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 $\theta \in \mathbb{R}^p$:

$$y_i = x_i^\top \theta + \epsilon_i, \quad i = 1, \ldots, n$$

- classical setting: $p$ fixed small, $n \to \infty$, lots of results.
Statistical inference 101

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

$$y_i = x_i^\top \theta + \epsilon_i, \quad i = 1, \ldots, n$$

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

$$p \sim 10^7, n \sim 10^3$$ $$p \sim 10^{10}, n \sim 10^8$$
High-dimensional challenge

- More unknown parameters than observations, ill-defined.
  - **structure**: effective number of unknown parameters is moderate.
    - $\theta$ is sparse: $\text{nnz}(\theta)$ small, but do not know which is which.
    - $\theta$ 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.
1. Introduction
2. Decomposing the Proximal Map
3. Approximation by the Proximal Average
4. Generalized Conditional Gradient
5. Post-PhD Extensions
Table of Contents

1 Introduction

2 Decomposing the Proximal Map

3 Approximation by the Proximal Average

4 Generalized Conditional Gradient

5 Post-PhD Extensions
Regularized loss minimization

Generic form for many ML problems:

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

- $\ell$ 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)

\[
\min_{w \in \mathbb{R}^p} \ell(w) + f(w) = \min_{w \in \mathbb{R}^p} \|A w - b\|^2 + \lambda \|w\|_1.
\]

Multiple benefits

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

Computationally?

- convex quadratic program
- but \( P \neq E \! \)
- especially when \( p \) is large.
Nonsmooth optimization

Generic subgradient descent:

\[ w_{t+1} = w_t - \eta [\nabla \ell(w_t) + \partial f(w_t)] \]

- guaranteed convergence, \( O(1/\epsilon^2) \);
- dense iterates;
- weak regularizing effect;
- and slow, very slow...

Second order methods (e.g. IPM) do not scale.
Moreau envelope and proximal map

Definition (Moreau’65)

\[
M^\eta_f(y) = \min_w \frac{1}{2\eta} \|w - y\|^2 + f(w)
\]

\[
P^\eta_f(y) = \arg\min_w \frac{1}{2\eta} \|w - y\|^2 + f(w)
\]
Some properties of the proximal map

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

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

For general $f$ (that decreases not too fast),
- $P^n_f(\cdot)$ is a nonempty compact set;
- $M^n_f(\cdot)$ is continuous;
- Still $\eta \downarrow 0 \implies M^n_f \uparrow f$. 
Proximal gradient (Fukushima & Mine’81)

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

\[y_t = w_t - \eta \nabla \ell(w_t); \quad \text{(forward)}\]

\[w_{t+1} = P_f^\eta(y_t). \quad \text{(backward)}\]

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

\[
[P_f^\eta(y)]_i = \text{sign}(y_i)(|y_i| - \eta)_+.
\]

- 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.
CONVEX PROGRAMMING IN HILBERT SPACE

BY A. A. GOLDSTEIN

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 orthonals, 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 \gamma \in C. Then \[ \| x - \gamma \|^2 = \| P(x) - \gamma \|^2. \]

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 subset \( \{ x \in C : f(x) \leq f(x_0) \} \), and let \( S \) be an open set containing the convex hull of S. Let f be a real-valued function on H and x_0 an arbitrary point of C. Then \[ \inf \{ f(x) : x \in S \} \]

is a minimum of f on C. A point x \in C will be called stationary if \( f(x - \rho \nabla f(x)) = f(x) \) for all \( \rho > 0 \); equivalently, when f is convex the linear functional \( f(x - \cdot) \) achieves a minimum on C at \( x \).

THEOREM. Assume \( f \) is bounded below. For each \( x \in \mathbb{R}, h \in H \) and for some \( p > 0 \), assume that \( f(x, h) \) exists in the sense of Fréchet, \[ f''(x, h, h) \] exist in the sense of Gateaux, and \[ f''(x, h, h) \leq \| h \|^2 / p_0. \]

Choose \( \sigma \) and \( p_0 \) satisfying \( 0 < \sigma \leq p_0 \), and \( \sigma \leq p_0 \leq 2p_0 - \sigma \). Set \( x_{n+1} = P(x_n - p_0 \nabla f(x_n)) \).

Then:

(i) The sequence \( x_n \) belongs to C, \( x_{n+1} - x_n \) converges to \( 0 \), and \( f(x_n) \) converges downward to a limit L.

(ii) If S is compact, \( x_n \) is a cluster point of \( \{ x_n \} \), and \( \nabla f \) is continuous in some neighborhood of \( x \), then \( x \) is a stationary point.

(iii) If S is convex and \( f''(x, h, h) \geq \| h \|^2 / p_0 \) for each \( x \in S \), \( h \in H \) and some \( p_0 \geq 0 \), then L = \inf \{ f(x) : x \in C \}.

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

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

710

A. A. GOLDSTEIN

(v) Assume (iii) with \( \mu \) positive and \( \nabla f \) bounded on S. Then \( f(x) = L \) for some \( x \in S \), \( x_n \) converges to \( x \), and \( x \) is unique.

PROOF. Assume \( x_n \) belongs to S and that \( x_n \) is not stationary. Let \( \nabla f(x_n) = \mu_1 \), \( x(x) = P(x_n - \mu \nabla f(x_n)) \), \( \delta(x) = x(x) - x_n \) and \( \Delta(x) = f(x) - f(x(x)) \).

If we notice that \( -\mu \nabla f(x_n) \) \( \leq \| x(x) \|^2 \) and invoke Taylor's theorem, we obtain \( \Delta(x) \geq \| x(x) \|^2 [p_0^{-1} f''(\Delta(x)) \delta(x) + \delta(x) \| / 2 \| \Delta(x) \|^2 \] [1] Here \( \delta(x) = x_n(x) \delta(x) \).

For some \( \rho \) sufficiently small and positive, \( \Delta(x) \) is positive and continuous. Let \( \beta \) denote the least positive \( \rho \) satisfying \( \Delta(x) = 0 \), if such exists. If \( \beta \) exists, \( \Delta(x) = 0 \) implies that \( \rho / 2 \rho_0 \leq \beta \). Thus if \( \beta \leq \rho \leq 2p_0 - \sigma \), \( \Delta(x) > 0 \) and \( \Delta(x) \in S \), whence \( \Delta(x) = \| x_n - x_n \|^2 / 4p_0 \) proving (i).

The proof of (ii) being straightforward, we proceed with the proof of (iii). Suppose that \( f(x) : x \in S \) and choose \( x \in C \) such that \( f(x) \). Then \( 0 > f(x) - f(x_n) \geq \| \nabla f(x_n) - x \|. \) If \( \inf \{ f(x) : x \in S \} = \beta \) were non-negative, a contradiction would be manifest. But the inequality \( \| x - x_n \|^2 + \| x - x_n \|^2 \leq \| x_n - x_n \|^2 + \| x_n - x_n \|^2 \) holds because either \( x_n - x_n \| x_n - x_n \) is the normal to C at \( x_n \), or it is 0. If the sequence \( x_n \) is bounded, clearly \( \beta = 0 \); otherwise choose a subsequence satisfying \( \| x_n - x_n \|^2 > x_n \). Then \( \beta \geq 0 \).

To prove (iv) we observe that \( f(x) \) is lower semi-continuous on S if and only if the set \( S_n = \{ x \in S : f(x) \leq m \} \) is closed in S for each \( m \).

Since \( f \) is convex and continuous, \( S_n \) is closed and convex, and is thus weakly closed. Hence \( j \) is weakly l.s.c. If \( x_n \) converges weakly to \( x \), then \( \inf f(x) = L = f(x) \).

Assume the hypotheses of (v). If \( \beta > 0 \), we may write that \( 0 > f(x_n) - f(x_n) - \| x_n - x_n \|^2 / 2 \beta \beta \), whence \( \| x_n - x_n \|^2 = 0 \) is bounded. Invoking again the supporting hyperplane at \( x_n + \beta x_n - x_n \), \( \| x - x_n \|^2 + \| x_n - x_n \|^2 \leq \| x_n - x_n \|^2 + \| x_n - x_n \|^2 \). Thus when \( \beta \) is sufficiently large \( \| x_n - x_n \|^2 = 0 \). There exists therefore \( z \in C \) minimizing f on C, and \( f(x) \geq f(z) + \| x - z \|^2 / (1/2) \). Thus \( \nabla f(x) = 0 \), \( f(x) = f(z) \). Therefore \( x \) is unique.
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)
Step 2: $P^n_f(y) = \arg\min_w \frac{1}{2\eta} \|y - w\|^2 + f(w)$
Structured sparsity: group

Group level sparse regularizer

\[ f(w) = \sum_i \| w \|_{g_i}. \]

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

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

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:

$$\min_{X \in \mathbb{R}^{m \times n}} \sum_{(i,j) \in O} (X_{ij} - Z_{ij})^2 + \lambda \| X \|_{\text{tr}} \cdot \ell(X) + \left\langle f(X) \right\rangle$$

- Can apply PG:

$$P_{\lambda \| \cdot \|_{\text{tr}}}^{\eta} (Y) = \sum_{k} (\sigma_k - \lambda \eta) + u_k v_k^\top.$$  

- 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.
1 Introduction

2 Decomposing the Proximal Map

3 Approximation by the Proximal Average

4 Generalized Conditional Gradient

5 Post-PhD Extensions
How to decompose?

- Typical structured sparse regularizers:

\[ f(\mathbf{w}) = \sum_i f_i(\mathbf{w}); \]

  - Also applies to ERM, each \( i \) is a sample.

- Key observation: each \( P_{f_i} \) is easy to compute.

- Can we compute \( P_f = P \sum_i f_i \) efficiently?

Theorem (Folklore)

\[ P_{f+g} = (P_{2f}^{-1} + P_{2g}^{-1})^{-1} \circ (2\mathbf{1}_d). \]

- 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

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

Theorem (Friedman et al.’07)

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

Theorem (Jenatton et al.’11)

$$P_{\sum_{i=1}^{k} \|\cdot\|_{g_i}} = P_{\|\cdot\|_{g_1}} \circ \cdots \circ P_{\|\cdot\|_{g_k}}.$$ 

Generalization

$$P_{f+g} \overset{?}{=} P_f \circ P_g \overset{?}{=} P_g \circ P_f.$$ 

But, is it even sensible?
**Theorem**

*On the real line, $\exists h$ such that $P_h = P_f \circ P_g$.***

**Example (But not so in general...)**

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

![Diagram showing $f = \iota\{x_1 = x_2\}$ and $g = \iota\{x_2 = 0\}$ on a 2D plane, illustrating the theorem's application.]
Nevertheless

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

\[
P_{f+g}(z) - z + \partial(f + g)(P_{f+g}(z)) \ni 0
\]
\[
P_g(z) - z + \partial g(P_g(z)) \ni 0
\]
\[
P_f(P_g(z)) - P_g(z) + \partial f(P_f(P_g(z))) \ni 0.
\]

- Adding the last two equations we obtain

\[
P_f(P_g(z)) - z + \partial g(P_g(z)) + \partial f(P_f(P_g(z))) \ni 0.
\]

**Theorem (Y’13a)**

A sufficient condition for \( P_{f+g}(z) = P_f(P_g(z)) \) is

\[
\forall \ y \in \text{dom} \ g, \ \partial g(P_f(y)) \supseteq \partial g(y).
\]
The rest is easy

- Find $f$ and $g$ that clinch our sufficient condition.
Result I: Start with “trivialities”

**Theorem (Y’13a)**

Fix $f \in \Gamma_0$. $P_{f+g} = P_f \circ P_g$ for all $g \in \Gamma_0$ if and only if

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

Asymmetry.

**Theorem (Y’13a)**

Fix $g \in \Gamma_0$. $P_{f+g} = P_f \circ P_g$ 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 \( \dim(H) \geq 2 \)):

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

ii). \( x \perp y \implies f(x + y) \geq f(y) \);

iii). For all \( z \in H \), \( P_f(z) = \lambda_z \cdot z \) for some \( \lambda_z \in [0, 1] \);

iv). \( 0 \in \text{dom} \ f \) and \( P_{f+\kappa} = P_f \circ P_\kappa \) 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)

\[ P_{\lambda \| \cdot \|_2^2 + \kappa} = P_{\lambda \| \cdot \|_2^2} \circ P_\kappa = \frac{1}{\lambda + 1} P_\kappa \]

double shrinkage

Example (Jenatton et al.’11)

Tree-structured (laminar system)

\[ P_{\sum_i \| \cdot \|_{g_i}} = P_{\| \cdot \|_{g_1}} \circ \cdots \circ P_{\| \cdot \|_{g_k}} \]
Characterizing the ball

\[ C \]

\[ 0 \]
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.

\[ P_{f+g} = P_f \circ P_g. \]

**Example (Friedman et al.’07)**

\[ P_{\| \cdot \|_1 + \| \cdot \|_{TV}} = P_{\| \cdot \|_1} \circ P_{\| \cdot \|_{TV}}. \]

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

“Always consider a problem under the minimum structure in which it makes sense.”
Summary

- Posed the question: $P_{f+g} \equiv P_f \circ P_g \equiv P_g \circ P_f$;
- Presented a sufficient condition: $\partial g(P_f(y)) \supseteq \partial g(y)$;
- “Trivial” case;
- Positive homogeneity and “roundness”;
- Comonotonicity and Choquet integral;
- Immediately useful if plugged into PG;

What if the sufficient condition fails?
Table of Contents

1. Introduction

2. Decomposing the Proximal Map

3. Approximation by the Proximal Average

4. Generalized Conditional Gradient

5. Post-PhD Extensions
More generally

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

- \( P_{f_i}^\eta \) easy to compute;
- \( f_i \) Lipschitz continuous.

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

\[ f_i(\mathbf{w}) = \| \mathbf{w} \|_{g_i} \text{ where } g_i \text{ is a group (subset) of variables.} \]

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

\[
[P_{f_i}^\eta(\mathbf{w})]_j = \begin{cases} 
  w_j, & j \not\in g_i \\
  (1 - \eta/\| \mathbf{w} \|_{g_i}) + w_j, & j \in g_i
\end{cases}
\]
Example (Graph-guided fused lasso, Kim & Xing’09)

Given some graph, we let $f_{ij}(w) = |w_i - w_j|$ 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 $P_{f_{ij}}^\eta$ is easy to compute:

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

Other examples abound.
Smoothing (Nesterov’05)

\[
M_f^\eta(y) = \min_w \frac{1}{2\eta} \|w - y\|^2 + f(w)
\]

\[
P_f^\eta(y) = \arg\min_w \frac{1}{2\eta} \|w - y\|^2 + f(w)
\]

Proposition (Nesterov’05)

If \( f \) is \( L \)-Lipschitz continuous, then
\[
0 \leq f - M_f^\eta \leq \frac{\eta L^2}{2}.
\]
In retrospect

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

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

- \( \text{dist}(\mathbf{w}, C) := \min_{\mathbf{z} \in C} \| \mathbf{w} - \mathbf{z} \| \), nonsmooth but Lipschitz continuous.
- Can smooth \( \text{dist} \) and apply gradient descent.
- But nobody does that, overkill.
- Can just use projected gradient.
A “naive” idea (Y’13b)

\[ \bar{f} = \sum \alpha_i f_i \]

\[ \downarrow \quad \text{as if have linearity?} \]

\[ P_{\bar{f}}^\eta \approx \sum \alpha_i P_{f_i}^\eta \]

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

There exists a unique function \( A^\eta \) such that \( P_{A^\eta}^\eta = \sum_i \alpha_i P_{f_i}^\eta \).
What mathematicians call a “picture”

\[ \Gamma_0 \ni f_i \xrightarrow{\text{onto}} M^n_{f_i} \in SS_{1/\eta} \]

\[ \Gamma_0 \ni A^n \xleftarrow{1-1} \sum_i \alpha_i M^n_{f_i} \in SS_{1/\eta} \]

\[ \nabla \]

\[ P^n_{A^n} \xleftarrow{} \sum_i \alpha_i P^n_{f_i} \]

- Not so easy to compute \( A^n \), but existence is enough.
The algorithm

**Dream**

1. \( z_t = w_t - \mu \nabla \ell(w_t) \)
2. \( w_{t+1} = P^\eta_f(z_t) \)

\[ \downarrow \]

\( \min_w \ell(w) + \bar{f}(w) \)

**Reality**

1. \( z_t = w_t - \mu \nabla \ell(w_t) \)
2. \( w_{t+1} = P^\eta_{A^\eta}(z_t) = \sum_i \alpha_i P^\eta_{f_i}(z_t) \)

\[ \downarrow \]

\( \min_w \ell(w) + A^\eta(w) \)

When are they close?

---

Y-L. Yu  
Fast gradient algos for structured sparsity  
June 05, 2015  38 / 60
Nonsmooth approximation

- How good the proximal average $A^n$ 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

$$0 \leq \bar{f} - A^n \leq \eta M^2 / 2.$$ 

- Proximal average is a tighter approximation than smoothing:

$$\sum_i \alpha_i M_{f_i}^n \leq A^n \leq \bar{f}.$$
An 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 $\epsilon$-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.

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

\[
PA\text{-PG: } \mathbf{w}_{t+1} = \sum_i \alpha_i \mathcal{P}_f^\eta(\mathbf{w}_t - \eta \nabla \ell(\mathbf{w}_t)).
\]
Experiment

\[ \varepsilon = \frac{1}{L} \]

\[ \text{PA-PG} \quad \text{PA-APG} \]

\[ \text{S-PG} \quad \text{S-APG} \]
Summary

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

1. Introduction

2. Decomposing the Proximal Map

3. Approximation by the Proximal Average

4. Generalized Conditional Gradient

5. Post-PhD Extensions
Conditional gradient (Frank-Wolfe’56)

\[
\min_{w \in C} \ell(w)
\]

- \(C\): compact convex;
- \(\ell\): smooth convex.

\[y_t \in \arg\min_{w \in C} \langle w, \nabla \ell(w_t) \rangle;\]

\[w_{t+1} = (1 - \eta) w_t + \eta y_t.\]

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

\[
\min_{a,b} a^2 + (b + 1)^2, \quad \text{s.t. } |a| \leq 1, 2 \geq b \geq 0
\]
An example

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

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

Can show \( \ell(w) - \ell(w^\star) = 4/t + o(1/t) \). PG converges in two iterations.

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

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

$$x_1 = (1, 1)$$

$$y_1 = (-1, 0)$$
An example

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

\[
x_1 = (1, 1)
\]

\[
y_1 = (-1, 0)
\]
An example

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

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

\[
\begin{align*}
    x_1 &= (1, 1) \\
    y_1 &= (-1, 0) \\
    y_2 &= (1, 0)
\end{align*}
\]
An example

\[
\min_{a,b} a^2 + (b + 1)^2, \quad \text{s.t. } |a| \leq 1, 2 \geq b \geq 0
\]
An example

\[
\min_{a,b} a^2 + (b + 1)^2, \quad \text{s.t. } |a| \leq 1, 2 \geq b \geq 0
\]
An example

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

\[y_1 = (-1, 0) \quad \text{and} \quad y_2 = (1, 0)\]

The figure shows the feasible region defined by the constraints and the points \(x_1 = (1, 1)\) and \(x_1 = (1, 0)\) as solutions.

The plot on the right illustrates the conditional gradient and the function \(\ell(w_t) - \ell(w^{\star}) = 4/t + o(1/t)\), indicating convergence in two iterations. Relevant references include Levitin-Polyak'66, Polyak'87, and Beck-Teboulle'04 for faster rates.
An example

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

Can show \( \ell(w_t) - \ell(w^*) = 4/t + o(1/t) \).

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 \( \ell(w_t) - \ell(w^*) = \frac{4}{t} + o(1/t) \).

PG converges in two iterations.

Refs: (Levitin-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

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

1. \( y_t \leftarrow \arg\min_{\|y\|_1 \leq 1} \langle y, \nabla \ell(w_t) \rangle , \quad \text{card}(y_t) = 1; \)
2. \( w_{t+1} \leftarrow (1 - \eta)w_t + \eta y_t, \quad \text{card}(w_{t+1}) \leq \text{card}(w_t) + 1. \)

Explicit control of the sparsity.

Later on, (Hazan’08; Jaggi-Sulovský’10) generalized the idea to SDPs.
Generalized conditional gradient

\[
\min_w \ell(w) + \lambda \cdot f(w)
\]

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

1. \(y_t \in \arg\min_w \langle w, \nabla \ell(w_t) \rangle + f(w)\);
2. \(w_{t+1} = (1 - \eta)w_t + \eta y_t\).

Our interest:
- \(f\) p.h. (e.g., a norm);
- Step 1 undefined.
Positive homogeneous regularizer

\[
\min_w \ell(w) + \lambda \cdot \kappa(w)
\]

- \(\ell\): smooth convex;
- \(\kappa\): positive homogeneous convex—gauge (not necessarily smooth).

Challenges:
- composite, with a nonsmooth term;
- unconstrained, hence unbounded domain;
- \(\kappa\) expensive to evaluate.

1. **Polar operator**: \(y_t \in \arg\min_{w: \kappa(w) \leq 1} \langle w, \nabla \ell(w_t) \rangle\);
2. **line search**: \(s_t \in \arg\min_{s \geq 0} \ell((1 - \eta)w_t + \eta s y_t) + \lambda \eta s\);
3. \(w_{t+1} = (1 - \eta)w_t + \eta s_t y_t\).
Convergence guarantee

\[ \min_w \ell(w) + \lambda \cdot \kappa(w) \]

**Theorem (ZYS’12)**

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

- Proof is simple: Line search is as good as knowing \( \kappa(w^*) \);
- Upper bound
  \[ \kappa((1 - \eta)w_t + \eta sy_t) \leq (1 - \eta)\kappa(w_t) + \eta\kappa(sy_t) \leq (1 - \eta)\kappa(w_t) + \eta s; \]
- Still too slow!
Local improvement

Assume some procedure (say \texttt{Local}) that can \textit{locally} solve

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

or some variation of it.

Combine \texttt{Local} with some \texttt{Global}?

Three conditions:

- \texttt{Local} cannot incur big overhead;
- cannot ruin \texttt{Global};
- easy to switch between \texttt{Local} and \texttt{Global}.

Refs: Burer-Monteiro’05; Mishra et al.’11; Laue’12
Case study: matrix completion with trace norm

**GLOBAL:**
\[
\min_X \sum_{(i,j) \in \mathcal{O}} (X_{ij} - Z_{ij})^2 + \lambda \cdot \|X\|_{tr}
\]

The only nontrivial step in GCG:
- Polar operator: \( Y_t \in \arg\min_Y \|Y\|_{tr} \leq 1 \langle Y, G_t \rangle \), 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)  
(b) Test NMAE vs time (semilogx)

MovieLens-100k, $\lambda = 20$

MovieLens-1m, $\lambda = 50$

MovieLens-10m, $\lambda = 50$
Summary

- Generalized conditional gradient for p.h. regularizer;
- $O(1/t)$ convergence rate;
- Combined \texttt{LOCAL} with GCG;
- Applied to matrix completion.
Table of Contents

1 Introduction

2 Decomposing the Proximal Map

3 Approximation by the Proximal Average

4 Generalized Conditional Gradient

5 Post-PhD Extensions
Prox-decomposition and isotonicity

Hölder’s inequality: $\langle x, y \rangle \leq \|x\|_r \|y\|_s$, $r \geq 1$, $1/r + 1/s = 1$

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

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

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 $\|y\|_{k,r}^\circ = \|z\|_s$, where $z := P_K(m) = \text{argmin}_{w_1 \geq w_2 \geq \cdots \geq w_k} \frac{1}{2} \|m - w\|_2^2$ 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)

Semantic Feature Extraction

Is it a “Horse Comp.”?

Why is it a “Horse Comp.”?

Comparing previous ADMM and ours

- obj gap: previous - ours
- run time: previous
- run time: ours

Input Matrix Size
0 500 1000 1500 2000 2500 3000 3500 4000 4500 5000

Running Time
0 32 64 96 128 160 192 224 256

- obj gap: previous - ours
- run time: previous
- run time: ours

June 05, 2015 57 / 60
Nonconvex proximal average (YZMX’14)

\[ P^n \sum_i \alpha_i f_i \approx \sum_i \alpha_i P^n_{f_i} \]
Approximate generalized conditional gradient

Pick $\kappa(y_t) \leq 1$ such that for some $\alpha \in (0, 1]$

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

**Theorem (YCZ’14)**

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

Thank you!