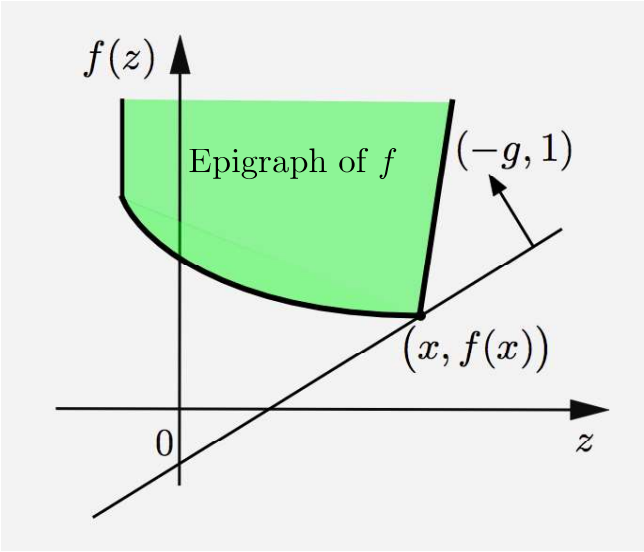
LECTURE 15

LECTURE OUTLINE

- Overview of properties of subgradients
- Subgradient methods
- Convergence analysis
- Reading: Section 2.2 of Convex Optimization Algorithms

SUBGRADIENTS - REAL-VALUED FUNCTIONS



- Let $f : \Re^n \mapsto (-\infty, \infty]$ be a convex function. A vector $g \in \Re^n$ is a subgradient of f at a point $x \in \text{dom}(f)$ if

$$f(z) \geq f(x) + (z - x)'g, \qquad \forall\ z \in \Re^n$$

The set of subgradients at x is the subdifferential $\partial f(x)$.

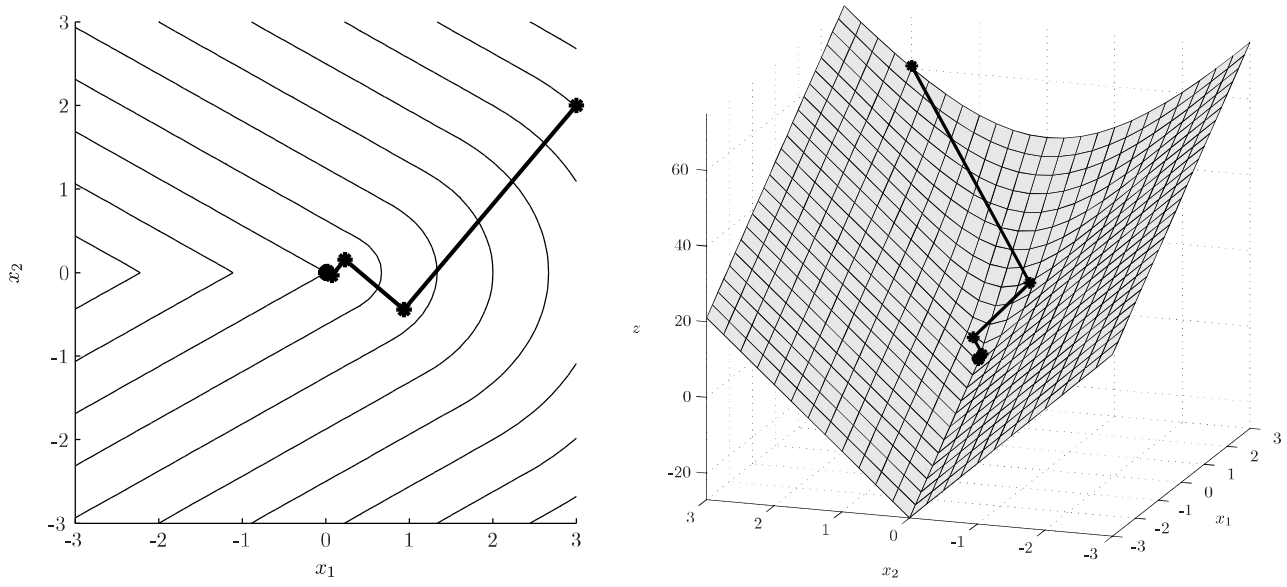
- If f is real-valued, $\partial f(x)$ is nonempty, convex, and compact for all x.

- $\cup_{x \in X} \partial f(x)$ is bounded if X is bounded.

- A single subgradient $g \in \partial f(x)$ is much easier to calculate than $\partial f(x)$ for many contexts involving dual functions and minimax.

MOTIVATION

- Consider minimization of convex f.
- Steepest descent at a point requires knowledge of the entire subdifferential at a point.
- Convergence failure of steepest descent



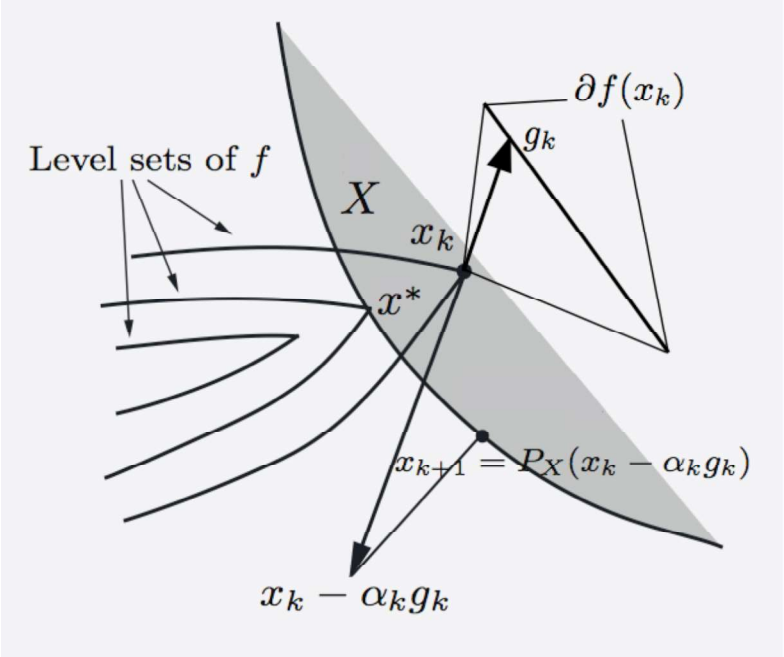
- Subgradient methods abandon the idea of computing the full subdifferential to effect cost function descent ...
- Move instead along the direction of an arbitrary subgradient

THE BASIC SUBGRADIENT METHOD

- Problem: Minimize convex function $f : \Re^n \mapsto \Re$ over a closed convex set X.
- Subgradient method:

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

where g_k is any subgradient of f at x_k, α_k is a positive stepsize, and $P_X(\cdot)$ is projection on X.



- Key geometric fact: By subgradient inequality

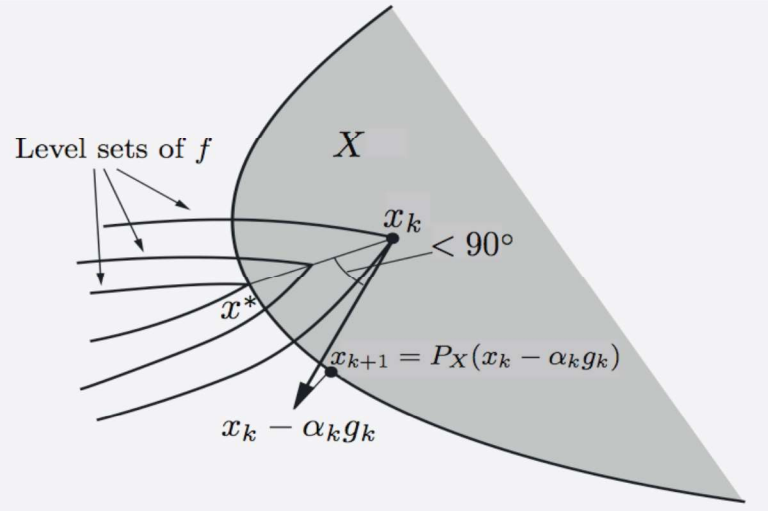
$$g_k'(y - x_k) < 0, \qquad \forall\ y \text{ such that } f(y) < f(x_k),$$

including all optimal y.

KEY PROPERTIES OF SUBGRADIENT METHOD

- For a small enough stepsize α_k, it reduces the Euclidean distance to the optimum.
- Nonexpansiveness of projection: For all x, y,

$$\|P_X(x) - P_X(y)\| \leq \|x - y\|$$



- Proposition: Let $\{x_k\}$ be generated by the subgradient method. Then, for all $y \in X$ and k:

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

and if $f(y) < f(x_k)$,

$$\|x_{k+1} - y\| < \|x_k - y\|,$$

for all α_k such that $0 < \alpha_k < \dfrac{2\big(f(x_k) - f(y)\big)}{\|g_k\|^2}$.

PROOF

• Proof of nonexpansive property: From the projection theorem

$$\big(z - P_X(x)\big)'\big(x - P_X(x)\big) \leq 0, \qquad \forall\ z \in X,$$

from which $\big(P_X(y) - P_X(x)\big)'\big(x - P_X(x)\big) \leq 0$. Similarly, $\big(P_X(x) - P_X(y)\big)'\big(y - P_X(y)\big) \leq 0$.

Adding and using the Schwarz inequality,

$$\begin{aligned}\big\|P_X(y) - P_X(x)\big\|^2 &\leq \big(P_X(y) - P_X(x)\big)'(y - x)\\ &\leq \big\|P_X(y) - P_X(x)\big\| \cdot \|y - x\|\end{aligned}$$

Q.E.D.

• Proof of proposition: Since projection is nonexpansive, we obtain for all $y \in X$ and k,

$$\begin{aligned}\|x_{k+1} - y\|^2 &= \big\|P_X\left(x_k - \alpha_k g_k\right) - y\big\|^2\\ &\leq \|x_k - \alpha_k g_k - y\|^2\\ &= \|x_k - y\|^2 - 2\alpha_k g_k'(x_k - y) + \alpha_k^2\|g_k\|^2\\ &\leq \|x_k - y\|^2 - 2\alpha_k\big(f(x_k) - f(y)\big) + \alpha_k^2\|g_k\|^2,\end{aligned}$$

where the last inequality follows from the subgradient inequality. **Q.E.D.**

CONVERGENCE MECHANISM

- Assume constant stepsize: $\alpha_k \equiv \alpha$
- Assume that $\|g_k\| \leq c$ for some c and all k.

Then for all optimal x^*

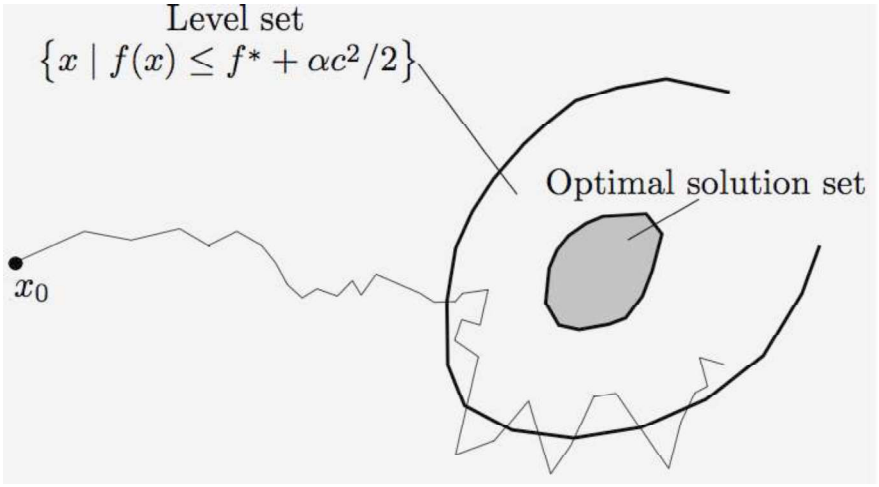
$$\|x_{k+1}-x^*\|^2 \leq \|x_k-x^*\|^2-2\alpha\big(f(x_k)-f(x^*)\big)+\alpha^2c^2$$

Thus the distance to the optimum decreases if

$$0 < \alpha < \frac{2\big(f(x_k) - f(x^*)\big)}{c^2}$$

or equivalently, if x_k does not belong to the level set

$$\left\{ x \;\middle|\; f(x) < f(x^*) + \frac{\alpha c^2}{2} \right\}$$



STEPSIZE RULES

- Constant Stepsize: $\alpha_k \equiv \alpha$.
- Diminishing Stepsize: $\alpha_k \to 0$, $\sum_k \alpha_k = \infty$
- Dynamic Stepsize:

$$\alpha_k = \frac{f(x_k) - f_k}{c^2}$$

where f_k is an estimate of f^*:

  - If $f_k = f^*$, makes progress at every iteration. If $f_k < f^*$ it tends to oscillate around the optimum. If $f_k > f^*$ it tends towards the level set $\{x \mid f(x) \leq f_k\}$.
  - f_k can be adjusted based on the progress of the method.

- Example of dynamic stepsize rule:

$$f_k = \min_{0 \leq j \leq k} f(x_j) - \delta_k,$$

and δ_k (the "aspiration level of cost reduction") is updated according to

$$\delta_{k+1} = \begin{cases} \rho\delta_k & \text{if } f(x_{k+1}) \leq f_k, \\ \max\{\beta\delta_k, \delta\} & \text{if } f(x_{k+1}) > f_k, \end{cases}$$

where $\delta > 0$, $\beta < 1$, and $\rho \geq 1$ are fixed constants.

SAMPLE CONVERGENCE RESULTS

- Let $\overline{f} = \inf_{k \geq 0} f(x_k)$, and assume that for some c, we have

$$c \geq \sup\{\|g\| \mid g \in \partial f(x_k), \ k \geq 0\}$$

- Proposition: Assume that α_k is fixed at some positive scalar α. Then:

  (a) If $f^* = -\infty$, then $\overline{f} = f^*$.

  (b) If $f^* > -\infty$, then

$$\overline{f} \leq f^* + \frac{\alpha c^2}{2}.$$

- Proposition: If α_k satisfies

$$\lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=0}^{\infty} \alpha_k = \infty,$$

then $\overline{f} = f^*$.

- Similar propositions for dynamic stepsize rules.
- Many variants ...

CONVERGENCE METHODOLOGY I

• Classical Contraction Mapping Theorem: Consider iteration $x_{k+1} = G(x_k)$, where $G : \Re^n \mapsto \Re^n$ is a contraction, i.e., for some $\rho < 1$

$$\big\|G(x) - G(y)\big\| \leq \rho\|x - y\|, \qquad \forall\ x, y \in \Re^n$$

where $\|\cdot\|$ is any norm. It converges to the unique fixed point of G.

• Can be used for gradient iterations with constant stepsize, but not subgradient iterations.

• Consider time varying contraction iteration $x_{k+1} = G_k(x_k)$, where

$$\big\|G_k(x)-G_k(y)\big\| \leq (1-\rho_k)\|x-y\|, \qquad \forall\ x, y \in \Re^n,$$

the G_k have a common fixed point, and

$$\rho_k \in (0, 1], \qquad \sum_{k=0}^{\infty} \rho_k = \infty$$

It converges to the unique common fixed point of G_k.

• Can be used for some time-varying gradient iterations, but not subgradient iterations.

CONVERGENCE METHODOLOGY II

• Supermartingale convergence (deterministic case): Let $\{Y_k\}$, $\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ be four nonnegative scalar sequences such that

$$Y_{k+1} \leq (1 + V_k)Y_k - Z_k + W_k, \qquad \forall\ k,$$

and

$$\sum_{k=0}^{\infty} W_k < \infty, \qquad \sum_{k=0}^{\infty} V_k < \infty$$

Then $\{Y_k\}$ converges and $\sum_{k=0}^{\infty} Z_k < \infty$.

• Supermartingale convergence (stochastic case): Let $\{Y_k\}$, $\{Z_k\}$, $\{W_k\}$, and $\{V_k\}$ be four nonnegative sequences of random variables, and let $\mathcal{F}_k$, $k = 0, 1, \ldots$, be sets of random variables such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Assume that

(1) For each k, Y_k, Z_k, W_k, and V_k are functions of the random variables in $\mathcal{F}_k$.

(2) $E\{Y_{k+1} \mid \mathcal{F}_k\} \leq (1 + V_k)Y_k - Z_k + W_k\ \forall\ k$

(3) There holds, $\sum_{k=0}^{\infty} W_k < \infty$, $\sum_{k=0}^{\infty} V_k < \infty$

Then $\{Y_k\}$ converges to some random variable Y, and $\sum_{k=0}^{\infty} Z_k < \infty$, with probability 1.

CONVERGENCE FOR DIMINISHING STEPSIZE

• Proposition: Assume that the optimal solution set X^* is nonempty, and that for some c and all k,

$$c^2 \left(1 + \min_{x^* \in X^*} \|x_k - x^*\|^2 \right) \geq \sup\{ \|g\|^2 \mid g \in \partial f(x_k) \}$$

and α_k satisfies

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Then $\{x_k\}$ converges to an optimal solution.

Proof: Write for any optimal x^*

$$\begin{aligned} \|x_{k+1} - x^*\|^2 \leq (1 + \alpha_k^2 c^2) \|x_k - x^*\|^2 \\ - 2\alpha_k \big(f(x_k) - f(x^*) \big) \\ + \alpha_k^2 c^2 \end{aligned}$$

Use the supermartingale convergence theorem.