Interior Methods for Optimization with application to Maximum Entropy problems

Sandia Saddle-Point Workshop

Santa Fe, New Mexico, Dec 3–6, 2003

Michael Saunders

SOL, Stanford University

saunders@stanford.edu

Entropy Collaborator: John Tomlin, IBM Almaden

Overview

- Least squares
- Regularization
- pdco.m (a MATLAB interior-point optimizer)
- Linear algebra
- Web traffic, Maximum entropy model

Theme

small numbers can have a BIG effeCT!

Least Squares

The classical saddle-point problem

$$\min_x \ \|Ax - b\|_2^2$$

$$\begin{pmatrix} I & A \\ A^T & \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}$$

- Assume $\|A\| \approx 1$
- A sparse or an operator for forming Ax, $A^T y$
- KKT system, equilibrium system

Is Gaussian Elimination safe?

$$\begin{pmatrix} I & A \\ A^T & \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}$$

Pivot on I
$\Rightarrow$ subtract $A^T\times$ row1 from row2:

$$\begin{pmatrix} I & A \\ & -A^TA \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ -A^Tb \end{pmatrix}$$

Scaled augmented system

$$r = \sigma s, \quad \begin{pmatrix} \sigma I & A \\ A^T & \end{pmatrix} \begin{pmatrix} s \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}$$

σ	condition
1	$\approx \mathrm{cond}(A)^2$
$\sigma_{\min}$	$\approx \mathrm{cond}(A)$

- Golub 1966, Björck 1967
- Previous MATLABs: Set $\sigma = \frac{1}{1000} \max |A_{ij}|$, use sparse LU
- With small σ, LU isn't tempted to pivot on σI

Caution

$$\begin{pmatrix} I & A \\ A^T & \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

$$\begin{pmatrix} \sigma I & A \\ A^T & \end{pmatrix} \begin{pmatrix} \frac{r}{\sigma} \\ x \end{pmatrix} = \begin{pmatrix} b \\ \frac{c}{\sigma} \end{pmatrix}$$

- Small σ isn't good if c or r large

Regularized Least Squares

$$\min \left\| \begin{pmatrix} A \\ \delta I \end{pmatrix} x - \begin{pmatrix} b \\ 0 \end{pmatrix} \right\|_2^2$$

$$\begin{pmatrix} \delta I & A \\ A^T & -\delta I \end{pmatrix} \begin{pmatrix} \frac{r}{\delta} \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}$$

$\mathrm{cond}(K) \approx \frac{\|A\|}{\delta}$ $\quad K \uparrow \quad$ $\mathrm{cond}(A)$ isn't relevant

- $P_1 K P_2^T = LU$ would be stable (sparse LU)
- $PKP^T = LU$ is **sufficiently** stable for **any** P if $\delta > 10^{-8}$
- $\Rightarrow PKP^T = LDL^{\mathsf{T}}$ **Indefinite Cholesky** $(D \not\succ 0)$

Regularized Least Squares II

$$(A^T A + \delta^2 I)x = A^T b + c$$

$$\begin{pmatrix} \delta I & A \\ A^T & -\delta I \end{pmatrix} \begin{pmatrix} \frac{r}{\delta} \\ x \end{pmatrix} = \begin{pmatrix} b \\ -\frac{c}{\delta} \end{pmatrix}$$

$$\min \left\| \begin{pmatrix} A \\ \delta I \end{pmatrix} x - \begin{pmatrix} b \\ \frac{c}{\delta} \end{pmatrix} \right\|_2^2$$

- $\delta \neq 0$ permits LS formulation when $c \neq 0$
- Indefinite LDL^T (used for LP in IBM's OSL)
- LSQR (CG method for general A, good stopping rules)

The Entropy Problem

$$\begin{array}{lll} \underset{x}{\text{minimize}} & \varphi(x) = \sum x_j \ln x_j & \\ \text{subject to} & Ax = b, & x > 0. \end{array}$$

$$A = \begin{bmatrix}
1 & 1 & & & & & & & & & & & & & -1 \\
-1 & & 1 & 1 & & & & & & & & & & & \\
 & -1 & & & 1 & 1 & & & & & & & & & \\
 & & -1 & & -1 & & 1 & 1 & & & & & & & \\
 & & & -1 & & -1 & & & 1 & 1 & 1 & & & & \\
 & & & & & & -1 & & -1 & & & 1 & & & \\
 & & & & & & & -1 & & -1 & & & 1 & & \\
 & & & & & & & & & & -1 & & & 1 & \\
 & & & & & & & & & & & -1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}$$

$$\min \sum x_j \ln x_j \quad \text{s.t.} \quad Ax = b$$

- Optimality conditions:

$$\begin{aligned} A^T y &= g(x) \qquad \leftarrow g_j = 1 + \ln x_j \\ Ax &= b \end{aligned}$$

- Erlander 1977, Eriksson 1981: eliminate x, apply Newton
- Newton's method for (x, y) together:

$$\begin{pmatrix} -X^{-1} & A^T \\ A & \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} g - A^T y \\ b - Ax \end{pmatrix} \leftarrow$$

- $X = \mathrm{diag}(x)$ (keep $x > 0$). Plausible method?

Regularized Entropy Problem

$$\begin{array}{ll} \underset{x,\,r}{\text{minimize}} & \sum x_j \ln x_j + \frac{1}{2}\|r/\delta\|^2 \\ \text{subject to} & Ax + r = b, \qquad x > 0. \end{array}$$

- Let y = Lagrange multipliers for $Ax + r = b$
- Optimal $r = \delta^2 y \;\Rightarrow\; Ax + \delta^2 y = b$
- $\delta \approx 10^{-3}$ for "equalities"
- Ideal for MATLAB primal-dual interior solver `pdco.m`
- $\delta > 0$ allows use of LSQR for Δy (CG solver, inexact Newton)

Matlab interior solver pdco.m

$$\begin{array}{ll} \underset{x,\,r}{\text{minimize}} & \varphi(x) + \tfrac{1}{2}\|\gamma x\|^2 + \tfrac{1}{2}\|r\|^2 \\ \text{subject to} & Ax + \delta r = b, \qquad \ell < x < u \end{array}$$

- x, y bounded, unique

$$\begin{cases} \gamma \approx 10^{-3} & \text{Tikhonov} \\ \delta \approx 10^{-3} & \text{“equalities”} \\ \delta = 1 & \text{least squares} \end{cases}$$

- $\varphi(x)$ convex, separable

$$\begin{cases} c^T x & \text{LP} \\ \sum x_j \ln x_j & \text{Maximum entropy} \\ - & \text{NNLS image restoration} \\ \|x\|_1 & \text{Basis Pursuit signal denoising} \end{cases}$$

Primal-Dual interior method

Sequence of subproblems with $\mu = 1, 0.1, \ldots, 10^{-7}$, say:

$$\begin{aligned} Ax + \delta^2 y &= b \\ z + A^T y &= \nabla\varphi(x) \\ Xz &= \mu e \qquad\qquad (x_j z_j = \mu) \end{aligned}$$

Newton:

$$\begin{pmatrix} -H & I & A^T \\ Z & X & \\ A & & -\delta^2 I \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta z \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_2 \\ r_3 \\ r_1 \end{pmatrix}$$

$X = \mathrm{diag}(x)$, $H = \nabla^2\varphi(x)$ diagonal

$Z = \mathrm{diag}(z)$, $r_1 = b - Ax - \delta^2 y$, etc

The Linear Algebra (LP Case)

Pivot on I:

$$\begin{pmatrix} I & & A^T \\ X & Z & \\ & A & \end{pmatrix} \begin{pmatrix} \Delta z \\ \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_2 \\ r_3 \\ r_1 \end{pmatrix}$$

$$\Rightarrow \quad \begin{pmatrix} I & & A^T \\ & Z & -XA^T \\ & A & \end{pmatrix} \begin{pmatrix} \Delta z \\ \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_2 \\ r_4 \\ r_1 \end{pmatrix}$$

The Linear Algebra II

$$\begin{pmatrix} -Z & XA^T \\ A & \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_5 \\ r_1 \end{pmatrix}$$

Let $\Delta x = X \Delta \bar{x}$ (so that $\Delta \bar{x}_j = \frac{\Delta x_j}{x_j}$ = relative change)

$$\begin{pmatrix} -ZX & XA^T \\ AX & \end{pmatrix} \begin{pmatrix} \Delta \bar{x} \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_5 \\ r_1 \end{pmatrix}$$

$$\approx \begin{pmatrix} -\mu I & XA^T \\ AX & \end{pmatrix} \begin{pmatrix} \Delta \bar{x} \\ \Delta y \end{pmatrix} = \begin{pmatrix} r_5 \\ r_1 \end{pmatrix}$$

$XZ \approx \mu I$ on "central path". Also $\sigma_{\min}(XA^T) \approx \mu$.

Least-Squares problem for Δy

$$\min_{\Delta y} \left\| \begin{pmatrix} DA^T \\ \delta I \end{pmatrix} \Delta y - \begin{pmatrix} Dw \\ \frac{r_1}{\delta} \end{pmatrix} \right\|_2$$

$$\begin{aligned} D &= (H + X^{-1}Z)^{-1/2} \quad \text{diagonal} \\ r_1 &= b - Ax - \delta^2 y \end{aligned}$$

- Set `atol` $= 0.001$ initially for LSQR
- Solve inexactly for Δy
- Get corresponding Δx and Δz exactly
- Reduce `atol` by 0.1 if necessary

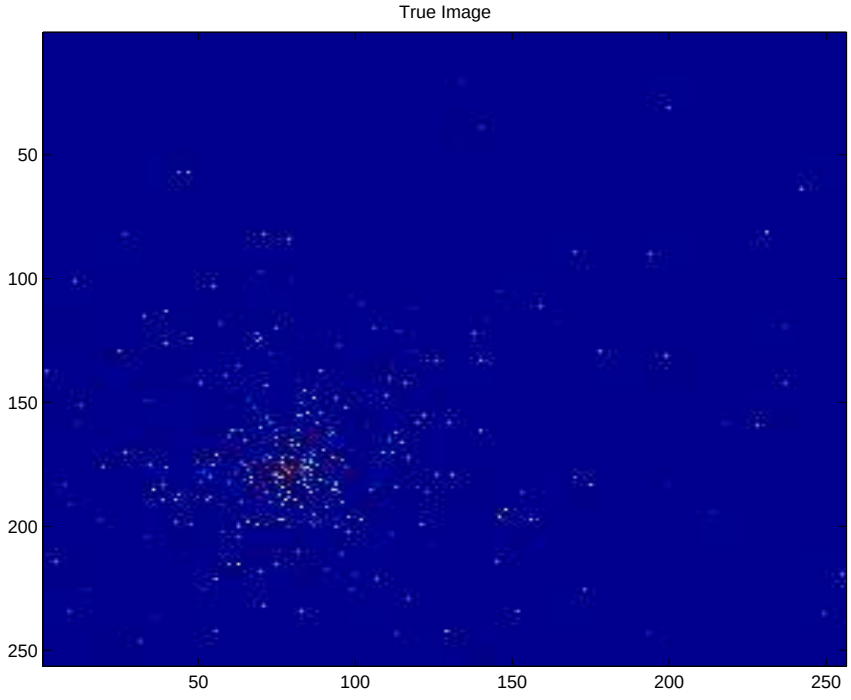
Image restoration: pdco log

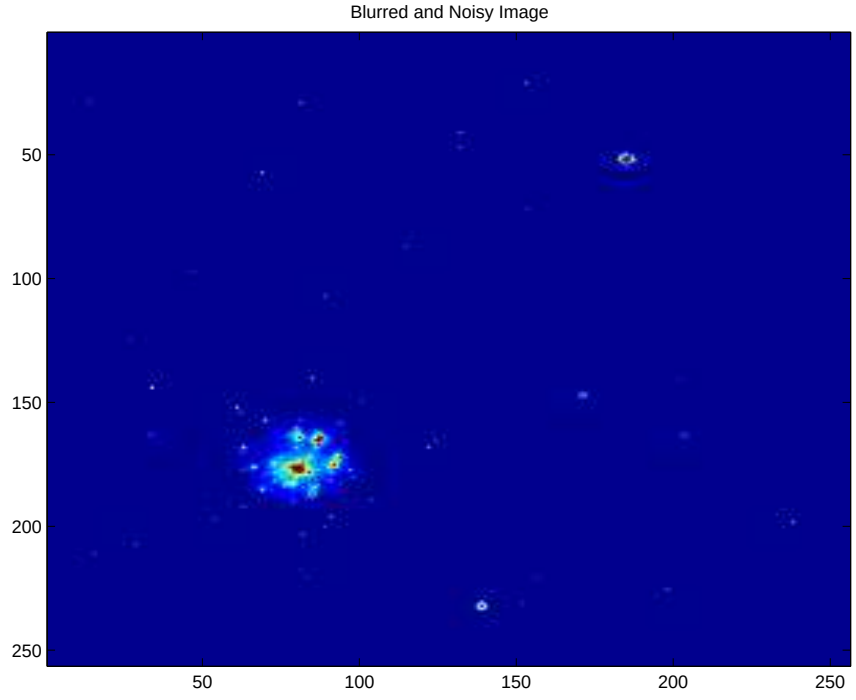
A is a 65536×65536 operator, Star image, NNLS

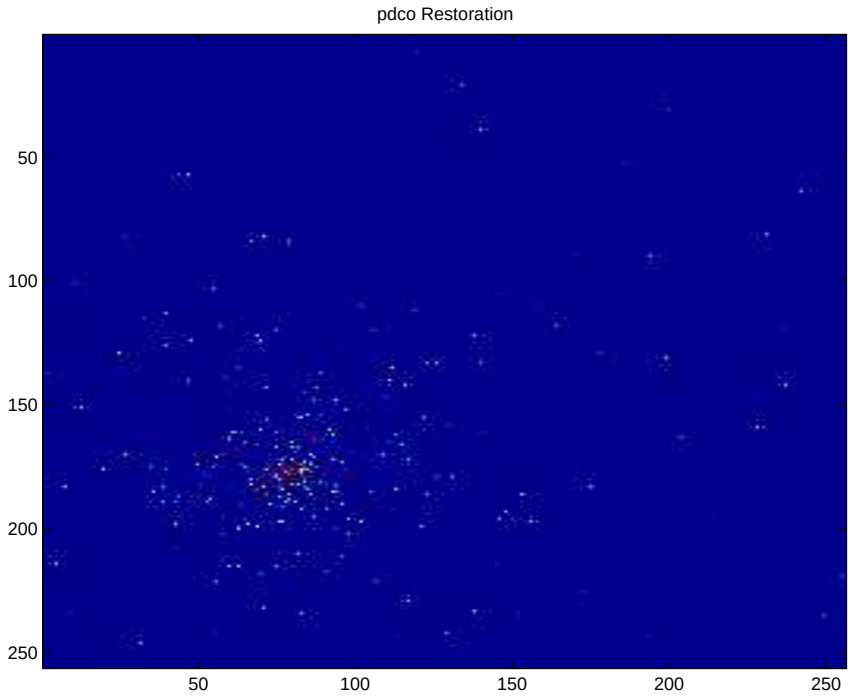
Itn	mu	step	Pinf	Dinf	Cinf	Objective	center	atol	LSQR	Inexact
0			-1.3	0.1	0.0	1.7286986e+05	20.0			
2	-4.6	0.737	-2.5	-1.1	-1.1	1.6151223e+06	107.9	-2.0	12	0.000
4	-5.7	0.621	-2.9	-2.1	-2.1	1.3339537e+06	113.1	-2.6	34	0.001
6	-6.0	0.469	-3.3	-2.6	-2.7	7.7504950e+05	99.7	-3.4	79	0.004
8	-6.0	0.368	-3.5	-3.1	-3.1	4.5536045e+05	86.9	-3.9	153	0.010
10	-6.0	0.331	-3.7	-3.4	-3.5	3.0379116e+05	87.5	-4.3	239	0.015
12	-6.0	0.335	-3.8	-3.8	-3.8	2.2427449e+05	97.0	-4.6	378	0.020
14	-6.0	0.289	-4.0	-4.1	-4.1	1.8493514e+05	84.6	-4.9	589	0.024
16	-6.0	0.307	-4.4	-4.4	-4.4	1.6051790e+05	65.9	-5.1	799	0.034
18	-6.0	0.360	-4.6	-4.8	-4.8	1.4541126e+05	51.1	-5.5	1179	0.032
19	-6.0	0.507	-4.7	-5.1	-5.1	1.4002536e+05	27.2	-5.6	1289	0.041
20	-6.0	0.654	-5.1	-5.6	-5.5	1.3612575e+05	16.5	-5.7	1567	0.037

```
max |x| =     0.994    max |y| =     7.399    max |z| =     0.949   scaled
max |x| = 31247.499    max |y| =    48.258    max |z| =     6.192 unscaled
PDitns  =        20   LSQRitns =     9166    time    =    5979.8     secs
```

Star Image Restoration







Entropy problem: pdco log

A is a 51000×662000 network matrix, nnz$(A) = 2$ million

Itn	mu	step	Pinf	Dinf	Cinf	Objective	center	atol	LSQR	Inexact
0			2.5	1.1	-6.7	-1.3403720e+01	1.0			
1	-5.0	0.267	2.4	1.1	-5.1	-1.3321172e+01	242.0	-3.0	5	0.001
2	-5.1	0.195	2.3	1.0	-5.3	-1.3220658e+01	36.9	-3.0	5	0.001
3	-5.2	0.431	2.1	0.9	-5.2	-1.2942743e+01	122.9	-3.0	5	0.001
4	-5.5	0.466	1.9	0.7	-5.3	-1.2711643e+01	41.8	-3.0	6	0.001
5	-5.7	0.671	1.4	0.2	-5.5	-1.2492935e+01	71.8	-3.0	9	0.001
6	-6.0	1.000	-0.0	-0.8	-5.8	-1.2367004e+01	2.7	-3.0	10	0.001
7	-6.0	1.000	-0.1	-2.3	-6.0	-1.2368200e+01	1.1	-3.0	9	0.002
8	-6.0	1.000	-1.1	-4.7	-6.0	-1.2367636e+01	1.0	-3.0	2	0.009
9	-6.0	1.000	-1.3	-5.7	-6.0	-1.2367655e+01	1.0	-3.0	7	0.015
10	-6.0	1.000	-2.5	-7.6	-6.0	-1.2367607e+01	1.0	-3.0	2	0.004
11	-6.0	1.000	-3.7	-8.6	-6.0	-1.2367609e+01	1.0	-3.5	8	0.004
12	-6.0	1.000	-5.9	-11.0	-6.0	-1.2367609e+01	1.0	-4.7	11	0.000

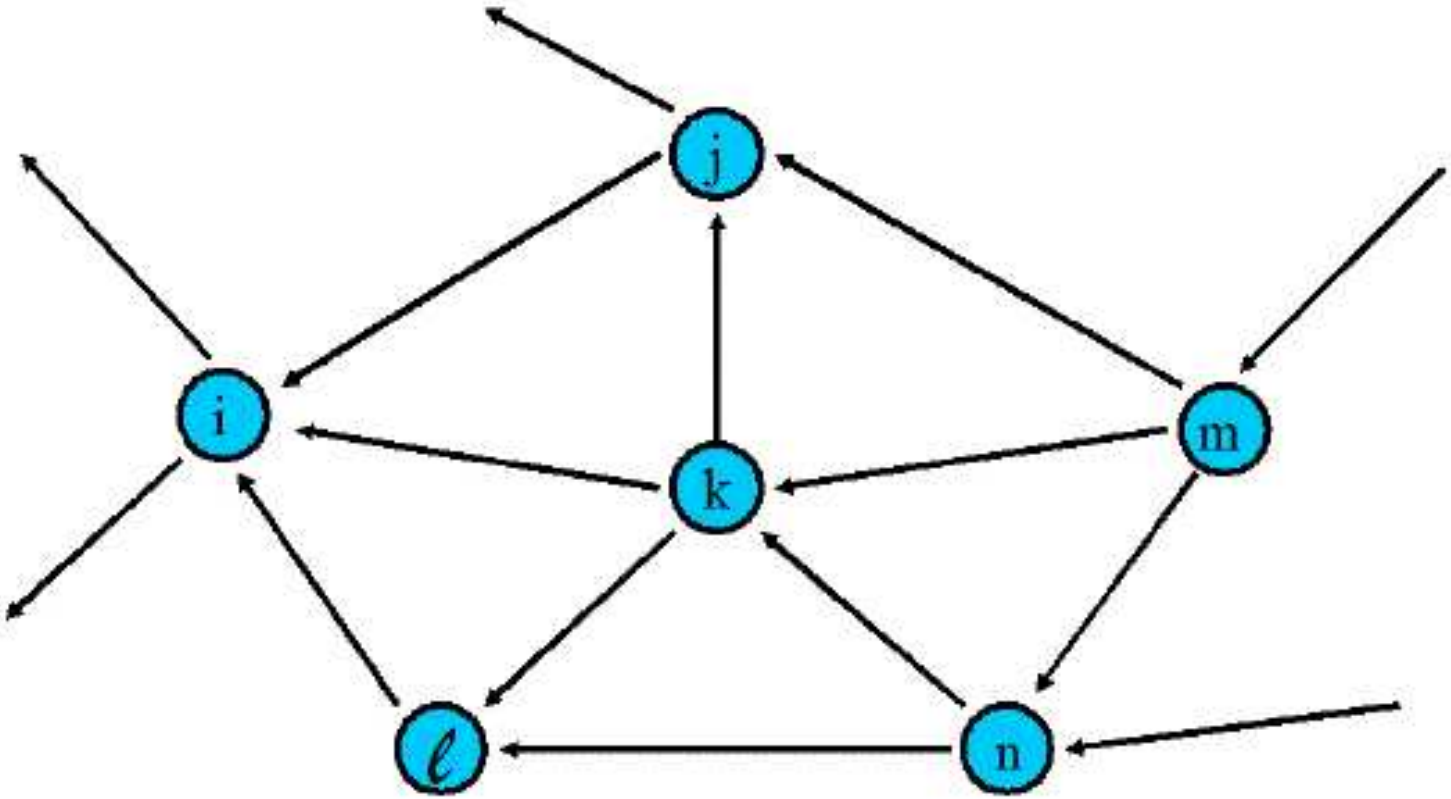
PDitns = 12 LSQRitns = 79 time = 101.4 (MATLAB)
22.4 (C++)

Maximum entropy models

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & S = -\sum x_j \ln x_j \\ \text{subject to} \quad & Ax = b, \quad x > 0 \end{aligned}$$

- Transportation planning
- Probabilistic Query Models for transaction data
- Natural Language Processing, Knowledge Management, etc.
- Model "traffic" on the World Wide Web

Network Graph



Graph model of the web:

$$G = (V,E)$$

Where V is set of vertices $(i,j,k,\ldots..)$ or nodes or pages

And E is the set of edges (i,j)

The Random Web Surfer

- Assume some notional "clock"
- At each clock tick the web surfer follows an out-link with some probability
- Model 1 (Google)
  Assume the probabilities are fixed
  e.g., probability of following an out-link from page i is $1/d_i$, where d_i is the out-degree
- Model 2 (John Tomlin, IBM)
  Let probabilities be variables to be determined

1: Markov Chain Model

- Assume transition probabilities p_{ij} from page i to page j are fixed at $1/d_i$ (constant for page i)
- Let $P = (\, p_{ij} \,)$. The stationary state of the Markov Chain defined by P is then x, where

$$x = P^T x$$

(dominant left eigenvector of the stochastic matrix P)

- The value x_i is the ideal PageRank of page i
(as used in Google, with additional features, to assign importance to a web page)

2: Network Flow Model

A richer class of models (John Tomlin, IBM)

Let variable y_{ij} be the number of surfers clicking on link (i, j) at each clock tick. Then

$$H_j = \sum_{i|(i,j)\in E} y_{ij}$$

is the number of hits per unit time at node j.

Conservation

$$\sum_{j|(i,j)\in E} y_{ij} - \sum_{j|(j,i)\in E} y_{ji} = 0 \quad (i = 1, \ldots, n)$$

$$Y = \sum_{i,j} y_{ij} = \sum_j H_j \quad \text{(total flow)}$$

Probabilistic Network Model

- Usually we prefer to work with normalized values (probabilities) $p_{ij} = y_{ij}/Y$. Constraints become

$$\sum_{j|(i,j)\in E} p_{ij} - \sum_{j|(j,i)\in E} p_{ji} = 0 \quad (i = 1, \ldots, n)$$

$$\sum_{i,j} p_{ij} = 1$$

- The PageRank model specifies a particular solution:

$$p_{ij} = \frac{H_i}{Y d_i} \quad \forall (i,j) \in E$$

They satisfy the conservation equations, but ... only one of infinitely many solutions

Entropy Objective

- Principle:
  In the absence of complete information about a probability distribution, choose the one that maximizes uncertainty, subject to whatever is known (Jaynes).
  This unique distribution is the (Shannon) entropy function.
- Hence we should

$$\text{maximize} \quad S = -\sum_{i,j} p_{ij} \log p_{ij}$$

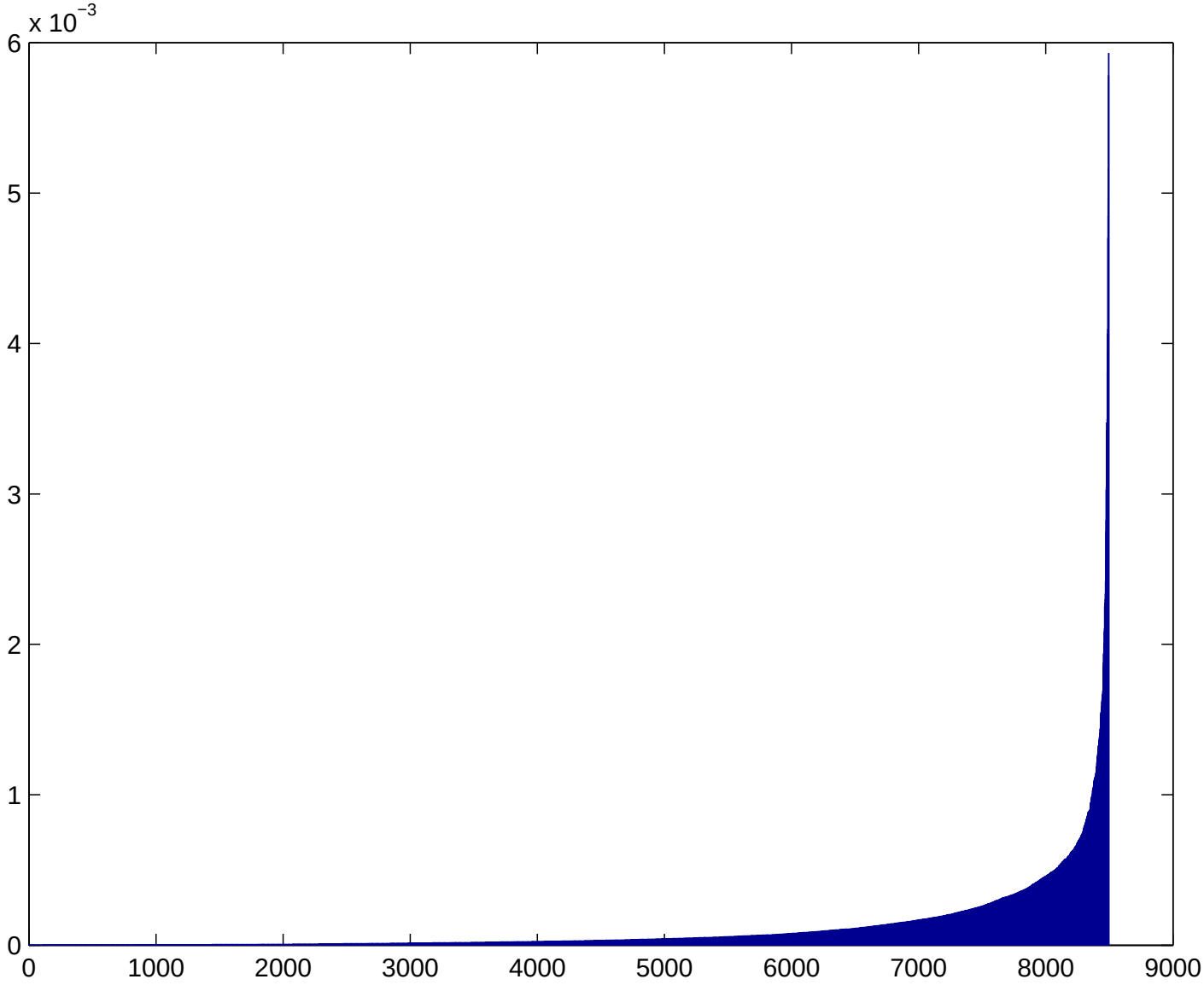
  subject to the constraints

Observations

Obj	Size of A		PD itns	LSQR itns	Time	
Ent	500 $\times$	8500	12	61	1	MATLAB
Ent	51000 $\times$	662000	12	79	101	MATLAB
LP	51000 $\times$	662000	24	1800	1000	MATLAB
Ent	1600000 $\times$	13000000	18	183	1100	C++, single

- Entropy models (and network constraints) are exceptionally friendly for interior methods
- pdco needs very few primal-dual iterations even with inexact search directions
- LSQR needs very few iterations to compute the inexact directions even near solution
- AA^T is very dense $\Rightarrow$ Bad for sparse Cholesky (e.g. CPLEX)

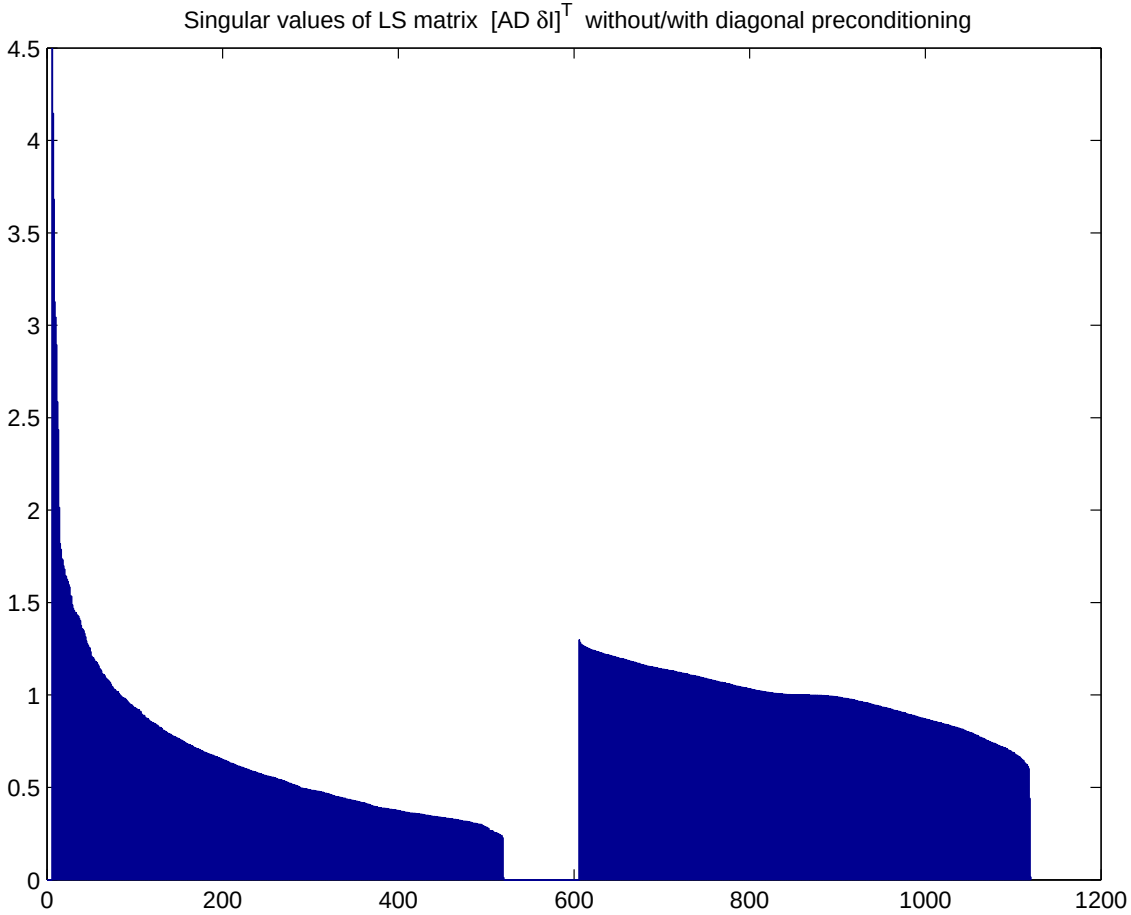
Distribution of x^*



A is 500×8500

Singular values at x^*

$\begin{pmatrix} DA^T \\ \delta I \end{pmatrix}$ $\begin{pmatrix} DA^T \\ \delta I \end{pmatrix}$ + diag preconditioning



Conclusions

- Network formulation more general than PageRank
- Primal-dual interior method effective for large entropy problems
- Further understanding needed
  but singular values tell a story
- Current implementations: O(1 million) variables
- For millions of nodes, need
  64-bit machines for in-core implementation
  distributed computation

Thanks

- Chris Paige, Gene Golub
- Shaobing Chen, David Donoho (signal denoising)
- Byunggyoo Kim, James Nagy (image restoration)
- John Tomlin (pdco, lsqr in C++)
- Sou Cheng Choi
- Kenneth Holmström (pdco, lsqr-Mex in TOMLAB)

[1] Åke Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[3] Jan Eriksson. *Algorithms for Entropy and Mathematical Programming*. PhD thesis, Dept of Mathematics, Linköping University, Sweden, 1981.

[4] A. Forsgren, P. E. Gill, and J. R. Shinnerl. Stability of symmetric ill-conditioned systems arising in interior methods for constrained optimization. *SIMAX*, 17:187–211, 1996.

[5] G. H. Golub and C. F. Van Loan. Unsymmetric positive definite linear systems. *LAA*, 28:85–98, 1979.

[6] Byunggyoo Kim. PhD thesis, Dept of Management Science and Engineering, Stanford University.

[7] C. C. Paige and M. A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM TOMS*, 8(1):43–71, 1982.

[8] M. A. Saunders. Cholesky-based methods for sparse least squares: The benefits of regularization. In L. Adams and J. L. Nazareth, editors, *Linear and Nonlinear Conjugate Gradient-Related Methods*, pages 92–100. SIAM Publications, Philadelphia, 1996.

[9] M. A. Saunders. pdco.m, a primal-dual interior solver for convex optimization. MATLAB software, http://www.stanford.edu/group/SOL/software/, 2003.

[10] J. A. Tomlin. A new paradigm for ranking pages on the World Wide Web. In *Proceedings of WWW2003, Budapest, Hungary*, pages ACM 1–58113–680–3/03/0005, 2003.

Sometimes small numbers aren't big enough

From **The Listener** (NZ TV Guide):

> He said the fee was increased from $5 to $20 because some people had complained it was not worth writing a cheque for $5.
>
> "The pedestrian count was not considered high enough to justify an overbridge", Helen Ritchie said. "And if there continues to be people knocked down on the crossing, the number of pedestrians will dwindle."