Math 273a: Optimization
Subgradient Methods

Instructor: Wotao Yin
Department of Mathematics, UCLA

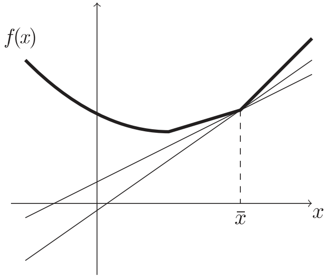
Fall 2015

online discussions on piazza.com

Nonsmooth convex function

Recall: For $\bar{\mathbf{x}} \subseteq \mathbb{R}^n$,

$$\partial f(\bar{\mathbf{x}}) := \{\mathbf{g} \in \mathbb{R}^n : f(\mathbf{y}) \geq f(\bar{\mathbf{x}}) + \langle \mathbf{g}, \mathbf{y} - \bar{\mathbf{x}} \rangle\}.$$



- If $f(\mathbf{x})$ is differentiable, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$ is a singleton.
- For any $\mathbf{x}$, $\mathbf{y}$, $\mathbf{g}_\mathbf{x} \in \partial f(\mathbf{x})$, and $\mathbf{g}_\mathbf{y} \in \partial f(\mathbf{y})$, we have

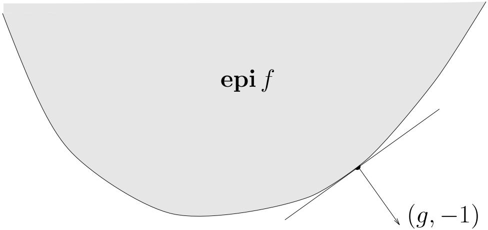
$$\langle \mathbf{g}_\mathbf{x} - \mathbf{g}_\mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \geq 0. \quad (\partial f \text{ is a monotonic point-to-set map})$$

Which functions have subgradients?

Theorem (Nesterov Thm 3.1.13)

*Let f be a **closed convex function and** $x_0 \in \operatorname{int}(\mathbf{dom}(f))$. Then $\partial f(x_0)$ is a nonempty bounded set.*

- Proof uses supporting hyperplanes of epigraph to show existence, and local Lipschitz continuity of convex functions to show boundedness.



1

---

[1] This and next Slide taken from Damek's lecture. Figure taken from Boyd and Vandenberghe, http://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf.

The converse

Lemma (Nesterov Lm 3.1.6)

If $\partial f(x) \neq \emptyset$ for all $x \in \mathbf{dom}(f)$, then f is convex.

Proof.

- $x, y \in \mathbf{dom}(f)$, $\alpha \in [0,1]$, $y_\alpha = (1-\alpha)x + \alpha y = x + \alpha(y-x)$, $g \in \partial f(y_\alpha)$.

$$f(y) \geq f(y_\alpha) + \langle g, y - y_\alpha \rangle = f(y_\alpha) + (1-\alpha)\langle g, y - x \rangle \tag{1}$$

$$f(x) \geq f(y_\alpha) + \langle g, x - y_\alpha \rangle = f(y_\alpha) - \alpha\langle g, y - x \rangle \tag{2}$$

- Multiply equation (1) by α and equation (2) by $(1-\alpha)$.
- Add them together to get

$$\alpha f(y) + (1-\alpha) f(x) \geq f(y_\alpha).$$

□

Compute subgradients: general rules

- **Differentiable functions:** $\partial f(x) = \{\nabla f(x)\}$.
- **Composition with affine map:** $\phi(x) = f(A(x) + b)$

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

- **Positive sums:** $\alpha, \beta > 0$, $f(x) = \alpha f_1(x) + \beta f_2(x)$.

$$\partial f(x) = \alpha \partial f_1(x) + \beta \partial f_2(x)$$

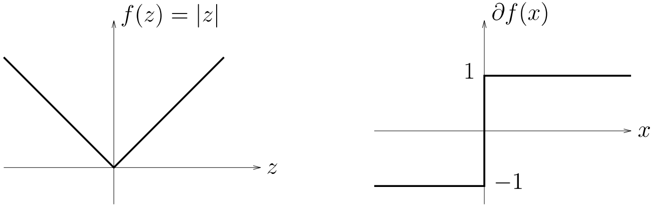
- **Maximums:** $f(x) = \max_{i \in \{1,\cdots,n\}} \{f_i(x)\}$

$$\partial f(x) = \mathrm{conv}\{\partial f_i(x) \mid f_i(x) = f(x)\}$$

Examples

- $f(x) = |x|$.

$$\partial f(x) \quad = \quad \begin{cases} \{\mathrm{sign}(x)\} & x \neq 0; \\ [-1, 1] & \text{otherwise} \end{cases}$$



[2]

[2] figure taken from Boyd and Vandenberghe,
http://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf.

Examples

- $f(\mathbf{x}) = \sum_{i=1}^{n} |\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i|$. Define

$$I_-(\mathbf{x}) = \{i | \langle \mathbf{a}_i, \mathbf{x} \rangle - b_i < 0\}$$
$$I_+(\mathbf{x}) = \{i | \langle \mathbf{a}_i, \mathbf{x} \rangle - b_i > 0\}$$
$$I_0(\mathbf{x}) = \{i | \langle \mathbf{a}_i, \mathbf{x} \rangle - b_i = 0\}.$$

Then

$$\partial f(\mathbf{x}) = \sum_{i \in I_+(\mathbf{x})} a_i - \sum_{i \in I_-(\mathbf{x})} a_i + \sum_{i \in I_0(\mathbf{x})} [-a_i, -a_i]$$

- $f(\mathbf{x}) = \max_{i \in \{1, \cdots, n\}} x_i$. Then

$$\partial f(\mathbf{x}) = \mathrm{conv}\{\mathbf{e}_i | x_i = f(\mathbf{x})\}$$
$$\partial f(0) = \mathrm{conv}\{\mathbf{e}_i | i \in \{1, \cdots, n\}\}$$

Examples

- $f(\mathbf{x}) = \|\mathbf{x}\|_2$. f is differential away from 0, so:

$$\partial f(\mathbf{x}) \quad = \quad \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \quad \mathbf{x} \neq 0.$$

At 0, go back to subgradient equation:

$$\|\mathbf{y}\|_2 \quad \geq \quad 0 + \langle \mathbf{g}, \mathbf{y} - 0 \rangle$$

Thus, $\mathbf{g} \in \partial f(0)$, if, and only if, $\frac{\langle \mathbf{g}, \mathbf{y} \rangle}{\|\mathbf{y}\|_2} \leq 1$ for all $\mathbf{y} \neq 0$. Thus, $\mathbf{g}$ is in the dual ball $B_2^*(0,1) = B_2(0,1)$.

- This is a common pattern!

Examples

- $f(\mathbf{x}) = \|\mathbf{x}\|_\infty = \max_{i\in\{1,\cdots,n\}} |x^{(i)}|$.

$$\partial f(\mathbf{x}) \quad = \quad \mathrm{conv}\{g^{(i)} \mid g^{(i)} \in \partial|x^{(i)}|, |x^{(i)}| = f(x)\}, \quad \mathbf{x} \neq 0.$$

Going back to subgradient equation

$$\|\mathbf{y}\|_\infty \quad \geq \quad 0 + \langle \mathbf{g}, \mathbf{y} \rangle$$

Thus, $\mathbf{g} \in \partial f(0)$, if, and only if, $\frac{\langle \mathbf{g},\mathbf{y}\rangle}{\|\mathbf{y}\|_\infty} \leq 1$ for all $\mathbf{y} \neq 0$. Thus, $\partial f(0)$ is the dual ball to the l_∞ norm: $B_1(0,1)$.

Examples

- $f(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$. Then

$$\partial f(\mathbf{x}) \quad = \quad \sum_{x_i > 0} \mathbf{e}_i - \sum_{x_i < 0} \mathbf{e}_i + \sum_{x_i = 0} [-\mathbf{e}_i, \mathbf{e}_i]$$

for all $\mathbf{x}$. Then

$$\partial f(0) = \sum_{i=1}^{n} [-\mathbf{e}_i, \mathbf{e}_i] = B_\infty(0, 1).$$

Optimality condition: 0 subgradient $\iff$ minimum

- Suppose that $0 \in \partial f(\mathbf{x})$.
- If f is smooth and convex, $0 \in \partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\} \implies \nabla f(\mathbf{x}) = 0$.
- In general: If $0 \in \partial f(\mathbf{x})$, then

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle 0, \mathbf{y} - \mathbf{x} \rangle \\ &= f(\mathbf{x}) \end{aligned}$$

for all $\mathbf{y} \in \mathbb{R}^n$.

- $\implies$ $\mathbf{x}$ is a minimum!
- Converse also true: $f(\mathbf{y}) \geq f(\mathbf{x}^*) + 0 = f(\mathbf{x}^*) + \langle 0, \mathbf{y} - \mathbf{x}^* \rangle$.

The subgradient method

Iteration:

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

where $\mathbf{g}^k \in \partial f(\mathbf{x}^k)$.

Questions:

- Applications?
- Are $f(\mathbf{x}^k)$ and $\|\mathbf{x}^k - \mathbf{x}^*\|$ monotonic?
- How to choose α_k?

Applications

- Finding a point in the intersection of closed convex sets

$$\text{minimize } f(\mathbf{x}) = \max\{\text{dist}(\mathbf{x}, C_1), \ldots, \text{dist}(\mathbf{x}, C_1)\}$$

  Subgradient: if $f(\mathbf{x}) = \text{dist}(\mathbf{x}, C_j)$ and $\mathbf{x} \notin C_j$, then

$$\mathbf{g} = \frac{\mathbf{x} - \text{Proj}_{C_j}(\mathbf{x})}{\|\mathbf{x} - \text{Proj}_{C_j}(\mathbf{x})\|} \in \partial_{\mathbf{x}} \text{dist}(\mathbf{x}, C_j).$$

- Minimizing non-smooth convex functions, e.g., piece-wise linear convex functions
- Dual subgradient method (generalizes the Uzawa algorithm), dual decomposition

(more to come in this lecture)

Convergence overview

- Typically, convergence is established by identifying a monotonically nonincreasing sequence, such as $f(\mathbf{x}^k) - f^*$ and $\|\mathbf{x}^k - \mathbf{x}^*\|^2$
- However, since the subgradient $g(\mathbf{x})$ is not continuous in $\mathbf{x}$, neither sequence is monotonic
- Instead, we will define the *total descent* and use it to bound $f(\mathbf{x}^k)$
- The choice of step sizes α_k is critical.

Monotonicity of $f(\mathbf{x}^k)$?

- The definition

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \quad \forall \mathbf{g} \in \partial f(\mathbf{x})$$

  yields

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \langle \mathbf{g}^{k+1}, \mathbf{x}^k - \mathbf{x}^{k+1} \rangle = f(\mathbf{x}^k) - \alpha_k \langle \mathbf{g}^{k+1}, \mathbf{g}^k \rangle.$$

- It is generally difficult to estimate $\langle \mathbf{g}^{k+1}, \mathbf{g}^k \rangle$ since $\mathbf{g}$ is not continuous. (No matter how close $\mathbf{x}^{k+1}$ is to $\mathbf{x}^k$, their subgradients $\mathbf{g}^{k+1}$ and $\mathbf{g}^k$ can be very different.) Therefore, we cannot guarantee $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$.

- **Note:** Taking the implicit iteration $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \mathbf{g}^{k+1}$ (the proximal method), we would ensure $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$. It is more expensive to compute though.

Monotonicity of $\|\mathbf{x}^k - \mathbf{x}^*\|^2$

- Let us assume that $\mathbf{x}^*$ exists. Then

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

- To have monotonicity: $\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \le \|\mathbf{x}^k - \mathbf{x}^*\|^2$, we need

$$-2\alpha_k \langle \mathbf{g}^k, \mathbf{x}^k - \mathbf{x}^* \rangle + \alpha_k^2 \|\mathbf{g}^k\|^2 \le 0 \iff \langle \mathbf{g}^k, \mathbf{x}^k - \mathbf{x}^* \rangle \ge \frac{\alpha_k}{2} \|\mathbf{g}^k\|^2.$$

- However, even at $\mathbf{x}^k \approx \mathbf{x}^*$, $\mathbf{g}^k$ may not vanish.
- Therefore, $\|\mathbf{x}^k - \mathbf{x}^*\|^2$ is generally not monotonic unless
  - $\|\mathbf{g}^k\| < G$ (commonly assumed for subgradient method), **and**
  - $\alpha_k = O(\|\mathbf{x}^k - \mathbf{x}^*\|)$, which is unrealistic to ensure since $\mathbf{x}^*$ is unknown.

- However, it is often easy to have an estimate on f^*.
  For example, $f^* \geq 0$ in many applications.

- The definition

  $$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \quad \forall \mathbf{g} \in \partial f(\mathbf{x})$$

  yields

  $$-\alpha_k \langle \mathbf{g}^k, \mathbf{x}^k - \mathbf{x}^* \rangle \leq -\alpha_k (f(\mathbf{x}^k) - f^*) \leq 0.$$

  Interpretation: the term $-\alpha_k \langle \mathbf{g}^k, \mathbf{x}^k - \mathbf{x}^* \rangle$ guarantees a sufficient descent by at least $-\alpha_k (f(\mathbf{x}^k) - f^*)$

- However, the ascending term $\alpha_k^2 \|\mathbf{g}^k\|^2$ can be as large as $\alpha_k^2 G^2$.

- After substitution, we get the bound

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

  Telescopic sum over k gives

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 = \|\mathbf{x}^0 - \mathbf{x}^*\|^2 - 2\sum_{i=0}^{k}\alpha_i(f(\mathbf{x}^i) - f^*) + \sum_{i=0}^{k}\alpha_i^2\|\mathbf{g}^i\|^2.$$

- Therefore,

$$\boxed{\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + 2\sum_{i=0}^{k}\alpha_i(f(\mathbf{x}^i) - f^*) \le \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \sum_{i=0}^{k}\alpha_i^2\|\mathbf{g}^i\|^2}$$

  which bounds the *total descent* $\sum_{i=0}^{k}\alpha_k(f(\mathbf{x}^i) - f^*)$ by the *total ascent* $\sum_{i=0}^{k}\alpha_i^2\|\mathbf{g}^i\|^2$. Clearly, α_k play the key role.

Step size and convergence

- By

$$\boxed{\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 + 2\sum_{i=0}^{k} \alpha_i (f(\mathbf{x}^i) - f^*) \le \|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \sum_{i=0}^{k} \alpha_i^2 \|\mathbf{g}^i\|^2}$$

and letting

$$f_{\mathrm{best}}^k = \min\{f(\mathbf{x}^i) : i = 0, 1, \ldots, k\}$$

$$(\text{thus, } f_{\mathrm{best}} - f^* \le f(\mathbf{x}^i) - f^*,\ i \le k)$$

$$\|\mathbf{g}\| \le G \quad (\text{equivalent to Lip. } f\text{: } |f(\mathbf{x}) - f(\mathbf{y})| \le G\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y})$$

we have

$$f_{\mathrm{best}}^k - f^* \le \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2\sum_{i=0}^{k} \alpha_i}.$$

We need *unbounded* $\sum_{i=0}^{k} \alpha_i$ and *bounded* $\sum_{i=0}^{k} \alpha_i$ as $k \to \infty$.

- To have $f_{\text{best}}^k - f^* \to 0$, we require, for example,
  - $\sum_{i=0}^{k} \alpha_i = \infty \quad$ and $\quad \sum_{i=0}^{k} \alpha_i^2 \leq \infty$;
  - or more weakly, $\sum_{i=0}^{k} \alpha_i = \infty \quad$ and $\quad \lim \alpha_k \to 0$. (the truncation trick)

Otherwise, $f_{\text{best}}^k \not\to f^*$ in general.

Fixed step size

- **Fixing** $\alpha_k \equiv \alpha$, we get

$$f_{\text{best}}^k - f^* \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + G^2 \sum_{i=0}^k \alpha^2}{2 \sum_{i=0}^k \alpha} = \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2\alpha(k+1)} + \frac{\alpha G^2}{2}.$$

- $f_{\text{best}}^k - f^* \to \alpha G^2/2 = O(\alpha)$.
- **while in the early stage**

$$k < \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\alpha^2 G^2},$$

  we have

$$\frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2\alpha(k+1)} + \frac{\alpha G^2}{2} \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\alpha(k+1)}$$

  and thus the (non-asymptotic, conditional) rate of convergence $O(\frac{1}{\alpha k})$.
- larger $\alpha \implies$ faster convergence, lower final accuracy
- smaller $\alpha \implies$ slower convergence, higher final accuracy

Fixed step "length" $\alpha_k = \alpha/\|\mathbf{g}^k\|$

- We have

$$f_{\mathrm{best}}^k - f^* \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2 + \sum_{i=0}^k \alpha_k^2 \|\mathbf{g}^k\|^2}{2\sum_{i=0}^k \alpha_k} = \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2 G}{2\alpha(k+1)} + \frac{\alpha G}{2}.$$

- $f_{\mathrm{best}}^k - f^* \to \alpha G/2 = O(\alpha)$, slightly better than with a fixed step size.
- while

$$k < \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{\alpha^2},$$

  we have the (non-asymptotic, conditional) rate of convergence $O(\frac{G}{\alpha k})$.

- larger $\alpha \implies$ faster convergence, lower final accuracy
- smaller $\alpha \implies$ slower convergence, higher final accuracy

Polyak step size

- Assume that f^* is known (not $\mathbf{x}^*$ though). Example: $f^* = 0$ in projection problems.) Set:

$$\alpha_k = \frac{f(\mathbf{x}^k) - f^*}{\|\mathbf{g}^k\|} = \arg\min \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha_k(f(\mathbf{x}^k) - f^*) + \alpha_k^2 \|\mathbf{g}^k\|^2$$

- Then,

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha_k(f(\mathbf{x}^k) - f^*) + \alpha_k^2 \|\mathbf{g}^k\|^2 = \|\mathbf{x}^k - \mathbf{x}^*\|^2 - \frac{(f(\mathbf{x}^k) - f^*)^2}{\|\mathbf{g}^k\|}$$

(so $\|\mathbf{x}^k - \mathbf{x}^*\|^2$ is monotonic) and thus after adding over k,

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}^0 - \mathbf{x}^*\|^2 - \frac{1}{G} \sum_{i=0}^{k} (f(\mathbf{x}^i) - f^*)^2.$$

- Finally,

$$f_{\text{best}}^k - f^* \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\| \sqrt{G}}{\sqrt{k+1}} = O(\frac{1}{\sqrt{k}}).$$

Oracle optimality

For an $O(\frac{1}{\sqrt{k}})$ method to guarantee $f_{\text{best}}^k - f^* \le \epsilon$, we need $O(1/\epsilon^2)$ iterations.

It this tight for the subgradient method? Answer: Yes.

Suppose $\mathbf{x}^{k+1}$ is computed by an *arbitrary* method as

$$\mathbf{x}^{k+1} \in \mathbf{x}^0 + \operatorname{span}\{\mathbf{g}^0, \mathbf{g}^1, \ldots, \mathbf{g}^k\}$$

where the oracle gives

- arbitrary $\mathbf{g}^k \in \partial f(\mathbf{x}^k)$ and
- $f(\mathbf{x}^k)$.

Theorem (Nesterov Thm 3.2.1)

There is a nonsmooth convex function with $\|\mathbf{g}\| \le G$ uniformly so that the above method obeys

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) \ge \frac{\|\mathbf{x}^0 - \mathbf{x}^*\| G}{2(1 + \sqrt{k+1})}.$$

The subgradient algorithm

- f is a proper closed convex function.
- **Problem:** $\text{minimize}_{\mathbf{x}} f(\mathbf{x})$
- **Algorithm:** pick any starting point $\mathbf{x}^1$
  - pick $\mathbf{g}^k \in \partial f(\mathbf{x}^k)$
  - set α_k (α_0/k, fixed size, fixed length, or Polyak)
  - $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k - \alpha_k \mathbf{g}^k$
  - $k \leftarrow k + 1$

  (monitor $f(\mathbf{x}^k)$ and f^k_{best}, especially if using fixed size, fixed length)

Variant: projected subgradient method

- f is a proper closed convex function.
- C is a nonempty closed convex set.
- **Problem:**

$$\operatorname*{minimize}_{\mathbf{x}} f(\mathbf{x}), \quad \text{subject to } x \in C.$$

- pick $\mathbf{g}^k \in \partial f(\mathbf{x}^k)$ and α_k, and apply

$$\mathbf{x}^{k+1} \leftarrow \mathrm{Proj}_C(\mathbf{x}^k - \alpha_k \mathbf{g}^k)$$

- since projection is *nonexpansive*,

$$\|\mathrm{Proj}_C(\mathbf{x}) - \mathrm{Proj}_C(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

  the analysis remains the same.

$$\begin{aligned}
\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 &= \|\mathrm{Proj}_C(\mathbf{x}^k - \alpha_k \mathbf{g}^k) - \mathrm{Proj}_C(\mathbf{x}^*)\|^2 \\
&\leq \|(\mathbf{x}^k - \alpha_k \mathbf{g}^k) - \mathbf{x}^*\|^2 \\
&= \|(\mathbf{x}^k - \mathbf{x}^*) - \alpha_k \mathbf{g}^k\|^2 \\
&= \|\mathbf{x}^k - \mathbf{x}^*\|^2 - 2\alpha_k \langle \mathbf{g}^k, \mathbf{x}^k - \mathbf{x}^* \rangle + \alpha_k^2 \|\mathbf{g}^k\|^2 \\
&= \ldots\ldots
\end{aligned}$$

Summary for subgradient methods

- **Universal**. It handles non-differentiable convex problems and, in particular, the dual problem of linearly constrained convex problems (later lectures)
- **No monotonicity** for either $f(\mathbf{x}^k)$ or $\mathrm{dist}(\mathbf{x}^k, \mathcal{X}^*)$ except for Polyak's step size (requiring known f^*)
- Convergence relies on **uniformly bounded subgradient** (or Lipschitz f)
- Rate of convergence $f^k_{\mathrm{best}} - f^*$ depends on the **step size**
  - Constant step size (or length) does *not* guarantee $f^k_{\mathrm{best}} \to f^*$
  - If we need $f^k_{\mathrm{best}} \to f^*$, use diminishing step sizes; the rate is at best $O(1/\sqrt{k})$
- Convergence is quite slow (but the method is widely applicable)
- Some non-smooth problems have better rates by other methods, e.g., prox-linear iteration, operator splitting, dual smoothing (later lectures)