

First order methods

FOR CONVEX OPTIMIZATION

Saketh (IIT Bombay)

Topics

- Part – I
 - *Optimal* methods for unconstrained convex programs
 - Smooth objective
 - Non-smooth objective
- Part – II
 - *Optimal* methods for constrained convex programs
 - Projection based
 - Frank-Wolfe based
 - Functional constraint based
 - Prox-based methods for structured non-smooth programs

Non-Topics

- Step-size schemes
- Bundle methods
- Stochastic methods
- Inexact oracles
- Non-Euclidean extensions (*Mirror-friends*)



Motivation

& EXAMPLE APPLICATIONS



Machine Learning Applications

- **Goal:** Construct $f : X \rightarrow Y$

Machine Learning Applications

- **Goal:** Construct $f : X \rightarrow Y$



Machine Learning Applications


- **Goal:** Construct $f : X \rightarrow Y$
- **Input data:** $\{ (x_1, y_1), \dots, (x_m, y_m) \}$

Machine Learning Applications

- **Goal:** Construct $f : X \rightarrow Y$
- **Input data:** $\{ (x_1, y_1), \dots, (x_m, y_m) \}$
- **Model:** $f(x) = w^T \phi(x)$

Machine Learning Applications

- **Goal:** Construct $f : X \rightarrow Y$
- **Input data:** $\{ (x_1, y_1), \dots, (x_m, y_m) \}$
- **Model:** $f(x) = w^T \phi(x)$
- **Algorithm:** Find simple functions that explain data


$$\min_{w \in \mathbb{R}^n} \Omega(w) + \sum_{i=1}^m l(w^T \phi(x_i), y_i)$$

Typical Program – Machine Learning

Smooth/Non-Smooth

$\min_{w \in \mathbb{R}^n}$

$\Omega(w)$

+

$$\sum_{i=1}^m l(w^T \phi(x_i), y_i)$$

Smooth/Non-Smooth

- Unconstrained
 , smooth functional constraints
- Smooth/Non-smooth/Composite Objectives

Typical Program – Machine Learning

Smooth/Non-Smooth

Smooth/Non-Smooth

Domain
e.g. simplex

$\min_{w \in W}$

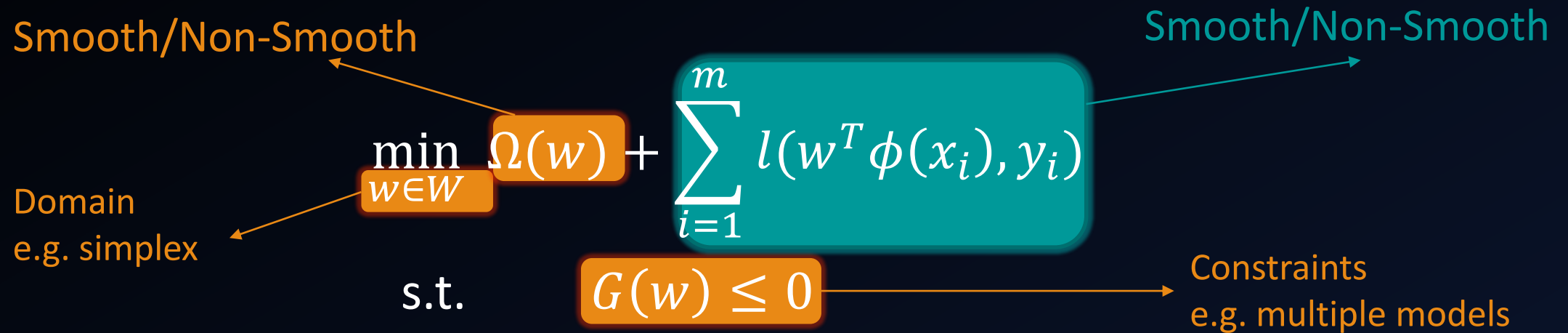
$\Omega(w)$

+

$$\sum_{i=1}^m l(w^T \phi(x_i), y_i)$$

- Unconstrained/Constrained
 - Simple domains, smooth functional constraints
- Smooth/Non-smooth/Composite Objectives

Typical Program – Machine Learning



- Unconstrained/Constrained
 - Simple domains, smooth functional constraints
- Smooth/Non-smooth/Composite Objectives

Scale is the issue!

- m, n as well as no. models may run into **millions!**
- Even a single iteration of IPM/Newton-variants is in-feasible.
- “Slower” but “cheaper” methods are the alternative
 - Decomposition based
 - **First order methods**

First Order Methods - Overview

- Iterative, gradient-like information, $O(n)$ per iteration 😊
- E.g. **Gradient method**, Cutting planes, Conjugate gradient
- Very old methods (1950s)
- Far slower than IPM:
 - Sub-linear rate 😞. (Not crucial for ML)
 - But (nearly) n -independent 😞
- Widely employed in state-of-the-art ML systems
- Choice of variant depends on problem structure

First Order Methods - Overview

- Iterative, gradient-like information, $O(n)$ per iteration 😊
- E.g. **Gradient method**, Cutting planes, Conjugate gradient
- Very old methods (1950s)
- Far slower than IPM:
 - Sub-linear rate 😞 . (Not crucial for ML)
 - But (nearly) n -independent 😊
- Widely employed in state-of-the-art ML systems
- Choice of variant depends on problem structure

Smooth un-constrained

$$\text{MIN}_{w \in \mathbb{R}^n} \sum_{i=1}^m (w^T \phi(x_i) - y_i)^2$$

Smooth Convex Functions

- Continuously differentiable
- Gradient is Lipschitz continuous

Smooth Convex Functions

- Continuously differentiable
- Gradient is Lipschitz continuous
 - $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
 - E.g. $g(x) \equiv x^2$ is not L-contr. over \mathbb{R} but is over $[0,1]$ with $L=2$
 - E.g. $g(x) \equiv |x|$ is L-contr. with $L=1$

Smooth Convex Functions

- Continuously differentiable
- Gradient is Lipschitz continuous
 - $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
 - E.g. $g(x) \equiv x^2$ is not L-contr. over R but is over $[0,1]$ with $L=2$
 - E.g. $g(x) \equiv |x|$ is L-contr. with $L=1$

Theorem: Let f be convex twice differentiable. Then

$$f \text{ is smooth with const. } L \Leftrightarrow f''(x) \preceq L I_n$$

Smooth Convex Functions

- Continuously differentiable
- Gradient is Lipschitz continuous
 - $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
 - E.g. $g(x) \equiv x^2$ is not L-contr. over R but is over $[0,1]$ with $L=2$
 - E.g. $g(x) \equiv |x|$ is L-contr. with $L=1$

$$\min_{w \in R^n} \sum_{i=1}^m (w^T \phi(x_i) - y_i)^2$$

is indeed smooth!

Theorem: Let f be convex twice differentiable. Then

$$f \text{ is smooth with const. } L \Leftrightarrow f''(x) \preceq L I_n$$

Gradient Method [Cauchy1847]

- Move iterate in direction of instantaneous decrease
 - $x_{k+1} = x_k - s_k \nabla f(x_k), s_k > 0$



Gradient Method

- Move iterate in direction of instantaneous decrease
 - $x_{k+1} = x_k - s_k \nabla f(x_k)$, $s_k > 0$
- Regularized minimization of first order approx.
 - $x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2$

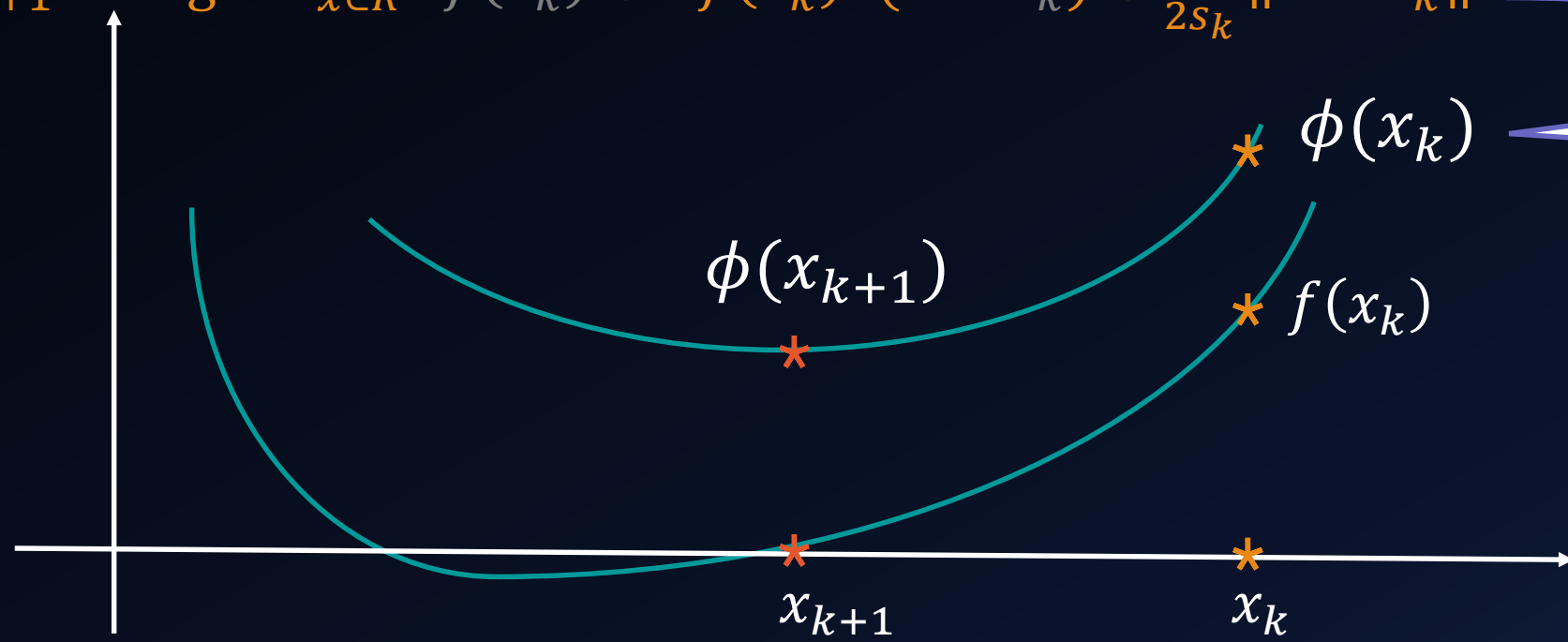
Gradient Method

- Move iterate in direction of instantaneous decrease

- $x_{k+1} = x_k - s_k \nabla f(x_k), s_k > 0$

- Regularized minimization of first order approx.

- $x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2$



Gradient Method

- Move iterate in direction of instantaneous decrease
 - $x_{k+1} = x_k - s_k \nabla f(x_k)$, $s_k > 0$
- Regularized minimization of first order approx.
 - $x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2$
- Various step-size schemes
 - Constant ($1/L$)
 - Diminishing ($s_k \downarrow 0, \sum s_k = \infty$)
 - Exact or back-tracking line search

Convergence rate – Gradient method

Theorem[Ne04]: If f is smooth with const. L and $s_k = \frac{1}{L}$, then gradient method generates x_k such that:

$$f(x_k) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{k + 4}.$$

Proof Sketch:

- $f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$
- $f(x_{k+1}) \geq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2L} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2$
- $\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{1}{L^2} \|\nabla f(x_k)\|^2$
- $\Delta_{k+1} \leq \Delta_k - \Delta_k^2 / r_0^2$ (Solve recursion)

Convergence rate – Gradient method

Theorem[Ne04]: If f is smooth with const. L and $s_k = \frac{1}{L}$, then gradient method generates x_k such that:

$$f(x_k) - f(x^*) \leq \frac{2L \|x_0 - x^*\|^2}{k + 4}.$$

Proof Sketch:

- $f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$
- $f(x_{k+1}) \geq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2L} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2$
- $\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{1}{L^2} \|\nabla f(x_k)\|^2$
- $\Delta_{k+1} \leq \Delta_k - \Delta_k^2 / \Delta_0^2$ (Solve recursion)

Majorization
minimization

Convergence rate – Gradient method

Theorem[Ne04]: If f is smooth with const. L and $s_k = \frac{1}{L}$, then gradient method generates x_k such that:

$$f(x_k) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{k + 4}.$$

Proof Sketch:

- $f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|^2$
- $f(x_{k+1}) \geq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2L} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2$
- $\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 - \frac{1}{L^2} \|\nabla f(x_k)\|^2$
- $\Delta_{k+1} \leq \Delta_k - \Delta_k^2 / r_0^2$ (Solve recursion)

Majorization
minimization

Comments on rate of convergence

- Sub-linear, very slow compared to IPM
- Applies to conjugate gradient and other traditional variants
- Sub-optimal (may be?):

Theorem[Ne04]: For any $k \leq \frac{n-1}{2}$, and any x_0 , there exists a smooth function f , with const. L , such that with any first order method, we have:

$$f(x_k) - f(x^*) \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}.$$

Proof Sketch: Choose function such that

$$x_k \in \text{lin}(\nabla f(x_0), \dots, \nabla f(x_{k-1})) \subset R^{k,n}$$

Comments on rate of convergence

- Sub-linear, very slow compared to IPM
- Applies to conjugate gradient and other traditional variants
- Sub-optimal (may be?):

Theorem[Ne04]: For any $k \leq \frac{n-1}{2}$, and any x_0 , there exists a smooth function f , with const. L , such that with **any first order method**, we have:

$$f(x_k) - f(x^*) \geq \frac{3L \|x_0 - x^*\|^2}{32(k+1)^2}.$$

Proof Sketch: Choose function such that

$$x_k \in \text{lin}(\nabla f(x_0), \dots, \nabla f(x_{k-1})) \subset R^{k,n}$$

Comments on rate of convergence

- Sub-linear, very slow compared to IPM
- Applies to conjugate gradient and other traditional variants
- Sub-optimal (may be?):

Theorem[Ne04]: For any $k \leq \frac{n-1}{2}$, and any x_0 , there exists a smooth function f , with const. L , such that with **any first order method**, we have:

$$f(x_k) - f(x^*) \geq \frac{3L \|x_0 - x^*\|^2}{32(k+1)^2}.$$

Proof Sketch: Choose function such that

$$x_k \in \text{lin}(\nabla f(x_0), \dots, \nabla f(x_{k-1})) \subset R^{k,n}$$

Comments on rate of convergence

Strongly convex: $O\left(\left(\frac{Q-1}{Q+1}\right)^{2k}\right)$

- Sub-linear, very slow compared to IPM
- Applies to conjugate gradient and other traditional variants
- Sub-optimal (may be?):

Strongly convex: $O\left(\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k}\right)$

Theorem[Ne04]: For any $k \leq \frac{n-1}{2}$, and any x_0 , there exists a smooth function f , with const. L , such that with any first order method, we have:

$$f(x_k) - f(x^*) \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}.$$

Proof Sketch: Choose function such that

$$x_k \in \text{lin}(\nabla f(x_0), \dots, \nabla f(x_{k-1})) \subset R^{k,n}$$

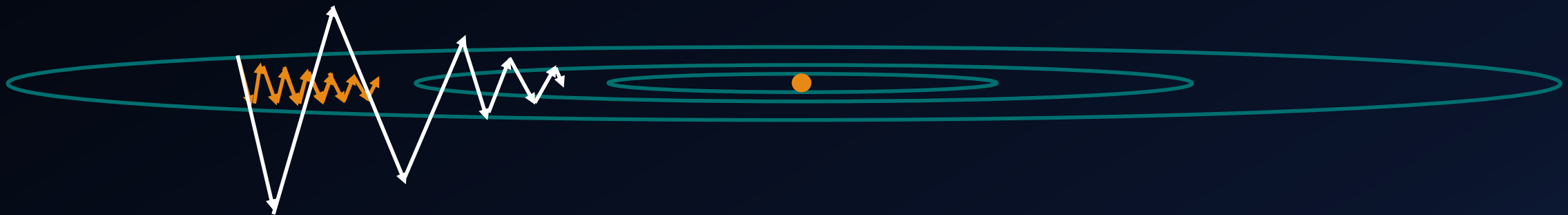
Intuition for non-optimality

- All variants are **descent** methods
- Descent essential for proof
- Overkill leading to **restrictive** movements
- Try non-descent alternatives!



Intuition for non-optimality

- All variants are **descent** methods
- Descent essential for proof
- Overkill leading to **restrictive** movements
- **Try non-descent alternatives!**



Towards optimality [Moritz Hardt]

Sub-optimal: $O\left(\left(1 - \frac{1}{Q}\right)^k\right)$

- $f(x) = \frac{1}{2}x^T Ax - bx ; x_0 = b$
- $x_k = x_{k-1} - \frac{1}{L}(Ax_{k-1} - b) = \sum_{i=0}^k \left(I - \frac{A}{L}\right)^i \frac{b}{L}$

Lemma[Mo12]: There is a (Chebyshev) poly. of degree $O\left(\sqrt{Q} \log^{1/\epsilon}\right)$ such that $p(0) = 1$ and $p(x) \leq \epsilon \forall x \in [\mu, L]$.

- Chebyshev poly. have two term recursive formula, hence we expect:
 - $x_k = x_{k-1} - s_{k-1}\nabla f(x_{k-1}) + \lambda_{k-1}\nabla f(x_{k-2})$, to be optimal (acceleration)

Towards optimality [Moritz Hardt]

- $f(x) = \frac{1}{2}x^T Ax - bx ; x_0 = b$
- $x_k = x_{k-1} - \frac{1}{L}(Ax_{k-1} - b) = \sum_{i=0}^k \left(I - \frac{A}{L}\right)^i \frac{b}{L}$

Sub-optimal: $O\left(\left(1 - \frac{1}{Q}\right)^k\right)$

Optimal: $O\left(\left(1 - \frac{1}{\sqrt{Q}}\right)^k\right)$

Lemma[Mo12]: There is a (Chebyshev) poly. of degree $O\left(\sqrt{Q} \log 1/\epsilon\right)$ such that $p(0) = 1$ and $p(x) \leq \epsilon \forall x \in [\mu, L]$.

- Chebyshev poly. have two term recursive formula, hence we expect:
 - $x_k = x_{k-1} - \lambda_{k-1} \nabla f(x_{k-1}) + \lambda_{k-1} \nabla f(x_{k-2})$, to be optimal (acceleration)

Towards optimality [Moritz Hardt]

- $f(x) = \frac{1}{2}x^T Ax - bx ; x_0 = b$

- $x_k = x_{k-1} - \frac{1}{L}(Ax_{k-1} - b) = \sum_{i=0}^k \left(I - \frac{A}{L}\right)^i \frac{b}{L}$

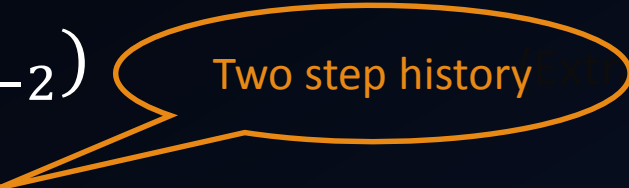
Sub-optimal: $O\left(\left(1 - \frac{1}{Q}\right)^k\right)$

Optimal: $O\left(\left(1 - \frac{1}{\sqrt{Q}}\right)^k\right)$

Lemma[Mo12]: There is a (Chebyshev) poly. of degree $O\left(\sqrt{Q} \log 1/\epsilon\right)$ such that $p(0) = 1$ and $p(x) \leq \epsilon \forall x \in [\mu, L]$.

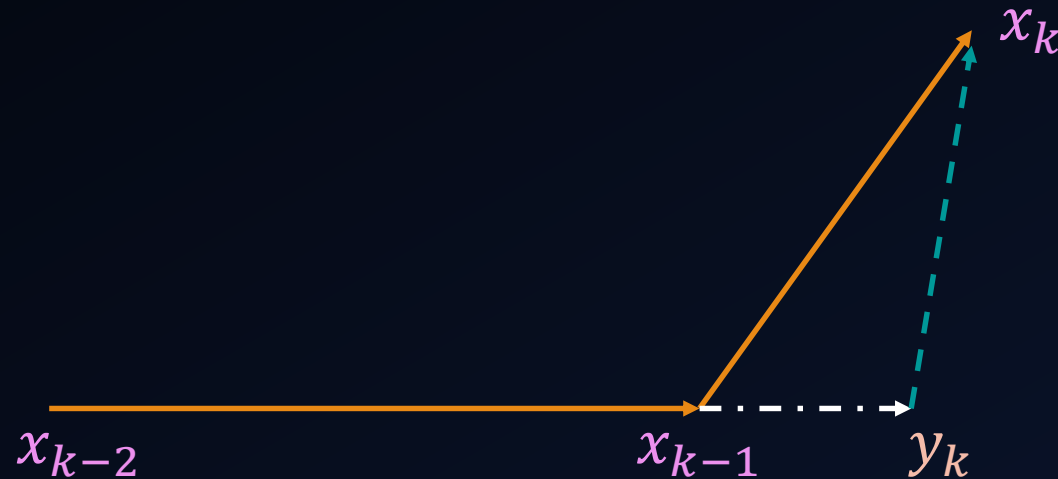
- Chebyshev poly. have two term recursive formula, hence we expect:
 - $x_k = x_{k-1} - s_{k-1}\nabla f(x_{k-1}) + \lambda_{k-1}\nabla f(x_{k-2})$, to be optimal (acceleration)

Accelerated Gradient Method [Ne83,88,Be09]

- $y_k = x_{k-1} + \frac{k-2}{k+1} (x_{k-1} - x_{k-2})$  Two step history (polation or momentum)
- $x_k = y_k - s_k \nabla f(y_k)$ (Usual gradient step)

Accelerated Gradient Method [Ne83,88,Be09]

- $y_k = x_{k-1} + \frac{k-2}{k+1} (x_{k-1} - x_{k-2})$ (Extrapolation or momentum)
- $x_k = y_k - s_k \nabla f(y_k)$ (Usual gradient step)



Rate of Convergence – Accelerated gradient

Theorem [Be09]: If f is smooth with const. L and $s_k = \frac{1}{L}$, then accelerated gradient method generates x_k such that:

$$f(x_k) - f(x^*) \leq \frac{2L \|x_0 - x^*\|^2}{(k+1)^2}.$$

Indeed optimal!

Proof Sketch:

- $f(x_k) \leq f(z) + L(x_k - z)^T(z - x_k) + \frac{L}{2} \|x_k - z\|^2 \quad \forall z \in \mathbb{R}^n$
- Convex combination at $z = x_k, z = x^*$ leads to:

$$\begin{aligned} \frac{(k+1)^2}{2L} (f(x_k) - f^*) + \|\bar{y}_k - x^*\|^2 &\leq \frac{(k)^2}{2L} (f(x_{k-1}) - f^*) + \|\bar{y}_{k-1} - x^*\|^2 \\ &\leq \|x_0 - x^*\|^2 \end{aligned}$$

Rate of Convergence – Accelerated gradient

Theorem [Be09]: If f is smooth with const. L and $s_k = \frac{1}{L}$, then accelerated gradient method generates x_k such that:

$$f(x_k) - f(x^*) \leq \frac{2L \|x_0 - x^*\|^2}{(k+1)^2}.$$

Proof Sketch:

- $f(x_k) \leq f(z) + L(x_k - y)^T(z - x_k) + \frac{L}{2} \|x_k - y\|^2 \quad \forall z \in R^n$

- Convex combination at $z = x_k, z = x^*$ leads to:

$$\begin{aligned} \frac{(k+1)^2}{2L} (f(x_k) - f^*) + \|\bar{y}_k - x^*\|^2 &\leq \frac{(k)^2}{2L} (f(x_{k-1}) - f^*) + \|\bar{y}_{k-1} - x^*\|^2 \\ &\leq \|x_0 - x^*\|^2 \end{aligned}$$

Rate of Convergence – Accelerated gradient

Theorem [Be09]: If f is smooth with const. L and $s_k = \frac{1}{L}$, then accelerated gradient method generates x_k such that:

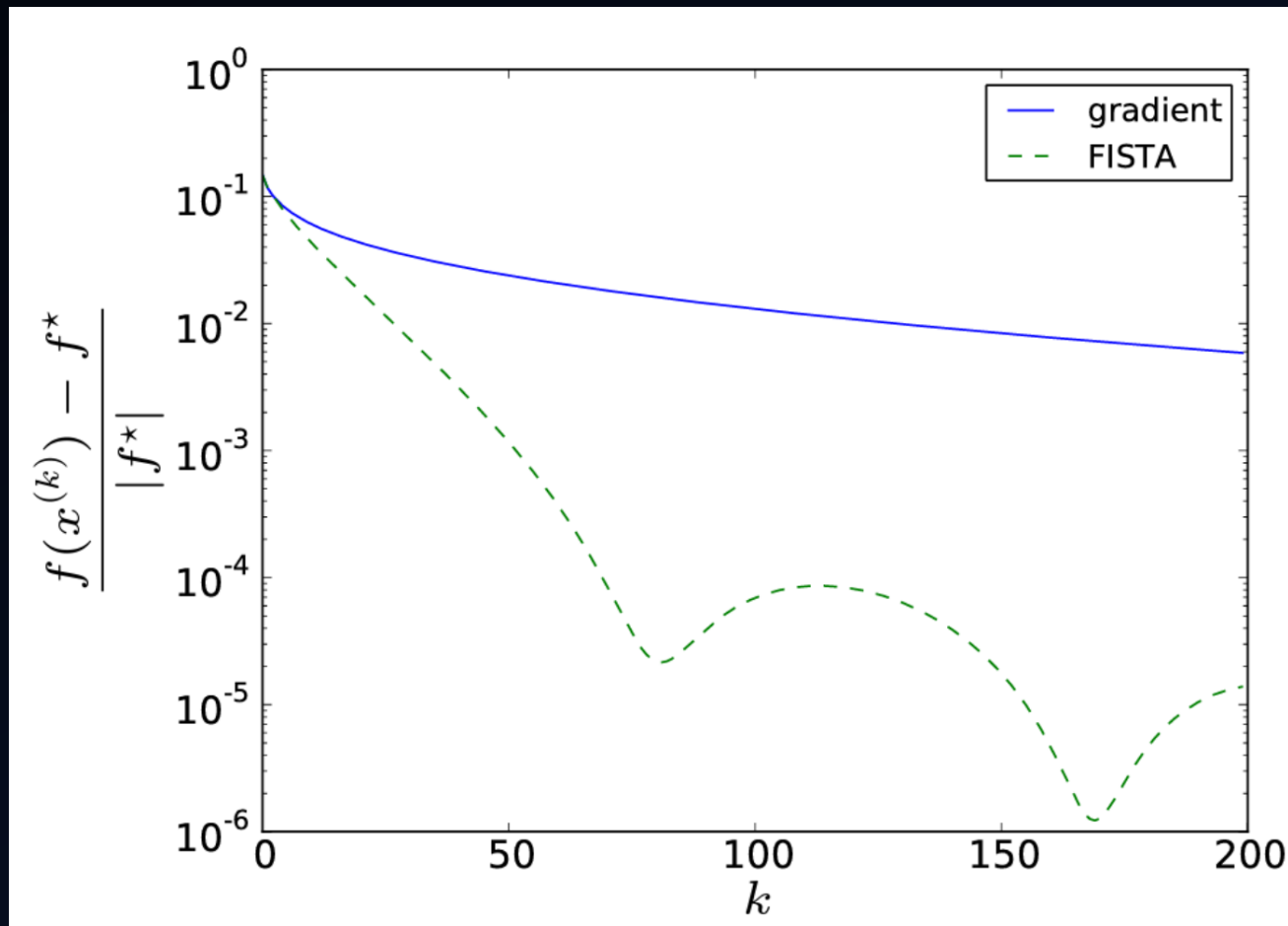
$$f(x_k) - f(x^*) \leq \frac{2L \|x_0 - x^*\|^2}{(k+1)^2}.$$

Proof Sketch:

- $f(x_k) \leq f(z) + L(x_k - y)^T(z - x_k) + \frac{L}{2} \|x_k - y\|^2 \quad \forall z \in R^n$
- Convex combination at $z = x_k, z = x^*$ leads to:
$$\frac{(k+1)^2}{2L} (f(x_k) - f^*) + \|\bar{y}_k - x^*\|^2 \leq \frac{(k)^2}{2L} (f(x_{k-1}) - f^*) + \|\bar{y}_{k-1} - x^*\|^2$$
$$\leq \|x_0 - x^*\|^2$$

A Comparison of the two gradient methods

$$\min_{x \in \mathbb{R}^{1000}} \log \left(\sum_{i=1}^{2000} e^{(a_i^T x + b_i)} \right)$$



Junk variants other than Accelerated gradient?

- Accelerated gradient is
 - Less robust than gradient method [Moritz Hardt]
 - Accumulates error with inexact oracles [De13]
- Who knows what will happen in your application?

Summary of un-constrained smooth convex programs

- **Gradient method** and friends: $\epsilon \approx O(1/k)$
 - Sub-linear and sub-optimal rate.
 - Additionally, strong convexity gives: $\epsilon \approx O\left(\left(\frac{Q-1}{Q+1}\right)^{2k}\right)$. Sub-optimal but linear rate.
- **Accelerated gradient methods**: $\epsilon \approx O(1/k^2)$
 - Sub-linear but optimal
 - $O(n)$ computation per iteration
 - Additionally, strong convexity gives: $\epsilon \approx O\left(\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k}\right)$. Optimal but still linear rate.

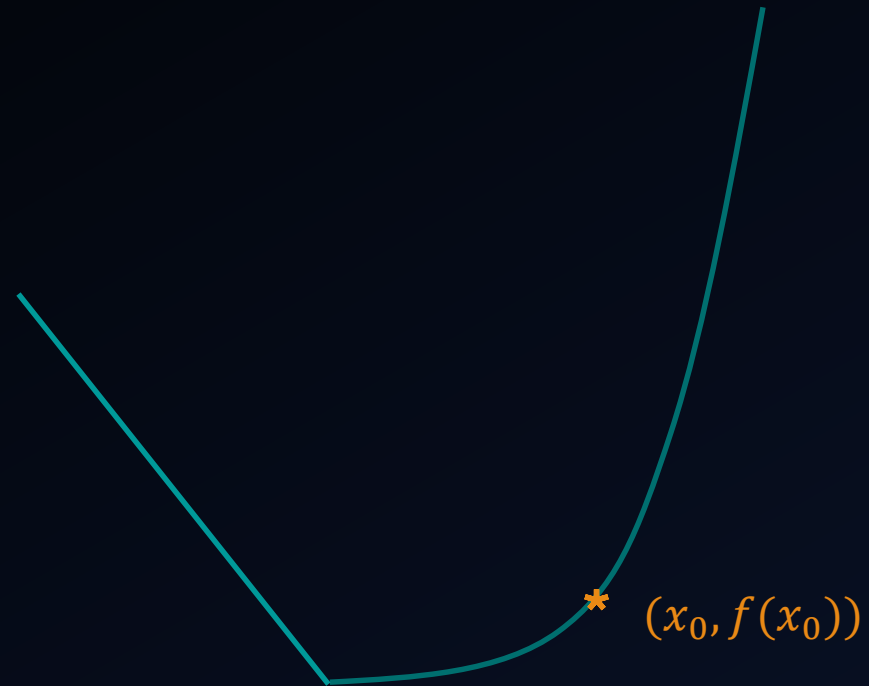
Summary of un-constrained smooth convex programs

- **Gradient method** and friends: $\epsilon \approx O(1/k)$
 - Sub-linear and sub-optimal rate.
 - Additionally, strong convexity gives: $\epsilon \approx O\left(\left(\frac{Q-1}{Q+1}\right)^{2k}\right)$. Sub-optimal but linear rate.
- **Accelerated gradient methods**: $\epsilon \approx O(1/k^2)$
 - Sub-linear but optimal
 - $O(n)$ computation per iteration
 - Additionally, strong convexity gives: $\epsilon \approx O\left(\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k}\right)$. Optimal but still linear rate.

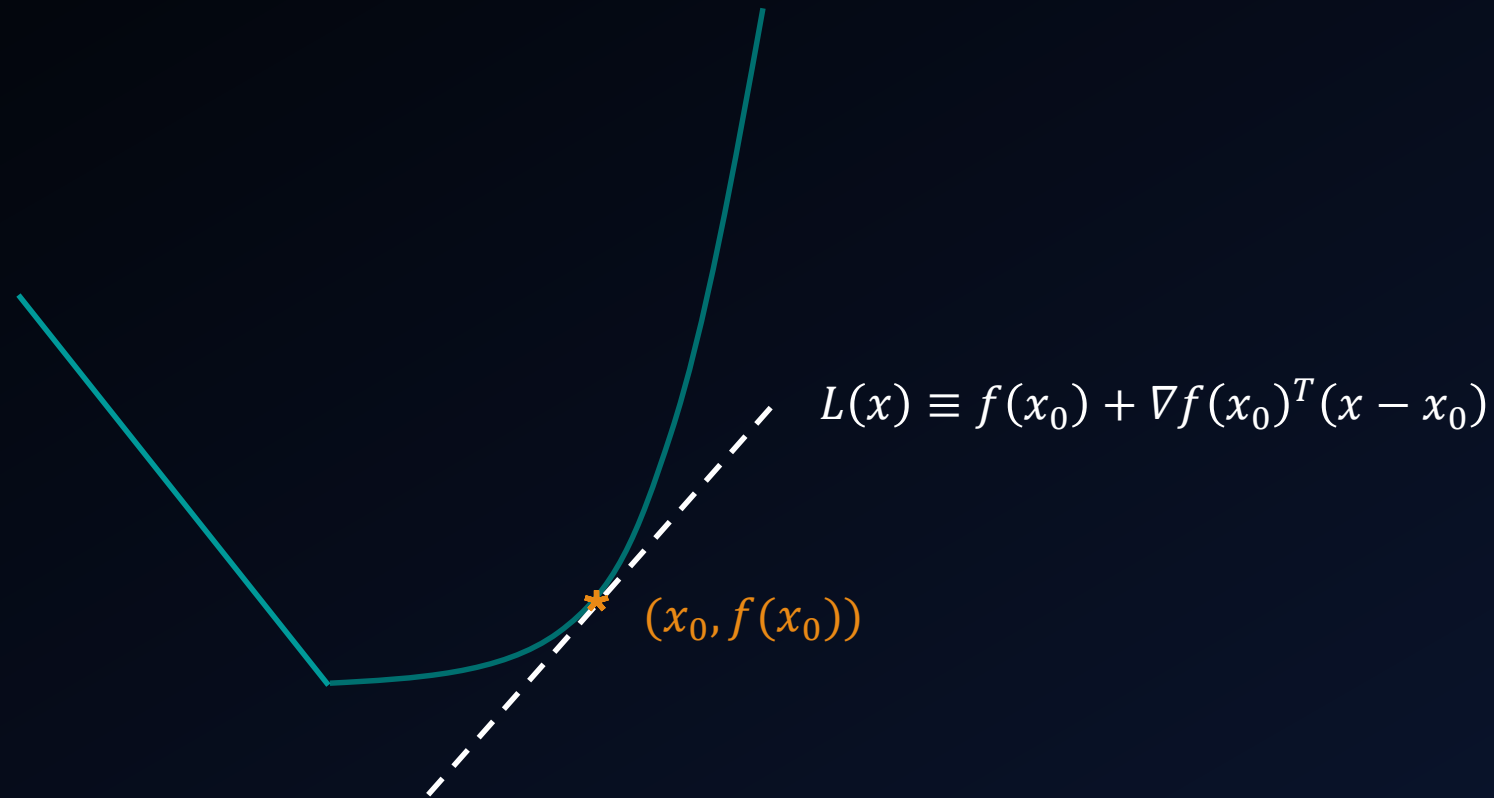
Non-smooth unconstrained

$$\text{MIN}_{w \in R^n} \sum_{i=1}^m |w' \phi(x_i) - y_i|$$

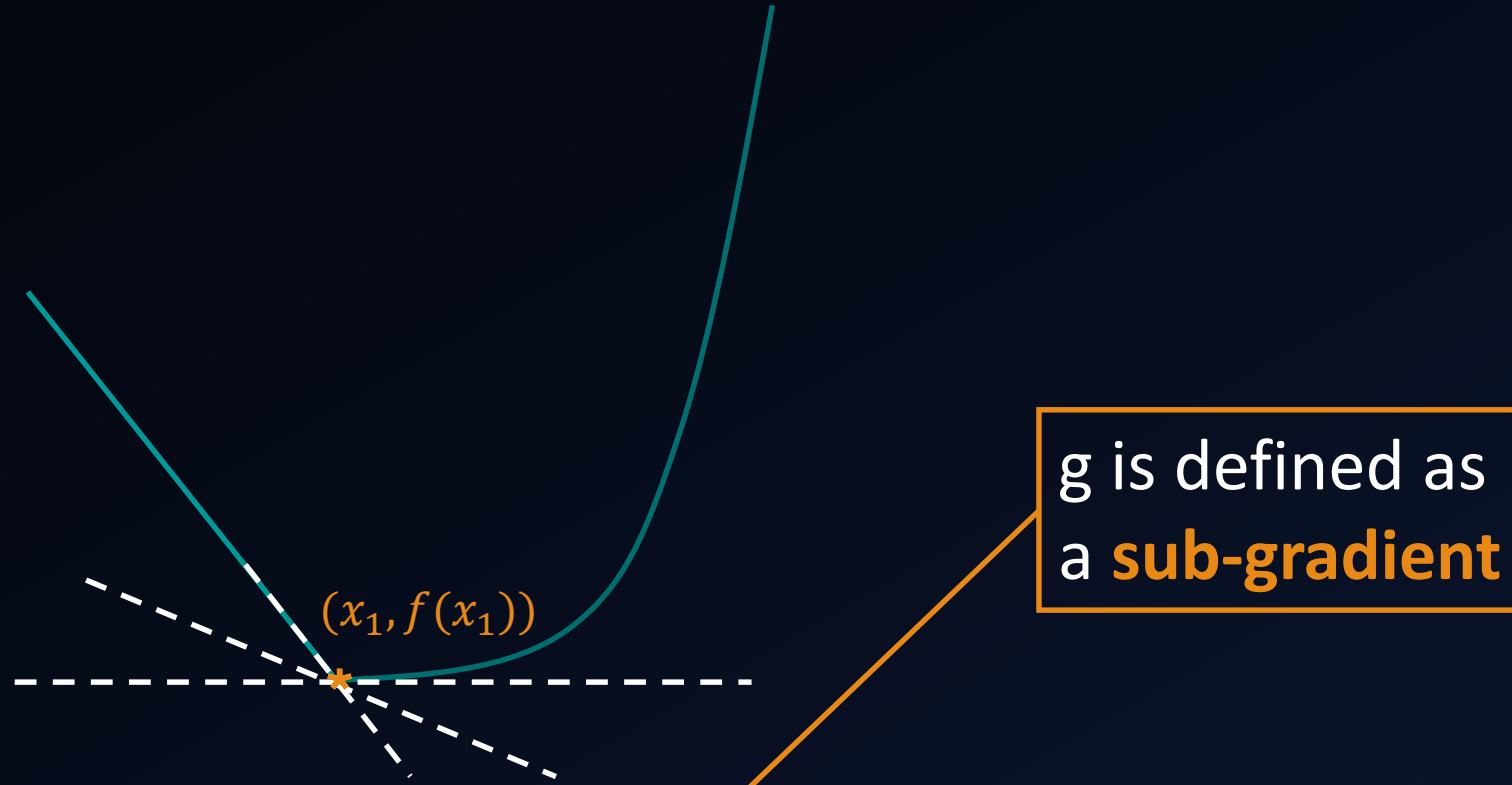
What is first order info?



What is first order info?



What is first order info?



g is defined as a **sub-gradient**

Canonical form: $L(x) \equiv f(x_1) + g^T(x - x_1)$.
Multiple g exist such that $L(x) \leq f(x) \forall x$

First Order Methods (Non-smooth)

Theorem: Let f be a closed convex function. Then

- At any $x \in ri(dom f)$, sub-gradient exists and set of all sub-gradients (denoted by $\partial f(x)$; sub-differential set) is closed convex.
- If f is differentiable at $x \in int(dom f)$, then gradient is the only sub-gradient.

Theorem: $x \in R^n$ minimizes f if and only if $0 \in \partial f(x)$.

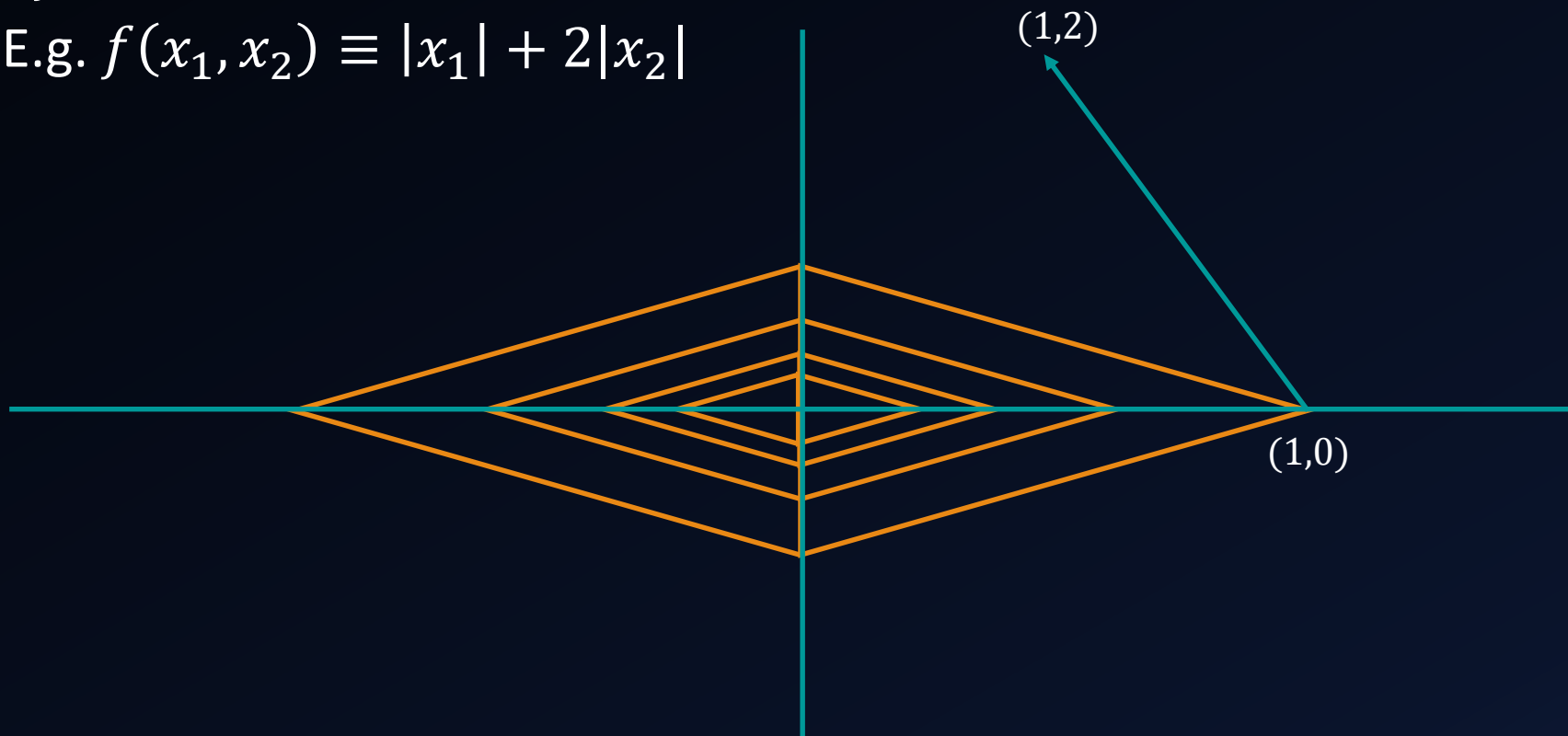
Sub-gradient Method

- Assume oracle that throws a sub-gradient.
- Sub-gradient method:

- $x_{k+1} = \operatorname{argmin}_{x \in \mathbb{R}^n} f(x_k) + g_f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2$
- $x_{k+1} = x_k - s_k g_f(x_k)$

Can sub-gradient replace gradient?

- No majorization minimization
- $-g_f(x)$ not even descent direction
 - E.g. $f(x_1, x_2) \equiv |x_1| + 2|x_2|$



How far can sub-gradient take?

Expect slower
than $O(1/k)$

How far can sub-gradient take?

Always exists!

Theorem[Ne04]: Let $\|x_0 - x^*\| \leq R$ and L the Lip. const. of f over this ball. Then sequence generated by sub-gradient descent satisfies:

$$\min_{i \in \{1, \dots, k\}} f(x_i) - f(x^*) \leq \frac{LR}{\sqrt{k+1}}.$$

Proof Sketch:

- $2s_k \Delta_k \leq r_k^2 - r_{k+1}^2 + s_k^2 \|g_k(x_k)\|^2$
- LHS $\leq \frac{r_0^2 + \sum_{i=0}^k s_i^2 \|g_i(x_i)\|^2}{2 \sum_{i=0}^k s_i} \leq \frac{R^2 + \sum_{i=0}^k s_i^2 L}{2 \sum_{i=0}^k s_i}$; Choose $s_k = R / \sqrt{k+1}$

How far can sub-gradient take?

Theorem[Ne04]: Let $\|x_0 - x^*\| \leq R$ and L the Lip. const. of f over this ball. Then sequence generated by sub-gradient descent satisfies:

$$\min_{i \in \{1, \dots, k\}} f(x_i) - f(x^*) \leq \frac{LR}{\sqrt{k+1}}.$$

Proof Sketch:

- $2s_k \Delta_k \leq r_k^2 - r_{k+1}^2 + s_k^2 \|g_f(x_k)\|^2$
- $\text{LHS} \leq \frac{r_0^2 + \sum_{i=0}^k s_i^2 \|g_f(x_k)\|^2}{2 \sum_{i=0}^k s_i} \leq \frac{R^2 + \sum_{i=0}^k s_i^2 L^2}{2 \sum_{i=0}^k s_i}$; Choose $s_k = R/\sqrt{k+1}$

Is this optimal?

Theorem[Ne04]: For any $k \leq n - 1$, and any x_0 such that

$\|x_0 - x^*\| \leq R$, there exists a convex f , with const. L over the ball, such that with **any first order method**, we have:

$$f(x_k) - f(x^*) \geq \frac{LR}{2(1 + \sqrt{k + 1})}.$$

Proof Sketch: Choose function such that

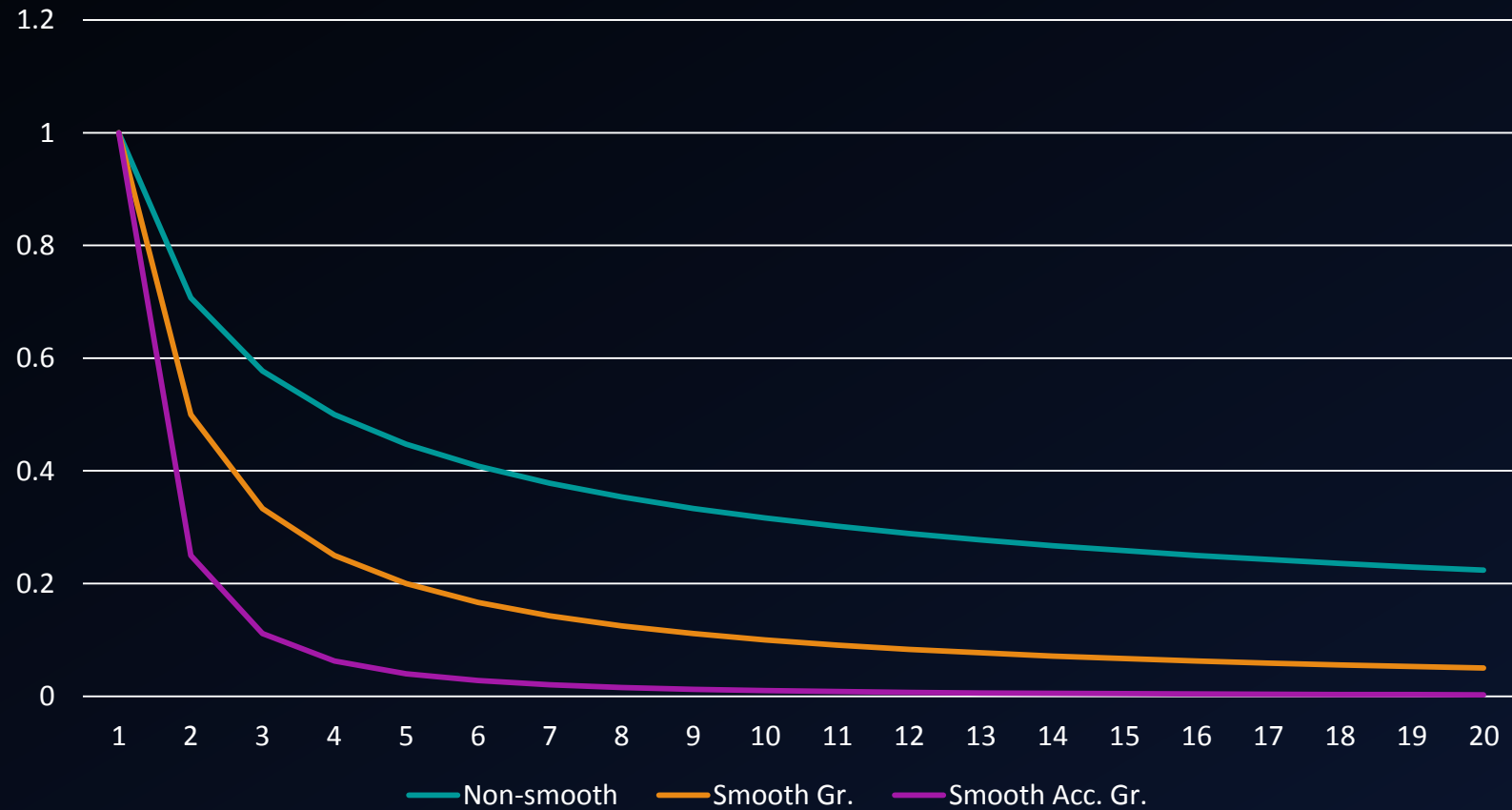
$$x_k \in \text{lin} \left(g_f(x_0), \dots, g_f(x_{k-1}) \right) \subset R^{k,n}$$

Summary of non-smooth unconstrained

- Sub-gradient descent method: $\epsilon \approx O\left(\frac{1}{\sqrt{k}}\right)$.
 - Sub-linear, slower than smooth case
 - But, optimal!
 - Can do better if additional structure (later)

Summary of Unconstrained Case

Chart Title



Bibliography

- **[Ne04]** Nesterov, Yurii. *Introductory lectures on convex optimization : a basic course*. Kluwer Academic Publ., 2004. <http://hdl.handle.net/2078.1/116858>.
- **[Ne83]** Nesterov, Yurii. *A method of solving a convex programming problem with convergence rate $O(1/k^2)$* . Soviet Mathematics Doklady, Vol. 27(2), 372-376 pages.
- **[Mo12]** Moritz Hardt, Guy N. Rothblum and Rocco A. Servedio. *Private data release via learning thresholds*. SODA 2012, 168-187 pages.
- **[Be09]** Amir Beck and Marc Teboulle. *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*. SIAM Journal of Imaging Sciences, Vol. 2(1), 2009. 183-202 pages.
- **[De13]** Olivier Devolder, François Glineur and Yurii Nesterov. *First-order methods of smooth convex optimization with inexact oracle*. Mathematical Programming 2013.