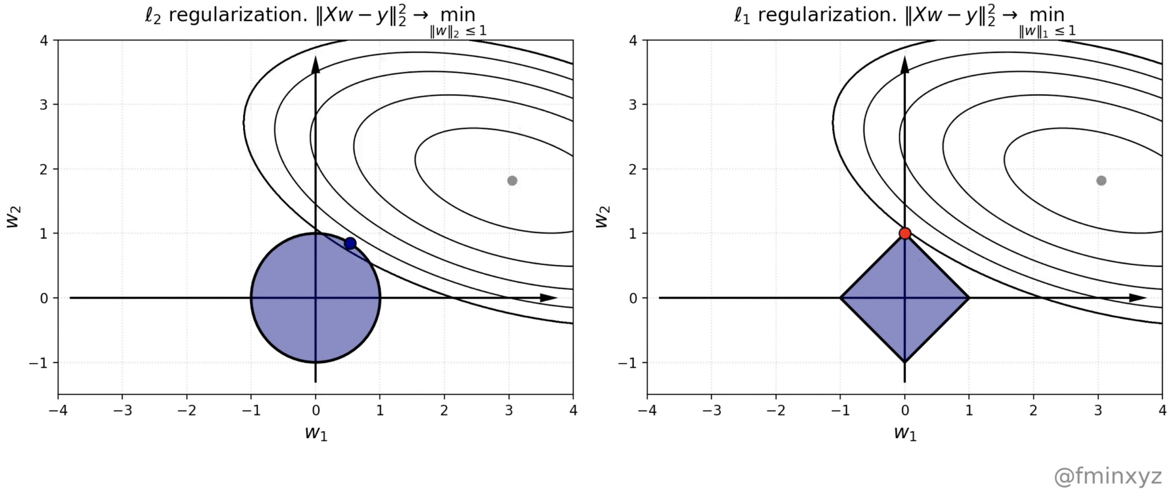
Subgradient Method. Specifics of non-smooth problems.

Daniil Merkulov

Optimization for ML. Faculty of Computer Science. HSE University

ℓ_1-regularized linear least squares

ℓ_1 induces sparsity



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.

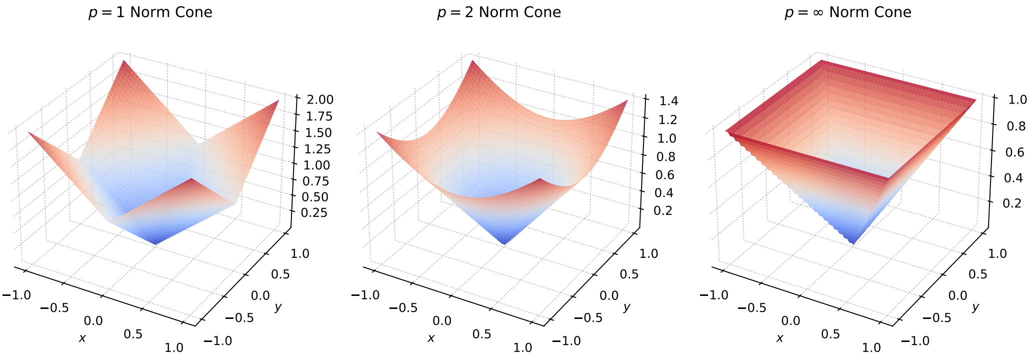


Figure 1: Norm cones for different p - norms are non-smooth

Wolfe's example

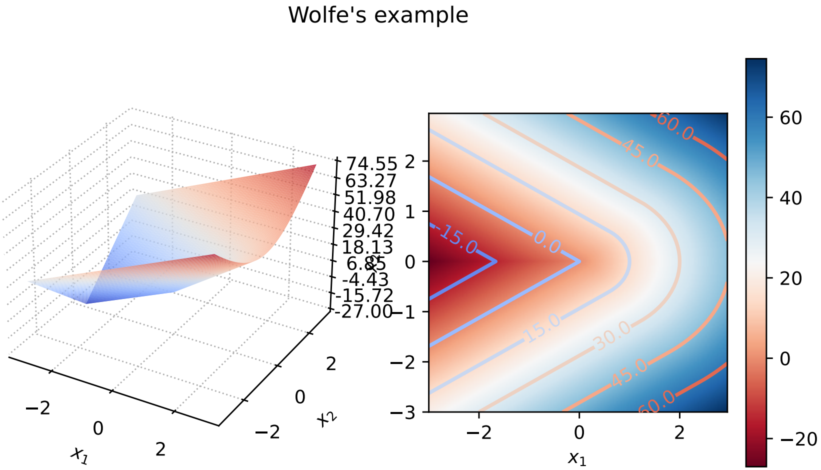


Figure 2: Wolfe's example. Open in Colab

Convex function linear lower bound

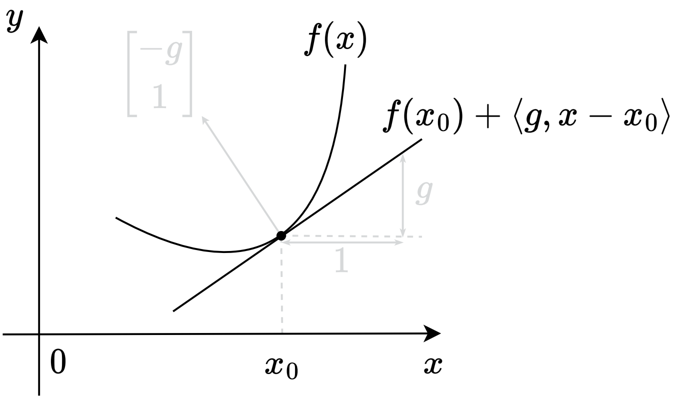


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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

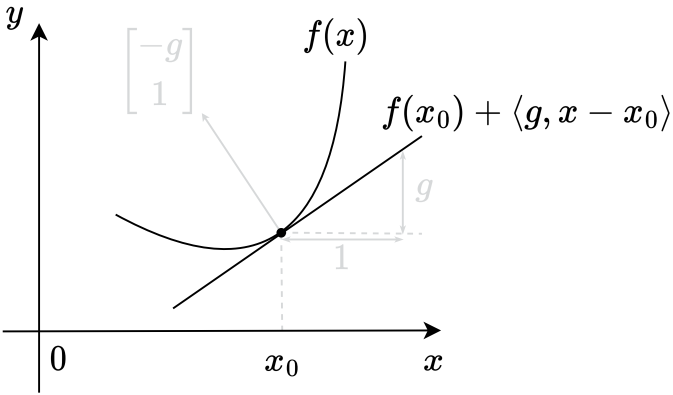


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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

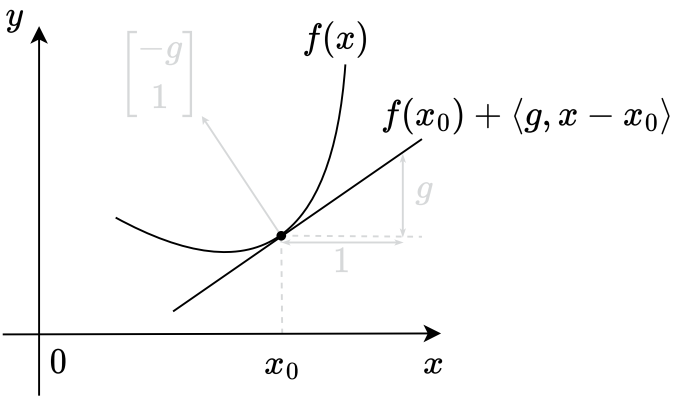


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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

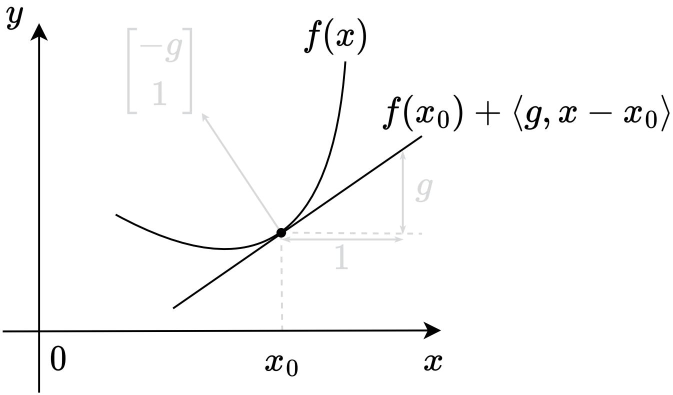


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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

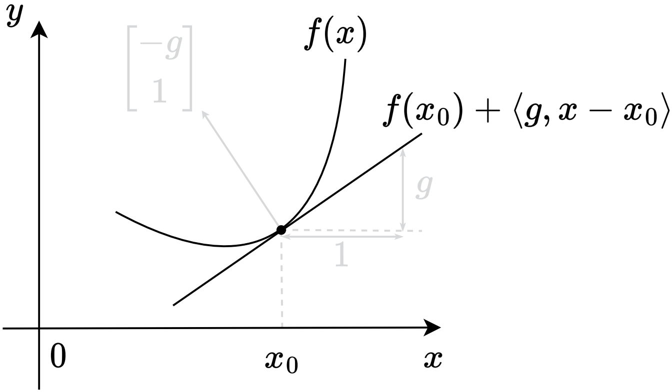


Figure 3: Taylor linear approximation serves as a global lower bound for a convex function

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 $\partial f(x_0)$.

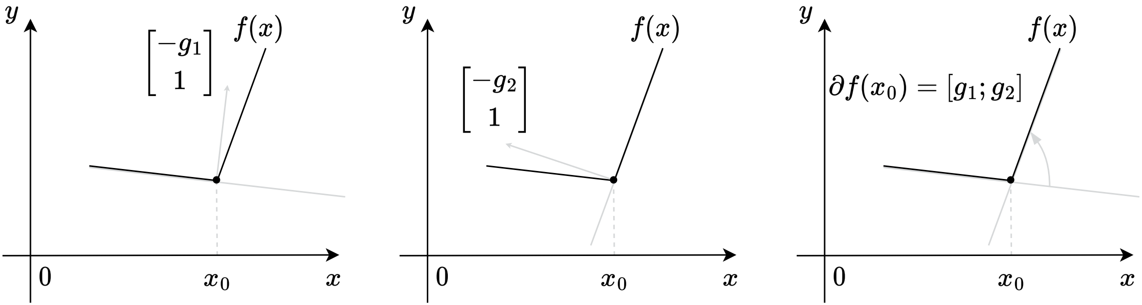


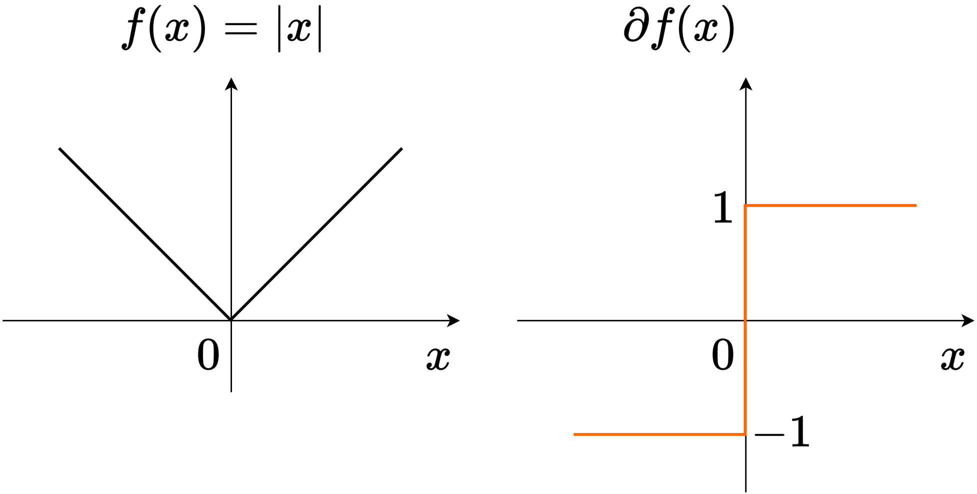
Figure 4: Subdifferential is a set of all possible subgradients

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 $x \in \partial f^*(z)$.

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, $g_k \in \partial f(x_k)$

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

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

Convergence bound

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

Convergence bound

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

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

Convergence bound

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

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

$$
\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle = \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k^2\|
$$

Convergence bound

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

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

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

Convergence bound

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

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

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

Convergence bound

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

- 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:

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

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

Convergence bound

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

- 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: $\langle g_k, x_k - x^* \rangle \leq f(x_k) - f(x^*) = f(x_k) - f^*$.

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

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

Convergence bound

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

- 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: $\langle g_k, x_k - x^* \rangle \leq f(x_k) - f(x^*) = f(x_k) - f^*$.
- We additionaly assume, that $\|g_k\|^2 \leq G^2$

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

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

Convergence bound

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

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

$$\sum_{k=0}^{T-1} 2\alpha_k \langle g_k, x_k - x^* \rangle = \|x_0 - x^*\|^2 - \|x_T - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k^2\|$$
$$\leq \|x_0 - x^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|g_k^2\|$$
$$\leq R^2 + G^2 \sum_{k=0}^{T-1} \alpha_k^2$$

- 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: $\langle g_k, x_k - x^* \rangle \leq f(x_k) - f(x^*) = f(x_k) - f^*$.
- We additionaly assume, that $\|g_k\|^2 \leq G^2$
- We use the notation $R = \|x_0 - x^*\|_2$

Convergence bound

Assuming $\alpha_k = \alpha$ (constant stepsize), we have:

$$
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T
$$

Convergence bound

Assuming $\alpha_k = \alpha$ (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by α gives $\alpha^* = \dfrac{R}{G}\sqrt{\dfrac{1}{T}}$ and

$\sum\limits_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$

Convergence bound

Assuming $\alpha_k = \alpha$ (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by α gives $\alpha^* = \frac{R}{G}\sqrt{\frac{1}{T}}$ and

$\sum\limits_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}$.

$$f(\overline{x}) - f^* = f\left(\frac{1}{T}\sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T}\left(\sum_{k=0}^{T-1}(f(x_k) - f^*)\right)$$

Convergence bound

Assuming $\alpha_k = \alpha$ (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by α gives $\alpha^* = \frac{R}{G}\sqrt{\frac{1}{T}}$ and

$\sum\limits_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}$.

$$\begin{aligned} f(\overline{x}) - f^* &= f\left(\frac{1}{T} \sum_{k=0}^{T-1} x_k \right) - f^* \leq \frac{1}{T} \left(\sum_{k=0}^{T-1} (f(x_k) - f^*) \right) \\ &\leq \frac{1}{T} \left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \end{aligned}$$

Convergence bound

Assuming $\alpha_k = \alpha$ (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by α gives $\alpha^* = \frac{R}{G}\sqrt{\frac{1}{T}}$ and

$\sum\limits_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}$.

$$\begin{aligned} f(\overline{x}) - f^* &= f\left(\frac{1}{T}\sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T}\left(\sum_{k=0}^{T-1} (f(x_k) - f^*)\right) \\ &\leq \frac{1}{T}\left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle\right) \\ &\leq GR\frac{1}{\sqrt{T}} \end{aligned}$$

Convergence bound

Assuming $\alpha_k = \alpha$ (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by α gives $\alpha^* = \frac{R}{G}\sqrt{\frac{1}{T}}$ and

$\sum\limits_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}$.

$$\begin{aligned}
f(\overline{x}) - f^* = f\left(\frac{1}{T}\sum_{k=0}^{T-1} x_k\right) - f^* &\leq \frac{1}{T}\left(\sum_{k=0}^{T-1} (f(x_k) - f^*)\right) \\
&\leq \frac{1}{T}\left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle\right) \\
&\leq GR\frac{1}{\sqrt{T}}
\end{aligned}$$

Convergence bound

Assuming $\alpha_k = \alpha$ (constant stepsize), we have:

$$
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T
$$

Minimizing the right-hand side by α gives $\alpha^* = \dfrac{R}{G}\sqrt{\dfrac{1}{T}}$ and

$\displaystyle\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$

$$
\begin{aligned}
f(\overline{x}) - f^* = f\left(\frac{1}{T}\sum_{k=0}^{T-1} x_k \right) - f^* &\leq \frac{1}{T}\left(\sum_{k=0}^{T-1} (f(x_k) - f^*) \right) \\
&\leq \frac{1}{T}\left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right) \\
&\leq GR\frac{1}{\sqrt{T}}
\end{aligned}
$$

Important notes:

- Obtaining bounds not for x_T but for the arithmetic mean over iterations $\overline{x}$ is a typical trick in obtaining estimates for methods where there is convexity but no monotonic decreasing at each iteration. There is no guarantee of success at each iteration, but there is a guarantee of success on average

Convergence bound

Assuming $\alpha_k = \alpha$ (constant stepsize), we have:

$$\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq \frac{R^2}{2\alpha} + \frac{\alpha}{2} G^2 T$$

Minimizing the right-hand side by α gives $\alpha^* = \frac{R}{G}\sqrt{\frac{1}{T}}$ and

$\sum\limits_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}.$

$$\begin{aligned}
f(\overline{x}) - f^* &= f\left(\frac{1}{T}\sum_{k=0}^{T-1} x_k\right) - f^* \leq \frac{1}{T}\left(\sum_{k=0}^{T-1}(f(x_k) - f^*)\right) \\
&\leq \frac{1}{T}\left(\sum_{k=0}^{T-1}\langle g_k, x_k - x^* \rangle\right) \\
&\leq GR\frac{1}{\sqrt{T}}
\end{aligned}$$

Important notes:

- Obtaining bounds not for x_T but for the arithmetic mean over iterations $\overline{x}$ is a typical trick in obtaining estimates for methods where there is convexity but no monotonic decreasing at each iteration. There is no guarantee of success at each iteration, but there is a guarantee of success on average
- To choose the optimal step, we need to know (assume) the number of iterations in advance. Possible solution: initialize T with a small value, after reaching this number of iterations double T and restart the algorithm. A more intelligent way: adaptive selection of stepsize.

Steepest subgradient descent convergence bound

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

Steepest subgradient descent convergence bound

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

Steepest subgradient descent convergence bound

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\
&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \circeq \\
\alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)}
\end{aligned}
$$

Steepest subgradient descent convergence bound

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\
&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \circeq \\
\alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\
&\circeq \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2}
\end{aligned}
$$

Steepest subgradient descent convergence bound

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\
&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \circeq \\
\alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\
&\circeq \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2} \\
\langle g_k, x_k - x^* \rangle^2 &= \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) \|g_k\|^2 \leq \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2
\end{aligned}
$$

Steepest subgradient descent convergence bound

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\
&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \overset{\circ}{=} \\
\alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\
&\overset{\circ}{=} \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2} \\
\langle g_k, x_k - x^* \rangle^2 &= \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) \|g_k\|^2 \leq \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) G^2 \\
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 &\leq \sum_{k=0}^{T-1} \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) G^2 \leq \left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right) G^2
\end{aligned}
$$

Steepest subgradient descent convergence bound

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\
&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \stackrel{\circ}{=} \\
\alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\
&\stackrel{\circ}{=} \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2} \\
\langle g_k, x_k - x^* \rangle^2 &= \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) \|g_k\|^2 \leq \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2 \\
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 &\leq \sum_{k=0}^{T-1} \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2 \leq \left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2 \right) G^2
\end{aligned}
$$

$$
\frac{1}{T} \left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right)^2 \leq \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 \leq R^2 G^2 \qquad \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}
$$

Steepest subgradient descent convergence bound

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\
&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \circeq \\
\alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\
&\circeq \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2} \\
\langle g_k, x_k - x^* \rangle^2 &= \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) \|g_k\|^2 \leq \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2 \\
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 &\leq \sum_{k=0}^{T-1} \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2 \leq \left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2 \right) G^2 \\
\frac{1}{T} \left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right)^2 &\leq \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 \leq R^2 G^2 \qquad \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}
\end{aligned}
$$

Steepest subgradient descent convergence bound

$$
\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \alpha_k g_k\|^2 = \\
&= \|x_k - x^*\|^2 + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \mathrel{\hat{=}} \\
\alpha_k &= \frac{\langle g_k, x_k - x^* \rangle}{\|g_k\|^2} \text{ (from minimizing right hand side over stepsize)} \\
&\mathrel{\hat{=}} \|x_k - x^*\|^2 - \frac{\langle g_k, x_k - x^* \rangle^2}{\|g_k\|^2} \\
\langle g_k, x_k - x^* \rangle^2 &= \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) \|g_k\|^2 \leq \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2 \\
\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 &\leq \sum_{k=0}^{T-1} \left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 \right) G^2 \leq \left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2 \right) G^2
\end{aligned}
$$

$$
\frac{1}{T} \left(\sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \right)^2 \leq \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle^2 \leq R^2 G^2 \qquad \sum_{k=0}^{T-1} \langle g_k, x_k - x^* \rangle \leq GR\sqrt{T}
$$

Which leads to exactly the same bound of $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ on the primal gap. In fact, for this class of functions, you can't get a better result than $\frac{1}{\sqrt{T}}$.

Convergence results

Theorem

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

$$f(\overline{x}) - f^* \leq \frac{G\|x_0 - x^*\|_2}{\sqrt{K}} \qquad \overline{x} = \frac{1}{K}\sum_{k=0}^{K-1} x_i$$

- $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ is slow, but already hits the lower bound ($\mathcal{O}\left(\frac{1}{T}\right)$ in the strongly convex case).

Convergence results

Theorem

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

$$f(\overline{x}) - f^* \leq \frac{G\|x_0 - x^*\|_2}{\sqrt{K}} \qquad \overline{x} = \frac{1}{K}\sum_{k=0}^{K-1} x_i$$

- $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ is slow, but already hits the lower bound ($\mathcal{O}\left(\frac{1}{T}\right)$ in the strongly convex case).
- Proved result requires pre-defined step size strategy, which is not practical (usually one cas just use several diminishes strategies).

Convergence results

Theorem

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

$$f(\overline{x}) - f^* \leq \frac{G\|x_0 - x^*\|_2}{\sqrt{K}} \qquad \overline{x} = \frac{1}{K}\sum_{k=0}^{K-1} x_i$$

- $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ is slow, but already hits the lower bound ($\mathcal{O}\left(\frac{1}{T}\right)$ in the strongly convex case).
- Proved result requires pre-defined step size strategy, which is not practical (usually one cas just use several diminishes strategies).
- There is no monotonic decrease of objective.

Convergence results

Theorem

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

$$
f(\overline{x}) - f^* \leq \frac{G\|x_0 - x^*\|_2}{\sqrt{K}} \qquad \overline{x} = \frac{1}{K}\sum_{k=0}^{K-1} x_i
$$

- $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ is slow, but already hits the lower bound ($\mathcal{O}\left(\frac{1}{T}\right)$ in the strongly convex case).
- Proved result requires pre-defined step size strategy, which is not practical (usually one cas just use several diminishes strategies).
- There is no monotonic decrease of objective.
- Convergence is slower, than for the gradient descent (smooth case). However, if we will go deeply for the problem structure, we can improve convergence (proximal gradient method).

Convergence results

Theorem

Let f be a convex G-Lipschitz function and $f_k^{\text{best}} = \min\limits_{i=1,\ldots,k} f(x^i)$. For a fixed step size α, subgradient method satisfies

$$\lim_{k\to\infty} f_k^{\text{best}} \leq f^* + \frac{G^2\alpha}{2}$$

Theorem

Let f be a convex G-Lipschitz function and $f_k^{\text{best}} = \min\limits_{i=1,\ldots,k} f(x^i)$. For a diminishing step size α_k (square summable but not summable. Important here that step sizes go to zero, but not too fast), subgradient method satisfies

$$\lim_{k\to\infty} f_k^{\text{best}} \leq f^*$$

Linear Least Squares with l_1-regularization

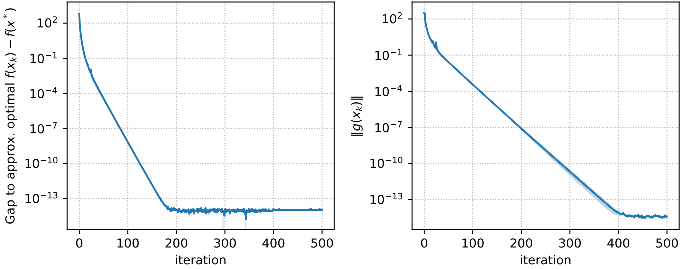
$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$$

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:

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

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

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