



**Non-smooth convex optimization. Lower  
bounds. Subgradient method.**

**Daniil Merkulov**

Optimization methods. MIPT

## Non-smooth problems

# $\ell_1$ -regularized linear least squares

$\ell_1$  induces sparsity

$\ell_2$  regularization.  $\|Xw - y\|_2^2 \rightarrow \min_{\|w\|_2 \leq 1}$



$\ell_1$  regularization.  $\|Xw - y\|_2^2 \rightarrow \min_{\|w\|_1 \leq 1}$



@fminxyz

# Norms are not smooth

$$\min_{x \in \mathbb{R}^n} f(x),$$

A classical convex optimization problem is considered. We assume that  $f(x)$  is a convex function, but now we do not require smoothness.



Figure 1: Norm cones for different  $p$  - norms are non-smooth

# Wolfe's example

Wolfe's example

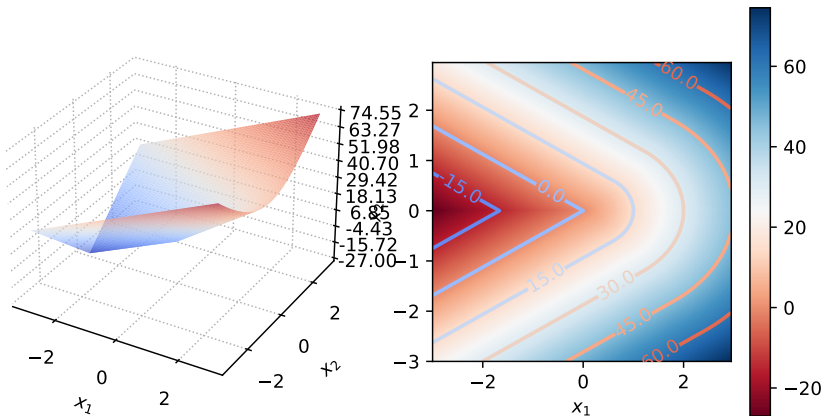


Figure 2: Wolfe's example. [Open in Colab](#)

## Subgradient calculus

## Convex function linear lower bound

An important property of a continuous convex function  $f(x)$  is that at any chosen point  $x_0$  for all  $x \in \text{dom } f$  the inequality holds:

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$



Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

## Convex function linear lower bound



An important property of a continuous convex function  $f(x)$  is that at any chosen point  $x_0$  for all  $x \in \text{dom } f$  the inequality holds:

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

for some vector  $g$ , i.e., the tangent to the function's graph is the *global* estimate from below for the function.

- If  $f(x)$  is differentiable, then  $g = \nabla f(x_0)$

Figure 3: Taylor linear approximation serves as a global lower bound for a convex function



## Convex function linear lower bound



An important property of a continuous convex function  $f(x)$  is that at any chosen point  $x_0$  for all  $x \in \text{dom } f$  the inequality holds:

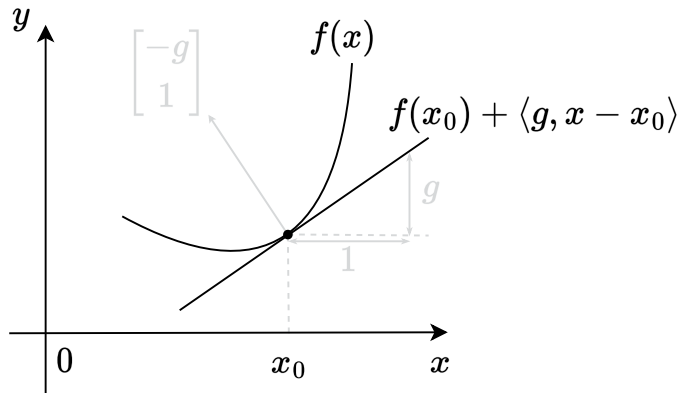
$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

for some vector  $g$ , i.e., the tangent to the function's graph is the *global* estimate from below for the function.

- If  $f(x)$  is differentiable, then  $g = \nabla f(x_0)$
- Not all continuous convex functions are differentiable.

Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

## Convex function linear lower bound



An important property of a continuous convex function  $f(x)$  is that at any chosen point  $x_0$  for all  $x \in \text{dom } f$  the inequality holds:

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

for some vector  $g$ , i.e., the tangent to the function's graph is the *global* estimate from below for the function.

- If  $f(x)$  is differentiable, then  $g = \nabla f(x_0)$
- Not all continuous convex functions are differentiable.

Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

## Convex function linear lower bound



An important property of a continuous convex function  $f(x)$  is that at any chosen point  $x_0$  for all  $x \in \text{dom } f$  the inequality holds:

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

for some vector  $g$ , i.e., the tangent to the function's graph is the *global* estimate from below for the function.

- If  $f(x)$  is differentiable, then  $g = \nabla f(x_0)$
- Not all continuous convex functions are differentiable.

We do not want to lose such a lovely property.

Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

## Subgradient and subdifferential

A vector  $g$  is called the **subgradient** of a function  $f(x) : S \rightarrow \mathbb{R}$  at a point  $x_0$  if  $\forall x \in S$ :

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

## Subgradient and subdifferential

A vector  $g$  is called the **subgradient** of a function  $f(x) : S \rightarrow \mathbb{R}$  at a point  $x_0$  if  $\forall x \in S$ :

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

The set of all subgradients of a function  $f(x)$  at a point  $x_0$  is called the **subdifferential** of  $f$  at  $x_0$  and is denoted by  $\partial f(x_0)$ .

## Subgradient and subdifferential

A vector  $g$  is called the **subgradient** of a function  $f(x) : S \rightarrow \mathbb{R}$  at a point  $x_0$  if  $\forall x \in S$ :

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

The set of all subgradients of a function  $f(x)$  at a point  $x_0$  is called the **subdifferential** of  $f$  at  $x_0$  and is denoted by  $\partial f(x_0)$ .

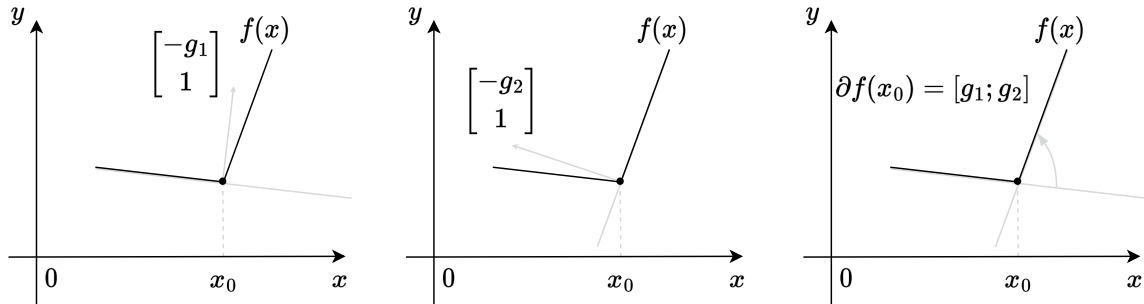


Figure 4: Subdifferential is a set of all possible subgradients

# Subgradient and subdifferential

Find  $\partial f(x)$ , if  $f(x) = |x|$

## Subgradient and subdifferential

Find  $\partial f(x)$ , if  $f(x) = |x|$

$$f(x) = |x|$$



$$\partial f(x)$$





## Subdifferential properties

- If  $x_0 \in \text{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set.

## Subdifferential properties

- If  $x_0 \in \mathbf{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set.
- The convex function  $f(x)$  is differentiable at the point  $x_0 \Rightarrow \partial f(x_0) = \{\nabla f(x_0)\}$ .

## Subdifferential properties

- If  $x_0 \in \mathbf{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set.
- The convex function  $f(x)$  is differentiable at the point  $x_0 \Rightarrow \partial f(x_0) = \{\nabla f(x_0)\}$ .
- If  $\partial f(x_0) \neq \emptyset \quad \forall x_0 \in S$ , then  $f(x)$  is convex on  $S$ .

## Subdifferential properties

- If  $x_0 \in \mathbf{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set.
- The convex function  $f(x)$  is differentiable at the point  $x_0 \Rightarrow \partial f(x_0) = \{\nabla f(x_0)\}$ .
- If  $\partial f(x_0) \neq \emptyset \quad \forall x_0 \in S$ , then  $f(x)$  is convex on  $S$ .

## Subdifferential properties

- If  $x_0 \in \text{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set.
- The convex function  $f(x)$  is differentiable at the point  $x_0 \Rightarrow \partial f(x_0) = \{\nabla f(x_0)\}$ .
- If  $\partial f(x_0) \neq \emptyset \quad \forall x_0 \in S$ , then  $f(x)$  is convex on  $S$ .

### i Subdifferential of a differentiable function

Let  $f : S \rightarrow \mathbb{R}$  be a function defined on the set  $S$  in a Euclidean space  $\mathbb{R}^n$ . If  $x_0 \in \text{ri}(S)$  and  $f$  is differentiable at  $x_0$ , then either  $\partial f(x_0) = \emptyset$  or  $\partial f(x_0) = \{\nabla f(x_0)\}$ . Moreover, if the function  $f$  is convex, the first scenario is impossible.

## Subdifferential properties

- If  $x_0 \in \text{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set.
- The convex function  $f(x)$  is differentiable at the point  $x_0 \Rightarrow \partial f(x_0) = \{\nabla f(x_0)\}$ .
- If  $\partial f(x_0) \neq \emptyset \quad \forall x_0 \in S$ , then  $f(x)$  is convex on  $S$ .

### i Subdifferential of a differentiable function

Let  $f : S \rightarrow \mathbb{R}$  be a function defined on the set  $S$  in a Euclidean space  $\mathbb{R}^n$ . If  $x_0 \in \text{ri}(S)$  and  $f$  is differentiable at  $x_0$ , then either  $\partial f(x_0) = \emptyset$  or  $\partial f(x_0) = \{\nabla f(x_0)\}$ . Moreover, if the function  $f$  is convex, the first scenario is impossible.

### Proof

1. Assume, that  $s \in \partial f(x_0)$  for some  $s \in \mathbb{R}^n$  distinct from  $\nabla f(x_0)$ . Let  $v \in \mathbb{R}^n$  be a unit vector. Because  $x_0$  is an interior point of  $S$ , there exists  $\delta > 0$  such that  $x_0 + tv \in S$  for all  $0 < t < \delta$ . By the definition of the subgradient, we have

$$f(x_0 + tv) \geq f(x_0) + t\langle s, v \rangle$$

## Subdifferential properties

- If  $x_0 \in \text{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set.
- The convex function  $f(x)$  is differentiable at the point  $x_0 \Rightarrow \partial f(x_0) = \{\nabla f(x_0)\}$ .
- If  $\partial f(x_0) \neq \emptyset \quad \forall x_0 \in S$ , then  $f(x)$  is convex on  $S$ .

### i Subdifferential of a differentiable function

Let  $f : S \rightarrow \mathbb{R}$  be a function defined on the set  $S$  in a Euclidean space  $\mathbb{R}^n$ . If  $x_0 \in \text{ri}(S)$  and  $f$  is differentiable at  $x_0$ , then either  $\partial f(x_0) = \emptyset$  or  $\partial f(x_0) = \{\nabla f(x_0)\}$ . Moreover, if the function  $f$  is convex, the first scenario is impossible.

### Proof

1. Assume, that  $s \in \partial f(x_0)$  for some  $s \in \mathbb{R}^n$  distinct from  $\nabla f(x_0)$ . Let  $v \in \mathbb{R}^n$  be a unit vector. Because  $x_0$  is an interior point of  $S$ , there exists  $\delta > 0$  such that  $x_0 + tv \in S$  for all  $0 < t < \delta$ . By the definition of the subgradient, we have

$$f(x_0 + tv) \geq f(x_0) + t\langle s, v \rangle$$

## Subdifferential properties

- If  $x_0 \in \text{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set. which implies:
- The convex function  $f(x)$  is differentiable at the point  $x_0 \Rightarrow \partial f(x_0) = \{\nabla f(x_0)\}$ .
- If  $\partial f(x_0) \neq \emptyset \quad \forall x_0 \in S$ , then  $f(x)$  is convex on  $S$ .

### i Subdifferential of a differentiable function

Let  $f : S \rightarrow \mathbb{R}$  be a function defined on the set  $S$  in a Euclidean space  $\mathbb{R}^n$ . If  $x_0 \in \text{ri}(S)$  and  $f$  is differentiable at  $x_0$ , then either  $\partial f(x_0) = \emptyset$  or  $\partial f(x_0) = \{\nabla f(x_0)\}$ . Moreover, if the function  $f$  is convex, the first scenario is impossible.

### Proof

1. Assume, that  $s \in \partial f(x_0)$  for some  $s \in \mathbb{R}^n$  distinct from  $\nabla f(x_0)$ . Let  $v \in \mathbb{R}^n$  be a unit vector. Because  $x_0$  is an interior point of  $S$ , there exists  $\delta > 0$  such that  $x_0 + tv \in S$  for all  $0 < t < \delta$ . By the definition of the subgradient, we have

$$f(x_0 + tv) \geq f(x_0) + t\langle s, v \rangle$$

$$\frac{f(x_0 + tv) - f(x_0)}{t} \geq \langle s, v \rangle$$

for all  $0 < t < \delta$ . Taking the limit as  $t$  approaches 0 and using the definition of the gradient, we get:

$$\langle \nabla f(x_0), v \rangle = \lim_{t \rightarrow 0; 0 < t < \delta} \frac{f(x_0 + tv) - f(x_0)}{t} \geq \langle s, v \rangle$$

2. From this,  $\langle s - \nabla f(x_0), v \rangle \geq 0$ . Due to the arbitrariness of  $v$ , one can set

$$v = -\frac{s - \nabla f(x_0)}{\|s - \nabla f(x_0)\|},$$

leading to  $s = \nabla f(x_0)$ .



## Subdifferential properties

- If  $x_0 \in \text{ri}(S)$ , then  $\partial f(x_0)$  is a convex compact set. which implies:
- The convex function  $f(x)$  is differentiable at the point  $x_0 \Rightarrow \partial f(x_0) = \{\nabla f(x_0)\}$ .
- If  $\partial f(x_0) \neq \emptyset \quad \forall x_0 \in S$ , then  $f(x)$  is convex on  $S$ .

### i Subdifferential of a differentiable function

Let  $f : S \rightarrow \mathbb{R}$  be a function defined on the set  $S$  in a Euclidean space  $\mathbb{R}^n$ . If  $x_0 \in \text{ri}(S)$  and  $f$  is differentiable at  $x_0$ , then either  $\partial f(x_0) = \emptyset$  or  $\partial f(x_0) = \{\nabla f(x_0)\}$ . Moreover, if the function  $f$  is convex, the first scenario is impossible.

### Proof

1. Assume, that  $s \in \partial f(x_0)$  for some  $s \in \mathbb{R}^n$  distinct from  $\nabla f(x_0)$ . Let  $v \in \mathbb{R}^n$  be a unit vector. Because  $x_0$  is an interior point of  $S$ , there exists  $\delta > 0$  such that  $x_0 + tv \in S$  for all  $0 < t < \delta$ . By the definition of the subgradient, we have

$$f(x_0 + tv) \geq f(x_0) + t\langle s, v \rangle$$

$$\frac{f(x_0 + tv) - f(x_0)}{t} \geq \langle s, v \rangle$$

for all  $0 < t < \delta$ . Taking the limit as  $t$  approaches 0 and using the definition of the gradient, we get:

$$\langle \nabla f(x_0), v \rangle = \lim_{t \rightarrow 0; 0 < t < \delta} \frac{f(x_0 + tv) - f(x_0)}{t} \geq \langle s, v \rangle$$

2. From this,  $\langle s - \nabla f(x_0), v \rangle \geq 0$ . Due to the arbitrariness of  $v$ , one can set

$$v = -\frac{s - \nabla f(x_0)}{\|s - \nabla f(x_0)\|},$$

leading to  $s = \nabla f(x_0)$ .

3. Furthermore, if the function  $f$  is convex, then according to the differential condition of convexity  $f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$  for all  $x \in S$ . But by definition, this means  $\nabla f(x_0) \in \partial f(x_0)$ .

# Subdifferential calculus

**i** Moreau - Rockafellar theorem (subdifferential of a linear combination)

Let  $f_i(x)$  be convex functions on convex sets  $S_i$ ,  $i = \overline{1, n}$ . Then if  $\bigcap_{i=1}^n \text{ri}(S_i) \neq \emptyset$  then the function

$f(x) = \sum_{i=1}^n a_i f_i(x)$ ,  $a_i > 0$  has a subdifferential

$\partial_S f(x)$  on the set  $S = \bigcap_{i=1}^n S_i$  and

$$\partial_S f(x) = \sum_{i=1}^n a_i \partial_{S_i} f_i(x)$$

# Subdifferential calculus

**i** Moreau - Rockafellar theorem (subdifferential of a linear combination)

Let  $f_i(x)$  be convex functions on convex sets  $S_i$ ,  $i = \overline{1, n}$ . Then if  $\bigcap_{i=1}^n \text{ri}(S_i) \neq \emptyset$  then the function

$f(x) = \sum_{i=1}^n a_i f_i(x)$ ,  $a_i > 0$  has a subdifferential

$\partial_S f(x)$  on the set  $S = \bigcap_{i=1}^n S_i$  and

$$\partial_S f(x) = \sum_{i=1}^n a_i \partial_{S_i} f_i(x)$$

**i** Dubovitsky - Milutin theorem (subdifferential of a point-wise maximum)

Let  $f_i(x)$  be convex functions on the open convex set  $S \subseteq \mathbb{R}^n$ ,  $x_0 \in S$ , and the pointwise maximum is defined as  $f(x) = \max_i f_i(x)$ . Then:

$$\partial_S f(x_0) = \text{conv} \left\{ \bigcup_{i \in I(x_0)} \partial_S f_i(x_0) \right\}, \quad I(x) = \{i \in [1, n] \mid f_i(x) = f(x)\}$$

# Subdifferential calculus

- $\partial(\alpha f)(x) = \alpha \partial f(x)$ , for  $\alpha \geq 0$

# Subdifferential calculus

- $\partial(\alpha f)(x) = \alpha \partial f(x)$ , for  $\alpha \geq 0$
- $\partial(\sum f_i)(x) = \sum \partial f_i(x)$ ,  $f_i$  - convex functions

# Subdifferential calculus

- $\partial(\alpha f)(x) = \alpha \partial f(x)$ , for  $\alpha \geq 0$
- $\partial(\sum f_i)(x) = \sum \partial f_i(x)$ ,  $f_i$  - convex functions
- $\partial(f(Ax + b))(x) = A^T \partial f(Ax + b)$ ,  $f$  - convex function

# Subdifferential calculus

- $\partial(\alpha f)(x) = \alpha \partial f(x)$ , for  $\alpha \geq 0$
- $\partial(\sum f_i)(x) = \sum \partial f_i(x)$ ,  $f_i$  - convex functions
- $\partial(f(Ax + b))(x) = A^T \partial f(Ax + b)$ ,  $f$  - convex function
- $z \in \partial f(x)$  if and only if  $x \in \partial f^*(z)$ .

## Subgradient Method



# Algorithm

A vector  $g$  is called the **subgradient** of the function  $f(x) : S \rightarrow \mathbb{R}$  at the point  $x_0$  if  $\forall x \in S$ :

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

# Algorithm

A vector  $g$  is called the **subgradient** of the function  $f(x) : S \rightarrow \mathbb{R}$  at the point  $x_0$  if  $\forall x \in S$ :

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

The idea is very simple: let's replace the gradient  $\nabla f(x_k)$  in the gradient descent algorithm with a subgradient  $g_k$  at point  $x_k$ :

$$x_{k+1} = x_k - \alpha_k g_k,$$

where  $g_k$  is an arbitrary subgradient of the function  $f(x)$  at the point  $x_k$ ,  $g_k \in \partial f(x_k)$

# Algorithm

A vector  $g$  is called the **subgradient** of the function  $f(x) : S \rightarrow \mathbb{R}$  at the point  $x_0$  if  $\forall x \in S$ :

$$f(x) \geq f(x_0) + \langle g, x - x_0 \rangle$$

The idea is very simple: let's replace the gradient  $\nabla f(x_k)$  in the gradient descent algorithm with a subgradient  $g_k$  at point  $x_k$ :

$$x_{k+1} = x_k - \alpha_k g_k,$$

where  $g_k$  is an arbitrary subgradient of the function  $f(x)$  at the point  $x_k$ ,  $g_k \in \partial f(x_k)$

Note that the **subgradient method is not guaranteed to be a descent method**; the negative subgradient need not be a descent direction, or the step size may cause  $f(x_{k+1}) > f(x_k)$ .

That is why we usually track the best value of the objective function

$$f_k^{\text{best}} = \min_{i=1, \dots, k} f(x_i).$$

# Convergence bound

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^* - \alpha_k g_k\|^2 =$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*))\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) \\ 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) \\ 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2\end{aligned}$$



## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*))\end{aligned}$$

$$2\alpha_k (f(x_k) - f(x^*)) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2$$

Let us sum the obtained inequality for  $k = 0, \dots, T-1$ :

$$\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) \leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*))\end{aligned}$$

$$2\alpha_k (f(x_k) - f(x^*)) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2$$

Let us sum the obtained inequality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*))\end{aligned}$$

$$2\alpha_k (f(x_k) - f(x^*)) \leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2$$

Let us sum the obtained inequality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

## Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) \\ 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2\end{aligned}$$

Let us sum the obtained inequality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

- Let's write down how close we came to the optimum  $x^* = \arg \min_{x \in \mathbb{R}^n} f(x) = \arg f^*$  on the last iteration:

# Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) \\ 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2\end{aligned}$$

Let us sum the obtained inequality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

- Let's write down how close we came to the optimum  $x^* = \arg \min_{x \in \mathbb{R}^n} f(x) = \arg f^*$  on the last iteration:
- For a subgradient:  $\langle g_k, x^* - x_k \rangle \leq f(x^*) - f(x_k)$ .

# Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) \\ 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2\end{aligned}$$

Let us sum the obtained inequality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

- Let's write down how close we came to the optimum  $x^* = \arg \min_{x \in \mathbb{R}^n} f(x) = \arg f^*$  on the last iteration:
- For a subgradient:  $\langle g_k, x^* - x_k \rangle \leq f(x^*) - f(x_k)$ .
- We additionally assume that  $\|g_k\|^2 \leq G^2$

# Convergence bound

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) \\ 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2\end{aligned}$$

Let us sum the obtained inequality for  $k = 0, \dots, T-1$ :

$$\begin{aligned}\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) &\leq \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k\|^2 \\ &\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2\end{aligned}$$

- Let's write down how close we came to the optimum  $x^* = \arg \min_{x \in \mathbb{R}^n} f(x) = \arg f^*$  on the last iteration:
- For a subgradient:  $\langle g_k, x^* - x_k \rangle \leq f(x^*) - f(x_k)$ .
- We additionally assume that  $\|g_k\|^2 \leq G^2$
- We use the notation  $R = \|x_0 - x^*\|_2$

## Convergence bound

- Finally, note:

$$\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) \geq \sum_{k=0}^{T-1} 2\alpha_k (f_k^{\text{best}} - f(x^*)) = (f_k^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} 2\alpha_k$$



## Convergence bound

- Finally, note:

$$\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) \geq \sum_{k=0}^{T-1} 2\alpha_k (f_k^{\text{best}} - f(x^*)) = (f_k^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} 2\alpha_k$$

- Which leads to the basic inequality:

$$f_k^{\text{best}} - f(x^*) \leq \frac{R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2}{2 \sum_{k=0}^{T-1} \alpha_k}$$

## Convergence bound

- Finally, note:

$$\sum_{k=0}^{T-1} 2\alpha_k (f(x_k) - f(x^*)) \geq \sum_{k=0}^{T-1} 2\alpha_k (f_k^{\text{best}} - f(x^*)) = (f_k^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} 2\alpha_k$$

- Which leads to the basic inequality:

$$f_k^{\text{best}} - f(x^*) \leq \frac{R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2}{2 \sum_{k=0}^{T-1} \alpha_k}$$

- From this point we can see, that if the stepsize strategy is such that

$$\sum_{k=0}^{T-1} \alpha_k^2 < \infty, \quad \sum_{k=0}^{T-1} \alpha_k = \infty,$$

then the subgradient method converges (step size should be decreasing, but not too fast).

# Different step size strategies



## Different step size strategies



## Convergence bound. Non-smooth convex case. Constant step size

### Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a fixed step size  $\alpha$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{R^2}{2\alpha k} + \frac{\alpha}{2}G^2$$

- Note, that with any constant step size, the first term of the right-hand side is decreasing, but the second term stays constant.

## Convergence bound. Non-smooth convex case. Constant step size

### Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a fixed step size  $\alpha$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{R^2}{2\alpha k} + \frac{\alpha}{2}G^2$$

- Note, that with any constant step size, the first term of the right-hand side is decreasing, but the second term stays constant.
- Some versions of the subgradient method (e.g., diminishing nonsummable step lengths) work when the assumption on  $\|g_k\|_2 \leq G$  doesn't hold; see <sup>1</sup> or <sup>2</sup>.

---

<sup>1</sup>B. Polyak. Introduction to Optimization. Optimization Software, Inc., 1987.

<sup>2</sup>N. Shor. Minimization Methods for Non-differentiable Functions. Springer Series in Computational Mathematics. Springer, 1985.

## Convergence bound. Non-smooth convex case. Constant step size

### i Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a fixed step size  $\alpha$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{R^2}{2\alpha k} + \frac{\alpha}{2} G^2$$

- Note, that with any constant step size, the first term of the right-hand side is decreasing, but the second term stays constant.
- Some versions of the subgradient method (e.g., diminishing nonsummable step lengths) work when the assumption on  $\|g_k\|_2 \leq G$  doesn't hold; see <sup>1</sup> or <sup>2</sup>.
- Let's find the optimal step size  $\alpha$  that minimizes the right-hand side of the inequality.

<sup>1</sup>B. Polyak. Introduction to Optimization. Optimization Software, Inc., 1987.

<sup>2</sup>N. Shor. Minimization Methods for Non-differentiable Functions. Springer Series in Computational Mathematics. Springer, 1985.

## Convergence bound. Non-smooth convex case. Constant step size

### Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a fixed step size  $\alpha = \frac{R}{G} \sqrt{\frac{1}{k}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR}{\sqrt{k}}$$

- This version requires knowledge of the number of iterations in advance, which is not usually practical.



## Convergence bound. Non-smooth convex case. Constant step size

### Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a fixed step size  $\alpha = \frac{R}{G} \sqrt{\frac{1}{k}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR}{\sqrt{k}}$$

- This version requires knowledge of the number of iterations in advance, which is not usually practical.
- It is interesting to mention, that if you want to find the optimal stepsizes for the whole sequence  $\alpha_0, \alpha_1, \dots, \alpha_{k-1}$ , you will get the same result.

## Convergence bound. Non-smooth convex case. Constant step size

### Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a fixed step size  $\alpha = \frac{R}{G} \sqrt{\frac{1}{k}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR}{\sqrt{k}}$$

- This version requires knowledge of the number of iterations in advance, which is not usually practical.
- It is interesting to mention, that if you want to find the optimal stepsizes for the whole sequence  $\alpha_0, \alpha_1, \dots, \alpha_{k-1}$ , you will get the same result.
- Why? Because the right-hand side is convex and **symmetric** function of  $\alpha_0, \alpha_1, \dots, \alpha_{k-1}$ .

## Convergence bound. Non-smooth convex case. Constant step length

### i Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a fixed step length  $\gamma = \alpha_k \|g_k\|_2$ , i.e.  $\alpha_k = \frac{\gamma}{\|g_k\|_2}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR^2}{2\gamma k} + \frac{G\gamma}{2}$$

- Note, that for the subgradient method, we typically can not use the norm of the subgradient as a stopping criterion (imagine  $f(x) = |x|$ ). There are some variants of more advanced stopping criteria, but the convergence is so slow, so typically we just set a maximum number of iterations.

# Convergence bound. Non-smooth convex case. Practical strategy

## Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a diminishing step size strategy  $\alpha_k = \frac{R}{G\sqrt{k+1}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR(2 + \ln k)}{4\sqrt{k+1}}$$

### 1. Bounding sums:

# Convergence bound. Non-smooth convex case. Practical strategy

## i Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a diminishing step size strategy  $\alpha_k = \frac{R}{G\sqrt{k+1}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR(2 + \ln k)}{4\sqrt{k+1}}$$

### 1. Bounding sums:

$$\sum_{k=0}^{T-1} \alpha_k^2 = \frac{R^2}{G^2} \sum_{k=1}^T \frac{1}{k} \leq \frac{R^2}{G^2} (1 + \ln T);$$

# Convergence bound. Non-smooth convex case. Practical strategy

## i Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a diminishing step size strategy  $\alpha_k = \frac{R}{G\sqrt{k+1}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR(2 + \ln k)}{4\sqrt{k+1}}$$

### 1. Bounding sums:

$$\sum_{k=0}^{T-1} \alpha_k^2 = \frac{R^2}{G^2} \sum_{k=1}^T \frac{1}{k} \leq \frac{R^2}{G^2} (1 + \ln T); \quad \sum_{k=0}^{T-1} \alpha_k = \frac{R}{G} \sum_{k=1}^T \frac{1}{\sqrt{k}} \geq \frac{R}{G} \int_1^{T+1} \frac{1}{\sqrt{t}} dt = \frac{2R}{G} (\sqrt{T+1} - 1).$$

## Convergence bound. Non-smooth convex case. Practical strategy

### Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a diminishing step size strategy  $\alpha_k = \frac{R}{G\sqrt{k+1}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR(2 + \ln k)}{4\sqrt{k+1}}$$

#### 1. Bounding sums:

$$\sum_{k=0}^{T-1} \alpha_k^2 = \frac{R^2}{G^2} \sum_{k=1}^T \frac{1}{k} \leq \frac{R^2}{G^2} (1 + \ln T); \quad \sum_{k=0}^{T-1} \alpha_k = \frac{R}{G} \sum_{k=1}^T \frac{1}{\sqrt{k}} \geq \frac{R}{G} \int_1^{T+1} \frac{1}{\sqrt{t}} dt = \frac{2R}{G} (\sqrt{T+1} - 1).$$

#### 2. We drop the last $-1$ in the upper bound above and use the basic inequality:

# Convergence bound. Non-smooth convex case. Practical strategy

## i Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a diminishing step size strategy  $\alpha_k = \frac{R}{G\sqrt{k+1}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR(2 + \ln k)}{4\sqrt{k+1}}$$

### 1. Bounding sums:

$$\sum_{k=0}^{T-1} \alpha_k^2 = \frac{R^2}{G^2} \sum_{k=1}^T \frac{1}{k} \leq \frac{R^2}{G^2} (1 + \ln T); \quad \sum_{k=0}^{T-1} \alpha_k = \frac{R}{G} \sum_{k=1}^T \frac{1}{\sqrt{k}} \geq \frac{R}{G} \int_1^{T+1} \frac{1}{\sqrt{t}} dt = \frac{2R}{G} (\sqrt{T+1} - 1).$$

### 2. We drop the last $-1$ in the upper bound above and use the basic inequality:

$$f_T^{\text{best}} - f(x^*) \leq \frac{R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2}{2 \sum_{k=0}^{T-1} \alpha_k}$$



## Convergence bound. Non-smooth convex case. Practical strategy

### i Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a diminishing step size strategy  $\alpha_k = \frac{R}{G\sqrt{k+1}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR(2 + \ln k)}{4\sqrt{k+1}}$$

#### 1. Bounding sums:

$$\sum_{k=0}^{T-1} \alpha_k^2 = \frac{R^2}{G^2} \sum_{k=1}^T \frac{1}{k} \leq \frac{R^2}{G^2} (1 + \ln T); \quad \sum_{k=0}^{T-1} \alpha_k = \frac{R}{G} \sum_{k=1}^T \frac{1}{\sqrt{k}} \geq \frac{R}{G} \int_1^{T+1} \frac{1}{\sqrt{t}} dt = \frac{2R}{G} (\sqrt{T+1} - 1).$$

#### 2. We drop the last $-1$ in the upper bound above and use the basic inequality:

$$f_T^{\text{best}} - f(x^*) \leq \frac{R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2}{2 \sum_{k=0}^{T-1} \alpha_k} \leq \frac{R^2 + R^2(1 + \ln T)}{4 \frac{R}{G} (\sqrt{T+1})}$$

## Convergence bound. Non-smooth convex case. Practical strategy

### i Theorem

Let  $f$  be a convex  $G$ -Lipschitz function and  $R = \|x_0 - x^*\|_2$ . For a diminishing step size strategy  $\alpha_k = \frac{R}{G\sqrt{k+1}}$ , subgradient method satisfies

$$f_k^{\text{best}} - f(x^*) \leq \frac{GR(2 + \ln k)}{4\sqrt{k+1}}$$

#### 1. Bounding sums:

$$\sum_{k=0}^{T-1} \alpha_k^2 = \frac{R^2}{G^2} \sum_{k=1}^T \frac{1}{k} \leq \frac{R^2}{G^2} (1 + \ln T); \quad \sum_{k=0}^{T-1} \alpha_k = \frac{R}{G} \sum_{k=1}^T \frac{1}{\sqrt{k}} \geq \frac{R}{G} \int_1^{T+1} \frac{1}{\sqrt{t}} dt = \frac{2R}{G} (\sqrt{T+1} - 1).$$

#### 2. We drop the last $-1$ in the upper bound above and use the basic inequality:

$$f_T^{\text{best}} - f(x^*) \leq \frac{R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2}{2 \sum_{k=0}^{T-1} \alpha_k} \leq \frac{R^2 + R^2(1 + \ln T)}{4 \frac{R}{G} (\sqrt{T+1})} = \frac{GR(2 + \ln T)}{4\sqrt{T+1}}$$

## Non-smooth strongly convex case



Non-smooth  
Convex



Non-smooth  
 $\mu$  - strongly convex

## Non-smooth strongly convex case



Non-smooth  
Convex

$$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$$



Non-smooth  
 $\mu$  - strongly convex

$$\mathcal{O}\left(\frac{1}{k}\right)$$

## Non-smooth strongly convex case

### Theorem

Let  $f$  be  $\mu$ -strongly convex on a convex set and  $x, y$  be arbitrary points. Then for any  $g \in \partial f(x)$ ,

$$\langle g, x - y \rangle \geq f(x) - f(y) + \frac{\mu}{2} \|x - y\|^2.$$

1. For any  $\lambda \in [0, 1)$ , by  $\mu$ -strong convexity,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2.$$

## Non-smooth strongly convex case

### i Theorem

Let  $f$  be  $\mu$ -strongly convex on a convex set and  $x, y$  be arbitrary points. Then for any  $g \in \partial f(x)$ ,

$$\langle g, x - y \rangle \geq f(x) - f(y) + \frac{\mu}{2} \|x - y\|^2.$$

1. For any  $\lambda \in [0, 1)$ , by  $\mu$ -strong convexity,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2.$$

2. By the subgradient inequality at  $x$ , we have

$$f(\lambda x + (1 - \lambda)y) \geq f(x) + \langle g, \lambda x + (1 - \lambda)y - x \rangle \quad \rightarrow \quad f(\lambda x + (1 - \lambda)y) \geq f(x) - (1 - \lambda) \langle g, x - y \rangle.$$

## Non-smooth strongly convex case

### i Theorem

Let  $f$  be  $\mu$ -strongly convex on a convex set and  $x, y$  be arbitrary points. Then for any  $g \in \partial f(x)$ ,

$$\langle g, x - y \rangle \geq f(x) - f(y) + \frac{\mu}{2} \|x - y\|^2.$$

1. For any  $\lambda \in [0, 1)$ , by  $\mu$ -strong convexity,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2.$$

2. By the subgradient inequality at  $x$ , we have

$$f(\lambda x + (1 - \lambda)y) \geq f(x) + \langle g, \lambda x + (1 - \lambda)y - x \rangle \rightarrow f(\lambda x + (1 - \lambda)y) \geq f(x) - (1 - \lambda) \langle g, x - y \rangle.$$

3. Thus,

$$f(x) - (1 - \lambda) \langle g, x - y \rangle \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2$$

$$(1 - \lambda)f(x) \leq (1 - \lambda)f(y) + (1 - \lambda) \langle g, x - y \rangle - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2$$

$$f(x) \leq f(y) + \langle g, x - y \rangle - \frac{\mu}{2} \lambda \|x - y\|^2$$

## Non-smooth strongly convex case

### i Theorem

Let  $f$  be  $\mu$ -strongly convex on a convex set and  $x, y$  be arbitrary points. Then for any  $g \in \partial f(x)$ ,

$$\langle g, x - y \rangle \geq f(x) - f(y) + \frac{\mu}{2} \|x - y\|^2.$$

1. For any  $\lambda \in [0, 1)$ , by  $\mu$ -strong convexity,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2.$$

2. By the subgradient inequality at  $x$ , we have

$$f(\lambda x + (1 - \lambda)y) \geq f(x) + \langle g, \lambda x + (1 - \lambda)y - x \rangle \rightarrow f(\lambda x + (1 - \lambda)y) \geq f(x) - (1 - \lambda) \langle g, x - y \rangle.$$

3. Thus,

$$f(x) - (1 - \lambda) \langle g, x - y \rangle \leq \lambda f(x) + (1 - \lambda)f(y) - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2$$

$$(1 - \lambda)f(x) \leq (1 - \lambda)f(y) + (1 - \lambda) \langle g, x - y \rangle - \frac{\mu}{2} \lambda(1 - \lambda) \|x - y\|^2$$

$$f(x) \leq f(y) + \langle g, x - y \rangle - \frac{\mu}{2} \lambda \|x - y\|^2$$

4. Letting  $\lambda \rightarrow 1^-$  gives  $f(x) \leq f(y) + \langle g, x - y \rangle - \frac{\mu}{2} \|x - y\|^2 \rightarrow \langle g, x - y \rangle \geq f(x) - f(y) + \frac{\mu}{2} \|x - y\|^2.$



## Convergence bound. Non-smooth strongly convex case.

### Theorem

Let  $f$  be a  $\mu$ -strongly convex function (possibly non-smooth) with minimizer  $x^*$  and bounded subgradients  $\|g_k\| \leq G$ . Using the step size  $\alpha_k = \frac{2}{\mu(k+1)}$ , the subgradient method guarantees for  $k > 0$  that:

$$f_k^{\text{best}} - f(x^*) \leq \frac{2G^2}{\mu k}.$$

1. We start with the method formulation as before:

## Convergence bound. Non-smooth strongly convex case.

### Theorem

Let  $f$  be a  $\mu$ -strongly convex function (possibly non-smooth) with minimizer  $x^*$  and bounded subgradients  $\|g_k\| \leq G$ . Using the step size  $\alpha_k = \frac{2}{\mu(k+1)}$ , the subgradient method guarantees for  $k > 0$  that:

$$f_k^{\text{best}} - f(x^*) \leq \frac{2G^2}{\mu k}.$$

1. We start with the method formulation as before:

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^* - \alpha_k g_k\|^2 =$$

## Convergence bound. Non-smooth strongly convex case.

### Theorem

Let  $f$  be a  $\mu$ -strongly convex function (possibly non-smooth) with minimizer  $x^*$  and bounded subgradients  $\|g_k\| \leq G$ . Using the step size  $\alpha_k = \frac{2}{\mu(k+1)}$ , the subgradient method guarantees for  $k > 0$  that:

$$f_k^{\text{best}} - f(x^*) \leq \frac{2G^2}{\mu k}.$$

1. We start with the method formulation as before:

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle\end{aligned}$$

## Convergence bound. Non-smooth strongly convex case.

### Theorem

Let  $f$  be a  $\mu$ -strongly convex function (possibly non-smooth) with minimizer  $x^*$  and bounded subgradients  $\|g_k\| \leq G$ . Using the step size  $\alpha_k = \frac{2}{\mu(k+1)}$ , the subgradient method guarantees for  $k > 0$  that:

$$f_k^{\text{best}} - f(x^*) \leq \frac{2G^2}{\mu k}.$$

1. We start with the method formulation as before:

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) - \alpha_k \mu \|x_k - x^*\|^2\end{aligned}$$

## Convergence bound. Non-smooth strongly convex case.

### Theorem

Let  $f$  be a  $\mu$ -strongly convex function (possibly non-smooth) with minimizer  $x^*$  and bounded subgradients  $\|g_k\| \leq G$ . Using the step size  $\alpha_k = \frac{2}{\mu(k+1)}$ , the subgradient method guarantees for  $k > 0$  that:

$$f_k^{\text{best}} - f(x^*) \leq \frac{2G^2}{\mu k}.$$

1. We start with the method formulation as before:

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) - \alpha_k \mu \|x_k - x^*\|^2 \\ &= (1 - \mu\alpha_k) \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*))\end{aligned}$$

## Convergence bound. Non-smooth strongly convex case.

### Theorem

Let  $f$  be a  $\mu$ -strongly convex function (possibly non-smooth) with minimizer  $x^*$  and bounded subgradients  $\|g_k\| \leq G$ . Using the step size  $\alpha_k = \frac{2}{\mu(k+1)}$ , the subgradient method guarantees for  $k > 0$  that:

$$f_k^{\text{best}} - f(x^*) \leq \frac{2G^2}{\mu k}.$$

1. We start with the method formulation as before:

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) - \alpha_k \mu \|x_k - x^*\|^2 \\ &= (1 - \mu\alpha_k) \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) \\ 2\alpha_k (f(x_k) - f(x^*)) &\leq (1 - \mu\alpha_k) \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2\end{aligned}$$

## Convergence bound. Non-smooth strongly convex case.

### Theorem

Let  $f$  be a  $\mu$ -strongly convex function (possibly non-smooth) with minimizer  $x^*$  and bounded subgradients  $\|g_k\| \leq G$ . Using the step size  $\alpha_k = \frac{2}{\mu(k+1)}$ , the subgradient method guarantees for  $k > 0$  that:

$$f_k^{\text{best}} - f(x^*) \leq \frac{2G^2}{\mu k}.$$

1. We start with the method formulation as before:

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \\ &\leq \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) - \alpha_k \mu \|x_k - x^*\|^2 \\ &= (1 - \mu\alpha_k) \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k (f(x_k) - f(x^*)) \\ 2\alpha_k (f(x_k) - f(x^*)) &\leq (1 - \mu\alpha_k) \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 + \alpha_k^2 \|g_k\|^2 \\ f(x_k) - f(x^*) &\leq \frac{1 - \mu\alpha_k}{2\alpha_k} \|x_k - x^*\|^2 - \frac{1}{2\alpha_k} \|x_{k+1} - x^*\|^2 + \frac{\alpha_k}{2} \|g_k\|^2 \end{aligned}$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:



## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

$$\sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq 0 - \frac{\mu(T-1)T}{4} \|x_T - x^*\|^2 + \frac{1}{\mu} \sum_{k=0}^{T-1} \|g_k\|^2$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

$$\sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq 0 - \frac{\mu(T-1)T}{4} \|x_T - x^*\|^2 + \frac{1}{\mu} \sum_{k=0}^{T-1} \|g_k\|^2 \leq \frac{G^2 T}{\mu}$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

$$\sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq 0 - \frac{\mu(T-1)T}{4} \|x_T - x^*\|^2 + \frac{1}{\mu} \sum_{k=0}^{T-1} \|g_k\|^2 \leq \frac{G^2 T}{\mu}$$
$$(f_{T-1}^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} k = \sum_{k=0}^{T-1} k (f_{T-1}^{\text{best}} - f(x^*))$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

$$\sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq 0 - \frac{\mu(T-1)T}{4} \|x_T - x^*\|^2 + \frac{1}{\mu} \sum_{k=0}^{T-1} \|g_k\|^2 \leq \frac{G^2 T}{\mu}$$

$$(f_{T-1}^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} k = \sum_{k=0}^{T-1} k(f_{T-1}^{\text{best}} - f(x^*)) \leq \sum_{k=0}^{T-1} k(f(x_k) - f(x^*))$$



## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

$$\sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq 0 - \frac{\mu(T-1)T}{4} \|x_T - x^*\|^2 + \frac{1}{\mu} \sum_{k=0}^{T-1} \|g_k\|^2 \leq \frac{G^2 T}{\mu}$$

$$(f_{T-1}^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} k = \sum_{k=0}^{T-1} k(f_{T-1}^{\text{best}} - f(x^*)) \leq \sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq \frac{G^2 T}{\mu}$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

$$\sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq 0 - \frac{\mu(T-1)T}{4} \|x_T - x^*\|^2 + \frac{1}{\mu} \sum_{k=0}^{T-1} \|g_k\|^2 \leq \frac{G^2 T}{\mu}$$

$$(f_{T-1}^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} k = \sum_{k=0}^{T-1} k(f_{T-1}^{\text{best}} - f(x^*)) \leq \sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq \frac{G^2 T}{\mu}$$

$$f_{T-1}^{\text{best}} - f(x^*) \leq \frac{G^2 T}{\mu \sum_{k=0}^{T-1} k}$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

$$\sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq 0 - \frac{\mu(T-1)T}{4} \|x_T - x^*\|^2 + \frac{1}{\mu} \sum_{k=0}^{T-1} \|g_k\|^2 \leq \frac{G^2 T}{\mu}$$

$$(f_{T-1}^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} k = \sum_{k=0}^{T-1} k(f_{T-1}^{\text{best}} - f(x^*)) \leq \sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq \frac{G^2 T}{\mu}$$

$$f_{T-1}^{\text{best}} - f(x^*) \leq \frac{G^2 T}{\mu \sum_{k=0}^{T-1} k} = \frac{2G^2 T}{\mu T(T-1)}$$

## Convergence bound. Non-smooth strongly convex case. Proof

2. Substitute the step size  $\alpha_k = \frac{2}{\mu(k+1)}$  into the inequality:

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu(k+1)} \|g_k\|^2$$

$$f(x_k) - f(x^*) \leq \frac{\mu(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu k} \|g_k\|^2$$

$$k(f(x_k) - f(x^*)) \leq \frac{\mu k(k-1)}{4} \|x_k - x^*\|^2 - \frac{\mu k(k+1)}{4} \|x_{k+1} - x^*\|^2 + \frac{1}{\mu} \|g_k\|^2$$

3. Summing up the inequalities for all  $k = 0, 1, \dots, T-1$ , we get:

$$\sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq 0 - \frac{\mu(T-1)T}{4} \|x_T - x^*\|^2 + \frac{1}{\mu} \sum_{k=0}^{T-1} \|g_k\|^2 \leq \frac{G^2 T}{\mu}$$

$$(f_{T-1}^{\text{best}} - f(x^*)) \sum_{k=0}^{T-1} k = \sum_{k=0}^{T-1} k(f_{T-1}^{\text{best}} - f(x^*)) \leq \sum_{k=0}^{T-1} k(f(x_k) - f(x^*)) \leq \frac{G^2 T}{\mu}$$

$$f_{T-1}^{\text{best}} - f(x^*) \leq \frac{G^2 T}{\mu \sum_{k=0}^{T-1} k} = \frac{2G^2 T}{\mu T(T-1)} \quad f_k^{\text{best}} - f(x^*) \leq \frac{2G^2}{\mu k}.$$

## Summary. Subgradient method

Problem Type	Stepsize Rule	Convergence Rate	Iteration Complexity
Convex & Lipschitz problems	$\alpha \sim \frac{1}{\sqrt{k}}$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$
Strongly convex & Lipschitz problems	$\alpha \sim \frac{1}{k}$	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathcal{O}\left(\frac{1}{\varepsilon}\right)$

## Numerical experiments

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
m=1000, n=100,  $\lambda=0$ ,  $\mu=0$ ,  $L=10$ . Optimal sparsity: 0.0e+00



Figure 6: Smooth convex case. Sublinear convergence, no convergence in domain

## Numerical experiments

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
m=1000, n=100,  $\lambda=0.1$ ,  $\mu=0$ ,  $L=10$ . Optimal sparsity: 1.0e-02



Figure 7: Non-smooth convex case. Small  $\lambda$  value imposes non-smoothness. No convergence with constant step size

## Numerical experiments

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
 $m=1000, n=100, \lambda=1, \mu=0, L=10$ . Optimal sparsity:  $7.0 \times 10^{-2}$



Figure 8: Non-smooth convex case. Larger  $\lambda$  value reveals non-monotonicity of  $f(x_k)$ . One can see that a smaller constant step size leads to a lower stationary level.



## Numerical experiments

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
 $m=100$ ,  $n=100$ ,  $\lambda=1$ ,  $\mu=0$ ,  $L=10$ . Optimal sparsity:  $2.3 \times 10^{-1}$

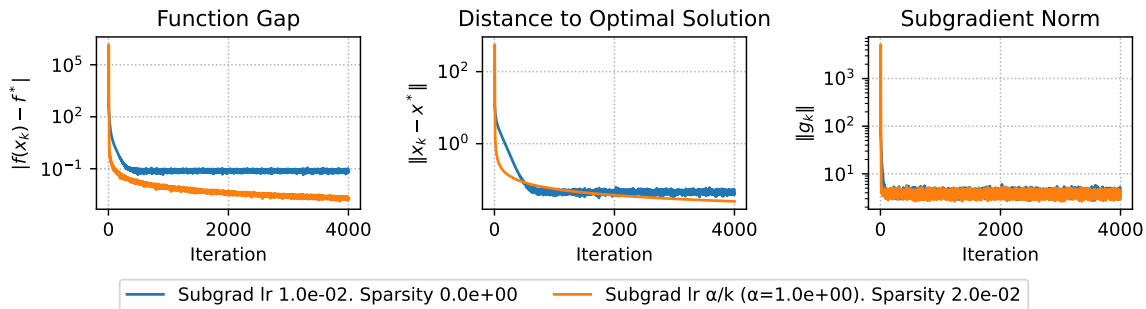


Figure 9: Non-smooth convex case. Diminishing step size leads to the convergence for the  $f_k^{\text{best}}$

## Numerical experiments

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
 $m=100, n=100, \lambda=1, \mu=0, L=10$ . Optimal sparsity:  $2.3e-01$



Figure 10: Non-smooth convex case.  $\frac{\alpha_0}{\sqrt{k}}$  step size leads to the convergence for the  $f_k^{\text{best}}$

## Numerical experiments

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
 $m=100, n=100, \lambda=1, \mu=0, L=10$ . Optimal sparsity:  $2.3e-01$



Figure 11: Non-smooth convex case.  $\frac{\alpha_0}{\sqrt{k}}$  step size leads to the convergence for the  $f_k^{\text{best}}$

## Numerical experiments

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
 $m=100, n=100, \lambda=1, \mu=1, L=10$ . Optimal sparsity:  $2.0 \times 10^{-1}$



Figure 12: Non-smooth strongly convex case.  $\frac{\alpha_0}{k}$  step size leads to the convergence for the  $f_k^{\text{best}}$

## Numerical experiments

$$f(x) = \frac{1}{2m} \|Ax - b\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A \in \mathbb{R}^{m \times n}, \quad \lambda \left( \frac{1}{m} A^T A \right) \in [\mu; L].$$

Linear Least Squares with  $\ell_1$  Regularization (LASSO).  
m=100, n=100,  $\lambda=1$ ,  $\mu=1$ , L=10. Optimal sparsity: 2.0e-01



Figure 13: Non-smooth strongly convex case.  $\frac{\alpha_0}{\sqrt{k}}$  step size works worse

## Numerical experiments

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i(A_i x))) + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A_i \in \mathbb{R}^n, \quad b_i \in \{-1, 1\}$$

Binary Logistic Regression with  $\ell_1$  Regularization.  
 $m=300$ ,  $n=50$ ,  $\lambda=0.1$ . Optimal sparsity:  $8.6e-01$



Figure 14: Logistic regression with  $\ell_1$  regularization

## Numerical experiments

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i(A_i x))) + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A_i \in \mathbb{R}^n, \quad b_i \in \{-1, 1\}$$

Binary Logistic Regression with  $\ell_1$  Regularization.  
 $m=300$ ,  $n=50$ ,  $\lambda=0.1$ . Optimal sparsity:  $8.6e-01$



Figure 15: Logistic regression with  $\ell_1$  regularization

## Numerical experiments

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i(A_i x))) + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A_i \in \mathbb{R}^n, \quad b_i \in \{-1, 1\}$$

Binary Logistic Regression with  $\ell_1$  Regularization.  
 $m=300$ ,  $n=50$ ,  $\lambda=0.25$ . Optimal sparsity:  $9.6e-01$

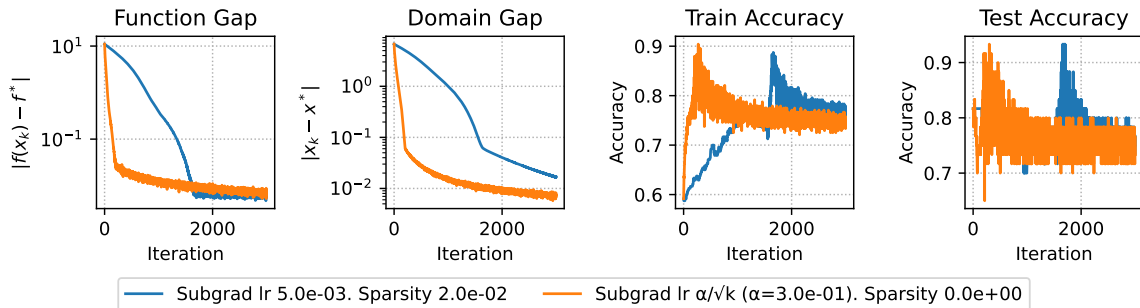


Figure 16: Logistic regression with  $\ell_1$  regularization



## Numerical experiments

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i(A_i x))) + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A_i \in \mathbb{R}^n, \quad b_i \in \{-1, 1\}$$

Binary Logistic Regression with  $\ell_1$  Regularization.  
 $m=300$ ,  $n=50$ ,  $\lambda=0.25$ . Optimal sparsity:  $9.6e-01$



Figure 17: Logistic regression with  $\ell_1$  regularization

## Numerical experiments

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i(A_i x))) + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A_i \in \mathbb{R}^n, \quad b_i \in \{-1, 1\}$$

Binary Logistic Regression with  $\ell_1$  Regularization.  
m=300, n=50,  $\lambda=0.27$ . Optimal sparsity: 1.0e+00



Figure 18: Logistic regression with  $\ell_1$  regularization

## Numerical experiments

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log(1 + \exp(-b_i(A_i x))) + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^n}, \quad A_i \in \mathbb{R}^n, \quad b_i \in \{-1, 1\}$$

Binary Logistic Regression with  $\ell_1$  Regularization.  
m=300, n=50,  $\lambda=0.27$ . Optimal sparsity: 1.0e+00



Figure 19: Logistic regression with  $\ell_1$  regularization

## Lower bounds

## Lower bounds

convex (non-smooth) <sup>3</sup>	smooth (non-convex) <sup>4</sup>	smooth & convex <sup>5</sup>	smooth & strongly convex (or PL) <sup>1</sup>
$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\frac{1}{k^2}\right)$	$\mathcal{O}\left(\frac{1}{k^2}\right)$	$\mathcal{O}\left(\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right)$
$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right)$	$k_\varepsilon \sim \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right)$	$k_\varepsilon \sim \mathcal{O}\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$

<sup>3</sup>Nesterov, Lectures on Convex Optimization

<sup>4</sup>Carmon, Duchi, Hinder, Sidford, 2017

<sup>5</sup>Nemirovski, Yudin, 1979

# Black box iteration

The iteration of gradient descent:

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) \\&= x^{k-1} - \alpha^{k-1} \nabla f(x^{k-1}) - \alpha^k \nabla f(x^k) \\&\vdots \\&= x^0 - \sum_{i=0}^k \alpha^{k-i} \nabla f(x^{k-i})\end{aligned}$$

# Black box iteration

The iteration of gradient descent:

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) \\&= x^{k-1} - \alpha^{k-1} \nabla f(x^{k-1}) - \alpha^k \nabla f(x^k) \\&\vdots \\&= x^0 - \sum_{i=0}^k \alpha^{k-i} \nabla f(x^{k-i})\end{aligned}$$

Consider a family of first-order methods, where

$$\begin{aligned}x^{k+1} &\in x^0 + \text{span} \{ \nabla f(x^0), \nabla f(x^1), \dots, \nabla f(x^k) \} && f - \text{smooth} \\x^{k+1} &\in x^0 + \text{span} \{ g_0, g_1, \dots, g_k \}, \text{ where } g_i \in \partial f(x^i) && f - \text{non-smooth}\end{aligned} \tag{1}$$

## Black box iteration

The iteration of gradient descent:

$$\begin{aligned}x^{k+1} &= x^k - \alpha^k \nabla f(x^k) \\&= x^{k-1} - \alpha^{k-1} \nabla f(x^{k-1}) - \alpha^k \nabla f(x^k) \\&\vdots \\&= x^0 - \sum_{i=0}^k \alpha^{k-i} \nabla f(x^{k-i})\end{aligned}$$

Consider a family of first-order methods, where

$$\begin{aligned}x^{k+1} &\in x^0 + \text{span} \{ \nabla f(x^0), \nabla f(x^1), \dots, \nabla f(x^k) \} && f - \text{smooth} \\x^{k+1} &\in x^0 + \text{span} \{ g_0, g_1, \dots, g_k \}, \text{ where } g_i \in \partial f(x^i) && f - \text{non-smooth}\end{aligned} \tag{1}$$

To construct a lower bound, we need to find a function  $f$  from the corresponding class such that any method from the family 1 will work at least as slowly as the lower bound.



## Non-smooth convex case

### i Theorem

There exists a function  $f$  that is  $G$ -Lipschitz and convex such that any method 1 satisfies

$$\min_{i \in [1, k]} f(x^i) - \min_{x \in \mathbb{B}(R)} f(x) \geq \frac{GR}{2(1 + \sqrt{k})}$$

for  $R > 0$  and  $k \leq n$ , where  $n$  is the dimension of the problem.

## Non-smooth convex case

### Theorem

There exists a function  $f$  that is  $G$ -Lipschitz and convex such that any method 1 satisfies

$$\min_{i \in [1, k]} f(x^i) - \min_{x \in \mathbb{B}(R)} f(x) \geq \frac{GR}{2(1 + \sqrt{k})}$$

for  $R > 0$  and  $k \leq n$ , where  $n$  is the dimension of the problem.

**Proof idea:** build such a function  $f$  that, for any method 1, we have

$$\text{span}\{g_0, g_1, \dots, g_k\} \subset \text{span}\{e_1, e_2, \dots, e_i\}$$

where  $e_i$  is the  $i$ -th standard basis vector. At iteration  $k \leq n$ , there are at least  $n - k$  coordinate of  $x$  are 0. This helps us to derive a bound on the error.

## Non-smooth case (proof)

Consider the function:

$$f(x) = \beta \max_{i \in [1, k]} x[i] + \frac{\alpha}{2} \|x\|_2^2,$$

where  $\alpha, \beta \in \mathbb{R}$  are parameters, and  $x[1 : k]$  denotes the first  $k$  components of  $x$ .

## Non-smooth case (proof)

Consider the function:

$$f(x) = \beta \max_{i \in [1, k]} x[i] + \frac{\alpha}{2} \|x\|_2^2,$$

where  $\alpha, \beta \in \mathbb{R}$  are parameters, and  $x[1 : k]$  denotes the first  $k$  components of  $x$ .

### Key Properties:

- The function  $f(x)$  is  $\alpha$ -strongly convex due to the quadratic term  $\frac{\alpha}{2} \|x\|_2^2$ .

## Non-smooth case (proof)

Consider the function:

$$f(x) = \beta \max_{i \in [1, k]} x[i] + \frac{\alpha}{2} \|x\|_2^2,$$

where  $\alpha, \beta \in \mathbb{R}$  are parameters, and  $x[1 : k]$  denotes the first  $k$  components of  $x$ .

### Key Properties:

- The function  $f(x)$  is  $\alpha$ -strongly convex due to the quadratic term  $\frac{\alpha}{2} \|x\|_2^2$ .
- The function is non-smooth because the first term introduces a non-differentiable point at the maximum coordinate of  $x$ .

## Non-smooth case (proof)

Consider the function:

$$f(x) = \beta \max_{i \in [1, k]} x[i] + \frac{\alpha}{2} \|x\|_2^2,$$

where  $\alpha, \beta \in \mathbb{R}$  are parameters, and  $x[1 : k]$  denotes the first  $k$  components of  $x$ .

### Key Properties:

- The function  $f(x)$  is  $\alpha$ -strongly convex due to the quadratic term  $\frac{\alpha}{2} \|x\|_2^2$ .
- The function is non-smooth because the first term introduces a non-differentiable point at the maximum coordinate of  $x$ .

## Non-smooth case (proof)

Consider the function:

$$f(x) = \beta \max_{i \in [1, k]} x[i] + \frac{\alpha}{2} \|x\|_2^2,$$

where  $\alpha, \beta \in \mathbb{R}$  are parameters, and  $x[1 : k]$  denotes the first  $k$  components of  $x$ .

### Key Properties:

- The function  $f(x)$  is  $\alpha$ -strongly convex due to the quadratic term  $\frac{\alpha}{2} \|x\|_2^2$ .
- The function is non-smooth because the first term introduces a non-differentiable point at the maximum coordinate of  $x$ .

Consider the subdifferential of  $f(x)$  at  $x$ :

$$\begin{aligned} \partial f(x) &= \partial \left( \beta \max_{i \in [1, k]} x[i] \right) + \partial \left( \frac{\alpha}{2} \|x\|_2^2 \right) \\ &= \beta \partial \left( \max_{i \in [1, k]} x[i] \right) + \alpha x \\ &= \beta \text{conv} \left\{ e_i \mid i : x[i] = \max_j x[j] \right\} + \alpha x \end{aligned}$$

## Non-smooth case (proof)

Consider the function:

$$f(x) = \beta \max_{i \in [1, k]} x[i] + \frac{\alpha}{2} \|x\|_2^2,$$

where  $\alpha, \beta \in \mathbb{R}$  are parameters, and  $x[1 : k]$  denotes the first  $k$  components of  $x$ .

### Key Properties:

- The function  $f(x)$  is  $\alpha$ -strongly convex due to the quadratic term  $\frac{\alpha}{2} \|x\|_2^2$ .
- The function is non-smooth because the first term introduces a non-differentiable point at the maximum coordinate of  $x$ .

Consider the subdifferential of  $f(x)$  at  $x$ :

$$\begin{aligned} \partial f(x) &= \partial \left( \beta \max_{i \in [1, k]} x[i] \right) + \partial \left( \frac{\alpha}{2} \|x\|_2^2 \right) \\ &= \beta \partial \left( \max_{i \in [1, k]} x[i] \right) + \alpha x \\ &= \beta \text{conv} \left\{ e_i \mid i : x[i] = \max_j x[j] \right\} + \alpha x \end{aligned}$$

It is easy to see, that if  $g \in \partial f(x)$  and  $\|x\| \leq R$ , then

$$\|g\| \leq \alpha R + \beta$$

Thus,  $f$  is  $\alpha R + \beta$ -Lipschitz on  $B(R)$ .



## Non-smooth case (proof)

Next, we describe the first-order oracle for this function. When queried for a subgradient at a point  $x$ , the oracle returns

$$\alpha x + \gamma e_i,$$

where  $i$  is the *first* coordinate for which  $x[i] = \max_{1 \leq j \leq k} x[j]$ .

- We ensure that  $\|x^0\| \leq R$  by starting from  $x^0 = 0$ .

## Non-smooth case (proof)

Next, we describe the first-order oracle for this function. When queried for a subgradient at a point  $x$ , the oracle returns

$$\alpha x + \gamma e_i,$$

where  $i$  is the *first* coordinate for with  $x[i] = \max_{1 \leq j \leq k} x[j]$ .

- We ensure that  $\|x^0\| \leq R$  by starting from  $x^0 = 0$ .
- When the oracle is queried at  $x^0 = 0$ , it returns  $e_1$ . Consequently,  $x^1$  must lie on the line generated by  $e_1$ .

## Non-smooth case (proof)

Next, we describe the first-order oracle for this function. When queried for a subgradient at a point  $x$ , the oracle returns

$$\alpha x + \gamma e_i,$$

where  $i$  is the *first* coordinate for with  $x[i] = \max_{1 \leq j \leq k} x[j]$ .

- We ensure that  $\|x^0\| \leq R$  by starting from  $x^0 = 0$ .
- When the oracle is queried at  $x^0 = 0$ , it returns  $e_1$ . Consequently,  $x^1$  must lie on the line generated by  $e_1$ .
- By an induction argument, one shows that for all  $i$ , the iterate  $x^i$  lies in the linear span of  $\{e_1, \dots, e_i\}$ . In particular, for  $i \leq k$ , the  $k + 1$ -th coordinate of  $x_i$  is zero and due to the structure of  $f(x)$ :

$$f(x^i) \geq 0.$$

## Non-smooth case (proof)

- It remains to compute the minimal value of  $f$ . Define the point  $y \in \mathbb{R}^n$  as

$$y[i] = -\frac{\beta}{\alpha k} \quad \text{for } 1 \leq i \leq k, \quad y[i] = 0 \quad \text{for } k+1 \leq i \leq n.$$

## Non-smooth case (proof)

- It remains to compute the minimal value of  $f$ . Define the point  $y \in \mathbb{R}^n$  as

$$y[i] = -\frac{\beta}{\alpha k} \quad \text{for } 1 \leq i \leq k, \quad y[i] = 0 \quad \text{for } k+1 \leq i \leq n.$$

- Note, that  $0 \in \partial f(y)$ :

$$\begin{aligned} \partial f(y) &= \alpha y + \beta \text{conv} \left\{ e_i \mid i : y[i] = \max_j y[j] \right\} \\ &= \alpha y + \beta \text{conv} \{ e_i \mid i : y[i] = 0 \} \\ 0 &\in \partial f(y). \end{aligned}$$

## Non-smooth case (proof)

- It remains to compute the minimal value of  $f$ . Define the point  $y \in \mathbb{R}^n$  as

$$y[i] = -\frac{\beta}{\alpha k} \quad \text{for } 1 \leq i \leq k, \quad y[i] = 0 \quad \text{for } k+1 \leq i \leq n.$$

- Note, that  $0 \in \partial f(y)$ :

$$\begin{aligned} \partial f(y) &= \alpha y + \beta \text{conv} \left\{ e_i \mid i : y[i] = \max_j y[j] \right\} \\ &= \alpha y + \beta \text{conv} \{ e_i \mid i : y[i] = 0 \} \\ 0 &\in \partial f(y). \end{aligned}$$

- It follows that the minimum value of  $f = f(y) = f(x^*)$  is

$$f(y) = -\frac{\beta^2}{\alpha k} + \frac{\alpha}{2} \cdot \frac{\beta^2}{\alpha^2 k} = -\frac{\beta^2}{2\alpha k}.$$

## Non-smooth case (proof)

- It remains to compute the minimal value of  $f$ . Define the point  $y \in \mathbb{R}^n$  as

$$y[i] = -\frac{\beta}{\alpha k} \quad \text{for } 1 \leq i \leq k, \quad y[i] = 0 \quad \text{for } k+1 \leq i \leq n.$$

- Note, that  $0 \in \partial f(y)$ :

$$\begin{aligned} \partial f(y) &= \alpha y + \beta \text{conv} \left\{ e_i \mid i : y[i] = \max_j y[j] \right\} \\ &= \alpha y + \beta \text{conv} \{ e_i \mid i : y[i] = 0 \} \\ 0 &\in \partial f(y). \end{aligned}$$

- It follows that the minimum value of  $f = f(y) = f(x^*)$  is

$$f(y) = -\frac{\beta^2}{\alpha k} + \frac{\alpha}{2} \cdot \frac{\beta^2}{\alpha^2 k} = -\frac{\beta^2}{2\alpha k}.$$

- Now we have:

$$f(x^i) - f(x^*) \geq 0 - \left( -\frac{\beta^2}{2\alpha k} \right) \geq \frac{\beta^2}{2\alpha k}.$$

## Non-smooth case (proof)

We have:  $f(x^i) - f(x^*) \geq \frac{\beta^2}{2\alpha k}$ , while we need to prove that  $\min_{i \in [1, k]} f(x^i) - f(x^*) \geq \frac{GR}{2(1+\sqrt{k})}$ .



## Non-smooth case (proof)

We have:  $f(x^i) - f(x^*) \geq \frac{\beta^2}{2\alpha k}$ , while we need to prove that  $\min_{i \in [1, k]} f(x^i) - f(x^*) \geq \frac{GR}{2(1+\sqrt{k})}$ .

### Convex case

$$\alpha = \frac{G}{R} \frac{1}{1 + \sqrt{k}} \quad \beta = \frac{\sqrt{k}}{1 + \sqrt{k}}$$

$$\frac{\beta^2}{2\alpha} = \frac{GRk}{2(1 + \sqrt{k})}$$

Note, in particular, that  $\|y\|_2^2 = \frac{\beta^2}{\alpha^2 k} = R^2$  with these parameters

$$\min_{i \in [1, k]} f(x^i) - f(x^*) \geq \frac{\beta^2}{2\alpha k} = \frac{GR}{2(1 + \sqrt{k})}$$

## Non-smooth case (proof)

We have:  $f(x^i) - f(x^*) \geq \frac{\beta^2}{2\alpha k}$ , while we need to prove that  $\min_{i \in [1, k]} f(x^i) - f(x^*) \geq \frac{GR}{2(1+\sqrt{k})}$ .

### Convex case

$$\alpha = \frac{G}{R} \frac{1}{1 + \sqrt{k}} \quad \beta = \frac{\sqrt{k}}{1 + \sqrt{k}}$$

$$\frac{\beta^2}{2\alpha} = \frac{GRk}{2(1 + \sqrt{k})}$$

Note, in particular, that  $\|y\|_2^2 = \frac{\beta^2}{\alpha^2 k} = R^2$  with these parameters

$$\min_{i \in [1, k]} f(x^i) - f(x^*) \geq \frac{\beta^2}{2\alpha k} = \frac{GR}{2(1 + \sqrt{k})}$$

### Strongly convex case

$$\alpha = \frac{G}{2R} \quad \beta = \frac{G}{2}$$

Note, in particular, that  $\|y\|_2^2 = \frac{\beta^2}{\alpha^2 k} = \frac{G^2}{4\alpha^2 k} = R^2$  with these parameters

$$\min_{i \in [1, k]} f(x^i) - f(x^*) \geq \frac{G^2}{8\alpha k}$$

# Applications

# Linear Least Squares with $l_1$ -regularization

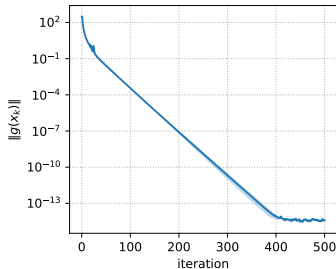
$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

The algorithm will be written as:

$$x_{k+1} = x_k - \alpha_k \left( A^\top (Ax_k - b) + \lambda \text{sign}(x_k) \right),$$

where the signum function is taken element-wise.

LLS with  $l_1$  regularization. 2 runs.  $\lambda = 1$



## Regularized logistic regression

Given  $(x_i, y_i) \in \mathbb{R}^p \times \{0, 1\}$  for  $i = 1, \dots, n$ , the logistic regression function is defined as:

$$f(\theta) = \sum_{i=1}^n \left( -y_i x_i^T \theta + \log(1 + \exp(x_i^T \theta)) \right)$$

This is a smooth and convex function with its gradient given by:

$$\nabla f(\theta) = \sum_{i=1}^n (y_i - s_i(\theta)) x_i$$

where  $s_i(\theta) = \frac{\exp(x_i^T \theta)}{1 + \exp(x_i^T \theta)}$ , for  $i = 1, \dots, n$ . Consider the regularized problem:

$$f(\theta) + \lambda r(\theta) \rightarrow \min_{\theta}$$

where  $r(\theta) = \|\theta\|_2^2$  for the ridge penalty, or  $r(\theta) = \|\theta\|_1$  for the lasso penalty.

# Support Vector Machines

Let  $D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \{\pm 1\}\}$

We need to find  $\theta \in \mathbb{R}^n$  and  $b \in \mathbb{R}$  such that

$$\min_{\theta \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2} \|\theta\|_2^2 + C \sum_{i=1}^m \max[0, 1 - y_i(\theta^\top x_i + b)]$$

# References

- Subgradient Methods Stephen Boyd (with help from Jaehyun Park)