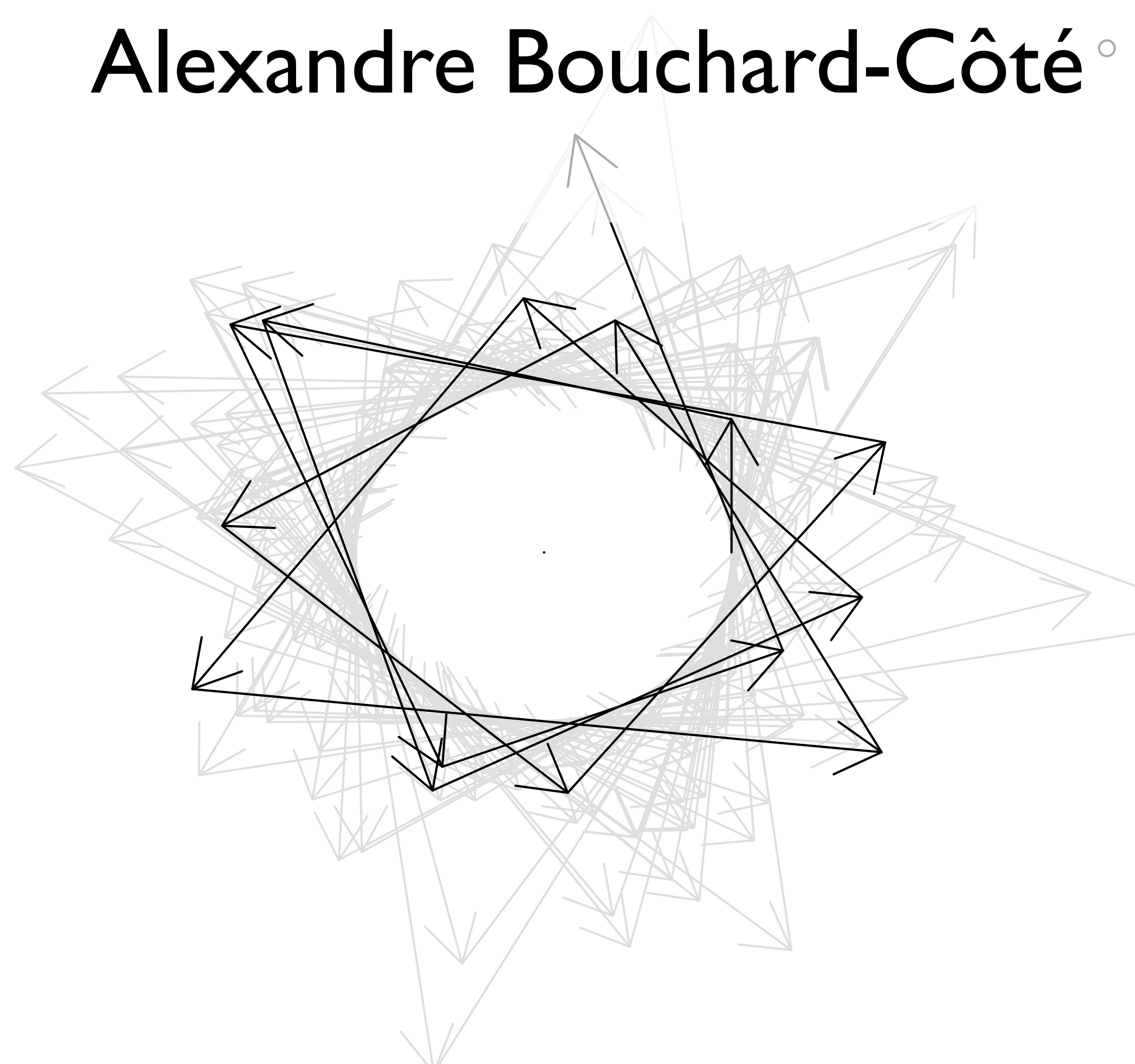
Monte Carlo methods

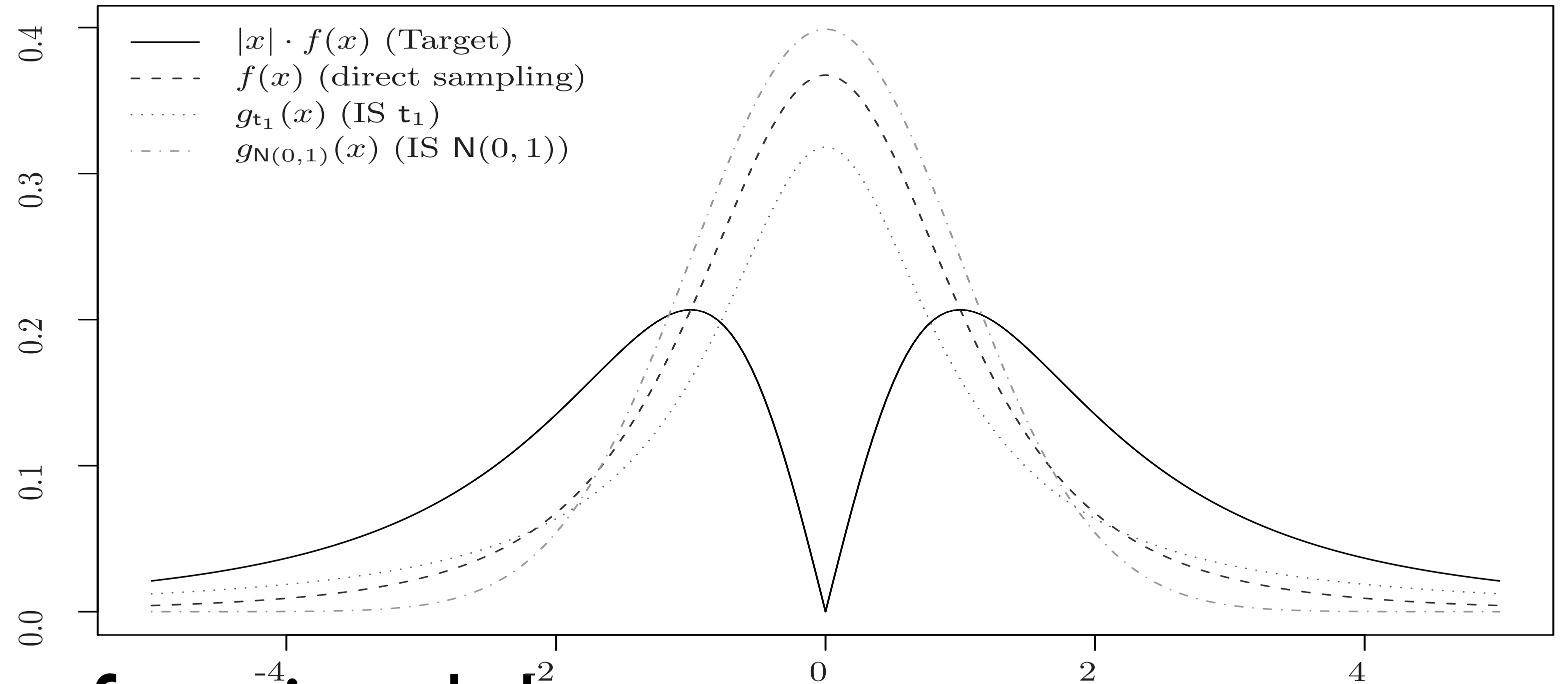
Alexandre Bouchard-Côté



Review / Q&A / Exercise solutions

Example/exercise

Ex 19



- Test function: |x|
- Target density: t-distribution, 3 degrees of freedom
- Compare (x-axis, 1-1500, y axis, partial sum, range of 100 replicates)
  - Simple MC
  - IS with t proposal, 1 degree of freedom
  - IS with normal proposal

Non-convergence?

- In the answer of Ex 19 (right), does IS still converges (albeit slowly)?
  - If not, construct an example where the following does not convergence (say in d) to a constant random variable?

$$\frac{1}{N}\sum_{i=1}^{N} X_i, \quad X_i \text{ iid}$$

NIS: Analysis of the asymptotic variance

Assume that $\mathbb{V}_q\left(\phi(X)w\left(X\right)\right) < \infty$ and $\mathbb{V}_q\left(w\left(X\right)\right) < \infty$ then

$$\sqrt{n}\left(\widehat{I}_n^{\mathrm{NIS}} - I\right) \overset{\mathrm{D}}{\to} \mathcal{N}\left(0, \sigma_{\mathrm{NIS}}^2\right)$$

Exercise: compute asymptotic variance

Tool: delta method

If: $$\sqrt{n}\left(Z_n - \mu\right) \overset{D}{\to} \mathcal{N}\left(0, \Sigma\right).$$

Then: $$\sqrt{n}\left(g\left(Z_n\right) - g\left(\mu\right)\right) \to \mathcal{N}\left(0, \nabla^T g\left(\mu\right)\ \Sigma\ \nabla g\left(\mu\right)\right).$$

NIS: Analysis of asymptotic *bias*

Prop 22

$$Assume\ \ that\ \mathbb{V}_q\left(\phi(X)w\left(X\right)\right) < \infty\ \ and\ \mathbb{V}_q\left(w\left(X\right)\right) < \infty\ \ then$$

$$\begin{aligned}\lim_{n\to\infty} n\mathbb{E}_q\left(\widehat{I}_n^{NIS} - I\right) &= -cov_q\left(\phi(X)w\left(X\right), w\left(X\right)\right) + \mathbb{V}_q(w\left(X\right))I \\ &= -\int\left(\phi\left(x\right) - I\right)\frac{\pi^2\left(x\right)}{q\left(x\right)}dx.\end{aligned}$$

- Consequence: asymptotically, the bias is negligible compared to the variance

IS and RS in high dimensions

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

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

and

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

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

Rejection sampling (RS)

- 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

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

Importance sampling

- We have

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

- For the 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 $\sigma^2 > 1/2 \Rightarrow$ Exponential variance increase.

- 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 $\exp(d)n^{-1/2}$.

$\Rightarrow$ The rate is indeed independent of d but the constant explodes.

Diagnostic for IS

Building Monte Carlo confidence interval for IS

Prop 24

- Bias asymptotically negligible, use asymptotic variance
- As in first exercise: for a 95% confidence interval, use

$$I_n \pm 1.96\sqrt{\sigma^2_{\text{asympt}}/n}$$

- The asymptotic variance is...
  - for BIS: $\sigma^2_{\text{IS}} := \mathbb{V}_q\left(\phi(X)w\left(X\right)\right)$
  - for NIS: $\sigma^2_{\text{NIS}} = \int \left(\phi\left(x\right) - I\right)^2 \frac{\pi^2\left(x\right)}{q\left(x\right)} dx$
- In both cases, replace unknowns by estimators...

Effective sampling size (ESS)

- Note with method from previous slide we need to fix a test function
  - On one hand this is good since performance can depend on the test function in general
    - For example: rare events
  - But often in practice performance more affected by discrepancy between target and proposal
  - Also, often have several test functions in mind
- So it's useful to have diagnostic depending only on the weights: use it to create the *particles*, ie pairs (x, w), then apply all the test functions to it

Effective sampling size (ESS)

- Relative ESS: constructed from unnormalized weights as follows

$$\frac{(E_q[\tilde{W}])^2}{E_q[\tilde{W}^2]} \approx \frac{(\frac{1}{n}\sum \tilde{W}^{(i)})^2}{\frac{1}{n}\sum(\tilde{W}^{(i)})^2} \quad \text{(Eq 25)}$$

  - Between [0, 1]
- ESS: multiply by number of particles
  - Interpretation and caveats: roughly, how many equivalent iid samples in terms of asymptotic variance - details in Owen 9.3
- Theoretical justification: more application of delta method, see Kong 1992, *A note on importance sampling using standardized weights*

Markov chain Monte Carlo

Motivation

- Methods we have seen so far (Simple MC, RS, IS)...
  - do not scale well in *d* (except for a few special cases)
  - often work poorly in combinatorial spaces

MCMC: main ideas

- We have LLNs and CLTs for Markov chains
  - Question: how to characterize the limits? (we cannot do it with the law of an arbitrary X_i as in iid case)
  - Answer: use the stationary law instead
- We can design and simulate Markov chains with a prescribed stationary distribution π
  - Even if we do not know the normalization of π

Towards MC LLN&CLT: Finite MC review

- Let $\mathbb{X}$ be finite, w.l.o.g. $\mathbb{X} := \{1, 2, ..., p\}$, then $(X_t)_{t \geq 1}$ is a Markov chain if

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

- We restrict ourselves to 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 so-called Markov transition kernel is

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

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

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

- We can also define the m-transition matrix K^m as

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

- Chapman-Kolmogorov equation:

$$K^{m+n} = K^m K^n.$$

- We obtain

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

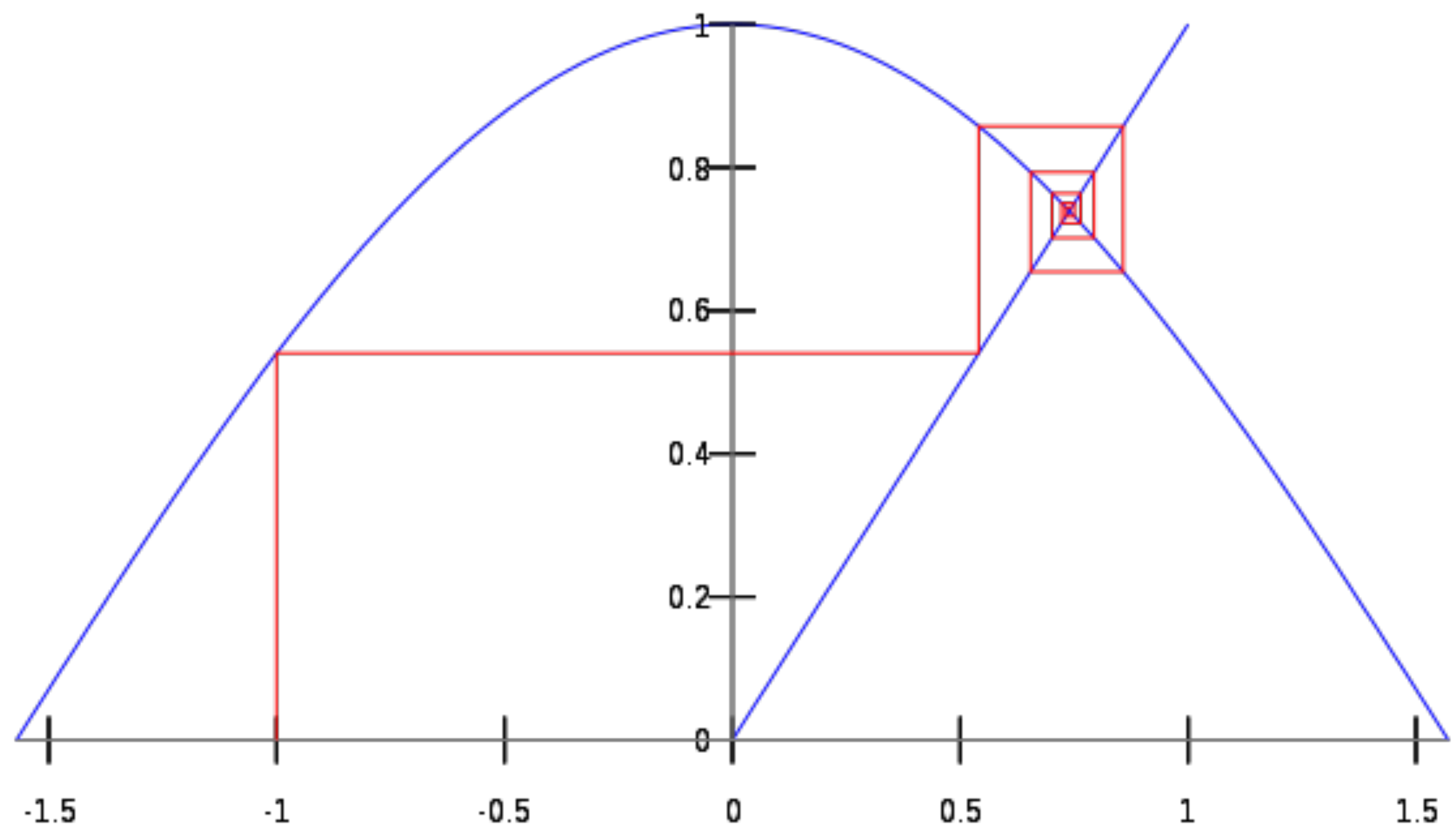
i.e. in standard vector-matrix multiplication

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

and recursively $\mu_{t+m} = \mu_t K^m$.

Stationarity/invariance

Fixed points of the transition kernels



- **Definition:** A distribution π is said to be *invariant* or *stationary* for a Markov kernel, K, if $\pi K = \pi$.
- If there exists t such that $X_t \sim \pi$ where π is a stationary distribution, then $X_{t+s} \sim \pi K^s = \pi$ for all $s \in \mathbb{N}$. (Note that this tells us nothing about the correlation between the states or their joint distribution.)
- *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.

- **Definition:** A Markov kernel K is π-reversible if

$$\forall x, y \in \mathbb{X}: \; \pi_x K_{xy} = \pi_y K_{yx}.$$

- **Lemma**: If K is π-reversible then K is π-invariant.
- **Proof**. Indeed we have

$$\sum_{x \in \mathbb{X}} \pi_x K_{xy} = \sum_{x \in \mathbb{X}} \pi_y K_{yx} = \pi_y,$$
$$\text{i.e . } (\pi K)_y = \pi_y$$

- Reversibility means that the statistics of the time-reversed version of the process match those of the process in the forward distribution, K_θ is π-reversible as $\pi_1 K_{\theta,12} = \frac{1}{2}(1-\theta) = \pi_2 K_{\theta,21}$.

- Let $P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$.

  We have $\pi P = \pi$ for $\pi = (1/2, 1/3, 1/6)$.
- P cannot be π reversible as

$$1 \rightarrow 3 \rightarrow 2 \rightarrow 1$$

  is a possible sequence whereas

$$1 \rightarrow 2 \rightarrow 3 \rightarrow 1$$

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

- All finite Markov chains have at least one stationary distribution but not all stationary distributions are also limiting distributions.
- **Example**

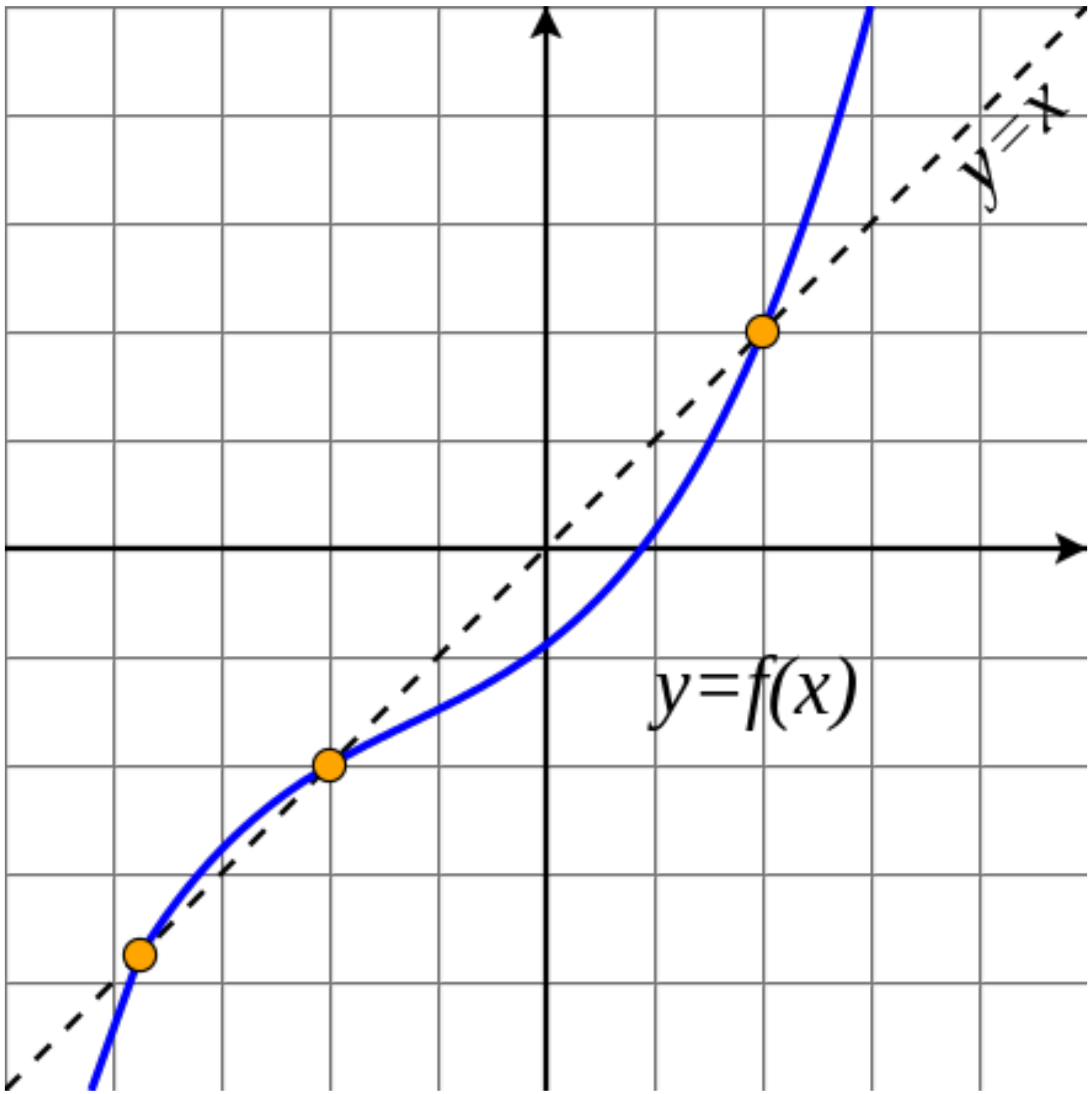
$$P = \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0.4 & 0.6 \\ 0 & 0 & 0.2 & 0.8 \end{pmatrix}$$

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 initial state we get a different stationary distribution.

Intuition



- **Definition:** A Markov chain is said to be irreducible if all the states communicate with each other, that is $\forall x, y \in \mathbb{X}$ $\inf \left\{ t : K^t_{xy} > 0 \right\} < \infty$.

- **Definition:** An irreducible Markov chain is aperiodic if there exists $x \in \mathbb{X}$ such that

$$\gcd \{ s \geq 1 : K^s_{xx} > 0 \} = 1$$

where gcd denotes the greatest common divisor.

- *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.

- **Proposition**: If a finite state-space Markov chain is irreducible then it has a unique stationary distribution and

$$\widehat{I}_n := \frac{1}{n}\sum_{t=1}^{n} \phi\left(X_t\right) \to I := \sum_{x\in\mathbb{X}} \phi\left(x\right) \pi(x).$$

- **Proposition**: If a finite state-space Markov chain is irreducible and aperiodic, then there exists $0 \leq \alpha < 1$ such that

$$\frac{1}{2}\left|\mathbb{P}(\, X_t = x|\, X_1) - \pi(x)\right| \leq \alpha^t.$$

- *Remark*: Aperiodicity is not required for the averages to converge to the expectation; e.g. take K_0.

This result (convergence of marginals) is not as useful to us

Exercise

- Construct an irreducible discrete Markov chain
- Compute a Monte Carlo average with test function = indicator on one of the states
- Try to make an educated analytical guess for the numerical value of asymptotic variance
- Approximate numerically the asymptotic variance

Why we need a CLT

- As before with IS, we want:
  - to determine when we have enough samples
  - to compare the running time of competing methods

Hint for the exercise

Consider an irreducible chain then

$$\lim_{n\to\infty} n\mathbb{V}_\pi\left(\widehat{I}_n\right) = \mathbb{V}_\pi\left(\phi\left(X_1\right)\right) + 2\sum_{k=1}^{\infty}\underbrace{\mathsf{Cov}_\pi\left(\phi\left(X_1\right), \phi\left(X_{k+1}\right)\right)}_{:=C(k)}$$

Proof: We have $\mathbb{E}_\pi\left(\widehat{I}_n\right) = I$ and

$$n\mathbb{V}_\pi\left(\widehat{I}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\underbrace{\mathsf{Cov}_\pi\left(\phi\left(X_i\right), \phi\left(X_j\right)\right)}_{=C(i-j)}$$

$$= \frac{1}{n}\sum_{k=-n+1}^{n-1} C\left(k\right)\times\underbrace{\left(\#\ \text{pairs} : i-j=k\right)}_{=n-|k|}$$

$$= \sum_{k=-n+1}^{n-1}\left(1-\frac{|k|}{n}\right)C\left(k\right) = \sum_{k=-\infty}^{\infty}\max\left(0, 1-\frac{|k|}{n}\right)C\left(k\right)$$

Now, specialize the Cov(...) expression for the setup of Exercise 34