Fast gradient algorithms for structured sparsity

Yao-Liang Yu

PhD: University of Alberta
Now: Carnegie Mellon University

June 5, CAIAC 2015 @ Dalhousie

Statistical inference 101

To estimate unknown parameter $\theta \in \mathbb{R}^p$:

$$y_i = \boldsymbol{x}_i^\top \theta + \epsilon_i, \quad i = 1, \ldots, n$$

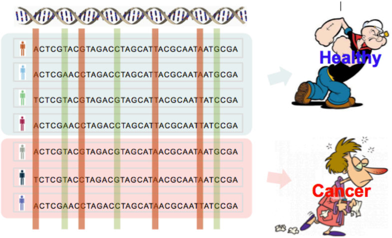
- classical setting: p fixed small, $n \to \infty$, lots of results.

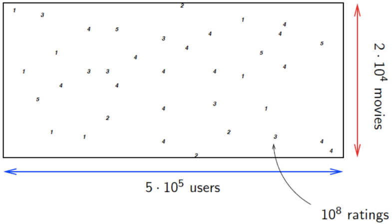
Statistical inference 101

To estimate unknown parameter $\theta \in \mathbb{R}^p$:

$$y_i = \mathbf{x}_i^\top \theta + \epsilon_i, \quad i = 1, \ldots, n$$

- classical setting: p fixed small, $n \to \infty$, lots of results.
- modern setting:





$p \sim 10^7, n \sim 10^3$ $\qquad$ $p \sim 10^{10}, n \sim 10^8$

High-dimensional challenge

- More unknown parameters than observations, ill-defined.
  - structure: effective number of unknown parameters is moderate.
    - ★ θ is sparse: $\mathrm{nnz}(\theta)$ small, but do not know which is which.
    - ★ θ as a matrix is low-rank, but do not know the column/row spaces.
- Extremely large scale, takes forever to run.
  - first order grad alg: scales (sub)linearly with problem size.
- Ideally, want algorithm to exploit structure for faster convergence.
  - open the blackbox.

  - contributions of this thesis lie in.

Table of Contents

Regularized loss minimization

Generic form for many ML problems:

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \ell(\boldsymbol{w}) + f(\boldsymbol{w}), \quad \text{where}$$

- ℓ is the loss/-likelihood function, usually smooth;
- f is the regularizer, usually nondifferentiable;
  - structure inducing

Special interest:

- sparsity (structure);
- computational efficiency.

The LASSO (Tibshirani'96)

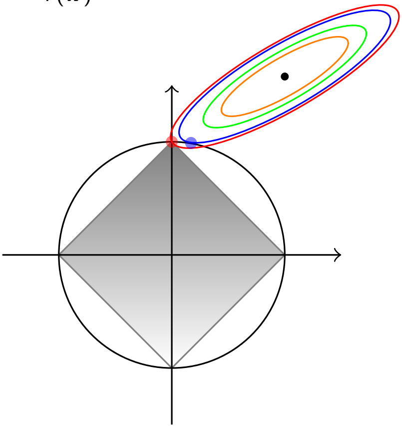
$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \underbrace{\|A\boldsymbol{w} - \boldsymbol{b}\|^2}_{\ell(\boldsymbol{w})} + \underbrace{\lambda\|\boldsymbol{w}\|_1}_{f(\boldsymbol{w})}.$$

Multiple benefits

- interpretability;
- complexity control;
- storage saving;
- perfect recovery;
- etc.

Computationally?

- convex quadratic program
- but $\mathrm{P} \neq \mathrm{E}$!
- especially when p is large.

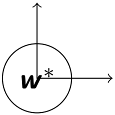


Nonsmooth optimization

Generic subgradient descent:

$$\boxed{\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta[\nabla \ell(\boldsymbol{w}_t) + \partial f(\boldsymbol{w}_t)]}$$

- guaranteed convergence, $O(1/\epsilon^2)$;



- dense iterates;
- weak regularizing effect;
- and slow, very slow...

Naum Zuselevich Shor (1937–2006)

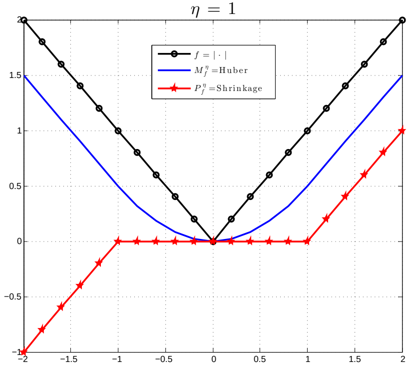
Second order methods (*e.g.* IPM) do not scale.

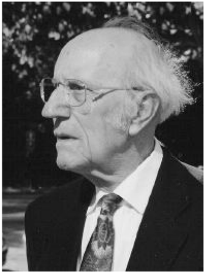
Moreau envelope and proximal map

Definition (Moreau'65)

$$\mathrm{M}_f^{\eta}(\boldsymbol{y}) = \min_{\boldsymbol{w}} \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{y}\|^2 + f(\boldsymbol{w})$$

$$\mathrm{P}_f^{\eta}(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{y}\|^2 + f(\boldsymbol{w})$$

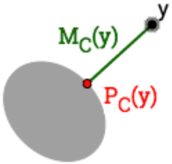




Jean Jacques Moreau, 1923–2014

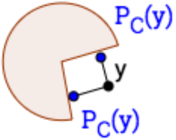
Some properties of the proximal map

- For $f(\boldsymbol{w}) = \iota_C(\boldsymbol{w}) := \begin{cases} 0, & \boldsymbol{w} \in C \\ \infty, & \text{otherwise} \end{cases}$,
  - $\mathsf{P}_f^\eta(\cdot)$ is the usual Euclidean projection onto C;
  - $\mathsf{M}_f^\eta(\cdot)$ is the (squared) distance function;
  - Both well-defined as long as C is closed.



- For f convex (and closed),
  - $\mathsf{P}_f^\eta(\cdot)$ is a nonexpansion: $\|\mathsf{P}_f^\eta(\boldsymbol{x}) - \mathsf{P}_f^\eta(\boldsymbol{y})\| \leq \|\boldsymbol{x} - \boldsymbol{y}\|$;
  - $\mathsf{M}_f^\eta(\cdot)$ is continuously differentiable;
  - $\eta \downarrow 0 \implies \mathsf{M}_f^\eta \uparrow f$.

- For general f (that decreases not too fast),
  - $\mathsf{P}_f^\eta(\cdot)$ is a nonempty compact set;
  - $\mathsf{M}_f^\eta(\cdot)$ is continuous;
  - Still $\eta \downarrow 0 \implies \mathsf{M}_f^\eta \uparrow f$.



Proximal gradient (Fukushima & Mine'81)

$$\min_{\boldsymbol{w}\in\mathbb{R}^m} \ell(\boldsymbol{w}) + f(\boldsymbol{w}), \quad \text{where} \quad \ell \in \mathcal{C}^1.$$

1. $\boldsymbol{y}_t = \boldsymbol{w}_t - \eta\nabla\ell(\boldsymbol{w}_t)$; (forward)
2. $\boldsymbol{w}_{t+1} = \mathsf{P}_f^\eta(\boldsymbol{y}_t)$. (backward)

For $f = \|\cdot\|_1$, obtain the shrinkage operator

$$[\mathsf{P}^\eta_{\|\cdot\|_1}(\boldsymbol{y})]_i = \mathrm{sign}(y_i)(|y_i| - \eta)_+.$$

- much faster, $O(1/\epsilon)$, can be accelerated;
- generalization of projected gradient: $f = \iota_C$;
- reveals the sparsity-inducing property.

Refs: Combettes & Wajs'05; Beck & Teboulle'09; Duchi & Singer'09; Nesterov'13; etc.

The good old days

CONVEX PROGRAMMING IN HILBERT SPACE

BY A. A. GOLDSTEIN[1]

Communicated by V. Klee, May 1, 1964

This note gives a construction for minimizing certain twice-differentiable functions on a closed convex subset C, of a Hilbert Space, H. The algorithm assumes one can constructively "project" points onto convex sets. A related algorithm may be found in Cheney-Goldstein [1], where a constructive fixed-point theorem is employed to construct points inducing a minimum distance between two convex sets. In certain instances when such projections are not too difficult to construct, say on spheres, linear varieties, and orthants, the method can be effective. For applications to control theory, for example, see Balakrishnan [2], and Goldstein [3].

In what follows P will denote the "projection" operator for the convex set C. This operator, which is well defined and Lipschitzian, assigns to a given point in H its closest point in C (see, e.g., [1]). Take $x \in H$ and $y \in C$. Then $[x-y,\ P(x)-y] \geqq \|P(x)-y\|^2$. In the nontrivial case this inequality is a consequence of the fact that C is supported by a hyperplane through $P(x)$ with normal $x-P(x)$. Let f be a real-valued function on H and x_0 an arbitrary point of C. Let S denote the level set $\{x \in C: f(x) \leqq f(x_0)\}$, and let $\hat{S}$ be any open set containing the convex hull of S. Let $f'(x,\ \cdot) = [\nabla f(x),\ \cdot\,]$ signify the Fréchet derivative of f at x. A point z in C will be called stationary if $P(z-\rho\nabla f(z)) = z$ for all $\rho > 0$; equivalently, when f is convex the linear functional $f'(z,\ \cdot)$ achieves a minimum on C at z.

THEOREM. *Assume f is bounded below. For each $x \in \hat{S}$, h in H and for some $\rho_0 > 0$, assume that $f'(x, h)$ exists in the sense of Fréchet, $f''(x, h, h)$ exists in the sense of Gâteaux, and $|f''(x, h, h)| \leqq \|h\|^2/\rho_0$. Choose σ and ρ_k satisfying $0 < \sigma \leqq \rho_0$ and $\sigma \leqq \rho_k \leqq 2\rho_0 - \sigma$. Set $x_{k+1} = P(x_k - \rho_k \nabla f(x_k))$. Then:*

(i) *The sequence x_k belongs to S, $(x_{k+1}-x_k)$ converges to 0, and $f(x_k)$ converges downward to a limit L.*

(ii) *If S is compact, z is a cluster point of $\{x_k\}$, and ∇f is continuous in some neighborhood of z, then z is a stationary point. If z is unique, x_k converges to z, and z minimizes f on C.*

(iii) *If S is convex and $f''(x, h, h) \geqq \mu\|h\|^2$ for each $x \in S$, $h \in H$ and some $\mu \geqq 0$, then $L = \inf\{f(x) : x \in C\}$.*

(iv) *Assume* (iii) *with S bounded. Weak cluster points of $\{x_k\}$ minimize f on C.*

[1] Present address, University of Washington, Seattle. This research was supported by grant AF-AFOSR-62-348.

(v) *Assume* (iii) *with μ positive and ∇f bounded on S. Then $f(z) = L$ for some z in S, x_k converges to z, and z is unique.*

PROOF. Assume x_k belongs to S and that x_k is not stationary. Let $\nabla f(x_k) = \nabla f_k$, $x(\rho) = P(x_k - \rho\nabla f_k)$, $\delta(\rho) = x(\rho) - x_k$ and $\Delta(\rho) = f(x_k) - f(x(\rho))$. If we notice that $-\rho[\nabla f_k,\ \delta(\rho)] \geqq \|\delta(\rho)\|^2$ and invoke Taylor's theorem, we obtain $\Delta(\rho) \geqq \|\delta(\rho)\|^2\{\rho^{-1} - f''(\xi(\rho),\ \delta(\rho),\ \delta(\rho))/2\|\delta(\rho)\|^2\}$. Here $\xi(\rho) = x_k + t\delta(\rho)$ with $t \in (0, 1)$. For some ρ sufficiently small and positive, $\Delta(\rho)$ is positive and continuous. Let β denote the least positive ρ satisfying $\Delta(\rho) = 0$, if such exists. If β exists, $\Delta(\beta) = 0$ implies that $\beta \geqq 2\rho_0$. Thus if $\sigma \leqq \rho \leqq 2\rho_0 - \sigma$, $\Delta(\rho) > 0$ and $x(\rho) \in S$, whence $\Delta(\rho_k) \geqq \|x_{k+1} - x_k\|^2\sigma/4\rho_0^2$, proving (i).

The proof of (ii) being straightforward, we proceed with the proof of (iii). Suppose that $L \neq \inf\{f(x) : x \in C\}$ and choose $z \in C$ such that $f(z) < L$. Then $0 > f(z) - f(x_k) \geqq [\nabla f_k,\ z - x_k]$. If $\liminf[\nabla f_k,\ z - x_k] = \beta$ were non-negative, a contradiction would be manifest. But the inequality $[\rho_k\nabla f_k,\ z - x_{k+1}] \geqq [x_k - x_{k+1},\ z] + [x_{k+1},\ x_{k+1} - x_k]$ holds because either $x_k - \rho_k\nabla f_k - x_{k+1}$ is the normal to C at x_{k+1}, or it is 0. If the sequence x_k is bounded, clearly $\beta = 0$; otherwise choose a subsequence satisfying $\|x_{k+1}\| > \|x_k\|$. Then $\beta \geqq 0$.

To prove (iv) we observe that f is lower semi-continuous on S if and only if the set $S_m = \{x \in S: f(x) \leqq m\}$ is closed in S for each m. Since f is convex and continuous, S_m is closed and convex, and is thus weakly closed. Hence f is weakly l.s.c. If x_k converges weakly to z, then $\liminf f(x_k) = L \geqq f(z)$.

Assume the hypotheses of (v). If $s > k$, we may write that $0 > f(x_s) - f(x_k) \geqq [\nabla f_k,\ x_s - x_k] + (1/2)\mu\|x_s - x_k\|^2$, whence $\{x_s\}$ is bounded. Invoking again the supporting hyperplane at x_{k+1}, $[\rho_k\nabla f_k,\ x_s - x_k] \geqq [\rho_k\nabla f_k,\ x_{k+1} - x_k] + [x_{k+1} - x_k,\ x_{k+1} - x_s]$. Thus when k is sufficiently large $\|x_s - x_k\| < \epsilon$. There exists therefore $z \in S$ minimizing f on C, and $f(x) \geqq f(z) + [\nabla f(z),\ x - z] + (1/2)\mu\|x - z\|^2$. Since $[\nabla f(z),\ x - z] \geqq 0$, $f(x) - f(z) \geqq (1/2)\mu\|x - z\|^2$; and therefore z is unique.

REFERENCES

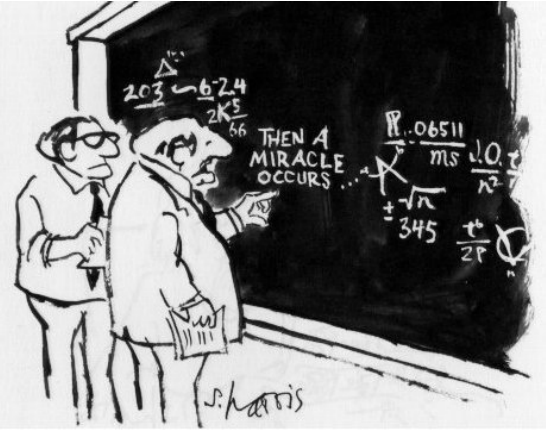
1. E. W. Cheney and A. A. Goldstein, *Proximity maps for convex sets*, Proc. Amer. Math. Soc. 10 (1959), 448–450.
2. A. V. Balakrishnan, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, J. SIAM Control Ser. A 1 (1963), 109–127.
3. A. A. Goldstein, *Minimizing functionals on Hilbert space*, Computer methods in optimization problems, Academic Press, New York, 1964, pp. 159–165.

UNIVERSITY OF TEXAS

Modern significance & rediscovery

- Donoho & Johnstone (90s), wavelet shrinkage;
- Starck, Donoho, and Candès (2003), astronomical image representation;
- Figueiredo & Nowak (2003), image restoration;
- Daubechies, Defrise, and De Mol (2004), inverse problem.
- Many many more...

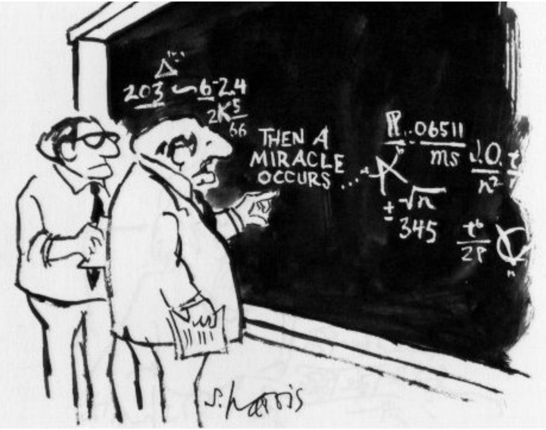
However...



"I think you should be more explicit here in step two."

from *What's so Funny about Science?* by Sidney Harris (1977)

However...



"I think you should be more explicit here in step two."

from *What's so Funny about Science?* by Sidney Harris (1977)

$$\text{Step 2:}\ \mathsf{P}_f^{\eta}(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2\eta}\|\boldsymbol{y} - \boldsymbol{w}\|^2 + f(\boldsymbol{w})$$

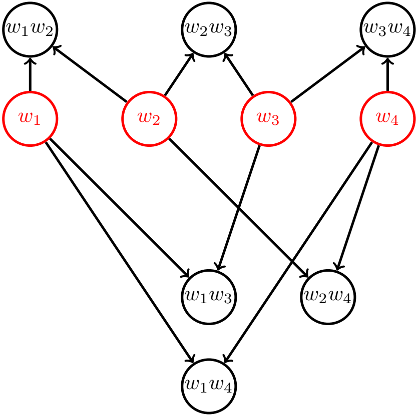
Structured sparsity: group

Group level sparse regularizer

$$f(\boldsymbol{w}) = \sum_i \|\boldsymbol{w}\|_{\mathrm{g}_i}.$$

For P_f, when groups are

- non-overlapping: decouple;
- tree structured: decompose;
- arbitrary?



Refs: Bakin'99; Yuan & Lin'06; Zhao et al.'09; etc.

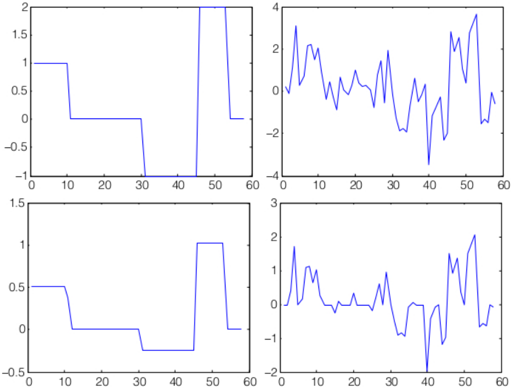
Structured sparsity: graph

Neighborhood sparse regularizer

$$f(\boldsymbol{w}) = \sum_{\{i,j\}\in E} |w_i - w_j|.$$

For P_f, when graph is

- a chain: DP;
- arbitrary?
- vector valued?

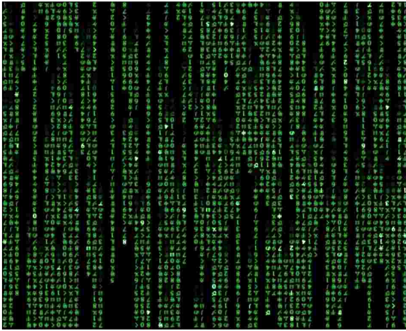


Refs: Tibshirani et al.'05; Kim et al.'09; Kim & Xing'09; Hoefling'10; etc.

Structured sparsity: matrix

- Matrix completion:

$$\min_{X \in \mathbb{R}^{m \times n}} \underbrace{\sum_{(i,j) \in \mathcal{O}} (X_{ij} - Z_{ij})^2}_{\ell(X)} + \underbrace{\lambda \|X\|_{\mathrm{tr}}}_{f(X)}.$$



- Can apply PG:

$$\mathrm{P}^{\eta}_{\lambda\|\cdot\|_{\mathrm{tr}}}(Y) = \sum_k (\sigma_k - \lambda\eta)_+ \boldsymbol{u}_k \boldsymbol{v}_k^\top.$$

- Require full SVD in each step.

Refs: Candès & Recht'09; Cai et al.'10; Pong et al.'10; Toh & Yun'10; Ma et al.'11; etc.

Learned so far

- Proximal gradient is simple, efficient, and structure-friendly.
  - easily parallelizable, can randomize, can block-wise.
- But backward step (proximal map) not always easy/cheap.
  - decompose;
  - approximate;
  - bypass proximal gradient;
- Constant theme: exploit the structure of your problem!
  - statistically;
  - and computationally.

Table of Contents

How to decompose?

- Typical structured sparse regularizers:

$$f(\boldsymbol{w}) = \sum_i f_i(\boldsymbol{w});$$

  - Also applies to ERM, each i is a sample.
- Key observation: each P_{f_i} is easy to compute.
- Can we compute $\mathrm{P}_f = \mathrm{P}_{\sum_i f_i}$ efficiently?

Theorem (Folklore)

$$\mathrm{P}_{f+g} = (\mathrm{P}_{2f}^{-1} + \mathrm{P}_{2g}^{-1})^{-1} \circ (2\mathrm{Id}).$$

- Not directly useful;
- Can numerically reduce to P_f and P_g (Combettes et al.'11);
- But a two-loop routine can be as slow as subgradient (Villa et al.'13).

Two previous results

$$\|\boldsymbol{w}\|_{\mathrm{TV}} = \sum_{i=1}^{p} |w_i - w_{i+1}| .$$

Theorem (Friedman et al.'07)

$$\mathsf{P}_{\|\cdot\|_1 + \|\cdot\|_{\mathrm{TV}}} = \mathsf{P}_{\|\cdot\|_1} \circ \mathsf{P}_{\|\cdot\|_{\mathrm{TV}}}.$$

Theorem (Jenatton et al.'11)

$$\mathsf{P}_{\sum_{i=1}^{k} \|\cdot\|_{\mathsf{g}_i}} = \mathsf{P}_{\|\cdot\|_{\mathsf{g}_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{\mathsf{g}_k}}.$$

Generalization

$$\mathsf{P}_{f+g} \stackrel{?}{=} \mathsf{P}_f \circ \mathsf{P}_g \stackrel{?}{=} \mathsf{P}_g \circ \mathsf{P}_f.$$

But, is it even sensible?

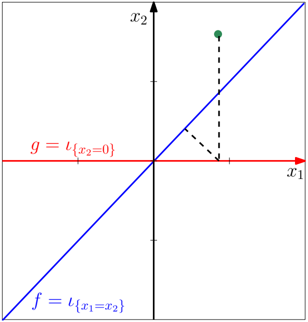
Good news and bad news

Theorem

On the real line, $\exists h$ *such that* $\mathsf{P}_h = \mathsf{P}_f \circ \mathsf{P}_g$.

Example (But not so in general...)

Consider $\mathbb{R}^2$, and let $f = \iota_{\{x_1 = x_2\}}, g = \iota_{\{x_2 = 0\}}$.



Nevertheless

- Can ask the decomposition to hold for many but not all cases.
- Setting the subdifferential to 0:

$$\begin{aligned} \mathsf{P}_{f+g}(\boldsymbol{z}) - \boldsymbol{z} + \partial(f+g)(\mathsf{P}_{f+g}(\boldsymbol{z})) &\ni 0 \\ \mathsf{P}_{g}(\boldsymbol{z}) - \boldsymbol{z} + \partial g(\mathsf{P}_{g}(\boldsymbol{z})) &\ni 0 \\ \mathsf{P}_{f}(\mathsf{P}_{g}(\boldsymbol{z})) - \mathsf{P}_{g}(\boldsymbol{z}) + \partial f(\mathsf{P}_{f}(\mathsf{P}_{g}(\boldsymbol{z}))) &\ni 0. \end{aligned}$$

- Adding the last two equations we obtain

$$\mathsf{P}_{f}(\mathsf{P}_{g}(\boldsymbol{z})) - \boldsymbol{z} + \partial g(\mathsf{P}_{g}(\boldsymbol{z})) + \partial f(\mathsf{P}_{f}(\mathsf{P}_{g}(\boldsymbol{z}))) \ni 0.$$

Theorem (Y'13a)

A sufficient condition for $\mathsf{P}_{f+g}(\boldsymbol{z}) = \mathsf{P}_{f}\big(\mathsf{P}_{g}(\boldsymbol{z})\big)$ is

$$\forall\ \boldsymbol{y} \in \operatorname{dom} g,\ \ \partial g(\mathsf{P}_{f}(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y}).$$

The rest is easy

- Find f and g that clinch our sufficient condition.

Result I: Start with "trivialities"

Theorem (Y'13a)

Fix $f \in \Gamma_0$. $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ for all $g \in \Gamma_0$ if and only if

- $\dim(\mathcal{H}) \geq 2$; *$f \equiv c$ or $f = \iota_{\{\mathbf{w}\}} + c$ for some $c \in \mathbb{R}$ and $\mathbf{w} \in \mathcal{H}$;*
- $\dim(\mathcal{H}) = 1$ *and $f = \iota_C + c$ for some closed and convex set C and $c \in \mathbb{R}$.*

Asymmetry.

Theorem (Y'13a)

Fix $g \in \Gamma_0$. $\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g$ for all $f \in \Gamma_0$ if and only if g is a continuous affine function.

Result II: Positive homogeneity and "roundness"

Theorem (Y'13a)

Let $f \in \Gamma_0$. The following are equivalent (provided $\dim(\mathcal{H}) \geq 2$):

i). $f = h(\|\cdot\|)$ *for some increasing function* $h : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$;

ii). $\boldsymbol{x} \perp \boldsymbol{y} \implies f(\boldsymbol{x} + \boldsymbol{y}) \geq f(\boldsymbol{y})$;

iii). *For all* $\boldsymbol{z} \in \mathcal{H}$, $\mathsf{P}_f(\boldsymbol{z}) = \lambda_{\boldsymbol{z}} \cdot \boldsymbol{z}$ *for some* $\lambda_{\boldsymbol{z}} \in [0, 1]$;

iv). $\mathbf{0} \in \operatorname{dom} f$ *and* $\mathsf{P}_{f+\kappa} = \mathsf{P}_f \circ \mathsf{P}_\kappa$ *for all p.h. functions* $\kappa \in \Gamma_0$.

- Include and generalize many results;
- Connects to the representer theorem in kernel methods (YCSS'13).

More implications

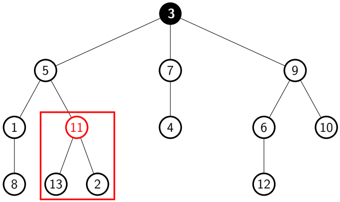
Example (Elastic net, Zou & Hastie'05)

$$\mathsf{P}_{\lambda\|\cdot\|_2^2+\kappa} = \mathsf{P}_{\lambda\|\cdot\|_2^2} \circ \mathsf{P}_{\kappa} = \frac{1}{\lambda+1}\mathsf{P}_{\kappa} \qquad \text{double shrinkage}$$

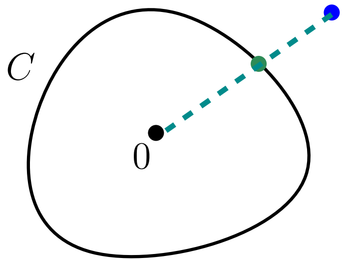
Example (Jenatton et al.'11)

Tree-structured (laminar system)

$$\mathsf{P}_{\sum_i \|\cdot\|_{g_i}} = \mathsf{P}_{\|\cdot\|_{g_1}} \circ \cdots \circ \mathsf{P}_{\|\cdot\|_{g_k}}.$$



Characterizing the ball



Result III: Comonotonicity and Choquet integral

Initially case by case for many polyhedral regularizers.

Theorem (Y'13a)

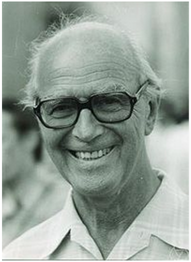
Let f be permutation invariant and g be the Choquet integral of some submodular set function.

$$\mathsf{P}_{f+g} = \mathsf{P}_f \circ \mathsf{P}_g.$$

Example (Friedman et al.'07)

$$\mathsf{P}_{\|\cdot\|_1 + \|\cdot\|_{\mathsf{TV}}} = \mathsf{P}_{\|\cdot\|_1} \circ \mathsf{P}_{\|\cdot\|_{\mathsf{TV}}}.$$

- $\|\cdot\|_1$: permutation invariant;
- $\|\cdot\|_{\mathsf{TV}}$: Choquet integral of something.



Gustave Choquet (1915–2006)

"Always consider a problem under the minimum structure in which it makes sense."

Summary

- Posed the question: $\mathsf{P}_{f+g} \stackrel{?}{=} \mathsf{P}_f \circ \mathsf{P}_g \stackrel{?}{=} \mathsf{P}_g \circ \mathsf{P}_f$;
- Presented a sufficient condition: $\partial g(\mathsf{P}_f(\boldsymbol{y})) \supseteq \partial g(\boldsymbol{y})$;
- "Trivial" case;
- Positive homogeneity and "roundness";
- Comonotonicity and Choquet integral;
- Immediately useful if plugged into PG;

$\boxed{\text{What if the sufficient condition fails?}}$

Table of Contents

More generally

Recall: typical structured sparse regularizers: $\bar{f} = \sum_i \alpha_i f_i$

- $\mathrm{P}_{f_i}^{\eta}$ easy to compute;
- f_i Lipschitz continuous.

Example (Overlapping group lasso, Zhao et al.'09)

$$f_i(\boldsymbol{w}) = \|\boldsymbol{w}\|_{\mathrm{g}_i} \text{ where } \mathrm{g}_i \text{ is a group (subset) of variables.}$$

- When the groups overlap arbitrarily, $\mathrm{P}_{\bar{f}}^{\eta}$ cannot be easily computed;
- Each f_i is 1-Lipschitz continuous *w.r.t.* $\|\cdot\|$;
- The proximal map $\mathrm{P}_{f_i}^{\eta}$ is simply a re-scaling:

$$[\mathrm{P}_{f_i}^{\eta}(\boldsymbol{w})]_j = \begin{cases} w_j, & j \notin \mathrm{g}_i \\ (1 - \eta/\|\boldsymbol{w}\|_{\mathrm{g}_i})_+ w_j, & j \in \mathrm{g}_i \end{cases}.$$

Example cont'

Example (Graph-guided fused lasso, Kim & Xing'09)

Given some graph, we let $f_{ij}(\boldsymbol{w}) = |w_i - w_j|$ for every edge $\{i, j\}$.

- For a general graph, the proximal map of the regularizer $\bar{f} = \sum_{\{i,j\}\in E} \alpha_{ij} f_{ij}$ can not be easily computed;
- Each f_{ij} is 1-Lipschitz continuous *w.r.t.* the Euclidean norm;
- The proximal map $\mathsf{P}^{\eta}_{f_{ij}}$ is easy to compute:

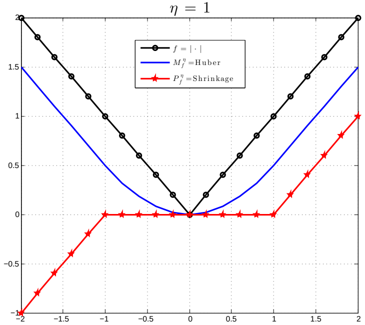
$$[\mathsf{P}^{\eta}_{f_{ij}}(\boldsymbol{w})]_s = \begin{cases} w_s, & s \notin \{i, j\} \\ w_s - \mathrm{sign}(w_i - w_j)\min\{\eta, |w_i - w_j|/2\}, & s \in \{i, j\} \end{cases}.$$

Other examples abound.

Smoothing (Nesterov'05)

$$\mathrm{M}_f^\eta(\boldsymbol{y}) = \min_{\boldsymbol{w}} \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{y}\|^2 + f(\boldsymbol{w})$$

$$\mathrm{P}_f^\eta(\boldsymbol{y}) = \operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{y}\|^2 + f(\boldsymbol{w})$$



Proposition (Nesterov'05)

If f is L-Lipschitz continuous, then $0 \le f - \mathrm{M}_f^\eta \le \eta L^2/2$.

In retrospect

Suppose want: $$\min_{\boldsymbol{w} \in C} \ell(\boldsymbol{w}).$$

Same for large $\lambda > 0$: $$\min_{\boldsymbol{w}} \ell(\boldsymbol{w}) + \lambda \cdot \mathrm{dist}(\boldsymbol{w}, C)$$

- $\mathrm{dist}(\boldsymbol{w}, C) := \min_{\boldsymbol{z} \in C} \|\boldsymbol{w} - \boldsymbol{z}\|$, nonsmooth but Lipschitz continuous.
- Can smooth dist and apply gradient descent.
- But nobody does that, overkill.
- Can just use projected gradient.

A "naive" idea (Y'13b)

$$\bar{f} = \sum_i \alpha_i f_i$$

$\Downarrow$ as if have linearity?

$$\mathsf{P}^{\eta}_{\bar{f}} \approx \sum_i \alpha_i \mathsf{P}^{\eta}_{f_i}$$

Definition (Proximal Average, Moreau'65; Bauschke et al.'08)

There exists a unique function A^{η} such that $\mathsf{P}^{\eta}_{\mathsf{A}^{\eta}} = \sum_i \alpha_i \mathsf{P}^{\eta}_{f_i}$.

What mathematicians call a "picture"

$$
\begin{tikzcd}
\Gamma_0 \ni f_i \arrow[r, "\text{onto}"] & \mathsf{M}^\eta_{f_i} \in \mathsf{SS}_{1/\eta} \arrow[d, "\text{convex}"] \\
\Gamma_0 \ni \mathsf{A}^\eta \arrow[d, "\nabla"'] & \sum_i \alpha_i \mathsf{M}^\eta_{f_i} \in \mathsf{SS}_{1/\eta} \arrow[l, "\text{1-1}"'] \arrow[d, "\nabla"] \\
\mathsf{P}^\eta_{\mathsf{A}^\eta} \arrow[r, Leftrightarrow] & \sum_i \alpha_i \mathsf{P}^\eta_{f_i}
\end{tikzcd}
$$

- Not so easy to compute A^η, but existence is enough.

The algorithm

Dream

1. $\mathbf{z}_t = \mathbf{w}_t - \mu \nabla \ell(\mathbf{w}_t)$
2. $\mathbf{w}_{t+1} = \mathsf{P}^{\eta}_{\bar{f}}(\mathbf{z}_t)$

$\Downarrow$

$$\min_{\mathbf{w}} \ell(\mathbf{w}) + \bar{f}(\mathbf{w})$$

Reality

1. $\mathbf{z}_t = \mathbf{w}_t - \mu \nabla \ell(\mathbf{w}_t)$
2. $\mathbf{w}_{t+1} = \mathsf{P}^{\eta}_{\mathsf{A}^{\eta}}(\mathbf{z}_t) = \sum_i \alpha_i \mathsf{P}^{\eta}_{f_i}(\mathbf{z}_t)$

$\Downarrow$

$$\min_{\mathbf{w}} \ell(\mathbf{w}) + \mathsf{A}^{\eta}(\mathbf{w})$$

When are they close?

Nonsmooth approximation

- How good the proximal average A^η approximates $\bar{f}$?

Proposition (Uniform lower approximation)

Assuming f_i is M_i-Lipschitz continuous, and $M := \sum_i \alpha_i M_i^2$, then

$$0 \le \bar{f} - \mathsf{A}^\eta \le \eta M^2/2.$$

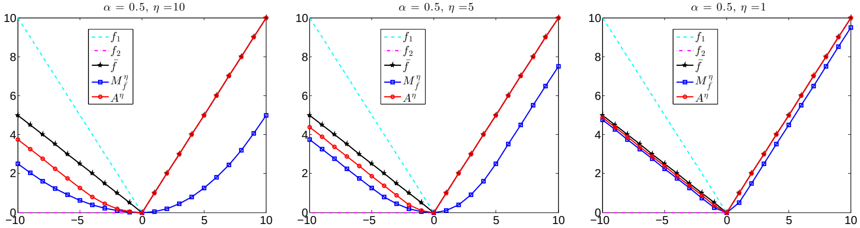
- Proximal average is a tighter approximation than smoothing:

$$\sum_i \alpha_i \mathsf{M}_{f_i}^\eta \le \mathsf{A}^\eta \le \bar{f}.$$

An example

Example

Consider $f_1(x) = |x|$, and $f_2(x) = \max\{x, 0\}$.



- The proximal average is smooth iff some f_i is;
- Essentially we de-smooth Nesterov's approximation.

Convergence guarantee

Theorem (Y'13b)

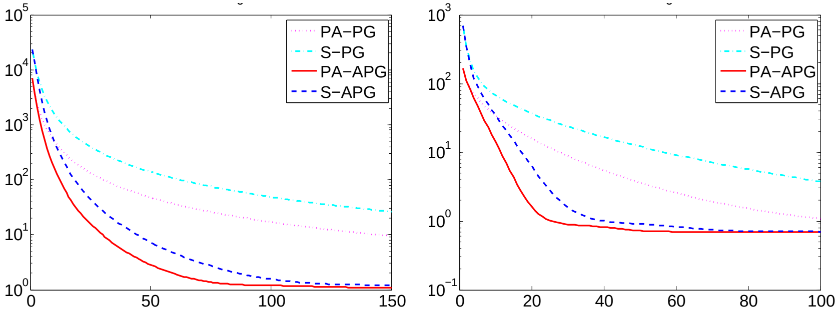
Using a suitable step size, we get an ϵ-accurate solution in at most $O(\sqrt{\max\{L_0, L^2/(2\epsilon)\}}\sqrt{1/\epsilon})$ steps.

- Improves Nesterov's complexity $O(\sqrt{L_0 + L^2/(2\epsilon)}\sqrt{1/\epsilon})$ by removing secondary term;
- No overhead, same assumption, strict improvement;
- Simple update rule.

$$\text{S-PG: } \boldsymbol{w}_{t+1} = \frac{\eta L_0}{1+\eta L_0}\left[\boldsymbol{w}_t - \frac{1}{L_0}\nabla\ell(\boldsymbol{w}_t)\right] + \frac{1}{1+\eta L_0}\sum_i \alpha_i \mathrm{P}^{\eta}_{f_i}(\boldsymbol{w}_t),$$

$$\text{PA-PG: } \boldsymbol{w}_{t+1} = \sum_i \alpha_i \mathrm{P}^{\eta}_{f_i}(\boldsymbol{w}_t - \eta\nabla\ell(\boldsymbol{w}_t)).$$

Experiment



PA−PG
S−PG
PA−APG
S−APG
10^5
10^4
10^3
10^2
10^1
10^0
0
50
100
150
PA−PG
S−PG
PA−APG
S−APG
10^3
10^2
10^1
10^0
10^{-1}
0
20
40
60
80
100

Summary

- Linear approximation of the proximal map;
- Improved convergence guarantee;
- Retain nonsmoothness (to some extent);
- How to combine regularizers?

Table of Contents

Conditional gradient (Frank-Wolfe'56)

$$\min_{\boldsymbol{w} \in C} \ \ell(\boldsymbol{w})$$

- C: compact convex;
- ℓ: smooth convex.

$$\boxed{\begin{array}{ll} 1 & \boldsymbol{y}_t \in \operatorname*{argmin}_{\boldsymbol{w} \in C} \langle \boldsymbol{w}, \nabla \ell(\boldsymbol{w}_t) \rangle; \\ 2 & \boldsymbol{w}_{t+1} = (1-\eta)\boldsymbol{w}_t + \eta \boldsymbol{y}_t. \end{array}}$$

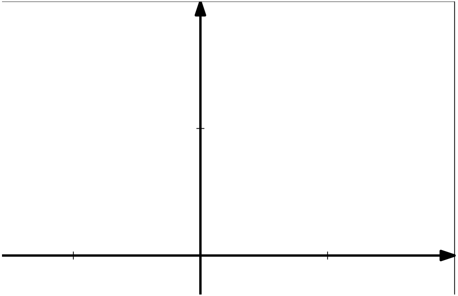
Gained much recent attention due to

- its simplicity;
- the greedy nature in step 1.

Refs: Zhang'03; Clarkson'10; Hazan'08; Jaggi-Sulovsky'10; etc.

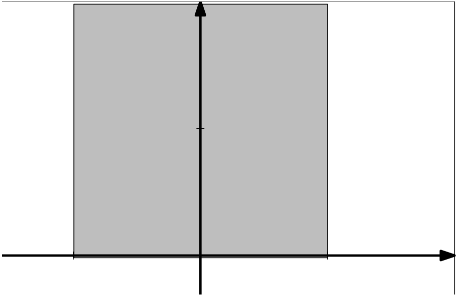
An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



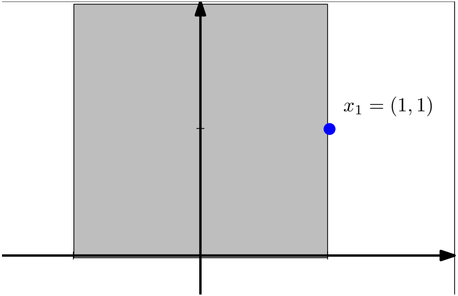
An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



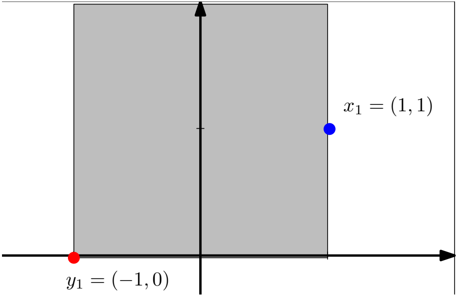
An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



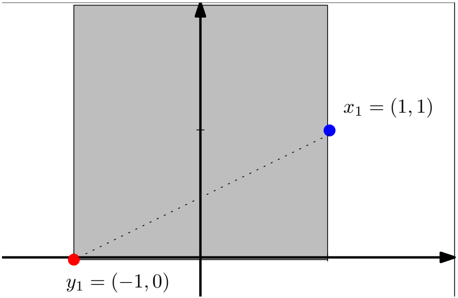
An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



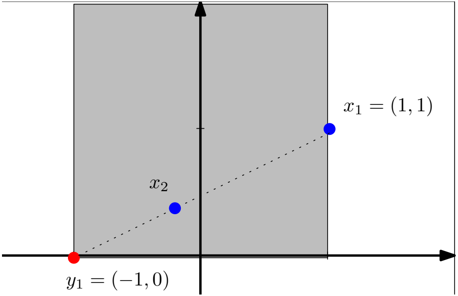
An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



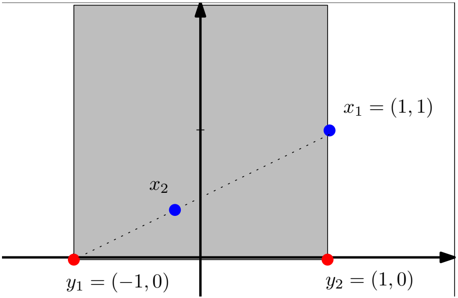
An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



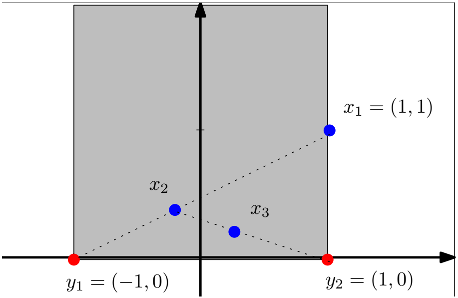
An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



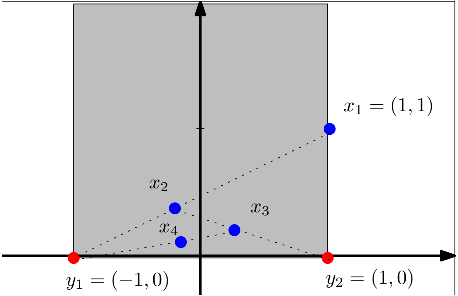
An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



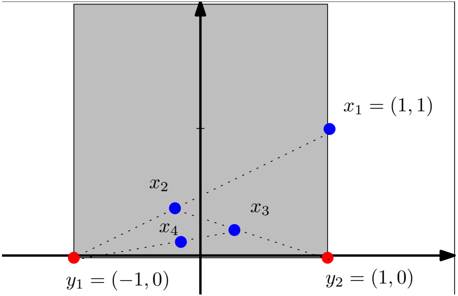
An example

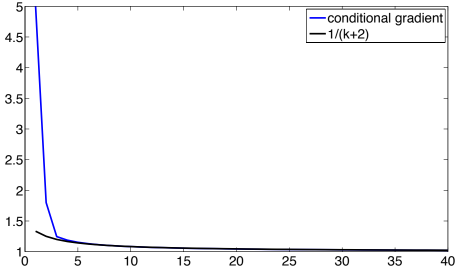
$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \le 1, 2 \ge b \ge 0$$



An example

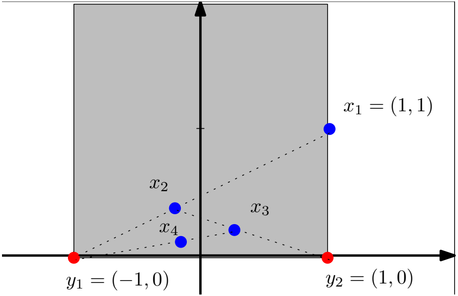
$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$

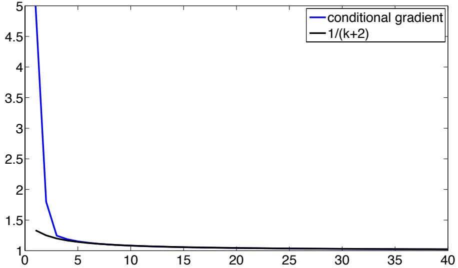




An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



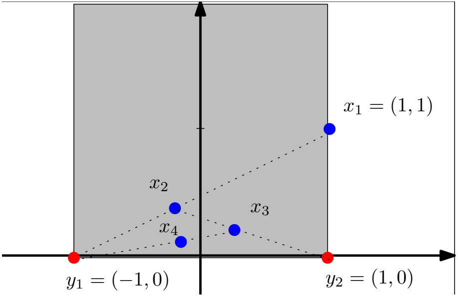


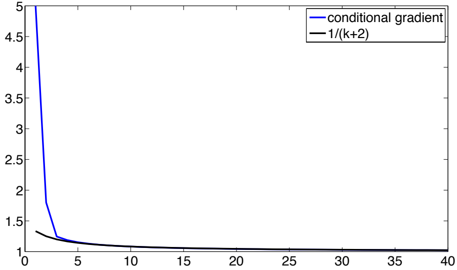
Can show $\ell(\boldsymbol{w}_t) - \ell(\boldsymbol{w}^\star) = 4/t + o(1/t)$.

PG converges in two iterations.

An example

$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$





Can show $\ell(\boldsymbol{w}_t) - \ell(\boldsymbol{w}^\star) = 4/t + o(1/t)$.

PG converges in two iterations.

Refs: (Levtin-Polyak'66; Polyak'87; Beck-Teboulle'04) for faster rates.

The revival of CG: sparsity!

The revived popularity of conditional gradient is due to (Clarkson'10; Shalev-Shwartz-Srebro-Zhang'10), both focusing on

$$\min_{\boldsymbol{w}:\ \|\boldsymbol{w}\|_1 \le 1} \ell(\boldsymbol{w}).$$

1. $\boldsymbol{y}_t \leftarrow \underset{\|\boldsymbol{y}\|_1 \le 1}{\operatorname{argmin}} \langle \boldsymbol{y}, \nabla \ell(\boldsymbol{w}_t) \rangle,$ $\qquad \mathtt{card}(\boldsymbol{y}_t) = 1;$
2. $\boldsymbol{w}_{t+1} \leftarrow (1-\eta)\boldsymbol{w}_t + \eta \boldsymbol{y}_t,$ $\qquad \mathtt{card}(\boldsymbol{w}_{t+1}) \le \mathtt{card}(\boldsymbol{w}_t) + 1.$

Explicit control of the sparsity.

Later on, (Hazan'08; Jaggi-Sulovsky'10) generalized the idea to SDPs.

Generalized conditional gradient

$$\min_{\boldsymbol{w}} \ \ell(\boldsymbol{w}) + \lambda \cdot f(\boldsymbol{w})$$

- composite, with a nonsmooth term;
- unconstrained, hence unbounded domain;
- first studied by Mine & Fukushima'81 and then Bredies et al.'09;
- generalizes CG.

$$\boxed{\begin{array}{ll} ❶ & \boldsymbol{y}_t \in \mathop{\mathrm{argmin}}_{\boldsymbol{w}} \langle \boldsymbol{w}, \nabla \ell(\boldsymbol{w}_t) \rangle + f(\boldsymbol{w}); \\ ❷ & \boldsymbol{w}_{t+1} = (1-\eta)\boldsymbol{w}_t + \eta \boldsymbol{y}_t. \end{array}}$$

Our interest:

- f p.h. (e.g., a norm);
- Step 1 undefined.

Positive homogeneous regularizer

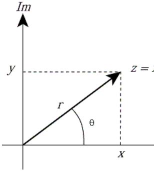
$$\min_{\boldsymbol{w}} \ \ell(\boldsymbol{w}) + \lambda \cdot \kappa(\boldsymbol{w})$$

- ℓ: smooth convex;
- κ: positive homogeneous convex—gauge (not necessarily smooth).

Challenges:

- composite, with a nonsmooth term;
- unconstrained, hence unbounded domain;
- κ expensive to evaluate.

1. Polar operator: $\boldsymbol{y}_t \in \underset{\boldsymbol{w}:\kappa(\boldsymbol{w})\leq 1}{\operatorname{argmin}} \ \langle \boldsymbol{w}, \nabla\ell(\boldsymbol{w}_t)\rangle$;
2. line search: $s_t \in \underset{s\geq 0}{\operatorname{argmin}} \, \ell((1-\eta)\boldsymbol{w}_t + \eta s \boldsymbol{y}_t) + \lambda\eta s$;
3. $\boldsymbol{w}_{t+1} = (1-\eta)\boldsymbol{w}_t + \eta s_t \boldsymbol{y}_t$.



Convergence guarantee

$$\min_{\boldsymbol{w}} \; \ell(\boldsymbol{w}) + \lambda \cdot \kappa(\boldsymbol{w})$$

Theorem (ZYS'12)

If ℓ and κ have bounded level sets and $\ell \in C^1$, then GCG converges at rate $O(1/t)$, where the constant is independent of λ.

- Proof is simple: Line search is as good as knowing $\kappa(\boldsymbol{w}^\star)$;
- Upper bound

$$\kappa((1-\eta)\boldsymbol{w}_t + \eta s \boldsymbol{y}_t) \le (1-\eta)\kappa(\boldsymbol{w}_t) + \eta\kappa(s\boldsymbol{y}_t) \le (1-\eta)\kappa(\boldsymbol{w}_t) + \eta s;$$

- Still too slow!

Local improvement

Assume some procedure (say LOCAL) that can *locally* solve

$$\min_{\boldsymbol{w}} \ell(\boldsymbol{w}) + \lambda \cdot \kappa(\boldsymbol{w}),$$

or some variation of it.

Combine LOCAL with some GLOBAL?

Three conditions:

- LOCAL cannot incur big overhead;
- cannot ruin GLOBAL;
- easy to switch between LOCAL and GLOBAL.

Refs: Burer-Monteiro'05; Mishra et al.'11; Laue'12

Case study: matrix completion with trace norm

GLOBAL: $$\min_{X} \sum_{(i,j)\in\mathcal{O}} (X_{ij} - Z_{ij})^2 + \lambda \cdot \|X\|_{\mathrm{tr}}$$

The only nontrivial step in GCG:

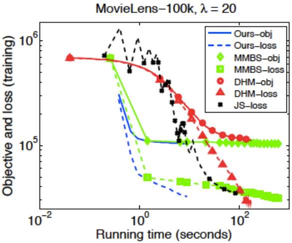
- Polar operator: $Y_t \in \underset{\|Y\|_{\mathrm{tr}}\leq 1}{\mathrm{argmin}} \langle Y, G_t \rangle$, *dominating* singular vectors.

In contrast, PG requires the *full* SVD of $-G_t$.

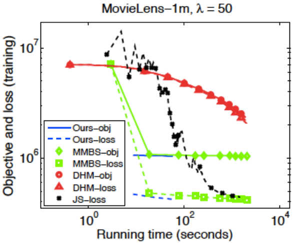
LOCAL (Srebro'05): $$\min_{B,W} \sum_{(i,j)\in\mathcal{O}} ((BW)_{ij} - Z_{ij})^2 + \lambda/2 \cdot (\|B\|_F^2 + \|W\|_F^2).$$

- Not jointly convex in B and W;
- But smooth in B and W;
- Y_t in GCG is rank-1 hence $X_t = BW$ is of rank at most t.

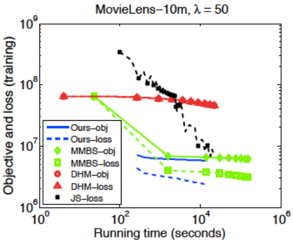
Case study: experiment



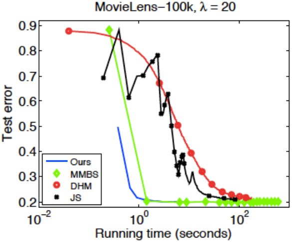
(a) Objective & loss vs time (loglog)



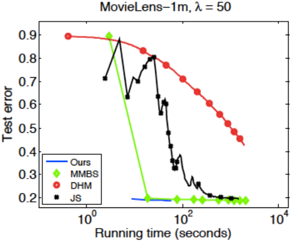
(a) Objective & loss vs time (loglog)



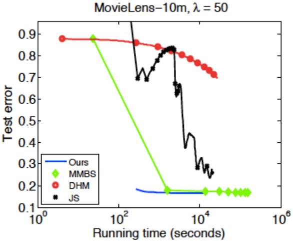
(a) Objective & loss vs time (loglog)



(b) Test NMAE vs time (semilogx)



(b) Test NMAE vs time (semilogx)



(b) Test NMAE vs time (semilogx)

Summary

- Generalized conditional gradient for p.h. regularizer;
- $O(1/t)$ convergence rate;
- Combined LOCAL with GCG ;
- Applied to matrix completion.

Table of Contents

Prox-decomposition and isotonicity

Hölder's inequality: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\|_r \|\boldsymbol{y}\|_s, \;\; r \geq 1, \; 1/r + 1/s = 1$

Ky Fan's norm $\|\boldsymbol{x}\|_{k,r} := \sqrt[r]{\sum_{i=1}^{k} |x|_{(i)}^r}$.

$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\|_{k,r} \cdot ???$, i.e., dual norm $\|\boldsymbol{y}\|_{k,r}^{\circ} := \max_{\|\boldsymbol{x}\|_{k,r} \leq 1} \langle \boldsymbol{x}, \boldsymbol{y} \rangle = ?$

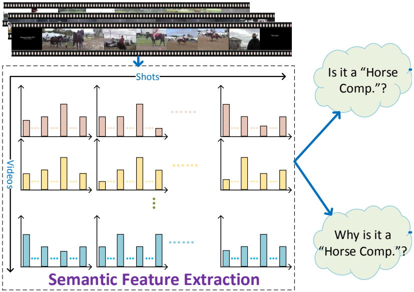
First shown in (Mudholkar et al, 1984).

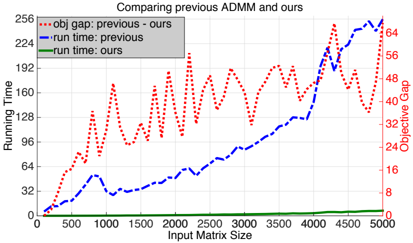
Theorem (YYX'15)

For any $r \geq 1$ and $1/r + 1/s = 1$, the dual Ky Fan norm $\|\boldsymbol{y}\|_{k,r}^{\circ} = \|\boldsymbol{z}\|_s$, where $\boldsymbol{z} := \mathrm{P}_{\mathcal{K}}(\boldsymbol{m}) = \operatorname*{argmin}_{w_1 \geq w_2 \geq \cdots \geq w_k} \frac{1}{2}\|\boldsymbol{m} - \boldsymbol{w}\|_2^2$ and

$$m_i = \begin{cases} |y|_{(i)}, & i = 1, \ldots, k-1 \\ \sum_{j=k}^{p} |y|_{(j)}, & i = k \end{cases}.$$

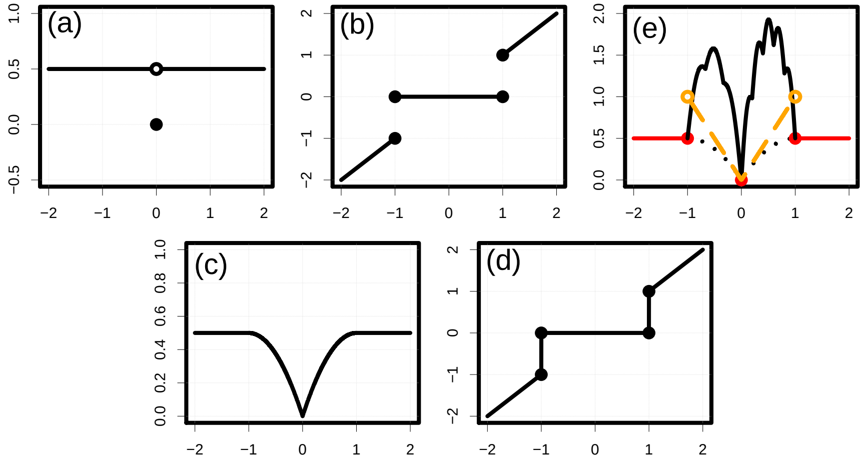
Video event detection and recounting (CYYH'15)





Nonconvex proximal average (YZMX'14)

$$\mathsf{P}^{\eta}_{\sum_i \alpha_i f_i} \overset{?}{\approx} \sum_i \alpha_i \mathsf{P}^{\eta}_{f_i}$$



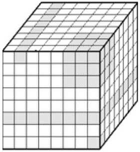
Approximate generalized conditional gradient

Pick $\kappa(\boldsymbol{y}_t) \leq 1$ such that for some $\alpha \in (0, 1]$

$$\langle \boldsymbol{y}_t, \nabla \ell(\boldsymbol{w}_t) \rangle \leq \alpha \cdot \min_{\boldsymbol{y}:\kappa(\boldsymbol{y}) \leq 1} \langle \boldsymbol{y}, \nabla \ell(\boldsymbol{w}_t) \rangle .$$

Theorem (YCZ'14)

Assume $\ell \geq 0$. Equipped with an α-approximate PO, GCG "converges" to an α-approximate solution at the rate $O(1/t)$.



Thank you!