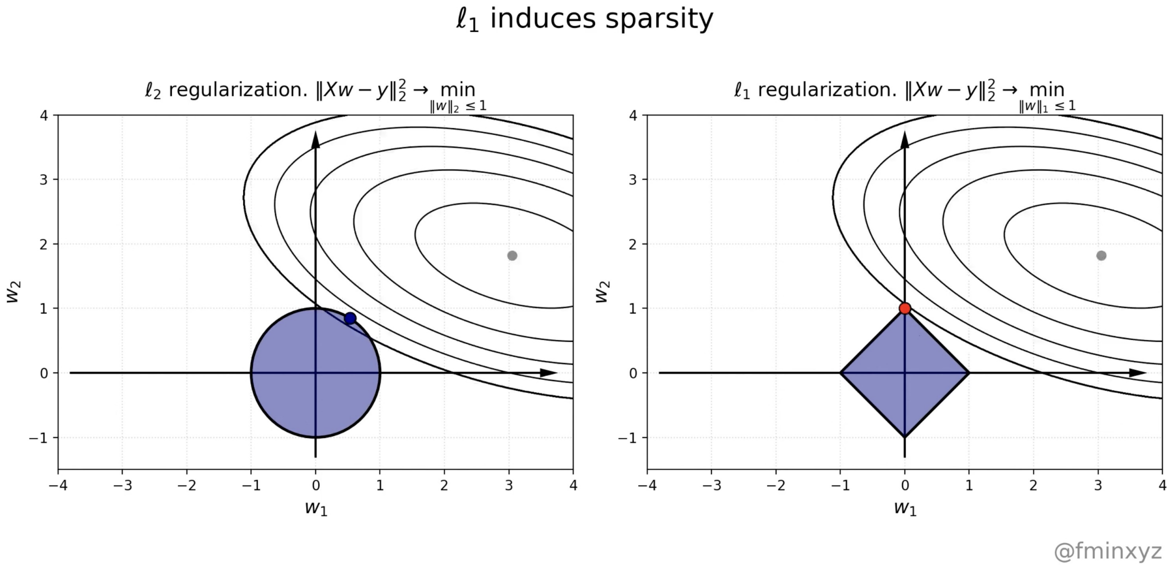
Non-smooth convex optimization. Lower bounds. Subgradient method.

Daniil Merkulov

Optimization methods. MIPT

Non-smooth problems

ℓ_1-regularized linear least squares



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.

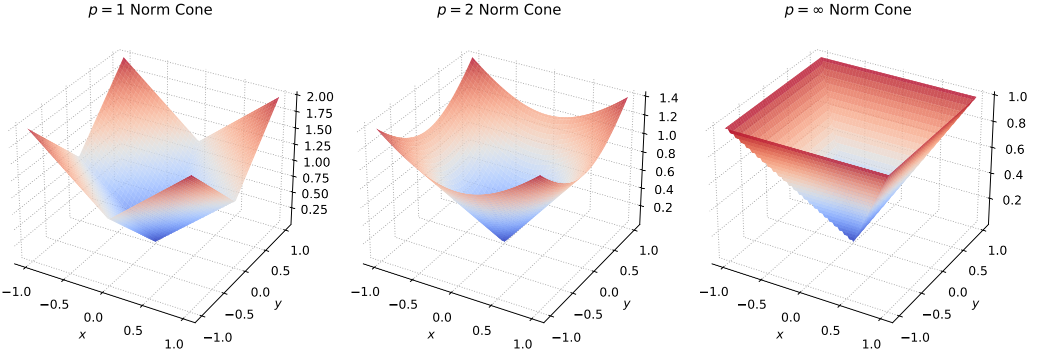


Figure 1: Norm cones for different p - norms are non-smooth

Wolfe’s example

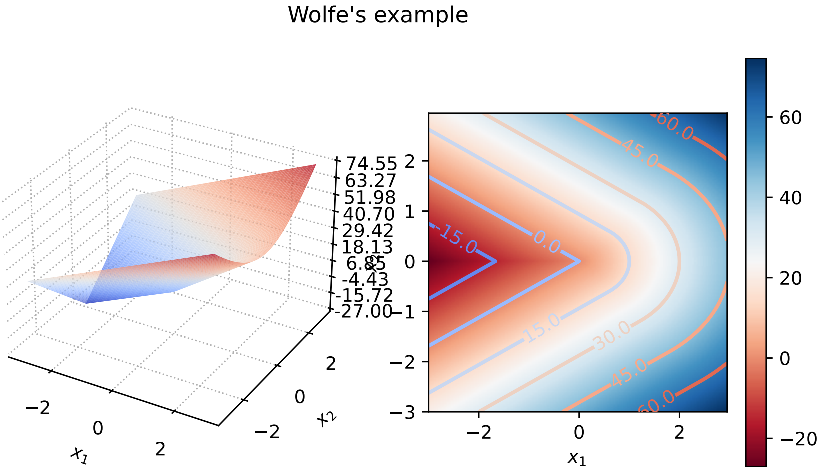


Figure 2: Wolfe’s example. Open in Colab

Subgradient calculus

Convex function linear lower bound

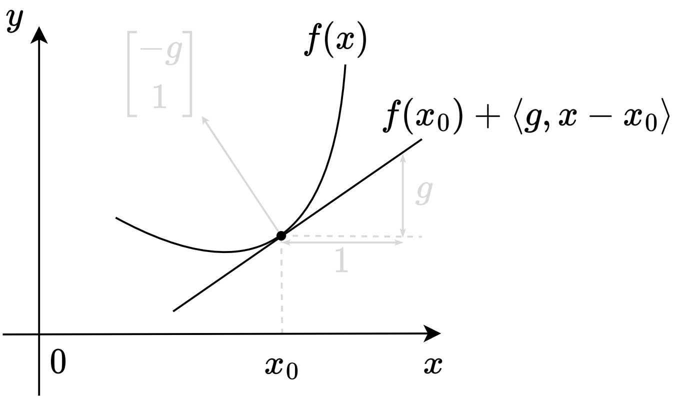


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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:

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

Convex function linear lower bound

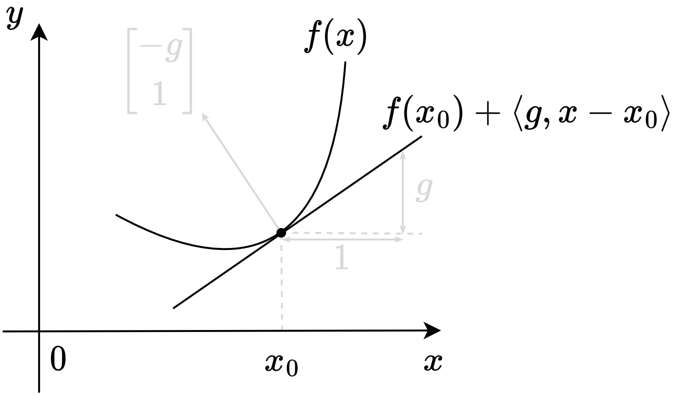


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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 graph of the function is the *global* estimate from below for the function.

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

Convex function linear lower bound

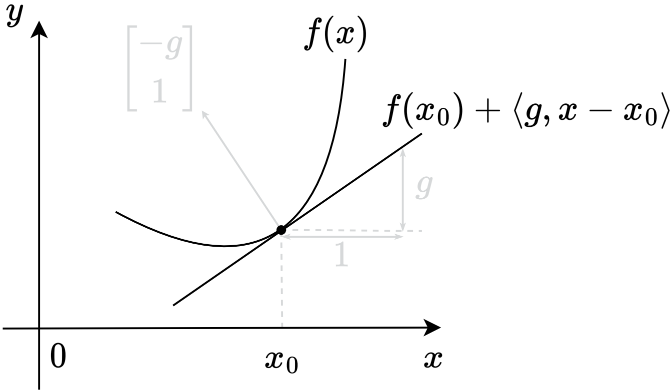


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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 graph of the function 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.

Convex function linear lower bound

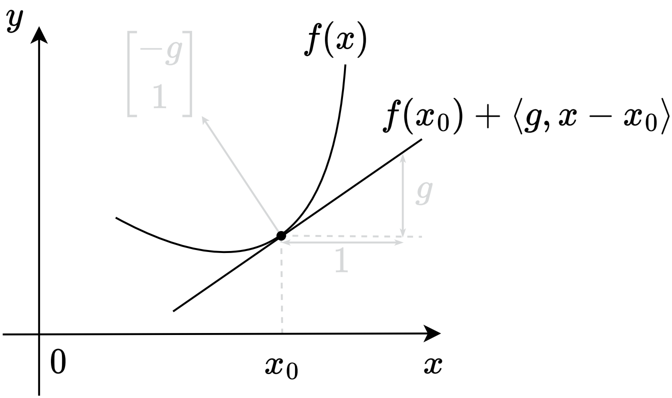


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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 graph of the function 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.

Convex function linear lower bound

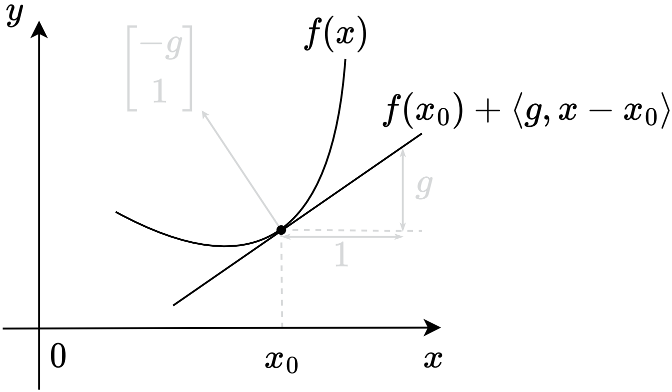


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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 graph of the function 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 wouldn't want to lose such a nice property.

Subgradient and subdifferential

A vector g is called the **subgradient** of a function $f(x) : S \to \mathbb{R}$ 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 $f(x) : S \to \mathbb{R}$ 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 $f(x) : S \to \mathbb{R}$ 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 $\partial f(x_0)$.

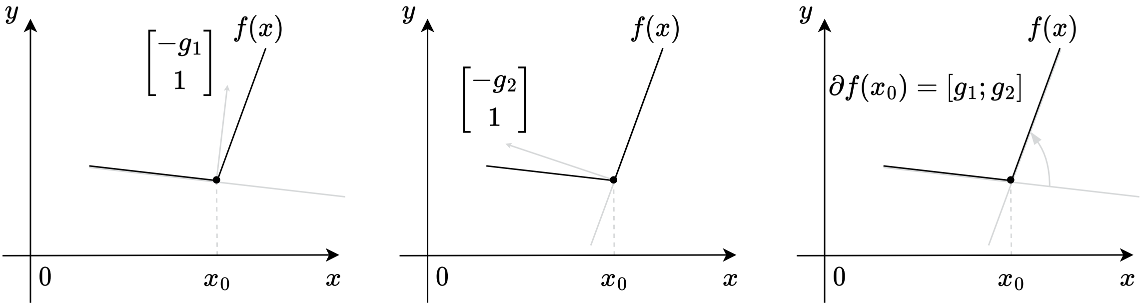


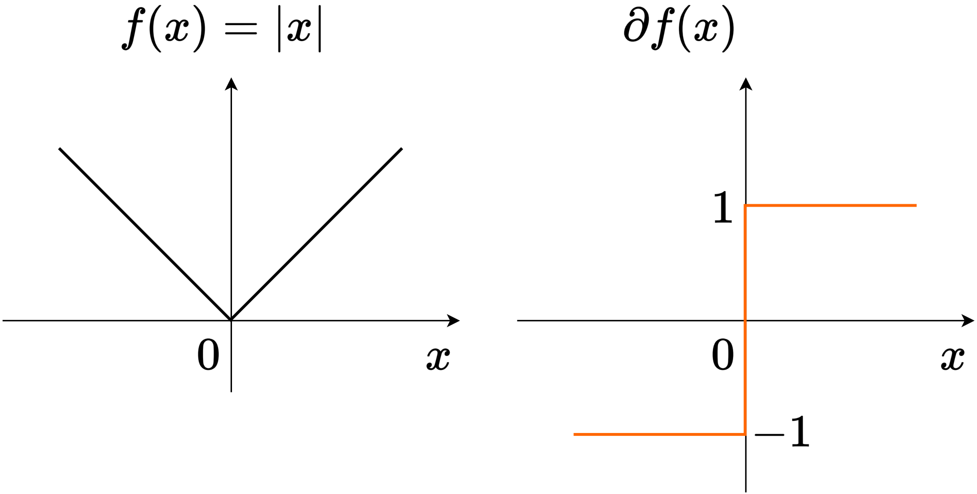
Figure 4: Subdifferential is a set of all possible subgradients

Subgradient and subdifferential

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

Subgradient and subdifferential

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



Subdifferential properties

- If $x_0 \in \mathbf{ri}(S)$, 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 $x_0 \in \mathbf{ri}(S)$, 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 of a differentiable function

Let $f : S \to \mathbb{R}$ be a function defined on the set S in a Euclidean space $\mathbb{R}^n$. If $x_0 \in \mathbf{ri}(S)$ 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 $x_0 \in \mathbf{ri}(S)$, 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 of a differentiable function

Let $f : S \to \mathbb{R}$ be a function defined on the set S in a Euclidean space $\mathbb{R}^n$. If $x_0 \in \mathbf{ri}(S)$ 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 $x_0 \in \mathbf{ri}(S)$, 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 of a differentiable function

Let $f : S \to \mathbb{R}$ be a function defined on the set S in a Euclidean space $\mathbb{R}^n$. If $x_0 \in \mathbf{ri}(S)$ 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 $x_0 \in \mathbf{ri}(S)$, 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 of a differentiable function

Let $f : S \to \mathbb{R}$ be a function defined on the set S in a Euclidean space $\mathbb{R}^n$. If $x_0 \in \mathbf{ri}(S)$ 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

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

which implies:

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

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

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

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 $x_0 \in \mathbf{ri}(S)$, 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 of a differentiable function

Let $f : S \to \mathbb{R}$ be a function defined on the set S in a Euclidean space $\mathbb{R}^n$. If $x_0 \in \mathbf{ri}(S)$ 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

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

which implies:

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

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

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

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

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 $\bigcap\limits_{i=1}^n \mathbf{ri}(S_i) \neq \emptyset$ then the function $f(x) = \sum\limits_{i=1}^n a_i f_i(x),\ a_i > 0$ has a subdifferential $\partial_S f(x)$ on the set $S = \bigcap\limits_{i=1}^n S_i$ and

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

Subdifferential calculus

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 $\bigcap\limits_{i=1}^n \mathbf{ri}(S_i) \neq \emptyset$ then the function $f(x) = \sum\limits_{i=1}^n a_i f_i(x),\ a_i > 0$ has a subdifferential $\partial_S f(x)$ on the set $S = \bigcap\limits_{i=1}^n S_i$ and

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

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 $f(x) = \max\limits_i f_i(x)$. Then:

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

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 $f(x) : S \to \mathbb{R}$ 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 $f(x) : S \to \mathbb{R}$ 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 $f(x) : S \to \mathbb{R}$ 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 a descent method**, i.e. negative subgradient direction is not necessarily a descent direction or the stepsize can be such $f(x_{k+1}) > f(x_k)$.

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

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

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

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

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

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

Convergence bound

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

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

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

Convergence bound

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

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

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

Convergence bound

- 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:

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

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

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

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 $k = 0, \ldots, T-1$:

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

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

Convergence bound

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

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

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

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

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

Convergence bound

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

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

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

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

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

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

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:

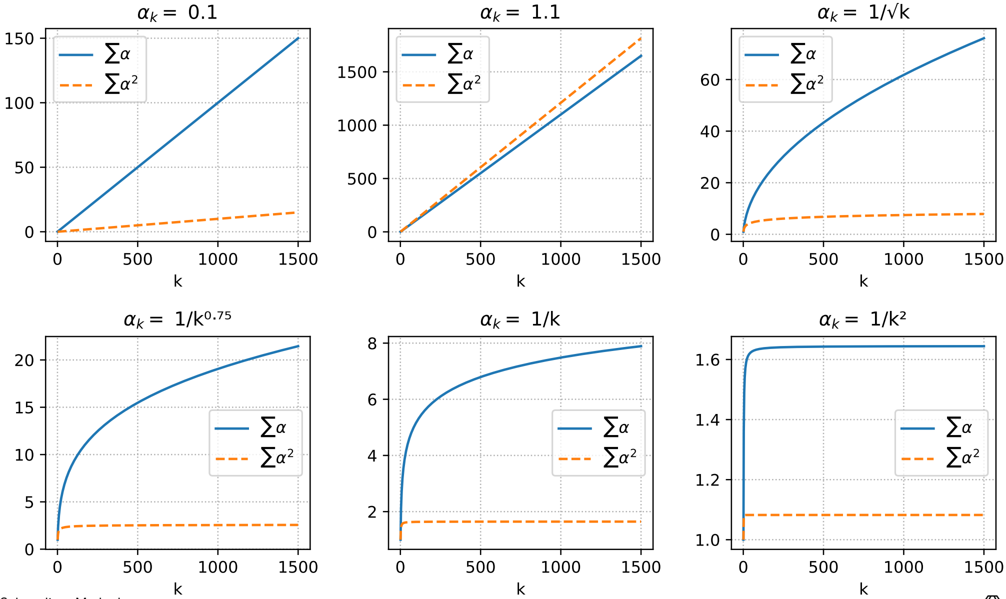
$$\boxed{f_k^{\text{best}} - f(x^*) \leq \frac{R^2 + G^2 \sum\limits_{k=0}^{T-1} \alpha_k^2}{2 \sum\limits_{k=0}^{T-1} \alpha_k}}$$

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

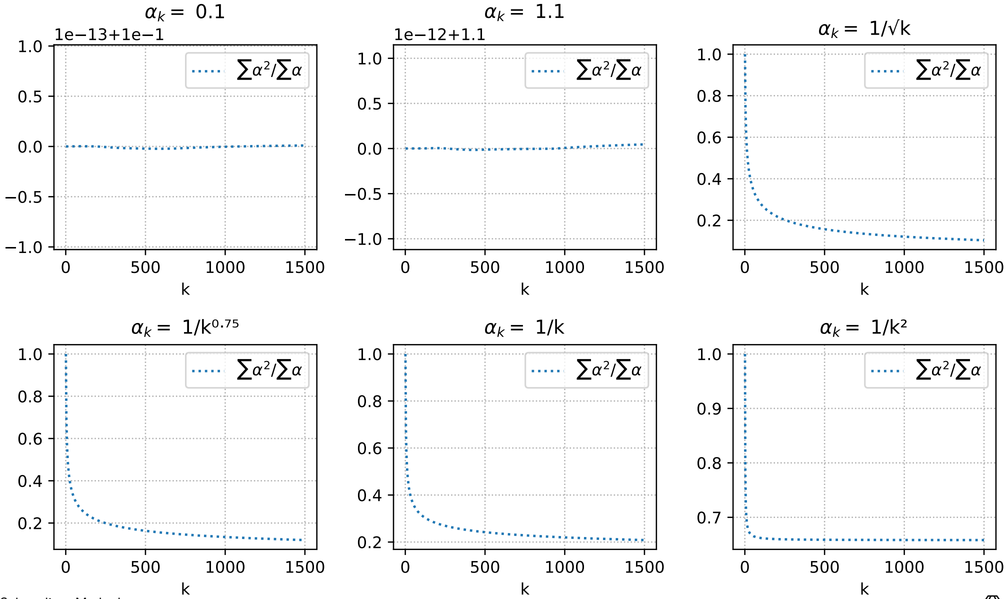
$$\sum_{k=0}^{T-1} \alpha_k^2 \leq \infty, \quad \sum_{k=0}^{T-1} \alpha_k = \infty,$$

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

Different step size strategies



Different step size strategies



Convergence bound. Constant step size

Theorem

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

Theorem

Let f be a convex G-Lipschitz function and $R = \|x_0 - x^*\|_2$. For a fixed step size α, 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 assumption on $\|g_k\|_2 \leq G$ doesn't hold; see [1] or [2].

[1] B. Polyak. Introduction to Optimization. Optimization Software, Inc., 1987.
[2] N. Shor. Minimization Methods for Non-differentiable Functions. Springer Series in Computational Mathematics. Springer, 1985.

Convergence bound. Constant step size

Theorem

Let f be a convex G-Lipschitz function and $R = \|x_0 - x^*\|_2$. For a fixed step size α, 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 assumption on $\|g_k\|_2 \leq G$ doesn't hold; see [1] or [2].
- Let's find the optimal step size α that minimizes the right-hand side of the inequality.

[1] B. Polyak. Introduction to Optimization. Optimization Software, Inc., 1987.
[2] N. Shor. Minimization Methods for Non-differentiable Functions. Springer Series in Computational Mathematics. Springer, 1985.

Convergence bound. 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. 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 $\alpha_0, \alpha_1, \ldots, \alpha_{k-1}$, you will get the same result.

Convergence bound. 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 $\alpha_0, \alpha_1, \ldots, \alpha_{k-1}$, you will get the same result.
- Why? Because the right-hand side is convex and **symmetric** function of $\alpha_0, \alpha_1, \ldots, \alpha_{k-1}$.

Convergence bound. Constant step length

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. 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. 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:

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

Convergence bound. 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:

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

Convergence bound. 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:

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

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

Convergence bound. 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:

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

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

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

Convergence bound. 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:

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

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

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

Convergence bound. 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

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

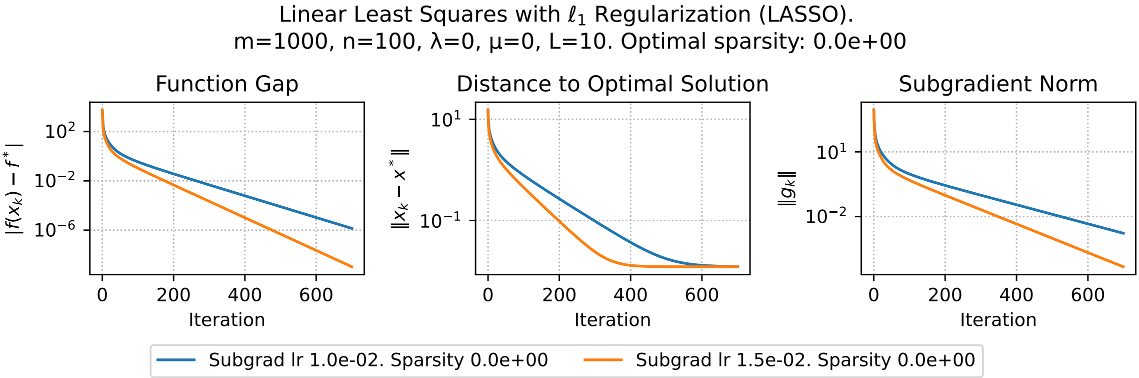
1. Bounding sums:

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

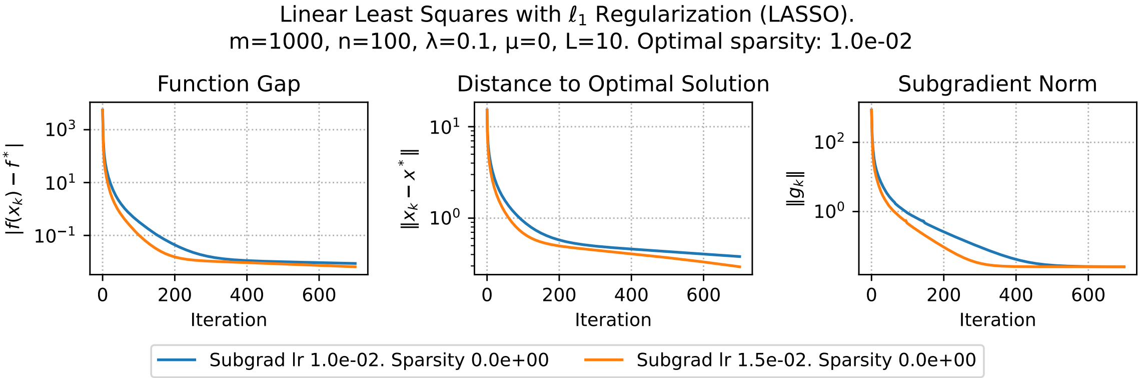
2. We drop the last -1 in the upper bound above and use the basic inequality:

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

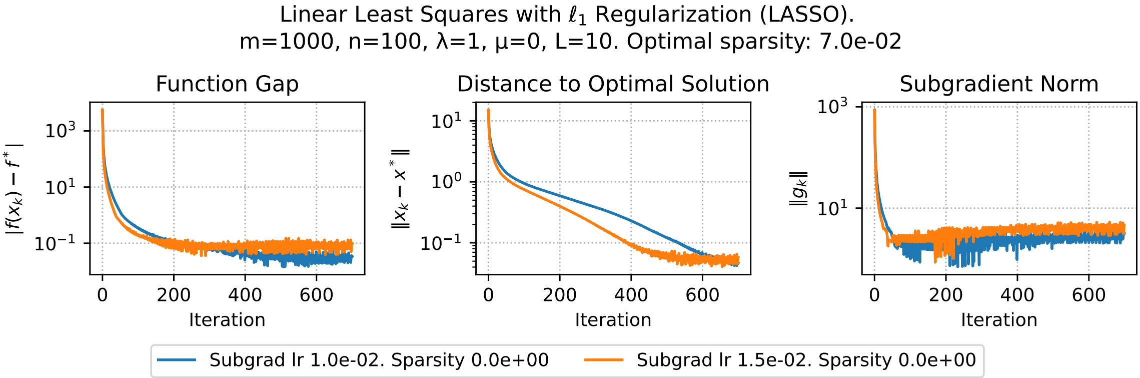
Numerical experiments



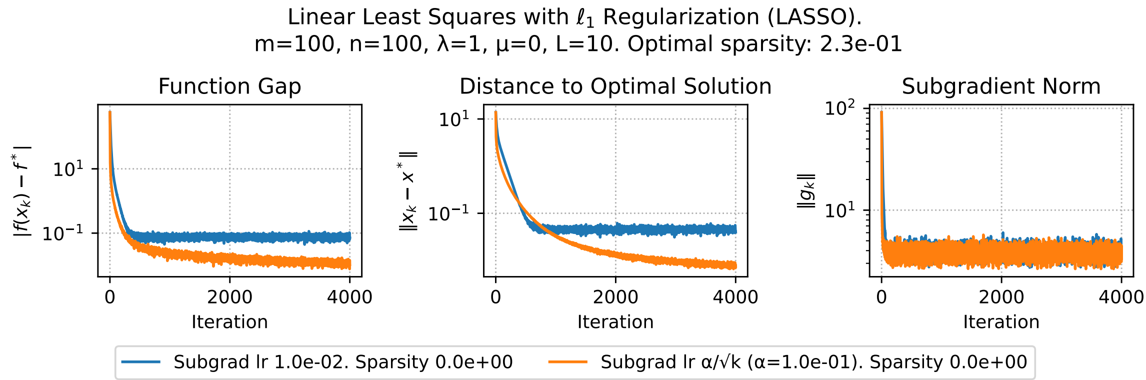
Numerical experiments



Numerical experiments



Numerical experiments



Lower bounds

Lower bounds

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

[3] Nesterov, Lectures on Convex Optimization
[4] Carmon, Duchi, Hinder, Sidford, 2017
[5] 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

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

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

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

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

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.

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

$$\text{span}\left\{g_0, g_1, \ldots, g_k\right\} \subset \text{span}\left\{e_1, e_2, \ldots, e_i\right\}$$

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 α-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 α-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 α-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 α-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 α-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 with $x[i] = \max_{1 \leq j \leq k} x[j]$.

- 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 $\{e_1, \ldots, e_i\}$. 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

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

Non-smooth case (proof)

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

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

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

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

Non-smooth case (proof)

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

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

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

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

- 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

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

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

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

- 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 $\min\limits_{i \in [1,k]} f(x^i) - f(x^*) \geq \frac{GR}{2(1+\sqrt{k})}$.

Non-smooth case (proof)

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

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 $\min\limits_{i \in [1,k]} f(x^i) - f(x^*) \geq \frac{GR}{2(1+\sqrt{k})}$.

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

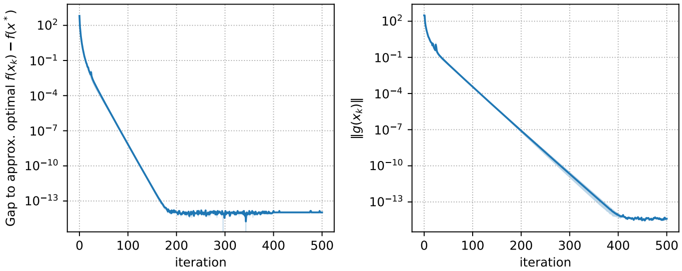
$$
\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1
$$

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 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:

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

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

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

References

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