Lecture Notes for IAP 2005 Course
Introduction to Bundle Methods

Alexandre Belloni*

Version of February 11, 2005

1 Introduction

Minimizing a convex function over a convex region is probably the core problem in the Nonlinear Programming literature. Under the assumption of the function of a differentiable function of interest, several methods have been proposed and successively studied. For example, the Steepest Descent Method consists of performing a line search along a descent direction given by minus the gradient of the function at the current iterate.

One of the most remarkable applications of minimizing convex functions reveals itself within duality theory. Suppose we are interested in solving the following problem

$$(P)\begin{cases} \max\limits_{y} & g(y) \\ & h_i(y) = 0 \quad i = 1, \ldots, m \\ & y \in \mathcal{D} \end{cases},$$

which is known to be hard (for example, a large instance of the TSP or a large-scale linear programming). One can define the following auxiliary function

$$f(x) = \max_{y} \; g(y) + \langle x, h(y) \rangle \quad y \in \mathcal{D},$$

which is convex without any assumption on $\mathcal{D}$, g, or h. Due to duality theory, the function f will always give an upper bound for the original maximization problem. Also, the dual information within this scheme is frequently very useful to build primal heuristics for nonconvex problems. In practice[1], it is easy to obtain solutions at most 5% worse than the unknown optimum value[2].

There is an intrinsic cost to be paid in the previous construction. The function f is defined implicitly, and evaluating it may be a costly operation. Moreover, the differentiability is lost in general[3]. One important remark is that

*Operation Research Center, M.I.T. (belloni@mit.edu)

[1] Of course, we cannot prove that in general, but it holds for a large variety of problems in real-world applications.

[2] In fact, the quality of the solution tends to be much better, i.e., smaller than 1% in the majority of the problems.

[3] This tends to be the rule rather than the exception in practice.

even though differentiability is lost, one can easily compute a substitute for the gradient called *subgradient* with no additional cost[4] beyond the evaluation of f.

This motivates an important abstract framework to work with. The framework consists of minimizing a convex function (possible nondifferentiable) given by an oracle. That is, given a point in the λ, the oracle returns the value $f(x)$ and a subgradient s.

Within this framework, several methods have also been proposed and many of them consists of adapting methods for differentiable functions by replacing gradients with subgradients. Unfortunately, there are several drawbacks with such procedure. Unlike in the differentiable case, minus the subgradient may not be a descent direction. In order to guarantee that a direction is indeed a descent direction, one needs to know the complete set of subgradients[5], which is a restrictive assumption in our framework[6]. Another drawback is the stopping criterium. In the differentiable case, one can check if $\nabla f(x) = 0$ or $\|\nabla f(x)\| < \varepsilon$. However, the nondifferentiablility of f imposes new challenges as shown in the following 1-dimensional example.

Example 1.1 *Let $f(x) = |x|$ where $\lambda \in \mathbb{R}$. One possible oracle could have the following rule for computing subgradients*

$$\partial f(x) \ni s = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}.$$

Note that the oracle returns a subgradient such that $\|s\| \geq 1$ for all x. In this case, there is no hope of a simple stopping criteria like $s = 0$ would work.

The goal of these notes is to give a brief but formal introduction to a family of methods for this framework called Bundle Methods.

2 Notation

We denote the inner product by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the norm induced by it. The ball centered at x with radius r is denoted by $B(x, r) = \{y \in \mathbb{R}^n : \|y - x\| \leq r\}$. The vector of all ones is denoted by e. Also, $\mathbf{span}(\{d^i\}_{i=1}^k) = \{\sum_{i=1}^k \alpha_i d^i : \alpha^i \in \mathbb{R},\ i = 1, \ldots, k\}$ denotes the set of linear combinations of $\{d^i\}_{i=1}^k$. If S is a set, $\mathbf{int}\ S = \{y \in \mathbb{R}^n : \exists r > 0,\ B(y, r) \subset S\}$ is the interior of S, $\mathbf{cl}\ S = \{y \in \mathbb{R}^n : \exists \{y^k\}_{k \geq 1} \subset S,\ y^k \to y\}$ is the closure of S, and $\partial S = \mathbf{cl}\ S \setminus \mathbf{int}\ S$ is the boundary of S.

3 Some Convex Analysis

Definition 3.1 *A set C is said to be convex if for all $x, y \in C$, and all $\alpha \in [0, 1]$, we have that*

$$\alpha x + (1 - \alpha) y \in C.$$

This definition can be extended for funtions on $\mathbb{R}^n$ as follows.

[4] See Appendix for this derivation and the proof that f is convex.

[5] This set is called subdifferential of f at x, $\partial f(x)$.

[6] Not only in theory but also in practice the complete knowledge of the subdifferential is much harder to obtain.

Definition 3.2 *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ is a convex function if for all $x, y \in \mathbb{R}^n$, $\alpha \in [0, 1]$,*

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

We also define the domain of f as the points where f is finite-valued, i.e., $\mathbf{dom}(f) = \{x \in \mathbb{R}^n : f(x) < \infty\}$. f is proper if $\mathbf{dom}(f) \neq \emptyset$.

Definition 3.3 *The indicator function of a set C is defined as*

$$I_C(x) = \begin{cases} 0, & x \in C \\ +\infty, & x \notin C \end{cases}.$$

Note that C is convex if and only if $I_C : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex function.

In some proofs, it is sometimes convenient to work with convex sets instead of functions. So, we define the following auxiliary object.

Definition 3.4 *Define the epigraph of a function f as*

$$\mathbf{epi}(f) = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : x \in \mathbf{dom}(f), t \geq f(x)\}.$$

Lemma 3.1 *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a function. Then, f is convex if and only if $\mathbf{epi}(f)$ is a convex set.*

Proof. ($\Rightarrow$) Let $(x, t), (y, s) \in \mathbf{epi}(f)$. Thus, $t \geq f(x)$ and $s \geq f(y)$ by definition. Take any $\alpha \in [0, 1]$, we have that

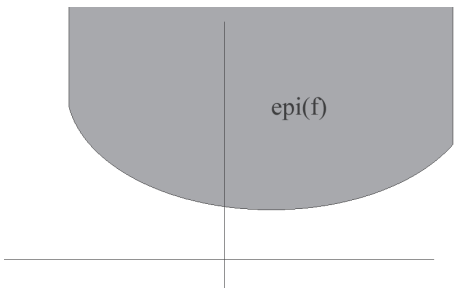
$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \leq \alpha t + (1-\alpha)s$$

which proves that $(\alpha x + (1-\alpha)y, \alpha t + (1-\alpha)s) \in \mathbf{epi}(f)$.

($\Leftarrow$) Consider $x, y \in \mathbf{dom}(f)$ and $\alpha \in [0, 1]$. Note that $(x, f(x))$ and $(y, f(y)) \in \mathbf{epi}(f)$. Since $\mathbf{epi}(f)$ is convex, $(\alpha x + (1-\alpha)y, \alpha f(x) + (1-\alpha)f(y)) \in \mathbf{epi}(f)$. By definition of $\mathbf{epi}(f)$,

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

■



Throughout these notes, we will assume that our convex functions have a closed epigraph. In fact, we have the following equivalence:

Theorem 3.1 *Let $f : \mathbb{R}^n \to [-\infty, +\infty]$ be an arbitrary function. Then, the following conditions are equivalent:*

(*i*) *f is lower semi-continuous throughout $\mathbb{R}^n$;*

(*ii*) *$\{x \in \mathbb{R}^n | f(x) \leq \alpha\}$ is closed for every $\alpha \in \mathbb{R}$;*

(*iii*) *The epigraph of f is a closed set in $\mathbb{R}^{n+1}$.*

We also state some basic facts from Convex Analysis Theory that can be found in [11].

Theorem 3.2 *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex function. Then, f is continuous in* **int dom**(f).

4 The Subdifferential of a Convex Function

Now, we define one of the main object of our theory, the subdifferential of a convex function. Later, we will also prove some additional results about the subdifferential that we do not use but are very insightful.

Definition 4.1 *The subdifferential of a convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at a point x is defined as*

$$\partial f(x) = \{s \in \mathbb{R}^n : f(y) \geq f(x) + \langle s, y - x \rangle \quad \textit{for all } y \in \mathbb{R}^n\}.$$

Remark 4.1 *We point out that the subdifferential is a set of linear functionals, so it lives on the dual space of the space that contains* **dom**(f). *In the case of $\mathbb{R}^n$ we do have an one-to-one correspondence between the dual space and the original space.*

Lemma 4.1 *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex function, and $x \in$* **int dom**(f). *Then, $\partial f(x) \neq \emptyset$.*

Proof. Since f is convex, **epi**(f) is also convex. Also, $x \in$ **dom**(f) implies that $f(x) < \infty$.

Using the Hahn-Banach Theorem[7] there exists a linear functional in $\tilde{s} = (s, s_{n+1}) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$\langle \tilde{s}, (x, f(x)) \rangle \leq \langle \tilde{s}, (y, t) \rangle \quad \text{for all } (y, t) \in \textbf{epi}(f),$$

where the "extended" scalar product is defined as $\langle (s, s_{n+1}), (x, t) \rangle = \langle s, x \rangle + s_{n+1}t$.

We need to consider three cases for s_{n+1}.

If $s_{n+1} < 0$, we have that

$$\langle s, x \rangle - |s_{n+1}| f(x) \leq \langle s, y \rangle - |s_{n+1}| t \quad \text{for all } (y, t) \in \textbf{epi}(f).$$

Letting $t \nearrow +\infty$ we obtain a contradiction since the left hand side is a constant.

If $s_{n+1} = 0$, we have that

$$\langle s, x \rangle \leq \langle s, y \rangle \quad \text{for all } y \in \textbf{dom}(f).$$

[7] The version of the Separating Hyperplane Theorem for convex sets.

Noting that $x \in \mathbf{int}\ \mathbf{dom}(f)$, this is also a contradiction and we can also reject this case.

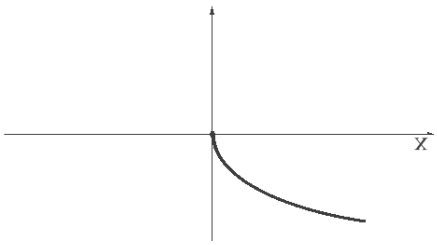
So, $s_{n+1} > 0$, and due to homogeneity, we can assume $s_{n+1} = 1$. By definition of the epigraph, $(y, f(y)) \in \mathbf{epi}(f)$ for all $y \in \mathbf{dom}(f)$, which implies that

$$\langle s, x\rangle + f(x) \leq \langle s, y\rangle + f(y) \qquad \therefore \qquad f(y) \geq f(x) + \langle -s, y - x\rangle .$$

Thus, $(-s) \in \partial f(x)$. ■

Example 4.1 *We cannot relax the assumption of* $x \in \mathbf{int}\ \mathbf{dom}(f)$ *in general.* $f(x) = -\sqrt{x}$ *is a convex function on* $\mathbb{R}_+$*, but* $\partial f(0) = \emptyset$*. Note that we always have that*

$$\mathbf{int}\ \mathbf{dom}(f) \subseteq \mathbf{dom}(\partial f) \subseteq \mathbf{dom}(f).$$



At this point, it is useful to stop and relate the subdifferential to the regular case where f is differentiable.

Lemma 4.2 *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be differentiable at* x*. Then,* $\partial f(x) = \{\nabla f(x)\}$.

Proof. Take $s \in \partial f(x)$ and an arbitrary $d \in \mathbb{R}^n$ and $t > 0$. By definition,

$$f(x + td) \geq f(x) + \langle s, x + td - x\rangle = f(x) + t\,\langle s, d\rangle$$

Thus, $\dfrac{f(x + td) - f(x)}{t} \geq \langle s, d\rangle$. Now, letting $t \searrow 0$, the left hand side has a unique limit since x is differentiable. Finally,

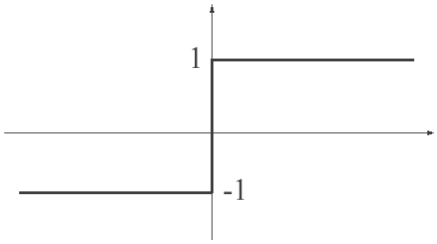
$$\langle \nabla f(x), d\rangle \geq \langle s, d\rangle \quad \text{for all} \quad d \in \mathbb{R}^n,$$

which implies that $s = \nabla f(x)$. ■

The proof of the last Lemma shows that the subgradients are coming from all the possible partial derivatives. In general, the subdifferential is not a function as the gradient is for differentiable functions. ∂f is a correspondence, a point-set mapping.

Example 4.2 *Consider the function* $f(x) = |x|$ *for* $x \in \mathbb{R}$*. The subdifferential of* f *in this example is*

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



We obeserve that at zero, where the function fails to be differentiable, we have many possible values for subgradients.

Definition 4.2 *Let $M : \mathbb{R}^n \to \mathcal{B}(\mathbb{R}^n)$ be a point-set mapping. M is said to be upper semi-continuous if for every convergent sequence $(x_k, s_k) \to (\bar{x}, \bar{s})$ such that $s_k \in M(x_k)$ for every $k \in \mathbb{N}$, we have that $\bar{s} \in M(\bar{x})$. M is said to be lower semi-continuous if for every $\bar{s} \in M(\bar{x})$ and every sequence $x_k \to \bar{x}$, there exists a sequence $\{y_k\}_{k\geq 1}$ such that $y_k \in M(x_k)$ and $y_k \to \bar{y}$. Finally, M is said to be continuous if it is both upper and lower semi-continuous.*

Remark 4.2 *Another important notion of continuity for point-set mappings are inner and outter continuity. In finite dimensional spaces, this notion coincides with our lower and upper continuity definitions.*

The study of the subdifferential through the eyes of the theory of correspondences clarifies some important properties and difficulties of working with that operator.

Lemma 4.3 *The subdifferential of a proper convex function f is upper semi-continuous and convex valued.*

Proof. Consider a sequence $(x^k, s^k) \to (\bar{x}, \bar{s})$ where $s^k \in \partial f(x^k)$. Thus, for every $k \in \mathbb{N}$ and every $y \in \mathbb{R}^n$,

$$f(y) \geq f(x^k) + \left\langle s^k, y - x^k \right\rangle$$

Taking limits, $f(x^k) \to \bar{f} \geq f(\bar{x})$, and we obtain

$$f(y) \geq f(\bar{x}) + \langle \bar{s}, y - \bar{x} \rangle .$$

Thus, $\bar{s} \in \partial f(\bar{x})$.

For the second statement, let $s^1, s^2 \in \partial f(x)$, $\alpha \in [0, 1]$, and the result follows since the $\langle \cdot, \cdot \rangle$ is a linear operator. ■

Example 4.3 *From our Example 4.2, it is easy to see that the subdifferential can fail to be lower semi-continuous. For instance, take $(\bar{x}, \bar{s}) = (0, 0)$ and $x^k = \frac{1}{k}$. The sequence of $\{s^k\}$ must be chosen to be identically equal to 1 which is not converging to zero.*

In order to introduce an equivalent characterization of the subdifferential of f, we need the following lemma.

Lemma 4.4 *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper lower semicontinuous convex function. Then, the directional derivatives of f are well defined for every $d \in \mathbb{R}^n$,*

$$f'(x;d) = \lim_{t \searrow 0} \frac{f(x+td) - f(x)}{t}.$$

Proof. Let $0 < t' < t$. Thus, let $\alpha = \frac{t'}{t}$ so that

$$x + t'd = \alpha(x + td) + (1-\alpha)x.$$

Using the convexity of f,

$$\begin{aligned} \frac{f(x+t'd) - f(x)}{t'} &= \frac{f(\alpha(x+td) + (1-\alpha)x) - f(x)}{\alpha t} \leq \frac{\alpha f(x+td) + (1-\alpha)f(x) - f(x)}{\alpha t} \\ &= \frac{\alpha}{\alpha} \frac{f(x+td) - f(x)}{t}, \end{aligned}$$

so the ratio is non increasing in t. Thus, as a monotonic sequence its limit (possibly $+\infty$) is well defined. ■

Even though the directional derivaties are a local characteristic, it does have some global properties due to the convexity of f. This is illustraded by the following alternative definition of the subdifferential of f.

Lemma 4.5 *The subdifferential of f at x can be written as*

$$\partial f(x) = \{s \in \mathbb{R}^n : f'(x;d) \geq \langle s, d \rangle \text{ for every } d \in \mathbb{R}^n\} \tag{1}$$

Proof. By definition, $s \in \partial f(x)$ if and only if

$$\begin{array}{l} f(y) \geq f(x) + \langle s, y - x \rangle \text{ for every } y \in \mathbb{R}^n \\ f(x+td) \geq f(x) + \langle s, td \rangle \text{ for every } d \in \mathbb{R}^n \text{ and } t > 0 \\ \frac{f(x+td)-f(x)}{t} \geq \langle s, d \rangle \text{ for every } d \in \mathbb{R}^n \text{ and } t > 0. \end{array}$$

Since the left hand side is decreasing as $t \to 0$ by Lemma 4.4,

$$\partial f(x) = \{s \in \mathbb{R}^n : f'(x;d) \geq \langle s, d \rangle \text{ for every } d \in \mathbb{R}^n\}.$$

■

Thus, the directional derivatives at x bound all the projections of subgradients, in the subdifferential of f at x, on the corresponding direction. The next result establish the converse.

Theorem 4.1 *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a lower semi-continuous convex function. If $x \in \mathbf{dom}(f)$,*

$$f'(x;d) = \max\{\langle s, d \rangle : s \in \partial f(x)\}.$$

Proof. Without loss of generality, assume that $x = 0$, $f(x) = 0$ and $\|d\| = 1$. Denote $d := d^1$, and let $\{d^1, d^2, \ldots, d^n\}$ be an orthogonal basis for $\mathbb{R}^n$. Let $H^k = \mathbf{span}\{d^1, d^2, \ldots, d^k\}$.

We start by defining a linear functional on H^1, $\lambda_1 : H^1 \to \mathbb{R}$ as

$$\lambda_1(y) = f(x) + \langle f'(x;d)d, y - x \rangle.$$

Using Lemma 4.4, for $t > 0$

$$f(x+td) - f(x) \geq tf'(x;d) = t\left\langle f'(x;d)d, d\right\rangle = \lambda_1(x+td) - f(x).$$

Assume that there exists $\bar{t} > 0$ such that $f(x - |\bar{t}|d) < \lambda_1(x - |\bar{t}|d)$. By convexity, for any $t \in (0, \bar{t}\,]$,

$$\begin{aligned} f(x-|t|d) &\leq \left(\frac{|t|}{|\bar{t}|}\right) f(x-|\bar{t}|d) + \left(1 - \frac{|t|}{|\bar{t}|}\right) f(x) \\ &< \left(\frac{|t|}{|\bar{t}|}\right) \lambda_1(x-|\bar{t}|d) + \left(1 - \frac{|t|}{|\bar{t}|}\right) \lambda_1(x) \\ &= \lambda_1(x-|t|d). \end{aligned}$$

So,

$$\begin{aligned} f'(x;-d) = \lim_{t\to 0} \tfrac{f(x-|t|d)-f(x)}{t} &\leq \lim_{t\to 0} \tfrac{\lambda_1(x-|t|d)-f(x)}{t} \\ &= \lim_{t\to 0} \tfrac{f(x)+\left\langle f'(x;d)d, -|t|d\right\rangle - f(x)}{t} \\ &= -f'(x;d). \end{aligned}$$

If $f'(x;-d) = -f'(x;d)$, $f : H^1 \to \mathbb{R}$ is differentiable at x with gradient equal to $f'(x;d)$, and using Lemma 4.2, we proved that $f(y) \geq \lambda_1(y)$ for all $y \in H^1$.

Thus, we can assume that $f'(x;-d) < -f'(x;d)$. Also, due to Lemma 4.4, we can write that

$$f(x-|t|d) = f(x)+|t|f'(x;-d)+o(|t|) \text{ and } f(x+|t|d) = f(x)+|t|f'(x;d)+o(|t|).$$

To obtain that

$$\begin{aligned} f(x) &\leq \tfrac{1}{2}f(x-|t|d) + \tfrac{1}{2}f(x+|t|d) \\ &= \tfrac{1}{2}\left(f(x) + |t|f'(x;-d) + o(|t|) + f(x) + |t|f'(x;d) + o(|t|)\right) \\ &= f(x) + \tfrac{|t|}{2}\left(f'(x;-d) + f'(x;d)\right) + o(|t|) \\ &< f(x), \end{aligned}$$

if we make t small enough since $f'(x;-d)+f'(x;d) < 0$. This is a contradiction. So, for every $t \in \mathbb{R}$, we have $f(x+td) \geq \lambda_1(x+td)$.

After the construction of the first step, we will proceed by induction. Assume that for $k < n$, we have a function $\lambda_k : H^k \to \mathbb{R}$ such that

$$\lambda_k(x) = f(x), \text{ and } \lambda_k(y) \geq f(y) \text{ for all } y \in H^k.$$

Assuming $f(x) = 0$ for convenience, we have that λ_k is a linear function, and let $z = \frac{d^{k+1}}{\|d^{k+1}\|}$.
Suppose $w, v \in H^k$ and $\alpha > 0, \beta > 0$.

$$\begin{aligned} \beta\lambda_k(w) + \alpha\lambda_k(v) &= \lambda_k(\beta w + \alpha v) \\ &= (\alpha+\beta)\lambda_k\left(\tfrac{\beta}{\alpha+\beta}w + \tfrac{\alpha}{\alpha+\beta}v\right) \\ &\leq (\alpha+\beta)f\left(\tfrac{\beta}{\alpha+\beta}w + \tfrac{\alpha}{\alpha+\beta}v\right) \\ &= (\alpha+\beta)f\left(\left[\tfrac{\beta}{\alpha+\beta}\right](w-\alpha z) + \left[\tfrac{\alpha}{\alpha+\beta}\right](v+\beta z)\right) \\ &\leq \beta f(w-\alpha z) + \alpha f(v+\beta z). \end{aligned}$$

Thus,

$$\beta\left[-f(w-\alpha z)+\lambda_k(w)\right] \leq \alpha\left[f(v+\beta z)-\lambda_k(v)\right]$$
$$\frac{1}{\alpha}\left[-f(w-\alpha z)+\lambda(w)\right] \leq \frac{1}{\beta}\left[f(v+\beta z)-\lambda_k(v)\right]$$

$$\sup_{w\in H^k,\alpha>0}\frac{1}{\alpha}\left[-f(w-\alpha z)+\lambda_k(w)\right] \leq a \leq \inf_{v\in H^k,\beta>0}\frac{1}{\beta}\left[f(v+\beta z)-\lambda_k(v)\right]$$

We extend λ_k to H^{k+1} by defining $\lambda_{k+1}(z)=a$. Thus, assuming $t>0$,

$$\begin{aligned}\lambda_{k+1}(tz+\tilde{x}) &= t\lambda_{k+1}(z)+\lambda_{k+1}(\tilde{x}) = ta+\lambda_k(tx)\\ &\leq t\left(\tfrac{1}{t}f(\tilde{x}+tz)-\tfrac{\lambda_k(\tilde{x})}{t}\right)+\lambda_k(\tilde{x})\\ &\leq f(tz+\tilde{x}),\end{aligned}$$

and the case of $t<0$ is similar.

So, the gradient of λ_n will define an element $\bar{s}\in\partial f(x)$. Note that

$$\max\{\langle s,d\rangle : s\in\partial f(x)\} \geq \langle \bar{s},d\rangle = \left\langle \sum_{i=1}^{n}\alpha_i d^i, d\right\rangle = f'(x;d),$$

where the last equality follows by the construction of λ_1. ■

We finish this section with an easy but useful lemma regarding sums of functions

Lemma 4.6 *Let $f_1:\mathbb{R}^n\to\overline{\mathbb{R}}$ and $f_2:\mathbb{R}^n\to\overline{\mathbb{R}}$ be two convex functions. If $x\in$ **int dom**$(f_1)\cap$ **int dom**$(f_2)\neq\emptyset$, then,*

$$\partial f_1(x)+\partial f_2(x)=\partial(f_1+f_2)(x).$$

Proof. ($\subseteq$)Let $x\in$ **int dom**$(f_1)\cap$ **int dom**$(f_2)\neq\emptyset$ and $s^i\in\partial f_i(x)$, $i=1,2$. Then

$$f_i(y)\geq f_i(x)+\left\langle s^i, y-x\right\rangle$$

and,

$$f_1(y)+f_2(y)\geq f_1(x)+f_2(x)+\left\langle s^1+s^2, y-x\right\rangle.$$

($\supseteq$) Let $f:=f_1+f_2$, and take $s\in\partial f(x)$, and let $H=\partial f_1(x)+\partial f_2(x)$. Since $x\in$ **int dom**(f), we have that $\partial f_1(x)$ and $\partial f_2(x)$ are bounded, closed and convex sets. Thus, since H is a closed convex set. Using the Separating Hyperplane, there exists d, $\|d\|=1$, such that

$$\langle s,d\rangle > \langle s^s,d\rangle \qquad \text{for all } s^s\in H.$$

Since $x\in$ **int dom**$(f_1)\cap$**int dom**(f_2), there exists $\eta>0$ such that $x+\eta d\in$ **int dom**$(f_1)\cap$**int dom**(f_2). Since f is convex, we can compute the directional derivatives

$$\begin{aligned}(f_1+f_2)'(x;d) &= \lim_{t\to0}\tfrac{f_1(x+td)+f_2(x+td)-f_1(x)-f_2(x)}{t}\\ &= \lim_{t\to0}\tfrac{f_1(x+td)-f_1(x)}{t}+\lim_{t\to0}\tfrac{f_2(x+td)-f_2(x)}{t}\\ &= f_1'(x;d)+f_2'(x;d).\end{aligned}$$

Combining this with Theorem 4.1, we have that

$$\max\{\langle s, d\rangle : s \in \partial(f_1 + f_2)(x)\} = \max\{\langle s, d\rangle : s \in \partial f_1(x) + \partial f_2(x)\}$$

A contradiction since $\bar{s} \in \partial(f_1 + f_2)(x)$ such that $\langle \bar{s}, d\rangle > f_1'(x; d) + f_2'(x; d)$.

■

5 The ε-Subdifferential of a Convex Function

Additional theoretical work is needed to recover the lower semi-continuity property. We will construct an approximation of the subdifferential which has better properties. Among these properties, we will be able to prove the continuity of our new point-set mapping.

Definition 5.1 *For any $\varepsilon > 0$, the ε-subdifferential of a convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ at a point x is the point-to-set mapping*

$$\partial_\varepsilon f(x) = \{s \in \mathbb{R}^n : f(y) \geq f(x) + \langle s, y - x\rangle - \varepsilon, \quad \text{for all } y \in \mathbb{R}^n\}.$$

Remark 5.1 *For any $\varepsilon > 0$, the following properties are easy to verify:*

- $\partial f(x) \subseteq \partial_\varepsilon f(x)$;
- $\partial_\varepsilon f(x)$ *is a closed convex set;*
- $\partial f(x) = \bigcap_{\varepsilon > 0} \partial_\varepsilon f(x)$;

Now, we will revisit Example 4.2.

Example 5.1 *Let $f(x) = |x|$ with $x \in \mathbb{R}$ and let $\varepsilon > 0$ be fixed. By definition,*

$$\partial_\varepsilon f(x) = \{s \in \mathbb{R} : |y| \geq |x| + s(y - x) - \varepsilon, \quad \text{for all } y \in \mathbb{R}\}$$

First consider $y - x > 0$. Then,

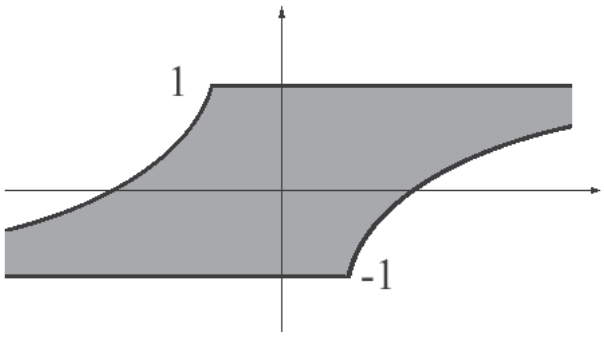
$$s \leq \frac{|y| - |x| + \varepsilon}{y - x} \Rightarrow s \leq \begin{cases} 1, & x > 0 \\ -1 + \frac{\varepsilon}{-x}, & x < 0 \end{cases}$$

On the other hand, if $y - x < 0$ we have that,

$$s \geq \frac{|y| - |x| + \varepsilon}{y - x} \Rightarrow s \geq \begin{cases} -1, & x > 0 \\ 1 - \frac{\varepsilon}{x}, & x < 0 \end{cases}$$

obtaining

$$\partial_\varepsilon f(x) = \begin{cases} \left[-1, -1 - \frac{\varepsilon}{x}\right], & x < -\frac{\varepsilon}{2} \\ [-1, 1], & x \in \left[\frac{\varepsilon}{2}, \frac{\varepsilon}{2}\right] \\ \left[1, 1 - \frac{\varepsilon}{x}\right], & x > \frac{\varepsilon}{2} \end{cases}$$



Before addressing the continuity of the operator, we answer the question of whether a subgradient at a some point is an approximation for subgradients at another point.

Proposition 5.1 *Let $x, x' \in \mathbf{dom}(f)$, and $s' \in \partial f(x')$. Then,*

$$s' \in \partial_\varepsilon f(x) \quad \textit{if and only if} \quad f(x') \geq f(x) + \langle s', x' - x \rangle - \varepsilon.$$

Proof. ($\Rightarrow$) From the definition of $\partial_\varepsilon f(x)$ using $y = x'$.
($\Leftarrow$) Since $s' \in \partial f(x')$,

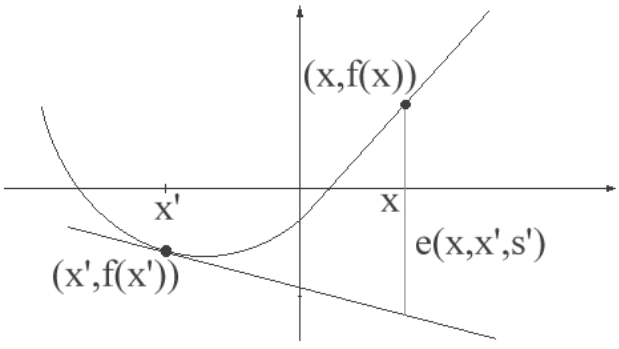
$$\begin{array}{rcl} f(y) & \geq & f(x') + \langle s', y - x' \rangle \\ & = & f(x) + \langle s', y - x \rangle + [f(x') - f(x) + \langle s', x - x' \rangle] \\ & \geq & f(x) + \langle s', y - x \rangle - \varepsilon \end{array}$$

where the last inequality follows from the assumption. So, $s' \in \partial_\varepsilon f(x)$. ■

This motivates the following definition, which will play an important role in the Bundle Methods to be defined later.

Definition 5.2 *For a triple $(x, x', s') \in \mathbf{dom}(f) \times \mathbf{dom}(f) \times \mathbb{R}^n$, the* linearization error *made at x, when f is linearized at x' with slope s', is the number*

$$\tilde{e}(x, x', s') := f(x) - f(x') - \langle s', x - x' \rangle .$$



Corollary 5.1 *Let $s' \in \partial_\eta f(x')$. Then, $s' \in \partial_\varepsilon f(x)$ if*
$f(x') \geq f(x) + \langle s', x' - x \rangle - \varepsilon + \eta$ *or, equivalently,* $\tilde{e}(x, x', s') + \eta \leq \varepsilon$.

6 Continuity of $\partial_\varepsilon f(\cdot)$

This section proves the continuity of the ε-subdifferential. To do so, we will need the following lemma:

Lemma 6.1 *Assume that* $\mathbf{dom}(f) \neq \emptyset$, $\delta > 0$, *and* $B(x,\delta) \subseteq \mathbf{int}\ \mathbf{dom}(f)$. *If* f *is* L*-Lipschitzian on* $B(x,\delta)$, *then, for any* $\delta' < \delta$, $s \in \partial_\varepsilon f(y)$, *and* $y \in B(x,\delta')$, *we have that*

$$\|s\| < L + \frac{\varepsilon}{\delta - \delta'}$$

Proof. Assume $s \neq 0$ and let $z := y + (\delta - \delta')\dfrac{s}{\|s\|}$. By definition of $s \in \partial_\varepsilon f(y)$

$$f(z) \geq f(y) + \langle s, z - y\rangle - \varepsilon$$

Now, note that $z \in B(x,\delta)$ and $y \in B(x,\delta)$. Thus, the Lipschitz constant L holds and we obtain that

$$L\|z - y\| \geq f(z) - f(y) \geq \langle s, z - y\rangle - \varepsilon$$

Using that $\|z - y\| = \|(\delta - \delta')\dfrac{s}{\|s\|}\| = (\delta - \delta')$,

$$(\delta - \delta')L \geq \left\langle s, (\delta - \delta')\frac{s}{\|s\|}\right\rangle - \varepsilon = \|s\|(\delta - \delta') - \varepsilon$$

we obtain the desired inequality. ■

Theorem 6.1 *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be a convex Lipschitz function on* $\mathbb{R}^n$. *Then, there exists a constant* $K > 0$ *such that for all* $x, x' \in \mathbb{R}^n$, $\varepsilon, \varepsilon' > 0$, $s \in \partial_\varepsilon f(x)$, *there exists* $s' \in \partial_{\varepsilon'} f(x')$ *satisfying*

$$\|s - s'\| \leq \frac{K}{\min\{\varepsilon, \varepsilon'\}}\left(\|x - x'\| + |\varepsilon - \varepsilon'|\right)$$

Proof. Since $\partial_\varepsilon f(x)$ and $\partial_{\varepsilon'} f(x)$ are convex sets[8], it is sufficient to show that for every d,

$$\max\{\langle s, d\rangle : s \in \partial_\varepsilon f(x)\} - \max\{\langle s', d\rangle : s' \in \partial_{\varepsilon'} f(x)\} < \frac{K\left(\|x - x'\| + |\varepsilon - \varepsilon'|\right)}{\min\{\varepsilon, \varepsilon'\}}.$$

So, we will fix a vector d, $\|d\| = 1$. Define the ε-directional derivate[9] of f at x in d as,

$$f'_\varepsilon(x; d) := \inf_{t>0} \frac{f(x + td) - f(x) + \varepsilon}{t}, \tag{2}$$

where the quotient on the right hand side will be denoted by

$$q_\varepsilon(x, t) = \frac{f(x + td) - f(x) + \varepsilon}{t}.$$

[8] Otherwise we could separate a point $s \in \partial_\varepsilon f(x)$ from $\partial_{\varepsilon'} f(x)$ using the Separating Hyperplane

[9] Note that the definition does not involve the $\lim_{t\to 0}$. It is the infimum of the values of the quocients $q_\varepsilon(x, t)$.

We note that Theorem 4.1 can be adapted to prove

$$f'_\varepsilon(x;d) = \max\{\langle s, d\rangle : s \in \partial_\varepsilon f(x)\}.$$

For any $\eta > 0$, there exists $t_\eta > 0$ such that

$$q_\varepsilon(x, t_\eta) \leq f'_\varepsilon(x, d) + \eta \tag{3}$$

Using Lemma 6.1 with $\delta \nearrow \infty$, we have that $f'_\varepsilon(x, d) \leq L$, and so $q_\varepsilon(x, t) < L + \eta$. On the other hand,

$$q_\varepsilon(x, t_\eta) = \frac{f(x + t_\eta d) - f(x) + \varepsilon}{t_\eta} \geq -L + \frac{\varepsilon}{t_\eta}$$

Therefore,

$$\frac{1}{t_\eta} \leq \frac{2L + \eta}{\varepsilon}. \tag{4}$$

Thus, using Equations (2) and (3)

$$\begin{aligned} f'_{\varepsilon'}(x';d) - f'_\varepsilon(x;d) - \eta &\leq q_{\varepsilon'}(x', t_\eta) - q_\varepsilon(x, t_\eta) \\ &= \frac{f(x' + t_\eta d) - f(x + t_\eta d) + f(x) - f(x') + \varepsilon' - \varepsilon}{t_\eta} \\ &\leq \frac{2L\|x - x'\| + |\varepsilon - \varepsilon'|}{t_\eta} \\ &\leq \frac{2L + \eta}{\varepsilon}(2L\|x - x'\| + |\varepsilon - \varepsilon'|) \end{aligned}$$

where the last inequality is obtained using Equation (4). Also, since $\eta > 0$ is arbitrary and one can interchange x and x', we obtain

$$|f'_{\varepsilon'}(x';d) - f'_\varepsilon(x;d)| \leq \frac{2L}{\min\{\varepsilon, \varepsilon'\}}(2L\|x - x'\| + |\varepsilon' - \varepsilon|)$$

The constant of interest can be chosen to be $K = \max\{2L, 4L^2\}$. ■

Corollary 6.1 *The correspondence $M : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathcal{B}(\mathbb{R}^n)$ defined by*

$$(x, \varepsilon) \mapsto \partial_\varepsilon f(x)$$

is a continuous correspondence.

Proof. The previous Lemma establishes the lower semi-continuity for any $\varepsilon > 0$.

The upper semi-continuity follows directly from f being lower semi-continuous. Take a sequence $(x^k, s^k, \varepsilon_k) \to (\bar{x}, \bar{s}, \bar{\varepsilon})$, then for all $y \in \mathbb{R}^n$,

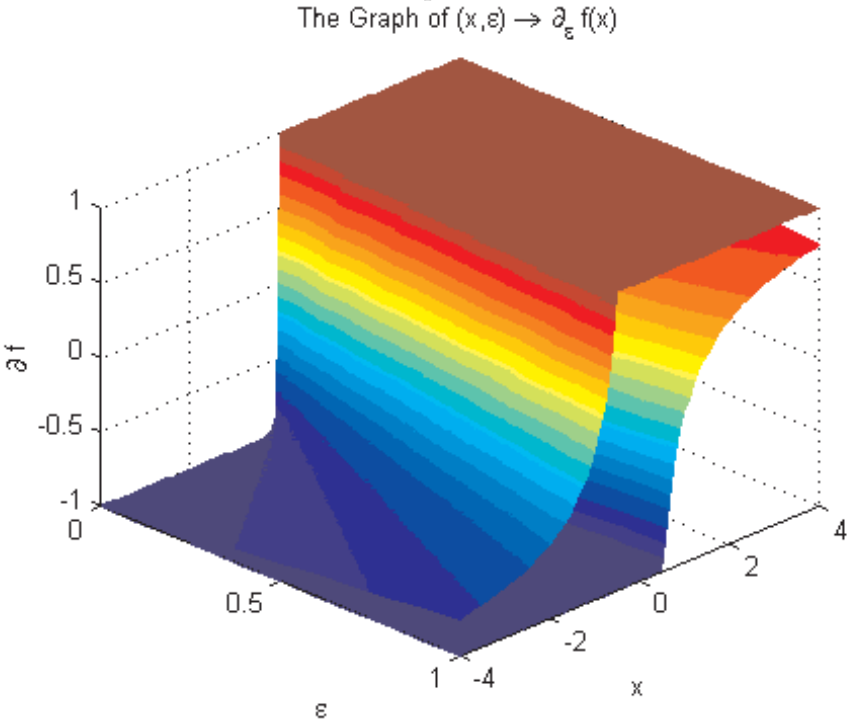
$$f(y) \geq f(x^k) + \left\langle s^k, y - x^k \right\rangle - \varepsilon_k \quad \text{for every } k.$$

Since f is lower semi-continuous, taking the limit of $k \to \infty$,

$$f(y) \geq f(\bar{x}) + \langle \bar{s}, y - \bar{x}\rangle - \bar{\varepsilon} \quad \text{for every } y \in \mathbb{R}^n.$$

Thus, $\bar{s} \in \partial_{\bar{\varepsilon}} f(\bar{x})$ and M is upper semi-continuous. ■

Example 6.1



7 Motivation through the Moreau-Yosida Regularization

Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function. We define the Moreau-Yosida regularization of f for a fixed $\lambda > 0$ as

$$F(x) = \min_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{\lambda}{2}\|y - x\|^2 \right\}.$$

$$p(x) = \arg\min_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{\lambda}{2}\|y - x\|^2 \right\}.$$

The proximal point $p(x)$ is well defined since the function being minimized is strictly convex.

Proposition 7.1 *If $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a proper convex function, then F is a finite-valued, convex and everywhere differentiable function with gradient given by*

$$\nabla F(x) = \lambda(x - p(x)).$$

Moreover,

$$\|\nabla F(x) - \nabla F(x')\|^2 \leq \lambda \langle \nabla F(x) - \nabla F(x'), x - x' \rangle \quad \text{for all } x, x' \in \mathbb{R}^n$$

and

$$\|\nabla F(x) - \nabla F(x')\| \leq \lambda \|x - x'\| \quad \text{for all } x, x' \in \mathbb{R}^n.$$

Proof.

(Finite-valued). If f is a proper function, there exists $x \in \mathbb{R}^n$ such that $f(x) < \infty$. Thus, F is finite-valued since

$$F(x') \leq f(x) + \frac{\lambda}{2}\|x - x'\|^2 < \infty.$$

*(Convexity).*Take $x, x' \in \mathbb{R}^n$, and $\alpha \in [0,1]$, then

$$\begin{array}{rcl} \alpha F(x) + (1-\alpha)F(x') & = & \alpha f(p(x)) + (1-\alpha) f(p(x')) + \frac{\lambda}{2}\left(\alpha\|p(x) - x\|^2 + (1-\alpha)\|p(x') - x'\|^2\right) \\ & \geq & f(\alpha p(x) + (1-\alpha)p(x')) + \frac{\lambda}{2}\|\alpha p(x) + (1-\alpha)p(x') - (\alpha x + (1-\alpha)x')\|^2 \\ & \geq & F(\alpha x + (1-\alpha)x') \end{array}$$

(Differentiability). Fix an arbitrary direction d, $\|d\| = 1$, and let $t > 0$. By definition

$$\begin{array}{rcl} \frac{F(x+td)-F(x)}{t} & = & \dfrac{\min_y \left[f(y) + \frac{\lambda}{2}\|y - x - td\|^2\right] - \min_w \left[f(w) + \frac{\lambda}{2}\|w - x\|^2\right]}{t} \\ & \geq & \dfrac{\left[f(p(x+td)) + \frac{\lambda}{2}\|p(x+td) - x - td\|^2\right] - \left[f(p(x+td)) + \frac{\lambda}{2}\|p(x+td) - x\|^2\right]}{t} \\ & = & \dfrac{\lambda}{2}\dfrac{\|p(x+td) - x - td\|^2 - \|p(x+td) - x\|^2}{t} \\ & = & \dfrac{\lambda}{2}\dfrac{\|p(x+td) - p(x) + p(x) - x - td\|^2 - \|p(x+td) - p(x) + p(x) - x\|^2}{t} \\ & = & \dfrac{\lambda}{2}\dfrac{\|p(x) - x - td\|^2 - \|p(x) - x\|^2}{t} - \lambda\left\langle p(x+td) - p(x), d\right\rangle \end{array}$$

where we used that $F(x) \leq f(p(x+td)) + \frac{\lambda}{2}\|p(x+td) - x\|^2$. Now, taking the limit as $t \to 0$, $p(x+td) \to p(x)$ implies that

$$\lim_{t\to 0} \frac{F(x+td) - F(x)}{t} \geq \left\langle \lambda(x - p(x)), d\right\rangle.$$

On the other hand, since $F(x+td) \leq f(p(x)) + \frac{\lambda}{2}\|p(x) - x - td\|^2$,

$$\begin{array}{rcl} \frac{F(x+td)-F(x)}{t} & = & \dfrac{\min_y \left[f(y) + \frac{\lambda}{2}\|y - x - td\|^2\right] - \min_w \left[f(w) + \frac{\lambda}{2}\|w - x\|^2\right]}{t} \\ & \leq & \dfrac{\left[f(p(x)) + \frac{\lambda}{2}\|p(x) - x - td\|^2\right] - \left[f(p(x)) + \frac{\lambda}{2}\|p(x) - x\|^2\right]}{t} \\ & = & \dfrac{\lambda}{2}\dfrac{\|p(x) - x - td\|^2 - \|p(x) - x\|^2}{t} \end{array}$$

Again, taking the limit $t \to 0$,

$$\lim_{t\to 0} \frac{F(x+td) - F(x)}{t} \leq \left\langle \lambda(x - p(x)), d\right\rangle.$$

$$\begin{array}{rcl} \|\nabla F(x) - \nabla F(x')\|^2 & = & \lambda\left\langle \nabla F(x) - \nabla F(x'), x - p(x) - x' + p(x')\right\rangle \\ & = & \lambda\left\langle \nabla F(x) - \nabla F(x'), x - x'\right\rangle + \lambda\left\langle \nabla F(x) - \nabla F(x'), p(x') - p(x)\right\rangle \end{array}$$

We will prove that the last term is negative. To do so, we need the following (known as the monotonicity of $\partial f(x)$). Let $s \in \partial f(x)$ and $s' \in \partial f(x')$,

then $\langle s - s', x - x' \rangle \geq 0$. To prove that, consider the subgradient inequalities associated with s and s' applied to x' and x respectively,

$$f(x') \geq f(x) + \langle s, x' - x \rangle \text{ and } f(x) \geq f(x') + \langle s, x' - x \rangle .$$

Adding these inequalities we obtain $\langle s - s', x - x' \rangle \geq 0$.

Now, note that by optimality of $p(x)$ and $p(x')$,

$$0 \in \partial \left(f(p(x)) + \frac{\lambda}{2}\|p(x) - x\|^2 \right) \text{ and } 0 \in \partial \left(f(p(x')) + \frac{\lambda}{2}\|p(x') - x'\|^2 \right).$$

So, $\nabla F(x) = \lambda(x - p(x)) \in \partial f(p(x))$ and $\nabla F(x') = \lambda(x' - p(x')) \in \partial f(p(x'))$. Using the monotonicity of ∂f,

$$\langle \nabla F(x) - \nabla F(x'), p(x') - p(x) \rangle \leq 0.$$

The proof of the last inequality follows directly from the previous result,

$$\|\nabla F(x) - \nabla F(x')\|^2 \leq \lambda \langle \nabla F(x) - \nabla F(x'), x - x' \rangle \leq \lambda \|\nabla F(x) - \nabla F(x')\| \|x - x'\|$$

Therefore, $\|\nabla F(x) - \nabla F(x')\| \leq \lambda \|x - x'\|$. ■

We finish this section with a characterization of points of minimum of f and its regularization F.

Proposition 7.2 *The following statements are equivalent:*

(i) $\bar{x} \in \arg\min\{f(y) : y \in \mathbb{R}^n\}$;

(ii) $\bar{x} = p(\bar{x})$;

(iii) $\nabla F(\bar{x}) = 0$;

(iv) $\bar{x} \in \arg\min\{F(y) : y \in \mathbb{R}^n\}$;

(v) $f(\bar{x}) = f(p(\bar{x}))$;

(vi) $f(\bar{x}) = F(\bar{x})$;

Proof.

(i) $\Rightarrow$ *(ii)* $f(\bar{x}) = f(\bar{x}) + \frac{\lambda}{2}\|\bar{x} - \bar{x}\|^2 \leq f(y) + \frac{\lambda}{2}\|y - \bar{x}\|^2 \Rightarrow p(\bar{x}) = \bar{x}$.

(ii) $\Leftrightarrow$ *(iii)* $\bar{x} = p(\bar{x}) \Leftrightarrow \nabla F(\bar{x}) = \lambda(\bar{x} - p(\bar{x})) = 0$.

(iii) $\Leftrightarrow$ *(iv)* Since F is convex, $\nabla F(\bar{x}) = 0 \Leftrightarrow \bar{x} \in \arg\min\{F(y) : y \in \mathbb{R}^n\}$.

(iv) $\Rightarrow$ *(v)* Using *(ii)*, $\bar{x} = p(\bar{x})$ implies that $f(\bar{x}) = f(p(\bar{x}))$.

(v) $\Rightarrow$ *(vi)* $f(\bar{x}) = f(p(\bar{x})) \leq f(p(\bar{x})) + \frac{\lambda}{2}\|p(\bar{x}) - \bar{x}\|^2 = F(\bar{x}) \leq f(\bar{x})$.

(vi) $\Rightarrow$ *(i)* $F(\bar{x}) = f(\bar{x})$ implies that $p(\bar{x}) = \bar{x}$. Thus,

$$0 \in \partial f(\bar{x}) + \lambda(\bar{x} - p(x)) = \partial f(\bar{x}).$$

■

Remark 7.1 *During the past ten years, several authors tried to introduce second order information via the Moreau-Yosida regularization. It is possible to show that the Hessian $\nabla^2 F(x)$ exists if and only if the Jacobian $\nabla p(x)$ exists obtaining*

$$\nabla^2 F(x) = \lambda \left(I - \nabla p(x) \right) \quad \textit{for all } x \in \mathbb{R}^n.$$

8 Nondifferentiable Optimization Methods

In the next section, we will state a Bundle Method for minimizing a convex function $f : \mathbb{R}^n \to \mathbb{R}$ given only by an oracle. Before proceeding to discuss the Bundle Methods in detail, we introduce two other methods which will contextualize the Bundle Method in the literature.

8.1 Subgradient Method

This is probably the most used method for nondifferentiable convex optimization. Basically, it consists of replacing gradients with subgradients in the classical steepest descent method.

Subgradient Method (SM)

Step 1. Start with a x^0, set $k = 0$ and a tolerance $\varepsilon > 0$.

Step 2. Compute $f(x^k)$ and a subgradient $s^k \in \partial f(x^k)$.

Step 3. If $\|s^k\| < \varepsilon$, STOP.

Step 4. Define a stepsize $\theta^k > 0$.

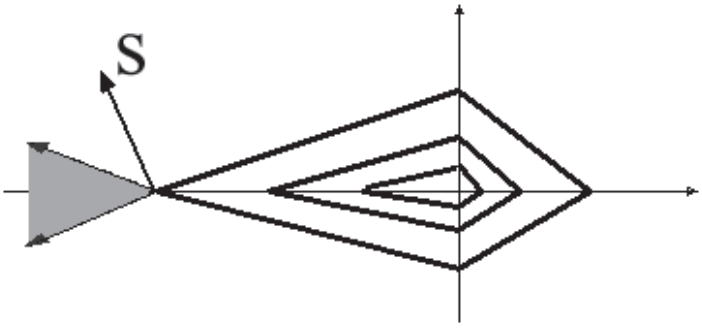
Step 5. Define $x^{k+1} = x^k - \dfrac{\theta^k s^k}{\|s^k\|}$.

Step 6. Set $k = k + 1$, and GOTO **Step 2**.

We refer to [12] for a detailed study. The choice of the sequence $\{\theta^k\}$ is crucial in the performance of this algorithm. There are choices which convergence is guaranteed. Let $\theta^k \to 0$ and $\sum_{k\geq 1} \theta^k = \infty$. Then, $f(x^k) \to f^*$ the optimal value. Moreover, if in addition the stepsize sequence satisfies $\sum_{k\geq 1}(\theta^k)^2 < \infty$, we have that $x^k \to \bar{x}$ a solution of the problem[10]. Several other rules have been proposed to speed up convergence especially when bounds on the optimal value are available.

An important difference between the differentiable case and the non-differentiable case concerns descent directions and line searches. A direction opposite to a particular subgradient does not need to be a descent direction (as opposed to minus the gradient for the differentiable case). In fact, a vector d is a descent direction for f at x only if

$$\langle d, s\rangle < 0 \text{ for all } s \in \partial f(x).$$



[10]Note that the second condition automatically implies in sublinear convergence.

Thus, traditional line searches first proposed for differentiable functions must be adapted before being applied.

Finally, the stopping criterion is too strict here. As illustrated by Example 4.2, this criterion may never be met even if we are at the solution.

8.2 Cutting Planes Method

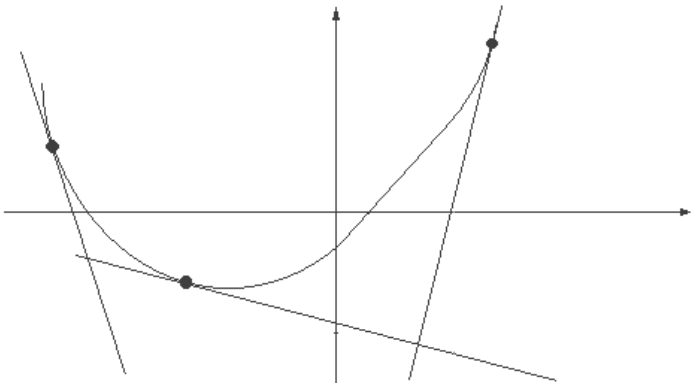
The simplicity of the previous algorithm comes at the price of ignoring past information. The information obtained by the oracle can be used to build not only descent directions, but also a model of the function f itself. This is also be used in the Bundle Methods but was first developed by the Cutting Plane Method.

Consider the bundle of information formed by $\{y^i, f(y^i), s^i \in \partial f(y^i)\}_{i=1}^{\ell}$. Remembering that for all $y \in \mathbb{R}^n$ and $i = 1, \ldots, \ell$,

$$f(y) \geq f(y^i) + \left\langle s^i, y - y^i \right\rangle,$$

we can construct the following piecewise-linear approximation of f,

$$\hat{f}_\ell(y) := \max_{i=1,\ldots,\ell} f(y^i) + \left\langle s^i, y - y^i \right\rangle. \tag{5}$$



By construction, we have that $\hat{f}_\ell(y) \leq f(y)$ for all $y \in \mathbb{R}^n$. Also, it is clear that $\hat{f}_\ell(y) \leq \hat{f}_{\ell+1}(y)$ for all $y \in \mathbb{R}^n$. There are at least two motivations for this cutting-plane model. First, it can be modelled by linear constraints, so optimizing the model reduces to a LP. The second motivation is the following lemma.

Lemma 8.1 *Let f be any lower semicontinuous convex function. Denote by $\mathcal{H}_f$ the set of affine functions minoring f, formally defined as*

$$\mathcal{H}_f = \{h : \mathbb{R}^n \to \mathbb{R} : h \text{ is an affine function such that } h(y) \leq f(y) \text{ for all } y \in \mathbb{R}^n\}.$$

Then,

$$f(x) = \sup\{h(x) : h \in \mathcal{H}_f\} \quad \text{and} \quad \partial f(x) = \{\nabla h(x) : h(x) = f(x),\ h \in \mathcal{H}_f\}.$$

Proof. The epigraph of f is closed. So, any point $(x,t) \notin \mathbf{epi}(f)$, can be strictly separated by an affine hyperplane $h_{(x,t)} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, defined as

$$h_{(x,t)}(y,s) = \langle s^x, x - y \rangle + (s - t),$$

from $\mathbf{epi}(f)$. In fact, we can take the supremum over $t \leq f(x)$ and to obtain,

$$h_x(y,s) = \langle s^x, x - y \rangle + s - f(x).$$

The semi-space $H_x = \{(y,s) \in \mathbb{R}^n \times \mathbb{R} : h_x(y,s) \geq 0\}$ contains $\mathbf{epi}(f)$. Thus, the intersection of all such hyperplanes is a closed convex set containing $\mathbf{epi}(f)$ and contained on $\mathbf{epi}(f)$ (since any point not in the epigraph can be separated). The result follows from the fact that arbitrary intersection of the epigraph of lower semicontinuous functions equals the epigraph of the maximum of these functions.

For the second statement, the direction ($\supseteq$) is by definition of $\partial f(x)$. To prove ($\subseteq$), note that any $s \in \partial f(x)$ induces an affine function

$$h_s(y) = f(x) + \langle s, y - x \rangle,$$

for which holds that $h_s(x) = f(x)$ and $h_s(y) \leq f(y)$ for all $y \in \mathbb{R}^n$. Thus, h_s is included in the supremum and $\nabla h_s(x) = s$. ■

In order for the method to be well defined, we still need to specify a compact set C and restrict our model to this set to ensure that every iterate is well defined.

Cutting Plane Method (CP)

Step 1. Let $\hat{\delta} > 0$ and consider a convex compact set C containing a minimizing point. Let $k = 1$, $y^1 \in C$, and define $\hat{f}_0 \equiv -\infty$.

Step 2. Compute $f(y^k)$ and a subgradient $s^k \in \partial f(y^k)$.

Step 3. Define $\delta_k := f(y^k) - \hat{f}_{k-1}(y^k) \geq 0$.

Step 4. If $\delta_k < \hat{\delta}$, STOP.

Step 5. Update Model $\hat{f}_k(y) := \max\{\hat{f}_{k-1}(y), f(x^k) + \langle s^k, y - y^k \rangle\}$.

Step 6. Compute $y^{k+1} \in \arg\min_{y \in C} \hat{f}_k(y)$.

Step 7. Set $k = k + 1$, and GOTO **Step 2**.

This method has several advantages when compared with the Subgradient Method. The model $\hat{f}_{k-1}$ provides us with a stopping criterion based on δ_k which did not exist for the Subgradient Method. The value of δ_k is called nominal decrease, the improvement predicted by the model in the objective function for the next iterate y^{k+1}. Since the model satisfies $\hat{f}_k(y) \leq f(y)$, we have that

$$\min_y \hat{f}_k(y) \leq \min_y f(y).$$

Thus, if the stopping criterion is satisfied we have that $f(y^k) - \hat{f}_{k-1}(y^k) < \hat{\delta}$ implying that

$$f(y^k) < \hat{\delta} + \hat{f}_k(y^k) = \hat{\delta} + \min_y \hat{f}_k(y) \leq \hat{\delta} + \min_y f(y).$$

This is achieved at the cost of solving a linear programming on **Step 5** in order to define the next point to be evaluated by the oracle. This definition of the next iterate also allows the algorithm to perform long steps on each iteration[11].

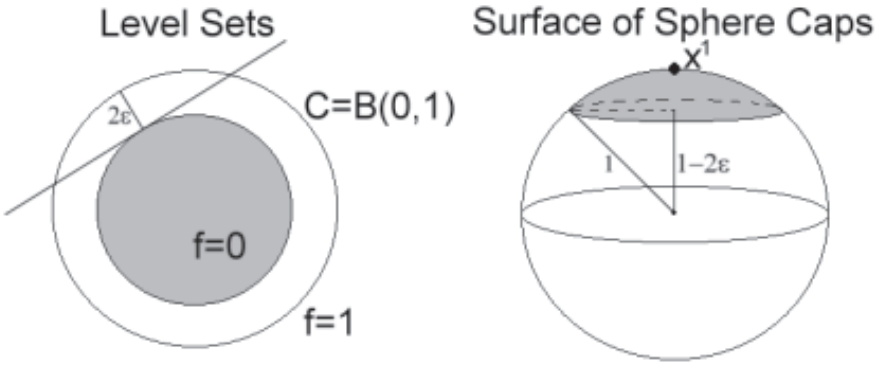
It is clear that the set C plays an important role in the performance of the algorithm, for example, we are requiring for C to containt some optimal solution. In fact, the convergence might depend a lot on this set. This is illustraded by the following example proposed by Nemirovskii.

Example 8.1 *For $0 < \varepsilon < 1$, consider the convex function*

$$f(x) := \max\{0, -1 + 2\varepsilon + \|x\|\},$$

and its subdifferential

$$\partial f(x) = \begin{cases} \{0\}, & \text{if } \|x\| < 1 - 2\varepsilon; \\ \{x/\|x\|\}, & \text{if } \|x\| > 1 - 2\varepsilon; \\ \mathbf{conv}\{0, x/\|x\|\}, & \text{if } \|x\| = 1 - 2\varepsilon. \end{cases}$$



Any point with norm smaller or equal to $1 - 2\varepsilon$ is optimal. Suppose that the solver for the subproblem in **Step 5** *always returns a solution with maximal norm. We initialize the algorithm with the set $C = B(0, 1)$, $k = 1$, $y^1 = 0 \in C$, $\hat{\delta} < 2\varepsilon$, and $\hat{f}_0 \equiv -\infty$.*

We obtain $f(y^1) = 0$ and $s^1 = 0$. Since $\delta_1 = +\infty$, we do not stop. Now, we update the model to obtain $\hat{f}_1 \equiv 0$, and compute the next candidate $y^2 \in C = \arg\min_C 0$. The oracle will return a solution with norm one. So, we have $f(y^2) = 2\varepsilon$ and $\delta_2 = 2\varepsilon$. Once again, the stopping criteria is not satisfied.

The model now is updated to

$$\hat{f}_2(y) = \max\{\hat{f}_1(y), 2\varepsilon + \langle y^2, y - y^2\rangle = \max\{0, -1 + 2\varepsilon + \langle y^2, y\rangle\}.$$

We observe that $\hat{f}_2$ will be zero except on the sphere cap $S_2 = \{y \in C : \langle x^1, y\rangle > 1 - 2\varepsilon\}$ (this is illstrated in Figure 8.1). From this observation, the computation of the next iterate becomes

$$y^3 \in \arg\min_{y \in C} \hat{f}_2(y) = C \setminus S_2.$$

[11] As it will become clear in the next example, this might be a drawback.

In the subsequent iterations, another sphere cap (exactly the same up to a rotation) will have positive values in the respective model. The next iterate is computed as

$$y^{k+1} \in \arg\min_{y \in C} \hat{f}_k(y) = C \setminus \bigcup_{i=2}^{k} S_i.$$

Consequently, as long as there exists vectors with norm one in $C \setminus \bigcup_{i=2}^{k} S_i$*, a solution with norm one will be returned by the solver, obtaining* $f(y^{k+1}) = 2\varepsilon$ *and the algorithm does not stop.*

Thus, the Cutting Plane will require at least $\mathsf{Vol}_{n-1}(\partial B(0,1))/\mathsf{Vol}_{n-1}(\partial S_2)$ *iterations before terminating. Denote by* $\nu_n = \mathsf{Vol}_{n-1}(\partial B(0,1))$ *the* $n-1$*-dimensional volume of the surface of a* n*-dimensional sphere,*

$$\mathsf{Vol}_{n-1}(\partial B(0,1)) = \nu_n = 2\nu_{n-1}\int_0^1 (1-t^2)^{\frac{n-1}{2}}dt \geq \nu_{n-1}\int_0^1 2t(1-t^2)^{\frac{n-1}{2}}dt \geq \frac{2}{n+1}\nu_{n-1},$$

$$\begin{aligned}\mathsf{Vol}_{n-1}(\partial S_2) = \nu_{n-1}\int_{1-2\varepsilon}^1 (1-t^2)^{\frac{n-1}{2}}dt &\leq \nu_{n-1}(1-(1-2\varepsilon)^2)^{(n-1)/2}(1-(1-2\varepsilon)) \\ &\leq \nu_{n-1}\left(4\varepsilon - 4\varepsilon^2\right)^{(n+1)/2}.\end{aligned}$$

So, we are required to perform at least

$$\frac{\mathsf{Vol}_{n-1}(\partial B(0,1))}{\mathsf{Vol}_{n-1}(\partial S_2)} \geq \frac{2}{n+1}\left(\frac{1}{4\varepsilon}\right)^{(n+1)/2}.$$

This is associated with the possibility of performing long steps and possible generating a zig-zag behavior of the iterates. Another important remark to be made is that the number of constraints in the LP is growing with the number of iterations.

9 Bundle Methods

Motivated by the theorems in the previous sections, we define another method which can be seen as a stabilization of the Cutting Plane Method.

We start by adding an extra point called the center, $\hat{x}^k$, to the bundle of information. We will still make use of the same piecewise-linear model for our function, $\hat{f}$, but no longer solve an LP on each iteration. Instead, we will compute the next iterate by computing the Moreau-Yosida regularization for $\hat{f}_k$ at $\hat{x}^k$.

Bundle Algorithm (BA)

Step 1. Let $\bar{\delta} > 0$, $m \in (0,1)$, $\hat{x}^0$, $y^0 = \hat{x}^0$, and $k = 0$. Compute $f(\hat{x}^0)$ and $s^0 \in \partial f(\hat{x}^0)$. Define $\hat{f}_0(y) = f(\hat{x}^0) + \langle s^0, y - \hat{x}^0 \rangle$.

Step 2. Compute the next iterate

$$y^{k+1} \in \arg\min_{y \in \mathbb{R}^n} \hat{f}_k(y) + \frac{\mu_k}{2}\|y - \hat{x}^k\|^2 \tag{6}$$

Step 3. Define $\delta_k := f(\hat{x}^k) - \left[\hat{f}_k(y^{k+1}) + \frac{\mu_k}{2}\|y^{k+1} - \hat{x}^k\|^2\right] \geq 0$.

Step 4. If $\delta_k < \bar{\delta}$ STOP.

Step 5. Compute $f(y^{k+1})$ and a subgradient $s^{k+1} \in \partial f(y^{k+1})$.

Step 6. If $f(\hat{x}^k) - f(y^{k+1}) \geq m\delta_k$, SERIOUS STEP (SS) $\hat{x}^{k+1} = y^{k+1}$.

Else, NULL STEP (NS) $\hat{x}^{k+1} = \hat{x}^k$.

Step 7. Update the Model

$$\hat{f}_{k+1}(y) := \max\{\hat{f}_k(y), f(y^{k+1}) + \langle s^{k+1}, y - y^{k+1}\rangle\}.$$

Step 8. Set $k = k + 1$, and goto **Step 2**.

The quadratic term in the relation (6) is responsible for interpreting this as a stabilization of the Cutting Planes Method. It will make the next iterate closer to the current center by avoiding drastic movements as in the case of Cutting Planes. The role of the parameter μ_k is exactly to control the trade off between minimizing the model $\hat{f}$ and staying close to a point $\hat{x}^k$ which is known to be good.

It will be important to highlight the subsequence of iterations where we performed a Serious Step. We denote such iterations by $K_s = \{k \in \mathbb{N} : k\text{th iteration was a SS}\}$. We also note that the sequence $\{f(\hat{x}^k)\}_{k\in K_s}$ is strictly decreasing due to the Serious Step test in **Step 6**. The SS test requires that the new center must improve by at least a fixed fraction m from the improvement δ_k predicted by the model.

This is a basic version of the method and several enhancements such as line searches, updates of the parameter μ_k, and other operations may also be incorporated.

9.1 A Dual View

Before we move to convergence proofs, it will be more convenient in theory and practice to work with the dual of the QP in (6).

Now, we will rewrite our model $\hat{f}_k$ more conveniently. Assuming that our model has $\ell \leq k$ pieces, we define the linearization errors at the center $\hat{x}^k$ to be

$$\tilde{e}_i := f(\hat{x}^k) - \left[f(y^i) - \langle s^i, \hat{x}^k - y^i\rangle\right] \quad \text{for } i = 1, \ldots, \ell.$$

Thus, our model becomes

$$\begin{aligned}
\hat{f}_k(y) &= \max_{i=1,\ldots,\ell}\{f(y^i) + \langle s^i, y - y^i\rangle\} \\
&= \max_{i=1,\ldots,\ell}\{f(\hat{x}^k) - \tilde{e}_i - \langle s^i, \hat{x}^k - y^i\rangle + \langle s^i, y - y^i\rangle\} \\
&= f(\hat{x}^k) + \max_{i=1,\ldots,\ell}\{-\tilde{e}_i + \langle s^i, y - \hat{x}^k\rangle\}
\end{aligned}$$

In fact, our bundle of information can be kept as

$$\left(\hat{x}^k, \left\{s^i, \tilde{e}^i\right\}_{i=1}^{\ell}\right),$$

since it is straightforward to update the linearization errors as we change the center.

Consider the quadratic programming problem which compute the new candidate y^{k+1}:

$$\min_{y\in\mathbb{R}^n} \hat{f}_k(y) + \frac{\mu_k}{2}\|y-\hat{x}^k\|^2 = \min_{y,r} \quad r + \frac{1}{2}\left\langle y-\hat{x}^k, y-\hat{x}^k\right\rangle \\ r \geq f(\hat{x}^k) - \tilde{e}_i + \left\langle s^i, y-\hat{x}^k\right\rangle \quad \text{for } i=1,\ldots,\ell. \tag{7}$$

Introducing multipliers $\alpha \in \mathbb{R}^\ell_+$, we can write the Lagrangian as

$$L(y,r,\alpha) = \left(1-\sum_{i=1}^{\ell}\alpha_i\right) r + \frac{\mu_k}{2}\|y-\hat{x}^k\|^2 + \sum_{i=1}^{\ell}\alpha_i(f(\hat{x}^k) - \tilde{e}_i + \left\langle s^i, y-\hat{x}^k\right\rangle).$$

Thus, by convexity, we have that

$$\min_{y,r}\max_{\alpha} L(y,r,\alpha) = \max_{\alpha}\min_{y,r} L(y,r,\alpha). \tag{8}$$

Since both sides are finite, we must have $\left(1-\sum_{i=1}^{\ell}\alpha_i\right)=0$. Also, optimality conditions on y impose that

$$\nabla_y L(y,r,\alpha) = 0 = \mu_k(y-\hat{x}^k) + \sum_{i=1}^{\ell}\alpha_i s^i \therefore \mu_k(y-\hat{x}^k) = -\sum_{i=1}^{\ell}\alpha_i s^i \tag{9}$$

Denoting by $\Delta^\ell = \{\alpha\in\mathbb{R}^\ell_+ : \sum_{i=1}^{\ell}\alpha_i = 1\}$ the simplex in $\mathbb{R}^\ell$, we plug equation (9) on (8). Then, our original QP is equivalent to

$$\max_{\alpha\in\Delta^\ell} \frac{\mu_k}{2}\left\|-\frac{\sum_{i=1}^{\ell}\alpha_i s^i}{\mu_k}\right\|^2 + \sum_{i=1}^{\ell}\alpha_i\left(f(\hat{x}^k) - \tilde{e}_i + \left\langle s^i, -\frac{\sum_{i=1}^{\ell}\alpha_i s^i}{\mu_k}\right\rangle\right)$$

$$= f(\hat{x}^k) + \max_{\alpha\in\Delta^\ell} -\frac{1}{2\mu_k}\left\|\sum_{i=1}^{\ell}\alpha_i s^i\right\|^2 - \sum_{i=1}^{\ell}\alpha_i\tilde{e}_i \tag{10}$$

It will be convenient to work with the convex combination of the linearization errors and subgradients given by the optimal solution of (10).

Definition 9.1 *Given an optimal solution $\alpha\in\Delta^\ell$ for (10) at iteration k, define the aggregated subgradient and aggregated linearization error respectively as*

$$\hat{s}^k = \sum_{i=1}^{\ell}\alpha_i s^i \qquad \textit{and} \qquad \hat{e}_k = \sum_{i=1}^{\ell}\alpha_i\tilde{e}_i.$$

Lemma 9.1 *Let $\alpha\in\Delta^\ell$ be a solution for (10). Then*

(i) $\hat{s}^k \in \partial\hat{f}_k(y^{k+1})$;

(ii) $\hat{f}_k(y^{k+1}) = f(\hat{x}^k) - \frac{1}{\mu_k}\left\|\hat{s}^k\right\|^2 - \hat{e}_k$;

(iii) $\delta_k = \frac{1}{2\mu_k}\|\hat{s}^k\|^2 + \hat{e}_k$.

Proof. (i) From equation (9), $\mu_k(y^{k+1} - \hat{x}^k) + \hat{s}^k = 0$. So,

$$-\mu_k(y^{k+1} - \hat{x}^k) = \hat{s}^k$$

and since y^{k+1} is optimal for (6),

$$0 \in \left(\partial \hat{f}_k(y^{k+1}) + \mu_k(y^{k+1} - \hat{x}^k)\right) \therefore -\mu_k(y^{k+1} - \hat{x}^k) \in \partial \hat{f}_k(y^{k+1}).$$

(ii) Due to convexity, there is no duality gap between (7) and (10). Thus,

$$\hat{f}_k(y^{k+1}) + \frac{\mu_k}{2}\|y^{k+1} - \hat{x}^k\|^2 = f(\hat{x}^k) - \frac{1}{2\mu_k}\|\hat{s}^k\|^2 - \hat{e}_k$$

$$\begin{aligned}\hat{f}_k(y^{k+1}) &= f(\hat{x}^k) - \tfrac{\mu_k}{2}\left\|\tfrac{-1}{\mu_k}\hat{s}^k\right\|^2 - \tfrac{1}{2\mu_k}\|\hat{s}^k\|^2 - \hat{e}_k \\ &= f(\hat{x}^k) - \tfrac{1}{\mu_k}\|\hat{s}^k\|^2 - \hat{e}_k.\end{aligned}$$

(iii) Using (ii) in the definition of δ_k,

$$\begin{aligned}\delta_k &= f(\hat{x}^k) - \hat{f}_k(y^{k+1}) - \tfrac{\mu_k}{2}\|y^{k+1} - \hat{x}^k\|^2 \\ &= f(\hat{x}^k) - \tfrac{\mu_k}{2}\|y^{k+1} - \hat{x}^k\|^2 - f(\hat{x}^k) + \tfrac{1}{\mu_k}\|\hat{s}^k\|^2 + \hat{e}_k \\ &= \tfrac{1}{2\mu_k}\|\hat{s}^k\|^2 + \hat{e}_k.\end{aligned}$$

■

Lemma 9.2 *For the aggregated subgradient and linearization error, it holds that*

$$\hat{s}^k \in \partial_{\hat{e}_k} f(\hat{x}^k).$$

Proof. Using Lemma 9.1 (i), $\hat{s}^k \in \partial \hat{f}_k(y^{k+1})$, and by construction $f \geq \hat{f}_k$. Therefore,

$$\begin{aligned}f(y) &\geq \hat{f}_k(y) \geq \hat{f}_k(y^{k+1}) + \left\langle \hat{s}^k, y - y^{k+1}\right\rangle \\ &= f(\hat{x}^k) - \tfrac{1}{\mu_k}\|\hat{s}^k\|^2 - \hat{e}_k + \left\langle \hat{s}^k, y - \hat{x}^k + \hat{x}^k - y^{k+1}\right\rangle \\ &= f(\hat{x}^k) + \left\langle \hat{s}^k, y - \hat{x}^k\right\rangle - \hat{e}_k + \left\langle \hat{s}^k, \hat{x}^k - y^{k+1}\right\rangle - \tfrac{1}{\mu_k}\|\hat{s}^k\|^2 \\ &= f(\hat{x}^k) + \left\langle \hat{s}^k, y - \hat{x}^k\right\rangle - \hat{e}_k,\end{aligned}$$

where we used that $\hat{x}^k - y^{k+1} = \frac{1}{\mu_k}\hat{s}^k$. ■

Lemmas 9.1 and 9.2 motivates our convergence proofs in the next section. If we manage to prove that $\delta_k \to 0$, we also proved that $\|\hat{s}^k\| \to 0$ and $\hat{e}_k \to 0$ (as long as the parameter μ_k is bounded). On the other hand, $\hat{s}^k \in \partial_{\hat{e}_k} f(\hat{x}^k)$. So, if $\hat{x}^k \to \bar{x}$, we obtain that $0 \in \partial f(\bar{x})$ since the underlying correspondence is upper semi-continuous.

9.2 Aggregation

In the Cutting Plane Method, the number of constraints in the linear programming problem is growing with the number of iterations. This is associated with the model $\hat{f}_k$, so we are subject to the same problem in the context of Bundle Methods. In fact, since we need to solve a quadratic probramming problem, we have a potentially more serious problem to deal with.

In order to avoid this problem, we can introduce an additional Step in (BA). Based on the optimal solution of (10), we will discard at least two linear pieces of our model, and introduce the gradient associated with the next iterate and another piece based on the aggregated subgradient and linearization error. Define the aggregated linear piece as

$$f_a(y) = f(\hat{x}^k) + \left\langle \hat{s}^k, y - \hat{x}^k \right\rangle - \hat{e}_k. \tag{11}$$

Lemma 9.3 *For f_a defined by (11), it holds that*

(i) $f_a(y) = \hat{f}_k(y^{k+1}) + \left\langle \hat{s}^k, y - y^{k+1} \right\rangle$,

(ii) $f_a(y) \leq \hat{f}_k(y)$.

Proof. (i) Using Lemma 9.1 (ii),

$$\begin{aligned} f_a(y) &= f(\hat{x}^k) + \left\langle \hat{s}^k, y - y^{k+1} \right\rangle + \left\langle \hat{s}^k, y^{k+1} - \hat{x}^k \right\rangle - \hat{e}_k \\ &= f(\hat{x}^k) + \left\langle \hat{s}^k, y - y^{k+1} \right\rangle + \left\langle \hat{s}^k, \tfrac{-\hat{s}^k}{\mu_k} \right\rangle + \hat{f}_k(y^{k+1}) - f(\hat{x}^k) + \tfrac{1}{\mu_k}\|\hat{s}^k\|^2 \\ &= \hat{f}_k(y^{k+1}) + \left\langle \hat{s}^k, y - y^{k+1} \right\rangle. \end{aligned}$$

$$\begin{aligned} (ii) \quad f_a(y) &= f(\hat{x}^k) + \left\langle \hat{s}^k, y - \hat{x}^k \right\rangle - \hat{e}_k \\ &= f(\hat{x}^k) + \textstyle\sum_{i=1}^{\ell} \alpha_i \left(-\tilde{e}_i + \left\langle s^i, y - \hat{x}^k \right\rangle\right) \\ &\leq \hat{f}_k(y^{k+1}) + \max_{i=1,\ldots,\ell}\{-\tilde{e}_i + \left\langle \hat{s}^k, y - \hat{x}^k \right\rangle\} = \hat{f}_k(y). \end{aligned}$$

■

Lemma 9.4 *Take an arbitrary function $\psi : \mathbb{R}^n \to \overline{\mathbb{R}}$ such that*

$$\psi(y) \geq f_a(y) \quad \textit{for all } y \in \mathbb{R}^n \qquad \textit{and} \qquad \psi(y^{k+1}) = f_a(y^{k+1})$$

Then,

$$y^{k+1} = \arg\min_{y \in \mathbb{R}^n} \psi(y) + \frac{\mu_k}{2}\|y - \hat{x}^k\|^2.$$

Proof. Using Lemma 9.3 (i),

$$\begin{aligned} \psi(y) + \tfrac{\mu_k}{2}\|y - \hat{x}^k\|^2 &\geq f_a(y) + \tfrac{\mu_k}{2}\|y - \hat{x}^k\|^2 \\ &= \hat{f}_k(y^{k+1}) + \left\langle \hat{s}^k, y - y^{k+1} \right\rangle + \tfrac{\mu_k}{2}\|y - \hat{x}^k\|^2, \end{aligned}$$

where equality holds if $y = y^{k+1}$. The optimality conditions for minimizing the right hand side is

$$\mu_k(y^* - \hat{x}^k) = -\hat{s}^k.$$

Thus, $y^* = y^{k+1}$, implying that y^{k+1} minimizes the desired expression. ■

As it will become clear in the convergence proofs, the only properties that are required by the model are exactly $f(y^k) = \hat{f}_k(y^k)$ and $\hat{f}_k(y) \geq f_a(y)$.

10 Convergence Analysis

In this section, we will assume that the algorithm (BA) is used over a lower-semi continuous function f, finite-valued on $\mathbb{R}^n$. The general case proofs are similar but one needs to keep track of the Normal Cone induced by the set $\mathbf{dom}(f)$. Here, we will apply (BA) with the parameter $\bar{\delta} = 0$, that is, (BA) loops forever. Our analysis is divided in two excludent cases:

- (BA) performs an infinite number of Serious Steps, i.e., $|K_s| = +\infty$;
- (BA) performs a finite number of Serious Steps and then only Null Steps.

Lemma 10.1 *Consider the (BA) and denote by $\bar{f} = \min_x f(x) > -\infty$. Then,*

$$(0 \leq) \sum_{k \in K_s} \delta_k \leq \frac{f(\hat{x}^0) - \bar{f}}{m} < +\infty.$$

Proof. Take $k \in K_s$. Thus, since the SS test was satisfied,

$$f(\hat{x}^k) - f(y^{k+1}) = f(\hat{x}^k) - f(\hat{x}^{k+1}) > m\delta_k.$$

Let k' be the next index in K_s which is ordered increasingly. Again, by definition,

$$f(\hat{x}^{k'}) - f(\hat{x}^{k'+1}) = f(\hat{x}^{k+1}) - f(\hat{x}^{k'+1}) \geq m\delta_{k'}.$$

Summing over K_s,

$$\sum_{k \in K_s} m\delta_k \leq \sum_{k \in K_s} f(\hat{x}^k) - f(\hat{x}^{k+1}) \leq f(\hat{x}^0) - \bar{f}$$

■

Lemma 10.2 *Suppose that $f^* = \lim_{k \in K_s} f(\hat{x}^k) > -\infty$ and $|K_s| = \infty$.*

(i) If $\sum_{k \in K_s} \frac{1}{\mu_k} = \infty$, then zero is a cluster point of $\{\hat{s}^k\}_{k \in K_s}$, that is, $\liminf \|\hat{s}^k\| = 0$.

(ii) If $\mu_k \geq c > 0$ and $\emptyset \neq \arg\min_x f(x)$, then $\{\hat{x}^k\}_{k \in K_s}$ is bounded.

Proof. *(i)* Using Lemma 9.1 *(iii)*, $\delta_k \geq \delta_k - \hat{e}_k - \frac{1}{2\mu_k}\|\hat{s}^k\|^2 \geq 0$.Also, Lemma 10.1 states that

$$\sum_{k \in K_s} \frac{\|\hat{s}^k\|^2}{2\mu_k} < \sum_{k \in K_s} \delta_k \leq \frac{f(\hat{x}^0) - f^*}{m}.$$

Thus, $\hat{s}^k \to 0$ over $k \in K_s$.

(ii) Let $\bar{x} \in \arg\min_y f(y)$. By definition $f(\bar{x}) \leq f(y)$ for all $y \in \mathbb{R}^n$. Now, for $k \in K_s$,

$$\begin{aligned}
\|\bar{x} - \hat{x}^{k+1}\|^2 &= \|\bar{x} - \hat{x}^k\|^2 + 2\left\langle \bar{x} - \hat{x}^k, \hat{x}^k - \hat{x}^{k+1} \right\rangle + \|\hat{x}^k - \hat{x}^{k+1}\|^2 \\
&= \|\bar{x} - \hat{x}^k\|^2 + \tfrac{2}{\mu_k}\left\langle \bar{x} - \hat{x}^k, \hat{s}^k \right\rangle + \tfrac{1}{\mu_k^2}\|\hat{s}^k\|^2 \\
&= \|\bar{x} - \hat{x}^k\|^2 + \tfrac{2}{\mu_k}\left(\left\langle \bar{x} - \hat{x}^k, \hat{s}^k \right\rangle + \tfrac{1}{2\mu_k}\|\hat{s}^k\|^2 \right) \\
&= \|\bar{x} - \hat{x}^k\|^2 + \tfrac{2}{\mu_k}\left(\hat{f}_k(\bar{x}) - f(\hat{x}^k) + \hat{e}_k + \tfrac{1}{2\mu_k}\|\hat{s}^k\|^2 \right) \\
&\leq \|\bar{x} - \hat{x}^k\|^2 + \tfrac{2}{\mu_k}\left(f(\bar{x}) - f(\hat{x}^k) + \delta_k \right) \\
&\leq \|\bar{x} - \hat{x}^k\|^2 + \tfrac{2}{\mu_k}\delta_k
\end{aligned}$$

Thus,

$$\|\bar{x} - \hat{x}^{k+1}\|^2 \leq \|\bar{x} - \hat{x}^0\|^2 + 2\sum_{i=1}^{k+1} \frac{\delta_i}{\mu_i} \leq \|\bar{x} - \hat{x}^0\|^2 + \frac{2}{c} \sum_{k \in K_s} \delta_k$$

which is bounded by Lemma 10.1. Thus, $\{\hat{x}^k\}_{k \in K_s}$ is bounded. ■

Theorem 10.1 *Suppose that $f^* = \lim_{k \in K_s} f(\hat{x}^k) > -\infty$, $\bar{\delta} = 0$, and $|K_s| = \infty$. If the sequence μ_k is bounded from above and away from zero, then $\{\hat{x}^k\}_{k \in K_s}$ has at least one cluster point which is optimal.*

Proof. By Lemma 10.1, $0 \leq \hat{e}_k \leq \delta_k \to 0$ for $k \in K_s$. Lemma 9.2 states that

$$\hat{s}^k \in \partial_{\hat{e}_k} f(\hat{x}^k) \quad \text{for every } k \in K_s.$$

Lemma 10.2(i) implies that there exists a subsequence $\{\hat{s}^{n_k}\}_{k \geq 1}$ converging to zero. Since $\{\hat{x}^k\}_{k \in K_s}$ is bounded by Lemma 10.2 (ii), $\{\hat{x}^{n_k}\}_{k \geq 1}$ is also bounded. Taking a subsequence if necessary, $\hat{x}^{n_k} \to \bar{x}$. So we have

$$(\hat{x}^{n_k}, \hat{s}^{n_k}, \hat{e}_{n_k}) \to (\bar{x}, 0, 0).$$

Corollary 6.1 ensures that the correspondence $(x, \varepsilon) \mapsto \partial_\varepsilon f(x)$ is continuous. So, $0 \in \partial f(\bar{x})$. ■

Now we move on to the second case, where (BA) performs one last Serious Step at iteration k_0, and then a sequence of Null Steps for all $k \geq k_0 + 1$.

Lemma 10.3 *Let $\hat{x}^{k_0}$ be the last Serious Step, and $\{y^{k+1}\}_{k \geq k_0}$ the sequence of Null Steps. Then, for all $k > k_0$ and $y \in \mathbb{R}^n$,*

$$f(\hat{x}^{k_0}) - \delta_k + \frac{\mu_k}{2}\|y - y^{k+1}\|^2 = \hat{f}_k(y^{k+1}) + \langle \hat{s}^k, y - y^{k+1} \rangle + \frac{\mu_k}{2}\|y - \hat{x}^{k_0}\|^2.$$

Proof.

$$\begin{aligned} \|y - \hat{x}^{k_0}\|^2 &= \|y - y^{k+1} + y^{k+1} - \hat{x}^{k_0}\|^2 \\ &= \|y - y^{k+1}\|^2 + 2\langle y - y^{k+1}, y^{k+1} - \hat{x}^{k_0} \rangle + \|y^{k+1} - \hat{x}^{k_0}\|^2 \\ &= \|y - y^{k+1}\|^2 - \tfrac{2}{\mu_k}\langle y - y^{k+1}, \hat{s}^k \rangle + \|y^{k+1} - \hat{x}^{k_0}\|^2. \end{aligned}$$

Using the definition of $\delta_k = f(\hat{x}^{k_0}) - \hat{f}_k(y^{k+1}) - \frac{\mu_k}{2}\|y^{k+1} - \hat{x}^{k_0}\|^2$,

$$\begin{aligned} f(\hat{x}^{k_0}) - \delta_k + \tfrac{\mu_k}{2}\|y - y^{k+1}\|^2 &= \hat{f}_k(y^{k+1}) + \tfrac{\mu_k}{2}\left(\|y^{k+1} - \hat{x}^{k_0}\|^2 + \|y - y^{k+1}\|^2\right) \\ &= \hat{f}_k(y^{k+1}) + \langle y - y^{k+1}, \hat{s}^k \rangle + \tfrac{\mu_k}{2}\|y - \hat{x}^{k_0}\|^2. \end{aligned}$$

■

Theorem 10.2 *Let $\hat{x}^{k_0}$ be the iterate generated by the last Serious Step performed by (BA), and denote as $\{y^{k+1}\}_{k \geq k_0}$ the sequence of iterates corresponding to the Null Steps.*
If $\{\mu_k\}_{k > k_0}$ is nondecreasing, it holds that $\delta_k \to 0$.

Proof. Let $y = y^{k+2}$ on Lemma 10.3,

$$\begin{aligned} f(\hat{x}^{k_0}) - \delta_k + \tfrac{\mu_k}{2}\|y^{k+2} - y^{k+1}\|^2 &= \hat{f}_k(y^{k+1}) + \left\langle y^{k+2} - y^{k+1}, \hat{s}^k \right\rangle + \tfrac{\mu_k}{2}\|y^{k+2} - \hat{x}^{k_0}\|^2 \\ &= f_a(y^{k+2}) + \tfrac{\mu_k}{2}\|y^{k+2} - \hat{x}^{k_0}\|^2 \\ &= \hat{f}_{k+1}(y^{k+2}) + \tfrac{\mu_k}{2}\|y^{k+2} - \hat{x}^{k_0}\|^2 \\ &\leq \hat{f}_{k+1}(y^{k+2}) + \tfrac{\mu_{k+1}}{2}\|y^{k+2} - \hat{x}^{k_0}\|^2 \\ &= f(\hat{x}^{k_0}) - \delta_{k+1}, \end{aligned}$$

where the last inequality follows since $\mu_k \leq \mu_{k+1}$. So, rearranging the previous relation,

$$\delta_k \geq \delta_{k+1} + \frac{\mu_k}{2}\|y^{k+2} - y^{k+1}\|^2. \tag{12}$$

Next, we will show that the sequence $\{y^k\}$ is bounded. Using one more time Lemma 10.3 with $y = \hat{x}^{k_0}$,

$$f(\hat{x}^{k_0}) - \delta_k + \frac{\mu_k}{2}\|\hat{x}^{k_0} - y^{k+1}\|^2 = \hat{f}_k(y^{k+1}) + \left\langle \hat{s}^k, x^{k_0} - y^{k+1} \right\rangle = \hat{f}_k(\hat{x}^{k_0}) \leq f(\hat{x}^{k_0}).$$

Thus, $\|\hat{x}^{k_0} - y^{k+1}\|^2 \leq \dfrac{2\delta_k}{\mu_k} \leq \dfrac{2\delta_{k_0}}{\mu_{k_0}}$ since δ_k is decreasing and μ_k is nondecreasing. So, $\{y^k\}$ is bounded.

Finally, to show that $\delta_k \searrow 0$, note that the Serious Step test fails for every $k > k_0$. Let C be a Lipschitz constant for f and $\hat{f}_k$ in $B(\hat{x}^{k_0}, \frac{\delta_{k_0}}{\mu_{k_0}}))$. Combining

$$-m\delta_k \leq f(y^{k+1}) - f(x^{k_0}) \qquad \text{and} \qquad \delta_k \leq f(\hat{x}^{k_0}) - \hat{f}_k(y^{k+1})$$

we obtain that

$$\begin{aligned} (1-m)\delta_k &\leq f(y^{k+1}) - \hat{f}_k(y^{k+1}) \\ &= f(y^{k+1}) - f(y^k) + \hat{f}_k(y^k) - \hat{f}_k(y^{k+1}) \\ &\leq 2C\|y^{k+1} - y^k\| \end{aligned}$$

where we used that $f(y^k) = \hat{f}_k(y^k)$. Combining this with relation (12),

$$\delta_k - \delta_{k+1} \geq \frac{\mu_k}{2}\|y^{k+2} - y^{k+1}\|^2 \geq \frac{(1-m)^2}{8C^2}\mu_k\delta_k^2 \geq \frac{(1-m)^2}{8C^2}\mu_{k_0}\delta_{k+1}^2.$$

Thus, summing up in $k \geq k_0$,

$$\frac{(1-m)^2}{8C^2}\mu_{k_0}\sum_{k\geq k_0}\delta_k^2 \leq \sum_{k\geq k_0}(\delta_k - \delta_{k+1}) \leq \delta_{k_0},$$

which implies that $\delta_k \to 0$. ■

Theorem 10.3 *Suppose that (BA) performs one last Serious Step at iteration k_0 corresponding to an iterate $\hat{x}^{k_0}$, and we set $\bar{\delta} = 0$. If $\{\mu_k\}_{k\geq k_0}$ is nondecreasing, $\hat{x}^{k_0}$ is an optimal solution.*

Proof. The assumptions allow us to invoke Theorem 10.2, so $\delta_k \to 0$ imply that $\hat{e}_k \to 0$ and $\|\hat{s}^k\| \to 0$ by Lemma 9.1 (*iii*). Again, Lemma 9.2 states that

$$\hat{s}^k \in \partial_{\hat{e}_k} f(\hat{x}^{k_0}) \quad \text{ for all } k > k_0.$$

Corollary 6.1 ensures that the correspondence $(x, \varepsilon) \mapsto \partial_\varepsilon f(x)$ is continuous. So, $0 \in \partial f(\hat{x}^{k_0})$. ■

11 Notes on the QP Implementation

One important question in practice concerns running times. Running times can be approximated by the number of calls to the oracle. This might be accurate in many applications but not all of them. Towards a more precise approximation, we need to take into account the effort necessary to define the next iterate to be evaluated by the oracle.

Method	Cost Per Iteration	# of Iterations
Subgradient	Updade a Vector	Very High
Cutting Planes	LP (growing)	Medium/High
Bundle Method	QP (bounded)	Small/Medium

Table 11 clarifies the trade off between all the three methods mentioned so far. In practice, we will encounter applications where evaluating the oracle will be a very expensive operation and also will find applications where the oracle returns a solution instantenauslly.

For the Bundle Methods to be competitive, the QP must be solved relatively fast. This is the subject of this section, i.e., to derive a specialized QP solver which takes advantage of the particular structure of our problem. We will be following [6] and refer to it for more details.

Let G be the matrix $n \times \ell$ formed by taking the subgradients s^i as columns, and $\tilde{e}$ a vector formed by the linearization errors. Consider the primal version of our QP (6),

$$(QPP) \left\{ \begin{array}{ll} \min_{r,y} & r + \frac{\mu_k}{2}\|y - \hat{x}^k\|^2 \\ & re \geq -\tilde{e} + G^T(y - \hat{x}^k) \end{array} \right.$$

Now, rewriting our dual version,

$$(QPD) \left\{ \begin{array}{ll} f(\hat{x}^k) - \frac{1}{\mu_k}\min_\alpha & \frac{1}{2}\left\langle \alpha, G^T G\alpha \right\rangle + \mu_k \left\langle \alpha, \tilde{e} \right\rangle \\ & \langle e, \alpha \rangle = 1 \\ & \alpha \in \mathbb{R}^\ell_+ \end{array} \right.$$

In practice, it is common for $n > 10000$ while $\ell < 100$. So, it is not surprising that all specialized QP codes within our framework focus on the dual version. Also, active set strategies seem to be more suitable in our framework since the QP called in consecutive iterations tend to be very similar, differing only by a few linear pieces.

Denote $\mathcal{B} = \{1, 2, \ldots, \ell\}$ as the complete set of indices of linear pieces in our model. For any $B \subset \mathcal{B}$, B is a base which induces a submatrix G_B of G (and a subvector $\tilde{e}_B$ of $\tilde{e}$) containing the columns (components) whose indices are in B. It will be convenient to define for any $B \subset \mathcal{B}$,

$$Q_B = G_B^T G_B \succcurlyeq 0 \qquad \text{and} \qquad \tilde{e}_B \geq 0.$$

We start by defining the QP restricted to the base B as

$$(QPD_B) \qquad \min \left\{ \frac{1}{2}\left\langle \alpha, Q_B\alpha \right\rangle + \left\langle \tilde{e}_B, \alpha \right\rangle : \alpha \geq 0, \langle e, \alpha \rangle = 1 \right\},$$

where we included the factor μ_k in $\tilde{e}_B$ for convenience throughout this section. The basic step of our QP solver will be to solve the following relaxation of (QPD_B),

$$(RQPD_B) \qquad \min\left\{\frac{1}{2}\langle\alpha, Q_B\alpha\rangle + \langle\tilde{e}_B, \alpha\rangle : \langle e, \alpha\rangle = 1\right\}.$$

The optimality conditions for $(RQPD_B)$ can be summarized with the following system of equations

$$(KT) \qquad \begin{bmatrix} Q_B & -e \\ e^T & 0 \end{bmatrix}\begin{bmatrix} \alpha_B \\ \rho \end{bmatrix} = \begin{bmatrix} -\tilde{e} \\ 1 \end{bmatrix}$$

Solving this linear system will be the core of our algorithm and we defer some additional comments until the next subsection. For now, we assume that we can either find the unique solution for the system, or a direction w_B such that $Q_B w_B = 0$, $\langle e, w_B\rangle = 0$.

Now we state the algorithm.

01.$B = \{1\}, \alpha = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
02.while $(\exists h \in \mathcal{B} \setminus B$ such that $r < \langle s^h, d\rangle - \tilde{e}_h)$
03. $B := B \cup \{h\}$
04. do
05. if $\left((KT)\text{ has a unique solution } \begin{bmatrix} \bar{\alpha}_B \\ \rho \end{bmatrix}\right)$ then
06. if $(\bar{\alpha}_B \geq 0)$
07. $\alpha = \begin{bmatrix} \bar{\alpha}_B \\ 0 \end{bmatrix}$
08. $w_B = 0$
09. else
10. $w_B = \bar{\alpha}_B - \alpha$
11. else (Let w_B be a descent direction s.t. $Q_B w_B = 0$, $\langle e, w_B\rangle = 0$)
12. if $(w_B \neq 0)$
13. $\eta = \min\left\{-\frac{\alpha_h}{w_h} : w_h < 0, h \in B\right\}$
14. $\alpha = \alpha + \eta \begin{bmatrix} w_B \\ 0 \end{bmatrix}$
15. for all $(h \in B : \alpha_h = 0)$
16. $B := B \setminus \{h\}$
17. while $(w_B \neq 0)$
18.end while

where we relate primal and dual variables by the following relations

$$d = -G\alpha \qquad r = -\|d\|^2 - \langle\tilde{e}, \alpha\rangle.$$

Now we will prove that our algorithm converges. First we need the following proposition.

Proposition 11.1 *In the inner loop, we always have $\eta > 0$.*

Proof.

Consider the inner loop defined between lines $04-17$. If the current inner iteration is not the first coming from the outer loop, by construction $\alpha_B > 0$ since we removed all zero components of the base B. By definition of η, it is strictly positive.

So we can assume that we are in the first iteration of the inner loop, with $B := B' \cup \{h\}$, $\alpha_B = [\alpha_{B'}^T \quad 0]^T$, where $\alpha_{B'} > 0$ and is optimal for $QPD_{B'}$. Also, $r < \langle s^h, d\rangle - \tilde{e}_h$, where $d = -G_{B'}\alpha_{B'}$ and $r = -\|d\|^2 - \langle \tilde{e}_{B'}, \alpha_{B'}\rangle$.

Optimality conditions of $\alpha_{B'}$ for $(QPD_{B'})$ imply that

$$-\rho e = G_{B'}^T d - \tilde{e}_{B'}.$$

Multiplying by $\alpha_{B'}$ from the left, we obtain

$$-\rho \langle \alpha_{B'}, e\rangle = \langle \alpha_{B'}, G_{B'}^T d\rangle - \langle \alpha_{B'}, \tilde{e}_{B'}\rangle$$
$$-\rho = -\|d\|^2 - \langle \tilde{e}_{B'}, \alpha_{B'}\rangle = r.$$

Consider any feasible direction from α_B, i.e., $w^T = \begin{bmatrix} w'^T & w_h \end{bmatrix}$.Then, the improvement by moving η on this direction from α_B is given by

$$\frac{1}{2}\langle \alpha_B + \eta w, Q_B(\alpha_B + \eta w)\rangle + \langle \tilde{e}_B, \alpha_B + \eta w\rangle - \frac{1}{2}\langle \alpha_B, Q_B\alpha_B\rangle + \langle \tilde{e}_B, \alpha_B\rangle =$$

$$= \eta \begin{bmatrix} w' & w_h \end{bmatrix} \begin{bmatrix} Q_{B'} & G_{B'}s^h \\ (G_{B'}s^h)^T & \langle s^h, s^h\rangle \end{bmatrix} \begin{bmatrix} \alpha_{B'} \\ 0 \end{bmatrix} + \eta \begin{bmatrix} w' & w_h \end{bmatrix} \begin{bmatrix} \tilde{e}_{B'} \\ \tilde{e}_h \end{bmatrix} + \frac{1}{2}\eta^2 \langle w, Q_B w\rangle .$$

The optimality conditions for minimizing in w are

$$\eta^2 Q_B w + \eta \begin{bmatrix} \tilde{e}_{B'} \\ \tilde{e}_h \end{bmatrix} + \eta \begin{bmatrix} Q_{B'}\alpha_{B'} \\ \langle s^h, G_{B'}\alpha_{B'}\rangle \end{bmatrix} + \lambda e = 0$$

Multiplying by w from the right,

$$\eta^2 \langle w, Q_B w\rangle + \eta \left\langle w, \begin{bmatrix} Q_{B'}\alpha_{B'} + \tilde{e}_{B'} \\ \langle s^h, G_{B'}\alpha_{B'}\rangle + \tilde{e}_h \end{bmatrix} \right\rangle + \lambda \langle w, e\rangle = 0$$

First, assume that $\langle w, Q_B w\rangle > 0$ and using that $re = -Q_{B'}\alpha_{B'} - \tilde{e}_{B'}$,

$$0 > \eta \left\langle w, \begin{bmatrix} -re \\ \langle s^h, G_{B'}\alpha_{B'}\rangle + \tilde{e}_h \end{bmatrix} \right\rangle = \eta(-r\langle w', e\rangle + w_h(\langle s^h, G_{B'}\alpha_{B'}\rangle + \tilde{e}_h)).$$

Since w is a feasible direction, $0 = \langle w, e\rangle = \langle w', e\rangle + w_h$, which implies $\langle e, w'\rangle = -w_h$. Thus, using $d = -G_{B'}\alpha_{B'}$,

$$0 > \eta(rw_h + w_h(-\langle s^h, d\rangle + \tilde{e}_h)) = \eta w_h(r - \langle s^h, d\rangle + \tilde{e}_h),$$

which implies that $w_h > 0$.

Assuming that the optimal direction is such that $\langle w, Q_B w\rangle = 0$, we also have that $Q_B w = 0$ and $G_B w = 0$ since $Q_B \succcurlyeq 0$. Since w is a descent direction (α_B cannot be optimal since we introduced another violated linear piece),

$$
\begin{aligned}
0 &> \eta \left\langle w, \begin{bmatrix} \tilde{e}_{B'} \\ \tilde{e}_h \end{bmatrix} \right\rangle \\
&= \eta \left(\langle w', \tilde{e}_{B'} \rangle + w_h \tilde{e}_h \right) \\
&= \eta \left(\left\langle w', G_{B'}^T d - re \right\rangle + w_h \tilde{e}_h \right) \\
&= \eta \left(\langle G_{B'} w', d \rangle - r \langle w', e \rangle + w_h \tilde{e}_h \right) \\
&= \eta (-w_h \langle s^h, d \rangle + r w_h + w_h \tilde{e}_h) \\
&= \eta w_h (r - \langle s^h, d \rangle + \tilde{e}_h),
\end{aligned}
$$

where we used that $0 = G_B w = G_{B'} w' + w_h s^h$. Thus, $w_h > 0$ since $(r - \langle s^h, d \rangle + \tilde{e}_h) < 0$. This implies that η can be chosen to be positive. ■

Theorem 11.1 *The algorithm finishes in a finite number of steps.*

Proof. First observe that on any inner iteration, at least one item is deleted from the base by definition of η. So, the inner loop cannot loop indefinetly.

Moreover, it is impossible for the same base B to appear twice at the beginning of an outer iteration. This is true since α_B is optimal for (QPD_B) and at least a strict improvement has been performed ($\eta > 0$ due to the previous Lemma).

Although exponential, the number of possible bases is finite. So, the algorithm terminates. ■

11.1 Solving the (KT) System

As mentioned before, the details of this procedure can be found in [6]. Here, we will just discuss general ideas. The main point is to keep a lower trapezoidal factorization of the matrix Q_B of the current (QPD_B). That is,

$$
\bar{L}_B = \begin{bmatrix} L_{B'} & 0 \\ V^T & 0 \end{bmatrix} \qquad \text{such that} \qquad \bar{L}_B \bar{L}_B^T = Q_B
$$

where $L_{B'}$ is a lower triangular matrix with all positive diagonal entries. It is possible to show that the submatrix V^T can only have up to two rows which allow us to derive specific formulae for computing the solution of the system (KT) or an infinite direction.

12 Applications

12.1 Basic Duality Theory

This section complements the material in the Introduction and it is included here for sake of completeness. Consider our problem of interest, the primal, known to be hard to solve as mentioned in the Introduction.

$$
(P) \begin{cases} \max\limits_{y} & g(y) \\ & h_i(y) = 0 \qquad i = 1, \ldots, m \\ & y \in \mathcal{D} \end{cases},
$$

We dualize the equality constraints to construct a dual function f given by

$$f(x) = \begin{array}{ll} \max\limits_{y} & g(y) + \langle x, h(y)\rangle \\ & y \in \mathcal{D} \end{array} .$$

As claimed before, given $x, w \in \mathbb{R}^n$, $\alpha \in [0,1]$,

$$\begin{array}{rcll} f(\alpha x + (1-\alpha)w) & = & \max\limits_{y\in\mathcal{D}} & g(y) + \langle \alpha x + (1-\alpha)w, h(y)\rangle \\ & = & \max\limits_{y^1,y^2} & \alpha\left[g(y^1) + \left\langle x, h(y^2)\right\rangle\right] + (1-\alpha)\left[g(y^2) + \left\langle w, h(y^2)\right\rangle\right] \\ & & & y^1 = y^2 \\ & & & y^1 \in \mathcal{D}, y^2 \in \mathcal{D} \\ & \leq & \max\limits_{y^1,y^2} & \alpha\left[g(y^1) + \left\langle x, h(y^1)\right\rangle\right] + (1-\alpha)\left[g(y^2) + \left\langle w, h(y^2)\right\rangle\right] \\ & & & y^1 \in \mathcal{D}, y^2 \in \mathcal{D} \\ & = & \alpha \max\limits_{y\in\mathcal{D}} [g(y) + \langle x, h(y)\rangle] + (1-\alpha) \max\limits_{y\in\mathcal{D}} [g(y) + \langle w, h(y)\rangle] \\ & = & \alpha f(x) + (1-\alpha) f(w). \end{array}$$

Also, using that the feasible set of (P) is always contained in $\mathcal{D}$,

$$\begin{array}{rcll} f(x) & = & \max\limits_{y} & g(y) + \langle x, h(y)\rangle \\ & & & y \in \mathcal{D} \\ & \geq & \max\limits_{y} & g(y), \\ & & & h_i(y) = 0 \quad i = 1, \ldots, m \\ & & & y \in \mathcal{D} \end{array} .$$

we have that $f(x) \geq (P)$ for any x. Thus,

$$(D) \left\{ \begin{array}{ll} \min\limits_{x} & f(x) \\ & x \in \mathbb{R}^n \end{array} \right.$$

is a upper bound for (P).

To obtain (D) we need to minimize the convex function f. Note that f is given implicitly by a maximization problem. So, it is hard to obtain additional structure for f in general. In fact, one cannot assume that f is differentiable. In order to apply any method based on oracles, we need to be able to compute subgradients. It turns out that, within this application, subgradients are no harder to compute than evaluating the function.

Let $y(x) \in \mathcal{D}$ achieve the maximum in the definition of $f(x)$. So,

$$\begin{array}{rcl} f(x) & = & f(x - w + w) = g(y(x)) + \langle h(y(x)), x - w + w\rangle \\ & = & g(y(x)) + \langle h(y(x)), w\rangle + \langle h(y(x)), x - w\rangle \\ & \leq & f(w) + \langle h(y(x)), x - w\rangle . \end{array}$$

Thus, $f(w) \geq f(x) + \langle h(y(x)), w - x\rangle$ and $h(y(x)) \in \partial f(x)$.

12.2 Held and Karp Bound

The Held and Karp Bound is one of the most celebrated applications of duality theory. It applies duality theory and combinatorial techniques to the most studied problem in Combinatorial Optimization, the Traveling Salesman Problem (TSP).

Let $G = G(V_n, E_n)$ be the undirected complete graph induced by a set of nodes V_n and let the set of edges be E_n. For each edge of the graph, a cost c_e is given, and we associate a binary variable p_e which equals one if we use the edge e in the solution and equals zero otherwise. For every set $S \subset V_n$, we denote by $\delta(S) = \{(i,j) \in E_n : i \in S, j \in V_n \setminus S\}$ the edges with exactly one endpoint in S, and $\gamma(S) = \{(i,j) \in E_n : i \in S,\ j \in S\}$. Finally, for every set of edges $A \subset E_n$, $p(A) = \sum_{e \in A} p_e$. The TSP formulation of Dantzig, Fulkerson and Johnson is given by

$$\begin{cases} \min\limits_{p} & \sum\limits_{e \in E_n} c_e p_e \\ & p(\delta(\{j\})) = 2\,, \text{ for every } j \in V_n \\ & p(\delta(S)) \geq 2\,,\ S \subset V_n, S \neq \emptyset \\ & p_e \in \{0,1\}\,,\ e \in E_n. \end{cases} \tag{13}$$

In order to derive the Held and Karp bound, we first conveniently define the set of 1-trees of G. Assume that one (any) given vertex of G is set aside, say vertex v_1. In association, consider one (any) spanning tree of the subgraph of G induced by the vertices $V_n \setminus \{v_1\}$. An 1-tree of G is obtained by adding any two edges incident on v_1 to that spanning tree. Let $\mathcal{X}$ be the convex hull of the incidence vectors of all the 1-Trees of G just introduced. Then, (13) is equivalent to

$$\begin{cases} \min\limits_{p} & \sum\limits_{e \in E_n} c_e p_e \\ & p(\delta(\{j\})) = 2\,, \text{ for every } j \in V_n \quad (*) \\ & p \in \mathcal{X} \end{cases} \tag{14}$$

Now, the Held and Karp bound is obtained by maximizing a dual function which arises from dualizing constraints (14.*),

$$f(x) = \min_{p} \sum_{i=1}^{n-1} \sum_{j>i}^{n} (c_{ij} + x_i + x_j) p_{ij} - 2 \sum_{i=1}^{n} x_i$$
$$p \in \mathcal{X}.$$

We point out that computing a minimum spanning trees, a spanning tree with minimum cost, is solvable in polynomial time. So, we can efficiently optimize over $\mathcal{X}$ and evaluate $f(x)$ for any given x.

In order to obtain tighter bounds, we introduce a family of facet inequalities associated with the polytope of (13) called the r-Regular t-Paths Inequalities. More precisely, we choose sets of verticess $H_1, H_2, \ldots, H_{r-1}$ and $T_1, T_2, \ldots, T_t$, called "handles" and "teeth" respectively, which satisfy the following relations:

$H_1 \subset H_2 \subset \cdots \subset H_{r-1}$

$H_1 \cap T_j \neq \phi$ for $j = 1, \ldots, t$

$T_j \setminus H_{r-1} \neq \phi$ for $j = 1, \ldots, t$

$(H_{i+1} \setminus H_i) \subset \cup_{j=1}^{t} T_j$ for $1 \leq i \leq r-2$.

The corresponding p-Regular t-Path inequality is given by

$$\sum_{i=1}^{r-1} y(\gamma(H_i)) + \sum_{j=1}^{t} p(\gamma(T_j)) \leq \sum_{i=1}^{r-1} |H_i| + \sum_{j=1}^{t} |T_j| - \frac{t(r-1)+r-1}{2}.$$

As we introduce these inequalities as additional constraints in (14), we may improve on the Held and Karp bound. In order to keep the approach computable, we also need to dualize these inequalities. However, the number of such inequalities is exponential in the size of the problem and it would be impossible to consider all of them at once. A dynamic scheme is needed here. That is, we will consider only a small subset of these inequalities on each iteration. We will introduce and remove inequalities from this subset based on primal and dual information obtained by the algorithm. We refer to [3] for a complete description and proofs of this Dynamic Bundle Method.

Based on the 1-tree solution obtained in the current iteration, we will try to identify violated inequalities to be added in our subset of inequalities. This is done through the use of a Separation Oracle. For this particular family, one can rely only on heuristics, since no efficient exact separation procedure is known. Fortunately, the Separation Oracle is called to separate inequalities from 1-tree structures. In this case, thanks to integrality and the existence of exactly one cycle, the search for violated inequalities is much easier than for general sets (which is the case for linear relaxations). Essentially, we search first for any node with three or more edges on the 1-tree, and then try to expand an odd number of paths from such a node.

12.3 LOP

The Linear Ordering Problem (LOP) is another example of a NP-Hard combinatorial problem that can be successifully addressed by Lagrangian relaxation. LOP consists in placing elements of a finite set N in sequential order. Also, if object i is placed before object j, we incur in a cost c_{ij}. The objective is to find the order with minimum cost. Applications related to the LOP are triangularization of input-output matrices in Economics, dating artifacts in Archeology, Internet search engines, among others (see [7]).

The LOP Linear Integer Programming formulation in [7] uses a set of binary variables $\{y_{ij} : (i,j) \in N \times N\}$. If object i is placed before object j, $y_{ij} = 1$ and $y_{ji} = 0$.

$$\begin{cases} \min\limits_{y} & \sum\limits_{(i,j):i\neq j} c_{ij} y_{ij} & & \\ & y_{ij} + y_{ji} = 1, & \text{for every pair } (i,j) & \\ & y_{ij} \in \{0,1\}, & \text{for every pair } (i,j) & \\ & y_{ij} + y_{jk} + y_{ki} \leq 2, & \text{for every triple } (i,j,k) & (*) \end{cases} \tag{15}$$

The 3-cycle inequalities in constraint (*) above have a huge cardinality, but one expects that only a small subset of them to be active on the optimal solution. These inequalities are the candidates for a dynamic bundle method proposed in [3], that is, we will keep only a subset of these inequalities on each iteration which is updated based on the solutions of the relaxed subproblems. We associate a multiplier x_{ijk} to each one of them. After relaxing these inequalities, we obtain a dual function which is concave in x.

Now, observe that the formulation (15) without the 3-cycle inequalities decomposes into smaller problems. In fact, given any x, we can evaluate the dual function by solving $(N^2-N)/2$ subproblems with only two variables each. More precisely,

$$\begin{cases} f(x) = \displaystyle\sum_{(i,j):i<j} & \min\limits_{\{y_{ij},y_{ji}\}} \tilde{c}_{ij}y_{ij} + \tilde{c}_{ji}y_{ji} \\ & y_{ij} + y_{ji} = 1 \\ & y_{ij}, y_{ji} \in \{0,1\} \end{cases}$$

where $\tilde{c}_{ij} = c_{ij} - \sum_{k\in N\setminus\{i,j\}} x_{ijk}$ depends on the multiplier x.

During the computation of $f(x)$, we obtain a relaxed solution $y(x)$. In order to select additional inequalities, we seach of 3-cycles inequalities violated by $y(x)$. The Separation Procedure is therefore easy: it just consists of checking any triple of indices, a task that can be done in constant time for each triple.

Computational experiments were performed on the standard LOLIB instances to compare the Bundle Methods and Subgradient Method. They confirmed the quality of the directions generated by the Bundle Methods. The (SM) used more than 1500 oracle calls on average while the (BM) needed less than 350 oracle calls on average.

12.4 Symmetry of a Polytope

Given a closed convex set S and a point $x \in S$, define the symmetry of S about x as follows:

$$\text{sym}(x,S) := \max\{\alpha \geq 0 : x + \alpha(x-y) \in S \text{ for every } y \in S\}\ , \qquad (16)$$

which intuitively states that $\text{sym}(x,S)$ is the largest scalar α such that every point $y \in S$ can be reflected through x by the factor α and still lie in S. The symmetry value of S then is:

$$\text{sym}(S) := \max_{x\in S}\ \text{sym}(x,S)\ , \qquad (17)$$

and x^* is a *symmetry point* of S if x^* achieves the above supremum. S is *symmetric* if $\text{sym}(S) = 1$. We refer to [4] for a more complete description of the properties of the symmetry function.

Our interest lies in computing an ε-approximate symmetry point of S, which is a point $x \in S$ that satisfies:

$$\text{sym}(x,S) \geq (1-\varepsilon)\text{sym}(S)\ .$$

In [4], it was established that the symmetry function is quasiconcave. Unfortunately, an oracle depends not only on the set, but also on its representation. It seems a hard problem for general convex sets.

Here, we restrict ourselves to the case that S is polyhedral. The convex set S is assumed to be represented by a finite number of linear inequalities, that is, $S := \{x \in \mathbb{R}^n : Ax \leq b\}$. We will show that there is still enough structure to obtain a constructive characterization which will allow us to solve the problem through the use of Bundle Methods (see [4] for a polynomial time algorithm based on Interior Point Method for this problem).

Let $\bar{x} \in S$ be given, and let $\alpha \geq \text{sym}(\bar{x},S)$. Then from the definition of $\text{sym}(\cdot,S)$ in (16) we have:

$$A(\bar{x} + v) \leq b \quad \Rightarrow \quad A(\bar{x} - \alpha v) \leq b \ ,$$

which we restate as:

$$Av \leq b - A\bar{x} \quad \Rightarrow \quad -\alpha A_{i\cdot} v \leq b_i - A_{i\cdot}\bar{x} \ , i = 1, \ldots, m \ . \tag{18}$$

Now, we can apply a theorem of the alternative to each of the $i = 1, \ldots, m$ above implications in (18). Then (18) is true if and only if there exists an $m \times m$ matrix Λ of multipliers that satisfies:

$$\begin{array}{c} \Lambda A = -\alpha A \\ \Lambda(b - A\bar{x}) \leq b - A\bar{x} \\ \Lambda \geq 0 \ . \end{array} \tag{19}$$

Here "$\Lambda \geq 0$" is componentwise[12] for all m^2 components of Λ.

This characterization implies that $\text{sym}(\bar{x}, S) \geq \alpha$ if and only if the system (19) has a feasible solution, i.e., (19) is a complete characterization of the α-level set of the symmetry function of S. So, we can state the $\text{sym}(S)$ as the optimal objective value of the following optimization problem

$$\begin{array}{ll} \max\limits_{x,\Lambda,\alpha} & \alpha \\ \text{s.t.} & \Lambda A = -\alpha A \\ & \Lambda(b - Ax) \leq b - Ax \\ & \Lambda \geq 0 \ , \end{array} \tag{20}$$

and any solution $(x^*, \Lambda^*, \alpha^*)$ of (20) satisfies $\text{sym}(S) = \alpha^*$ and $\text{sym}(x^*, S) = \text{sym}(S)$. Unfortunately, this formulation is not a linear programming problem since x and Λ are multiplying each other.

In order to apply the Bundle Methods, we will build a two-level scheme. For each x^0 fixed, computing its symmetry is reduced to solve a linear programming problem. Thus, the oracle for our problem becomes

$$\begin{array}{lll} \text{sym}(x^0, S) = \max\limits_{\Lambda,\alpha} & \alpha & \\ s.t. & \Lambda A = -\alpha A & (i) \\ & \Lambda b + \alpha A x^0 \leq b - A x^0 & (ii) \\ & \Lambda \geq 0 & (iii). \end{array} \tag{21}$$

Now, from any solution of (21), we can generate a suporting hyperplane for the level set $\text{sym}(x^0, S)$ at x^0. Let μ^* be an optimal Lagrange multiplier associated with the constraints (21).(ii), then

$$-(1 + \alpha^*)\mu^{*T} A \in \partial \text{sym}(x^0, S). \tag{22}$$

This hyperplane will play the role of the subgradient in our previous analysis. Also, we will use the following fact in our analysis.

Remark 12.1 *If x^0 is not optimal, for every symmetry point x^* we have that*

$$\left\langle x^0 - x^*, (1 + \alpha^*)\mu^{*T} A \right\rangle < 0$$

since x^* is in the interior of the upper level set of sym(x^0, S).

[12]It is not a semidefinite positive constraint.

We cannot use the standard cutting plane model in this problem. Since $f(\cdot) = -\text{sym}(\cdot, S)$ is not convex, its epigraph may not be convex and the cutting plane model would cut the epigraph of f. To deal with the quasi-convexity of f, one needs to adapt the cutting plane model. In our case, we will ignore the value of the function given by the oracle. The model will be

$$\hat{f}(y) = \max_{i=1,\ldots,\ell} \{\langle s^i, y - y^i \rangle\}.$$

Remark 12.1 implies that for any symmetry point x^*, $\hat{f}(x^*) < 0$ if $\{y^i\}_{i=1}^{\ell}$ does not contain any optimal point.

12.5 Semidefinite Programing (SDP)

Semidefinite Programming (SDP) was the most exciting development in mathematical programming in the 1990's. We point out two among several reasons for that. First, the modelling power within the (SDP) framework has proved to be remarkable. Combinatorial optimization, control theory, and eigenvalue optimization are a few examples of fields were (SDP) had a drastic impact. Second, due to the development of interior point methods (IPMs), it was proved polynomial time complexity for (SDP).

Although IPMs enjoy an incredible success for (SDP) of moderate size, it does have limitations for large-scale instances. This is exactly where Bundle Methods can be an attractive approach.

We will be following [9] closely. First we introduce the needed notation for this section. Let S^n denote the set of all $n \times n$ symmetric matrices. $M \in S^n$ is said to be semidefinite positive if for every $d \in \mathbb{R}^n$,

$$\langle d, Md \rangle \geq 0.$$

Denote by S_+^n the set of all $n \times n$ symmetric semidefinite matrices. Given $A, B \in S^n$, $A \succcurlyeq B$ means that $A - B \in S_+^n$. The inner product defined on S^n is the natural extension of the usual inner product for $\mathbb{R}^n$,$\langle A, B \rangle = \mathbf{tr}(B^T A) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} B_{ij}$. A linear operator $\mathcal{A} : S^n \to \mathbb{R}^m$ and its adjoint[13] $\mathcal{A}^* : \mathbb{R}^m \to S^n$ are defined as

$$\mathcal{A}X = \begin{bmatrix} \langle A_1, X \rangle \\ \langle A_2, X \rangle \\ \vdots \\ \langle A_m, X \rangle \end{bmatrix}$$

and $\mathcal{A}^* y = \sum_{i=1}^m y_i A_i$, where $A_i \in S^n$ for $i = 1, \ldots, m$.

Now, we can define the standard primal (SDP) problem as

$$(P) \begin{cases} \max\limits_X & \langle C, X \rangle \\ & \mathcal{A}X = b \\ & X \succcurlyeq 0 \end{cases},$$

[13] Defined to induce $\langle \mathcal{A}X, y \rangle = \langle X, \mathcal{A}^* y \rangle$ for all $y \in \mathbb{R}^m$ and $X \in S^n$.

and the dual problem can be constructed as

$$(D)\begin{cases}\min\limits_{y,Z} & \langle b, y\rangle \\ & Z = \mathcal{A}^* y - C \\ & Z \succcurlyeq 0.\end{cases}$$

For convenience, we will assume that some constraint qualification does hold, implying in no duality gap between the primal and dual formulations, and for every optimal primal-dual pair of solutions (X^*, y^*, Z^*), we have that

$$X^* Z^* = 0.$$

Moreover, we assume that

$$\mathcal{A}X = b \text{ implies } \mathbf{tr}(X) = 1$$

Thus, a redundant constraint, $\mathbf{tr}(X) = 1$, can be added to the primal problem yielding the following dual equivalent to (D)

$$\begin{array}{ll}\min\limits_{y,\lambda,Z} & \lambda + \langle b, y\rangle \\ & Z = \mathcal{A}^* y + \lambda I - C \\ & Z \succcurlyeq 0.\end{array}$$

Since $\mathbf{tr}(X^*) = 1$ for any primal optimal solution, we must have $X^* \neq 0$. Therefore, any optimal dual solution (y^*, Z^*) must have Z^* to be a singular matrix, or equivalently, $\lambda_{\max}(-Z^*) = 0$. Rewriting our dual problem, we have

$$\begin{array}{ll}\min\limits_{y} & \lambda_{\max}(C - \mathcal{A}^* y) + \langle b, y\rangle \\ & y \in \mathbb{R}^m,\end{array}$$

which is an eigenvalue problem. For completeness, we recall some results regarding this problem. The function

$$\lambda_{\max} = \max\{\langle W, X\rangle : \mathbf{tr}(W) = 1, W \succcurlyeq 0\}$$

is a convex function on S^n_+, differentiable only if the maximal eigenvalue of X has multiplicity one. Unfortunately, when optimizing eigenvalue functions, the optimal is generelly attained at matrices whose maximal eigenvalue has multipicity greater than one. Using Lemma 8.1, the subdifferential of $\lambda_{\max}(\cdot)$ at X is exactly

$$\partial\lambda_{\max}(X) = \{W \succcurlyeq 0 : \langle W, X\rangle = \lambda_{\max}(X), \mathbf{tr}(W) = 1\}.$$

In particular, consider any $v \in \mathbb{R}^n$ with norm one contained in the eigensubspace of the maximal eigenvalue of X. Then, $W = vv^T$ is such that

$$\mathbf{tr}(W) = \mathbf{tr}(vv^T) = \mathbf{tr}(v^T v) = v^T v = 1,$$

and

$$\langle W, X\rangle = \left\langle vv^T, X\right\rangle = \langle v, Xv\rangle = \langle v, \lambda_{\max} v\rangle = \lambda_{\max}.$$

Thus, $vv^T \in \partial\lambda_{\max}$.

Returning to the function of interest,

$$f(y) = \lambda_{\max}(C - \mathcal{A}^* y) + \langle b, y \rangle ,$$

its subdifferential at y can be computed as

$$\partial f(y) = \{b - \mathcal{A}W : \langle W, C - \mathcal{A}^* y \rangle = \lambda_{\max}(C - \mathcal{A}^* y),\ \mathbf{tr}(W) = 1,\ W \succcurlyeq 0\}.$$

We also note that the set of subgradients is bounded, and thus our function f is Lipschitz.

Given a y, the oracle for f can be implemented through standard methods such as the Lanczos method.

12.5.1 Spectral Bundle Methods

Recently, a different version called Spectral Bundle Method was introduced for solving SDP problems. A different model was used to approximate f instead of the "classical" cutting plane model $\hat{f}$ to take advantage of the particular structure. Our proofs required mild assumptions on the model itself (as noted in Section 9.2). The convergence proofs are not affected at all.

Consider the auxiliary function

$$L(W, y) := \langle C - \mathcal{A}^* y, W \rangle + \langle b, y \rangle .$$

Using Lemma 8.1, we can rewrite our function of interest as

$$f(y) = \max\{L(W, y) : W \succcurlyeq 0, \mathbf{tr}(W) = 1\}.$$

This will be used to motivate a new lower approximation for f. For any set $\mathcal{W} \subset \{W \in S^n : W \succcurlyeq 0, \mathbf{tr}(W) = 1\}$, we have that

$$f(y) \geq \hat{f}_{\mathcal{W}}(y) := \max\{L(W, y) : W \in \mathcal{W}\}.$$

The Spectral Bundle Method uses the following definition for $\mathcal{W}$. For $r < n$, consider a (orthogonal) matrix $P \in \mathbb{R}^{n \times r}$ such that $P^T P = I_r \in \mathbb{R}^{r \times r}$, and $\overline{W} \in S^n_+$ with $\mathbf{tr}(\overline{W}) = 1$. The set of semidefinite matrices that we will take the supremum over is

$$\hat{\mathcal{W}} = \{\alpha \overline{W} + PVP^T : \alpha + \mathbf{tr}(V) = 1,\ \alpha \geq 0,\ V \succcurlyeq 0\}.$$

The new minorant model $\hat{f}$ is defined as

$$\hat{f}(y) = \max\{L(W, y) : W \in \hat{\mathcal{W}}\}$$

which now depends on P and $\overline{W}$. We note that not only we have $\hat{f}(y) \leq f(y)$, but also if for some $\hat{y} \in \mathbb{R}^m$, $vv^T \in \hat{\mathcal{W}}$ for some eigenvector v associated with $\lambda_{\max}(C - \mathcal{A}^* \hat{y})$, then $\hat{f}(\hat{y}) = f(\hat{y})$. For example, if v is a column of P or v belongs to the range of P.

As motivated in the last paragraph, the idea is for P to contain subgradient information from the current center and from previous iterations. This generates a more accurated model than a cutting-plane model based on the same

subgradients. The cost associated with this gain is paid in the computation of the next iterate. Instead of the "classical" quadratic problem, we need to solve

$$\min_y \max_{W\in\hat{\mathcal{W}}} L(W,y) + \frac{\mu_k}{2}\|y-\hat{x}^k\|^2,$$

which includes a semidefinite constraint in the definition of $\hat{\mathcal{W}}$. Since y is unconstrained, the problem can be simplified using duality. Interchanging max and min,

$$\min_y \max_{W\in\hat{\mathcal{W}}} L(W,y) + \frac{\mu_k}{2}\|y-\hat{x}^k\|^2 = \max_{W\in\hat{\mathcal{W}}} \min_y L(W,y) + \frac{\mu_k}{2}\|y-\hat{x}^k\|^2,$$

we can now optimize in y for any fixed W. The first order condition,

$$\nabla_y \left(L(W,y) + \frac{\mu_k}{2}\|y-\hat{x}^k\|^2 \right) = b - \mathcal{A}W + \mu_k(y-\hat{x}^k) = 0,$$

is used to eliminate the variable y. We obtain

$$\max_{W\in\hat{\mathcal{W}}} \left\langle C - \mathcal{A}^*\hat{x}^k, W \right\rangle + \left\langle b, \hat{x}^k \right\rangle - \frac{1}{2\mu_k} \left\langle \mathcal{A}W - b, \mathcal{A}W - b \right\rangle, \tag{23}$$

a semidefinite program with quadratic cost function which can be solved by interior point methods.

Now, the trade-off in the choice of r, the number of columns of P, is clear. We want to increase r to obtain a better model but we also need to keep it relatively small so we can still solve (23) through interior point methods. Aggregation is an important feature to control r and still ensure convergence.

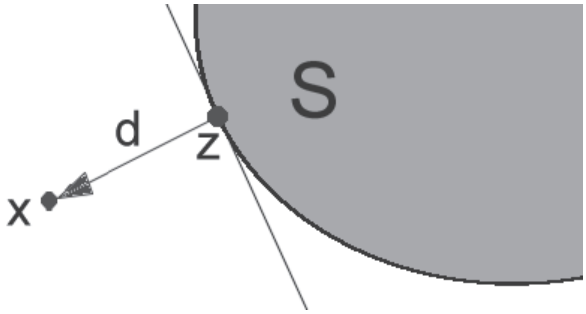
When [9] was published, interior-point methods were restricted to SDP problems with semidefinite blocks of sizes up to 500, while the Spectral Bundle method was capable to deal problems with semidefinite blocks of size up to 3000.

A Appendix: Separating Hyperplane (geometric version of Hahn-Banach)

Theorem A.1 *Given a nonempty closed convex set $S \subset \mathbb{R}^n$. Let $x \in \mathbb{R}^n$, $x \notin S$, there exists $d \in \mathbb{R}^n$ such that*

$$\langle d, x\rangle > \langle d, y\rangle \qquad \text{for all } y \in S.$$

Proof. Since S is closed, there exists $z \in \partial S$ such that $\|z - x\| = \inf\{\|y - x\| : y \in S\}$.



We claim that $\langle x - z, y - z\rangle \leq 0$ for all $y \in S$. Suppose that there exists z such that $\langle x - z, y - z\rangle > 0$. For $0 \leq t \leq 1$, let $s^t := z + t(y - z)$, so

$$\|s^t - x\|^2 = \|z - x\|^2 - 2t\langle x - z, y - z\rangle + t^2\|y - z\|^2,$$

which is smaller than $\|z - x\|^2$ for $t > 0$ small enough. This is a contradiction with z being the point of minimum distance.

Thus, for $d := x - z$,

$$\langle d, y - z\rangle \leq 0 \qquad \text{for all } y \in S.$$

So,

$$\langle d, y\rangle \leq \langle d, z\rangle = \langle d, x - d\rangle = \langle d, x\rangle - \langle d, d\rangle .$$

■

Corollary A.1 *Given a nonempty closed convex set $S \subseteq \mathbb{R}^n$. If $x \in \partial S$, there exists d such that*

$$\langle d, x\rangle \geq \langle d, y\rangle \qquad \text{for all } y \in S.$$

Proof. Take a sequence $x^k \to x$, $x^k \notin S$. Applying the Separating Hyperplane Theorem, there exists d^k, $\|d^k\| = 1$, such that

$$\langle d^k, x^k\rangle > \langle d^k, y\rangle \text{ for all } y \in S.$$

Since $\{d^k\}$ is a bounded sequence, there exists a convergent subsequence $d^{n_k} \to d$, $\|d\| = 1$. So,

$$\langle d, x\rangle \geq \langle d, y\rangle \text{ for all } y \in S.$$

■

B Appendix: Hahn-Banach for real vector spaces

Theorem B.1 (Hahn-Banach for real vector spaces). *Let X be a Banach space and p a convex functional on X, then there exists a linear functional λ on X such that*

$$p(x) \geq \lambda(x) \qquad \forall x \in X$$

Proof. Take $0 \in X$, and define $\tilde{X} = \{0\}$, $\tilde{\lambda}(0) = p(0)$.
If $\exists z \in X$, $z \notin \tilde{X}$, extend $\tilde{\lambda}$ from $\tilde{X}$ to the subspace generated by $\tilde{X}$ and z,

$$\tilde{\lambda}(tz + \tilde{x}) = t\tilde{\lambda}(z) + \tilde{\lambda}(\tilde{x}).$$

Suppose $x_1, x_2 \in \tilde{X}$ and $\alpha > 0, \beta > 0$.

$$\begin{aligned}
\beta\lambda(x_1) + \alpha\lambda(x_2) &= \lambda(\beta x_1 + \alpha x_2) \\
&= (\alpha+\beta)\lambda\left(\tfrac{\beta}{\alpha+\beta}x_1 + \tfrac{\alpha}{\alpha+\beta}x_2\right) \\
&\leq (\alpha+\beta)p\left(\tfrac{\beta}{\alpha+\beta}x_1 + \tfrac{\alpha}{\alpha+\beta}x_2\right) \\
&= (\alpha+\beta)p\left(\left[\tfrac{\beta}{\alpha+\beta}\right](x_1 - \alpha z) + \left[\tfrac{\alpha}{\alpha+\beta}\right](x_2 + \beta z)\right) \\
&\leq \beta p(x_1 - \alpha z) + \alpha p(x_2 + \beta z)
\end{aligned}$$

Thus,

$$\begin{aligned}
\beta\left[-p(x_1 - \alpha z) + \lambda(x_1)\right] &\leq \alpha\left[p(x_2 + \beta z) - \lambda(x_2)\right] \\
\frac{1}{\alpha}\left[-p(x_1 - \alpha z) + \lambda(x_1)\right] &\leq \frac{1}{\beta}\left[p(x_2 + \beta z) - \lambda(x_2)\right]
\end{aligned}$$

$$\sup_{x_1 \in X, \alpha > 0} \frac{1}{\alpha}\left[-p(x_1 - \alpha z) + \lambda(x_1)\right] \leq a \leq \inf_{x_2 \in X, \beta > 0} \frac{1}{\beta}\left[p(x_2 + \beta z) - \lambda(x_2)\right]$$

Define $\tilde{\lambda}(z) = a$, then assuming $t > 0$,

$$\begin{aligned}
\tilde{\lambda}(tz + \tilde{x}) &= t\tilde{\lambda}(z) + \tilde{\lambda}(\tilde{x}) = ta + \tilde{\lambda}(tx) \\
&\leq t\left(\tfrac{1}{t}p(\tilde{x} + tz) - \tfrac{\lambda(\tilde{x})}{t}\right) + \lambda(\tilde{x}) \\
&\leq p(tz + \tilde{x})
\end{aligned}$$

the case of $t < 0$ is similar. To extend for X, we will use the Zorn's lemma.
Let $\mathcal{E}$ be a collection of all extensions e of λ, $e(x) \leq p(x)$ for all $x \in X_e$. Where $e_1 \prec e_2$ if $X_{e_1} \subseteq X_{e_2}$ and $e_1(x) = e_2(x)$ in X_{e_1}. Thus, $\mathcal{E}$ is a partially ordered and $\mathcal{E} \neq \emptyset$.

If $\{e_s\}_{s \in J}$ is a totally ordered subset of $\mathcal{E}$, $\Lambda = \cup_{s \in J} X_{e_s}$ is a subspace (monotone union of subspaces) and

$$e : \Lambda \to \mathbb{R}, \qquad e(x) = e_s(x), \quad \text{if } \; x \in X_{e_s}.$$

e is well defined and $e_s \prec e$ for all $s \in J$. Then e is maximal for J. Thus, $\mathcal{E}$ must have a maximal element (Zorn's Lemma) $\tilde{\lambda}$.

So, $\tilde{\lambda}$ must be defined on X, otherwise we could extend it contradicting the fact that it is maximal. ■

References

[1] A. Belloni and A. Lucena, *Improving on the Held and Karp Bound*, Working Paper.

[2] A. Belloni and A. Lucena, *Lagrangian Heuristics to Linear Ordering*, Metaheuristics: Computer-Decision Making, Kluwer Academic Publishers, 2004 (Book).

[3] A. Belloni and C. Sagastizábal, *Dynamic Bundle Methods: Application to Combinatorial Optimization*, Submitted to Mathematical Programming (July 2004).

[4] A. Belloni and R. Freund, *Symmetry Points of a Convex Set: Basic Properties and Computational Complexity*, Submitted to Mathematical Programming (July 2004).

[5] J. F. Bonnans, J. Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal,*Numerical Optimization: Theoretical and Practical Aspects*, Universitext, Springer-Verlag, Berlin, 2003, xiv+423.

[6] A. Frangioni, *Solving Semidefinite Quadratic Problems Within Nonsmooth Optimization Algorithms*, Computers & Operations Research 23(11), p.1099 - 1118, 1996.

[7] M. Grötschel, M. Jünger, and G. Reinelt, *A cutting-plane algorithm for the linear ordering problem*,Operations Research, 32:1195-1220, 1984.

[8] M. Held and R. M. Karp, *The traveling-salesman problem and minimum spaning tress*, Operations Research, 17, 1138-1162, 1970.

[9] C. Helmberg and F. Rendl, *A Spectral Bundle Method for Semidefinite Programming*, SIAM J. Otimization, Vol. 10, No. 3, pp. 673-696.

[10] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I & II*, Grund. der math. Wiss, no. 305-306, Springer-Verlag, 1993.

[11] R. T. Rockafellar, *Convex Analysis* Princeton University Press, 1970.

[12] N. Shor, *Minimization methods for non-differentiable functions*,Springer-Verlag, Berlin, 1985.