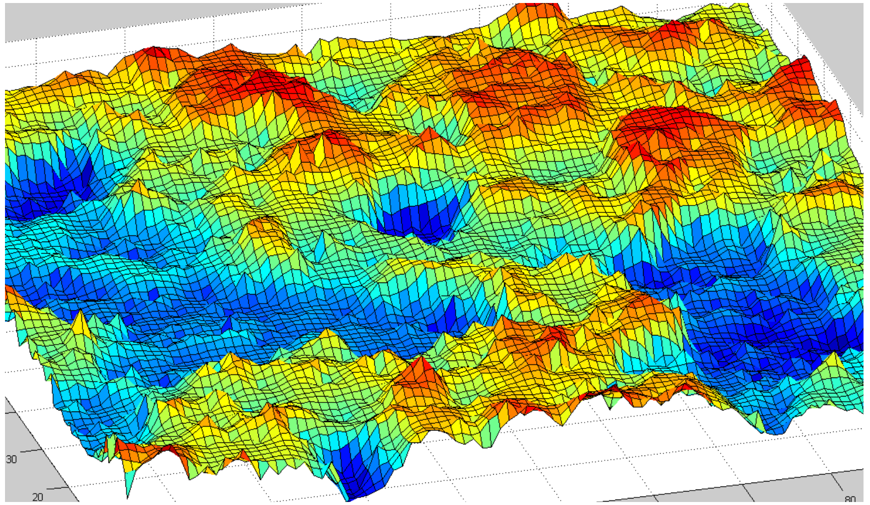
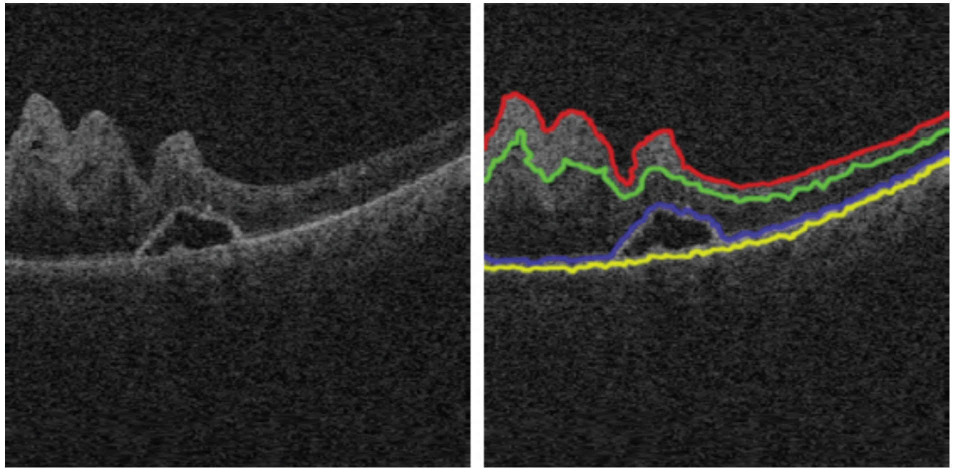
Sparse Newton's Method

Yu Cheng

May 3, 2016

Large Graphs





Spectral Graph Theory

Graph algorithms

Paths Flows
Cuts …

Combinatorial

Scientific computing

Matrices Eigenvalues
Iterative Methods …

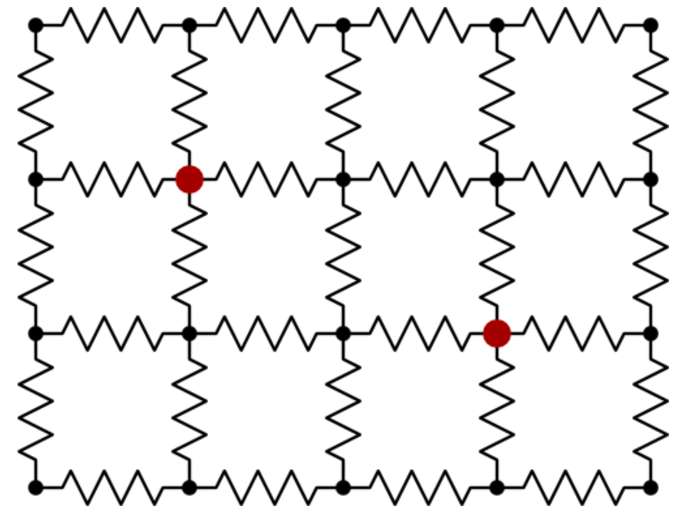
Numerical

Scalable Algorithms

Electric Flow

- Electric flow / SDD linear systems
  - Direct methods: $O(n^{2.37})$
  - Iterative methods: $O(m\sqrt{\kappa})$
- [ST'04, KMP '11, KOSZ '13, CKMPPRX'14]

$$O\left(m\sqrt{\log n}\ \log\frac{1}{\epsilon}\right)$$



n = size of the graph / dimension of the matrix
m = number of edges / number of non-zeros
ϵ = precision κ = condition number

Our Contributions

Graph algorithms

Spectral sparsification

Scientific computing

Newton's method

Sparse Newton's method

Our Contributions

First nearly-linear work and polylog depth algorithm for:

1. Sampling Gaussian MRF with SDD precision matrix

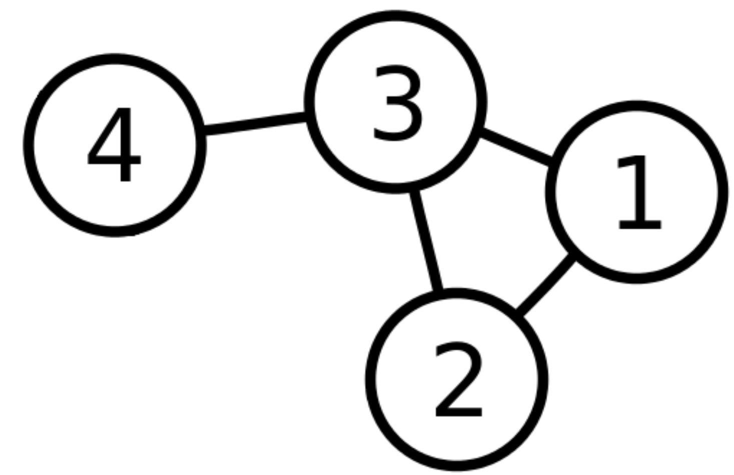
2. Computing pth power of SDDM matrices for $p \in [-1,1]$

3. Spectral Sparsification of Random-Walk Matrix Polynomials

Sampling Gaussian Graphical Models

Markov Random Field (MRF)

- Random variables described by an undirected graph
- Conditionally independence
- Gaussian MRF:
  - $\Pr(x|\Lambda, h) \propto \exp\left(-\frac{1}{2}x^T\Lambda x + h^T x\right)$
  - Precision matrix $\Lambda = \Sigma^{-1}$
  - Potential vector h



Previous Work

- Gibbs sampling $\Omega(n^2)$ work, highly sequential

Initialize $x^{(0)} \sim q(x)$
for iteration $i = 1, 2, \ldots$ **do**
$x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}, \ldots, X_D = x_D^{(i-1)})$
$x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}, \ldots, X_D = x_D^{(i-1)})$
$\vdots$
$x_D^{(i)} \sim p(X_D = x_D | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}, \ldots, X_D = x_{D-1}^{(i)})$
end for

Our Result

- For Gaussian MRFs with SDD precision matrices Λ

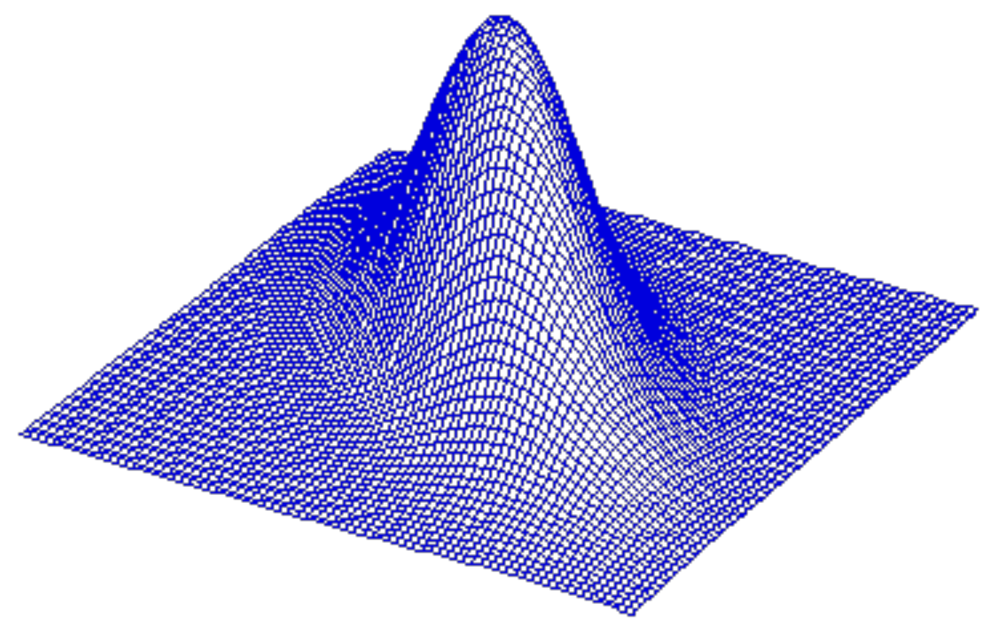
$$\Pr(x|\Lambda, h) \propto \exp\left(-\frac{1}{2}x^T\Lambda x + h^T x\right)$$

  - Work: $O(m \log^c n)$
  - Depth: $O(\log^c n)$
  - Randomness: $O(n)$

Sampling Gaussian MRF

- Input: precision matrix $\Lambda = \Sigma^{-1}$
  potential vector h

1. Compute mean $\mu = \Lambda^{-1} h$
2. Find $CC^T = \Lambda^{-1}$
3. Draw $z \sim \mathcal{N}(0, I)$
   Return $x = Cz + \mu$



Splitting

- Can assume $\Lambda = I - X$
  - $\rho(X) = \max|\lambda(X)| \leq 1 - \frac{1}{\kappa}$
  - $\kappa = \kappa(\Lambda) = \frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Lambda)}$ poly(n)

- Goal: $CC^T = (I - X)^{-1}$

$(I - X)^{-1/2}$: First Attempt

- Cholesky factorization of Λ^{-1}
  - $CC^T = (I - X)^{-1}$
  - $\Omega(n^{2.37})$ work

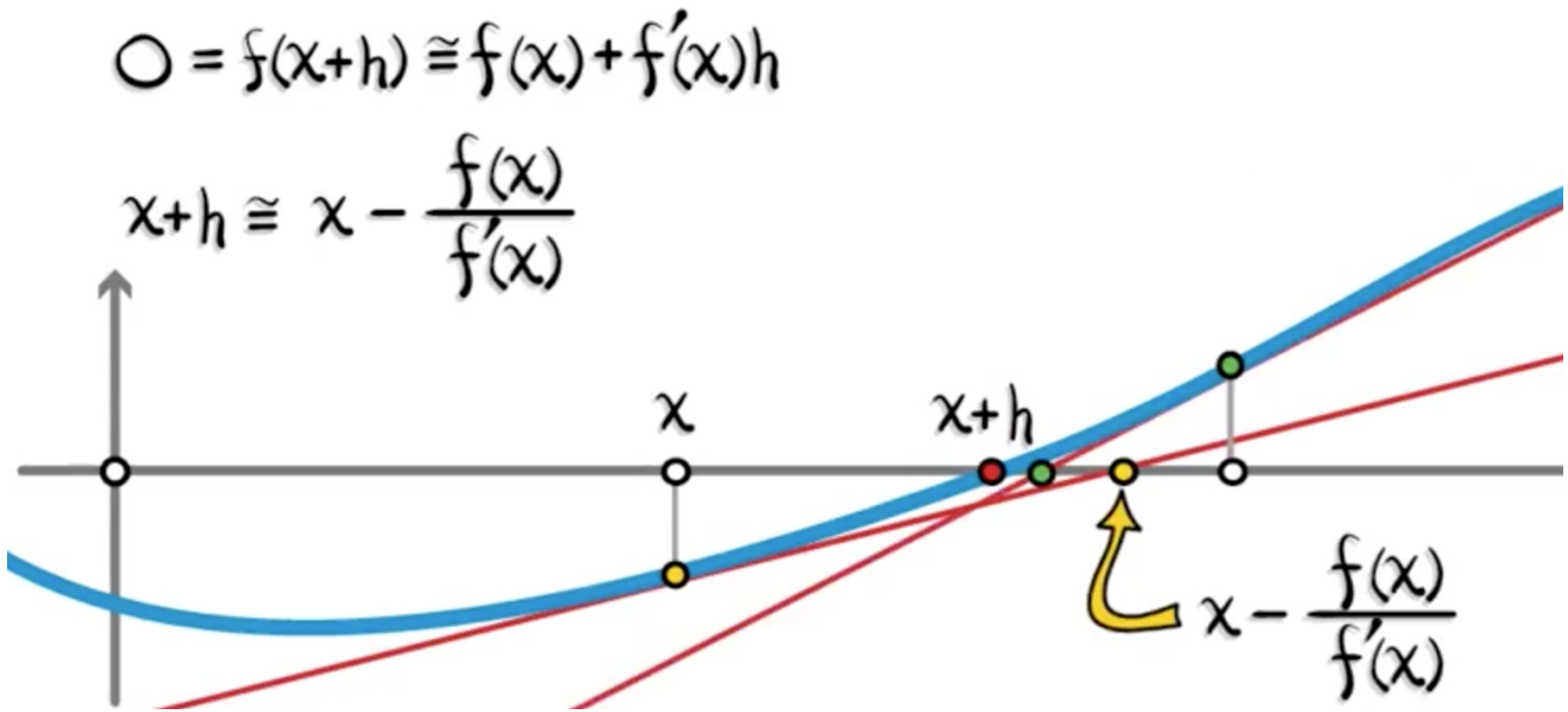
$$\begin{pmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{pmatrix} \begin{pmatrix} 2 & 6 & -8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{pmatrix}$$

- Edge-vertex incident matrices
  - $m \times n$ matrix B with $B^T B = (I - X)$
  - Use $C = B(I - X)^{-1}$
  - $O(m)$ random numbers

$(I - X)^{-1/2}$: Second Attempt

- Taylor expansion
- $(I - X)^{-1/2} = I + \frac{1}{2}X + \frac{3}{8}X^2 + \frac{5}{16}X^3 + \cdots$
- $\rho(X) \leq 1 - \frac{1}{\kappa}$ Need $O(\kappa)$ terms

Newton's Method



Newton's Method

Find the root of $f(y) = \frac{1}{y^2} - (1 - a) \implies y = \frac{1}{\sqrt{1-a}}$

$$y_0 = 1 \qquad y_{k+1} = y_k - \frac{f(y_k)}{f'(y_k)} = y_k\left[1 + \frac{1-(1-a)y_k^2}{2}\right]$$

$$r_k \triangleq 1 - (1 - a)y_k^2 \qquad r_0 = a$$

$$y_k = \prod_{i=0}^{k}\left(1 + \frac{r_i}{2}\right) \qquad 1 - r_{k+1} = \left(1 + \frac{r_k}{2}\right)^2 (1 - r_k)$$

$(I-X)^{-1/2}$: Newton's Method

$$(I-X)^{-1} = \left(I+\tfrac{1}{2}X\right)\left(I-\tfrac{3}{4}X^2-\tfrac{1}{4}X^3\right)^{-1}\left(I+\tfrac{1}{2}X\right)$$

Reduce factoring $(I-X)^{-1}$ to factoring $\left(I-\tfrac{3}{4}X^2-\tfrac{1}{4}X^3\right)^{-1}$

$$r_k \triangleq 1-(1-a)y_k^2$$

$$y_k = \prod_{i=0}^{k}\left(1+\frac{r_i}{2}\right)$$

$$r_0 = a$$

$$1-r_{k+1} = \left(1+\frac{r_k}{2}\right)^2(1-r_k)$$

$(I - X)^{-1/2}$: Newton's Method

$$(I - X_0)^{-1} = \left(I + \tfrac{1}{2}X_0\right)\left(I - \tfrac{3}{4}X_0^2 - \tfrac{1}{4}X_0^3\right)^{-1}\left(I + \tfrac{1}{2}X_0\right)$$

$$= \left(I + \tfrac{1}{2}X_0\right)(I - X_1)^{-1}\left(I + \tfrac{1}{2}X_0\right)$$

$$= \left(I + \tfrac{1}{2}X_0\right)\left(I + \tfrac{1}{2}X_1\right)\left(I - \tfrac{3}{4}X_1^2 - \tfrac{1}{4}X_1^3\right)^{-1}\left(I + \tfrac{1}{2}X_1\right)\left(I + \tfrac{1}{2}X_0\right)$$

$$X_0 = X \qquad X_{i+1} = \tfrac{3}{4}X_i^2 + \tfrac{1}{4}X_i^3$$

$$C = \left(I + \frac{X_0}{2}\right)\left(I + \frac{X_1}{2}\right)\ldots\left(I + \frac{X_d}{2}\right)$$

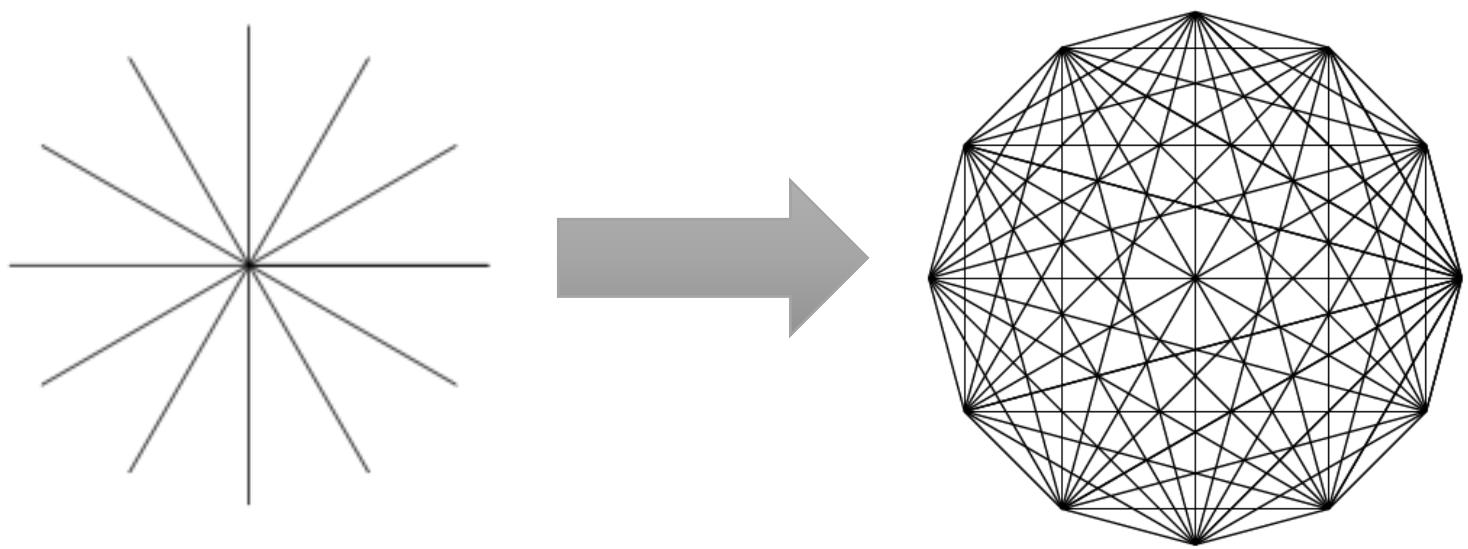
Factor Chain

- $X_0 = X$
- $X_{i+1} = \frac{3}{4}X_i^2 + \frac{1}{4}X_i^3$ $\qquad$ $\rho(X_{i+1}) = \rho\left(\frac{3}{4}X_i^2 + \frac{1}{4}X_i^3\right) \leq \rho(X_i)^2$
- $X_d \longrightarrow 0$

- $C = \left(I + \frac{X_0}{2}\right)\left(I + \frac{X_1}{2}\right) \ldots \left(I + \frac{X_d}{2}\right)$
- $CC^T = (I - X)^{-1}$

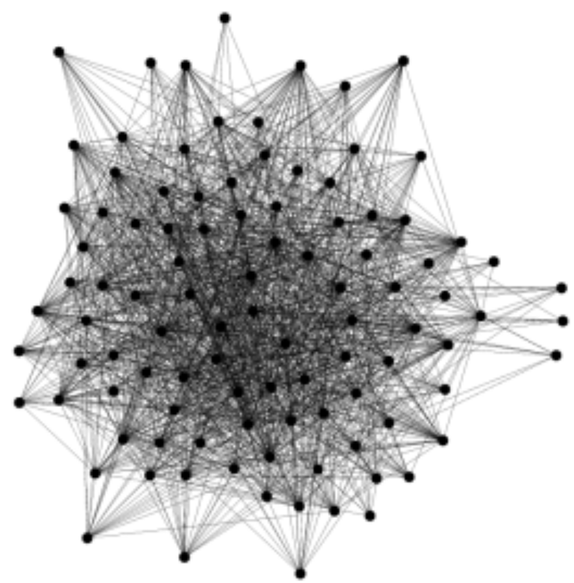
- Convergence: $d = \log\log(\kappa/\epsilon)$ $\qquad$ $\rho(X) \leq 1 - \frac{1}{\kappa}$

Preserving Sparsity



$I - X$ $\qquad$ $I - \frac{3}{4}X_i^2 + \frac{1}{4}X_i^3$

Spectral Sparsification



$$\Lambda \approx_\epsilon \tilde{\Lambda} \iff x^T \tilde{\Lambda} x \approx_\epsilon x^T \Lambda x, \qquad \forall x \in \mathbb{R}^n$$

Sparse Factor Chain

- $X_0 = X$
- ~~$X_{i+1} = \frac{3}{4}X_i^2 + \frac{1}{4}X_i^3$~~ $\quad I - X_{i+1} \approx_\epsilon I - \frac{3}{4}X_i^2 - \frac{1}{4}X_i^3$
- $X_d \longrightarrow 0$

- $C = \left(I + \frac{X_0}{2}\right)\left(I + \frac{X_1}{2}\right) \ldots \left(I + \frac{X_d}{2}\right)$
- $CC^T = (I - X)^{-1}$

- Convergence: $d = \log\log(\kappa/\epsilon)$

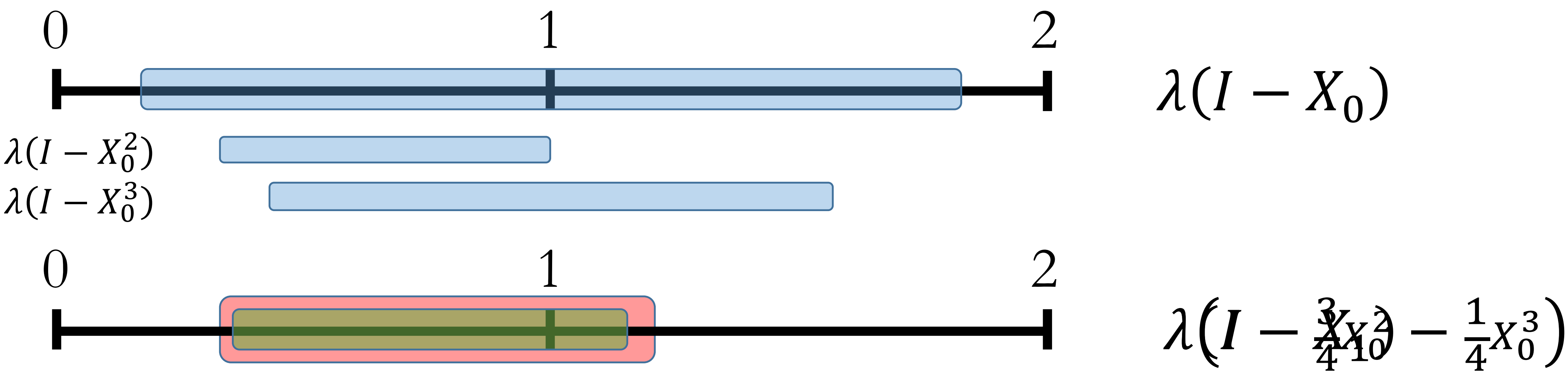
Sparse Factor Chain

- $X_0 = X$
- ~~$X_{i+1} = \frac{3}{4}X_i^2 + \frac{1}{4}X_i^3$~~ $\quad I - X_{i+1} \approx_\epsilon I - \frac{3}{4}X_i^2 - \frac{1}{4}X_i^3$
- $X_d \longrightarrow 0$

- $C = \left(I + \frac{X_0}{2}\right)\left(I + \frac{X_1}{2}\right) \dots \left(I + \frac{X_d}{2}\right)$
- $CC^T = (I - X)^{-1}$

- Convergence: ~~$d = \log\log(\kappa/\epsilon)$~~ ??

Chain Convergence



- $d = \log(\kappa/\epsilon)$ enough

Chain Accuracy

$$(I - X_0)^{-1} = \left(I + \tfrac{1}{2}X_0\right)\left(I - \tfrac{3}{4}X_0^2 - \tfrac{1}{4}X_0^3\right)^{-1}\left(I + \tfrac{1}{2}X_0\right)$$

$$\approx_{\epsilon_1} \left(I + \tfrac{1}{2}X_0\right)(I - X_1)^{-1}\left(I + \tfrac{1}{2}X_0\right)$$

$$= \left(I + \tfrac{1}{2}X_0\right)\left(I + \tfrac{1}{2}X_1\right)\left(I - \tfrac{3}{4}X_1^2 - \tfrac{1}{4}X_1^3\right)^{-1}\left(I + \tfrac{1}{2}X_1\right)\left(I + \tfrac{1}{2}X_0\right)$$

$$\approx_{\epsilon_2} \left(I + \tfrac{1}{2}X_0\right)\left(I + \tfrac{1}{2}X_1\right)(I - X_2)^{-1}\left(I + \tfrac{1}{2}X_1\right)\left(I + \tfrac{1}{2}X_0\right)$$

- Overall error $\sum_{i=1}^{d} \epsilon_i$, set $\epsilon_i = \epsilon/d$

Computing $(I - X)^p$

Computing *p*th Power

- Fundamental problem in numerical analysis

- Nick Higham's quoted an email from a power company

> *I have an Excel spreadsheet containing the transition matrix of how a company's [Standard & Poor's] credit rating changes from one year to the next. I'd like to be working in eighths of a year, so the aim is to find* ***the eighth root of the matrix****.*

Our Result

- $(I - X)^p = \left(I + \frac{1}{2}X\right)^{-p}\left(I - \frac{3}{4}X^2 - \frac{1}{4}X^3\right)^p\left(I + \frac{1}{2}X\right)^{-p}$
- How to evaluate $\left(I + \frac{1}{2}X\right)^{-p}$?
  - Taylor expansion!
  - $\left(I + \frac{X}{2}\right)^{-1/2} = I + \frac{1}{2}\left(\frac{X}{2}\right) - \frac{3}{8}\left(\frac{X}{2}\right)^2 + \frac{5}{16}\left(\frac{X}{2}\right)^3 + \cdots$
  - $O(\log \kappa)$ terms suffice

Summary

- Newton's Method + Spectral Sparsification

- First nearly linear work and polylog depth algorithm for
  - Gaussian Sampling
  - Computing $(I - X)^p$

- Dense matrix ≈ product of sparse matrices