Sampling by random walk

A: Random walks on Markov chains

The sampling methods we'll cover

1. Gibbs sampler
2. Metropolis-Hastings sampler
3. Langevin sampler

High-level approach: **random walk**.

Sampling by random walk

Want to sample from a distribution P over some space $\mathcal{X}$. This might be

- **discrete**, e.g., $\mathcal{X} = \{0, 1\}^N$ (say, binary images on N pixels), or
- **continuous**, e.g. $\mathcal{X} = \mathbb{R}^d$ (say, rainfall levels in d cities)

Difficulties:

- $\mathcal{X}$ might be huge or infinite: we cannot enumerate all outcomes.
- We might not be able to evaluate $P(x)$ explicitly for $x \in \mathcal{X}$ due to unknown normalization factor. But can often get ratios $P(x)/P(x')$.

Solution strategy: **random walk on $\mathcal{X}$**

- Start at any $x \in \mathcal{X}$
- Repeatedly move to a "nearby" state, with some transition probabilities
- After a while: the distribution over the current location is (close to) P

Random walks and Markov chains: the finite case

Random walk on a finite space $\mathcal{X}$

- **Let Q_t be the position ("state") at time t**
  Next state Q_{t+1} depends only on Q_t, not prior history: **Markov chain**

- **Random walk is defined by $|\mathcal{X}| \times |\mathcal{X}|$ transition matrix**

$$M(x, x') = M_{x,x'} = \Pr(Q_{t+1} = x' | Q_t = x)$$

- **Let π_t be the distribution of Q_t, so $\pi_t \in \Delta_{\mathcal{X}}$**

$$\pi_{t+1}(x) = \Pr(Q_{t+1} = x) = \sum_{x' \in \mathcal{X}} \Pr(Q_t = x') M_{x',x} = \sum_{x'} \pi_t(x') M_{x',x}$$

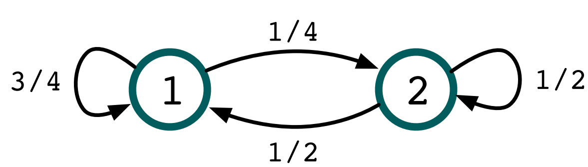
  In vector form:

$$\pi_{t+1}^T = \pi_t^T M = \pi_{t-1}^T M^2 = \cdots = \pi_o^T M^{t+1}$$

Stationary distribution

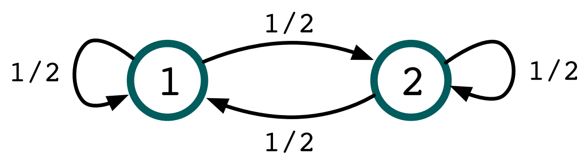
We say π is a **stationary distribution** if $\pi^T = \pi^T M$. Such a distribution always exists.

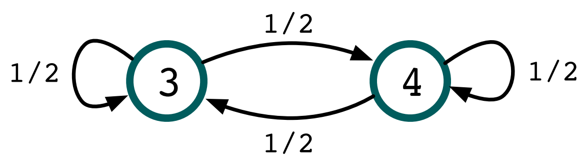
Determine the stationary distribution of this Markov chain:



Stationary distribution: issues

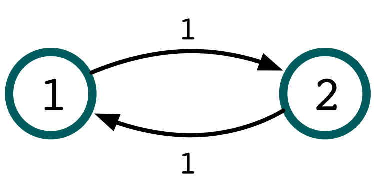
(1) There may be several stationary distributions





Stationary distribution: issues

(2) The random walk may not converge to the stationary distribution, even if it is unique



Irreducible, aperiodic Markov chains

Things become easier if the Markov chain is:

- **Irreducible:** the transition graph (nodes are states, directed edges are transitions with non-zero probability) is strongly connected.
- **Aperiodic:** there exists $k > 0$ such that $M^k(x, x') > 0$ for all x, x'.

Theorem. Any irreducible, aperiodic Markov chain has a unique stationary distribution π^*. For all $x, x' \in \mathcal{X}$,

$$\lim_{t \to \infty} M^t(x, x') = \pi^*(x').$$

Figuring out the stationary distribution

Lemma. If π satisfies the **detailed balance** condition:

$$\pi(x)M(x, x') = \pi(x')M(x', x) \quad \forall x, x' \in \mathcal{X}$$

then π is a stationary distribution of M.

B: The Gibbs sampler

Gibbs sampler

Finite state space $\mathcal{X} = \mathcal{X}_o^N$. Want to sample from a distribution $P > 0$ on $\mathcal{X}$.

- Start with any $x \in \mathcal{X}$
- Repeat:
  - Pick a coordinate $i \in \{1, 2, \ldots, N\}$ at random
  - Resample x_i from $P(X_i = x_i | x_{\setminus i})$

Check:

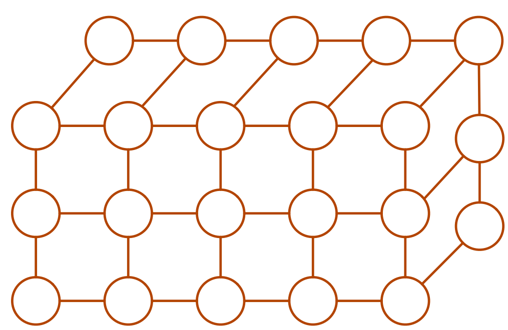
1. This is a Markov chain
2. It is irreducible and aperiodic
3. The stationary distribution is P

Example: Ising model

System of N particles arranged in a lattice.

- Each particle has a **spin** $X_i \in \{-1, +1\}$
- Overall configuration
  $X = (X_1, \ldots, X_N) \in \{-1, +1\}^N$

Probability $P(x) \propto e^{-U(x)}$



Energy of configuration x:

$$U(x) = -\sum_{i,j \text{ neighbors}} J_{ij} x_i x_j - \sum_i \beta_i x_i$$

- Ferromagnetic regime: $J_{ij} > 0$
- Statistical mechanics: Local interactions $\Longrightarrow$ macroscopic properties

Gibbs sampler for Ising model

Pick a particle $k \in [N]$ and resample its spin $X_k \in \{-1, +1\}$ while keeping everything else fixed.

Mixing time of a Markov chain

How many steps before the random walk gets close to the stationary distribution?

For Markov chain with transition matrix M, we can define **mixing time** $T_{\text{mix}}(\epsilon)$ as the smallest t for which

$$\max_{x \in \mathcal{X}} \| M^t(x, \cdot) - \pi^*(\cdot) \|_{TV} \leq \epsilon.$$

- The chain is **rapidly mixing** if T_{mix} is polynomial in dimension of $\mathcal{X}$
- The chain is **torpidly mixing** if T_{mix} is super-polynomial (e.g. exponential)

C: Metropolis-Hastings sampler

Metropolis-Hastings walk

Want to sample from distribution P on state space $\mathcal{X}$.

- We already have an irreducible, aperiodic Markov chain on it, with transition probabilities $M(x, x')$.
- But it doesn't have the right stationary distribution. How to modify it?

- Start with any $x \in \mathcal{X}$
- Repeat:
  - Pick a new state $x' \sim M(x, \cdot)$
  - Accept it with probability

$$\min\left(\frac{P(x')M(x', x)}{P(x)M(x, x')}, 1\right)$$

  else stay at x

Analyzing the stationary distribution

Theorem. The modified chain has stationary distribution P.

C: Langevin sampler

Brownian motion

- Random movement of particles in liquid/gas.
- Robert Brown, botanist: "pollen grains suspended in water perform a continual swarming motion" (1827).

Mathematical model: a *Gaussian process*.

- $B_0 = 0$
- For $t > s$,
  - $B_t - B_s$ is independent of B_s
  - $B_t - B_s \sim N(0, \omega^2(t-s))$
- $t \to B_t$ is almost surely continuous

Limit of a simple random walk with step size δ and time increment τ going to zero such that $\delta/\sqrt{\tau} \to \omega$.

Langevin diffusion

Suppose the target density on $\mathcal{X} = \mathbb{R}^d$ is

$$\pi(x) \propto e^{-U(x)}.$$

Langevin diffusion (X_t) is defined by stochastic differential equation

$$dX_t = -\nabla U(X_t)dt + \sqrt{2}dB_t$$

where (B_t) is d-dimensional Brownian motion.

Long-term distribution of (X_t):

- Suppose U is twice continuously differentiable and ∇U is Lipschitz.
- Then π is the unique stationary distribution of this process.

How can this process be simulated?

Discretizing the Langevin diffusion

Euler-Maruyama scheme for sampling diffusion paths:

$$X_{t+1} = X_t - \gamma_{t+1} \nabla U(X_t) + \sqrt{2\gamma_{t+1}} Z_{t+1}$$

where $Z_1, Z_2, \ldots$ are i.i.d. $N(0, I_d)$ and γ_t are step sizes.

- Related to stochastic gradient descent
- If step size is held constant ($\gamma_t = \gamma$):
  - Converges to a unique stationary distribution π_γ
  - But this isn't (necessarily) the same as π
- When step size is decreased: harder to analyze.

Metropolis-adjusted Langevin algorithm (MALA): Use Metropolis-Hastings to fix the bias, i.e., use the discretized diffusion as a proposal distribution.

Historical notes

Metropolis-Hastings sampler:

- Metropolis, Rosenbluth, Rosenbluth, Teller, Teller. *Equations of state calculations by fast computing machines*. Journal of Chemical Physics, 1953.
  Goal was to sample from $p(x) \propto e^{-E(x)/kT}$. Only considered symmetric proposal distributions.

- Hastings. *Monte Carlo sampling methods using Markov chains and their applications*. Biometrika, 1970.
  Generalized sampler.

Historical notes (cont'd)

Gibbs sampler:

- Formalization of Glauber dynamics in statistical physics.
- Geman, Geman. *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984.
- Gelfand, Smith. *Sampling based approaches to calculating marginal densities*. Journal of the American Statistical Association, 1990.

Historical notes (cont'd)

Langevin sampler:

- Paul Langevin (1872-1946) developed Langevin equation that described evolution of a system under a combination of determinstic and random forces.
- Grenander, Miller. *Representations of knowledge in complex systems*. Journal of the Royal Statistical Society, 1994.
- Besag. *Comments on "Representations of knowledge in complex systems"*. Same journal.