

CS794/CO673: Optimization for Data Science

Lec 06: Conditional Gradient

Yaoliang Yu



UNIVERSITY OF
WATERLOO

FACULTY OF MATHEMATICS
**DAVID R. CHERITON SCHOOL
OF COMPUTER SCIENCE**

September 30, 2022

Problem

Constrained smooth minimization:

$$f_{\star} = \inf_{\mathbf{w} \in C} f(\mathbf{w})$$

- f : smooth and possibly nonconvex
- C : (closed) bounded and convex
- Minimizer may or may not be attained
- Maximization is just negation
- Projection P_C is expensive to compute



...

Alice	1			4	
Bob		2	5		
Carol			4	5	
Dave	5				4
⋮					

Matrix Completion

$$\min_{X:\text{rank}(X)\leq k} \sum_{(i,j)\in\mathcal{O}} (A_{ij} - X_{ij})^2,$$

- rank is nonconvex (in fact, discrete valued)

$$\min_{X:\|X\|_{\text{tr}}\leq\lambda} \sum_{(i,j)\in\mathcal{O}} (A_{ij} - X_{ij})^2,$$

- $\|\cdot\|_{\text{tr}}$: trace norm, sum of singular values
- Let $X = U\Sigma V^{\top}$ be its **singular value decomposition**. Then,

$$P_{\|\cdot\|_{\text{tr}}}(X) = U \text{diag}(\gamma) V^{\top}, \quad \text{where } \gamma = P_{\|\cdot\|_1}(\sigma)$$

- Expensive operation: $O(nm^2)$

Sparsity

$$\min_{\mathbf{w}} \underbrace{\frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2}_{\ell} + \lambda \cdot \underbrace{\|\mathbf{w}\|_0}_r$$

- Balancing square error with sparsity
- ℓ is convex and L-smooth, r is nonsmooth and nonconvex

$$\min_{\mathbf{w}} \underbrace{\frac{1}{n} \|\mathbf{w}\mathbf{X} - \mathbf{y}\|_2^2}_{\ell} + \lambda \cdot \underbrace{\|\mathbf{w}\|_1}_r$$

- Convex relaxation: r is now convex but remains nonsmooth (crucial)

Indicator and Support

Recall that the indicator function of a set C is:

$$\iota_C(\mathbf{w}) = \begin{cases} 0, & \text{if } \mathbf{w} \in C \\ \infty, & \text{otherwise} \end{cases}$$

The **support** function of a set C is:

$$\sigma_C(\mathbf{w}^*) = \max_{\mathbf{w} \in C} \langle \mathbf{w}, \mathbf{w}^* \rangle = \max_{\mathbf{w}} \langle \mathbf{w}, \mathbf{w}^* \rangle - \iota_C(\mathbf{w})$$

- Always (closed) convex and positive homogeneous
- Any norm is a support function of the unit ball of its dual
- The subdifferential $\partial\sigma_C$ will play a crucial role

From Linear to Quadratic

- Suppose we have an algorithm to solve a linear program:

$$\min_{\mathbf{w} \geq \mathbf{0}} \langle \mathbf{w}, \mathbf{c} \rangle \quad \text{s.t.} \quad A\mathbf{w} = \mathbf{b}$$

- How do we solve a quadratic program?

$$\min_{\mathbf{w} \geq \mathbf{0}} \langle \mathbf{w}, A\mathbf{w} \rangle + \langle \mathbf{w}, \mathbf{c} \rangle \quad \text{s.t.} \quad A\mathbf{w} = \mathbf{b}$$

- The power of reduction: try to reduce quadratic to linear!

Algorithm 1: Conditional gradient (condgrad)

Input: $\mathbf{w}_0 \in C$

```
1 for  $t = 0, 1, \dots$  do
2    $\mathbf{z}_t \leftarrow \underset{\mathbf{z} \in C}{\operatorname{argmax}} \langle \mathbf{z}; -\nabla f(\mathbf{w}_t) \rangle$  // polar operator
3   choose step size  $\eta_t \in [0, 1]$ 
4    $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t)\mathbf{w}_t + \eta_t\mathbf{z}_t$  // convex combination
```

- The only nontrivial step in Line 2 has a linear objective
- It is in fact $\partial\sigma_C(-\mathbf{g})$ where $\mathbf{g} = \nabla f(\mathbf{w}_t)$
- We find a point in C that “correlates” the most with $-\nabla f(\mathbf{w}_t)$
- No projection to C needed: Line 4 remains in C

M. Frank and P. Wolfe (1956). “An Algorithm for Quadratic Programming”. *Naval Research Logistics Quarterly*, vol. 3, no. 1-2, pp. 95–110; V. F. Dem’yanov and A. M. Rubinov (1967). “The Minimization of a Smooth Convex Functional on a Convex Set”. *SIAM Journal on Control*, vol. 5, no. 2, pp. 280–294. [English translation of paper in *Vestnik Leningradskogo Universiteta, Seriya Matematiki, Mekhaniki i Astronomii* vol. 19, pp. 7–17, 1964].

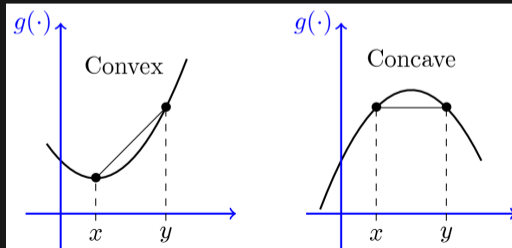
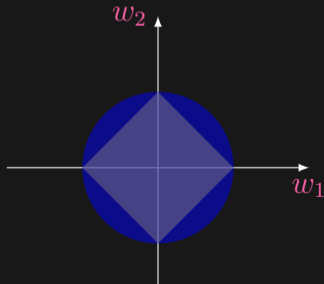
Definition: Extreme point

$\mathbf{w} \in C$ is an **extreme point** (of C) if it does not lie on the line segment of any two points in C . In other words, if $\mathbf{w} \in [\mathbf{w}_1, \mathbf{w}_2], \mathbf{w}_1, \mathbf{w}_2 \in C$ then $\mathbf{w} = \mathbf{w}_1 = \mathbf{w}_2$.

- For a convex set C , $\mathbf{w} \in C$ is an extreme point iff $C \setminus \{\mathbf{w}\}$ remains convex.

Theorem: Convex maximizer is at the boundary

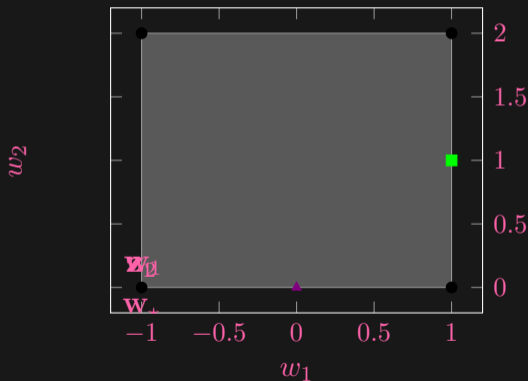
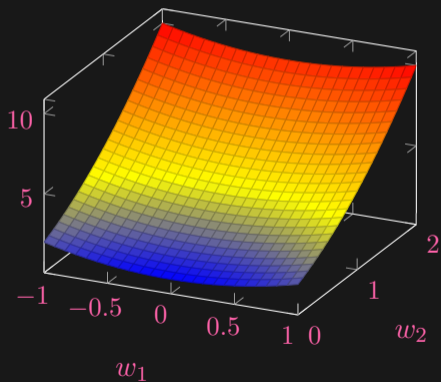
The maximizer of a convex f over C can always be chosen from the extreme points.



Consider the following simple problem:

$$\min_{\mathbf{w} \in C} w_1^2 + (w_2 + 1)^2 \quad \text{and} \quad C := \{\mathbf{w} : w_1 \in [-1, 1], w_2 \in [0, 2]\}.$$

The global minimizer is clearly at $\mathbf{w}_* = (0, 0)$, as shown below.



Let us see how the conditional gradient works on this toy problem:

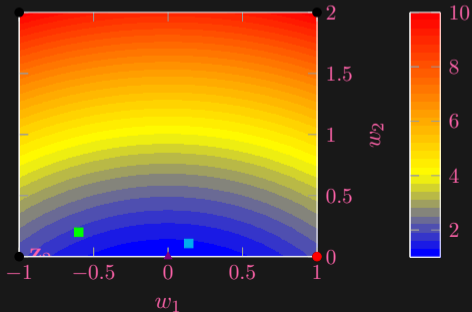
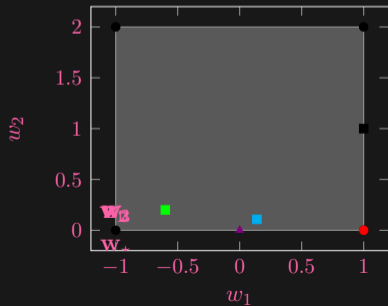
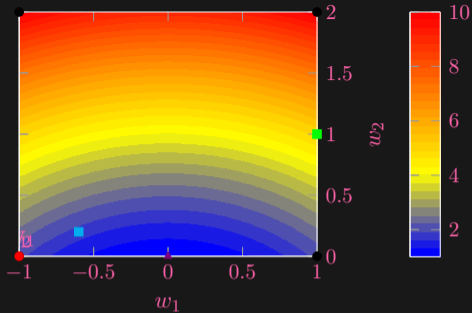
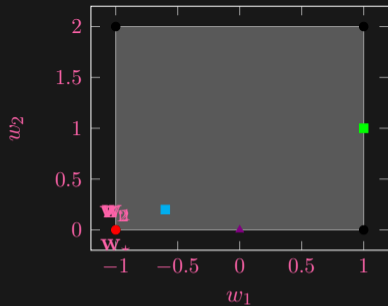
- We first identify the four extreme points of C as

$$\mathbf{z}_1 = (-1, 0), \quad \mathbf{z}_2 = (1, 0), \quad \mathbf{z}_3 = (1, 2), \quad \mathbf{z}_4 = (-1, 2).$$

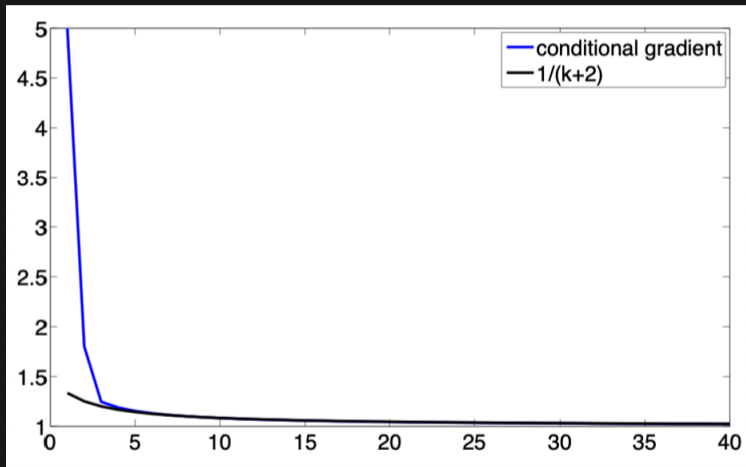
- Start with say $\mathbf{w}_1 = (1, 1)$, we compute the gradient $\nabla f(\mathbf{w}_1) = (2, 4)$.
- We pick the extreme point \mathbf{z} that maximizes $\langle \mathbf{z}; -\nabla f(\mathbf{w}_1) \rangle$. Clearly, \mathbf{z}_1 wins.
- Next, we find $\eta > 0$ to minimize $f((1 - \eta)\mathbf{w}_1 + \eta\mathbf{z}_1)$ by setting its derivative w.r.t. η to 0 :

$$\eta_1 = \eta = \frac{\langle \mathbf{w} + (0, 1), \mathbf{w} - \mathbf{z} \rangle}{\|\mathbf{w} - \mathbf{z}\|_2^2} = \frac{4}{5}.$$

- Lastly, we compute $\mathbf{w}_2 = (1 - \eta_1)\mathbf{w}_1 + \eta_1\mathbf{z}_1 = (-\frac{3}{5}, \frac{1}{5})$, and the process repeats.



Convergence rate closely follows $\Theta(1/t)$, while projected gradient converges in 2 iterations on this example!



Sparsity

Let $C := \{\mathbf{w} : \|\mathbf{w}\|_1 \leq \lambda\}$, whose polar operator reduces to

$$\mathbf{z}_t = \operatorname{argmax}_{\|\mathbf{z}\|_1 \leq \lambda} \langle \mathbf{z}; -\nabla f(\mathbf{w}_t) \rangle \ni -\lambda \mathbf{e}_i, \quad \text{where} \quad \langle \mathbf{e}_i; \nabla_i f(\mathbf{w}_t) \rangle = \max_j |\nabla_j f(\mathbf{w}_t)|.$$

- May choose \mathbf{e}_i to be the i -th standard basis (i.e. 1 at the i -th entry and 0 elsewhere)
- After t steps, the iterate \mathbf{w}_t has (added) at most t nonzeros! In comparison, after even a single iteration, projected gradient can result in a fully dense iterate!
- The resulting coordinate-wise update is a bit wasteful though: we compute the entire gradient ∇f only to find its minimum index and throw out everything else...

Sparsity in Rank

- For the matrix setting:

$$Z_t = \operatorname{argmax}_{\|Z\|_{\text{tr}} \leq \lambda} \langle Z; -\nabla f(W_t) \rangle = -\lambda \mathbf{u} \mathbf{v}^\top, \quad \text{where} \quad \mathbf{u}^\top \nabla f(W_t) \mathbf{v} = \|\nabla f(W_t)\|_{\text{sp}}$$

- After t steps, the iterate W_t has (added) rank at most t
- Computing the spectral norm, i.e. the largest singular value, costs $O(mn)$, an order of magnitude cheaper than projection
- Same for tensors

Theorem: convergence of conditional gradient

Suppose f is convex and L -smooth, and C is compact convex with bounded diameter ρ . Then, conditional gradient satisfies:

$$f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}) + \pi_t(1 - \eta_0)(f(\mathbf{w}_0) - f(\mathbf{w})) + \frac{L\rho^2}{2} \sum_{s=0}^t \frac{\pi_t}{\pi_s} \eta_s^2,$$

where $\pi_t := \prod_{s=1}^t (1 - \eta_s)$ with $\pi_0 := 1$.

- Setting $\eta_t = \frac{2}{t+2}$, we have $\eta_0 = 1$, $\pi_t = \frac{2}{(t+1)(t+2)}$ and

$$f(\mathbf{w}_t) - f(\mathbf{w}) \leq \langle \mathbf{w}_t - \mathbf{z}_t; \nabla f(\mathbf{w}_t) \rangle \leq \frac{2L\rho^2}{t+3},$$

where the initializer \mathbf{w}_0 , surprisingly, does not play any role.

The Proof

$$f(\mathbf{w}_{t+1}) - f(\mathbf{w}) = f((1 - \eta_t)\mathbf{w}_t + \eta_t\mathbf{z}_t) - f(\mathbf{w})$$

$$\text{(L-smoothness)} \leq f(\mathbf{w}_t) - f(\mathbf{w}) + \eta_t \langle \mathbf{z}_t - \mathbf{w}_t; \nabla f(\mathbf{w}_t) \rangle + \frac{\eta_t^2}{2} \underbrace{L \|\mathbf{w}_t - \mathbf{z}_t\|^2}_{\leq \rho^2}$$

$$\text{(optimality of } \mathbf{z}_t) \leq f(\mathbf{w}_t) - f(\mathbf{w}) + \eta_t \langle \mathbf{w} - \mathbf{w}_t; \nabla f(\mathbf{w}_t) \rangle + \frac{\eta_t^2}{2} L \rho^2$$

$$\text{(convexity of } f) \leq (1 - \eta_t)(f(\mathbf{w}_t) - f(\mathbf{w})) + \frac{\eta_t^2}{2} L \rho^2$$

Telescoping and collecting the terms we arrive at the claim

Discussions

- The rate $O(\frac{1}{t})$ is tight and cannot be improved (disappointing)

- Polar operator can be solved **approximately**

- additive error: $\langle \mathbf{z}_t, -\mathbf{g}_t \rangle \leq \max_{\mathbf{w} \in C} \langle \mathbf{w}, -\mathbf{g}_t \rangle - \epsilon_t$
- multiplicative error: $\langle \mathbf{z}_t, -\mathbf{g}_t \rangle \leq \frac{1}{\alpha_t} \cdot \max_{\mathbf{w} \in C} \langle \mathbf{w}, -\mathbf{g}_t \rangle$

- Choices of the step size η_t

- Open-loop rule: $\eta_t = \frac{2}{t+2}$, or more generally $\eta_t = \Theta(1/t)$.

- Cauchy's rule: $\eta_t \in \operatorname{argmin}_{0 \leq \eta \leq 1} f((1 - \eta)\mathbf{w}_t + \eta\mathbf{z}_t)$.

- Quadratic rule:

$$\eta_t = \operatorname{argmin}_{0 \leq \eta \leq 1} f(\mathbf{w}_t) + \eta_t \langle \mathbf{z}_t - \mathbf{w}_t; \nabla f(\mathbf{w}_t) \rangle + \frac{L^2 \eta_t^2 \|\mathbf{w}_t - \mathbf{z}_t\|^2}{2} = \left[\frac{\langle \mathbf{w}_t - \mathbf{z}_t; \nabla f(\mathbf{w}_t) \rangle}{L^2 \|\mathbf{w}_t - \mathbf{z}_t\|^2} \right]_0^1.$$

- Possible to accelerate

Extension to Composite

$$\min_{\mathbf{w}} f(\mathbf{w}), \quad \text{where } \ell(\mathbf{w}) + r(\mathbf{w})$$

Algorithm 2: Generalized conditional gradient (GCG)

Input: $\mathbf{w}_0 \in C$, functions ℓ and r

```
1 for  $t = 0, 1, \dots$  do
2    $\mathbf{z}_t \leftarrow \underset{\mathbf{z}}{\operatorname{argmin}} \langle \mathbf{z}; \nabla \ell(\mathbf{w}_t) \rangle + r(\mathbf{w})$  // conjugate of  $r$ 
3   choose step size  $\eta_t \in [0, 1]$ 
4    $\mathbf{w}_{t+1} \leftarrow (1 - \eta_t)\mathbf{w}_t + \eta_t \mathbf{z}_t$  // convex combination
```

T. Bonesky et al. (2007). "A Generalized Conditional Gradient Method for Nonlinear Operator Equations with Sparsity Constraints". *Inverse Problems*, vol. 23, no. 5, pp. 2041–2058; K. Bredies et al. (2009). "A Generalized Conditional Gradient Method and its Connection to an Iterative Shrinkage Method". *Computational Optimization and Applications*, vol. 42, pp. 173–193; Y. Yu et al. (2017). "Generalized Conditional Gradient for Structured Sparse Estimation". *Journal of Machine Learning Research*, vol. 18, pp. 1–46.

Totally Corrective

- Inspecting the conditional gradient algorithm we realize that

$$\mathbf{w}_{t+1} \in \text{conv}\{\mathbf{w}_0, \mathbf{z}_1, \dots, \mathbf{z}_t\},$$

where the extreme points \mathbf{z}_k are repeatedly identified and averaged.

- One immediate, natural idea is to replace the next iterate \mathbf{w}_{t+1} as the best approximation in the entire convex hull:

$$\mathbf{w}_{t+1} = \underset{\mathbf{w} \in \text{conv}\{\mathbf{w}_0, \mathbf{z}_1, \dots, \mathbf{z}_t\}}{\text{argmin}} \quad f(\mathbf{w}).$$

- Potentially much faster, but more expensive in each step
- Can restrict memory size, even to 2

G. Meyer (1974). "Accelerated Frank–Wolfe Algorithms". *SIAM Journal on Control*, vol. 12, no. 4, pp. 655–655; C. A. Holloway (1974). "An extension of the Frank and Wolfe method of feasible directions". *Mathematical Programming*, vol. 6, pp. 14–27.

