = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T-1}\sum_{t=1}^{T}(X_{m,t} - \bar{X}_{m,\cdot})^2 \quad B = \frac{1}{M-1}\sum_{m=1}^{M}(\bar{X}_{m,\cdot} - \bar{X}_{\cdot,\cdot})^2 \quad V = \left(1 - \frac{1}{T}\right)W + B$$

W and V should both converge to true variance of target distribution, V would be unbiased if starting points were drawn from target, and W usually smaller than V. So plot sqrt(V/W) and compare with 1, it should converge to 1 as we approach true distribution. If we don't approach it then we've messed up.

Parallel Tempering: run N chains targeting different versions of pi of 'increasing difficulty'.

Introduce **inverse temperatures** $0 < \gamma_1 < \gamma_2 < \ldots < \gamma_N = 1.$ **,** introduce tempered distributions π^{γ_n}, and N chains for each. For y ~ 0 , pi^y is considered easier to sample because the variations are smaller.

If we want modes of pi, we can just use high temperature chains. If we want to sample from pi, can use high temperature chains to improve mixing of the chain target pi.

Parallel tempering proposes swaps between the different temperature chains.

Joint chain targeting $\pi^{\gamma_1} \otimes \pi^{\gamma_2} \otimes \ldots \otimes \pi^{\gamma_N}$. Occasionally perform swap: Sample indices k_1, k_2 uniformly in $\{1, \ldots, N\}$. with acceptance probability $\min\left(1, \frac{\pi^{\gamma_{k_1}}(x_{k_2})\pi^{\gamma_{k_2}}(x_{k_1})}{\pi^{\gamma_{k_1}}(x_{k_1})\pi^{\gamma_{k_2}}(x_{k_2})}\right)$. exchange the value of x_{k_1} and x_{k_2}. . *The swap moves preserve detailed balance*, doesn't change joint target distribution, and the N-th chain still targets pi. Doesn't work well in high dimensions when different modes have different shapes

Hamiltonian Monte Carlo (HMC):

$\mu(q) \propto \exp(-U(q))$ U(x) is **potential energy.**

$\exp(-U(q) - K(p)) = \exp(-U(q)) \exp\left(-\frac{p^T M^{-1} p}{2}\right)$, decomposition shows **position q** and **momentum p** variables are independent, K(p) is **Kinetic Energy** $\frac{\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}}{2}$

Hamiltonian Energy need be constant $H(\mathbf{q}, \mathbf{p}) := \frac{\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}}{2} + U(\mathbf{q}) = K(\mathbf{p}) + U(\mathbf{q}).$,

as $\frac{d}{dt} H(\mathbf{q}(t), \mathbf{p}(t)) = \mathbf{p}(t)^T \mathbf{M}^{-1} \frac{d}{dt}\mathbf{p}(t) + \left(\frac{d}{dt}\mathbf{q}(t)\right)^T \nabla U(\mathbf{q}) = 0$

Rewriting:

$\frac{d}{dt}\mathbf{q} = +\nabla_{\mathbf{p}} H(\mathbf{q}, \mathbf{p}),$
$\frac{d}{dt}\mathbf{p} = -\nabla_{\mathbf{q}} H(\mathbf{q}, \mathbf{p}).$ t $\mathbf{z} := \begin{pmatrix} \mathbf{q} \\ \mathbf{p} \end{pmatrix}$ **J canonical structure matrix** $\mathbf{J} := \begin{pmatrix} \mathbf{0} & \mathbf{I}_d \\ -\mathbf{I}_d & \mathbf{0} \end{pmatrix}$

$$\frac{d}{dt}\mathbf{z} = \mathbf{J}\nabla_{\mathbf{z}} H(\mathbf{z}).$$

Canonical Hamiltonian Equations

Symplectic map: A smooth map $\boldsymbol{\Psi} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ s.t:

Jacobian $\nabla_z \boldsymbol{\Psi}(z)$ satisfies that $[\nabla \boldsymbol{\Psi}(\mathbf{z})]^T \mathbf{J}^{-1} \nabla \boldsymbol{\Psi}(\mathbf{z}) = \mathbf{J}^{-1}$

for all $\mathbf{z}$ in $\mathbb{R}^{2d}$, where $\mathbf{J} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_d \\ -\mathbf{I}_d & \mathbf{0} \end{pmatrix}$ is the canonical structure matrix. This satisfies that $\mathbf{J}^{-1} = \mathbf{J}^T$.

Proposition: Symplectic Maps are volume preserving (so det(J) = 1)

Proof: infinitesimal cube $[z_1, z_1 + \delta] \times \ldots \times [z_{2d}, z_{2d} + \delta]$ maps to thing, volume $\det(\nabla_{\mathbf{z}} \boldsymbol{\Psi}(\mathbf{z}))|\delta^{2d}.$

By product rule $\det(\nabla_z \boldsymbol{\Psi}(z))^2 \det(\mathbf{J}^{-1}) = \det(\mathbf{J}^{-1}),$ $|\det(\nabla_z \boldsymbol{\Psi}(z))| = 1$ since $|\det(\mathbf{J}^{-1})| = 1$, so infinitesimal volumes are preserved, same can be obtained for non-infinitesimal by integration

Proposition: Hamiltonian flow is symplectic (and volume preserving)

Proof: Step 1 Let $\Psi_{t,H}(\mathbf{z})$ denote the flow-map of the Hamiltonian dynamics with Hamiltonian $H(\mathbf{z})$, i.e. $\Psi_{t,H}(\mathbf{z}(0)) = \mathbf{z}(t)$.

We need to understand the behaviour of the Jacobian $\frac{\partial}{\partial \mathbf{z}}\Psi_{t,H}(\mathbf{z})$.

We are going to describe the evolution of this Jacobian in time

Let $\mathbf{z}(0)$ be an initial point, and $\bar{\mathbf{z}}(0) = \mathbf{z}(0) + \delta\mathbf{z}(0)$ be another nearby initial point.

The differences of two paths at time t can be written as

$$\delta\mathbf{z}(t) := \bar{\mathbf{z}}(t) - \mathbf{z}(t) = \Psi_{t,H}(\bar{\mathbf{z}}(0)) - \Psi_{t,H}(\mathbf{z}(0)) = \frac{\partial}{\partial \mathbf{z}(0)}\Psi_{t,H}(\mathbf{z}(0)) \cdot \delta\mathbf{z}(0) + o(\delta\mathbf{z}(0)).$$

Different in t $\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{\partial}{\partial \mathbf{z}(0)}\Psi_{t,H}(\mathbf{z}(0)) \cdot \delta\mathbf{z}(0)\right] = \frac{\partial}{\partial \mathbf{z}(0)}\left(\frac{d}{dt}\Psi_{t,H}(\mathbf{z}(0))\right) \cdot \delta\mathbf{z}(0) = \frac{\partial}{\partial \mathbf{z}(0)}(\mathbf{J}\nabla H(\mathbf{z}(t))) \cdot \delta\mathbf{z}(0)$

$$= \frac{\partial}{\partial \mathbf{z}(0)}(\mathbf{J}\nabla H(\Psi_{t,H}(\mathbf{z}(0)))) \cdot \delta\mathbf{z}(0) = \mathbf{J}\nabla^2 H(\mathbf{z}(t)) \cdot \left[\frac{\partial}{\partial \mathbf{z}(0)}\Psi_{t,H}(\mathbf{z}(0)) \cdot \delta\mathbf{z}(0)\right]$$

Step 2 $\mathbf{F}(t) := \frac{\partial}{\partial \mathbf{z}(0)}\Psi_{t,H}(\mathbf{z}(0))$ = Jaconbian at time t, as variation equations hold for every direction $\frac{d}{dt}\mathbf{F} = \mathbf{J}\nabla^2 H(\mathbf{z}(t))\mathbf{F}$, $\mathbf{F}(0) = \mathbf{I}_{2d}$. , enough then to verify $\mathbf{F}(t)^T\mathbf{J}^{-1}\mathbf{F}(t) = \mathbf{J}^{-1}$.

Step 3 $\mathbf{F}(0) = \mathbf{I}_{2d}$, it holds for $t = 0$.

$$\frac{d}{dt}(\mathbf{F}(t)^T\mathbf{J}^{-1}\mathbf{F}(t)) = (\mathbf{J}\nabla^2 H(\mathbf{z}(t))\mathbf{F}(t))^T\mathbf{J}^{-1}\mathbf{F}(t) + \mathbf{F}(t)^T\mathbf{J}^{-1}\mathbf{J}\nabla^2 H(\mathbf{z}(t))\mathbf{F}(t)$$

$$= \mathbf{F}(t)^T\nabla^2 H(\mathbf{z}(t))\mathbf{J}^T\mathbf{J}^{-1}\mathbf{F}(t) + \mathbf{F}(t)^T\mathbf{J}^{-1}\mathbf{J}\nabla^2 H(\mathbf{z}(t))\mathbf{F}(t) = \mathbf{0}, \text{ since } \mathbf{J}^T\mathbf{J}^{-1} = -\mathbf{J}\mathbf{J}^{-1} = -\mathbf{I}_{2d} \text{ and } \mathbf{J}^{-1}\mathbf{J} = \mathbf{I}_{2d}.$$

So Hamiltonian flow is symplectic, so volume preserving by previous proposition

Proposition *The distribution $\pi(d\mathbf{z}) \propto \exp(-H(\mathbf{z}))d\mathbf{z}$ is stationary with respect to the Hamiltonian flow.*

Proof Let the density of π be equal $\pi(\mathbf{z}) = \frac{\exp(-H(\mathbf{z}))}{C}$ for some normalising constant C.

Let $\mathbf{Z}_0 \sim \pi$, and $\mathbf{Z}_t = \Psi_{t,H}(\mathbf{Z}(0))$. Because $\Psi_{t,H}$ is 1-to-1 mapping

$$\mathbf{Z}_t \sim \pi_t, \quad \pi_t(z_t) = \pi(z_0)\left|\det F(t)\right|, \quad z_t = \Psi_{t,H}(z_0).$$

By volume and Hamiltonian preserving, $|\det F(t)| = 1$ and $e^{-H(z_0)} = e^{-H(z_t)}$. Then $\pi_t(z_t) = \pi(z_t)$ so $\mathbf{Z}_t \sim \pi$.

- Hamilton's equations themselves do not define an ergodic Markov chain, as they keep the Hamiltonian invariant.

Proposition *Let $\pi(\mathbf{q}) \propto \exp(-U(\mathbf{q}))$. Let $T \sim \nu_T$, with positive density on an interval $[0, \tau]$ for $\tau > 0$.*

Let K be the Markov kernel for the position variables corresponding to

1. *sampling a random momentum $\mathbf{p}$ from $\mathcal{N}(0, I_d)$, then*
2. *running the Hamiltonian dynamics started at $(\mathbf{q}, \mathbf{p})$ up to time T sampled from ν_T (independently at each step),*
3. *and finally discarding the momentum variable.*

Suppose that U is continuously differentiable on $\mathbb{R}^d$, and satisfies that $\sup_{\mathbf{q}} \|\nabla^2 U(\mathbf{q})\| \leq L$, and $\inf_{\mathbf{q}\in\mathbb{R}^d} U(\mathbf{q}) > -\infty$. Then K is strongly π-irreducible.

Discretizing Hamilton's equations:

Explicit Scheme: $\mathbf{p}(t+\epsilon) = \mathbf{p}(t) - \epsilon\nabla U(\mathbf{q}(t))$ $\mathbf{q}(t+\epsilon) = \mathbf{q}(t) + \epsilon\mathbf{M}^{-1}\mathbf{p}(t).$

Modified Explicit Scheme: $\mathbf{p}(t+\epsilon) = \mathbf{p}(t) - \epsilon\nabla U(\mathbf{q}(t))\,\mathbf{q}(t+\epsilon) = \mathbf{q}(t) + \epsilon\mathbf{M}^{-1}\mathbf{p}(t+\epsilon)$

Leapfrog (Stormer-Verlet Scheme): $\mathbf{p}(t+\epsilon/2) = \mathbf{p}(t) - \frac{\epsilon}{2}\nabla U(\mathbf{q}(t))$

$\mathbf{q}(t+\epsilon) = \mathbf{q}(t) + \epsilon\mathbf{M}^{-1}\mathbf{p}(t+\epsilon/2).$ $\mathbf{p}(t+\epsilon) = \mathbf{p}(t+\epsilon/2) - \frac{\epsilon}{2}\nabla U(\mathbf{q}(t+\epsilon))$

Note: as it is an approximation we don't really expect Hamiltonian energy to always be exactly the same, so we need to recheck the stable solution. Can try to stop it from running away by rejecting the next step if energy is too different

Proposition: The above 3 schemes are symplectic, and hence volume preserving

Proof:

- By the chain rule for Jacobians, if $\Psi(\mathbf{z}) = \Psi_1(\Psi_2(\mathbf{z}))$, then $\nabla_{\mathbf{z}}\Psi(\mathbf{z}) = \nabla\Psi_1(\Psi_2(\mathbf{z}))) \cdot \nabla\Psi_2(\mathbf{z}))$.
- Assuming Ψ_1 and Ψ_2 are symplectic, then $[\nabla\Psi(\mathbf{z})]^T\mathbf{J}^{-1}\nabla_{\mathbf{z}}\Psi(\mathbf{z}) =$

$$(\nabla\Psi_2(\mathbf{z}))^T(\nabla\Psi_1(\Psi_2(\mathbf{z})))^T\mathbf{J}^{-1}\nabla\Psi_1(\Psi_2(\mathbf{z}))\cdot\nabla\Psi_2(\mathbf{z})) = \mathbf{J}^{-1},$$

By induction this holds for more than two maps, so need to show single step in the schemes

- For example, the step $\mathbf{p}(t+\epsilon) = \mathbf{p}(t) - \epsilon\nabla U(\mathbf{q}(t))$ corresponds to the map $\Psi(\mathbf{z}) = \begin{pmatrix}\mathbf{q}\\ \mathbf{p} - \epsilon\nabla U(\mathbf{q})\end{pmatrix}$, which has Jacobian $\nabla\Psi(\mathbf{z}) = \begin{pmatrix}\mathbf{I}_d & \mathbf{0}\\ -\epsilon\nabla^2 U(\mathbf{q}) & \mathbf{I}_d\end{pmatrix}$
- By direct calculation, $(\nabla\Psi(\mathbf{z}))^T\mathbf{J}^{-1}\nabla\Psi(\mathbf{z})$

$$= \begin{pmatrix}\mathbf{I}_d & -\epsilon\nabla^2 U(\mathbf{q})\\ \mathbf{0} & \mathbf{I}_d\end{pmatrix}\begin{pmatrix}\mathbf{0} & -\mathbf{I}_d\\ \mathbf{I}_d & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{I}_d & \mathbf{0}\\ -\epsilon\nabla^2 U(\mathbf{q}) & \mathbf{I}_d\end{pmatrix} = \begin{pmatrix}\mathbf{0} & -\mathbf{I}_d\\ \mathbf{I}_d & \mathbf{0}\end{pmatrix} = \mathbf{J}^{-1}.$$

So symplecticity holds

Hamilton Monte Carlo

- Let $\Psi_\epsilon : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the Leapfrog map, and for some $L \in \mathbb{N}$, we denote by Ψ_ϵ^L the L times composition of the Leapfrog map.
- Let $\mathbf{N} : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ denote the map that negates the momentum, i.e. $\mathbf{N}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{q}\\ -\mathbf{p}\end{pmatrix}$.
- Let $\mathbf{P}_R$ be the Markov kernel that resamples the momentum component $\mathbf{p}$ from a multivariate Gaussian distribution with covariance matrix $\mathbf{M}$.

1. resample the momentum component $\boldsymbol{p}$, i.e. apply the Markov kernel $\boldsymbol{P}_R$.
2. We propose a new position $\begin{pmatrix}\boldsymbol{q}^*\\ \boldsymbol{p}^*\end{pmatrix} = \boldsymbol{\Psi}\begin{pmatrix}\boldsymbol{q}\\ \boldsymbol{p}\end{pmatrix} := \boldsymbol{N}\left(\boldsymbol{\Psi}_\epsilon^L\begin{pmatrix}\boldsymbol{q}\\ \boldsymbol{p}\end{pmatrix}\right)$ by applying L Leapfrog steps and then flipping the momentum. This new position is accepted with probability

$$\min\left[1, \exp(H(\boldsymbol{q},\boldsymbol{p}) - H(\boldsymbol{q}^*,p^*))\right] = \min\left[1, \exp\left(U(\boldsymbol{q}) - U(\boldsymbol{q}^*) + \frac{1}{2}\boldsymbol{p}^T\boldsymbol{M}^{-1}\boldsymbol{p} - \frac{1}{2}(\boldsymbol{p}^*)^T\boldsymbol{M}^{-1}\boldsymbol{p}^*\right)\right].$$

Proposition: pi is invariant wrt Markov kernel above, and is reversible wrt second step of the Markov Kernel

Proof:

- Since $\mathbf{q}$ and $\mathbf{p}$ are independent according to the target distribution π, it is clear that the first step keeps the target invariant.

Let $\mathbf{P}_2$ denote the Markov kernel corresponding to the second step (a combination of a deterministic step and Metropolis-Hastings accept-reject step), then we are going to check that π is reversible with respect to $\mathbf{P}_2$.

- Recall that in general a Markov kernel K is reversible with respect to a distribution π on state space $\mathbb{X}$ if for every bounded measurable function $f : \mathbb{X}^2 \to \mathbb{R}$, we have

$$\int \int f(\mathbf{x}, \mathbf{y})\pi(d\mathbf{x})K(\mathbf{x}, d\mathbf{y}) = \int \int f(\mathbf{x}, \mathbf{y})\pi(d\mathbf{y})K(\mathbf{y}, d\mathbf{x}).$$

$\mathbf{P}_2(\mathbf{x}, d\mathbf{y})$ is non-zero only for $\mathbf{y} = \Psi(\mathbf{x})$ and $\mathbf{y} = \mathbf{x}$.

$$\int \int f(\mathbf{x}, \mathbf{y})\pi(d\mathbf{x})\mathbf{P}_2(\mathbf{x}, d\mathbf{y}) = \int \int f(\mathbf{x}, \Psi(\mathbf{x})) \min[1, e^{H(\mathbf{x}) - H(\Psi(\mathbf{x}))}]\pi(d\mathbf{x}) + \int \int f(\mathbf{x}, \mathbf{x}) \left(1 - \min[1, e^{H(\mathbf{x}) - H(\Psi(\mathbf{x}))}]\right) \pi(d\mathbf{x})$$

- Let $\mathbf{y} = \Psi(\mathbf{x})$, then $\mathbf{x} = \Psi(\mathbf{y}) = \Psi(\Psi(\mathbf{x}))$ As volume preserving:

$$\pi(d\mathbf{y}) = \pi(d\mathbf{x}) \times \underbrace{|\det(\nabla\Psi)|}_{=1} \frac{\exp(-H(\mathbf{y}))}{\exp(-H(\mathbf{x}))} = \pi(d\mathbf{x}) \cdot e^{H(\mathbf{x}) - H(\mathbf{y})}$$

The first part of the previous sum can be written: $\int \int f(\mathbf{x}, \Psi(\mathbf{x})) \min[1, e^{H(\mathbf{x}) - H(\Psi(\mathbf{x}))}]\pi(d\mathbf{x})$

$$= \int \int f(\Psi(\mathbf{y}), \mathbf{y}) \min[1, e^{H(\Psi(\mathbf{y})) - H(\mathbf{y})}]\pi(d\mathbf{x}) = \int \int f(\Psi(\mathbf{y}), \mathbf{y}) \min[1, e^{H(\Psi(\mathbf{y})) - H(\mathbf{y})}] \cdot e^{H(\mathbf{y}) - H(\Psi(\mathbf{y}))}\pi(d\mathbf{y})$$

$$= \int \int f(\Psi(\mathbf{y}), \mathbf{y}) \min[1, e^{H(\mathbf{y}) - H(\Psi(\mathbf{y}))}]\pi(d\mathbf{y}).$$

The second part satisfies that

$$\int \int f(\mathbf{x}, \mathbf{x}) \left(1 - \min[1, e^{H(\mathbf{x}) - H(\Psi(\mathbf{x}))}]\right) \pi(d\mathbf{x}) = \int \int f(\mathbf{y}, \mathbf{y}) \left(1 - \min[1, e^{H(\mathbf{y}) - H(\Psi(\mathbf{y}))}]\right) \pi(d\mathbf{y}).$$

Combining $\int \int f(\mathbf{x}, \mathbf{y})\pi(d\mathbf{x})K(\mathbf{x}, d\mathbf{y}) = \int \int f(\mathbf{x}, \mathbf{y})\pi(d\mathbf{y})K(\mathbf{y}, d\mathbf{x})$

How to choose epsilon and L: new Markov kernel is a mixture of kernels where pi is stationary, so pi is still stationary for it. Using random period length means we can avoid issues that sometimes happen because the deterministic nature of Hamiltonian mechanics

Concentration in High Dimensions:
Suppose that Z ~N (0,Id) is a d dimensional standard normal random vector. Then the Euclidean norm of Z satisfies that for every t ≥0, $\Pr(|\|\mathbf{Z}\| - \sqrt{d}| \geq t) \leq C \exp\left(-\frac{t^2}{C}\right)$, C absolute constant

This means that with high probability, $\|\mathbf{Z}\| = \sqrt{d} + O(1)$, so most of the probability is concentrated around sphere of radius d^0.5, SImilar examples for other functions (other than Gaussian)

- In general, if the Hessian of the target potential satisfies that $\mu\mathbf{I}_d \preceq \nabla^2 U(\mathbf{x}) \preceq L\mathbf{I}_d$ for some $0 < \mu < L < \infty$ (strongly convex and smooth potential), and we let $H_{\min} := \inf_{\mathbf{z}} H(\mathbf{z})$, then it is possible to show that

$$\Pr\left(\left|\sqrt{H(\mathbf{z}) - H_{\min}} - \mathbb{E}\sqrt{H(\mathbf{z}) - H_{\min}}\right| \geq t\right) \leq C \exp\left(-\frac{t^2}{C}\right)$$

Hamiltonian is close to constant high probability density area, and HMC is very efficient in exploring this potentially complicated set automatically.

Optimal Scaling:

Random Walk Metropolis $\mathbf{q}^* = \mathbf{q} + \sigma W.$, **MALA** $\mathbf{q}^* = \mathbf{q} + \sigma W + \frac{\sigma^2}{2}\nabla \log \pi(\mathbf{q})$

HMC: $\mathbf{p}(t+\epsilon/2) = \mathbf{p}(t) - \frac{\epsilon}{2}\nabla U(\mathbf{q}(t))$ $\mathbf{q}^* = \mathbf{q} + \epsilon \mathbf{M}^{-1}\mathbf{p}(t+\epsilon/2).$

$\mathbf{p}(t+\epsilon) = \mathbf{p}(t+\epsilon/2) - \frac{\epsilon}{2}\nabla U(\mathbf{q}(t+\epsilon))$ If $\mathbf{M} = \mathbf{I}$ and $\epsilon = \sigma$ then one step HMC reduces to MALA.

Maximizing the expected squared jumping distance (ESJD): $\mathbb{E}[\|X^{(t+1)} - X^{(t)}\|^2]$

E.g. Squared jumping distance

$$\|X^{(t+1)} - X^{(t)}\|^2 = \begin{cases} \|\mathbf{q}^* - \mathbf{q}\|^2, & \text{w.p. } \left(1 \wedge \frac{\pi(\mathbf{q}^*)p(\mathbf{q}|\mathbf{q}^*)}{\pi(\mathbf{q})p(\mathbf{q}^*|\mathbf{q})}\right) \\ 0, & \text{otherwise.} \end{cases}$$

Dimension dependence on step size :

- RWM: $\mathcal{O}(d^{-1})$
- MALA: $\mathcal{O}(d^{-1/3})$
- HMC: $\mathcal{O}(d^{-1/4})$

Optimal Acceptance Rates:

- RWM: 0.234 (see Roberts, et. al., 1997)
- MALA: 0.574 (see Roberts and Rosenthal, 1998)
- HMC: 0.651 (see Beskos, et. al., 2013)

Hamilton Monte Carlo basically non examinable

Hidden Markov Models:

Time Series: models st $Y_t \mid y_{1:t-1} \sim p(dy_t \mid y_{1:t-1}, \theta)$, **,** def $y_{k:l} = (y_k, \ldots, y_l)$

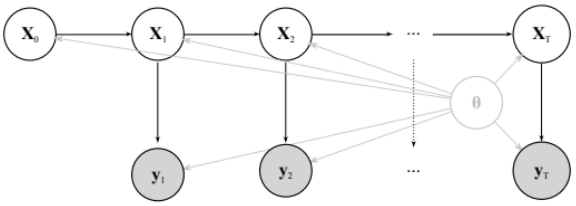
Likelihood: $\forall \theta \in \Theta \quad p(y_{1:t} \mid \theta) = p(y_1 \mid \theta) \prod_{k=2}^{t} p(y_k \mid y_{1:k-1}, \theta).$,

can put prior on theta and consider sampling from posterior

HMMs assume there is a hidden/unobsvered X valued Markov Process Xt, with $X_0 \mid \theta \sim p(dx_0 \mid \theta)$, it is time homogenous so that $X_t \mid x_{t-1}, \theta \sim p(dx_t \mid x_{t-1}, \theta)$, so

$$p(x_{0:t} \mid \theta) = p(x_0 \mid \theta) \prod_{k=1}^{t} p(x_k \mid x_{1:k-1}, \theta) = p(x_0 \mid \theta) \prod_{k=1}^{t} p(x_k \mid x_{k-1}, \theta)$$

Observations are dependent on our unobserved Markov process Xt, *and conditional dependent on it* $Y_t \mid x_t, \theta \sim p(dy_t \mid x_t, \theta)$ i.e. the distribution of Y_t is independent of $(X_k)_{k \neq t}$ conditional upon $X_t = x_t$.



With a prior $p(\theta)$ on the parameters, the joint density is:

$$p(x_{0:T}, y_{1:T}, \theta) = p(\theta)p(x_0 \mid \theta) \prod_{t=1}^{T} p(x_t \mid x_{t-1}, \theta) \prod_{t=1}^{T} p(y_t \mid x_t, \theta).$$

$\Rightarrow$ $p(y_{1:t} \mid x_{1:t}, \theta) = \prod_{k=1}^{t} p(y_k \mid x_k, \theta)$ (cond. independent)

Major goal 1: $p(x_{0:T}, y_{1:T}, \theta)$ or p(xt|y1:t, theta) (the **filtering distribution**)

Major goal 2: do **parameter estimation**

Linear Gaussian Models: $\mathbb{X} = \mathbb{R}^{d_x}$ and $\mathbb{Y} = \mathbb{R}^{d_y}$, $X_t = AX_{t-1} + \varepsilon_t \quad \varepsilon \sim \mathcal{N}(0, \Sigma_x)$, need not be diagonal covariance, $Y_t = CX_t + \eta_t$, $\eta_t \sim \mathcal{N}(0, \Sigma_y)$, , can get distributions X1:t anc Y1:t by **Kalman Recursions,** and likelihood parameters can be evaluated exactly

General Case: $$p(x_{0:t}|y_{1:t},\theta) = \frac{p(x_{0:t},y_{1:t}\mid\theta)}{p(y_{1:t}\mid\theta)} = \frac{p(x_{0:t}\mid\theta)\,p(y_{1:t}|x_{0:t},\theta)}{p(y_{1:t}\mid\theta)}$$

$$p(x_{0:t}\mid\theta) = p(x_0\mid\theta)\prod_{k=1}^{t} p(x_k|x_{k-1},\theta) \quad p(y_{1:t}\mid x_{0:t},\theta) = \prod_{k=1}^{t} p(y_k\mid x_k,\theta)$$

Marginal Likelihood $$p(y_{1:t}\mid\theta) = \int_{\mathbb{X}^t} p(x_{0:t},y_{1:t}\mid\theta)\,dx_{0:t}.$$ highD, hard to get

Propositlon: *The posterior* $p(x_{0:t}|y_{1:t},\theta)$ *satisfies*

$$p(x_{0:t}|y_{1:t},\theta) = p(x_{0:t-1}|y_{1:t-1},\theta)\frac{p(x_t|x_{t-1},\theta)\,p(y_t|x_t,\theta)}{p(y_t|y_{1:t-1},\theta)}$$ (where dropping theta)

$$p(y_t|y_{1:t-1}) = \int p(x_{0:t-1}|y_{1:t-1},)\,p(x_t|x_{t-1})\,p(y_t|x_t)\,dx_{0:t}.$$

Proof

$$p(x_{0:t}|y_{1:t}) = p(x_{0:t-1},y_{1:t-1})\,p(x_t|x_{t-1})\,p(y_t|x_t)$$

$$p(y_{1:t}) = p(y_{1:t-1})\,p(y_t|y_{1:t-1}), \quad \text{and so}$$

$$p(x_{0:t}|y_{1:t}) = \frac{p(x_{0:t},y_{1:t})}{p(y_{1:t})} = \underbrace{\frac{p(x_{0:t-1},y_{1:t-1})}{p(y_{1:t-1})}}_{p(x_{0:t-1}|y_{1:t-1})}\frac{p(x_t|x_{t-1})\,p(y_t|x_t)}{p(y_t|y_{1:t-1})}$$ and the expression for $p(y_t|y_{1:t-1})$ follows

In general, this means filtering problem is intractable

$$\int \varphi(x_t)p(x_t\mid y_{1:t},\theta)dx_t = \int \varphi(x_t)p(x_{0t}\mid y_{1:t},\theta)dx_{0:t} = \frac{\int \varphi(x_t)p(x_0\mid\theta)\prod_{k=1}^{t}p(x_k\mid x_{k-1},\theta)\prod_{k=1}^{t}p(y_k\mid x_k,\theta)dx_{0:t}}{p(y_{1:t}|\theta)}$$

Numerator is (t x dim(X)) high dimensional integral, denominator is the marginal likelihood which is also intractable. This *double intractability* means we can't even compute it pointwise for metropolis hastings

$$p(y_{1:t}\mid\theta) = \int p(x_{0:t},y_{1:t}\mid\theta)dx_{0:t} = \int p(x_0\mid\theta)\prod_{k=1}^{t}p(x_k\mid x_{k-1},\theta)\prod_{k=1}^{t}p(y_k\mid x_k,\theta)dx_{0:t}$$

Historical approach consists of Gibbs Sampling joint space of (theta, X0:t), alternative between sample from $\theta \mid x_{0:t},y_{1:t},$ with conditional distribution

$$p(\theta\mid x_{0:t},y_{1:t}) \propto p(\theta)p(x_{0:t},y_{1:t}\mid\theta) = p(\theta)p(x_0\mid\theta)\prod_{k=1}^{t}p(x_k\mid x_{k-1},\theta)\prod_{k=1}^{t}p(y_k\mid x_k,\theta)$$

And sampling from $p(x_{0:t}\mid y_{1:t},\theta)$ by iteratively sampling xk given x_{k-1},y_k,x_{k+1} and θ.

$$p(x_k \mid x_{-k}, y_{1:t}, \theta) = p(x_k \mid x_{k-1}, y_k, x_{k+1}, \theta) \propto p(x_k \mid x_{k-1}, \theta) p(y_k, x_{k+1} \mid x_k, \theta)$$
$$= p(x_k \mid x_{k-1}, \theta) p(x_{k+1} \mid x_k, \theta) p(y_k \mid x_k, \theta)$$

Sometimes we can evaluate this pointwise, and then use MH to update components. As the X0:t are highly correlated Gibbs will fail

Objects of interest:

Various by-products of the joint posterior $p(\theta, x_{0:t} \mid y_{1:t})$:

- prediction under parameter uncertainty through $p(x_{t+1} \mid y_{1:t})$,

$$p(x_{t+1} \mid y_{1:t}) = \int_\Theta \int_{\mathcal{X}^{t+1}} \underbrace{p(x_{t+1} \mid x_t, \theta)}_{\text{transition}} \underbrace{p(d\theta, dx_{0:t} \mid y_{1:t})}_{\text{joint posterior}}.$$

- predictive checking through $\mathbb{P}(Y_{t+1} \leq y_{t+1} \mid y_{1:t})$,

$$\mathbb{P}(Y_{t+1} < y_{t+1} \mid y_{1:t}) = \int_\Theta \int_{\mathcal{X}} \int_{-\infty}^{y_{t+1}} p(dy \mid x_{t+1}, \theta) p(dx_{t+1}, d\theta \mid y_{1:t}).$$

- sequential model comparison through $p(y_{1:t})$.

$$p(y_{1:t}) = \int_\Theta \int_{\mathcal{X}^{t+1}} p(d\theta, dx_{0:t}, y_{1:t})$$

Importance Sampling: Know $I = \int \varphi(X) d\pi(X),$, approximate

$\hat{I}^{NIS} = \sum_{i=1}^N W^i \varphi(X^i), W^i = \frac{w(X^i)}{\sum_{i=1}^N w(X^i)}$ and approximate $\pi(x) \approx \pi^N(x) = \sum_{i=1}^N W^i \delta_{X^i}(x)$

Where X^i are sampled from q and $w(x) = \gamma(x)/q(x)$. , know q isn't write density, just a proposal we can draw from

Sequential Importance Sampling: Key observation that **we can write w(X) as recursion**

$$w(X_{1:t}) = \frac{\gamma(X_{1:t})}{q(X_{1:t})} = \frac{\gamma(X_t|X_{1:t-1})\gamma(X_{1:t-1})}{q(X_t|X_{1:t-1})q(X_{1:t-1})} \text{ Call } w_t = w(X_{1:t}) \text{ and } \omega_t = \frac{\gamma(X_t|X_{1:t-1})}{q(X_t|X_{1:t-1})}.$$

Then $w_t = w_{t-1} \times \omega_t$,

Need 2 additional assumptions: That q, as well as pi, is markovian. And that pi is the

$$q_1(x_1),\ q_{t|t-1}(X_t|X_{t-1}),\ \text{then } \omega_t = \frac{\gamma(X_t|X_{t-1})}{q_{t|t-1}(X_t|X_{t-1})}$$

probability of interest' Moreover, we focus on $\pi(X_{1:t}) = p(X_{1:t}|Y_{1:t})$

Sampling Algorithm for SIS:

<u>At time t =1:</u> • Sample $X_1^i \sim q_1(\cdot)$. **Compute** $w_1^i = \frac{\mu(X_1^i) g(y_1 \mid X_1^i)}{q_1(X_1^i)}.$

<u>At time t>1:</u> • Sample $X_t^i \sim q_{t|t-1}(\cdot \mid X_{t-1}^i)$. **compute** $w_t^i = w_{t-1}^i \times \omega_t^i$

$$= w_{t-1}^i \times \frac{f\left(X_t^i \mid X_{t-1}^i\right) g\left(y_t \mid X_t^i\right)}{q_{t|t-1}(X_t^i \mid X_{t-1}^i)}.$$

New Xi is from Markov chain of q, wt is , the top is out gamma

Outcome: get sequence $X^i = X_{1:t}^i$ and weights $w^i = w_t^i = w(X_{1:t}^i)$ so

$\pi^N(x) = \sum_{i=1}^N W^i \times \delta_{X^i}(x), \quad W^i = \frac{w^i}{\sum_{i=1}^N w^j}$ is an approximation of pi^N. For example,

$$p(y_t|y_{1:t-1}) = \int f(x_t|x_{t-1})g(y_t|x_t)p(x_{1:t-1}|y_{1:t-1})dx_{1:t-1}dx_t$$
$$= \int \omega_t(x)p(x_{1:t-1}|y_{1:t-1})q_{t|t-1}(x_t|x_{t-1})dx_{1:t-1}dx_t$$
$$\approx p^N(y_t|y_{1:t-1}) := \sum_{i=1}^{N} \omega_t(X^i) \times W_{t-1}^i$$

and $p(y_{1:t}) \approx p^N(y_{1:t}) := p^N(y_1)\prod_{k=2}^{t} p^N(y_k|y_{1:k-1})$, see this by integrating p(x1:t,y1:t)d(x1:t), using the recursion formula given before

Prior Proposal: $q_1(x_1) = \mu(x_1),$ $q_{t|t-1}(x_t \mid x_{t-1}) = f(x_t \mid x_{t-1}).$ $\Rightarrow$ $\omega(x_{t-1}, x_t) = g(y_t \mid x_t)$

Proposal blindly propogates without considering yi, We can implement SIS as soon as we can sample from the hidden process (Xt)t≥1 and evaluate g(y |x) pointwise.

Optimal Proposals: Choose proposal that minimises variance of $(\omega_t^i)_{i=1}^N$

$q_{t|t-1}^{\text{opt}}(x_t|x_{t-1}) = p(x_t|x_{t-1}, y_t) = \frac{f(x_t|x_{t-1})\, g(y_t|x_t)}{p(y_t|x_{t-1})}$ uses observation yt to guide propagation of xt, $\Rightarrow$ $\omega_t^{\text{opt}}(x_{t-1}, x_t) = p(y_t|x_{t-1})$, indep of xt

(proof of optimality in slides) <u>problem:</u> Can't always implement it, e.g. get omega with no closed-form formulae. *Solve via gaussian approximation*

Curse of Dimensionality: SIS as a special type of IS is still cursed by dimensionality

Resampling: motivation: propagating particles with zero weight wastes computational resources,
so at time t, select particle with high weights and remove particles with low weights, spending computational budget on most promising parts

Obtain equally weighted sample $(N^{-1}, \bar{X}^i)$ from a weighted sample (w^i, X^i). resample on empirical measure with input $\pi^N(x) = \left(\sum w^j\right)^{-1} \sum w^i \delta_{X^i}(x)$ output

$\bar{\pi}^N(x) = N^{-1} \sum \delta_{\bar{X}^i}(x).$

Multinomial resampling:
Draw ancestry vector
$A^{1:N} = \left(A^1, \ldots, A^N\right) \in \{1, \ldots, N\}^N$ independently from a categorical distribution $A^{1:N} \overset{\text{i.i.d}}{\sim} \text{Cat}\left(w^1, \ldots, w^N\right)$,

Define $\bar{X}^i$ to be X^{A^i} for all $i \in \{1, \ldots, N\}$. X^{A^i} is said to be the "parent" or "ancestor" of $\bar{X}^i$.

Return $\bar{X} = \left(\bar{X}^1, \ldots, \bar{X}^N\right)$

<u>Forwards version:</u> Draw 'offspring vector' $O^{1:N} = \left(O^1, \ldots, O^N\right) \in \{0, \ldots, N\}^N$

$O^{1:N} \sim \text{Multinomial}\left(N; w^1, \ldots, w^N\right)$ s.t. $\forall i \in \{1, \ldots, N\}$ $\mathbb{E}[O^i] = N\frac{w^i}{\sum_{j=1}^N w^j}$ and $\sum_{i=1}^{N} O^i = N.$

Each particle X^i is replicated O^i times , to create sample $\bar{X}$:
Return $\bar{X} = \left(\bar{X}^1, \ldots, \bar{X}^N\right)$

Assumed Requirements $\mathbb{E}[O^i] = N\frac{w^i}{\sum_{j=1}^N w^j}$, or $\mathbb{P}[A^i = k] = \frac{w^k}{\sum_{j=1}^N w^j}$, sometimes called unbiasedness as ⇒ $\frac{1}{N}\sum_{k=1}^N \varphi(\bar{X}^k) = \frac{1}{N}\sum_{k=1}^N O^k \varphi(X^k)$ has expectation $\sum_{k=1}^N \frac{w^k}{\sum_{j=1}^N w^j}\varphi(X^k)$

Therefore, $\bar{\pi}^N$, on average, behaves as the original random measure $\pi^N(x) = \sum W^i \times \delta_{X^i}(x)$. It has more variance!

Sequential Monte Carlo Algorithm

- *At time* $t = 1$ • Sample $X_1^i \sim q_1(\cdot)$ • Compute the weights

$$w_1^i = \frac{\mu(X_1^i)g(y_1 \mid X_1^i)}{q_1(X_1^i)}.$$

- *At time* $t \geq 2$ • Resample $\left(w_{t-1}^i, X_{1:t-1}^i\right) \to \left(N^{-1}, \overline{X}_{1:t-1}^i\right)$
- Sample $X_t^i \sim q_{t|t-1}(\cdot \mid \bar{X}_{t-1}^i)$, $X_{1:t}^i := \left(\bar{X}_{1:t-1}^i, X_t^i\right)$ • Compute the weights

$$w_t^i = \omega_t^i = \frac{f\left(X_t^i \mid X_{t-1}^i\right) g\left(y_t \mid X_t^i\right)}{q_{t|t-1}(X_t^i \mid X_{t-1}^i)}.$$

The resampling breaks the recursion we had in SIS

The cost: **Path Degeneracy:** Sometimes one ancestor leads to more than one offspring, *decreasing diversity*.

Particle approximation of filtering $p(x_t \mid y_{1:t}, \theta)$: $\frac{1}{\sum_{j=1}^N w_t^j}\sum_{i=1}^N w_t^i \delta_{X_t^i}(dx_t)$

After resampling $\frac{1}{N}\sum_{i=1}^N \delta_{\bar{X}_t^i}(dx_t)$ Particle approximation of path filtering $p(x_{1:t} \mid y_{1:t}, \theta)$:

$\frac{1}{\sum_{j=1}^N w_t^j}\sum_{i=1}^N w_t^i \delta_{X_{1:t}^i}(dx_{1:t})$, • Particle filters approximate well $p(x_t \mid y_{1:t})$ but not $p(x_s \mid y_{1:t})$ for $s \ll t$.

Usually we only care about (pxt|y1:t), so not too bad

Likelihood Estimation: At time 1 (using sample approximation)

$$p^N(y_1) = \frac{1}{N}\sum_{i=1}^N \omega_1^i \xrightarrow[N\to\infty]{\text{a.s.}} \int \frac{\mu(x_1)g(y_1 \mid x_1)}{q_1(x_1)} q_1(x_1)\,dx_1 = p(y_1).$$

At time t (using after sampling approximation) $p^N(y_t \mid y_{1:t-1}) = \frac{1}{N}\sum_{i=1}^N \omega_t^i$

$$\xrightarrow[N\to\infty]{\text{a.s.}} \int \omega(x_{t-1}, x_t) q_{t|t-1}(x_t \mid x_{t-1}) p(x_{t-1} \mid y_{1:t-1}) dx_{t-1:t} = p(y_t \mid y_{1:t-1})$$

where $\omega(x_{t-1}, x_t) = (f(x_t \mid x_{t-1})g(y_t \mid x_t))/(q_{t|t-1}(x_t \mid x_{t-1}))$

Leads to estimator $p^N(y_{1:t}) = p^N(y_1)\prod_{s=2}^{t} p^N(y_s \mid y_{1:s-1}) = \prod_{s=1}^{t}\frac{1}{N}\sum_{i=1}^{N}\omega_s^i \xrightarrow[N\to\infty]{a.s.} p(y_{1:t}).$

Surprisingly unbiased $\mathbb{E}\left[p^N(y_{1:t})\right] = p(y_{1:t}),$ whereas for t>=2 $\mathbb{E}\left[p^N(y_t \mid y_{1:t-1})\right] \neq p(y_t \mid y_{1:t-1}).$

(i.e all of the parts are biased but the whole thing isn't)

Theoretical Results: Consistency as N →∞ is simple to prove, as each step (propagation, weighting, resampling) is itself consistent

Convergence results include Central Limit Theorems and non-asymptotic results.

Consider $I(\varphi_t) = \int \varphi_t(x_{1:t})p(x_{1:t} \mid y_{1:t})dx_{1:t}.$

- L_p-bound on the path space: $\mathbb{E}\left[\left|I^N(\varphi_t) - I(\varphi_t)\right|^p\right]^{1/p} \leq \frac{B(t)c(p)\,\|\varphi_t\|_\infty}{\sqrt{N}},$

$\sqrt{N}\left(I^N(\varphi_t) - I(\varphi_t)\right) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma_t^2\right)$ As expected, $B(t)$ and σ_t^2 typically grow exponentially fast with t. This is the path degeneracy problem.

Consider instead $I(\varphi_t) = \int \varphi_t(x_t)p(x_t \mid y_{1:t})dx_t.$ L_p-bound:

$\mathbb{E}\left[\left|I^N(\varphi_t) - I(\varphi_t)\right|^p\right]^{1/p} \leq \frac{B_1 c(p)\,\|\varphi_t\|_\infty}{\sqrt{N}}$ $\sqrt{N}\left(I^N(\varphi_t) - I(\varphi_t)\right) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma_t^2\right)$

For filtering estimates, error is independent of time t: $\sigma_t^2 < \sigma_{\max}^2$ for all t, and B_1 independent of t.

Consider estimator of marginal likelihood $p^N(y_{1:t}) = \prod_{s=1}^{t}\frac{1}{N}\sum_{i=1}^{N} w_s^i.$

Unbiasedness $\mathbb{E}\left[p^N(y_{1:t})\right] = p(y_{1:t}).$ Non-asymptotic relative variance $\mathbb{E}\left(\left(\frac{p^N(y_{1:t})}{p(y_{1:t})} - 1\right)^2\right) \leq \frac{B_3 t}{N}.$

Choose $N = \mathcal{O}(t)$ to control the relative variance.

Curse of dimensionality: Still cursed by dimensionality

Complexity: Propagating and weighting the particles is $\mathcal{O}(N)$

Multinomial resampling is $\mathcal{O}(N)$ if the uniforms are generated in sorted order.

The memory cost is $\mathcal{O}(N)$ if only the latest particles are stored.

The memory cost is at most $\mathcal{O}(Nt)$ if the paths are stored

Sequential Monte Carlo in Generic Models:

$$\pi_T(\theta) \propto \pi_0(\theta)\prod_{s=1}^{T} p(y_s \mid \theta).$$

Want to sample from , *not imposing any structure on theta here* Idea: sample from Pi0 and apply SIS-like recursion

If $w_t^i = \frac{\pi_t(\theta^i)}{\pi_0(\theta^i)}$ then $w_t^i = \frac{\pi_{t-1}(\theta^i)}{\pi_0(\theta^i)}\frac{\pi_t(\theta^i)}{\pi_{t-1}(\theta^i)} = w_{t-1}^i\omega_t^i, \quad \text{with } \omega_t^i = \frac{\pi_t(\theta^i)}{\pi_{t-1}(\theta^i)}.$

Note all these methods work even when we haven't normalised weights. Incremental weights satisfy $\omega^i \propto \frac{\pi_t(\theta)}{\pi_{t-1}(\mu)} \propto p(y_t \mid \theta)$

Algorithm: • Here X corresponds exactly to θ. No time structure.

- *At time* $t = 0$ • Sample $X^i \sim \pi_0(\cdot)$ for $i \in \{1, \ldots, N\}$ $\forall i \in \{1, \ldots, N\}$ $w_0^i = \frac{1}{N}$.
- *At time* $t \geq 1$ $w_t^i = w_{t-1}^i \times \omega_t^i = w_{t-1}^i \times \frac{\pi_t(X^i)}{\pi_{t-1}(X^i)}$ At all times, $(w_t^i, X^i)_{i=1}^N$ approximates π_t.

Not sampling again at each step:, so exacerbate problems of path degeneracy, ending up with very few unique values

SIS with resampling for posterior inference:

- *At time* $t = 0$ Sample $X_0^i \sim \pi_0(\cdot)$ for $i \in \{1, \ldots, N\}$ $\forall i \in \{1, \ldots, N\}$ $w_0^i = \frac{1}{N}$.
- *At time* $t \geq 1$ Resample $\left(w_{t-1}^i, X_{t-1}^i\right) \to \left(N^{-1}, \overline{X}_{t-1}^i\right)$ Define $X_t^i = \overline{X}_{t-1}^i$ $w_t^i = \omega_t^i = \frac{\pi_t(X_t^i)}{\pi_{t-1}(X_t^i)}$.

Suppose we want to use MH for p(theta|y1:t), need to compute acceptance ratio which needs p(y1:t,theta), but that's hard to compute. **BUT** can get unbiased estimators of p(y1:t,theta), P^N(y1:t,theta) based on Sequential Monte Carlo (SMC)

Pseudo Marginal Metropolis Hasting: Uses estimates of $\tilde{\pi}(x)$

- Starting with $X^{(1)}$, and $Z^{(1)} \geq 0$ such that $\mathbb{E}(Z^{(1)}) = \tilde{\pi}(X^{(1)})$, for $t = 2, 3, \ldots$

1. Sample $X' \sim q\left(\cdot \mid X^{(t-1)}\right)$ 2. Estimate $\tilde{\pi}(X')$ by Z', such that $\mathbb{E}(Z') = \tilde{\pi}(X')$.

3. Compute $$\alpha\left(X' \mid X^{(t-1)}\right) = \min\left(1, \frac{Z' \, q\left(X^{(t-1)} \mid X'\right)}{Z^{(t-1)} \, q\left(X' \mid X^{(t-1)}\right)}\right)$$

4. Sample $U \sim \mathcal{U}_{[0,1]}$. If $U \leq \alpha\left(X' \mid X^{(t-1)}\right)$, set $(X^{(t)}, Z^{(t)}) = (X', Z')$, otherwise set $(X^{(t)}, Z^{(t)}) = (X^{(t-1)}, Z^{(t-1)})$.

Exact limiting law

- For any x, denote by Z_x a non-negative unbiased estimator of $\tilde{\pi}(x)$, with distribution $g(\cdot \mid x) \equiv g_x$.
- If $\mathbb{V}_{g(\cdot|x)}(Z_x/\tilde{\pi}(x)) \ll 1$, then the algorithm $\approx$ original Metropolis–Hastings.
- Thus the generated chain $(X^{(t)})_{t \geq 1}$ goes to $\approx \pi$.
- In fact, the limiting law of $(X^{(t)})_{t \geq 0}$ is exactly $\pi \ldots !$

For general: introduce extended target distribution $\bar{\pi}(x, z) \propto z \times g_x(z)$

Introduce proposal $\bar{q}\left(x', z' \mid x, z\right) = q(x' \mid x) g_{x'}(z')$. ⇒ MH acceptance ratio is

$$\min\left(1, \frac{\bar{\pi}(x', z')}{\bar{\pi}(x, z)} \frac{\bar{q}\left(x, z \mid x', z'\right)}{\bar{q}\left(x', z' \mid x, z\right)}\right) = \min\left(1, \frac{z'}{z} \frac{q(x \mid x')}{q(x' \mid x)}\right)$$ **This is the algorithm above**

Is standard MH targeting $\bar{\pi}(x, z)$. Distribution of X if X,Z following $\bar{\pi}(x, z)$. :

Integrating X out $\bar{\pi}(x) \propto \int z g_x(z) dz = \mathbb{E}_{g_x}[Z_x] = \tilde{\pi}(x)$, so marginal is pi

- Thus if the Markov chain $(X^{(t)}, Z^{(t)})_{t\geq 0}$ converges to $\bar{\pi}$, then the first component $(X^{(t)})$ converges to the first marginal of $\bar{\pi}$, which is π.

- Therefore pseudo-marginal Metropolis–Hastings is *exact*.

Metropolis Hastings ratio for PMMH algorithm

- Recall that we have

$p(\theta, x_{1:t}|y_{1:t}) = \frac{1}{p(y_{1:t})} p(\theta) p(x_{1:t}|\theta) p(y_{1:t}|x_{1:t}, \theta)$ where

$p(y_{1:t}) = \int p(\theta) p(y_{1:t}|\theta) d\theta$ and $p(y_{1:t}|\theta) = \int p(y_{1:t}|x_{1:t}, \theta) p(x_{1:t}|\theta) dx_{1:t}$

- Therefore, $p(\theta|y_{1:t}) = \frac{p(\theta)p(y_{1:t}|\theta)}{p(y_{1:t})}$ and $p(y_{1:t}|\theta) \approx p^N(y_{1:t}|\theta)$

- Here $z = p(y_{1:t}|\theta)$ is approximated by $p^N(y_{1:t}|\theta)$ where the randomness is induced by the sampled paths $X_{1:t}$ (targeting the posterior).

- Parameters are drawn from the Kernel $q(\theta'|\theta)$ and $X'_{1:t}$ from a suitable θ'-SMC sampler, the M-H ratio is $r = \frac{p(\theta')p^N_{x'}(y_{1:t}|\theta')q(\theta|\theta')}{p(\theta)p^N_x(y_{1:t}|\theta)q(\theta'|\theta)}$

Gibbs Samplers for HMM: not great

- Sampling from $p(x_{0:t}|y_{1:t}, \theta)$ can be done by iteratively sampling x_k given x_{k-1}, y_k, x_{k+1} and θ.

$$p(x_k | x_{-k}, y_{1:t}, \theta) = p(x_k | x_{k-1}, y_k, x_{k+1}, \theta) \propto p(x_k | x_{k-1}, \theta) p(y_k, x_{k+1} | x_k, \theta)$$
$$= p(x_k | x_{k-1}, \theta) p(x_{k+1} | x_k, \theta) p(y_k | x_k, \theta)$$

- In which case, we could use Metropolis–Hastings to update each component of $X_{0:t}$ given the others.
- By definition, the components of $X_{0:t}$ are highly correlated, thus this Gibbs sampling approach will fail (remember the bivariate normal!).

"Idealised" Gibbs Sampler

- To sample from $p(\theta, x_{1:T}|y_{1:T})$, an MCMC strategy consists of using the following block Gibbs sampler.

At iteration $i = 0$

- Sample $\theta(0)$, $X_{1:T}(0)$ and its ancestral lineage $B_{1:T}(0)$ arbitrarily.

At iteration $i \geq 1$

- Sample $X_{1:T}(i) \sim p_{\theta(i-1)}(x_{1:T}|y_{1:T})$.
- Sample $\theta(i) \sim p(\theta|y_{1:T}, X_{1:T}(i))$

- **Problem**: We do not know how to sample from $p_\theta(x_{1:T}|y_{1:T})$
- Naive particle approximation where $X_{1:T}(i) \sim \hat{p}(x_{1:T}|y_{1:T}, \theta(i))$ is substituted to $X_{1:T}(i) \sim p(x_{1:T}|y_{1:T}, \theta(i))$ is obviously incorrect.

Particle Gibbs Sampler:

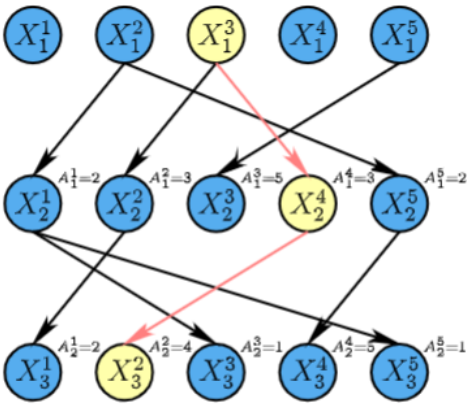
At iteration $i = 0$ • Sample $\theta(0)$, $X_{1:T}(0)$ and its ancestral lineage $B_{1:T}(0)$ arbitrarily.

At iteration $i \geq 0$ • Sample $\theta(i) \sim p(\theta | y_{1:T}, X_{1:T}(i-1))$.

- Run a conditional SMC algorithm for $\theta(i)$ consistent with the $X_{1:T}(i-1)$ and its ancestral lineage $B_{1:T}(i-1)$.
- Re-sample the trajectories at T and define uniformly an index in $k \in 1:N$. Then, pull a trajectory from that index X_T^k and going back it time through its genealogy B_t to define $X_{1:T}(i)$.

- **Proposition**. Assume that the 'ideal' Gibbs sampler chain is ergodic then under very weak assumptions the particle Gibbs sampler chain is ergodic and admits $p(\theta, x_{1:T} | y_{1:T})$ as an invariant distribution **for any** $N \geq 2$.

Conditional SMC



•In our case we use SMC to approximate $\hat{\pi}(X|\theta)$ and then update θ with a MH move. As we pull a trajectory we go from approximation to 'true' distribution.

- Suppose we have an approximation for the posterior over (θ, X), $\hat{\pi}(\theta, X)$ based on n particles X^i.
- We can conceive an enlarged space with an index k and define $\hat{\pi}(\theta, X, k)$ as selecting one of the particles.
- If K is random It can be shown that (θ, X^K) distributes according to $\pi(\theta, X)$.