Optimization in Machine Learning

Daniel L Boley
University of Minnesota

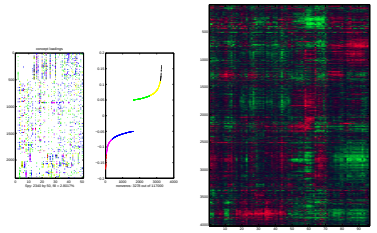
How Convex Optimization plays a big role in Big Data.

NSF Grant 1319749

Discovery Problems

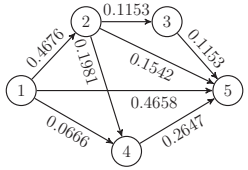
- Many traditional pattern discovery problems: extract hidden patterns in data, by finding an approximate "low-complexity" representation.
  - Text documents (news, laws, WWW documents).
  - Gene expression profiles
  - Attributes for individual people, transactions, locations, ecosystems, ...
  - Images

tabular



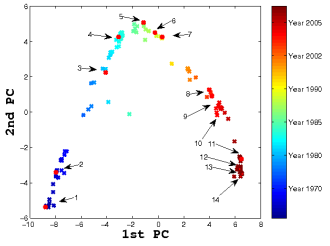
  - Gene-gene or protein-protein interaction networks
  - WWW connectivity graph
  - Computer inter-connect in Internet
  - People-people affinities in Social Media

graph



- Datasets are large, subject to noise and sampling bias.

- Goals include seeking underlying signal through the noise.
- Many bigger examples to be seen later in this symposium.



Data with Special Structure

- Extract properties of Networks - Graph Properties
  - represent connectivity between entities with links.
  - partitioning: optimize a "cut" objective function.
  - graphical models: Signal at each node depends only on neighbors:
    want to recover [unknown] connections from observations.
  - methods used: spectral methods, random walk models, optimization.
    Many of these methods are approximate relaxations of a hard combinatorial problem to a tractible optimization problem.
- Extract Low Complexity Patterns via Sparse Representation
  - seek to represent each data sample as a combination of a few components out of a given dictionary.
  - discover low complexity representation of data.
  - the components found can be a reduced set of features, or identify the general class of a data sample.
  - alternative: seek to eliminate noise by projecting observed data onto a smaller space represented by a sparse model.
  - traditional dimensionality reduction can yield hard-to-interpret components.
  - sparse model can often yield a classifier.
- Most (but not all) of this talk is on the latter.
- Show [Convex] Optimization plays central supporting role in Big Data.

Outline

1. Introduction – Early Ideas
2. Sparsity via Convex Relaxation
3. Variations on Sparsity Model
4. Convex Optimization: First Order Methods
5. Matrix-based Linear Convergence Bound (my work)
6. Conclusions

Early ideas: Basis Pursuit, Compressive Sensing

- sparse solutions: finding which components to select: expensive combinatorial problem.
- Observation: Can often recover sparse solutions using a convex relaxation.
- Empirical Observations in Geophysics
  - Claerbout & Muir 1973.: informal observation that $\ell 1$ regularization works.
  - Taylor, Banks, McCoy 1979. Specific experiments and explicit LP algorithm. (using $\ell 1$ error + $\ell 1$ penalty).
- Formal analysis
  - Chen, Donoho, Saunders 1998. Comparison with alternative greedy methods.
  - Tropp 2006. Theoretical recovery guarantees.
- Exploded into large literature. Here we will see only highlights.

Example: recover spike train

[Taylor, Banks, and McCoy, 1979]

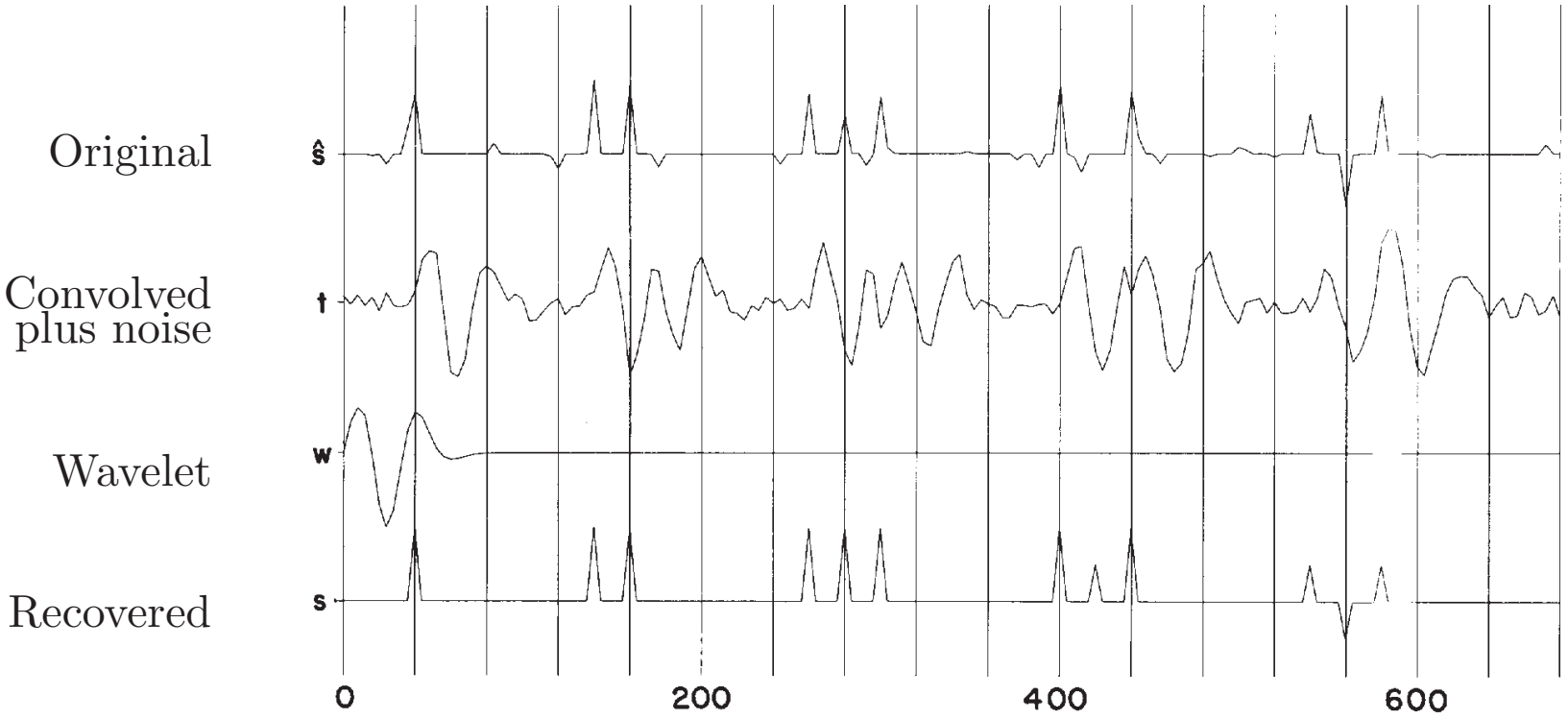
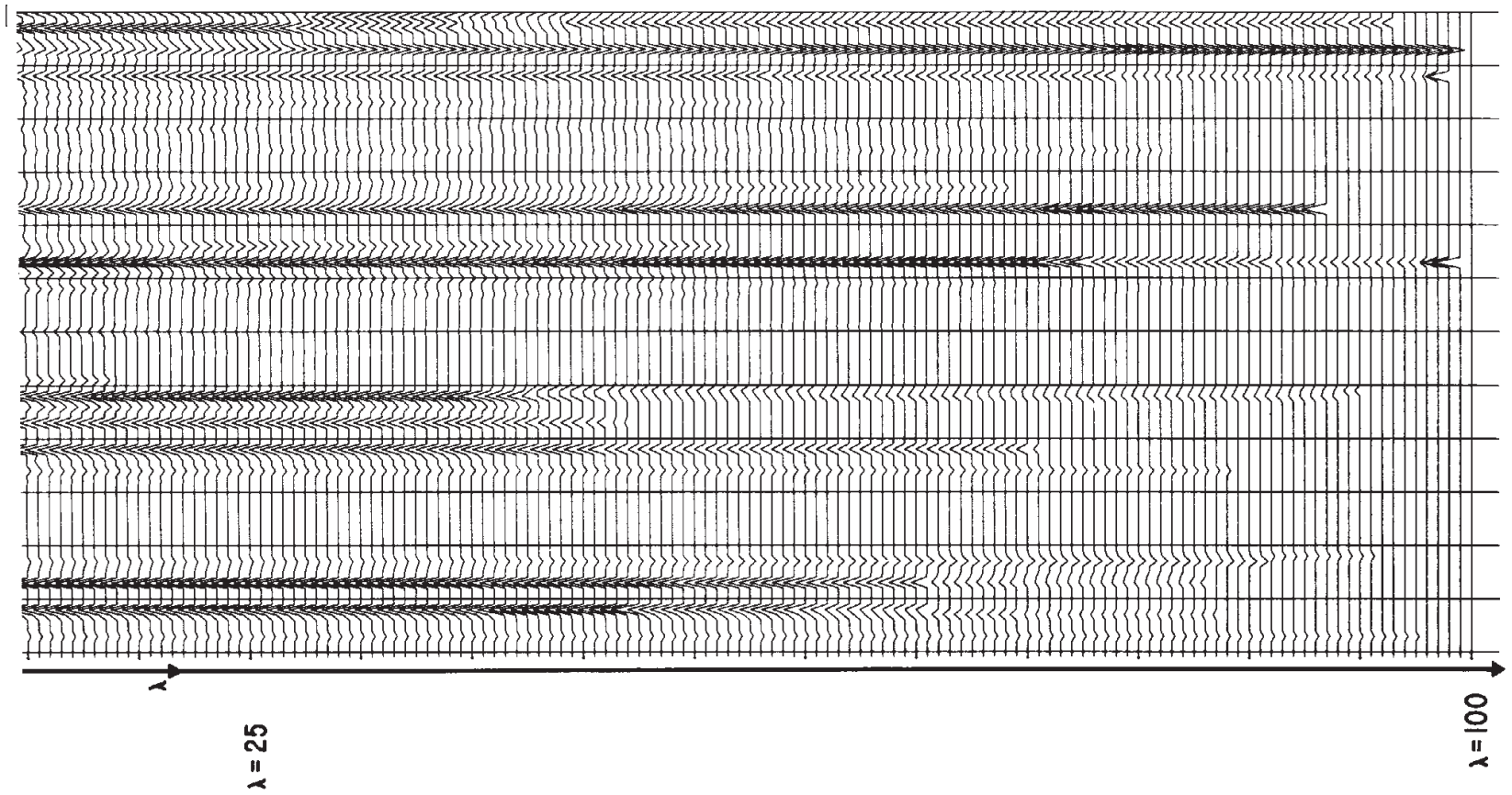


FIG. 2. Synthetic spike train extraction. This example shows the assumed spike train **s** and wavelet **w** convolved and sufficient random noise **n** added so that the trace $\mathbf{t} = \mathbf{s} * \mathbf{w} + \mathbf{n}$ has a signal-to-noise ratio of 4. The extracted spike train $\hat{\mathbf{s}}$ is for the case $\mathbf{l} = 25$.

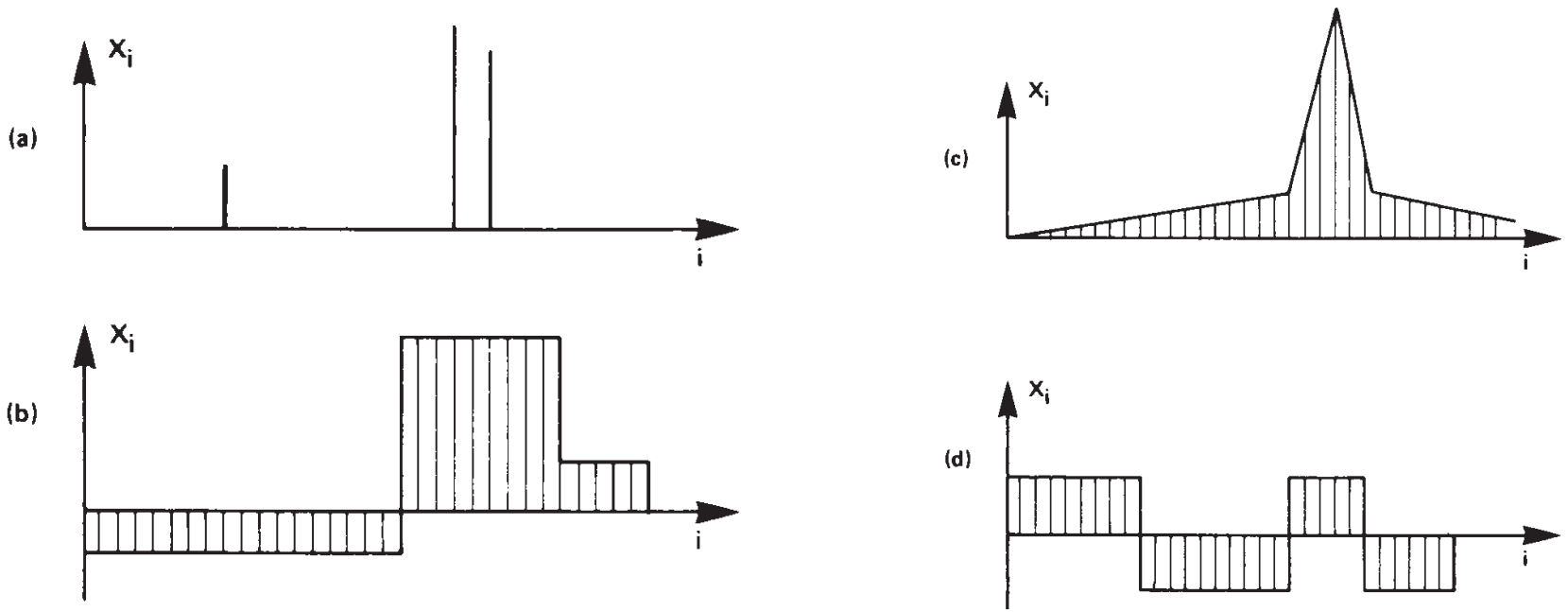
Example: vary regularization

under regularized $\Longleftarrow$ ..... $\Longrightarrow$ over regularized



Generalize: Variations on Total Variation

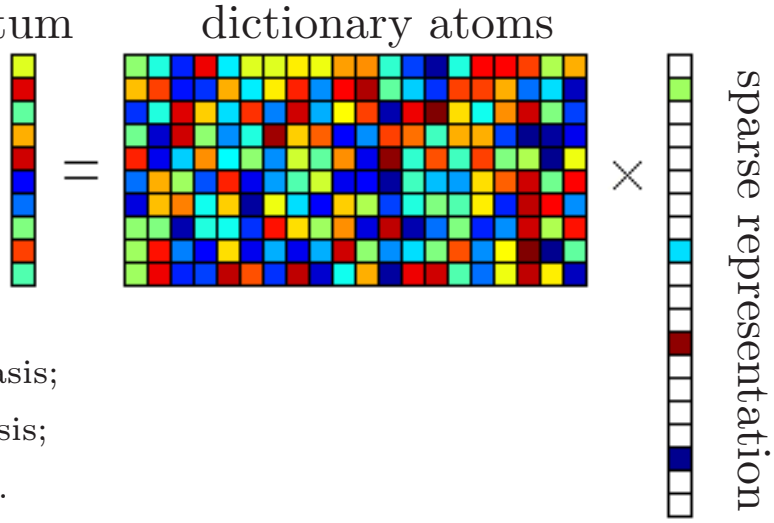
Min $\ell 1$ norm (segment) [Claerbout and Muir, 1973] min $\ell 1$ of 2nd diff's (min variation)



Min $\ell 1$ norm of 1st diff's

min $\ell\infty$ norm (threshold)

Constructing Sparse Basis



- Matching Pursuit: [Mallat and Zhang, 1993]
  - Greedy algorithm: try every column not already in your basis;
  - evaluate quality of new column if it were added to your basis;
  - add "best" column to your basis, and repeat until satisfied.
- Basis Pursuit [Chen, Donoho, and Saunders, 2001]
  - Minimize $\lambda\|\mathbf{x}\|_0$ s.t. $A\mathbf{x} = \mathbf{b}$, or softened to: $\|\mathbf{b} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0$.
  - Difficulty: this is a NP-hard combinatorial problem.
  - Relax to $\lambda\|\mathbf{x}\|_1$ s.t. $A\mathbf{x} = \mathbf{b}$, or softened to $\|\mathbf{b} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$.
  - Relaxed problem is convex, so solvable more efficiently.
  - LASSO, LARS: Solve soft problem for all λ fast [Tibshirani, 1996].
- Non-linear Problem
  - Use Newton's method: inner loop $\equiv$ LASSO Problem.

Convex Relaxation $\Longrightarrow$ LASSO

- Variations: Basis Pursuit, Compressed Sensing, "small error + sparse".
- Add penalty for number of nonzeros with weight λ:

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_0.$$

- Relax hard combinatorial problem into easier convex optimization problem.

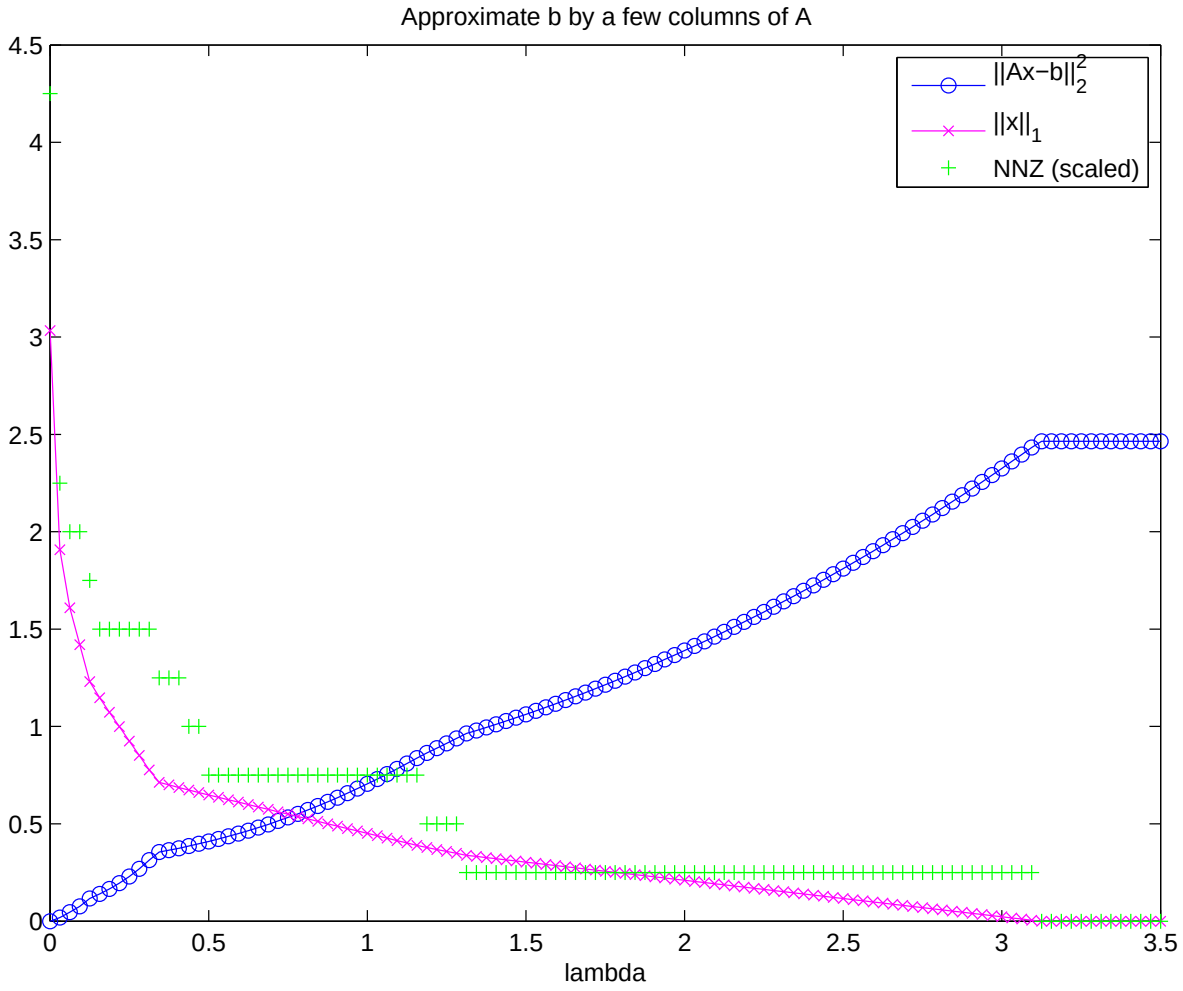
$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1,$$

- or convert to constrained problem:

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_1 \leq \texttt{tol}.$$

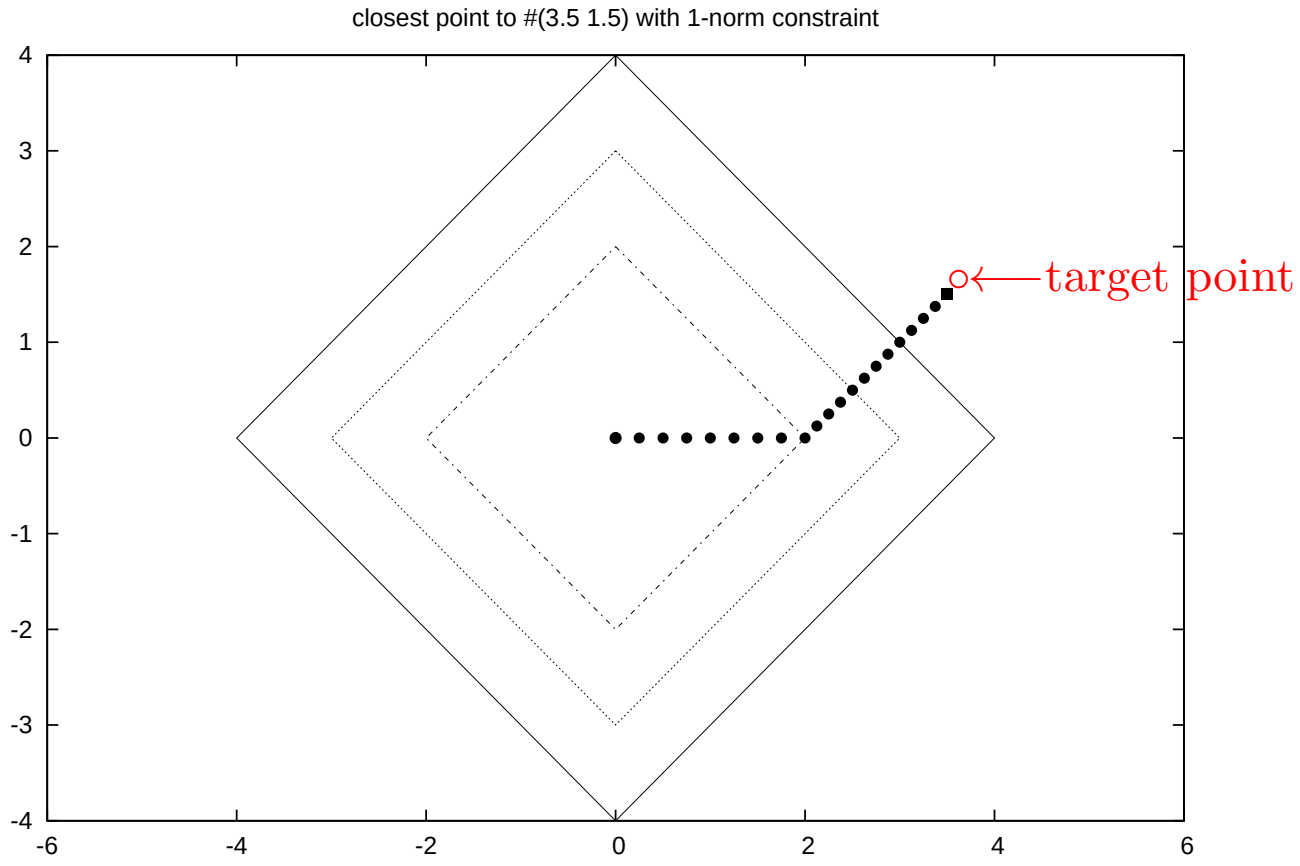
- Vary parameter λ or `tol`, to explore the trade-off between "small error" and "sparse".

Example: 17 signals with 10 time points



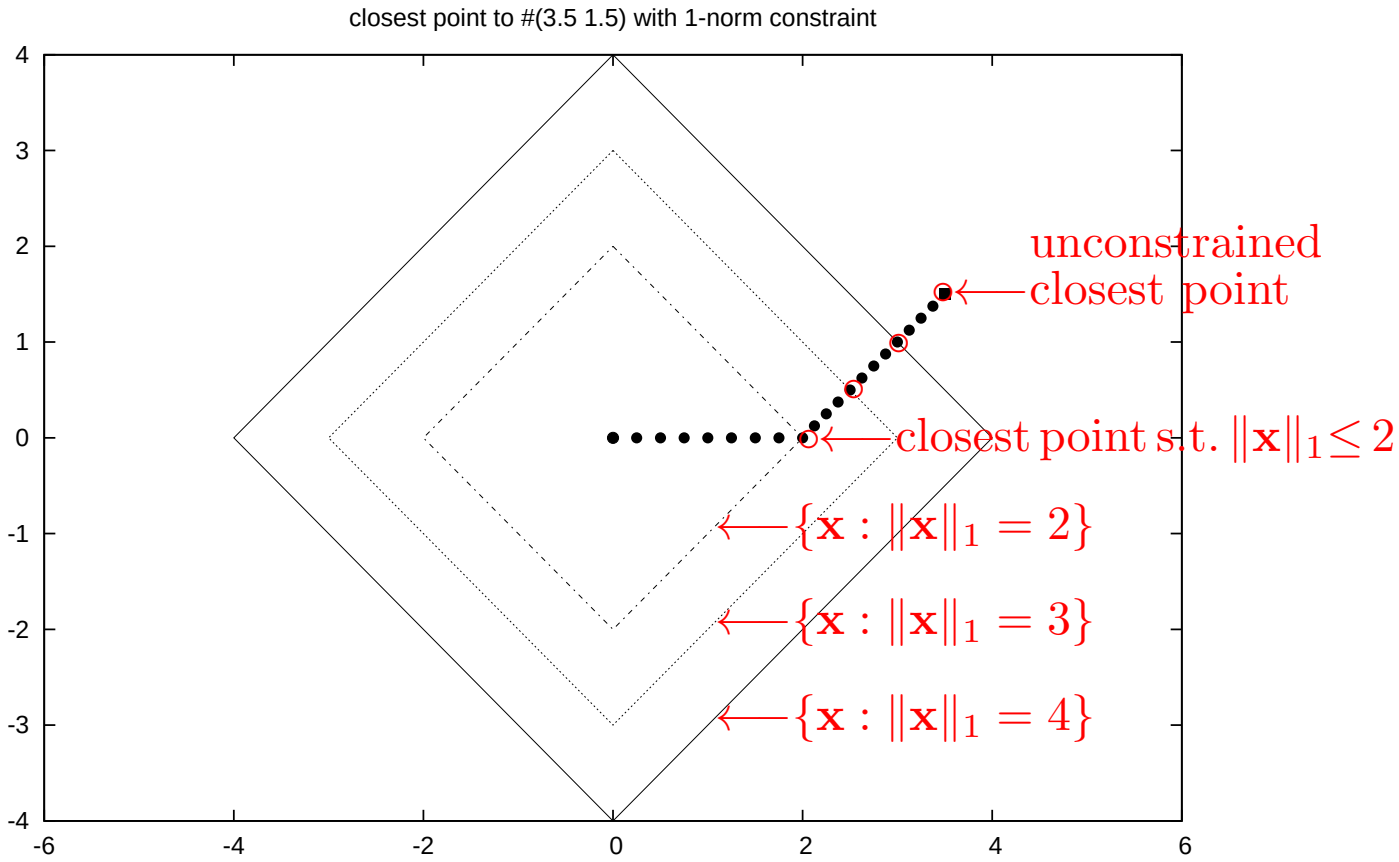
- As λ grows, the error grows, fill (#non-zeros) shrinks.
- Can follow purple line for all λ fast.

Motivation: find closest sparse point



- Find closest point to target . . . subject to $\ell 1$ norm constraint.

Motivation: find closest sparse point



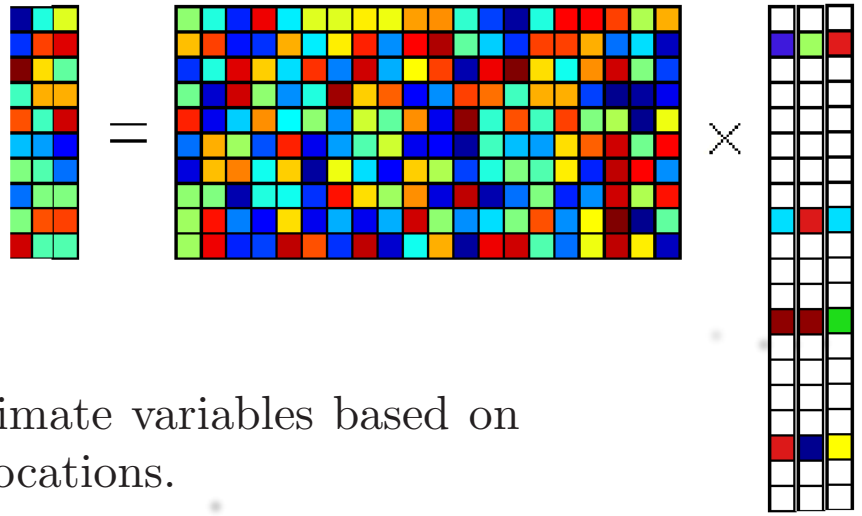
- Tighten limit on $\|\mathbf{x}\|_1 \Longrightarrow$ drive the coordinates toward zero.
- As soon as one coordinate reaches zero, it is removed, and the remaining coordinates are driven to zero.
- Shrinkage operator.

Outline

1. Introduction – Early Ideas
2. Sparsity via Convex Relaxation
3. Variations on Sparsity Model
4. Convex Optimization: First Order Methods
5. Matrix-based Linear Convergence Bound (my work)
6. Conclusions

Group Lasso

- Use a group norm penalty to fit many samples at once.
- Example: $\min_X \|AX - B\|_F^2 + \lambda \|X\|_{1,2}$,
  where $\|X\|_{1,2} = \sum_i \|X_{i,:}\|_2$ (1-norm of row 2-norms).
- Goal, fit several columns of B using as few columns of A as possible for the entire fit.
- Contrast with ordinary LASSO, where we minimize the number of columns of A separately for each column of B. [Yuan and Lin, 2007; Friedman, Hastie, and Tibshirani, 2010]

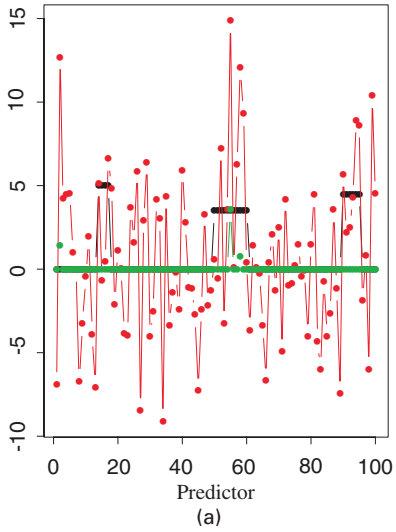


- Example: want to predict multiple climate variables based on climate variables at small number of locations.

Fused Lasso

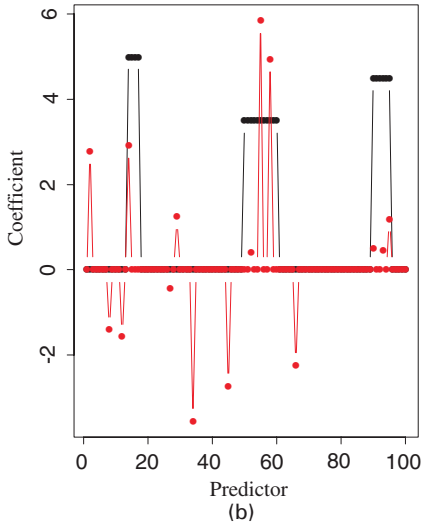
- Like LASSO, but impose also penalty on differences with neighbors.
- Balance error, sparsity, variations. [Tibshirani, Saunders, Rosset, Zhu, and Knight, 2005]

original
regression
thresholded



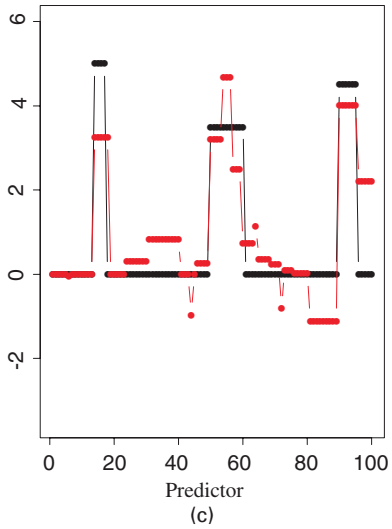
too noisy

original
LASSO



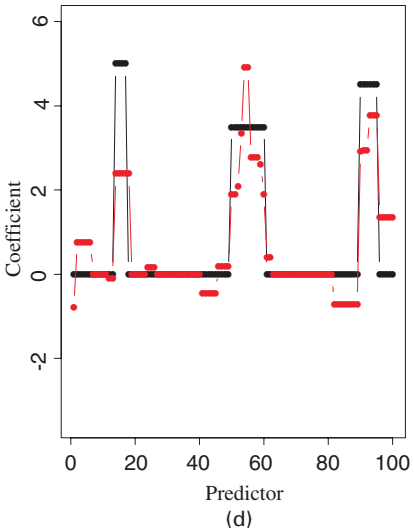
spikey

original
fusion: pure
differences



overshoot

original
fused:
balanced



good compromise

- Use neighbors in 2 dimensions to segment images.

Sparse Inverse Covariance Recovery

- Gaussian Markov Random Field:
  - an undirected graph with random variables on each vertex;
  - the probability distribution of a random variable depends only on the adjacent random variables;
  - 2 non-adjacent random variables are independent conditioned on all other variables;
  - the overall probability distribution is Gaussian with a stationary mean & covariance matrix;
  - Zero entries in the precision matrix (inverse of covariance matrix) mark non-adjacent vertices.
- Goal is to recover a sparse precision matrix closely matching a given sample covariance matrix S.
- Max Likelihood Estimate of inverse covariance is [Hsieh, Sustik, Dhillon, and Ravikumar, 2012]

$$\min_X \underbrace{-\log\det X + \operatorname{tr}(SX)}_{\substack{\text{KL divergence between} \\ \text{two centered gaussians}}} + \underbrace{\lambda\|\operatorname{vec}(X)\|_1}_{\substack{\text{regularization} \\ \text{term}}}$$

- E.G. reveal far-flung climate dependencies.

Matrix Completion

- Arises in: recommender systems, Netflix prize, fill-in missing data.
- **Given** partially filled matrix M, with set Ω of indices of filled entries.
  Find $X = \text{argmin}_X \text{rank}\, X$ s.t. $X_{ij} = M_{ij}$ for $(i,j) \in \Omega$.
- Convex relaxation [Candés and Recht, 2008]:
  $\min_X \|X\|_*$ s.t. $X_{ij} = M_{ij}$ for $(i,j) \in \Omega$
  where $\|X\|_* =$ sum of singular values of $X = \|(\sigma_1(X); \sigma_2(x); \ldots)\|_1$.
- Alternative formulation (often more efficient, less memory intensive)
  - $\min_X \|UV^T - M\|_\Omega^2$ s.t. U, V each has k columns.
  - where $\|X\|_\Omega$ denotes the F-norm summed over only indices in Ω.
  - Biconvex: convex in U, V individually: use alternating least squares.

    [Jain, Netrapalli, and Sanghavi, 2013] [NSrebro, Rennie, and Jaakkola, 2005]

Computer Vision

Covariance Descriptor: eliminate brightness variation.

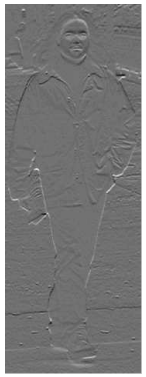
Raw Image

Image

first derivatives

x-grad

grad-mag



y-grad

grad-dir

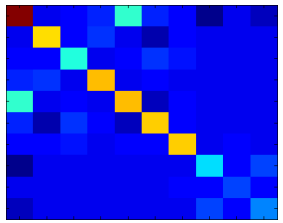
second derivatives

Dxx

Dxy

Dyy

pixel by pixel descriptor



Covariance descriptor

Covariance of pixel-wise features on 7×7 patch

Want to reduce covariance descriptors to linear combination of "small" dictionary of descriptors.

Optimization Setup for Covariances

Want to use small dictionary of descriptors to represent all descriptors in an image [Sivalingam, Boley, Morellas, and Papanikolopoulos, 2010, 2011].

- S = a raw covariance matrix,
  $\mathbf{x}$ = vector of unknown coefficients.
  $\mathcal{A} = (A_1, A_2, \ldots, A_k)$ = collection of dictionary atoms.
  $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ = vector of unknown coefficients.

- Goal: Approximate $S \approx A_1 x_1 + \cdots + A_k x_k = \mathcal{A} \cdot \mathbf{x}$.

- Use "logdet" divergence as measure of discrepancy:
  $D_{\mathrm{ld}}(\mathcal{A} \cdot \mathbf{x}, S) = tr((\mathcal{A} \cdot \mathbf{x})S^{-1}) - \log\det((\mathcal{A} \cdot \mathbf{x})S^{-1}) - n.$

- Logdet divergence measures relative entropy between two different zero-mean multivariate Gaussians.

Optimization Problem for Covariances

[Sivalingam, Boley, Morellas, and Papanikolopoulos, 2010, 2011]

- Leads to optimization problem

$$\min_{\mathbf{x}} \quad \underbrace{\sum_i x_i tr(A_i) - \log\det\left[\sum_i x_i A_i\right]}_{\text{Dist}(\mathcal{A}\cdot\mathbf{x},S)} + \underbrace{\lambda \sum_i x_i}_{\text{sparsity}}$$

$$\begin{aligned} \text{s.t.} \quad & \mathbf{x} \geq 0 \\ & \textstyle\sum_i x_i A_i \succeq 0 \quad \text{(positive semi-definite)} \\ & \textstyle\sum_i x_i A_i \preceq S \quad \text{(residual positive semi-def.)} \end{aligned}$$

- This is in a standard form for a MaxDet problem.
- The sparsity term is a relaxation of true desired penalty: # nonzeros in $\mathbf{x}$.
- Convex problem solvable by many solvers.

Semidefinite Programming

- Like an LP, but with semidefinite constraints:

$$\min_{\mathbf{x}} \mathbf{c}^T\mathbf{x} \text{ s.t. } A_0 + \sum_i x_i A_i \succeq 0,$$

  for A_i given symmetric matrices.

- Convex, hence amenable to "efficient" convex methods.

- Many problems can be expressed in this form:
  LPs, QCQPs, Minimize max eigenvalue/singular value.

  - e.g. $(A\mathbf{x}+\mathbf{b})^T(A\mathbf{x}+\mathbf{b}) - \mathbf{c}^T\mathbf{x} - d \leq 0 \Longleftrightarrow \begin{bmatrix} I & A\mathbf{x}+\mathbf{b} \\ (A\mathbf{x}+\mathbf{b})^T & \mathbf{c}^T\mathbf{x}+d \end{bmatrix} \succeq 0.$
    [LVandenberghe and Boyd, 1996]

Additional Formulations

- Dantzig Selector $\min_{\mathbf{x}} \|\mathbf{x}\|_1$ s.t. $\|\nabla\phi(\mathbf{x})\|_\infty = \|A^T(A\mathbf{x}-\mathbf{b})\|_\infty \leq \text{tol}$ [Candès and Tao, 2005]
- Ky-Fan k-norm: sum of top k singular values.
- Schatten p norm: ordinary p norm of vector of singular values.
- Atomic norm: if $\mathcal{A}$ is a bounded [possibly finite or countable] set of vectors s.t. $\mathbf{a} \in \mathcal{A} \Longrightarrow -\mathbf{a} \in \mathcal{A}$, then $\|\mathbf{v}\| = \inf t : \mathbf{v}/t \in \text{conv\_hull}\,\mathcal{A}$ is a norm.
  - Many choices for $\mathcal{A}$: columns of I, all unit-norm rank-one matrices or tensors, ...
  - general theory of recovery guarantees from sample data based on *gaussian width* of unit ball for given norm. [Chandrasekaran, Recht, Parrilo, and Willsky, 2012]
- Clustering with Soft Must-link & Cannot-link Constraints
  ℓ_1-norm penalty on contraint violations. [Kawale and Boley, 2013].
- Sparse Inverse Covariance Recovery [Hsieh, Sustik, Dhillon, and Ravikumar, 2012] (Non-linear: Newton method w/ Armijo line search. CD on inner LASSO problem).

Outline

1. Introduction – Early Ideas
2. Sparsity via Convex Relaxation
3. Variations on Sparsity Model
4. Convex Optimization: First Order Methods
5. Matrix-based Linear Convergence Bound (my work)
6. Conclusions

First Order Methods

- First order method: often only tractible methods on large problems.
- ADMM: Alternating Direction Method of Multipliers
  - $\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x}) + \mathbb{I}(\mathbf{z})$ s.t. $\mathbf{x} = \mathbf{z}$
  - minimize *wrt* $\mathbf{x}$, then *wrt* $\mathbf{z}$, then gradient ascept in dual.
  - simplest first order method applicable also to constrained problems.
- F/ISTA (Fast Shrinkage/Threshold Alg.: special case of proximal methods)
  - $\min_{\mathbf{x}} f(\mathbf{x}) + r(\mathbf{x})$: f = objective, r = regularizer.
  - If $r = 0$, reduces to steepest descent.
- CD: Coordinate Descent.
  - Applicable to problems like F/ISTA problem.
  - If $r = 0$, reduces to Gauss-Seidel
- (Hard Thresholding)
- (Alternating Relaxation)
- (Frank-Wolfe, using the dual norm)

Three Convergence Regimes

- Analyse on model problem(s).
- When the zero–non-zero structure of the iterate remains the same from one iteration to the next, the mapping from one iterate to the next is a stationary linear operator, denoted $\mathbf{M}_{\text{aug}}^{[k]}$.
- Largest eigenvalue of operator is 1. Convergence behavior depends on the remaining eigenvalues. The possibilities are [Boley, 2013; Tao, Boley, and Zhang, 2015]:

[A] The eigenvalue $\lambda = 1$ of $\mathbf{M}_{\text{aug}}^{[k]}$ is simple. We get [local] linear convergence.

[B] The eigenvalue $\lambda = 1$ of $\mathbf{M}_{\text{aug}}^{[k]}$ is double, but only with one eigenvector. We get "constant-step" convergence.

[C] The eigenvalue $\lambda = 1$ of $\mathbf{M}_{\text{aug}}^{[k]}$ is double with a complete set of eigenvectors. We get linear convergence to a solution, possibly non-unique.

- Locally near optimum: smooth problems like model problems.

Toy Example - ADMM on an Linear Program

ADMM is also applicable in presence of constraints, like an LP .
LP not strictly convex, so good example to exhibit the regimes.

Simple resource allocation model:
- v_1 = rate of cheap process (e.g. fermentation),
- v_2 = rate of costly process (e.g. respiration).

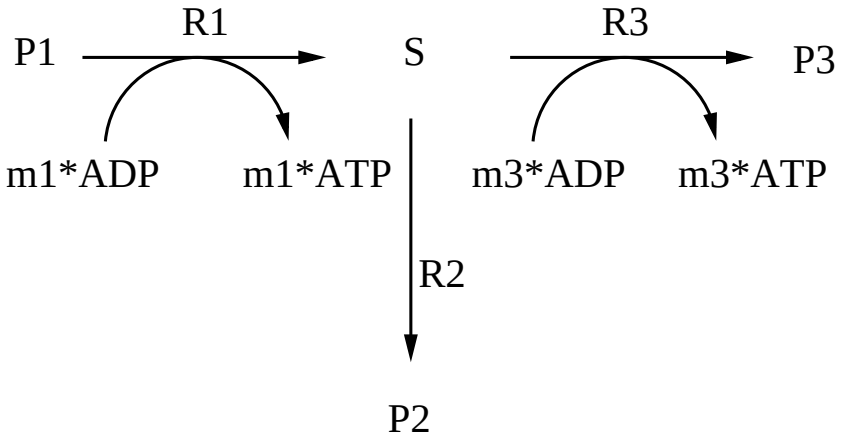
[Schuster, Boley, Möller, Stark, and Kaleta, 2015] Warburg effect

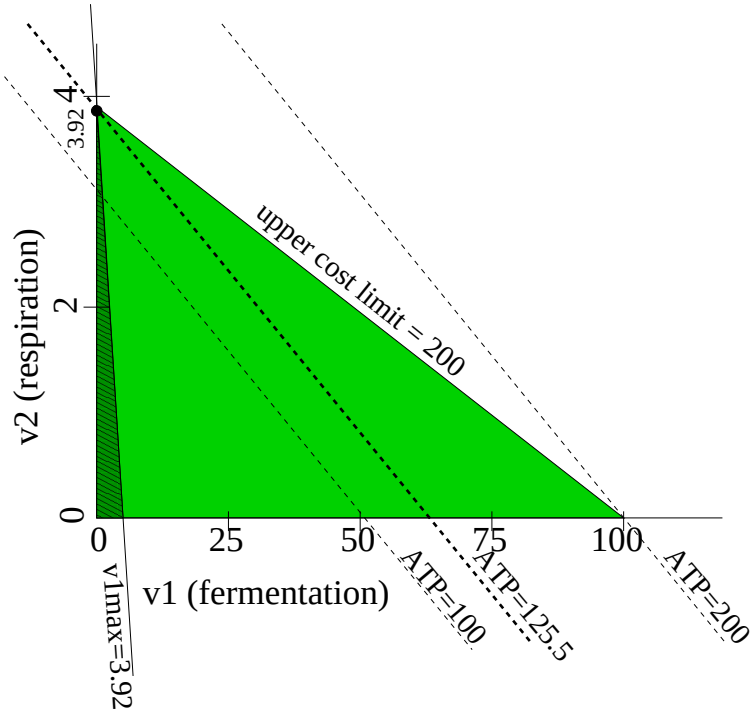
$$
\begin{array}{lllll}
\text{maximize}_{\mathbf{v}} & +2v_1 + 30v_2 & & & (\textit{desired end product production}) \\
\text{subject to} & v_1 \quad + v_2 & & \leq\ v_{0,max} & (\textit{limit on raw material}) \\
& 2v_1 \quad + 50v_2 & & \leq\ 200 & (\textit{internal capacity limit}) \\
& v_1 \geq 0 & v_2 \geq 0 & & (\textit{irreversibility of reactions})
\end{array}
$$

Put into standard form:

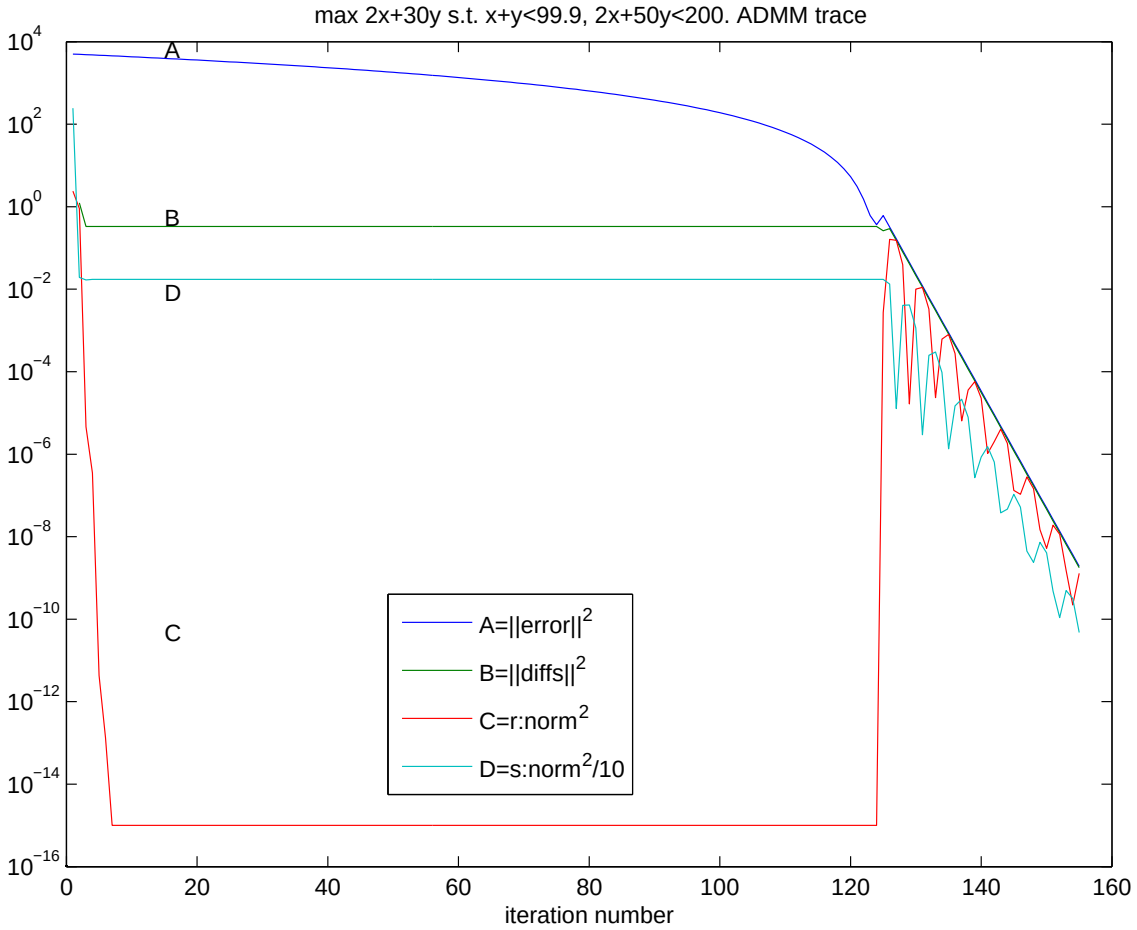
$$
\begin{array}{lllll}
\text{minimize}_{\mathbf{v}} & -2v_1 - 30v_2 & & & (\textit{desired end product production}) \\
\text{subject to} & v_1 \quad + v_2 & + \ v_3 = & v_{0,max} & (\textit{limit on raw material}) \\
& 2v_1 \quad + 50v_2 & + \ v_4 = & 200 & (\textit{internal capacity limit}) \\
& v_1 \geq 0 & v_2 \geq 0 & & (\textit{irreversibility of reactions}) \\
& v_3 \geq 0 & v_4 \geq 0 & & (\textit{slack variables})
\end{array}
\tag{1}
$$

Resource Allocation LP



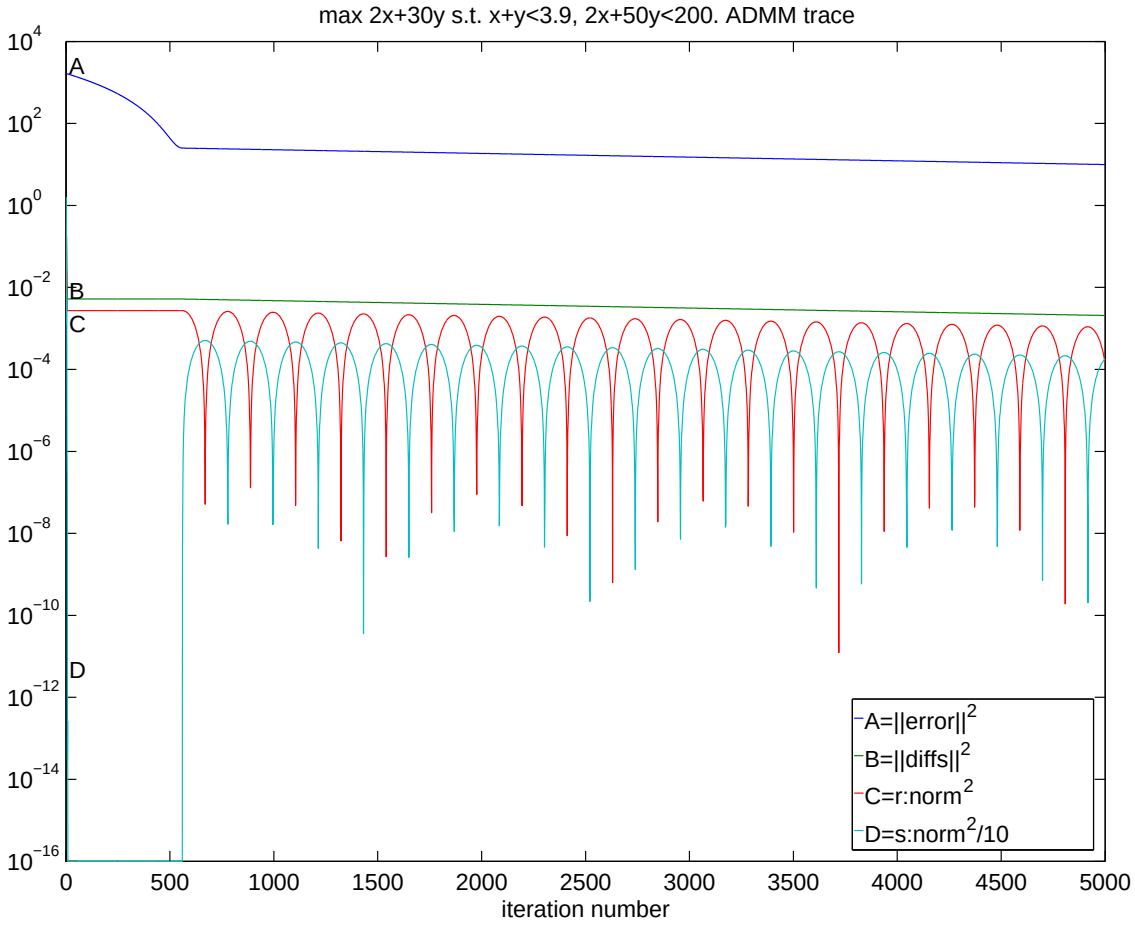


Typical Convergence Behavior $v_{0,max} = 99.9$



ADMM on Example 1: typical behavior. Curves: A: error $\|(\mathbf{z}^{[k]} - \mathbf{u}^{[k]}) - (\mathbf{z}^* - \mathbf{u}^*)\|^2$. B: $\|(\mathbf{z}^{[k]} - \mathbf{u}^{[k]}) - (\mathbf{z}^{[k-1]} - \mathbf{u}^{[k-1]})\|^2$. C: $\|(\mathbf{x}^{[k]} - \mathbf{z}^{[k]})\|^2$. D: $\|(\mathbf{z}^{[k]} - \mathbf{z}^{[k-1]})\|^2/10$ (D is scaled by 1/10 just to separate it from the rest).

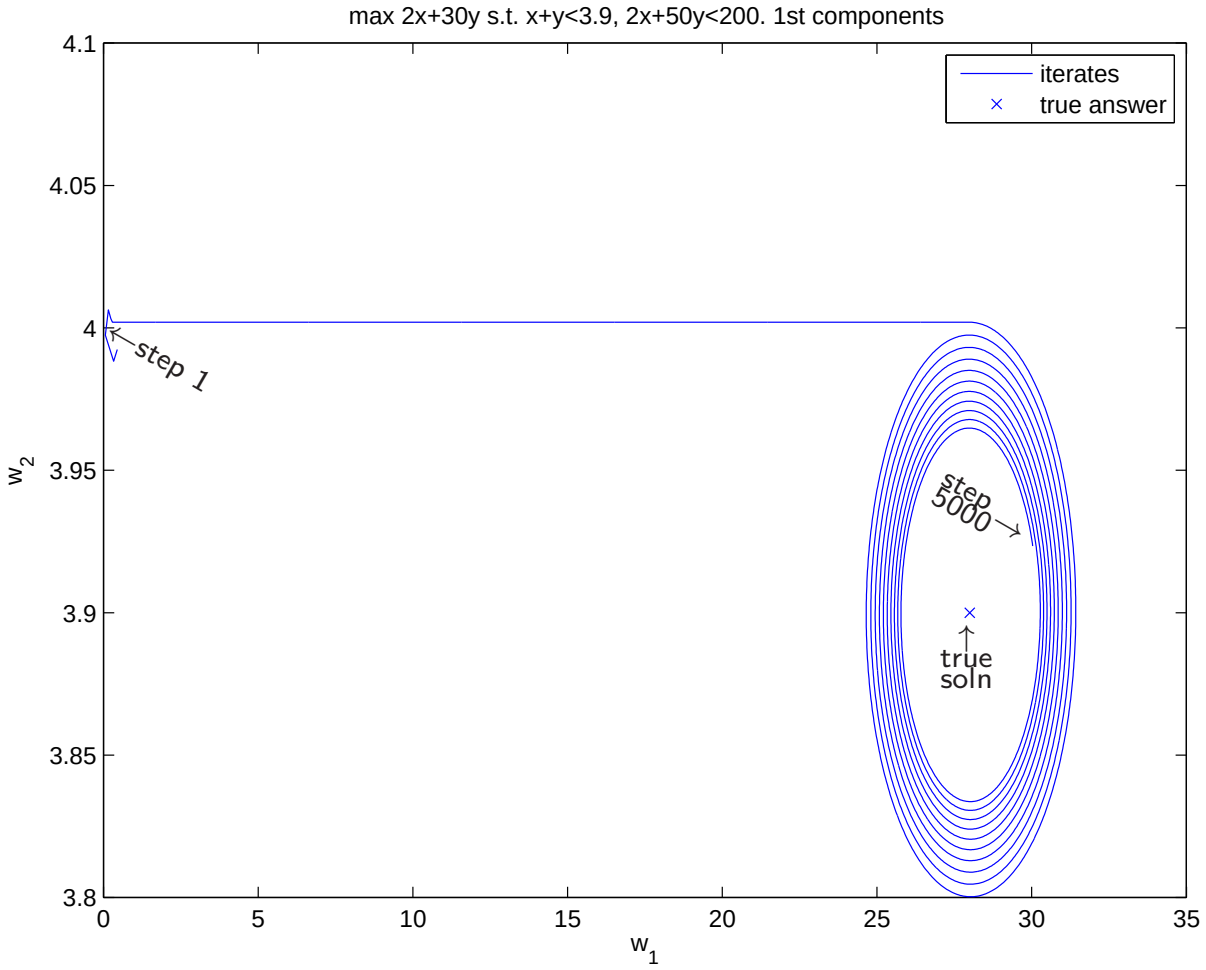
Toy Example with $v_{0,max} = 3.9$



ADMM on Example 2: slow linear convergence.

Second largest eigenvalue $= \sigma(M) = 0.999896$. convergence is very slow: $-1/\log_{10}(\sigma(M)) = 22135$ iterations needed per decimal digit of accuracy.

Convergence Of Modified Toy Example



Convergence behavior of first two components of $\mathbf{w}^{[k]}$ for Example 2, showing the initial straight line behavior (initial regime [b]) leading to the spiral (final regime [a]).

Outline

1. Introduction – Early Ideas
2. Sparsity via Convex Relaxation
3. Variations on Sparsity Model
4. Convex Optimization: First Order Methods
5. Matrix-based Linear Convergence Bound (my work)
6. Conclusions

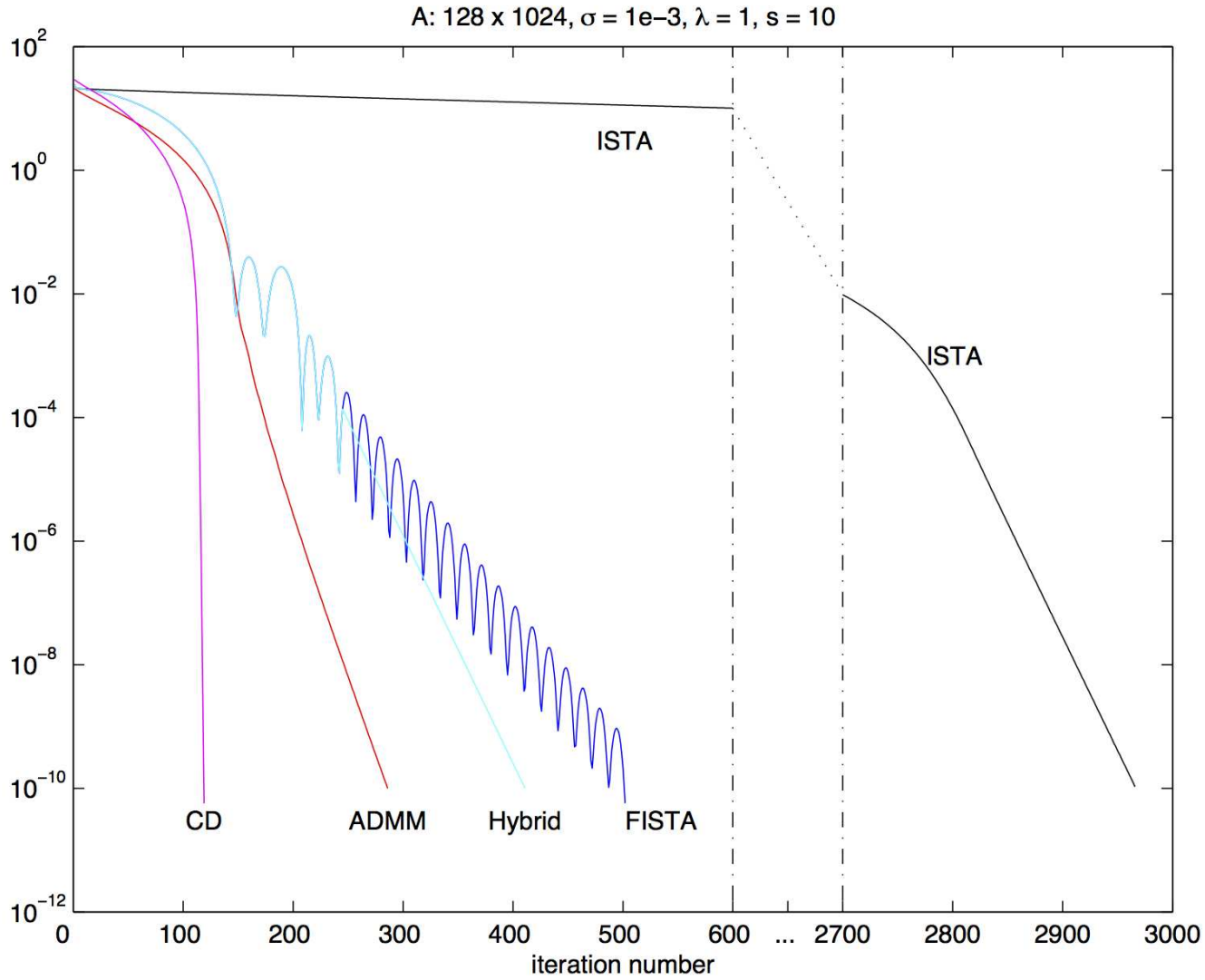
Model Problem

Return to Model LASSO problem: ℓ_1-regularized least squares:

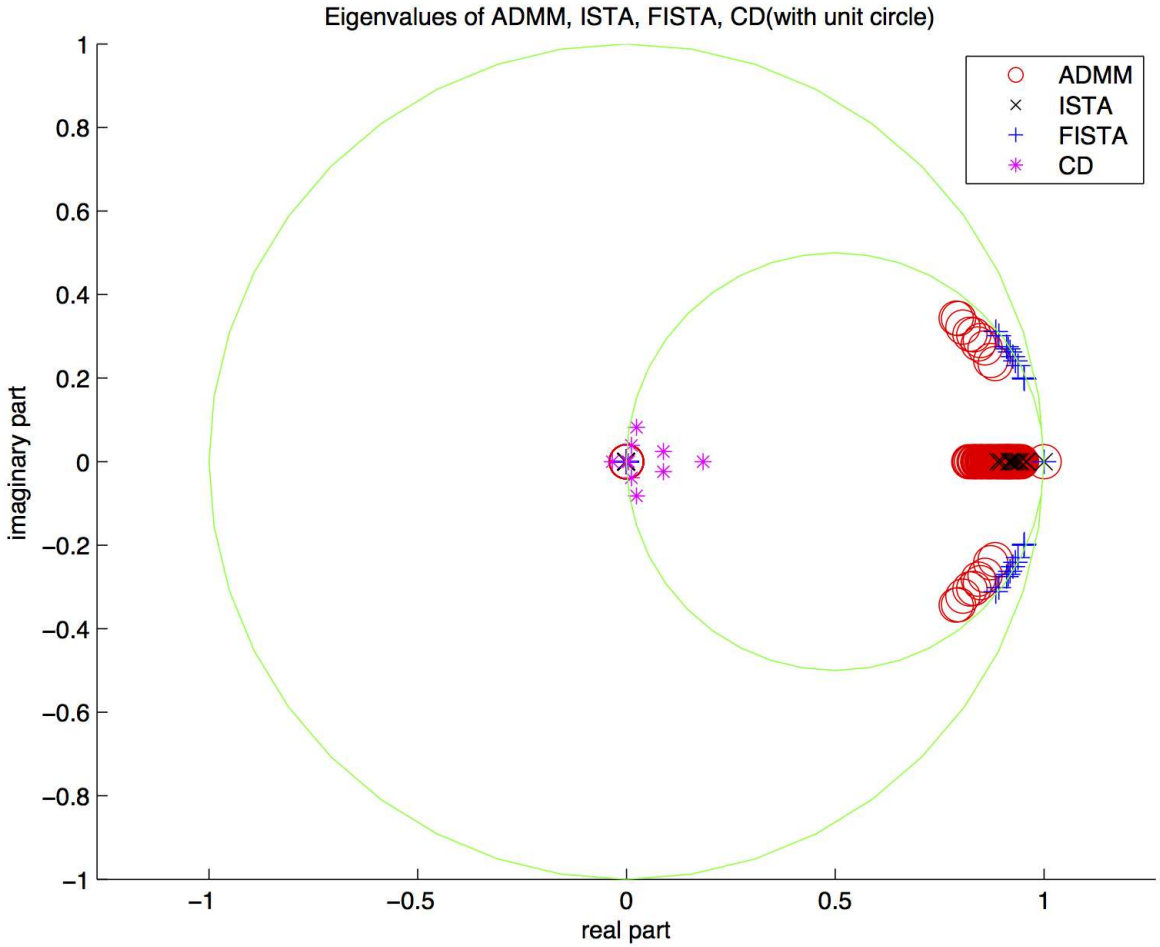
$$\min_{\mathbf{x}\in\mathbb{R}^n} \tfrac{1}{2}\|A\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

- $A \in \mathbb{R}^{m\times n}$ is a short-flat matrix (i.e. $n > m$) with full row rank, $\mathbf{b}$ is a given vector, and λ is a positive scalar.
- General interior point methods do not scale to the large-scale data problems encountered in practice.
- Most popular algorithms proposed include Alternating Direction Method of Multipliers (ADMM), Iterative Shrinkage Thresholding Algorithm (ISTA) and its accelerated version Fast ISTA (FISTA), and the cyclic Coordinate Descent method (CD). All of them has been shown to enjoy the sublinear convergence.

Compare Convergence Behavior

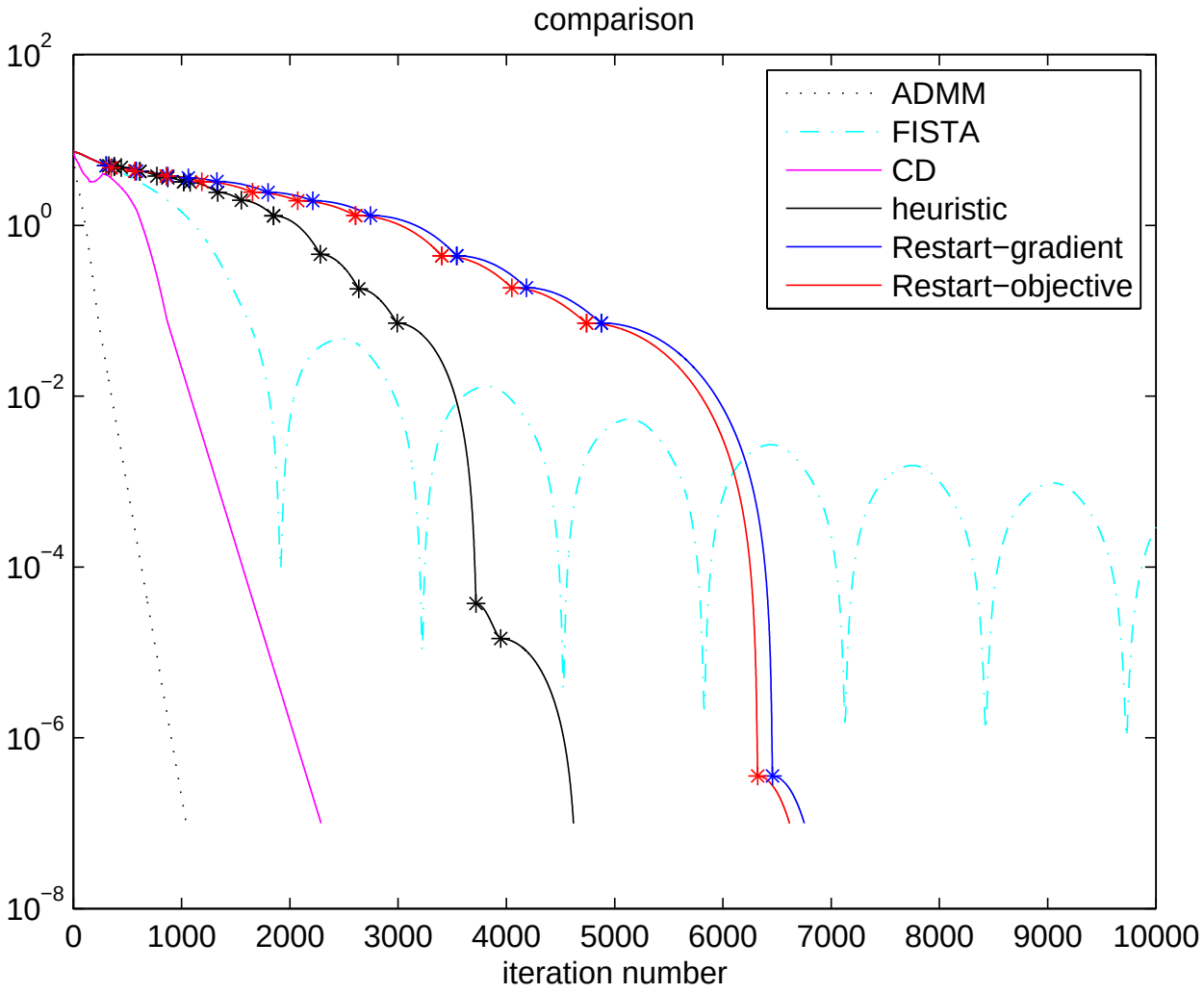


Eigenvalues



Spectrum of M, R, N.

Heuristic Algorithms



Outline

1. Introduction – Early Ideas
2. Sparsity via Convex Relaxation
3. Variations on Sparsity Model
4. Convex Optimization: First Order Methods
5. Matrix-based Linear Convergence Bound (my work)
6. Conclusions

Conclusions

- Convex Optimization: recently discovered tool for many machine learning problems.
- Effective first order methods exist to solve them.
- Convergence Guarantees: still active research area.
  - ADMM, (F)ISTA, CD converge linearly when close enough to optimal solution.
  - FISTA can be slow down compared to ISTA towrds the end.
  - ISTA stagnates during the initial iterations, and FISTA during later iterations.
  - CD $\Leftrightarrow$ Gauss-Seidel iteration, a preconditioned Richardson iteration $\Leftrightarrow$ ISTA.
- Take-Away: Optimization has an essential supporting role in Big Data.

Outline

1. Introduction – Early Ideas
2. Sparsity via Convex Relaxation
3. Variations on Sparsity Model
4. Convex Optimization: First Order Methods
5. Matrix-based Linear Convergence Bound (my work)
6. Conclusions

THANK YOU!

References

K. P. Bennett and C. Campbell. Support vector machines: Hype or hallelujah? In *SIGKDD Explorations*, volume 2 #2. ACM, 2000. URL http://www.acm.org/sigs/sigkdd/explorations/issue2-2/bennett.pdf.

D. Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *SIAM J. Optim.*, 23(4):2183–2207, 2013.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011. doi: doi:10.1561/2200000. http://www.stanford.edu/~boyd/papers/admm/.

E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35:23132351, 2005.

E. J. Candés and B. Recht. Exact matrix completion via convex optimization. *Found. of Comput. Math.*, 9:717–772, 2008.

V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Foundations Comput. Math.*, 12(6):805–849, 2012.

S. Shaobing Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43:129–159, January 2001. ISSN 0036-1445. doi: 10.1137/S003614450037906X. URL http://portal.acm.org/citation.cfm?id=588736.588850.

J. F. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38:826–844, 1973.

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and sparse group lasso. arXiv:1001.0736, 2010.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, October 2010.

Y. Han, Z. Sun, T. Tan, and Y. Hao. Palmprint recognition based on regional rank correlation of directional features. In Massimo Tistarelli and Mark Nixon, editors, *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 587–596. Springer Berlin / Heidelberg, 2009.

B. He and X. Yuan. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. http://www.optimization-online.org/DB_HTML/2012/01/3318.html, 2012.

C. J Hsieh, M. A. Sustik, I. S Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. In *NIPS*, 2012.

P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 665–674, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2029-0. doi: 10.1145/2488608.2488693. URL http://doi.acm.org/10.1145/2488608.2488693.

J. Kawale and D. Boley. Constrained spectral clustering using L1 regularization. In *SIAM Data Mining Conf. SDM 13*, pages 103–111, 2013.

Y. Li, Z. Zhang, and D. Boley. the routing continuum from shortest-path to all-path: A unifying theory. In *The 31st Int'l Conference on Distributed Computing Systems (ICDCS 2011)*, pages 847–856. IEEE, 2011.

J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proc. CACSD Conf.*, Taipei, Taiwan, 2004. URL http: //users.isy.liu.se/johanl/yalmip. http://users.isy.liu.se/johanl/yalmip .

L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, March 1996.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12): 3397 –3415, December 1993. ISSN 1053-587X. doi: 10.1109/78.258082.

N. Srebro, J. Rennie, and T. Jaakkola. Maximum margin matrix factorization. In *NIPS*, volume 17, 2005.

A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In Tomas Pajdla and Jiri Matas, editors, *Computer Vision - ECCV 2004*, volume 3022 of *Lecture Notes in Computer Science*, pages 71–84. Springer Berlin / Heidelberg, 2004.

H. Palaio, C. Maduro, K. Batista, and J. Batista. Ground plane velocity estimation embedding rectification on a particle filter multi-target tracking. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 825 –830, May 2009. doi: 10.1109/ROBOT.2009.5152610.

Y. Pang, Y. Yuan, and X. Li. Effective feature extraction in high-dimensional space. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 38(6):1652 –1656, Dec. 2008. ISSN 1083-4419. doi: 10.1109/TSMCB.2008.927276.

F. Porikli and T. Kocak. Robust license plate detection using covariance descriptor in a neural network framework. In *Video and Signal Based Surveillance, 2006. AVSS '06. IEEE International Conference on*, pages 107 –107, Nov. 2006. doi: 10.1109/AVSS.2006.100.

S. Schuster, D. Boley, P. Möller, H. Stark, and C. Kaleta. Mathematical models for explaining the Warburg effect: a review focussed on ATP and biomass production. *Biochemical Society Transactions*, 43(6):1187–1194, 2015.

X. Shi, W. Fan, and P. S. Yu. Efficient semi-supervised spectral co-clustering with constraints. In *ICDM*, pages 1043–1048, 2010.

R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Tensor sparse coding for region covariances. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *European Conf. on Comp. Vision (ECCV 2010)*, volume 6314 of *LNCS*, pages 722–735. Springer, 2010.

R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Papanikolopoulos. Positive definite dictionary learning for region covariances. In *IEEE Int'l Conf. on Comp. Vision (ICCV 2011)*, pages 1013–1019, 2011.

S. Tao, D. Boley, and S. Zhang. Local linear convergence of ISTA and FISTA on the LASSO problem. *SIAM J Optim.*, 26(1): 313–336, 2015.

H. L. Tay or, S. C. Banks, and J. F. McCoy. Devolution with the ℓ_1 norm. *Geophysics*, 44:39–52, 1979.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused LASSO. *J. R. Statist. Soc. B*, pages 91–108, 2005.

O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, June 2007. doi: 10.1109/CVPR.2007.383197.

O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In Ales Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision ECCV 2006*, volume 3952 of *Lecture Notes in Computer Science*, pages 589–600. Springer Berlin / Heidelberg, 2006.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Royal Statistical Society, Series B*, 68 (1):49–67, 2007.