An Introduction to Markov Chain Monte Carlo

STATS 217: Introduction to Stochastic Processes I
Summer 2026

Skyler Wu, Stanford University Department of Statistics
skylerw@stanford.edu

Outline for Today

1. **Motivation:** sampling from arbitrary distributions is important but hard.
2. **Solution:** Markov Chain Monte Carlo (MCMC, Metropolis), in pictures.
3. **Math:** How/why does MCMC work?
   - Stationarity — “Once correct, remains correct.”
   - Detailed Balance + Reversibility — “A quick way to prove stationarity.”
   - Ergodicity — “Can we actually get to the correct stationary distribution?”
4. **Metropolis-Hastings MCMC:** a concrete example, in detail.
5. **Frontiers:** more powerful MCMC algorithms.

Motivation: sampling from arbitrary distributions is important, but very hard.

1. In statistics, our bread and butter is computing expectations:
   - Suppose X has PMF/PDF π. Then, $\mathbb{E}[f(X)] = \sum_x f(x)\pi(x)$ if discrete, else $\int f(x)p(x)dx$ if continuous.
   - But … what if the sum/integral does not have a closed-form solution? 😭
   - Fine … let's just numerically-approximate using the trapezoid rule or something with Δx … 🤓
     - But … what if X is high-dimensional (e.g., a size-1000 random vector)? Good luck.
     - We're going to need a whole bunch of trapezoids … 😭
     - One more thing: we might only know π up to a normalizing constant.
     - Example (Bayesian statistics): $\pi(x) := p(x \mid \mathbf{y}) = p(\mathbf{y} \mid x)p(x)/p(\mathbf{y})$, where $p(\mathbf{y}) = \int p(\mathbf{y}, x)dx$ is too hard to compute.
   - Idea ("Monte Carlo"): just draw many samples $x_1, \ldots, x_n \overset{i.i.d.}{\sim} \pi$ and approximate $\mathbb{E}[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.
     - But not everything has an `np.random.(…)` function! We can't always directly sample i.i.d.! 😔
2. **Big Idea: design a Markov chain $X_0, \ldots, X_n$ whose "long-run" distribution is π. This is MCMC.**

OK, what does that really mean … practically?

1. **Big Idea of MCMC: design a Markov chain $X_0, \ldots, X_n$ whose "long-run" distribution is π.**
    - Markov chain (in practice): specify two things.
        - **Initial condition:** X_0 (could be random $\sim \pi_0$, or deterministic).
        - **Transition probability function:** $p(x, y) := P(X_{n+1} = y \mid X_n = x)$ for all n.
    - OK what does "long-run" distribution mean here?
        - We want π to be the **unique, stationary distribution** of our Markov chain.
2. Great, what can I practically do with this setup?
    1. Run Markov chain for large n, approximate using **sampled states****: $\mathbb{E}_\pi[f(X)] \approx n^{-1} \sum_{i=1}^{n} f(X_i)$.
    2. Theoretical guarantees: $n^{-1} \sum_{i=1}^{n} f(X_i) \xrightarrow{p} \mathbb{E}_\pi[f(X)]$ as $n \to \infty$.

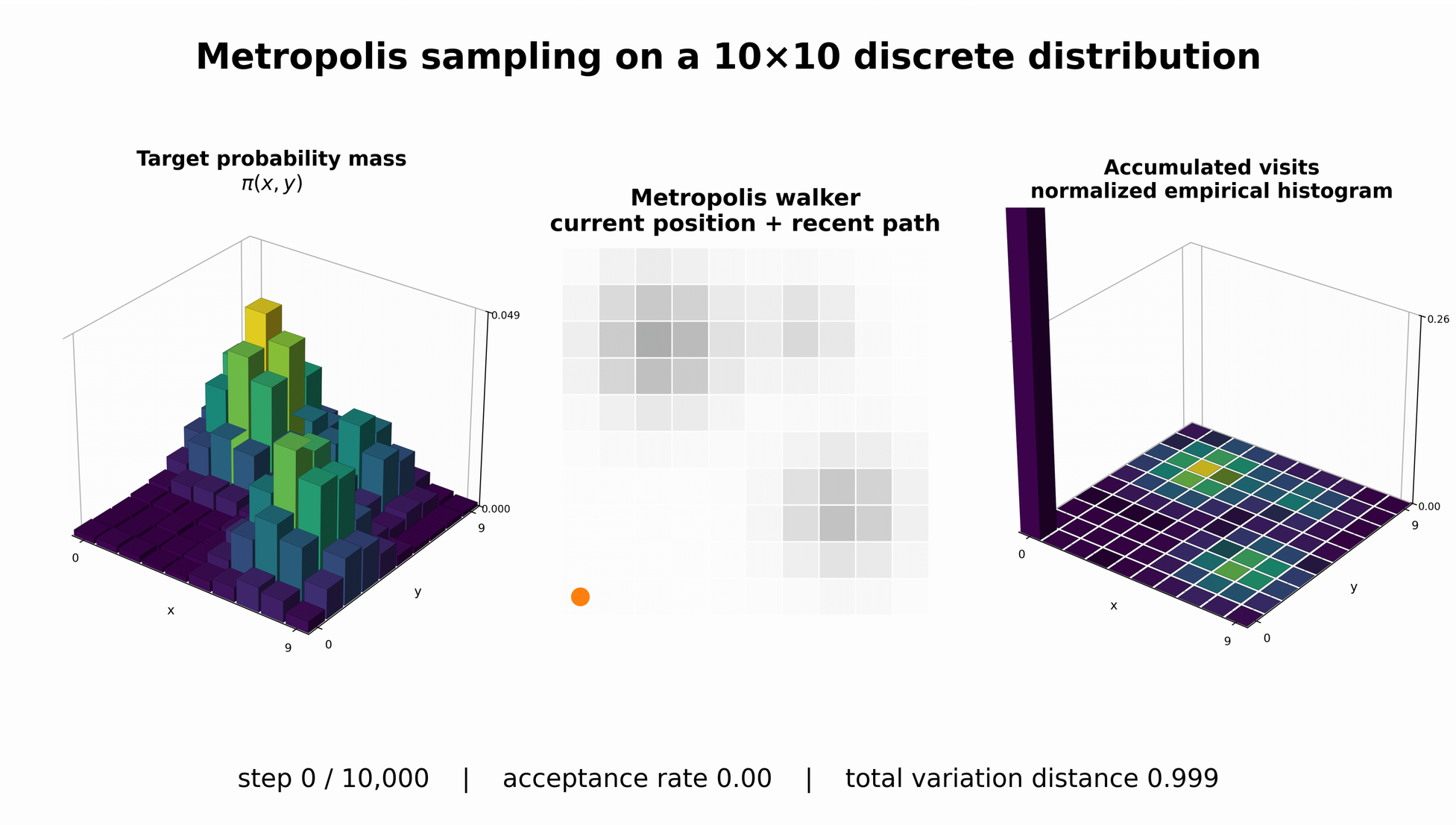
Great, we can approximate expectations. Why should I care?

1. **To recap, we can, with MCMC:**
   - **Design a Markov chain $X_0, \ldots, X_n$ whose unique, stationary distribution is π.**
   - **Use the sampled states to approximate, for any f, the expectation $\mathbb{E}_\pi[f(X)]$ for $X \sim \pi$ … even if we only know $\pi(\cdot)$ up to a normalizing constant.**
2. **OK, how is this useful in the real world?**
   - **Healthcare:** suppose $\theta \sim \pi$ is a random variable representing how many additional years of life a cancer patient can live after taking a new treatment. *Randomness could come from different patient characteristics, metabolic pathways, immune responses, etc.*
     - I might want to estimate $\mathbb{E}_\pi[\theta]$ — on average, how many more years of life can this treatment give a patient?
     - But, a point estimate might not be enough — $\mathbb{E}_\pi[\mathbf{1}(\theta \leq 0)]$ tells me the probability that this treatment might harm a patient.
     - Why not a confidence/credible interval? Because the distribution of θ may not have a nice form.
   - **Finance:** suppose $X \sim \pi$ is a random variable representing a stock price, and π is the modeled distribution of that stock price in 1 day. Can't sample directly from π, else everyone would be rich.
     - I might want to know $\mathbb{E}_\pi[\mathbf{1}(X \leq c)]$, the probability my stock price drops below c.
     - Intuitively, sample many simulated draws $X_1, \ldots, X_n \sim \pi$ and see how many of them $\leq c$.

OK, all this math is confusing me. I want street English.

1. **Street Intuition Version:** MCMC means:
    - Initialize a **<hopefully intelligent>**** random walker at X_0.
    - Have it random-walk thru the **<state space>** (i.e., all possible values of X) for n time steps, where n is large, and record where it visited as **<samples>** $X_1, \ldots, X_n$.
    - Hopefully, have it visit regions of state space x with frequency **<corresponding exactly in the limit>** to $\pi(x)$ (i.e., go to high-$\pi(x)$ regions more often, and low-$\pi(x)$ regions less often, but **not never** … to **<properly explore>** the state space.) — governed by **<transition probability>** $p(x, y)$.
    - Use samples $X_1, \ldots, X_n$ as approximate draws from **<target distribution>** π to estimate expectations of interest and/or do more simulations.

MCMC in pictures: a picture is a thousand words … a GIF can be hundreds of thousands of words depending on the frame rate.



*Here, $\mathbf{x} = (x, y)$. MCMC is designed for high-dimensional problems!

Mathematical/Statistical Requirements of MCMC targeting π (1)

1. **Stationarity** — "Once in correct distribution, remains in correct distribution."
    - π is a *stationary distribution* for Markov chain P (think of as a transition matrix, with transition probability function p) **if and only if** $\pi P = \pi$.
        - Written out: **for any states** x, y of our Markov chain P, $\sum_x \pi(x)p(x,y) = \pi(y)$, where $p(x,y) := P(X_{n+1} = y \mid X_n = x)$ for all n.
    - Interpretation: "if the chain currently has distribution π, then after one more step it will still have distribution π."
    - An easier way for me: if $X_n \sim \pi$ (marginally), then $X_{n+1} \sim \pi$ (marginally), for any n.
    - But … for many problems (esp. continuous/infinite), stationarity on its own is **very hard to check** because of the summation (over potentially infinite items).
        - As an aside, in continuous settings, replace all PMFs with PDFs and all sums with integrals.

Mathematical/Statistical Requirements of MCMC targeting π (2)

2. **Detailed Balance + Reversibility** — "A quick way to prove stationarity."

   - π satisfies the *detailed balance condition* if **for any states** x, y of our Markov chain P, $\pi(x)p(x, y) = \pi(y)p(y, x)$.

     - A Markov chain P is *with reversible with respect to* π if the *detailed balance condition* above holds.

     - Equivalently, the detailed balance condition says **for any states** x, y of our Markov chain P: $P(X_{n+1} = y, X_n = x) = P(X_{n+1} = x, X_n = y)$.

     - Also, equivalently: if the chain is **started in stationarity** as $X_0 \sim \pi$, then $(X_0, \ldots, X_n)$ has the same (joint) distribution as $(X_n, \ldots, X_0)$.

       - Intuitively: "the process looks the same forwards and backwards in time."

   - Interpretation of Detailed Balance: the "probability flow" from x to y equals the "probability flow" from y to x.

   - Importance: the *detailed balance condition* implies stationarity! (See next slide).

Mathematical/Statistical Requirements of MCMC targeting π (3)

3. **Proof:** detailed balance implies stationarity.
   - By definition, if π satisfies the *detailed balance condition,* then **for any states** x, y of our Markov chain P, $\pi(x)p(x,y) = \pi(y)p(y,x)$.
   - Summing over all x, we have:
     $$\sum_x \pi(x)p(x,y) = \sum_x \pi(y)p(y,x) = \pi(y)\sum_x p(y,x) = \pi(y).$$
   - The last equality occurs because from y, the chain **must go somewhere.**
   - **Implication:** *detailed balance* is a sufficient (but not necessary) condition for *stationarity*. Usually, it's good/simple enough for us to check.

Mathematical/Statistical Requirements of MCMC targeting π (4)

4. **Ergodicity (bonus)**— "Can we get to the correct stationary distribution?"
    - Stationarity (implied by detailed balance) tells us once we are at the target stationary distribution π, we will stay at π.
    - But, MCMC usually starts from some arbitrary or random X_0. How do we know that we that the distribution of X_n (as $n \to \infty$) will actually **ever** approach π?
5. **Formalization:** a Markov chain is *ergodic* if for any initial condition $X_0 = x$, the distribution of X_n (as $n \to \infty$, conditioning only on $X_0 = x$) converges to π (in TV distance).
6. **Theorem:** in finite state spaces, *ergodicity* is implied by three conditions:
    - Irreducible: from any state x, there exists some positive-probability path to any other state y: "no islands."
        - Intuitively, can the chain reach all relevant parts of the space?
    - Aperiodic: every state has period 1. *For finite irreducible chains, it suffices to check one state. Even lazier, in MCMC, so long as the chain has the option to <stay in the same place>, aperiodicity is automatically satisfied.*
        - Intuitively, does the chain avoid rigid cycles?
    - π is a stationary distribution.

Metropolis-Hastings MCMC: a concrete example.

1. Suppose we are in a discrete state space indexed by x and wish to sample from target π. Initialize $X_0 = x_0$ arbitrarily.
    - *Proving ergodicity will show us that initialization doesn't matter!*
2. For each sampling step $i = 1, \ldots, n$:
    - **Proposal:** at current state $X_{i-1} = x$, use a *proposal distribution* $q(x, \cdot)$ *(that we know how to sample from directly)* to draw a *proposal* $Y \sim q(x, \cdot)$ — i.e., $P(Y = y \mid X_{i-1} = x) = q(x, y)$.
        - In discrete settings, can pick q to be a discrete random walk (e.g., up, right, down, left with or without equal probabilities … but still ensure irreducibility — the accept/reject step will help us correct).
        - In continuous cases, oftentimes people use $Y \sim \mathcal{N}(\cdot \mid x, \epsilon^2)$, for step-size ϵ —i.e., a continuous random walk.
    - **Accept/Reject:** compute *acceptance probability* $\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}\right)$. *Note that ratio cancels out π's normalizing constant!*
        - Set $X_i = y$ with probability $\alpha(x, y)$ (i.e., "accept the proposal"), else set $X_i = X_{i-1} = x$ (i.e., "reject the proposal" / "stay in place").
        - In practice, we implement via a uniform "coin-flip": $U \sim \text{Unif}(0,1)$.
        - Intuitions: if we don't do accept/reject, the random walk proposal might not match the target distribution. Accept/reject is mandatory to **ensure detailed balance, and thus stationarity.** Also, think in terms of "encouraging movement towards higher mass regions."

Proving that Metropolis-Hastings satisfies detailed balance (1).

1. **Metropolis-Hastings Algorithm** (summarized, after initialization): for each sampling step $i = 1, \ldots, n$:
   - **Proposal:** draw a *proposal* $Y \sim q(x, \cdot\,)$ — i.e., $P(Y = y \mid X_{i-1} = x) = q(x, y)$.
   - **Accept/Reject:** compute *acceptance probability* $\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$.
     - Set $X_i = y$ with probability $\alpha(x, y)$, else set $X_i = X_{i-1} = x$.
2. **Proposition:** the transition probability $p(x, y) := P(X_{n+1} = y \mid X_n = x)$ of Metropolis-Hastings is:
   A. **If** $x \neq y$**:** then, $p(x, y) = q(x, y)\alpha(x, y)$.
   B. **If** $x = y$**:** then, $p(x, y) = 1 - \sum_{y' \neq x} q(x, y')\alpha(x, y')$.

Proving that Metropolis-Hastings satisfies detailed balance (2).

1. **Want to prove:** $\pi(x)p(x,y) = \pi(y)p(y,x)$ for any states x, y.
   - Define $A := \pi(x)q(x,y)$ and $B := \pi(y)q(y,x)$. Then, $\alpha(x,y) = \min\left(1, \frac{B}{A}\right)$.
   - **Super useful fact:** note that $A \min\left(1, \frac{B}{A}\right) = \min(A, B)$ for positive A, B.
   - If $x \neq y$, then $p(x,y) = q(x,y)\alpha(x,y)$ and $p(y,x) = q(y,x)\alpha(y,x)$:
     - Thus, $\pi(x)p(x,y) = \pi(x)q(x,y)\alpha(x,y) = A \min\left(1, \frac{B}{A}\right) = \min(A,B)$.
     - Also, $\pi(y)p(y,x) = \pi(y)q(y,x)\alpha(y,x) = B \min\left(1, \frac{A}{B}\right) = \min(B,A)$.
     - Thus, $\pi(x)p(x,y) = \pi(y)p(y,x)$ for any $x \neq y$.
   - If $x = y$, then $\pi(x)p(x,y) = \pi(y)p(y,x)$ automatically.
   - **Thus, Metropolis-Hastings preserves detailed balance!**

What about ergodicity of Metropolis-Hastings?

1. **Recall:** ergodicity is implied by (a) irreducibility; (b) aperiodicity; (c) π stationary. *We've just proven (c) via detailed balance.*

2. **Irreducibility + Aperiodicity:** dependent on our choice of *proposal distribution* $q(x, \cdot)$.

   - So long as our random walker can go from any state x to any other state y with nonzero probability (eventually, not necessarily in one step), then irreducibility is satisfied.

   - For aperiodicity, it suffices to show that at least one state has a self-loop (i.e., has nonzero probability of staying in place) — either $q(x, x) > 0$ for some or all x, and/or some proposals are rejected with positive probability.

3. If (a) irreducibility; (b) aperiodicity; and (c) π stationary, then our Metropolis-Hasting algorithm's outputted samples will converge in the limit to the target distribution π, *regardless of initial condition* X_0.

More Powerful MCMC Algorithms

1. **What's the catch with Metropolis-Hastings?** A few points:
   - From animation, Metropolis-Hastings explores the state space very slowly: you need to wait for the random walker to slowly make its way there.
   - This was a toy 10x10 grid. What if our state space were continuous and/or high-dimensional (e.g., $D = 10000+$)?
   - Formally, Metropolis-Hasting's <convergence rate> is non-zero, but very slow.
   - Autocorrelation: even though we output n samples, they are not independent — in fact, they are highly, highly correlated: "inefficient sampling."
2. **Idea**: use a more intelligent proposal distribution that leverages information about π itself. I'm a huge fan of MALAtang 🤓
   - **Metropolis-Adjusted Langevin Dynamics** (continuous): same accept/reject step, but propose $y = x_{t-1} + \epsilon \nabla_x \log \pi(x_{t-1}) + \sqrt{2\epsilon}\xi_t$ for $\xi_t \sim \mathcal{N}(0, I)$. "*Metropolis-Adjusted" is just a fancy term for accept/reject step.*
   - **Intuition:** proposal is
     A. Gradient ascent with step-size ϵ to move towards high-density regions.
     B. Noisy perturbation to ensure ergodicity (irreducibility, intuitively).
   - **Impact:** much faster exploration of state space than Metropolis-Hastings with Normal random walk.
3. **Even fancier MCMC algorithms** (continuous**)**:** Hamiltonian Monte Carlo (HMC), No U-Turn Sampler (NUTS) — same core idea as MALA, but with much fancier physics integration to decrease autocorrelation + promote even faster state-space exploration.

Questions?

Happy to chat after class, too.