Math 273a: Optimization
Cutting plane and bundle methods

Instructor: Wotao Yin

Department of Mathematics, UCLA

Fall 2015

online discussions on piazza.com

Subgradient method

Iteration:

$$x^{k+1} \leftarrow x^k - \alpha^k p^k$$

where $p^k \in \partial f(x^k)$.

Applications:

- find $x^* \in \bigcap_{i=1}^m C_i$ by $\min f(x) = \max\{\mathrm{dist}(x, C_1), \ldots, \mathrm{dist}(x, C_m)\}$
- minimize non-smooth convex functions, e.g., SVM with hinge loss
- dual ascent method (typically, non-smooth), dual decomposition

Step size and convergence: assumption $\|p^k\| \leq G$ uniformly

- fix $\alpha^k \equiv \alpha$. While $k < O(\frac{1}{\alpha^2 G^2})$, $f^k_{\text{best}} - f^* \leq O(\frac{1}{\alpha k})$.
  Larger $\alpha \Rightarrow$ faster, less accurate. Smaller $\alpha \Rightarrow$ slower, more accurate.
- diminishing α_k: $\lim \alpha_k \to 0$ and $\sum_k \alpha_k = \infty$, then $f^k_{\text{best}} - f^* \leq O(\frac{1}{\sqrt{k}})$.

A negative subgradient may not be a descent direction!

Consider

$$f(x) = |x_1| + 2|x_2|$$

At $x = (1, 0)$,

$$\partial f(x) = \{(1, \alpha)^T : \alpha \in [-2, 2]\}.$$

$d = -(1, 2)^T \in -\partial f(x)$ but for any small $\alpha > 0$,

$$f(x + \alpha d) = |1 - \alpha| + 2|\alpha| > 1 = f(x).$$

Consequences:

- iterative monotonicity of f^k is not generally guaranteed
- line search (highly effective in gradient descent) may not help here

Cutting plane

Theorem (Nesterov'03 Thm 3.1.16)

Let $x_0 \in \mathbb{R}^n$. Then all $g \in \partial f(x_0)$ define supporting hyperplanes to the lower level set $\mathcal{L}_f(f(x_0)) = \{x | f(x) < f(x_0)\}$:

$$\langle g, x_0 - x \rangle \geq f(x_0) - f(x) \geq 0.$$

- Thus, each $g \in \partial f(x_0)$ cuts the search space for x^* in half:

$$\langle g, x_0 - x^* \rangle \geq 0$$

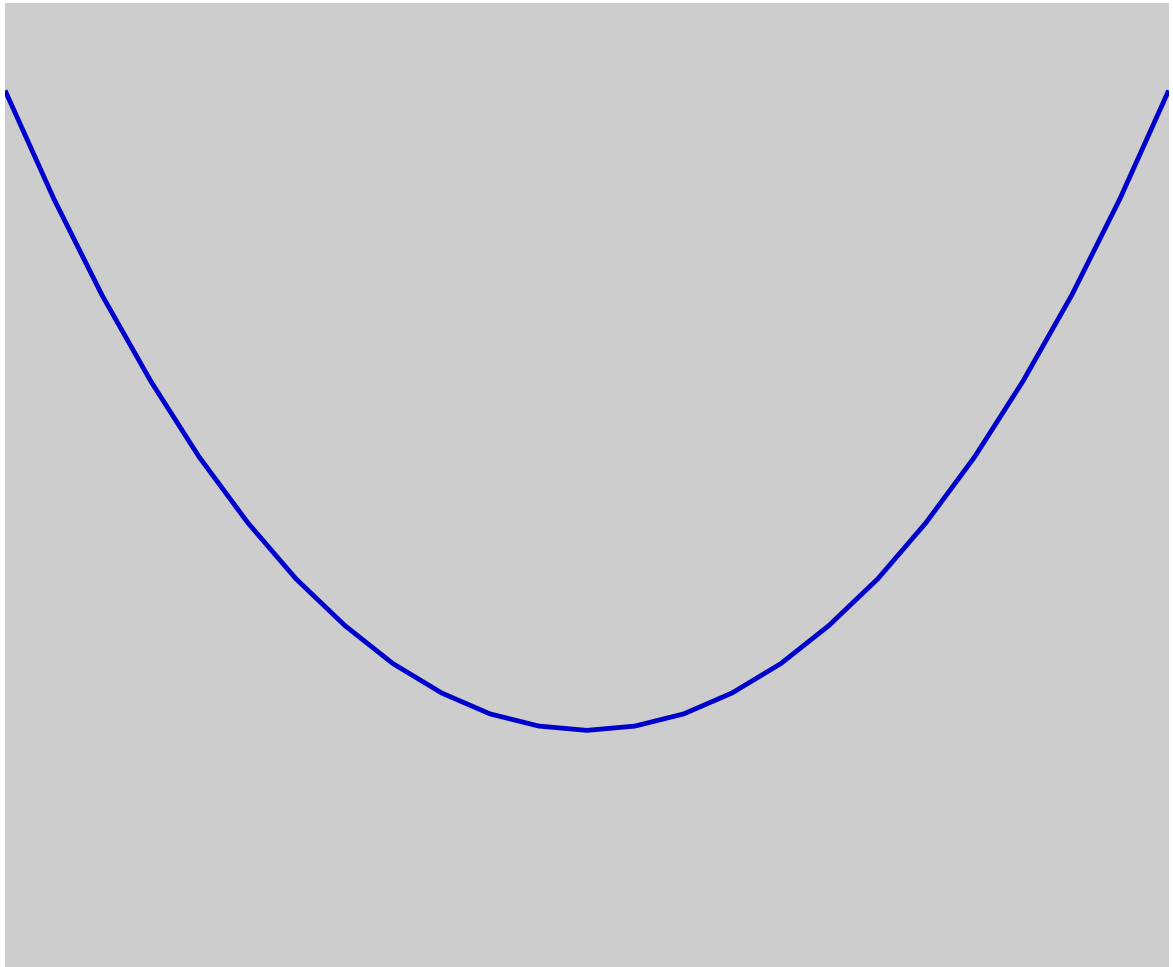
Theorem

Define $\mathcal{H}_f = \{$affine function h such that $h(x) \leq f(x)\ \forall\ x \in \mathbb{R}^n\}$. Then

$$f(x) = \sup\{h(x) : h \in \mathcal{H}_f\}.$$

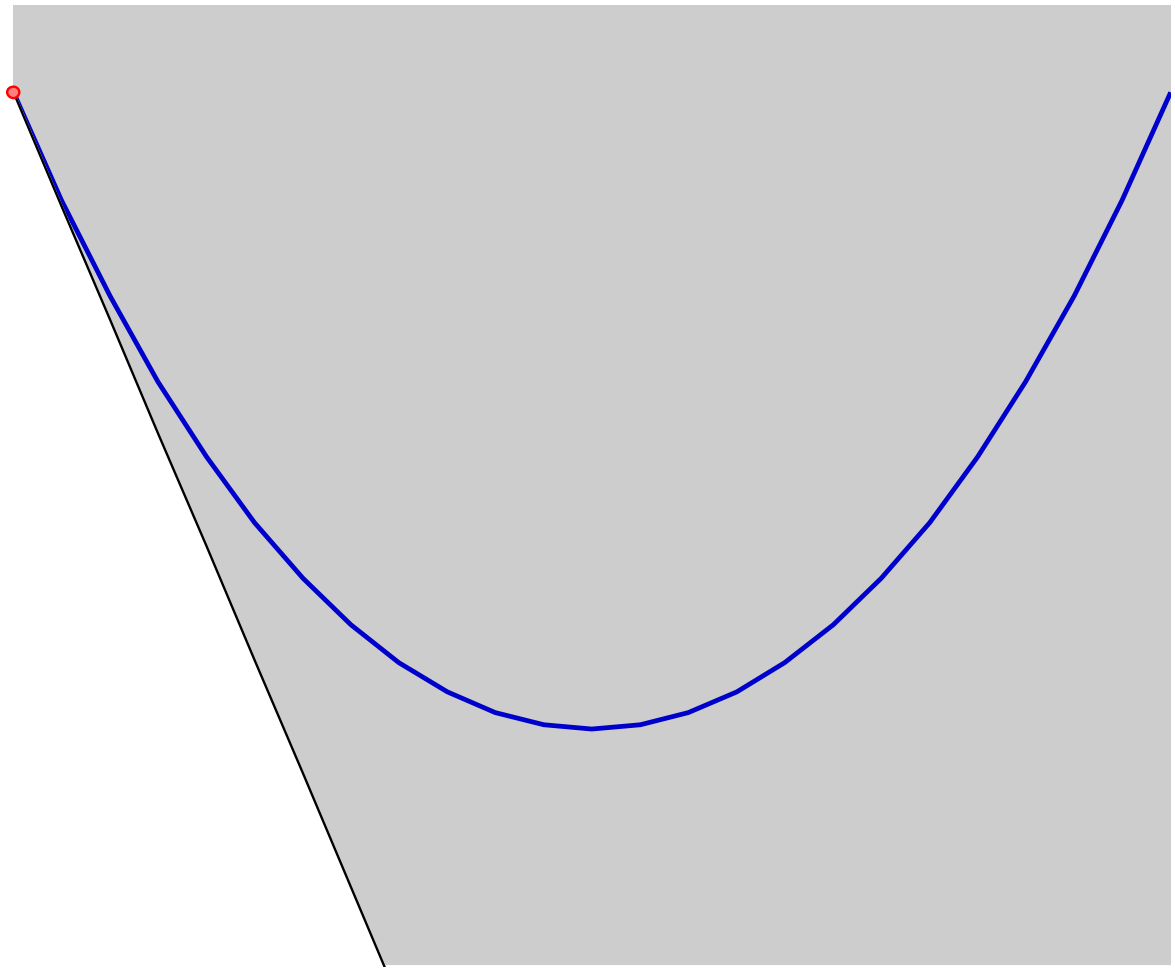
- These results motivate *the cutting plane method*.

Cutting plane method: a demonstration[1]



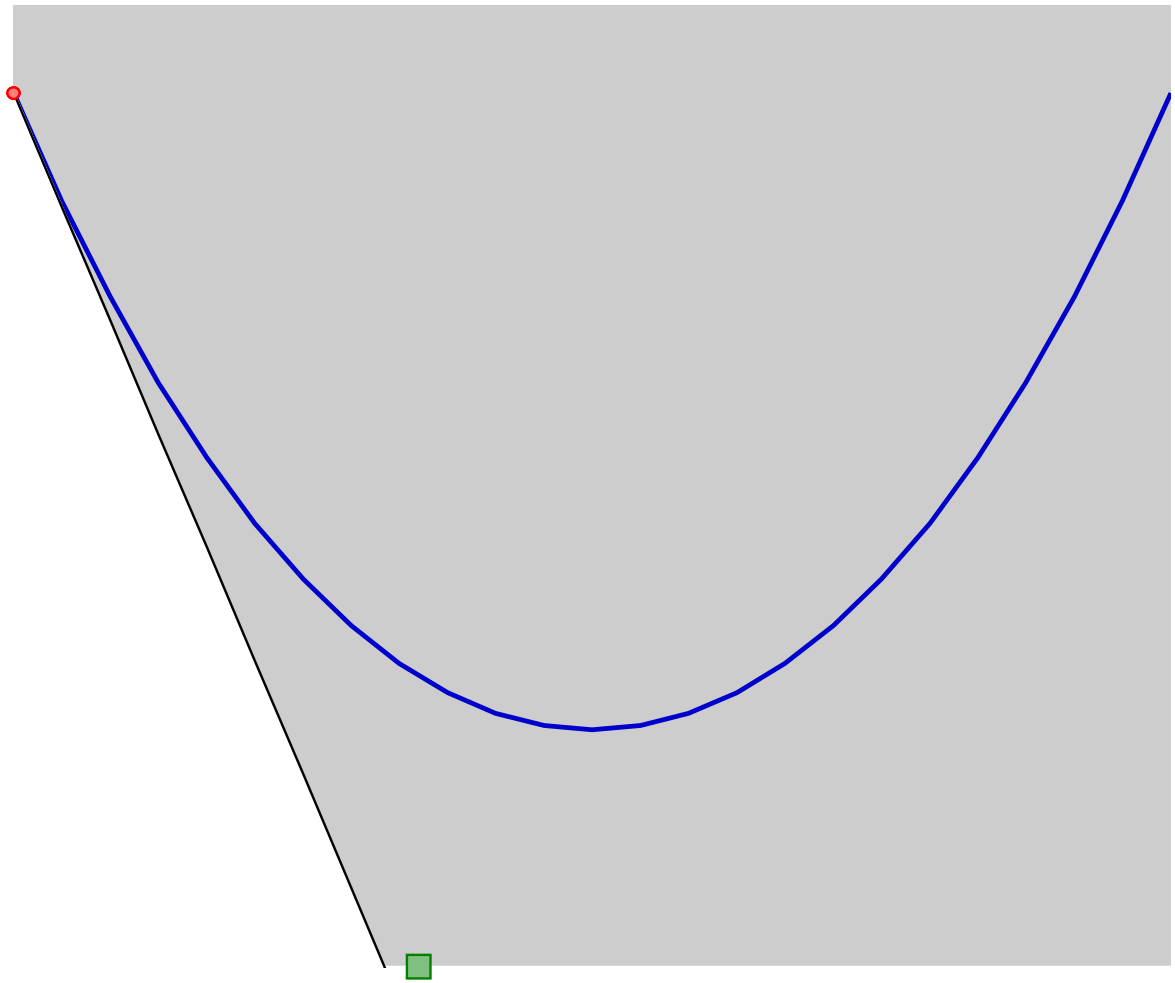
[1] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Cutting plane method: a demonstration[2]



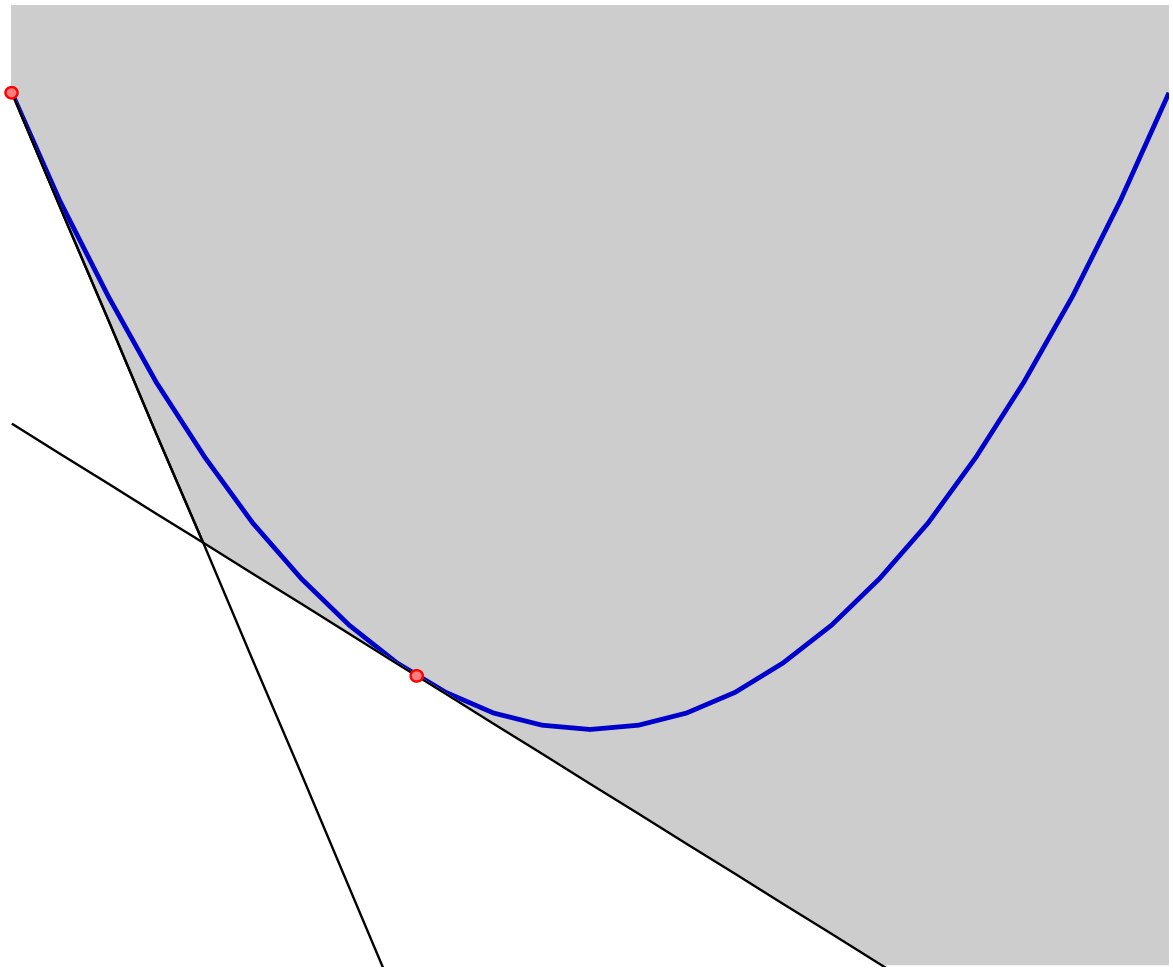
[2] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Cutting plane method: a demonstration[3]



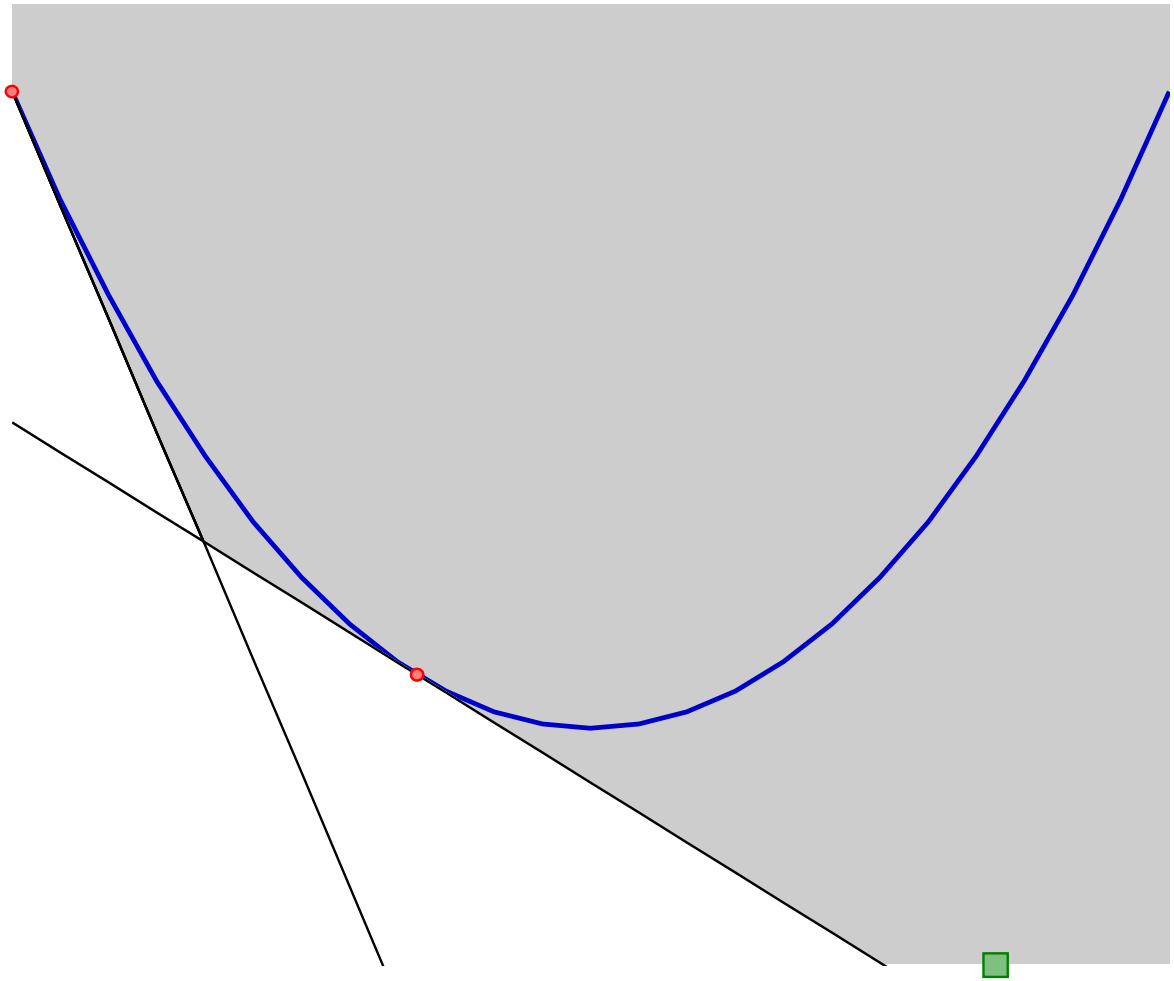
[3] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Cutting plane method: a demonstration[4]



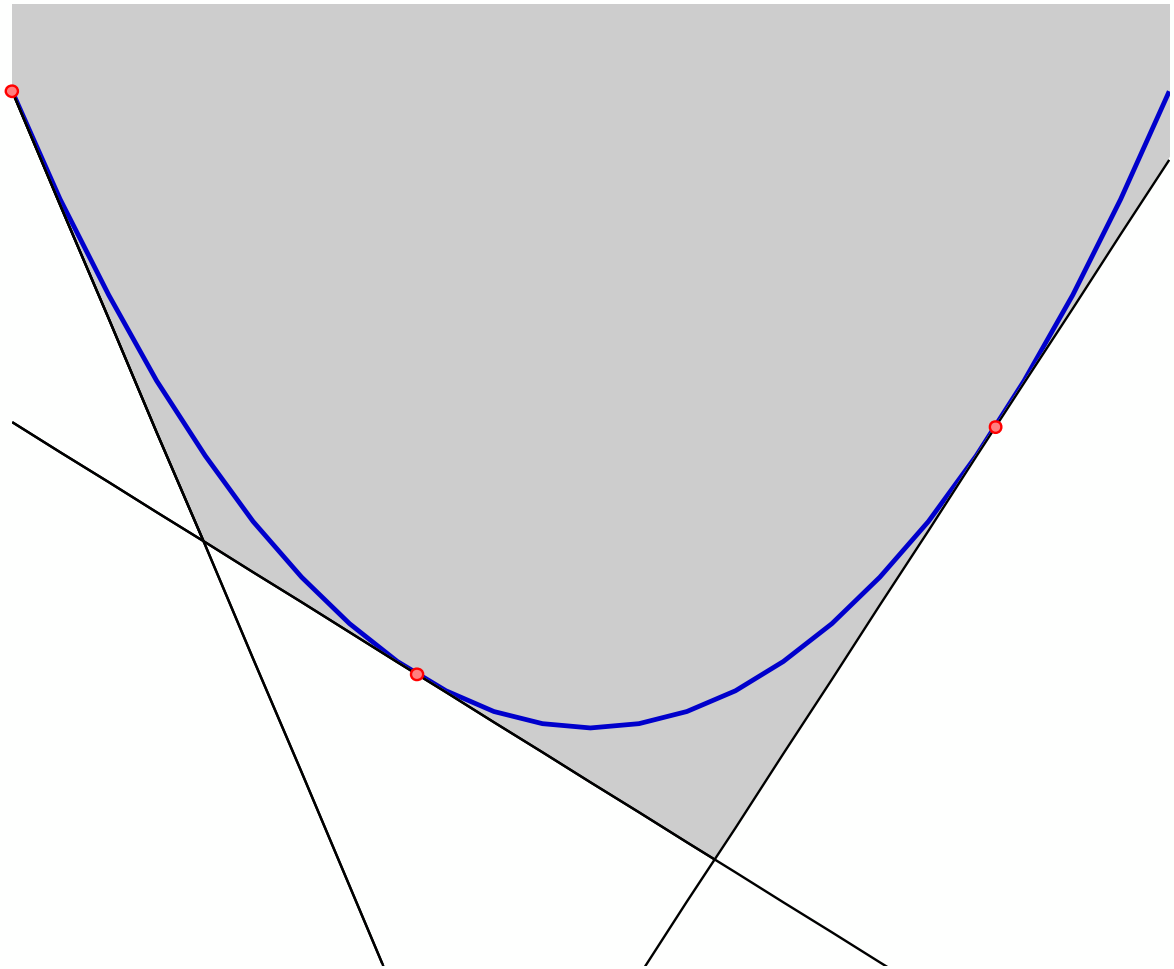
[4] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Cutting plane method: a demonstration[5]



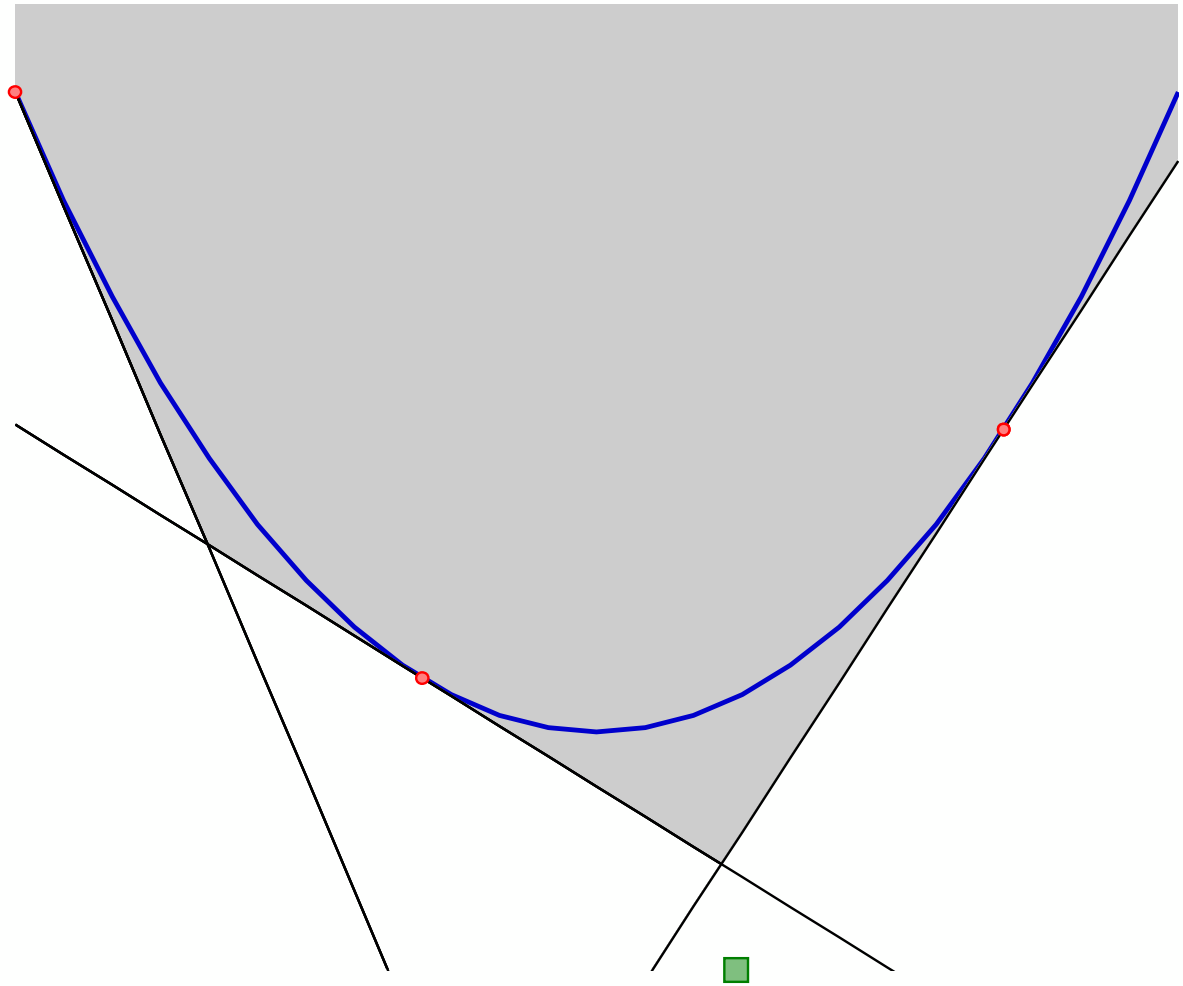
[5] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Cutting plane method: a demonstration[6]



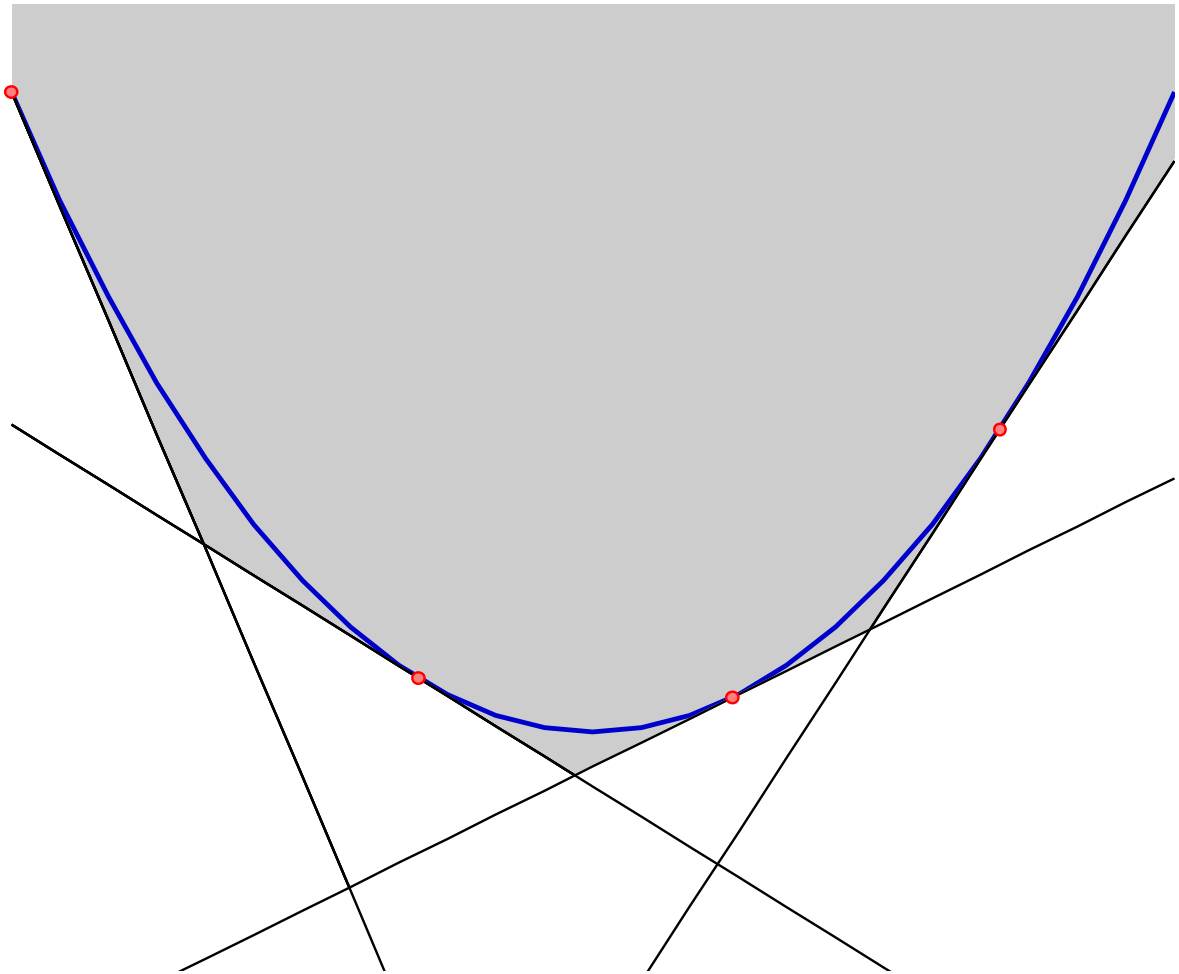
[6] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Cutting plane method: a demonstration[7]



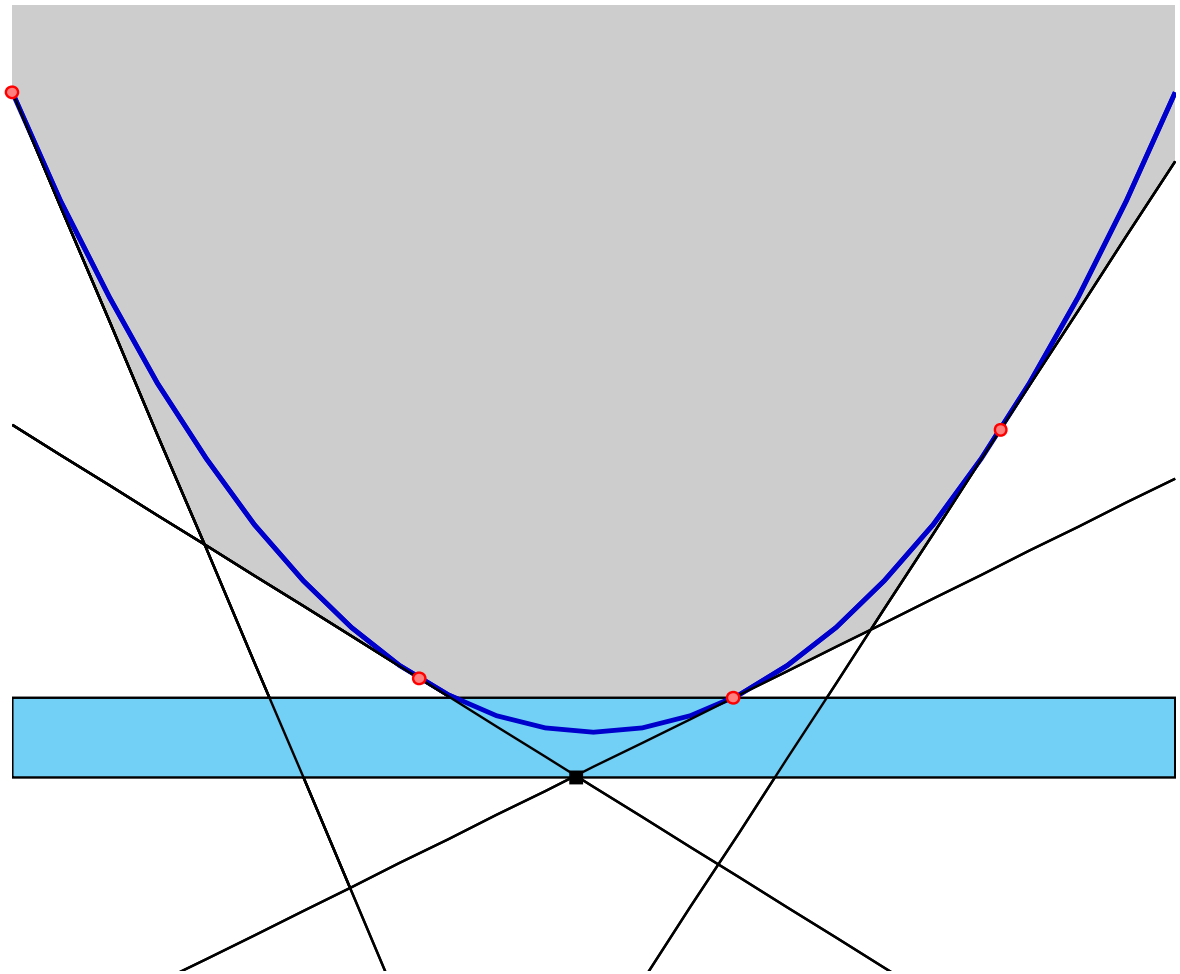
[7] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Cutting plane method: a demonstration[8]



[8] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Cutting plane method: a demonstration[9]



[9] http://learning.stat.purdue.edu/wiki/_media/courses/fall2011/cs590/optimization3.pdf

Initialize compact set C containing the minimizer, tolerance $\epsilon > 0$, $k = 1$, $x^1 \in C$, and $h_0 = -\infty$.

Iterate:

1. compute $p^k \in \partial f(x^k)$;
2. construct piece-wise affine function

$$h_k(x) = \max\{h_{k-1}(x), f(x^k) + \langle p^k, x - x^k \rangle\};$$

3. find $x^{k+1} \in \arg\min_{x \in C} h_k(x)$;
4. compute $\epsilon_k = f(x^k) - h_k(x^k)$;
5. if $\epsilon_k < \epsilon$, STOP; otherwise, continue with $k \leftarrow k + 1$.

Remarks:

- at every iteration, $x^{k+1} \in \arg\min_{x \in C} h_k(x)$ is an LP since h_k is piece-wise maximum of affine functions
- the LP' size increases with k!
- the stopping condition is reliable
  if $f(x^k) - h_k(x^k) < \epsilon$, then since

$$\min h_k(x) \leq \min f(x)$$

  for any k, we have

$$f(x^k) \leq h_k(x^k) + \epsilon = \min h_k(x) + \epsilon \leq \min f(x) + \epsilon.$$

  It is more reliable than the subgradient method, which often uses unreliable $\|p^k\|$.
- but, it possibly takes big, zig-zagging steps

Bundle methods

A bundle is referred to as $\{x^k, f(x^k), p^k\}$ where $p^k \in \partial f(x^k)$.

Initialize: tolerance $\epsilon > 0$, $\gamma \in (0,1)$, $k = 1$, $\hat{x}^1 = x^1 \in C$, and $h_0 = -\infty$.

Iterate:

1. compute $p^k \in \partial f(x^k)$;
2. construct piece-wise affine function

$$h_k(x) = \max\{h_{k-1}(x), f(x^k) + \langle p^k, x - x^k \rangle\};$$

3. find $x^{k+1} \in \arg\min_{x \in C} h_k(x) + \frac{\mu_k}{2} \|x - \hat{x}^k\|^2$;
4. compute $\epsilon_k = f(\hat{x}^k) - [h_k(x^{k+1}) + \frac{\mu_k}{2} \|x^{k+1} - \hat{x}^k\|^2]$;
5. if $\epsilon_k < \epsilon$, STOP; else, continue;
6. if $f(\hat{x}^k) - f(x^{k+1}) \geq m\epsilon_k$, then *serious step* $\hat{x}^{k+1} \leftarrow x^{k+1}$; else, *null step* $\hat{x}^{k+1} \leftarrow \hat{x}^k$;
7. $k \leftarrow k + 1$.

Remarks:

- Bundle algorithm (BA) is a stabilized cutting plane algorithm
- next iterate is closer to the current $\hat{x}^k$ to avoid drastic moves
- let $K_{\mathsf{s}} := \{k : \text{a } \textit{serious} \text{ step is taken at iteration } k\}$
- step 6 ensures strictly decreasing $f(\hat{x}^k)$, $k \in K_{\mathsf{serious}}$
- the presented BA is a basic version; several enhancements exist
- convergence under non-growing # of constraints
- has convergence analysis assuming bounded μ_k

Method	Per-itr cost	Iteration #
Subgradient	Update a vector	Very high
Cutting plane	LP	Medium/High
Bundle method	QP	Small/Medium

Karush-Kuhn-Tucker conditions

Theorem (Kuhn-Tucker, Nesterov'03 Thm 3.1.17)

Let f_i, $i = 0, \cdots, m$, be C^1 convex functions.

- *Suppose there exists $\overline{x}$ such that $f_i(\overline{x}) < 0$, for $i = 1, \cdots, m$.*

Then a point x^* is a solution to

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, f_0(x) \quad \text{s.t. } f_i(x) < 0,\ i = 1, \cdots, m, \tag{1}$$

if, and only if, there exists $\lambda_i \geq 0$, such that

$$\nabla f_0(x^*) + \sum_{\{i | f_i(x) = 0\}} \lambda_i \nabla f_i(x^*) = 0$$

Karush-Kuhn-Tucker conditions

Proof.

- Suppose x^* is a solution to (1), and $f^* := f(x^*)$.
- Define

$$\phi(x) = \max_{x \in \mathbb{R}^n} \{f(x) - f^*, f_1(x), \ldots, f_m(x)\}.$$

- Suppose $\hat{x}$ is a minimum of ϕ. If $\phi(\hat{x}) < 0$, then $f(\hat{x}) < f^*$ and $f_i(\hat{x}) < 0$.

Thus, $\hat{x}$ produces a strictly smaller objective value than x^* does. It is also feasible. This is clearly a contradiction.

- Thus, $\phi(x^*) = 0$, and x^* is the minimum of ϕ.

(Cont.) □

Karush-Kuhn-Tucker conditions

Proof (Cont.)

- How can we get an expression for x^*?
- Subgradients! x^* is a minimum of ϕ if, and only if,

$$0 \in \partial\phi(x^*) = \text{conv}\{\nabla f_i(x^*) : i \in I_0\},$$

where $I_0 = \{0\} \cup \{i | f_i(x^*) = 0\}$.

- This is true if, and only if, there exists $\alpha_i \geq 0$, $i \in I_0$, such that $\alpha_0 + \sum_{i \in I_0} \alpha_i = 1$ and

$$\alpha_0 \nabla f(x^*) + \sum_{i \in I_0} \alpha_i \nabla f_i(x^*) = 0.$$

- If $\alpha_0 \neq 0$, we're done (divide by α_0).

(Cont.) □

Karush-Kuhn-Tucker conditions

Proof (Cont.)

- Suppose that $\alpha_0 = 0$. Remember $\bar{x}$, in the interior of the feasible set?

$$\begin{aligned}\sum_{i \in I_0} \alpha_i f_i(\bar{x}) &= \sum_{i \in I_0} \alpha_i \left(f_i(x^*) + \langle \nabla f(x^*), \bar{x} - x^* \rangle \right) \\ &= 0.\end{aligned}$$

But $f_i(\bar{x}) < 0$ for all $i \in I_0$, and there exists $\alpha_i > 0$.

- This is a contradiction
- $\implies \lambda_i = \frac{\alpha_i}{\alpha_0} \geq 0.$ □

Lagrangian

- The expression

$$L(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

is called the *Lagrangian* of (1). Note that we restrict $\lambda \geq 0$.

- Lagrangian is a relaxation: If $f_i(x) \leq 0$ $i = 1, \ldots, m$, then $L(x, \lambda) < f_0(x)$.

- For fixed $\lambda \geq 0$,

$$\inf_x L(x, \lambda) \leq f_0(x^*).$$

- Note that

$$\sup_{\lambda \geq 0} L(x, \lambda) \quad = \quad \begin{cases} f_0(x) & \text{if } f_i(x) < 0 \text{ for all } i > 0; \\ \infty & \text{otherwise} \end{cases}$$

- Thus,

$$\inf_x \sup_{\lambda \geq 0} L(x, \lambda) = f_0(x^*).$$

Lagrangian

- If we set

$$\nabla_x L(x, y) = \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) = 0$$

we get a constraint for a characterization of x_λ for each $\lambda \geq 0$.

- The previous theorem shows that $x^* = x_{\lambda_{x^*}}$ for some $\lambda \geq 0$.
- In particular, it showed that $\lambda_i = 0$ for all $f_i(x) < 0$. Thus,

$$\begin{aligned} L(x^*, \lambda_{x^*}) &= f_0(x^*) + \sum_{i=1}^{n} \lambda_{x^*,i} f_i(x^*) \\ &= f_0(x^*) \end{aligned}$$

Strong duality

- If we take the supremum

$$\begin{aligned} f_0(x^*) &= L(x^*, \lambda_{x^*}) \\ &= \sup_{\lambda \geq 0} L(x_\lambda, \lambda) \\ &= \sup_{\lambda \geq 0} \inf_x L(x, \lambda) \\ &\leq \inf_x \sup_{\lambda \geq 0} L(x, \lambda) \\ &= f_0(x^*). \end{aligned}$$

i.e.

$$\sup_{\lambda \geq 0} \inf_x L(x, \lambda) = \inf_x \sup_{\lambda \geq 0} L(x, \lambda)$$

This is called *strong duality*.

- Key to result is the existence of $\bar{x}$ such that $f_i(\bar{x}) < 0$. (Slater's condition)

Strong duality

- Strong duality says problem (1) is equivalent to the *dual problem*:

$$\sup_{\lambda \geq 0} \inf_{x} L(x, \lambda)$$

- Introduce $g(\lambda) = \inf_x L(x, \lambda)$, which is called the *dual function*.

- Since g is the infimum of a family of linear functions (infimum over x, linear in λ), $g(\lambda)$ is concave, regardless of the structure of f_0.

- Sometimes, the dual problem is easier to solve than original problem. It takes care of constraints $f_i(x) < 0, i > 0$, implicitly, though introduce constraints $\lambda \geq 0$.

- We'll come back to duality at a later lecture.