GAUTHAM NARAYAN

ASTR 496: FOUNDATIONS OF DATA SCIENCE IN ASTRONOMY

EFFECTIVE SAMPLING TECHNIQUES, WEEK 6

HERE'S WHERE WE ARE AT

CLASS SCHEDULE FALL 2025 (subject to revision)

- **Aug 28**
  First steps, crash course in python. **NO CLASS AUG 26.**
- **Sep 2, 4**
  Probability distributions, descriptive statistics, the Central Limit theorem and when it doesn't hold, robust statistics, and hypothesis testing (ICVG Ch. 3, FB Ch. 2). **CLASS OVER ZOOM THIS WEEK.**
- **Sep 9, 11**
  Statistical inference, frequentist properties such as unbiasedness & the Cramér–Rao bound, consistency, asymptotic limits, mean-squared errors (ICVG Ch. 4, FB Ch. 3)
- **Sep 16, 18**
  Maximum likelihood estimation and applications, ranting about minimizing χ^2 (ICVG Ch. 4). **CLASS OVER ZOOM ON SEP 18.**
- **Sep 23, 25**
  Regression & Inference: ordinary least squares, generalized least squares, orthogonal distance regression vs generative modeling of data (ICVG Ch. 8, FB Ch. 7)
- **Sep 30, Oct 2**
  Bayes in practice, sampling and Markov Chain Monte Carlo methods (ICVG Ch. 5)
- **Oct 7, 9**
  Building models, effective sampling techniques, estimating parameters & uncertainties, posterior predictive checks, other MCMC wizardry (ICVG Ch. 8). **Midterm exam posted.**

RECAP

- You should be comfortable with the basic idea behind MCMC sampling
  - **rejection sampling** and Metropolis-Hastings:
    - Random walks are robust but inefficient - suppress random walk behavior to improve efficiency at the cost of complexity (interpretability) and applicability
  - **burn-in**
  - **reversibility**
  - **ergodicity**
  - **diagnostics** for convergence, mixing and number of independent samples, visualizing a corner plot
  - This statement should now make sense:
    "Well that's easy, MCMC generates samples from the posterior distribution by constructing an ergodic, reversible Markov-chain that has as its equilibrium distribution the target posterior distribution. Questions?" - Thomas Wiecki

MCMC VARIANTS

- **Affine-invariant MC (emcee)** - works great as long as posterior is "nice" after affine transformation
  - counter-examples: Rosenbrock function, eggbox
- **Parallel-tempering (now, ptemcee)** - adds chains at multiple temperatures (we care about T=1)
    - connection to simulated annealing
    - computationally more intensive, even with a low number of dimensions

QUESTIONS ABOUT ANYTHING THUS FAR?

RECALL

Once you've picked an algorithm and random seed, provided your likelihood and priors don't change, Markov Chains are specified by two things: **Starting Position and Transition Probability**

$$=x$$

$$X_1, X_2, \ldots, X_n, X_{n+1}, X_n, \ldots$$

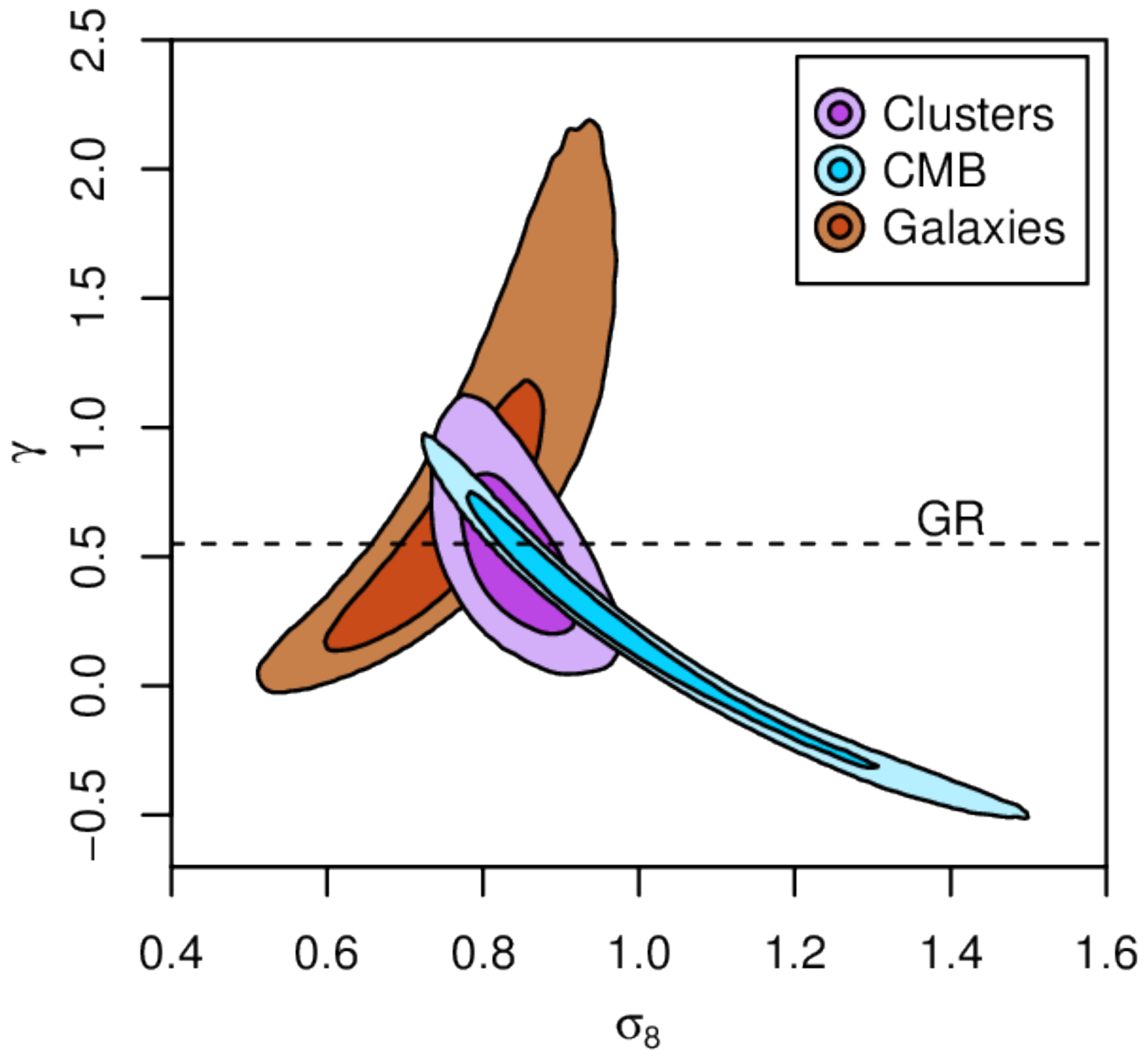
$$P(X_{n+1} \mid X_n = x) = P(X_{n+2} = x \mid X_{n+1})$$

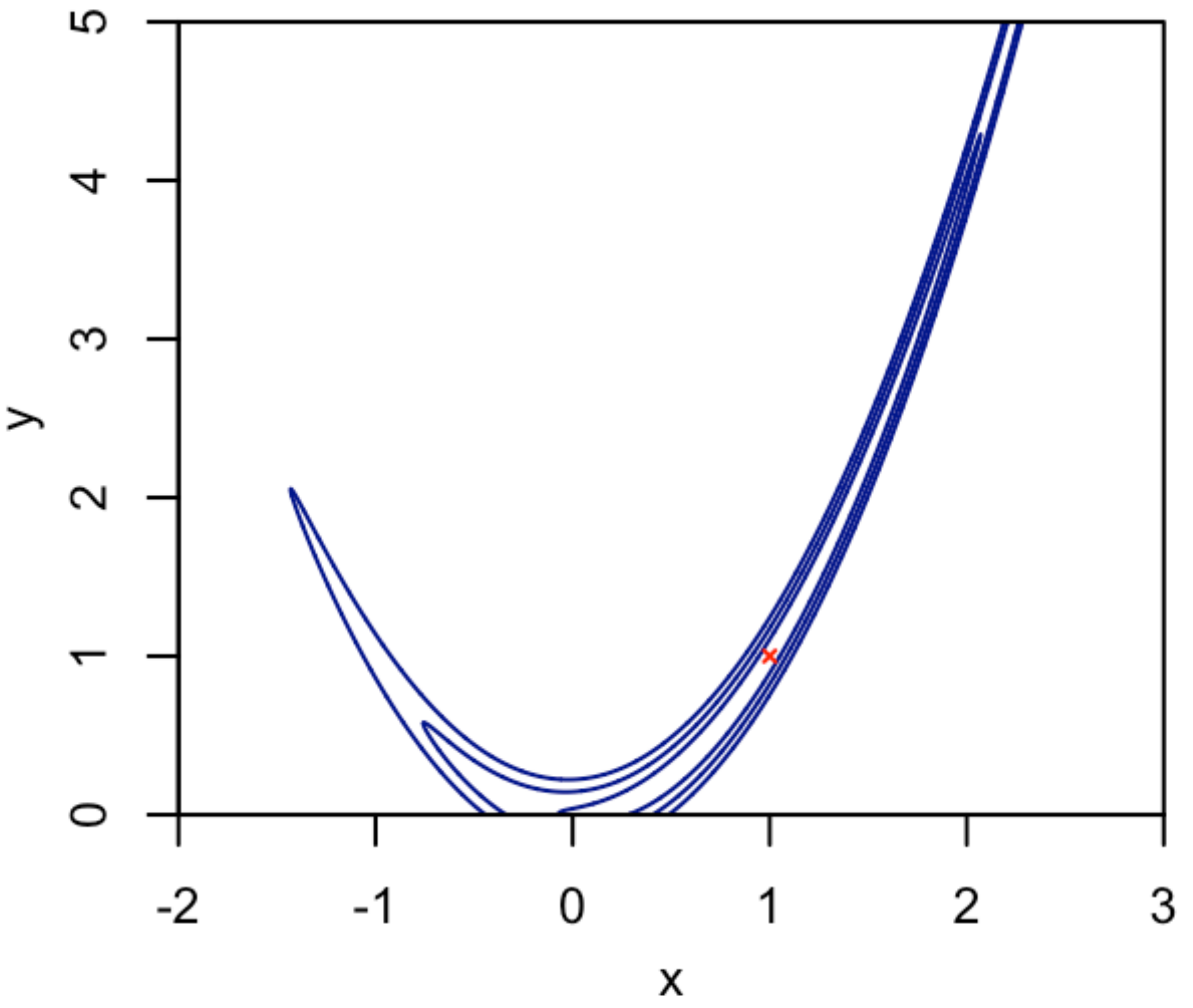
$\sim$*time-reversal invariant*

EMCEE/GOODMAN–WEARE WALKERS

- Unlike regular Metropolis-Hastings, you **don't have to specify a step-size** (just the initial positions of the walkers).
- Algorithm starts many walkers at different positions, and the transition probabilities are set by ensemble of walkers - this allows the algorithm to adapt to linear-rescaling along any dimension: "affine-invariant"
- What do we do if the posterior can't be rescaled linearly?

NON-LINEAR POSTERIORS AREN'T HYPOTHETICAL FOR ASTROPHYSICS





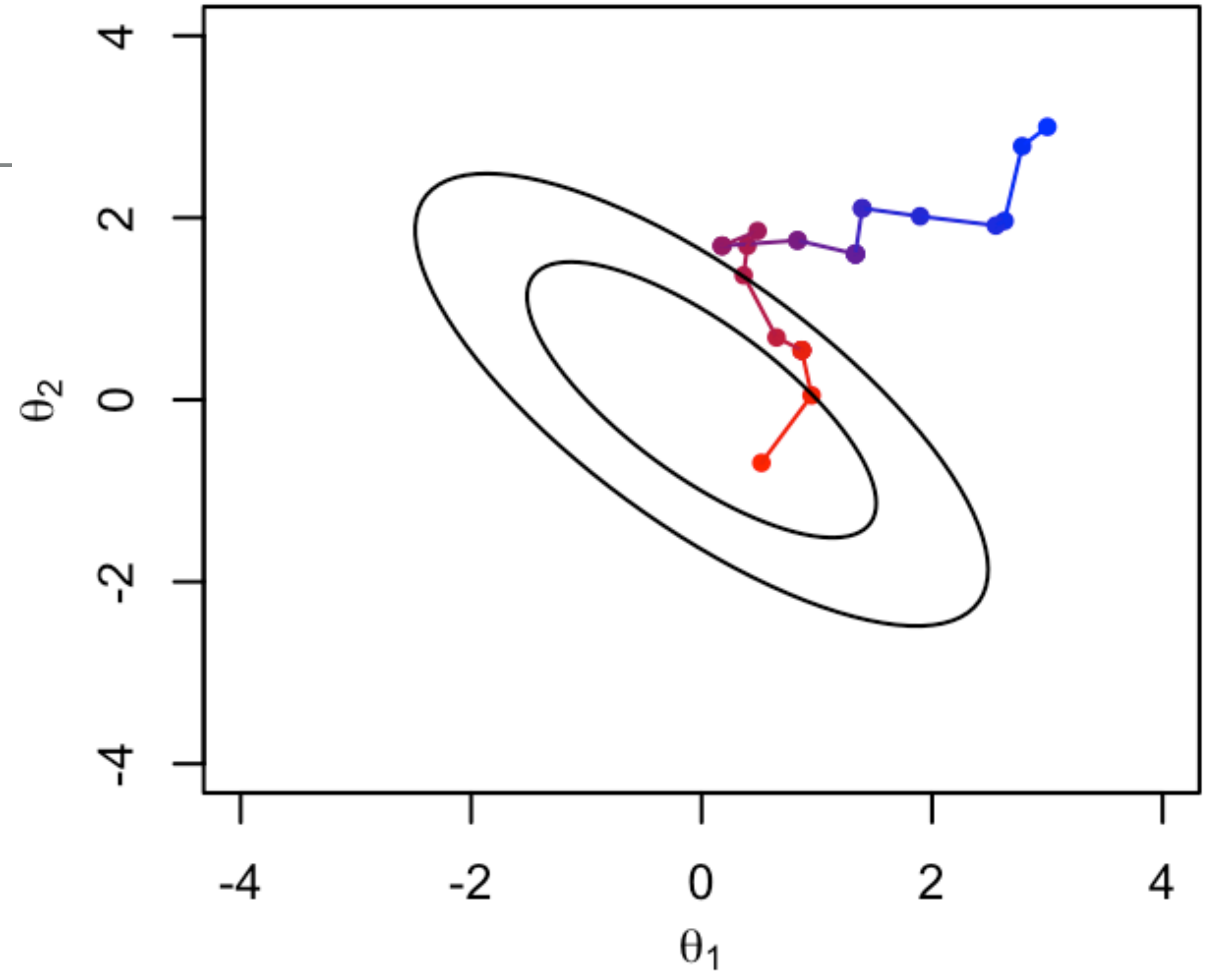
MH can be very ineffective here because there is **no one right step size.**
Affine-invariant is good for figuring out the step size that works with **linear transformations** of the posterior, but the **posterior here can't be transformed linearly.**

SPEEDING THINGS UP

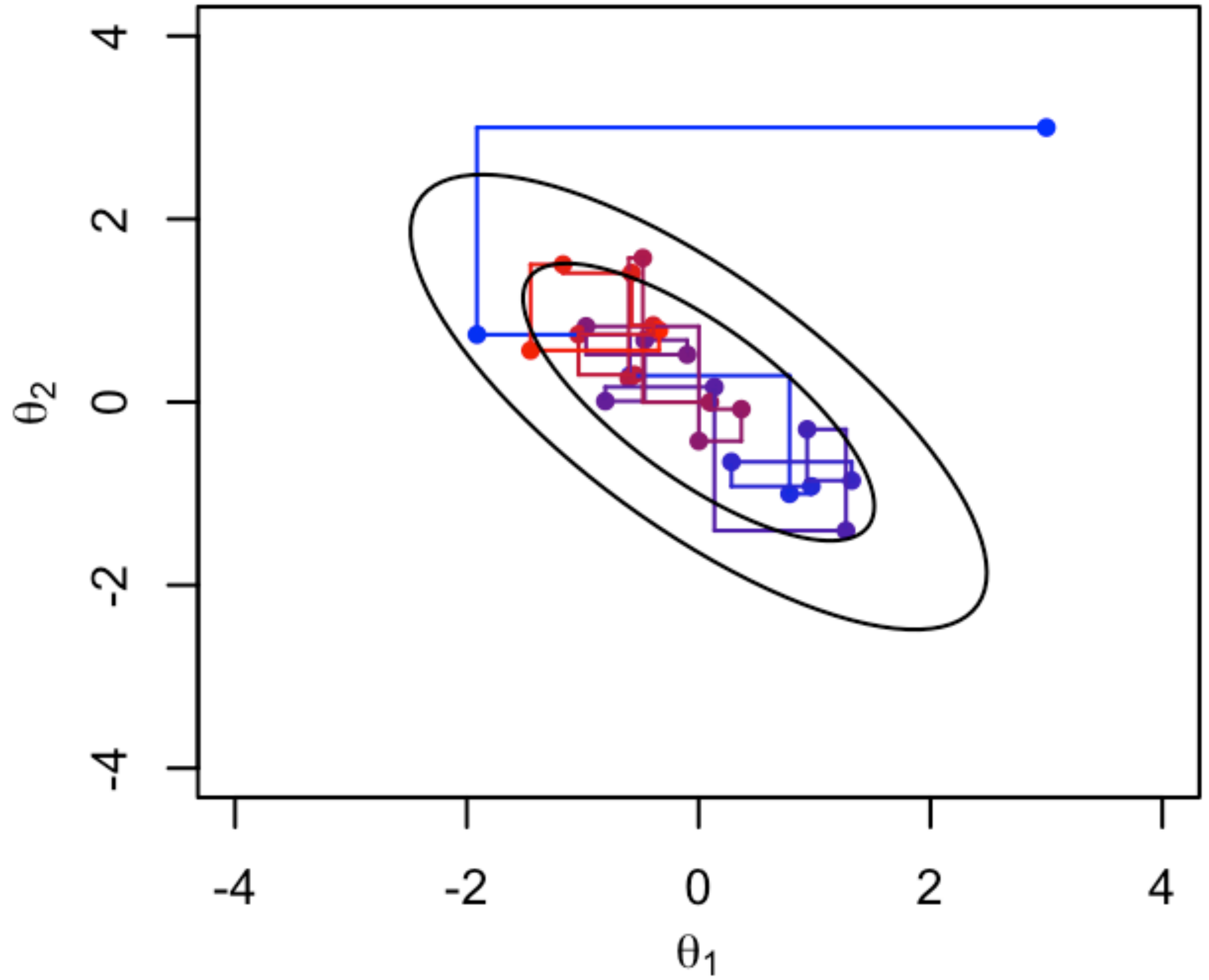
- Broadly speaking, we can try to:
  - tailor algorithms to specific classes of PDF
  - look for ways to make the general samplers more intelligent
- **We can also use different samplers for different subsets of parameters** - the only rule is that every parameter must get updated somehow.
- **Gibbs Sampling** is a specialization of Metropolis-Hastings:
  - Instead of making a general proposal in all dimensions, we cycle through the parameters proposing changes to one at a time
  - A proposal for θ_i is from the fully conditional posterior $p(\theta_i|\theta_{-i},x)$, where $-i$ means all subscripts other than i.

GIBBS SAMPLING

- while we want more samples
- propose theta1 | theta2, theta3, …, data
- accept/reject theta1
- propose theta2 | theta1, theta3, …, data
- accept/reject theta2
- …



- **See difference vs regular Metropolis-Hastings at right**
- Why is a random drunk walking in one specific direction at a time better than just taking a random step???
- In general, this is not obviously an improvement to proposing changes to all θ simultaneously.



WHY/WHEN DOES GIBBS SAMPLING WORK

- **Something interesting happens if the fully conditional likelihood and prior are conjugate**
- For some likelihood functions, if you choose a certain prior, **the posterior ends up being in the same distribution as the prior**. Such a prior then is called a **Conjugate Prior.**
- i.e.
- $P(\theta)$ such that $P(\theta|D)=P(\theta)$
- i.e. we know the conditional posterior exactly!

USING INDEPENDENT SAMPLES FROM THE POSTERIOR AS PROPOSALS

- If we use independent samples of the conditional posterior as proposals, then the Metropolis-Hastings acceptance ratio becomes

$$\frac{p(x')g(x \mid x')}{p(x)g(x' \mid x)} = \frac{p(x')p(x)}{p(x)p(x')} = 1$$

- and every proposal is automatically accepted! i.e.
- draw theta1 from p(theta1| theta2, theta3, ..., data)
- draw theta2 from p(theta2| theta1, theta3, ..., data)

SO WHAT ARE CONJUGATE PRIORS

- Beta posterior
- Beta prior * Bernoulli likelihood → Beta posterior
- Beta prior * Binomial likelihood → Beta posterior
- Beta prior * Negative Binomial likelihood → Beta posterior
- Beta prior * Geometric likelihood → Beta posterior
- **Gamma posterior (often useful for population selection effects)**
- **Gamma prior * Poisson likelihood → Gamma posterior**
- **Gamma prior * Exponential likelihood → Gamma posterior**
- **Normal posterior (often useful for fitting)**
- **Normal prior * Normal likelihood (mean) → Normal posterior**

WHEN YOU CAN USE GIBBS, IT'S THE BEST CHOICE

- Gibbs Sampling Pros:
  - No cycles "wasted" on rejected proposals
  - No pesky tuning of the proposal scale

- Gibbs Sampling Cons:
  - Only works for conjugate or partially conjugate models (hence must choose conjugate priors)

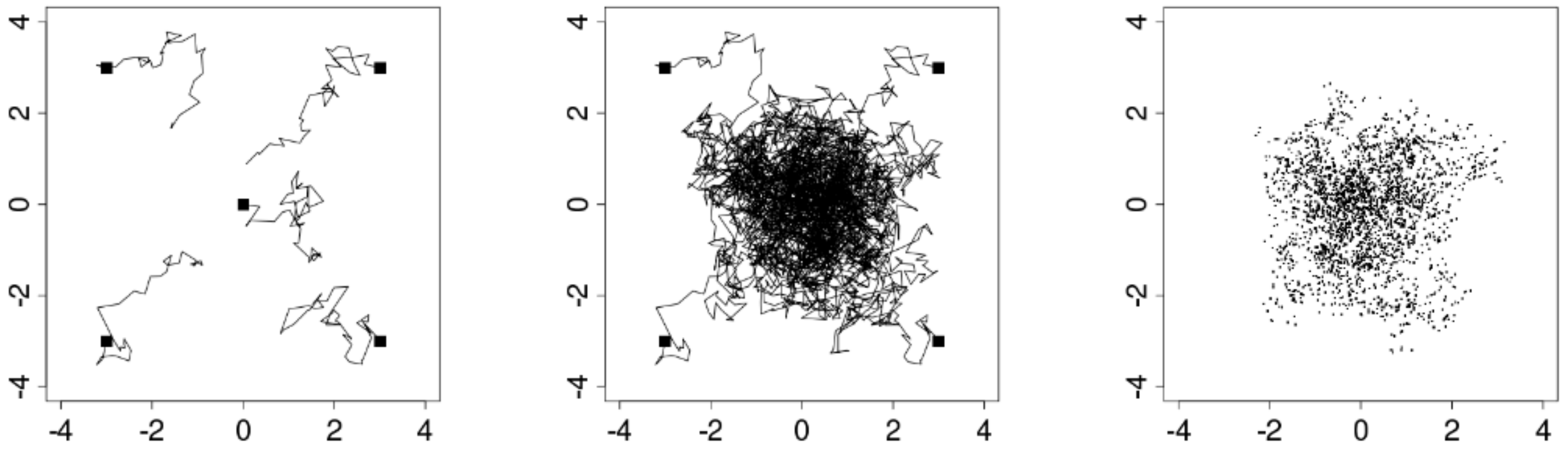
http://chi-feng.github.io/mcmc-demo/app.html?algorithm=RandomWalkMH&target=banana

http://chi-feng.github.io/mcmc-demo/app.html?algorithm=GibbsSampling&target=banana

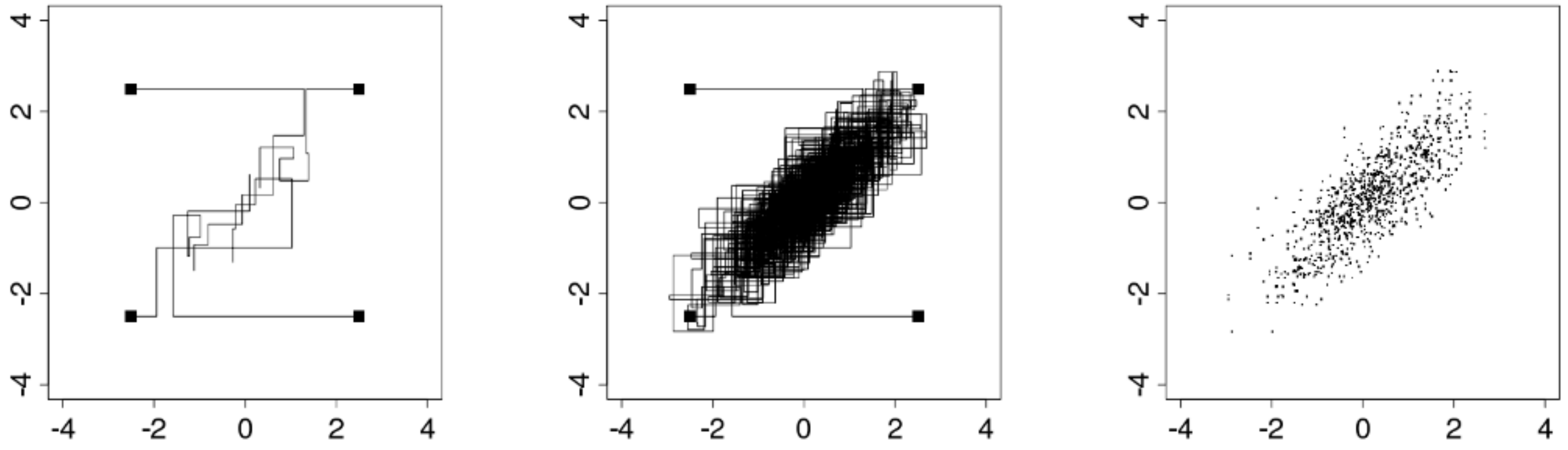
ADDING DERIVATIVES

- Thus far, all of the methods we've looked at are popular with astrophysicists for a simple reason - we need to specify only likelihoods and priors, NOT THEIR DERIVATIVES
  - recall, optimizers could use the gradient information if available to speed up finding a solution. Can we do something similar to MCMC?
- Reducing the correlation between successive states is key to improving the accuracy of MCMC approximations.
- MCMC samplers tend to exhibit so-called "random walk" behavior - meander to and fro as they sample from the target distribution.
- Using well-chosen transformations and large moves can improve mixing performance (e.g. Affine-invariant or "Stretch moves" or Gibbs sampling, but often they are hard to construct for complex distributions on high-dimensional spaces.

Random walk behavior of Metropolis-Hastings on a bivariate normal target distribution



Random walk behavior of Gibbs sampling on a bivariate normal target distribution



(figure from Gelman et al. (2013), BDA3, Chapter 11)

HAMILTONIAN MONTE-CARLO

- Hamiltonian Monte Carlo (HMC) employs a **dynamics approach** to more quickly traverse the space and thus improve MCMC mixing
- Assume we can compute the gradient of the log density, **∇** log p(x).
  - **Analogous to gradient-based optimization methods, HMC uses gradients to improve MCMC mixing.**
  - Sample an auxiliary variable $v \in R_d$ where $v_i|x \sim N(0, 1/m_i)$ independently for i = 1, . . . , d. This might seem like a nuisance parameters. I'm using $1/m_i$ but I really just mean some sigma of a Gaussian
  - Jointly transform (x, v) in a way that leaves p(x, v) roughly constant by using Hamiltonian dynamics
  - Use a Metropolis–Hastings step to accept or reject the transformed (x, v)

PHYSICAL INTERPRETATION

- The transformation of (x, v) is done by running a dynamical system with Hamiltonian H(x, v) forward in time, where

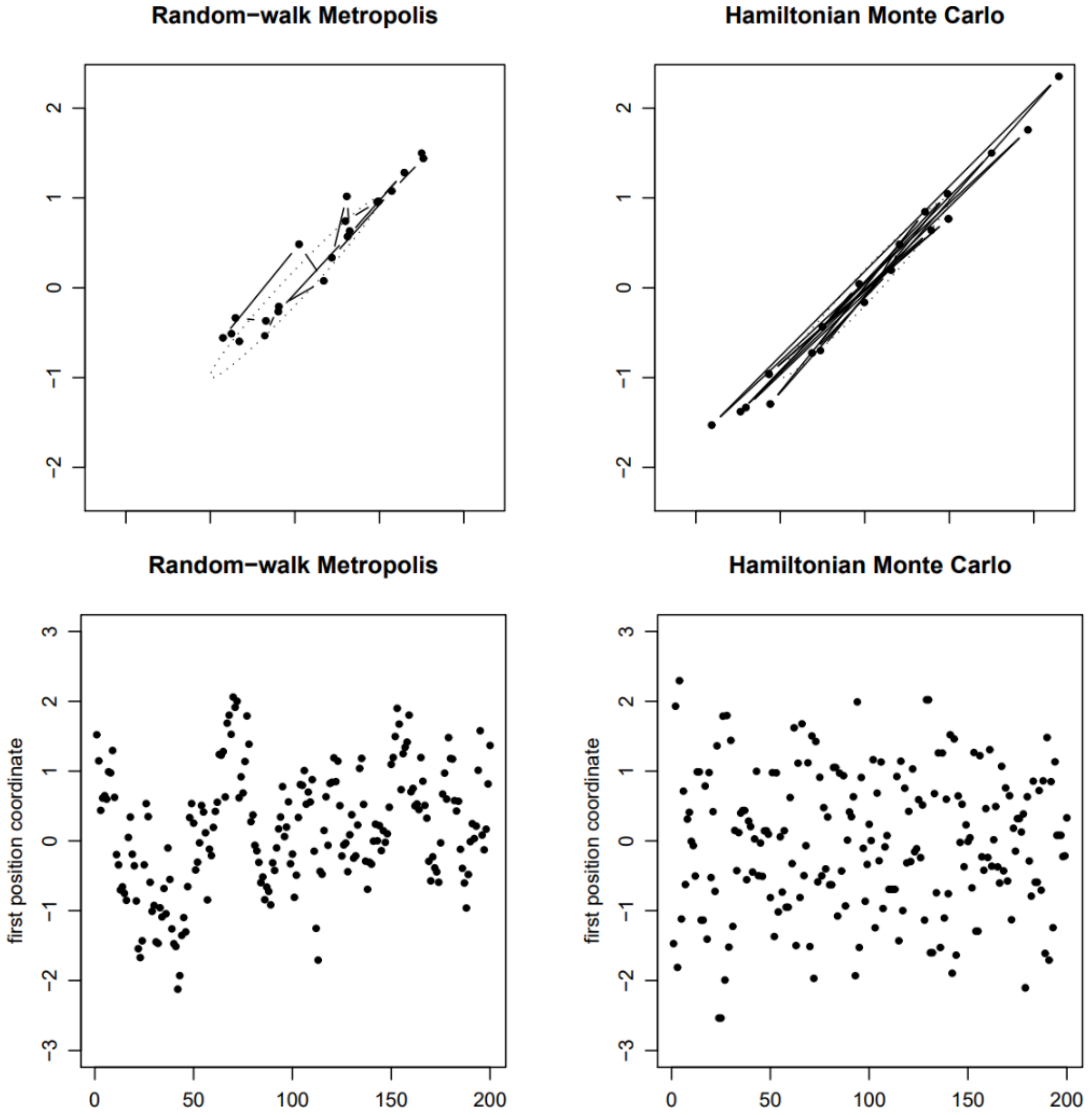
$$H(x, v) = -\log p(x) + \frac{1}{2}\sum_{i=1}^{d} m_i v_i^2$$

- Intuition: x moves like a ball rolling on the surface – log p(x)
- Physical interpretation:
- $x_1, \ldots, x_d$ = position coordinates
- $v_1, \ldots, v_d$ = momemntum coordinates
- -log p(x) = potential energy
- $\sum_i \frac{1}{2} m_i v_i^2$ = kinetic energy

PHYSICAL INTERPRETATION

- The Hamiltonian represents the total energy of the system:
  - H = Total energy = Potential energy + Kinetic energy.
- By conservation of energy, H remains constant as the dynamical system evolves over time.
- Thus, $p(x, v) \propto \exp(-H(x, v))$ also remains constant as (x, v) evolves according to the dynamical system.
- To gain some intuition for how the system evolves, first suppose $p(x)$ is flat in some region. Then $\nabla \log p(x) = 0$, so there is zero acceleration and consequently, x will move at constant velocity through this region.
- Meanwhile, if $p(x)$ is not flat, then force $= \nabla \log p(x)$ means that x is accelerating in the direction of the gradient, i.e., it is accelerating towards a region of higher density

http://arogozhnikov.github.io/2016/12/19/markov_chain_monte_carlo.html


Random−walk Metropolis
Hamiltonian Monte Carlo
Random−walk Metropolis
Hamiltonian Monte Carlo
first position coordinate
first position coordinate

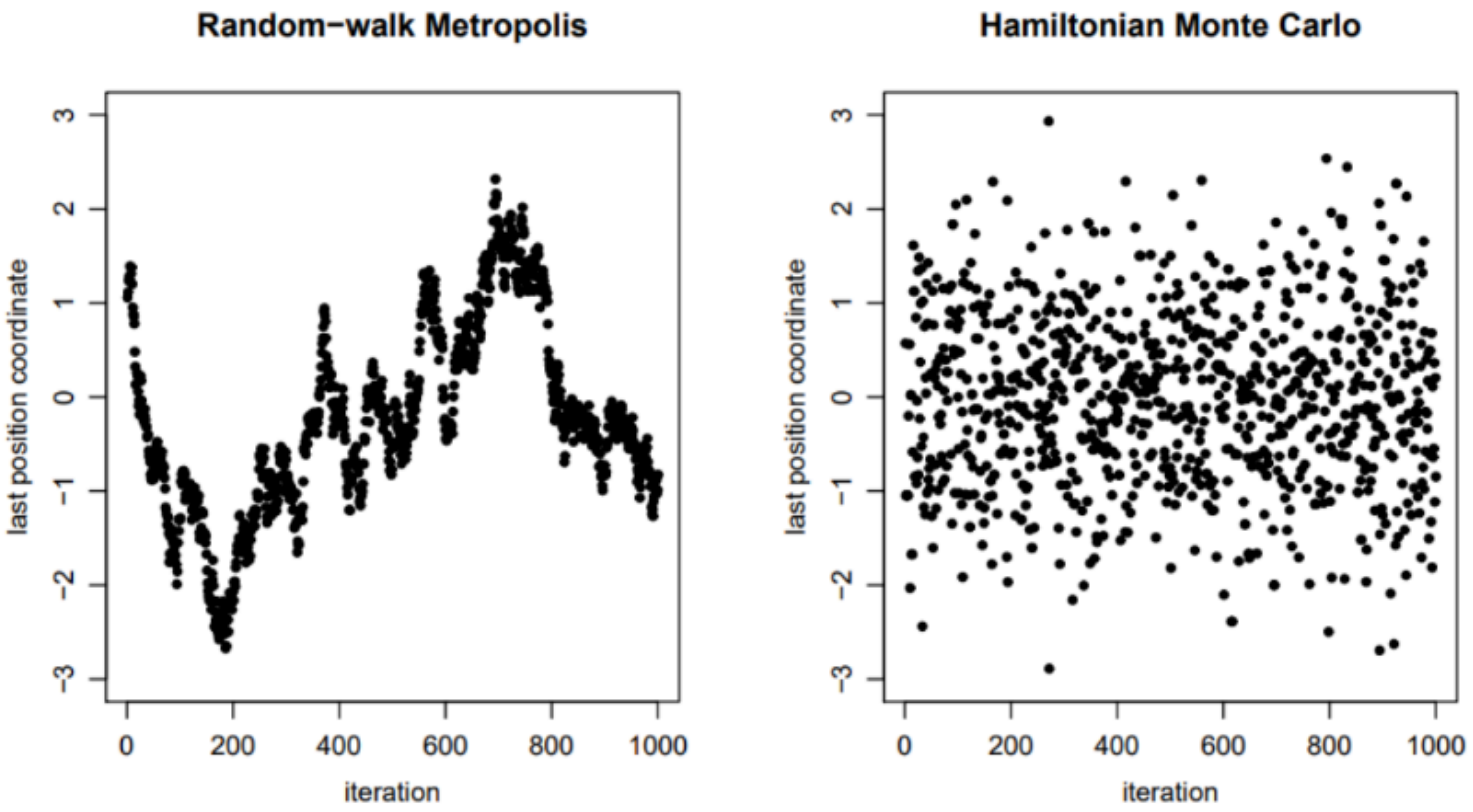


Figure 6: Values for the variable with largest standard deviation for the 100-dimensional example, from a random-walk Metropolis run and an HMC run with $L = 150$. To match computation time, 150 updates were counted as one iteration for random-walk Metropolis.

(figure from Neal (2011))