Lecture #14 The Frank-Wolfe method

EI 260

(Conditional gradient method)

- Algorithm
  - Geometric understanding
  - Examples
- Convergence analysis
  - L-smooth convex functions

The FW algorithm:

$$\boxed{\begin{aligned} &\text{minimize} && f(\underline{x}) \\ &\text{s.to} && \underline{x} \in C \end{aligned}}$$

- f is differentiable, L- smooth convex function
- $C \subseteq \text{dom}(f)$ is convex and closed set

FW has two steps:

local linear approximation

① Direction finding :

$$\overbrace{f(x) \simeq f(x_t) + \nabla f^T(x_t)(\underline{s} - \underline{x}_t)}^{\text{local linear approximation}}$$

Solves a linear optimization over a convex set

$$\underline{s}_t = \arg\min_{\underline{s} \in C} \nabla f^T(x_t)\, \underline{s}$$

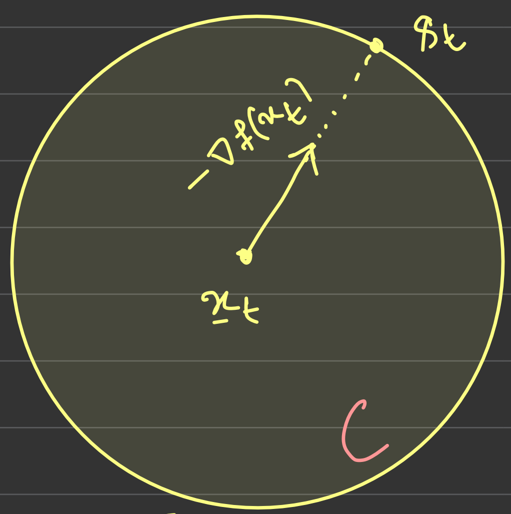
② update :

$$x_{t+1} = (1 - \gamma_t)\, \underline{x}_t + \gamma_t\, \underline{s}_t \in C$$

③ Step size : $\gamma_t = \frac{2}{t+1}$

Examples:

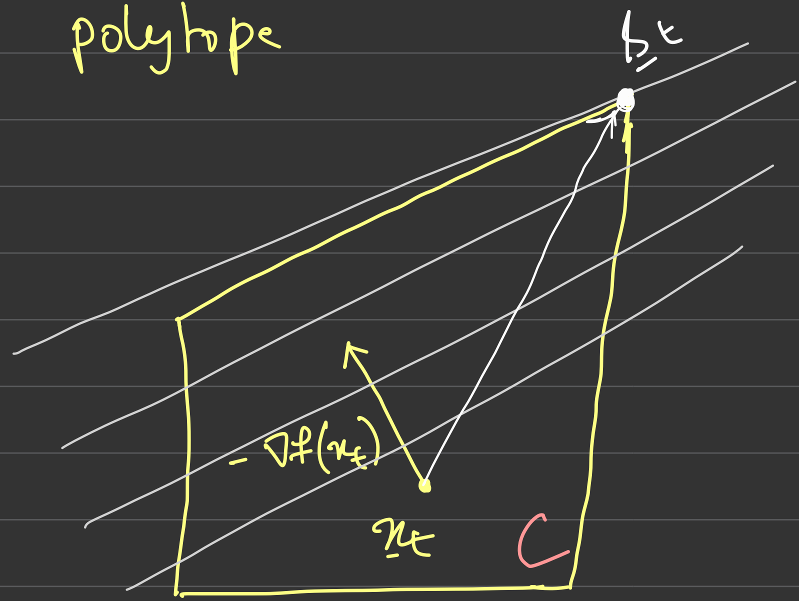
① C is a sphere centered at origin



$s_t \in \arg\min_{s \in C} \nabla f^T(x_t)\, s$

In this example, FW follows GD trajectory

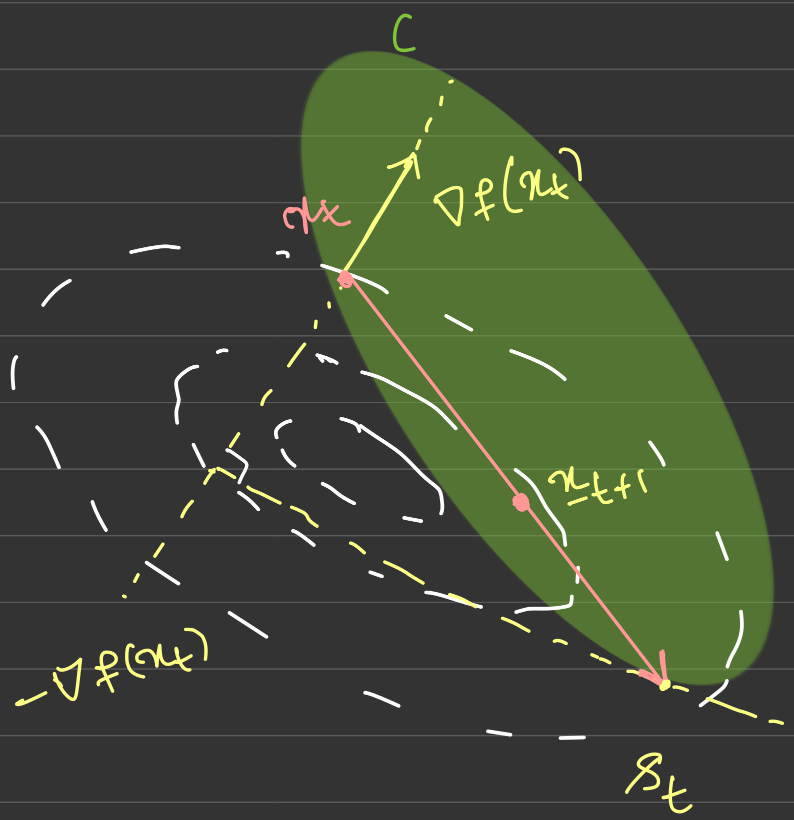
② polytope



Linear program

FW doesn't always follow -ve gradient

Another example:



Projected Gradient descent vs. FW

① Projection is replaced by linear optimization

② Both require gradient computation

Norm constraints:

$$C = \{ \underline{x} : \|\underline{x}\| \le k \}$$ for an arbitrary norm $\| \cdot \|$

$$s_t \in \underset{\|\underline{s}\| \le k}{\arg\min} \; \nabla f^T(\underline{x}_t)\, \underline{s}$$

$$= -k \underset{\|s\| \le t}{\arg\max} \; \nabla f^T(\underline{x}_t)\, \underline{s}$$

$$= \underbrace{-k\, \partial \| \nabla f(\underline{x}_t) \|_*}_{\text{Subgradient of the dual norm}}$$

$$\min \; f(x) \quad \text{s.t.} \quad \|\underline{x}\|_1 \le k$$

<u>l1 - norm:</u>

$$\underline{s}_t \in -k \, \partial \, \|\nabla f(\underline{x}_t)\|_\infty$$

$$i_k \in \underset{i=1,\ldots,d}{\arg\max} \; |\nabla_i f(\underline{x}_t)|$$

<u>FW update:</u>

$$\underline{x}_{t+1} = (1-\gamma_t)\,\underline{x}_t - \gamma_t \cdot k \,\mathrm{sgn}\left(\nabla_{i_k} f(\underline{x}_t)\right) \underline{e}_{i_k}$$

- Greedy co-ordinate descent (simpler than projection onto l_1 - ball)

Can, be applied for non-convex problems:

$$\underset{\underline{x}}{\text{minimize}} \quad -\underline{x}^T Q \underline{x} \quad \text{s.t.} \quad \|\underline{x}\|_2 \leq 1$$

with $Q \succ 0$

$$\underline{s}_t \in -\partial \|\nabla f(\underline{x}_t)\|_2$$

$$= -\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|_2} = \frac{Q \underline{x}_t}{\|Q \underline{x}_t\|_2}$$

$$\Rightarrow \underline{x}_{t+1} = (1-\gamma_t)\underline{x}_t + \gamma_t \frac{Q \underline{x}_t}{\|Q \underline{x}_t\|_2}$$

set $\gamma_t = 1 \quad \left[= \underset{0 \leq \gamma \leq 1}{\arg\min} \; f\left((1-\gamma)x_t + \gamma \frac{Q x_t}{\|Q x_t\|_2}\right)\right]$

$$\Rightarrow \underline{x}_{t+1} = \frac{Q \underline{x}_t}{\|Q \underline{x}_t\|_2}$$

: power method to find the leading eigen vector of Q

Convergence result:

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be convex and L-smooth and

$D = \mathrm{diameter}(C) = \sup_{\underline{x}, \underline{y} \in C} \| \underline{x} - \underline{y} \|$. With $\gamma_t = \frac{2}{t+1}$,

FW satisfies

$$f(x_T) - f(x^*) \leq \frac{2 L D^2}{T+1}$$

Sublinear convergence : $O(1/T)$ [same as projected gradient descent]

ε-accuracy : $O(1/\varepsilon)$

Proof: $f(\underline{y}) \leq f(\underline{x}) + \nabla f^T(\underline{x})(\underline{y} - \underline{x}) + \frac{L}{2} \|\underline{x} - \underline{y}\|^2$

$y = x_{t+1}$; $x = x_t$

$$f(\underline{x}_{t+1}) - f(\underline{x}_t) \leq \nabla f^T(x_t)(\underline{x}_{t+1} - \underline{x}_t) + \frac{L}{2} \|\underline{x}_{t+1} - x_t\|^2$$

$$\underline{x}_{t+1} = x_t + \gamma_t (s_t - x_t)$$

$$f(x_{t+1}) - f(\underline{x}_t) \leq \gamma_t \underbrace{\nabla f^T(x_t)(\underline{s}_t - \underline{x}_t)}_{\leq \nabla f^T(x_t)\underline{x}^*} + \frac{L}{2}\gamma_t^2 \underbrace{\|\underline{s}_t - x_t\|^2}_{\leq \frac{L}{2}\gamma_t^2 D^2}$$

$$s_t \in \arg\min_{s \in C} \nabla f^T(x_t) s$$

$$f(x_{t+1}) - f(\underline{x}_t) \leq \gamma_t \underbrace{\nabla f^T(x_t)[\underline{x}^* - x_t]}_{\text{convexity}} + \frac{L}{2}\gamma_t^2 D^2$$

$$\leq \gamma_t [f(\underline{x}^*) - f(\underline{x}_t)] + \frac{L}{2}\gamma_t^2 D^2$$

$$f(x_{t+1}) - f(\underline{x}^*) \leq (1 - \gamma_t)[f(x_t) - f(\underline{x}^*)] + \frac{L}{2}\gamma_t^2 D^2$$

$$\Delta_{t+1} \le (1-\gamma_t)\,\Delta_t + \frac{L}{2}\gamma_t^2 D^2$$

Our claim:

$$\Delta_t \le \frac{2LD^2}{T+1} \qquad \gamma_t = \frac{2}{t+1}$$

Proof by induction:

① Base case $t = 1$: $\Delta_2 \le 0 + \frac{L}{2}D^2 \le \frac{2}{3}LD^2 \quad [\gamma_1 = 1]$

② Inductive hypothesis: assume the upper bound is true for all T

③ Need to show it holds for t

①
$$\Delta_{t+1} \le (1-\gamma_t)\,\Delta_t + \frac{L}{2}\gamma_t^2 D^2$$

$$\le \left(1 - \frac{2}{t+1}\right)\frac{2LD^2}{t+1} + \frac{L}{2}\cdot\frac{4}{(t+1)^2}D^2$$

$$= \frac{t-1}{t+1}\cdot\frac{2LD^2}{t+1} + \frac{2LD^2}{(t+1)^2}$$

$$= \frac{2LD^2}{(t+1)^2}\left[t-1+1\right]$$

$$= \frac{2LD^2}{(t+1)} \cdot \frac{t}{(t+1)} \leq \frac{2LD^2}{(t+1)}$$

∎

- FW updates are affine invariant :

Suppose $\underline{x} = A\underline{x}'$ and $F(\underline{x}') = f(A x')$

for non singular A. Then

$$\underline{s}' = \underset{z \in A^{-1}C}{\arg\min} \; \nabla F^T(x')\, z$$

$$(x')^+ = (1-\gamma)\,\underline{x}^+ + \gamma\, s'$$

multiplying by A produces same updates as that from f.

- In general, stronge convexity does not improve convergence of FW.

  - Additional conditions on the constraint set.

  - α - strongly convex set yield linear convergence.

Duality gap: (not covered; only for reference)

Constrained problem: $\underset{\underline{x}}{\text{minimize}} \quad f(x) + \underbrace{\mathbb{I}_C(x)}_{\text{Indicator function}}$

$$I_C(x) = \begin{cases} 0, & \text{if } \underline{x} \in C \\ +\infty, & \text{o.w.} \end{cases}$$

Recall: $f^*(y) = \sup_{\underline{x}} \{ \underline{x}^T \underline{y} - f(\underline{x}) \}$

$$\underset{\underline{x}}{\text{minimize}} \ \underset{\underline{u}}{\text{maximize}} \left[\underline{x}^T \underline{u} - f^*(u) \right] + \mathbb{I}_C(x)$$

$$= \underset{\underline{u}}{\text{maximize}} \ \underset{\underline{x}}{\text{minimize}} \left[\mathbb{I}_C(x) + \underline{x}^T \underline{u} \right] - f^*(\underline{u})$$

$$\left.= \underset{\underline{u}}{\text{maximize}} \ - \mathbb{I}_C^*(-\underline{u}) - f^*(\underline{u}) \right\} \text{dual problem}$$

Duality gap between $\underline{x}$ and u:

$$f(x) + f^*(u) + I_C^*(-u) \geqslant \underline{x}^T \underline{u} + I_C^*(-u)$$

At $\underline{x} = \underline{x}_k$ and $\underline{u} = \nabla f(\underline{x}_k)$

$$\nabla f^T(\underline{x}_k)\, x_k + \max_{s \in C} - \nabla f^T(x_k)\, s = \nabla f^T(\underline{x}_k)(\underline{x}_k - s_k)$$

In fact,

$$f(\underline{x}_k) - f(\underline{x}^*) \geqslant \nabla f^T(\underline{x}_k)(\underline{x}_k - s_k)$$