http://www.math.nyu.edu/faculty/goodman/teaching/MonteCarlo15/

Class notes: Monte Carlo methods
Week 4, An example, Bayesian statistics
Jonathan Goodman
April 14, 2015

1 An example

It is possible to understand one non-trivial example in detail, the heat bath/Gibbs sampler applied to a multi-variate normal. We will find expressions for all the eigenvalues of the transition operator in terms of the eigenvalues of a related *Gauss Seidel* matrix. This analysis suggests ways find faster (i.e. shorter auto-correlation times) MCMC algorithms for Gaussians. There is much ongoing research looking for ways to apply Gaussian insights to non-Gaussian problems.

Suppose $f(x)$ is a probability density in $\mathbb{R}^n$, and that $P(x,y) = P(x \to y)$ is the transition probability density of a Markov chain. We have seen in an assignment that P defines a self adjoint operator on the space of functions u with $\mathrm{E}_f\left[u(X)^2\right] < \infty$. An eigenfunction is a function that satisfies

$$\lambda u(x) = \int P(x,y)u(y)\,dy = \mathrm{E}_P[\,u(X_{n+1}) \mid X_n = x] \ . \tag{1}$$

We write this as $Pu = \lambda u$, where P is the *integral operator* defined by the right hand side above. An exercise shows that if $Pu = \lambda u$ and $Pv = \mu v$ with $\lambda \neq \mu$, then $\langle u, v\rangle = \mathrm{E}_f[\,u(X)v(X)] = 0$. A somewhat technical argument shows that if P has no delta function component, then there is a complete set of eigenfunctions. A delta function on the diagonal of P corresponds to a non-zero probability that $X_{n+1} = X_n$. This happens in rejection based methods (Metropolis), but not for heat bath methods.

The analysis of heat bath on Gaussians has two parts. The first part shows that the heat bath method produces MCMC iterates that satisfy

$$X_{n+1} = AX_n + BZ_n \ , \tag{2}$$

where $Z_n \sim \mathcal{N}(0, I)$ are independent multivariate standard normals. The important thing is that the MCMC process turns out to be linear. The second part is an analysis of linear processes. The eigenvalues of the corresponding P may be expressed in terms of the eigenvalues of A. It turns out (as pointed out to me by Persi Diaconis) that this analysis applies even if the Z_n are not Gaussian.

We start by identifying the integral kernel $P(x,y)$ for Gaussian iteration. This amounts to identifying the conditional PDF of X_{n+1} conditional on $X_n = x$. Clearly X_{n+1} is Gaussian with mean x. The covariance matrix is $C = BB^t$ (famous fact about multivariate normals).

$$P(x,y) = \frac{1}{Z} e^{-\frac{1}{2}(y-Ax)^t C^{-1}(y-Ax)} \ . \tag{3}$$

In other applications the normalization constant would depend on x, but for the Gaussian, the normalization constant depends on $\det(C)$, which is independent of x.

1.1 Heat bath on general Gaussians, Gauss Seidel

The multivariate Gaussian density will be

$$f(x) = \frac{1}{Z} e^{-\frac{1}{2} x^t H x} = \frac{1}{Z} \exp\left\{ \frac{-1}{2} \left[\sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j H_{ij} \right] \right\} . \tag{4}$$

The heat bath/Gibbs sampler algorithm cycles through the components, replacing the value by the conditional mean given all the other components. It can be confusing that one MCMC step $X_n \to X_{n+1}$ consists d individual component steps. We try to make this clearer with the following notation. One MCMC sweep will be $X \to Y$. In this and the next paragraph X_i with be component i of $X = (X_1, \ldots, X_d)$. The first step of the sweep replaces X_1 with Y_1, the next replaces X_2 with Y_2, and so on. The d steps of the sweep are

$$\begin{aligned} (X_1, X_2, \ldots, X_d) &\overset{\text{step } 1}{\to} (Y_1, X_2, \ldots, X_d) \\ &\overset{\text{step } 2}{\to} (X_1, Y_2, X_3 \ldots, X_d) \overset{\text{step } 3}{\to} \cdots \\ &\overset{\text{step } d-1}{\to} (Y_1, Y_2, \ldots, Y_{d-1}, X_d) \overset{\text{step d}}{\to} (Y_1, Y_2, \ldots, Y_{d-1}, X_d) \end{aligned}$$

Step i uses replaces X_i using the new values $Y_1, \ldots, Y_{i-1}$, and the old values $X_{i+1}, \ldots, X_d$.

The individual step i uses a sample of one variable conditional density

$$f_i(x_i \mid Y_1, \ldots, Y_{i-1}, X_{i+1}, \ldots, X_d) .$$

Of course, the formula for f_i is the same as the formula for f, except for a normalization constant that is independent of x_i. Some Gaussian tricks help us identify f_i in the Gaussian case (4). The exponent in (4) is a quadratic function of x, so the exponent in f_i is a quadratic function of x_i. A quadratic is determined by three quantities: its minimum value, its symmetry point (the minimizer), and the coefficient of x_i^2. For our purposes, the minimum value is irrelevant, because it can be absorbed by the normalization constant. The coefficient of x_i^2 is $\frac{1}{2} H_{ii}$. Therefore $\frac{1}{H_{ii}}$ is the conditional variance of X_i. For Gaussians, the minimizing value of the exponent is the maximizing value of the PDF, which is the conditional mean. The derivative of the exponent of (4) with respect to x_i is

$$H_{ii} x_i + \sum_{j<i} H_{ij} Y_j + \sum_{j>i} H_{ij} X_j .$$

The conditional mean of X_i comes from setting this to zero:

$$\overline{X}_i = \frac{-1}{H_{ii}} \left(\sum_{j<i} H_{ij} Y_j + \sum_{j>i} H_{ij} X_j \right) .$$

The new X_i should be Gaussian with this mean, and variance $\frac{1}{H_{ii}}$. These criteria are satisfied by the formula

$$Y_i = \frac{-1}{H_{ii}} \left(\sum_{j<i} H_{ij} Y_j + \sum_{j>i} H_{ij} X_j \right) + \frac{1}{\sqrt{H_{ii}}} Z_i \; . \tag{5}$$

As usual, the Z_i are independent standard normals. To do an MCMC sweep $X \to Y$, we do

```
for (i = 1, ..., d) { apply (5);}
```

The formulas (5) make Y_i a linear function of the earlier Y_j, the later X_j, and Z_i. Of course, the earlier Y_j themselves are linear functions of X variables and the earlier Z_j. Therefore, Y_i is a linear function of the X_j (all of them), and the Z_j with $j \leq i$. This shows that the MCMC step has the form (2) as claimed.

The heat bath equations (5) are related to the *Gauss Seidel* iterative method for solving systems of equations of the form

$$Hx = b \; . \tag{6}$$

The Gauss Seidel algorithm produces a sequence of iterates $x_n \to x$ as $n \to \infty$. The $x_n \to x_{n+1}$ iteration consists of a sweep through the components of x_n. At step i of the sweep, you solve equation i for variable i. The formula is

$$x_{n+1,i} = \frac{1}{H_{ii}} \left(b_i - \sum_{j<i} H_{ij} x_{n+1,j} - \sum_{j>i} H_{ij} x_{n,j} \right) \; . \tag{7}$$

On the right we have the new values $x_{n+1,j}$ for $j < i$. These correspond to components that are updated before component i. The components $j > i$ are updated after component i. If we set $b = 0$ in the Gauss Seidel iteration (7) and $Z = 0$ in the heat bath iteration (5), the iterations are the same.

There is a simple argument to show that the Gauss Seidel iteration converges to the correct solution, provided H is symmetric and positive definite. This argument is related to the the derivation of the heat bath iteration (5). It is a *variational* argument, one that argues from maximizing or minimizing something. The *variational principle* for the linear equation system (6) is that it is equivalent to minimizing the "energy" function

$$\phi(x) = \frac{1}{2} x^t H x - x^t b \; . \tag{8}$$

Indeed, minimizing ϕ by setting the gradient to zero leads to (using $H^t = H$ here) $\nabla \phi = Hx - b = 0$. The Gauss Seidel formula (7) is what you get if you minimize $\phi(x)$ over the single component x_i. In fact, $\phi(x)$ is a quadratic function of x_i. The calculation that leads from minimizing over x_i to (7) is the same as the one that led to (5). Since H is positive definite, $\phi(x)$ has a unique

global minimum. The Gauss Seidel iteration $x_n \to x_{n+1}$ consists of d steps (7), each of which reduces ϕ. Therefore $\phi(x_{n+1}) \leq \phi(x_n)$.

A closer examination shows two things. One is that ϕ is strictly decreasing: $\phi(x_{n+1}) < \phi(x_n)$ unless $Hx_n = b$. The second is an eigenvalue gap that implies that $x_n \to x$ exponentially (albeit with a possibly small exponential factor). This discussion is simpler if $b = 0$. The trick for reducing to the case $b = 0$ is to look at the error $y_n = x_n - x$. Of course, the program uses only x_n because x is unknown. But the analysis can use y_n. The rate at which $x_n \to x$ is the same as the rate at which $y_n \to 0$. We can write $Hx = b$ in the form

$$x_i = \frac{1}{H_{ii}} \left(b_i - \sum_{j<i} H_{ij} x_j - \sum_{j>i} H_{ij} x_j \right) .$$

We subtract this from (7), and get

$$y_{n+1,i} = \frac{-1}{H_{ii}} \left(\sum_{j<i} H_{ij} y_{n+1,j} + \sum_{j>i} H_{ij} y_{n,j} \right) .$$

This is a linear iteration, which may be written in the abstract form $y_{n+1} = Ay_n$, where A is the matrix in (2). It is the same as (5) if you set $Z = 0$. So now we know that A is a $d \times d$ matrix so that $A^n y \to 0$ as $n \to \infty$ for any y. This implies that if μ is an eigenvalue of A, then $|\mu| < 1$. Since there are finitely many eigenvalues, there must be a *spectral gap*

$$g = 1 - \max |\mu| > 0 .$$

This implies that the iterates $y_n = A^n y_0$ converge to zero exponentially as $n \to \infty$.

The strict decrease of ϕ is another route to exponential decay of y_n. It is easy to see that unless $Hx_n = b$, then $\phi(x_{n+1}) < \phi(x_n)$, which is strict decrease. Look at the first i where $x_{n+1,i} \neq x_{n,i}$. If there is no such i, then $x_{n+1} = x_n$, which means that every component of $Hx_n = b$ is satisfied. But x_i changes in the Gauss Seidel algorithm only if changing x_i reduces ϕ. If only x_i changes, then $\phi(x_{n+1}) < \phi(x_n)$. If other components of x change, then even more so $\phi(x_{n+1}) < \phi(x_n)$. We look to y_n to make this argument more quantitative. These satisfy the Gauss Seidel iteration formulas with $b = 0$. Therefore, consider the energy with $b = 0$, which we write as $\phi(y) = \frac{1}{2} y^t H y$. The argument just given shows that unless $y_n = 0$, we have $\psi(y_{n+1}) < \psi(y_n)$. Since H is positive definite, the set of y with $\psi(y) = 1$ is compact. Therefore

$$1 - \gamma = \max_{\psi(y_n)=1} \psi(y_{n+1})$$

is well defined, with $\gamma > 0$. This implies that $\psi(y_{n+1}) \leq (1 - \gamma)\psi(y_n) \leq (1 - \gamma)^n \psi(y_0)$. This is another way to see the exponential convergence of Gauss Seidel. The argument has the same weakness as the previous eigenvalue argument. It uses compactness to infer the existence of a constant without giving any information about how small γ might be. In many applications, Gauss Seidel iteration converges so slowly that it is impractical.

1.2 Gaussian lattice free fields

One definition of a *field* in physics is something whose value depends on coordinates. Electric and magnetic fields are examples. The strength of the electric field changes from place to place and time to time. This is more or less what mathematicians call a *function*. A *random field* is a function whose values are random. Brownian motion is an example.

A generic field depending on continuous coordinates cannot be represented in the computer using finitely many numbers. Instead, a continuum field may be approximated by a *lattice* field, which is defined when the coordinates are points in a discrete lattice. A mathematician normally would write a function $u(x)$ (or something like that). If we did that, x would be the coordinates and $u(x)$ would be the random value at location x. We do not use this notation here, to keep this section notationally consistent with the rest of the section. Here, the coordinates will be called $i = (i_1, \ldots, i_k)$ for a k dimensional field. We write $x_i = x_{i_1,\ldots,i_k}$ for the value of the field at location i. We suppose that the coordinates are integer valued with range $1 \leq i_j \leq L$. There are L^k distinct *lattice sites* i. The lattice field x is determined by the L^k values x_i, so we write $x \in \mathbb{R}^k$. A one dimensional lattice field has values $x_1, \ldots x_L$, and may be thought of a string of beads. A two dimensional lattice field has values $x_{11}, \ldots, x_{1L}, x_{21}, \ldots, x_{LL}$. The graph of a two dimensional lattice field defines a random two dimensional surface in 3D.

A Gaussian field is a random field with a Gaussian density $f(x) = \frac{1}{Z} e^{-\beta\phi(x)}$, where ϕ is a quadratic function of x. A common energy function is a discrete version of the Dirichlet integral

$$\int |\nabla u|^2$$

Instead of the gradient, we use the lattice gradient. In 2D this is

$$(x_{i_1+1,i_2} - x_{i_1,i_2})^2 + (x_{i_1,i_2+1} - x_{i_1,i_2})^2$$

The first term is thought of as the energy in the horizontal bond connecting neighboring lattice sites (i_1, i_2) and $(i_1 + 1, i_2)$. The second term is the bond energy for the vertical bond between (i_1, i_2) and $(i_1, i_2 + 1)$ The lattice energy is the sum of these bond energy over all the bonds in the lattice

$$\phi(x) = \frac{1}{2} \sum_{i_1,i_2} \left[(x_{i_1+1,i_2} - x_{i_1,i_2})^2 + (x_{i_1,i_2+1} - x_{i_1,i_2})^2 \right] . \tag{9}$$

This is sometimes called the Gaussian *free* field, as opposed to an *interacting* field that would have cubic, quartic or other terms in its energy function.

The field energy expression (9) is incomplete in that it does not say what to do at the boundary of the lattice. One possibility is *Dirichlet* boundary conditions, where the field values on sites that neighbor lattice sites are set to zero:

$$x_{0,i_2} = 0 \ , \quad x_{i_1,0} = 0 \ , \quad x_{L+1,i_2} = 0 \ , \quad x_{i_1,L+1} = 0 \ .$$

These values matter if the sum over bonds (9) includes bonds such that connect interior point to boundary points, such as the bond between $(0, i_2)$ and $(1, i_2)$. *Periodic* boundary conditions are also possible. These suppose that x_i is a periodic function of i, so that for example $x_{i_1+L,i_2} = x_{i_1,i_2}$. In practice, we implement periodic boundary conditions by copying "active" values such as $x_{0,i_2} = x_{L,i_2}$. We use this value when computing the bond energy contribution $(x_{1,i_2} - x_{0,i_2})^2$. The point of periodic boundary conditions is that all lattice sites are the same. There is in effect no "boundary".

The heat bath/Gauss Seidel algorithm is local for this *nearest neighbors* quadratic energy function. Suppose X is a current sample field and we want to resample X_{i_1,i_1}. For this, we need to know how $\phi(X)$ depends on X_{i_1,i_1}. There are four terms in the sum (9) that depend on X_{i_1,i_1}, so we can write

$$\begin{aligned}\phi(X) = \frac{1}{2}\Big[&(X_{i_1+1,i_2} - X_{i_1,i_2})^2 + (X_{i_1,i_2+1} - X_{i_1,i_2})^2 \\ &+ (X_{i_1-1,i_2} - X_{i_1,i_2})^2 + (X_{i_1,i_2-1} - X_{i_1,i_2})^2\Big] \\ &+ \quad \text{terms independent of } X_{i_1,i_2} \ .\end{aligned}$$

The conditional mean of X_{i_1,i_2}, given the values of the neighbors, is found by minimizing over X_{i_1,i_2}. Setting the derivative with respect to X_{i_1,i_2} gives

$$\overline{X}_{i_1,i_2} = \frac{1}{4}\left(X_{i_1+1,i_2} + X_{i_1,i_2+1} + X_{i_1-1,i_2} + X_{i_1,i_2-1}\right) \ .$$

To find the variance, we look for the coefficient of $X^2_{i_1,i_2}$, which we can see by letting this go to infinite keeping the other values fixed. The coefficient is $\frac{1}{2}4 = 2$. Therefore, we may write the conditional density as

$$f(x_{i_1,i_2}|\text{rest}) = \frac{1}{Z} e^{-\frac{1}{2}4\beta\left(x_{i_1,i_2} - \overline{X}_{i_1,i_2}\right)^2} \ .$$

One heat bath sweep is implemented, roughly, as

```
sig = sqrt(4.*beta);     // sigma = standard deviation of the noise
for i = 1, ..., L {
   for j = 1, ..., L {
      Xb = .25*( X[i+1,j] + X[i-1,j] + X[i,j+1] + X[i,j-1]; // for "X bar"
      X[i,j] = Xb + sig*N_samp();      // N_samp gives standard normals
     }
  }
```

Note that there is only one copy of the array. All changes are made "in place". When `X[i,j]` is updated, the values `X[i-1,j]` and `X[i,j-1]` are new for this sweep, while `X[i+1,j]` and `X[i,j+1]` are old from the previous sweep.

1.3 Adjoints

If you have a matrix, A, that is not symmetric, the eigenvalue/eigenvector problem for A^t may be easier or harder (or, anyway, different) from the problem

for A. Of course, the eigenvalues must be the same. But if we evaluate the eigenvalues by identifying the eigenvectors, then A or A^t may be easier. As an example of this phenomenon, suppose P is the $n \times n$ transition matrix for a discrete Markov chain: $P_{jk} = \mathrm{P}(j \to k)$. It is easy see that $\lambda = 1$ is an eigenvalue of P, but harder with P^t. For P, note that $\sum_k P_{jk} = 1$ for every j (the walker must go somewhere in one step). Therefore, if $u_k = 1$ for all k, we have

$$(Pu)_j = \sum_k P_{jk} u_k = \sum_k P_{jk} = 1 \ , \quad \text{for all } j.$$

This verifies that $Pu = u$, and that u is an eigenvector for eigenvalue $\lambda = 1$. On the other hand, finding an eigenvector for P^t is the same as finding a row vector, f, so that $fP = f$, which is

$$\sum_j f_j P_{jk} = f_k \ .$$

Such an eigenvector is an invariant probability distribution for the Markov chain. Its existence is harder to prove directly.

For Markov chains with a continuous state space, the transition distribution is given by a transition probability density, $P(x, y)$, so that for "any" set B,

$$\mathrm{P}(X_{n+1} \in B \mid X_n = x) = \int_{y \in B} P(x, y) \, dy \ .$$

This P defines a linear *operator* by saying that $Pu = v$ means that

$$v(x) = \int P(x, y) u(y) \, dy \ .$$

A function, u, is an eigenfunction if it satisfies the eigenvalue equation (1). The adjoint operator of P, with respect to the L^2 inner product, satisfies

$$\langle f, Pu \rangle = \langle P^* f, u \rangle$$

$$\int f(x) \left(\int P(x, y) u(y) \, dy \right) dx = \int \left(\int f(x) P^*(x, y) \, dx \right) u(y) \, dy \ .$$

On the right side of the second line, f is before P^* to make the formulas look like the matrix formulas, where f would be a row vector. It is clear from this that if $g = P^* f$, then

$$g(y) = \int f(x) P(x, y) \, dy \ . \tag{10}$$

In the matrix case, the adjoint action of P on column vectors was an action on row vectors. The numbers P_{jk} are the same, but the sum is over j instead of k. In the L^2 case here, the values $P(x, y)$ are the same, but we integrate over x instead of y.

We look for eigenvalues of P either by looking directly at (1), or by looking for adjoint eigenfunctions for (10)

$$\lambda g(y) = \int g(x) P(x, y) \, dy \ . \tag{11}$$

1.4 Eigenvalues for linear Markov chains

For the linear Gaussian transition (2), the transition "matrix", the transition probability density, is given by (3). The eigenvalues of P turn out to be products of eigenvalues of A. Suppose $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a list of non-negative integers. In this context, we call α a *multi index*. Let $\mu_1, \ldots, \mu_d$ be the eigenvalues of A. For any multi-index, α, there is an eigenvalue of P

$$\lambda_\alpha = \mu^\alpha = \mu_1^{\alpha_1} \cdots \mu_d^{\alpha_d} \ . \tag{12}$$

The expression μ^α is *multi-index notation*.

We prove (12) be writing the eigenfunctions of P^* and verifying that they are correct. These eigenfunctions are built from *Hermite polynomials*, so we start by explaining them. The verification that they are eigenfunctions is simple, once we know what they are. We start with the one variable case. The multi-variate case is a natural generalization.

Standard uni-variate Hermite polynomials may be defined by the *Rodrigues* formula

$$\left(\frac{d}{dx}\right)^n e^{-\frac{1}{2}x^2} = H_n(x) e^{-\frac{1}{2}x^2} \ . \tag{13}$$

The first few are found directly:

$$\begin{aligned} e^{-\frac{1}{2}x^2} &\xrightarrow{\partial_x} -x\, e^{-\frac{1}{2}x^2} \\ &\xrightarrow{\partial_x} \left(x^2 - 1\right) e^{-\frac{1}{2}x^2} \\ &\xrightarrow{\partial_x} \left(-x^3 + 3x\right) e^{-\frac{1}{2}x^2} \\ &\xrightarrow{\partial_x} \left(x^4 - 6x^2 + 3\right) e^{-\frac{1}{2}x^2} \\ &\xrightarrow{\partial_x} \left(-x^5 + 10x^3 - 15x\right) e^{-\frac{1}{2}x^2} \ . \end{aligned}$$

This gives

$$\begin{aligned} H_0(x) &= 1 \\ H_1(x) &= -x \\ H_2(x) &= x^2 - 1 \\ H_3(x) &= -x^3 + 3x \\ H_4(x) &= x^4 - 6x^2 + 3 \\ H_5(x) &= -x^5 + 10x^3 - 15x \end{aligned}$$

An induction argument shows that $H_n(x)$ is a polynomial of degree n with leading coefficient ± 1. Since we are looking for eigenfunctions, it is common to ignore the signs, and write, for example $H_1(x) = x$ and $H_3(x) = x^3 - 3x$.

There are many things to say about Hermite polynomials. One is the orthogonality relation

$$\int_{-\infty}^{\infty} H_n(x) H_m(x) e^{-\frac{1}{2}x^2}\, dx = 0 \ , \quad \text{if } n \neq m \ . \tag{14}$$

This orthogonality, plus the requirement that

$$H_n(x) = \pm x^n + \text{polynomial degree } n-1 \ ,$$

determines the H_n completely. Another is the *interlacing* property of the zeros. $H_n(x)$ has n real roots, $x_{n,1} < \cdots < x_{n,n}$. The interlacing property is $x_{n+1,i} < x_{n,i} < x_{n+1,i+1}$, which holds for all $n > 0$ and all $i = 1, \ldots, n$. It indicates that $H_n(x)$ becomes more oscillatory as a function of x (goes back and forth across zero more) as n increases. A third is the recurrence relation

$$H_{n+1}(x) + xH_n(x) + nH_{n-1}(x) = 0 \ . \tag{15}$$

The Rodrigues formula (13) implies the recurrence relation (15). This follows from the following calculation

$$\begin{aligned}
\left(\frac{d}{dx}\right)^{n+1} e^{-\frac{1}{2}x^2} &= \left(\frac{d}{dx}\right)^n \left[\frac{d}{dx} e^{-\frac{1}{2}x^2}\right] \\
&= -\left(\frac{d}{dx}\right)^n \left[x e^{-\frac{1}{2}x^2}\right] \\
&= -x\left(\frac{d}{dx}\right)^n e^{-\frac{1}{2}x^2} - n\left(\frac{d}{dx}\right)^{-n-1} e^{-\frac{1}{2}x^2} \\
H_{n+1}(x)e^{-\frac{1}{2}x^2} &= -xH_n(x)e^{-\frac{1}{2}x^2} - nH_{n-1}(x)e^{-\frac{1}{2}x^2} \ .
\end{aligned}$$

The recurrence relation allows a proof by induction of the interlacing property. Suppose we know that the zeros of H_{n-1} interlace the zeros of H_n. This implies that the numbers $H_{x-1}(x_{n,j})$ alternate in sign as a function of j. Between $x_{n,j}$ and $x_{n,j+1}$, there is a single root of $H_{n-1}(x)$. The recurrence relation, applied at the zeros of $H_n(x)$ gives

$$H_{n+1}(x_{n,j}) = -nH_{n-1}(x_{n,j}) \ .$$

This implies that $H_{n+1}(x)$ has at least one real root between $x_{n,j}$ and $x_{n,j+1}$. Since H_{n+1} has at most $n+1$ real roots, and there are $n-1$ such intervals, we learn about $n-1$ real roots in this way. It may be that $H_{n+1}(x)$ has only $n-1$ real roots, or that some interval has three instead of one roots of H_{n+1}. We rule out this possibility by looking at the behavior of $H_n(x)$ and $H_{n+1}(x)$ for $x < x_{n,1}$ and for $x > x_{n,n}$. It is possible to show (using signs of things) there is at least one real root of H_{n+1} in each of these intervals. It is possible to prove the orthogonality property (14) from the Rodrigues formula by induction. But there is a better proof of orthogonality based on the fact that the H_n are eigenfunctions of some operator.

Back to the eigenvalue problem, consider first the scalar recurrence

$$X_{n+1} = aX_n + bZ_n \ ,$$

with $-1 < a < 1$. The transition kernel formula (3) simplifies to

$$P(x,y) = \frac{1}{Z} e^{-\frac{1}{2b^2}(y-ax)^2} \ . \tag{16}$$

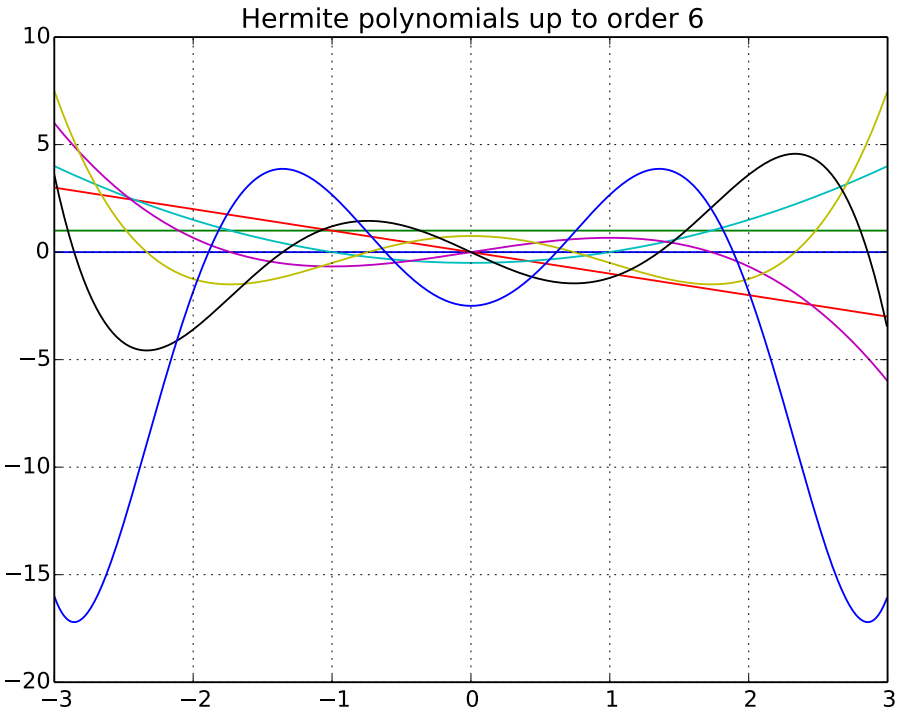

Figure 1: Plots of Hermite polynomials, featuring $H_0(x) \equiv 1$, and $\frac{1}{k}H_k(x)$ for $k = 1, \ldots, 5$.

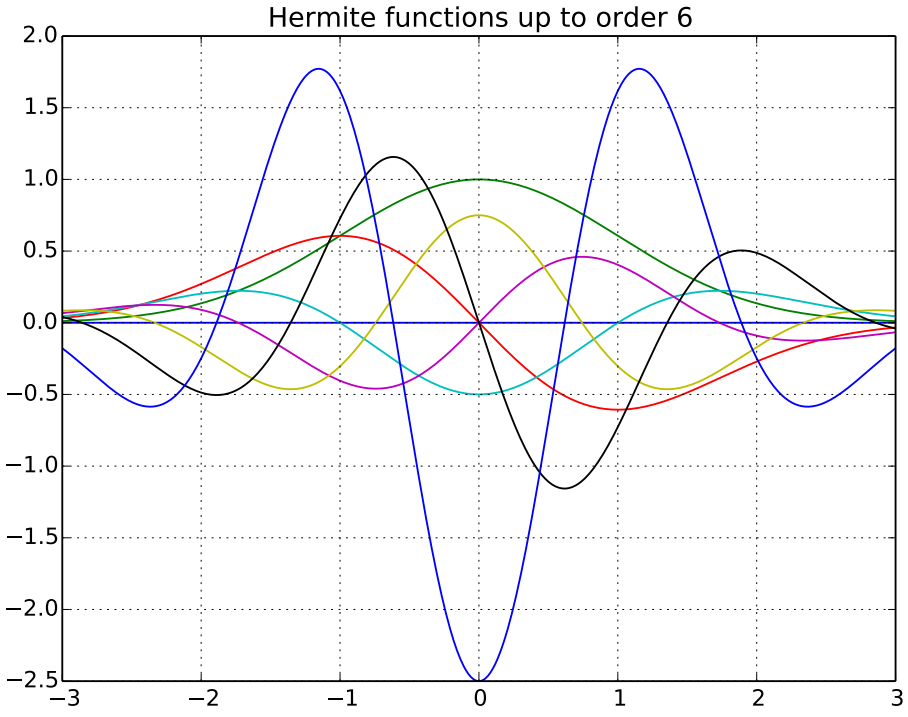

Figure 2: Hermite functions $\frac{1}{k}h_k(x)$, where $h_k(x) = H_k(x)e^{-\frac{1}{2}x^2}$.

Suppose $f(x)$ is the steady state density, which is Gaussian in this case. Although we do not need it here, it is easy to find the steady state variance as follows. If $\sigma^2 = \mathrm{var}(X_n) = \mathrm{var}(X_{n+1})$, then the fact that Z_n is independent of X_n implies that

$$\sigma^2 = a^2\sigma^2 + b^2 \ ,$$

which gives the variance formula for the scalar case

$$\sigma^2 = \frac{b^2}{1-a^2} \ .$$

Simple scaling implies that X_n scales linearly with b, so the variance scales as b^2. The denominator expresses the fact that the variance goes to infinity as a approaches 1.

We simply check that the functions $h_n(x) = \partial_x^n f(x)$ are adjoint eigenfunctions of the integral operator with kernel (16).

$$a^n h_n(y) = \int h_n(x) P(x,y)\, dx \ .$$

We do not need the Gaussian form, only the more general $P(x,y) = K(ax-y)$. The derivation is based on $\partial_x K(ax-y) = aK'(ax-y)$, and $\partial_y K(ax-y) = -K'(ax-y)$. We "do the math" with $n=1$. The result for larger n is similar.

$$\begin{aligned}
\int h_1(x) K(ax-y)\, dx &= \int \left(\partial_x f(x)\right) K(ax-y)\, dx \\
&= -\int f(x) \partial_x K(ax-y)\, dx \\
&= -a \int f(x) K(ax-y)'\, dx \\
&= a \int f(x) \partial_y K(ax-y)\, dx \\
&= a\partial_y \left(\int f(x) K(ax-y)\, dx \right) \\
&= a\partial_y f(y) \\
&= a h_1(y) \ .
\end{aligned}$$