Optimization for Machine Learning

(Problems; Algorithms - B)

Suvrit Sra
Massachusetts Institute of Technology

PKU Summer School on Data Science (July 2017)

Recap

♡ Convex sets, convex functions, some challenges

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \leq \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \leq \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \leq \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.
- ♡ Converge as $O(e^{-k})$ for strongly convex; AGM attains lower-bd.

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \leq \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.
- ♡ Converge as $O(e^{-k})$ for strongly convex; AGM attains lower-bd.
- ♡ Constrained optimization: $\min f(x)$ s.t. $x \in \mathcal{X}$

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \le \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.
- ♡ Converge as $O(e^{-k})$ for strongly convex; AGM attains lower-bd.
- ♡ Constrained optimization: $\min f(x)$ s.t. $x \in \mathcal{X}$
- ♡ Optimality condition: $\langle \nabla f(x^*), x - x^* \rangle \ge 0$ for all $x \in \mathcal{X}$

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \leq \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.
- ♡ Converge as $O(e^{-k})$ for strongly convex; AGM attains lower-bd.
- ♡ Constrained optimization: $\min f(x)$ s.t. $x \in \mathcal{X}$
- ♡ Optimality condition: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$
- ♡ Frank-Wolfe algorithm, using $\min_{z \in \mathcal{X}} \langle \nabla f(x^k), z \rangle$

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \leq \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.
- ♡ Converge as $O(e^{-k})$ for strongly convex; AGM attains lower-bd.
- ♡ Constrained optimization: $\min f(x)$ s.t. $x \in \mathcal{X}$
- ♡ Optimality condition: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$
- ♡ Frank-Wolfe algorithm, using $\min_{z \in \mathcal{X}} \langle \nabla f(x^k), z \rangle$
- ♡ Projected gradient, $x \leftarrow P_{\mathcal{X}}(x - \alpha \nabla f(x))$

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d\rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \leq \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.
- ♡ Converge as $O(e^{-k})$ for strongly convex; AGM attains lower-bd.
- ♡ Constrained optimization: $\min f(x)$ s.t. $x \in \mathcal{X}$
- ♡ Optimality condition: $\langle \nabla f(x^*), x - x^*\rangle \geq 0$ for all $x \in \mathcal{X}$
- ♡ Frank-Wolfe algorithm, using $\min_{z \in \mathcal{X}} \langle \nabla f(x^k), z\rangle$
- ♡ Projected gradient, $x \leftarrow P_{\mathcal{X}}(x - \alpha \nabla f(x))$
- ♡ Stochastic programming: $\min_x F(x) := \mathbb{E}_\xi[f(x,\xi)]$

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d \rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \leq \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.
- ♡ Converge as $O(e^{-k})$ for strongly convex; AGM attains lower-bd.
- ♡ Constrained optimization: $\min f(x)$ s.t. $x \in \mathcal{X}$
- ♡ Optimality condition: $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for all $x \in \mathcal{X}$
- ♡ Frank-Wolfe algorithm, using $\min_{z \in \mathcal{X}} \langle \nabla f(x^k), z \rangle$
- ♡ Projected gradient, $x \leftarrow P_{\mathcal{X}}(x - \alpha \nabla f(x))$
- ♡ Stochastic programming: $\min_x F(x) := \mathbb{E}_\xi[f(x, \xi)]$
- ♡ SA/SGD: $x^{k+1} = x^k - \alpha_k g_k$, where $\mathbb{E}[g_k] = \nabla F(x^k)$

Recap

- ♡ Convex sets, convex functions, some challenges
- ♡ Minimizing $f(x)$ via descent $x \leftarrow x + \alpha d$ ($\langle \nabla f, d\rangle < 0$)
- ♡ $\nabla f(x^*) = 0$ necessary for optimality; sufficient for convex
- ♡ Gradient descent ensures $f(x^k) - f(x^*) \le \epsilon$ in $O(1/\epsilon)$ iterations (we wrote this as: $f(x^k) - f(x^*) = O(1/k)$).
- ♡ Lower bound: $O(1/k^2)$; attained by Nesterov's accelerated gradient method.
- ♡ Converge as $O(e^{-k})$ for strongly convex; AGM attains lower-bd.
- ♡ Constrained optimization: $\min f(x)$ s.t. $x \in \mathcal{X}$
- ♡ Optimality condition: $\langle \nabla f(x^*), x - x^*\rangle \ge 0$ for all $x \in \mathcal{X}$
- ♡ Frank-Wolfe algorithm, using $\min_{z\in\mathcal{X}} \langle \nabla f(x^k), z\rangle$
- ♡ Projected gradient, $x \leftarrow P_{\mathcal{X}}(x - \alpha \nabla f(x))$
- ♡ Stochastic programming: $\min_x F(x) := \mathbb{E}_\xi[f(x,\xi)]$
- ♡ SA/SGD: $x^{k+1} = x^k - \alpha_k g_k$, where $\mathbb{E}[g_k] = \nabla F(x^k)$
- ♡ Finite-sum: $\frac{1}{n}\sum_i f_i(x)$; $x^{k+1} = x^k - \alpha_k \nabla f_{i_k}(x^k)$, where $i_k \sim \mathrm{U}([n])$

Example: joint convexity

- Show that $f(w, X) := w^T X^{-1} w$ is jointly convex (in $w \in \mathbb{R}^n$ and $X \succ 0$, i.e., positive definite)

Example: joint convexity

- Show that $f(w, X) := w^T X^{-1} w$ is jointly convex (in $w \in \mathbb{R}^n$ and $X \succ 0$, i.e., positive definite)

Let us prove via midpoint convexity. So we show that

$$f\left(\frac{w+v}{2}, \frac{A+B}{2}\right) \leq \frac{1}{2} f(w, A) + \frac{1}{2} f(v, B).$$

Example: joint convexity

- Show that $f(w, X) := w^T X^{-1} w$ is jointly convex (in $w \in \mathbb{R}^n$ and $X \succ 0$, i.e., positive definite)

Let us prove via midpoint convexity. So we show that

$$f\left(\frac{w+v}{2}, \frac{A+B}{2}\right) \leq \frac{1}{2} f(w, A) + \frac{1}{2} f(v, B).$$

In other words, we show that

$$\left\langle \frac{w+v}{2}, \left(\frac{A+B}{2}\right)^{-1} \frac{w+v}{2} \right\rangle \leq \frac{1}{2} f(w, A) + \frac{1}{2} f(v, B),$$

which simplifies to showing that (**verify!**)

Example: joint convexity

$$w^T A^{-1} w + v^T B^{-1} v \geq (w+v)^T (A+B)^{-1} (w+v). \qquad (\star)$$

Example: joint convexity

$$w^T A^{-1} w + v^T B^{-1} v \geq (w+v)^T (A+B)^{-1} (w+v). \qquad (\star)$$

Recall the Schur complement lemma, i.e., $\begin{bmatrix} P & Q \\ Q^T & R \end{bmatrix} \succeq 0$ iff $P \succeq QR^{-1}Q^T$ (we essentially proved this in Lecture 1).

Example: joint convexity

$$w^T A^{-1} w + v^T B^{-1} v \geq (w+v)^T (A+B)^{-1} (w+v). \qquad (\star)$$

Recall the Schur complement lemma, i.e., $\begin{bmatrix} P & Q \\ Q^T & R \end{bmatrix} \succeq 0$ iff $P \succeq QR^{-1}Q^T$ (we essentially proved this in Lecture 1).

Thus, since $w^T A^{-1} w \geq w^T A^{-1} w$, we have

$$\begin{bmatrix} w^T A^{-1} w & w^T \\ w & A \end{bmatrix} \succeq 0,$$

Example: joint convexity

$$w^T A^{-1} w + v^T B^{-1} v \geq (w+v)^T (A+B)^{-1} (w+v). \qquad (\star)$$

Recall the Schur complement lemma, i.e., $\begin{bmatrix} P & Q \\ Q^T & R \end{bmatrix} \succeq 0$ iff $P \succeq QR^{-1}Q^T$ (we essentially proved this in Lecture 1).

Thus, since $w^T A^{-1} w \geq w^T A^{-1} w$, we have

$$\begin{bmatrix} w^T A^{-1} w & w^T \\ w & A \end{bmatrix} \succeq 0, \text{ similarly, } \begin{bmatrix} v^T B^{-1} v & v^T \\ v & B \end{bmatrix} \succeq 0.$$

Since sum of PD matrices is PD, this implies that

$$\begin{bmatrix} w^T A^{-1} w + v^T B^{-1} v & w^T + v^T \\ w + v & A + B \end{bmatrix} \succeq 0.$$

Taking Schur complements of this matrix, we obtain $(\star)$.
Thus, we have proved $f(w, X) = w^T X^{-1} w$ is jointly convex.

Nonsmooth functions

Power of nonsmooth functions

Write constrained problem as unconstrained

$$\min \quad f(x) \quad \text{s.t. } x \in \mathcal{X}$$

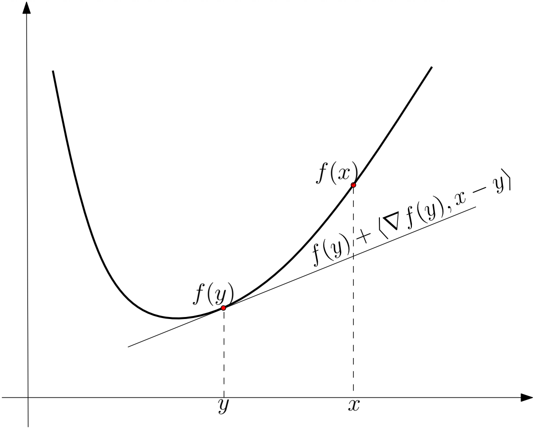
Power of nonsmooth functions

Write constrained problem as unconstrained

$$
\begin{aligned}
&\min \quad f(x) \quad \text{s.t. } x \in \mathcal{X} \\
&\min \quad f(x) + \mathbb{1}_{\mathcal{X}}(x),
\end{aligned}
$$

where $\mathbb{1}_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and $+\infty$ otherwise.

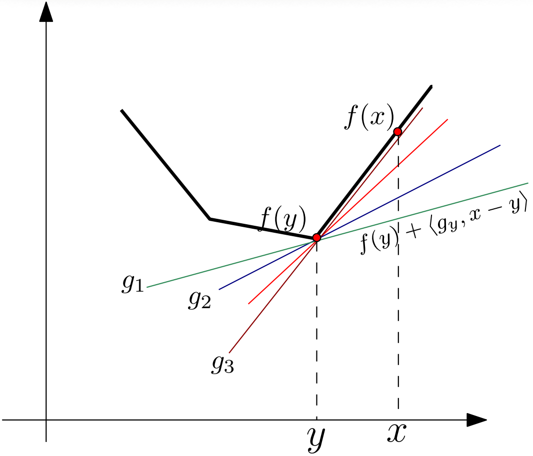
Subgradients: global underestimators



$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

Hence $\nabla f(y) = 0$ implies that y is global min.

Subgradients: global underestimators



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

If one of the $g = 0$, then y a global min.

Subgradients – basic facts

- f is convex, differentiable: $\nabla f(y)$ the **unique** subgradient at y
- A vector g is a subgradient at a point y if and only if $f(y) + \langle g, x - y \rangle$ is **globally** smaller than $f(x)$.
- Usually, **one** subgradient costs approx. as much as $f(x)$

Subgradients – basic facts

- f is convex, differentiable: $\nabla f(y)$ the **unique** subgradient at y
- A vector g is a subgradient at a point y if and only if $f(y) + \langle g, x - y\rangle$ is **globally** smaller than $f(x)$.
- Usually, **one** subgradient costs approx. as much as $f(x)$
- Determining all subgradients at a given point — difficult.
- Subgradient calculus—major achievement in convex analysis
- Fenchel-Young inequality: $f(x) + f^*(s) \geq \langle s, x\rangle$
  tight at a subgradient

Rules for subgradients

Subgradient for pointwise sup

$$f(x) := \sup_{y \in \mathcal{Y}} \quad h(x, y)$$

$$\boxed{\text{Getting } \partial f(x) \text{ is complicated!}}$$

Subgradient for pointwise sup

$$f(x) := \sup_{y \in \mathcal{Y}} \quad h(x, y)$$

$$\boxed{\text{Getting } \partial f(x) \text{ is complicated!}}$$

Simple way to obtain some $g \in \partial f(x)$:

Subgradient for pointwise sup

$$f(x) := \sup_{y \in \mathcal{Y}} \quad h(x, y)$$

$$\boxed{\text{Getting } \partial f(x) \text{ is complicated!}}$$

Simple way to obtain some $g \in \partial f(x)$:

- Pick any y^* for which $h(x, y^*) = f(x)$

Subgradient for pointwise sup

$$f(x) := \sup_{y \in \mathcal{Y}} \quad h(x, y)$$

$$\boxed{\text{Getting } \partial f(x) \text{ is complicated!}}$$

Simple way to obtain some $g \in \partial f(x)$:

- Pick any y^* for which $h(x, y^*) = f(x)$
- Pick any subgradient $g \in \partial h(x, y^*)$

Subgradient for pointwise sup

$$f(x) := \sup_{y \in \mathcal{Y}} \quad h(x, y)$$

$$\boxed{\text{Getting } \partial f(x) \text{ is complicated!}}$$

Simple way to obtain some $g \in \partial f(x)$:

- Pick any y^* for which $h(x, y^*) = f(x)$
- Pick any subgradient $g \in \partial h(x, y^*)$
- This $g \in \partial f(x)$

Subgradient for pointwise sup

$$f(x) := \sup_{y \in \mathcal{Y}} \quad h(x, y)$$

$$\boxed{\text{Getting } \partial f(x) \text{ is complicated!}}$$

Simple way to obtain some $g \in \partial f(x)$:

- Pick any y^* for which $h(x, y^*) = f(x)$
- Pick any subgradient $g \in \partial h(x, y^*)$
- This $g \in \partial f(x)$

$$\begin{aligned} h(z, y^*) \quad &\geq \quad h(x, y^*) + g^T(z - x) \\ h(z, y^*) \quad &\geq \quad f(x) + g^T(z - x) \end{aligned}$$

Subgradient for pointwise sup

$$f(x) := \sup_{y \in \mathcal{Y}} \quad h(x, y)$$

$$\boxed{\text{Getting } \partial f(x) \text{ is complicated!}}$$

Simple way to obtain some $g \in \partial f(x)$:

- Pick any y^* for which $h(x, y^*) = f(x)$
- Pick any subgradient $g \in \partial h(x, y^*)$
- This $g \in \partial f(x)$

$$\begin{aligned} h(z, y^*) &\geq h(x, y^*) + g^T(z - x) \\ h(z, y^*) &\geq f(x) + g^T(z - x) \\ f(z) &\geq h(z, y) \qquad \text{(because of sup)} \\ f(z) &\geq f(x) + g^T(z - x). \end{aligned}$$

Example

Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \le i \le n} (a_i^T x + b_i).$$

This f a max (in fact, over a finite number of terms)

Example

Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \le i \le n} (a_i^T x + b_i).$$

This f a max (in fact, over a finite number of terms)

- Suppose $f(x) = a_k^T x + b_k$ for some index k

Example

Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \le i \le n} (a_i^T x + b_i).$$

This f a max (in fact, over a finite number of terms)

- Suppose $f(x) = a_k^T x + b_k$ for some index k
- Here $f(x; y) = f_k(x) = a_k^T x + b_k$, and $\partial f_k(x) = \{\nabla f_k(x)\}$

Example

Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \le i \le n} (a_i^T x + b_i).$$

This f a max (in fact, over a finite number of terms)

- Suppose $f(x) = a_k^T x + b_k$ for some index k
- Here $f(x; y) = f_k(x) = a_k^T x + b_k$, and $\partial f_k(x) = \{\nabla f_k(x)\}$
- Hence, $a_k \in \partial f(x)$ works!

Subgradient of expectation

Suppose $f = \mathbf{E}f(x, u)$, where f is convex in x for each u (an r.v.)

$$f(x) := \int f(x, u)p(u)du$$

Subgradient of expectation

Suppose $f = \mathbf{E}f(x,u)$, where f is convex in x for each u (an r.v.)

$$f(x) := \int f(x,u)p(u)du$$

- For each u choose any $g(x,u) \in \partial_x f(x,u)$

Subgradient of expectation

Suppose $f = \mathbf{E}f(x,u)$, where f is convex in x for each u (an r.v.)

$$f(x) := \int f(x,u)p(u)du$$

- For each u choose any $g(x,u) \in \partial_x f(x,u)$
- Then, $g = \int g(x,u)p(u)du = \mathbf{E}g(x,u) \in \partial f(x)$

Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each f_i cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each f_i cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each f_i cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

- For $i = 1$ to n, compute $g_i \in \partial f_i(x)$

Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each f_i cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

- For $i = 1$ to n, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$

Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each f_i cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

- For $i = 1$ to n, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$
- Set $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$; this $g \in \partial f(x)$

Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each f_i cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

- For $i = 1$ to n, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$
- Set $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$; this $g \in \partial f(x)$
- Compare with $\nabla f(x) = J \nabla h(x)$, where J matrix of $\nabla f_i(x)$

Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each f_i cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

- For $i = 1$ to n, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$
- Set $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$; this $g \in \partial f(x)$
- Compare with $\nabla f(x) = J \nabla h(x)$, where J matrix of $\nabla f_i(x)$

Exercise: Verify $g \in \partial f(x)$ by showing $f(z) \geq f(x) + g^T(z - x)$

References for subgradients

1. R. T. Rockafellar. *Convex Analysis*
2. S. Boyd (Stanford); EE364b Lecture Notes.

Subdifferential*

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

- ♣ If $x \in$ relative interior of $\operatorname{dom} f$, then $\partial f(x)$ nonempty

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

- ♣ If $x \in$ relative interior of $\mathrm{dom}\, f$, then $\partial f(x)$ nonempty
- ♣ If f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

- ♣ If $x \in$ relative interior of $\operatorname{dom} f$, then $\partial f(x)$ nonempty
- ♣ If f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- ♣ If $\partial f(x) = \{g\}$, then f is differentiable and $g = \nabla f(x)$

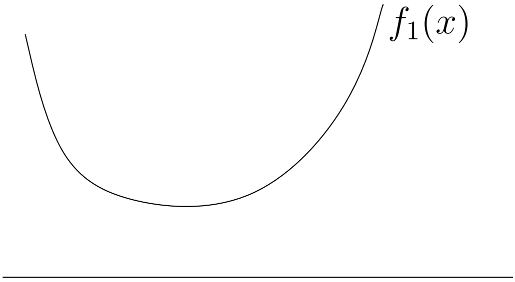
Exercise: What is $\partial f(x)$ for the *ReLU* function: $\max(0, x)$?

Subdifferential – example

$\boxed{f(x) := \max(f_1(x), f_2(x))\text{; both } f_1, f_2 \text{ convex, differentiable}}$

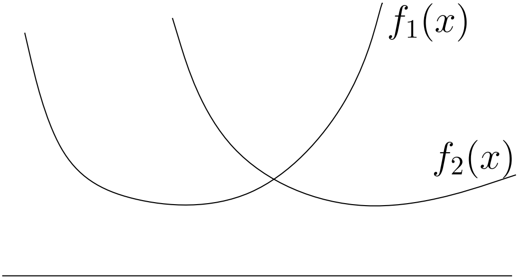
Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



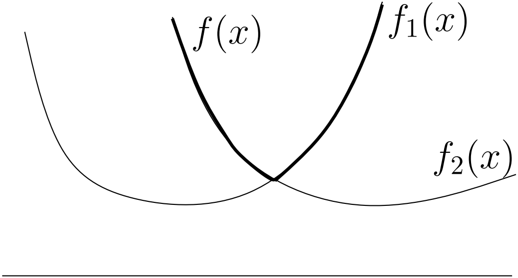
Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



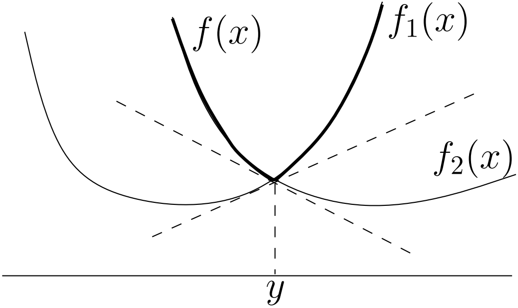
Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



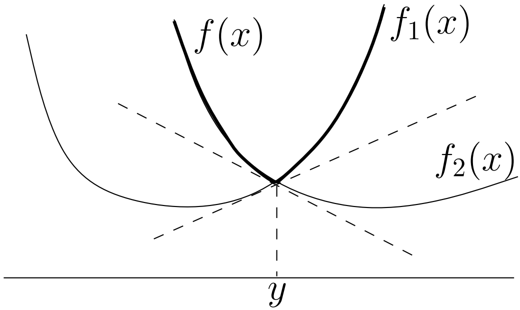
Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



Subdifferential – example

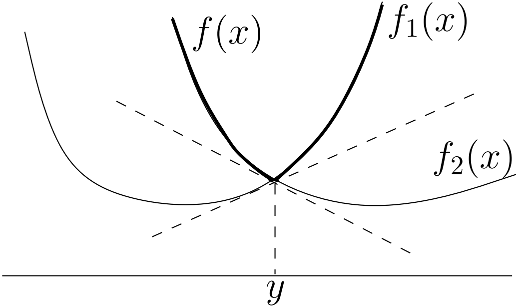
$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



$\star$ $f_1(x) > f_2(x)$: unique subgradient of f is $f_1'(x)$

Subdifferential – example

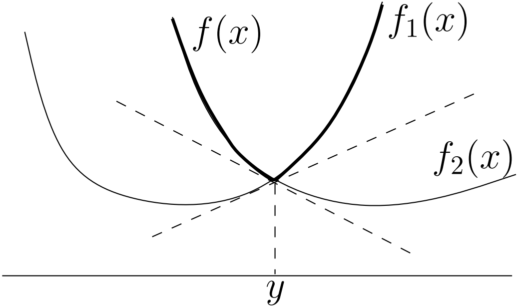
$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



- $\star$ $f_1(x) > f_2(x)$: unique subgradient of f is $f_1'(x)$
- $\star$ $f_1(x) < f_2(x)$: unique subgradient of f is $f_2'(x)$

Subdifferential – example

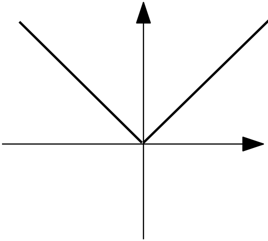
$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



- ⋆ $f_1(x) > f_2(x)$: unique subgradient of f is $f_1'(x)$
- ⋆ $f_1(x) < f_2(x)$: unique subgradient of f is $f_2'(x)$
- ⋆ $f_1(y) = f_2(y)$: subgradients, the segment $[f_1'(y), f_2'(y)]$ (imagine all supporting lines turning about point y)

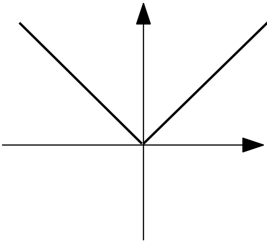
Subdifferential for abs value

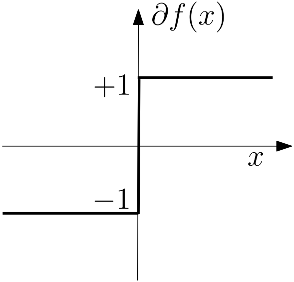
$$\boxed{f(x) = |x|}$$



Subdifferential for abs value

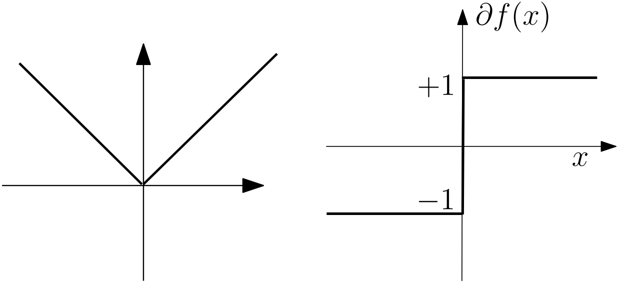
$$\boxed{f(x) = |x|}$$





Subdifferential for abs value

$$\boxed{f(x) = |x|}$$



$$\partial|x| = \begin{cases} -1 & x < 0, \\ +1 & x > 0, \\ [-1, 1] & x = 0. \end{cases}$$

Subdifferential for Euclidean norm

Example. $f(x) = \|x\|_2$. Then,

$$\partial f(x) := \begin{cases} x/\|x\|_2 & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

Subdifferential for Euclidean norm

Example. $f(x) = \|x\|_2$. Then,

$$\partial f(x) := \begin{cases} x/\|x\|_2 & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

Proof.

$$\begin{aligned} \|z\|_2 &\geq \|x\|_2 + \langle g,\, z - x\rangle \\ \|z\|_2 &\geq \langle g,\, z\rangle \\ &\Longrightarrow \|g\|_2 \leq 1. \end{aligned}$$

Example: difficulties

Example. A convex function need not be subdifferentiable everywhere. Let

$$f(x) := \begin{cases} -(1 - \|x\|_2^2)^{1/2} & \text{if } \|x\|_2 \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

f diff. for all x with $\|x\|_2 < 1$, but $\partial f(x) = \emptyset$ whenever $\|x\|_2 \geq 1$.

Subdifferential calculus

- Finding one subgradient within $\partial f(x)$
- Determining entire subdifferential $\partial f(x)$ at a point x
- Do we have the chain rule?

Subdifferential calculus

- If f is differentiable, $\partial f(x) = \{\nabla f(x)\}$
- **Scaling** $\alpha > 0$, $\partial(\alpha f)(x) = \alpha \partial f(x) = \{\alpha g \mid g \in \partial f(x)\}$
- **Addition***: $\partial(f + k)(x) = \partial f(x) + \partial k(x)$ (set addition)
- **Chain rule***: Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^m \to \mathbb{R}$, and $h : \mathbb{R}^n \to \mathbb{R}$ be given by $h(x) = f(Ax + b)$. Then,

$$\partial h(x) = A^T \partial f(Ax + b).$$

- **Chain rule***: $h(x) = f \circ k$, where $k : X \to Y$ is diff.

$$\partial h(x) = \partial f(k(x)) \circ Dk(x) = [Dk(x)]^T \partial f(k(x))$$

- **Max function***: If $f(x) := \max_{1 \leq i \leq m} f_i(x)$, then

$$\partial f(x) = \operatorname{conv} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

  convex hull over subdifferentials of "active" functions at x
- **Conjugation:** $z \in \partial f(x)$ if and only if $x \in \partial f^*(z)$

* — can fail to hold without precise assumptions.

Example: breakdown

It can happen that $\partial(f_1 + f_2) \neq \partial f_1 + \partial f_2$

Example. Define f_1 and f_2 by

$$f_1(x) := \begin{cases} -2\sqrt{x} & \text{if } x \geq 0, \\ +\infty & \text{if } x < 0, \end{cases} \quad \text{and} \quad f_2(x) := \begin{cases} +\infty & \text{if } x > 0, \\ -2\sqrt{-x} & \text{if } x \leq 0. \end{cases}$$

Then, $f = \max\{f_1, f_2\} = \mathbb{1}_{\{0\}}$, whereby $\partial f(0) = \mathbb{R}$
But $\partial f_1(0) = \partial f_2(0) = \emptyset$.

However, $\partial f_1(x) + \partial f_2(x) \subset \partial(f_1 + f_2)(x)$ always holds.

Subdifferential – example

Example. $f(x) = \|x\|_\infty$. Then,

$$\partial f(0) = \text{conv}\left\{\pm e_1, \ldots, \pm e_n\right\},$$

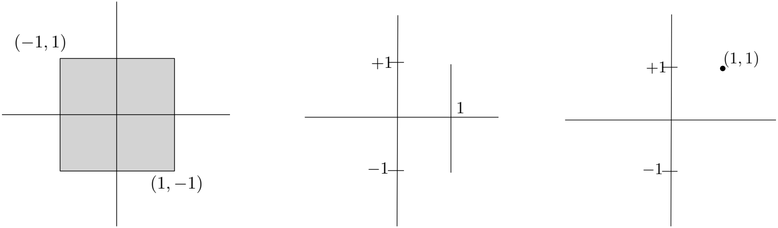
where e_i is i-th canonical basis vector.

To prove, notice that $f(x) = \max_{1 \le i \le n} \left\{|e_i^T x|\right\}$

Then use, *chain rule* and *max rule* and $\partial | \cdot |$

Subdifferential - example (Boyd)

Example. Let $f(x) = \max\left\{s^T x \mid s_i \in \{-1, 1\}\right\}$ (2^n members)



∂f at $x = (0,0)$ ∂f at $x = (1,0)$ ∂f at $x = (1,1)$

Optimality via subdifferentials

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \geq f(x) + \langle 0, y - x \rangle \quad \forall y$,

Optimality via subdifferentials

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \le f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \ge f(x) + \langle 0, y - x \rangle \quad \forall y, \leftrightarrow 0 \in \partial f(x)$.

Example: constrained smooth problem

Constrained smooth problem

$$
\begin{aligned}
&\min \quad f(x) \quad \text{s.t. } x \in \mathcal{X} \\
&\min \quad f(x) + \mathbb{1}_{\mathcal{X}}(x).
\end{aligned}
$$

Example: constrained smooth problem

Constrained smooth problem

$$\begin{array}{lll} \min & f(x) & \text{s.t. } x \in \mathcal{X} \\ \min & f(x) + \mathbb{1}_{\mathcal{X}}(x). & \end{array}$$

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom} f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle$ $\forall y \in \mathcal{X}$.
- **Normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \{g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x \rangle \quad \forall y \in \mathcal{X}\}$$

Thus: $\min f(x) \quad \text{s.t. } x \in \mathcal{X}$:

◇ If f is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$

Example: constrained smooth problem

Constrained smooth problem

$$\begin{array}{lll} \min & f(x) & \text{s.t. } x \in \mathcal{X} \\ \min & f(x) + \mathbb{1}_{\mathcal{X}}(x). & \end{array}$$

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom} f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle$ $\forall y \in \mathcal{X}$.
- **Normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \{g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x \rangle \quad \forall y \in \mathcal{X}\}$$

Thus: $\min f(x) \quad \text{s.t. } x \in \mathcal{X}$:

♢ If f is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$

♢ $-\nabla f(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*) \iff \langle \nabla f(x^*), y - x^* \rangle \geq 0$ for all $y \in \mathcal{X}$.

Subgradient methods

Subgradient method

$$x^{k+1} = x^k - \alpha_k g^k$$

where $g^k \in \partial f(x^k)$ is **any** subgradient

Subgradient method

$$x^{k+1} = x^k - \alpha_k g^k$$

where $g^k \in \partial f(x^k)$ is **any** subgradient

Stepsize $\alpha_k > 0$ must be chosen

Subgradient method

$$x^{k+1} = x^k - \alpha_k g^k$$

where $g^k \in \partial f(x^k)$ is **any** subgradient

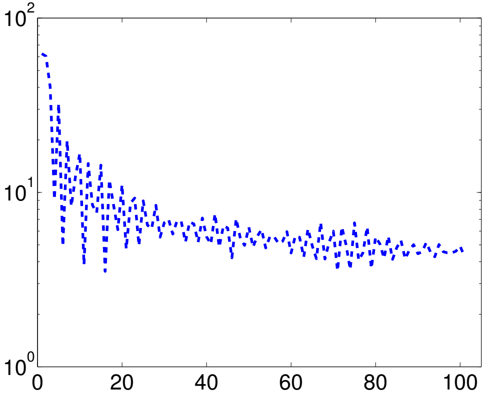
Stepsize $\alpha_k > 0$ must be chosen

- Method generates sequence $\{x^k\}_{k \geq 0}$
- Does this sequence converge to an optimal solution x^*?
- If yes, then how fast?
- What if have constraints: $x \in \mathcal{X}$?

Example: Lasso problem

$$\min \quad \tfrac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$$

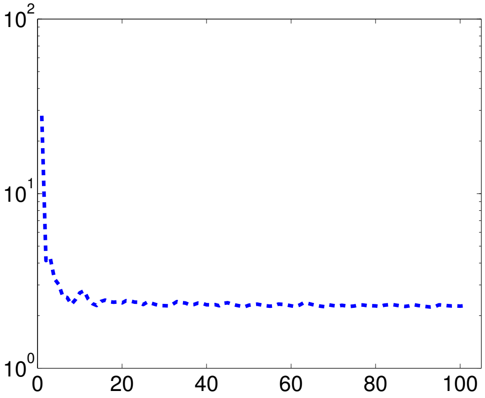
$$x^{k+1} = x^k - \alpha_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))$$



Example: Lasso problem

$$\min \quad \tfrac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$$

$$x^{k+1} = x^k - \alpha_k(A^T(Ax^k - b) + \lambda\,\mathrm{sgn}(x^k))$$



(More careful implementation)

Subgradient method – stepsizes

- **Constant** Set $\alpha_k = \alpha > 0$, for $k \geq 0$
- **Scaled constant** $\alpha_k = \alpha / \|g^k\|_2 \quad (\|x^{k+1} - x^k\|_2 = \alpha)$

Subgradient method – stepsizes

- **Constant** Set $\alpha_k = \alpha > 0$, for $k \geq 0$
- **Scaled constant** $\alpha_k = \alpha/\|g^k\|_2 \quad (\|x^{k+1} - x^k\|_2 = \alpha)$
- **Square summable but not summable**

$$\sum\nolimits_k \alpha_k^2 < \infty, \qquad \sum\nolimits_k \alpha_k = \infty$$

- **Diminishing scalar**

$$\lim_k \alpha_k = 0, \qquad \sum\nolimits_k \alpha_k = \infty$$

- **Adaptive stepsizes** (not covered)

Not a descent method!
Work with best f^k so far: $f^k_{\min} := \min_{0 \leq i \leq k} f^i$

Convergence analysis

Assumptions

- Min is attained: $f^\star := \inf_x f(x) > -\infty$, with $f(x^*) = f^\star$

Convergence analysis

Assumptions

- Min is attained: $f^\star := \inf_x f(x) > -\infty$, with $f(x^*) = f^\star$
- Bounded subgradients: $\|g\|_2 \le G$ for all $g \in \partial f$
  ($f(x) - f(y) = \langle g_\xi, x - y\rangle$; use Cauchy-Schwarz or Hölder)

Convergence analysis

Assumptions

- Min is attained: $f^\star := \inf_x f(x) > -\infty$, with $f(x^*) = f^\star$
- Bounded subgradients: $\|g\|_2 \le G$ for all $g \in \partial f$
  ($f(x) - f(y) = \langle g_\xi, x - y\rangle$; use Cauchy-Schwarz or Hölder)
- Bounded domain: $\|x^0 - x^*\|_2 \le R$

Convergence analysis

Assumptions

- Min is attained: $f^\star := \inf_x f(x) > -\infty$, with $f(x^*) = f^\star$
- Bounded subgradients: $\|g\|_2 \le G$ for all $g \in \partial f$
  ($f(x) - f(y) = \langle g_\xi, x - y\rangle$; use Cauchy-Schwarz or Hölder)
- Bounded domain: $\|x^0 - x^*\|_2 \le R$

$$\boxed{\text{Convergence results for: } f^k_{\min} := \min_{0\le i\le k} f^i}$$

Subgradient method – convergence

Lyapunov function: **Distance to x^***, not function values

Subgradient method – convergence

Lyapunov function: Distance to x^*, not function values

$$\|x^{k+1} - x^*\|_2^2 \quad = \quad \|x^k - \alpha_k g^k - x^*\|_2^2$$

Subgradient method – convergence

Lyapunov function: Distance to x^*, not function values

$$\begin{aligned} \|x^{k+1} - x^*\|_2^2 &= \|x^k - \alpha_k g^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\langle \alpha_k g^k, x^k - x^* \rangle \end{aligned}$$

Subgradient method – convergence

Lyapunov function: Distance to x^*, not function values

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \alpha_k g^k - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\langle \alpha_k g^k,\, x^k - x^* \rangle \\
&\leq \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\alpha_k (f(x^k) - f^\star),
\end{aligned}$$

since $f^\star = f(x^*) \geq f(x^k) + \langle g^k,\, x^* - x^k \rangle$

Subgradient method – convergence

Lyapunov function: Distance to x^*, not function values

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|x^k - \alpha_k g^k - x^*\|_2^2 \\ &= \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\langle \alpha_k g^k, x^k - x^* \rangle \\ &\leq \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\alpha_k (f(x^k) - f^\star),\end{aligned}$$

since $f^\star = f(x^*) \geq f(x^k) + \langle g^k, x^* - x^k \rangle$

Apply same argument to $\|x^k - x^*\|_2^2$ recursively

Subgradient method – convergence

Lyapunov function: **Distance to x^***, not function values

$$\begin{aligned}\|x^{k+1}-x^*\|_2^2 &= \|x^k-\alpha_k g^k - x^*\|_2^2\\ &= \|x^k-x^*\|_2^2+\alpha_k^2\|g^k\|_2^2-2\langle \alpha_k g^k,\, x^k-x^*\rangle\\ &\le \|x^k-x^*\|_2^2+\alpha_k^2\|g^k\|_2^2-2\alpha_k(f(x^k)-f^\star),\end{aligned}$$

since $f^\star = f(x^*) \ge f(x^k) + \langle g^k,\, x^*-x^k\rangle$

Apply same argument to $\|x^k-x^*\|_2^2$ recursively

$$\|x^{k+1}-x^*\|_2^2 \le \|x^0-x^*\|_2^2+\sum\nolimits_{t=1}^{k}\alpha_t^2\|g^k\|_2^2-2\sum\nolimits_{t=1}^{k}\alpha_t(f^t-f^\star).$$

Subgradient method – convergence

Lyapunov function: **Distance to x^***, not function values

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 &= \|x^k - \alpha_k g^k - x^*\|_2^2 \\
&= \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\langle \alpha_k g^k,\, x^k - x^* \rangle \\
&\le \|x^k - x^*\|_2^2 + \alpha_k^2 \|g^k\|_2^2 - 2\alpha_k (f(x^k) - f^\star),
\end{aligned}$$

since $f^\star = f(x^*) \ge f(x^k) + \langle g^k,\, x^* - x^k \rangle$

Apply same argument to $\|x^k - x^*\|_2^2$ recursively

$$\|x^{k+1} - x^*\|_2^2 \le \|x^0 - x^*\|_2^2 + \sum\nolimits_{t=1}^{k} \alpha_t^2 \|g^k\|_2^2 - 2 \sum\nolimits_{t=1}^{k} \alpha_t (f^t - f^\star).$$

Now use our convenient assumptions!

Subgradient method – convergence

$$\|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2 \sum_{t=1}^{k} \alpha_t (f^t - f^\star).$$

- To get a bound on the last term, simply notice (for $t \leq k$)

$$f^t \geq f^t_{\min} \geq f^k_{\min} \qquad \text{since} \quad f^t_{\min} := \min_{0 \leq i \leq t} f(x^i)$$

Subgradient method – convergence

$$\|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2 \sum_{t=1}^{k} \alpha_t (f^t - f^\star).$$

- To get a bound on the last term, simply notice (for $t \leq k$)

$$f^t \geq f^t_{\min} \geq f^k_{\min} \qquad \text{since} \quad f^t_{\min} := \min_{0 \leq i \leq t} f(x^i)$$

- Plugging this in yields the bound

$$2 \sum_{t=1}^{k} \alpha_t (f^t - f^\star) \geq 2(f^k_{\min} - f^\star) \sum_{t=1}^{k} \alpha_t.$$

Subgradient method – convergence

$$\|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2\sum_{t=1}^{k} \alpha_t (f^t - f^\star).$$

- To get a bound on the last term, simply notice (for $t \leq k$)

$$f^t \geq f^t_{\min} \geq f^k_{\min} \qquad \text{since} \quad f^t_{\min} := \min_{0 \leq i \leq t} f(x^i)$$

- Plugging this in yields the bound

$$2\sum_{t=1}^{k} \alpha_t (f^t - f^\star) \geq 2(f^k_{\min} - f^\star) \sum_{t=1}^{k} \alpha_t.$$

- So that we finally have

$$0 \leq \|x^{k+1} - x^*\|_2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2(f^k_{\min} - f^\star) \sum_{t=1}^{k} \alpha_t$$

Subgradient method – convergence

$$\|x^{k+1} - x^*\|_2^2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2\sum_{t=1}^{k} \alpha_t (f^t - f^\star).$$

- To get a bound on the last term, simply notice (for $t \leq k$)

$$f^t \geq f^t_{\min} \geq f^k_{\min} \qquad \text{since} \quad f^t_{\min} := \min_{0\leq i\leq t} f(x^i)$$

- Plugging this in yields the bound

$$2\sum_{t=1}^{k} \alpha_t (f^t - f^\star) \geq 2(f^k_{\min} - f^\star) \sum_{t=1}^{k} \alpha_t.$$

- So that we finally have

$$0 \leq \|x^{k+1} - x^*\|_2 \leq R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2 - 2(f^k_{\min} - f^\star) \sum_{t=1}^{k} \alpha_t$$

$$\boxed{f^k_{\min} - f^\star \leq \frac{R^2 + G^2 \sum_{t=1}^{k} \alpha_t^2}{2\sum_{t=1}^{k} \alpha_t}}$$

Subgradient method – convergence

$$\boxed{f_{\min}^k - f^\star \le \frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2 \sum_{t=1}^k \alpha_t}}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^\star$ for the different choices of stepsize that we mentioned.

Subgradient method – convergence

$$\boxed{f_{\min}^k - f^\star \leq \frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2 \sum_{t=1}^k \alpha_t}}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^\star$ for the different choices of stepsize that we mentioned.

Constant step: $\alpha_k = \alpha$; We obtain

$$f_{\min}^k - f^\star \leq \frac{R^2 + G^2 k \alpha^2}{2k\alpha}$$

Subgradient method – convergence

$$\boxed{f_{\min}^k - f^\star \leq \frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2\sum_{t=1}^k \alpha_t}}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^\star$ for the different choices of stepsize that we mentioned.

Constant step: $\alpha_k = \alpha$; We obtain

$$f_{\min}^k - f^\star \leq \frac{R^2 + G^2 k\alpha^2}{2k\alpha} \to \frac{G^2\alpha}{2} \quad \text{as } k \to \infty.$$

Subgradient method – convergence

$$\boxed{f_{\min}^k - f^\star \le \frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2\sum_{t=1}^k \alpha_t}}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^\star$ for the different choices of stepsize that we mentioned.

Constant step: $\alpha_k = \alpha$; We obtain

$$f_{\min}^k - f^\star \le \frac{R^2 + G^2 k\alpha^2}{2k\alpha} \to \frac{G^2\alpha}{2} \quad \text{as } k \to \infty.$$

Square summable, not summable: $\sum_k \alpha_k^2 < \infty$, $\sum_k \alpha_k = \infty$

Subgradient method – convergence

$$\boxed{f_{\min}^k - f^\star \leq \frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2\sum_{t=1}^k \alpha_t}}$$

Exercise: Analyze $\lim_{k\to\infty} f_{\min}^k - f^\star$ for the different choices of stepsize that we mentioned.

Constant step: $\alpha_k = \alpha$; We obtain

$$f_{\min}^k - f^\star \leq \frac{R^2 + G^2 k\alpha^2}{2k\alpha} \to \frac{G^2\alpha}{2} \quad \text{as } k \to \infty.$$

Square summable, not summable: $\sum_k \alpha_k^2 < \infty$, $\sum_k \alpha_k = \infty$
As $k \to \infty$, numerator $< \infty$ but denominator $\to \infty$; so $f_{\min}^k \to f^*$

In practice, fair bit of stepsize tuning needed, e.g. $\alpha_k = a/(b+k)$

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should k be?

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \le \varepsilon$, how big should k be?
- Optimize the bound for α_t

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should k be?
- Optimize the bound for α_t
- We want

$$\frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2 \sum_{t=1}^k \alpha_t} \leq \varepsilon$$

Subgradient method – convergence

- Suppose we want $f_{\min}^k - f^* \leq \varepsilon$, how big should k be?
- Optimize the bound for α_t
- We want
$$\frac{R^2 + G^2 \sum_{t=1}^k \alpha_t^2}{2 \sum_{t=1}^k \alpha_t} \leq \varepsilon$$
- Largest possible $\alpha_t \propto 1/\sqrt{t}$
- Number of steps $k = (RG/\varepsilon)^2 = O(\frac{1}{\varepsilon^2})$

Exercise

Support vector machines

- Let $\mathcal{D} := \{(x_i, y_i) \mid x_i \in \mathbb{R}^n, y_i \in \{\pm 1\}\}$
- We wish to find $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that

$$\min_{w,b} \quad \tfrac{1}{2}\|w\|_2^2 + C\sum\nolimits_{i=1}^{m} \max[0, 1 - y_i(w^T x_i + b)]$$

- Derive and implement a subgradient method
- Plot evolution of objective function
- Experiment with different values of $C > 0$
- Plot and keep track of $f_{\min}^k := \min_{0 \le t \le k} f(x^t)$

Subgradient method – exercise

- Let $a \in \mathbb{R}^n$ be a given vector.
- Let $f(x) = \sum_i |x - a_i|$, i.e., $f : \mathbb{R} \to \mathbb{R}_+$
- Implement different subgradient methods to minimize f
- Also keep track of $f_{\text{best}}^k := \min_{0 \le i < k} f(x_i)$

Subgradient method – exercise

- Let $a \in \mathbb{R}^n$ be a given vector.
- Let $f(x) = \sum_i |x - a_i|$, i.e., $f : \mathbb{R} \to \mathbb{R}_+$
- Implement different subgradient methods to minimize f
- Also keep track of $f_{\text{best}}^k := \min_{0 \le i < k} f(x_i)$

Exercise: Implement the above in Matlab. Report a plot of $f(x_k)$ values; also try to guess what optimum is being found.

♡ *Hint:* Here we can use $\partial(f(x) + g(x)) = \partial f(x) + \partial g(x)$

♡ *Hint:* $|x - c|$ is not diff. at $x = c$; there subgrad is $[-1, 1]$

♡ *Hint:* It might help to try solving this for an integer valued vector a

Polyak's stepsize

- Assume f^* is known (or can be estimated). Then use

$$\alpha_k = \frac{f^k - f^*}{\|g^k\|_2^2}$$

Polyak's stepsize

- Assume $f^\star$ is known (or can be estimated). Then use

$$\alpha_k = \frac{f^k - f^\star}{\|g^k\|_2^2}$$

- Motivation: recall bound

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2\alpha_k(f^k - f^\star) + \alpha_k^2\|g^k\|^2$$

and minimize RHS.

Polyak's stepsize

- Assume $f^\star$ is known (or can be estimated). Then use

$$\alpha_k = \frac{f^k - f^\star}{\|g^k\|_2^2}$$

- Motivation: recall bound

$$\|x^{k+1} - x^*\|^2 \le \|x^k - x^*\|^2 - 2\alpha_k(f^k - f^\star) + \alpha_k^2\|g^k\|^2$$

  and minimize RHS.

- Let's plug in α_k:

$$\|x^{k+1} - x^*\|^2 \le \|x^k - x^*\|^2 - \frac{(f^k - f^\star)^2}{\|g_k\|^2}$$

Polyak's stepsize

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \frac{(f^k - f^\star)^2}{\|g_k\|^2}$$

Polyak's stepsize

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \frac{(f^k - f^\star)^2}{\|g_k\|^2}$$

- **Observation 1** $\|x^k - x^*\|$ decreases

Polyak's stepsize

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \frac{(f^k - f^\star)^2}{\|g_k\|^2}$$

- **Observation 1** $\|x^k - x^*\|$ decreases
- Recursion:

$$\sum_{k=1}^{K} \frac{(f^k - f^\star)^2}{\|g^k\|^2} \quad \leq \quad \|x^1 - x^*\|^2 \quad \leq \quad R^2$$

Polyak's stepsize

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \frac{(f^k - f^\star)^2}{\|g_k\|^2}$$

- **Observation 1** $\|x^k - x^*\|$ decreases
- Recursion:

$$\sum_{k=1}^{K} \frac{(f^k - f^\star)^2}{\|g^k\|^2} \quad \leq \quad \|x^1 - x^*\|^2 \quad \leq \quad R^2$$

- Now use $\|g^k\| \leq G$

$$\sum_{k=1}^{K} (f^k - f^\star)^2 \leq R^2 G^2$$

Polyak's stepsize

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \frac{(f^k - f^\star)^2}{\|g_k\|^2}$$

- **Observation 1** $\|x^k - x^*\|$ decreases
- Recursion:

$$\sum_{k=1}^{K} \frac{(f^k - f^\star)^2}{\|g^k\|^2} \quad \leq \quad \|x^1 - x^*\|^2 \quad \leq \quad R^2$$

- Now use $\|g^k\| \leq G$

$$\sum_{k=1}^{K} (f^k - f^\star)^2 \leq R^2 G^2$$

- **Observation 2** $f^k \to f^\star$

Polyak's stepsize

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - \frac{(f^k - f^\star)^2}{\|g_k\|^2}$$

- **Observation 1** $\|x^k - x^*\|$ decreases
- Recursion:

$$\sum\nolimits_{k=1}^{K} \frac{(f^k - f^\star)^2}{\|g^k\|^2} \;\leq\; \|x^1 - x^*\|^2 \;\leq\; R^2$$

- Now use $\|g^k\| \leq G$

$$\sum\nolimits_{k=1}^{K} (f^k - f^\star)^2 \leq R^2 G^2$$

- **Observation 2** $f^k \to f^\star$
- for accuracy ε, need $K = (RG/\varepsilon)^2$

Nonsmooth convergence rates

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$

Nonsmooth convergence rates

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.

Nonsmooth convergence rates

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$

Nonsmooth convergence rates

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial |x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.

Nonsmooth convergence rates

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.
- This behavior typical for the subgradient method which exhibits $O(1/\sqrt{k})$ convergence in general

Nonsmooth convergence rates

- Let $\phi(x) = |x|$ for $x \in \mathbb{R}$
- Subgradient method $x^{k+1} = x^k - \alpha_k g^k$, where $g^k \in \partial|x^k|$.
- If $x^0 = 1$ and $\alpha_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}$ (this stepsize is known to be optimal), then $|x^k| = \frac{1}{\sqrt{k+1}}$
- Thus, $O(\frac{1}{\epsilon^2})$ iterations are needed to obtain ϵ-accuracy.
- This behavior typical for the subgradient method which exhibits $O(1/\sqrt{k})$ convergence in general

Can we do better in general?

Nonsmooth convergence rates

Theorem. (Nesterov.) Let $\mathcal{B} = \{x \mid \|x - x^0\|_2 \le D\}$. Assume, $x^* \in \mathcal{B}$. There exists a convex function f in $C_L^0(\mathcal{B})$ (with $L > 0$), such that for $0 \le k \le n-1$, the lower-bound

$$f(x^k) - f(x^*) \ge \frac{LD}{2(1+\sqrt{k+1})},$$

holds for **any algorithm** that generates x^k by linearly combining the previous iterates and subgradients.

Exercise: So design problems where we can do better!