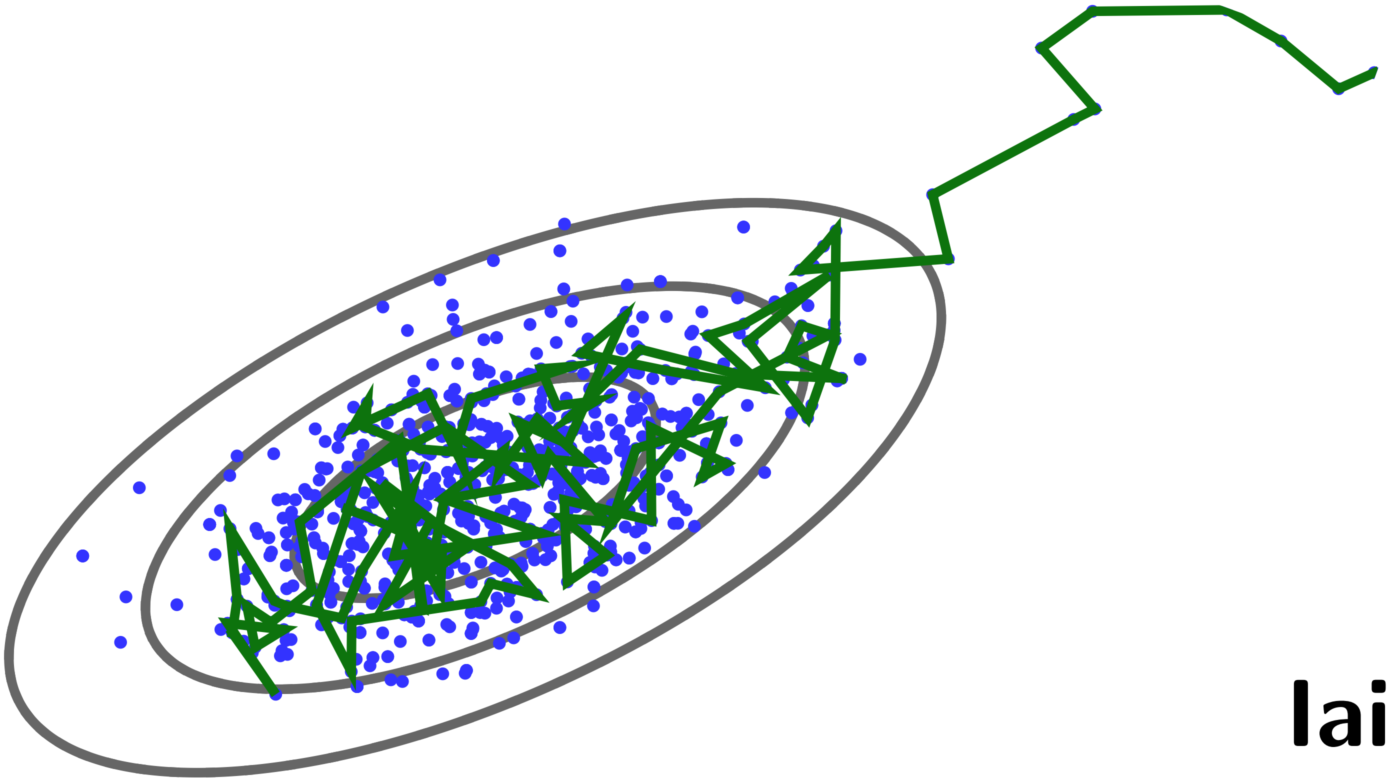
Monte Carlo Inference Methods

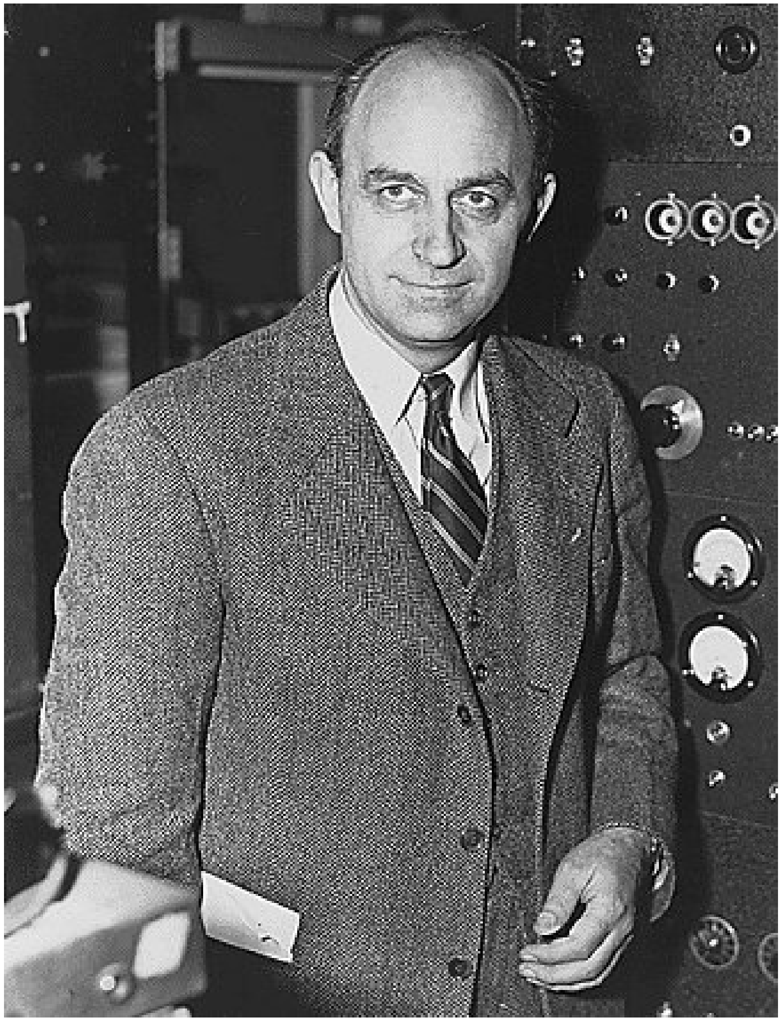


Iain Murray

University of Edinburgh

http://iainmurray.net

Monte Carlo and Insomnia



Enrico Fermi (1901–1954) took great delight in astonishing his colleagues with his remarkably accurate predictions of experimental results. . .

. . . his “guesses” were really derived from the statistical sampling techniques that he used to calculate with whenever insomnia struck!

—*The beginning of the Monte Carlo method*, N. Metropolis

Overview

Gaining insight from random samples

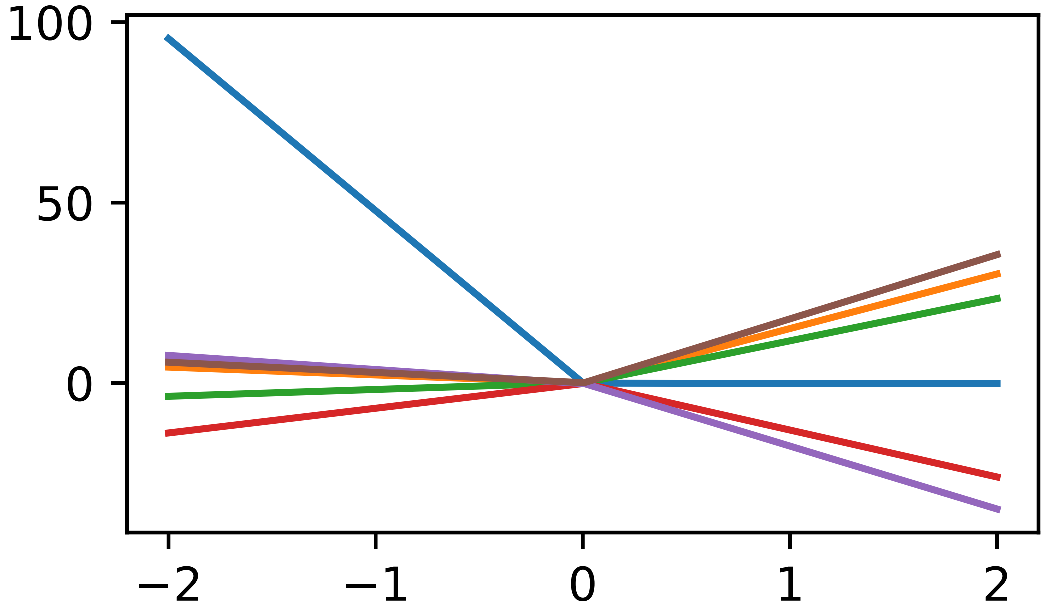
Inference / Computation

What does my data imply? What is still uncertain?

Sampling methods:

Importance, Rejection, Metropolis–Hastings, Gibbs

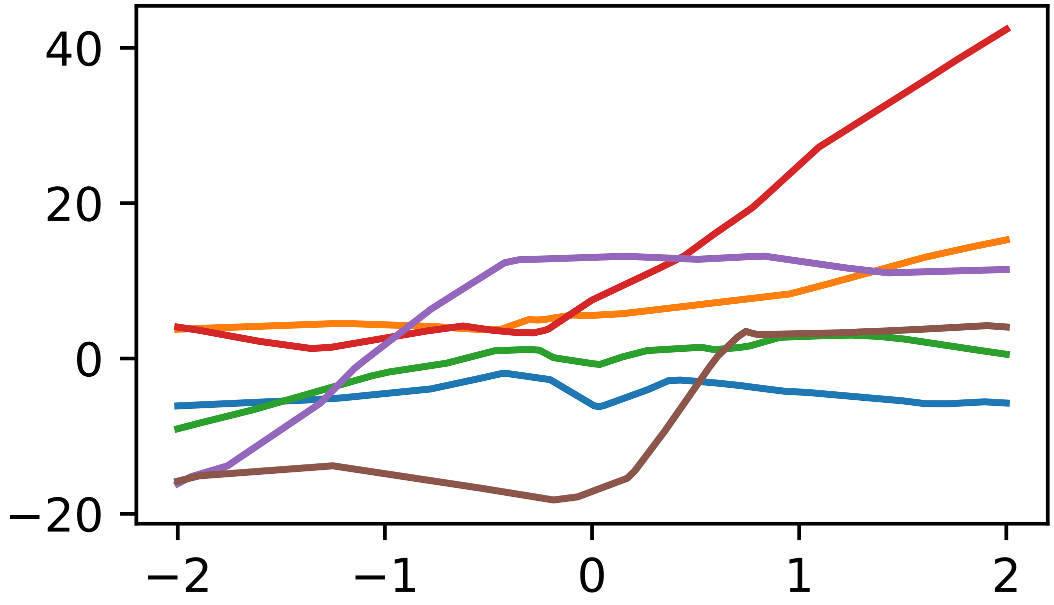
Practical issues / Debugging



```
import numpy as np
import matplotlib.pyplot as plt
plt.clf()

def relu(a): return np.fmax(0, a)
def linear(a): return a
def neural_net(X, activations, weights):
    for (act, W) in zip(activations, weights):
        X = act(np.dot(X, W))
    return X

xx = np.linspace(start=-2, stop=2, num=1000)[:, None]
for sample in range(6):
    randn = np.random.randn
    weights = [randn(1, 20), randn(20, 10), randn(10, 10), randn(10, 1)]
    activations = [relu] * (len(weights) - 1) + [linear]
    ff = neural_net(xx, activations, weights)
    plt.plot(xx, ff)
plt.show()
```



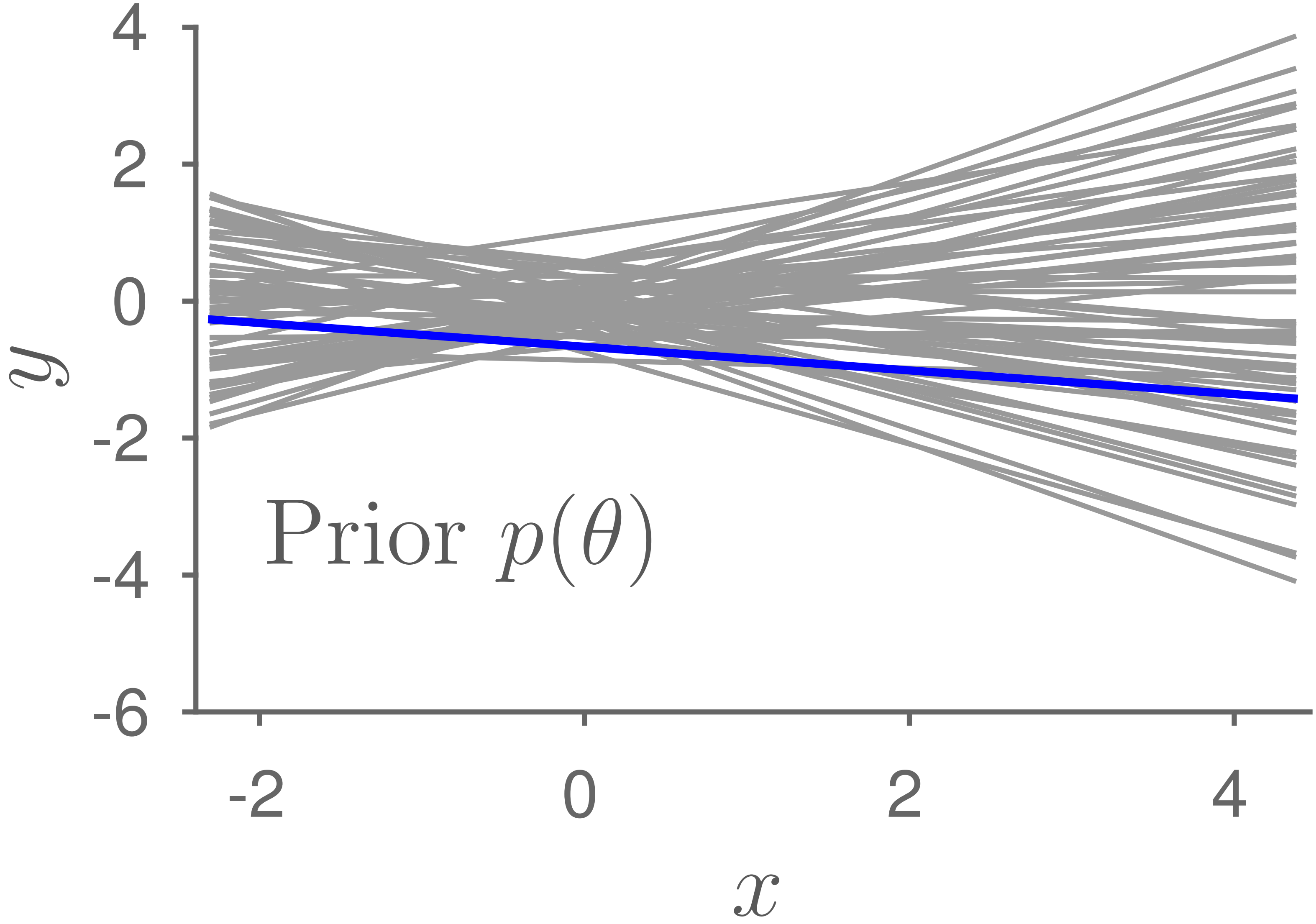
```
import numpy as np
import matplotlib.pyplot as plt
plt.clf()

def relu(a): return np.fmax(0, a)
def linear(a): return a
def neural_net(X, activations, weights, biases):
    for (act, W, bb) in zip(activations, weights, biases):
        X = act(np.dot(X, W) + bb[None, :])
    return X

xx = np.linspace(start=-2, stop=2, num=1000)[:, None]
for sample in range(6):
    randn = np.random.randn
    weights = [randn(1, 20), randn(20, 10), randn(10, 10), randn(10, 1)]
    activations = [relu] * (len(weights) - 1) + [linear]
    biases = [randn(W.shape[1]) for W in weights]
    ff = neural_net(xx, activations, weights, biases)
    plt.plot(xx, ff)
plt.show()
```

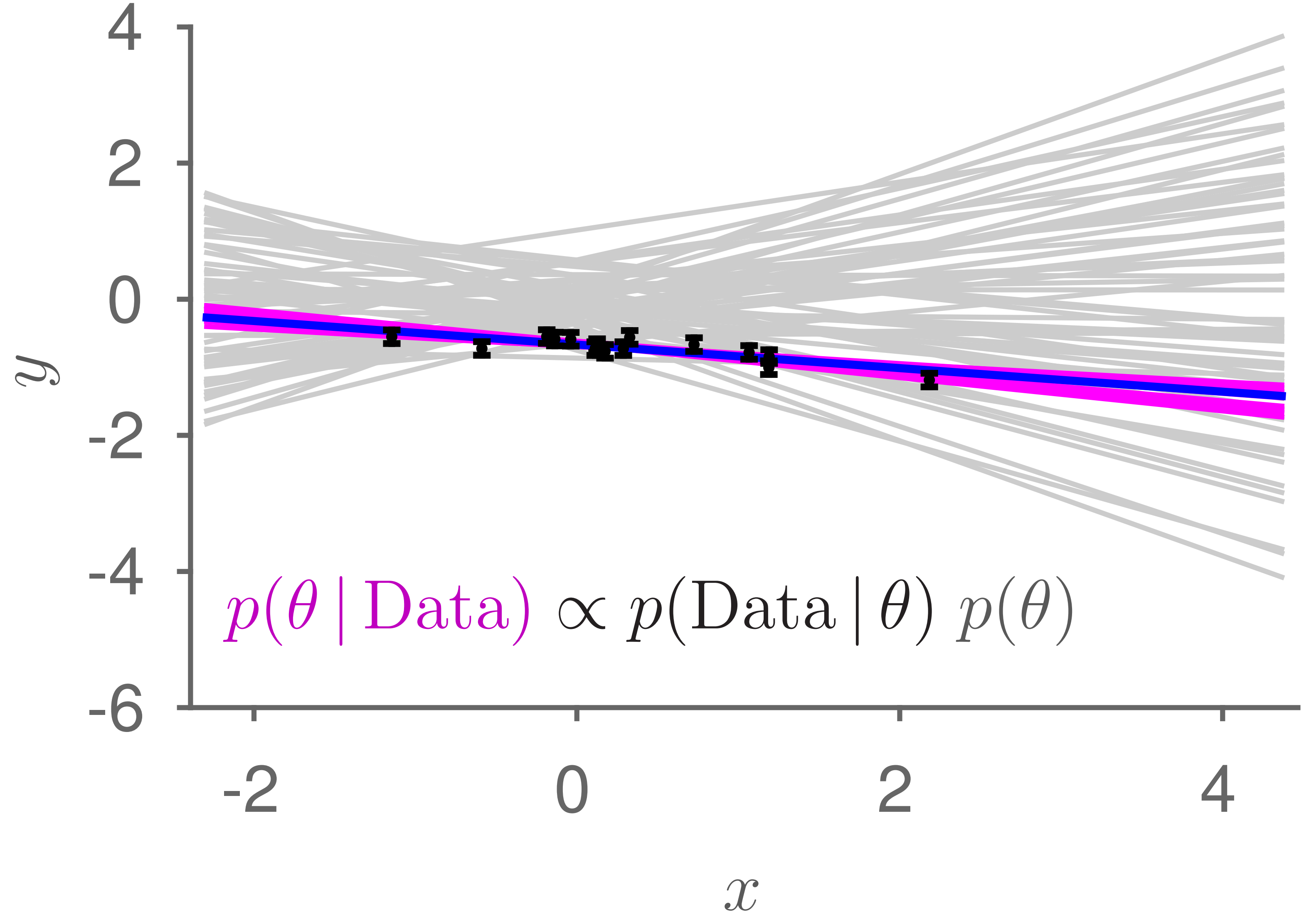
Linear regression

$y = \theta_1 x + \theta_2, \qquad p(\theta) = \mathcal{N}(\theta;\, 0, 0.4^2 I)$



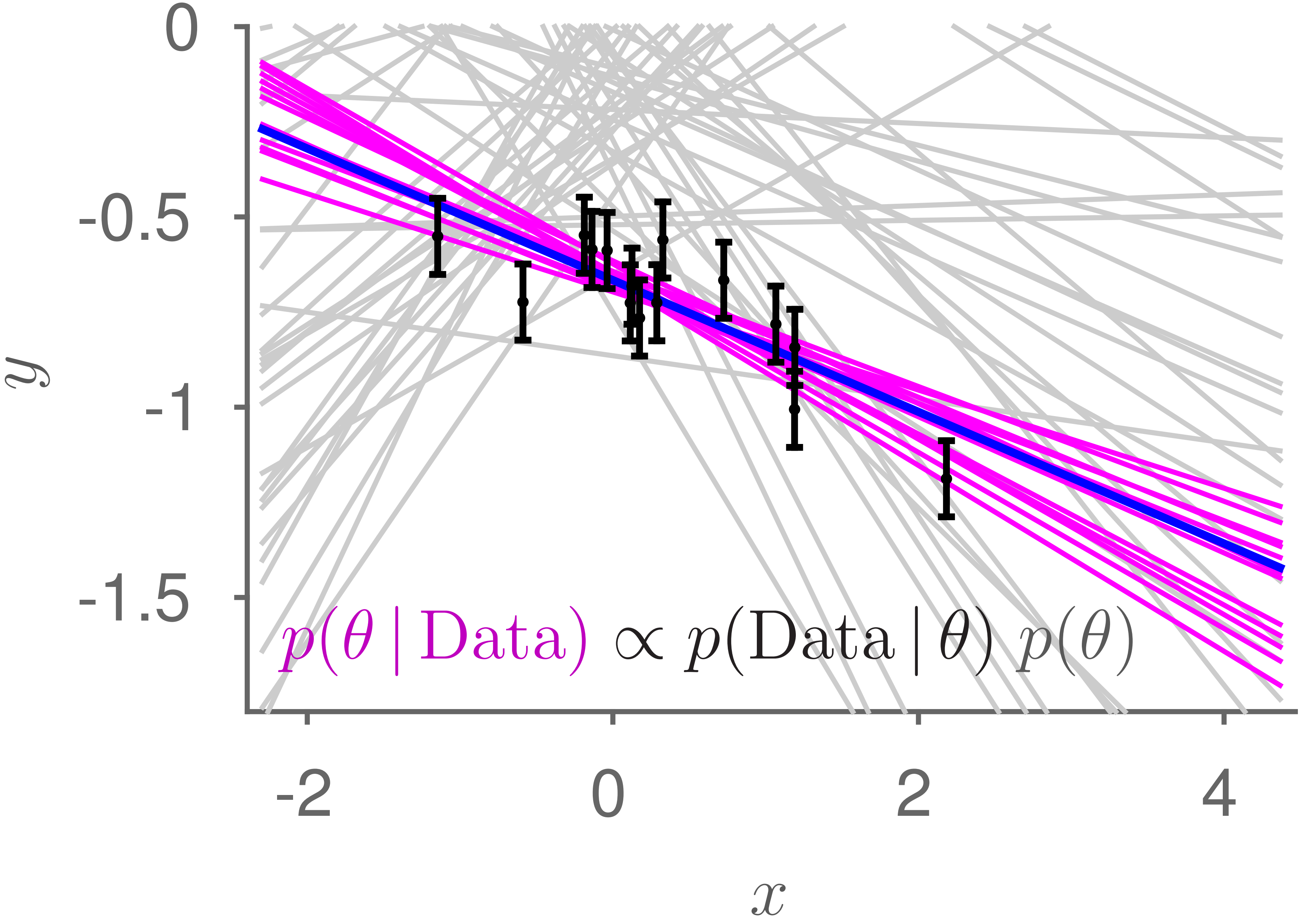
Linear regression

$$y^{(n)} = \theta_1 x^{(n)} + \theta_2 + \epsilon^{(n)}, \qquad \epsilon^{(n)} \sim \mathcal{N}(0, 0.1^2)$$

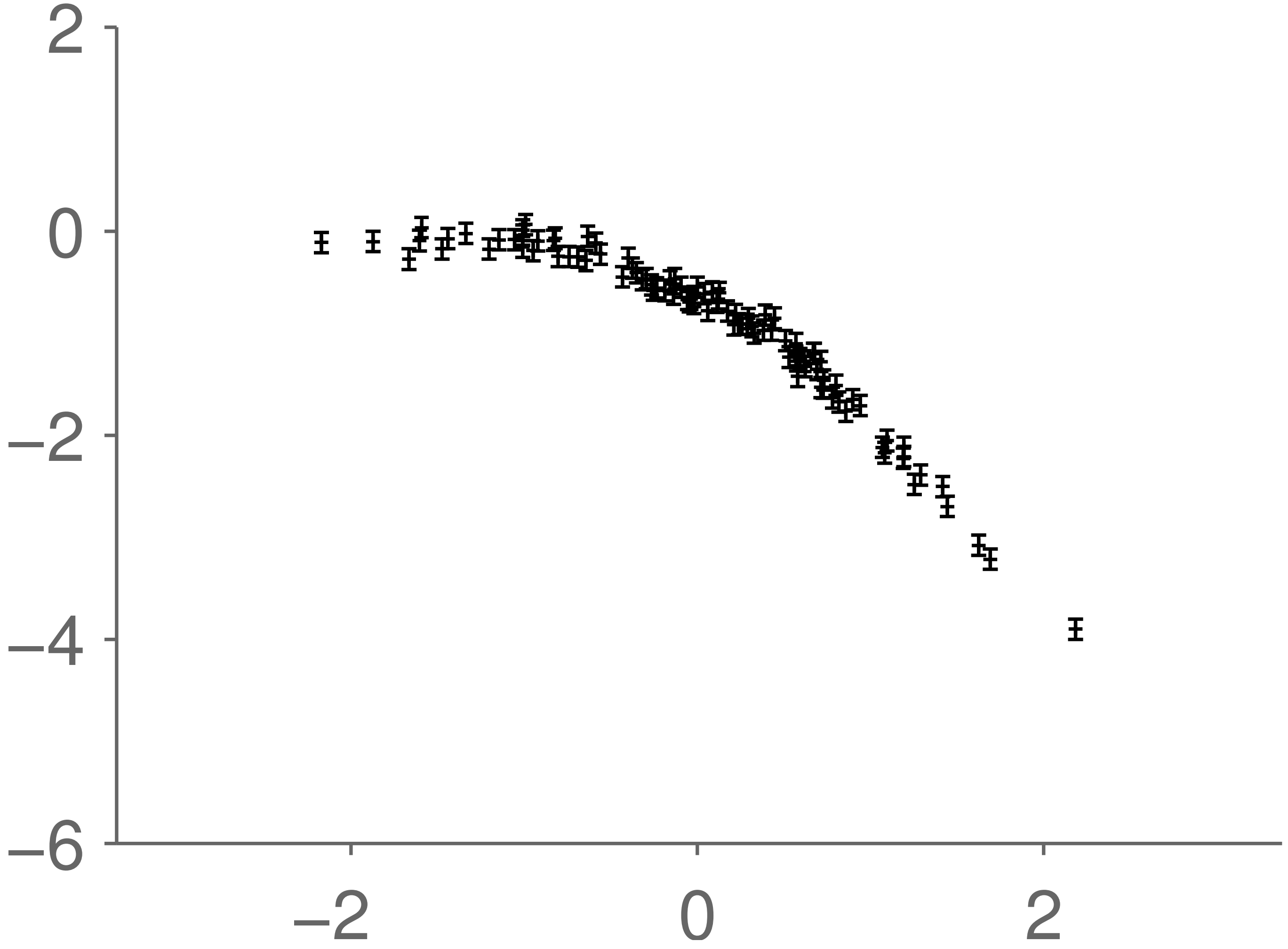


$$p(\theta \,|\, \mathrm{Data}) \propto p(\mathrm{Data} \,|\, \theta)\; p(\theta)$$

Linear regression (zoomed in)



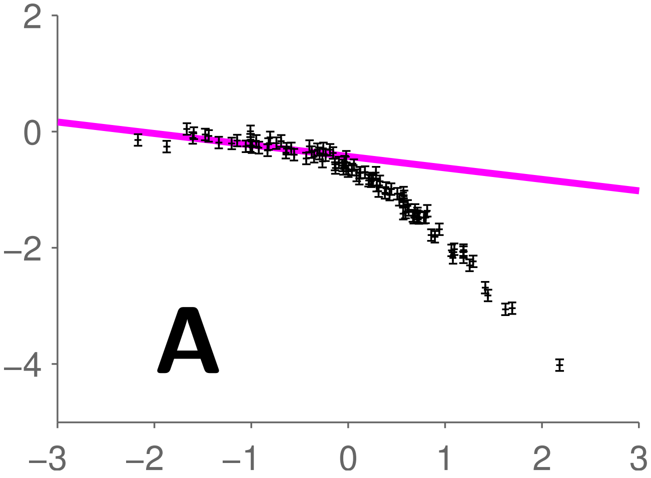
Model mismatch

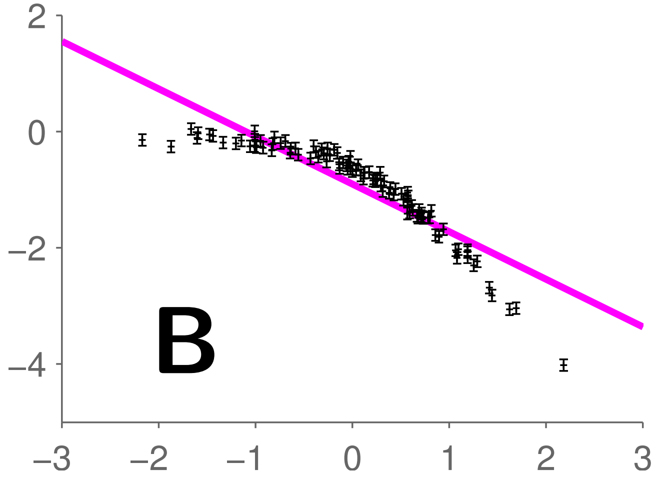


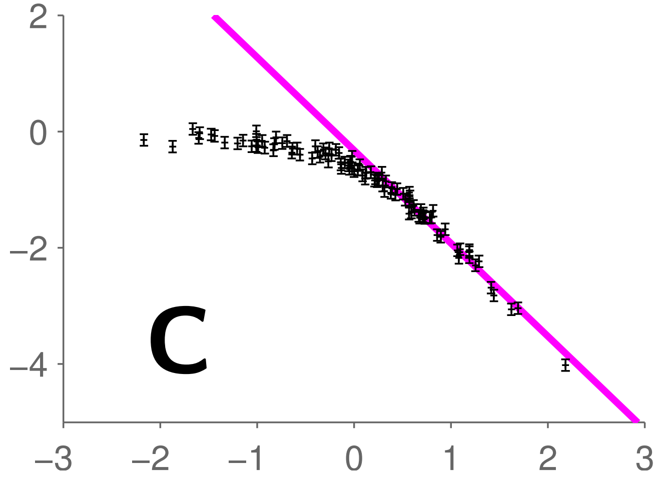
What will Bayesian linear regression do?

Quiz

Given a (wrong) linear assumption, which explanations are typical of the posterior distribution?





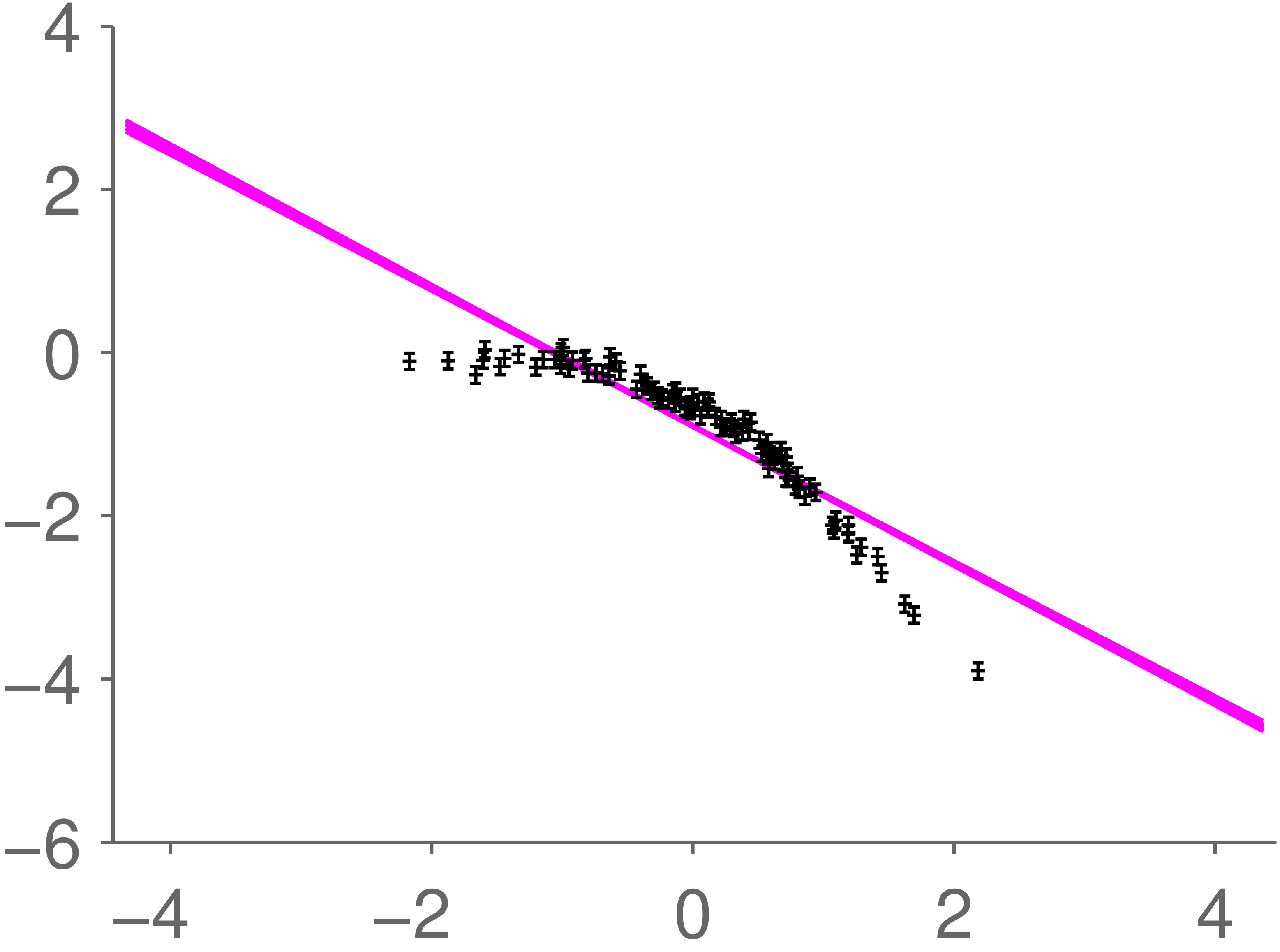


D All of the above

E None of the above

Z Not sure

‘Underfitting’



Posterior *very* certain despite blatant misfit. Peaked around least bad option.

Roadmap

— Looking at samples

— **Monte Carlo computations**

— How to actually get the samples

Simple Monte Carlo Integration

$$\int f(\theta)\, \pi(\theta)\, \mathrm{d}\theta = \text{“average over } \pi \text{ of } f\text{”}$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} f(\theta^{(s)}), \quad \theta^{(s)} \sim \pi$$

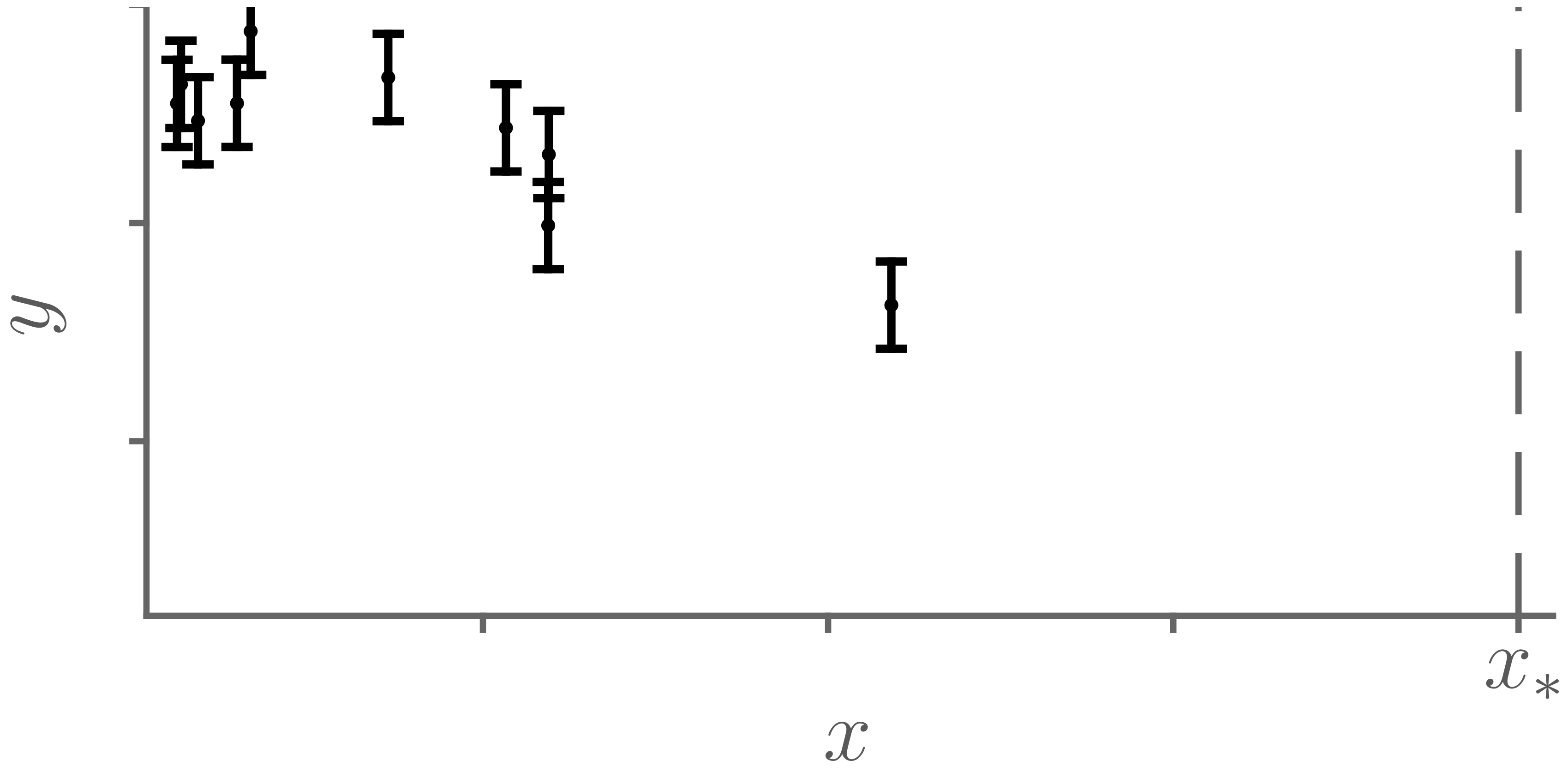
Unbiased

Variance $\sim 1/S$

Prediction

$$p(y_* \mid x_*, \mathcal{D}) = \int p(y_* \mid x_*, \theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta$$

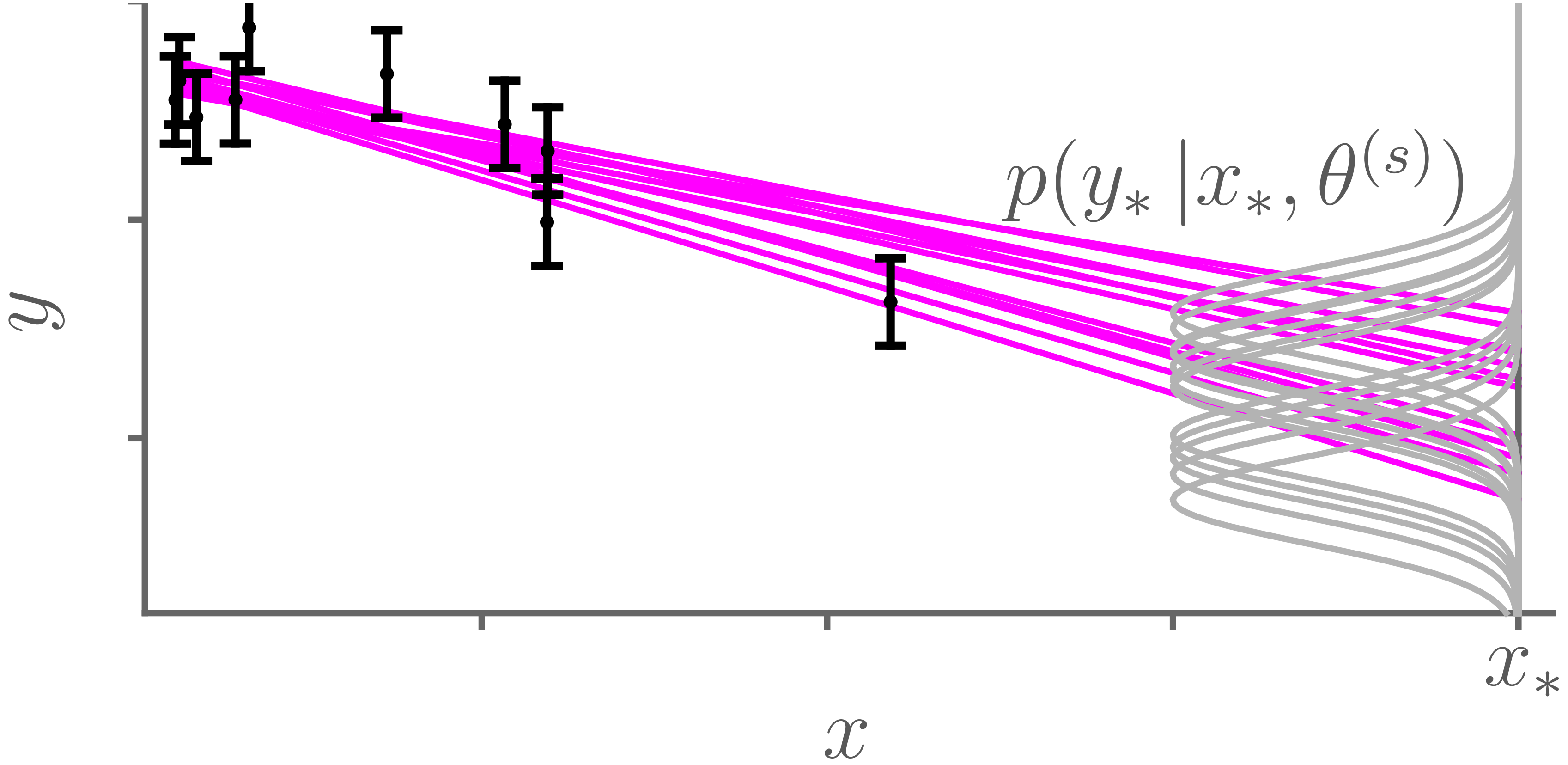
$$\approx \frac{1}{S} \sum_s p(y_* \mid x_*, \theta^{(s)}), \quad \theta^{(s)} \sim p(\theta \mid \mathcal{D})$$



Prediction

$$p(y_* \,|\, x_*, \mathcal{D}) = \int p(y_* \,|\, x_*, \theta)\, p(\theta \,|\, \mathcal{D})\, \mathrm{d}\theta$$

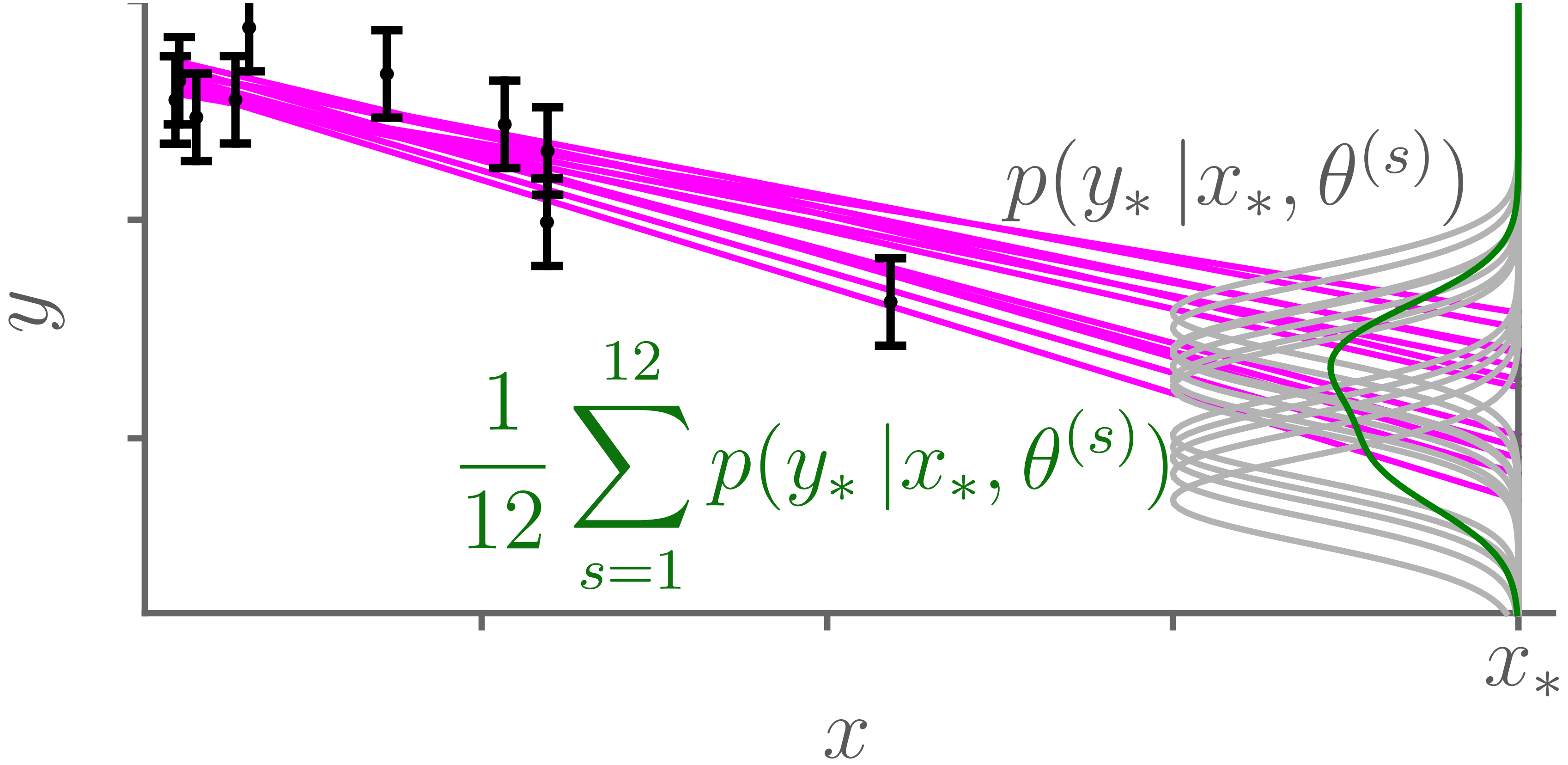
$$\approx \frac{1}{S} \sum_s p(y_* \,|\, x_*, \theta^{(s)}), \qquad \theta^{(s)} \sim p(\theta \,|\, \mathcal{D})$$



Prediction

$$p(y_* \,|\, x_*, \mathcal{D}) = \int p(y_* \,|\, x_*, \theta)\, p(\theta \,|\, \mathcal{D})\, \mathrm{d}\theta$$

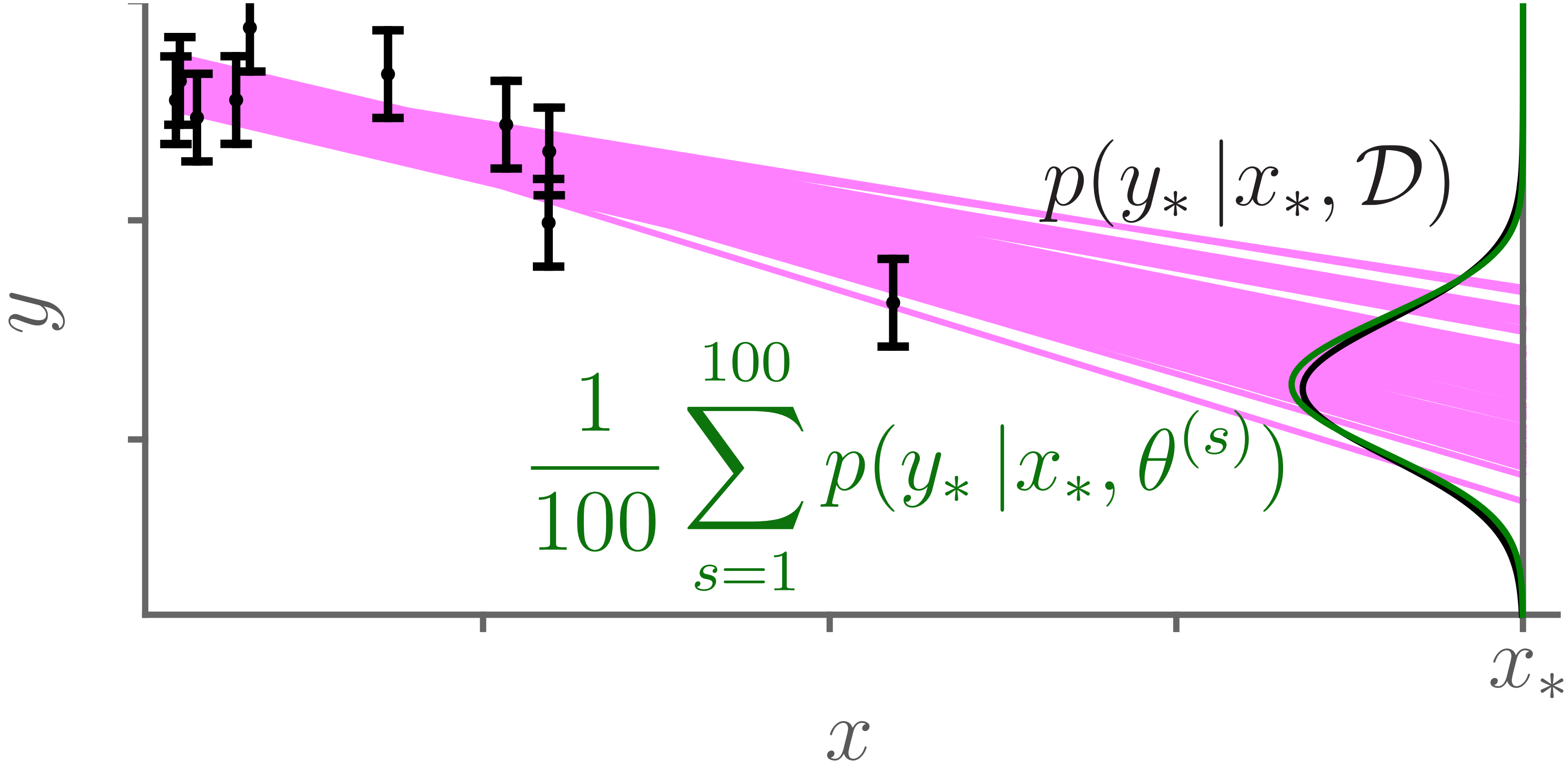
$$\approx \frac{1}{S} \sum_s p(y_* \,|\, x_*, \theta^{(s)}), \quad \theta^{(s)} \sim p(\theta \,|\, \mathcal{D})$$



Prediction

$$p(y_* \,|\, x_*, \mathcal{D}) = \int p(y_* \,|\, x_*, \theta)\, p(\theta \,|\, \mathcal{D})\; \mathrm{d}\theta$$

$$\approx \frac{1}{S} \sum_s p(y_* \,|\, x_*, \theta^{(s)}), \qquad \theta^{(s)} \sim p(\theta \,|\, \mathcal{D})$$



More interesting models

— Noisy function values: $y^{(n)} \sim p(y \,|\, f(\mathbf{x}^{(n)}; \mathbf{w}), \Sigma)$

— What weights and noise are plausible? $p(\Sigma)$, $p(\mathbf{w} \,|\, \alpha)$

— Some observations corrupted?

if $z^{(n)}\!=\!1$, then $y^{(n)} \sim p(y \,|\, 0, \Sigma_N)$, $\quad \ldots\, p(z\!=\!1) = \epsilon$

. . . A lot of choices. Joint probability:

$$\left[\prod_n p(y^{(n)} \,|\, z^{(n)}, \mathbf{x}^{(n)}, \mathbf{w}, \Sigma, \Sigma_N)\, p(z^{(n)} \,|\, \epsilon)\, p(\mathbf{x}^{(n)})\right] p(\epsilon)\, p(\Sigma_N)\, p(\Sigma)\, p(\mathbf{w} \,|\, \alpha)\, p(\alpha)$$

Inference

Observe data: $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}$

Unknowns: $\theta = \{\mathbf{w}, \alpha, \epsilon, \Sigma, \{z^{(n)}\}, \ldots\}$

$$p(\theta \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \theta)\, p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D}, \theta)$$

Target distribution

$$\pi(\theta) = \frac{\pi^*(\theta)}{\mathcal{Z}},$$

For inference: $\pi^*(\theta) = p(\mathcal{D}, \theta)$

Marginalization

Interested in particular parameter θ_i

$$p(\theta_i \,|\, \mathcal{D}) = \int p(\theta \,|\, \mathcal{D}) \; \mathrm{d}\theta_{\setminus i}$$

Sampling solution:

— Sample everything: $\theta^{(s)} \sim p(\theta \,|\, \mathcal{D})$

— $\theta_i^{(s)}$ comes from marginal $p(\theta_i \,|\, \mathcal{D})$

(But see also ‘Rao–Blackwellization’)

Roadmap

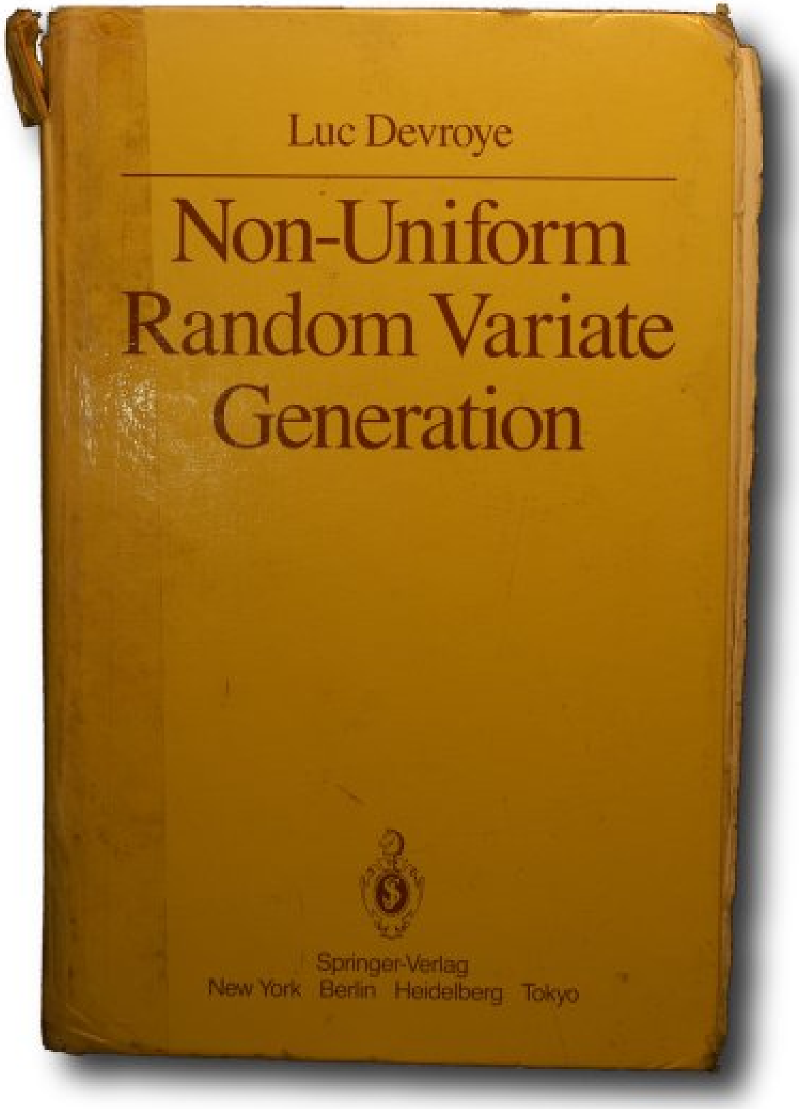
— Looking at samples

— Monte Carlo computations

— **How to actually get the samples**
Standard generators, Markov chains

— Practical issues

Sampling simple distributions


Luc Devroye
Non-Uniform Random Variate Generation
Springer-Verlag
New York Berlin Heidelberg Tokyo

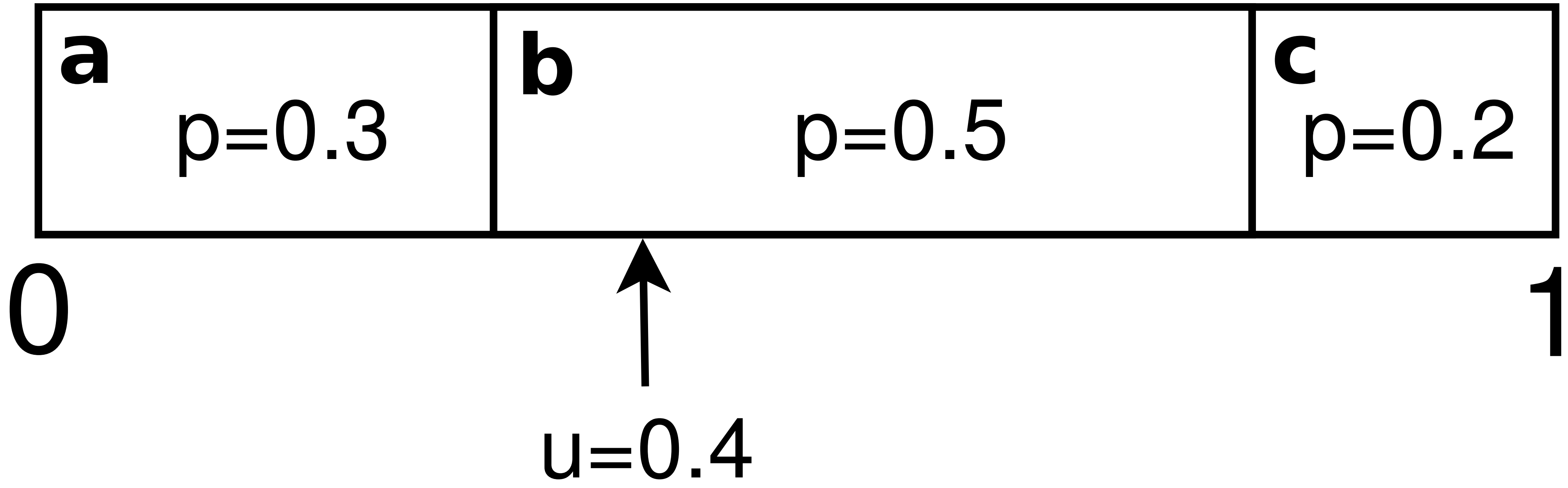
Use library routines for univariate distributions

(and some other special cases)

This book (free online) explains how some of them work

http://luc.devroye.org/rnbookindex.html

Sampling discrete values



$$u \sim \mathrm{Uniform}[0, 1]$$

$$u = 0.4 \quad \Rightarrow \quad \theta = \mathsf{b}$$

Large number of samples? 1) Alias method, Devroye book; 2) Ex 6.3 MacKay book

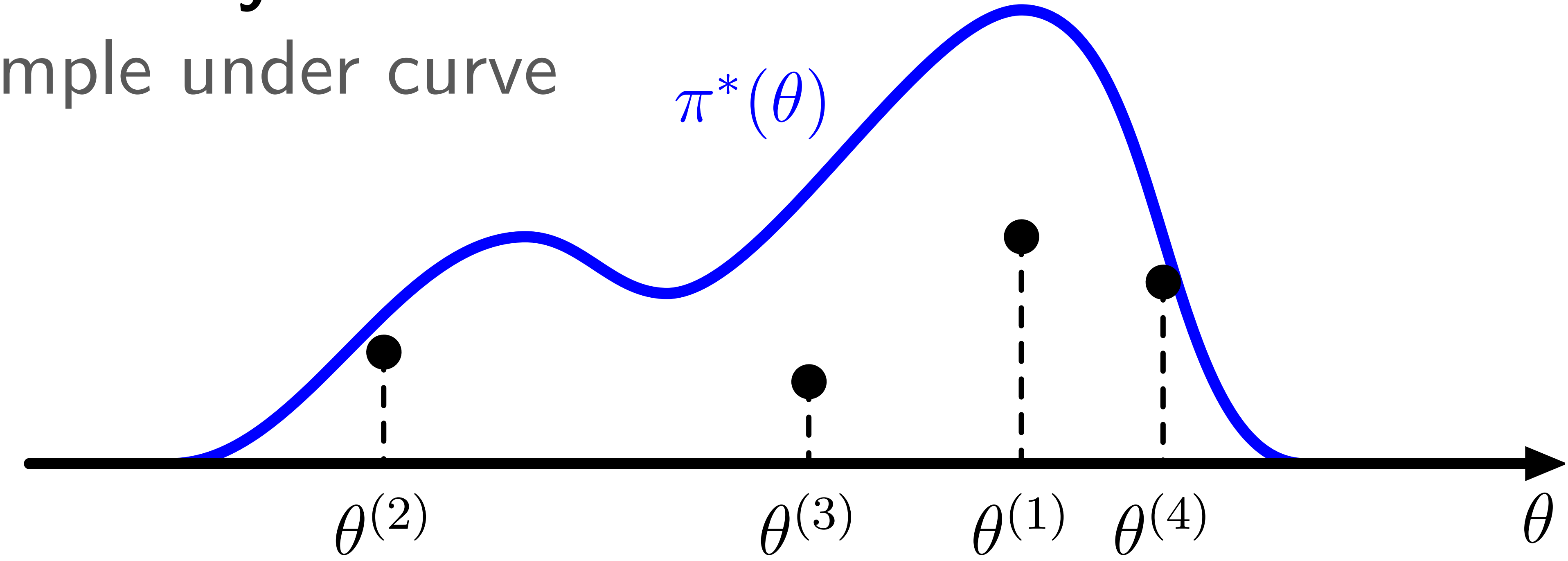
Sampling from a density

Math: $A^{(s)} \sim \mathrm{Uniform}[0,1], \quad \theta^{(s)} = \Phi^{-1}(A^{(s)})$

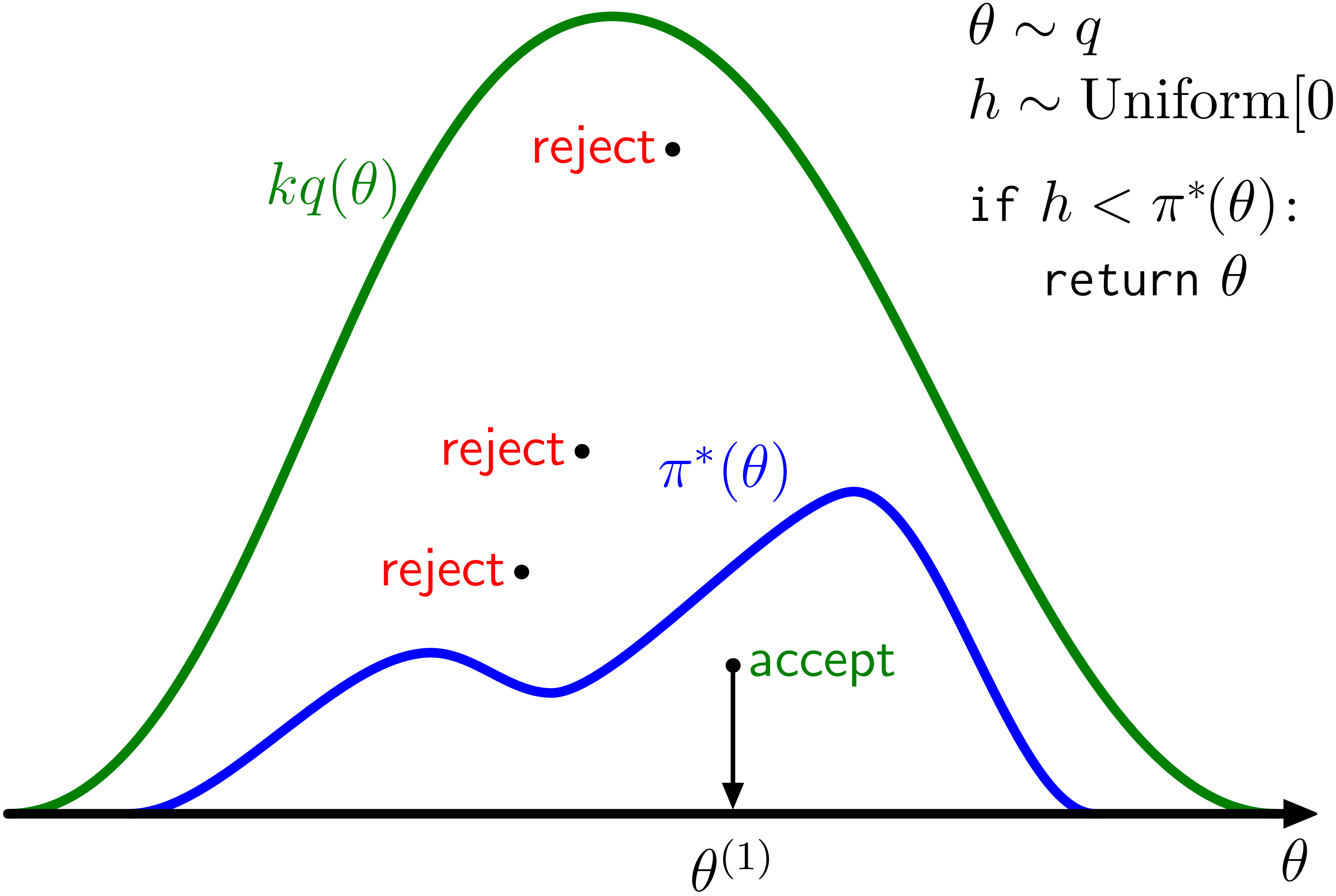
where cdf $\Phi(\theta) = \int_{-\infty}^{\theta} \pi(\theta')\, \mathrm{d}\theta'$

Geometry:

sample under curve



Rejection Sampling



```
while true:
    $\theta \sim q$
    $h \sim \text{Uniform}[0, kq(\theta)]$
    if $h < \pi^*(\theta)$:
        return $\theta$
```

Importance Sampling

Rewrite integral: expectation under simple distribution q:

$$\int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta \;=\; \int f(\theta)\,\frac{\pi(\theta)}{q(\theta)}\,q(\theta)\,\mathrm{d}\theta,$$

$$\approx\; \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)})\,\frac{\pi(\theta^{(s)})}{q(\theta^{(s)})}, \qquad \theta^{(s)} \sim q$$

Unbiased if $q(\theta)\!>\!0$ where $\pi(\theta)\!>\!0$. Can have infinite variance.

Importance Sampling 2

Can't evaluate $\pi(\theta) = \dfrac{\pi^*(\theta)}{\mathcal{Z}}, \quad \mathcal{Z} = \int \pi^*(\theta)\,\mathrm{d}\theta$

Alternative version:

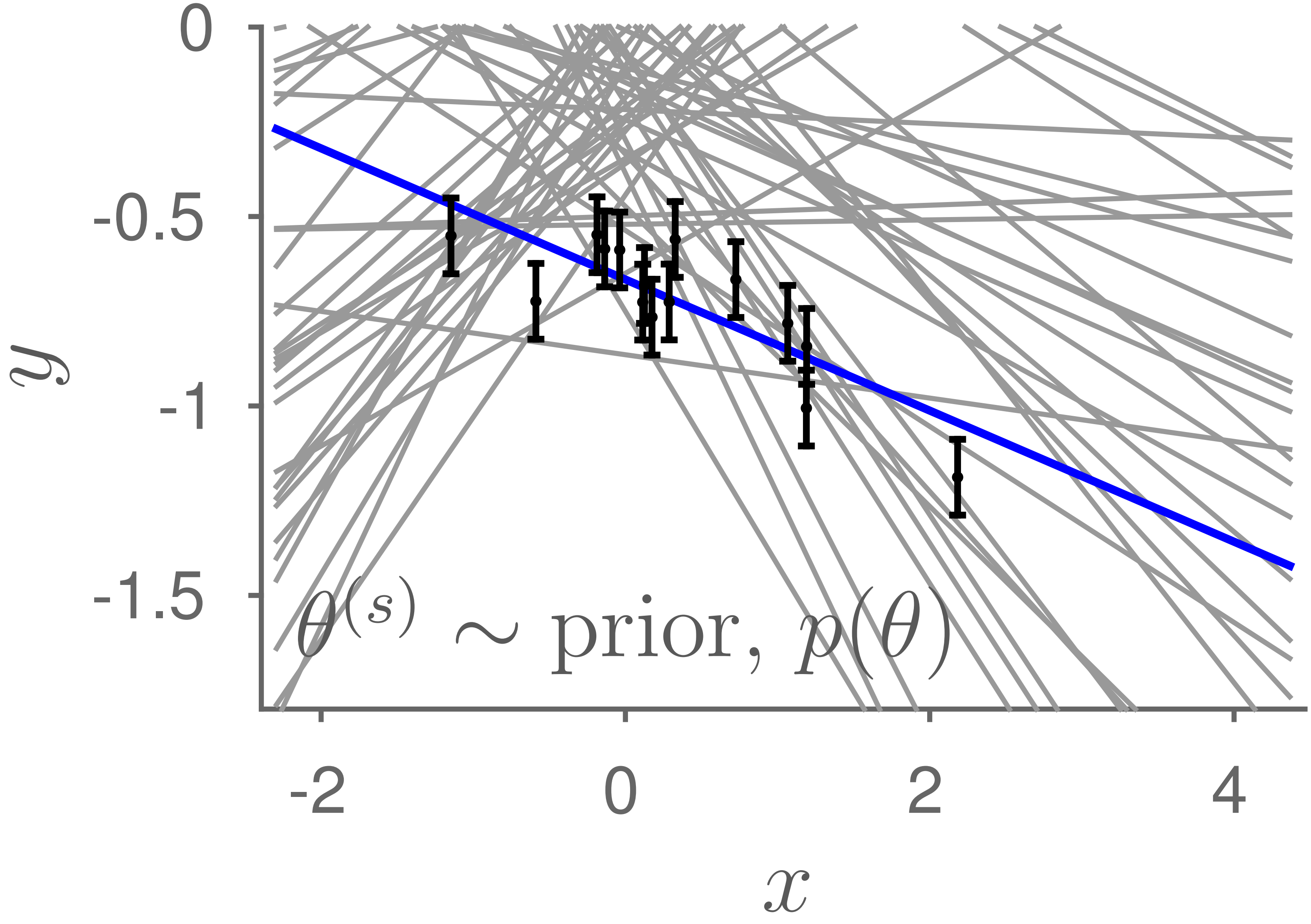
$$\theta^{(s)} \sim q, \qquad r^{*(s)} = \frac{\pi^*(\theta^{(s)})}{q(\theta^{(s)})}, \qquad r^{(s)} = \frac{r^{*(s)}}{\sum_{s'} r^{*(s')}}$$

Biased but consistent estimator:

$$\int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta \;\approx\; \sum_{s=1}^{S} f(\theta^{(s)})\,r^{(s)}$$

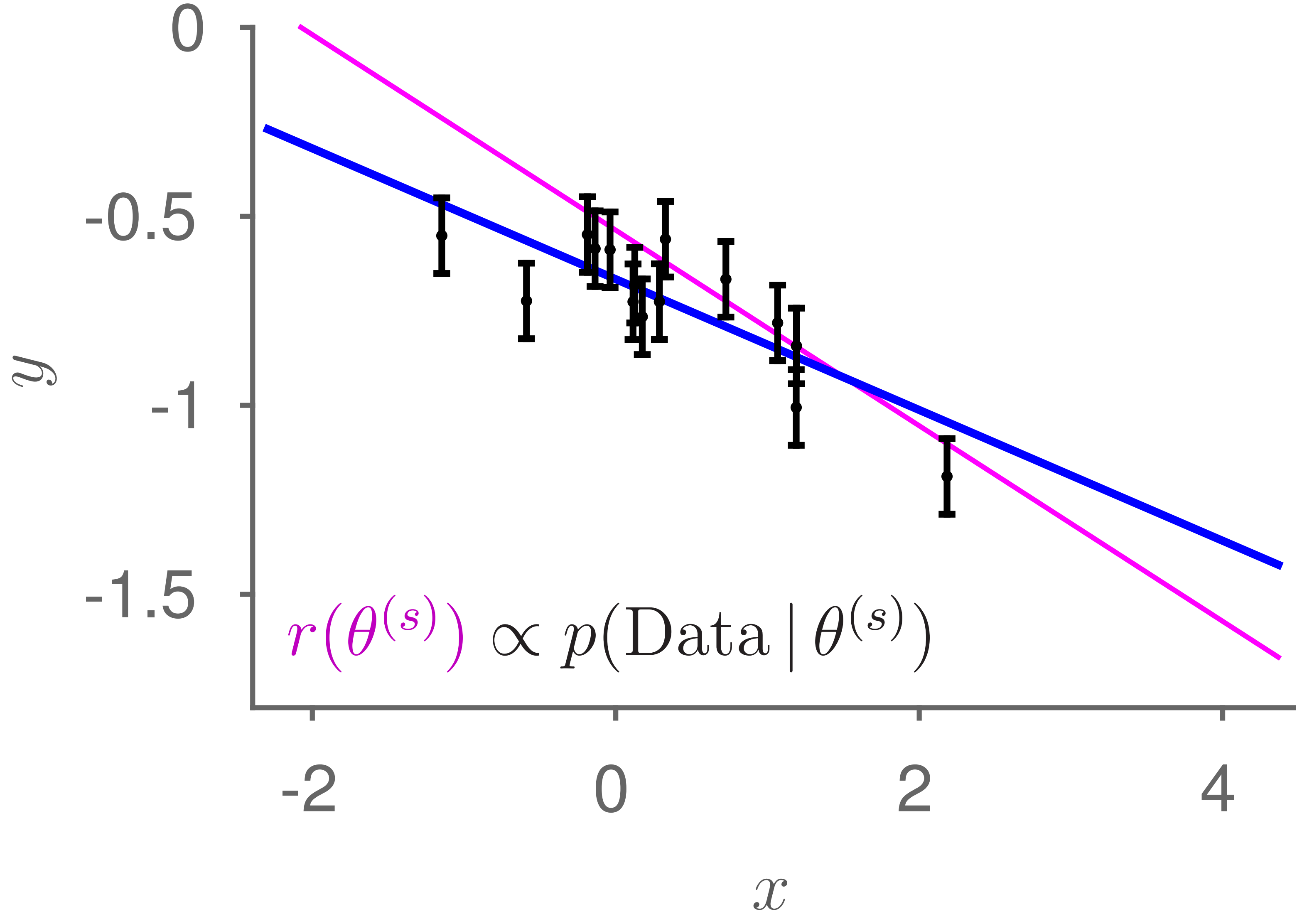
Linear regression

60 samples from prior:



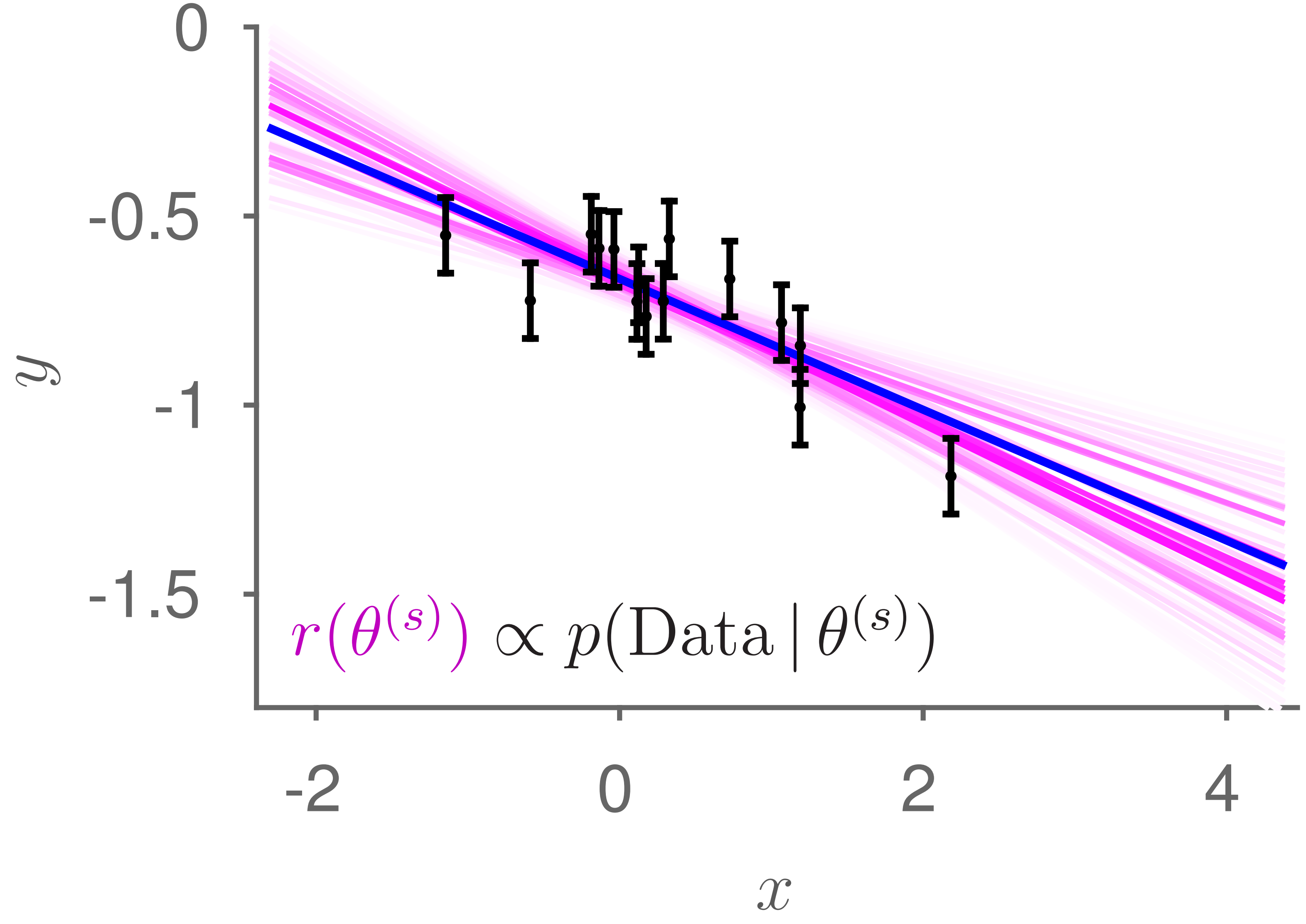
Linear regression

60 samples from prior, importance reweighted:



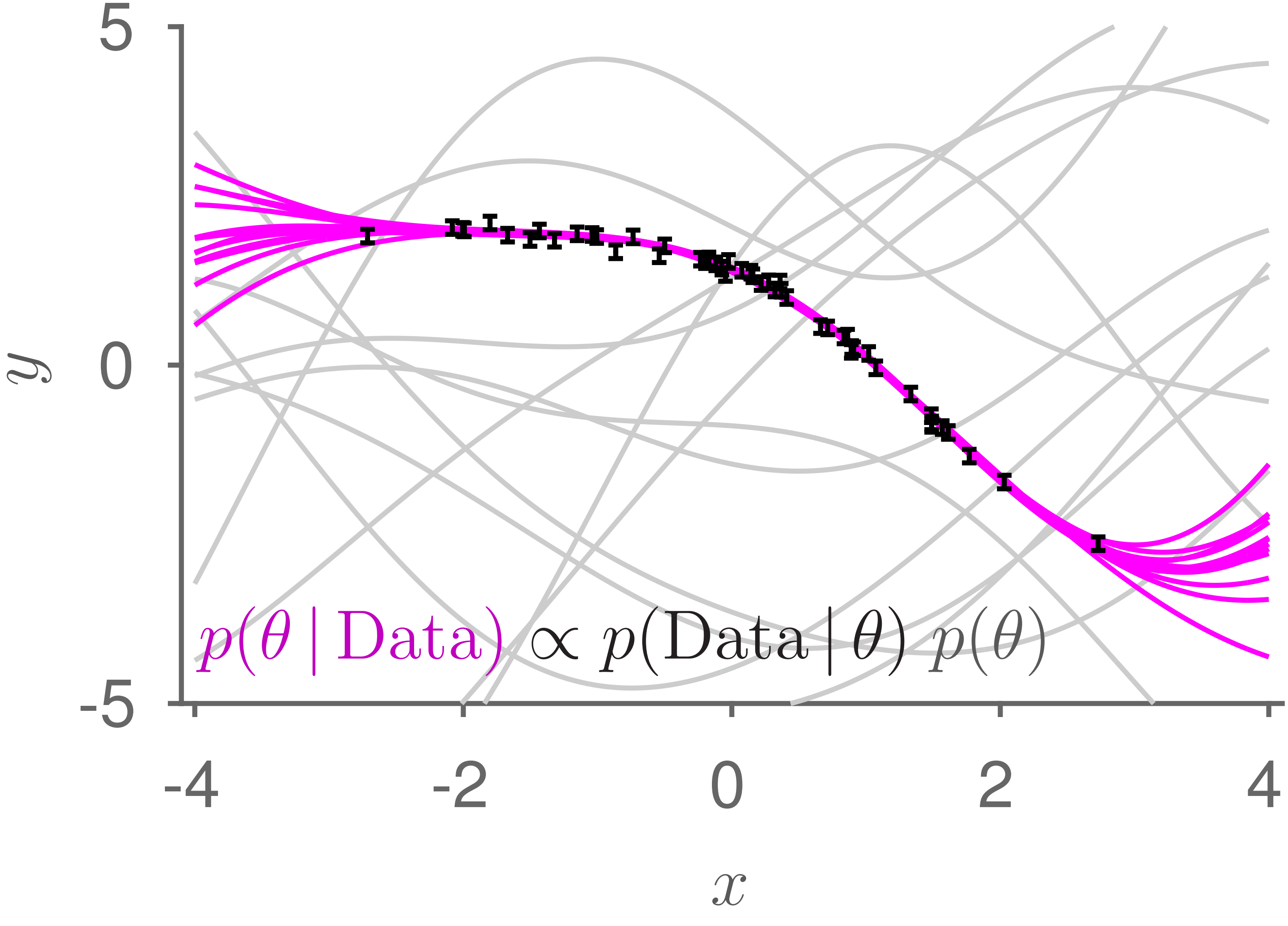
Linear regression

10,000 samples from prior, importance reweighted:



High dimensional θ

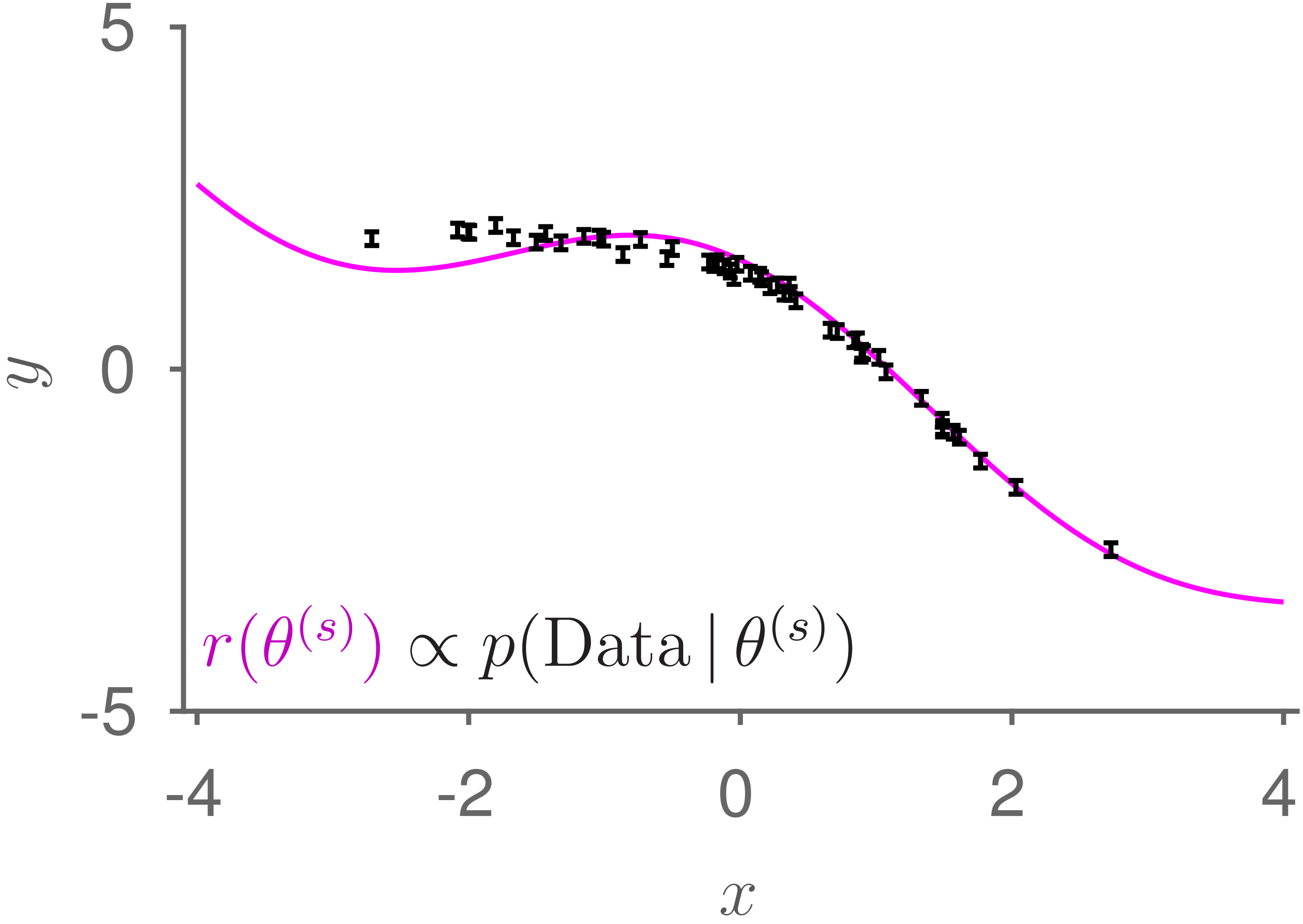
12 curves from prior and 12 curves from posterior



$p(\theta \mid \mathrm{Data}) \propto p(\mathrm{Data} \mid \theta)\, p(\theta)$

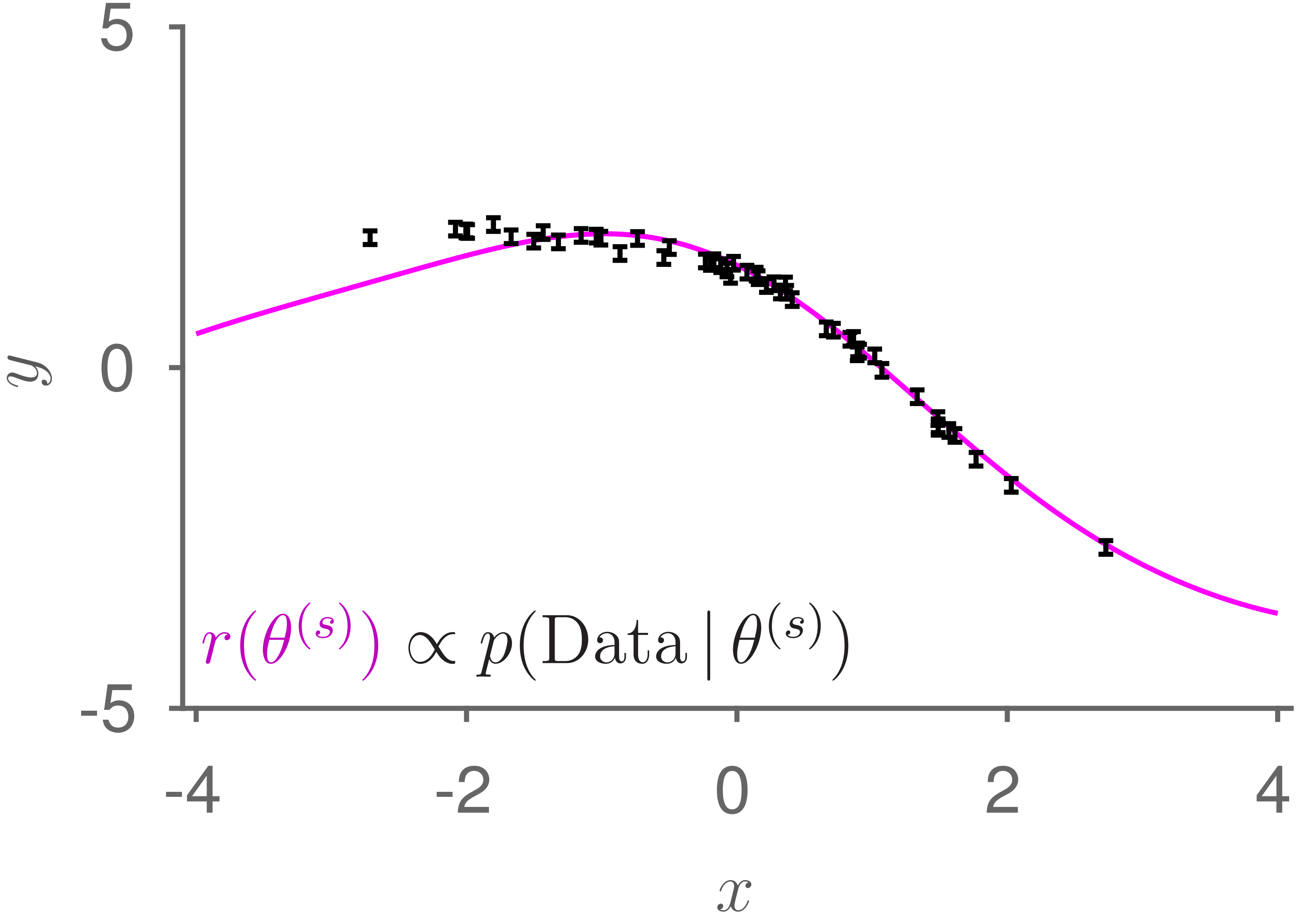
High dimensional θ

10,000 samples from prior, importance reweighted:



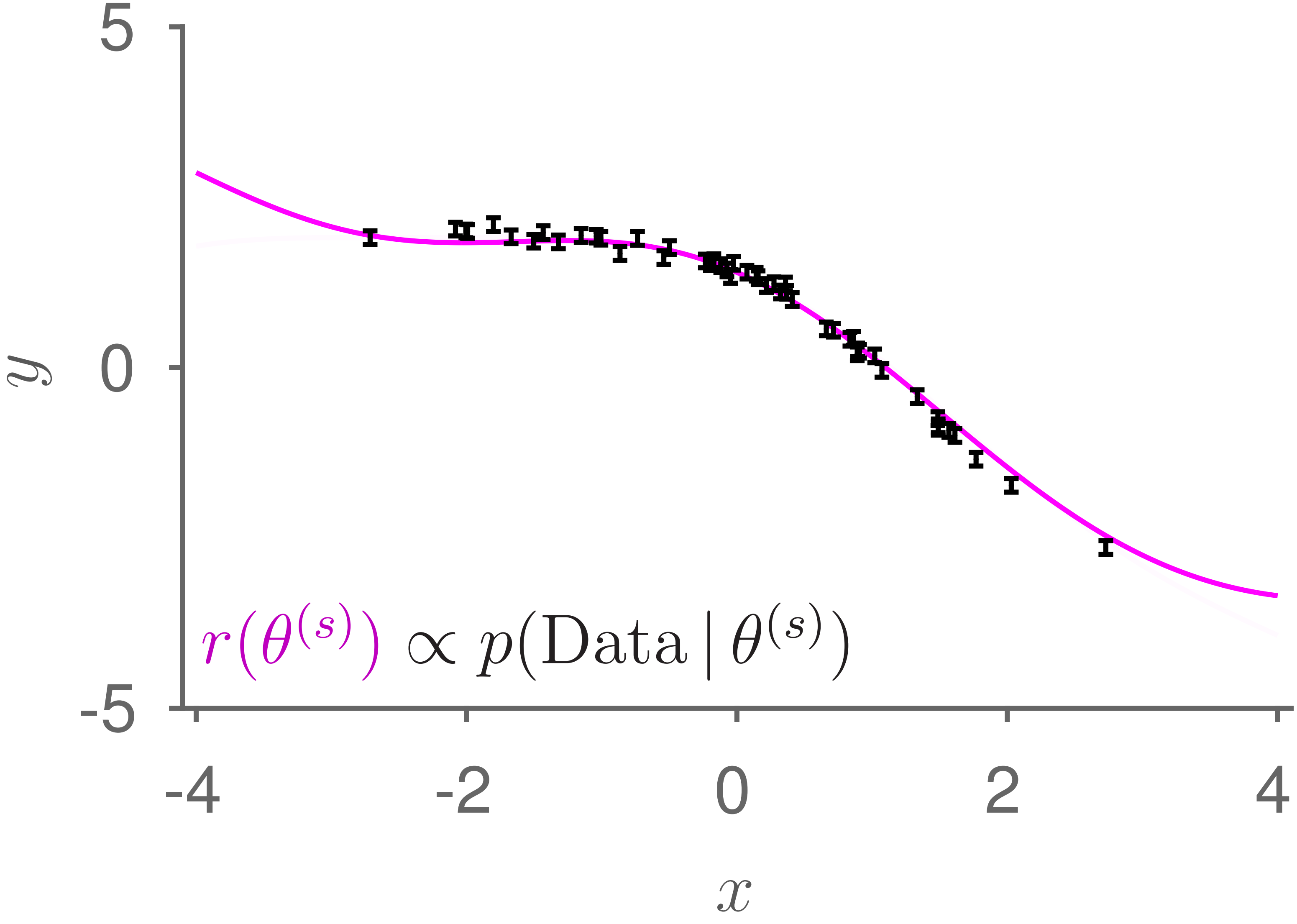
High dimensional θ

100,000 samples from prior, importance reweighted:



High dimensional θ

1,000,000 samples from prior, importance reweighted:



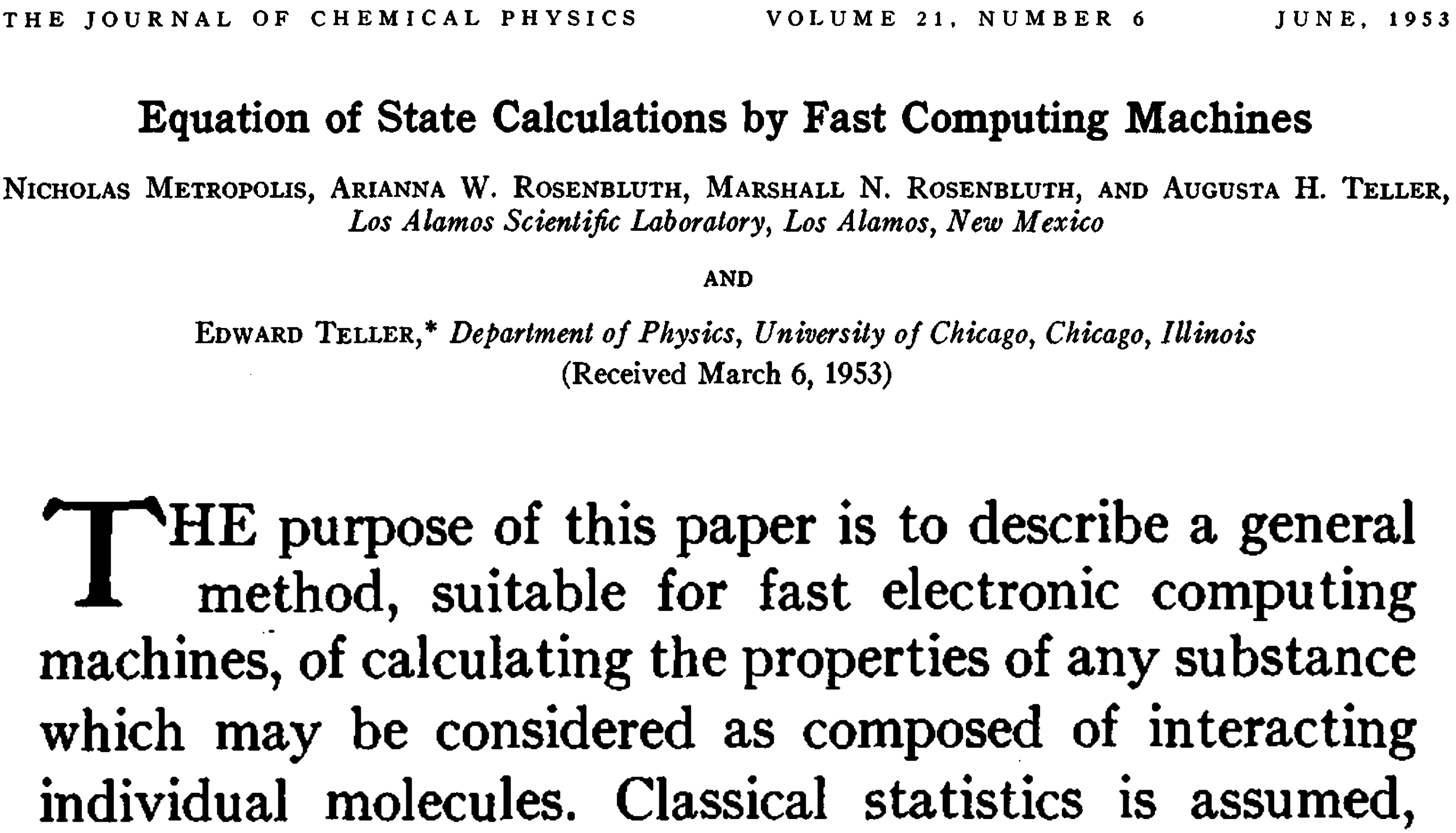
Roadmap

— Looking at samples

— Monte Carlo computations

— **How to actually get the samples**
Standard generators, Markov chains

— Practical issues



THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

THE purpose of this paper is to describe a general method, suitable for fast electronic computing machines, of calculating the properties of any substance which may be considered as composed of interacting individual molecules. Classical statistics is assumed,

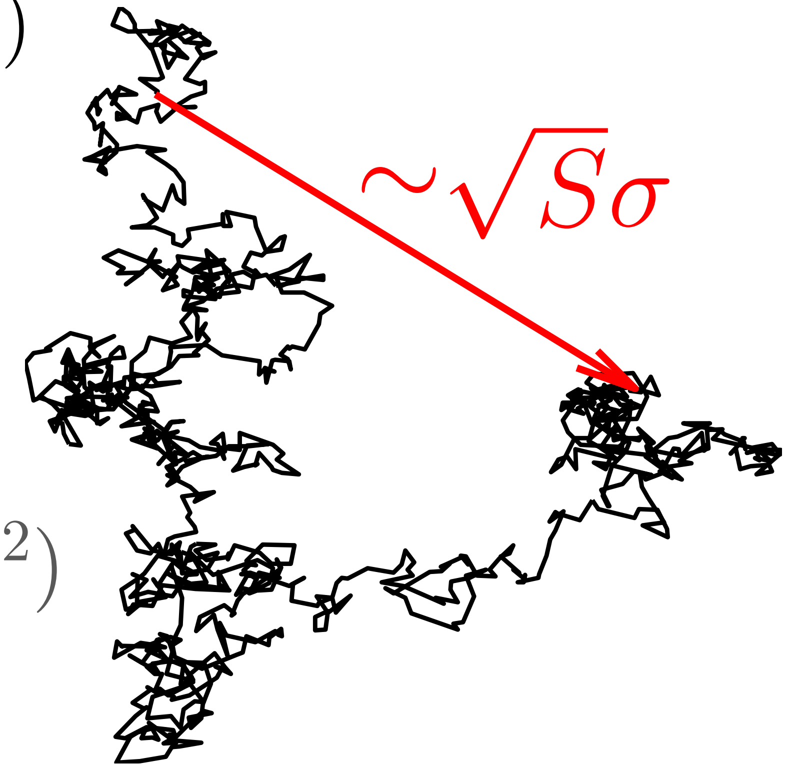
>30,000 citations

Marshall Rosenbluth's account:

http://dx.doi.org/10.1063/1.1887186

Markov chains

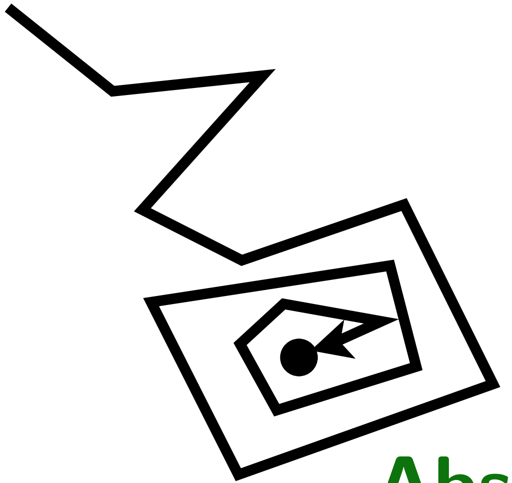
$$p(\theta^{(s+1)} \,|\, \theta^{(1)} \ldots \theta^{(s)}) = T(\theta^{(s+1)} \!\leftarrow\! \theta^{(s)})$$



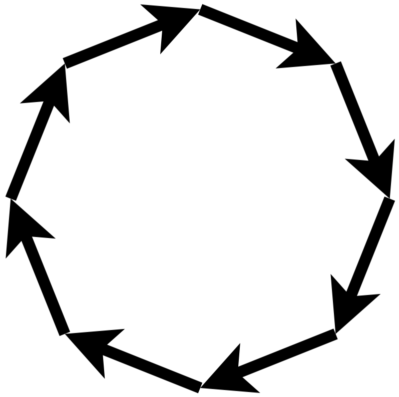
Divergent

e.g., random walk diffusion

$$T(\theta^{(s+1)} \!\leftarrow\! \theta^{(s)}) = \mathcal{N}(\theta^{(s+1)};\, \theta^{(s)}, \sigma^2)$$

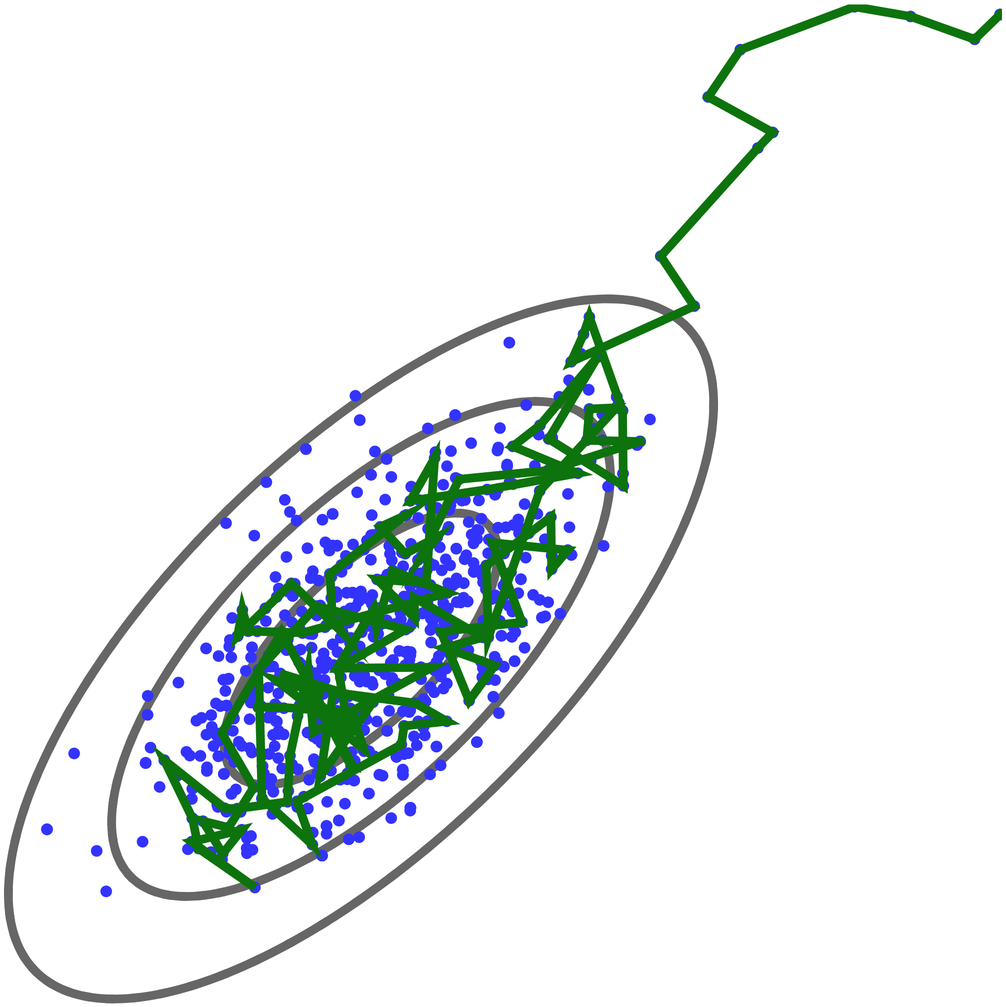


Absorbing states



Cycles

Markov chain equilibrium $\pi(\theta)$



$$\lim_{s\to\infty} p(\theta^{(s)}) = \pi(\theta^{(s)})$$

'Ergodic'

if true for all $\theta^{(s=0)}$

(other definitions of ergodic exist)

Possible to get anywhere in K steps,

($T^K(\theta'\!\leftarrow\!\theta) > 0$ for all pairs) $\Rightarrow$ no cycles or islands

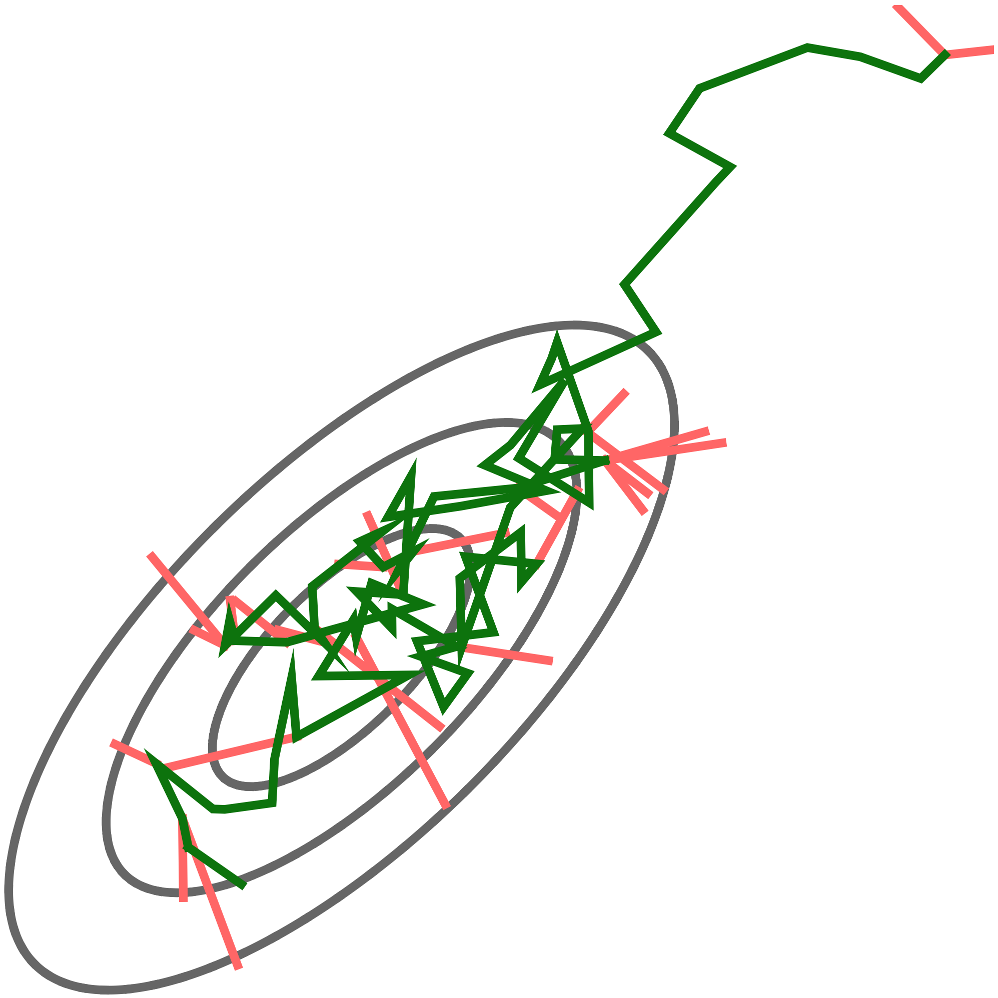
Invariant/stationary condition

If $\theta^{(s)}$ is a sample from π,

$\theta^{(s+1)}$ is also a sample from π.

$$p(\theta') = \int T(\theta' \!\leftarrow\! \theta)\, \pi(\theta)\; \mathrm{d}\theta = \pi(\theta')$$

Metropolis–Hastings

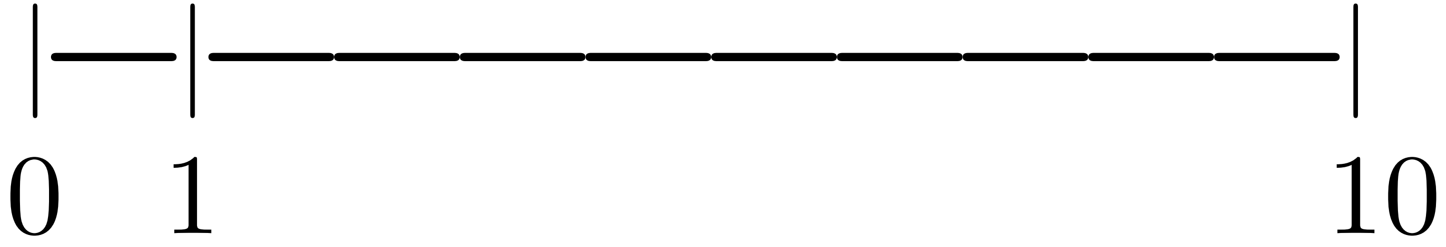


$$\theta' \sim q(\theta'; \theta^{(s)})$$

```
if accept:
    $\theta^{(s+1)} \leftarrow \theta'$
else:
    $\theta^{(s+1)} \leftarrow \theta^{(s)}$
```

$$P(\text{accept}) = \min\left(1,\ \frac{\pi^*(\theta')\, q(\theta^{(s)}; \theta')}{\pi^*(\theta^{(s)})\, q(\theta'; \theta^{(s)})}\right)$$

Example / warning



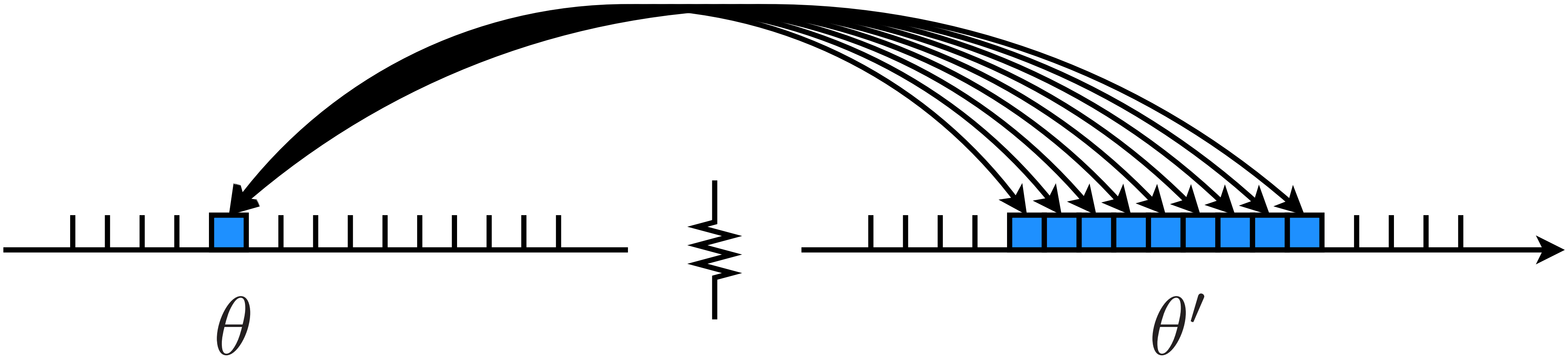
Proposal: $$\begin{cases} \theta^{(s+1)} = 9\theta^{(s)} + 1, & 0 < \theta^{(s)} < 1 \\ \theta^{(s+1)} = (\theta^{(s)} - 1)/9, & 1 < \theta^{(s)} < 10 \end{cases}$$

Accept move with probability:

$$\min\left(1,\ \frac{\pi^*(\theta')\, q(\theta;\theta')}{\pi^*(\theta)\, q(\theta';\theta)}\right) = \min\left(1,\ \frac{\pi^*(\theta')}{\pi^*(\theta)}\right)$$

(Wrong!)

Example / warning



Accept θ' with probability:

$$\min\left(1,\ \frac{q(\theta;\theta')\,\pi^*(\theta')}{q(\theta';\theta)\,\pi^*(\theta)}\right) \;=\; \min\left(1,\ \frac{1}{1/9}\,\frac{\pi^*(\theta')}{\pi^*(\theta)}\right)$$

Really, Green (Biometrika 82:4, 1995):

$$\min\left(1,\ \left|\frac{\partial\theta'}{\partial\theta}\right|\frac{\pi^*(\theta')}{\pi^*(\theta)}\right) \;=\; \min\left(1,\ 9\,\frac{\pi^*(\theta')}{\pi^*(\theta)}\right)$$

Python code for demo

```
import numpy as np

def metropolis(init, log_pstar_fn, iters, sigma):
    init = np.array(init)
    samples = np.zeros((iters, init.size))

    state = init
    Lp_state = log_pstar_fn(state)
    for ss in range(iters):
        # Propose
        prop = state + (sigma * np.random.randn(state.shape))
        Lp_prop = log_pstar_fn(prop)
        if np.random.rand() < np.exp(Lp_prop - Lp_state):
            # Accept
            state = prop
            Lp_state = Lp_prop
        samples[ss, :] = state.ravel()

    return samples
```

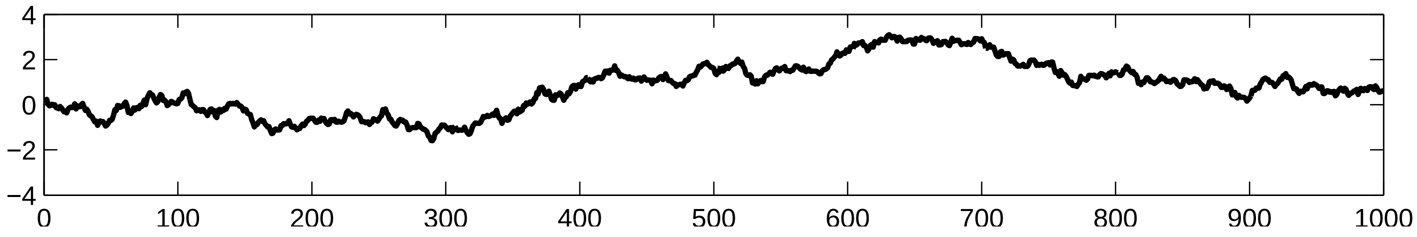
Step-size demo

Explore $\mathcal{N}(0, 1)$ with different step sizes σ

```
import matplotlib.pyplot as plt
def sigma(s): plt.plot(metropolis(0, lambda x: -0.5*x*x, 1000, s))
```

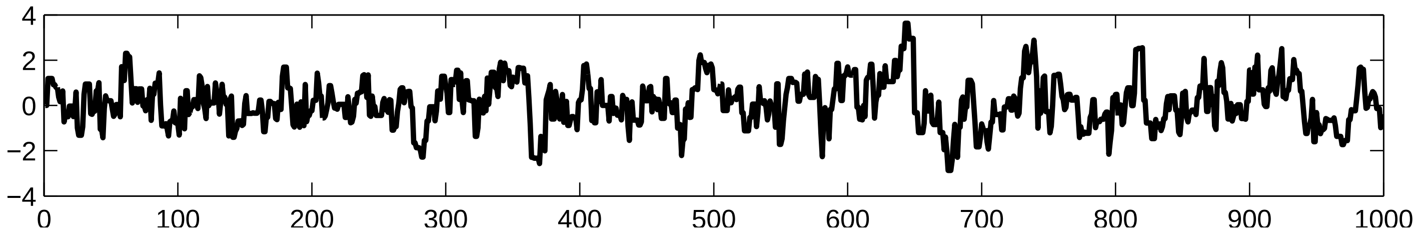
sigma(0.1)

99.8% accepts



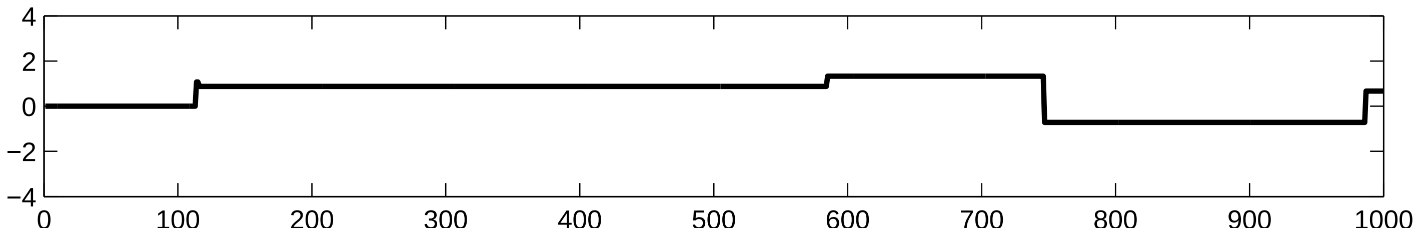
sigma(1)

68.4% accepts

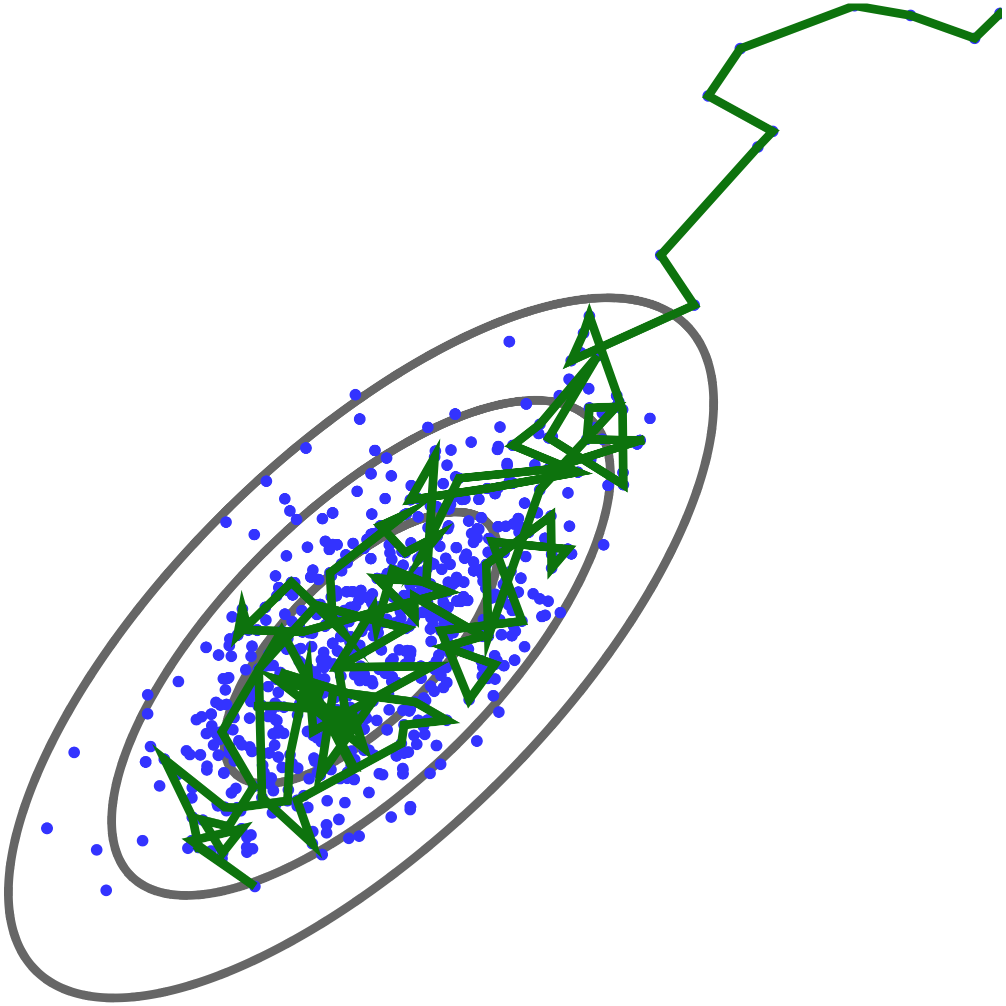


sigma(100)

0.5% accepts



Markov chain Monte Carlo (MCMC)



User chooses $\pi(\theta)$

Explore some plausible θ

For large s:

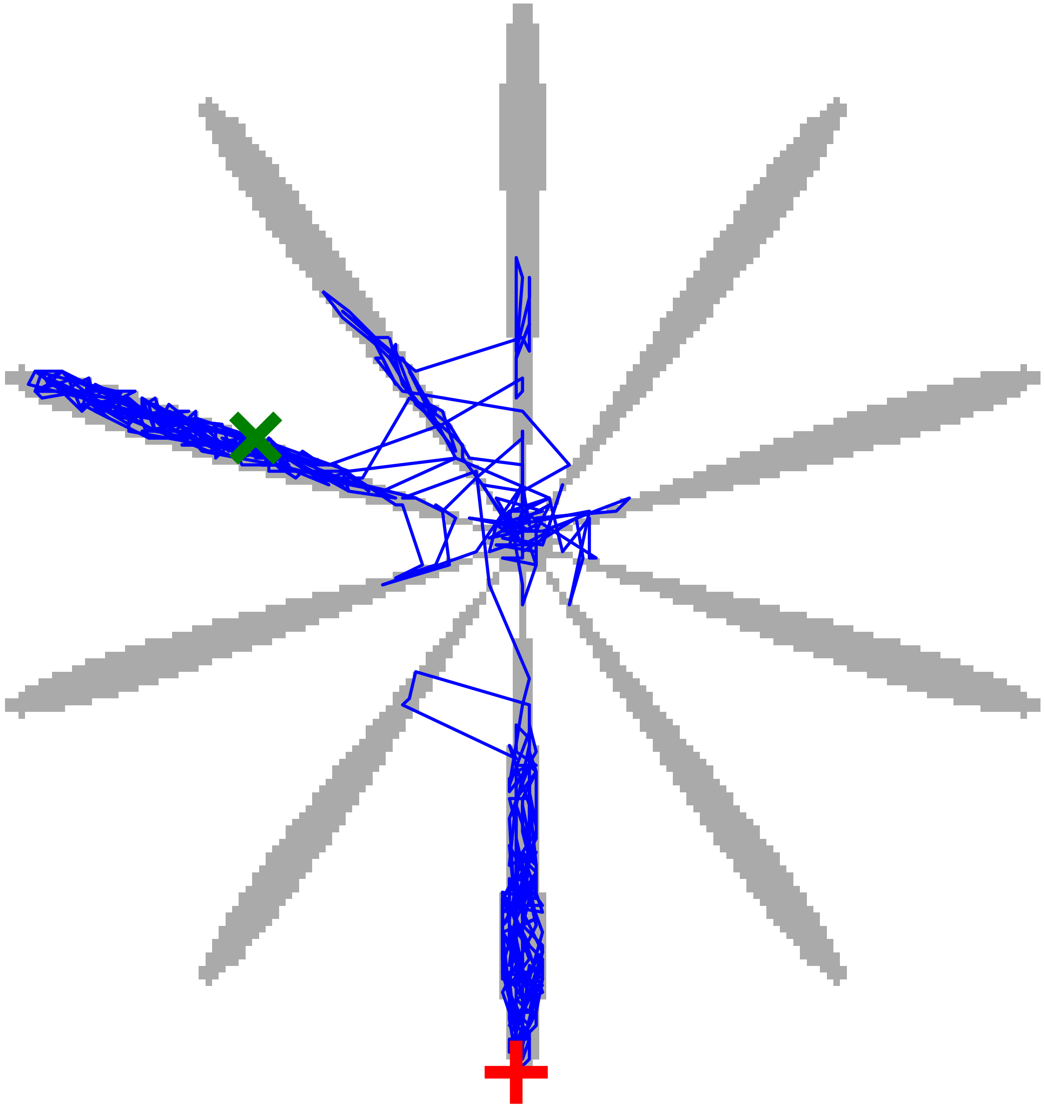
$$p(\theta^{(s)}) = \pi(\theta^{(s)})$$

$$p(\theta^{(s+1)}) = \pi(\theta^{(s+1)})$$

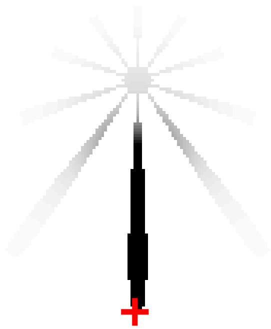
$$\mathbb{E}\left[f(\theta^{(s)})\right] = \mathbb{E}\left[f(\theta^{(s+1)})\right] = \int f(\theta)\pi(\theta)\,\mathrm{d}\theta = \lim_{S\to\infty}\frac{1}{S}\sum_s f(\theta^{(s)})$$

Markov chain mixing

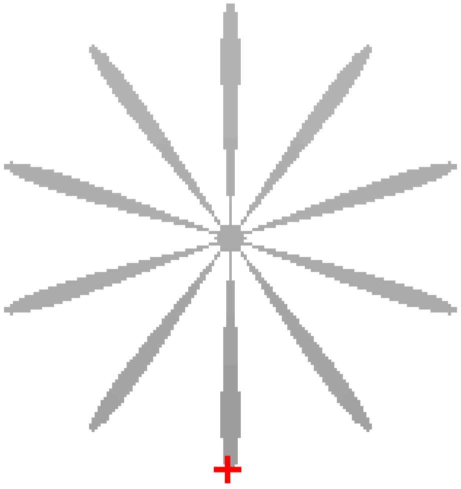
Initialization + → 2000 steps → x ‘sample’



$p(\theta^{(s=100)})$



$p(\theta^{(s=2000)}) \approx \pi(\theta)$



Creating an MCMC scheme

M–H gives $T(\theta' \!\leftarrow\! \theta)$ with invariant π

Lots of options for $q(\theta'; \theta)$:

— Local diffusions

— Appproximations of π

— Update one variable or all?

— ...

Multiple valid operators T_A, T_B, T_C, ...

Composing operators

If $p(\theta^{(1)}) = \pi(\theta^{(1)})$

$$\theta^{(2)} \sim T_A(\cdot \leftarrow \theta^{(1)}) \quad \Rightarrow \quad p(\theta^{(2)}) = \pi(\theta^{(2)})$$

$$\theta^{(3)} \sim T_B(\cdot \leftarrow \theta^{(2)}) \quad \Rightarrow \quad p(\theta^{(3)}) = \pi(\theta^{(3)})$$

$$\theta^{(4)} \sim T_C(\cdot \leftarrow \theta^{(3)}) \quad \Rightarrow \quad p(\theta^{(4)}) = \pi(\theta^{(4)})$$

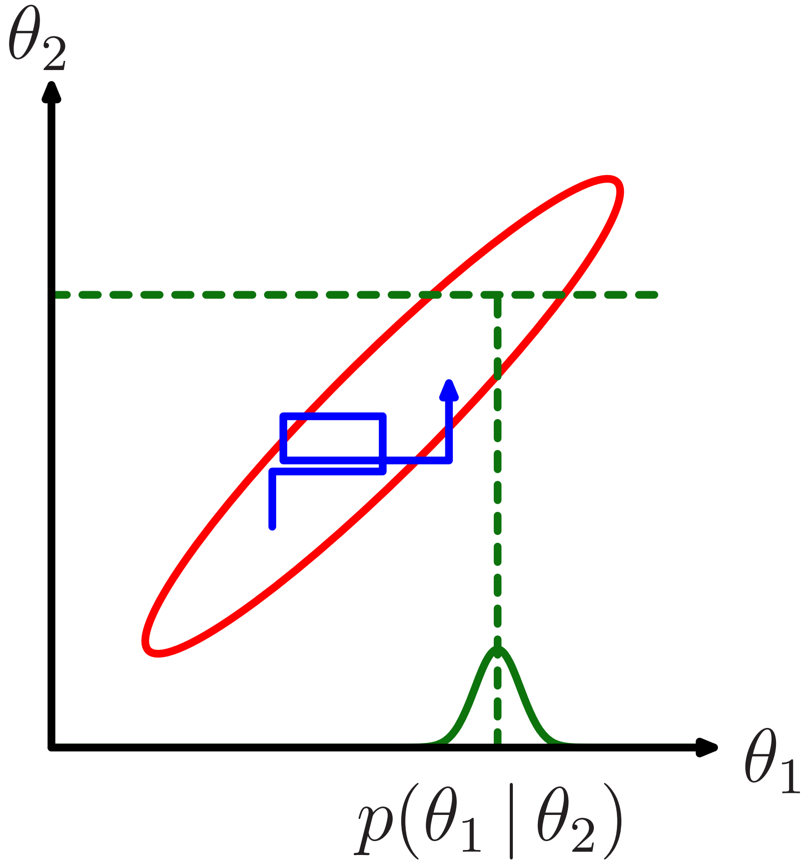
Composition $T_C T_B T_A$ leaves π invariant

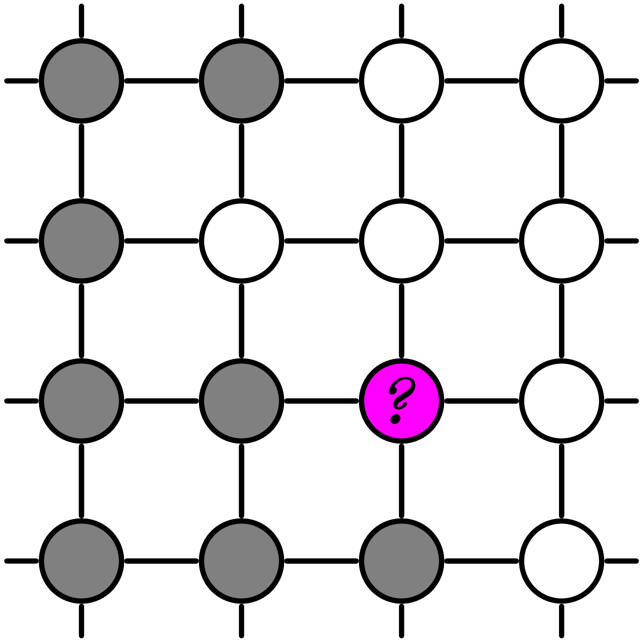
Valid MCMC method if ergodic overall

Gibbs sampling

Pick variables in turn or randomly,

and resample $p(\theta_i \,|\, \theta_{j\neq i})$





$$T_i(\theta' \!\leftarrow\! \theta) = p(\theta'_i \,|\, \theta_{j\neq i})\, \delta(\theta'_{j\neq i} - \theta_{j\neq i})$$

LHS adapted from Fig 11.11 Bishop PRML

Gibbs sampling correctness

$$p(\theta) = p(\theta_i \,|\, \theta_{\setminus i})\, p(\theta_{\setminus i})$$

Simulate by drawing $\theta_{\setminus i}$, then $\theta_i \,|\, \theta_{\setminus i}$

Draw $\theta_{\setminus i}$: sample θ, throw initial θ_i away

Other methods

Hamiltonian/Hybrid Monte Carlo (HMC) and NUTS

Available in PyMC and Stan

Slice Sampling (Neal 2003, MacKay textbook)

(Elliptical Slice Sampling for GPs)

Ensemble methods (e.g., emcee)

Run optimizer first. Easy to use software wins!

Roadmap

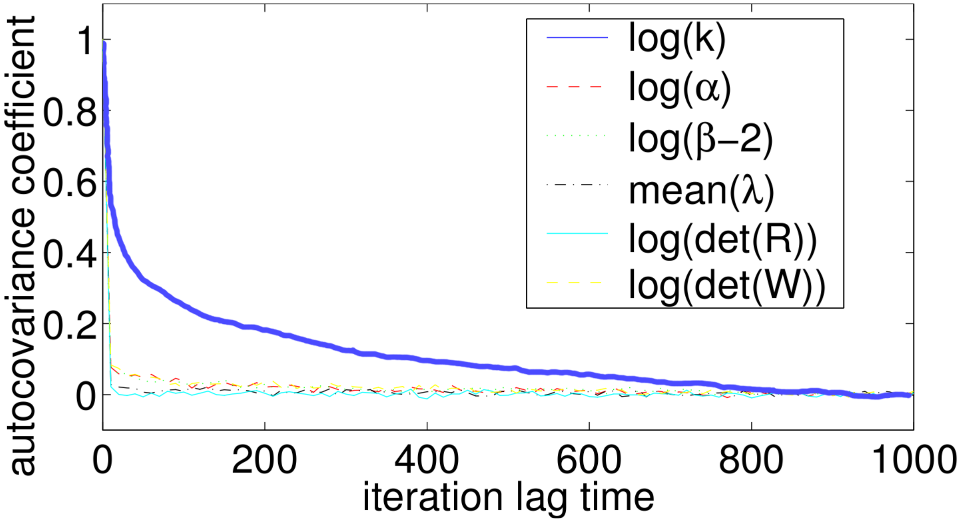
— Looking at samples

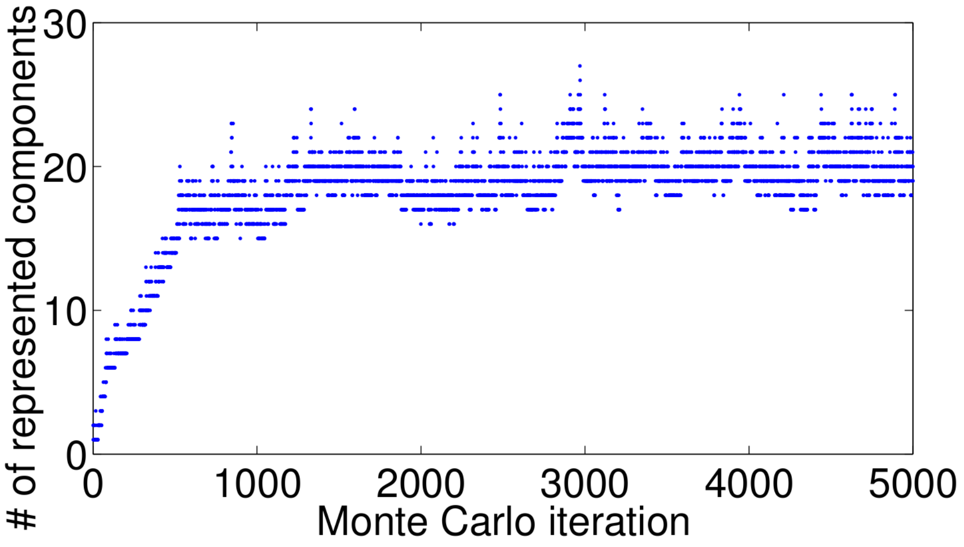
— Monte Carlo computations

— How to actually get the samples
Standard generators, Markov chains

— **Practical issues**

Empirical diagnostics





Rasmussen (2000)

Recommendations

For diagnostics:

Standard software packages like `R-CODA`

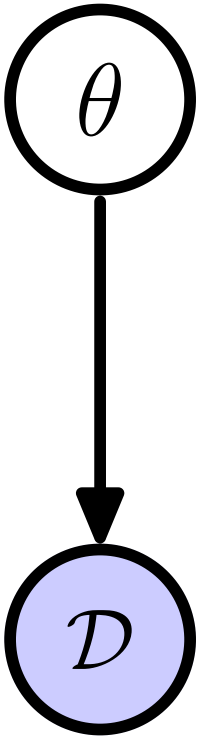
For opinion on thinning, multiple runs, burn in, etc.

Practical Markov chain Monte Carlo

Charles J. Geyer, *Statistical Science.* 7(4):473–483, 1992.

`http://www.jstor.org/stable/2246094`

Getting it right



We write MCMC code to update $\theta \,|\, \mathcal{D}$

Idea: also write code to sample $\mathcal{D} \,|\, \theta$

Both codes leave $p(\theta, \mathcal{D})$ invariant

Run codes alternately. Check θ's match prior

Geweke, *JASA*, 99(467):799–804, 2004.

Other consistency checks

Do I get the right answer on tiny versions of my problem?

Can I make good inferences about synthetic data drawn from my model?

Posterior Model checking: Gelman et al. Bayesian Data Analysis textbook and papers.

Summary

Write down the probability of everything $p(\mathcal{D}, \theta)$

Condition on data, $\mathcal{D}$,
explore unknowns θ by MCMC

Samples give plausible explanations
— Look at them
— Average their predictions

Which method?

Simulate / sample with known distribution:

Exact samples, rejection sampling

Posterior distribution, small, noisy problem

Importance sampling

Posterior distribution, interesting problem

Could start with MCMC

Slice sampling, M–H if careful, Gibbs if clever

Hamiltonian methods, HMC, uses gradients

Meta methods: AIS, SMC, Nested Sampling, ...

Acknowledgements / References

My first reading: MacKay's book and Neal's review:

www.inference.phy.cam.ac.uk/mackay/itila/

www.cs.toronto.edu/∼radford/review.abstract.html

Handbook of Markov Chain Monte Carlo

(Brooks, Gelman, Jones, Meng eds, 2011)

http://www.mcmchandbook.net/HandbookSampleChapters.html

Practical examples

My exercise sheet:

http://iainmurray.net/teaching/09mlss/

BUGS examples and more in STAN:

https://github.com/stan-dev/example-models/

Kaggle entry:

http://iainmurray.net/pub/12kaggle_dark/

Appendix slides

Blocking / Collapsing

Infer $\theta = (\mathbf{w}, \mathbf{z})$ given $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}$. Model:

$$\mathbf{w} \sim \mathcal{N}(0, I)$$

$$z_n \sim \text{Bernoulli}(0.1)$$

$$y^{(n)} \sim \begin{cases} \mathcal{N}(\mathbf{w}^\top \mathbf{x}^{(n)},\, 0.1^2) & z_n = 0 \\ \mathcal{N}(0,\, 1) & z_n = 1 \end{cases}$$

Block Gibbs: resample $p(\mathbf{w} \,|\, \mathbf{z}, \mathcal{D})$ and $p(\mathbf{z} \,|\, \mathbf{w}, \mathcal{D})$

Collapsing: run MCMC on $p(\mathbf{z} \,|\, \mathcal{D})$ or $p(\mathbf{w} \,|\, \mathcal{D})$

Auxiliary variables

Collapsing: analytically integrate variables out

Auxiliary methods: introduce extra variables;
integrate by MCMC

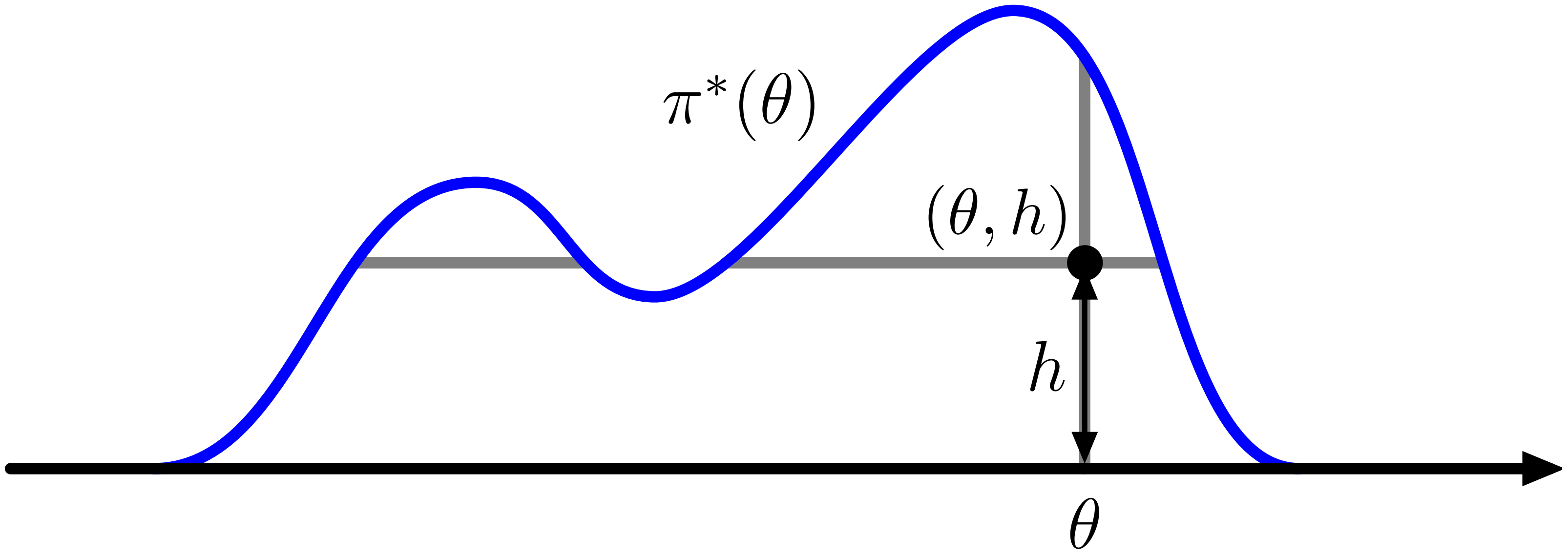
Explore: $\pi(\theta, h)$, where

$$\int \pi(\theta, h)\, \mathrm{d}h = \pi(\theta)$$

Swendsen–Wang, Hamiltonian Monte Carlo (HMC), Slice Sampling, Pseudo-Marginal methods. . .

Slice sampling idea

Sample uniformly under curve $\pi^*(\theta) \propto \pi(\theta)$

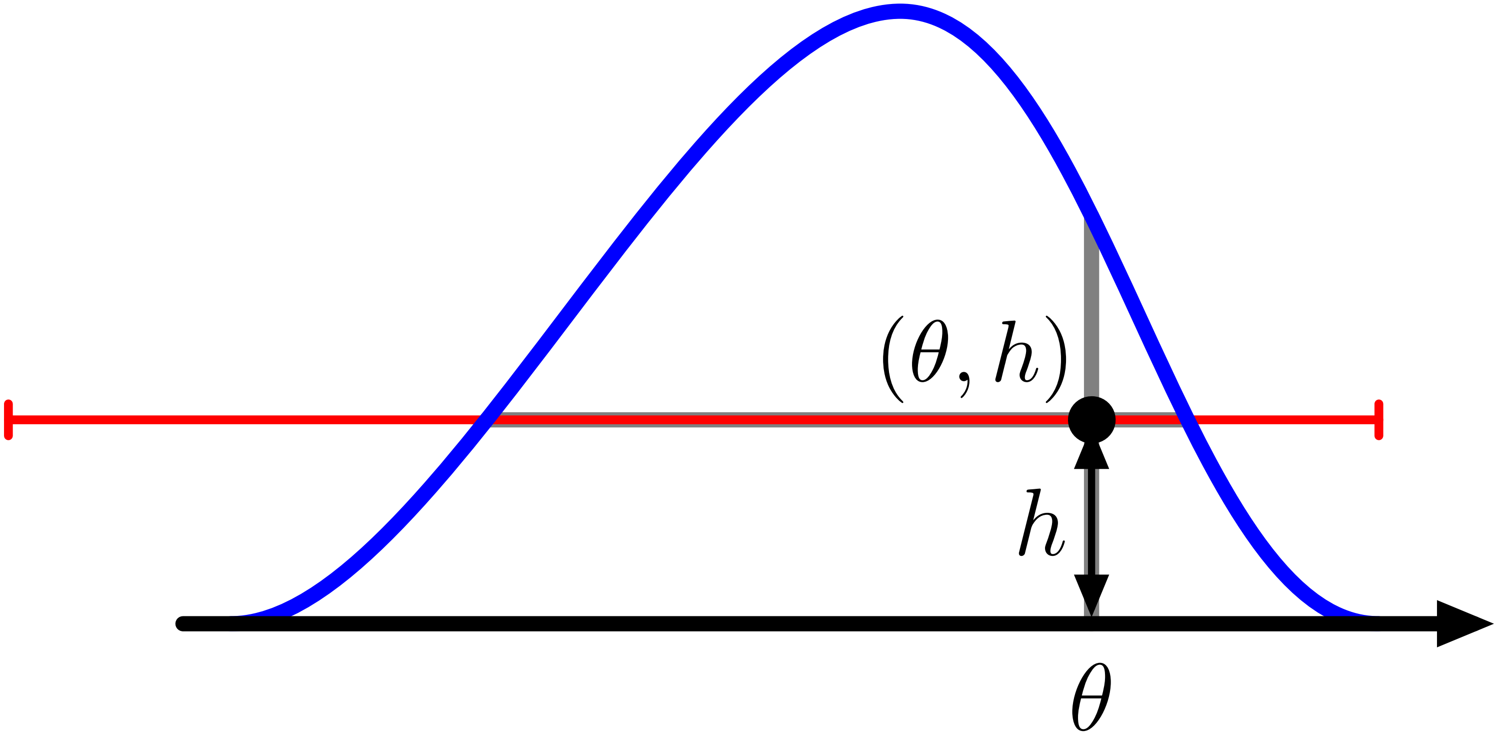


$$p(h \mid \theta) = \text{Uniform}[0, \pi^*(\theta)]$$

$$p(\theta \mid h) \propto \begin{cases} 1 & \pi^*(\theta) \geq h \\ 0 & \text{otherwise} \end{cases} = \text{“Uniform on the slice”}$$

Slice sampling

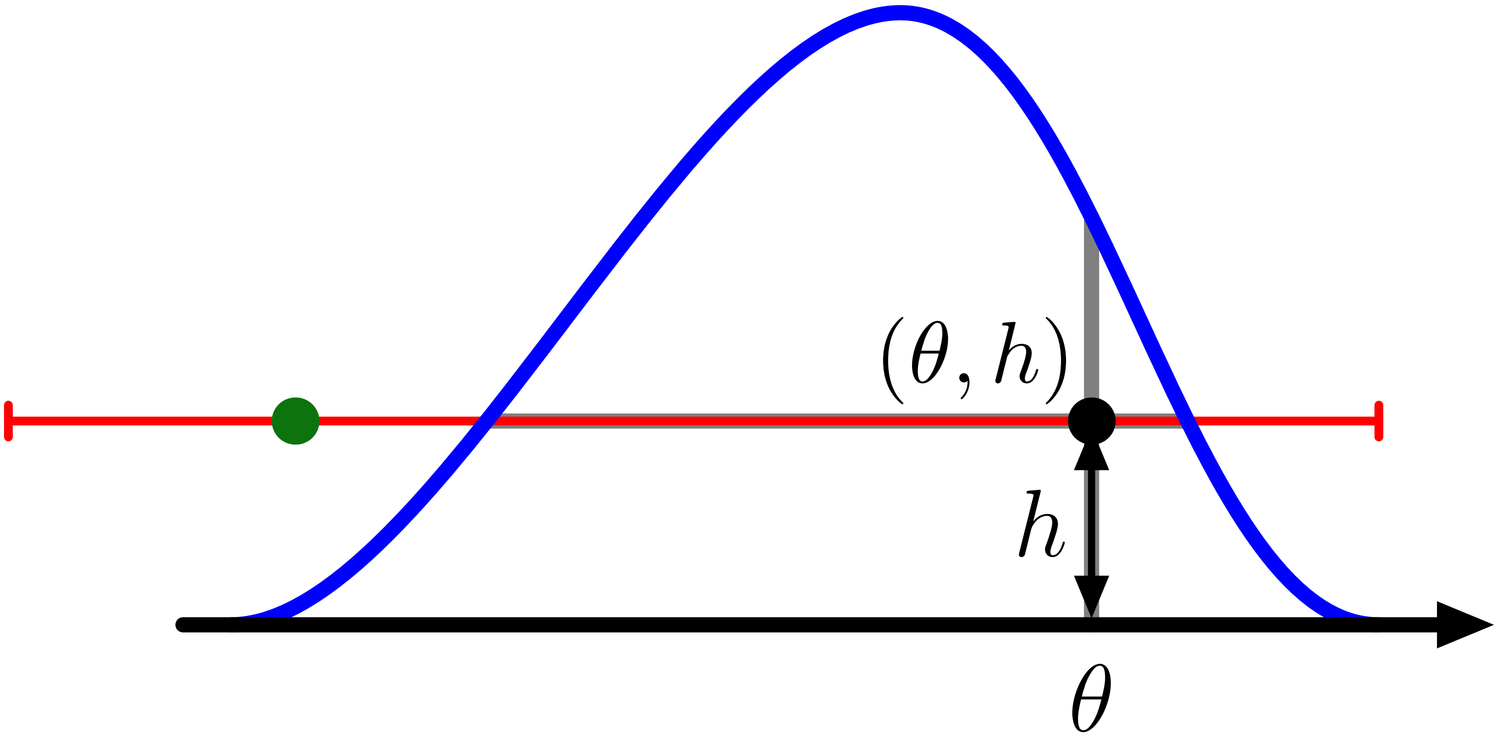
Unimodal conditionals



Rejection sampling $p(\theta \,|\, h)$ using broader uniform

Slice sampling

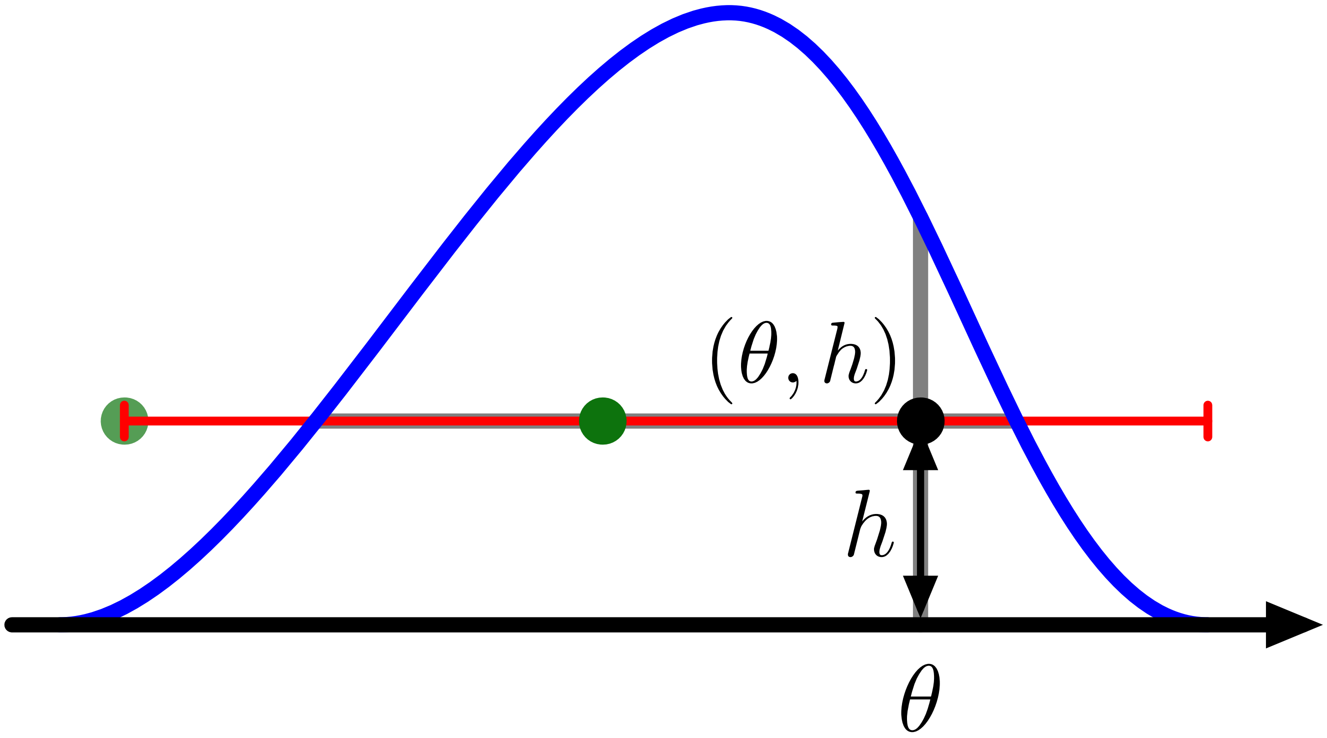
Unimodal conditionals



Adaptive rejection sampling $p(\theta \,|\, h)$

Slice sampling

Unimodal conditionals

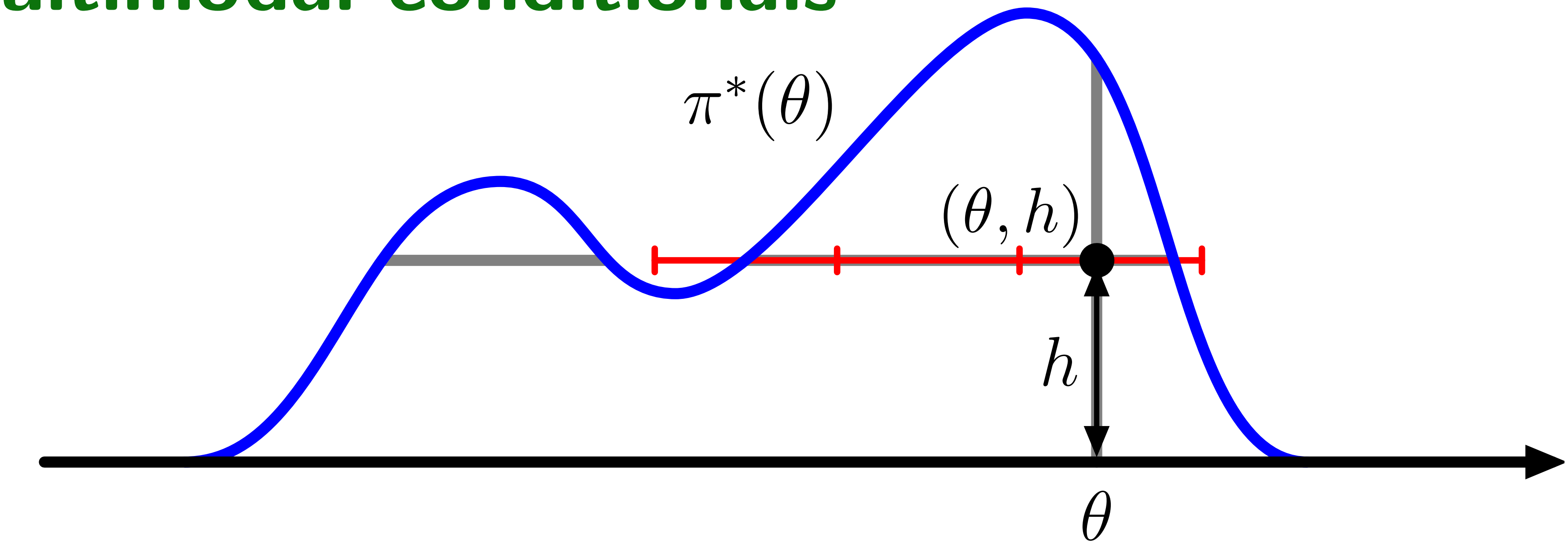


Quickly find new θ

No rejections recorded

Slice sampling

Multimodal conditionals



Use updates that leave $p(\theta \,|\, h)$ invariant:

— place bracket randomly around point
— linearly step out until ends are off slice
— sample on bracket, shrinking as before

Slice sampling

Advantages of slice-sampling:

- Easy — only requires $\pi^*(\theta) \propto \pi(\theta)$
- No rejections
- Step sizes adaptive

Other versions:

Neal (2003): `http://www.cs.toronto.edu/~radford/slice-aos.abstract.html`

Elliptical Slice Sampling: `http://iainmurray.net/pub/10ess/`

Pseudo-Marginal Slice Sampling: `http://arxiv.org/abs/1510.02958`

Reverse operator

$$R(\theta \!\leftarrow\! \theta') = \frac{T(\theta' \!\leftarrow\! \theta)\, \pi(\theta)}{\int T(\theta' \!\leftarrow\! \theta)\, \pi(\theta)\; \mathrm{d}\theta} = \frac{T(\theta' \!\leftarrow\! \theta)\, \pi(\theta)}{\pi(\theta')}$$

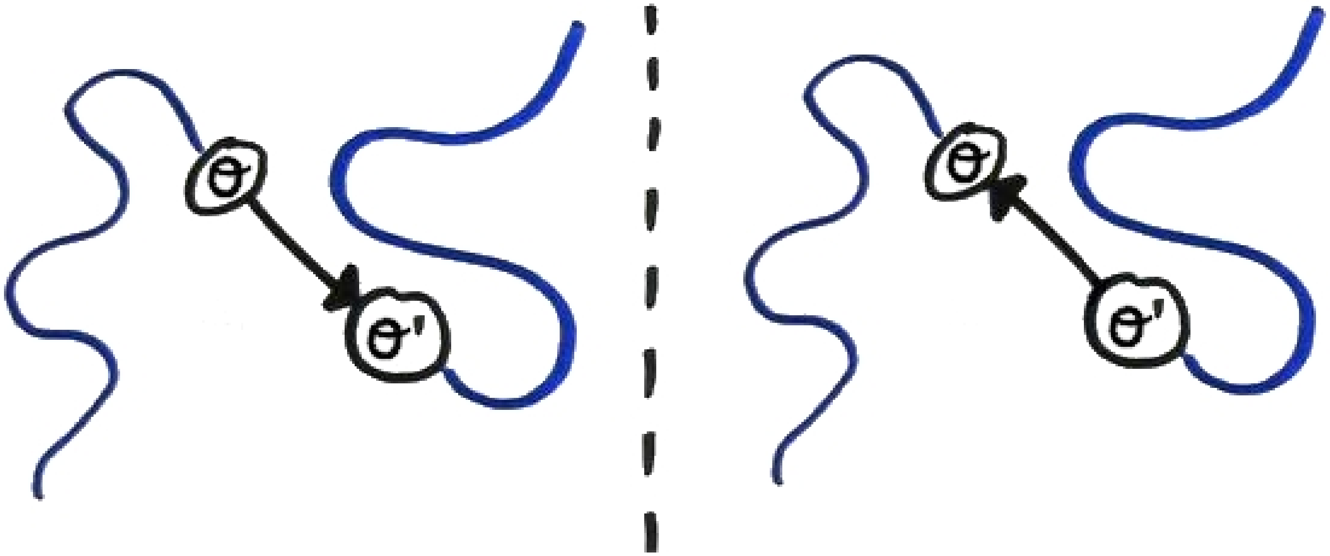
Necessary condition:

$$T(\theta' \!\leftarrow\! \theta)\, \pi(\theta) \;=\; R(\theta \!\leftarrow\! \theta')\, \pi(\theta'), \qquad \forall \theta, \theta'$$

If $R = T$, known as **detailed balance** (not necessary)

Balance condition

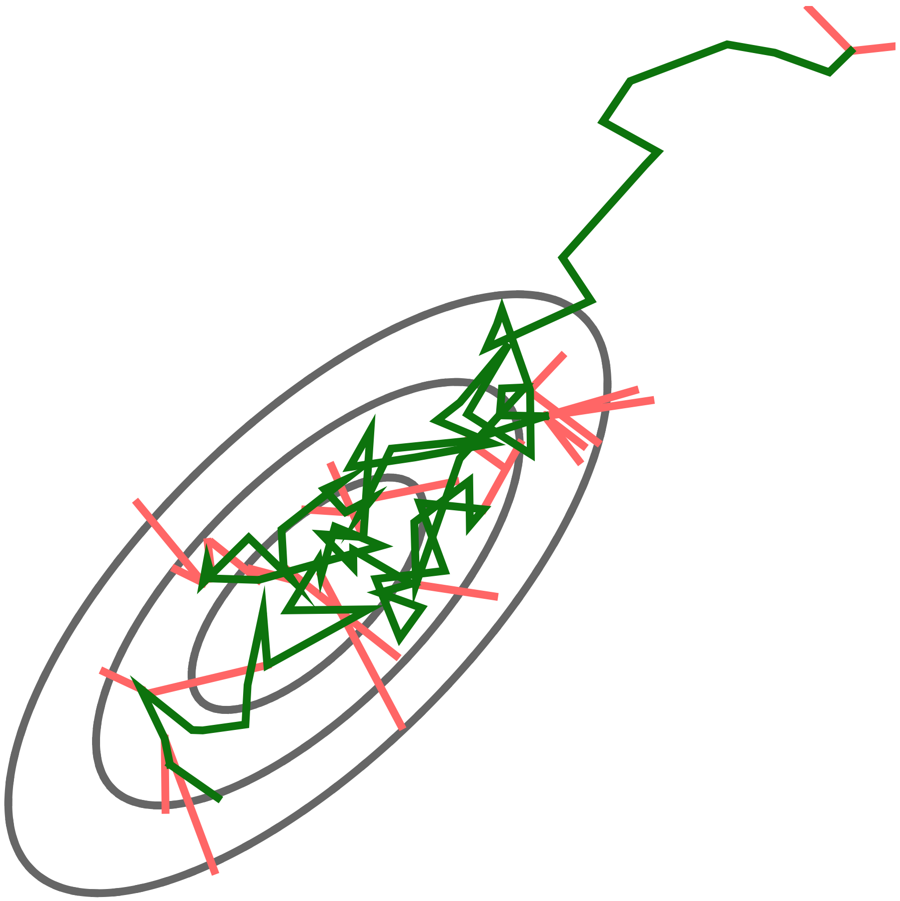
$$T(\theta' \leftarrow \theta)\,\pi(\theta) \;=\; R(\theta \leftarrow \theta')\,\pi(\theta')$$



Implies that $\pi(\theta)$ is left invariant:

$$\int T(\theta' \!\leftarrow\! \theta)\,\pi(\theta)\;\mathrm{d}\theta \;=\; \pi(\theta')\,\cancelto{1}{\int R(\theta \!\leftarrow\! \theta')\;\mathrm{d}\theta}$$

Metropolis–Hastings



$$\theta' \sim q(\theta'; \theta^{(s)})$$

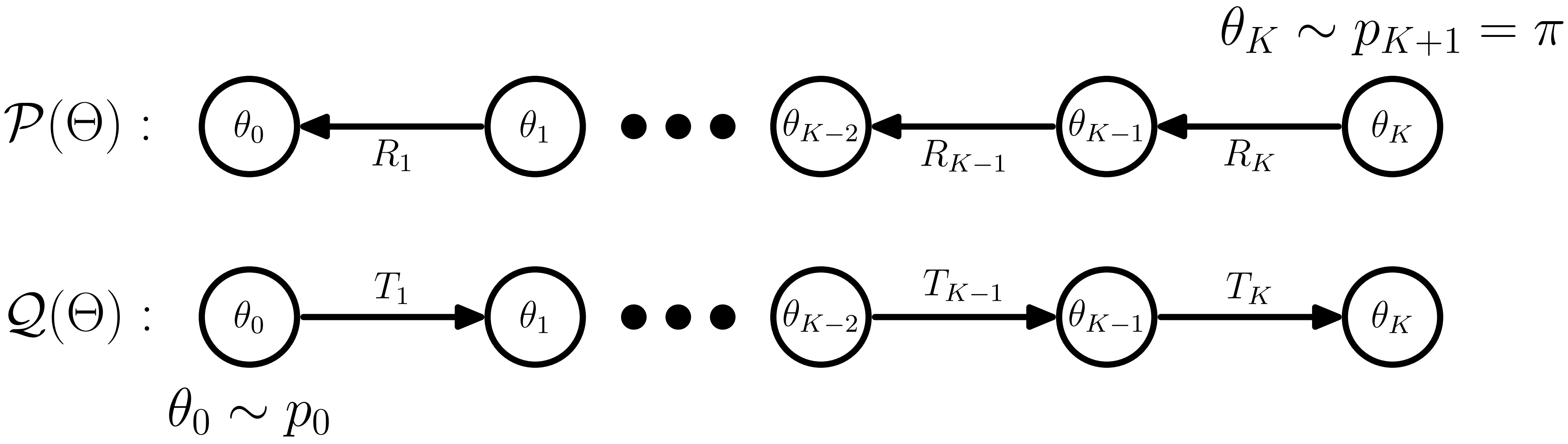
```
if accept:
    $\theta^{(s+1)} \leftarrow \theta'$
else:
    $\theta^{(s+1)} \leftarrow \theta^{(s)}$
```

Transitions satisfy detailed balance:

$$T(\theta' \!\leftarrow\! \theta) = \begin{cases} q(\theta';\theta)\,\min\left(1,\, \frac{\pi(\theta')\,q(\theta;\theta')}{\pi(\theta)\,q(\theta';\theta)}\right) & \theta' \neq \theta \\ \ldots & \theta' = \theta \end{cases}$$

$T(\theta' \!\leftarrow\! \theta)\pi(\theta)$ symmetric in θ, θ'

Annealed Importance Sampling

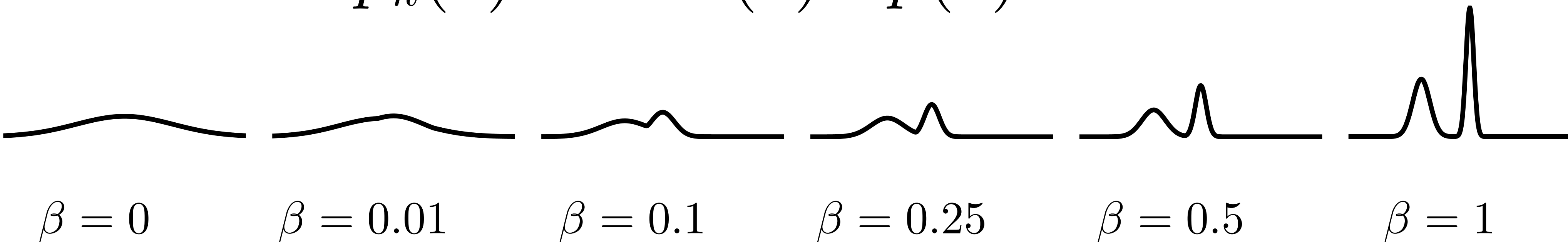


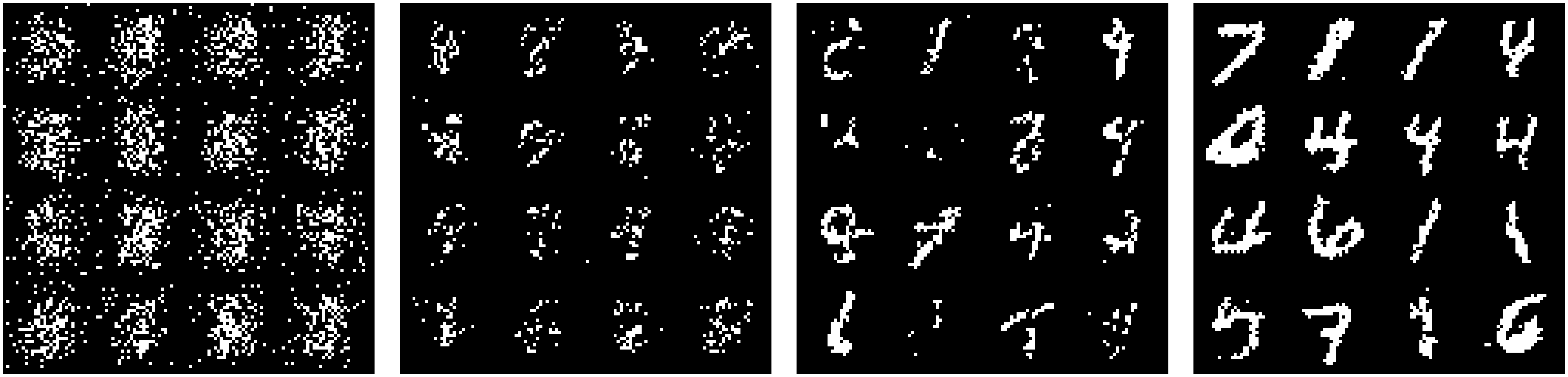
$$\mathcal{P}(\Theta) = \frac{\pi^*(\theta_K)}{\mathcal{Z}} \prod_{k=1}^{K} R_k(\theta_{k-1};\theta_k), \qquad \mathcal{Q}(\Theta) = \pi(\theta_0) \prod_{k=1}^{K} T_k(\theta_k;\theta_{k-1})$$

Standard importance sampling of $\mathcal{P}(\Theta) = \frac{\mathcal{P}^*(\Theta)}{\mathcal{Z}}$ with $\mathcal{Q}(\Theta)$

Annealing / Tempering

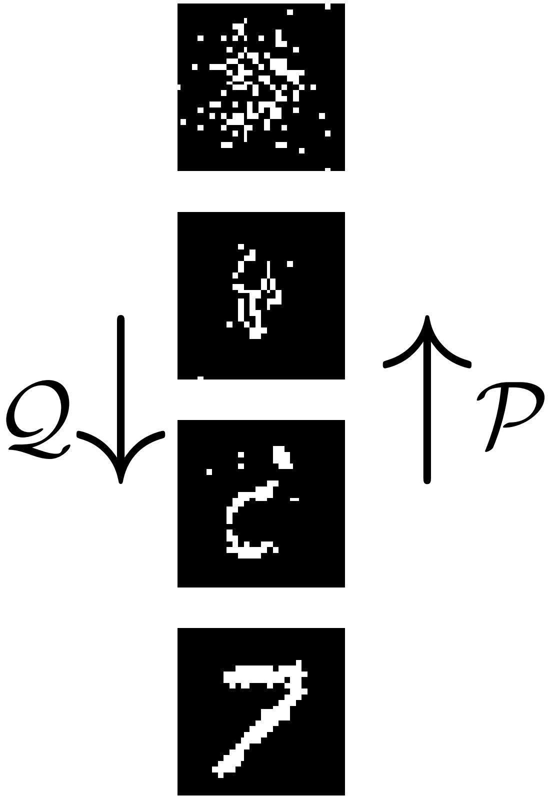
$$p_k(\theta) \;\propto\; \pi^*(\theta)^{\beta_k}\, p(\theta)^{(1-\beta_k)}$$





$1/\beta$ = “temperature”

Annealed Importance Sampling

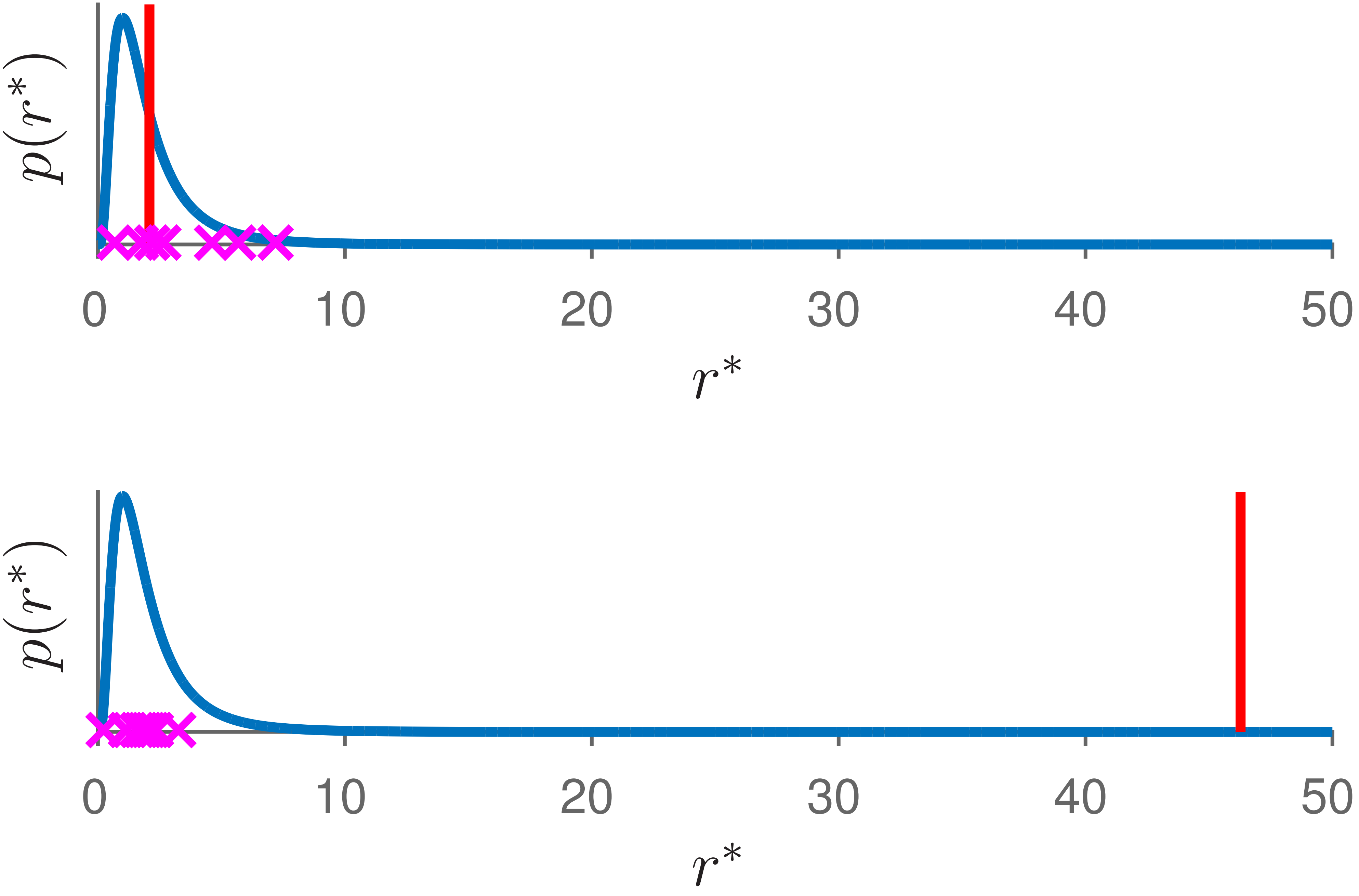


$$r^{*(s)} = \frac{\mathcal{P}^*(\Theta^{(s)})}{\mathcal{Q}(\Theta^{(s)})}, \qquad \Theta^{(s)} \sim \mathcal{Q}$$

$$\mathcal{Z} \approx \frac{1}{S}\sum_{s=1}^{S} r^{*(s)}$$

$$\int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta \;\approx\; \sum_{s} f(\theta^{(s)})\,r^{(s)}$$

Unbiased positive estimators



Importance sampling in MCMC

AIS for proposals → *tempered transitions* (Neal, 1996)

Recent work: *particle MCMC* and *particle Gibbs*

(uses more general Sequential Monte Carlo (SMC) methods)

Free to combine with more traditional updates

Matlab/Octave code for demo

```
function samples = metropolis(init, log_pstar_fn, iters, sigma)

D = numel(init);
samples = zeros(D, iters);

state = init;
Lp_state = log_pstar_fn(state);
for ss = 1:iters
    % Propose
    prop = state + sigma*randn(size(state));
    Lp_prop = log_pstar_fn(prop);
    if rand() < exp(Lp_prop - Lp_state)
        % Accept
        state = prop;
        Lp_state = Lp_prop;
    end
    samples(:, ss) = state(:);
end
```

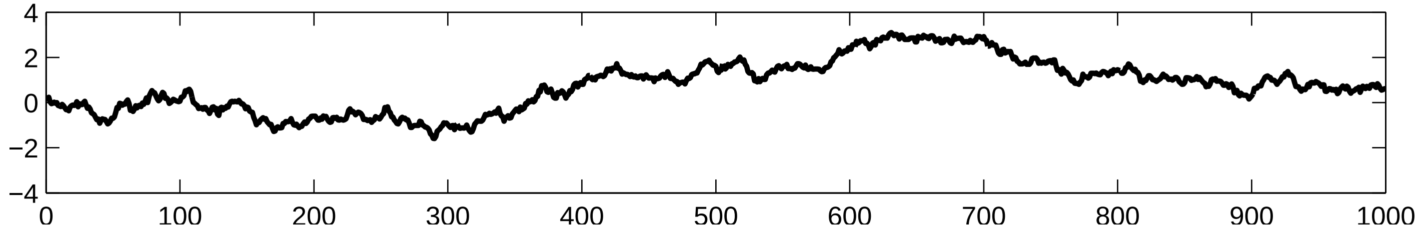
Step-size demo

Explore $\mathcal{N}(0, 1)$ with different step sizes σ

```
sigma = @(s) plot(metropolis(0, @(x)-0.5*x*x, 1e3, s));
```

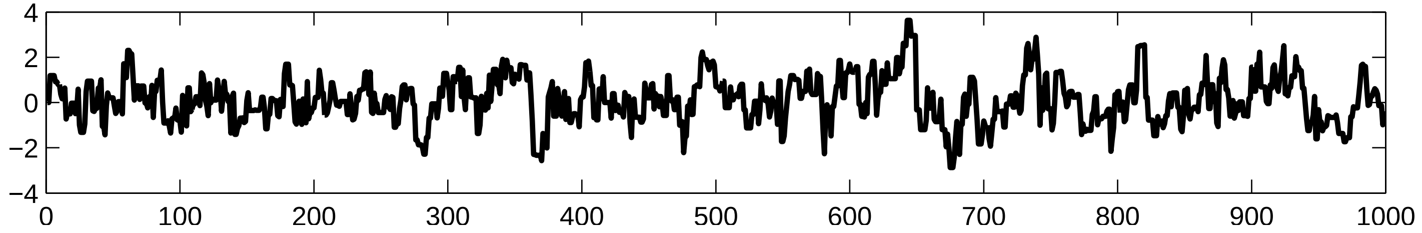
`sigma(0.1)`

99.8% accepts



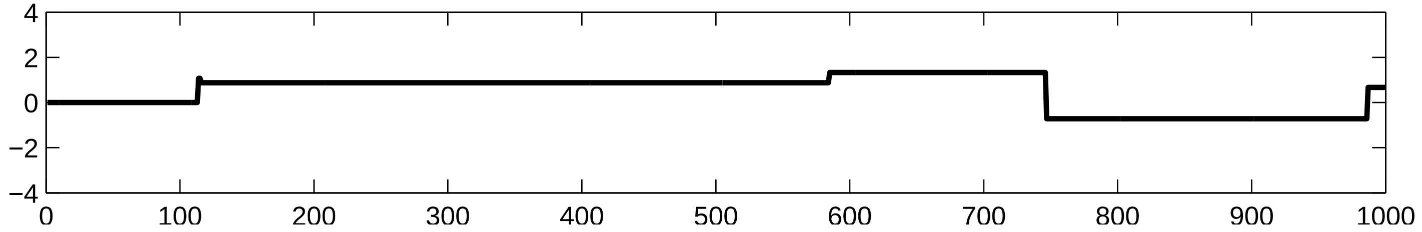
`sigma(1)`

68.4% accepts

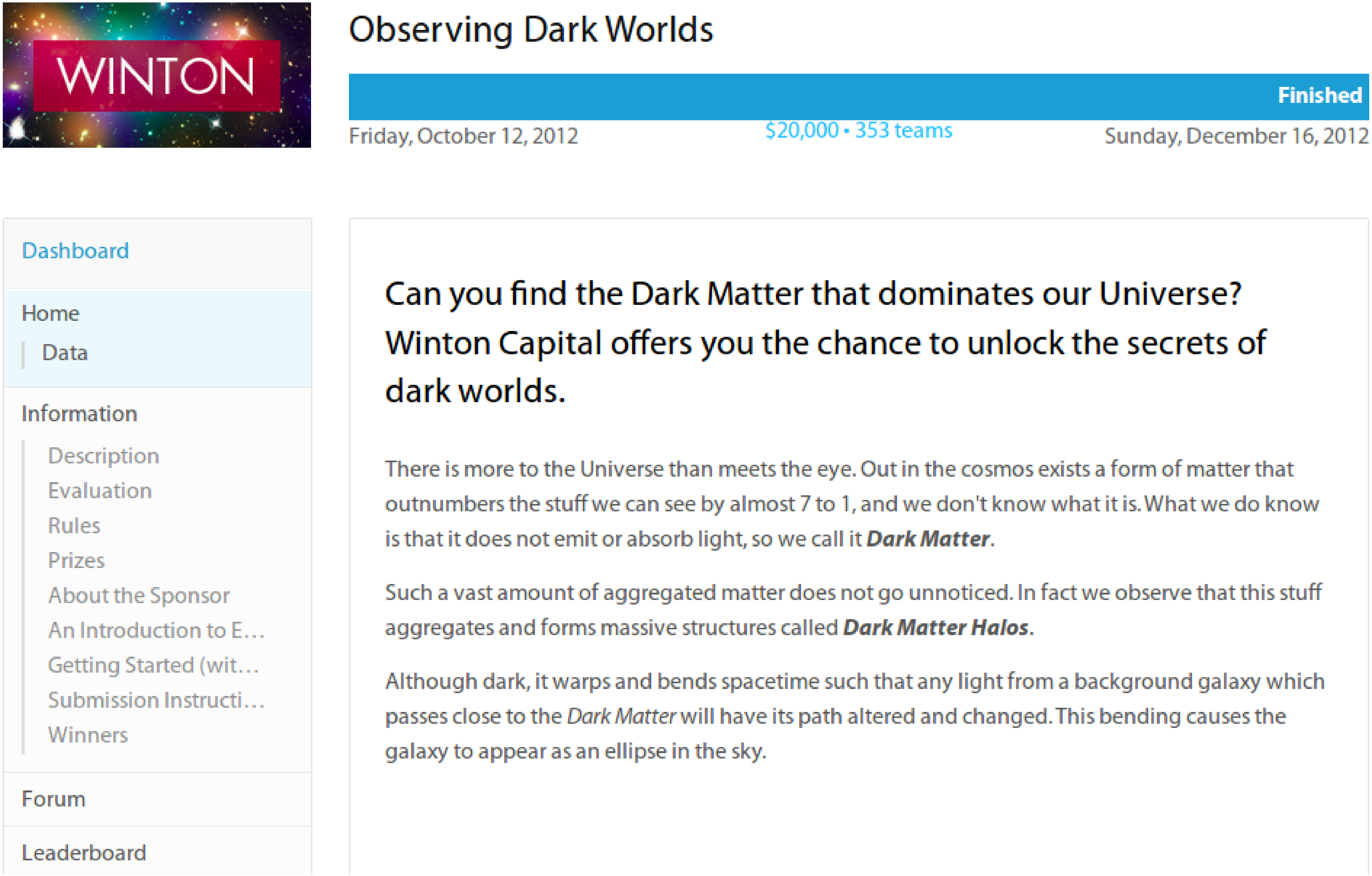


`sigma(100)`

0.5% accepts



http://www.kaggle.com/c/DarkWorlds



Observing Dark Worlds

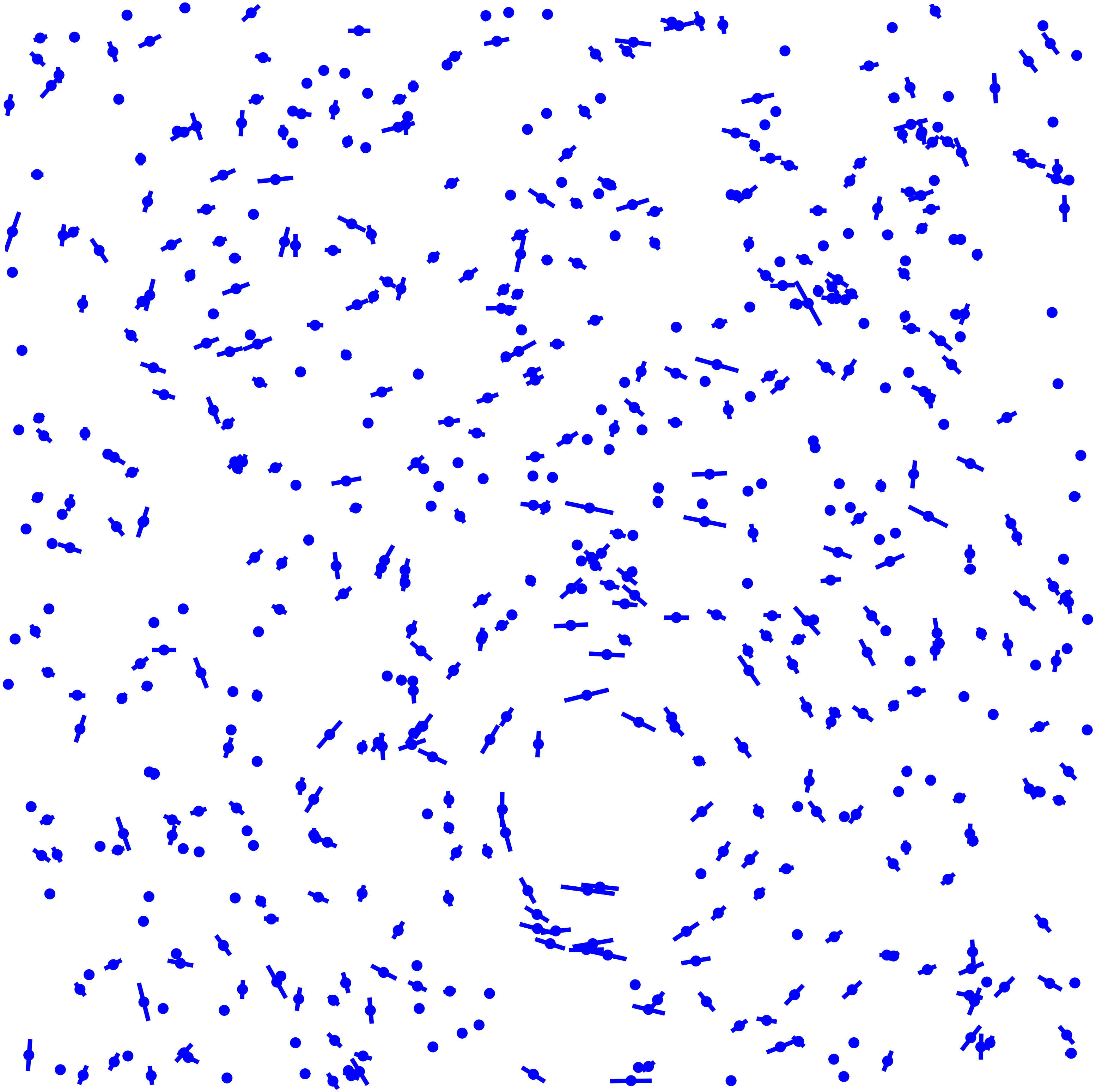
Friday, October 12, 2012 — $20,000 • 353 teams — Sunday, December 16, 2012

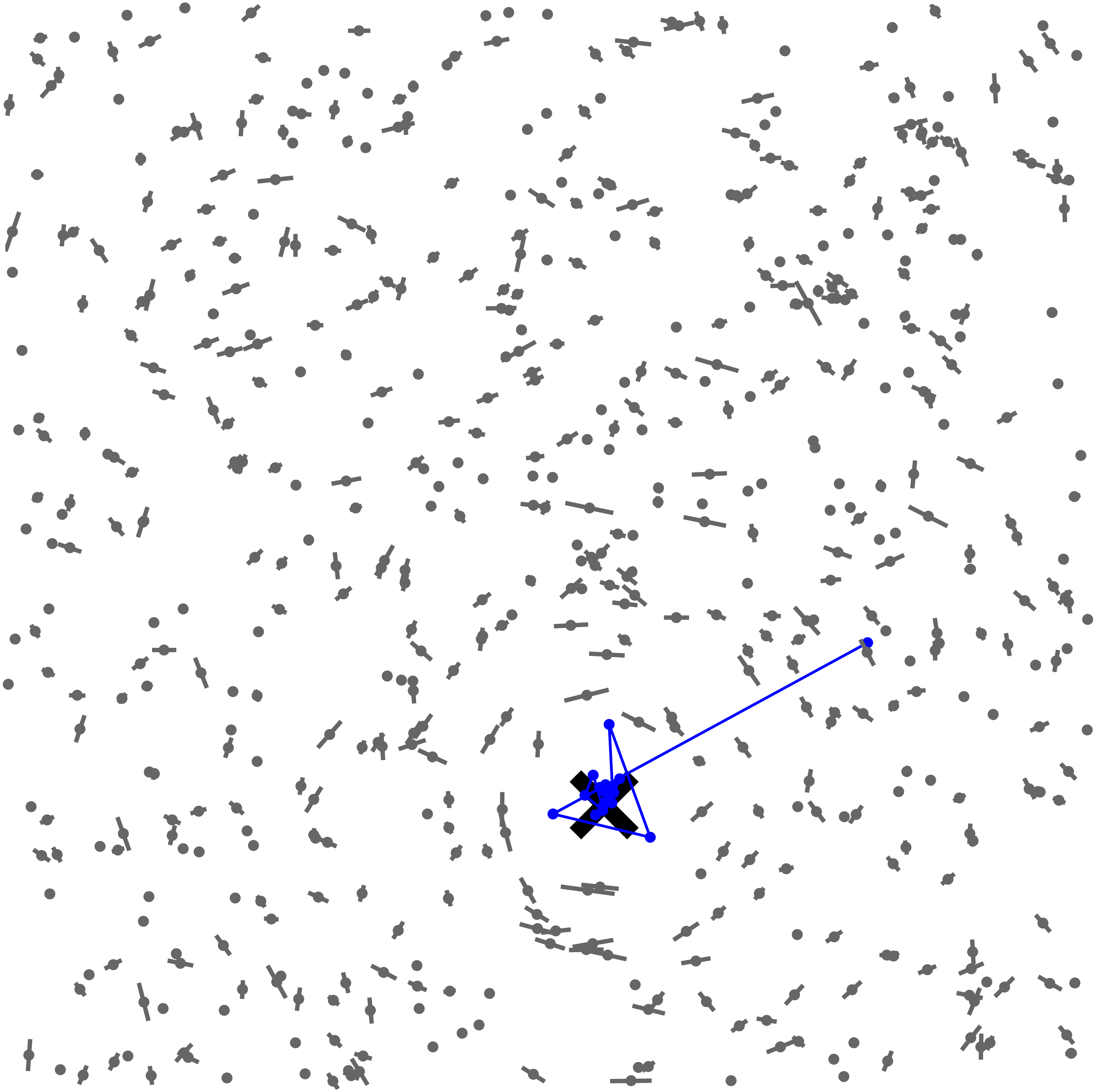
Can you find the Dark Matter that dominates our Universe? Winton Capital offers you the chance to unlock the secrets of dark worlds.

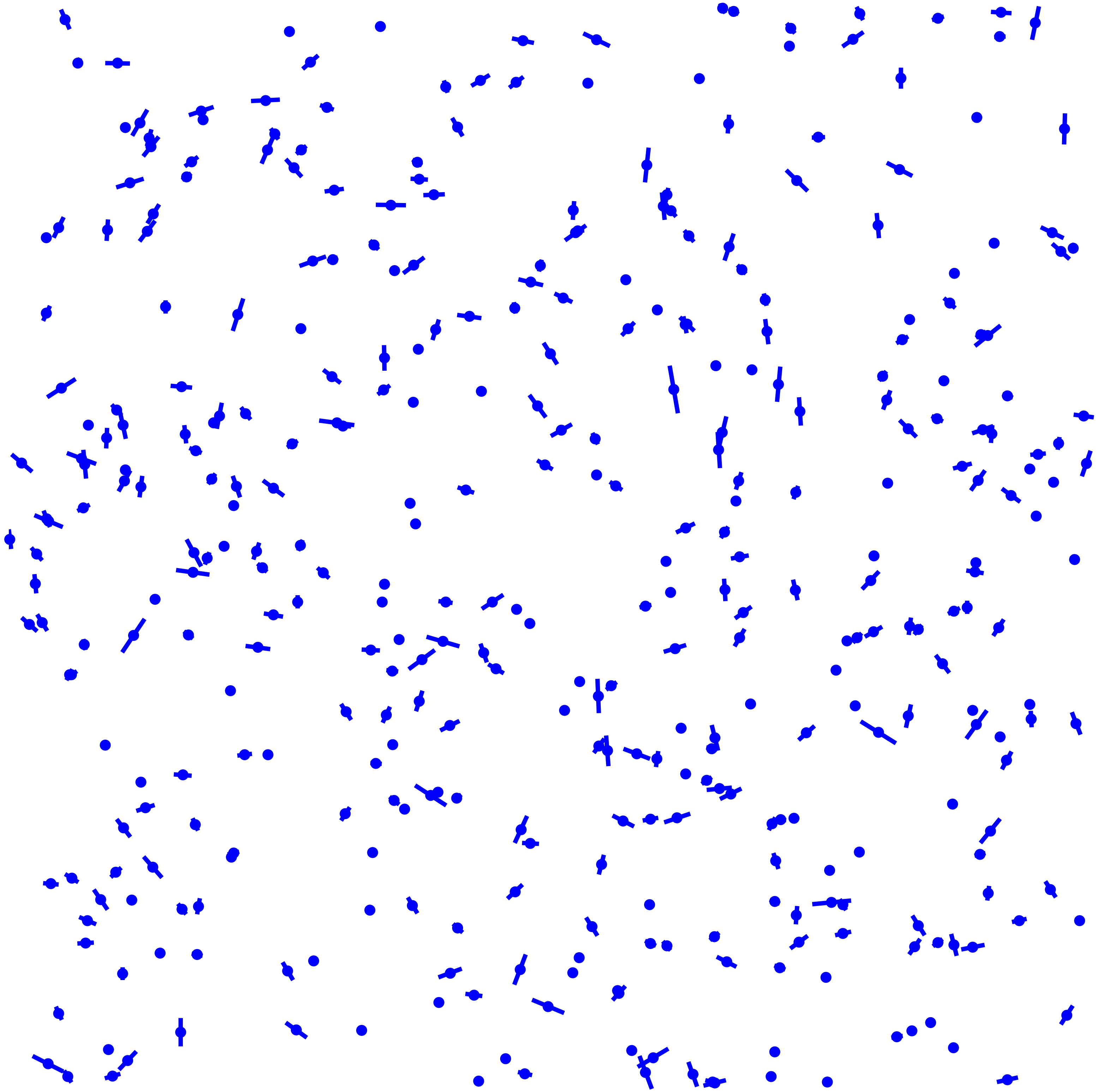
There is more to the Universe than meets the eye. Out in the cosmos exists a form of matter that outnumbers the stuff we can see by almost 7 to 1, and we don't know what it is. What we do know is that it does not emit or absorb light, so we call it ***Dark Matter***.

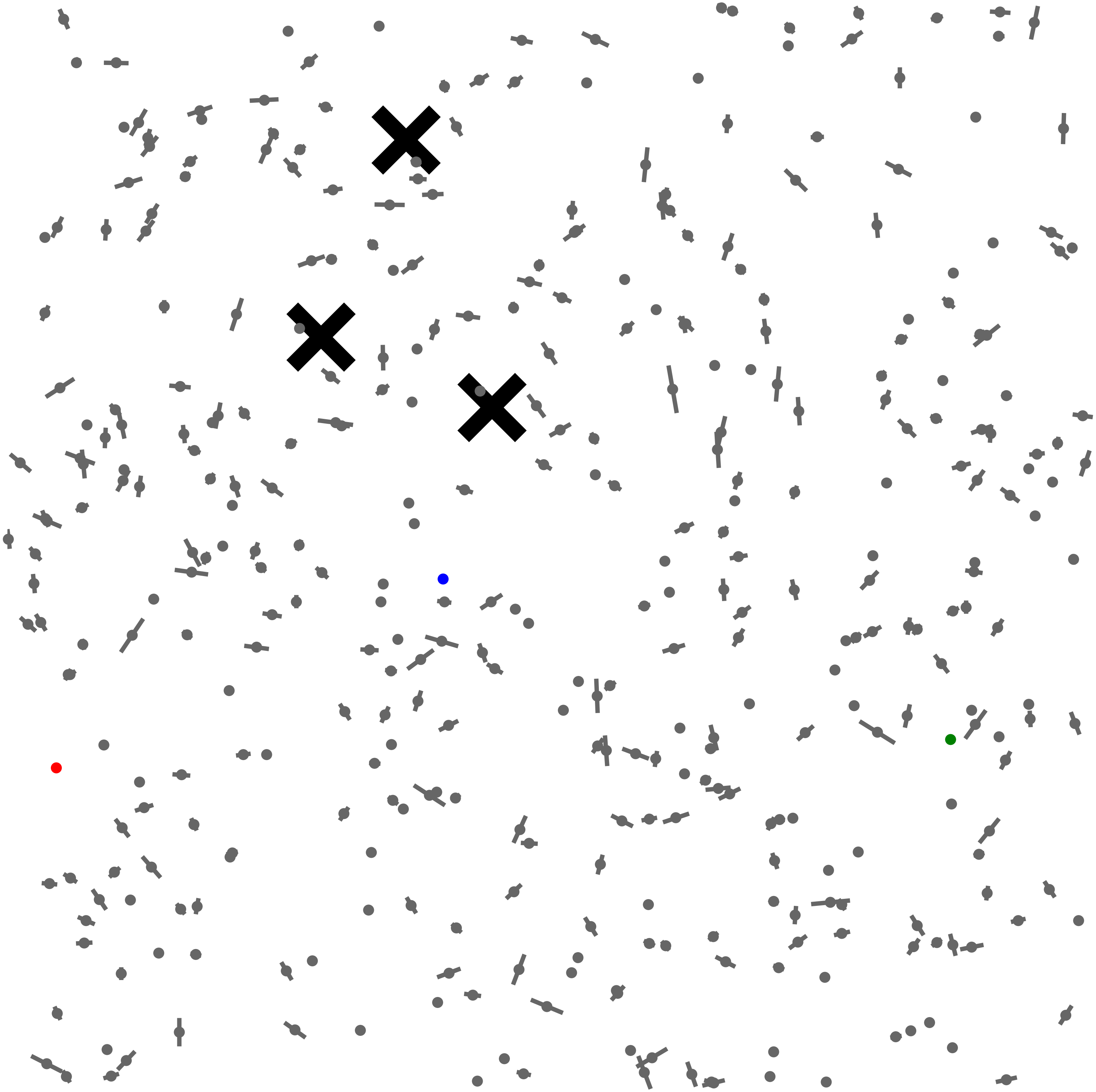
Such a vast amount of aggregated matter does not go unnoticed. In fact we observe that this stuff aggregates and forms massive structures called ***Dark Matter Halos***.

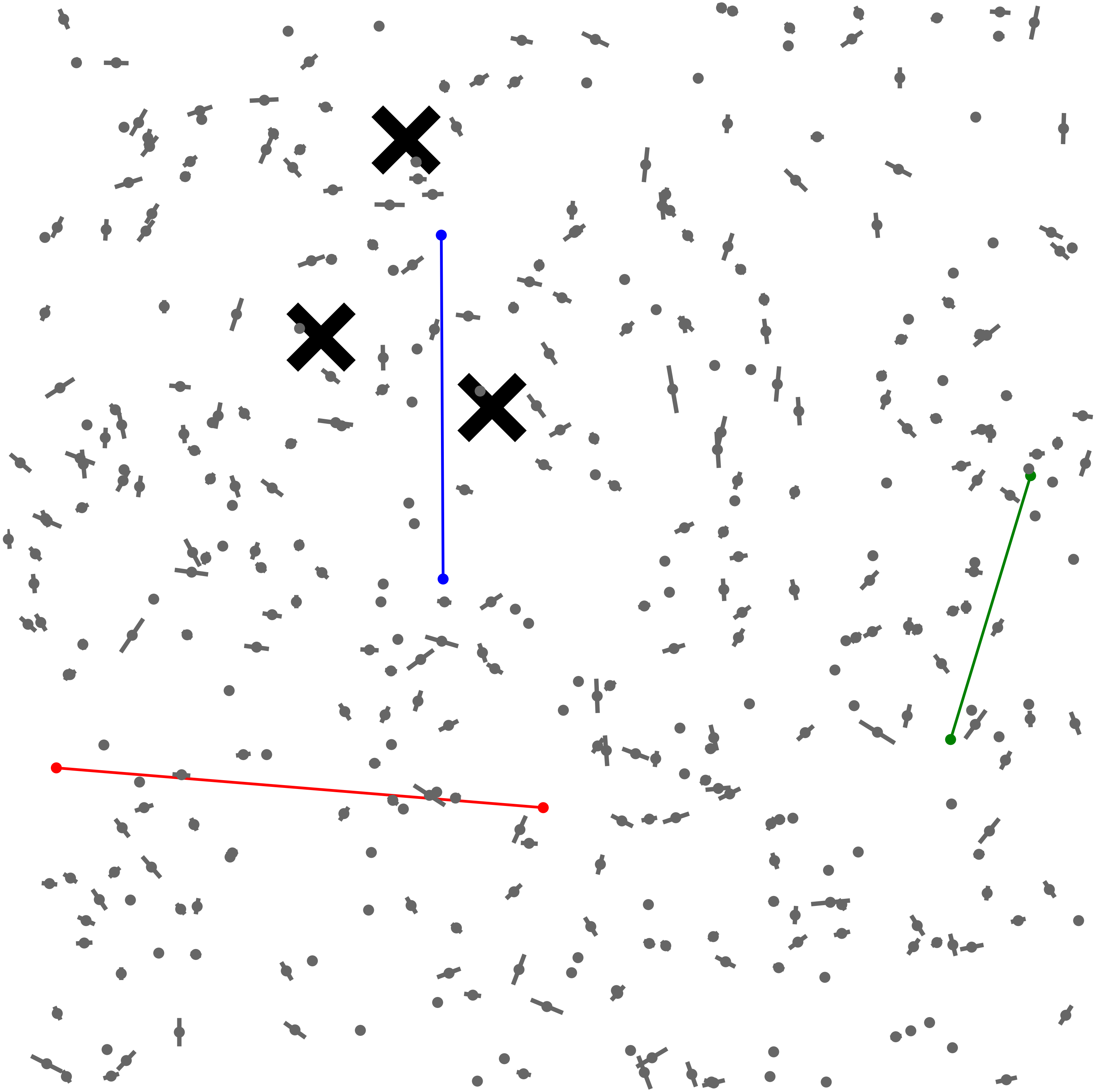
Although dark, it warps and bends spacetime such that any light from a background galaxy which passes close to the *Dark Matter* will have its path altered and changed. This bending causes the galaxy to appear as an ellipse in the sky.

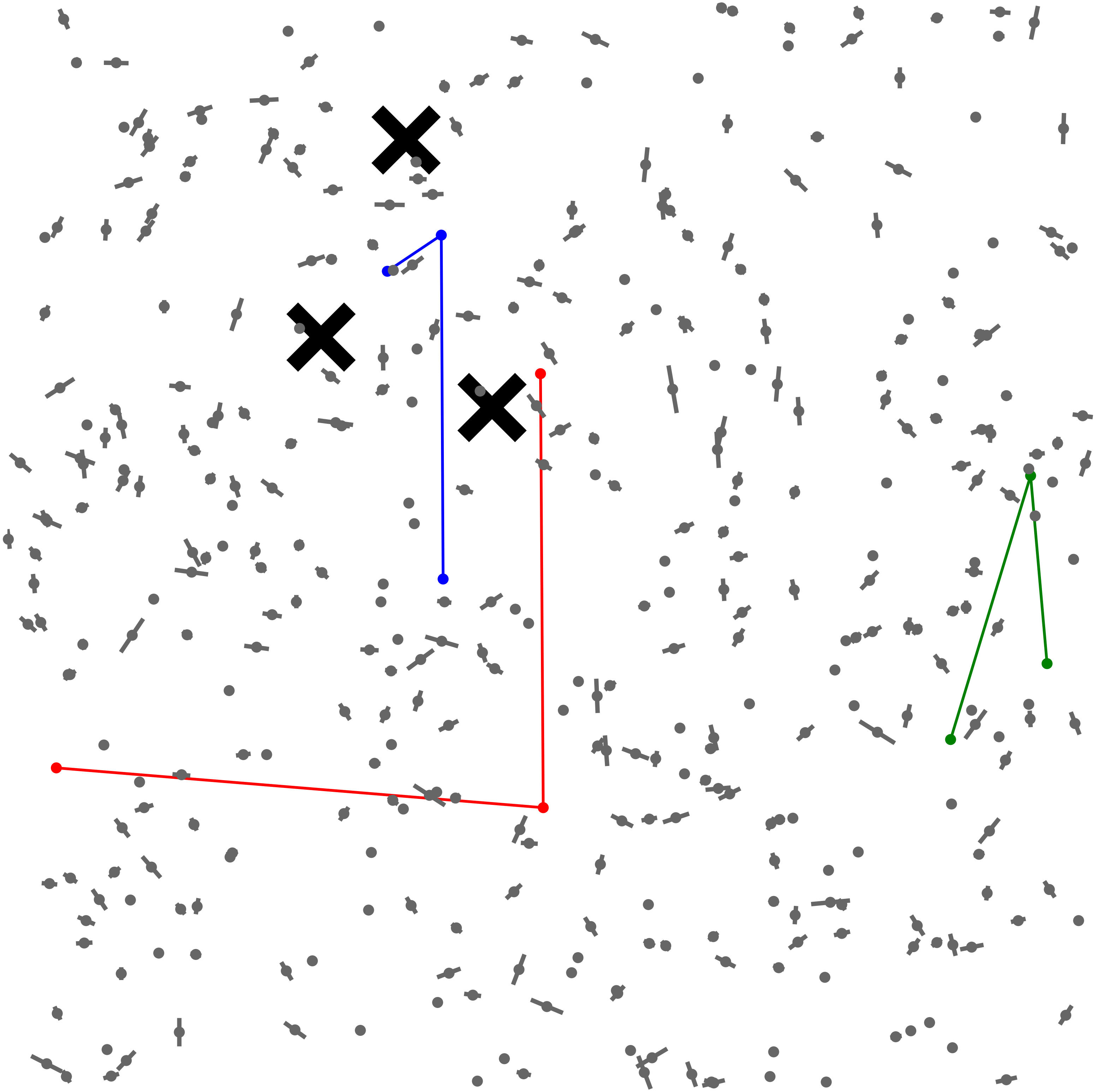


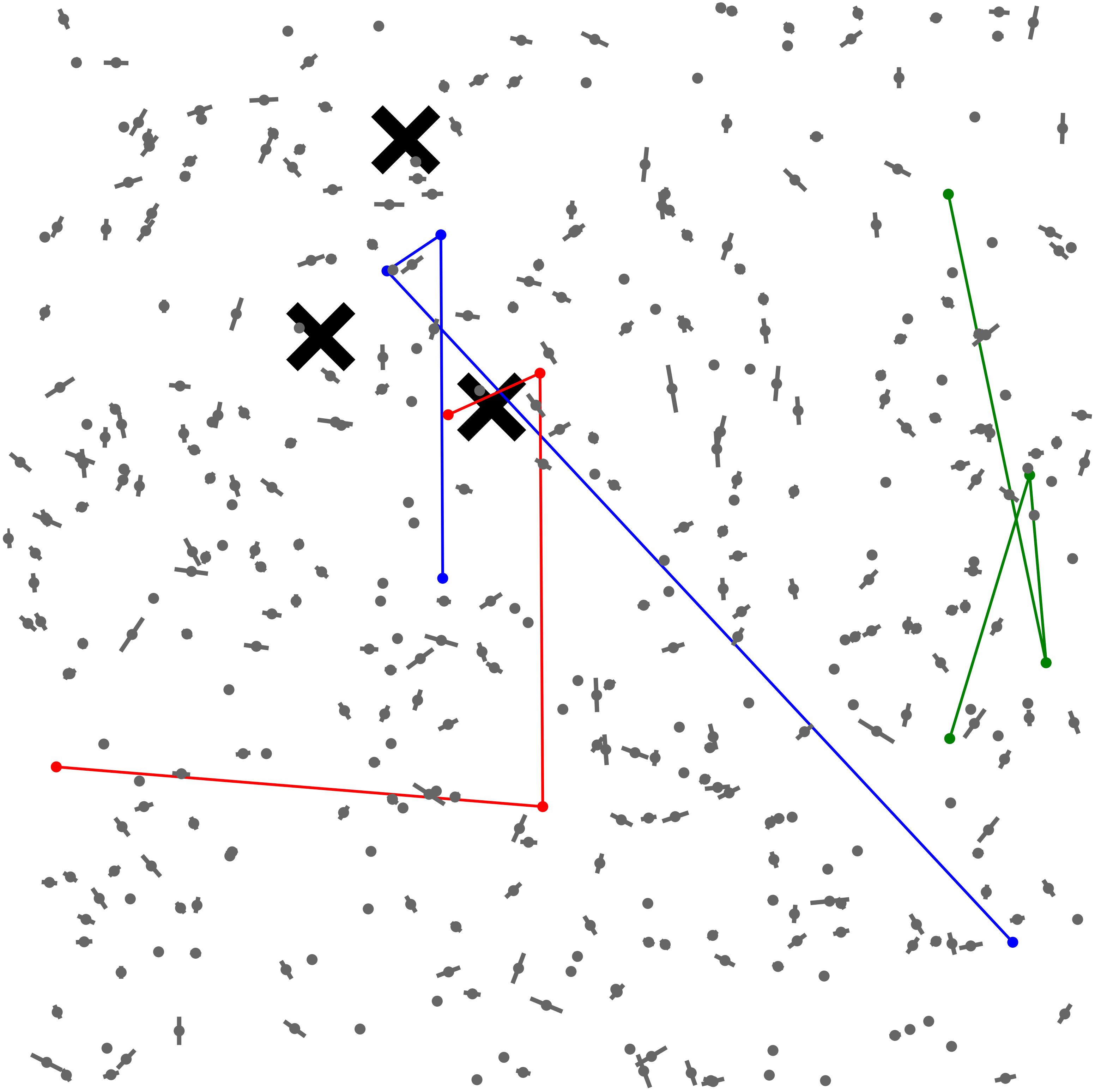


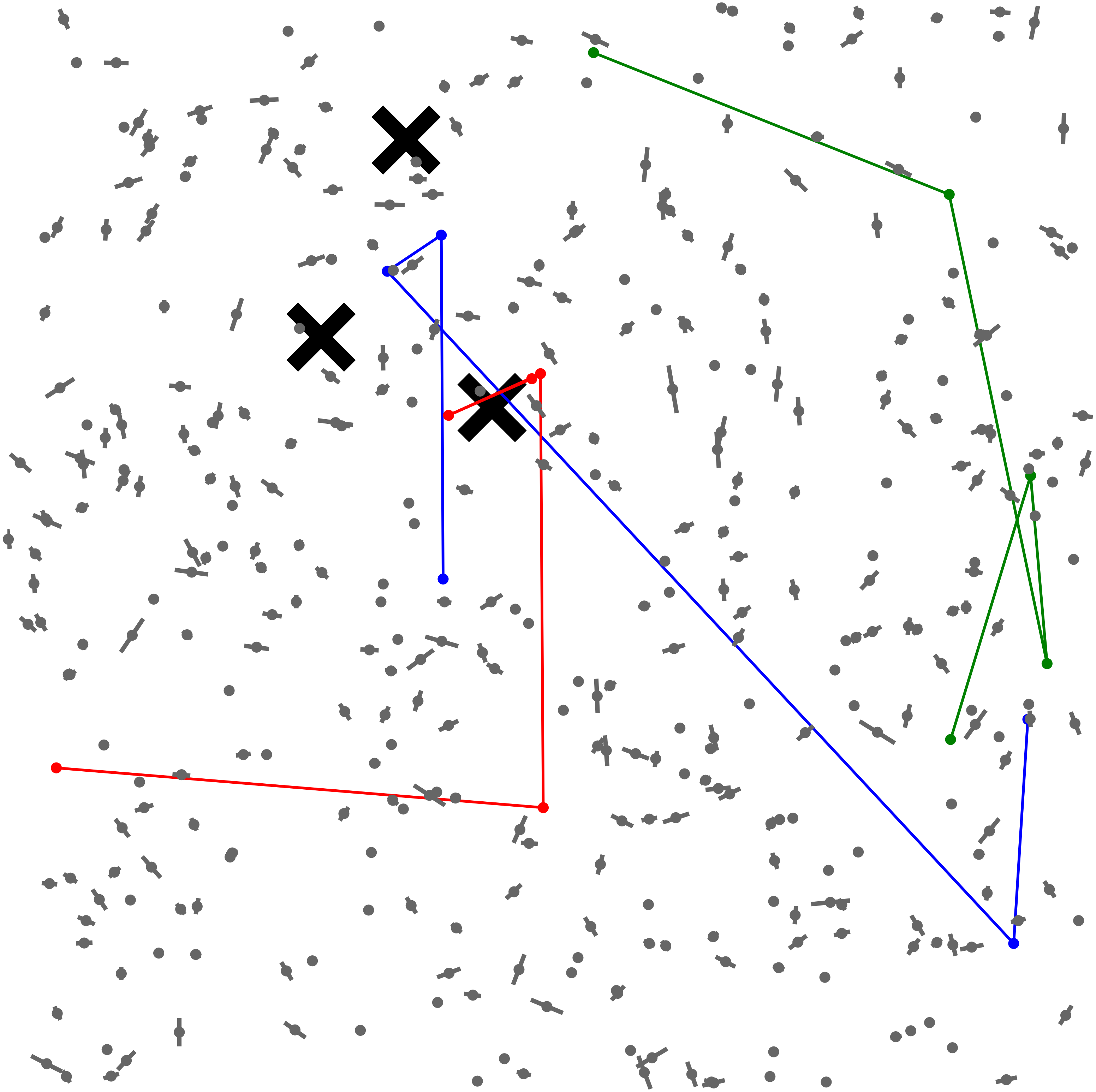


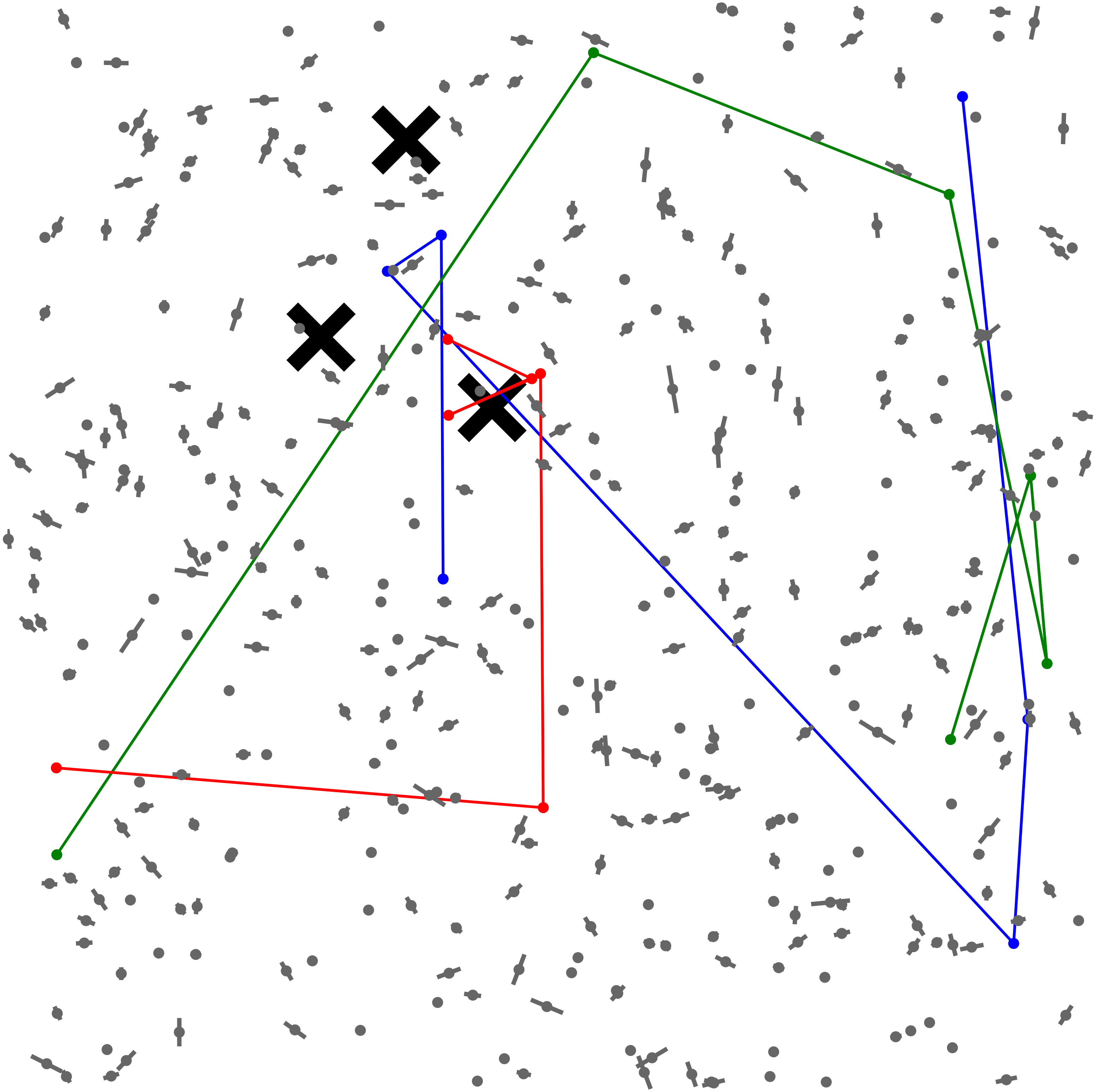


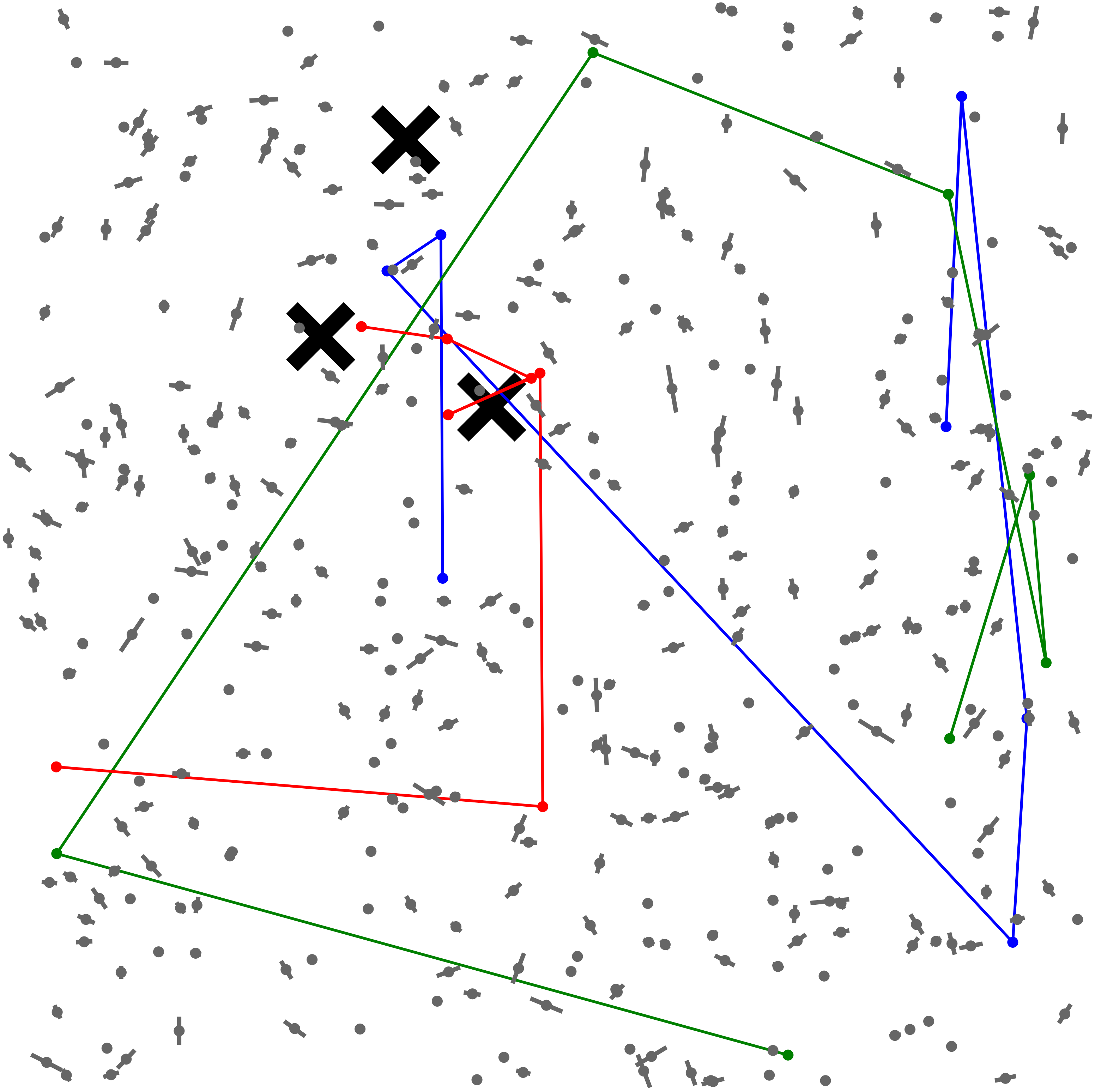


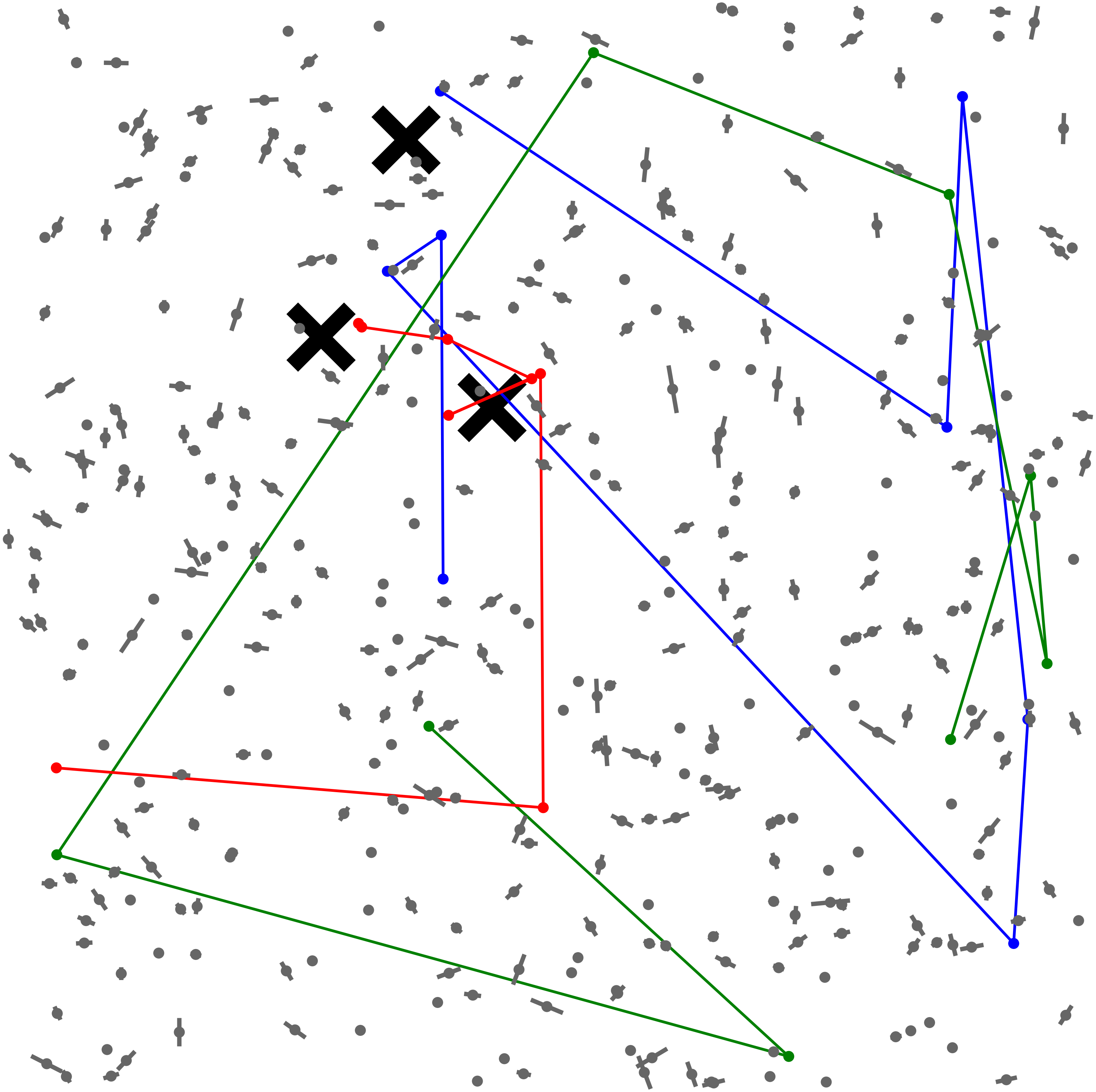


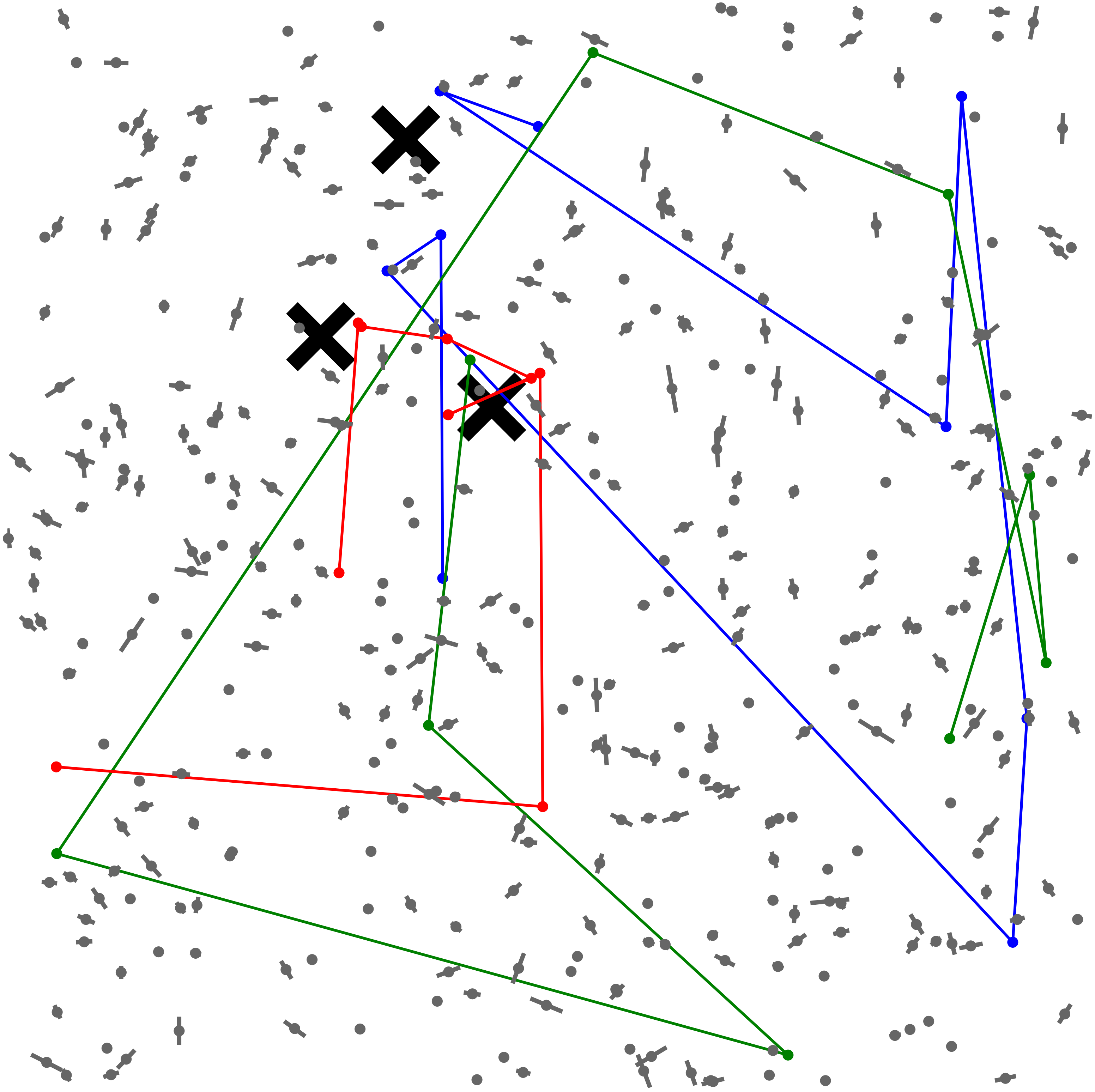


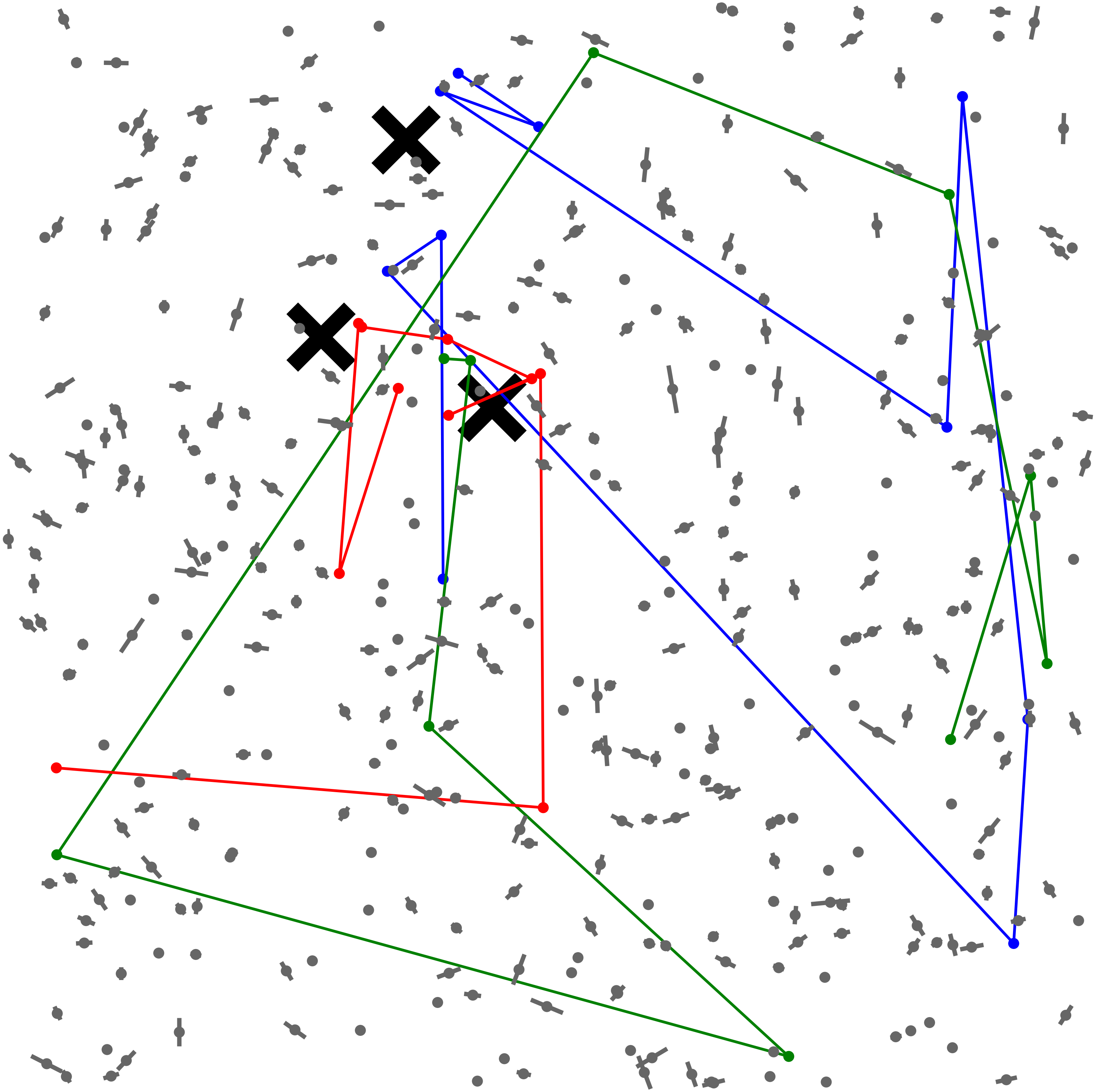


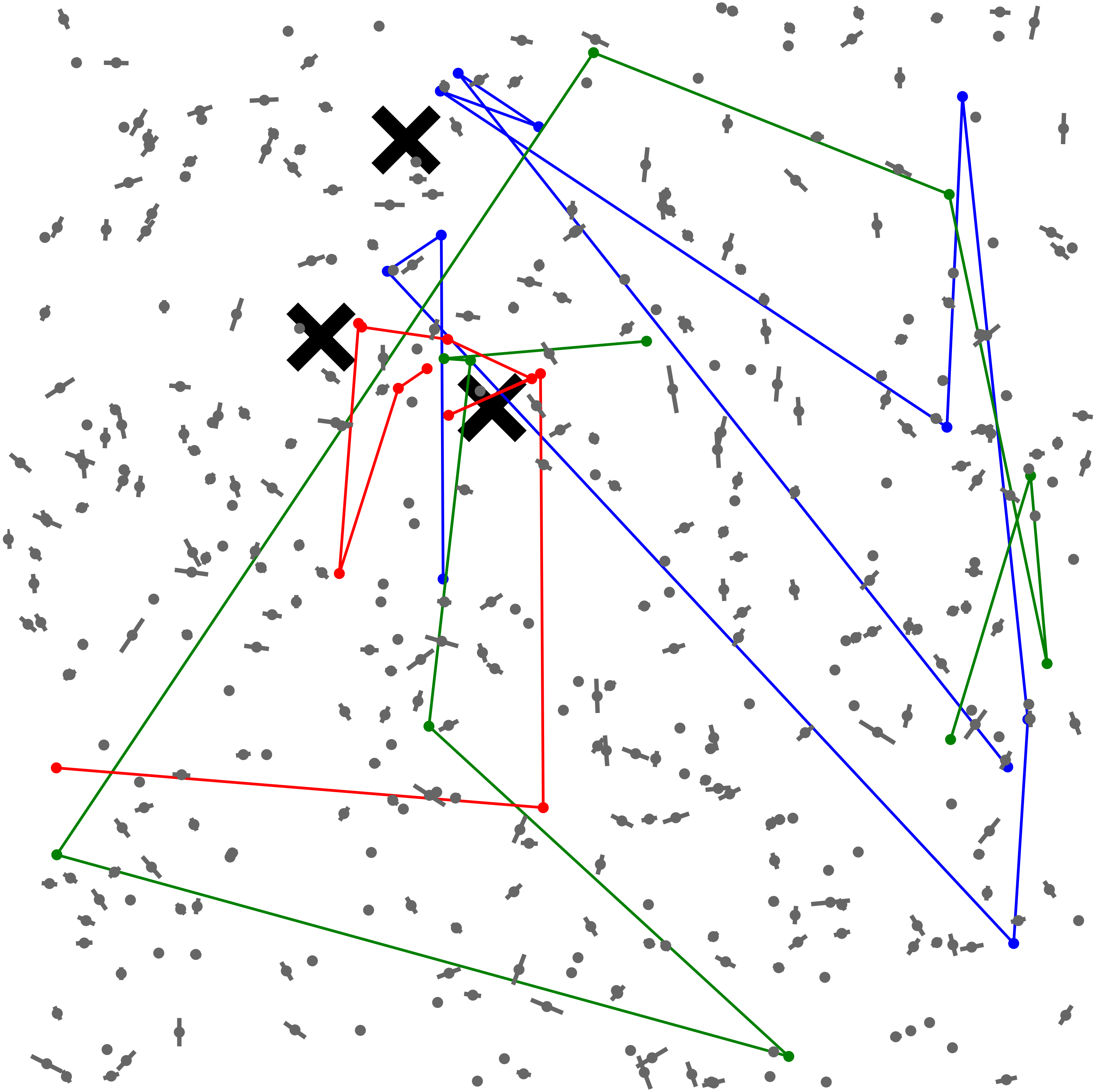


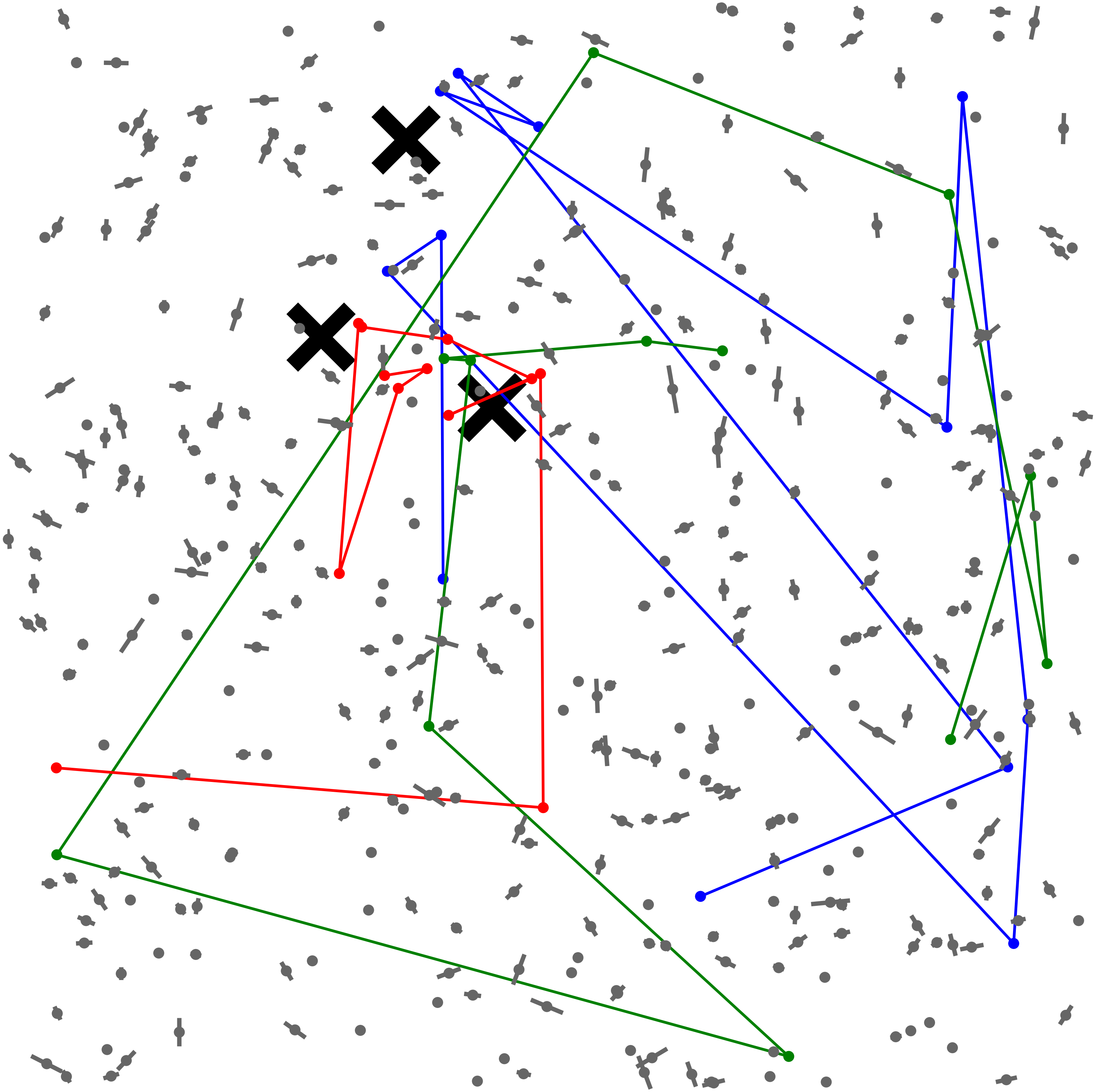


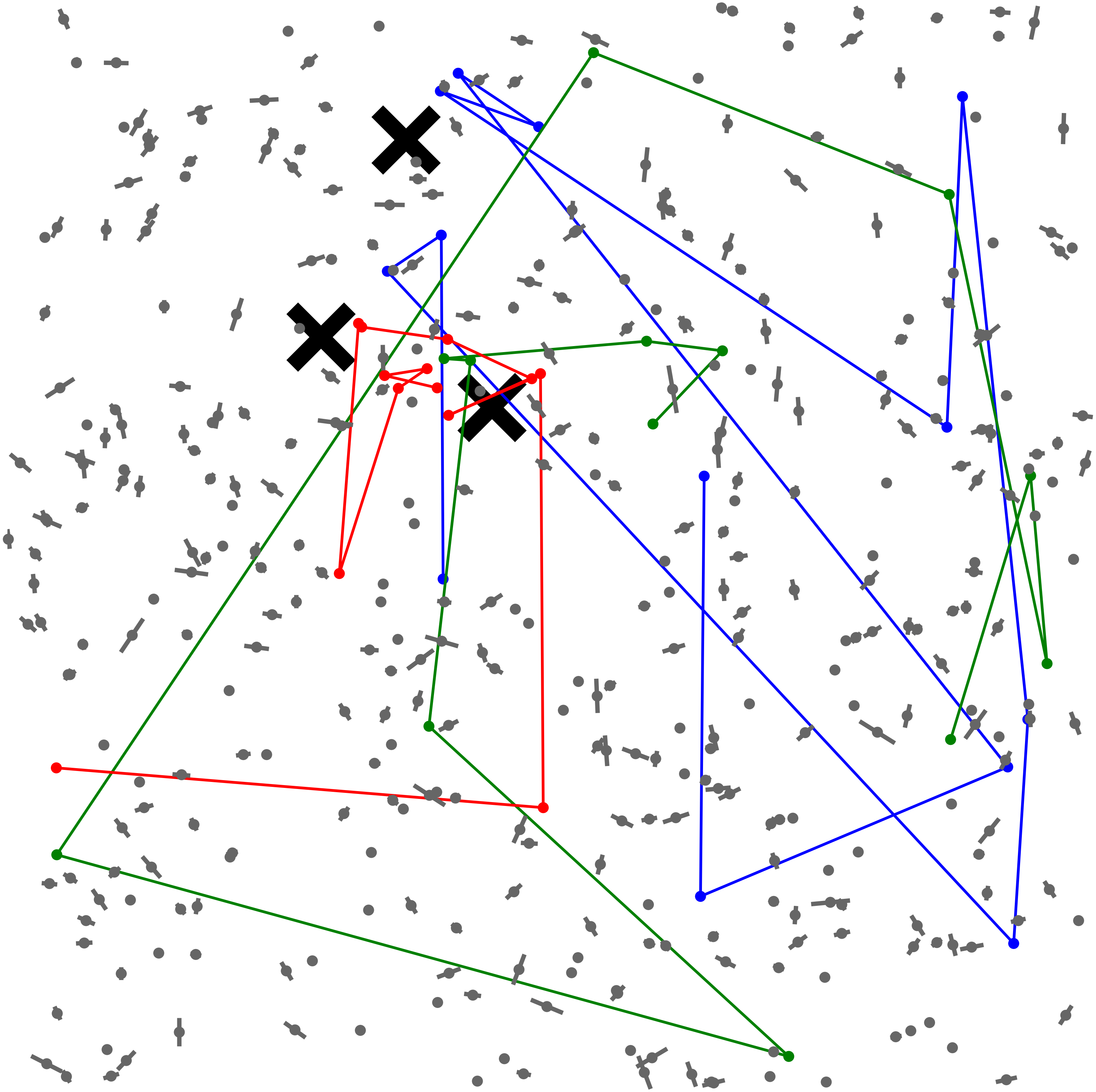


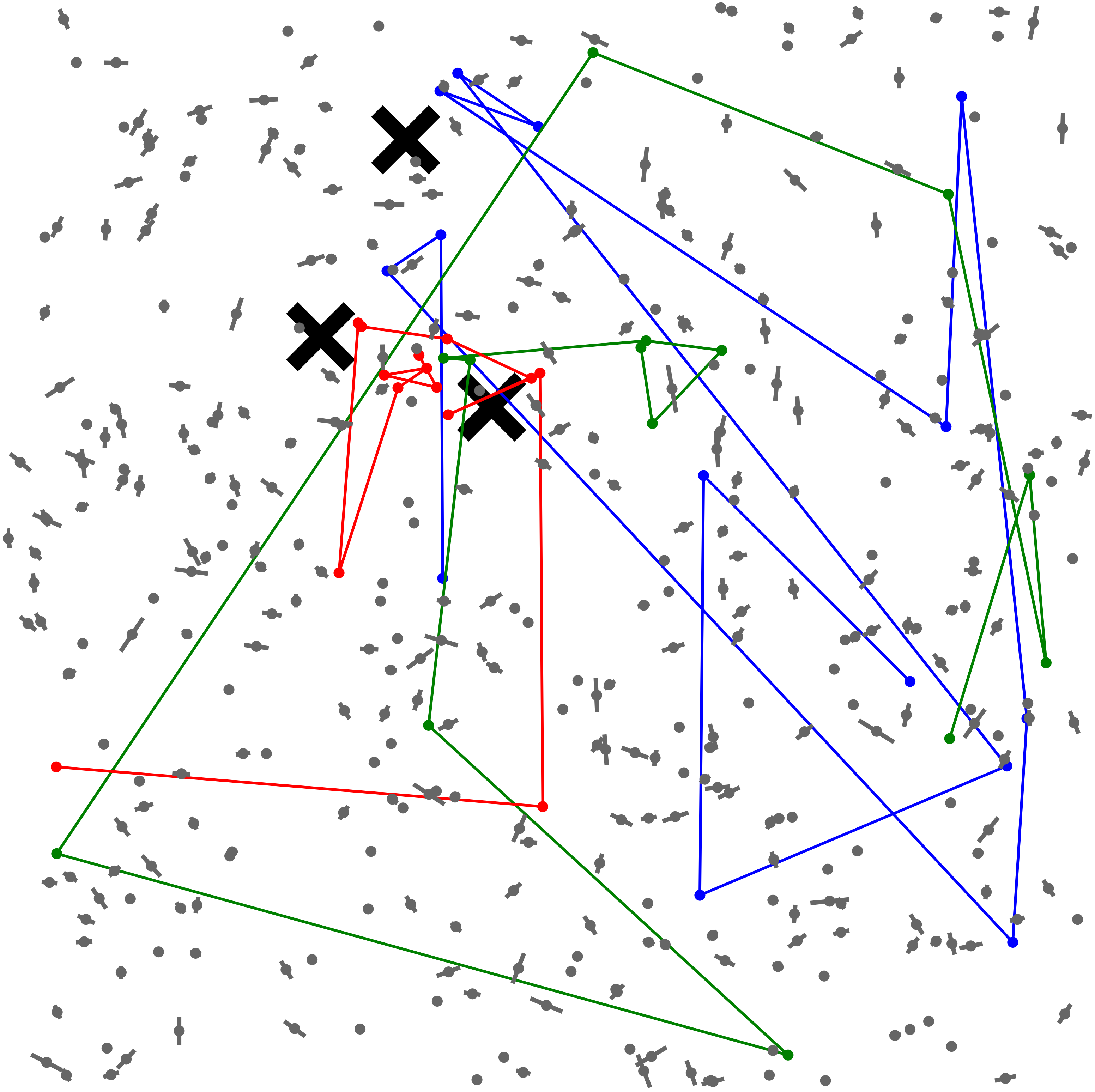


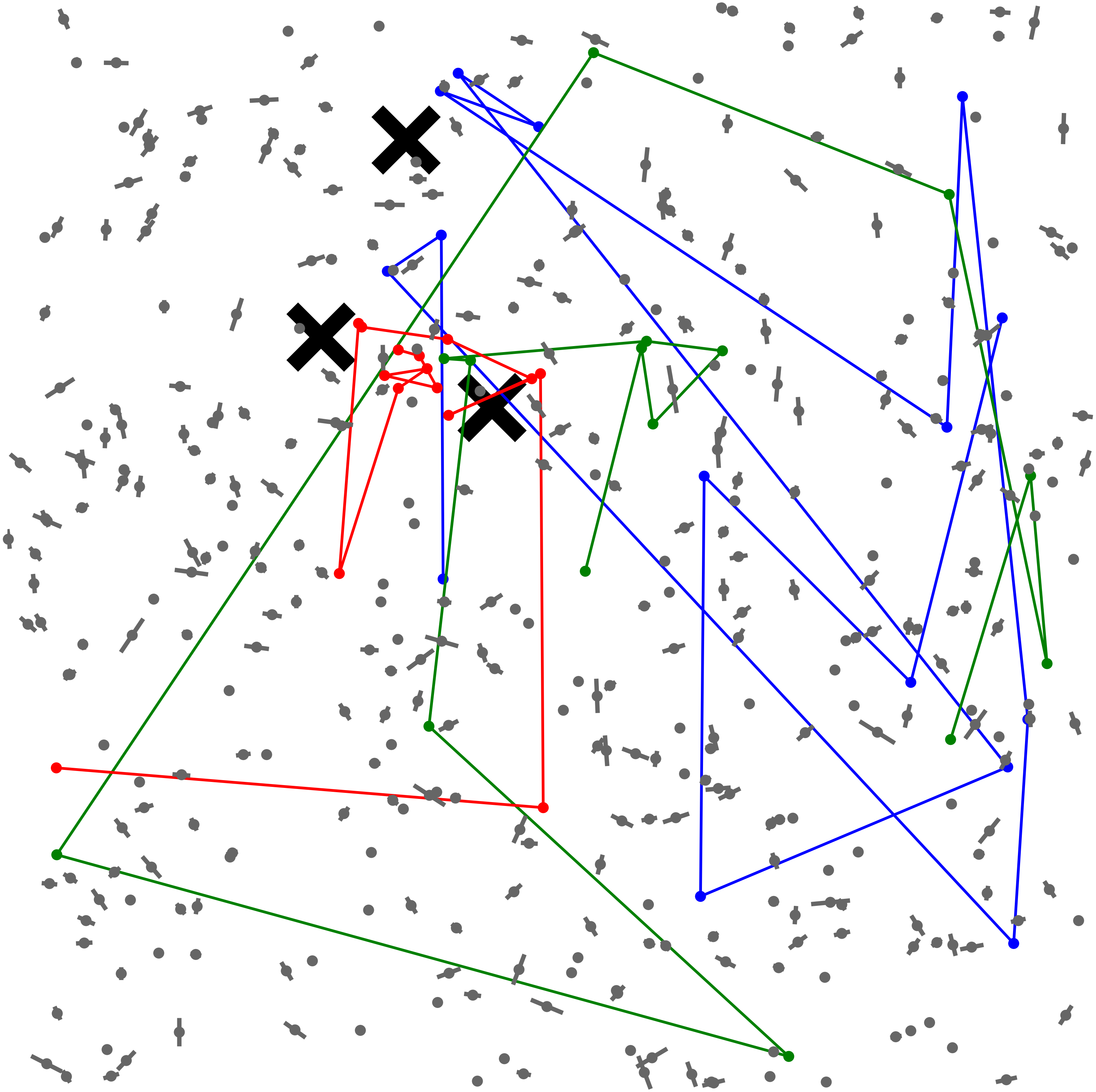












■ ■ ■