EECS 559 Optimization Methods for SIPML

Lecture 10 – Frank-Wolfe (Conditional Gradient) Method

Instructor: Prof. Qing Qu (qingqu@umich.edu)

Lecture Agenda

- Method Introduction
- Application Examples
- Convergence Analysis

Lecture Agenda

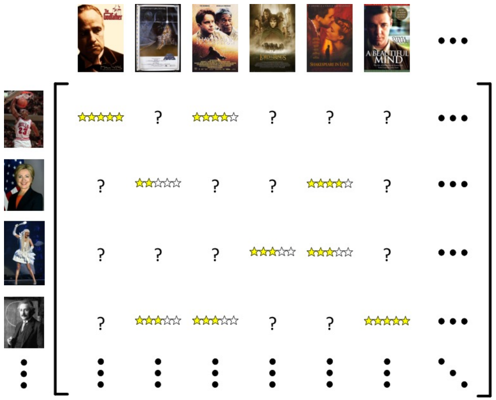
- Method Introduction
- Application Examples
- Convergence Analysis

Solving Unconstraint Problems

- **Stable low-rank matrix completion**

Given incomplete observation $\boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$ and $\Omega \in [n_1] \times [n_2]$, we want to recover a low-rank matrix $\boldsymbol{X}_\star$ from noise $\boldsymbol{Z}$ corruption:

$$\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X}_\star) + \boldsymbol{Z}$$



We solve the following unconstrained problem:

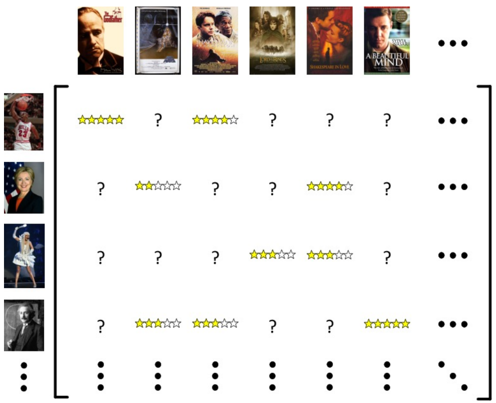
$$\min_{\boldsymbol{X}} \frac{1}{2} \|\mathcal{P}_\Omega(\boldsymbol{X}) - \boldsymbol{Y}\|_F^2 + \lambda \|\boldsymbol{X}\|_* .$$

Solving Unconstraint Problems

- **Stable low-rank matrix completion**

We solve the following unconstrained problem:

$$\min_{\boldsymbol{X}} \ \frac{1}{2} \left\| \mathcal{P}_{\Omega}(\boldsymbol{X}) - \boldsymbol{Y} \right\|_F^2 + \lambda \left\| \boldsymbol{X} \right\|_* .$$



- Methods: proximal gradient, ADMM, etc.
- For each iteration, needs to evaluate

$$\mathrm{prox}_{\lambda \|\cdot\|_*}(\boldsymbol{W}) = \boldsymbol{U} \mathcal{S}_{\mu}(\boldsymbol{\Sigma}) \boldsymbol{V}^\top, \quad \text{with} \quad \boldsymbol{W} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top.$$

which is usually quite expensive $O(n_1 n_2 \min\{n_1, n_2\})$.

Constrained Problem Formulations

- **Stable low-rank matrix completion**

We consider a variant of *constrained* problem formulation:

$$\min_{\boldsymbol{X}} \underbrace{\frac{1}{2}\left\|\mathcal{P}_{\Omega}(\boldsymbol{X})-\boldsymbol{Y}\right\|_F^2,}_{f(\boldsymbol{X})} \quad \text{s.t.} \quad \underbrace{\|\boldsymbol{X}\|_* \leq \tau}_{\boldsymbol{X}\in\mathcal{C}}.$$

We introduce a new optimization strategy, reducing per iteration complexity from $O(n_1 n_2 \min\{n_1, n_2\})$ to $O(n_1 n_2 r)$, with $r = \operatorname{rank}(\boldsymbol{X}_\star)$.

Constrained Problem Formulations

- **Noisy sparse recovery**

Given the measurement $\boldsymbol{y}$ and sensing matrix $\boldsymbol{A}$, we want to recover a *sparse* $\boldsymbol{x}_\star$ from

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_\star + \boldsymbol{n}.$$

We can consider a *constrained* variant of Lasso problem:

$$\min_{\boldsymbol{x}} \underbrace{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2}_{f(\boldsymbol{x})}, \quad \text{s.t.} \quad \underbrace{\|\boldsymbol{x}\|_1 \le \tau}_{\boldsymbol{x} \in \mathcal{C}}.$$

Structured Constrained Optimization

$$\boxed{\min_{\boldsymbol{x}} \; f(\boldsymbol{x}), \quad \text{s.t.} \quad \boldsymbol{x} \in \mathcal{C}.}$$

- $f : \mathbb{R}^n \mapsto R$ is a *convex, continuously differentiable* function, whose gradient $\nabla f(\boldsymbol{x})$ is L-Lipschitz

$$\left\| \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}') \right\|_2 \leq L \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_2, \quad \forall \boldsymbol{x}, \; \boldsymbol{x}' \in \mathbb{R}^n.$$

- The constraint set $\mathcal{C}$ is a *compact, convex set* with a diameter

$$\operatorname{diam}(\mathcal{C}) := \sup \left\{ \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|_2 \mid \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C} \right\}.$$

Minimizing a 1st Order Approximation

Interpretation of Frank-Wolfe method:

- At each point $\boldsymbol{x}_k$, we form a first-order approximation of f :

$$f(\boldsymbol{v}) \approx \widehat{f}(\boldsymbol{v}, \boldsymbol{x}_k) := f(\boldsymbol{x}_k) + \langle \boldsymbol{v} - \boldsymbol{x}_k, \nabla f(\boldsymbol{x}_k) \rangle .$$

- Minimize the approximation $\widehat{f}(\boldsymbol{v}, \boldsymbol{x}_k)$ over $\boldsymbol{v} \in \mathcal{C}$ by

$$\boldsymbol{v}_k = \arg\min_{\boldsymbol{v} \in \mathcal{C}} \widehat{f}(\boldsymbol{v}, \boldsymbol{x}_k) = \arg\min_{\boldsymbol{v} \in \mathcal{C}} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}_k) \rangle .$$

- Take a step in the direction $\boldsymbol{w}_k = \boldsymbol{v}_k - \boldsymbol{x}_k$:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \gamma_k \boldsymbol{w}_k = (1 - \gamma_k) \boldsymbol{x}_k + \gamma_k \boldsymbol{v}_k$$

Frank-Wolfe Method

- At each iteration, we generate a new point $\boldsymbol{v}_k$ by solving

$$\boldsymbol{v}_k \in \arg\min_{\boldsymbol{v} \in \mathcal{C}} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}_k) \rangle$$

- We then set

$$\boldsymbol{x}_{k+1} = (1 - \gamma_k)\boldsymbol{x}_k + \gamma_k \boldsymbol{v}_k \in \mathcal{C}.$$

with $\gamma_k \in (0,1)$ determined by line search or $\gamma_k = \frac{2}{k+2}$.

Frank-Wolfe Method

Frank-Wolfe Method (FW)

Problem Class:

$$\begin{array}{ll} \min_{\boldsymbol{x}} & f(\boldsymbol{x}), \\ \text{subject to} & \boldsymbol{x} \in \mathsf{C}. \end{array}$$

$f : \mathbb{R}^n \to \mathbb{R}$ convex, differentiable, $\nabla f(\boldsymbol{x})$ L-Lipschitz.
C a compact convex set.

Basic Iteration: Repeat

$$\boldsymbol{v}_k \in \arg\min_{\boldsymbol{v} \in \mathsf{C}} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}_k) \rangle,$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \gamma_k (\boldsymbol{v}_k - \boldsymbol{x}_k),$$

with $\gamma_k = \frac{2}{k+2}$.

Convergence Guarantee:

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) \leq \frac{2L \operatorname{diam}^2(\mathsf{C})}{k+2}.$$

Solving Frank-Wolfe Subproblem

The core of Frank-Wolfe method is to solve the following

$$\boxed{\boldsymbol{v}_k \;=\; \arg\min_{\boldsymbol{v}\in\mathcal{C}} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}_k)\rangle\,.}$$

- The constrained problem can be difficult to solve in general;
- However, by utilizing the structure of $\mathcal{C}$, it can be solved efficiently in many cases.

Lecture Agenda

- Method Introduction
- **Application Examples**
- Convergence Analysis

Example I: Stable Matrix Completion

$$\min_{\boldsymbol{X}} \underbrace{\frac{1}{2}\left\|\mathcal{P}_{\Omega}(\boldsymbol{X})-\boldsymbol{Y}\right\|_F^2}_{f(\boldsymbol{X})}, \quad \text{s.t.} \quad \underbrace{\|\boldsymbol{X}\|_* \leq \tau}_{\boldsymbol{X}\in\mathcal{C}}.$$

- **Frank-Wolfe Method:**

$$\boldsymbol{V}_k = -\tau \boldsymbol{u}_1\boldsymbol{v}_1^\top \in \arg\min_{\boldsymbol{V}\in\mathcal{C}} \left\langle \boldsymbol{V}, \nabla f(\boldsymbol{X}_k)\right\rangle, \ \nabla f(\boldsymbol{X}_k) = \sum_{i=1}^{n_1} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top.$$

$$\boldsymbol{X}_{k+1} = (1-\gamma_k)\boldsymbol{X}_k + \gamma_k \boldsymbol{V}_k.$$

Example I: Nuclear Norm Ball

- **Stable low-rank matrix completion**

$$\min_{\boldsymbol{X}} \underbrace{\frac{1}{2}\left\|\mathcal{P}_{\Omega}(\boldsymbol{X}) - \boldsymbol{Y}\right\|_F^2}_{f(\boldsymbol{X})}, \quad \text{s.t.} \quad \underbrace{\left\|\boldsymbol{X}\right\|_* \le \tau}_{\boldsymbol{X}\in\mathcal{C}}.$$

The subproblem is of the form

$$\min_{\boldsymbol{X}} \left\langle \nabla f(\boldsymbol{X}_0), \boldsymbol{X} \right\rangle, \quad \text{s.t.} \quad \left\|\boldsymbol{X}\right\|_* \le \tau.$$

Example I: Stable Matrix Completion

Given a matrix $\boldsymbol{G}$, consider the problem

$$\min_{\boldsymbol{X}} \ \langle \boldsymbol{G}, \boldsymbol{X} \rangle \,, \quad \text{s.t.} \quad \|\boldsymbol{X}\|_* \ \le \ \tau.$$

- The problem has a *closed-form* solution

$$\boxed{\boldsymbol{X}_\star \ = \ -\tau \boldsymbol{u}_1 \boldsymbol{v}_1^\top \,,}$$

where the singular value decomposition $\boldsymbol{G} = \sum_{i=1}^n \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$.

- Computing the leading singular value can be much cheaper than the full SVD: $O(n_1 n_2)$ vs. $O(n_1 n_2 \min\{n_1, n_2\})$.

Example I: Stable Matrix Completion

Algorithm 8.8 Frank-Wolfe for Stable Matrix Completion

Problem: given $\boldsymbol{Y} \in \mathbb{R}^{n_1 \times n_2}$ and $\Omega \subseteq [n_1] \times [n_2]$,

$$\min_{\boldsymbol{X}} \frac{1}{2} \|\mathcal{P}_\Omega[\boldsymbol{X}] - \boldsymbol{Y}\|_F^2 \quad \text{subject to} \quad \|\boldsymbol{X}\|_* \leq \tau.$$

Input: $\boldsymbol{X}_0 \in \mathbb{R}^{n_1 \times n_2}$ satisfying $\|\boldsymbol{X}_0\|_* \leq \tau$.

for $(k = 0, 1, 2, \ldots, K-1)$ **do**

$\quad (\boldsymbol{u}_1, \sigma_1, \boldsymbol{v}_1) \leftarrow \text{LeadSV}\left(\mathcal{P}_\Omega\left[\boldsymbol{X}_k - \boldsymbol{Y}\right]\right)$.

$\quad \boldsymbol{V}_k \leftarrow -\tau \boldsymbol{u}_1 \boldsymbol{v}_1^*$.

$\quad \boldsymbol{X}_{k+1} \leftarrow \frac{k}{k+2} \boldsymbol{X}_k + \frac{2}{k+2} \boldsymbol{V}_k$.

end for

Output: $\boldsymbol{X}_\star \leftarrow \boldsymbol{X}_K$.

Example II: Noisy Sparse Recovery

$$\min_{\boldsymbol{x}} \underbrace{\frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2}_{f(\boldsymbol{x})}, \quad \text{s.t.} \quad \underbrace{\|\boldsymbol{x}\|_1 \leq \tau}_{\boldsymbol{x}\in\mathcal{C}}.$$

- **Frank-Wolfe Method:**

$$\boldsymbol{v}_k = -\tau\sigma_i\boldsymbol{e}_i \in \arg\min_{\boldsymbol{v}\in\mathcal{C}} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}_k)\rangle, \quad \nabla f(\boldsymbol{x}) = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}).$$

$$\boldsymbol{x}_{k+1} = (1-\gamma_k)\boldsymbol{x}_k + \gamma_k\boldsymbol{v}_k.$$

Closely related to greedy methods for sparse approximations such as matching pursuit (MP), and orthogonal MP (OMP).

Example II: ℓ_1-Norm Ball

- **Noisy sparse recovery**

$$\min_{\boldsymbol{x}} \underbrace{\frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\right\|_2^2}_{f(\boldsymbol{x})}, \quad \text{s.t.} \quad \underbrace{\left\|\boldsymbol{x}\right\|_1 \le \tau}_{\boldsymbol{x}\in\mathcal{C}}.$$

The subproblem is of the form

$$\min_{\boldsymbol{x}} \left\langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} \right\rangle, \quad \text{s.t.} \quad \left\|\boldsymbol{x}\right\|_1 \le \tau.$$

Example II: Noisy Sparse Recovery

Given a vector $\boldsymbol{g}$, consider the problem

$$\min_{\boldsymbol{x}} \ \langle \boldsymbol{g}, \boldsymbol{x} \rangle , \quad \text{s.t.} \quad \|\boldsymbol{x}\|_1 \ \leq \ \tau.$$

The problem has a *closed-form* solution:

$$\boxed{\boldsymbol{x}_\star \ = \ -\tau \sigma_i \boldsymbol{e}_i,}$$

where $\boldsymbol{e}_i$ is the i-th standard basis vector with $i = \arg\max_j |g_j|$ and $\sigma_i = \text{sign}(g_i)$.

Example II: Noisy Sparse Recovery

Algorithm 8.9 Frank-Wolfe for Noisy Sparse Recovery

Problem: given $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$,

$$\min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{x}\|_1 \leq \tau.$$

Input: $\boldsymbol{x}_0 \in \mathbb{R}^n$ satisfying $\|\boldsymbol{x}_0\|_1 \leq \tau$.

for $(k = 0, 1, 2, \ldots, K-1)$ **do**

  $\boldsymbol{r}_k \leftarrow \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{y}$.

  $i_k \leftarrow \arg\max_i |\boldsymbol{a}_i^* \boldsymbol{r}_k|$.

  $\sigma \leftarrow \operatorname{sign}\left(\boldsymbol{a}_{i_k}^* \boldsymbol{r}_k\right)$.

  $\boldsymbol{v}_k \leftarrow -\tau \sigma \boldsymbol{e}_{i_k}$.

  $\boldsymbol{x}_{k+1} \leftarrow \frac{k}{k+2}\boldsymbol{x}_k + \frac{2}{k+2}\boldsymbol{v}_k$.

end for

Output: $\boldsymbol{x}_\star \leftarrow \boldsymbol{x}_K$.

Example III: Nonconvex Problems

Example: Given a symmetric $\boldsymbol{Q}$, find the leading eigenvector of $\boldsymbol{Q}$ via

$$\min_{\boldsymbol{x}} -\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}, \quad \text{s.t.} \quad \|\boldsymbol{x}\|_2 \leq 1.$$

- **Frank-Wolfe Method:**

$$\boldsymbol{v}_k = \arg\min_{\|\boldsymbol{x}\|_2 \leq 1} \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} \rangle = -\frac{\nabla f(\boldsymbol{x}_k)}{\|\nabla f(\boldsymbol{x}_k)\|_2} = \frac{\boldsymbol{Q}\boldsymbol{x}_k}{\|\boldsymbol{Q}\boldsymbol{x}_k\|_2}.$$

$$\boldsymbol{x}_{k+1} = (1-\gamma_k)\boldsymbol{x}_k + \gamma_k \boldsymbol{v}_k.$$

Example III: Nonconvex Problems

Example: Given a symmetric $\boldsymbol{Q}$, find the leading find the leading eigenvector $\boldsymbol{Q}$ via

$$\min_{\boldsymbol{x}} -\boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x}, \quad \text{s.t.} \quad \|\boldsymbol{x}\|_2 \leq 1.$$

- **Frank-Wolfe Method:**

$$\gamma_k \;=\; \arg\min_{0 \leq \gamma \leq 1} f\left((1-\gamma)\boldsymbol{x}_k + \gamma \frac{\boldsymbol{Q}\boldsymbol{x}_k}{\|\boldsymbol{Q}\boldsymbol{x}_k\|_2} \right) = 1.$$

- This gives power method: $\boldsymbol{x}_{k+1} \;=\; \boldsymbol{Q}\boldsymbol{x}_k / \|\boldsymbol{Q}\boldsymbol{x}_k\|_2 .$

Lecture Agenda

- Method Introduction
- Application Examples
- Convergence Analysis

Convergence of Frank-Wolfe

Theorem. (Convergence of Frank-Wolfe, Jaggi'13)
Let f be convex and L-smooth. With $\gamma_k = \frac{2}{k+2}$, one has

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) \le \frac{2L \cdot \mathrm{diam}^2(\mathcal{C})}{k+2},$$

with $\mathrm{diam}(\mathcal{C}) = \sup_{\boldsymbol{x},\boldsymbol{x}'} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$.

- For compact constraint set $\mathcal{C}$, Frank-Wolfe attains ε-accuracy with $O(1/\varepsilon)$ iterations, or the convergence rate is $O(1/k)$.
- For nonconvex function f, the rate reduces to $O(1/\sqrt{k})$.

Proof of Convergence

Proof. For the ease of notation, $d := \mathrm{diam}(\mathcal{C}) = \sup_{\boldsymbol{x},\boldsymbol{x}' \in C} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$

Note that

$$\boldsymbol{x}_{k+1} = (1-\gamma_k)\boldsymbol{x}_k + \gamma_k \boldsymbol{v}_k$$

$$\boldsymbol{v}_k = \arg\min_{\boldsymbol{v} \in \mathcal{C}} \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v} \rangle$$

1. Given that f is L-smooth: $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{x}'\|_2$, $\forall\ \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{C}$

$$\Rightarrow \quad \begin{aligned} f(\boldsymbol{x}_{k+1}) &\leq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_{k+1} - \boldsymbol{x}_k \rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 \\ &\leq f(\boldsymbol{x}_k) + \gamma_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v}_k - \boldsymbol{x}_k \rangle + \frac{L}{2}\gamma_k^2 \|\boldsymbol{v}_k - \boldsymbol{x}_k\|_2^2 \\ &\leq f(\boldsymbol{x}_k) + \gamma_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v}_k - \boldsymbol{x}_k \rangle + \frac{L}{2}\gamma_k^2 d^2 \end{aligned}$$

Proof of Convergence

2. Given f is convex, suppose $\boldsymbol{x}_\star$ is optimal in $\mathcal{C}$,

$$\begin{aligned} f(\boldsymbol{x}_\star) &\geq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_\star - \boldsymbol{x}_k \rangle \\ &\geq f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v}_k - \boldsymbol{x}_k \rangle \end{aligned}$$

$$\Rightarrow \quad \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v}_k - \boldsymbol{x}_k \rangle \leq -\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)\right)$$

Combine 1 & 2, we find

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_\star) &\leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) + \gamma_k \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{v}_k - \boldsymbol{x}_k \rangle + \frac{\gamma_k^2 L}{2} d^2 \\ &\leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) - \gamma_k \left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)\right) + \frac{\gamma_k^2 L}{2} d^2 \\ &\leq (1 - \gamma_k)\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)\right) + \frac{\gamma_k^2 L}{2} d^2 \end{aligned}$$

Proof of Convergence

Based upon the above, we can prove by **induction**.

- First, for k=1, we have

$$\varepsilon_1 = f(\boldsymbol{x}_1) - f(\boldsymbol{x}_\star) \le \frac{\gamma_0^2 L}{2} d^2 \le \frac{Ld^2}{2} \le \frac{2}{3} Ld^2 = \frac{2Ld^2}{2+1}$$

- Second, let us assume $\varepsilon_l \le \frac{2}{2+l} Ld^2$ for all $l = 1, \cdots, k$. We want to show that it also holds for k+1.

$$\varepsilon_{k+1} \le \frac{2}{2+(k+1)} Ld^2$$

Proof of Convergence

To show above, we have

$$
\begin{aligned}
\varepsilon_{k+1} &\leq (1-\gamma_k)\varepsilon_k + \frac{\gamma_k^2 L}{2} d^2 \\
&\leq (1-\gamma_k)\frac{2}{2+k} L d^2 + \frac{\gamma_k^2 L}{2} d^2 \\
&= \left(1 - \frac{2}{k+2}\right)\frac{2}{k+2} L d^2 + \frac{L}{2}\left(\frac{2}{k+2}\right)^2 d^2 \\
&\leq \frac{2(k+1)}{(k+2)^2} L d^2 \leq \frac{2}{(k+1)+2} L d^2
\end{aligned}
$$

which finishes our prove.

Example III: Nonconvex Problems

More generally, we have

Theorem (Nonconvex Convergence, Lacoste-Julien'16)
Suppose f is L-smooth but not necessarily convex. Then the Frank-Wolfe method with line search produces iterates that

$$g_k \;\leq\; \frac{L}{\sqrt{k+1}} \left(f(\boldsymbol{x}_0) - f(\boldsymbol{x}_\star) \right),$$

where $g_k = \min_{0 \leq t \leq k} \langle \boldsymbol{v}_t - \boldsymbol{x}_t, -\nabla f(\boldsymbol{x}_t) \rangle$.

Strongly Convex Problems?

Claim: the convergence rate of Franke-Wolfe *cannot* be improved even when f is strongly convex.

A negative example:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \ \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x}, \quad \text{s.t.} \quad \boldsymbol{x} \in \underbrace{\mathrm{Conv}\left\{\boldsymbol{a}_1, \cdots \boldsymbol{a}_k\right\}}_{\mathcal{C}}.$$

- Suppose $\text{interior}\,(\mathcal{C}) \neq \emptyset$
- Suppose the optimal point $\boldsymbol{x}_\star$ lies on the boundary of $\mathcal{C}$, and it is not an extreme point.

Strongly Convex Problems?

Theorem. (Canon & Cullum'68)
Let $\{\boldsymbol{x}_k\}_{k\geq 1}$ be Frank-Wolfe iterates with exact line search for solving the problem above. Then there exists an initial point $\boldsymbol{x}_0$, such that for every $\varepsilon > 0$,

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star) \geq \frac{1}{k^{1+\varepsilon}}, \quad \text{for infinitely many } k.$$

- Choose $\boldsymbol{x}_0 \in \text{interior}(\mathcal{C})$ obeying $f(\boldsymbol{x}_0) < \min_i f(\boldsymbol{a}_i)$.
- In general, *cannot* improve $O(1/k)$ convergence guarantees.

A Tight Upper Bound on the Rate of Convergence of Frank-Wolfe Algorithm. M. D. Canon and C. D. Cullum, SIAM Journal on Control, 1968.

An Example of Positive Results

To achieve linear convergence, one needs additional assumptions:

Definition. (strongly convex set)
A set $\mathcal{C}$ is said to be μ-strongly convex if $\forall \lambda \in [0,1]$,
and $\forall \boldsymbol{x}, \boldsymbol{z} \in \mathcal{C}$:

$$\mathcal{B}\left(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{z}, \frac{\mu}{2}\lambda(1-\lambda)\left\|\boldsymbol{x}-\boldsymbol{z}\right\|_2^2\right) \in \mathcal{C},$$

where $\mathcal{B}(\boldsymbol{x}, r) := \{\boldsymbol{z} \in \mathbb{R}^n \mid \|\boldsymbol{z}-\boldsymbol{x}\|_2 \leq r\}$.

- Example: ℓ_2-ball.

An Example of Positive Results

Theorem. (Levitin & Polyak '66)
Suppose f is convex and L-smooth, and $\mathcal{C}$ is μ-strongly convex. Suppose $\|\nabla f(\boldsymbol{x})\|_2 \geq c > 0$ for all $\boldsymbol{x} \in \mathcal{C}$. Then under mild conditions, Frank-Wolfe with exact line search converges linearly.

- Study of the linear convergence conditions for Frank-Wolfe method is still an active research area (Lacoste-Julien'15).

Constrained Minimization Methods. E.S. Levitin and B.T. Polyak, USSR Computational Mathematics and Mathematical Physics, 1966.

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Chapter 8.6**)

- *Conditional Gradient Algorithms for Rank-one Matrix Approximations with a Sparsity Constraint*. Ronny Luss, Marc Teboulle, SIAM Review, 2013. https://arxiv.org/abs/1107.1163

- *Revisiting Frank-Wolfe: Projection-free Sparse Convex Optimization*. Martin Jaggi, ICML, 2013. http://proceedings.mlr.press/v28/jaggi13.html

Further Readings

- *On the Global Linear Convergence of Frank-Wolfe Optimization Variants.* Simon Lacoste-Julien, Martin Jaggi, NeurIPS'15, 2015. *https://arxiv.org/abs/1511.05932*

- *A Tight Upper Bound on the Rate of Convergence of Frank-Wolfe Algorithm.* M. D. Canon and C. D. Cullum, SIAM Journal on Control, 1968.

- *Constrained Minimization Methods.* E.S. Levitin and B.T. Polyak, USSR Computational Mathematics and Mathematical Physics, 1966.