STAT 513/413: Lecture 16
Markov chain Monte Carlo: the technology

(Markov chains with continuous state space, Metropolis algorithm, Gibbs sampler)

First, a bit about continuous setting

While Markov chains with finite state spaces can be useful, and also are very nice to deal with in: especially, when learning something, it is often easier to work in the discrete setting. However, Markov chains with continuous state spaces make for most of the interesting applications... Thus, in continuous setting, typically:

- the state space S is continuous, like, $\mathbb{R}^p$

- the distribution of X_t given $X_{t-1} = x$ is given by a density $p_x(y)$

It is then not practical to organize $p_x(y)$ in to a matrix

Sums become integrals: if X_{t-1} has *density* $\pi_{t-1}(x)$, then

$$\pi_t(y) = \sum_x \pi_{t-1}(x) P_{xy}$$

First, a bit about continuous setting

While Markov chains with finite state spaces can be useful, and also are very nice to deal with in: especially, when learning something, it is often easier to work in the discrete setting. However, Markov chains with continuous state spaces make for most of the interesting applications... Thus, in continuous setting, typically:

- the state space S is continuous, like, $\mathbb{R}^p$

- the distribution of X_t given $X_{t-1} = x$ is given by a density $p_x(y)$

It is then not practical to organize $p_x(y)$ in to a matrix

Sums become integrals: if X_{t-1} has *density* $\pi_{t-1}(x)$, then

$$\pi_t(y) = \sum_x \pi_{t-1}(x) p_{xy}$$

First, a bit about continuous setting

While Markov chains with finite state spaces can be useful, and also are very nice to deal with in: especially, when learning something, it is often easier to work in the discrete setting. However, Markov chains with continuous state spaces make for most of the interesting applications... Thus, in continuous setting, typically:

- the state space S is continuous, like, $\mathbb{R}^p$

- the distribution of X_t given $X_{t-1} = x$ is given by a density $p_x(y)$

It is then not practical to organize $p_x(y)$ in to a matrix

Sums become integrals: if X_{t-1} has *density* $\pi_{t-1}(x)$, then

$$\pi_t(y) = \int \pi_{t-1}(x) p_x(y)\, dx$$

First, a bit about continuous setting

While Markov chains with finite state spaces can be useful, and also are very nice to deal with in: especially, when learning something, it is often easier to work in the discrete setting. However, Markov chains with continuous state spaces make for most of the interesting applications... Thus, in continuous setting, typically:

- the state space S is continuous, like, $\mathbb{R}^p$

- the distribution of X_t given $X_{t-1} = x$ is given by a density $p_x(y)$

It is then not practical to organize $p_x(y)$ in to a matrix

Sums become integrals: if X_{t-1} has *density* $\pi_{t-1}(x)$, then

$$\pi_t(y) = \int \pi_{t-1}(x) p_x(y) \, dx$$

and the invariant density is then defined via

$$\pi(y) = \int \pi(x) p_x(y) \, dx$$

It may not be that easy to solve such kind of equations in π, and we may want to learn some smart moves

Some remarks about theory we do not cover here

What determines a (homogeneous) Markov chain?

It is the state space S, the transition probabilities/densities $p_x(y)$ defined for every $x, y \in S$ (organized into the transition matrix P for finite S)

and the initial probability/density giving the distribution of X_0

Invariant probability/density:

it always exists for finite S

when it does exist, *in many practical cases* it is unique

and a limit of *every* (for every given initial π_0) sequence of probabilities/densities π_i

What constitutes "many practical cases" is a topic of mathematical theory of Markov chains which we do not cover here

Instead, we just believe (from now on) that all the cases we deal with belong to those "many practical" ones

And now: reversible Markov chains

A Markov chain/its transition probability/density is called reversible with respect to a function $\ell(x) \geqslant 0$, if

$$\ell(x)p_x(y) = \ell(y)p_y(x) \qquad \text{for all } x, y \in S$$

This is "continuous" formulation; the discrete one is analogous. Note that the identity trivially holds for $x = y$; also, if it is satisfied by $\ell(x)$, it is also satisfied by $c\ell(x)$.

Integrating the identity in x, we get

$$\int \ell(x)p_x(y)\,dx = \int \ell(y)p_y(x)\,dx = \ell(y)\int p_y(x)\,dx = \ell(y)$$

So, if $\ell(x)$ is a density, or can be made into one by multiplying by some constant c, then we have an invariant density

It can be made into one if $\qquad 0 < \displaystyle\int \ell(x)\,dx = \frac{1}{c} < +\infty$

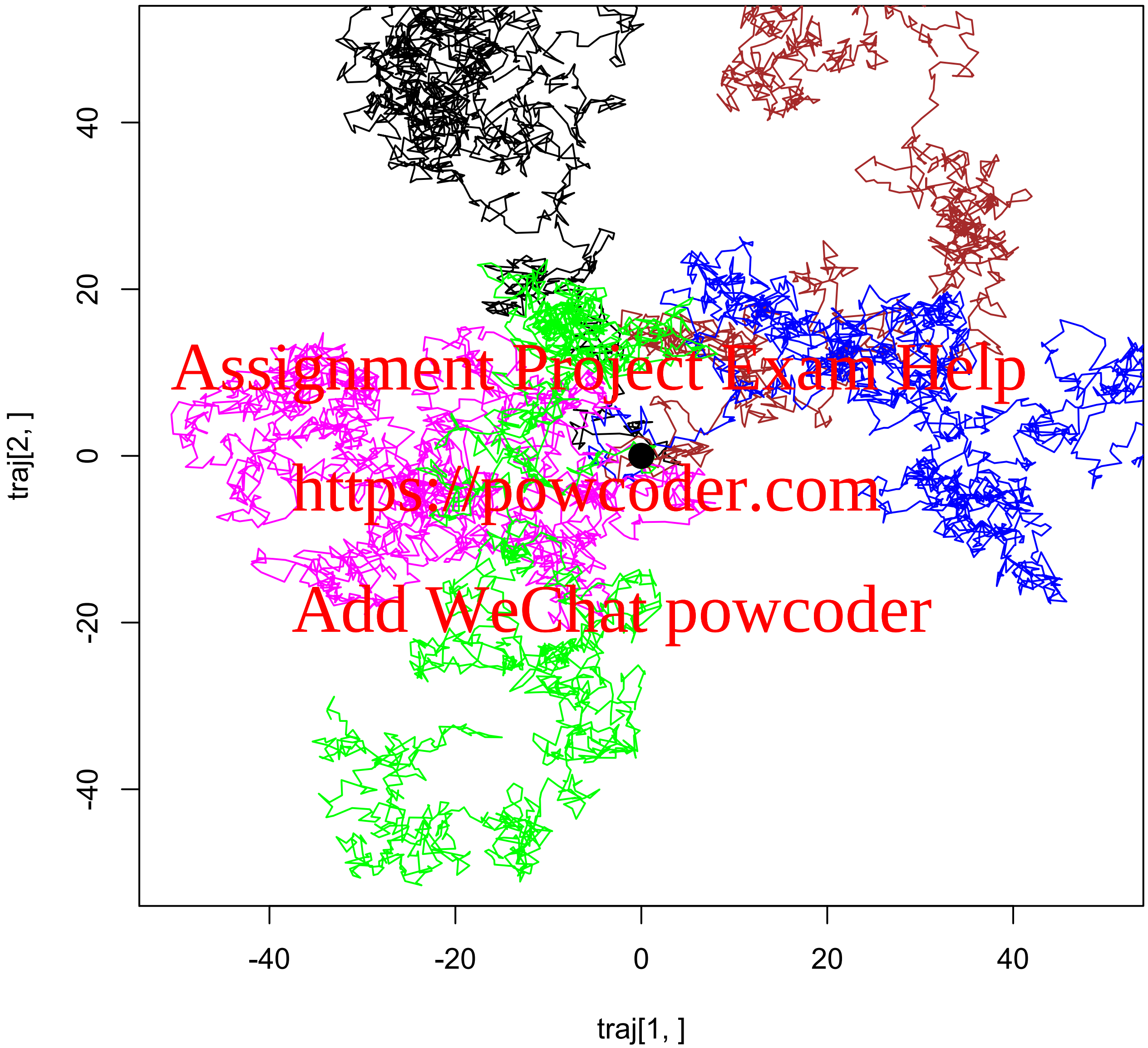
The first inequality excludes $\ell(x)$ identically equal to 0 (with respect to which the Markov chain is trivially reversible). The second inequality is crucial here - but it obviously holds true if we somewhat know that $\ell(x)$ is a constant multiple of a probability density.

Another example: wandering in the plane

$$X_0 = (0, 0) \qquad\qquad p_x(y) = \frac{1}{2\pi} e^{-\frac{1}{2}\|y-x\|^2}$$

```
> inn=matrix(0,2,2000)
> inn[1,2:ncol(inn)]=rnorm(ncol(inn)-1)
> inn[2,2:ncol(inn)]=rnorm(ncol(inn)-1)
> traj=rbind(cumsum(inn[1,]),cumsum(inn[2,]))
> plot(traj[1,],traj[2,],type="l",xlim=c(-50,50),ylim=c(-50,50))
> inn[1,2:ncol(inn)]=rnorm(ncol(inn)-1)
> inn[2,2:ncol(inn)]=rnorm(ncol(inn)-1)
> traj=rbind(cumsum(inn[1,]),cumsum(inn[2,]))
> lines(traj[1,],traj[2,],col="brown")
> inn[1,2:ncol(inn)]=rnorm(ncol(inn)-1)
> inn[2,2:ncol(inn)]=rnorm(ncol(inn)-1)
> traj=rbind(cumsum(inn[1,]),cumsum(inn[2,]))
> lines(traj[1,],traj[2,],col="blue")
...
> points(0,0,pch=16,cex=2)
```

A couple of realizations (trajectories)



One-dimensional case (easier to analyze)

$$X_0 = 0 \qquad p_x(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-x)^2}$$

```
> plot(y=1:1000,x=cumsum(c(0,rnorm(999))),type="l",xlim=c(-60,60))
> lines(y=1:1000,x=cumsum(c(0,rnorm(999))),col="brown")
> lines(y=1:1000,x=cumsum(c(0,rnorm(999))),col="magenta")
> lines(y=1:1000,x=cumsum(c(0,rnorm(999))),col="green")
> lines(y=1:1000,x=cumsum(c(0,rnorm(999))),col="blue")
```

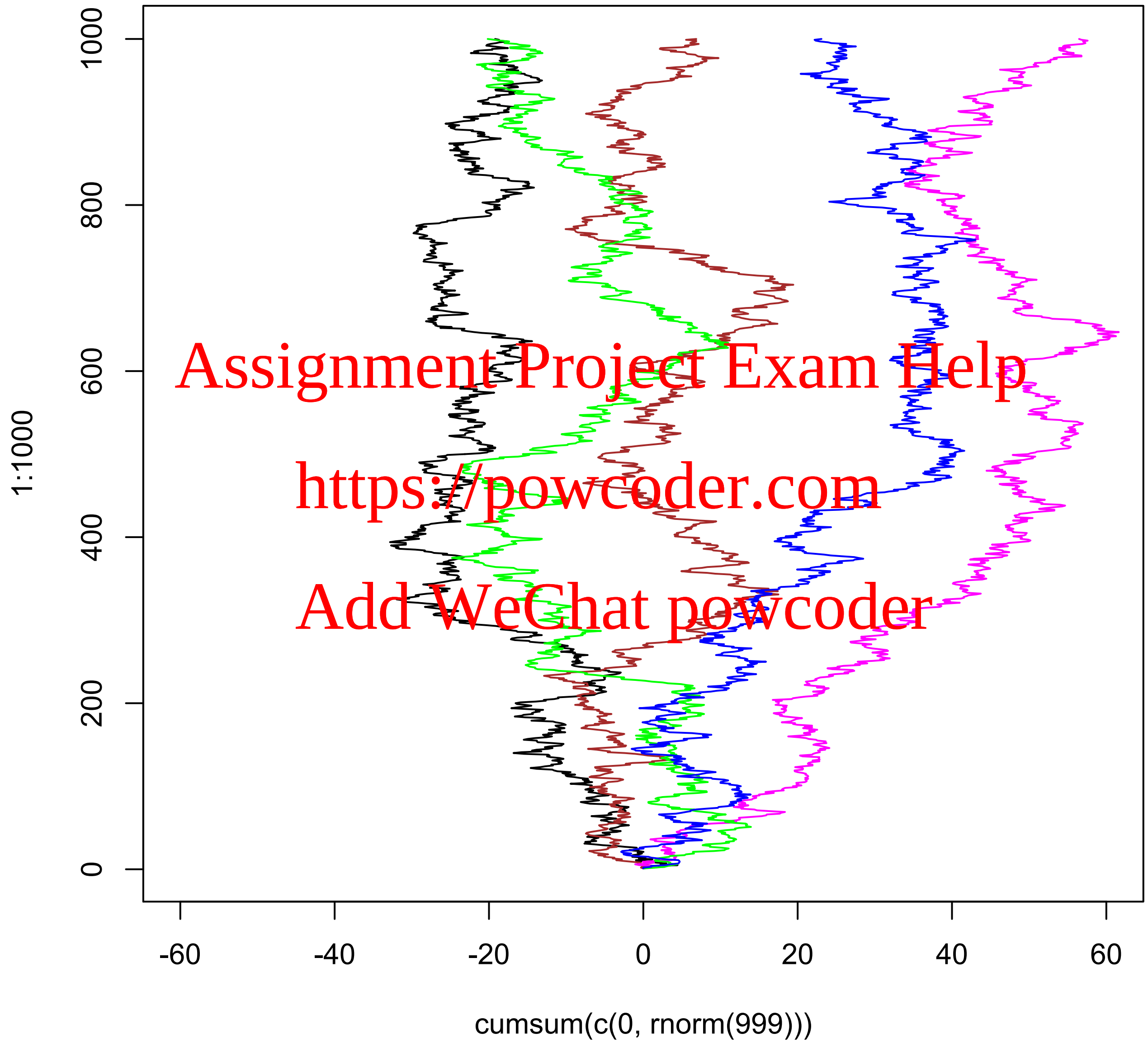

It is reversible with $\ell(x) = 1$: $\qquad p_x(y) = p_y(x)$

A Markov chain/transition probability/density with this property (reversible with respect to a constant) is called *symmetric*

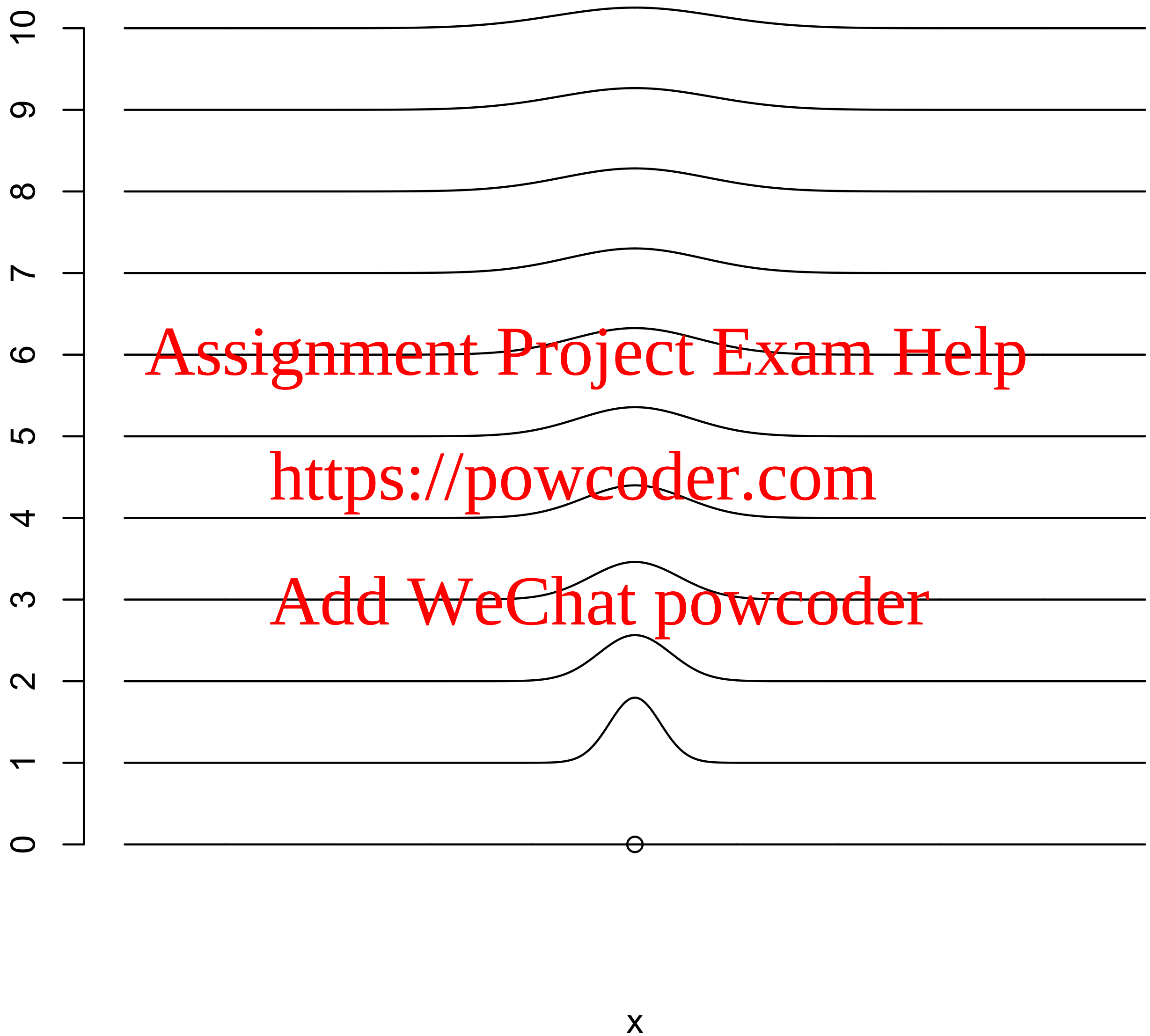
However, $\ell(x) = 1$ is not a constant multiple of any density on $\mathbb{R}$

Note that $P[X_0 = 0] = 1$, and otherwise, the distribution of X_t is normal, with mean 0 and standard deviation t

A couple of realizations (trajectories)



The dissipative evolution of distributions

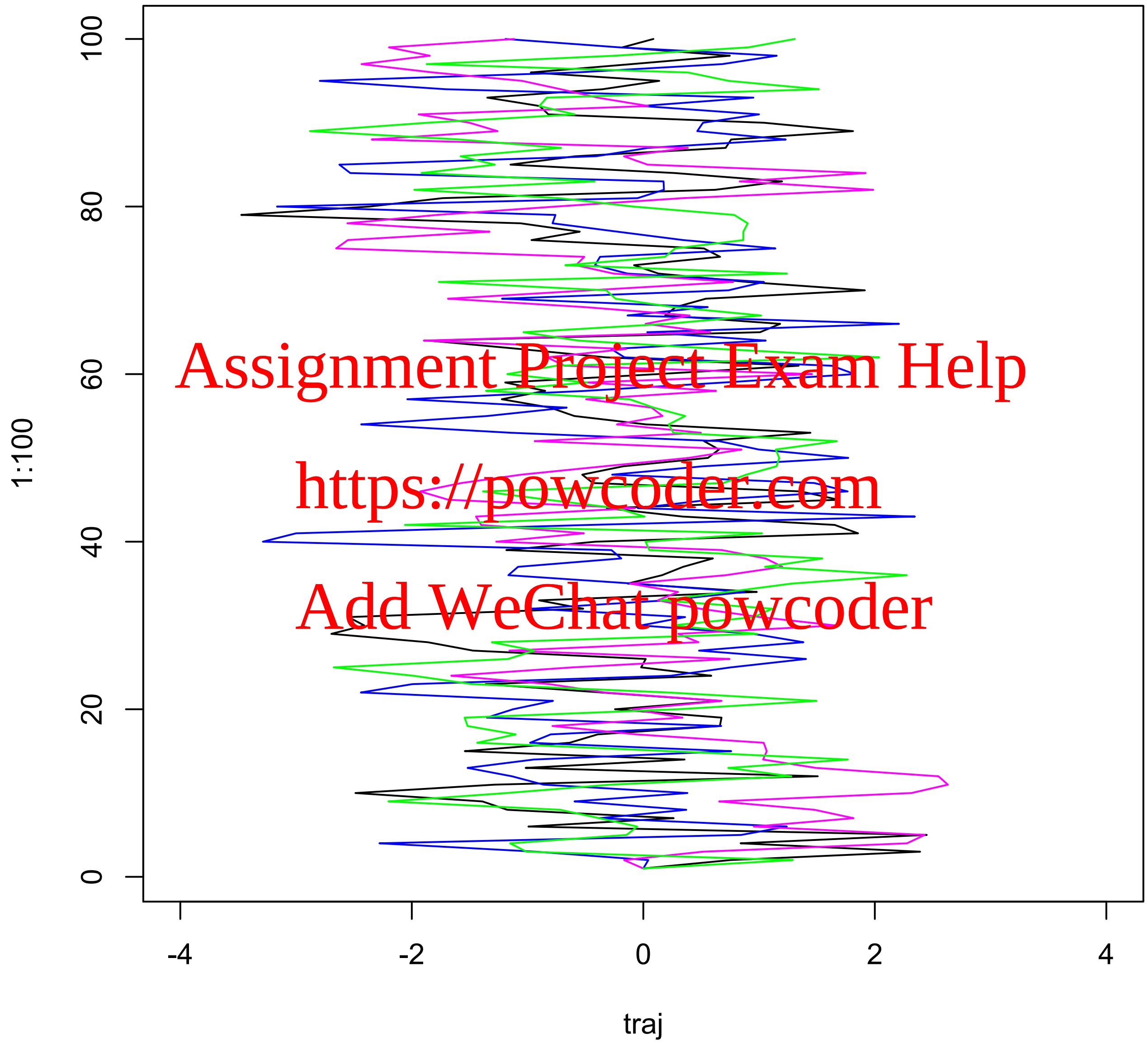


“Return-home” modification

$$X_0 = 0 \qquad p_x(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\frac{x}{2})^2}$$

```
> traj=numeric(100)
> for (k in (2:100)) traj[k] = rnorm(1,traj[k-1]/2)
> plot(y=1:100,x=traj,type="l",xlim=c(-4,4))
> for (k in (2:100)) traj[k] = rnorm(1,traj[k-1]/2)
> lines(y=1:100,x=traj,col="blue")
> for (k in (2:100)) traj[k] = rnorm(1,traj[k-1]/2)
> lines(y=1:100,x=traj,col="magenta")
> for (k in (2:100)) traj[k] = rnorm(1,traj[k-1]/2)
> lines(y=1:100,x=traj,col="green")
```

Realizations (trajectories) now



And it is reversible

This time with $\ell(x) = \mathrm{e}^{-\frac{3}{8}x^2}$

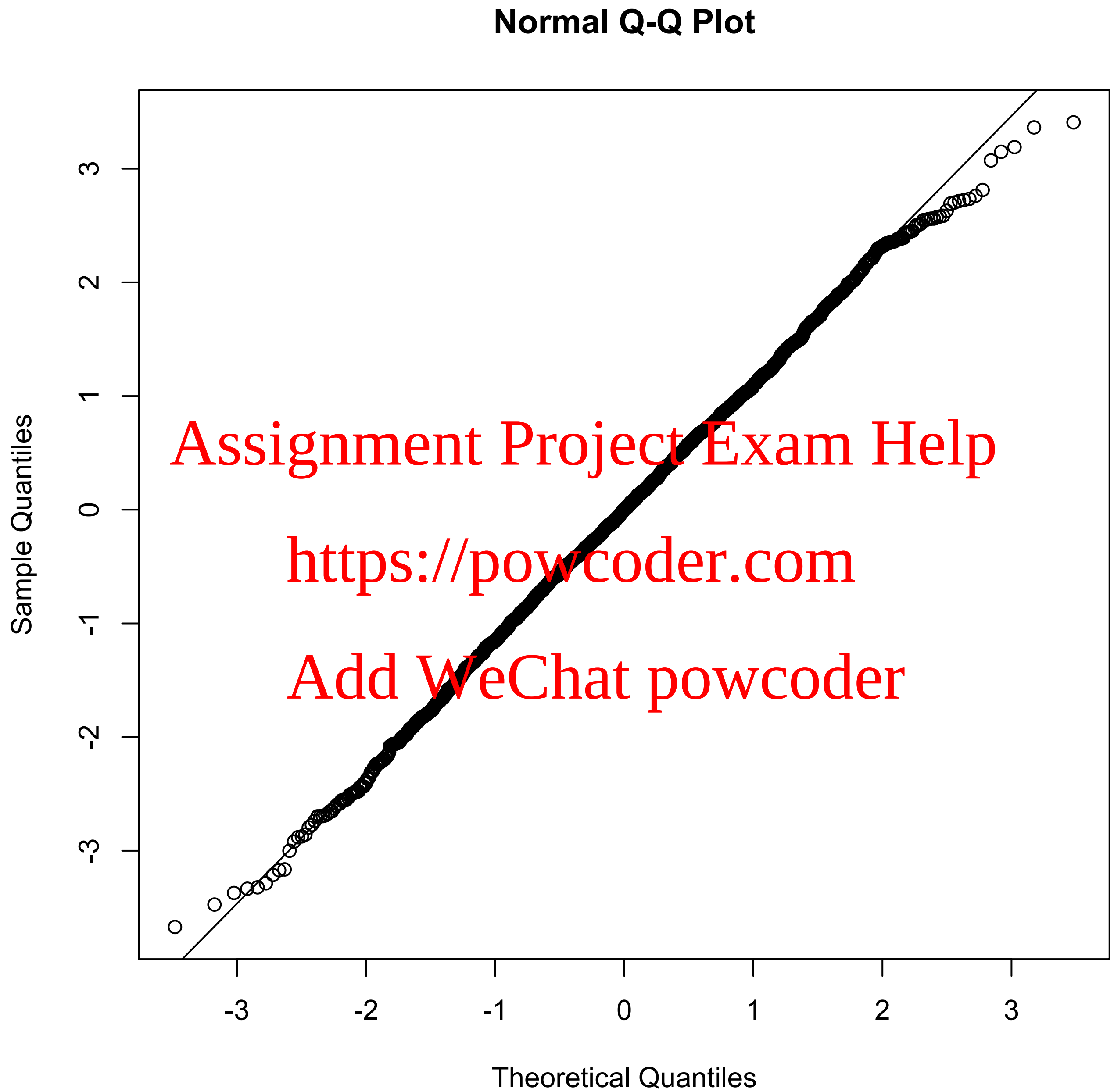
Verify that: $$\mathrm{e}^{-\frac{3}{8}x^2}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(y-\frac{x}{2})^2} = \mathrm{e}^{-\frac{3}{8}y^2}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}(x-\frac{y}{2})^2}$$

And $\ell(x)$ is this time a multiple of the density $$\frac{1}{\sqrt{\frac{4}{3}}\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2(\frac{4}{3})}}$$

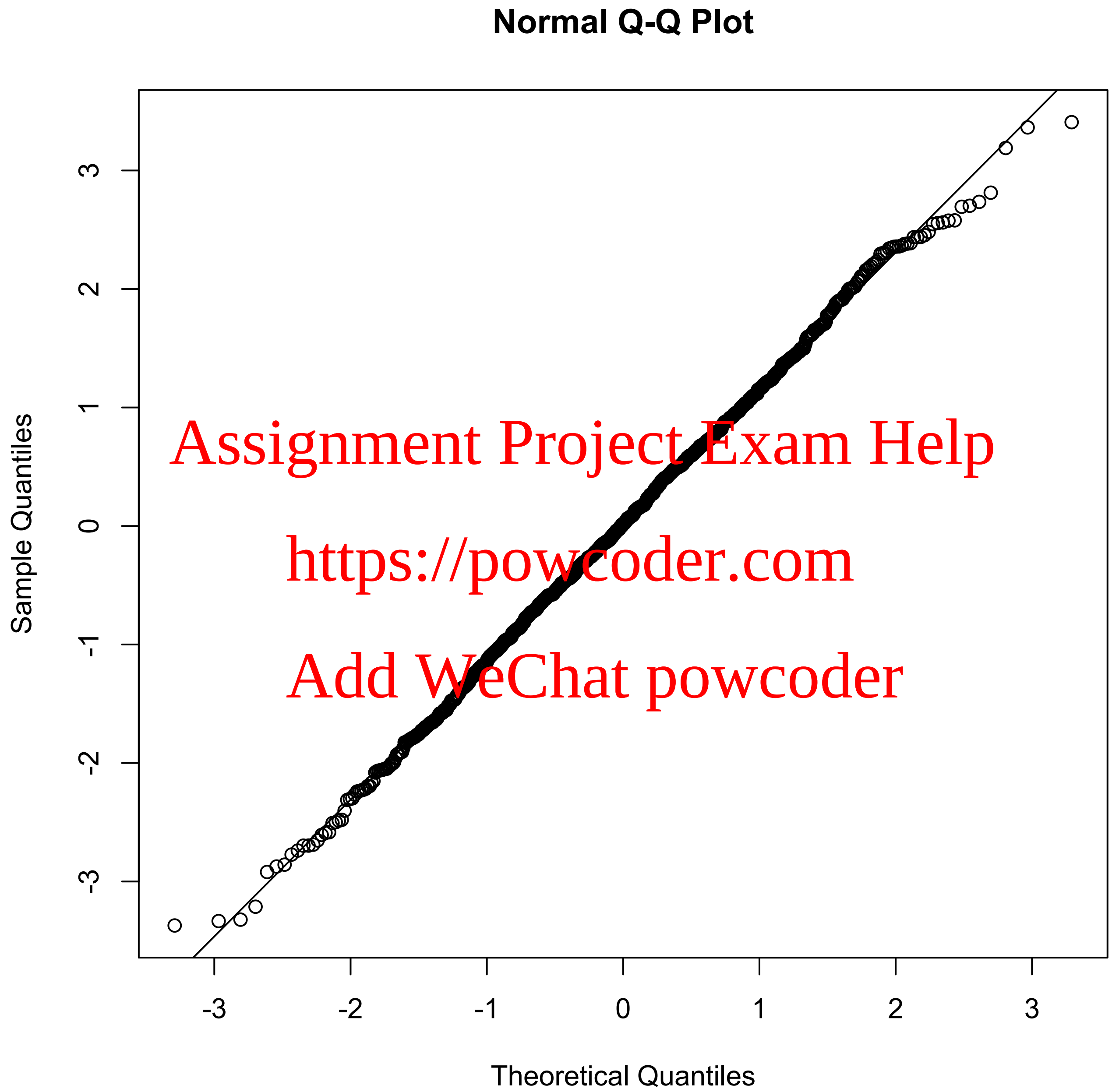
Hence this is an (and we believe that the) invariant density


```
> traj=numeric(2000)
> for (k in (2:2000)) traj[k] = rnorm(1,traj[k-1]/2)
> qqnorm(traj)
> abline(0,sqrt(4/3))
> qqnorm(traj[1000:2000])
> abline(0,sqrt(4/3))
```

Check it out: all of them



Now only those starting with $k = 1000$ - better?



A general recipe: Metropolis (discrete version)

A way to construct a Markov chain with prescribed invariant distribution - known only up to a constant: let $\ell(x) > 0$ be given

Discrete version: Suppose that $\sum_x \ell(x) = \frac{1}{c} < \infty$

Take an arbitrary function $q_x(y)$ of $x, y \in S$ such that

$$q_x(y) = q_y(x) \text{ for all } x, y \in S \quad \text{and} \quad \sum_y q_x(y) \leqslant 1 \text{ for all } x \in S$$

Construct then a new Markov chain such that

$$p_x(y) = \begin{cases} q_x(y) \min\left\{1, \dfrac{\ell(y)}{\ell(x)}\right\} & \text{for } y \neq x \\ 1 - \sum_{y \neq x} p_x(y) & \text{for } y = x \end{cases}$$

How does it work?

- we are in x (for X_{t-1}); we would like to see where to move for X_t
- the "subprobability" function $q_x(y)$ is offering us various y's, each with probability $q_y(x)$

  (subprobability because $\sum_y q_x(y) \leqslant 1$, possibly < 1)
- thus, we generate $y_1, y_2, \ldots$ with probability $q_x(y_1), q_x(y_2), \ldots$

  if none of those gets generated (only the case with sum < 1), then we stay at x

- so, in the previous step we generated y; but we don't move there (make it X_t) yet

- instead, we look at the ratio $r = \ell(y)/\ell(x)$:
- if $r \geqslant 1$, we move to y: $X_t = y$
- if $r < 1$, then we move to y with probability r:

  we generate 1 with probability r (and 0 with probability $1-r$), and move to y ($X_t = y$) if 1 was generated

  otherwise, we again stay at x: $X_t = x$

A general recipe: Metropolis (continuous version)

A way to construct a Markov chain with prescribed invariant distribution - known only up to a constant: let $\ell(x) > 0$ be given

Continuous version: Suppose that $\displaystyle\int \ell(x)\,dx = \frac{1}{c} < \infty$

Take an arbitrary function $q_x(y)$ of $x, y \in S$ such that

$q_x(y) = q_y(x)$ for all $x, y \in S$ and $\displaystyle\int q_x(y)\,dy \leqslant 1$ for all $x \in S$

Construct then a new Markov chain such that for all $x, y \in S$,

- the transition density is $$p_x(y) = \begin{cases} q_x(y) \min\left\{1, \dfrac{\ell(y)}{\ell(x)}\right\} & \text{for } y \neq x \\ 1 - \displaystyle\sum_{y \neq x} p_x(y) & \text{for } y = x \end{cases}$$

How does it work?

Pretty much in the same way as the discrete case, especially when $q_x(y)$ is a density for each x: then, given x in the first step, we generate y out of the distribution with density $q_x(y)$

If $\displaystyle\int q_x(y)\,dy = c < 1$, then   with probability $1 - c$ we stay in x

- and with probability c we generate y as a random number out of density $q_x(y)/c$

The rest is the same: we move to y with probability $r = \ell(y)/\ell(x)$, otherwise we stay in x

Note the subtle difference to the acceptance/rejection method: in the latter, the rejection means we start over, we look for a different random number, we didn't generate it this time. Here, the rejection means we do not move to y, but we stay in x - which is X_t, so we *always* generate a number

Equation of State Calculations by Fast Computing Machines by Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller

Another recipe: Gibbs sampler (discrete version)

Gibbs sampler works always for the state spaces that are multidimensional, that is $\mathbb{R}^p$ with $p \geqslant 2$

For learning sake, we explain the algorithm for $p = 2$, in a hope that the generalization to general p will be then clear

Again, we are to construct a Markov chain on $\mathbb{R}^2$, with a prescribed invariant distribution $\ell(x_1, x_2)$, known only up to a multiplicative constant. The algorithm works as follows: at each step, we start from (x_1, x_2) and generate (y_1, y_2) following the transition probability $p_{(x_1,x_2)}(y_1, y_2)$ such that

$$p_{(x_1,x_2)}(x_1, y_2) = \frac{1}{2}\frac{\ell(x_1, y_2)}{\int \ell(x_1, y)\, dy} \qquad \text{that is, } p_{(x_1,x_2)}(y_1, y_2) \text{ if } x_1 = y_1$$

$$p_{(x_1,x_2)}(y_1, x_2) = \frac{1}{2}\frac{\ell(y_1, x_2)}{\int \ell(y, x_2)\, dy} \qquad \text{that is, } p_{(x_1,x_2)}(y_1, y_2) \text{ if } x_2 = y_2$$

$$p_{(x_1,x_2)}(y_1, y_2) = 0 \qquad \text{if neither } x_1 = y_1 \text{ nor } x_2 = y_2$$

and finally $\quad p_{(x_1,x_2)}(x_1, x_2) = 1 - \sum_{(z_1,z_2) \neq (x_1,x_2)} p_{(x_1,x_2)}(z_1, z_2)$

How does it work?

Gibbs sampler - in the continuous version - is typically applied in cases when the ratios appearing in the definition of transition probabilities correspond to conditional densities that are convenient to generate random numbers from. Suppose that $c\ell(x_1, x_2)$ is a density of a random vector (X_1, X_2). The conditional density of X_2 given $X_1 = x_1$ is then

$$f(y|x_1) = \frac{c\ell(x_1, y)}{\int c\ell(x_1, y)\, dy} = \frac{\ell(x_1, y)}{\int \ell(x_1, y)\, dy}$$

- and this is the quantity appearing in transition probabilities of the Gibbs sampler. While integrating ℓ in all variables may pose a problem, integrating it in just single variables may be feasible

Once we have $f(y|x_1)$, and also $g(y|x_2)$, the conditional density of X_2 given $X_1 = x_1$, and know how to generate random numbers following these densities, the Gibbs sampler works as follows (without any acceptance/rejection taking place). We give the recipe for the general, p-dimensional case

The general recipe

Now, it is not only x_1 and x_2, but $x_1, x_2, \ldots, x_p$. The case with $p = 2$ is a special case. The transition probabilities change accordingly: $1/2$ is replaces by $1/p$. We need to know how to generate random numbers from p conditional densities, conditional densities of i-th coordinate random variables given all others are fixed.

First, we select randomly the coordinate from $1, 2, \ldots, p$, each with the same probability p.

Second: keeping all other coordinates fixed, we generate new i-th coordinate as a random number following the conditional density described above.

We start from some convenient initial value and repeat first and second step until the generated sequence is long enough. Note that there is no acceptance/rejection in this algorithm.

Concluding remarks

The book of Rizzo brings detailed recipes for several instances of Markov chain Monte Carlo algorithms, and also code for selected examples of those. Metropolis algorithm is covered as a special case of slightly more general Metropolis-Hastings algorithm. Another special case covered by the textbook is what is called there Independence Sampler.

There are several versions of Gibbs sampler in the literature, all of them giving equivalent results. The more widespread is a version that does not sample i randomly, but instead "cycles" over i, taking consecutively coordinates $1, 2, \dots, n$ and then repeating the process. The book of Rizzo gives another slightly different version of Gibbs sampler: the jumps in all coordinates are accomplished "simultaneously" (this corresponds to taking every p-th number generated by the previous version)