J Optim Theory Appl manuscript No.
(will be inserted by the editor)

Approximate Level Method for Nonsmooth Convex Minimization

Peter Richtárik

the date of receipt and acceptance should be inserted later

Abstract In this paper, we propose and analyze an approximate variant of the *level method* of Lemaréchal, Nemirovskii and Nesterov, which is a bundle method for minimizing nonsmooth convex functions. The main work per iteration of the level method is spent on (i) minimizing a piecewise-linear model of the objective function and on (ii) projecting onto the intersection of the feasible region and a polyhedron arising as a level set of the model function. We show that, by replacing exact computations in both cases by *approximate computations*, in *relative scale*, the theoretical iteration complexity increases only by a small factor depending on the approximation level. Our analysis gives a smooth interpolation between the approximate and the exact methods, retaining the original complexity in the exact case.

Keywords Level method · Approximate projections in relative scale · Nonsmooth convex minimization · Sensitivity analysis · Large-scale optimization · Bundle methods

Mathematics Subject Classification (2000) 65K05 · 90C06 · 90C25

1 Introduction

In this paper, we consider the basic convex optimization problem of the form

$$f^* := \min_{x \in Q} f(x), \tag{1}$$

Communicated by Marcin Studniarski

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 author(s).

The author wishes to thank Yurii Nesterov for enlightening discussions and suggestions that greatly helped to improve the paper, and to Robert Chares and two anonymous referees for carefully reading the manuscript.

Peter Richtárik
School of Mathematics, University of Edinburgh, King’s Buildings, Edinburgh, EH9 3JZ, United Kingdom
E-mail: peter.richtarik@ed.ac.uk

where $Q \subset \mathbb{R}^n$ is a compact convex set and f is Lipschitz continuous and convex with $Q \subseteq \mathrm{dom}\, f$. We assume that all information available to us about f be 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 $x_0, \dots, x_k \in Q$, it is natural to condense it into the following single object:

$$\hat{f}_k(x) := \max_{0 \le i \le k} \{ f(x_i) + \langle f'(x_i), x - x_i \rangle \}. \tag{2}$$

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). In *Kelley's method* [1], for example, the next iterate x_{k+1} is simply chosen to be the 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 Kelley's method is exponential in the dimension (see Section 3.3.2 in [2]). Several versions of *bundle methods* [3], [4], 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 λ_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 [5], 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 Kelley'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 Kelley's idea in practice. Also, the theoretical complexity no longer depends on the dimension of the problem. In fact, in order to produce an ε-minimizer of (1), it suffices to take

$$N \le \frac{4L^2D^2}{\varepsilon^2} \tag{3}$$

iterations, where $D = \mathrm{Diam}(Q)$ and L is the Lipschitz constant of f. It is well known 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* [6], level method is in many situations much better in practice. For examples of other related methods we refer the reader to [7–12].

Contribution. The main work at every iteration of the level method is spent on (i) minimizing a piecewise-linear model of the objective function and on (ii) projecting onto the level set of the model function. In this paper we show that by replacing exact computations in both cases by *approximate computations*, in *relative scale*[1] (in a certain sense which will be precisely defined in Section 2.2), the theoretical iteration complexity (3) increases only by a small factor depending on the level of approximation (for more detail see Theorem 4.1 and the subsequent discussion). In the exact case our result reduces to (3). That is, our analysis gives a smooth interpolation between the approximate and the exact methods, retaining the original complexity in the exact case. In other words, while spending less work on the subproblems, the new approach still retains the good theoretical guarantees of the level method.

[1] Results on convergence in relative scale are rare in the continuous optimization literature; for some recent work see [13–15].

We show that for the first subproblem, a precision proportional to the current gap and *independent* of the target accuracy ε of the master convex problem is completely satisfactory (see Section 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

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

Observe that

$$\rho = O\left(\frac{\varepsilon^4}{L^4D^4}\right).$$

Our goal in this work is to describe the essential core of the theory in a concise manner. Extensions (say to sets Q of special structure, to the situation with unknown L and D and so on), applications and numerical implementation will be the focus of a follow-up work.

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 $\mathbb{R}^n$ as column vectors and the entries of $x \in \mathbb{R}^n$ are denoted by $x = (x^{(1)}, \ldots, x^{(n)})^T$. For x and y in $\mathbb{R}^n$, $\langle x, y\rangle$ 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 x, 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 Q, let us denote the *minimal value* of the model function (2), resp. the *record value* of the objective function, by

$$\hat{f}_k^* := \min_{x\in Q} \hat{f}_k(x), \quad \text{resp.} \quad f_k^* := \min_{0\le i\le k} f(x_i). \tag{4}$$

Note that the following relations hold for all k:

$$\hat{f}_k \le f, \qquad \hat{f}_k \le \hat{f}_{k+1}, \quad \text{and} \quad \hat{f}_k^* \le \hat{f}_{k+1}^* \le f^* \le f_{k+1}^* \le f_k^*. \tag{5}$$

The first inequality states that the model function always underestimates f, the second says that the model function grows as we add new cutting-planes to it. Observe that due to the last set of inequalities we know that the quantity

$$\delta_k := f_k^* - \hat{f}_k^* \tag{6}$$

is nonincreasing, and that we can stop once it gets bellow the target error tolerance ε.

The level method at every iteration solves two subproblems: (i) minimization of the model function and (ii) Euclidean projection onto a certain level set of the model function. In Sections 2.1 and 2.2 we formally describe acceptable approximate solutions of these two subproblems and then in Section 2.3 proceed with describing our algorithm in more detail. We postpone the question of *how* to obtain these approximate solutions in practice until Section 5.

2.1 Minimizing the Model Function

We 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 $x_k^* \in Q$ such that

$$\tilde{f}_k^* := \hat{f}_k(x_k^*) \leq (1-\gamma)\hat{f}_k^* + \gamma f_k^*. \tag{7}$$

Note that the choice $\gamma = 0$, which is the case in the level method, corresponds to finding the exact minimizer. If we define

$$\tilde{\delta}_k := f_k^* - \tilde{f}_k^*, \tag{8}$$

then (7) is equivalent to $(1-\gamma)\delta_k \leq \tilde{\delta}_k$. Furthermore, from $\hat{f}_k^* \leq f_k^*$ and (7) we get $\tilde{f}_k^* \geq \hat{f}_k^*$, which in turn gives $\tilde{\delta}_k \leq \delta_k$. Putting these two observations together, we obtain the following useful bounds:

$$(1-\gamma)\delta_k \leq \tilde{\delta}_k \leq \delta_k, \qquad k \geq 0. \tag{9}$$

In words, the point x_k^* approximately "closes the gap" δ_k, in *relative scale*, with accuracy governed by the parameter γ. The true gap δ_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

$$\tilde{\delta}_k \leq (1-\gamma)\varepsilon \tag{10}$$

implies

$$f_k^* - f^* \leq \delta_k \overset{(9)}{\leq} \frac{\tilde{\delta}_k}{1-\gamma} \overset{(10)}{\leq} \varepsilon,$$

and hence it is a good stopping criterion for our method. The following relations will be useful later in the complexity analysis section:

$$(1-\gamma)\tilde{\delta}_k \overset{(9)}{\leq} (1-\gamma)\delta_k \leq (1-\gamma)\delta_i \overset{(9)}{\leq} \tilde{\delta}_i, \qquad k \geq i. \tag{11}$$

2.2 Projection Subproblem

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

$$l_k(\alpha) := (1-\alpha)\tilde{f}_k^* + \alpha f_k^*, \tag{12}$$

and consider the *level set*

$$\mathscr{L}_k(\alpha) := \{x \in Q \;:\; \hat{f}_k(x) \leq l_k(\alpha)\}. \tag{13}$$

Note that the classical level method of Lemaréchal, Nemirovskii and Nesterov [5] uses $\hat{f}_k^*$ instead of $\tilde{f}_k^*$ in the definition of the level value, which corresponds to the choice $\gamma = 0$ in our setting. In our method, the next iterate x_{k+1} is 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 2.1 (Approximate Projection) *Let C be a convex set, $x \notin C, z \in C$ and $\rho \geq 0$. We say that z is a ρ-*approximate projection *of x onto C, if*

$$\|x-z\|^2 \leq (1+\rho) \inf_{y \in C} \|x-y\|^2. \tag{14}$$

Let us remark at this point that in Section 3 we will establish a simple inequality that holds for approximate projections; this will be useful in our analysis. Then, in Section 5.2, we show that approximate projections can be computed, in principle, using interior point methods.

2.3 The Algorithm

Having described the method in previous subsections in some detail, we now proceed to a formal description.

```
Approximate Level Method

(1) Input:
    $f,\ Q,\ L,\ D,\ x_0,\ \varepsilon > 0$

(2) Choice of parameters:
    (a) $0 < \alpha < 1,\ 0 < \gamma < 1,\ \beta > 1$
    (b) Set projection accuracy to $\rho = \frac{\omega^2}{2\omega+1}$, where $\omega = \frac{(1-\gamma)^4(1-\alpha)^2\varepsilon^2}{\beta L^2 D^2}$

(3) For $k \geq 0$ iterate:
    (a) Compute $f_k^*$ and $\tilde{f}_k^*$ and set $\tilde{\delta}_k = f_k^* - \tilde{f}_k^*$
    (b) STOP if $\tilde{\delta}_k \leq (1-\gamma)\varepsilon$
    (c) Compute $x_{k+1}$ as an $\rho$-approximate projection of $x_k$ onto $\mathscr{L}_k(\alpha)$
```

The first two steps of the method describe the input (see Table 1) and the choice of parameters (see Table 2). For a discussion about reasonable choice of parameters please read the discussion following Theorem 4.1. The main iterative part is described in step 3. Note that the stopping 3(b) criterion corresponds to (10).

3 Approximate Projection Inequality

The analysis of the level method applied to problem (1) (Section 2.2.1 in [2] or Section 3.3.3 in [2]) 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, which can be successfully substituted into the original analysis, yielding the desired sensitivity result.

object	meaning
f	objective function
Q	feasible set
L	Lipschitz constant of f
D	(upper bound on the) diameter of Q
x_0	an initial feasible point
ε	target accuracy of the master problem (1)

Table 1 Input data.

parameter	meaning
α	parameter defining the level set
β	an auxiliary parameter indirectly controlling ρ
γ	relative accuracy with which we minimize the model function
ρ	relative accuracy with which we compute projections

Table 2 Parameters of the algorithm.

The main goal of this section is to show that condition (14) implies an inequality of the form

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

for certain $\omega=\omega(\rho)$. Note that if $\rho=0$, 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 ρ.

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

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

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

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

$$p^* := p^*(x,r,y) := \max\{\|z-y\|^2 \,:\, z\in\mathscr{B}(x,r)\cap\mathscr{H}(x)\}, \tag{15}$$

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

Lemma 3.1 *Fix $0\neq x\in\mathbb{R}^n$, $r\ge\|x\|$, $y\in\mathscr{H}(x)$, let*

$$R := \sqrt{r^2-\|x\|^2}, \tag{16}$$

and by $\hat{y}$ denote the projection of y onto $\partial\mathscr{H}(x)$, i.e.,

$$\hat{y} = y - \frac{\langle y,x\rangle}{\|x\|^2}x. \tag{17}$$

Then

$$p^*(x,r,y) = R^2+\|y\|^2+2R\|\hat{y}\|. \tag{18}$$

Proof For $r = \|x\|$ the statement is trivial since $\mathscr{B} \cap \mathscr{H} = \{0\}$ and hence in problem (15) we have $z^* = 0$ and $p^* = \|y\|^2$. From now on assume that $r > \|x\|$. First, note that the objective function can be written as

$$\begin{aligned} \|z-y\|^2 &= \|z-x\|^2 + 2\langle z-x, x-y\rangle + \|x-y\|^2 \\ &= \|z-x\|^2 + 2\langle z, x-y\rangle - \|x\|^2 + \|y\|^2. \end{aligned} \tag{19}$$

Case 1. Assume $\hat{y} = 0$; that is, $y = tx$ for some $t \le 0$. In this case $\langle z, y\rangle \ge 0$ for all $z \in \mathscr{H}$. Therefore, in view of (19), all feasible points z satisfy

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

with equality precisely when $z \in \partial\mathscr{B} \cap \partial\mathscr{H}$. Note that this agrees with (18).

Case 2. Assume $\hat{y} \neq 0$. It follows from (19) that, if all optimal solutions z^* of

$$q^* := \max_{\substack{\|z-x\|^2 \le r^2 \\ \langle z, x\rangle \le 0}} \langle x-y, z\rangle \tag{20}$$

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

$$p^* = r^2 + 2q^* - \|x\|^2 + \|y\|^2. \tag{21}$$

Indeed, we will show that the Lagrange multiplier λ at any optimal point z^* of (20) corresponding to the first inequality is positive, and hence $\|z^* - x\| = r$. The Lagrangean dual of (20) is (there is no duality gap)

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

where

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

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

$$q^* = \min_{\lambda > 0} \frac{1}{4\lambda}\|\hat{y}\|^2 + \lambda R^2. \tag{22}$$

Its minimizer is $\lambda^* = \frac{\|\hat{y}\|}{2R}$ and substituting this into (22) and q^* into (21) gives (18). □

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

Lemma 3.2 *Let $0 \neq x \in \mathbb{R}^n$ and $\rho \ge 0$. Then for $r^2 := (1+\rho)\|x\|^2$ and*

$$\omega := \rho + \sqrt{\rho^2 + \rho}, \tag{23}$$

we have the following inequality

$$\|x-z\|^2 + \|z-y\|^2 \le (1+\omega)\|x-y\|^2, \quad y \in \mathscr{H}(x), \quad z \in \mathscr{H}(x) \cap \mathscr{B}(x,r). \tag{24}$$

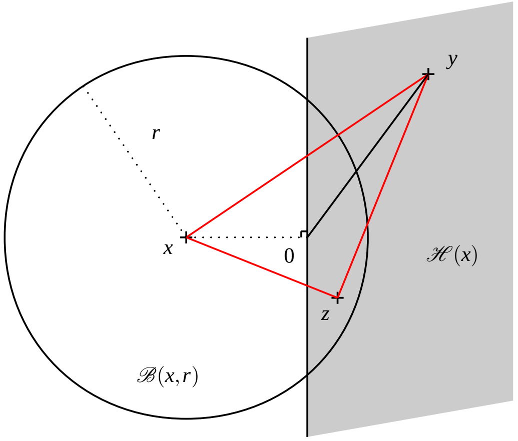

Fig. 1 Lemma 3.2 – approximate projection.

Proof Fixing arbitrary $y \in \mathscr{H}$, Lemma 3.1 implies that

$$\begin{aligned}\max_{z\in\mathscr{H}\cap\mathscr{B}} \|x-z\|^2+\|z-y\|^2 \;&\le\; \max_{z\in\mathscr{H}\cap\mathscr{B}} \|x-z\|^2+\max_{z\in\mathscr{H}\cap\mathscr{B}} \|z-y\|^2\\ &\overset{(18)}{=} r^2+(R^2+\|y\|^2+2R\|\hat{y}\|)\\ &\overset{(16)}{=} (1+2\rho)\|x\|^2+\|y\|^2+2\rho^{1/2}\|x\|\|\hat{y}\|.\end{aligned}$$

It thus remains to argue that the last expression is upper-bounded by $(1+\omega)\|x-y\|^2$. Using (17) and (23), this inequality can be equivalently written as

$$(\omega-2\rho)\|x\|^2-2\rho^{1/2}[\|x\|^2\|y\|^2-\langle x,y\rangle^2]^{1/2}+\omega\|y\|^2-2(1+\omega)\langle x,y\rangle \ge 0. \tag{25}$$

Observe, however, that for $\omega \ge 2\rho$ we have

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

Since $\langle x,y\rangle \le 0$, notice that this inequality is *stronger* than (25), provided that $\omega \ge 2\rho$ and $\sqrt{\omega(\omega-2\rho)} \ge \rho^{1/2}$. Solving for ω in terms of ρ in the latter quadratic gives (23). □

We can now proceed to the main result of this section.

Theorem 3.1 (Approximate Projection Inequality) *Let C be a convex set and x a point not lying in this set. If $z \in C$ is a* ρ-approximate projection *of x onto C, with $\rho \ge 0$, and $\omega = \omega(\rho)$ is given by* (23), *then*

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

Proof By appropriate shifting we can without loss of generality assume that the projection point be the origin. We now apply Lemma 3.2 and note that $C \subset \mathscr{H}$ since $\partial\mathscr{H}$ is a supporting hyperplane to C at the origin. □

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 (23) by

$$\omega = (\sqrt{2} + 1)\rho^{1/2}. \tag{26}$$

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 (23) by

$$\omega = (\sqrt{2} + 1)\rho. \tag{27}$$

4 Iteration Complexity Analysis

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

Lemma 4.1 says that if the values of the (presumably easily computable) gap $\tilde{\delta}_i$, for $i = k, \ldots, p$, 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 $\mathscr{L}_i(\alpha)$ for $i = k, \ldots, p$. This property will be exploited in Lemma 4.3, which is in turn used in the proof of our main result.

Lemma 4.1 (cf. Lemma 3.3.1, [2]) *If for $i = k, \ldots, p$ we have $\tilde{\delta}_p \geq (1 - \alpha)\tilde{\delta}_i$, then*

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

Proof For any $i \in \{k, \ldots, p\}$ we have

$$\hat{f}_i(x_p^*) \overset{(5)}{\leq} \hat{f}_p(x_p^*) = \tilde{f}_p^* \overset{(8)}{=} f_p^* - \tilde{\delta}_p \leq f_p^* - (1 - \alpha)\tilde{\delta}_i \overset{(5)}{\leq} f_i^* - (1 - \alpha)\tilde{\delta}_i \overset{(8), (12)}{\leq} l_i(\alpha).$$

The claim now follows by comparing the resulting inequalities with the definition of the level sets (13). □

The statement and proof is analogous to that of Lemma 3.3.1 in [2], which is formulated with exact gaps δ_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 Q. The following result says that if the current gap is large, then the size of the next step will also be large.

Lemma 4.2 (cf. Lemma 3.3.2, [2]) *If $\{x_k\}$ is a sequence of points generated by the approximate level method, then*

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

Proof Indeed,

$$\begin{aligned} f(x_k) - (1 - \alpha)\tilde{\delta}_k &\geq f_k^* - (1 - \alpha)\tilde{\delta}_k \\ &\overset{(8), (12)}{=} l_k(\alpha) \\ &\geq \hat{f}_k(x_{k+1}) \\ &\overset{(2)}{\geq} f(x_k) + \langle f'(x_k), x_{k+1} - x_k \rangle \\ &\geq f(x_k) - L\|x_{k+1} - x_k\|. \end{aligned}$$

□

Boundedness of the feasible set Q is only needed in the last two results.

Lemma 4.3 *Let ω be chosen as in Theorem 3.1. If for some $p \geq k$ we have*

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

then

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

Proof In view of Lemma 4.1, point x_p^* lies in $\mathscr{L}_i(\alpha)$ for all $i = k, \dots, p$. We can therefore individually for each i use Theorem 3.1 with $x = x_i$, $C = \mathscr{L}_i(\alpha)$, $z = x_{i+1}$ and $y = x_p^*$. This together with Lemma 4.2 and the inequality $\tilde{\delta}_i \geq (1-\gamma)\tilde{\delta}_p$ (see (11)) implies

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

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 and rearranging the terms. □

We are now ready to state the main result of this paper, the iteration complexity of the Approximate Level Method.

Theorem 4.1 (Iteration Complexity) *Let $\varepsilon, \alpha, \beta, \gamma$ be such that $\varepsilon > 0$, $0 < \alpha < 1$, $\beta > 1$ and $0 \leq \gamma < 1$, and let*

$$\rho = \frac{\omega^2}{2\omega + 1} \quad \textit{with} \quad \omega = \frac{(1-\alpha)^2(1-\gamma)^4\varepsilon^2}{\beta L^2D^2}. \tag{28}$$

Then the level method with ρ-approximate projections produces an ε-approximate minimizer of problem (1) *after no more than*

$$N = \left\lfloor \frac{\beta}{(\beta-1)(1-\gamma)^4} \times \frac{1}{\alpha(1-\alpha)^2(2-\alpha)} \times \frac{L^2D^2}{\varepsilon^2} \right\rfloor \tag{29}$$

iterations.

Proof The proof closely follows that of Theorem 3.3.1 in [2] with the main difference being that we use Lemma 4.3 instead of Lemma 3.3.3 in [2]. Assume that

$$\delta_N > \varepsilon. \tag{30}$$

Let $p(0) = N$ and inductively define $p(1), p(2), \dots$ in the following way. If $p(j)$ is well-defined and the set

$$S_j := \left\{ i \in \{0, 1, \dots, p(j) - 1\} \,:\, \tilde{\delta}_{p(i)} > \frac{\tilde{\delta}_{p(j)}}{1-\alpha} \right\}.$$

is nonempty, then let $p(j+1) = \max\{i \,:\, i \in S_j\}$. If S_j is empty then let $l := j$ and stop the process having defined $p(0), p(1), \dots, p(l)$.

Now define $k(j) = p(j+1)+1$ for $j = 0, \dots, l-1$, and finally put $k(l) = 0$. Note that we have constructed a partition of $\{0,1,\dots,N\}$ (in reverse order) into $l+1$ sets as follows

$$\{0,1,\dots,N\} = \underbrace{\{k(l),\dots,p(l)\}}_{I_l} \cup \underbrace{\{k(l-1),\dots,p(l-1)\}}_{I_{l-1}} \cup \cdots \cup \underbrace{\{k(0),\dots,p(0)\}}_{I_0}. \tag{31}$$

Note that, by construction, for each $j = 0, \dots, l$ we have

$$\tilde{\delta}_{p(j)} \geq (1-\alpha)\tilde{\delta}_i, \quad i \in I_j, \tag{32}$$

and

$$\tilde{\delta}_{p(j)} > \frac{\tilde{\delta}_{p(j-1)}}{1-\alpha} > \frac{\tilde{\delta}_{p(0)}}{(1-\alpha)^j} = \frac{\tilde{\delta}_N}{(1-\alpha)^j} \overset{(9)}{\geq} \frac{(1-\gamma)\delta_N}{(1-\alpha)^j} \overset{(30)}{>} \frac{(1-\gamma)\varepsilon}{(1-\alpha)^j}. \tag{33}$$

Moreover, for $i = 0, \dots, N$ we have

$$\tilde{\delta}_i \overset{(11)}{\geq} (1-\gamma)\delta_N > (1-\gamma)\varepsilon \overset{(28)}{=} \frac{\sqrt{\beta\omega}LD}{(1-\gamma)(1-\alpha)} \geq \frac{\sqrt{\omega}LD}{(1-\gamma)(1-\alpha)}. \tag{34}$$

Relations (32), (33) and (34), together with the fact that ω and ρ satisfy (23), allow us to use Lemma 4.3 individually on each of the partitions to get the desired result:

$$\begin{aligned}
N+1 &\overset{(31)}{=} \sum_{j=0}^{l} (p(j)-k(j)+1) \\
&\overset{\text{(Lemma 4.3)}}{\leq} \sum_{j=0}^{l} \frac{L^2D^2}{(1-\gamma)^2(1-\alpha)^2\tilde{\delta}_{p(j)}^2 - \omega L^2 D^2} \\
&\overset{(33)}{\leq} \frac{L^2D^2}{(1-\gamma)^2(1-\alpha)^2} \sum_{j=0}^{l} \frac{1}{\frac{(1-\gamma)^2\varepsilon^2}{(1-\alpha)^{2j}} - \frac{\omega L^2 D^2}{(1-\alpha)^2(1-\gamma)^2}} \\
&\overset{(28)}{=} \frac{L^2D^2}{(1-\gamma)^2(1-\alpha)^2} \sum_{j=0}^{l} \frac{1}{\frac{(1-\gamma)^2\varepsilon^2}{(1-\alpha)^{2j}} - \frac{(1-\gamma)^2\varepsilon^2}{\beta}} \\
&\leq \frac{L^2D^2}{(1-\gamma)^4(1-\alpha)^2} \sum_{j=0}^{l} \frac{\beta(1-\alpha)^{2j}}{(\beta-(1-\alpha)^{2j})\,\varepsilon^2} \\
&< \frac{\beta L^2 D^2}{(\beta-1)\,(1-\gamma)^4(1-\alpha)^2\varepsilon^2} \sum_{j=0}^{\infty} (1-\alpha)^{2j} \\
&= \underbrace{\frac{\beta}{(\beta-1)(1-\gamma)^4(1-\alpha)^2(1-(1-\alpha)^2)} \frac{L^2D^2}{\varepsilon^2}}_{c}.
\end{aligned}$$

We have thus shown that if $\delta_N > \varepsilon$, it follows that $N+1 < c$. Conversely, if $N \geq c-1$ (which holds if $N \geq \lfloor c \rfloor$), we have $\delta_N \leq \varepsilon$. □

Looking at (29), the optimal choice of the parameters is $\beta \to +\infty$ (this implies the exact projections case since then $\omega \to 0$ and hence $\rho \to 0$, see (28)), $\gamma = 0$ (this corresponds to minimizing the model function exactly, see (7)) and $\alpha = 1 - \frac{1}{\sqrt{2}}$. In this case our analysis gives the complexity estimate

$$N \leq \frac{4L^2D^2}{\varepsilon^2}, \tag{35}$$

recovering the existing result for the level method (3).

If we want to be using approximate projections ($\beta < +\infty$, $\gamma \ll 1$) and/or minimizing the model function approximately only ($\gamma > 0$), we will have to pay, in comparison with the worst-case estimate for the exact method (35), by an increase in the number of iterations by the constant factor

$$\theta(\beta,\gamma) := \frac{\beta}{(\beta-1)(1-\gamma)^4}.$$

Table 3 lists the values of $\theta(\beta,\gamma)$ for several choices of the constants β and γ.

β/γ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	3.05	4.88	8.33	15.43	32.00	78.13	246.91	1250.00	20000.00
4	2.03	3.26	5.55	10.29	21.33	52.08	164.61	833.33	13333.33
8	1.74	2.79	4.76	8.82	18.29	44.64	141.09	714.29	11428.57
16	1.63	2.60	4.44	8.23	17.07	41.67	131.69	666.67	10666.67
32	1.57	2.52	4.30	7.96	16.52	40.32	127.44	645.16	10322.58
64	1.55	2.48	4.23	7.84	16.25	39.68	125.42	634.92	10158.73
128	1.54	2.46	4.20	7.78	16.13	39.37	124.43	629.92	10078.74
256	1.53	2.45	4.18	7.75	16.06	39.22	123.94	627.45	10039.22

Table 3 Factors $\theta(\beta,\gamma)$ increasing the complexity of the approximate level method as compared to the (exact) level method.

The effect of large γ is very strong and hence γ should be kept small. The dependence on β is very weak. For instance, if we use $\beta = 16$ and $\gamma = 0.1$, we need 63% more iterations, in theoretical worst case, as compared to the number of iteration required by the level method.

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. Our discussion is not meant to be exhaustive.

5.1 Minimizing the Model Function

Consider any optimization method M for minimizing a (convex) function g on Q with *guaranteed and computable* iteration complexity. That is, we assume that for any $\kappa > 0$, M is accompanied with an explicit formula for the number of iterations $N(\kappa) = N(\kappa, g, y_0)$ needed for the method 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 κ. Let us start with a simple observation about this setup.

Lemma 5.1 *For $\kappa > 0$ and $0 < \gamma < 1$, at least one of the following conditions is satisfied*

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

Proof Observe that if

$$g(y_0) - g(y_{N(\gamma\kappa)}) \geq \kappa, \tag{36}$$

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 (36) does not hold, then

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

finishing the proof. □

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

Corollary 5.1 *If we choose*

$$g := \hat{f}_k, \qquad y_0 := \arg\min_{0\leq i\leq k} f(x_i), \qquad \textit{and} \qquad \kappa := \frac{\varepsilon}{1+\gamma},$$

whence $g^* = \hat{f}_k^*$ and $g(y_0) = f_k^*$, then either inequality (7) *holds for $x_k^* = y_{N(\gamma\kappa)}$, or y_0 is an ε-solution of* (1)*, or both.*

This means that we either find a point x_k^* satisfying (7) in $N = N(\varepsilon\gamma/(1+\gamma))$ iterations of method M, or the best current iterate is ε-optimal for our master problem. It is likely, although we do not provide computational results in this paper, 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 iteration or after a fixed number of iterations whether condition (36) is satisfied, in which case we stop.

If a self-concordant barrier for the set 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 ϑ) of $C = \mathscr{L}_k(\alpha)$ be available. This is the case, for instance, when Q is polyhedral. By x_C we denote the *analytic center* of C. If, for instance, C is represented as

$$C = \{x \in \mathbb{R}^k \ : \ a_i^T x \leq b_i, \ i = 1, \ldots, m\}$$

for some vectors $a_i \in \mathbb{R}^k$ and $b \in \mathbb{R}^m$, and we assume that C has nonempty interior, then the analytic center of C is the minimizer of the logarithmic barrier function

$$\Psi(x) := -\sum_{i=1}^{m} \log(b_i - a_i^T x).$$

Further, let

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

which is the Minkowski function of C with pole at x_C. For more details about these notions please refer to [16].

To lighten up the notation in what follows, let

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

We are interested in finding a ρ-approximate minimizer of g on C, in relative scale, as defined by the inequality (14).

Theorem 5.1 *If the stopping criterion* (10) *is not satisfied, then the path-following interior-point method of Section 3.2 of* [16], *as applied to the problem of minimizing g on C and initialized at some point $z \in \operatorname{int} C$, outputs a point x satisfying*

$$g(x) \le (1+\rho) g_* \tag{37}$$

after no more than

$$N = \mathscr{O}(1)\sqrt{\vartheta} \ln\left(\frac{2\vartheta}{\rho'(1-\pi(z))}\right) \tag{38}$$

Newton steps, where

$$\rho' = \frac{\rho}{\left(1 + \frac{LD}{(1-\alpha)\tilde{\delta}_k}\right)^2 - 1} \ge \frac{\rho}{\left(1 + \frac{LD}{(1-\alpha)(1-\gamma)\varepsilon}\right)^2 - 1}. \tag{39}$$

Proof By Theorem 3.2.1 in [16], in N iterations of the IPM we obtain point x such that

$$g(x) - g_* \le \rho'(g^* - g_*). \tag{40}$$

The triangle inequality $\sqrt{g^*} \le \sqrt{g_*} + D$ and the estimate $\sqrt{g_*} \ge \frac{1}{L}(1-\alpha)\tilde{\delta}_k$ (see Lemma 4.2) imply

$$\frac{g^*}{g_*} \le \left(1 + \frac{LD}{(1-\alpha)\tilde{\delta}_k}\right)^2. \tag{41}$$

The relation (37) then follows by combining (40) and (41). The inequality in (39) is a consequence of the assumption that the stopping criterion is not satisfied. □

If we want to use Theorem 5.1 in the framework of our approximate level method, we need to be able to ensure that inequality (37) holds. Therefore, a computable upper bound on the number of steps N given by (38) is needed. The constant term in (38) 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—an algorithm capable to give a certificate that (37) is satisfied.

On the other hand, observe that the strong dependence of ρ (and ρ') on ε 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.

6 Concluding Remarks

We have shown that it is possible to extend the exact level method to the inexact case, in which approximate projections and approximate minimization of the model function, both in relative scale, are performed. Moreover, under our worst-case iteration complexity results, this was done at the cost of a small multiplicative factor only, diminishing to one as the level of exactness of both computations increases. This work can thus be viewed as a sensitivity analysis of the level method to the level of exactness of the two main computations involved. We have presented the core of the idea only: extensions, generalizations, practical considerations and numerical experiments are beyond the scope of this paper and are left for future research.

References

1. Kelley, J.E.: The cutting plane method for solving convex programs. Journal of the SIAM **8**, 703–712 (1960)
2. Nesterov, Y.: Introductory Lectures on Convex Optimization. A Basic Course, *Applied Optimization*, vol. 87. Kluwer Academic Publishers, Boston (2004)
3. Lemaréchal, C.: Nonsmooth optimization and descent methods. IIASA Research Report 78-4 (1978)
4. Kiwiel, K.C.: An aggregate subgradient method for nonsmooth convex minimization. Mathematical Programming **27**, 320–341 (1983)
5. Lemaréchal, C., Nemirovskii, A., Nesterov, Y.: New variants of bundle methods. Mathematical Programming **69**(1), 111–147 (1995). DOI http://dx.doi.org/10.1007/BF01585555
6. Shor, N.Z.: Minimization Methods for Non-differentiable Functions. Springer Series in Computational Mathematics. Springer (1985)
7. Polyak, B.T.: Minimization of unsmooth functionals. USSR Computational Mathematics and Mathematical Physics **9**(3), 14–29 (1969)
8. Kim, S., Ahn, H., Cho, S.C.: Variable target value subgradient method. Mathematical Programming **49**(1–3), 359–369 (1991). DOI 10.1007/BF01588797
9. Kiwiel, K.C.: The efficiency of subgradient projection methods for convex optimization, part i: General level methods. SIAM Journal on Control and Optimization **34**, 660–676 (1996)
10. Kiwiel, K.C.: The efficiency of subgradient projection methods for convex optimization, part ii: Implementations and extensions. SIAM Journal on Control and Optimization **34**, 677–697 (1996)
11. Cegielski, A.: A method of projection onto an acute cone with level control in convex minimization. Mathematical Programming **85**(3), 469–490 (1999)
12. Cegielski, A., Dylewski, R.: Residual selection in a projection method for convex minimization problems. Optimization **52**, 211–220 (2003). DOI 10.1080/0233193031000079883
13. Nesterov, Y.: Unconstrained convex minimization in relative scale. CORE Discussion Paper #2003/96 (November 2003)
14. Richtárik, P.: Improved algorithms for convex minimization in relative scale. Tech. rep. (2009)
15. Richtárik, P.: Simultaneously solving seven optimization problems in relative scale. Tech. rep. (2009)
16. Nesterov, Y., Nemirovski, A.: Interior-Point Polynomial Algorithms in Convex Programming, *SIAM Studies in Applied Mathematics*, vol. 13. SIAM, Philadelphia, PA (1994)