Advanced Simulation - Lecture 5

George Deligiannidis

January 28th, 2019

Normalised Importance Sampling

Standard IS has limited applications in statistics as it requires knowing $\pi(x)$ and $q(x)$ exactly.

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

$$\widetilde{w}(x) = \frac{\widetilde{\pi}(x)}{\widetilde{q}(x)}.$$

An alternative identity is

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

SLLN for NIS

Proposition (SLLN for NIS)

Let $X_1, ..., X_n \overset{i.i.d.}{\sim} q$ *and assume that* $\mathbb{E}_q(|\varphi(X)|\, w(X)) < \infty$*.*
Then

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

is strongly consistent.

Proof.

Divide numerator and denominator by n. Both converge almost surely by the strong law of large numbers. $\square$

BUT, for finite n $\widehat{I}_n^{\text{NIS}}$ is **biased**, see notes Chapter 3.

CLT for NIS

Proposition

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

$$\sqrt{n}(\widehat{I}_n^{NIS} - I) \Rightarrow \mathcal{N}(0, \sigma_{NIS}^2),$$

where

$$\begin{aligned}\sigma_{NIS}^2 &:= \mathbb{V}_q\Big(\big[\varphi(X)w(X)) - Iw(X)\big]\Big) \\ &= \int \frac{\pi(x)^2\left(\varphi(x) - I\right)^2}{q(x)}\mathrm{d}x.\end{aligned}$$

Proof

Proof.

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

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

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

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

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

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

Proof ctd...

Proof.

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

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

□

Alternatively, use Delta method.

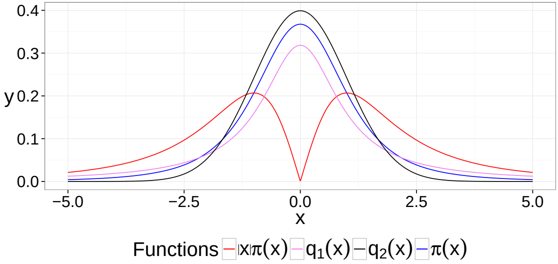
Toy Example: t-distribution

- We want to compute $I = \mathbb{E}_{\pi}(|X|)$ where $\pi(x) \propto \left(1 + x^2/3\right)^{-2}$ (t_3-distribution).

(a) Directly sample from π.

(b) Use $q_1(x) = g_{t_1}(x) \propto \left(1 + x^2\right)^{-1}$ (t_1-distribution).

(c) Use $q_2(x) \propto \exp\left(-x^2/2\right)$ (normal).



Toy Example: t-distribution

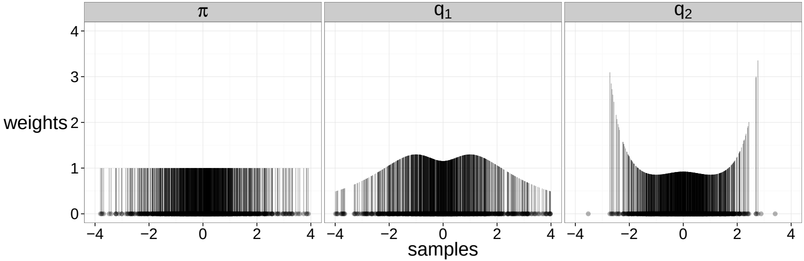


Figure: Sample weights obtained for 1000 realisations of X_i, from the different proposal distributions.

Toy Example: t-distribution

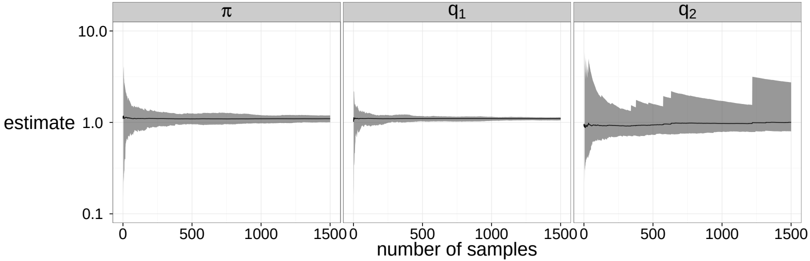


Figure: Estimates $\widehat{I_n}$ of I obtained after 1 to 1500 samples. The grey shaded areas correpond to the range of 100 independent replications.

Variance of importance sampling estimators

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

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

- Asymptotic Variance:

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

- Thus the asymptotic variance can be estimated consistently with

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

Variance of importance sampling estimators

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

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

- Asymptotic Variance:

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

- Thus the asymptotic variance can be estimated consistently with

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

Diagnostics

- Importance sampling works well when all weights roughly equal.
- If dominated by one $\widetilde{w}(X_j)$,

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

  The "effective sample size" is one.

- To how many unweighted samples correspond our weighted samples of size n? Solve for n_e in

$$\frac{1}{n}\mathbb{V}_{as}\left(\widehat{I}_n^{\text{NIS}}\right) = \frac{\sigma^2}{n_e},$$

  where σ^2/n_e corresponds to the variance of an unweighted sample of size n_e.

Diagnostics

- We solve by matching $\varphi(X_i) - \widehat{I}^{\mathrm{NIS}}$ with $\varphi(X_i) - I \approx \sigma$ as if they were i.i.d samples:

$$\frac{1}{n}\frac{\frac{1}{n}\sum_{i=1}^{n}\widetilde{w}(X_i)^2\left(\varphi(X_i) - \widehat{I}_n^{\mathrm{NIS}}\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{w}(X_i)\right)^2} \approx \frac{\sigma^2}{n_e}$$

$$\text{i.e.} \qquad \frac{1}{n}\frac{\frac{1}{n}\sum_{i=1}^{n}\widetilde{w}(X_i)^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{w}(X_i)\right)^2} = \frac{1}{n_e}.$$

- The solution is

$$n_e = \frac{\left(\sum_{i=1}^{n}\widetilde{w}(X_i)\right)^2}{\sum_{i=1}^{n}\widetilde{w}(X_i)^2},$$

and is called the effective sample size.

Rejection and Importance Sampling in High Dimensions

- **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

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

- How do Rejection sampling and Importance sampling scale in this context?

Performance of Rejection Sampling

- We have

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

  for $\sigma > 1$.

- Acceptance probability is

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

  i.e. exponential degradation of performance.

- For $d = 100$, $\sigma = 1.2$, we have

$$\mathbb{P}(X \text{ accepted}) \approx 1.2 \times 10^{-8}.$$

Performance of Importance Sampling

- We have

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

- Variance of the weights:

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

where $\sigma^4/\left(2\sigma^2 - 1\right) > 1$ for any $\sigma^2 > 1/2$.

- For $d = 100$, $\sigma = 1.2$, we have

$$\mathbb{V}_q[w(X)] \approx 1.8 \times 10^4.$$

Wait a minute. . .

Lecture 1:

- Simpson's rule for approximating integrals: error in $\mathcal{O}(n^{-1/d})$.

Lecture 2:

- Monte Carlo for approximating integrals: error in $\mathcal{O}(n^{-1/2})$ with rate independent of d.

And now:

- Importance Sampling standard deviation in the Gaussian example in $\exp(d)n^{-1/2}$.

The rate is indeed independent of d but the "constant" (in n) explodes exponentially (in d).

Markov chain Monte Carlo

- Revolutionary idea introduced by Metropolis et al., J. Chemical Physics, 1953.

- **Key idea**: Given a target distribution π, build a Markov chain $(X_t)_{t\geq 1}$ such that, as $t \to \infty$, $X_t \sim \pi$ and

$$\frac{1}{n}\sum_{t=1}^{n} \varphi(X_t) \to \int \varphi(x)\,\pi(x)\,dx$$

  when $n \to \infty$ e.g. almost surely.

- Also central limit theorems with a rate in $1/\sqrt{n}$.

- In some cases the constant (in n) does not explode exponentially with the dimension d, but polynomially.

Markov chain Monte Carlo

- Revolutionary idea introduced by Metropolis et al., J. Chemical Physics, 1953.

- **Key idea**: Given a target distribution π, build a Markov chain $(X_t)_{t \geq 1}$ such that, as $t \to \infty$, $X_t \sim \pi$ and

$$\frac{1}{n} \sum_{t=1}^{n} \varphi(X_t) \to \int \varphi(x) \pi(x) \, dx$$

  when $n \to \infty$ e.g. almost surely.

- Also central limit theorems with a rate in $1/\sqrt{n}$.

- In some cases the constant (in n) does not explode exponentially with the dimension d, but polynomially.

Markov chains - discrete space

- Let $\mathbb{X}$ be discrete, e.g. $\mathbb{X} = \mathbb{Z}$.
- $(X_t)_{t\geq 1}$ is a Markov chain if

  $$\mathbb{P}(X_t = x_t | X_1 = x_1, ..., X_{t-1} = x_{t-1}) = \mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}).$$

  The future is conditionally independent of the past given the present.

- Homogeneous Markov chains:

  $$\forall m \in \mathbb{N} : \mathbb{P}(X_t = y | X_{t-1} = x) = \mathbb{P}(X_{t+m} = y | X_{t+m-1} = x).$$

- The Markov transition kernel is a stochastic **matrix**

  $$K(i, j) = K_{ij} = \mathbb{P}(X_t = j | X_{t-1} = i).$$

Markov chains - discrete space

- Let $\mu_t(x) = \mathbb{P}(X_t = x)$, the chain rule yields

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_t = x_t) = \mu_1(x_1) \prod_{i=2}^{t} K_{x_{i-1}x_i}.$$

- The m-transition matrix K^m as

$$K_{ij}^m = \mathbb{P}(X_{t+m} = j | X_t = i).$$

- Chapman-Kolmogorov equation:

$$K_{ij}^{m+n} = \sum_{k \in \mathbb{X}} K_{ik}^m K_{kj}^n.$$

- We obtain

$$\mu_{t+1}(j) = \sum_i \mu_t(i) K_{ij}$$

i.e. using "linear algebra notation",

$$\mu_{t+1} = \mu_t K.$$

Roadmap

- We will see that we can choose the transition matrix K such that if $\mu_0 = \pi$ then $\mu_t = \pi$ for all t.
- In practice we will have $\mu_0 \neq \pi$;
- We will see that under certain conditions, not matter what μ_0 is, $\mu_t \to \pi$ in **total variation**.
- This is enough to guarantee us a law of large numbers and a central limit theorem;
- Making this convergence precise, e.g. in terms of the dimension, is still an active research area.

Irreducibility and aperiodicity

Definition

A Markov chain is said to be irreducible if all the states communicate with each other, that is

$$\forall x, y \in \mathbb{X} \quad \min\left\{t : K^t_{xy} > 0\right\} < \infty.$$

A state x has period $d(x)$ defined as

$$d(x) = \gcd\left\{s \geq 1 : K^s_{xx} > 0\right\}.$$

An irreducible chain is aperiodic if all states have period 1.

Example: $K_\theta = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$ is irreducible if $\theta \in [0,1)$ and aperiodic if $\theta \in (0,1)$. If $\theta = 0$, the gcd is 2.

Transience and recurrence

Introduce the number of visits to x:

$$\eta_x := \sum_{k=1}^{\infty} \mathbb{1}\{X_k = x\}.$$

Definition

A state x is termed transient if:

$$\mathbb{E}_x(\eta_x) < \infty,$$

where $\mathbb{E}_x$ refers to the law of the chain starting from x.
A state is called recurrent otherwise and

$$\mathbb{E}_x(\eta_x) = \infty.$$

If a finite state chain is irreducible, then either all states are recurrent or transient.

Invariant distribution

Definition

A distribution π is invariant for a Markov kernel K, if

$$\pi K = \pi.$$

Note: if there exists t such that $X_t \sim \pi$, then

$$X_{t+s} \sim \pi$$

for all $s \in \mathbb{N}$.
Example: for any $\theta \in [0, 1]$

$$K_\theta = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}$$

admits the invariant distribution

$$\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Detailed balance

Definition

A Markov kernel K satisfies detailed balance for π if

$$\forall x, y \in \mathbb{X} : \; \pi(x) K_{xy} = \pi(y) K_{yx}.$$

Lemma

If K satisfies detailed balance for π then K is π-invariant.

If K satisfies detailed balance for π then the Markov chain is reversible, i.e. at stationarity,

$$\forall x, y \in \mathbb{X} : \quad \mathbb{P}(X_t = x, X_{t+1} = y) = \mathbb{P}(X_t = x, X_{t-1} = y).$$

Lack of reversibility

- Let $P = \left(\begin{array}{ccc} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{array} \right)$.
- Check $\pi P = \pi$ for $\pi = (1/2, 1/3, 1/6)$.
- P cannot be π reversible as

$$1 \to 3 \to 2 \to 1$$

  is a possible sequence whereas

$$1 \to 2 \to 3 \to 1$$

  is not (as $P_{2,3} = 0$).
- Detailed balance does not hold as $\pi_2 P_{23} = 0 \neq \pi_3 P_{32}$.

Remarks

- All finite space Markov chains have at least one stationary distribution but not all stationary distributions are also limiting distributions.
- $$P = \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$$

  Two left eigenvectors of eigenvalue 1:

  $$\begin{aligned} \pi_1 &= (1/4, 3/4, 0, 0)\,, \\ \pi_2 &= (0, 0, 1/4, 3/4) \end{aligned}$$

  depending on the initial state, two different stationary distributions.

Equilibrium

Proposition

If a discrete space Markov chain is aperiodic and irreducible and admits an invariant distribution $\pi(\cdot)$, then

$$\forall x \in \mathbb{X} \quad \mathbb{P}_\mu \left(X_t = x \right) \xrightarrow[t \to \infty]{} \pi(x),$$

for any starting distribution μ.

- In the Monte Carlo perspective, we will be primarily interested in convergence of empirical averages, such as

$$\widehat{I}_n = \frac{1}{n} \sum_{t=1}^{n} \varphi \left(X_t \right) \xrightarrow[n \to \infty]{a.s.} I = \sum_{x \in \mathbb{X}} \varphi \left(x \right) \pi(x).$$

- Before turning to these "ergodic theorems", let us consider continuous spaces.

Markov chains - continuous space

- 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,

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

  The future is conditionally independent of the past given the present.

- 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.

Markov chains - continuous space

- Denoting μ_1 the pdf of X_1, we obtain directly

$$\mathbb{P}(X_1 \in A_1, ..., X_t \in A_t)$$
$$= \int_{A_1 \times \cdots \times A_t} \mu_1(x_1) \prod_{k=2}^{t} K(x_{k-1}, x_k)\, dx_1 \cdots dx_t.$$

- Denoting by μ_t the distribution of X_t, Chapman-Kolmogorov equation reads

$$\mu_t(y) = \int_{\mathbb{X}} \mu_{t-1}(x) K(x, y) dx$$

and similarly for $m > 1$

$$\mu_{t+m}(y) = \int_{\mathbb{X}} \mu_t(x) K^m(x, y) dx$$

where

$$K^m(x_t, x_{t+m}) = \int_{\mathbb{X}^{m-1}} \prod_{k=t+1}^{t+m} K(x_{k-1}, x_k)\, dx_{t+1} \cdots dx_{t+m-1}.$$

Example

- Consider the autoregressive (AR) model

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

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

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

- We also have

$$X_{t+m} = \rho^m X_t + \sum_{k=1}^{m} \rho^{m-k} V_{t+k}$$

so in the Gaussian case

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

with $\tau_m^2 = \tau^2 \sum_{k=1}^{m} \left(\rho^2\right)^{m-k} = \tau^2 \frac{1-\rho^{2m}}{1-\rho^2}$.

Irreducibility 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

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

Otherwise the chain is aperiodic.

Recurrence and Harris Recurrence

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

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

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

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

Harris recurrence is stronger than recurrence.

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

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

where f is a bounded measurable function.

Detailed balance

In practice it is easier to check the detailed balance condition:

$$\forall x, y \in \mathbb{X} \quad \pi(x)K(x,y) = \pi(y)K(y,x)$$

Lemma

If detailed balance holds, then π is invariant for K and K is π-reversible.

Example: the Gaussian AR process is π-reversible, π-invariant for

$$\pi(x) = \mathcal{N}\left(x; 0, \frac{\tau^2}{1-\rho^2}\right)$$

when $|\rho| < 1$.

Law of Large Numbers

Theorem

Suppose the Markov chain $\{X_i; i \geq 0\}$ is π-irreducible, with invariant distribution π, and suppose that $X_0 = x$.
Then for any π-integrable function $\varphi : \mathbb{X} \to \mathbb{R}$:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \varphi(X_i) = \int_{\mathbb{X}} \varphi(w)\,\pi(w)\,\mathrm{d}w$$

almost surely, for $\pi-$ almost every x.

*If the chain in addition is Harris recurrent then this holds for **every** starting value x.*

Convergence

Theorem

Suppose the kernel K is π-irreducible, π-invariant, aperiodic. Then, we have

$$\lim_{t\to\infty}\int_{\mathbb{X}}\left|K^t(x,y)-\pi(y)\right|dy=0$$

for $\pi-$almost all starting values x.

Under some additional conditions, one can prove that there exists a $\rho<1$ and a function $M:\mathbb{X}\to\mathbb{R}^+$ such that for all measurable sets A and all n

$$|K^n(x,A)-\pi(A)|\leq M(x)\rho^n.$$

The chain is then said to be **geometrically ergodic**.

Central Limit Theorem

Theorem

Under regularity conditions, for a Harris recurrent, π-invariant Markov chain, we can prove

$$\sqrt{t}\left[\frac{1}{t}\sum_{i=1}^{t}\varphi\left(X_i\right)-\int_{\mathbb{X}}\varphi\left(x\right)\pi\left(x\right)\mathrm{d}x\right]\xrightarrow[t\to\infty]{\mathcal{D}}\mathcal{N}\left(0,\sigma^2\left(\varphi\right)\right),$$

where the asymptotic variance can be written

$$\sigma^2\left(\varphi\right)=\mathbb{V}_\pi\left[\varphi\left(X_1\right)\right]+2\sum_{k=2}^{\infty}\mathbb{C}ov_\pi\left[\varphi\left(X_1\right),\varphi\left(X_k\right)\right].$$

This formula shows that (positive) correlations increase the asymptotic variance, compared to i.i.d. samples for which the variance would be $\mathbb{V}_\pi(\varphi(X))$.

Central Limit Theorem

Example: for the AR Gaussian model,
$\pi(x) = \mathcal{N}\left(x; 0, \tau^2/(1-\rho^2)\right)$ for $|\rho| < 1$ and

$$\mathbb{C}\mathrm{ov}\left(X_1, X_k\right) = \rho^{k-1}\mathbb{V}\left[X_1\right] = \rho^{k-1}\frac{\tau^2}{1-\rho^2}.$$

Therefore with $\varphi(x) = x$,

$$\sigma^2(\varphi) = \frac{\tau^2}{1-\rho^2}\left(1 + 2\sum_{k=1}^{\infty}\rho^k\right) = \frac{\tau^2}{1-\rho^2}\frac{1+\rho}{1-\rho} = \frac{\tau^2}{(1-\rho)^2},$$

which increases when $\rho \to 1$.

Markov chain Monte Carlo

- We are interested in sampling from a distribution π, for instance a posterior distribution in a Bayesian framework.

- Markov chains with π as invariant distribution can be constructed to approximate expectations with respect to π.

- For example, the Gibbs sampler generates a Markov chain targeting π defined on $\mathbb{R}^d$ using the full conditionals

$$\pi(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d).$$

Gibbs Sampling

- Assume you are interested in sampling from

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

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

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

  $\vdots$

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

  $\vdots$

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

Gibbs Sampling

A few questions one can ask about this algorithm:

- Is the joint distribution π uniquely specified by the conditional distributions $\pi_{X_i|X_{-i}}$?
- Does the Gibbs sampler provide a Markov chain with the correct stationary distribution π?
- 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

Theorem

Consider a distribution whose density $\pi\left(x_1, x_2, ..., x_d\right)$ *is such that*

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

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

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

The condition above is known as the **positivity condition**.

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

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

Sufficient for the Gibbs sampler to be irreducible.

Proof of Hammersley-Clifford Theorem

Proof.

We have

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

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.

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

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

Proof.

By $z \in \mathrm{supp}(\pi)$ 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

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

By iterating we have the theorem. $\square$

Example: Non-Integrable Target

- Consider the following conditionals on $\mathbb{R}^+$

$$\pi_{X_1|X_2}(x_1|\, x_2) = x_2 \exp(-x_2 x_1)$$
$$\pi_{X_2|X_1}(x_2|\, x_1) = x_1 \exp(-x_1 x_2).$$

  We might expect that these full conditionals define a joint probability density $\pi(x_1, x_2)$.

- Hammersley-Clifford would give

$$\begin{aligned}\pi(x_1, x_2, ..., x_d) &\propto \frac{\pi_{X_1|X_2}(x_1|\, z_2)}{\pi_{X_1|X_2}(z_1|\, z_2)} \frac{\pi_{X_2|X_1}(x_2|\, x_1)}{\pi_{X_2|X_1}(z_2|\, x_1)} \\ &= \frac{z_2 \exp(-z_2 x_1)\, x_1 \exp(-x_1 x_2)}{z_2 \exp(-z_2 z_1)\, x_1 \exp(-x_1 z_2)} \propto \exp(-x_1 x_2).\end{aligned}$$

- However $\iint \exp(-x_1 x_2)\, dx_1 dx_2 = \infty$ so $\pi_{X_1|X_2}(x_1|\, x_2) = x_2 \exp(-x_2 x_1)$ and $\pi_{X_2|X_1}(x_1|\, x_2) = x_1 \exp(-x_1 x_2)$ are not compatible.

Example: Positivity condition violated

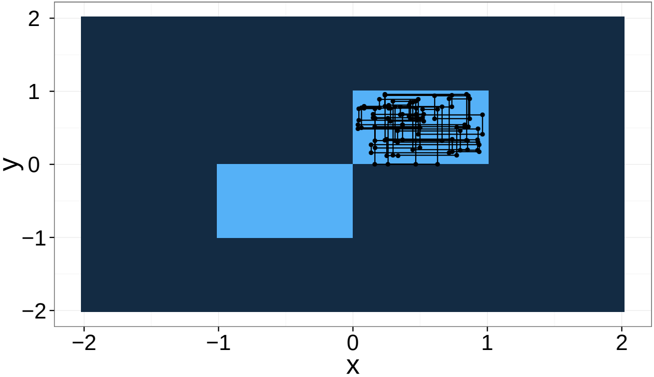


Figure: Gibbs sampling targeting $\pi(x, y) \propto \mathbf{1}_{[-1,0]\times[-1,0]\cup[0,1]\times[0,1]}(x, y)$.

Invariance of the Gibbs sampler I

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 $d > 2$:

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

Proposition

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

Invariance of the Gibbs sampler II

Proof for $d = 2$.

We have

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

$\square$