Frank-Wolfe algorithm – iterative 1st order optimization algorithm for constrained convex optimization

M. Frank, P. Wolfe [1956]

a.k.a. Conditional gradients method,
reduced gradient method,
convex combination algorithm.

$$\min f(x) \quad \text{s.t.} \quad x \in D$$

D: compact convex set
$f: D \to \mathbb{R}$ convex differentiable real-valued function

Algorithm: Let $k \leftarrow 0$, and x_0 be any point in D

Step 1: (Direction finding subproblem): Find s_k solving

$$\min_{s \in D} \quad s^T \nabla f(x_k)$$

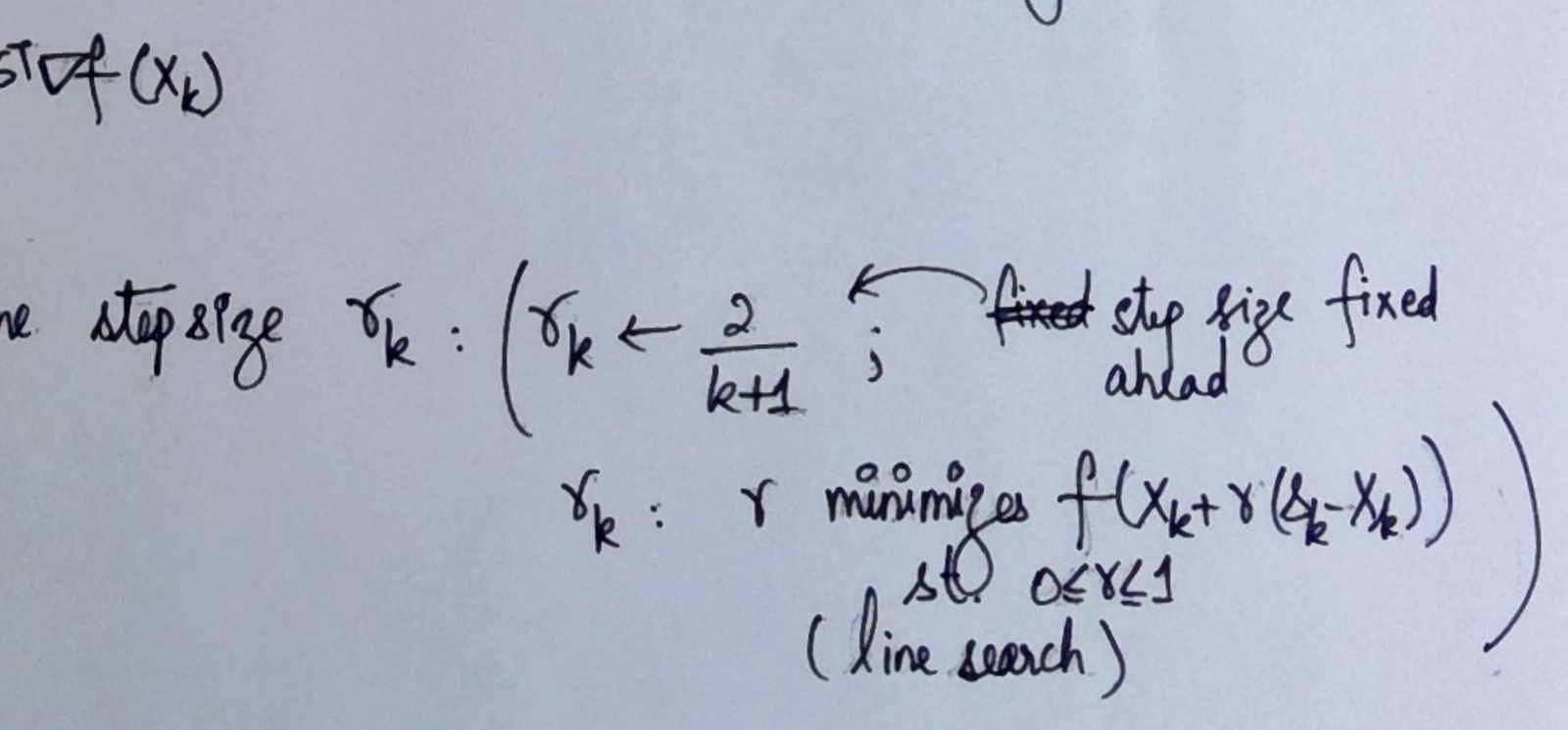
Step 2: Determine step size γ_k: ($\gamma_k \leftarrow \frac{2}{k+1}$; ~~fixed~~ step size fixed ahead

γ_k: γ minimizes $f(x_k + \gamma(s_k - x_k))$ s.t. $0 \le \gamma \le 1$ (line search))

Step 3: Update: $x_{k+1} \leftarrow x_k + \gamma_k(s_k - x_k)$; $k \leftarrow k+1$

Iterate.

What's happening?



- No projection step!
- Convergence is sublinear in general. Error is $O(1/k)$ after k iterations as long as the gradient is Lipschitz continuous.
- Iterates are sparse combinations of extremal points of the feasible set. When the constraints are linear, the subproblem becomes an LP! (linear program)

What happens when $\mathcal{C} = \{x : \|x\| \le t\}$ for a norm $\|\cdot\|$?

Then $s \in \underset{\|s\| \le t}{\operatorname{argmin}} \nabla f(x_{k-1})^T s$

$$= -t \left(\underset{\|s\| \le 1}{\operatorname{argmax}} \nabla f(x_{k-1})^T s \right)$$

$$= -t \, \partial \|\nabla f(x_{k-1})\|_* \qquad \|\cdot\|_* \text{ is the corresponding dual norm.}$$

Convergence Analysis

$$M = \max_{y, x, s \in \mathcal{C}, \; y = (1-\gamma)x + \gamma s} \frac{2}{\gamma^2} \left(f(y) - f(x) - \nabla f(x)^T (y - x) \right)$$

(Bounded curvature) $M < \infty$

Note: $\gamma \in [0, 1]$. What happens when $f(\cdot)$ is linear? ($M = 0$)

$f(y) - f(x) - \nabla f(x)^T (y - x)$ is called Bregman divergence defined by f.

Theorem: Conditional gradient method using fixed step sizes $\gamma_k = \frac{2}{(k+1)}$ $k = 1, 2, \cdots$ satisfies

$$f(x^k) - f(x^*) \le \frac{2M}{k+2}$$

What does this mean?

iteration to achieve ε error is $O(1/\varepsilon)$

Convergence rate matches that of projected gradient when ∇f is Lipschitz.
But are the assumptions comparable?

fact: If ∇f is Lipschitz with Lipschitz constant L, then $M \le \text{diam}^2(C) \times L$ where

$$\text{diam}(C) = \max_{x,s \in C} \|x - s\|_2$$

Recall:

$$f(y) - f(x) - \nabla f(x)^T (y-x) \le \frac{L}{2} \|y-x\|_2^2$$ ← L lipschitz condition (one of the several equivalent)

Max. over all $y = ((1-\gamma)x + \gamma s)$ and mul by $2/\gamma^2$

$$M \le \max_{x,b,y \in C,\ y=(1-\gamma)x+\gamma s} \left(\frac{2}{\gamma^2}\right) \cdot \frac{L}{2} \|y-x\|_2^2 = \max_{x,s \in C} L\|x-s\|_2^2$$

Essentially assuming bounded curvature is no stronger than assuming lipschitz.

Convergence

key inequality $f(x_k) \le f(x_{k-1}) - \gamma_k g(x_{k-1}) + \frac{\gamma_k^2 M}{2}$

$$g(x) = \max_{s \in C} \nabla f(x)^T (x-s)$$

$x^+ \leftarrow x_k$
$x \leftarrow x_{k-1}$

$$f(x^+) = f(x + \gamma(s-x))$$
$$\le f(x) + \gamma \nabla f(x)^T (s-x) + \frac{\gamma^2 M}{2}$$
$$= f(x) - \gamma g(x) + \frac{\gamma^2 M}{2}$$

Let $h(x) = f(x) - f(x^*)$

Claim $g(x) \ge h(x)$ always.

aside

$$f(y) \ge f(x) + \nabla f(x)^T (y-x)$$
$$f(y) - f(x) \ge \nabla f(x)^T (y-x)$$
$$f(x^*) - f(x) \ge \nabla f(x)^T (x^* - x)$$
$$f(x) - f(x^*) \le \nabla f(x)^T (x - x^*) \le \max_s \nabla f(x)^T (x-s)$$

$$h(x^+) \leq h(x) - \gamma g(x) + \frac{\gamma^2 M}{2}$$

$$h(x^+) \leq h(x)(1-\gamma) + \frac{\gamma^2 M}{2}$$

Proof is left as homework.

$$\Rightarrow h(x_k) \leq \frac{2M}{k+2}$$