

CORE DISCUSSION PAPER

2008/83

# Approximate level method

Peter RICHTÁRIK<sup>1</sup>

December 2008

## Abstract

In this paper we propose and analyze a variant of the *level method* [4], which is an algorithm for minimizing nonsmooth convex functions. The main work per iteration is spent on 1) minimizing a piecewise-linear model of the objective function and on 2) projecting onto the intersection of the feasible region and a polyhedron arising as a level set of the model. We show that by replacing exact computations in both cases by *approximate computations*, in *relative scale*, the theoretical iteration complexity increases only by the factor of four. This means that while spending less work on the subproblems, we are able to retain the good theoretical properties of the level method.

**Keywords:** level method, approximate projections in relative scale, nonsmooth convex optimization, sensitivity analysis, large-scale optimization.

---

<sup>1</sup>Université catholique de Louvain, Center for Operations Research and Econometrics (CORE) and Department of Mathematical Engineering (INMA), B-1348 Louvain-la-Neuve, Belgium. E-mail: peter.richtarik@uclouvain.be  
The research results presented in this paper have been supported by a grant “Action de recherche concertée ARC 04/09-315” from the “Direction de la recherche scientifique - Communauté française de Belgique”.

This paper presents research results of the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister’s Office, Science Policy Programming. The scientific responsibility is assumed by the authors.

# 1 Introduction

**Motivation.** In this paper we consider the basic convex optimization problem of the form

$$f^* = \min_{x \in \mathcal{Q}} f(x), \quad (1.1)$$

where  $\mathcal{Q} \subset \mathbf{R}^n$  is a compact convex set and  $f$  is Lipschitz continuous with  $\mathcal{Q} \subseteq \text{dom } f$ . We will assume that all information available to us about  $f$  is given by a first-order oracle. That is, for all feasible points  $x$  we have access to  $f(x)$  and  $f'(x)$  only, the latter being an arbitrary subgradient of  $f$  at  $x$ . Having collected this information about  $f$  for points  $x_0, \dots, x_k \in \mathcal{Q}$ , it is natural to condense it into the following single object:

$$\hat{f}_k(x) \stackrel{\text{def}}{=} \max_{0 \leq i \leq k} f(x_i) + \langle f'(x_i), x - x_i \rangle. \quad (1.2)$$

Note that  $\hat{f}_k$  is a piecewise-linear and convex *model* of  $f$ , always underestimating it.

There are several approaches in the literature for exploiting this object to design algorithmic schemes for solving (1.1). In *Kelly's method* [1], for example, the next iterate  $x_{k+1}$  is chosen to be simply a minimizer of the model function. It is known, however, that this strategy leads to an unstable method with bad practical and theoretical performance. In fact, simple examples can be constructed for which the number of iterations needed by Kelly's method is exponential in the dimension (see Section 3.3.2 in [5]). Several versions of *bundle methods* [3], [2], on the other hand, pick  $x_{k+1}$  to be the minimizer of the model function penalized by a simple quadratic of the form  $\frac{1}{2}\lambda_k\|x - u_k\|^2$ , where  $\lambda_k$  is the current "penalty parameter" and  $u_k$  the current "prox-center". It appears that finding good updating strategies for the former is not as easy as for the latter. *Level method*, developed by Lemaréchal, Nemirovski and Nesterov [4], sets the next iterate to be the *exact* projection of the current point  $x_k$  onto a certain level set of the model function. The level value is chosen to be smaller than the best of the function values observed so far but higher than the minimum of the model (setting it equal to this minimum corresponds to Kelly's strategy), which also has to be computed *exactly*. It turns out that the level value can be updated in a very simple way, as a fixed convex combination of the two changing bounds mentioned above. As a consequence, the method depends only on the choice of a single parameter. One of the effects of this approach is that of stabilizing Kelly's idea in practice. Also, the theoretical complexity no longer depends on the dimension of the problem. In fact, in order to produce an  $\varepsilon$ -minimizer of (1.1), it suffices to take

$$N \leq \frac{4L^2D^2}{\varepsilon^2} \quad (1.3)$$

iterations, where  $D = \text{Diam}(\mathcal{Q})$  and  $L$  is the Lipschitz constant of  $f$ . It appears that this complexity is optimal, uniformly in the dimension. Although this is also the case, for example, with the simple but practically inefficient *subgradient method* [7], level method is much better in practice.

**Contribution.** The main work at every iteration of the level method is spent on 1) minimizing a piecewise-linear model of the objective function and on 2) projecting onto the level set of the model. In this paper we show that by replacing exact computations in both cases by

*approximate computations*, in *relative scale* (in a certain sense which will be precisely defined later), the theoretical iteration complexity (1.3) increases only by the factor of four. This means that while spending less work on the subproblems, the new approach still retains the good theoretical guarantees of the level method.

We show that for the first subproblem, a precision proportional to the current gap and *independent* of the target accuracy  $\varepsilon$  of the master convex problem is completely satisfactory (see Subsection 2.1). In a certain sense this is to be expected as the computed minimum enters the algorithm only through the level value, which can be set to *any*, albeit fixed, convex combination of the the minimum and the best upper bound. For the second subproblem, our analysis requires that the projections be made with relative accuracy

$$\rho = \frac{w^2}{2w + 1}, \quad \text{where} \quad w = \mathcal{O}\left(\frac{\varepsilon^2}{L^2 D^2}\right).$$

Observe that  $\rho = \mathcal{O}(\varepsilon^4/L^4 D^4)$ .

**Contents.** The paper is organized as follows. In Section 2 we give a brief formal description of our version of the level method. In Section 3 we study approximate projections and derive a technical inequality which will be used in the iteration complexity analysis, contained in Section 4. Finally, in Section 5 we comment on possible approaches to finding approximate solutions to the two subproblems.

**Notation.** We treat vectors of  $\mathbf{R}^n$  as column vectors and the entries of  $x \in \mathbf{R}^n$  are denoted by  $x = (x^{(1)}, \dots, x^{(n)})^T$ . For  $x$  and  $y$  in  $\mathbf{R}^n$ ,  $\langle x, y \rangle$  is the standard inner product:

$$\langle x, y \rangle = \sum_{i=1}^n x^{(i)} y^{(i)} = x^T y.$$

By  $\|x\|$  we denote the standard Euclidean norm of vector  $x \in \mathbf{R}^n$ , i.e.  $\|x\| = \langle x, x \rangle^{1/2}$ . More notation will be introduced at the spot in text where needed.

## 2 Approximate level method

For a sequence of points  $\{x_i\}$  in  $\mathcal{Q}$ , let us denote the *minimal value* of the model (1.2), resp. *record value* of the objective function, by

$$\hat{f}_k^* = \min_{x \in \mathcal{Q}} \hat{f}_k(x), \quad \text{resp.} \quad f_k^* = \min_{0 \leq i \leq k} f(x_i). \quad (2.1)$$

Note that the following relations hold for all  $k$ :

$$\hat{f}_k \leq f, \quad \hat{f}_k \leq \hat{f}_{k+1}, \quad \text{and} \quad \hat{f}_k^* \leq \hat{f}_{k+1}^* \leq f^* \leq f_{k+1}^* \leq f_k^*. \quad (2.2)$$

The first inequality states that the model always underestimates  $f$ , the second says that the model grows as we add new cutting-planes to it. Observe that due to the last set of inequalities we know that the quantity  $\delta_k = f_k^* - \hat{f}_k^*$  is decreasing, and that we can stop once it gets bellow the target accuracy  $\varepsilon$ .

Level method at every iteration solves two subproblems: 1) minimization of the model function and 2) Euclidean projection onto a certain level set of the model function. In the following two subsections we formally describe acceptable approximate solutions of these two subproblems and then proceed with proposing our algorithm. We postpone the question of *how* to obtain these approximate solutions in practice until Section 5.

## 2.1 Minimizing the model

We will assume that the minimal value of the model function is at every iteration computed only *approximately* in the following sense. We fix a parameter  $0 \leq \gamma < 1$  and obtain a point  $x_k^* \in \mathcal{Q}$  such that

$$\tilde{f}_k^* \stackrel{\text{def}}{=} \hat{f}_k(x_k^*) \leq (1 - \gamma)\hat{f}_k^* + \gamma f_k^*. \quad (2.3)$$

The choice  $\gamma = 0$ , which is the case in the level method, corresponds to finding the exact minimizer. Note that if we define  $\tilde{\delta}_k = f_k^* - \tilde{f}_k^*$ , then condition (2.3) is equivalent to (the first inequality in)

$$(1 - \gamma)\delta_k \leq \tilde{\delta}_k \leq \delta_k. \quad (2.4)$$

This means that the point  $x_k^*$  approximately “closes the gap”  $\delta_k$ , in *relative scale*, with accuracy governed by the parameter  $\gamma$ . The true gap  $\delta_k$  is assumed to be hard to compute, while the approximate gap  $\tilde{\delta}_k$  is thought to be easier to obtain.

Note that the inequality

$$\tilde{\delta}_k \leq (1 - \gamma)\varepsilon \quad (2.5)$$

implies  $f_k^* - f^* \leq \delta_k \stackrel{(2.4)}{\leq} \tilde{\delta}_k / (1 - \gamma) \stackrel{(2.5)}{\leq} \varepsilon$ , and hence it is a good stopping criterion for our method.

The following relations will be useful later in the analysis

$$(1 - \gamma)\tilde{\delta}_i \leq \tilde{\delta}_k, \quad i > k. \quad (2.6)$$

To see why (2.6) holds, it suffices to notice that

$$\tilde{\delta}_i \stackrel{(2.4)}{\leq} \delta_i \leq \delta_k \leq \tilde{\delta}_k / (1 - \gamma).$$

## 2.2 Projection subproblem

Further, we choose a *level parameter*  $0 < \alpha < 1$ , define the *level value* by

$$l_k(\alpha) = (1 - \alpha)\tilde{f}_k^* + \alpha f_k^*,$$

and consider the *level set*

$$\mathcal{L}_k(\alpha) = \{x \in \mathcal{Q} : \hat{f}_k(x) \leq l_k(\alpha)\}.$$

Note that level method uses  $\hat{f}_k^*$  instead of  $\tilde{f}_k^*$  in the definition of the level value (which corresponds to the  $\gamma = 0$  choice). The next iterate  $x_{k+1}$  will be chosen as an approximate Euclidean projection, in relative scale, of the previous iterate  $x_k$  onto the level set. Level method instead works with exact projections. Let us define the concept more formally.

**Definition 1 (Approximate projection)** *Let  $C$  be a convex set,  $x \notin C$  and  $z \in C$ . We say that  $z$  is a  $\rho$ -approximate projection of  $x$  onto  $C$  if*

$$\|x - z\|^2 \leq (1 + \rho) \inf_{y \in C} \|x - y\|^2. \quad (2.7)$$

## 2.3 Algorithm

Figure 1 summarizes the input data and Figure 2 lists the parameters defining our method.

object	meaning
$f$	objective function
$\mathcal{Q}$	feasible set
$L$	Lipschitz constant of $f$
$D$	(upper bound on the) diameter of $\mathcal{Q}$
$x_0$	an initial feasible point
$\varepsilon$	target accuracy of the master problem (1.1)

Figure 1: Input data.

parameter	meaning
$\alpha$	parameter defining the level set
$\beta$	a correction parameter (can be set to $\beta = 1$ if $\gamma > 0$ )
$\gamma$	relative accuracy with which we minimize the model
$\rho$	relative accuracy with which we compute projections

Figure 2: Parameters of the algorithm.

Our variant of the level method for solving problem (1.1) is the following.

<b>Approximate level method</b>	
(1) <b>Input:</b>	$f, \mathcal{Q}, L, D, x_0, \varepsilon > 0$
(1) <b>Choice of parameters:</b>	$0 < \alpha < 1$ and $0 < \gamma < 1, \beta = 1$ or $\gamma = 0, \beta > 1$
(3) <b>Preprocessing:</b>	Set projection accuracy to $\rho = \frac{\omega^2}{2\omega+1}$ where $\omega = \frac{(1-\gamma)^4(1-\alpha)^2\varepsilon^2}{\beta L^2 D^2}$
(4) <b>For <math>k \geq 0</math> iterate:</b>	<ul style="list-style-type: none"> <li>(a) Compute <math>f_k^*</math> and <math>\tilde{f}_k^*</math> and set <math>\tilde{\delta}_k = f_k^* - \tilde{f}_k^*</math></li> <li>(b) STOP if <math>\tilde{\delta}_k \leq (1 - \gamma)\varepsilon</math></li> <li>(c) Compute <math>x_{k+1}</math> as an <math>\rho</math>-approximate projection of <math>x_k</math> onto <math>\mathcal{L}_k(\alpha)</math></li> </ul>

A reasonable choice of the parameters is  $\alpha = \gamma = 1 - \frac{1}{\sqrt{2}}$  and  $\beta = 1$ . The argument leading to this choice follows from the complexity estimate given in Theorem 2.

### 3 Approximate projection inequality

The analysis of the level method applied to problem (1.1) (Section 2.2.1 in [5] or Section 3.3.3 in [5]) makes use of the first-order necessary optimality conditions for the projection subproblem. The projection point has to be exact for the analysis to go through. In this section we will construct optimality conditions that hold at an approximate minimizer, i.e. an approximate projection point. This leads to a relaxed inequality that can be successfully substituted into the original analysis, yielding the desired sensitivity result.

The main goal of this section is to show that condition (2.7) implies an inequality of the form

$$\|x - z\|^2 + \|z - y\|^2 \leq (1 + \omega)\|x - y\|^2, \quad y \in C,$$

for certain  $\omega = \omega(\rho)$ . In the case  $\rho = 0$ , we can choose  $\omega = 0$ , which follows from the first order necessary conditions for the projection problem. Our goal is to generalize this for positive values of  $\rho$ .

To make the exposition in the rest of this section lighter, it will be useful to establish some notation. For vector  $x \in \mathbf{R}^n$  and a scalar  $r$  denote

$$\begin{aligned} \mathcal{B}(x, r) &= \{s : \|s - x\| \leq r\}, \\ \partial\mathcal{B}(x, r) &= \{s : \|s - x\| = r\}, \\ \mathcal{H}(x) &= \{s : \langle s, x \rangle \leq 0\}, \text{ and} \\ \partial\mathcal{H}(x) &= \{s : \langle s, x \rangle = 0\}. \end{aligned}$$

We will use this full notation in the statements of the theorems and resort to the simpler form  $\mathcal{B}, \partial\mathcal{B}, \mathcal{H}$  and  $\partial\mathcal{H}$  in the proofs.

In our first lemma we compute the optimal value of the problem

$$p^* = p^*(x, r, y) \stackrel{\text{def}}{=} \max\{\|z - y\|^2 : z \in \mathcal{B}(x, r) \cap \mathcal{H}(x)\}, \quad (3.1)$$

for a triple  $(x, r, y)$  satisfying a certain condition.

**Lemma 1** *Fix  $0 \neq x \in \mathbf{R}^n, r > \|x\|$  and  $y \in \mathcal{H}(x)$ . Let  $\hat{y}$  denote the projection of  $y$  onto  $\partial\mathcal{H}(x)$ , and  $R = \sqrt{r^2 - \|x\|^2}$ . Then*

$$p^*(x, r, y) = R^2 + \|y\|^2 + 2R\|\hat{y}\|. \quad (3.2)$$

**Proof:**

First notice that the objective function can be written as

$$\begin{aligned} \|z - y\|^2 &= \|z - x\|^2 + 2\langle z - x, x - y \rangle + \|x - y\|^2 \\ &= \|z - x\|^2 + 2\langle z, x - y \rangle - \|x\|^2 + \|y\|^2. \end{aligned} \quad (3.3)$$

Case 1. Assume  $\hat{y} = 0$ ; that is,  $y = tx$  for some  $t \leq 0$ . In this case  $\langle z, y \rangle \geq 0$  for all  $z \in \mathcal{H}$ . Therefore, in view of (3.3), all feasible points  $z$  satisfy

$$\|z - y\|^2 \leq r^2 - \|x\|^2 + \|y\|^2,$$

with equality precisely when  $z \in \partial\mathcal{B} \cap \partial\mathcal{H}$ . Notice that this is in agreement with (3.2).

Case 2. Assume  $\hat{y} \neq 0$ . It follows from (3.3) that if all optimal solutions  $z^*$  of

$$q^* = \max_{\substack{\|z-x\|^2 \leq r^2 \\ \langle z, x \rangle \leq 0}} \langle x - y, z \rangle, \quad (3.4)$$

satisfy  $\|z^* - x\| = r$ , then

$$p^* = r^2 + 2q^* - \|x\|^2 + \|y\|^2. \quad (3.5)$$

Indeed, we will show that the Lagrange multiplier  $\lambda$  at any optimal point  $z^*$  of (3.4) corresponding to the first inequality is positive, and hence  $\|z^* - x\| = r$ . The Lagrangean dual of (3.4) is

$$q^* = \min_{\lambda, \mu \geq 0} \Phi(\lambda, \mu),$$

where

$$\Phi(\lambda, \mu) = \begin{cases} \infty & \text{if } \lambda = 0, y + (\mu - 1)x \neq 0 \\ 0 & \text{if } \lambda = 0, y + (\mu - 1)x = 0 \\ \frac{1}{4\lambda} \|y - (2\lambda + 1 - \mu)x\|^2 + \lambda R^2 & \text{if } \lambda \neq 0. \end{cases}$$

Since we assume that  $\hat{y} \neq 0$ , we cannot have  $y + (\mu - 1)x = 0$  for any  $\mu$  and hence the optimal  $\lambda$  must be positive. Note that for any fixed  $\lambda > 0$ , the value of  $\Phi(\lambda, \mu)$  is minimized with  $\mu$  such that  $y - (2\lambda + 1 - \mu)x = \hat{y}$ . Hence we can instead solve the following one-dimensional convex problem:

$$q^* = \min_{\lambda > 0} \frac{1}{4\lambda} \|\hat{y}\|^2 + \lambda R^2. \quad (3.6)$$

Its minimizer is  $\lambda^* = \frac{\|\hat{y}\|}{2R}$  and substituting this into (3.6) and  $q^*$  into (3.5) gives (3.2).  $\square$

The main result of this section is a simple consequence of the following lemma.

**Lemma 2** *Let  $0 \neq x \in \mathbf{R}^n$  and  $\rho \geq 0$ . Then for  $r^2 = (1 + \rho)\|x\|^2$  and*

$$\omega = \rho + \sqrt{\rho^2 + \rho}, \quad (3.7)$$

*we have the following inequality*

$$\|x - z\|^2 + \|z - y\|^2 \leq (1 + \omega)\|x - y\|^2, \quad y \in \mathcal{H}(x), \quad z \in \mathcal{H}(x) \cap \mathcal{B}(x, r). \quad (3.8)$$

**Proof:**

Fixing arbitrary  $y \in \mathcal{H}$ , Lemma 1 implies that

$$\begin{aligned} \max_{z \in \mathcal{H} \cap \mathcal{B}} \|x - z\|^2 + \|z - y\|^2 &\leq \max_{z \in \mathcal{H} \cap \mathcal{B}} \|x - z\|^2 + \max_{z \in \mathcal{H} \cap \mathcal{B}} \|z - y\|^2 \\ &= r^2 + (R^2 + \|y\|^2 + 2R\|\hat{y}\|) \\ &= 2r^2 - \|x\|^2 + \|y\|^2 + 2\rho^{1/2}\|x\|\|\hat{y}\|. \end{aligned}$$

It thus remains to argue that the last expression is upper-bounded by  $(1 + \omega)\|x - y\|^2$ . A straightforward substitution and simplification yields the following equivalent inequality:

$$(\omega - 2\rho)\|x\|^2 - 2\rho^{1/2}[\|x\|^2\|y\|^2 - \langle x, y \rangle^2]^{1/2} + \omega\|y\|^2 - 2(1 + \omega)\langle x, y \rangle \geq 0. \quad (3.9)$$

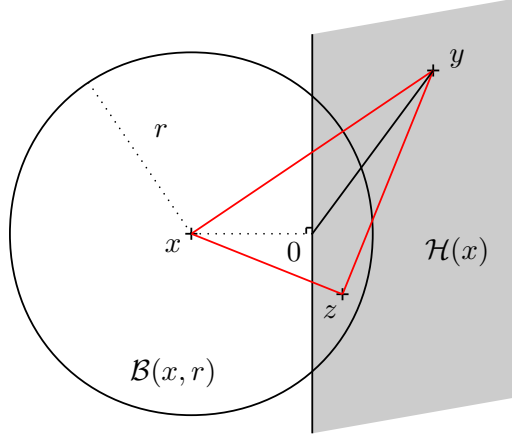


Figure 3: Lemma 2 – approximate projection.

Notice, however, that for  $\omega \geq 2\rho$  we have

$$(\omega - 2\rho)\|x\|^2 - 2\sqrt{\omega(\omega - 2\rho)}\|x\|\|y\| + \omega\|y\|^2 = \left(\sqrt{\omega - 2\rho}\|x\| - \sqrt{\omega}\|y\|\right)^2 \geq 0$$

Notice that this inequality is *stronger* than (3.9), provided that  $\omega \geq 2\rho$  and  $\sqrt{\omega(\omega - 2\rho)} \geq \rho^{1/2}$ . Solving for  $\omega$  in terms of  $\rho$  in the latter quadratic gives (3.7).  $\square$

Note that for  $\rho \leq 1$  we have the estimate  $\rho + \sqrt{\rho^2 + \rho} \leq \sqrt{\rho} + \sqrt{\rho + \rho}$ , and hence we can replace (3.7) by

$$\omega = (\sqrt{2} + 1)\rho^{1/2}. \quad (3.10)$$

On the other hand, if  $\rho > 1$ , we have  $\rho + \sqrt{\rho^2 + \rho} \leq \rho + \sqrt{2\rho^2}$ , and so we can replace (3.7) by

$$\omega = (\sqrt{2} + 1)\rho. \quad (3.11)$$

**Theorem 1 (Approximate projection inequality)** *Let  $C$  be a convex set and  $x$  a point not lying in this set. If  $z \in C$  is an  $\rho$ -approximate projection of  $x$  onto  $C$  and  $\omega = \omega(\rho)$  is given by (3.7) (or (3.10) if  $\rho \leq 1$  or (3.11) if  $\rho > 1$ ), then*

$$\|x - z\|^2 + \|z - y\|^2 \leq (1 + \omega)\|x - y\|^2, \quad y \in C.$$

**Proof:**

By appropriate shifting we can wlog assume that the projection point is the origin. We now apply Lemma 2 and note that  $C \subset H$  since  $\partial H$  is a supporting hyperplane to  $C$  at the origin.

## 4 Complexity analysis

In this section we modify the analysis of the level method by replacing exact minimization of the model by *approximate minimization* and exact projection onto the level set by *approximate projection*, as described in the previous section.

Lemma 3 says that if the values of the (presumably easily computable) gap  $\tilde{\delta}_i$ , for  $i = k, \dots, p$ , stay above a certain fraction of the “initial” value  $\tilde{\delta}_k$ , i.e. if there is not enough progress from iteration  $k$  to iteration  $p$ , then the point  $x_p^*$  must necessarily lie in the intersection of the level sets  $\mathcal{L}_i(\alpha)$  for  $i = k, \dots, p$ . This property will be exploited in Lemma 5, which is in turn used in the proof of our main result.

**Lemma 3 (cf. Lemma 3.3.1, [5])** *If for  $i = k, \dots, p$  we have  $\tilde{\delta}_p \geq (1 - \alpha)\tilde{\delta}_i$ , then*

$$x_p^* \in \bigcap_{i=k}^p \mathcal{L}_i(\alpha).$$

**Proof:**

For such  $i$  we get

$$\hat{f}_i(x_p^*) \leq \hat{f}_p(x_p^*) = \tilde{f}_p^* = f_p^* - \tilde{\delta}_p \leq f_p^* - (1 - \alpha)\tilde{\delta}_i \leq f_i^* - (1 - \alpha)\tilde{\delta}_i \leq l_i(\alpha).$$

□

The statement and proof is analogous to that of Lemma 3.3.1 in [5], which is formulated with exact gaps  $\delta_i$  instead. The latter is thus recovered as a special case with  $\gamma = 0$ .

Note that the Lipschitz constant  $L$  of  $f$  is an upper bound on the norms of all subgradients of  $f$  evaluated at points of  $\mathcal{Q}$ . The following result says that if the current gap is large, then the size of the next step will also be large.

**Lemma 4 (cf. Lemma 3.3.2, [5])** *If  $\{x_k\}$  is a sequence of points generated by the level method, then*

$$\|x_{k+1} - x_k\| \geq \frac{(1 - \alpha)\tilde{\delta}_k}{L}.$$

**Proof:**

Indeed,

$$\begin{aligned} f(x_k) - (1 - \alpha)\tilde{\delta}_k &\geq f_k^* - (1 - \alpha)\tilde{\delta}_k \\ &= l_k(\alpha) \\ &\geq \hat{f}_k(x_{k+1}) \\ &\geq f(x_k) + \langle f'(x_k), x_{k+1} - x_k \rangle \\ &\geq f(x_k) - L\|x_{k+1} - x_k\|. \end{aligned}$$

□

Boundedness of the feasible set  $\mathcal{Q}$  is only needed in the last two results.

**Lemma 5** *Let  $\omega$  be chosen as in Theorem 1. If for some  $p \geq k$  we have*

$$\tilde{\delta}_p > \frac{\sqrt{\omega}LD}{(1-\alpha)(1-\gamma)}, \quad \text{and} \quad \tilde{\delta}_p \geq (1 - \alpha)\tilde{\delta}_i, \quad i = k, \dots, p,$$

then

$$p - k + 1 \leq \frac{L^2 D^2}{(1 - \alpha)^2 (1 - \gamma)^2 \tilde{\delta}_p^2 - \omega L^2 D^2}.$$

**Proof:**

In view of Lemma 3, point  $x_p^*$  lies in  $\mathcal{L}_i(\alpha)$  for all  $i = k, \dots, p$ . We can therefore individually for each  $i$  use Theorem 1 with  $x = x_i$ ,  $C = \mathcal{L}_i(\alpha)$ ,  $z = x_{i+1}$  and  $y = x_p^*$ . This together with Lemma 4 and the inequality  $\tilde{\delta}_i \geq (1 - \gamma)\tilde{\delta}_p$  (see (2.6)) implies

$$\|x_{i+1} - x_p^*\|^2 \leq (1 + \omega)\|x_i - x_p^*\|^2 - \|x_{i+1} - x_i\|^2 \leq (1 + \omega)\|x_i - x_p^*\|^2 - \frac{(1 - \alpha)^2(1 - \gamma)^2\tilde{\delta}_p^2}{L^2}.$$

After rearranging the terms and summing up these inequalities for  $i = k, \dots, p$  we get

$$(p - k + 1)\frac{(1 - \alpha)^2(1 - \gamma)^2\tilde{\delta}_p^2}{L^2} - \omega \sum_{i=k}^p \|x_i - x_p^*\|^2 \leq \|x_k - x_p\|^2.$$

The result now easily follows by replacing the norms in the last expression by  $D$ .  $\square$

This is the main result of this paper.

**Theorem 2** *Let  $\varepsilon, \alpha, \beta, \gamma$  be such that  $\varepsilon > 0 < \alpha < 1$ ,  $0 \leq \gamma < 1$  and  $\beta \geq 1$  ( $\beta > 1$  if  $\gamma = 0$ ), and let*

$$\rho = \frac{\omega^2}{2\omega + 1} \quad \text{with} \quad \omega = \frac{(1 - \alpha)^2(1 - \gamma)^4\varepsilon^2}{\beta L^2 D^2}. \quad (4.1)$$

*Then the level method with  $\rho$ -approximate projections produces an  $\varepsilon$ -approximate minimizer of problem (1.1) after no more than*

$$N = \left\lceil \frac{\beta L^2 D^2}{(1 - \gamma)^2(\beta - (1 - \gamma)^2)\alpha(1 - \alpha)^2(2 - \alpha)\varepsilon^2} \right\rceil \quad (4.2)$$

*iterations.*

**Proof:**

The proof closely follows that of Theorem 3.3.1 in [5] with the exception that we use Lemma 5 instead of Lemma 3.3.3 in [5]. Assume that  $\delta_N > \varepsilon$ . Let  $p(0) = N$  and inductively define  $p(j + 1)$ , for  $j = 0, \dots, l$ , to be the *largest* number from the index set  $\mathcal{I} = \{N, N - 1, \dots, 1\}$  for which  $\tilde{\delta}_{p(j+1)} > \tilde{\delta}_{p(j)}/(1 - \alpha)$ . Having done that, define  $k(j) = p(j + 1) + 1$  for all  $j = 0, \dots, l - 1$ , and finally put  $k(l) = 0$ . Note that by construction we have partitioned the index set (in reverse order) as follows

$$\mathcal{I} = \{p(0), \dots, k(0)\} \cup \{p(1), \dots, k(1)\} \cup \dots \cup \{p(l), \dots, k(l)\},$$

so that

$$\tilde{\delta}_{p(j)} \geq (1 - \alpha)\tilde{\delta}_i, \quad i = k(j), \dots, p(j), \quad j = 0, \dots, l - 1, \quad (4.3)$$

$$\tilde{\delta}_{p(j)} \geq \frac{\tilde{\delta}_{p(j-1)}}{1 - \alpha} \geq \frac{\tilde{\delta}_{p(0)}}{(1 - \alpha)^j} > \frac{\varepsilon}{(1 - \alpha)^j}, \quad (4.4)$$

and also

$$\tilde{\delta}_i \stackrel{(2.4)}{\geq} (1 - \gamma)\delta_N > (1 - \gamma)\varepsilon \stackrel{(4.1)}{=} \frac{\sqrt{\beta\omega}LD}{(1 - \gamma)(1 - \alpha)} \geq \frac{\sqrt{\omega}LD}{(1 - \gamma)(1 - \alpha)}, \quad i = 0, \dots, N. \quad (4.5)$$

Relations (4.3), (4.4) and (4.5), together with the fact that  $\omega$  and  $\rho$  satisfy (3.7), allow us to use Lemma 5 individually on each of the partitions to get the desired result:

$$\begin{aligned}
N + 1 = \sum_{j=0}^l (p(j) - k(j) + 1) &\stackrel{\text{(Lemma 5)}}{\leq} \sum_{j=0}^l \frac{L^2 D^2}{(1-\gamma)^2 (1-\alpha)^2 \tilde{\delta}_{p(j)}^2 - \omega L^2 D^2} \\
&\stackrel{(4.4)}{\leq} \frac{L^2 D^2}{(1-\gamma)^2 (1-\alpha)^2} \sum_{j=0}^{\infty} \frac{1}{\frac{\varepsilon^2}{(1-\alpha)^{2j}} - \frac{\omega L^2 D^2}{(1-\alpha)^2 (1-\gamma)^2}} \\
&\stackrel{(4.1)}{=} \frac{L^2 D^2}{(1-\gamma)^2 (1-\alpha)^2} \sum_{j=0}^{\infty} \frac{1}{\frac{\varepsilon^2}{(1-\alpha)^{2j}} - \frac{\varepsilon^2}{\beta/(1-\gamma)^2}} \\
&= \frac{L^2 D^2}{(1-\gamma)^2 (1-\alpha)^2} \sum_{j=0}^{\infty} \frac{\frac{\beta}{(1-\gamma)^2} (1-\alpha)^{2j}}{\left(\frac{\beta}{(1-\gamma)^2} - (1-\alpha)^{2j}\right) \varepsilon^2} \\
&\leq \frac{\beta L^2 D^2}{\left(\frac{\beta}{(1-\gamma)^2} - 1\right) (1-\gamma)^4 (1-\alpha)^2 \varepsilon^2} \sum_{j=0}^{\infty} (1-\alpha)^{2j} \\
&= \frac{\beta L^2 D^2}{(1-\gamma)^2 (\beta - (1-\gamma)^2) (1-\alpha)^2 (1 - (1-\alpha)^2)} \frac{1}{\varepsilon^2}.
\end{aligned}$$

□

Parameter  $\beta$  is needed only to safeguard the  $\gamma = 0$  case. If  $\gamma > 0$ , we can set  $\beta = 1$ . The expressions involving  $\alpha$  and  $\gamma$  in (4.2) then become identical, and optimizing for  $\alpha$  (resp.  $\gamma$ ) yields  $\alpha = \gamma = 1 - \frac{1}{\sqrt{2}}$ . For this choice of the parameters we get the following complexity estimate

$$N \leq \frac{16L^2 D^2}{\varepsilon^2}.$$

## 5 Solving the subproblems

We have shown that, in theory, one does not lose anything by solving the two principal subproblems (steps (4a) and (4c)) of the level method only approximately. However, we have not described *how* to perform these approximate computations. In this section we outline some possible approaches.

### 5.1 Minimizing the model

Consider any optimization method  $M$  for minimizing a (convex) function  $g$  on  $\mathcal{Q}$  with *guaranteed and computable* iteration complexity. That is, we assume that for any  $\kappa > 0$ ,  $M$  is accompanied with a formula for the number of iterations  $N(\kappa) = N(\kappa, g, y_0)$  needed to find a feasible point  $y_{N(\kappa)}$ , starting from the initial iterate  $y_0$ , for which the residual  $g(y_{N(\kappa)}) - g^*$  is at most  $\kappa$ . Let us start with a simple observation about this setup.

**Lemma 6** *For  $\kappa > 0$  and  $0 < \gamma < 1$ , one of the following conditions is satisfied*

$$(i) \quad g(y_{N(\gamma\kappa)}) \leq (1-\gamma)g^* + \gamma g(y_0),$$

(ii)  $g(y_0) \leq g^* + (1 + \gamma)\kappa$ .

**Proof:**

Observe that if

$$g(y_0) - g(y_{N(\gamma\kappa)}) \geq \kappa, \tag{5.1}$$

then

$$\begin{aligned} g(y_{N(\gamma\kappa)}) &\leq g^* + \gamma\kappa \leq g^* + \gamma(g(y_0) - g(y_{N(\gamma\kappa)})) \\ &\leq g^* + \gamma(g(y_0) - g^*) = (1 - \gamma)g^* + \gamma g(y_0). \end{aligned}$$

On the other hand, if condition (5.1) does not hold then

$$g(y_0) < \kappa + g(y_{N(\gamma\kappa)}) \leq g^* + (1 + \gamma)\kappa.$$

□

Applying this result to the model function, we obtain the following corollary.

**Theorem 3** *If we choose  $g \equiv \hat{f}_k$ ,  $y_0 = \arg \min_{0 \leq i \leq k} f(x_i)$  (whence  $g^* = \hat{f}_k^*$  and  $g(y_0) = f_k^*$ ) and  $\kappa = \varepsilon/(1 + \gamma)$ , then either inequality (2.3) holds for  $x_k^* = y_{N(\gamma\kappa)}$ , or  $y_0$  is an  $\varepsilon$ -solution of (1.1).*

This means that we either find a point  $x_k^*$  satisfying (2.3) in  $N = N(\varepsilon\gamma/(1 + \gamma))$  iterations of method  $M$ , or the best current iterate is  $\varepsilon$ -optimal for our master problem. It is likely that in practical computations we do not need to run method  $M$  for the full number of iterations  $N$ . Instead, we can check at every step whether condition (5.1) is satisfied, in which case we stop.

If a self-concordant barrier for the set  $\mathcal{Q}$  is available, we can use an interior-point method in place of  $M$ .

## 5.2 Projection subproblem

In this section we outline how one can, in principle, solve the approximate projection problem at iteration  $k$  using an interior-point method (IPM). For this we need to assume that a self-concordant barrier (with parameter  $\vartheta$ ) of  $C = \mathcal{L}_k(\alpha)$  is available. This is the case, for instance, when  $\mathcal{Q}$  is polyhedral. By  $x_C$  we denote the analytic center of  $C$  (minimizer of the barrier of  $C$ ) and let

$$\pi(z) \stackrel{\text{def}}{=} \inf\{t : x_C + t^{-1}(z - x_C) \in C\},$$

which is the Minkowski function of  $C$  with pole at  $x_C$ .

To lighten up the notation in what follows, let

$$g(x) \stackrel{\text{def}}{=} \|x - x_k\|^2, \quad g_* \stackrel{\text{def}}{=} \min_{x \in C} g(x) > 0, \quad \text{and} \quad g^* \stackrel{\text{def}}{=} \max_{x \in C} g(x).$$

We are interested in finding a  $\rho$ -approximate minimizer of  $g$  on  $C$ , in relative scale, as defined by the inequality (2.7).

**Theorem 4** *If the stopping criterion (2.5) is not satisfied, then the path-following interior-point method of Section 3.2 of [6], as applied to the problem of minimizing  $g$  on  $C$  and initialized at some point  $z \in \text{int } C$ , outputs a point  $x$  satisfying*

$$g(x) \leq (1 + \rho)g_* \quad (5.2)$$

after no more than

$$N = \mathcal{O}(1)\sqrt{\vartheta} \ln \left( \frac{2\vartheta}{\rho'(1 - \pi(z))} \right) \quad (5.3)$$

Newton steps, where

$$\rho' = \frac{\rho}{\left(1 + \frac{LD}{(1-\alpha)\delta_k}\right)^2 - 1} \geq \frac{\rho}{\left(1 + \frac{LD}{(1-\alpha)(1-\gamma)\varepsilon}\right)^2 - 1}. \quad (5.4)$$

**Proof:**

By Theorem 3.2.1 in [6], in  $N$  iterations of the IPM we obtain point  $x$  such that

$$g(x) - g_* \leq \rho'(g^* - g_*). \quad (5.5)$$

The triangle inequality  $\sqrt{g^*} \leq \sqrt{g_*} + D$  and the estimate  $\sqrt{g_*} \geq \frac{1}{L}(1 - \alpha)\tilde{\delta}_k$  (see Lemma 4) imply

$$\frac{g^*}{g_*} \leq \left(1 + \frac{LD}{(1 - \alpha)\tilde{\delta}_k}\right)^2. \quad (5.6)$$

The relation (5.2) then follows by combining (5.5) and (5.6). The inequality in (5.4) is a consequence of the assumption that the stopping criterion is not satisfied.  $\square$

If we want to use Theorem 4 in the framework of our approximate level method, we need to be able to ensure that inequality (5.2) holds. Therefore, a computable upper bound on the number of steps  $N$  given by (5.3) is needed. The constant term in (5.3) depends only on the parameters of the IPM algorithm and can be evaluated (a reasonable choice of the parameters makes this term equal to 7.36). All that remains is the availability of an interior point  $z$  of  $C$  for which we have a reasonable positive lower bound on  $1 - \pi(z)$ . This seems to be a difficult task. It is desirable to design a method which is free of this complication — algorithm capable to give a certificate that (5.2) is satisfied.

On the other hand, observe that the strong dependence of  $\rho$  (and  $\rho'$ ) on  $\varepsilon$  does not pose any problem for an IPM as this quantity appears under a logarithm. Since the dimension of the subproblem grows with increasing iteration count  $k$  of the master program, it would be interesting to develop a first-order method for solving the approximate projection subproblem. Eventually, executing even a single iteration of an IPM becomes impossible due to memory limitations.

### Acknowledgements

The author wishes to thank Yurii Nesterov for enlightening discussions and suggestions that greatly helped to improve the paper, and to Robert Chares for careful reading of an early draft.

## References

- [1] J. E. Kelley. The cutting plane method for solving convex programs. *Journal of the SIAM*, 8:703–712, 1960.
- [2] K. C. Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Mathematical Programming*, 27:320–341, 1983.
- [3] C. Lemaréchal. Nonsmooth optimization and descent methods. *IIASA Research Report 78-4*, 1978.
- [4] C. Lemaréchal, A. Nemirovskii, and Yu. Nesterov. New variants of bundle methods. *Mathematical Programming*, 69(1):111–147, 1995.
- [5] Yu. Nesterov. *Introductory Lectures on Convex Optimization. A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, 2004.
- [6] Yu. Nesterov and A. Nemirovski. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *SIAM Studies in Applied Mathematics*. 1994.
- [7] N. Z. Shor. *Minimization Methods for Non-differentiable Functions*. Springer Series in Computational Mathematics. Springer, 1985.