Advanced Simulation - Lecture 5

Patrick Rebeschini

January 29th, 2018

Limits of standard Monte Carlo methods

- Monte Carlo methods yield convergence rates in $1/\sqrt{n}$, which is independent of the dimension d.
- On close inspection, the error still depends on d, through the constant in front of the rate.
- Unfortunately that "constant" (in n) typically explodes exponentially with d.
- Markov chain Monte Carlo methods yield errors which explodes only polynomially in d, at least under some conditions.

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

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

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

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

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

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

- 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

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

Markov chains - discrete space

- Let $\mu_t(x) = \mathbb{P}\left(X_t = x\right)$, 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”,

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

Irreducibility and aperiodicity

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

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

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

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

- 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:

$$\eta_x := \sum_{k=1}^{\infty} \mathbb{1}_x \left(X_k \right).$$

- For a Markov chain, a state x is termed transient if:

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

where $\mathbb{E}_x$ refers to the law of the chain starting from x.

- A state is called recurrent otherwise and

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

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

Detailed balance

- 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 $P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$.
- 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 has an invariant distribution, 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,

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

- We have

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

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

- $K : \mathbb{X}^2 \rightarrow \mathbb{R}$ 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 process such that

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

- 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

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

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

- Given a distribution 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

$$\eta_A = \sum_{k=1}^{\infty} \mathbb{I}_A \left(X_k \right).$$

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

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

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

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

- Harris recurrence is stronger than recurrence.

Invariant Distribution and Reversibility

- 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

$$\forall f \quad \int\int f(x, y)\pi(x) K(x, y)\, dx dy = \int\int f(y, x)\pi(x) K(x, y)\, dx dy$$

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 $|\rho| < 1$.

Selected asymptotic results

- **Theorem**. If K is a π-irreducible, π-invariant Markov kernel, then for any integrable function $\varphi : \mathbb{X} \to \mathbb{R}$:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \varphi\left(X_i\right) = \int_{\mathbb{X}} \varphi\left(x\right) \pi\left(x\right) dx$$

almost surely, for $\pi-$ almost all starting value x.

- **Theorem**. If K is a π-irreducible, π-invariant, Harris recurrent Markov chain, then for any integrable function $\varphi : \mathbb{X} \to \mathbb{R}$:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} \varphi\left(X_i\right) = \int_{\mathbb{X}} \varphi\left(x\right) \pi\left(x\right) dx$$

almost surely, for any starting value x.

Selected asymptotic results

- **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 value x.

- Under some additional conditions, one can prove that a chain is geometrically ergodic, i.e. there exists $\rho < 1$ and a function $M : \mathbb{X} \to \mathbb{R}^+$ such that for all measurable set A:

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

  for all $n \in \mathbb{N}$. In other words, we can obtain a rate of convergence.

Central Limit Theorem

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

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

  where the asymptotic variance can be written

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

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

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

- Therefore with $\varphi\left(x\right) = x$,

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

which increases when $\rho \to 1$.