Optimization for Machine Learning
CSCI-599

Lecture 3: Constrained and Non-Smooth Gradient Descent

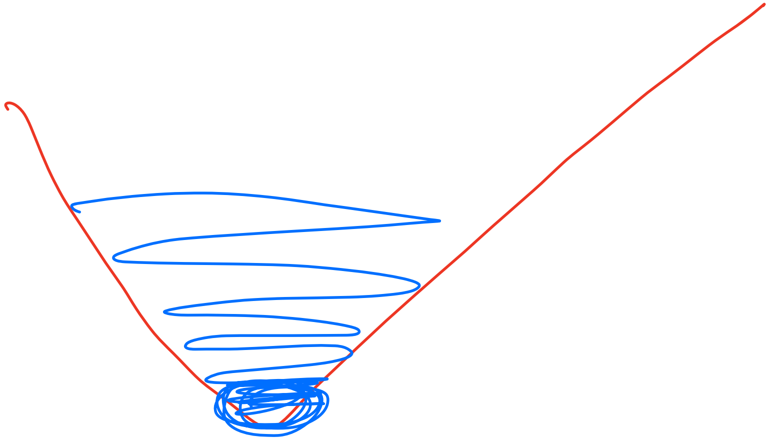
Sai Praneeth Karimireddy

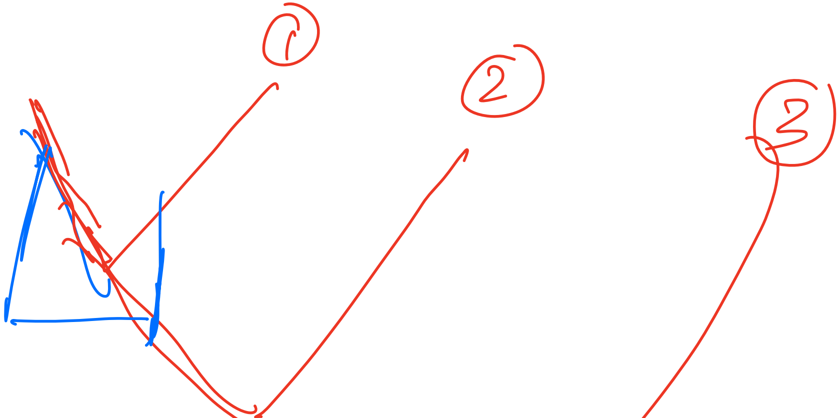
USC – `https://spkreddy.org/optmlspring2025.html`

February 3, 2025

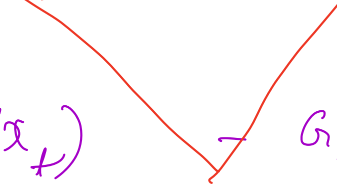
Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Same as vanilla case for smooth functions, but now for any h for which we can compute the proximal mapping.





Recap



$$x_{t+1} = x_t - \gamma \nabla f(x_t) \quad \to \text{GD}$$

① f is B-Lip $\iff |f(x)-f(y)| \le B\|x-y\|_2 \ \forall x,y$

$$\|\nabla f(x)\|_2 \le B \quad \forall x$$

$$\gamma = \frac{R}{B\sqrt{T}} \Rightarrow f(\bar{x}_T) - f^* \le \frac{BR}{\sqrt{T}}$$

$$\bar{x}_T = \frac{1}{T}\sum_t x_t$$

② f is L-smooth $\Rightarrow \|\nabla f(x_t)\|^2 \le 2L\left(f(x_t) - f(x_{t+1})\right)$

$$\gamma = \frac{1}{L} \Rightarrow f(x_T) - f^* \le \frac{LR}{T}$$

③ f is L-smooth + μ-strong cvx $\Rightarrow$

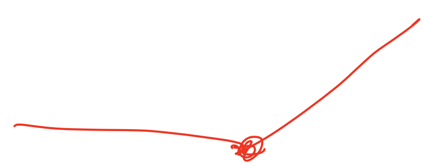
$$\gamma = \frac{1}{L} \Rightarrow \|x_T - x^*\|_2^2 \le \left(1 - \frac{\mu}{L}\right)^T \|x_0 - x^*\|_2^2$$

$$f(x) = \underbrace{\frac{1}{2} x^T A x - b^T x}_{\text{cvx}} + \underbrace{\lambda \|x\|_1}_{\text{cvx}}$$

Lasso Regression

ReLU

$$f(x) = \underline{\max(\cancel{Ax-b}\; x, 0)}$$



Subgradients

What if f is not differentiable?

If f is diff $\Rightarrow$

$\nabla f(x)$ is a subgrad

Definition

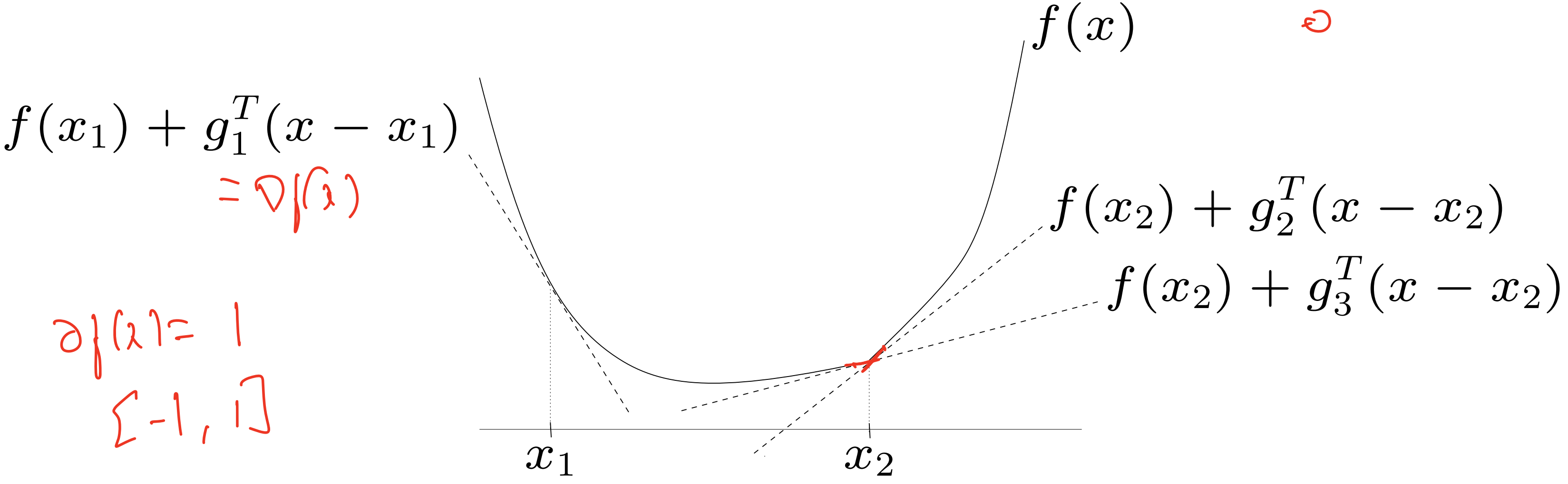
$\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{y} \in \mathbf{dom}(f)$$

$f(x) = |x|$

$x = 1 \qquad \partial f(x) = 1$

$x = 0 \qquad [-1, 1]$

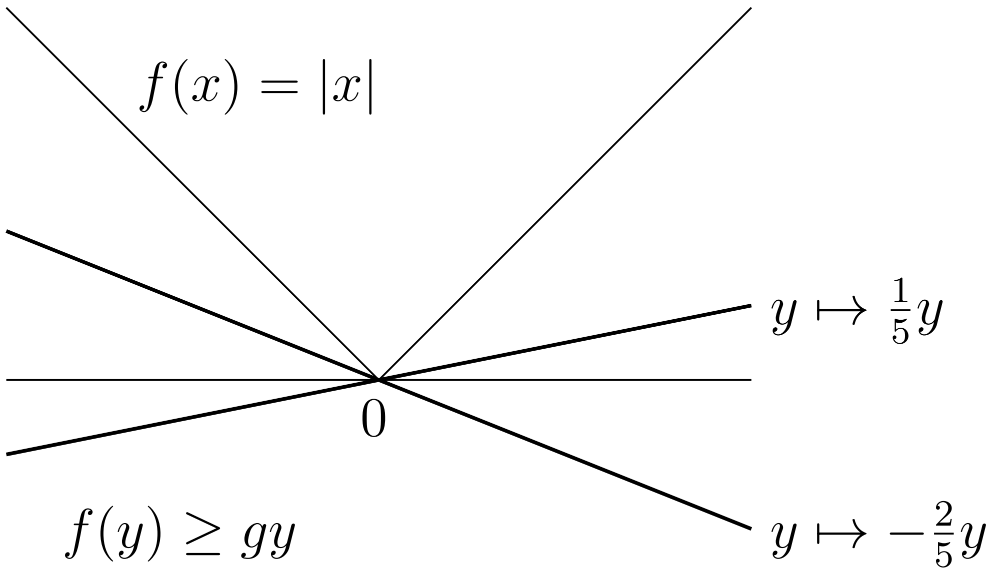


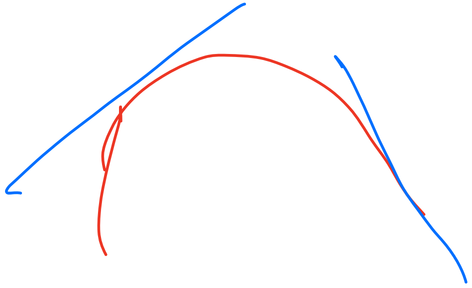
$\partial f(\mathbf{x}) \subseteq \mathbb{R}^d$ is the subdifferential, the set of subgradients of f at $\mathbf{x}$.

Subgradients II

$f(x) = -x^2$

Example:



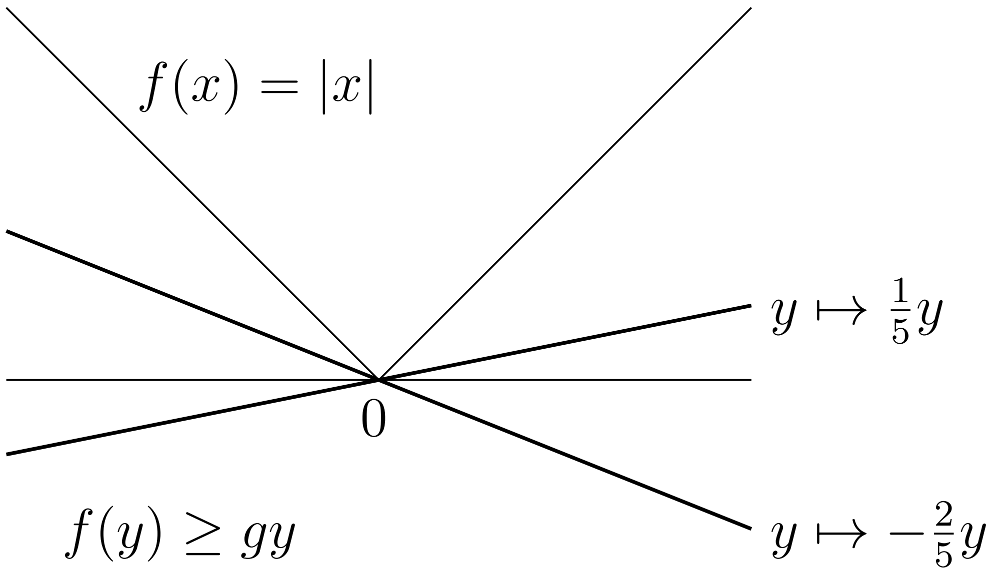


Subgradient condition at $x = 0$:

f is cvx $\Rightarrow$ subgrad exists

Subgradients II

Example:



Subgradient condition at $x = 0$: $f(y) \geq f(0) + g(y - 0) = gy$.

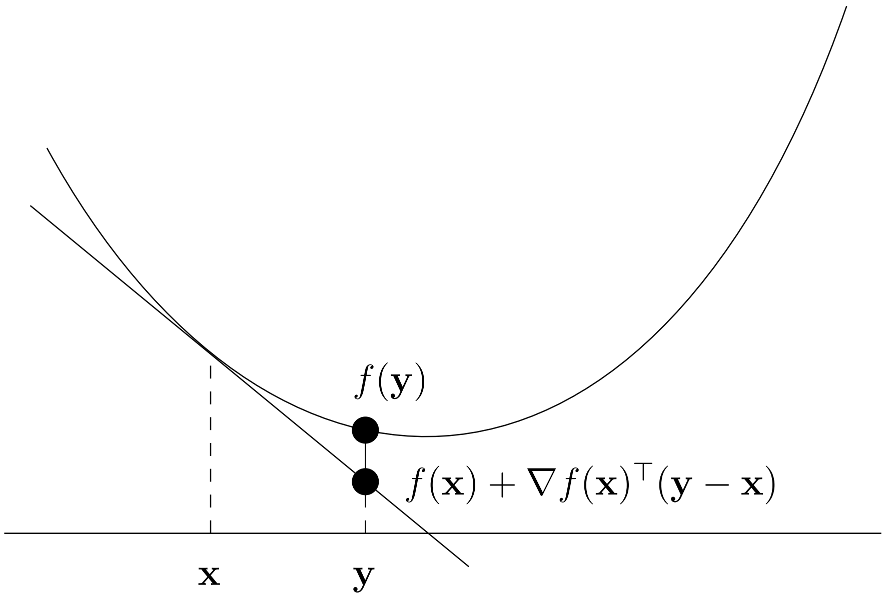
$\partial f(0) = [-1, 1]$

Subgradients III

Lemma (Exercise 28)

If $f : \mathbf{dom}(f) \to \mathbb{R}$ *is differentiable at* $\mathbf{x} \in \mathbf{dom}(f)$, *then* $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$.

Either exactly one subgradient $\nabla f(\mathbf{x})$...



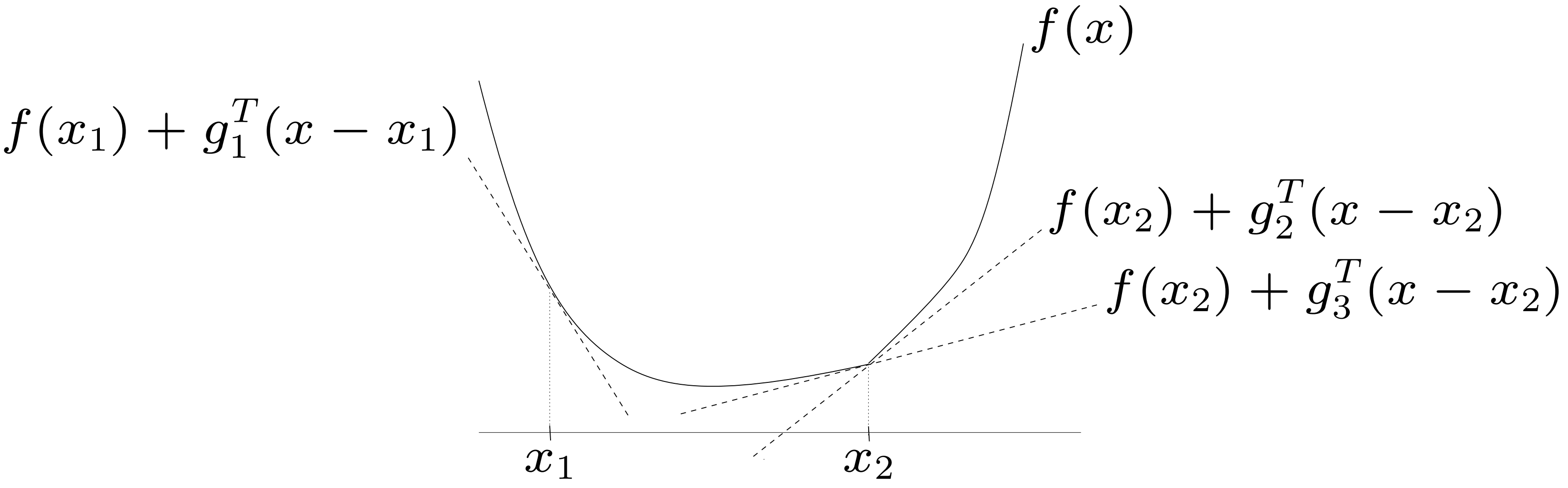
...or no subgradient at all.

Subgradient characterization of convexity

"convex = subgradients everywhere"

Lemma (Exercise 29)

A function $f : \mathbf{dom}(f) \to \mathbb{R}$ *is convex if and only if* $\mathbf{dom}(f)$ *is convex and* $\partial f(\mathbf{x}) \neq \emptyset$ *for all* $\mathbf{x} \in \mathbf{dom}(f)$.



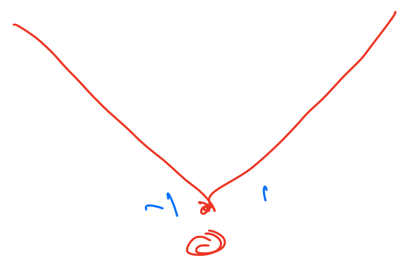
→ Let $g(x_t)$ be "a" subgradient of f at x_t

— $x_{t+1} = x_t - \gamma g(x_t)$ subgradient descent

x^* is opt of $\min_x f(x)$

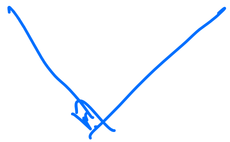
if $\nabla f(x^*) = 0$

$f(x) = |x|$



$0 \in g(x^*) \Rightarrow x^*$ is optimum

Convex and Lipschitz = bounded subgradients



Lemma (Exercise 30)

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ *be convex,* $\mathbf{dom}(f)$ *open,* $B \in \mathbb{R}_+$*. Then the following two statements are equivalent.*

(i) $\|\mathbf{g}\| \leq B$ *for all* $\mathbf{x} \in \mathbf{dom}(f)$ *and all* $\mathbf{g} \in \partial f(\mathbf{x})$.

(ii) $|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$.

Subgradient optimality condition

Lemma

Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ *and* $\mathbf{x} \in \mathbf{dom}(f)$*. If* $\mathbf{0} \in \partial f(\mathbf{x})$*, then* $\mathbf{x}$ *is a global minimum.*

Proof.

By definition of subgradients, $\mathbf{g} = \mathbf{0} \in \partial f(\mathbf{x})$ gives

Subgradient optimality condition

Lemma

Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ *and* $\mathbf{x} \in \mathbf{dom}(f)$*. If* $\mathbf{0} \in \partial f(\mathbf{x})$*, then* $\mathbf{x}$ *is a global minimum.*

Proof.

By definition of subgradients, $\mathbf{g} = \mathbf{0} \in \partial f(\mathbf{x})$ gives

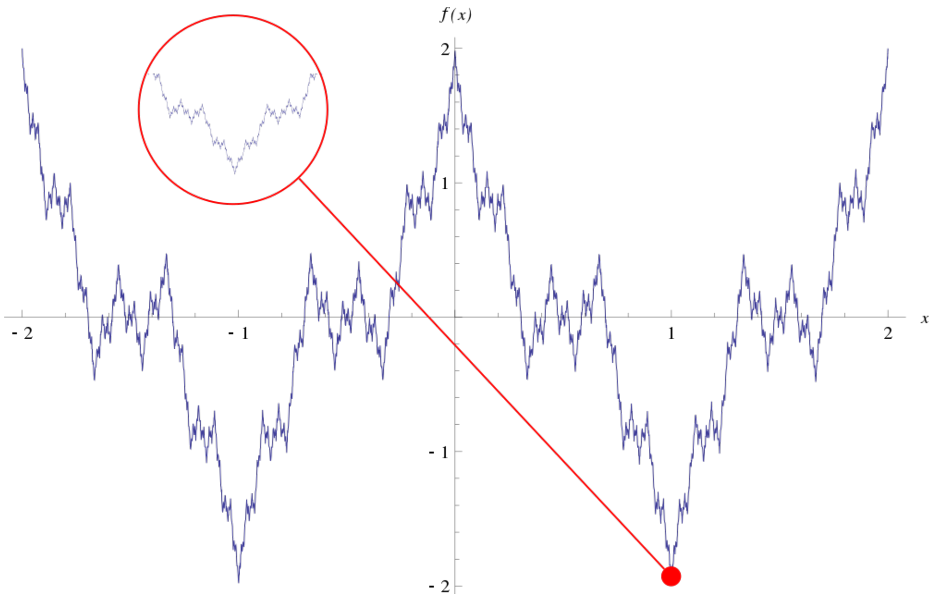
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) = f(\mathbf{x})$$

for all $\mathbf{y} \in \mathbf{dom}(f)$, so $\mathbf{x}$ is a global minimum. □

Differentiability of convex functions

How “wild” can a non-differentiable convex function be?

Weierstrass function: a function that is continuous everywhere but differentiable nowhere



https://commons.wikimedia.org/wiki/File:WeierstrassFunction.svg

Differentiability of convex functions

Theorem ([?, Theorem 25.5])

A convex function $f : \mathbf{dom}(f) \to \mathbb{R}$ is differentiable almost everywhere.

In other words:

- Set of points where f is non-differentiable has measure 0 (no volume).
- For all $\mathbf{x} \in \mathbf{dom}(f)$ and all $\varepsilon > 0$, there is a point $\mathbf{x}'$ such that $\|\mathbf{x} - \mathbf{x}'\| < \varepsilon$ and f is differentiable at $\mathbf{x}'$.

The subgradient descent algorithm

Subgradient descent: choose $\mathbf{x}_0 \in \mathbb{R}^d$.

$$\text{Let } \mathbf{g}_t \in \partial f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t$$

for **times** $t = 0, 1, \ldots,$ and **stepsizes** $\gamma_t \geq 0$.

Stepsize can vary with time!

This is possible in (projected) gradient descent as well, but so far, we didn't need it.

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and B-Lipschitz continuous with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \le R$. Choosing the constant stepsize

$$\gamma := \frac{R}{B\sqrt{T}},$$

subgradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \frac{RB}{\sqrt{T}}.$$

Proof is identical to the one of Theorem 2.1, except...

- In vanilla analyis, now use $\mathbf{g}_t \in \partial f(\mathbf{x}_t)$ instead of $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$.
- Inequality $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ now follows from subgradient property instead of first-order charaterization of convexity.

$$\boxed{\sum g_t^T (x_t - x^*)} \le \frac{1}{2\gamma}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2 \right) + \frac{\gamma}{2} \sum \|g_t\|^2 \le B$$

$$\ge \sum f(x_t) - f(x^*)$$

Optimality of first-order methods

With all the convergence rates we have seen so far, a very natural question to ask is if these rates are best possible or not. Surprisingly, the rate can indeed not be improved in general.

Theorem (Nesterov)

For any $T \leq d-1$ and starting point $\mathbf{x}_0$, there is a function f in the problem class of B-Lipschitz functions over $\mathbb{R}^d$, such that any (sub)gradient method has an objective error at least

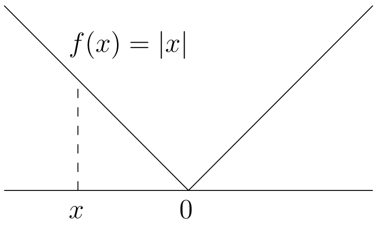
$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \geq \frac{RB}{2(1+\sqrt{T+1})} .$$

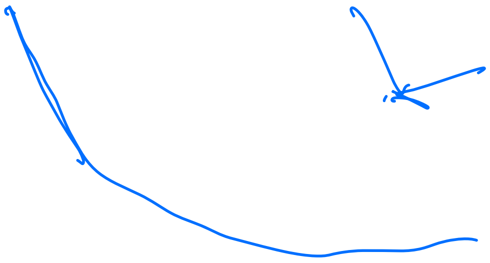
Smooth (non-differentiable) functions?

smoothness

differentiability

They don't exist (Exercise 31)!

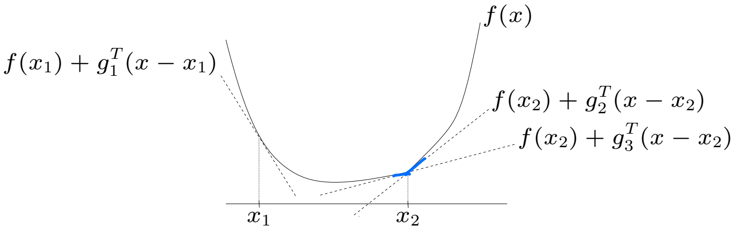




At 0, graph can't be below a tangent paraboloid.

Can we still improve over $O(1/\varepsilon^2)$ steps for Lipschitz functions?

Yes, if we also require strong convexity (graph is above not too flat tangent paraboloids).



Optimality

x^* is opt of $\min_x f(x)$ f is cvx

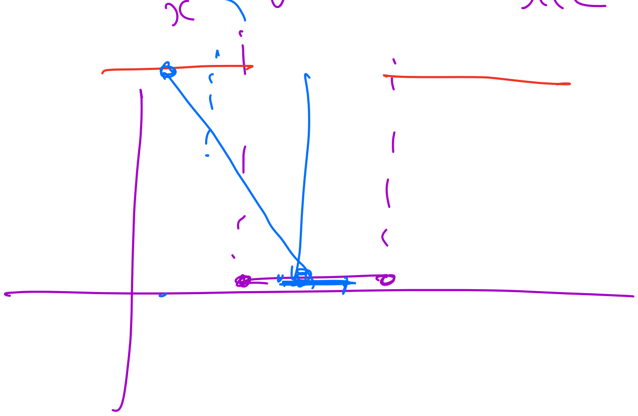
if $0 \in \text{subgrad}(x^*)$

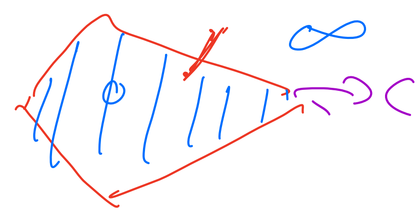
$\min_{x \in C} f(x)$ constrained opt

$|f(x) - f(y)| \leq B \cdot \|x - y\|_2$

$\underbrace{\tilde{f}(x)}_{\text{cvx}} = \underbrace{f(x)}_{\text{cvx}} + \underline{1(x \in C)}$

$\min_x \tilde{f}(x) \Leftrightarrow \min_{x \in C} f(x)$





we need $\underline{\text{Lip assumption}}$

+ smooth

f is cvx is not sufficient

for convergence