Lectures 4:
Algorithms for the Lagrangian dual problem

Michael Patriksson

2010-09-08

Subgradients of convex functions

- Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a convex function
  A vector $\mathbf{p} \in \mathbb{R}^n$ is a *subgradient* of f at $\mathbf{x} \in \mathbb{R}^n$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{p}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}), \qquad \mathbf{y} \in \mathbb{R}^n \tag{1}$$

- The set of such vectors $\mathbf{p}$ defines the *subdifferential* of f at $\mathbf{x}$, and is denoted $\partial f(\mathbf{x})$

- $\partial f(\mathbf{x})$ is the collection of "slopes" of the function f at $\mathbf{x}$

- *For every* $\mathbf{x} \in \mathbb{R}^n$, $\partial f(\mathbf{x})$ *is a non-empty, convex, and compact (i.e., closed and bounded) set*

Subgradients of convex functions, II

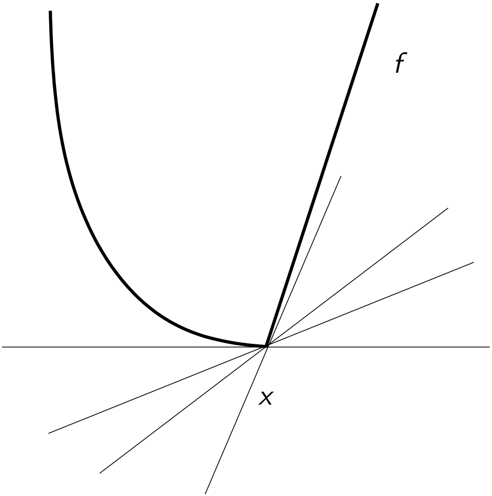


Figure: Four possible slopes of the convex function f at x

Subgradients of convex functions, III

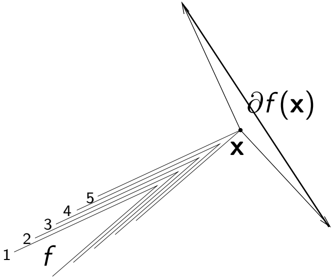


Figure: The subdifferential of a convex function f at $\mathbf{x}$
f is indicated by level curves

- *The convex function f is differentiable at $\mathbf{x}$ if there exists exactly one subgradient of f at $\mathbf{x}$, which then equals the gradient of f at $\mathbf{x}$, $\nabla f(\mathbf{x})$*

Differentiability of the Lagrangian dual function

- Consider the problem

$$f^* = \operatorname*{infimum}_{\mathbf{x}} f(\mathbf{x}), \tag{2a}$$
$$\text{subject to } \mathbf{x} \in X, \tag{2b}$$
$$g_i(\mathbf{x}) \leq 0, \qquad i = 1, \ldots, m, \tag{2c}$$

and assume that

$$f, g_i \, (\forall i) \text{ continuous; } X \text{ nonempty and compact} \tag{3}$$

- The set of solutions to the *Lagrangian subproblem*

$$X(\boldsymbol{\mu}) = \arg \min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\mu})$$

is non-empty and compact for every $\boldsymbol{\mu} \in \mathbb{R}^m$

Subgradients and gradients of q

- *Suppose that* (3) *holds* (f*,* g_i*,* $\forall i$ *continuous;* X *nonempty and compact) in the problem* (2):
  $f^* = \inf_{\mathbf{x}} \{ f(\mathbf{x}) \mid \mathbf{x} \in X, g_i(\mathbf{x}) \leq 0, \; i = 1, \dots, m \}$
- The dual function q is *finite*, *continuous*, and *concave* on $\mathbb{R}^m$. If its supremum over $\mathbb{R}^m_+$ is attained, then the optimal solution set therefore is closed and convex
- Let $\boldsymbol{\mu} \in \mathbb{R}^m$. If $\mathbf{x} \in X(\boldsymbol{\mu})$, then $\mathbf{g}(\mathbf{x})$ is a subgradient to q at $\boldsymbol{\mu}$, that is, $\mathbf{g}(\mathbf{x}) \in \partial q(\boldsymbol{\mu})$
- *Proof.* Let $\bar{\boldsymbol{\mu}} \in \mathbb{R}^m$ be arbitrary. We have that

$$
\begin{aligned}
q(\bar{\boldsymbol{\mu}}) &= \operatorname*{infimum}_{\mathbf{y} \in X} L(\mathbf{y}, \bar{\boldsymbol{\mu}}) \leq f(\mathbf{x}) + \bar{\boldsymbol{\mu}}^{\mathrm{T}} \mathbf{g}(\mathbf{x}) \\
&= f(\mathbf{x}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{g}(\mathbf{x}) + \boldsymbol{\mu}^{\mathrm{T}} \mathbf{g}(\mathbf{x}) \\
&= q(\boldsymbol{\mu}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{g}(\mathbf{x})
\end{aligned}
$$

Subgradients and gradients of q, cont'd

- Recall the *subgradient inequality* (1) for a *convex* function f: $\mathbf{p}$ is a subgradient of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{p}^{\mathrm{T}}(\mathbf{y} - \mathbf{x}), \qquad \mathbf{y} \in \mathbb{R}^n$$

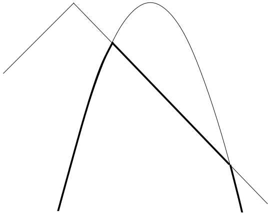
- The function $\mathbf{y} \mapsto f(\mathbf{x}) + \mathbf{p}^{\mathrm{T}}(\mathbf{y} - \mathbf{x})$ is linear w.r.t. $\mathbf{y}$ and *underestimates* $f(\mathbf{y})$ over $\mathbb{R}^n$
- Here, we have a *concave* function q and the opposite inequality: $\mathbf{g}(\mathbf{x})$ is a subgradient (actually, supgradient) of q at $\boldsymbol{\mu}$ if $\mathbf{x} \in X(\boldsymbol{\mu})$:

$$q(\bar{\boldsymbol{\mu}}) \leq q(\boldsymbol{\mu}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{g}(\mathbf{x}), \qquad \bar{\boldsymbol{\mu}} \in \mathbb{R}^m$$

- The function $q(\boldsymbol{\mu}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{g}(\mathbf{x})$ is linear w.r.t. $\bar{\boldsymbol{\mu}}$ and *overestimates* $q(\boldsymbol{\mu})$ over $\mathbb{R}^m$

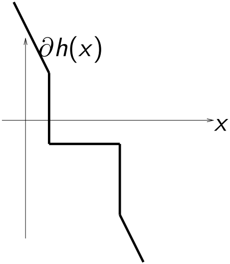
Example

- A pointwise minimum of any collection of concave functions is concave
- Let $h(x) = \min\{h_1(x), h_2(x)\}$, where $h_1(x) = 4 - |x|$ and $h_2(x) = 4 - (x-2)^2$
- Then, $h(x) = \begin{cases} 4 - x, & 1 \leq x \leq 4, \\ 4 - (x-2)^2, & x \leq 1, \ x \geq 4 \end{cases}$



Example, cont'd

- h is non-differentiable at $x = 1$ and $x = 4$, since its graph has non-unique supporting hyperplanes there



$$\partial h(x) = \begin{cases} \{-1\}, & 1 < x < 4 \\ \{4 - 2x\}, & x < 1, \ x > 4 \\ [-1, 2]\,, & x = 1 \\ [-4, -1]\,, & x = 4 \end{cases}$$

- The subdifferential is here either a singleton (at differentiable points) or an interval (at non-differentiable points)

Monotonicity of the subdifferential

- Recall from convex analysis that differentiable convex functions have a monotone gradient: if $f : \mathbb{R}^n \to \mathbb{R}$ is convex and in C^1 then for any pair $(\mathbf{x}, \mathbf{y})$ of vectors in $\mathbb{R}^n$ it holds that

$$[\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})]^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) \geq 0$$

- Case of $n = 1$: $[f'(x) - f'(y)](x - y) \geq 0$: f' is increasing
- For general convex functions we have an extended notion of monotonicity of its subdifferential: for every pair of subgradients $\boldsymbol{\xi}_x \in \partial f(\mathbf{x})$ and $\boldsymbol{\xi}_y \in \partial f(\mathbf{y})$ it holds that

$$[\boldsymbol{\xi}_x - \boldsymbol{\xi}_y]^{\mathrm{T}}(\mathbf{x} - \mathbf{y}) \geq 0$$

- For general concave functions the reverse inequality of course holds; see the figure of ∂h above!

The Lagrangian dual problem

- Let $\boldsymbol{\mu} \in \mathbb{R}^m$. Then, $\partial q(\boldsymbol{\mu}) = \operatorname{conv} \{ \mathbf{g}(\mathbf{x}) \mid \mathbf{x} \in X(\boldsymbol{\mu}) \}$
- Let $\boldsymbol{\mu} \in \mathbb{R}^m$. The dual function q is differentiable at $\boldsymbol{\mu}$ *if and only if* $\{ \mathbf{g}(\mathbf{x}) \mid \mathbf{x} \in X(\boldsymbol{\mu}) \}$ is a singleton set. Then,

  $$\nabla q(\boldsymbol{\mu}) = \mathbf{g}(\mathbf{x}),$$

  for every $\mathbf{x} \in X(\boldsymbol{\mu})$
- Holds in particular if the Lagrangian subproblem has a unique solution (the solution set $X(\boldsymbol{\mu})$ is a singleton)
  True, e.g., when X is convex, f strictly convex on X, and g_i convex on X $\forall i$ (e.g., f strictly convex quadratic, X polyhedral, g_i linear)

How do we write the subdifferential of h?

- If $h(\mathbf{x}) = \min_{i=1,\dots,m} h_i(\mathbf{x})$, where each function h_i is concave and differentiable on $\mathbb{R}^n$, then h is a concave function on $\mathbb{R}^n$
- Define the set $\mathcal{I}(\mathbf{x}) \subseteq \{1, \dots, m\}$ by the active segments at $\mathbf{x}$:

$$\begin{cases} i \in \mathcal{I}(\mathbf{x}) & \text{if } h(\mathbf{x}) = h_i(\mathbf{x}), \\ i \notin \mathcal{I}(\mathbf{x}) & \text{if } h(\mathbf{x}) < h_i(\mathbf{x}), \end{cases} \qquad i \in \{1, \dots, m\}$$

- Then, the subdifferential $\partial h(\mathbf{x})$ is the *convex hull* of the gradients $\{\nabla h_i(\mathbf{x}) \mid i \in \mathcal{I}(\mathbf{x})\}$:

$$\partial h(\mathbf{x}) = \left\{ \xi = \sum_{i \in \mathcal{I}(\mathbf{x})} \lambda_i \nabla h_i(\mathbf{x}) \,\middle|\, \sum_{i \in \mathcal{I}(\mathbf{x})} \lambda_i = 1;\ \lambda_i \geq 0, i \in \mathcal{I}(\mathbf{x}) \right\}$$

Optimality conditions for the dual problem

- For a differentiable, concave function h it holds that

$$\mathbf{x}^* \in \arg\max_{\mathbf{x} \in \mathbb{R}^n} h(\mathbf{x}) \quad \Longleftrightarrow \quad \nabla h(\mathbf{x}^*) = \mathbf{0}^n$$

- Theorem: *Assume that h is concave on $\mathbb{R}^n$. Then,*

$$\mathbf{x}^* \in \arg\max_{\mathbf{x} \in \mathbb{R}^n} h(\mathbf{x}) \quad \Longleftrightarrow \quad \mathbf{0}^n \in \partial h(\mathbf{x}^*)$$

- *Proof.*

  Suppose that $\mathbf{0}^n \in \partial h(\mathbf{x}^*) \Longrightarrow h(\mathbf{x}) \leq h(\mathbf{x}^*) + (\mathbf{0}^n)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*)$ for all $\mathbf{x} \in \mathbb{R}^n$, that is, $h(\mathbf{x}) \leq h(\mathbf{x}^*)$ for all $\mathbf{x} \in \mathbb{R}^n$

  Suppose that $\mathbf{x}^* \in \arg\max_{\mathbf{x} \in \mathbb{R}^n} h(\mathbf{x}) \Longrightarrow$ $h(\mathbf{x}) \leq h(\mathbf{x}^*) = h(\mathbf{x}^*) + (\mathbf{0}^n)^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^*)$ for all $\mathbf{x} \in \mathbb{R}^n$, that is, $\mathbf{0}^n \in \partial h(\mathbf{x}^*)$ $\square$

Optimality conditions for the dual problem, cont'd

- The example: $0 \in \partial h(1) \implies x^* = 1$
- For optimization with constraints the KKT conditions are generalized:

$$\mathbf{x}^* \in \arg\max_{\mathbf{x} \in X} h(\mathbf{x}) \quad \iff \quad \partial h(\mathbf{x}^*) \cap N_X(\mathbf{x}^*) \neq \emptyset,$$

where $N_X(\mathbf{x}^*)$ is the normal cone to X at $\mathbf{x}^*$, that is, the conical hull of the active constraints' normals at $\mathbf{x}^*$

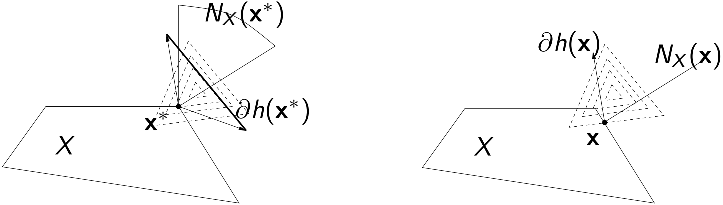


Figure: An optimal solution $\mathbf{x}^*$ A non-optimal solution $\mathbf{x}$

Optimality conditions for the dual problem, cont'd

- The dual problem has only sign conditions $\boldsymbol{\mu} \geq \mathbf{0}^m$
- Consider the dual problem

$$q^* = \underset{\boldsymbol{\mu} \geq \mathbf{0}^m}{\text{maximize}} \; q(\boldsymbol{\mu})$$

- $\boldsymbol{\mu}^* \geq \mathbf{0}^m$ is then optimal *if and only if* there exists a subgradient $\mathbf{g} \in \partial q(\boldsymbol{\mu}^*)$ for which the following holds:

$$\mathbf{g} \leq \mathbf{0}^m; \quad \mu_i^* g_i = 0, \; i = 1, \ldots, m$$

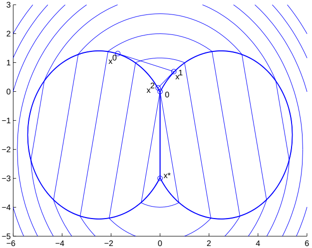
- Compare with a one-dimensional max-problem (h concave):

$$x^* \geq 0 \text{ is optimal} \quad \Longleftrightarrow \quad h'(x^*) \leq 0; \quad x^* \cdot h'(x^*) = 0$$

- Note: it is not possible in general to obtain the whole subdifferential, only a representative $\mathbf{g} \in \partial q(\boldsymbol{\mu})$ from the constraint function value $\mathbf{g}(\mathbf{x}(\boldsymbol{\mu}))$ of a Lagrangian subproblem solution $\mathbf{x}(\boldsymbol{\mu})$. Hence, verifying optimality is considerably more difficult than in the differentiable case

Methods for convex nondifferentiable optimization: A caution

- Continuous convex functions are differentiable almost everywhere; use methods from smooth optimization!
- No! Big mistake!
- Take $f(\mathbf{x}) := \max\{-5x_1 + x_2; x_1^2 + x_2^2 + 4x_2; 5x_1 + x_2\}$



- Optimum at $(0, -3)^{\mathrm{T}}$ with $f^* = -3$, but starting at $\mathbf{x}^0$, using steepest descent with an exact line search leads to $\mathbf{x} = (0, 0)^{\mathrm{T}}$ with $f = 0$

A subgradient method for the dual problem

- Subgradient methods extend gradient projection methods from C^1 to general convex (or, concave) functions, generating a sequence of dual vectors in $\mathbb{R}^m_+$ using a single subgradient in each iteration

- The simplest type of iteration has the form

$$\begin{aligned} \boldsymbol{\mu}^{k+1} &= \mathrm{Proj}_{\mathbb{R}^m_+}[\boldsymbol{\mu}^k + \alpha_k \mathbf{g}^k] \\ &= [\boldsymbol{\mu}^k + \alpha_k \mathbf{g}^k]_+ \\ &= (\mathrm{maximum}\,\{0, (\boldsymbol{\mu}^k)_i + \alpha_k (\mathbf{g}^k)_i\})_{i=1}^m, \end{aligned} \tag{4}$$

where k is the iteration counter, $\mathbf{g}^k \in \partial q(\boldsymbol{\mu}^k)$ is an arbitrary subgradient, and $\alpha_k > 0$ is a step length

A subgradient method for the dual problem, cont'd

- We often write $\mathbf{g}^k = \mathbf{g}(\mathbf{x}^k)$, where $\mathbf{x}^k \in \arg\min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\mu}^k)$
- Main difference to C^1 case: an arbitrary subgradient $\mathbf{g}^k$ *may not be an ascent direction!*

$\Rightarrow$ Cannot make line searches; must use predetermined step lengths α_k

- *Suppose that* $\boldsymbol{\mu} \in \mathbb{R}^m_+$ *is not optimal in* $\max_{\boldsymbol{\mu} \geq \mathbf{0}^m} q(\boldsymbol{\mu})$ *Then, for every optimal solution* $\boldsymbol{\mu}^* \in U^*$

$$\|\boldsymbol{\mu}^{k+1} - \boldsymbol{\mu}^*\| < \|\boldsymbol{\mu}^k - \boldsymbol{\mu}^*\|$$

holds for every step length α_k *in the interval*

$$\alpha_k \in (0, 2[q^* - q(\boldsymbol{\mu}^k)]/\|\mathbf{g}^k\|^2)$$

A subgradient method for the dual problem, cont'd

- Why? Let $\mathbf{g} \in \partial q(\bar{\boldsymbol{\mu}})$, and let U^* be the set of optimal solutions to $\max_{\boldsymbol{\mu} \geq \mathbf{0}^m} q(\boldsymbol{\mu})$. Then,

$$U^* \subseteq \{\, \boldsymbol{\mu} \in \mathbb{R}^m \mid \mathbf{g}^{\mathrm{T}}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \geq 0 \,\}$$

  In other words, $\mathbf{g}$ defines a half-space that contains the set of optimal solutions

- Good news: If the step length α_k is small enough we get closer to the set of optimal solutions!

Each (sub)gradient defines a halfspace containing the optimal set

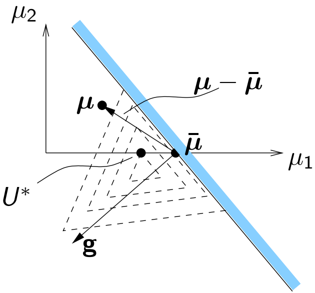
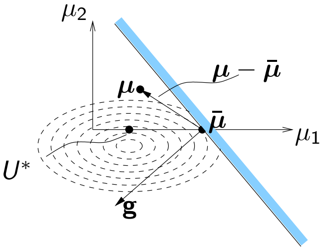


Figure: q non-differentiable



q differentiable

$$\mathbf{g} \in \partial q(\bar{\boldsymbol{\mu}}) \quad \Rightarrow \quad U^* \subseteq \{\, \boldsymbol{\mu} \in \mathbb{R}^m \mid \mathbf{g}^{\mathrm{T}}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \geq 0 \,\}$$

Each (sub)gradient defines a halfspace containing the optimal set

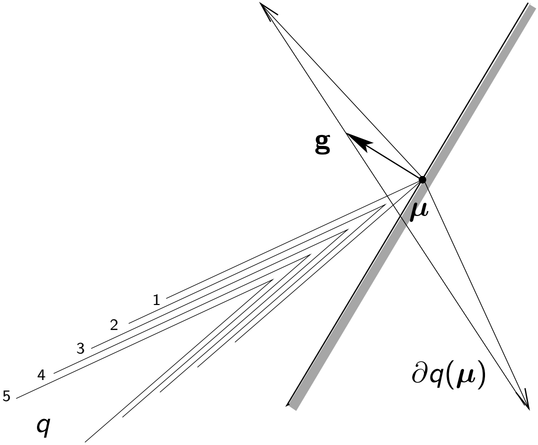


Figure: The half-space defined by a subgradient $\mathbf{g} \in q(\boldsymbol{\mu})$
Note that this subgradient is *not an ascent direction*

Polyak's step length rule

- Choose the step length α_k such that

$$\sigma \leq \alpha_k \leq 2[q^* - q(\boldsymbol{\mu}^k)]/\|\mathbf{g}^k\|^2 - \sigma, \quad k = 1, 2, \dots \tag{5}$$

- $\sigma > 0 \Rightarrow$ step lengths α_k don't converge to 0, or converges to a too large value

- Bad news: Utilizes knowledge of the optimal value q^*!

- But: q^* can be replaced by an approximation $\bar{q}_k \geq q^*$

The divergent series step length rule

- Choose the step lengths α_k such that

$$\alpha_k > 0, \; k = 1, 2, \dots \; ; \quad \lim_{k \to \infty} \alpha_k = 0; \quad \sum_{s=1}^{\infty} \alpha_s = +\infty \tag{6}$$

- Additional condition often added:

$$\sum_{s=1}^{\infty} \alpha_s^2 < +\infty \tag{7}$$

Convergence results

- *Suppose that f and $\mathbf{g}$ are continuous, X is compact, $\exists \mathbf{x} \in X : \mathbf{g}(\mathbf{x}) < \mathbf{0}$, and consider the problem*

$$f^* = \inf\{ f(\mathbf{x}) \,|\, \mathbf{x} \in X, \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \} \tag{8}$$

(a) *Let $\{\boldsymbol{\mu}^k\}$ be generated by the method on p. 17, under the Polyak step length rule (5), where $\sigma > 0$ is small*
Then, $\{\boldsymbol{\mu}^k\} \to \boldsymbol{\mu}^* \in U^*$

(b) *Let $\{\boldsymbol{\mu}^k\}$ be generated by the method on p. 17, under the divergent series step length rule (6)*
Then, $\{q(\boldsymbol{\mu}^k)\} \to q^*$, and $\{\mathrm{dist}_{U^*}(\boldsymbol{\mu}^k)\} \to 0$

(c) *Let $\{\boldsymbol{\mu}^k\}$ be generated by the method on p. 17, under the divergent series step length rule (6), (7)*
Then, $\{\boldsymbol{\mu}^k\} \to \boldsymbol{\mu}^* \in U^*$

Application to the Lagrangian dual problem

1. Given $\boldsymbol{\mu}^k \geq \mathbf{0}^m$
2. Solve the Lagrangian subproblem: $\min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\mu}^k)$
3. Let an optimal solution to this problem be $\mathbf{x}^k = \mathbf{x}(\boldsymbol{\mu}^k)$
4. Calculate $\mathbf{g}(\mathbf{x}^k) \in \partial q(\boldsymbol{\mu}^k)$
5. Take a step $\alpha_k > 0$ in the direction of $\mathbf{g}(\mathbf{x}^k)$ from $\boldsymbol{\mu}^k$, according to a step length rule
6. Set any negative components of this vector to $0 \Rightarrow \boldsymbol{\mu}^{k+1}$
7. Let $k := k + 1$ and repeat from 2

Additional algorithms

- We can choose the subgradient more carefully, to obtain *ascent* directions
- Gather several subgradients at nearby points $\boldsymbol{\mu}^k$ and solve quadratic programming problems to find the best convex combination of them (*Bundle methods*)
- Pre-multiply the subgradient by some positive definite matrix $\Rightarrow$ methods similar to Newton methods (*Space dilation methods*)
- Pre-project the subgradient vector (onto the tangent cone of $\mathbb{R}^m_+$) $\Rightarrow$ step direction is a *feasible direction* (*Subgradient-projection methods*)

More to come . . .

- Discrete optimization: The size of the duality gap, and the relation to the continuous relaxation
- Convexification
- Primal feasibility heuristics
- Global optimality conditions for discrete optimization (and general problems)