Smoothed  Optimization

Wotao Yin (CAAM, Rice University)

December 2012 - Peking U

Sparse optimization

The optimization that processes "large" data sets for "sparse" solutions.

("Sparse" means having very few nonzeros and other structures as well.)

Some applications

- Signal decomposition
- Structure discover
- Signal recovery, compressive sensing
- (many more ...)

Example: motion separation

Goal: to separate machine motion from human motion

(wmv)

(Joint with W.Deng, S.Jiang, et al) Model extending robust PCA:

$$\underset{X,P,Z}{\text{minimize}}\, \mu_1\|X\|_* + \mu_2\|\theta\|_1 + \|Z\|_1, \quad \text{s.t. } X + D\theta + Z = \text{input video.}$$

X: static; $D\theta$: background and reg. motion, Z irreg. motion

Some applications

- Signal decomposition
- Structure discover
- Signal recovery, compressive sensing
- (many more ...)

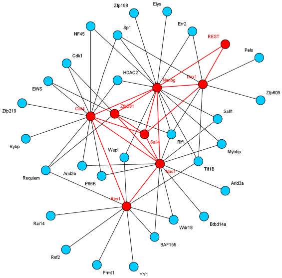
Example: latent variable graphical model selection

Graphical model: a statistical model defined on a graph

Nodes: Gaussian random variables $X = [X_1, X_2, \ldots]$;

Edges: a missing edge means conditional independence;

Inverse covariance matrix Σ_X^{-1}: $(\Sigma_X^{-1})_{ij} \neq 0$ if r.v. X_i and X_j are not conditional independent.



(from Hulei Xu)

Applications: gene regulatory (molecular reaction) network, stocks.

Example: latent variable graphical model selection

Chandrasekaran-Parrilo-Willsky'10: $X = [\text{Observed Hidden}] = [X_O \; X_H]$.

Assume **sparse**

$$\Sigma_X^{-1} = \begin{bmatrix} R_{OO} & R_{OH} \\ R_{HO} & R_{HH} \end{bmatrix}.$$

The **observed inverse co-variance of** X_O is the Schur complement

$$\Sigma_{X_O}^{-1} = R_{OO} - R_{OH} R_{HH}^{-1} R_{HO} = \text{sparse} - \text{low-rank}.$$

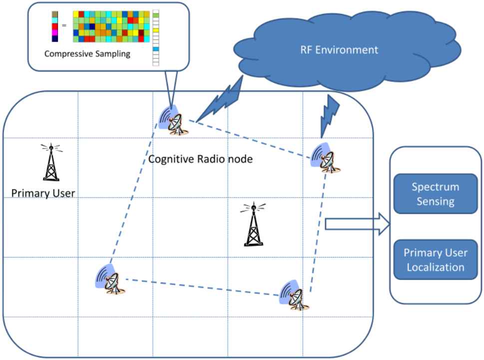
Model:

$$\underset{R,S,L}{\text{minimize}}\, \ell(R; \hat{\Sigma}_X) + \alpha \|S\|_1 + \beta \operatorname{tr}(L) \quad \text{s.t. } R = S - L, R \succeq 0, L \succeq 0.$$

Some applications

- Signal decomposition
- Structure discover
- Signal recovery, compressive sensing
- (many more ...)

Wireless spectrum sensing



Goal: to identify wireless bands in use and locate their sources, in real time

Model:

$$\text{minimize} \;\; \textsf{fitting} + \textsf{spatial sparsity} + \textsf{spectrum sparsity}$$

Computational approaches

Off-the-shelf: subgradient descent, LP/SOCP/SDP

Computational approaches

Off-the-shelf: subgradient descent, LP/SOCP/SDP

Smoothing: in order to apply gradient-based methods, approximate ℓ_1 by

- $\ell_{1+\epsilon}$-norm
- $\sum_i \sqrt{x_i^2 + \epsilon}$
- Huber-norm.

Computational approaches

Off-the-shelf: subgradient descent, LP/SOCP/SDP

Smoothing: in order to apply gradient-based methods, approximate ℓ_1 by

- $\ell_{1+\epsilon}$-norm
- $\sum_i \sqrt{x_i^2 + \epsilon}$
- Huber-norm.

Splitting: turn a problem into multiple simpler subproblems

- Operator splitting: GPSR/SPGL1/FPC/SpaRSA/FISTA/...
- Variable splitting: ADMM/split Bregman/...

Computational approaches

Off-the-shelf: subgradient descent, LP/SOCP/SDP

Smoothing: in order to apply gradient-based methods, approximate ℓ_1 by

- $\ell_{1+\epsilon}$-norm
- $\sum_i \sqrt{x_i^2 + \epsilon}$
- Huber-norm.

Splitting: turn a problem into multiple simpler subproblems

- Operator splitting: GPSR/SPGL1/FPC/SpaRSA/FISTA/...
- Variable splitting: ADMM/split Bregman/...

Non-convex: ℓ_p-minimization ($p < 1$), reweighted ℓ_1/LS, non-convex priors

Computational approaches

Off-the-shelf: subgradient descent, LP/SOCP/SDP

Smoothing: in order to apply gradient-based methods, approximate ℓ_1 by

- $\ell_{1+\epsilon}$-norm
- $\sum_i \sqrt{x_i^2 + \epsilon}$
- Huber-norm.

Splitting: turn a problem into multiple simpler subproblems

- Operator splitting: GPSR/SPGL1/FPC/SpaRSA/FISTA/...
- Variable splitting: ADMM/split Bregman/...

Non-convex: ℓ_p-minimization ($p < 1$), reweighted ℓ_1/LS, non-convex priors

Non-optimization approaches: greedy, message passing, ...

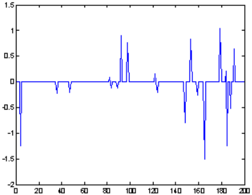
Sparse optimization

- existed a long time (seismic, TV, sparse SVM, kSVD, soft-thresholding);
- started to grow more quickly upon the arrival of CS, $m = O(k \log(n/k))$;
- has become a distinct area in optimization;
- interacts with a variety of other areas.

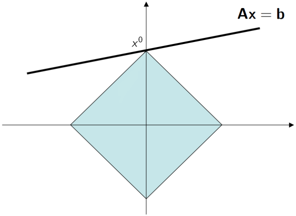
Goal of the rest of talk: smoothing without tuning the smoothing parameter.

Sharp corners and sparse solution

The level-set of ℓ_1-norm has sharp corners, giving sparse sol.



$$\min\{\|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b}\}$$

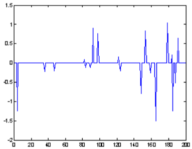


Smoothing

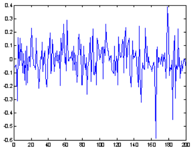
Smoothing (e.g., $\ell_{1+\epsilon}$-norm, $\sum_i \sqrt{x_i^2+\epsilon}$, Huber-norm) rounds the corners and "kills" solution sparsity.

We "smooth" ℓ_1 by adding ℓ_2^2:

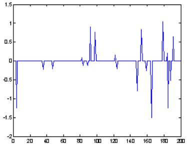
$$\|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2.$$



$\min\{\|\mathbf{x}\|_1 : \mathbf{Ax} = \mathbf{b}\}$



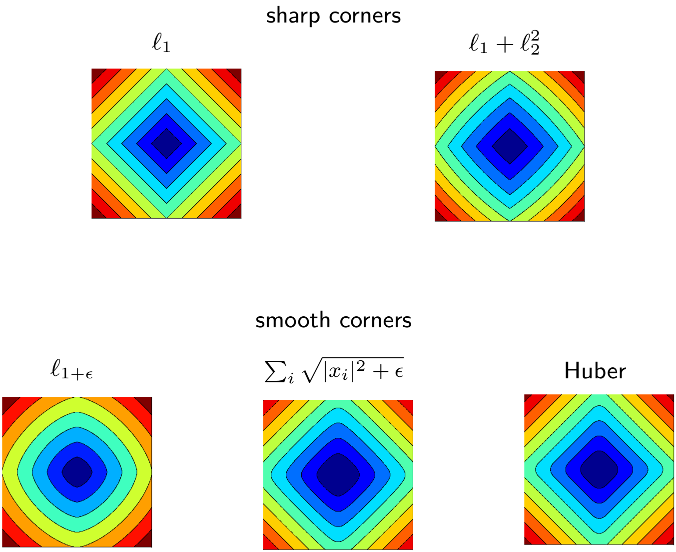
$\min\{\|\mathbf{x}\|_2^2 : \mathbf{Ax} = \mathbf{b}\}$



$\min\{\|\mathbf{x}\|_1 + \frac{1}{25}\|\mathbf{x}\|_2^2 : \mathbf{Ax} = \mathbf{b}\}$

Exactly the same as ℓ_1 solution

Related to: Tikhonov, linearized Bregman, elastic net.


sharp corners
ℓ_1
$\ell_1 + \ell_2^2$
smooth corners
$\ell_{1+\epsilon}$
$\sum_i \sqrt{|x_i|^2 + \epsilon}$
Huber

Exact regularization

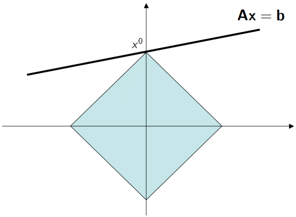
Theorem (Friedlander and Tseng [2007], Yin [2010])

There exists a finite $\alpha^0 > 0$ such that whenever $\alpha > \alpha^0$, the solution to

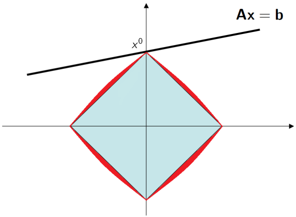
$$\textit{(L1+LS)} \qquad \text{minimize } \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 \quad \textit{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}$$

is also a solution to

$$\textit{(L1)} \qquad \text{minimize } \|\mathbf{x}\|_1 \quad \textit{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}.$$



L1



L1+LS

Compressive sensing (CS) recovery guarantees

- Goal: reliably recover a sparse signal from its linear measurements
- Question: how many linear measurements do I need?
- A typical form of condition for minimizing $\|\mathbf{x}\|_1$

$$\#\text{measurements} \geq C \cdot f(\text{signal dim}, \text{signal sparsity}).$$

- After adding $\frac{1}{2\alpha}\|\mathbf{x}\|_2^2$, the condition becomes

$$\#\text{measurements} \geq (C + O(1/\alpha)) \cdot f(\text{signal dim}, \text{signal sparsity}).$$

Example: RIP-based recovery guarantees

Definition (Candes and Tao [2005])

The *restricted isometry property* (RIP) constant δ_k is the smallest value such that

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_k)\|\mathbf{x}\|_2^2$$

holds for all k-sparse vectors $\mathbf{x} \in \mathbb{R}^n$.

Importance: A few classes of $\mathbf{A} \in \mathbb{R}^{m \times n}$ (e.g., those sampled i.i.d. from sub-Gaussian distributions) have RIP with a sufficiently small δ_k if

$$m \geq O(k \log(n/k)).$$

Example: RIP-based recovery guarantees

Theorem (exact recovery, Lai and Yin [2012])

Under the assumptions

1. $\mathbf{x}^0$ *is* k*-sparse, and* $\mathbf{A}$ *satisfies RIP with* $\delta_{2k} \leq 0.4404$*, and*
2. $\alpha \geq 10\|\mathbf{x}^0\|_\infty$,

(L1+LS) *uniquely recovers* $\mathbf{x}^0$.

Example: RIP-based recovery guarantees

For approximately sparse signals and/or noisy measurements, solve:

$$\underset{\mathbf{x}}{\text{minimize}} \left\{ \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \le \sigma \right\} \tag{1}$$

Example: RIP-based recovery guarantees

For approximately sparse signals and/or noisy measurements, solve:

$$\underset{\mathbf{x}}{\text{minimize}} \left\{ \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \le \sigma \right\} \tag{1}$$

Theorem (stable recovery, Lai and Yin [2012])

Let $\mathbf{x}^0$ be an arbitrary vector, $\mathcal{S} = \{$largest k entries of $\mathbf{x}^0\}$, and $\mathcal{Z} = \mathcal{S}^C$. Let $\mathbf{b} := A\mathbf{x}^0 + \mathbf{n}$, where $\mathbf{n}$ is arbitrary noisy. If $\mathbf{A}$ satisfies RIP with $\delta_{2k} \le 0.3814$ and $\alpha \ge 10\|\mathbf{x}^0\|_\infty$, then the solution $\mathbf{x}^*$ of (1) with $\sigma = \|\mathbf{n}\|_2$ satisfies

$$\|\mathbf{x}^* - \mathbf{x}^0\|_2 \le \bar{C}_1 \cdot \|\mathbf{n}\|_2 + \bar{C}_2 \cdot \|\mathbf{x}^0_{\mathcal{Z}}\|_1/\sqrt{k},$$

where $\bar{C}_1$, and $\bar{C}_2$ are constants depending on δ_{2k}.

Other types of conditions:

- null space property (NSP),
- spherical section property (SSP),
- RIPless property,
- ...

They together offer a large variety of CS matrices.

$\alpha \geq 10\|\mathbf{x}^0\|_\infty$ also enables these properties to provide recovery guarantees.

Summary of properties

$$\text{(L1+LS)} \quad \text{minimize } \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 \quad \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}$$

- Exact regularization: get exact ℓ_1 solution if $\frac{1}{2\alpha}$ is small enough..
- CS recovery is stable given $\alpha \geq 10\|\mathbf{x}^0\|_\infty$ plus typical conditions on $\mathbf{A}$.
- Dual is unconstrained and C^1: gradient-based methods applicable
- Restricted strong convexity: weaker than strong convexity, applies to more functions, yet gives the same optimization complexity.

Summary of properties

$$\text{(L1+LS)} \quad \text{minimize}\, \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 \quad \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}$$

- Exact regularization: get exact ℓ_1 solution if $\frac{1}{2\alpha}$ is small enough..
- CS recovery is stable given $\alpha \geq 10\|\mathbf{x}^0\|_\infty$ plus typical conditions on $\mathbf{A}$.
- Dual is unconstrained and C^1: gradient-based methods applicable
- Restricted strong convexity: weaker than strong convexity, applies to more functions, yet gives the same optimization complexity.

Lagrangian duality

Convex problem: $\operatorname{minimize}_{\mathbf{x}} h(\mathbf{x}) \quad$ s.t. $\mathbf{A}\mathbf{x} = \mathbf{b}$.

Lagrange relaxation: $\mathcal{L}(\mathbf{x}; \mathbf{y}) = h(\mathbf{x}) + \mathbf{y}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$.

Dual function: $f(\mathbf{y}) = \operatorname{minimize}_{\mathbf{x}} \mathcal{L}(\mathbf{x}; \mathbf{y})$.

Lagrange dual problem:

$$\underset{\mathbf{y}}{\operatorname{maximize}}\, f(\mathbf{y}) \quad \text{or} \quad \underset{\mathbf{y}}{\operatorname{minimize}} -f(\mathbf{y}).$$

(This talk uses the latter dual problem, i.e., the min problem.)

Given optimal $\mathbf{y}^*$, under some conditions, recover $\mathbf{x}^* \leftarrow \operatorname{minimize}_{\mathbf{x}} \mathcal{L}(\mathbf{x}; \mathbf{y}^*)$.

The Lagrangian dual of (L1+LS)

Primal:

$$\underset{\mathbf{x}}{\text{minimize}}\, \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 \quad \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}$$

The Lagrangian dual of (L1+LS)

Primal:

$$\underset{\mathbf{x}}{\text{minimize}}\, \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 \quad \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}$$

Lagrange dual:

$$\underset{\mathbf{y}}{\text{minimize}}\, -\mathbf{b}^\top \mathbf{y} + \frac{\alpha}{2}\|\mathbf{A}^\top \mathbf{y} - \mathrm{Proj}_{[-1,1]^n}(\mathbf{A}^\top \mathbf{y})\|_2^2.$$

The Lagrangian dual of (L1+LS)

Primal:

$$\underset{\mathbf{x}}{\text{minimize}}\, \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 \quad \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{b}$$

Lagrange dual:

$$\underset{\mathbf{y}}{\text{minimize}}\, -\mathbf{b}^\top \mathbf{y} + \frac{\alpha}{2}\|\mathbf{A}^\top \mathbf{y} - \mathrm{Proj}_{[-1,1]^n}(\mathbf{A}^\top \mathbf{y})\|_2^2.$$

Theorem (*Convex Analysis*, Rockafellar [1970])

If a convex program has a strictly convex objective, it has a unique solution and its Lagrangian dual program is differentiable.

Therefore, let us apply gradient descent to the dual.

Dual objective: $f(\mathbf{y}) = -\mathbf{b}^\top \mathbf{y} + \frac{\alpha}{2}\|\mathbf{A}^\top \mathbf{y} - \mathrm{Proj}_{[-1,1]^n}(\mathbf{A}^\top \mathbf{y})\|_2^2$;

Introduce: $\mathrm{shrink}(\mathbf{z}) = \mathbf{z} - \mathrm{Proj}_{[-1,1]^n}(\mathbf{z})$; an entry-wise operator

Gradient: $\nabla f(\mathbf{y}) = -\mathbf{b} + \alpha \mathbf{A}\,\mathrm{shrink}(\mathbf{A}^\top \mathbf{y})$;

Dual gradient descent: $\mathbf{y}^k \leftarrow \mathbf{y}^{k-1} - c\nabla f(\mathbf{y})$;

(dual ascent is more typical, but we use the minimization dual problem)

Recover: $\mathbf{x}^k = \alpha\,\mathrm{shrink}(\mathbf{A}^\top \mathbf{y}^k)$.

Linearized Bregman vs. FISTA

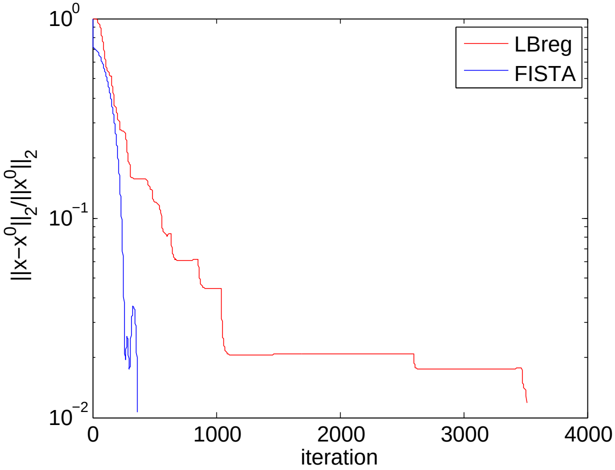
Compare

- Linearized Bregman, i.e., dual gradient descent applied to (L1+LS).
- FISTA (Beck-Teboulle'09), primal accelerated prox-linear iteration.

(They have comparable per-iteration complexities.)

Simulation:

- $\mathbf{A}$ is 256×512 i.i.d. Gaussian;
- $\mathbf{x}^0$ has 50 Gaussian nonzeros;
- Gaussian noise is added to $\mathbf{A}\mathbf{x}^0$.



Linearized Bregman is much slower than FISTA!

Can we accelerate linearized Bregman?

Two classical classes of functions

- $\mathcal{F}_L(\mathbb{R}^n)$: convex, C^1, Lipschitz ∇f:

$$\|\nabla f(u) - \nabla f(v)\| \leq L\|u - v\|, \quad \forall u, v \in \mathbb{R}^n,$$

  where $L > 0$.

- $\mathcal{S}_{L,\mu}(\mathbb{R}^n)$: the *subclass* of $\mathcal{F}_L(\mathbb{R}^n)$ with strongly convex f:

$$\langle \nabla f(u) - \nabla f(v), u - v \rangle \geq \mu \|u - v\|^2, \quad \forall u, v \in \mathbb{R}^n,$$

  where $L \geq \mu > 0$.

worst-case # of iterations to reach ϵ-accuracy in objective

function class	gradient descent complexity	optimal complexity
$\mathcal{F}_L$	$O\left(\frac{L}{\epsilon}\right)$	$O\left(\sqrt{\frac{L}{\epsilon}}\right)$
$\mathcal{S}_{L,\mu}$	$O\left(\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$

A new class of function

- $\mathcal{R}_{L,\nu}(\mathbb{R}^n)$: the subclass of $\mathcal{F}_L(\mathbb{R}^n)$ with *restricted* strongly cvx f:

$$\langle \nabla f(u) - \nabla f(v), u - v \rangle \geq \nu \|u - v\|^2, \quad \forall u \in \mathbb{R}^n, v = \mathrm{Proj}_{X^*}(u),$$

where X^* is the set of minimizers of f, assumed to be non-empty.

A new class of function

- $\mathcal{R}_{L,\nu}(\mathbb{R}^n)$: the subclass of $\mathcal{F}_L(\mathbb{R}^n)$ with *restricted* strongly cvx f:

$$\langle \nabla f(u) - \nabla f(v), u - v \rangle \geq \nu \|u - v\|^2, \quad \forall u \in \mathbb{R}^n, v = \mathrm{Proj}_{X^*}(u),$$

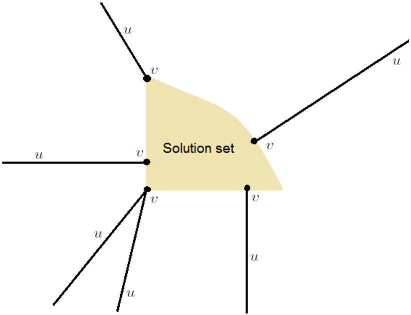
  where X^* is the set of minimizers of f, assumed to be non-empty.

In comparison,

- $\mathcal{S}_{L,\mu}(\mathbb{R}^n)$: the subclass of $\mathcal{F}_L(\mathbb{R}^n)$ with strongly convex f:

$$\langle \nabla f(u) - \nabla f(v), u - v \rangle \geq \mu \|u - v\|^2, \quad \forall u, v \in \mathbb{R}^n.$$

Illustration:

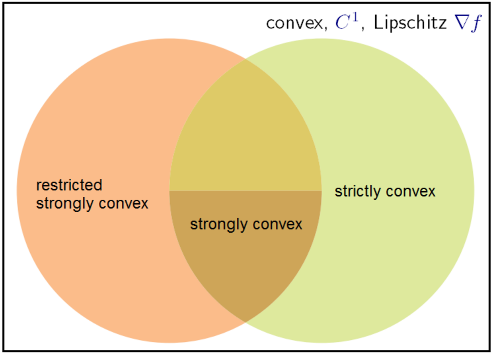


The curvature inequality

$$\langle \nabla f(u) - \nabla f(v), u - v \rangle \geq \nu \|u - v\|^2$$

only holds between a point u and its project v to the solution set.

It does *not* need to hold between any two points on the same ray or two points across different rays.


convex, C^1, Lipschitz ∇f
restricted strongly convex
strongly convex
strictly convex

Examples of $\mathcal{R}_{L,\nu}$

- (Lai-Yin'10) Dual objective of (L1+LS)

$$-\mathbf{b}^\top \mathbf{y} + \frac{\alpha}{2}\|\mathbf{A}^\top \mathbf{y} - \mathrm{Proj}_{[-1,1]^n}(\mathbf{A}^\top \mathbf{y})\|_2^2.$$

  is in $\mathcal{R}_{L,\nu}$ with $\nu > 0$ depending on α, nonzeros of $\mathbf{x}^*$, and spectral properties of $\mathbf{A}$. An explicit formula is available.

- (Zhang-Yin-Cheng'12) Let $g \in \mathcal{S}_{L,\mu}$, and define

$$f(x) = g(Ex) + c^\top x.$$

  Then $f \in \mathcal{R}_{L,\nu}$ with $\nu = \mu / \|E^\dagger\|^2$ as long as $c \in \mathrm{Range}(E)$.

- If a function is strictly convex and has restricted Lipschitz subgradient,

$$L\langle p - q, x\rangle \geq \|p - q\|^2, \ \forall p \in \partial f(x), \ q = \mathrm{Proj}_{\partial f(0)}(p),$$

  its convex conjugate is $\mathcal{R}_{L^{-1},\cdot}$.

Theorem (Zhang-Yin-Cheng'12)

If $f \in \mathcal{R}_{L,\nu}(\mathbb{R}^n)$, then gradient descent with step size $\frac{1}{L}$, then

$$\operatorname{dist}(x^k, X^*) = O\left((1-(\nu/L))^{k/2}\right)$$

and

$$f(x^k) - f^* = O\left((1-(\nu/L))^k\right).$$

Hence, it reaches an ϵ-solution in at most

$$O\left(\frac{L}{\nu}\log\frac{1}{\epsilon}\right) \text{ iterations.}$$

Theorem (Zhang-Yin-Cheng'12)

If $f \in \mathcal{R}_{L,\nu}(\mathbb{R}^n)$, then Nesterov's accelerated gradient descent with fixed restart reaches an ϵ-solution in at most

$$O\left(\sqrt{\frac{L}{\nu}}\log\frac{1}{\epsilon}\right) \text{ iterations.}$$

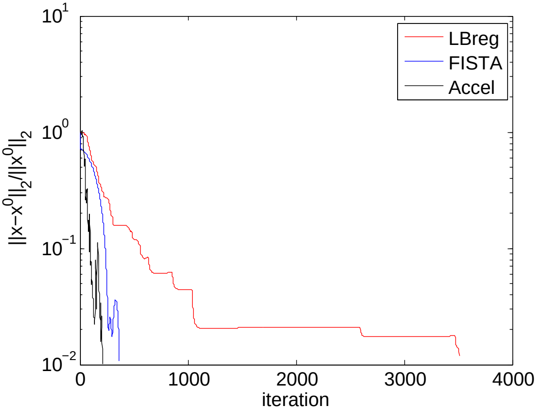
worst-case # of iterations to reach ϵ-accuracy in objective

function class	gradient descent complexity	optimal complexity
$\mathcal{F}_L$	$O\left(\frac{L}{\epsilon}\right)$	$O\left(\sqrt{\frac{L}{\epsilon}}\right)$
$\mathcal{R}_{L,\nu}$	$O\left(\frac{L-\nu}{\nu}\log\frac{1}{\epsilon}\right)$	$O\left(\sqrt{\frac{L}{\nu}}\log\frac{1}{\epsilon}\right)$
$\mathcal{S}_{L,\mu}$	$O\left(\frac{L-\mu}{\mu}\log\frac{1}{\epsilon}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$

$\mathcal{R}_{L,\nu}$ is weaker than $\mathcal{S}_{L,\mu}$ but enjoys similar rates of convergence.

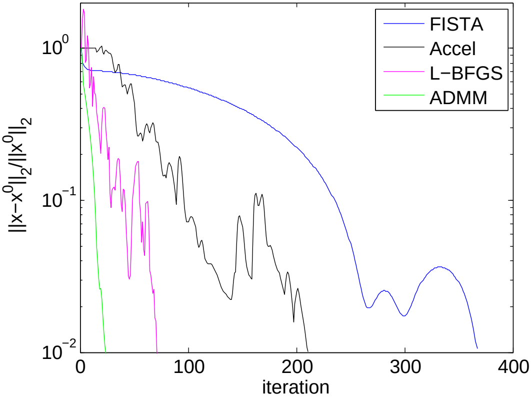
Numerical simulation

- LBreg: dual gradient descent
- FISTA: accelerated primal prox-linear iteration
- Accel: accelerated dual gradient descent



Numerical simulation

- LBreg: dual gradient descent
- FISTA: accelerated primal prox-linear iteration
- Accel: Nesterovized dual gradient descent
- L-BFGS: limited-memory quasi-Newton (use approx. 2nd-order info)
- Split-Bregman/ADMM (Moller-Yang-Osher'11)



Summary

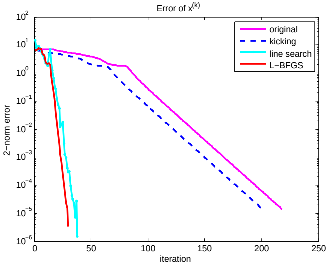
Adding $\|\mathbf{x}\|_2^2$ helps sparse optimization:

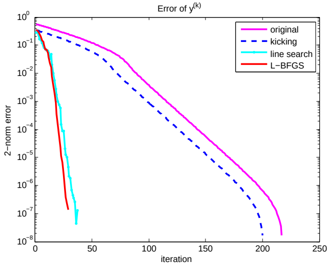
- Exact regularization and CS recovery guarantees;
- Dual is unconstrained and C^1, gradient-based methods applicable;
- Restricted strong convexity: weaker than strong convexity, but gives the same complexity;

Sparse Bernoulli Signal Test

Compare

- dual gradient descent
- dual gradient descent + kicking
- BB-step with nonmonotone line search
- L-BFGS





Global linear convergence rate

The global linear convergence rate is C^k and

$$C \approx 1 - \frac{\omega^2}{\kappa^2}$$

where

$$\omega = \min_{i \in \mathrm{supp}(\mathbf{x}^*)} \frac{|\mathbf{x}_i^*|/\alpha}{1 + |\mathbf{x}_i^*|/\alpha}$$

$$\kappa = \min \left\{ \frac{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}{\lambda_{\min}^{++}(\mathbf{C}^\top \mathbf{C})} : \mathbf{C} \text{ is subset of columns of } \mathbf{A} \right\}$$

Acknowledgements: AFOSR, DoD, NSF, ONR

References:

M.P. Friedlander and P. Tseng. Exact regularization of convex programs. *SIAM Journal on Optimization*, 18(4):1326–1350, 2007.

W. Yin. Analysis and generalizations of the linearized Bregman method. *SIAM Journal on Imaging Sciences*, 3(4):856–877, 2010.

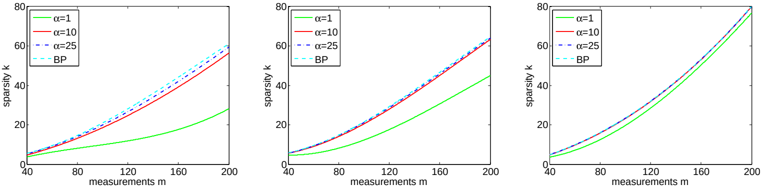
E. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

M.-J. Lai and W. Yin. Augmented ℓ_1 and nuclear-norm models with a globally linearly convergent algorithm. *Rice University CAAM Technical Report TR12-02*, 2012. [pdf].

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

Compare solution quality

$$\text{minimize } \|\mathbf{x}\|_1 \quad \text{v.s.} \quad \text{minimize } \|\mathbf{x}\|_1 + \frac{1}{2\alpha}\|\mathbf{x}\|_2^2 \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{b}$$



± 1 sparse **Gaussian** sparse **Power-law** sparse

Level curves of relative-error 10^{-3}. Higher is better.