Lecture 20

Better Iterative Methods; Google and Markov Chains

David Semeraro

University of Illinois at Urbana-Champaign

November 12, 2013

Applying Iterations

Beginning at the k^{th} iteration...

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_0^k \\ x_1^k \\ x_2^k \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

Table : Iterative Methods

Jacobi	Gauss-Seidel
$x_0^{k+1} = \frac{x_1^k + b_0}{2}$	$x_0^{k+1} = \frac{x_1^k + b_0}{2}$
$x_1^{k+1} = \frac{x_0^k + x_2^k + b_1}{2}$	$x_1^{k+1} = \frac{x_0^{k+1} + x_2^k + b_1}{2}$
$x_2^{k+1} = \frac{x_1^k + b_2}{2}$	$x_2^{k+1} = \frac{x_1^{k+1} + b_2}{2}$

Quadratic Form

$$f(x) = \frac{1}{2}x^T Ax - b^T x + c$$

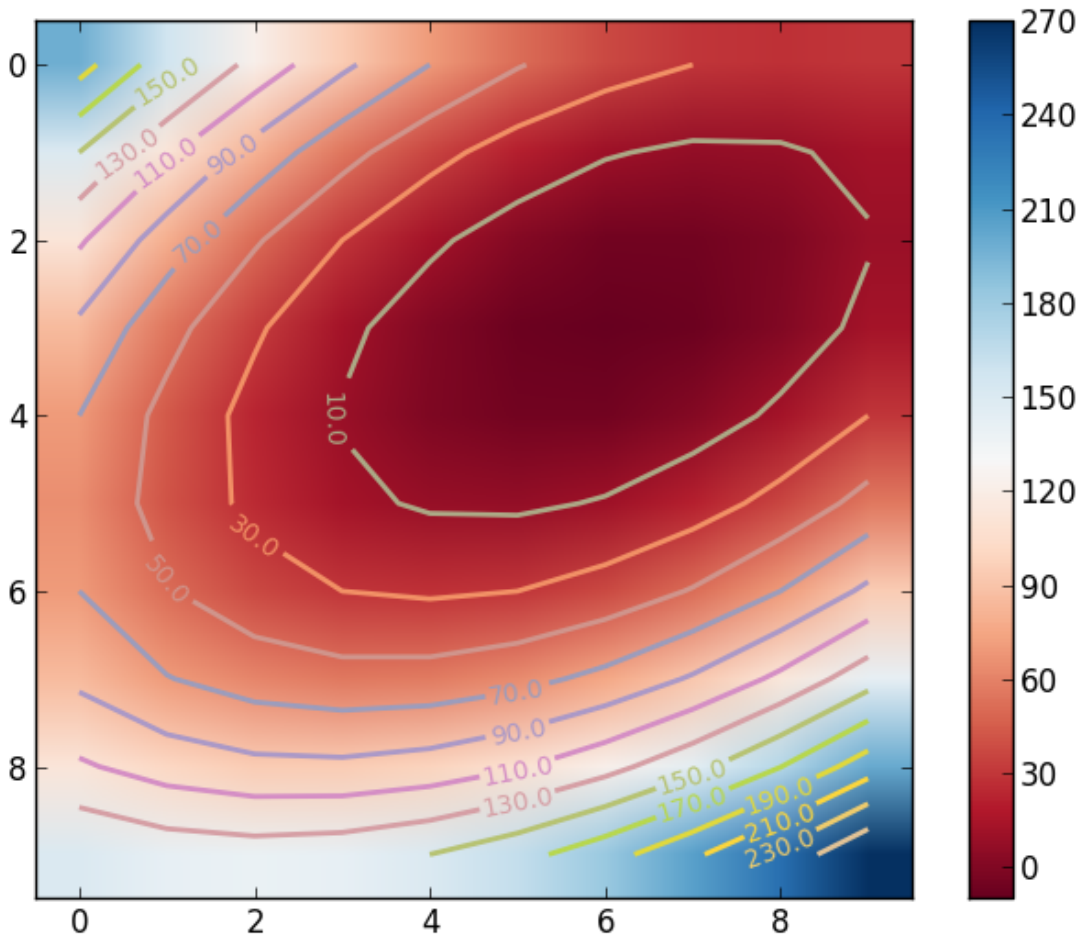
is a *quadratic* form. Example:

$$A = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ -8 \end{bmatrix} \quad c = 0$$

```
x,y = np.meshgrid(np.linspace(-6.,6.,n),np.linspace(-6.,6.,n))
z = f(x,y,n)
im = plt.imshow(z,cmap=plt.cm.RdBu)
cset = plt.contour(z,np.arange(-10.,300.,20.0),linewidths=2,
    cmap=plt.cm.Set2)
plt.clabel(cset,inline=True,fmt='%1.1f',fontsize=10)
plt.colorbar(im) # adding the colobar on the right
```

Quadratic Form

Gradient: $\nabla f(x)$ points uphill



Quadratic Form

Lets look at the i^{th} component of ∇f:

$$\begin{aligned} f(x+he_i) &= \frac{1}{2}(x+he_i)^T A(x+he_i) - b^T(x+he_i) + c \\ &\approx \frac{1}{2}(x^T Ax + he_i^T Ax + x^T Ahe_i) - b^T(x+he_i) + c \end{aligned}$$

so

$$\begin{aligned} \frac{f(x+he_i) - f(x)}{h} &= \frac{\frac{1}{2}(he_i^T Ax + x^T Ahe_i) - he_i^T b}{h} \\ &= \text{i}^{th} \text{ component of } \frac{1}{2}(Ax + A^T x) - b \end{aligned}$$

So

$$\nabla f = \frac{1}{2}A^T x + \frac{1}{2}Ax - b$$

So if A is symmetric, then $Ax = b$ at the minimum of f.

Quadratic Form

So

$$\nabla f = \frac{1}{2}A^T x + \frac{1}{2}Ax - b$$

So if A is symmetric, then $Ax = b$ at the minimum of f. Now, if

A is positive definite, f is concave up.
A is negative definite, f is concave down.
A is positive semi-definite, f is concave up (with a line as the minimum).
A is indefinite, f has a saddle. (Think $x_1^2 - x_2^2$)

Conjugate Gradients

- Suppose that A is $n \times n$ symmetric and positive definite.
- Since A is positive definite, $x^T Ax > 0$ for all $x \in \mathbb{R}^n$.
- Define a quadratic function

$$f(x) = \frac{1}{2}x^T Ax - x^T b$$

- It turns out that $-\nabla f = b - Ax = r$
- Optimization methods look in a "search direction" and pick the best step:

$$x_{k+1} = x_k + \alpha s_k$$

  Choose α so that $f(x_k + \alpha s_k)$ is minimized in the direction of s_k.
- Find α so that f is minimized:

$$0 = \frac{d}{d\alpha} f(x_{k+1}) = \nabla f(x_{k+1})^T \frac{d}{d\alpha} x_{k+1} = -r_{k+1}^T \frac{d}{d\alpha}(x_k + \alpha s_k) = -r_{k+1}^T s_k.$$

Steepest Descent

- Find α so that f is minimized in the direction of $\nabla f = r$:

$$0 = \frac{d}{d\alpha} f(x_{k+1}) = \nabla f(x_{k+1})^T \frac{d}{d\alpha} x_{k+1} = -r_{k+1}^T \frac{d}{d\alpha}(x_k + \alpha r_k) = -r_{k+1}^T r_k.$$

- Pick α so that $\nabla f(x_{k+1})$ and r_k are orthogonal.
- Since $\nabla f(x_{k+1}) = -r_{k+1}$ we want

$$\begin{aligned} r_{k+1}^T r_k &= 0 \\ (b - A(x_k + \alpha r_k))^T r_k &= 0 \\ (b - Ax_k)^T r_k - \alpha (Ar_k)^T r_k &= 0 \\ r_k^T r_k &= \alpha r_k^T A r_k \end{aligned}$$

- So, the optimal search parameter is

$$\alpha = -\frac{r_k^T r_k}{r_k^T A r_k}$$

So far, we've *motivated* a method. We haven't been careful about the step direction s_k yet.

- Notice: convergence is staggered

So far, we've *motivated* a method. We haven't been careful about the step direction s_k yet.

- Notice: convergence is staggered
- Better: step toward a principle axis

So far, we've *motivated* a method. We haven't been careful about the step direction s_k yet.

- Notice: convergence is staggered
- Better: step toward a principle axis
- Notice: axis defined by eigenvectors of A

So far, we've *motivated* a method. We haven't been careful about the step direction s_k yet.

- Notice: convergence is staggered
- Better: step toward a principle axis
- Notice: axis defined by eigenvectors of A
- Compromise: find a step direction that is A-orthogonal to the previous

So far, we've *motivated* a method. We haven't been careful about the step direction s_k yet.

- Notice: convergence is staggered
- Better: step toward a principle axis
- Notice: axis defined by eigenvectors of A
- Compromise: find a step direction that is A-orthogonal to the previous
- Notice: take a search closest to r that is conjugate.

So far, we've *motivated* a method. We haven't been careful about the step direction s_k yet.

- Notice: convergence is staggered
- Better: step toward a principle axis
- Notice: axis defined by eigenvectors of A
- Compromise: find a step direction that is A-orthogonal to the previous
- Notice: take a search closest to r that is conjugate.
- Result: converges in at most n steps

Conjugate Gradients

1 $x_0 =$ `initial guess`
2 $r_0 = b - Ax_0$
3 $s_0 = r_0$
4 `for` $k = 0, 1, 2, \ldots$
5 $\quad \alpha_k = \frac{r_k^T r_k}{s_k^T A s_k}$
6 $\quad x_{k+1} = x_k + \alpha_k s_k$
7 $\quad r_{k+1} = r_k - \alpha_k A s_k$
8 $\quad \beta_{k+1} = r_{k+1}^T r_{k+1} / r_k^T r_k$
9 $\quad s_{k+1} = r_{k+1} + \beta_{k+1} s_k$

Convergence

- A-norm of the error is minimized in each step
- Bound:

$$\|e_k\|_A \leqslant 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|e_0\|_A$$

- κ is the condition number
- for symmetric matrices, $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$

Preconditioning

We've seen iterations that replace A with a matrix Q such that $Q^{-1}A$ resembles the identity matrix. Conjugate Gradient takes a different approach and solves $Ax = b$ more directly.
Can we combine these ideas?
Preconditioning is transforming the original problem $Ax = b$ with another matrix Q and then solving that modified problem. For example

- Left Preconditioning: $QAx = Qb$
- Right Preconditioning: $AQw = b$, $Qw = x$
- Symmetric preconditioning: $Q^TAQw = Q^Tb$, $Qw = x$

For Conjugate Gradient, the matrix must be symmetric, so the last form is the most common.
There are methods that generalize conjugate gradient for nonsymmetric matrices. See CS 457 for more details

Coming Up

- Google, Markov Chains, intro to Monte Carlo Simulations
- Eigenvalues: SVD, Power Method
- Least-Squares
- Monte Carlo Simulations

Random Process

- Goal: model a random process in which a system transitions from one state to another at discrete time steps.
- At each time, say there are n states the system could be in.
- At time k, we model the system as a vector $x_k \in \Re^n$ (whose entries represent the probability of being in each of the n states).
- Here, $k = 0, 1, 2, \cdots$, and the "initial state" is x_0.

Definition

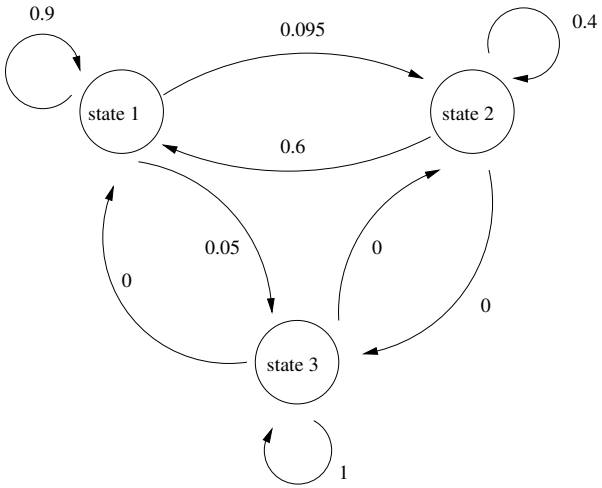
A probability vector is a vector in $\Re^n$ whose entries are nonnegative and sum to 1.

Markov Chains

- Markov chains can model the behavior of a system that depends only on the previous experiment or state.
- That is, the next state of the system depends only on the current state where the outcome of each experiment is one of a discrete set of states.
- Markov chains require a transition matrix, P, where $P(j, i)$ equals the probability of going from state i to state j.

Markov chain model of a Broken Machine

A certain office machine has three states: Working, Temporarily Broken, and Permanently Broken. Each day it is in one of these states. The *transition diagram* between the states is shown below.



On any particular day the probability that the machine is in any particular state is given by the probability vector.

$$x_k = \begin{bmatrix} p_0 \\ p_1 \\ p_2 \end{bmatrix}$$

Markov chain model of a Broken Machine

Given the transition diagram, the probability that the machine will be in a particular state on a subsequent day is.

$$x_{k+1} = \begin{bmatrix} .9 & .6 & 0 \\ .095 & .4 & 0 \\ .005 & 0 & 1 \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \end{bmatrix}$$

Therefore, the transition matrix is:

$$P = \begin{pmatrix} .9 & .6 & 0 \\ .095 & .4 & 0 \\ .005 & 0 & 1 \end{pmatrix}$$

Markov chain model of a Broken Machine

Let us say the initial state is the machine is working.

$$x_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

The probability that it will be in any particular state on day 2 is

$$x_{k+1} = \begin{bmatrix} .9 & .6 & 0 \\ .095 & .4 & 0 \\ .005 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} .9 \\ .095 \\ .005 \end{bmatrix}$$

And in a year the probability that the machine will be in any particular state is:

$$\begin{bmatrix} .9 & .6 & 0 \\ .095 & .4 & 0 \\ .005 & 0 & 1 \end{bmatrix}^{365} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.1779 \\ 0.0284 \\ 0.7937 \end{bmatrix}$$

Markov chain model of a Broken Machine

We see that the probability that the machine is permanently broken in 0.7937 and so there is about a 21% chance that the machine is still functional.

Definition

A Markov matrix is a square matrix M whose columns are probability vectors.

Definition

A Markov chain is a sequence of probability vectors $x_0, x_1, x_2, \cdots$ such that $x_{k+1} = Mx_k$ for some Markov matrix M.

Mardov chain model of a Broken Machine

For this Markov chain the steady state solution is the vector $\mathbf{v} = [0\ 0\ 1]'$, as we can see by computing the eigenvalues and eigenvectors. That is, $P\mathbf{v} = \mathbf{v}$.

State 3 (Permanently Broken) is called an *absorbing state*. No matter what state we start in, we will eventually end up in State 3 with probability 1, and once in State 3 we can never leave.

steady state

Theorem

If M is a Markov matrix, there exists a vector $x \neq 0$ such that $Mx = x$.

Perron-Frobenius Theorem

If M is a Markov matrix with all positive entries, then M has a unique steady-state vector, x. if x_0 is any initial state, then $x_k = M^k x_0$ converges to x as $k \to \infty$.

Randomly Walking with Google

- start at any webpage
- randomly select a link and follow
- repeat
- what are the outcomes?

The outcomes of such a random walk are:

- a dead end on a page with no outgoing links
- a cycle where you end up where you began: known as a *Markov chain* or *Markov process*.
- The limiting probability that an infinitely dedicated random surfer visits any particular page is its PageRank.
- A page has high rank if other pages with high rank link to it.

Random Walking with Google

- Let W be the set of Web pages that can reached by following a chain of hyperlinks starting from a page at Google.
- Let n be the number of pages in W.
- The set W actually varies with time, by the end of 2005, n was over 10 billion.
- Let G be the $n \times n$ connectivity matrix of W, that is, $G_{i,j}$ is 1 if there is a hyperlink from page i to page j and 0 otherwise.
- The matrix G is huge, but very sparse; its number of nonzeros is the total number of hyperlinks in the pages in W.

Google and Probability

- Let c_j and r_i be the column and row sums of G, respectively. That is,

$$c_j = \sum_i G_{i,j}, \qquad r_i = \sum_j G_{i,j}$$

- Then c_k and r_k are the indegree and outdegree of the k-th page. In other words, c_k is the number of links into page k and r_k is the number of links from page k.
- Let p be the fraction of time that the random walk follows a link.
- Google typically takes this to be $p = 0.85$.
- Then $1 - p$ is the fraction of time that an arbitrary page is chosen.

Google meets Markov

- Let A be an $n \times n$ matrix whose elements are $A_{i,j} = pG_{i,j}/c_j + \delta$ where $\delta = (1-p)/n$.
- This matrix is the transition matrix of the Markov chain of a random walk!
- Notice that A comes from scaling the connectivity matrix by its column sums.
- The j-th column is the probability of jumping from the j-th page to the other pages on the Web.

The problem

Can write A, the transition matrix, as

$$A = pGD + ez^T$$

where e is the vector of all ones and where ez^T account for dead linked pages and

$$D_{jj} = 1/c_j \text{ (or 0)} \quad z_j = \delta \text{ (or } 1/n)$$

Then $x = Ax$ can be written

$$(I - pGD)x = (z^T x)e = \gamma e$$

and we can scale x such that $\gamma = 1$

Eigenvectors and Google

Find $x = Ax$ and the elements of x are Google's PageRank. Remember $n > 10^{10}$ (as of 2005) and growing.

For any particular query, Google finds pages on the Web that match the query. The pages are then listed in the order of their PageRank.

Numerics?

- what does this have to do with numerical methods?
- large sparse matrices (although Google avoids this)
- solution to a linear system (eigenvalue problem)
- Markov chains...

The Markov process is important.

Why

- Central to the PageRank (and many many other applications in finance, science, informatics, etc) is that we randomly process something
- what we want to know is “on average” what is likely to happen
- what would happen if we have an infinite number of samples?
- next: eigenvalues, SVD, pagerank, Monte Carlo