Bundle-type methods for inexact data

Csaba Fabian
Kecskemét College

Bundle-type Methods for Inexact Data

Csaba I. Fábián*

Abstract

We treat such convex programming problems, the constraints and objective functions of which are not easy to handle: the more accurate data we want, the greater the computational effort.

We propose such bundle-type methods that keep balance between the amount of work invested into estimating the function data on the one hand, and into the optimization method on the other hand. At the beginning, only rough estimates of the function values and gradients are needed. As the optimum is gradually approached, more and more accurate data are computed.

The proposed methods inherit stability from bundle methods, moreover, the functions need not be smooth.

1 Introduction

We treat such convex programming problems, where the constraints and objective functions cannot be computed exactly. Let $f : \mathbb{R}^n \to \mathbb{R}$ be one of the functions, and $\boldsymbol{x}_0 \in \mathbb{R}^n$. We cannot construct any supporting hyperplane for f at $\boldsymbol{x}_0$. But we can construct approximating hyperplanes with a prescribed tolerance of $\delta > 0$. In other words, we can construct a linear function $l : \mathbb{R}^n \to \mathbb{R}$ such that

$$f \geq l \quad \text{and} \quad f(\boldsymbol{x}_0) - l(\boldsymbol{x}_0) \leq \delta. \tag{1}$$

Such functions often occur in stochastic programming problems. The more accurate data we want, the greater the computational effort.

The proposed solution methods are modifications of the Level Method and Constrained Level Method of Lemaréchal, Nemirovskii, and Nesterov (1995). We make the methods capable of handling inexact data. At the beginning, only rough estimates of the function values and gradients are needed. As the optimum is gradually approached, more and more accurate data are computed. Such heuristics are already employed in the code of Szántai (1988). The present procedures have the advantage that the required accuracy is known at each step, hence better coordination of efforts is possible. Moreover, the present procedures inherit stability from bundle-type methods.

In Section 2, we minimize a convex function over a polyhedron (or over some other 'simple' convex set). In Section 3, we minimize a convex function over the common part of the level sets of convex constrain functions.

2 Minimization over a Polyhedron

The aim is to minimize the convex function $f : \mathbb{R}^n \to \mathbb{R}$ over a bounded convex polyhedron X. Such an f is known to satisfy the Lipschitz condition relative to X (see Theorem 10.4 of Rockafellar (1970)). Assume that the polyhedron has the diameter D.

*Department of Operations Research, Loránd Eötvös University of Budapest. P.O.Box 120, Budapest, H-1518 Hungary. E-mail: fabian@cs.elte.hu. This work was supported by the Bolyai János Foundation of the Hungarian Academy of Sciences; by AKP, Research Fund of the Hungarian Academy of Sciences, project 98-70 2,1/31; and by OTKA, Hungarian National Fund for Scientific Research, projects T029775 and T032369.

Assume that given a feasible point $\boldsymbol{x}_0$ and a tolerance of $\delta > 0$, we can construct a linear function $l : \mathbb{R}^n \to \mathbb{R}$ such that (1) holds. Assume that the construction is such that l satisfies the Lipschitz condition with the constant Λ. (This is a reasonable assumption: Since l is an approximation of a supporting hyperplane of f, it inherits Lipschitz continuity from f.)

We construct a series of feasible solutions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots$ with improving objective values. In the construction of new iterates we use estimates of the objective value and subgradient at former iterates.

Let us survey the usual methods for the case when exact supporting hyperplanes can be computed. The classical cutting-plane method of Kelley (1960) is known as slow. (Slowness is the consequence of instability.) Bundle methods are refinements of the classical cutting-plane method. A comprehensive treatment of the bundle methods can be found in Hiriart-Urruty and Lemaréchal (1993). The idea is to maintain a stability center or prox-center, that is, to distinguish one of the iterates generated that far. The prox-center is updated every time a significantly better point was found. Roaming away from the current prox-center is penalized. Thus, bundle methods reduce the influence of the inaccuracy of the cutting-plane approximation, thereby reducing instability. More recently, Lemaréchal, Nemirovskii, and Nesterov (1995) proposed such variants of the bundle methods, that the prox-center can be forgotten about. (Though it exists implicitly, and attracts the path along which new iterates are visited.) The Level Method is one of those. We make the Level Method capable of handling inexact data.

In Section 2.1, the Level Method is briefly described, and the convergence results of Lemaréchal, Nemirovskii, and Nesterov are summed up. Though, for the sake of simplicity, we prove slightly weaker bounds. In Section 2.2, we show how to handle inexact data.

2.1 Description of the Level Method

In this section, we assume that exact supporting hyperplanes can be computed. Having generated the iterates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i \in X$, and using an oracle to compute the function values $f(\boldsymbol{x}_i)$ and the subgradients $\nabla f(\boldsymbol{x}_i)$, the supporting hyperplane approximation of f is

$$f_i(\boldsymbol{x}) = \max_{1 \le j \le i} \left\{ f(\boldsymbol{x}_j) + \left[\nabla f(\boldsymbol{x}_j)\right]^T (\boldsymbol{x} - \boldsymbol{x}_j) \right\} \qquad (\boldsymbol{x} \in X)$$

where f_i is a piecewise linear and convex function that inherits Lipschitz continuity from f. We have $f_i(\boldsymbol{x}_j) = f(\boldsymbol{x}_j)\ \ (j \le i)$ and $f_1 \le f_2 \le \ldots \le f_i \le f$.

The aim is to enclose the optimal solution in a contracting sequence of sets. If the iterates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots$ are properly selected, then the functions f_i are more and more accurate approximations of f in a contracting neighborhood of an optimal solution.

The classical cutting plane method computes the iterate $\boldsymbol{x}_{i+1}$ by minimizing f_i over X. The problem is that the inaccuracy of the approximation may have an excessive effect on the selection of new iterates.

The Level Method is a natural refinement of the classical cutting plane method. The best function value obtained until the ith step,

$$U_i = \min_{1 \le j \le i} f(\boldsymbol{x}_j)$$

is used as an upper bound for the optimum. A lower bound is the minimum of the ith approximating function:

$$L_i = \min_{\boldsymbol{x} \in X} f_i(\boldsymbol{x}).$$

The gap between the above bounds is $\Delta_i = U_i - L_i$.

The sequence of the upper bounds U_i is monotone decreasing, and the sequence of the lower bounds L_i is monotone increasing. Hence, the gap is tightening at each step.

Let $0 < \lambda < 1$ be some preset parameter. Consider the level set

$$X_i = \{ \boldsymbol{x} \in X \mid f_i(\boldsymbol{x}) \le L_i + \lambda \Delta_i \} .$$

The next iterate $\boldsymbol{x}_{i+1}$ is computed by projecting $\boldsymbol{x}_i$ onto the level set X_i. That is,

$$\boldsymbol{x}_{i+1} = \arg\min_{\boldsymbol{x}\in X_i} \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}),$$

where dist means the Euclidean distance. (Setting $\lambda = 0$ gives the classical method. With non-extremal setting, the level sets stabilize the procedure.)

2.1.1 The Framework of the Level Method

Initialization.

Set the stopping tolerance $\epsilon > 0$.

Set the parameter $\lambda \quad (0 < \lambda < 1)$.

Find a starting point $\boldsymbol{x}_1 \in X$.

Let $i := 1$ (iteration counter).

Update bundle.

Compute the function value $f(\boldsymbol{x}_i)$ and a subgradient $\nabla f(\boldsymbol{x}_i)$.

Define the upper cover $f_i(\boldsymbol{x}) := \max_{1\le j\le i} \left\{ f(\boldsymbol{x}_j) + [\nabla f(\boldsymbol{x}_j)]^T (\boldsymbol{x} - \boldsymbol{x}_j) \right\}$.

Compute the bounds $U_i := \min_{1\le j\le i} f(\boldsymbol{x}_j), \quad L_i := \min_{\boldsymbol{x}\in X} f_i(\boldsymbol{x})$,
and the gap $\Delta_i := U_i - L_i$.

If $\Delta_i < \epsilon$ then near-optimal solution found, stop.

Find new iterate.

Define the level set $X_i := \{\boldsymbol{x} \in X \mid f_i(\boldsymbol{x}) \le L_i + \lambda\Delta_i\}$,
and let $\boldsymbol{x}_{i+1}$ be the projection of $\boldsymbol{x}_i$ onto X_i.

Increment i.

$\rightarrow$ *Update bundle.*

2.1.2 Convergence Proofs.

We are going to compute an upper bound for the number of iterations required to decrease the gap under a prescribed tolerance. The key argument is the following

Proposition 1 *Consider a sequence of iterations at the end of which the gap has not been reduced much. Namely, let t and s be natural numbers, $t < s$, and assume that the following inequality holds*

$$(1-\lambda)\Delta_t \le \Delta_s. \tag{2}$$

Then, the number of iterations performed while getting from $\boldsymbol{x}_t$ to $\boldsymbol{x}_s$ cannot be greater than

$$\left(\frac{\Lambda D}{(1-\lambda)\Delta_s}\right)^2.$$

Proof. Consider the intervals

$$[\, L_t,\, U_t \,] \supseteq [\, L_s,\, U_s \,]\,,$$

the lengths of which are Δ_t and Δ_s, respectively. The point $L_t + \lambda\Delta_t$ divides the first interval into two subintervals, the upper one having a length of $(1-\lambda)\Delta_t$. According to the assumption (2), the upper subinterval cannot contain $[L_s, U_s]$ in the interior. Hence, we have $L_s \le L_t + \lambda\Delta_t$. Let us select a point

$$\boldsymbol{u}_s \in \{\, \boldsymbol{x} \in X \mid f_s(\boldsymbol{x}) \le L_s \,\}.$$

(Due to the definition of L_i, the above set is not empty.) Then, we have $\boldsymbol{u}_s \in X_t$. Similarly, each level set X_i $(t \le i \le s)$ contains $\boldsymbol{u}_s$.

The point $\boldsymbol{u}_s$ can be viewed as a stability center, though it does not lie on the path of the method. (That is, $\boldsymbol{u}_s$ is not among the iterates $\boldsymbol{x}_i$.) We are going to show that $\boldsymbol{u}_s$ indeed attracts the path along which new iterates are visited. By definition, the iterate $\boldsymbol{x}_{i+1}$ is the projection of $\boldsymbol{x}_i$ onto the level set X_i. Hence, the iterates are getting closer and closer to the point $\boldsymbol{u}_s$. Namely, from the properties of the projection, we have

$$\operatorname{dist}(\boldsymbol{x}_{i+1}, \boldsymbol{u}_s)^2 \;\le\; \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{u}_s)^2 \;-\; \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})^2 \qquad (t \le i \le s-1)\,. \tag{3}$$

We are going to find a lower bound for the decrease $\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})^2$. Obviously, we have

$$f_i(\boldsymbol{x}_i) = f(\boldsymbol{x}_i) \;\ge\; U_i \quad \text{and} \quad f_i(\boldsymbol{x}_{i+1}) \;\le\; L_i + \lambda\Delta_i \qquad (t \le i \le s-1)\,.$$

From the above inequalities, we obtain

$$f_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_{i+1}) \;\ge\; (U_i - L_i) - \lambda\Delta_i \;=\; (1-\lambda)\Delta_i \qquad (t \le i \le s-1)\,.$$

The function f_i is Lipschitz continuous with the constant Λ. It follows that

$$\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1}) \;\ge\; \frac{1}{\Lambda}\,(1-\lambda)\Delta_i\,.$$

By substituting this into the inequality (3), we obtain

$$\operatorname{dist}(\boldsymbol{x}_{i+1}, \boldsymbol{u}_s)^2 \;\le\; \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{u}_s)^2 - \left(\frac{1}{\Lambda}\,(1-\lambda)\Delta_i\right)^2 \qquad (t \le i \le s-1)\,.$$

We have $\Delta_i \ge \Delta_s$, hence, we may substitute Δ_s for Δ_i in the right-hand side of the above expression. Moreover, D being the diameter of the feasible domain, we have

$$\operatorname{dist}(\boldsymbol{x}_t, \boldsymbol{u}_s) \le D\,.$$

Hence, the number of steps performed while getting from $\boldsymbol{x}_t$ to $\boldsymbol{x}_s$, cannot be greater than

$$\frac{D^2}{\left(\frac{1}{\Lambda}\,(1-\lambda)\Delta_s\right)^2}\,.$$

□

Corollary 2 *Consider a sequence of iterations at the end of which the gap has not been reduced much. Namely, let $t < s$, and assume that the following inequality holds*

$$(1-\lambda)\Delta_t \le \Delta_s\,.$$

Then, the number of iterations performed while getting from $\boldsymbol{x}_t$ to $\boldsymbol{x}_s$, cannot be greater than

$$\left(\frac{\Lambda D}{(1-\lambda)^2\Delta_t}\right)^2.$$

Theorem 3 *To obtain a gap smaller than ϵ, it suffices to perform*

$$c(\lambda)\,\left(\frac{\Lambda D}{\epsilon}\right)^2$$

iterations, where $c(\lambda)$ is a constant that depends only on λ.

Proof. Starting from $\boldsymbol{x}_1$, consider the maximal sequence of iterations :

$$\boldsymbol{x}_1 \to \boldsymbol{x}_2\,,\ \ldots,\ \boldsymbol{x}_{s-1} \to \boldsymbol{x}_s\,,$$

at the end of which the gap has not been reduced much. That is, the following inequalities hold:

$$(1-\lambda)\Delta_1 \le \Delta_s\,, \quad (1-\lambda)\Delta_1 > \Delta_{s+1}\,.$$

The iteration $\boldsymbol{x}_s \to \boldsymbol{x}_{s+1}$ will be called *critical*.

The above construction is repeated starting from $\boldsymbol{x}_{s+1}$. Thus, the iterations are grouped into sequences, and the sequences are separated with critical iterations. Assume there are k sequences. Denote by $\Delta^{(\ell)}$ ($1 \le \ell \le k$) the gap at the start of the ℓth sequence. (E.g., we have $\Delta^{(1)} = \Delta_1$, $\Delta^{(2)} = \Delta_{s+1}$.) From the definitions, it follows that

$$(1-\lambda)\Delta^{(\ell)} \;>\; \Delta^{(\ell+1)} \qquad (\ell = 1, \ldots, k-1)\,.$$

We have $\Delta^{(k)} > \epsilon$. (Otherwise the process stops before the kth sequence.) It follows that

$$(1-\lambda)^{k-\ell}\,\Delta^{(\ell)} \;>\; \epsilon \qquad (\ell = 1, \ldots, k)\,. \tag{4}$$

Hence, from Corollary 2, it follows that the length of the ℓth sequence is not greater than

$$\left(\frac{\Lambda D}{(1-\lambda)^2\,\epsilon}\right)^2 (1-\lambda)^{2(k-\ell)} \qquad (\ell = 1, \ldots, k)\,.$$

The number of non-critical iterations can be computed by summing up the lengths of the sequences. An upper bound is

$$\left(\frac{\Lambda D}{(1-\lambda)^2\,\epsilon}\right)^2 \sum_{\ell=1}^{k} (1-\lambda)^{2(k-\ell)} \;\le\; \left(\frac{\Lambda D}{(1-\lambda)^2\,\epsilon}\right)^2 \sum_{(k-\ell)=0}^{\infty} (1-\lambda)^{2(k-\ell)}.$$

The number of the critical iterations, k, is negligible compared to the above expression. □

Remark 4 *A drawback of the method is the constantly increasing size of the linear and quadratic problems that need to be solved at each iteration. In order to overcome this, reasonable strategies are wanted to discard some of the cutting planes, or aggregate the information obtained.*

Lemaréchal, Nemirovskii, and Nesterov report on successful applications of the Level Method to a variety of problems. (The largest of which had 442 variables.) The following reduction strategy is applied: After each critical iteration, all the non-binding cutting planes are eliminated. Namely, if the sth iteration turns out to be critical, then they eliminate all the cutting planes which are inactive at the determination of the gap Δ_{s+1}. It can easily be seen that the theoretical efficiency estimates are preserved by the above reduction strategy.

Concerning practical behavior, the numbers of the iterations are reported to have increased by at most 15%. (The constant $\lambda = \frac{1}{2}$ was used in each test run.) Moreover, the cardinality of the bundle (i.e., the number of the cutting planes stored simultaneously) never exceeded $2n$. This implies that the number of steps between two consecutive critical iterations may be much smaller in practice than the theoretical bound in Proposition 1.

2.2 Handling Inexact Data

Suppose that we have constructed the approximate supporting hyperplanes $l_1, \ldots, l_i$ at the points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i$ and with the accuracy of $\delta_1, \ldots, \delta_i$, respectively. I.e., we have

$$f \geq l_j \quad \text{and} \quad f(\boldsymbol{x}_j) - l_j(\boldsymbol{x}_j) \;\leq\; \delta_j \qquad (1 \leq j \leq i).$$

Assume that each linear function l_j satisfies the Lipschitz condition with the constant Λ.

The ith approximating function is defined as

$$f_i(\boldsymbol{x}) = \max_{1 \leq j \leq i} l_j(\boldsymbol{x}) \qquad (\boldsymbol{x} \in X)$$

where f_i is a piecewise linear and convex function that inherits the Lipschitz continuity from the linear functions. Moreover, we have $f(\boldsymbol{x}_j) \;\leq\; l_j(\boldsymbol{x}_j) + \delta_j \;\leq\; f_i(\boldsymbol{x}_j) + \delta_j \;\; (j \leq i)$, and $f_1 \leq f_2 \leq \ldots \leq f_i \leq f$.

At the ith step, we have the following upper bound for the optimum:

$$U_i = \min_{1 \leq j \leq i} \; f_i(\boldsymbol{x}_j) + \delta_j \; .$$

A lower bound is the minimum of the ith approximating function:

$$L_i = \min_{\boldsymbol{x} \in X} \; f_i(\boldsymbol{x}) \, .$$

The sequence of the upper bounds U_i is monotone decreasing, and the sequence of the lower bounds L_i is monotone increasing. Hence the gap $\Delta_i = U_i - L_i$ is tightening at each step.

Let $0 < \lambda < 1$ be some preset parameter. Consider the level set

$$X_i = \{ \, \boldsymbol{x} \in X \mid f_i(\boldsymbol{x}) \leq L_i + \lambda \Delta_i \, \} \; .$$

The next iterate $\boldsymbol{x}_{i+1}$ is computed by projecting $\boldsymbol{x}_i$ onto the level set X_i. That is,

$$\boldsymbol{x}_{i+1} = \; \arg\min_{\boldsymbol{x} \in X_i} \; \mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x}).$$

The accuracy to be prescribed for the approximate supporting hyperplane at the new iterate: Let $\delta_1 > 0$ be selected arbitrarily. Let $\delta_{i+1} = \gamma \Delta_i \quad (i = 1, 2, \ldots)$, where γ is a constant parameter satisfying $0 < \gamma < (1 - \lambda)^2$.

Remark 5 *The sequence* $\delta_i \;\; (i = 2, 3, \ldots)$ *is monotone decreasing.*

2.2.1 The Framework of the Inexact Level Method

Initialization.

Determine the Lipschitz constant Λ.

Set the stopping tolerance $\epsilon > 0$.

Set the parameters λ and $\gamma \quad (0 < \lambda < 1, \; 0 < \gamma < (1 - \lambda)^2)$.

Find a starting point $\boldsymbol{x}_1 \in X$.

Set the starting accuracy $\delta_1 > 0$.

Let $i := 1$ (iteration counter).

Update bundle.

Construct the cutting plane l_i at the point $\boldsymbol{x}_i$ with the accuracy of δ_i. Moreover, l_i must satisfy the Lipschitz condition with the constant Λ.

Define the upper cover $f_i(\boldsymbol{x}) := \max_{1 \le j \le i} l_j(\boldsymbol{x})$.

Compute the bounds $U_i := \min_{1 \le j \le i} f_i(\boldsymbol{x}_j) + \delta_j, \quad L_i := \min_{\boldsymbol{x} \in X} f_i(\boldsymbol{x}),$

and the gap $\Delta_i := U_i - L_i$.

If $\Delta_i < \epsilon$, then near-optimal solution found, stop.

Find new iterate.

Define the level set $X_i := \{\boldsymbol{x} \in X \mid f_i(\boldsymbol{x}) \le L_i + \lambda\Delta_i\}$,

and let $\boldsymbol{x}_{i+1}$ be the projection of $\boldsymbol{x}_i$ onto X_i.

Let $\delta_{i+1} := \gamma\Delta_i$.

Increment i.

$\rightarrow$ *Update bundle.*

2.2.2 Convergence Proofs.

We are going to compute an upper bound for the number of iterations required to decrease the gap under a prescribed tolerance.

Proposition 6 *Consider a sequence of iterations at the end of which the gap has not been reduced much. Namely, let $t < s$, and assume that the following inequality holds*

$$(1-\lambda)\Delta_t \le \Delta_s \,. \tag{5}$$

Then, the number of iterations performed while getting from $\boldsymbol{x}_t$ to $\boldsymbol{x}_s$, cannot be greater than

$$1 \; + \; \left(\frac{\Lambda\, D}{\left(1 - \lambda - \frac{\gamma}{1-\lambda}\right) \Delta_s} \right)^2 .$$

Proof. We count the number of iterations while getting from $\boldsymbol{x}_{t+1}$ to $\boldsymbol{x}_s$. (The 1 in the above expression is for the first iteration, $\boldsymbol{x}_t \rightarrow \boldsymbol{x}_{t+1}$.)

Let us select a point $\boldsymbol{u}_s \in \{\, \boldsymbol{x} \in X \mid f_s(\boldsymbol{x}) \le L_s \,\}$. (Due to the definition of L_i, the above set is not empty.) As in the exact case, each level set X_i $(t+1 \le i \le s)$ contains $\boldsymbol{u}_s$. (See the proof of Proposition 1.) Since the iterate $\boldsymbol{x}_{i+1}$ is the projection of $\boldsymbol{x}_i$ onto the level set X_i, it follows that the iterates are getting closer and closer to the point $\boldsymbol{u}_s$. Specifically, we have

$$\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{u}_s)^2 - \operatorname{dist}(\boldsymbol{x}_{i+1}, \boldsymbol{u}_s)^2 \ge \operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1})^2 \ge \left(\frac{f_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_{i+1})}{\Lambda} \right)^2 \tag{6}$$

for $t+1 \le i \le s-1$. As for the right-hand side expression, we have

$$f_i(\boldsymbol{x}_i) \ge U_i - \delta_i \quad \text{and} \quad f_i(\boldsymbol{x}_{i+1}) \le L_i + \lambda\Delta_i.$$

From the above inequalities and the definition of δ_i, we obtain

$$f_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_{i+1}) \;\geq\; (U_i - L_i) - \lambda\Delta_i - \delta_i \;=\; (1-\lambda)\Delta_i - \gamma\Delta_{i-1}\,.$$

Here, we have $(1-\lambda)\Delta_{i-1} \leq \Delta_i$ as a consequence of the assumption (5). (That's why we need $t+1 \leq i$.) Hence the above inequality can be continued as

$$f_i(\boldsymbol{x}_i) - f_i(\boldsymbol{x}_{i+1}) \;\geq\; (1-\lambda)\Delta_i - \gamma\Delta_{i-1} \;\geq\; \left(1 - \lambda - \frac{\gamma}{1-\lambda}\right)\Delta_i\,.$$

Substituting this into the inequality (6), we obtain that

$$\operatorname{dist}(\boldsymbol{x}_i, \boldsymbol{u}_s)^2 - \operatorname{dist}(\boldsymbol{x}_{i+1}, \boldsymbol{u}_s)^2 \geq \left[\,\frac{1}{\Lambda}\left(1 - \lambda - \frac{\gamma}{1-\lambda}\right)\Delta_i\,\right]^2$$

holds for $t+1 \leq i \leq s-1$. The right-hand side expression of the above inequality is strictly positive due to the selection $\gamma < (1-\lambda)^2$. The proof can be finished like the proof of Proposition 1. □

Applying the arguments presented in Section 2.1.2 to the result of Proposition 6 instead of to that of Proposition 1, we obtain

Theorem 7 *To obtain a gap smaller than ϵ, it suffices to perform*

$$c(\lambda, \gamma)\;\left(\frac{\Lambda D}{\epsilon}\right)^2$$

iterations, where $c(\lambda, \gamma)$ is a constant that depends only on λ and γ.

Remark 8 *The bundle reduction strategy mentioned in Remark 4 can be applied to the inexact case as well.*

3 Constrained Minimization

Let $X \subset \mathbb{R}^n$ be a bounded convex polyhedron with diameter D. Let f and $g^1, \ldots, g^m$ be $X \to \mathbb{R}$ convex functions, each satisfying the Lipschitz condition with the constant Λ. The aim is to minimize $f(\boldsymbol{x})$ subject to $\boldsymbol{x} \in X,\ g^1(\boldsymbol{x}) \leq 0,\ \ldots,\ g^m(\boldsymbol{x}) \leq 0$.

Let us define the function

$$g = \left[\max_{1 \leq k \leq m} g^k\right]_+,$$

where $[.]_+$ means the positive part of real number. (Hence g is the upper cover of the functions $g^0, g^1, \ldots, g^m$, where $g^0 : X \to \mathbb{R}$ denotes the constant 0 function.) This is also a convex function satisfying the Lipschitz condition with the constant Λ. The problem can be stated in the form

$$\min f(\boldsymbol{x}) \quad \text{such that } \boldsymbol{x} \in X,\ g(\boldsymbol{x}) \leq 0. \tag{7}$$

Bundle-type methods can handle this problem, since they do not require smoothness. We assume that (7) is really a constrained problem and is consistent, i.e., g takes positive values as well as 0.

Assume that given a point $\boldsymbol{x}_0 \in X$, and a tolerance of $\delta > 0$, we can construct linear functions $l, z : \mathbb{R}^n \to \mathbb{R}$ such that

$$\begin{array}{lll} f \geq l & \text{and} & f(\boldsymbol{x}_0) - l(\boldsymbol{x}_0) \;\leq\; \delta, \\ g \geq z & \text{and} & g(\boldsymbol{x}_0) - z(\boldsymbol{x}_0) \;\leq\; \delta. \end{array}$$

Assume that the construction is such that l and z satisfy the Lipschitz condition with the constant Λ.

In Section 3.1, the Constrained Level Method is briefly described, and the convergence results of Lemaréchal, Nemirovskii, and Nesterov are summed up. (We made minor modifications for the sake of simplicity.) In Section 3.2, we show how to handle inexact data.

3.1 Description of the Constrained Level Method

In this section, we assume that exact supporting hyperplanes can be computed.

Let F denote the optimal objective value of problem (7). If F is known in advance, then the quality of an approximate solution $\boldsymbol{x} \in X$ can be measured by $\varepsilon(\boldsymbol{x}) = \max\left(f(\boldsymbol{x}) - F,\, g(\boldsymbol{x}) \right)$. E.g., $\varepsilon(\boldsymbol{x}) = 0$ means that $\boldsymbol{x}$ is optimal. The original problem can be solved through minimizing the convex ε function over X.

Approximating models are used for f, g, and F. Suppose that we have generated the iterates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i$, and called an oracle to compute function values and subgradients at these points. The approximating model of f is again

$$f_i(\boldsymbol{x}) = \max_{1\le j\le i} \left\{ f(\boldsymbol{x}_j) + \left[\nabla f(\boldsymbol{x}_j)\right]^T (\boldsymbol{x} - \boldsymbol{x}_j) \right\} \qquad (\boldsymbol{x} \in X)$$

where f_i is a piecewise linear and convex function that inherits Lipschitz continuity from f. We have $f_i(\boldsymbol{x}_j) = f(\boldsymbol{x}_j)\ \ (j \le i)$ and $f_1 \le f_2 \le \ldots \le f_i \le f$.

Similarly, the approximating model of g is

$$g_i(\boldsymbol{x}) = \left[\max_{1\le j\le i} \left\{ g(\boldsymbol{x}_j) + \left[\nabla g(\boldsymbol{x}_j)\right]^T (\boldsymbol{x} - \boldsymbol{x}_j) \right\} \right]_+ \qquad (\boldsymbol{x} \in X)$$

where g_i is a piecewise linear and convex function that inherits Lipschitz continuity from g. We have $g_i(\boldsymbol{x}_j) = g(\boldsymbol{x}_j)\ \ (j \le i)$ and $0 \le g_1 \le g_2 \le \ldots \le g_i \le g$.

A lower approximation for F is

$$F_i = \min\left\{ f_i(\boldsymbol{x}) \mid \boldsymbol{x} \in X,\, g_i(\boldsymbol{x}) \le 0 \right\},$$

for which $F_1 \le F_2 \le \ldots \le F_i \le F$.

Let us introduce the notation

$$T_i = \left\{ \left(f(\boldsymbol{x}_j),\, g(\boldsymbol{x}_j) \right) \mid 1 \le j \le i \right\}, \quad C_i = \mathrm{Conv}(T_i),$$

and let

$$H_i = \min_{(u,v)\in C_i} \max(\, u - F_i,\, v\,).$$

Assume that $(u_i^\star, v_i^\star) \in C_i$ minimizes the above expression. That is, $H_i = \max(\, u_i^\star - F_i,\, v_i^\star\,)$, where

$$(u_i^\star, v_i^\star) = \sum_{j=1}^{i} r_j^i \left(f(\boldsymbol{x}_j),\, g(\boldsymbol{x}_j) \right)$$

is a convex combination. Let $\boldsymbol{x}_i^\star = \sum_{j=1}^{i} r_j^i \boldsymbol{x}_j$. Then, obviously $\boldsymbol{x}_i^\star \in X$, and from the convexity of the functions f and g, it follows that $\varepsilon(\boldsymbol{x}_i^\star) \le H_i$. Hence, the aim is to direct the search for new iterates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots$ in such a manner that the sequence $H_1, \ldots, H_i, \ldots$ decreases sufficiently.

We have

$$H_i = \min_{(u,v)\in C_i} \max(\, u - F_i,\, v\,) = \min_{(u,v)\in C_i} \max_{0\le\alpha\le 1} \alpha(u - F_i) + (1-\alpha)v.$$

Using von Neumann's lemma, we can continue:

$$\min_{(u,v)\in C_i} \max_{0\le\alpha\le 1} \alpha(u - F_i) + (1-\alpha)v = \max_{0\le\alpha\le 1} \min_{(u,v)\in C_i} \alpha(u - F_i) + (1-\alpha)v.$$

Introducing the notation

$$h_i(\alpha) = \min_{(u,v)\in C_i} \alpha(u - F_i) + (1-\alpha)v,$$

we have

$$H_i = \max_{0\le\alpha\le 1} h_i(\alpha). \tag{8}$$

Let us examine the function h_i. From the definition of C_i we have

$$h_i(\alpha) = \min_{1 \le j \le i} \alpha(f(\boldsymbol{x}_j) - F_i) + (1-\alpha) g(\boldsymbol{x}_j), \tag{9}$$

where h_i is a concave function, and $h_1 \ge h_2 \ge \ldots$.

α can be viewed as a dual variable. We are going to tune it perpetually to the current bundle. Taking into account (8), we are going to select $\alpha_i \in [0, 1]$ such that $h_i(\alpha_i)$ be 'close enough' to H_i (closeness is formulated in (10), later):

Let us consider the interval $I_i = [\underline{\alpha}_i, \bar{\alpha}_i] \subseteq [0, 1]$ on which h_i takes non-negative values. Let $\hat{I}_i \subset I_i$ be a sub-interval 'well inside' I_i. Formally, let

$$\hat{I}_i = \left[\underline{\alpha}_i + \frac{\mu}{2}|I_i|, \; \bar{\alpha}_i - \frac{\mu}{2}|I_i| \right],$$

where $|I_i|$ is the length of I_i, and $0 < \mu < 1$ is some preset parameter. Due to the concavity of h_i, it easily follows that

$$h_i(\alpha_i) \ge \frac{1}{2}\mu H_i \tag{10}$$

holds for any $\alpha_i \in \hat{I}_i$. Specifically, the selection is made as follows: For $i = 1$, let $\alpha_1 = \frac{1}{2}(\bar{\alpha}_1 - \underline{\alpha}_1)$. For $i > 1$, let

$$\alpha_i = \begin{cases} \alpha_{i-1}, & \text{if} \quad \alpha_{i-1} \in \hat{I}_i, \\ \frac{1}{2}(\bar{\alpha}_i - \underline{\alpha}_i), & \text{otherwise.} \end{cases} \tag{11}$$

Having generated $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i$, the next iterate $\boldsymbol{x}_{i+1}$ is the projection of $\boldsymbol{x}_i$ onto the level set

$$X_i = \left\{ \boldsymbol{x} \in X \;\middle|\; \alpha_i(f_i(\boldsymbol{x}) - F_i) + (1-\alpha_i) g_i(\boldsymbol{x}) \le \lambda h_i(\alpha_i) \right\}, \tag{12}$$

where $0 < \lambda < 1$ is some preset parameter. Due to the definition of F_i, the set X_i is not empty.

3.1.1 The Framework of the Constrained Level Method

Initialization.

Set the stopping tolerance $\epsilon > 0$.

Set the parameters λ and μ $\quad (0 < \lambda, \mu < 1)$.

Find a starting point $\boldsymbol{x}_1 \in X$.

Let $i := 1$ (iteration counter).

Update bundle.

Compute the function values $f(\boldsymbol{x}_i), g(\boldsymbol{x}_i)$ and subgradients $\nabla f(\boldsymbol{x}_i), \nabla g(\boldsymbol{x}_i)$.

Define the upper covers

$$\begin{aligned} f_i(\boldsymbol{x}) &:= \max_{1 \le j \le i} \left\{ f(\boldsymbol{x}_j) + [\nabla f(\boldsymbol{x}_j)]^T (\boldsymbol{x} - \boldsymbol{x}_j) \right\}, \\ g_i(\boldsymbol{x}) &:= \left[\max_{1 \le j \le i} \left\{ g(\boldsymbol{x}_j) + [\nabla g(\boldsymbol{x}_j)]^T (\boldsymbol{x} - \boldsymbol{x}_j) \right\} \right]_+ . \end{aligned}$$

Compute the lower approximation of the optimum

$$F_i := \min \{ f_i(\boldsymbol{x}) \,|\, \boldsymbol{x} \in X,\, g_i(\boldsymbol{x}) \le 0 \} \,.$$

Define the function $h_i(\alpha) := \min\limits_{1 \le j \le i} \alpha(f(\boldsymbol{x}_j) - F_i) + (1-\alpha) g(\boldsymbol{x}_j)$,

and compute its maximum $H_i := \max\limits_{0 \le \alpha \le 1} h_i(\alpha)$.

If $H_i < \epsilon$, then near-optimal solution found, stop.

Tune dual variable.

Determine the interval $I_i = [\underline{\alpha}_i, \bar{\alpha}_i] \subseteq [0,1]$ on which h_i takes non-negative values.

Compute α_i:

– for $i = 1$, let $\alpha_1 := \frac{1}{2}(\bar{\alpha}_1 - \underline{\alpha}_1)$,

– for $i > 1$, let $\alpha_i := \begin{cases} \alpha_{i-1}, & \text{if} \quad \underline{\alpha}_i + \frac{\mu}{2}|I_i| \le \alpha_{i-1} \le \bar{\alpha}_i - \frac{\mu}{2}|I_i|, \\ \frac{1}{2}(\bar{\alpha}_i - \underline{\alpha}_i), & \text{otherwise.} \end{cases}$

Find new iterate.

Define the level set

$$X_i := \left\{ \boldsymbol{x} \in X \,\middle|\, \alpha_i f_i(\boldsymbol{x}) + (1-\alpha_i) g_i(\boldsymbol{x}) \le \alpha_i F_i + \lambda h_i(\alpha_i) \right\},$$

and let $\boldsymbol{x}_{i+1}$ be the projection of $\boldsymbol{x}_i$ onto X_i.

Increment i.

$\rightarrow$ *Update bundle.*

3.1.2 Convergence Proofs.

Proposition 9 *Consider a sequence of iterations in the course of which α_i does not change. Namely, let p and q be natural numbers, $p < q$, and assume that $\alpha_i = \hat{\alpha}$ $(p \le i < q)$. Let ϵ be a small positive tolerance, and $c(\lambda)$ the constant defined in Theorem 3.*

If we have $q - p > c(\lambda) \left(\frac{\Lambda D}{\epsilon}\right)^2$, then $h_q(\alpha_q) \le \epsilon$ holds.

Proof. Let us introduce the following functions:

$$e = \hat{\alpha} f + (1-\hat{\alpha}) g, \qquad \text{and} \qquad e_i = \hat{\alpha} f_i + (1-\hat{\alpha}) g_i \quad (p \le i < q).$$

These are convex functions, and Lipschitz continuous with the constant Λ. We have $e_i(\boldsymbol{x}_j) = e(\boldsymbol{x}_j)$ $(j \le i)$ and $e_p \le e_{p+1} \le \ldots \le e_q \le e$.

For $p \le i < q$, let

$$L_i = \hat{\alpha} F_i, \quad U_i = \min_{1 \le j \le i} e(\boldsymbol{x}_j), \quad \text{and} \quad \Delta_i = U_i - L_i.$$

Substituting the above expressions into (9), we obtain $\Delta_i = h_i(\hat{\alpha})$. Further, substituting this into the expression (12), we obtain

$$X_i = \{ \, \boldsymbol{x} \in X \mid e_i(\boldsymbol{x}) \le L_i + \lambda \Delta_i \, \} \,.$$

For $p \leq i < q$, the next iterate $\boldsymbol{x}_{i+1}$ is the projection of $\boldsymbol{x}_i$ onto X_i. We have $L_p \leq \ldots \leq L_q \leq U_q \leq \ldots \leq U_p$. Hence, the gap is tightening at each step.

The arguments presented in course of the convergence proof of the Level Method can be applied to the functions e and e_i $(p \leq i \leq q)$ instead of f and f_i $(i = 1, 2, \ldots)$. The only difference from the original form of the Level Method is the special selection of the lower levels $L_i = \hat{\alpha} F_i$. It is easy to check that the arguments presented in Section 2.1.2 remain valid if L_i satisfies the following two requirements:

(1) $L_i \leq U_i$ must hold, and

(2) there must exist $\boldsymbol{u}_i \in X$ such that $e_i(\boldsymbol{u}_i) \leq L_i$.

In the present case, any point in $\arg\min\{f_i(\boldsymbol{x}) | \boldsymbol{x} \in X, g_i(\boldsymbol{x}) \leq 0\}$ is feasible for the second requirement. □

Corollary 10 *Consider a sequence of iterations in the course of which α_i does not change. Namely, let p and q be natural numbers, $p < q$, and assume that $\alpha_i = \hat{\alpha}$ $(p \leq i < q)$. Let ϵ be a small positive tolerance, and $c(\lambda)$ the constant defined in Theorem 3.*

If we have $q - p > c(\lambda) \left(\frac{2\Lambda D}{\mu\epsilon}\right)^2$, *then* $\varepsilon(\boldsymbol{x}_q^\star) \leq \epsilon$ *holds.*

Proof. Use (10) and $\varepsilon(\boldsymbol{x}_q^\star) \leq H_q$. □

Theorem 11 *Let $\epsilon > 0$. To obtain $\varepsilon(\boldsymbol{x}_i^\star) \leq \epsilon$, it suffices to perform*

$$c(\mu, \lambda) \left(\frac{2\Lambda D}{\epsilon}\right)^2 \ln\left(\frac{2\Lambda D}{\epsilon}\right)$$

iterations, where $c(\mu, \lambda)$ is a constant that depends only on μ and λ.

Proof. We have proved that $H_i \geq \varepsilon(\boldsymbol{x}_i^\star)$. We are going to prove that H_i decreases substantially every time α_i changes. According to (8), H_i is the maximum of h_i over $[0, 1]$.

First, we prove that h_i is Lipschitz continuous: Since f_i is Lipschitz continuous with the constant Λ, we have $|f_i(\boldsymbol{x}_j) - F_i| \leq \Lambda D$. Moreover, $f_i(\boldsymbol{x}_j) = f(\boldsymbol{x}_j)$ $(j \leq i)$, hence $|f(\boldsymbol{x}_j) - F_i| \leq \Lambda D$. Similarly, g is Lipschitz continuous with the constant Λ, hence $|g(\boldsymbol{x}_j)| \leq \Lambda D$. (We have assumed that g takes the value 0 as well as positive values.) It follows that h_i is Lipschitz continuous with the constant $2\Lambda D$.

Let us consider the interval $I_i = [\underline{\alpha}_i, \bar{\alpha}_i] \subseteq [0, 1]$ on which h_i takes non-negative values. If i is such that the strict inclusion $I_i \subset [0, 1]$ holds, then obviously, either $h_i(\underline{\alpha}_i) = 0$ or $h_i(\bar{\alpha}_i) = 0$ must hold. Hence,

$$2\Lambda D\, |I_i| \geq H_i \geq \varepsilon(\boldsymbol{x}_i^\star) \tag{13}$$

holds for such i. Since we have $h_1 \geq h_2 \geq \ldots$, it follows that $I_1 \supseteq I_2 \supseteq \ldots$.

Starting from $\boldsymbol{x}_1$, consider the maximal sequence of iterations

$$\boldsymbol{x}_1 \to \boldsymbol{x}_2\,,\ \ldots,\ \boldsymbol{x}_{s-1} \to \boldsymbol{x}_s, \tag{14}$$

along which α does nor change. That is, we have $\alpha_1 = \ldots = \alpha_{s-1}$, and $\alpha_{s-1} \neq \alpha_s$.

The above construction is repeated starting from $\boldsymbol{x}_{s+1}$. Thus, the iterations are grouped into maximal sequences along which α is unchanged. Assume there are k sequences. Denote by $I^{(\ell)}$ $(1 \leq \ell \leq k)$ the interval at the start of the ℓth sequence. (E.g., we have $I^{(1)} = I_1$, $I^{(2)} = I_{s+1}$.) Moreover, let $I^{(k+1)}$ denote the interval after the kth sequence.

We are going to prove that the sequence of the lengths $|I^{(\ell)}|$ $(1 \leq \ell \leq k)$ decreases at a geometric rate: Let $\alpha^{(\ell)}$ $(1 \leq \ell \leq k)$ denote the α value unchanged along the ℓth sequence. Obviously, $\alpha^{(\ell)}$ cuts $I^{(\ell)}$ into two sub-intervals of equal length. The construction (11) is such that one of these sub-intervals must contain $I^{(\ell+1)}$, with the possible exception of a segment shorter then $\frac{1}{2}\mu|I^{(\ell+1)}|$. Hence,

$$\frac{1}{2}\,|I^{(\ell)}| \ > \ \left(1 - \frac{1}{2}\mu\right)\,|I^{(\ell+1)}| \qquad (\ell = 1, \ldots, k).$$

Since we have $|I^{(1)}| \leq 1$, it follows that

$$\left(\frac{1}{2-\mu}\right)^k > |I^{(k+1)}|.$$

An obvious consequence of the above inequality is that strict inclusion $I^{(k+1)} \subset [0,1]$ holds for $k = 1, 2, \ldots$. Hence, we can use (13) for estimating the accuracy. After performing k sequences of iterations, the error $\varepsilon(\boldsymbol{x}^\star)$ is less than

$$2\Lambda D \left(\frac{1}{2-\mu}\right)^k .$$

Hence, after performing

$$\frac{1}{\ln(2-\mu)} \ln\left(\frac{2\Lambda D}{\epsilon}\right)$$

sequences of iterations, the error $\varepsilon(\boldsymbol{x}^\star)$ is less than ϵ.

Finally, if a sequence of iterations (14) is longer than $c(\lambda)\,\frac{1}{\mu^2}\,\left(\frac{2\Lambda D}{\epsilon}\right)^2$, then the error $\varepsilon(\boldsymbol{x}^\star)$ decreases under ϵ in the course of this sequence (Corollary 10). Hence if the accuracy of ϵ has not yet been reached, then the total number of the iterations performed cannot exceed

$$\frac{1}{\ln(2-\mu)} \ln\left(\frac{2\Lambda D}{\epsilon}\right) c(\lambda)\,\frac{1}{\mu^2} \left(\frac{2\Lambda D}{\epsilon}\right)^2 .$$

□

3.2 Handling Inexact Data in Constrained Minimization

As in the exact case, let F denote the optimal objective value of problem (7). If F is known in advance, then the quality of an approximate solution $\boldsymbol{x} \in X$ can be measured by $\varepsilon(\boldsymbol{x}) = \max\left(f(\boldsymbol{x}) - F,\, g(\boldsymbol{x})\right)$. E.g., $\varepsilon(\boldsymbol{x}) = 0$ means that $\boldsymbol{x}$ is optimal. The original problem can be solved through minimizing the convex function ε over X.

Approximating models are used for f, g, and F. Suppose that we have found the iterates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i$. At $\boldsymbol{x}_j$, we have constructed the approximate supporting hyperplanes l_j and z_j for the functions f and g, respectively, with the accuracy of δ_j. I.e., we have

$$\begin{array}{lll} f \geq l_j & \text{and} & f(\boldsymbol{x}_j) - l_j(\boldsymbol{x}_j) \leq \delta_j; \\ g \geq z_j & \text{and} & g(\boldsymbol{x}_j) - z_j(\boldsymbol{x}_j) \leq \delta_j \end{array} \qquad (1 \leq j \leq i).$$

Assume that each linear function l_j and z_j satisfies the Lipschitz condition with the constant Λ.

The ith approximating functions is defined as

$$f_i(\boldsymbol{x}) = \max_{1\leq j\leq i} l_j(\boldsymbol{x}), \quad g_i(\boldsymbol{x}) = \left[\max_{1\leq j\leq i} z_j(\boldsymbol{x})\right]_+ \qquad (\boldsymbol{x} \in X)\,.$$

(Hence, g_i is the upper cover of the linear functions $z_0, z_1, \ldots, z_i$, where z_0 denotes the constant function 0.) f_i and g_i are piecewise linear and convex functions that inherit Lipschitz continuity from the linear functions. Moreover, we have

$$\begin{array}{rcll} f_1 \leq f_2 \leq \ldots \leq f_i \leq f & \text{and} & f(\boldsymbol{x}_j) \leq l_j(\boldsymbol{x}_j) + \delta_j \leq f_i(\boldsymbol{x}_j) + \delta_j & (j \leq i), \\ 0 \leq g_1 \leq g_2 \leq \ldots \leq g_i \leq g & \text{and} & g(\boldsymbol{x}_j) \leq z_j(\boldsymbol{x}_j) + \delta_j \leq g_i(\boldsymbol{x}_j) + \delta_j & (j \leq i). \end{array} \tag{15}$$

A lower approximation for F is

$$F_i = \min\left\{\, f_i(\boldsymbol{x}) \mid \boldsymbol{x} \in X,\, g_i(\boldsymbol{x}) \leq 0 \,\right\},$$

for which $F_1 \leq F_2 \leq \ldots \leq F_i \leq F$.

Let us introduce the notation

$$T_i = \{\, (\, f_i(\boldsymbol{x}_j) + \delta_j,\, g_i(\boldsymbol{x}_j) + \delta_j \,) \mid 1 \leq j \leq i \,\}\,, \quad C_i = \mathrm{Conv}(T_i),$$

and let

$$H_i = \min_{(u,v)\in C_i} \max(\, u - F_i,\, v \,).$$

Assume that $(u_i^\star, v_i^\star) \in C_i$ minimizes the above expression. That is, $H_i = \max(\, u_i^\star - F_i,\, v_i^\star \,)$, where

$$(u_i^\star, v_i^\star) = \sum_{j=1}^{i} r_j^i \,(\, f_i(\boldsymbol{x}_j) + \delta_j,\, g_i(\boldsymbol{x}_j) + \delta_j \,)$$

is a convex combination. Let $\boldsymbol{x}_i^\star = \sum_{j=1}^{i} r_j^i \boldsymbol{x}_j$. Then, obviously, $\boldsymbol{x}_i^\star \in X$. Moreover, we have

$$f(\boldsymbol{x}_i^\star) \;\leq\; \sum_{j=1}^{i} r_j^i f(\boldsymbol{x}_j) \;\leq\; \sum_{j=1}^{i} r_j^i \,(f_i(\boldsymbol{x}_j) + \delta_j) \;=\; u_i^\star.$$

(We used the convexity of f, and the inequality (15).) Similarly, we have

$$g(\boldsymbol{x}_i^\star) \;\leq\; v_i^\star.$$

It follows that

$$\varepsilon(\boldsymbol{x}_i^\star) = \;\max\,(\, f(\boldsymbol{x}_i^\star) - F,\, g(\boldsymbol{x}_i^\star) \,) \;\leq\; \max\,(\, \boldsymbol{u}_i^\star - F_i,\, \boldsymbol{v}_i^\star \,) \;=\; H_i.$$

Hence, the aim is to direct the search for new iterates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i, \ldots$ in such a manner that the sequence $H_1, \ldots, H_i, \ldots$ decreases sufficiently.

We have

$$H_i = \min_{(u,v)\in C_i} \max(\, u - F_i,\, v \,) = \min_{(u,v)\in C_i} \max_{0\leq\alpha\leq 1} \alpha(u - F_i) + (1-\alpha)v.$$

Using von Neumann's lemma, we can continue:

$$\min_{(u,v)\in C_i} \max_{0\leq\alpha\leq 1} \alpha(u - F_i) + (1-\alpha)v = \max_{0\leq\alpha\leq 1} \min_{(u,v)\in C_i} \alpha(u - F_i) + (1-\alpha)v.$$

Introducing the notation

$$h_i(\alpha) = \min_{(u,v)\in C_i} \alpha(u - F_i) + (1-\alpha)v,$$

we have

$$H_i = \max_{0\leq\alpha\leq 1} h_i(\alpha). \tag{16}$$

Let us examine the function h_i. From the definition of C_i we have

$$h_i(\alpha) = \min_{1\leq j\leq i} \alpha(f_i(\boldsymbol{x}_j) - F_i) + (1-\alpha)g_i(\boldsymbol{x}_j) + \delta_j, \tag{17}$$

where h_i is a concave function, and $h_1 \geq h_2 \geq \ldots$.

α can be viewed as a dual variable. We are going to tune it perpetually to the current bundle. Taking into account (16), we are going to select $\alpha_i \in [0, 1]$ such that $h_i(\alpha_i)$ be 'close enough' to H_i (closeness is formulated in (18), later):

Let us consider the interval $I_i = [\underline{\alpha}_i, \bar{\alpha}_i] \subseteq [0, 1]$ on which h_i takes non-negative values. Let $\hat{I}_i \subset I_i$ be a sub-interval 'well inside' I_i. Formally, let

$$\hat{I}_i = \left[\, \underline{\alpha}_i + \frac{\mu}{2}|I_i|,\; \bar{\alpha}_i - \frac{\mu}{2}|I_i| \,\right],$$

where $|I_i|$ is the length of I_i, and $0 < \mu < 1$ is some preset parameter. Due to the concavity of h_i, it easily follows that

$$h_i(\alpha_i) \geq \frac{1}{2}\mu H_i \tag{18}$$

holds for any $\alpha_i \in \hat{I}_i$. Specifically, the selection is made as follows: For $i = 1$, let $\alpha_1 = \frac{1}{2}(\bar{\alpha}_1 - \underline{\alpha}_1)$. For $i > 1$, let

$$\alpha_i = \begin{cases} \alpha_{i-1}, & \text{if} \quad \alpha_{i-1} \in \hat{I}_i, \\ \\ \frac{1}{2}(\bar{\alpha}_i - \underline{\alpha}_i), & \text{otherwise.} \end{cases} \tag{19}$$

Having generated $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i$, the next iterate $\boldsymbol{x}_{i+1}$ is the projection of $\boldsymbol{x}_i$ onto the level set

$$X_i = \left\{ \boldsymbol{x} \in X \;\middle|\; \alpha_i(f_i(\boldsymbol{x}) - F_i) + (1 - \alpha_i)g_i(\boldsymbol{x}) \leq \lambda h_i(\alpha_i) \right\} \tag{20}$$

where $0 < \lambda < 1$ is some preset parameter. Due to the definition of F_i, the set X_i is not empty.

The accuracy to be prescribed for the approximate supporting hyperplane at the new iterate: Let $0 < \delta_1$ be selected arbitrarily. Let $\delta_{i+1} = \gamma\, h_i(\alpha_i) \quad (i = 1, 2, \ldots)$, where γ is a constant parameter satisfying $0 < \gamma < (1 - \lambda)^2$.

Remark 12 *The sequence $\delta_i \quad (i = 2, 3, \ldots)$ is not necessarily monotone decreasing, but the fluctuation is curbed by $H_i \geq h_i(\alpha_i) \geq \frac{1}{2}\mu H_i$.*

3.2.1 The Framework of the Inexact Constrained Level Method

Initialization.

Determine the Lipschitz constant Λ.

Set the stopping tolerance $\epsilon > 0$.

Set the parameters λ, μ, and $\gamma \quad (0 < \lambda, \mu < 1;\; 0 < \gamma < (1 - \lambda)^2)$.

Find a starting point $\boldsymbol{x}_1 \in X$.

Set the starting accuracy $\delta_1 > 0$.

Let $i := 1$ (iteration counter).

Update bundle.

At the point $\boldsymbol{x}_i$, construct the approximate support functions l_i and z_i for f and g, respectively, with the accuracy of δ_i. Moreover, l_i and z_i must satisfy the Lipschitz condition with the constant Λ.

Define the upper covers $f_i(\boldsymbol{x}) := \max\limits_{1 \leq j \leq i} l_j(\boldsymbol{x}), \quad g_i(\boldsymbol{x}) := \left[\max\limits_{1 \leq j \leq i} z_j(\boldsymbol{x})\right]_+$.

Compute the lower approximation of the optimum

$$F_i := \min \left\{ f_i(\boldsymbol{x}) \,|\, \boldsymbol{x} \in X,\, g_i(\boldsymbol{x}) \leq 0 \right\}.$$

Define the function $h_i(\alpha) := \min\limits_{1 \leq j \leq i} \; \alpha(f_i(\boldsymbol{x}_j) - F_i) + (1 - \alpha)g_i(\boldsymbol{x}_j) + \delta_j,$

and compute its maximum $H_i := \max_{0 \le \alpha \le 1} h_i(\alpha)$.

If $H_i < \epsilon$, then near-optimal solution found, stop.

Tune dual variable.

Determine the interval $I_i = [\underline{\alpha}_i, \bar{\alpha}_i] \subseteq [0,1]$ on which h_i takes non-negative values.

Compute α_i:

– for $i = 1$, let $\alpha_1 := \frac{1}{2}(\bar{\alpha}_1 - \underline{\alpha}_1)$,

– for $i > 1$, let $\alpha_i := \begin{cases} \alpha_{i-1}, & \text{if } \underline{\alpha}_i + \frac{\mu}{2}|I_i| \le \alpha_{i-1} \le \bar{\alpha}_i - \frac{\mu}{2}|I_i|, \\ \frac{1}{2}(\bar{\alpha}_i - \underline{\alpha}_i), & \text{otherwise.} \end{cases}$

Find new iterate.

Define the level set

$$X_i := \left\{ \boldsymbol{x} \in X \,\middle|\, \alpha_i f_i(\boldsymbol{x}) + (1-\alpha_i) g_i(\boldsymbol{x}) \le \alpha_i F_i + \lambda h_i(\alpha_i) \right\},$$

and let $\boldsymbol{x}_{i+1}$ be the projection of $\boldsymbol{x}_i$ onto X_i.

Let $\delta_{i+1} := \gamma h_i(\alpha_i)$.

Increment i.

$\rightarrow$ *Update bundle.*

3.2.2 Convergence Proofs.

Proposition 13 *Consider a sequence of iterations in the course of which α_i does not change. Namely, let p and q be natural numbers, $p < q$, and assume that $\alpha_i = \hat{\alpha}$ $(p \le i < q)$. Let ϵ be a small positive tolerance, and $c(\lambda, \gamma)$ the constant defined in Theorem 7.*

If we have $q - p > c(\lambda, \gamma) \left(\frac{\Lambda D}{\epsilon}\right)^2$, then $h_q(\alpha_q) \le \epsilon$ holds.

Proof. Let us introduce the following functions:

$$e = \hat{\alpha} f + (1-\hat{\alpha}) g, \qquad \text{and} \qquad e_i = \hat{\alpha} f_i + (1-\hat{\alpha}) g_i \quad (p \le i < q).$$

These are convex functions, and Lipschitz continuous with the constant Λ. We have $e_p \le e_{p+1} \le \ldots \le e_q \le e$. From (15), it follows that $e(\boldsymbol{x}_j) \le e_i(\boldsymbol{x}_j) + \delta_j$ $(j \le i)$.

For $p \le i < q$, let

$$L_i = \hat{\alpha} F_i, \quad U_i = \min_{1 \le j \le i} e_i(\boldsymbol{x}_j) + \delta_j, \quad \text{and} \quad \Delta_i = U_i - L_i.$$

Substituting the above expressions into (17), we obtain $\Delta_i = h_i(\hat{\alpha})$. Further, substituting this into the expression (20), we obtain

$$X_i = \{ \boldsymbol{x} \in X \mid e_i(\boldsymbol{x}) \le L_i + \lambda \Delta_i \}.$$

For $p \le i < q$, the next iterate $\boldsymbol{x}_{i+1}$ is the projection of $\boldsymbol{x}_i$ onto X_i. We have $L_p \le \ldots \le L_q \le U_q \le \ldots \le U_p$. Hence, the gap is tightening at each step.

The arguments presented in course of the convergence proof of the inexact Level Method can be applied to the functions e and e_i $(p \le i \le q)$ instead of f and f_i $(i = 1, 2, \ldots)$. The only difference from the original form of the inexact Level Method is the special selection of the lower levels $L_i = \hat{\alpha} F_i$. It is easy to check that the arguments presented in Section 2.2.2 remain valid if L_i satisfies the following two requirements:

(1) $L_i \leq U_i$ must hold, and

(2) there must exist $\boldsymbol{u}_i \in X$ such that $e_i(\boldsymbol{u}_i) \leq L_i$.

In the present case, any point in $\arg\min\{f_i(\boldsymbol{x})|\boldsymbol{x} \in X, g_i(\boldsymbol{x}) \leq 0\}$ is feasible for the second requirement. $\square$

Applying the arguments presented in Section 3.1.2 to the result of Proposition 13 instead of to that of Proposition 9, we obtain

Theorem 14 *Let $\epsilon > 0$. To obtain $\varepsilon(\boldsymbol{x}_i^\star) \leq \epsilon$, it suffices to perform*

$$c(\mu, \lambda, \gamma) \left(\frac{2\Lambda D}{\epsilon}\right)^2 \ln\left(\frac{2\Lambda D}{\epsilon}\right)$$

iterations, where $c(\mu, \lambda, \gamma)$ is a constant that depends only on μ, λ, and γ.

4 Summary and Conclusions

The convergence results presented in this paper describe the data exchange between the convex optimization method and the *oracle* that returns function data related to a given point. (Nemirovsky and Yudin (1983) initiated research into this direction. See Nesterov (1997) for a state-of-art summary.) Stochastic objective and constraint functions are often very difficult to compute. Hence, the number of the function evaluations required is the crucial measure of the effectiveness of solution methods devised for such problems.

Another characteristic of stochastic problems is that estimate function data can often be computed effectively, but the more accurate data we want, the greater the computational effort. Using the inexact versions of the Level Method or the Constrained Level Method, one can keep balance between the amount of work invested into estimating the function data on the one hand, and into the optimization method on the other hand. (A good example is when the functions are integral functions of high dimensions: In order to get more and more accurate data, the grid must be refined constantly. Remarks 5 and 12 show that the inexact methods use more and more expensive data economically.)

It may happen that the optimization procedure must be terminated prematurely, and the prescribed accuracy cannot be reached due to time limits. Level-type methods provide usable (though less accurate) solutions even in such cases. The inexact level-type methods, moreover, use resources economically even in such cases.

Acknowledgement

I wish to express my thanks to András Prékopa for his suggestion that bundle methods should be examined from a stochastic programming point of view, and for his support in the project.

References

[1] Hiriart-Urruty, J-B. and C. Lemaréchal. (1993). *Convex Analysis and Minimization Algorithms.* Volume II: *Advanced Theory and Bundle Methods.* Springer-Verlag, Berlin.

[2] Kelley, J.E. (1960). The Cutting Plane Method for Solving Convex Programs. *Journal of the SIAM* **8**, 703-712.

[3] Lemaréchal, C., A. Nemirovskii, and Yu. Nesterov. (1995). New Variants of Bundle Methods. *Mathematical Programming* **69**, 111-147.

[4] Nemirovsky, A., and D. Yudin. (1983). *Informational Complexity and Efficient Methods for Solution of Convex Extremal Problems.* Wiley, New York.

[5] Nesterov, Yu. (1997). Interior-point Methods: An Old and a New Approach to Nonlinear Programming. *Mathematical Programming* **79**, 285-297.

[6] Rockafellar, R.T. (1970). *Convex Analysis*. Princeton University Press, Princeton.

[7] Szántai, T. (1988). A Computer Code for the Solution of Probabilistic-constrained Stochastic Programming Problems. In *Numerical Techniques for Stochastic Optimization*, Yu. Ermoliev and R.J-B. Wets (eds.). Springer-Verlag, New York.