Optimization for Machine Learning

Lecture 10: Frank-Wolfe Methods 6.881: MIT

Suvrit Sra
(Acknowledgements: Alp Yurtsever)
Massachusetts Institute of Technology

25 Mar, 2021

Motivation: constrained optimization

$$\min_{x \in \mathcal{M}} \quad f(x)$$

$\mathcal{M} \subseteq \mathbb{R}^d$ is convex and compact.

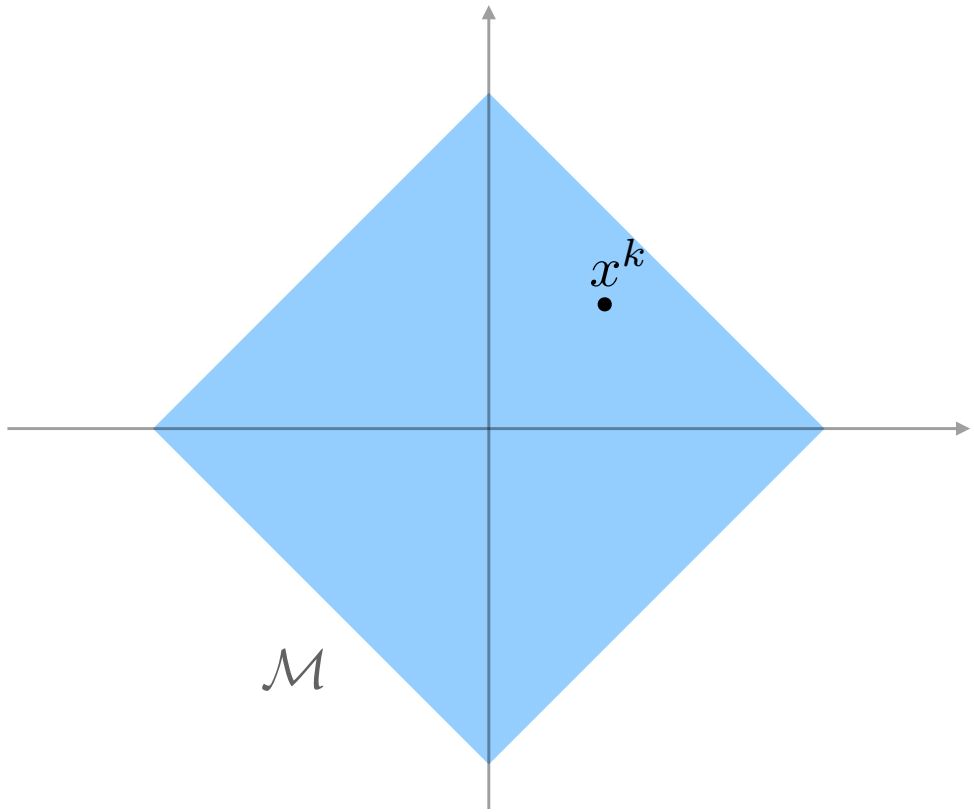
Motivation: constrained optimization

$$\min_{x \in \mathcal{M}} \quad f(x)$$

$\mathcal{M} \subseteq \mathbb{R}^d$ is convex and compact.

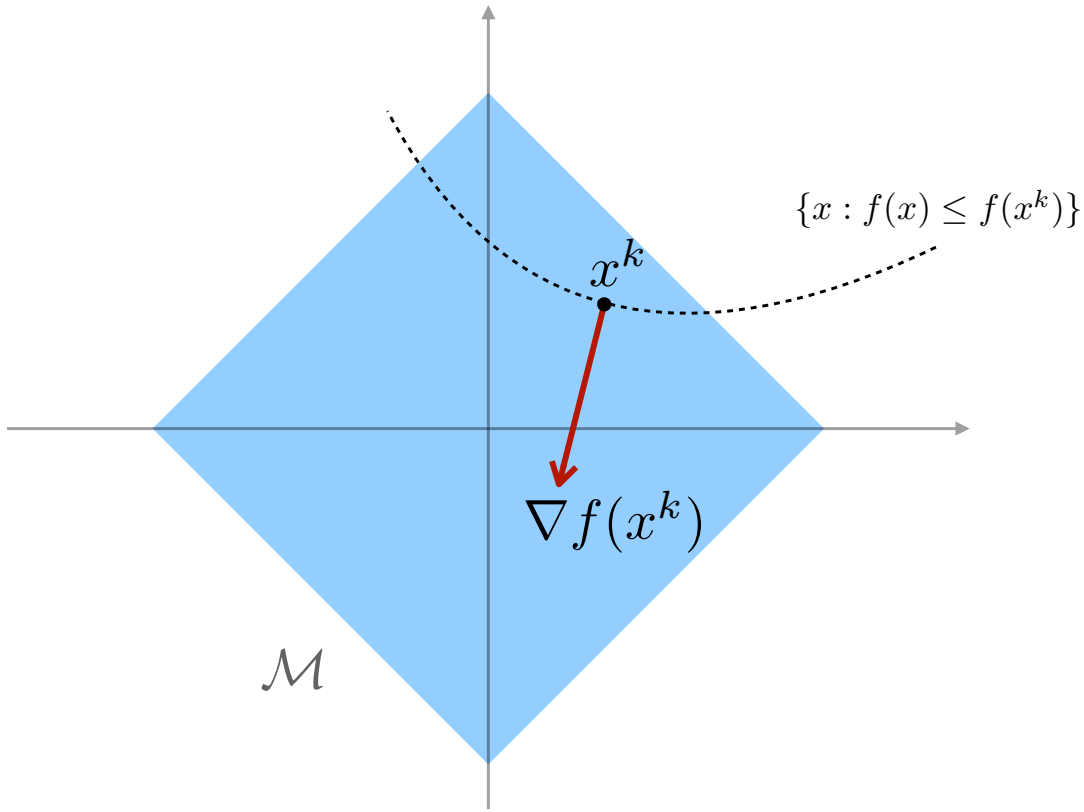
What ways have we seen so far?

Projected gradient method



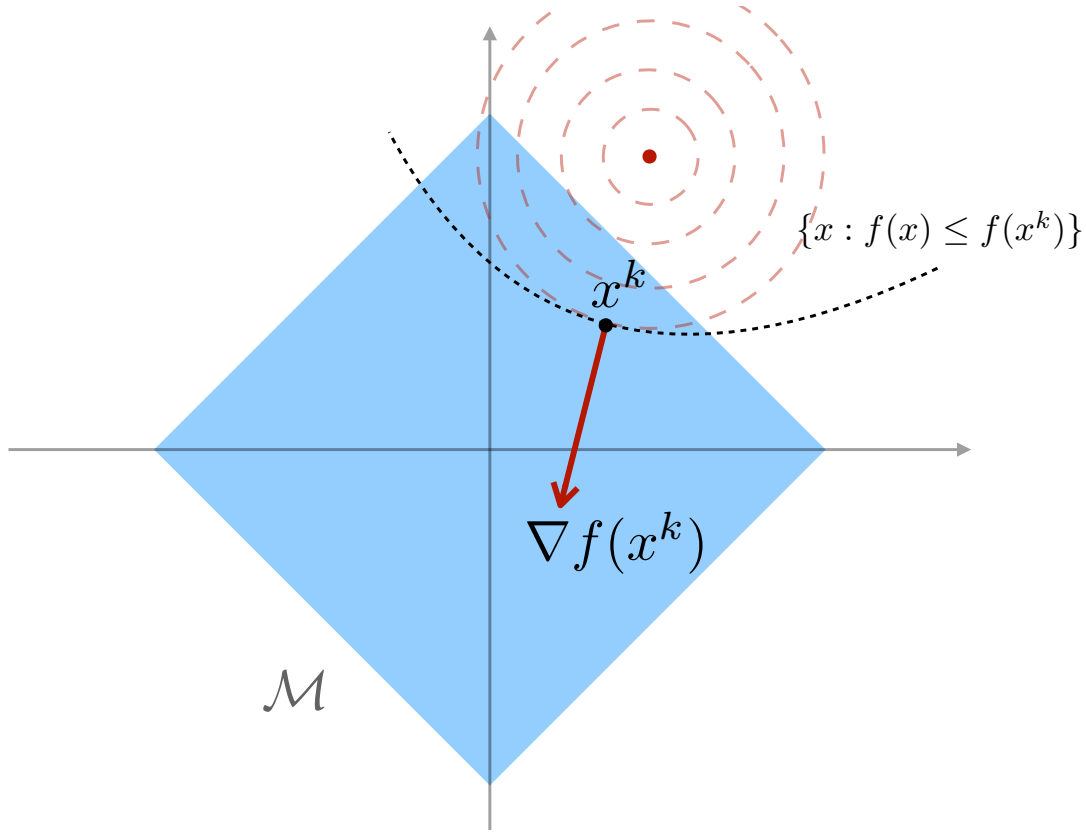
$$x^{k+1} = P_{\mathcal{M}}\left(x^k - \tfrac{1}{L}\nabla f(x^k)\right)$$
$$= \operatorname*{argmin}_{x\in\mathcal{M}} f(x^k) + \langle \nabla f(x^k),\, x - x^k\rangle + \tfrac{L}{2}\|x - x^k\|^2$$

Projected gradient method



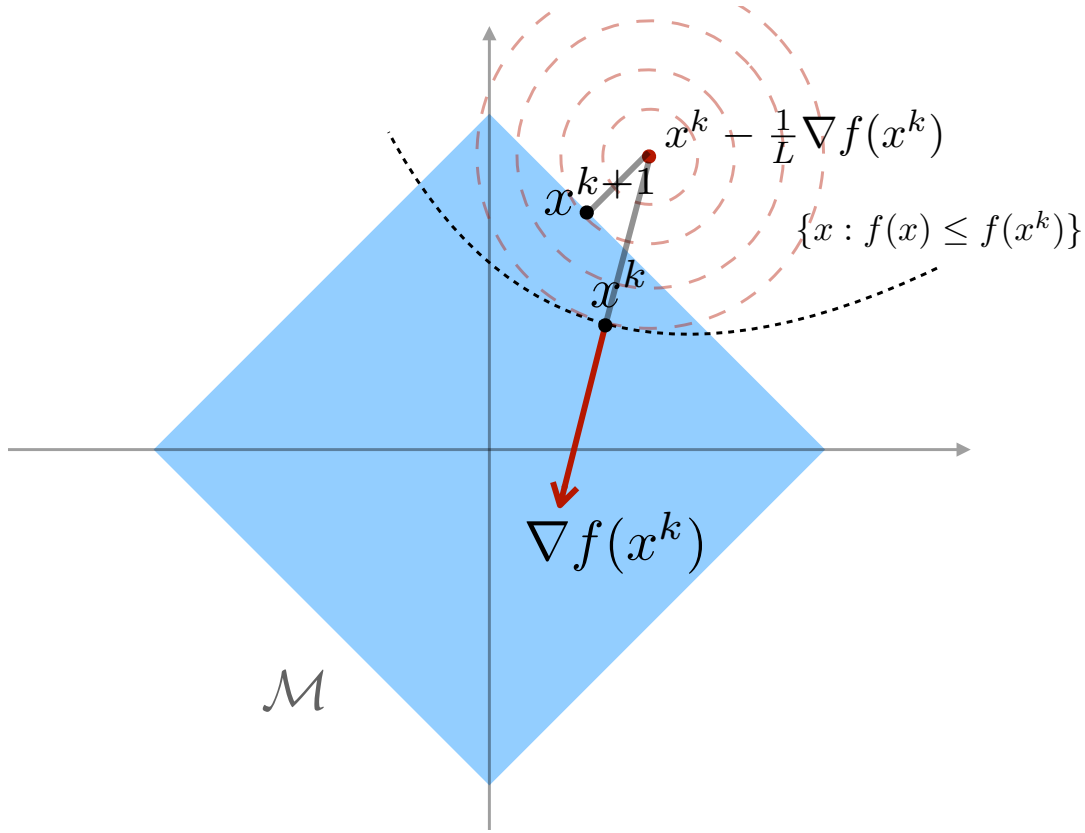
$$x^{k+1} = P_{\mathcal{M}}\left(x^k - \tfrac{1}{L}\nabla f(x^k)\right)$$
$$= \operatorname*{argmin}_{x \in \mathcal{M}} f(x^k) + \langle \nabla f(x^k),\, x - x^k \rangle + \tfrac{L}{2}\|x - x^k\|^2$$

Projected gradient method



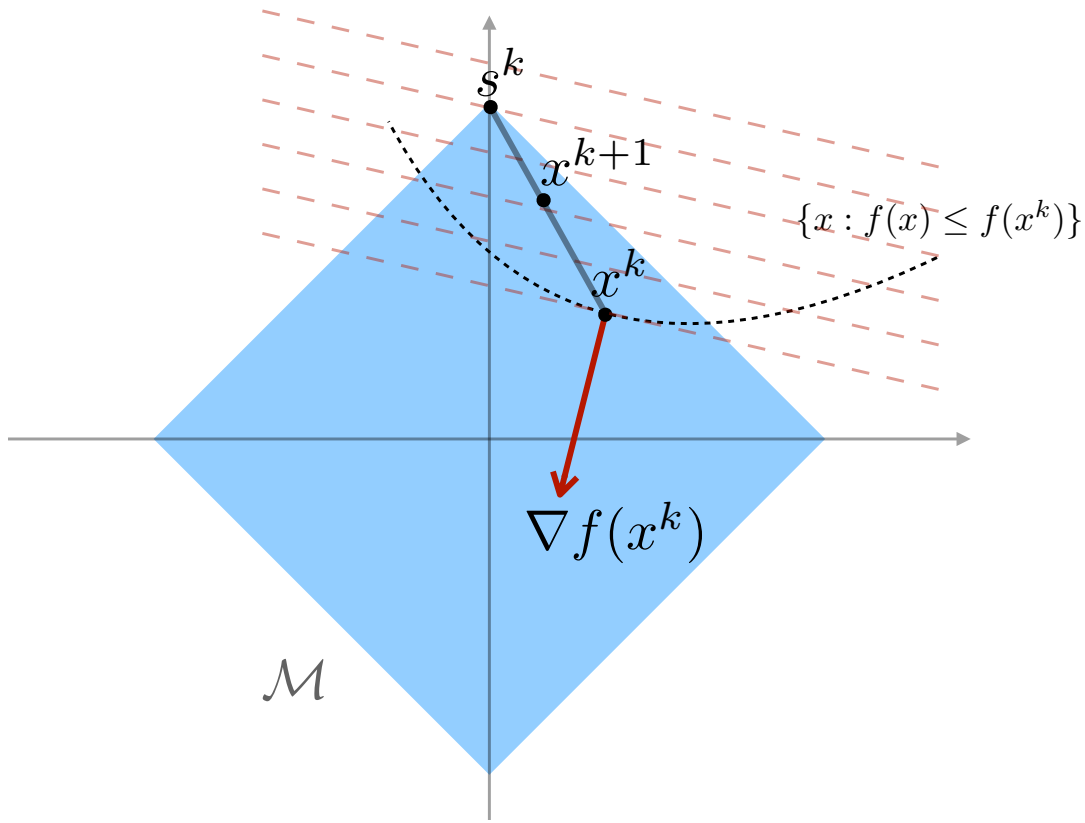
$$x^{k+1} = P_{\mathcal{M}}\left(x^k - \tfrac{1}{L}\nabla f(x^k)\right)$$
$$= \operatorname*{argmin}_{x \in \mathcal{M}} f(x^k) + \langle \nabla f(x^k),\, x - x^k \rangle + \tfrac{L}{2}\|x - x^k\|^2$$

Projected gradient method



$$x^{k+1} = P_{\mathcal{M}}\left(x^k - \tfrac{1}{L}\nabla f(x^k)\right)$$
$$= \operatorname*{argmin}_{x \in \mathcal{M}} f(x^k) + \langle \nabla f(x^k),\, x - x^k \rangle + \tfrac{L}{2}\|x - x^k\|^2$$

Frank-Wolfe aka Conditional Gradient method



$$s^k \in \operatorname*{argmin}_{x \in \mathcal{M}} f(x^k) + \langle \nabla f(x^k),\, x - x^k \rangle$$

$$x^{k+1} = x^k + \eta(s^k - x^k), \qquad (\eta \in [0, 1])$$

Frank-Wolfe method

$$\min_{x \in \mathcal{M}} \quad f(x)$$

1. Start with some guess $x^0 \in \mathcal{M}$
2. Form linear approximation of f at x^k:

$$\phi_f(x, x^k) := f(x^k) + \langle \nabla f(x^k),\, x - x^k \rangle$$

3. Find $s^k \in \operatorname{argmin}_{x \in \mathcal{M}} \phi_f(x, x^k)$
4. Take convex combination of x^k and s^k

$$x^{k+1} = x^k + \eta_k (s^k - x^k), \quad \eta_k \in [0, 1]$$

5. Repeat above procedure until $f(x^k) \leq f(x^*) + \varepsilon$.

Example: ℓ_1-norm regularization

$$\min_{\|x\|_1 \leq \gamma} \quad f(x).$$

Procedure for the linear minimization step:

1. Choose $i_k \in \operatorname{argmax}_{1 \leq i \leq d} |\nabla_i f(x^k)|$
2. Set $s_i^k = -\gamma \operatorname{sign}(\nabla_i f(x^k))$ for $i = i_k$ and 0 otherwise.

Example: ℓ_1-norm regularization

$$\min_{\|x\|_1 \leq \gamma} \quad f(x).$$

Procedure for the linear minimization step:

1. Choose $i_k \in \operatorname{argmax}_{1 \leq i \leq d} |\nabla_i f(x^k)|$
2. Set $s_i^k = -\gamma \operatorname{sign}(\nabla_i f(x^k))$ for $i = i_k$ and 0 otherwise.

- $\mathcal{O}(d)$ runtime.
- Coordinate / greedy updates.

Example: ℓ_p-norm regularization

$$\min_{\|x\|_p \leq \gamma} \quad f(x)$$

Procedure for the linear minimization step ($p \in (1, \infty)$):

Denote $\frac{1}{p} + \frac{1}{q} = 1$, then

1. Set $s_i^k = -\text{sign}(\nabla_i f(x^k)) \, |\nabla_i f(x^k)|^{q-1} \quad \text{for } i = 1, 2, \ldots, d.$
2. Rescale s^k so that $\|s^k\|_p = \gamma$.

Example: ℓ_p-norm regularization

$$\min_{\|x\|_p \leq \gamma} \quad f(x)$$

Procedure for the linear minimization step ($p \in (1, \infty)$):

Denote $\frac{1}{p} + \frac{1}{q} = 1$, then

1. Set $s_i^k = -\text{sign}(\nabla_i f(x^k)) \, |\nabla_i f(x^k)|^{q-1}$ for $i = 1, 2, \ldots, d$.
2. Rescale s^k so that $\|s^k\|_p = \gamma$.

- $\mathcal{O}(d)$ runtime.
- Projection in general is challenging (apart from $p = 1, 2, \infty$).

Example: Trace-norm regularization

$$\min_{\|X\|_{\text{tr}} \leq \gamma} \quad f(X)$$

Procedure for the linear minimization step:

1. Find top left and right singular vectors (u^k, v^k) of $\nabla f(X^k)$.
2. Set $S^k = -\gamma u^k v^{k^T}$.

Example: Trace-norm regularization

$$\min_{\|X\|_{\mathrm{tr}} \leq \gamma} \quad f(X)$$

Procedure for the linear minimization step:

1. Find top left and right singular vectors (u^k, v^k) of $\nabla f(X^k)$.
2. Set $S^k = -\gamma u^k v^{k^T}$.

- Rank-1 updates.
- Can be approximated efficiently using *power method*.
- Projection requires computing the SVD.

Example: Submodular polytope

Submodular function.

$g(A \cap B) + g(A \cup B) \leq g(A) + g(B)$ for all $A, B \subseteq [n]$

Example: Submodular polytope

Submodular function.
$g(A \cap B) + g(A \cup B) \leq g(A) + g(B)$ for all $A, B \subseteq [n]$
Submodular polyhedron. For $g(\emptyset) = 0$, the set

$$P_g := \left\{ x \in \mathbb{R}^n \mid \sum\nolimits_{i \in S} x_i \leq g(S),\ \forall S \subset [n] \right\}.$$

Example: Submodular polytope

Submodular function.
$g(A \cap B) + g(A \cup B) \leq g(A) + g(B)$ for all $A, B \subseteq [n]$
Submodular polyhedron. For $g(\emptyset) = 0$, the set

$$P_g := \left\{ x \in \mathbb{R}^n \mid \sum\nolimits_{i \in S} x_i \leq g(S),\ \forall S \subset [n] \right\}.$$

Linear minimization over P_g takes $O(n \log n)$ time using a greedy method. Projection is much harder.

Stepsize selection

- **Oblivious** Set $\eta_k = 2/(k+2)$, for $k \geq 0$

  (Not a descent method)

Stepsize selection

- **Oblivious** Set $\eta_k = 2/(k+2)$, for $k \geq 0$

  (Not a descent method)

- **Exact line-search** $\eta_k \in \operatorname*{argmin}_{\eta \in [0,1]} f(x^k + \eta(s^k - x^k))$

Stepsize selection

- **Oblivious** Set $\eta_k = 2/(k+2)$, for $k \geq 0$

  (Not a descent method)

- **Exact line-search** $\eta_k \in \operatorname*{argmin}_{\eta \in [0,1]} f(x^k + \eta(s^k - x^k))$

Exercise: $f(x) = \frac{1}{2}\|Ax - b\|^2$

$$\implies \eta_k = \text{clip}_{[0,1]}\left(\frac{\langle Ax^k - As^k,\, Ax^k - b\rangle}{\|Ax^k - As^k\|^2}\right)$$

Stepsize selection

- **Oblivious** Set $\eta_k = 2/(k+2)$, for $k \geq 0$

  (Not a descent method)

- **Exact line-search** $\eta_k \in \operatorname*{argmin}_{\eta \in [0,1]} f(x^k + \eta(s^k - x^k))$

  Exercise: $f(x) = \frac{1}{2}\|Ax - b\|^2$

$$\implies \eta_k = \text{clip}_{[0,1]}\left(\frac{\langle Ax^k - As^k,\, Ax^k - b\rangle}{\|Ax^k - As^k\|^2}\right)$$

- **Approx. line-search** Set $\eta_k = \text{clip}_{[0,1]}\left(\frac{\langle \nabla f(x^k),\, x^k - s^k\rangle}{LR^2}\right)$

Stepsize selection

- **Oblivious** Set $\eta_k = 2/(k+2)$, for $k \geq 0$

  (Not a descent method)

- **Exact line-search** $\eta_k \in \operatorname*{argmin}_{\eta \in [0,1]} f(x^k + \eta(s^k - x^k))$

  Exercise: $f(x) = \frac{1}{2}\|Ax - b\|^2$

$$\implies \eta_k = \operatorname{clip}_{[0,1]}\left(\frac{\langle Ax^k - As^k,\, Ax^k - b\rangle}{\|Ax^k - As^k\|^2}\right)$$

- **Approx. line-search** Set $\eta_k = \operatorname{clip}_{[0,1]}\left(\frac{\langle \nabla f(x^k),\, x^k - s^k\rangle}{LR^2}\right)$

- **Fully corrective*** Solve $x^{k+1} \in \operatorname*{argmin}_{x \in \operatorname{conv}\{s^0, s^1, \ldots, s^k\}} f(x)$

Convergence analysis

Theorem. Let $f \in C_L^1$ be cvx; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}$$

Convergence analysis

Theorem. Let $f \in C_L^1$ be cvx; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}$$

Proof. Let $\eta_k = 2/(k+2)$. Recall, $x^{k+1} = x^k + \eta_k(s^k - x^k)$.

Convergence analysis

Theorem. Let $f \in C_L^1$ be cvx; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}$$

Proof. Let $\eta_k = 2/(k+2)$. Recall, $x^{k+1} = x^k + \eta_k(s^k - x^k)$.

$$f(x^{k+1}) - f^*$$

Convergence analysis

Theorem. Let $f \in C_L^1$ be cvx; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}$$

Proof. Let $\eta_k = 2/(k+2)$. Recall, $x^{k+1} = x^k + \eta_k(s^k - x^k)$.

$$\begin{aligned} & f(x^{k+1}) - f^* \\ \leq \; & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L \|s^k - x^k\|^2 \end{aligned}$$

Convergence analysis

Theorem. Let $f \in C_L^1$ be cvx; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}$$

Proof. Let $\eta_k = 2/(k+2)$. Recall, $x^{k+1} = x^k + \eta_k(s^k - x^k)$.

$$\begin{aligned} & f(x^{k+1}) - f^* \\ \leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L \|s^k - x^k\|^2 \\ \leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L R^2 \end{aligned}$$

Convergence analysis

Theorem. Let $f \in C_L^1$ be cvx; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}$$

Proof. Let $\eta_k = 2/(k+2)$. Recall, $x^{k+1} = x^k + \eta_k(s^k - x^k)$.

$$\begin{aligned}
& f(x^{k+1}) - f^* \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L \|s^k - x^k\|^2 \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L R^2 \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, x^* - x^k \rangle + \tfrac{1}{2}\eta_k^2 L R^2
\end{aligned}$$

Convergence analysis

Theorem. Let $f \in C_L^1$ be cvx; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}$$

Proof. Let $\eta_k = 2/(k+2)$. Recall, $x^{k+1} = x^k + \eta_k(s^k - x^k)$.

$$\begin{aligned}
& f(x^{k+1}) - f^* \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L \|s^k - x^k\|^2 \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L R^2 \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, x^* - x^k \rangle + \tfrac{1}{2}\eta_k^2 L R^2 \\
\leq\ & f(x^k) - f^* - \eta_k (f(x^k) - f(x^*)) + \tfrac{1}{2}\eta_k^2 L R^2
\end{aligned}$$

Convergence analysis

Theorem. Let $f \in C_L^1$ be cvx; let $R = \max_{x,y \in \mathcal{M}} \|x - y\|$. Then,

$$f(x^k) - f^* \leq \frac{2LR^2}{k+1}$$

Proof. Let $\eta_k = 2/(k+2)$. Recall, $x^{k+1} = x^k + \eta_k(s^k - x^k)$.

$$\begin{aligned}
& f(x^{k+1}) - f^* \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L \|s^k - x^k\|^2 \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, s^k - x^k \rangle + \tfrac{1}{2}\eta_k^2 L R^2 \\
\leq\ & f(x^k) - f^* + \eta_k \langle \nabla f(x^k),\, x^* - x^k \rangle + \tfrac{1}{2}\eta_k^2 L R^2 \\
\leq\ & f(x^k) - f^* - \eta_k (f(x^k) - f(x^*)) + \tfrac{1}{2}\eta_k^2 L R^2 \\
=\ & (1 - \eta_k)(f(x^k) - f(x^*)) + \tfrac{1}{2}\eta_k^2 L R^2.
\end{aligned}$$

Verify: inductively, this leads to $f(x^k) - f^* \leq \frac{2LR^2}{k+1}$.

Invariance under affine transforms

$$\min_{x \in \mathcal{M}} \quad f(x)$$

Invariance under affine transforms

$$\min_{x \in \mathcal{M}} \quad f(x)$$

$$\min_{\hat{x} \in \hat{\mathcal{M}}} \quad \hat{f}(\hat{x})$$

- Re-parametrize the problem with a nonsingular matrix A

$$A : \hat{\mathcal{M}} \to \mathcal{M}, \qquad \hat{f}(\hat{x}) = f(A\hat{x})$$

Invariance under affine transforms

$$\min_{x \in \mathcal{M}} \quad f(x)$$

$$\min_{\hat{x} \in \hat{\mathcal{M}}} \quad \hat{f}(\hat{x})$$

- Re-parametrize the problem with a nonsingular matrix A

$$A : \hat{\mathcal{M}} \to \mathcal{M}, \qquad \hat{f}(\hat{x}) = f(A\hat{x})$$

- These two problems look completely same to Frank-Wolfe:

$$\operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle \nabla \hat{f}(\hat{x}^k),\, \hat{x} \rangle \equiv \operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle A^T \nabla f(A\hat{x}^k),\, \hat{x} \rangle$$

Invariance under affine transforms

$$\min_{x \in \mathcal{M}} \quad f(x)$$

$$\min_{\hat{x} \in \hat{\mathcal{M}}} \quad \hat{f}(\hat{x})$$

- Re-parametrize the problem with a nonsingular matrix A

$$A : \hat{\mathcal{M}} \to \mathcal{M}, \qquad \hat{f}(\hat{x}) = f(A\hat{x})$$

- These two problems look completely same to Frank-Wolfe:

$$\begin{aligned} \operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle \nabla \hat{f}(\hat{x}^k),\, \hat{x} \rangle &\equiv \operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle A^T \nabla f(A\hat{x}^k),\, \hat{x} \rangle \\ &\equiv \operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle \nabla f(A\hat{x}^k),\, A\hat{x} \rangle \end{aligned}$$

Invariance under affine transforms

$$\min_{x \in \mathcal{M}} \quad f(x)$$

$$\min_{\hat{x} \in \hat{\mathcal{M}}} \quad \hat{f}(\hat{x})$$

- Re-parametrize the problem with a nonsingular matrix A

$$A : \hat{\mathcal{M}} \to \mathcal{M}, \qquad \hat{f}(\hat{x}) = f(A\hat{x})$$

- These two problems look completely same to Frank-Wolfe:

$$\begin{aligned}
\operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle \nabla \hat{f}(\hat{x}^k),\, \hat{x} \rangle &\equiv \operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle A^T \nabla f(A\hat{x}^k),\, \hat{x} \rangle \\
&\equiv \operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle \nabla f(A\hat{x}^k),\, A\hat{x} \rangle \\
&\equiv A \cdot \operatorname*{argmin}_{x \in A^{-1}\hat{\mathcal{M}}} \langle \nabla f(A\hat{x}^k),\, x \rangle
\end{aligned}$$

Invariance under affine transforms

$$\min_{x \in \mathcal{M}} \quad f(x)$$

$$\min_{\hat{x} \in \hat{\mathcal{M}}} \quad \hat{f}(\hat{x})$$

- Re-parametrize the problem with a nonsingular matrix A

$$A : \hat{\mathcal{M}} \to \mathcal{M}, \qquad \hat{f}(\hat{x}) = f(A\hat{x})$$

- These two problems look completely same to Frank-Wolfe:

$$\begin{aligned}
\operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle \nabla \hat{f}(\hat{x}^k),\, \hat{x} \rangle &\equiv \operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle A^T \nabla f(A\hat{x}^k),\, \hat{x} \rangle \\
&\equiv \operatorname*{argmin}_{\hat{x} \in \hat{\mathcal{M}}} \langle \nabla f(A\hat{x}^k),\, A\hat{x} \rangle \\
&\equiv A \cdot \operatorname*{argmin}_{x \in A^{-1}\hat{\mathcal{M}}} \langle \nabla f(A\hat{x}^k),\, x \rangle \\
&\equiv A \cdot \operatorname*{argmin}_{x \in \mathcal{M}} \langle \nabla f(A\hat{x}^k),\, x \rangle
\end{aligned}$$

Curvature constant

$$\mathcal{C}_f = \sup_{\substack{x,s \in \mathcal{M},\ \eta \in [0,1] \\ y = x + \eta(s-x)}} \frac{2}{\eta^2} \Big(f(y) - f(x) - \langle y - x,\ \nabla f(x) \rangle \Big)$$

Then, $f(x^k) - f^* \leq \frac{2\mathcal{C}_f}{k+1}$.

Curvature constant

$$\mathcal{C}_f = \sup_{\substack{x,s\in\mathcal{M},\ \eta\in[0,1] \\ y=x+\eta(s-x)}} \frac{2}{\eta^2}\Big(f(y) - f(x) - \langle y-x,\ \nabla f(x)\rangle\Big)$$

Then, $f(x^k) - f^* \leq \frac{2\mathcal{C}_f}{k+1}$.

Exercise: Let $f \in C^1_L$, then $\mathcal{C}_f \leq LR^2$.

Curvature constant

$$\mathcal{C}_f = \sup_{\substack{x,s\in\mathcal{M},\ \eta\in[0,1] \\ y=x+\eta(s-x)}} \frac{2}{\eta^2}\Big(f(y) - f(x) - \langle y-x,\ \nabla f(x)\rangle\Big)$$

Then, $f(x^k) - f^* \leq \frac{2\mathcal{C}_f}{k+1}$.

Exercise: Let $f \in C^1_L$, then $\mathcal{C}_f \leq LR^2$.

- $\mathcal{C}_f$ is often unknown in practice

Curvature constant

$$\mathcal{C}_f = \sup_{\substack{x,s\in\mathcal{M},\ \eta\in[0,1] \\ y=x+\eta(s-x)}} \frac{2}{\eta^2}\Big(f(y) - f(x) - \langle y-x,\ \nabla f(x)\rangle\Big)$$

Then, $f(x^k) - f^* \leq \frac{2\mathcal{C}_f}{k+1}$.

Exercise: Let $f \in \mathrm{C}^1_L$, then $\mathcal{C}_f \leq LR^2$.

- $\mathcal{C}_f$ is often unknown in practice
- $\mathcal{C}_f$ does not depend on the choice of norm.
- $\mathcal{C}_f$ is invariant under affine transformation.

Stopping criterion

- (Optimality condition) Recall from Lecture 4:

$$\langle \nabla f(x^*),\, x^* - x \rangle \leq 0, \quad \forall x \in \mathcal{M}.$$

- (Definition) Frank-Wolfe gap / directional derivative:

$$\mathcal{G}_{FW}(x^k) = \max_{x \in \mathcal{M}} \langle \nabla f(x^k),\, x^k - x \rangle$$

  - $\mathcal{G}_{FW}(x) \geq 0$ for all $x \in \mathcal{M}$,
  - $\mathcal{G}_{FW}(x) = 0$ iff $x = x^*$.

Stopping criterion

- (Optimality condition) Recall from Lecture 4:

$$\langle \nabla f(x^*),\, x^* - x \rangle \leq 0, \quad \forall x \in \mathcal{M}.$$

- (Definition) Frank-Wolfe gap / directional derivative:

$$\mathcal{G}_{FW}(x^k) = \max_{x \in \mathcal{M}} \langle \nabla f(x^k),\, x^k - x \rangle = \langle \nabla f(x^k),\, x^k - s^k \rangle$$

  - $\mathcal{G}_{FW}(x) \geq 0$ for all $x \in \mathcal{M}$,
  - $\mathcal{G}_{FW}(x) = 0$ iff $x = x^*$.

Stopping criterion

- (Optimality condition) Recall from Lecture 4:

$$\langle \nabla f(x^*),\, x^* - x \rangle \leq 0, \quad \forall x \in \mathcal{M}.$$

- (Definition) Frank-Wolfe gap / directional derivative:

$$\mathcal{G}_{FW}(x^k) = \max_{x \in \mathcal{M}} \langle \nabla f(x^k),\, x^k - x \rangle = \langle \nabla f(x^k),\, x^k - s^k \rangle$$

  - $\mathcal{G}_{FW}(x) \geq 0$ for all $x \in \mathcal{M}$,
  - $\mathcal{G}_{FW}(x) = 0$ iff $x = x^*$.

Exercise: If f is convex, then $f(x^k) - f^* \leq \mathcal{G}_{FW}(x^k)$.

Can we accelerate FW?

Theorem. (Jaggi 2013) There exists a convex smooth function f such that $f(x) - f^* \leq \varepsilon$ requires $\Omega(\min\{n, 1/\varepsilon\})$ linear minimization steps.

Exercise: $f(x) = \|x\|^2$ and $\mathcal{M} = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\}$.

Can we accelerate FW?

Theorem. (Jaggi 2013) There exists a convex smooth function f such that $f(x) - f^* \leq \varepsilon$ requires $\Omega(\min\{n, 1/\varepsilon\})$ linear minimization steps.

Exercise: $f(x) = \|x\|^2$ and $\mathcal{M} = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\}$.

Theorem. (Canon & Cullum 1968) There exists a convex smooth function f such that $f(x) - f^* \leq \varepsilon$ requires $\Omega(\varepsilon^{-\frac{1}{1+\delta}})$ linear minimization steps for all $\delta > 0$.

Can we accelerate FW?

Theorem. (Jaggi 2013) There exists a convex smooth function f such that $f(x) - f^* \leq \varepsilon$ requires $\Omega(\min\{n, 1/\varepsilon\})$ linear minimization steps.

Exercise: $f(x) = \|x\|^2$ and $\mathcal{M} = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\}$.

Theorem. (Canon & Cullum 1968) There exists a convex smooth function f such that $f(x) - f^* \leq \varepsilon$ requires $\Omega(\varepsilon^{-\frac{1}{1+\delta}})$ linear minimization steps for all $\delta > 0$.

- In general, there is *no acceleration* for FW.
- f is strongly convex in both examples.

Can we accelerate FW?

Theorem. (Jaggi 2013) There exists a convex smooth function f such that $f(x) - f^* \leq \varepsilon$ requires $\Omega(\min\{n, 1/\varepsilon\})$ linear minimization steps.

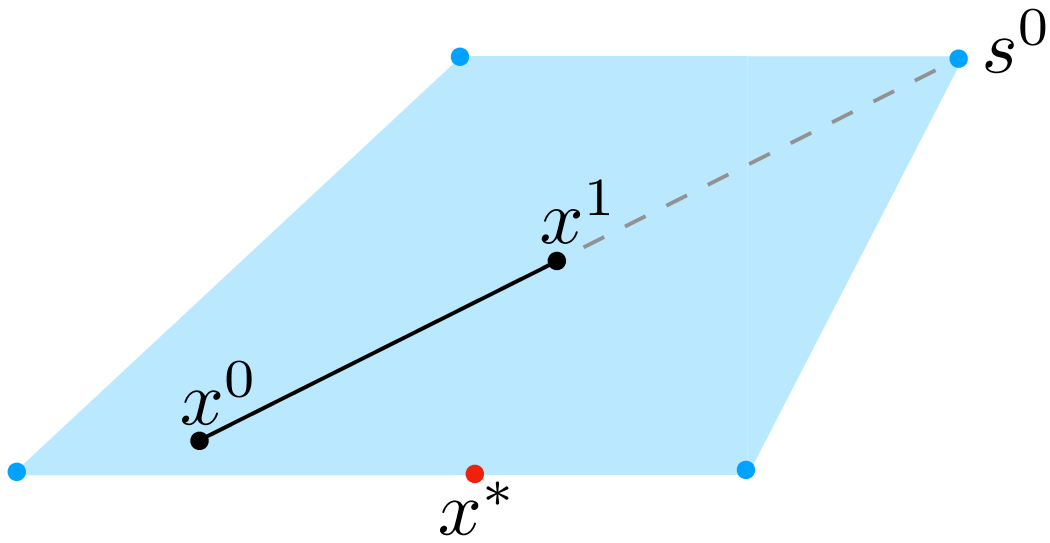
Exercise: $f(x) = \|x\|^2$ and $\mathcal{M} = \{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\}$.

Theorem. (Canon & Cullum 1968) There exists a convex smooth function f such that $f(x) - f^* \leq \varepsilon$ requires $\Omega(\varepsilon^{-\frac{1}{1+\delta}})$ linear minimization steps for all $\delta > 0$.

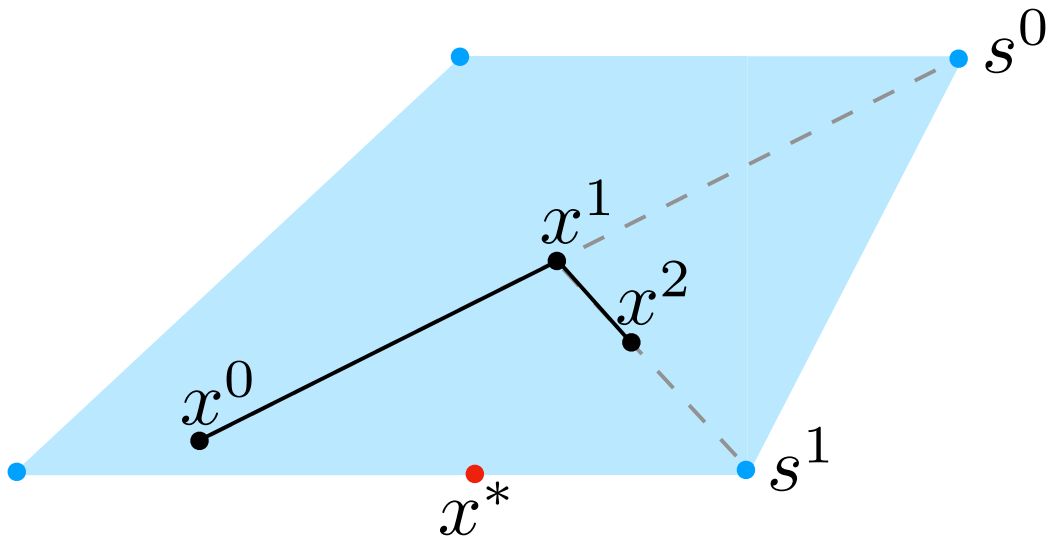
- In general, there is *no acceleration* for FW.
- f is strongly convex in both examples.
- Faster rates are known under stronger assumptions on $\mathcal{M}$.

Speeding up FW

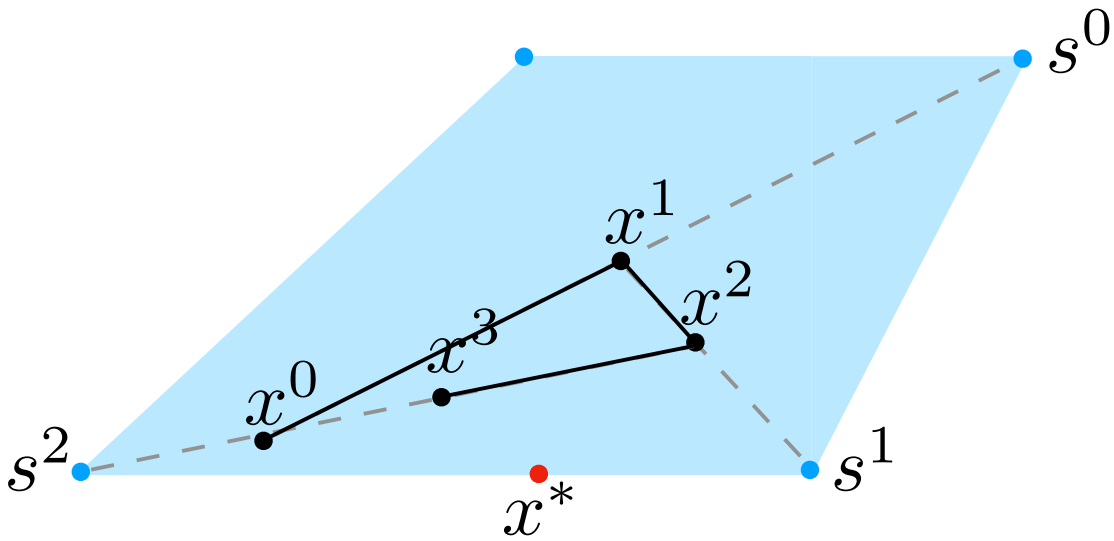
Zigzagging phenomenon



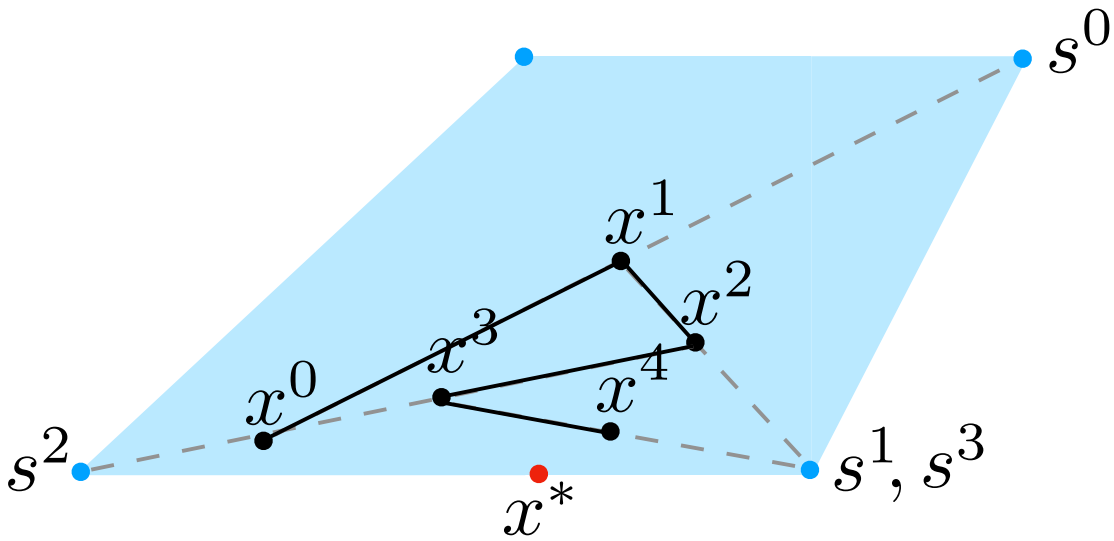
Zigzagging phenomenon



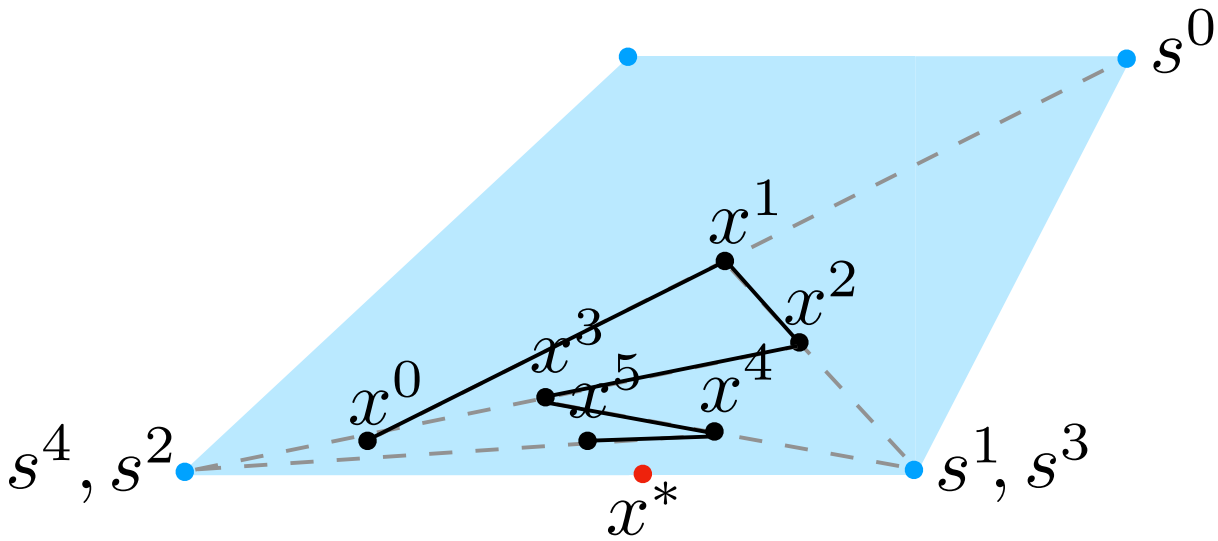
Zigzagging phenomenon



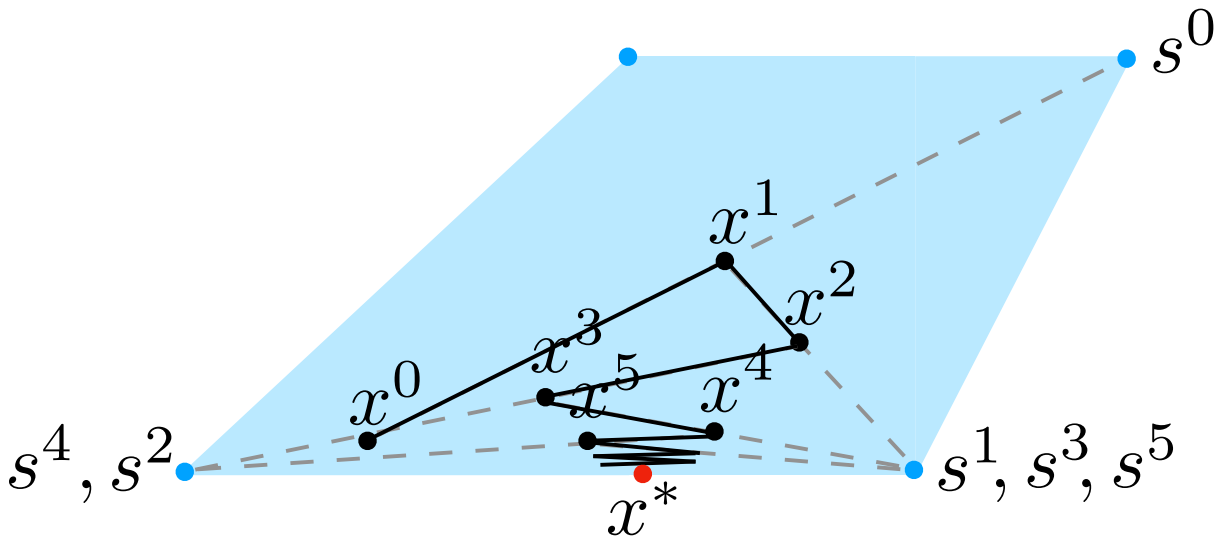
Zigzagging phenomenon



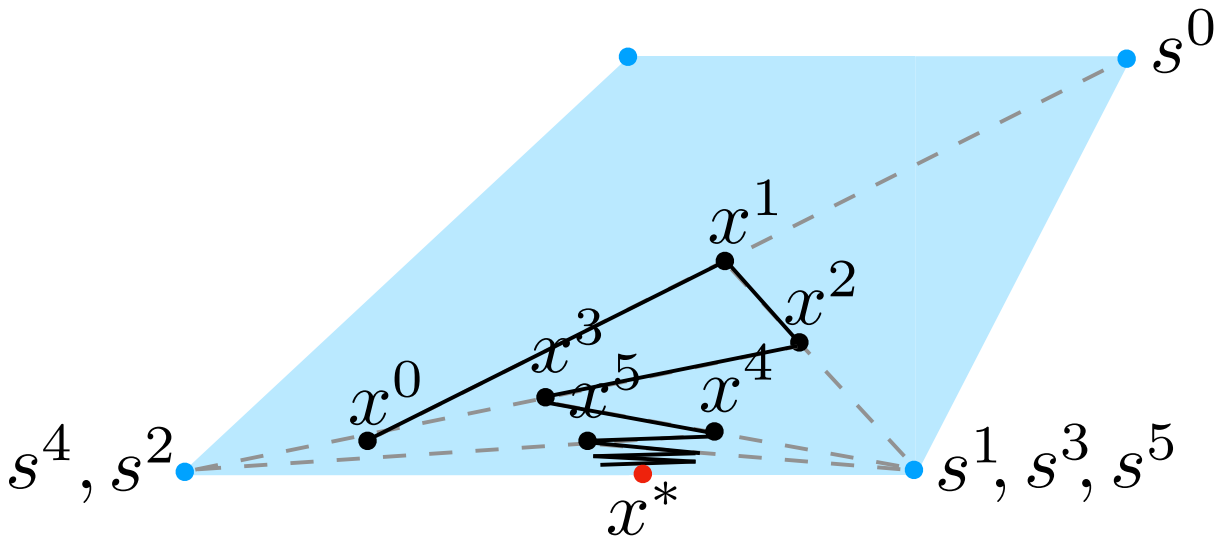
Zigzagging phenomenon



Zigzagging phenomenon



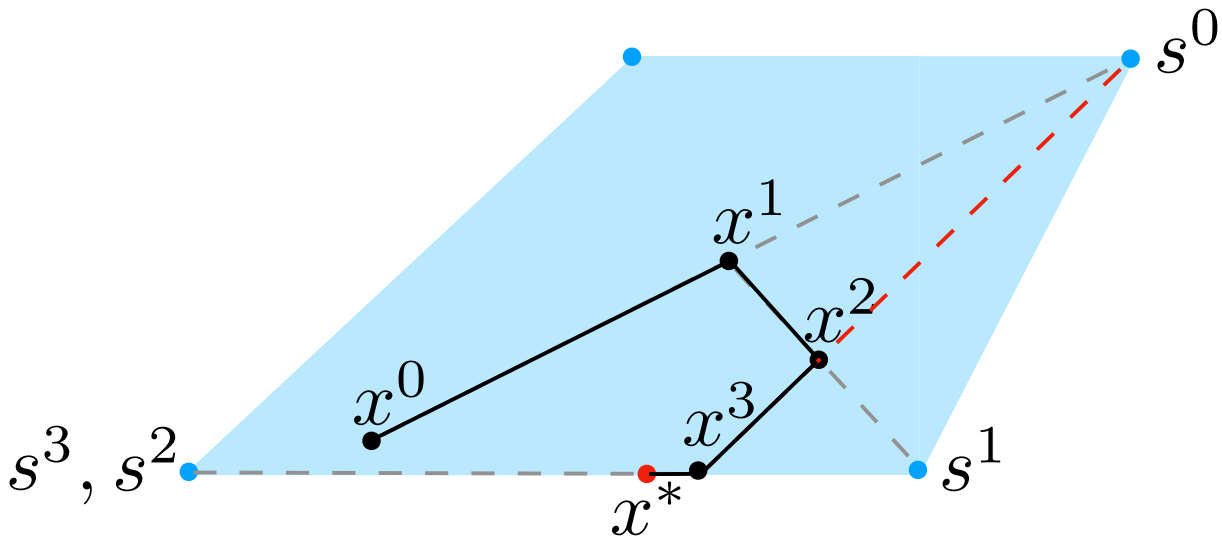
Zigzagging phenomenon



FW goes *towards* an extreme point at each iteration.

Idea: Incorporate steps that go *away* from an extreme point.

Key idea: away step



FW goes *towards* an extreme point at each iteration.

Idea: Incorporate steps that go *away* from an extreme point.

FW with away steps

Assumption: $\mathcal{M} = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is a *finite* set of vectors.

1. Let $x^0 \in \mathcal{A}$ and set the *active set* $\mathcal{S}^k = \{x^0\}$

FW with away steps

Assumption: $\mathcal{M} = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is a *finite* set of vectors.

1. Let $x^0 \in \mathcal{A}$ and set the *active set* $\mathcal{S}^k = \{x^0\}$
2. Find FW $s^k \in \text{argmin}_{x \in \mathcal{M}} \langle \nabla f(x^k), x \rangle$
3. Return if $\mathcal{G}_{FW}(x^k) = \langle \nabla f(x^k), x^k - s^k \rangle \leq \varepsilon$

FW with away steps

Assumption: $\mathcal{M} = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is a *finite* set of vectors.

1. Let $x^0 \in \mathcal{A}$ and set the *active set* $\mathcal{S}^k = \{x^0\}$
2. Find FW $s^k \in \text{argmin}_{x \in \mathcal{M}} \langle \nabla f(x^k), x \rangle$
3. Return if $\mathcal{G}_{FW}(x^k) = \langle \nabla f(x^k), x^k - s^k \rangle \leq \varepsilon$
4. Find away direction $v^k \in \text{argmax}_{x \in \mathcal{S}^k} \langle \nabla f(x^k), x \rangle$

FW with away steps

Assumption: $\mathcal{M} = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is a *finite* set of vectors.

1. Let $x^0 \in \mathcal{A}$ and set the *active set* $\mathcal{S}^k = \{x^0\}$
2. Find FW $s^k \in \text{argmin}_{x \in \mathcal{M}} \langle \nabla f(x^k), x \rangle$
3. Return if $\mathcal{G}_{FW}(x^k) = \langle \nabla f(x^k), x^k - s^k \rangle \leq \varepsilon$
4. Find away direction $v^k \in \text{argmax}_{x \in \mathcal{S}^k} \langle \nabla f(x^k), x \rangle$
5. If $\langle -\nabla f(x^k), s^k - x^k \rangle \geq \langle -\nabla f(x^k), x^k - v^k \rangle$, take a FW step

$$x^{k+1} = x^k + \eta_k (s^k - x^k)$$

$$\text{where } \eta_k \in \text{argmin}_{\eta \in [0,1]} f(x^k + \eta_k (s^k - x^k))$$

FW with away steps

Assumption: $\mathcal{M} = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is a *finite* set of vectors.

1. Let $x^0 \in \mathcal{A}$ and set the *active set* $\mathcal{S}^k = \{x^0\}$
2. Find FW $s^k \in \text{argmin}_{x \in \mathcal{M}} \langle \nabla f(x^k), x \rangle$
3. Return if $\mathcal{G}_{FW}(x^k) = \langle \nabla f(x^k), x^k - s^k \rangle \leq \varepsilon$
4. Find away direction $v^k \in \text{argmax}_{x \in \mathcal{S}^k} \langle \nabla f(x^k), x \rangle$
5. If $\langle -\nabla f(x^k), s^k - x^k \rangle \geq \langle -\nabla f(x^k), x^k - v^k \rangle$, take a FW step

$$x^{k+1} = x^k + \eta_k (s^k - x^k)$$

$$\text{where } \eta_k \in \text{argmin}_{\eta \in [0,1]} f(x^k + \eta_k (s^k - x^k))$$

6. Otherwise, take an away step

$$x^{k+1} = x^k + \eta_k (x^k - v^k)$$

$$\text{where } \eta_k \in \text{argmin}_{\eta \in [0, \eta_{\max}]} f(x^k + \eta_k (x^k - v^k))$$

FW with away steps

Assumption: $\mathcal{M} = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is a *finite* set of vectors.

1. Let $x^0 \in \mathcal{A}$ and set the *active set* $\mathcal{S}^k = \{x^0\}$
2. Find FW $s^k \in \text{argmin}_{x \in \mathcal{M}} \langle \nabla f(x^k), x \rangle$
3. Return if $\mathcal{G}_{FW}(x^k) = \langle \nabla f(x^k), x^k - s^k \rangle \leq \varepsilon$
4. Find away direction $v^k \in \text{argmax}_{x \in \mathcal{S}^k} \langle \nabla f(x^k), x \rangle$
5. If $\langle -\nabla f(x^k), s^k - x^k \rangle \geq \langle -\nabla f(x^k), x^k - v^k \rangle$, take a FW step

$$x^{k+1} = x^k + \eta_k (s^k - x^k)$$

$$\text{where } \eta_k \in \text{argmin}_{\eta \in [0,1]} f(x^k + \eta_k (s^k - x^k))$$

6. Otherwise, take an away step

$$x^{k+1} = x^k + \eta_k (x^k - v^k)$$

$$\text{where } \eta_k \in \text{argmin}_{\eta \in [0, \eta_{\max}]} f(x^k + \eta_k (x^k - v^k))$$

7. Update $\mathcal{S}^{k+1}$ and repeat the procedure.

FW with away steps - convergence

Assumptions:

- Let $f \in S^1_{L,\mu}$
- $\mathcal{M} = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is a *finite* set of vectors,
- (Diameter) $R = \max_{x,y \in \mathcal{M}} \|x - y\|$,
- (Facial dist.) $\Phi = \min_{F \in \text{faces}(\text{conv}(A))} \text{dist}(F, \text{conv}(A \backslash F))$,

Then,

$$f(x^k) - f^* \leq \left(1 - \frac{\mu}{L} \frac{\Phi^2}{4R^2}\right)^{k/2} (f(x^0) - f^*).$$

FW with away steps - convergence

Assumptions:

- Let $f \in S^1_{L,\mu}$
- $\mathcal{M} = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is a *finite* set of vectors,
- (Diameter) $R = \max_{x,y \in \mathcal{M}} \|x - y\|$,
- (Facial dist.) $\Phi = \min_{F \in \text{faces}(\text{conv}(A))} \text{dist}(F, \text{conv}(A \backslash F))$,

Then,

$$f(x^k) - f^* \leq \left(1 - \frac{\mu}{L} \frac{\Phi^2}{4R^2}\right)^{k/2} (f(x^0) - f^*).$$

$\frac{R^2}{\Phi^2}$ can be interpreted as the *condition number* of $\mathcal{M}$.

Does FW work with subgradients?

$$\min_{x_1^2+x_2^2 \le 1} \quad \max\{x_1, x_2\}$$

- Denote $f(x) = \max\{x_1, x_2\}$, and $\mathcal{M} = \{x \in \mathbb{R}^2 : \|x\| \le 1\}$.
- $x^* = [-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}]^T$.

Does FW work with subgradients?

$$\min_{x_1^2+x_2^2\leq 1} \quad \max\{x_1, x_2\}$$

- Denote $f(x) = \max\{x_1, x_2\}$, and $\mathcal{M} = \{x \in \mathbb{R}^2 : \|x\| \leq 1\}$.
- $x^* = [-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}]^T$.
- Let us replace the gradient with subgradient in FW

$$s^k \in \operatorname*{argmin}_{x\in\mathcal{M}} \langle g^k, x\rangle \text{ where } g^k \in \partial f(x^k)$$

$$x^{k+1} = x^k + \eta_k(s^k - x^k)$$

Does FW work with subgradients?

$$\min_{x_1^2+x_2^2\leq 1} \quad \max\{x_1, x_2\}$$

- Denote $f(x) = \max\{x_1, x_2\}$, and $\mathcal{M} = \{x \in \mathbb{R}^2 : \|x\| \leq 1\}$.
- $x^* = [-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}]^T$.
- Let us replace the gradient with subgradient in FW

$$s^k \in \operatorname*{argmin}_{x\in\mathcal{M}} \langle g^k, x\rangle \text{ where } g^k \in \partial f(x^k)$$

$$x^{k+1} = x^k + \eta_k(s^k - x^k)$$

- $g^k = \begin{cases} [1,0]^T & \text{if } x_1 \geq x_2 \\ [0,1]^T & \text{otherwise} \end{cases}$ and $s_k = -\frac{g^k}{\|g^k\|}$

Does FW work with subgradients?

$$\min_{x_1^2+x_2^2\leq 1} \quad \max\{x_1, x_2\}$$

- Denote $f(x) = \max\{x_1, x_2\}$, and $\mathcal{M} = \{x \in \mathbb{R}^2 : \|x\| \leq 1\}$.
- $x^* = [-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}]^T$.
- Let us replace the gradient with subgradient in FW

$$s^k \in \operatorname*{argmin}_{x\in\mathcal{M}} \langle g^k, x\rangle \text{ where } g^k \in \partial f(x^k)$$

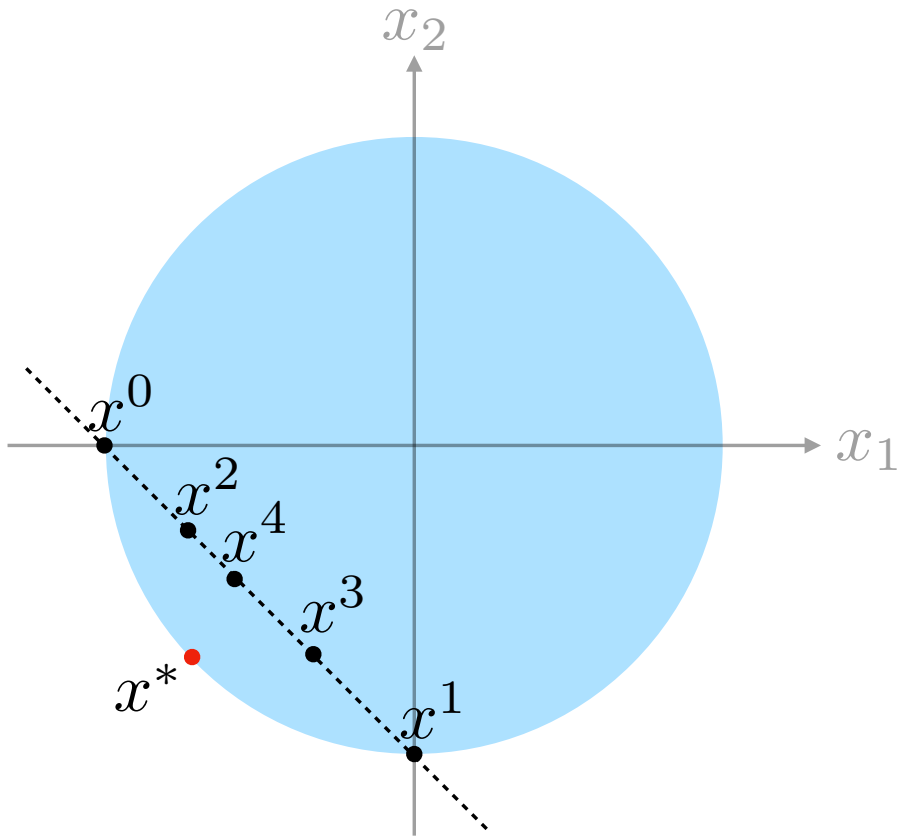
$$x^{k+1} = x^k + \eta_k(s^k - x^k)$$

- $g^k = \begin{cases} [1,0]^T & \text{if } x_1 \geq x_2 \\ [0,1]^T & \text{otherwise} \end{cases}$ and $s_k = -\frac{g^k}{\|g^k\|}$

$$\implies x^k \in \text{conv}(\{x^0, [-1,0]^T, [0,-1]^T\}) \not\ni x^* \quad \text{for all } k.$$

Does FW work with subgradients?

$$\min_{x_1^2+x_2^2\leq 1} \quad \max\{x_1, x_2\}$$



FW does not work for nonsmooth f.

(Some recent work (2020) extends FW to nonsmooth)

Nonconvex Frank-Wolfe

Assumptions.

- f is continuously differentiable, potentially nonconvex.
- $\mathcal{M}$ is convex and compact.

Then,

$$\min_{0 \le k \le K} \mathcal{G}_{FW}(x^k) \le \frac{\max\{2(f(x^0) - f^*), C_f\}}{\sqrt{K+1}}$$

Exercise: Supply a proof of the above rate.

Stochastic Frank-Wolfe

$$\min_{x \in \mathcal{M}} \quad \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Replacing ∇f with an unbiased estimator
does not directly work for FW.

Exercise. $\min_{x \in [-3,1]} \ \frac{1}{2}\big(f_1(x) + f_2(x)\big)$ where $\begin{cases} f_1(x) = \frac{1}{2}x^2 + 4x \\ f_2(x) = \frac{1}{2}x^2 - 4x \end{cases}$

Stochastic Frank-Wolfe

$$\min_{x \in \mathcal{M}} \quad \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Replacing ∇f with an unbiased estimator
does not directly work for FW.

Exercise. $\min_{x \in [-3,1]} \ \frac{1}{2}\big(f_1(x) + f_2(x)\big)$ where $\begin{cases} f_1(x) = \frac{1}{2}x^2 + 4x \\ f_2(x) = \frac{1}{2}x^2 - 4x \end{cases}$

So what should we do?

Stochastic Frank-Wolfe - analysis

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

$$\leq f(x^k) - f^* + \eta_k \underbrace{\langle \nabla f(x^k),\, s^k - x^k \rangle}_{\spadesuit} + \eta_k^2 \frac{LR^2}{2}$$

Stochastic Frank-Wolfe - analysis

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

$$\leq f(x^k) - f^* + \eta_k \underbrace{\langle \nabla f(x^k),\, s^k - x^k \rangle}_{\spadesuit} + \eta_k^2 \frac{LR^2}{2}$$

$$\spadesuit \quad = \quad \langle \nabla f_{i_k}(x^k),\, s^k - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle$$

Stochastic Frank-Wolfe - analysis

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

$$\leq f(x^k) - f^* + \eta_k \underbrace{\langle \nabla f(x^k),\, s^k - x^k \rangle}_{\spadesuit} + \eta_k^2 \frac{LR^2}{2}$$

$$\begin{aligned} \spadesuit &= \langle \nabla f_{i_k}(x^k),\, s^k - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\ &\leq \langle \nabla f_{i_k}(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \end{aligned}$$

Stochastic Frank-Wolfe - analysis

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

$$\leq f(x^k) - f^* + \eta_k \underbrace{\langle \nabla f(x^k),\, s^k - x^k \rangle}_{\spadesuit} + \eta_k^2 \frac{LR^2}{2}$$

$$\begin{aligned}
\spadesuit &= \langle \nabla f_{i_k}(x^k),\, s^k - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\
&\leq \langle \nabla f_{i_k}(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\
&= \langle \nabla f(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^* \rangle
\end{aligned}$$

Stochastic Frank-Wolfe - analysis

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

$$\leq f(x^k) - f^* + \eta_k \underbrace{\langle \nabla f(x^k),\, s^k - x^k \rangle}_{\spadesuit} + \eta_k^2 \frac{LR^2}{2}$$

$$\begin{aligned}
\spadesuit &= \langle \nabla f_{i_k}(x^k),\, s^k - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\
&\leq \langle \nabla f_{i_k}(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\
&= \langle \nabla f(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^* \rangle \\
&\leq f^* - f(x^k) + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^* \rangle
\end{aligned}$$

Stochastic Frank-Wolfe - analysis

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

$$\leq f(x^k) - f^* + \eta_k \underbrace{\langle \nabla f(x^k),\, s^k - x^k \rangle}_{\spadesuit} + \eta_k^2 \frac{LR^2}{2}$$

$$\begin{aligned}
\spadesuit &= \langle \nabla f_{i_k}(x^k),\, s^k - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\
&\leq \langle \nabla f_{i_k}(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\
&= \langle \nabla f(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^* \rangle \\
&\leq f^* - f(x^k) + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^* \rangle \\
&\leq f^* - f(x^k) + \underbrace{R\|\nabla f(x^k) - \nabla f_{i_k}(x^k)\|}_{\spadesuit\spadesuit}
\end{aligned}$$

Stochastic Frank-Wolfe - analysis

$$f(x^{k+1}) - f^* \leq f(x^k) - f^* + \langle \nabla f(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

$$\leq f(x^k) - f^* + \eta_k \underbrace{\langle \nabla f(x^k),\, s^k - x^k \rangle}_{\spadesuit} + \eta_k^2 \frac{LR^2}{2}$$

$$\begin{aligned}
\spadesuit &= \langle \nabla f_{i_k}(x^k),\, s^k - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\
&\leq \langle \nabla f_{i_k}(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^k \rangle \\
&= \langle \nabla f(x^k),\, x^* - x^k \rangle + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^* \rangle \\
&\leq f^* - f(x^k) + \langle \nabla f(x^k) - \nabla f_{i_k}(x^k),\, s^k - x^* \rangle \\
&\leq f^* - f(x^k) + \underbrace{R\|\nabla f(x^k) - \nabla f_{i_k}(x^k)\|}_{\spadesuit\spadesuit}
\end{aligned}$$

- When we take $\mathbb{E}$ of both sides, the remainder term is

$$\eta_k R\, \mathbb{E}[\|\nabla f(x^k) - \nabla f_{i_k}(x^k)\|]$$

Stochastic Frank-Wolfe - methods

Table from (Yurtsever, Sra, Cevher 2019):

	convex				non-convex			
	finite-sum		expectation		finite-sum		expectation	
	(*ifo*)	(*lmo*)	(*sfo*)	(*lmo*)	(*ifo*)	(*lmo*)	(*sfo*)	(*lmo*)
FW	$\mathcal{O}(n\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-1})$	-	-	$\mathcal{O}(n\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	-	-
CGS	$\mathcal{O}(n\epsilon^{-1/2})$	$\mathcal{O}(\epsilon^{-1})$	-	-	$\mathcal{O}(n\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	-	-
SFW	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$
SFW-1	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	-	-	-	-
Online-FW	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	-	-	-	-
SCGS	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-4})$	$\mathcal{O}(\epsilon^{-2})$
SVRF / SVFW	$\mathcal{O}(n\ln(\epsilon^{-1})+\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	-	-	$\mathcal{O}(n+n^{2/3}\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-10/3})$	$\mathcal{O}(\epsilon^{-2})$
STORC†	$\mathcal{O}(n\ln(\epsilon^{-1})+\epsilon^{-3/2})$	$\mathcal{O}(\epsilon^{-1})$	-	-	-	-	-	-
SPIDER-FW	$\mathcal{O}(n\ln(\epsilon^{-1})+\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(n^{1/2}\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$
SPIDER-CGS	$\mathcal{O}(n\ln(\epsilon^{-1})+\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(n^{1/2}\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$

Table 1: Comparison of conditional gradient methods for stochastic optimization. Contribution of *this work* is highlighted with blue font. See Section 6 for more details.

FW (Frank & Wolfe, 1956; Jaggi, 2013) , CGS (Lan & Zhou, 2016) , SFW (Hazan & Luo, 2016; Reddi et al., 2016) , SFW-1 (Mokhtari et al., 2018) , Online-FW (Hazan & Kale, 2012) , SCGS (Lan & Zhou, 2016) , SVRF / SVFW (Hazan & Luo, 2016; Reddi et al., 2016) , STORC (Hazan & Luo, 2016)