Sparse Optimization

Stephen Wright

University of Wisconsin-Madison

SIAM-OPT, Darmstadt, May 2011

1. Formulations and Applications
   - ℓ_1 and Sparsity
   - ℓ_1 in Compressed Sensing
   - Applications of ℓ_1
   - Beyond ℓ_1

2. Algorithms
   - Prox-Linear
   - Accelerated First-Order
   - Stochastic Gradient
   - Augmented Lagrangian

Slides: google my name + "Wisconsin" and follow the link.

Sparse Optimization: Motivation

Many applications need structured, approximate solutions of optimization formulations, rather than exact solutions.

- More Useful, More Credible
  - Structured solutions are easier to understand.
  - They correspond better to prior knowledge about the solution.
  - They may be easier to use and actuate.
  - Extract just the *essential* meaning from the data set, not the less important effects.
- Less Data Needed
  - Structured solution lies in lower-dimensional spaces $\Rightarrow$ need to gather / sample less data to capture it.
  - Choose good structure instead of "overfitting" to a particular sample.

The structural requirements have deep implications for how we formulate and solve these problems.

ℓ_1 and Sparsity

A common type of desired structure is sparsity: We would like the approx solution $x \in \mathbb{R}^n$ to have few nonzero components.

A sparse formulation of "$\min_x f(x)$" could be

Find an approximate minimizer $\bar{x} \in \mathbb{R}^n$ of f such that $\|x\|_0 \leq k$,

where $\|x\|_0$ denotes cardinality: the number of nonzeros in x.

Too Hard!

Use of $\|x\|_1$ has long been known to promote sparsity in x. Also,

- Can solve without discrete variables;
- It maintains convexity.

Regularized Formulations with ℓ_1

Weighted form:

$$\min\ f(x) + \tau\|x\|_1,$$

for some parameter $\tau \geq 0$. Generally, larger $\tau \Rightarrow$ sparser x.

ℓ_1**-constrained form** (variable selection):

$$\min f(x)\ \text{ subject to }\ \|x\|_1 \leq T,$$

for some $T > 0$. Generally, smaller $T \Rightarrow$ sparser x.

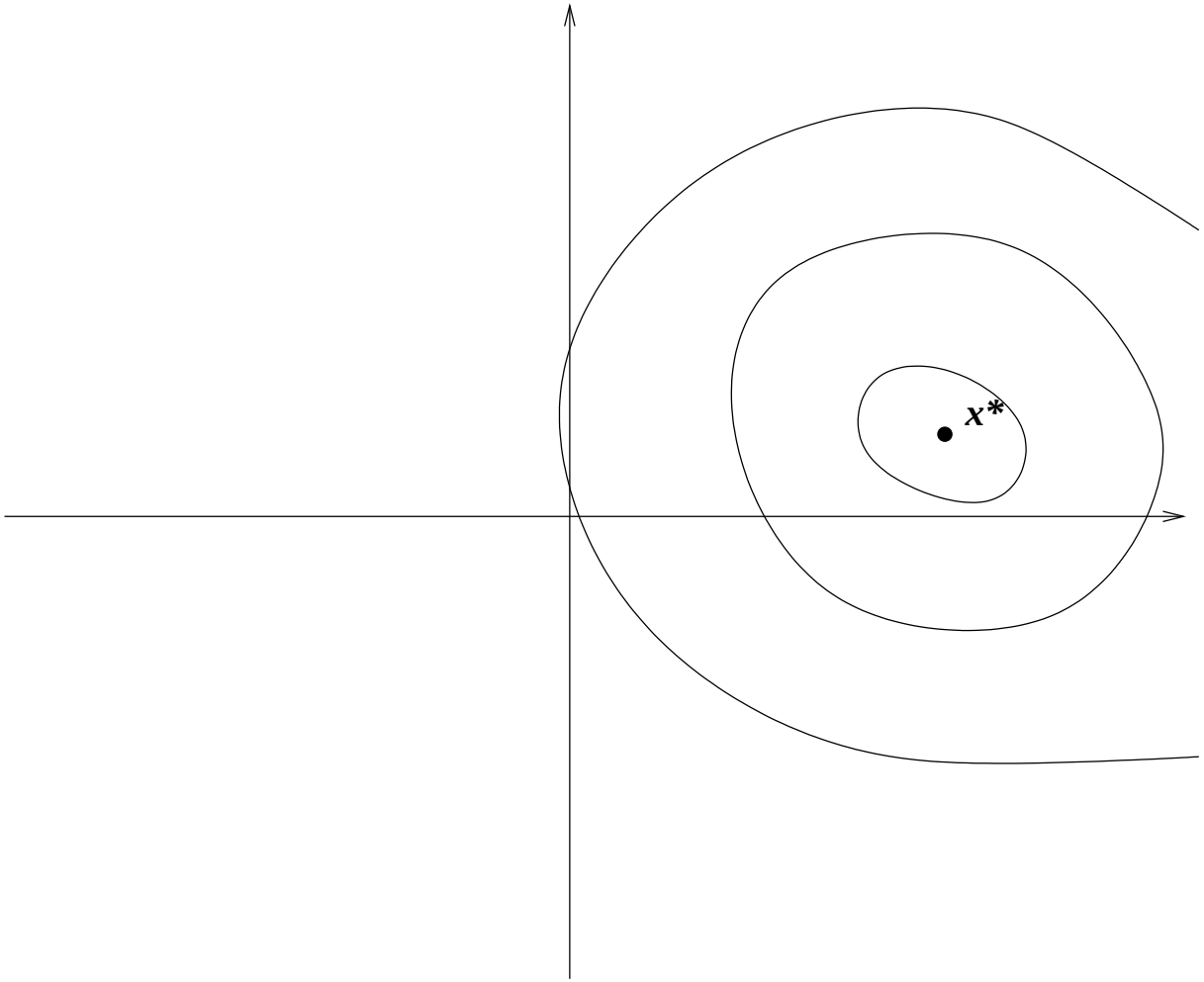
Function-constrained form:

$$\min\ \|x\|_1\ \text{ subject to }\ f(x) \leq \bar{f},$$

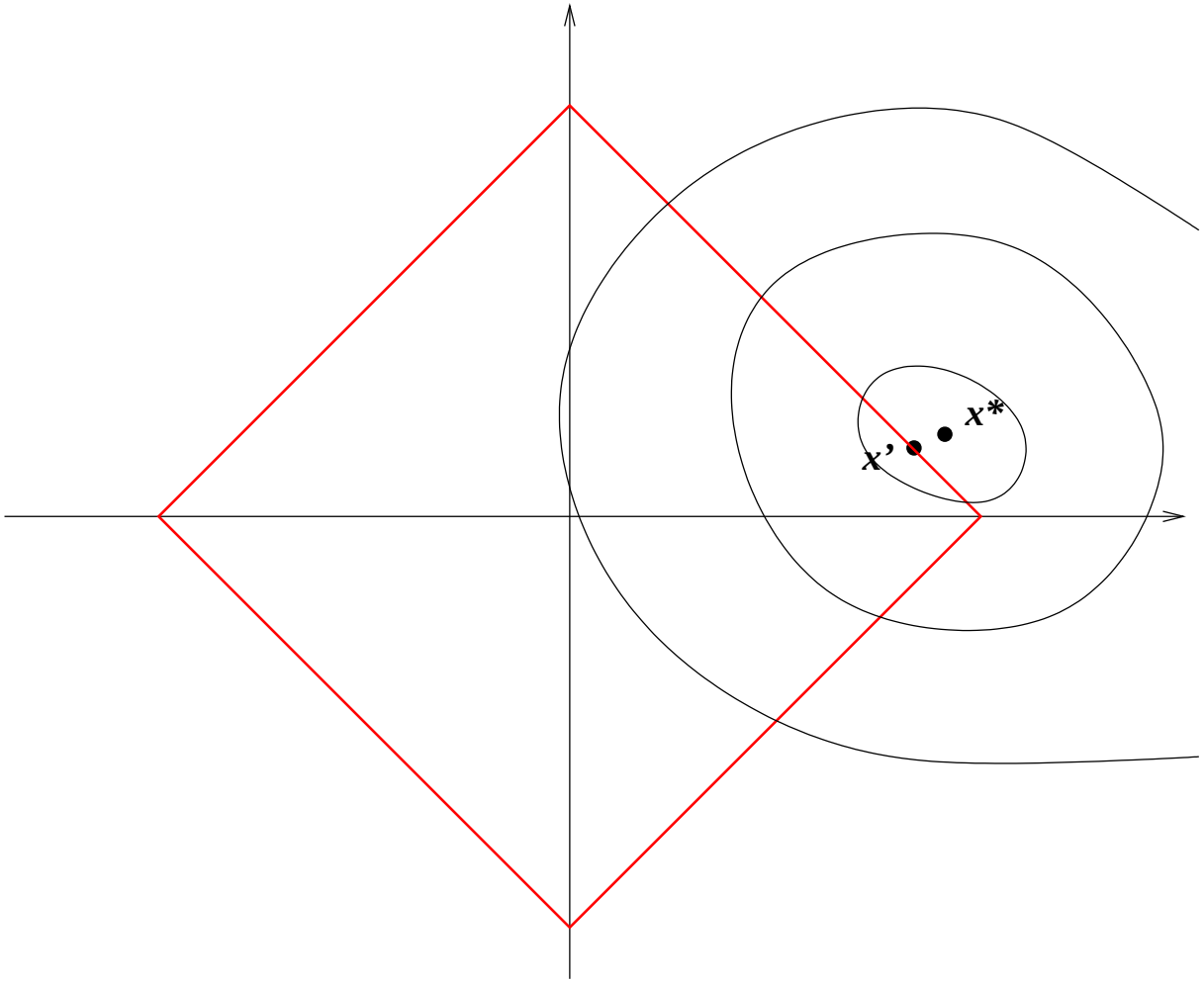
for some $\bar{f} \geq \min f$.

Can follow up with a "debiasing" phase in which the zero components are eliminated from the problem, and we minimize f itself over the support identified in the variable selection phase.

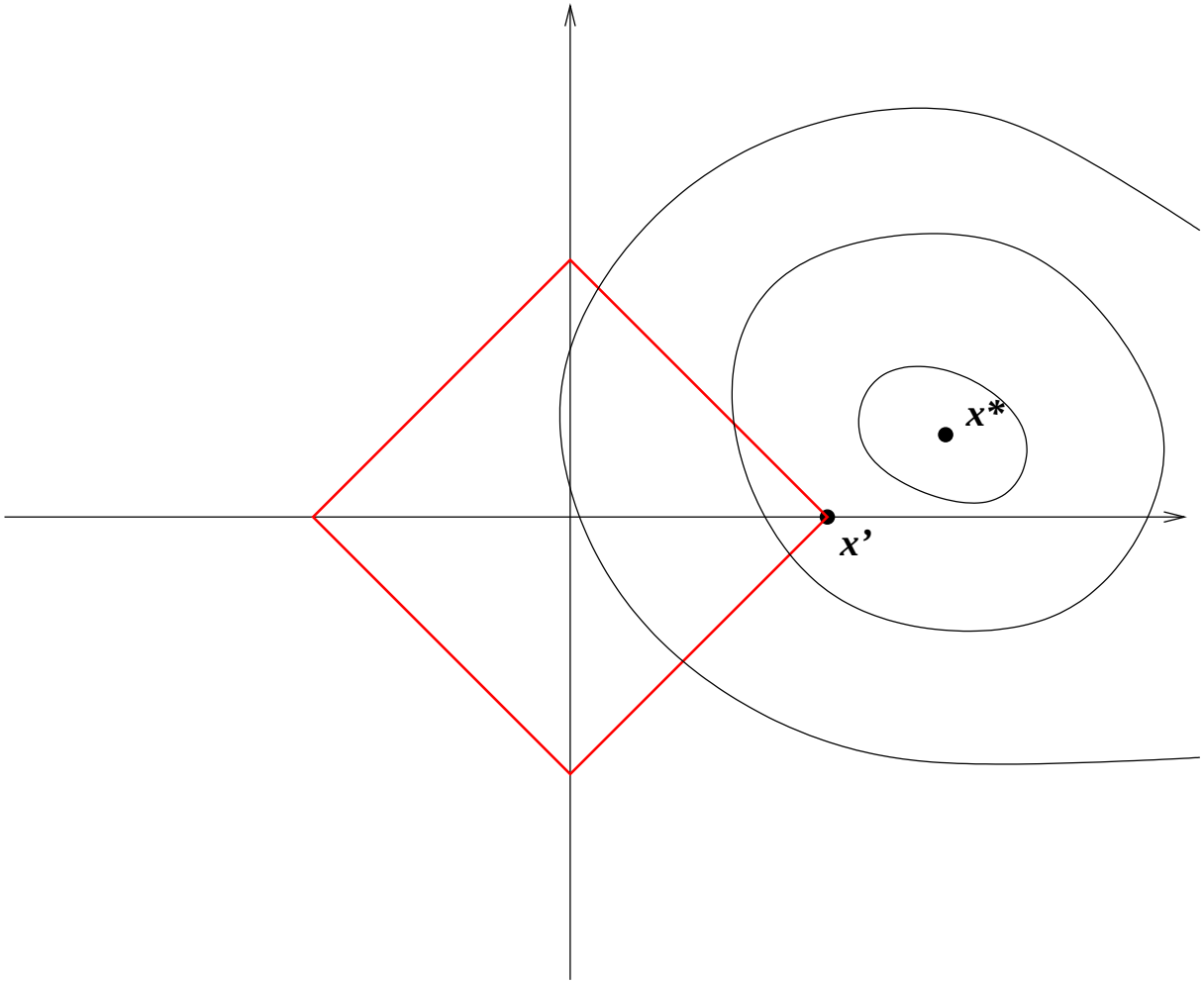
$\min f(x)$ s.t. $\|x\|_1 \leq T$: Effect of T



$\min f(x)$ s.t. $\|x\|_1 \leq T$: Effect of T



$\min f(x)$ s.t. $\|x\|_1 \leq T$: Effect of T



ℓ_1 Past and Present

ℓ_1-regularization used in statistics literature (robust estimation, regularized regression, basis pursuit) (Chen, Donoho, Saunders, 1998; Tibshirani, 1996).

Also in geophysical inversion literature (Claerbout and Muir (1973), Santosa and Symes (1986)), and elsewhere.

Heuristically, ℓ_1 often works - but is there rigorous justification?

ℓ_1 Past and Present

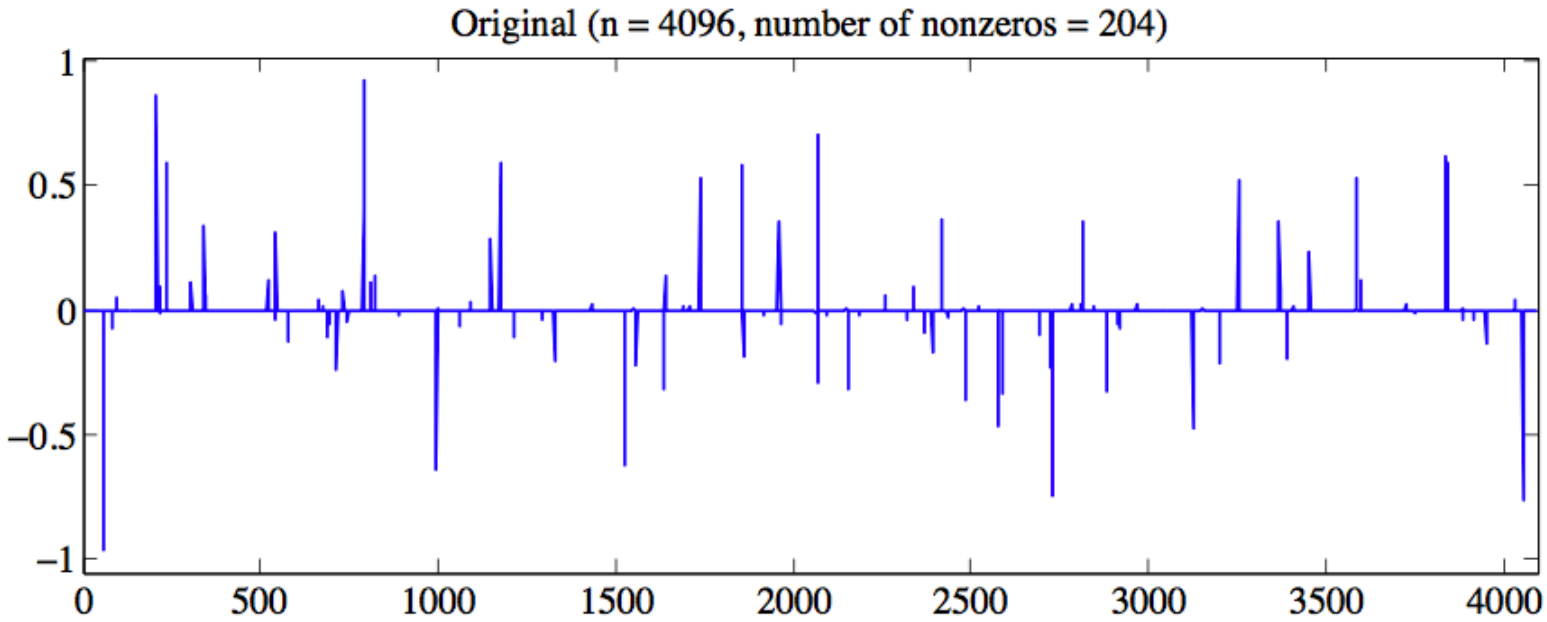
ℓ_1-regularization used in statistics literature (robust estimation, regularized regression, basis pursuit) (Chen, Donoho, Saunders, 1998; Tibshirani, 1996).

Also in geophysical inversion literature (Claerbout and Muir (1973), Santosa and Symes (1986)), and elsewhere.

Heuristically, ℓ_1 often works - but is there rigorous justification?

Compressed Sensing is a fundamental class of problems for which ℓ_1 can provably be used as a perfect surrogate for cardinality.

> *Recover $x \in \mathbb{R}^n$ from observations $y \in \mathbb{R}^m$ given $Ax = y$ with known sensing matrix $A \in \mathbb{R}^{m \times n}$.*
>
> *Additionally, know that x is sparse: $\|x\|_0 \ll n$.*


Original (n = 4096, number of nonzeros = 204)
1
0.5
0
−0.5
−1
0
500
1000
1500
2000
2500
3000
3500
4000

Compressed Sensing: Why Does ℓ_1 Work?

Elementary Analysis from W. Yin and Y. Zhang, *SIAG Views and News* 19 (2008), using Kashin (1977) and Garnaev and Gluskin (1984).

Suppose that $\bar{x}$ is the *minimum-cardinality solution* of the underdetermined linear equations $Ax = y$, where $A \in \mathbb{R}^{m \times n}$ with $m < n$.

$$\bar{x} = \arg\min \|x\|_0 \ \text{ s.t. } \ Ax = y.$$

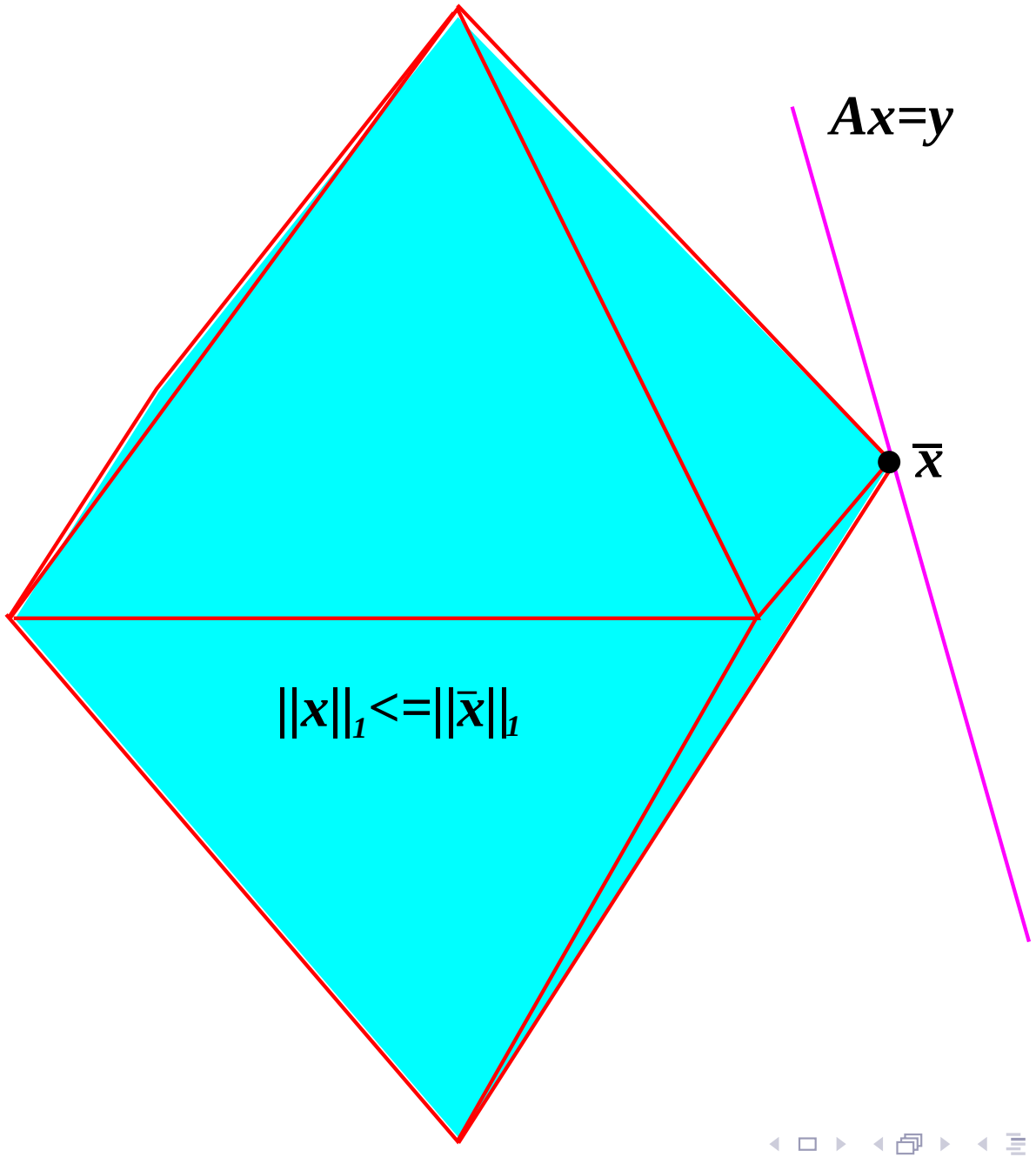
- $S \subset \{1, 2, \ldots, n\}$ be the support of $\bar{x}$;
- $k := \|\bar{x}\|_0 = |S|$;
- $Z = S^c$.

The 1-norm form is:

$$\min \|x\|_1 \ \text{ s.t. } \ Ax = y. \tag{1}$$

$\bar{x}$ solves this problem too provided

$$\|\bar{x} + v\|_1 \geq \|\bar{x}\|_1 \ \text{ for all } v \in N(A).$$


Ax=y
$\bar{x}$
$\|x\|_1 <= \|\bar{x}\|_1$

$$\begin{aligned}\|\bar{x}+v\|_1 &= \|\bar{x}_S+v_S\|_1+\|v_Z\|_1\\ &\geq \|\bar{x}_S\|_1+\|v_Z\|_1-\|v_S\|_1\\ &= \|\bar{x}\|_1+\|v\|_1-2\|v_S\|_1\\ &\geq \|\bar{x}\|_1+\|v\|_1-2\sqrt{k}\|v\|_2.\end{aligned}$$

Hence, $\bar{x}$ solves (1) provided that

$$\frac{1}{2}\frac{\|v\|_1}{\|v\|_2} \geq \sqrt{k} \text{ for all } v \in N(A).$$

In general we have only:

$$1 \leq \frac{\|v\|_1}{\|v\|_2} \leq \sqrt{n}$$

However this ratio tends to be significantly larger than 1 if v is restricted to a random subspace.

Specifically, if the elements of $A \in \mathbb{R}^{m \times n}$ are choosen iid from $N(0,1)$, we have with high probability that

$$\frac{\|v\|_1}{\|v\|_2} \geq \frac{C\sqrt{m}}{\sqrt{\log(n/m)}}, \quad \text{for all } v \in N(A),$$

for some constant C. (Concentration of measure.)

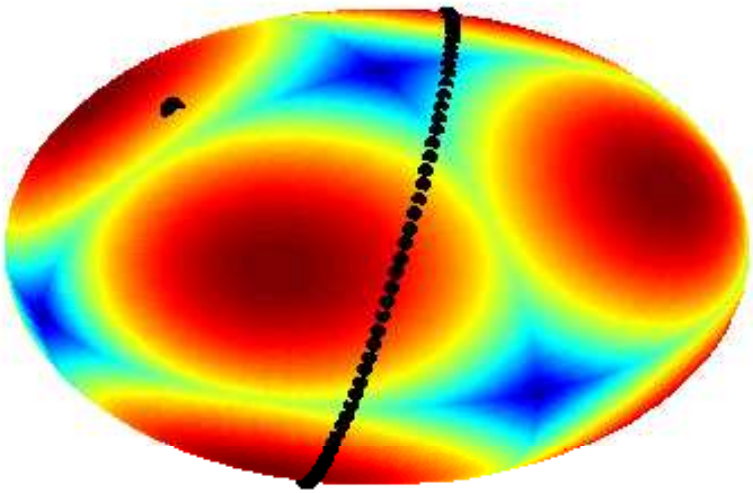
Thus, with high prob, $\bar{x}$ solves (1) if

$$m \geq \frac{4}{C^2} k \log n.$$

The number m of random linear observations (rows of A) is a multiple of $k \log n$ — typically much less than n.

Ratio $\|v\|_1/\|v\|_2$ in $\mathbb{R}^3$

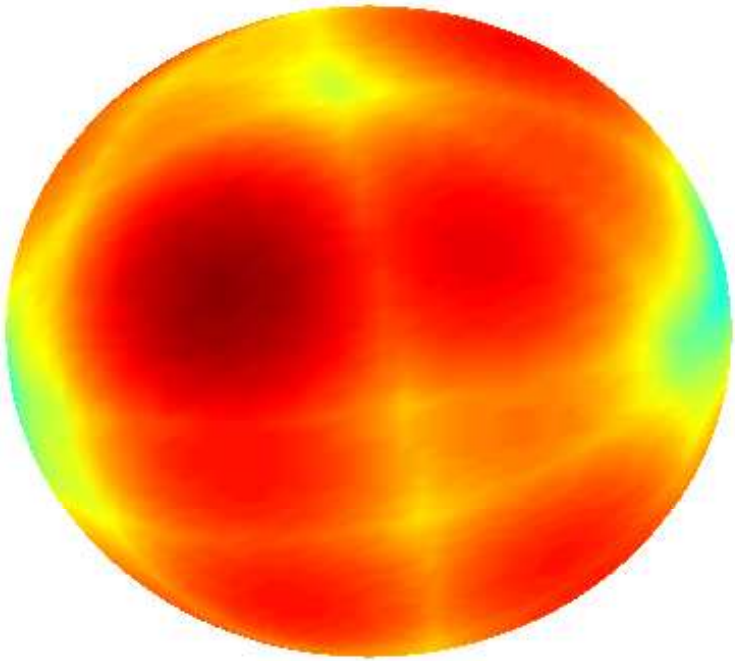
Plotting $\|v\|_1$ on sphere $\{v : \|v\|_2 = 1\}$. Blue: $\|v\|_1 \approx 1$. Red: $\|v\|_1 \approx \sqrt{3}$. (Ratio is smallest along the principal axes.)



Dot: $N(A)$ for a random $A \in \mathbb{R}^{2\times 3}$.
Equator: $N(A)$ for a random $A \in \mathbb{R}^{1\times 3}$.
(Both usually avoid smaller values of $\|v\|_1$.)

Ratio $\|v\|_1/\|v\|_2$ on Random Null Spaces

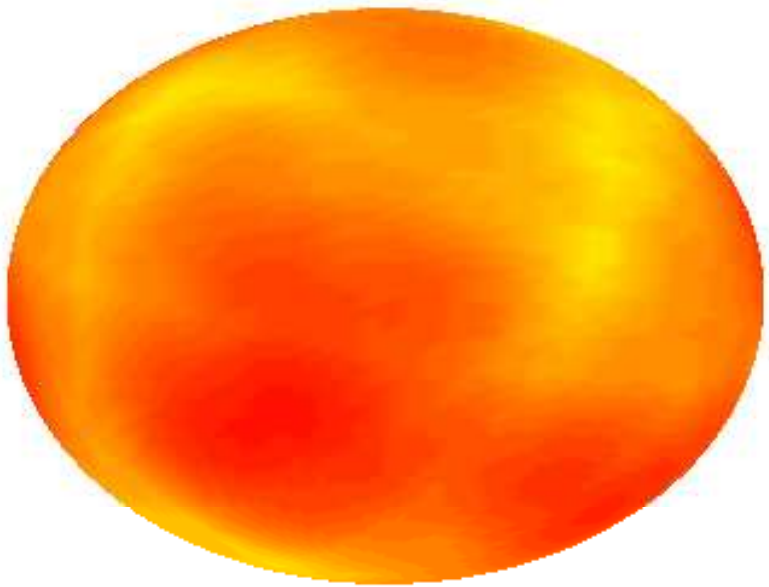
Random $A \in \mathbb{R}^{4\times 7}$, showing ratio $\|v\|_1$ for $v \in N(A)$ with $\|v\|_2 = 1$



Blue: $\|v\|_1 \approx 1$. Red: ratio $\approx \sqrt{7}$. Note that $\|v\|_1$ is well away from the lower bound of 1 over the whole nullspace.

Ratio $\|v\|_1/\|v\|_2$ on Random Null Spaces

The effect grows more pronounced as m/n grows.
Random $A \in \mathbb{R}^{17\times 20}$, showing ratio $\|v\|_1$ for $v \in N(A)$ with $\|v\|_2 = 1$.



Blue: $\|v\|_1 \approx 1$. Red: $\|v\|_1 \approx \sqrt{20}$. Note that $\|v\|_1$ is closer to upper bound throughout.

Other Analyses of ℓ_1 Formulations

- Donoho (2006): Similar elements to the above. Later study by Donoho, Tanner, others. (Bound $m \geq 2k \log n$ established.)
- Candès, Tao, Romberg (2004, 2006): Deterministic result based on a Restricted Isometry Property (RIP) of matrix A.
  - Requires column submatrices of A with $2k$ columns to be almost orthogonal. (Almost certainly true for random matrices.)

Other A have the properties required for reconstruction: e.g. Bernoulli random, random rows of discrete cosine / discrete Fourier transform.

Applications of ℓ_1

Sparse Basis Signal Representations. e.g. wavelet basis: $z = Wx$ where x is vector of wavelet coefficients and W is inverse wavelet transform. Formulation:

$$\min_x \frac{1}{2}\|y - LWx\|_2^2 + \tau\|x\|_1,$$

where L is linear observation operator. Allows for Gaussian noise in observations y.

Sparse Learning, Feature Selection. From data $x_i \in \mathbb{R}^n$, $i = 1, 2, 3, \ldots$ and outcomes y_i, $i = 1, 2, 3, \ldots$, learn a function f that predicts outcome y for a new vector x.

Want f to be plausible and possibly to depend on just a few components of x (features).

- LASSO
- regularized logistic regression
- sparse support vector machines

Applications of ℓ_1

Power Systems:

- Power distribution network can become "infeasible" after a disturbance (e.g. a transmission line failure).
- We may be interested in the "least disruptive fix" i.e. change power generation (on a *few* generation nodes) and / or shed load (on a *few* load nodes) to restore feasibility.

Face Recognition: (J. Wright et al, 2008)

Seismic Inversion: (Herrmann et al., 2007-)

Compressive Radar.

See Rice Compressed Sensing page `http://dsp.rice.edu/cs` for many other applications.

Other Applications, Other Structures

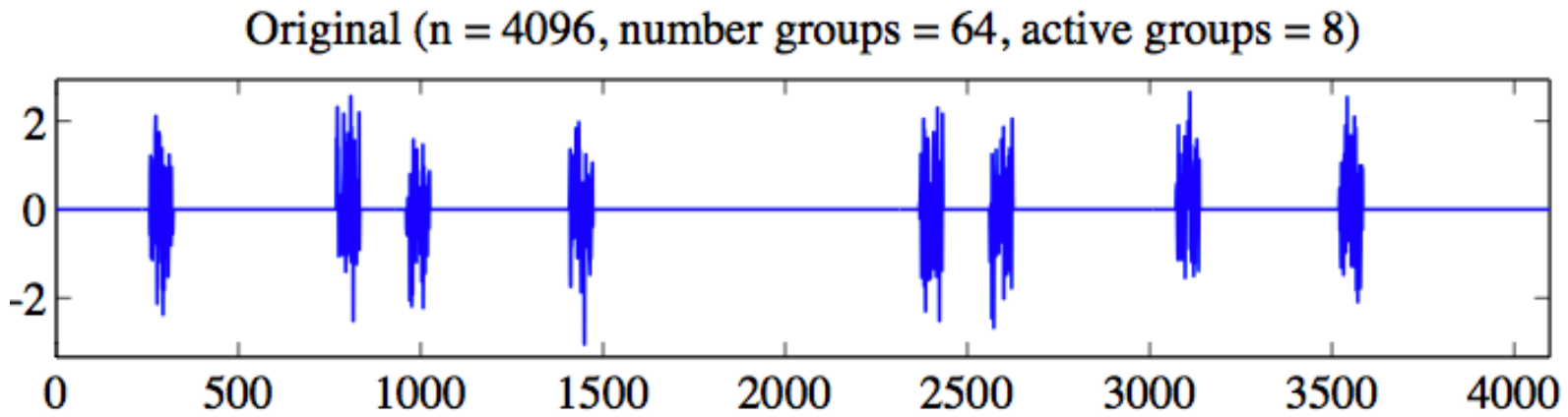
In many applications, the solution structure desired is different from simple sparsity — cannot be easily attained by the ℓ_1 regularizer.

What kinds of structures are common?

How can we choose regularizers that induce the desired structure, while retaining tractability of the optimization problem?

Group Sparsity

There may be a natural relationship between some components of x. We could thus group the components, and select or deselect at the group level.



Can use "sum of ℓ_∞" or "sum of ℓ_2" regularizers:

$$\sum_{k=1}^{m} \|x_{[k]}\|_\infty, \qquad \sum_{k=1}^{m} \|x_{[k]}\|_2,$$

where $[k]$ (for $k = 1, 2, \ldots, m$) represent subsets of the components of x.

(Turlach, Venables, Wright, 2005).

Image Processing

Natural images are not random! They tend to have large areas of near-constant intensity or color, separated by sharp edges.

Denoising: Given an image in which the pixels contain noise, find a "nearby natural image."

Can have Gaussian noise, "salt-and-pepper" noise, impulsive noise, etc.

(a) Cameraman: Clean

(b) Cameraman: Noisy

(c) Cameraman: Denoised

(d) Cameraman: Noisy

Total-Variation Regularization

Given intensity measures U_{ij} for $i, j = 1, 2, \ldots, N$ (a 2D grid), define the *variation* at grid point (i, j) as

$$\left\| \begin{bmatrix} U_{i+1,j} - U_{ij} \\ U_{i,j+1} - U_{ij} \end{bmatrix} \right\|_2.$$

(zero iff U_{ij}, $U_{i+1,j}$, $U_{i,j+1}$ all have the same intensity).

Total Variation obtained by summing across the grid:

$$\mathrm{TV}(U) := \frac{1}{N^2} \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \left\| \begin{bmatrix} U_{i+1,j} - U_{ij} \\ U_{i,j+1} - U_{ij} \end{bmatrix} \right\|_2.$$

Forces most grid points (i, j) to have the same intensities as their neighbors. (Rudin, Osher, Fatemi, 1992)

Denoising: Given observed intensities $F \in R^{N \times N}$, solve

$$\min_{U \in \mathbb{R}^{N \times N}} \frac{1}{2} \|U - F\|_F^2 + \tau \mathrm{TV}(U).$$

Medical Imaging

- X-ray computed tomography (CT); nuclear magnetic resonance (MRI) and its "real-time" and "functional" variants.
- Fewer measurements $\Rightarrow$ Less radiation (for CT), less time.
- Images are natural, and a prior may be available.

Formulation features:

- Use TV regularization to induce natural image.
- Can also use ℓ_1 regularization to penalize deviation from the prior.

Lustig (2008); Lauzier, Tang, Chen (2011).

Angiogram

Matrix Estimation / Completion

Given an $m \times n$ matrix M in which only certain elements are known:

$$\Omega \subset \{(i,j) \,|\, i = 1,2,\ldots,m, \;\; j = 1,2,\ldots,n\}.$$

Find a matrix X with "nice structure" such that $X_{ij} \approx M_{ij}$ for $(i,j) \in \Omega$.

Example: Netflix

Desirable structures:

- Low rank: induced by nuclear norm $\|X\|_*$ (sum of singular values) (Recht, Fazel, Parrilo, 2010) or "max norm" (Lee at al., 2010).
- Sparsity: Induced by element-wise 1-norm: $\sum_{i,j} |X_{ij}|$.
- Both: combine these regularizers.

Algorithms: Many Techniques Used

- Large-scale optimization: optimal first-order, gradient projection, second-order, continuation, coordinate relaxation, interior-point, augmented Lagrangian, conjugate gradient, semismooth Newton ...
- Nonsmooth optimization: cutting planes, subgradient methods, successive approximation, smoothing, prox-linear methods, ...
- Dual and primal-dual formulations / methods
- Numerical linear algebra
- Stochastic approximation, sampled-average approximation.
- Heuristics

Also a LOT of domain-specific knowledge about the problem structure and the type of solution demanded by the application.

Discuss just a few key techniques — but omit other important ones.

A Useful Setting

Formulate a regularized problem

$$f(x) + \tau c(x),$$

where

- **nominal objective** $f(x)$, e.g. fit to data;
- **regularization function** or **regularizer** $c(x)$ — usually convex and nonsmooth — to induce the desired structure in x.
- **regularization parameter** $\tau > 0$. Trades off between optimizing the nominal objective and the regularizer.

Prox-Linear Methods

For the setting

$$\min_x \ f(x) + \tau c(x).$$

At x^k, solve this subproblem for new iterate x^{k+1}:

$$x^{k+1} = \arg\min_z \nabla f(x^k)^T (z - x^k) + \tau c(z) + \frac{1}{2\alpha_k} \|z - x^k\|_2^2,$$

for some choice of $\alpha_k > 0$.

Works well when this subproblem is "easy" to formulate and solve.

- If c is vacuous, this reduces to gradient descent, with a line search.
- If $\alpha_k \leq 1/L$ ($L =$ Lipschitz constant for ∇f), get descent at each iteration and convergence.
- Adaptive α_k: can impose sufficient decrease criterion via backtracking; "Barzilai-Borwein" α_k for a nonmonotone approach.

Application to ℓ_1 Regularizer

For ℓ_1 regularization ($c(x) = \|x\|_1$), can solve the subproblem explicitly in $O(n)$ time: **("Shrink Operator")**

The other expensive steps at each iteration are computation of ∇f and computation of f (to test for acceptability of x^{k+1}).

- **Compressed Sensing.** $\nabla f(x) = A^T(Ax - y)$: the matrix-vector multiplications are often cheap (e.g. for discrete cosine transformation, chirp sensing). **Codes:** SpaRSA, FPC.
- **Logistic Regression.** Evaluation of ∇f less expensive after f has been evaluated. **Code:** LPS.

For other regularizers e.g. $\mathrm{TV}(x)$, the subproblem is nontrivial, so we may have to settle for an approximate solution. (This issue persists in alternating direction augmented Lagrangian approaches; see below.)

Application to Matrix Completion

Formulate matrix completion as

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2 + \tau \|X\|_*,$$

where $\mathcal{A}(X) = [A_i \bullet X]_{i=1,2,\ldots,p}$ and $\|X\|_*$ is the nuclear norm.

"Shrink operator" (subproblem) is

$$\min_Z \frac{1}{2\alpha_k}\|Z - Y^k\|_F^2 + \tau \|Z\|_*.$$

Can solve explicitly using a singular value decomposition of Y^k.

Code: SVT: Compute a partial SVD (largest singular values) using Lanczos. (Cai, Candès, Shen, 2008)

Same framework works with max-norm (Lee et al. 2010).

Prox-Linear: Theory and Practice

Convergence proved in the case of convex c using fairly standard analysis (monotone and nonmonotone line search variants, gradient descent). (e.g. Wright, Figueiredo, Nowak 2008)

Theory from forward-backward splitting methods also useful. (Combettes and Wajs, 2005)

Can extend the theory beyond convexity, to prox-regular functions. (Lewis and Wright, 2008)

In practice, speed of convergence depends heavily on τ.

- Larger τ (sparser solution): convergence often very fast;
- Smaller τ: can be miserably slow (or fails).

Continuation helps: solve for a decreasing sequence of τ values, using previous solution as the starting point for the current τ.

Other Enhancements

- Block-coordinate: Take steps in just a subset of components at each iteration (need only partial gradient). (Tseng and Yun, 2009; Wright, 2011)
- Estimation of the optimal manifold (i.e. the nonzero coefficients, in the case of ℓ_1) and consequent reduction of the search space. (Shi et al., 2008)
- Use (approximate) second-order information, e.g. in logistic regression (Byrd et al., 2010; Shi et al, 2008)

Accelerated First-Order Methods

Can exploit the research on methods for smooth convex optimization that use gradients, but do better than simply stepping in the negative gradient direction $-\nabla f(x)$. (Nesterov)

They generate two (or three) intertwined sequences. Typically:

- Get the next x-sequence iterate from a short gradient-descent step from the latest y-sequence element
- Get the next y-sequence element by extrapolating from the last two x-sequence iterates.

FISTA (Beck and Teboulle, 2008): min f, L = Lipschitz const for ∇f:

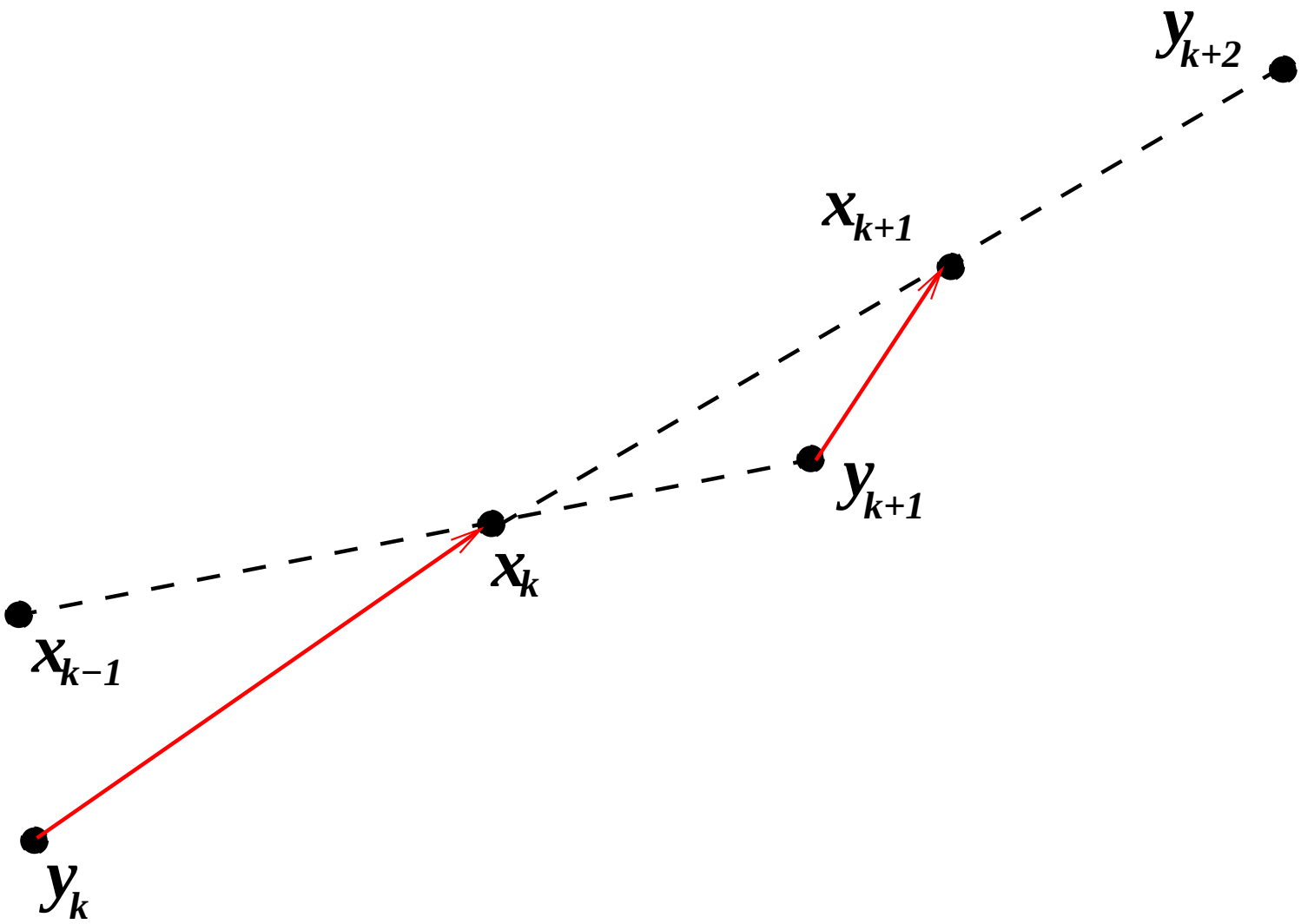
0: Choose x_0; set $y_1 = x_0$, $t_1 = 1$;

k: $x_k \leftarrow y_k - \frac{1}{L}\nabla f(y_k)$;

$t_{k+1} \leftarrow \frac{1}{2}\left(1 + \sqrt{1 + 4t_k^2}\right)$;

$y_{k+1} \leftarrow x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$.

Two Sequences: $\{x_k\}$ and $\{y_k\}$



Analysis is short but not very intuitive. $f(x_k)$ converges to its optimal value f^* at a "fast" sublinear rate of $O(1/k^2)$.

Can extend to the regularized problem min $f(x) + \tau c(x)$ by replacing the step $y_k \to x_k$ by the prox-linear subproblem, with $\alpha_k \equiv 1/L$.

Practically, less sensitive to τ than prox-linear.

Similar approaches can achieve a geometric rate when f is *strongly* convex. (Nesterov, 2004)

Stochastic Gradient Methods

For min $f(x)$ (f convex, nonsmooth), stochastic approximation methods (SA or SGD) may be useful when a cheap estimate of a subgradient $\partial f(x^k)$ is available:

$$x_{k+1} = x_k - \gamma_k g_k, \qquad E(g_k) \in \partial f(x_k),$$

for steplength $\gamma_k > 0$.

Exact evaluation of f or a subgradient may require a complete scan through the data — but g_k could be obtained from a single data element.

The machine learning community is very interested in these methods.

Acceptable solutions may be obtained without even looking at some of the data, if random sampling is done.

(Robbins and Monro, 1951)

Regularized Dual Averaging

(Nesterov, 2009) For min $f(x)$ with f convex, nonsmooth. Use subgradients $g_i \in \partial f(x_i)$ and average, to obtain

$$\bar{g}_k = \frac{1}{k} \sum_{i=1}^{k} g_i.$$

Step:

$$x_{k+1} := \min_x \bar{g}_k^T x + \frac{\gamma}{\sqrt{k}} \|x - x_1\|_2^2, \qquad \text{for some } \gamma > 0.$$

Possibly average the iterates $x_1, x_2, x_3, \ldots$ too.

Described for min $\frac{1}{T} \sum_{t=1}^{T} f_t(x) + \tau c(x)$ by Xiao (2010).

The (non-averaged) primal iterates can almost surely identify the optimal manifold on which x^* lies. e.g. when $c(x) = \|x\|_1$, identify the nonzero components. (Lee and Wright, 2010)

Augmented Lagrangian

A classical method: For closed convex Ω:

$$\min_{x \in \Omega} p(x) \text{ subject to } Ax = b.$$

Generate iterates x^k together with Lagrange multiplier estimates λ^k from:

- x^k is approximate solution of

$$\min_{x \in \Omega} p(x) + (\lambda^k)^T (Ax - b) + \frac{\mu_k}{2} \|Ax - b\|_2^2;$$

- update Lagrange multipliers:

$$\lambda^{k+1} = \lambda^k + \mu_k (Ax^k - b).$$

(If p convex, need only $\mu_k > 0$.)

Constrained Formulation

Given

$$\min_x \ f(x) + \tau c(x),$$

"duplicate" the variable and write as an equality constrained problem:

$$\min_{z,u} \ f(z) + \tau c(u) \ \text{ subject to } \ u = z.$$

Augmented Lagrangian:

$$(z^k, u^k) := \min_{z,u} \ f(z) + \tau c(u) + (\lambda^k)^T (u - z) + \frac{\mu_k}{2} \|u - z\|_2^2,$$
$$\lambda^{k+1} := \lambda^k + \mu_k (u^k - z^k).$$

The $\min_{z,u}$ problem is usually still too hard to solve (u and z are coupled via final penalty term). However can take alternating steps in z and u.

Alternating Directions

(Eckstein and Bertsekas, 1992)

$$z^k := \min_z f(z) + \tau c(u^{k-1}) + (\lambda^k)^T(u^{k-1} - z) + \frac{\mu_k}{2}\|u^{k-1} - z\|_2^2,$$

$$u^k := \min_u f(z^k) + \tau c(u) + (\lambda^k)^T(u - z^k) + \frac{\mu_k}{2}\|u - z^k\|_2^2,$$

$$\lambda^{k+1} := \lambda^k + \mu_k(u^k - z^k).$$

Approximate minimization for z and u may now be much simpler. e.g. for compressed sensing:

- One of these minimizations is the "shrink operator" (easy);
- The other is linear system with coefficient matrix $(A^TA + \sigma I)$ (solve approximately).

The subproblems are not vastly different from prox-linear subproblems:

- λ^k is asymptotically similar to the gradient term in prox-linear;
- the quadratic term is the same.

Use in Sparse Optimization

Extensions and variants of these ideas have been much studied recently, and applied in various contexts. Examples:

- Compressed sensing (Yang and Zhang, 2009; **Code:** YALL1), (Goldfarb, Ma, Scheinberg, 2010)
- Image processing (Goldstein and Osher, 2008; Figueiredo and Bioucas-Dias, 2010, 2011)
- Video processing, matrix completion, sparse principal components (Goldfarb, Ma, Scheinberg, 2010).

Can be melded with (accelerated) first-order methods (Goldfarb, Ma, Scheinberg, 2010).

To Conclude, Some Observations

- Sparse optimization has wealth of diverse applications.
- Formulations are key, particularly design of regularizers.
- Exciting forum for algorithm design:
  - Assembling known tools
  - Designing and analyzing new tools
  - Fitting to the application and context.
- The interdisciplinary nature of optimization is especially evident in sparse optimization!

A Very Incomplete Bibliography

- M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems, " *IEEE Transactions on Image Processing* 20 (2011), pp. 681–695.
- A. Beck and M. Teboulle, "A fast iterative shrinkage-threshold algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences* 2 (2009), pp. 183–202.
- J. Bioucas-Dias and M. Figueiredo, "Multiplicative noise removal using variable splitting and constrained optimization," *IEEE Transactions on Image Processing*, 19 (2010), pp. 1720–1730.
- R. Byrd, G. Chin, W. Neveitt, and J. Nocedal, "On the use of stochastic Hessian information in unconstrained optimization," 2010.
- J.-F. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," Technical Report, 2008.
- V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, "The convex geometry of linear inverse problems," Technical Report, December 2010.
- S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing* 20 (1998), pp. 33–61.
- P. Combettes and V. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling and Simulation* 4 (2005), pp. 1168–1200.
- J. Eckstein and D. Bertsekas, "On the Douglas-Rachford splitting method and the proximal-point algorithm for maximal monotone operators," *Mathematical Programming* 55 (1992), pp. 293–318.
- M. Figueiredo, R. Nowak, and S. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE Journal on Selected Topics in Signal Processing*, 1 (2007), pp. 586–597.

- D. Goldfarb, S. Ma, and K. Scheinberg, "Fast alternating linearization methods for minimizing the sum of two convex functions," submitted, 2010.
- T. Goldstein and S. Osher, "The split Bregman method for ℓ_1 regularized problems," *SIAM Journal on Imaging Sciences* 2 (2009), pp. 323–343.
- E. Hale, W. Yin, and Y. Zhang, "A fixed-point continuation method for -minimization: Methodology and convergence," *SIAM Journal on Optimization* 19 (2008), pp. 1107–1130.
- M. Hintermüller and K. Kunisch, "Total bounded variation regularization as a bilaterally constrained optimization problem," *SIAM Journal on Applied Mathematics* 64 (2004), pp. 1311–1333.
- M. Hintermüller and G. Stadler, "An infeasible primal-dual algorithm for total bounded variation-based inf-convolution-type image restoration," *SIAM Journal on Scientific Computing* 28 (2006), pp. 1–23.
- P. Lauzier, J. Tang, G.-H. Chen, "Prior image constrained compressed sensing (PICCS)," submitted, 2011.
- J. Lee, B. Recht, R. Salakhutdinov, N. Srebro, and J. Tropp, "Practical large-scale optimization for max-norm regularization," NIPS, 2010.
- S. Lee and S. Wright, "Implementing algorithms for signal and image reconstruction on graphical processing units," 2008.
- S. Lee and S. Wright, "Manifold Identification of Dual Averaging Methods for Regularized Stochastic Online Learning," ICML, 2011.
- A. Lewis and S. Wright, "A proximal method for composite minimization," 2008.
- M. Lustig, "Sparse MRI," Ph.D. Thesis, Stanford University, 2008.
- Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer, 2004.

- Y. Nesterov, "Primal-dual subgradient methods for convex programs," *Mathematical Programming B* 120 (2009), pp. 221–259.
- N. Rao, R. Nowak, S. Wright, N. Kingsbury, "Convex approaches to model avelet sparsity," ICIP, 2011.
- B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions to linear matrix equations via nuclear norm minimization," *SIAM Review* 52 (2010), pp. 471–501.
- H. Robbins and S. Monro, "A Stochastic approximation method," *Annals of Mathematical Statistics* 22 (1951), pp. 400–407.
- L. Rudin, S. Osher, E. Fatemi, "Nonlinear total variation based noise removal algorithms," Physica D 60 (1992), pp. 259–268.
- W. Shi, G. Wahba, S. Wright, K. Lee, R. Klein, and B. Klein, "LASSO-Patternsearch algorithm with application to opthalmology data," *Statistics and its Interface* 1 (2008), pp. 137–153.
- R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society B* 58 (1996), pp. 267–288.
- P. Tseng and S. Yun, "A coordinate descent method for nonsmooth separable minimization," Math Prog B (2009), pp. 387–423.
- B. Turlach, W. Venables, and S. Wright, "Simultaneous variable selection," Technometrics 47 (2005), pp. 349–363.
- E. van den Berg and M. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM Journal of Scientific Computing*, 31 (2008), pp. 890–912.
- S. Wright, "Accelerated Block-Coordinate Relaxation for Regularized Optimization," 2010.
- S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing* 57 (2009), pp. 2479–2493.
- L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *Journal of Machine Learning Research* 11 (2010), pp. 2543-2596.
- J. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1 problems in compressive