Generalized Conditional Gradient and Its Applications

Yaoliang Yu

University of Alberta

UBC – Kelowna, 04/18/13

Table of Contents

Regularized Loss Minimization

Generic form for many ML problems:

$$\min_{w} f(w) + \lambda \cdot h(w), \quad \text{where}$$

- f is the loss function;
- h is the regularizer;

Assuming f and h to be convex/smooth

- Interior point method;
- Mirror descent / Proximal gradient;
- Averaging gradient;
- Conditional gradient.

Machine Learning Examples

Example (Matrix Completion)

$$\min_{X \in \mathbb{R}^{m \times n}} \frac{1}{2} \sum_{ij \in \mathcal{O}} (X_{ij} - Z_{ij})^2 + \lambda \cdot \|X\|_{\mathrm{tr}}$$

- Netflix problem;
- Covariance matrix estimation; etc.

Example (Group Lasso)

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|A\mathbf{w} - \mathbf{b}\|_2^2 + \lambda \cdot \sum_{g \in \mathcal{G}} \|\mathbf{w}\|_g$$

- Statistical estimation;
- Inverse problem;
- Denoising; etc.

Interesting case: m, n or d are extremely large.

Conditional gradient (Frank-Wolfe'56)

Consider $$\min_{x \in C} \; f(x),$$

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

1. $y_t \in \operatorname*{argmin}_{x \in C} \langle x, \nabla f(x_t) \rangle$;
2. $x_{t+1} = (1 - \eta) x_t + \eta y_t$.

(Frank-Wolfe'56; Canon-Cullum'68) proved that CG converges at $\Theta(1/t)$.

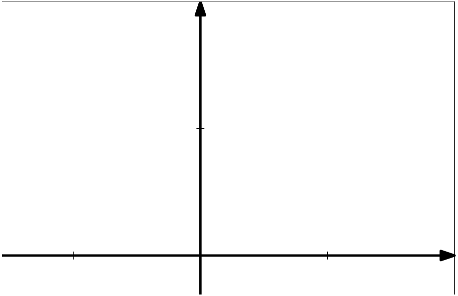
Gained much recent attention due to

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

Refs: (Zhang'03; Clarkson'10; Hazan'08; Jaggi-Sulovsky'10; Bach'12; etc.)

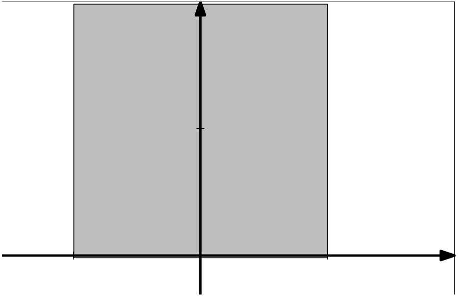
An Example

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



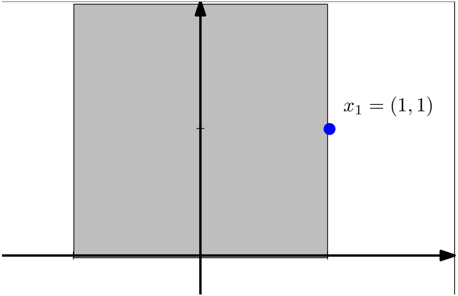
An Example

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



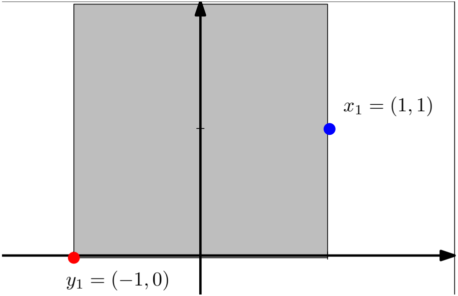
An Example

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



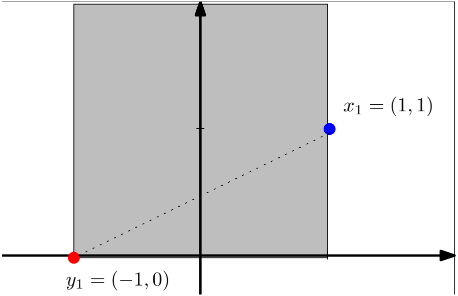
An Example

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



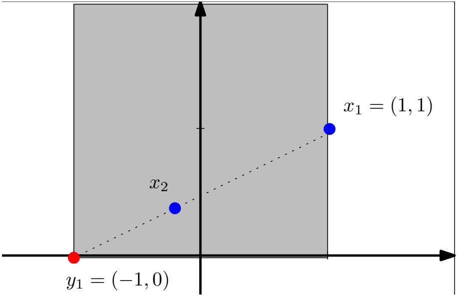
An Example

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



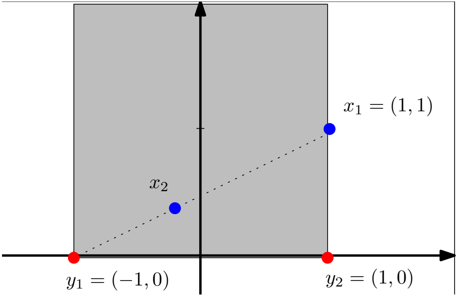
An Example

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



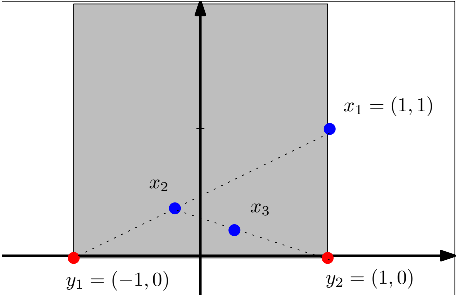
An Example

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



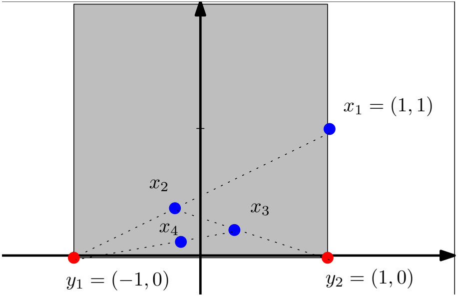
An Example

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



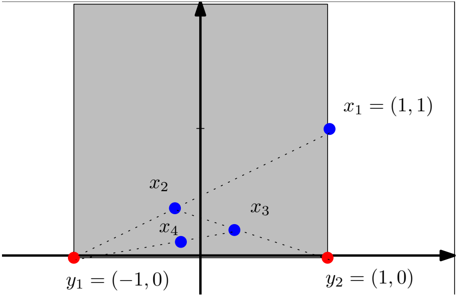
An Example

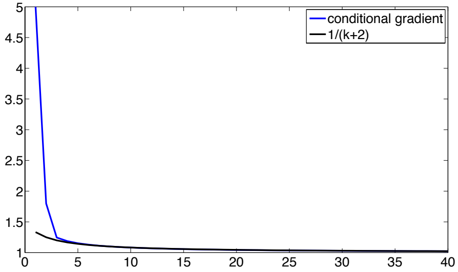
$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$



An Example

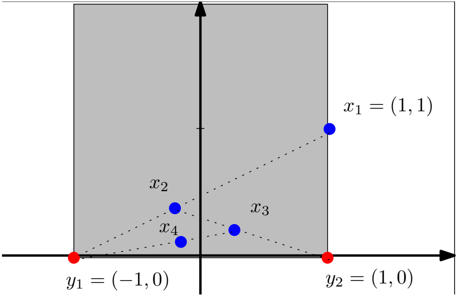
$$\min_{a,b} a^2 + (b+1)^2, \text{ s.t. } |a| \leq 1, 2 \geq b \geq 0$$

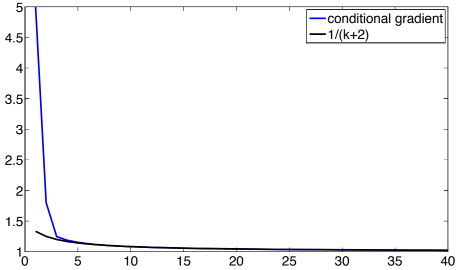




An Example

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



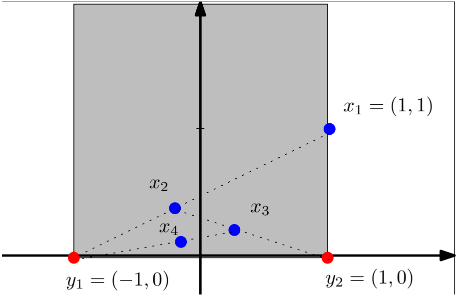


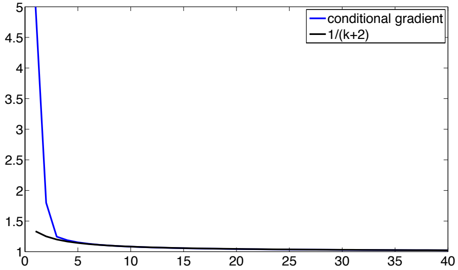
Can show $f(x_k) - f(x^\star) = 4/k + o(1/k)$.

Projected gradient 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 $f(x_k) - f(x^\star) = 4/k + o(1/k)$.

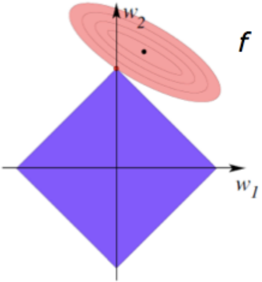
Projected gradient converges in two iterations.

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

$$\min_{x:\ \|x\|_1 \le 1} f(x).$$



1. $y_t \leftarrow \operatorname*{argmin}_{\|y\|_1 \le 1} \langle y; \nabla f(x_t) \rangle$, $\quad \mathtt{card}(y_t) = 1;$
2. $x_{t+1} \leftarrow (1-\eta)x_t + \eta y_t$, $\quad \mathtt{card}(x_{t+1}) \le \mathtt{card}(x_t) + 1.$

Explicit control of the sparsity. $\quad 1/\epsilon$ vs. $1/\sqrt{\epsilon}$.

Sparsity, more generally structure, is the key to the success of ML.

Table of Contents

Generalized conditional gradient

Consider $$\min_x \; f(x) + \lambda \cdot \kappa(x),$$

- f: smooth convex;
- κ: gauge (not necessarily smooth).

Important distinction:

- composite, with a non-smooth term;
- unconstrained, hence unbounded domain.

1. Polar operator: $y_t \in \operatorname*{argmin}_{x:\kappa(x)\le 1} \langle x, \nabla f(x_t)\rangle$;
2. line search: $s_t \in \operatorname*{argmin}_{s \ge 0} f((1-\eta)x_t + \eta s y_t) + \lambda \eta s$;
3. $x_{t+1} = (1-\eta)x_t + \eta s_t y_t$.

Convergence Rate

$$\min_x \; f(x) + \lambda \cdot \kappa(x)$$

Theorem (Zhang-Y-Schuurmans'12)

If f and κ have bounded level sets and $f \in \mathcal{C}^1$, 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 $\kappa(x^*)$;
- Note that we upper bound $\kappa((1-\eta)x_t + \eta s y_t) \leq (1-\eta)\kappa(x_t) + \eta s$;
- Still too slow!

Local improvement

Assume some procedure (say BFGS) that can *locally* minimize the nonsmooth problem $\min_x f(x) + \lambda \cdot \kappa(x)$, or some variation of it.

Combine this local procedure with some globally convergent routine?

Two conditions:

- The local procedure cannot incur big overhead;
- Cannot ruin the globally convergent routine.

Both are met by the GCG.

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

Case study: Matrix completion with trace norm

Consider
$$\min_X \frac{1}{2} \sum_{ij \in \mathcal{O}} (X_{ij} - Z_{ij})^2 + \lambda \cdot \|X\|_{\mathrm{tr}}.$$

- $\|\cdot\|_{\mathrm{tr}}$ is the convex hull of rank on the unit ball $\{X : \|X\|_{\mathrm{sp}} \leq 1\}$.

The only nontrivial step in GCG:

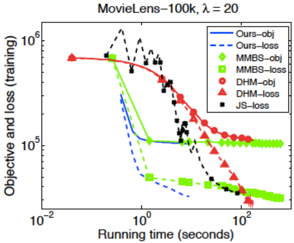
- Polar operator: $Y_t \in \operatorname*{argmin}_{\|Y\|_{\mathrm{tr}} \leq 1} \langle Y, G_t \rangle$, amounts to the dominating singular vectors of $-G_t$.

In contrast, popular gradient methods need the *full* SVD of $-G_t$.

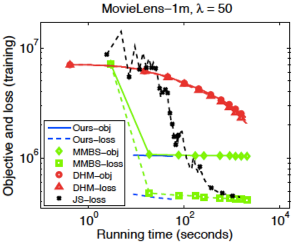
Variation: $$\frac{1}{2} \min_{U,V} \sum_{ij \in \mathcal{O}} ((UV)_{ij} - Z_{ij})^2 + \lambda \cdot (\|U\|_F^2 + \|V\|_F^2).$$

- Not jointly convex in U and V;
- But smooth in U and V;
- Y_t in GCG is rank-1 hence $X_t = UV$ is of rank at most t.

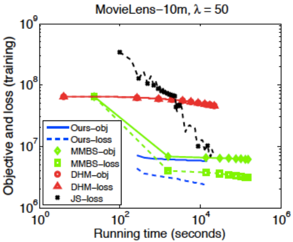
Case study: Experiment



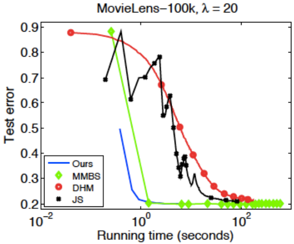
(a) Objective & loss vs time (loglog)



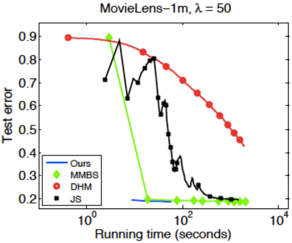
(a) Objective & loss vs time (loglog)



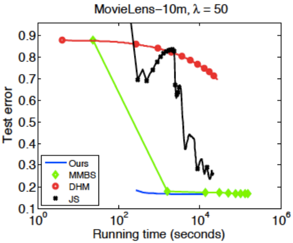
(a) Objective & loss vs time (loglog)



(b) Test NMAE vs time (semilogx)



(b) Test NMAE vs time (semilogx)



(b) Test NMAE vs time (semilogx)

Interpretation

Dictionary learning problem:

$$\min_{D \in \mathbb{R}^{m\times r}, \Phi \in \mathbb{R}^{r\times n}} L(X, D\Phi).$$

- Many applications: NMF, sparse coding, topic model...
- Not *jointly* convex, in fact NP-hard for fixed r;

Convexify by *not* constraining the rank *explicitly*: relax r!

Refs: (Bengio et al'05; Bach-Mairal-Ponce'08; Zhang-Y-White-Sch'10)

Convexification

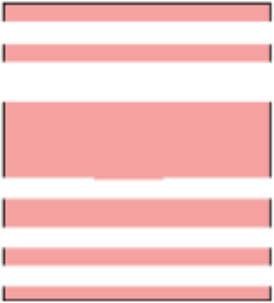
$$\min_{D,\Phi} L(X, D\Phi) + \lambda \cdot \Omega(\Phi).$$

- Let $D_{:i}$ have unit norm (say ℓ_2);
- Put row-wise norm on Φ: *implicitly* constraining the rank;
- Rewrite $\hat{X} := D\Phi = \sum_i \|\Phi_{i:}\| \cdot D_{:i} \frac{\Phi_{i:}}{\|\Phi_{i:}\|}$;
- Reformulate

$$\min_{\hat{X}} L(X, \hat{X}) + \lambda \cdot \kappa(\hat{X}) \quad \text{where}$$

$$\kappa(X) = \inf\{\textstyle\sum_i \sigma_i : X = \sum_i \sigma_i \cdot D_{:i} \frac{\Phi_{i:}}{\|\Phi_{i:}\|}\};$$

- Can apply GCG now, PO: $\min_{\mathbf{d},\phi} \mathbf{d}^\top G_t \frac{\phi}{\|\phi\|}$.



Setting both norms to ℓ_2, we recover the matrix completion example.

Table of Contents

Computing the Polar

The complexity of GCG is packed into the PO:

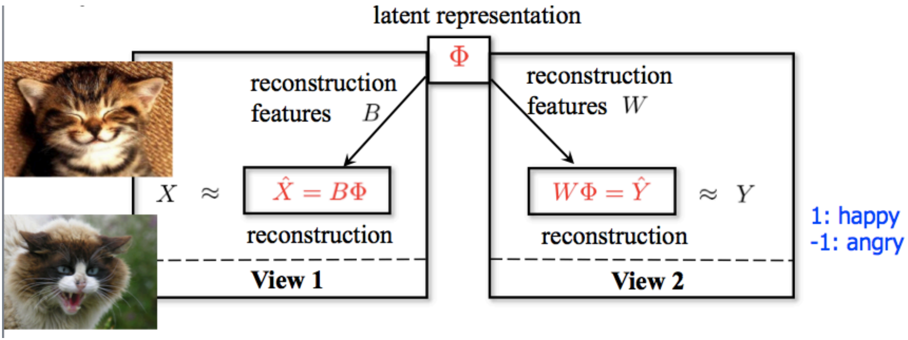
$$\left\{ \min_{x:\kappa(x)\leq 1} \langle g, x\rangle \right\} = -\kappa^\circ(-g).$$

Recall that in the dictionary learning problem:

$$\left\{ \min_{\mathbf{d},\phi} \ \mathbf{d}^\top G \frac{\phi}{\|\phi\|} \right\} = -\left\{ \max_{\mathbf{d}} \ \|G^\top \mathbf{d}\|^\circ \right\}$$

Can easily become computationally intractable!

Multi-view Learning



Partition $\mathbf{d} = \begin{bmatrix} \mathbf{b} \\ \mathbf{w} \end{bmatrix}$ and constrain their norms respectively.

Harder than single-view, but still doable (White-Y-Zhang-Sch'12):

$$\max_{\|\mathbf{b}\|=1, \|\mathbf{w}\|=1} \begin{bmatrix} \mathbf{b}^\top & \mathbf{w}^\top \end{bmatrix} GG^\top \begin{bmatrix} \mathbf{b} \\ \mathbf{w} \end{bmatrix} = \operatorname{tr}\left(GG^\top \begin{bmatrix} \mathbf{b} \\ \mathbf{w} \end{bmatrix} \begin{bmatrix} \mathbf{b}^\top & \mathbf{w}^\top \end{bmatrix} \right)$$

$$\frac{2(2+1)}{2} > 2.$$

Reducing PO to Proximal

Consider the group regularizer:

$$\Psi(\mathbf{w}) = \sum_g \|\mathbf{w}\|_g .$$

Its polar

$$\Psi^\circ(\mathbf{u}) = \inf\left\{ \max_g \|\mathbf{z}^g\|_g^\circ : \sum_g \mathbf{z}^g = \mathbf{u} \right\}$$

does not seem to be easy to compute.

Theorem

For any gauge Ω, its polar $\Omega^\circ(\mathbf{y})$ equals the smallest $\zeta \geq 0$ s.t.

$$\left\{ \min_{\Omega^\circ(\mathbf{x}) \leq \zeta} \|\mathbf{y} - \mathbf{x}\|_2^2 \right\} = \|\mathbf{y}\|_2^2 - 2 \cdot \mathrm{prox}_{\zeta\Omega}(\mathbf{y}) = 0,$$

where $\mathrm{prox}_f(\mathbf{y}) = \min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 + f(\mathbf{x})$ *and* $\mathrm{Prox}_f(\mathbf{y})$ *denotes the (unique) minimizer.*

Proximal Gradient

Consider

$$\min_{x \in C} f(x), \quad \text{where} \quad f \in \mathcal{C}_L^1.$$

$$\boxed{x_{t+1} = \operatorname*{argmin}_{x \in C} f(x_t) + \langle x - x_t, \nabla f(x_t) \rangle + \tfrac{L}{2} \|x - x_t\|_2^2.}$$

More generally

$$\min_{x \in C} f(x) + g(x), \quad \text{where} \quad f \in \mathcal{C}_L^1.$$

$$\boxed{\begin{aligned} x_{t+1} &= \operatorname*{argmin}_{x \in C} f(x_t) + \langle x - x_t, \nabla f(x_t) \rangle + \tfrac{L}{2} \|x - x_t\|_2^2 + g(x) \\ &= \operatorname*{argmin}_{x \in C} g(x) + \tfrac{L}{2} \|x - (x_t - \tfrac{1}{L} \nabla f(x_t))\|_2^2 \end{aligned}}$$

Decomposing the Proximal

How to compute the proximal operator for $\Psi(\mathbf{w}) = \sum_g \|\mathbf{w}\|_g$?

Theorem (NEW?)

$\mathrm{Prox}_{\Omega+\Phi} = \mathrm{Prox}_\Phi \circ \mathrm{Prox}_\Omega$ *for all gauges* Ω *iff* $\Phi = c\|\cdot\|_2$ *for some* $c \geq 0$.

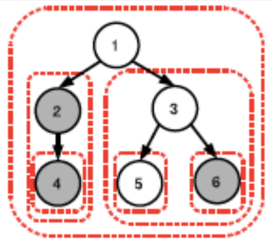
Corollary (Jenatton et al'11)

Let $\mathcal{G}$ *be a collection of tree-structured groups, that is, either* $g \subseteq g'$ *or* $g' \subseteq g$ *or* $g \cap g' = \emptyset$. *Then*

$$\mathrm{Prox}_{\sum_i \|\cdot\|_{g_i}} = \mathrm{Prox}_{\|\cdot\|_{g_1}} \circ \cdots \circ \mathrm{Prox}_{\|\cdot\|_{g_m}},$$

where we arrange the groups so that

$$g_i \subset g_j \implies i > j.$$



$\mathrm{Prox}_{2\Omega} = \mathrm{Prox}_\Omega \circ \mathrm{Prox}_\Omega$? More generally $\mathrm{Prox}_{\Omega+\Phi} = f(\mathrm{Prox}_\Omega, \mathrm{Prox}_\Phi)$?

Table of Contents

Conclusions

We have

- introduced the GCG;
- discussed efficient computations of PO;
- applied to matrix completion, group Lasso, etc.

Further questions

- when the PO is "hard"?
- nonsmooth loss?
- online? stochastic?

Thank you !