ON SOME VERSIONS OF SUBSPACE OPTIMIZATION METHODS WITH INEXACT GRADIENT INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

It is well-known that accelerated gradient first order methods possess optimal complexity estimates for the class of convex smooth minimization problems. In many practical situations, it makes sense to work with inexact gradients. However, this can lead to the accumulation of corresponding inexactness in the theoretical estimates of the rate of convergence. We propose some modification of the methods for convex optimization with inexact gradient based on the subspace optimization sush as Nemirovski's Conjugate Gradients and Sequential Subspace Optimization. We research the methods convergence for different condition of inexactness both in gradient value and accuracy of subspace optimization problems. Besides this, we investigate generalization of this result to the class of quasar-convex (weakly-quasi-convex) functions.

INTRODUCTION

The first-order methods are an important class of approaches for optimization problems. They have different advantages: simple implementation, usually low cost of iterations and high performance for wide class of functions (Beck (2017), Gasnikov (2017), Polyak (1987)). Nevertheless, there are different areas where method access to only inexact gradient: the gradient free optimization in infinite dimensional spaces Vasilyev (2002), inverse problems Kabanikhin (2011), saddle-point problems Lan (2020), L. Hien (2023). Therefore, such methods are interesting for many researchers Devolder (2013), O. Devolder & Nesterov (2014), d'Aspremont (2008), Polyak (1987), Vasin et al. (2021).

Further, note that there are well-known results about convergence of first order methods and optimality of accelerated methods for class of convex functions d'Aspremont et al. (2021), Bubeck (2015), Nemirovsky & Yudin (1979a), Barre et al. (2020), Nemirovsky & Yudin (1979b). On the other hand, nonconvex optimization appears in many practical problems. Especially, interest to such problems is growing because of deep learning Lan (2020), Goodfellow et al. (2016). One of possible expansion of convexity is quasar-convexity (or weakly quasar convexity). This class and non-convex examples are described in S. Guminov (2008); Hardt et al. (2016); Hinder et al. (2020).

This paper continues research of the first order methods based on subspace optimization. Such methods were considered in S. Guminov (2008); Kuruzov & Stonyakin (2021). One of such methods is Sequential Subspace Optimization (SESOP) Narkiss & Zibulevsky (2005). This method searches sequentially minima on subspaces and converges to the solution. Recently, several interesting properties were demonstrated for this method. Especially, S. Guminov (2008) contains proof for convergence of SESOP for quasar-convex case. It demonstrates that this method converges with rate similar to accelerated rates. Besides, Kuruzov & Stonyakin (2021) proofs the convergence in the case of inexact gradient. Moreover, it states that this method does not accumulate error in contrast to other known accelerated methods. However, there were no works devoted to complexity of such methods in terms of gradient calculations but not iterations. In this work, we propose to use ellipsoid method Nemirovski et al. (2010); Gladin et al. (2020) for auxiliary problems. We demonstrate theoretical complexity under condition of convexity and with inexactness because of inexact gradient and inexact solution of subproblems.

Besides of SESOP, we consider generalization of Nemirovski's Conjugate Gradient method. In the Section 3.1 we research it convergence for quasar-convex functions that meet quadratic growth condition. For this method we also demonstrated its non-accumulation of the additive gradient

inexactness. Nevertheless, note that quality of obtained by CG method solution can be degraded to $O\left(\sqrt{\frac{\varepsilon}{\mu}}\delta_1\right)$. Also, we prove that for obtaining quality ε we need $O\left(\frac{1}{\varepsilon}\right)$ iterations for enough small δ_1. The main disadvantage of the result in this section is requirement for enough large norm of gradient on each iteration. But we introduce stop condition that guarantee compromise between quality of solution and complexity of algorithm.

Our contributions are the following:

1. Linear convergence for inexact CG method in non-convex case with. We generalize proof of convergence for Nemirovski's conjugate gradient method with inexact gradient. Moreover, we propose stopping rule to approach required quality.

2. Complexity of auxiliary problems for SESOP and CG methods in convex case. It is natural to solve low-dimensional subproblems in these methods by low-dimensional methods. We consider Ellipsoid Method and Multidimensional dichotomy for these problems and estimate complexity for convex case.

1 Problem Statement

Let us consider minimization problems of convex and L-smooth function f ($\|\cdot\|$ is a usual Euclidean norm)

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\| \quad \forall x, y \in \mathbb{R}^n \tag{1}$$

with an inexact gradient $g : \mathbb{R}^n \to \mathbb{R}^n$:

$$\|g(x) - \nabla f(x)\| \leqslant \delta, \tag{2}$$

where $L > 0$ and $\delta > 0$.

The result for convergence per iteration are formulated for quasar-convex problems.

Definition 1. *Assume that $\gamma \in (0, 1]$ and let x^* be a minimizer of the differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. The function f is γ-quasar-convex with respect to x^* if for all $x \in \mathbb{R}^n$,*

$$f(x^*) \geq f(x) + \frac{1}{\gamma}\langle \nabla f(x), x^* - x\rangle. \tag{3}$$

This class generalize convex functions. It is also known as weakly-quasi convex functions. In the case of $\gamma = 1$ it is well-known star-convexity. Hinder et al. (2020) demonstrates that functions $f(x) = (x^2 + 1/8)^{1/6}$ are quasar-convex but not convex or star-convex. Besides, work Wang & Wibisono (2023) demonstrates that some problems of training general linear models are quasar-convex.

Also, in this work, we will consider generalizations of strong convexity. The first considered condition is PL-condition.

Definition 2. *The differentiable function f satisfies the Polyak-Łojasiewicz condition (for brevity, we write PL-condition) for some constant $\mu > 0$:*

$$f(x) - f^* \leqslant \frac{1}{2\mu}\|\nabla f(x)\|^2 \quad \forall\, x \in \mathbb{R}^n, \tag{4}$$

where $f^* = f(x_*)$ is the value of the function f at one of the exact solutions x_* of the optimization problem under consideration.

The first present of this condition was in Polyak (1963). Recently (see Karimi et al. (2016); Belkin (2021)), it was proven that many practical problems satisfy this condition. Especially, it holds for over-parameterized non-linear systems.

Moreover, we propose results about works of well-known Conjugate-Gradient methods for more weak condition that PL-condition - quadratic growth condition (see Karimi et al. (2016)).

Definition 3. *The differentiable function f satisfies the quadratic growth condition (for brevity, we write QC-condition) for some constant $\mu > 0$:*

$$f(x) - f^* \geqslant \frac{\mu}{2}\|x - x^*\|^2 \quad \forall\, x \in \mathbb{R}^n, \tag{5}$$

where $f^* = f(x_*)$ is the value of the function f at one of the exact solutions x_* of the optimization problem under consideration.

2 SUBSPACE OPTIMIZATION METHODS

Let us present SESOP (Sequential Subspace Optimization) method from S. Guminov (2008); Narkiss & Zibulevsky (2005); Kuruzov & Stonyakin (2021) (see Algorithm 1). The first step of this method is constructing of subspace for further optimization. On each iteration there are three important directions: gradient at current point, direction from start point to current and weighted sum of gradients from all iterations. Note, all this directions can be calculated with only $x^k, x^0, g(x^k)$. In other words, this method does not require too much additional memory in comparison with other first order methods like Gradient Descent.

Algorithm 1 A modification of the SESOP method with an inexact gradient

Require: objective function f with an inexact gradient g, initial point x_0, number of iterations T.

1: $w_0 = 1$
2: **for** $k = 0, \dots, T-1$ **do**
3: Construct Subspace: $\mathbf{d}_k^0 = g(x_k), \quad \mathbf{d}_k^1 = x_k - x_0, \quad \mathbf{d}_k^2 = \sum\limits_{i=0}^{k} \omega_i g(x_i)$.
4: Find the optimal step

$$\tau_k \leftarrow \arg\min_{\tau \in \mathbb{R}^3} f\left(x_k + \sum_{i=1}^{3} \tau_1 d_k^{i-1}\right) \tag{6}$$

5: $x_{k+1} \leftarrow x_k + \sum\limits_{i=1}^{3} \tau_1 d_k^{i-1}$
6: Update $w : w_{k+1} = \frac{1}{2} + \sqrt{\frac{1}{4} + w_k^2}$
return x_T

The next step is optimization on three-dimensional subspace. It is the most complex step. Note, even if the original problem is quasar convex, the auxiliary problem may not have good properties. It is the bottleneck of this method, and it will be discussed in Section 4 for convex case.

The final steps are calculation of new point and update of new weight for direction of weighted gradient sum. This step does not require additional computations. The convergence per iterations for this method is presented in Narkiss & Zibulevsky (2005). This result were recently generalized for quasar-convex case S. Guminov (2008). Further, it was proven that this method is robust for inexactness in gradient in Kuruzov & Stonyakin (2021).

Algorithm 2 A modification of Nemirovski's Conjugate Gradient method with an inexact gradient

Require: objective function f with an inexact gradient g, initial point x_0, number of iterations T.

1: $\mathbf{q}_0 = 0$
2: **for** $k = 1, \dots, T-1$ **do**
3: Solve 2-dimensional problem

$$\hat{x}_k \leftarrow \arg\min_{x \in \mathcal{X}_k} f(x), \quad \text{where } \mathcal{X}_k = x_0 + \text{Lin}(x_k - x_0, \mathbf{q}_k) \tag{7}$$

4: Make gradient step: $x_k = \hat{x}_k - \frac{1}{2L} g(\hat{x}_k)$
5: $\mathbf{q}_k = \mathbf{q}_{k-1} + g(\hat{x}_k)$
return x_T

Another considered in S. Guminov (2008) method was Nemirovski's Conjugate Gradient Method (see Algorithm 2). It is well-known method with enough high performance in practice. Besides, it is known that this method has close form for quadratic minimization problem and it is optimal for them. There are different variants of generalization of such method for non-quadratic problem. In this paper, we consider Nemirovski's Conjugate Gradient Method Nemirovsky & Yudin (1979a). In Nemirovsky & Yudin (1979a) theoretical convergence rate of CG was consequence of the following properties: 1) smoothness of function, 2) strong-convexity of function, 3) orthogonality of gradient at current point and direction from start point to current, 4) orthogonality of gradient at current point and sum of gradients from all previous iterations.

In S. Guminov (2008), it was proven that this method converge if to replace strong-convexity condition by quasar-convexity and quadratic growth condition. The last two conditions are consequence of optimization on 2-dimensional subspace (see Step 3 in Algorithm 2). Nevertheless, this orthogonality will be inexact in the case of inexact solution of the auxiliary subproblem. The next section is devoted to this problem

3 CONVERGENCE OF CG METHOD

3.1 CONVERGENCE WITH INEXACT GRADIENT

In work S. Guminov (2008) the authors obtained the result for convergence rate of Nemirovski's Conjugate Gradient Method. In this work, we show that the method 2 can work with additively inexact gradient too when its inexactness is not large. To do this, we need the following auxiliary lemma.

Lemma 4. *Let the objective function f be L-smooth and γ-quasar-convex with respect to x^*. Also for the inexact gradient $g : \mathbb{R}^n \to \mathbb{R}^n$ there is some constant $\delta_1 \geq 0$ such that for all $x \in \mathbb{R}^n$:*

$$\|g(x) - \nabla f(x)\| \leq \delta_1. \tag{8}$$

Then the following inequality holds:

$$\|q_T\| \leq 3\delta_1 T + \left(\sum_{k=0}^{T} \|g(\hat{x}_k)\|^2\right)^{\frac{1}{2}}. \tag{9}$$

Using Lemma 4 we can generalize the result of Theorem 2 from work S. Guminov (2008) for the case of inexact gradient. Finally, we have the following result.

Theorem 5. *Let the objective function f be L-smooth and γ-quasar-convex with respect to x^*. Also for the inexact gradient $g : \mathbb{R}^n \to \mathbb{R}^n$ there is some constant $\delta_1 \geq 0$ such that for all $x \in \mathbb{R}^n$ $\|g(x) - \nabla f(x)\| \leq \delta_1$. Moreover, function satisfied condition of quadratic growth $f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|$. Then if on all iterations $\|g(\hat{x}_k)\| \geq 2\delta_1$ the CG obtain x_T such that*

$$f(x_T) - f^* \leq \beta\epsilon_0 + \frac{4}{\gamma}\sqrt{\frac{2\varepsilon_0}{\mu}}\delta_1. \tag{10}$$

after

$$T = \left\lceil \frac{2}{\gamma\beta}\sqrt{\frac{2(1-\beta)L}{\mu}} \right\rceil.$$

iterations, where $R = \|x^* - x_0\|$, $\epsilon_0 = f(x_0) - f^*$ and any constant parameter $\beta \in (0, 1)$.

Remark 6. *Note, that quadratic growth condition is met when the object function f satisfies well-known PL-condition equation 4. It's well known Nesterov & Polyak (2006); Karimi et al. (2016); Gasnikov (2017) that under additional smoothness assumptions standard non-accelerated iterative methods for such functions(Gradient Descent, Cubic Regularized Newton method etc.) converge as if f to be μ-strongly convex function. For accelerated methods, such results are not known. So we were motivated to find such additional sufficient conditions that guarantee convergence for properly chosen accelerated methods. In this section we observe that such a condition could be α-weakly-quasi-convexity of f.*

Remark 7. *Note, if function f meets PL-condition equation 4 and we will stop our method when $\|g(x_k)\| \leq 2\delta_1$ then we have that $f(x_k) - f^* \leq \frac{4\delta_1^2}{\mu}$. Note, that in work Vorontsova E.A. & F.S. (2021) authors proved that there are no methods that can converges better than $O\left(\frac{\delta_1^2}{\mu}\right)$ in general case.*

3.2 NEMIROVSKI'S CG METHOD WITH RESTARTS

It is well-known that restart technique can significantly improve convergence of conjugate gradient methods. To use that, we need to run the algorithm for T iterations, and after that start the same algorithm from the final point after T iterations (see Algorithm 3).

Similar to S. Guminov (2008), we can obtain the following result.

Algorithm 3 Restarted Nemirovski's Conjugate Gradient method with an inexact gradient

Require: objective function f with an inexact gradient g, initial point x_0, number of iterations T, number of restarts K.

1: $\mathbf{q}_0 = 0$
2: **for** $k = 1, \ldots, K-1$ **do**
3: Run Algorithm 2 for start point x_{k-1} and T iterations:

$$x_k = \text{CG}(f, g, x_{k-1}, T)$$

return x_K

Theorem 8. *Let the objective function f be L-smooth and γ-quasar-convex with respect to x^*. Also for the inexact gradient $g : \mathbb{R}^n \to \mathbb{R}^n$ there is some constant $\delta_1 \geq 0$ such that for all $x \in \mathbb{R}^n$ $\|g(x) - \nabla f(x)\| \leq \delta_1$ and $\delta_1^2 \leq \frac{\gamma^2\alpha^2\mu\varepsilon}{32}$ for some $\alpha \in (0,1)$. Moreover, function satisfied the condition of quadratic growth $f(x) - f^* \geq \frac{\mu}{2}\|x - x^*\|$. Then if on all iterations condition $\|g(x_k)\| \geq 2\delta_1$ is met the CG obtain outer point $\hat{x}$ such that*

$$f(\hat{x}) - f^* \leq \varepsilon. \tag{11}$$

after

$$K = \left\lceil \frac{2}{1-\alpha} \log \frac{1}{\varepsilon} \right\rceil$$

restarts and

$$T = \left\lceil \frac{8}{\gamma}\sqrt{\frac{L}{\mu}}\frac{\sqrt{1+\alpha}}{1-\alpha} \right\rceil,$$

iterations, where $\epsilon_0 = f(x_0) - f^*$.

Proof. We have that the method may degrade the quality on function for enough large δ_1. At the same time, in the case

$$\frac{4}{\gamma}\sqrt{\frac{2\varepsilon_0}{\mu}}\delta_1 \leq \alpha\varepsilon_0, \tag{12}$$

for some constant $\alpha \in (0,1)$ we have that

$$f(x_T) - f^* \leq \tilde{\beta}\epsilon_0$$

after

$$T = \left\lceil \frac{8}{\gamma}\sqrt{\frac{L}{\mu}}\frac{\sqrt{1+\alpha}}{1-\alpha} \right\rceil.$$

iterations, where $\tilde{\beta} = \frac{1+\alpha}{2}$. Note, that condition equation 12 can be rewritten in the following form:

$$\delta_1^2 \leq \frac{\gamma^2\alpha^2\mu\varepsilon_0}{32}. \tag{13}$$

So, to approach quality ε we need to require condition equation 13 for ε in the following form:

$$\delta_1^2 \leq \frac{\gamma^2\alpha^2\mu\varepsilon}{32}. \tag{14}$$

In this case, after

$$K = \left\lceil \frac{\log \varepsilon}{\log \frac{1-\alpha}{2}} \right\rceil \leq \left\lceil \frac{2}{1-\alpha} \log \frac{1}{\varepsilon} \right\rceil$$

restarts the method obtains a point x_{TK} such that:

$$f(x_{TK}) - f^* \leq \varepsilon.$$

□

Remark 9. *Generally, we can obtain that after K restarts and KT general number of iterations we obtain the point $\hat{x}$ such that*

$$f(\hat{x}_T) - f^* \leq \beta^N \varepsilon_0 + \left(\sum_{j=0}^{N-1} \beta^j \right) \frac{4}{\gamma} \sqrt{\frac{2\epsilon_0}{\mu}} \delta_1,$$

or

$$f(x_T) - f^* \leq \beta^N \epsilon_0 + \frac{4}{\gamma(1-\beta)} \sqrt{\frac{2\epsilon_0}{\mu}} \delta_1.$$

Here we can see it cannot be guaranteed that the Nemirovski's Conjugate Gradient method will converge to a quality better than $O\left(\sqrt{\frac{\varepsilon_0}{\mu}}\delta_1\right)$.

Remark 10. *The algorithm 2 requires the total number of gradient computations $O\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$ to approach quality ε.*

As we mentioned above, the first-order methods can not approach quality better than $O(\frac{\delta_1^2}{\mu})$ for strong-convex function. Consequently, it is true for functions that meet PL-condition equation 4 or quadratic growth condition equation 5. So, let us consider estimate $f(x_k) - f^* \leq \frac{\delta_1^2}{\mu}$ acceptable for the function level and *agree to terminate algorithm 2 if the condition $\|g(x_k)\| \leq \frac{8}{\gamma}\delta_1$ is satisfied.*

Finally, let us state the following results about work of method with stop condition.

Theorem 11. *Let the objective function f be L-smooth and γ-quasar-convex with respect to x^*. Also for the inexact gradient $g : \mathbb{R}^n \to \mathbb{R}^n$ there is some constant $\delta_1 \geq 0$ such that for all $x \in \mathbb{R}^n$ $\|g(x) - \nabla f(x)\| \leq \delta_1$. Moreover, function satisfied PL-condition equation 4.*

Let one of the following alternatives hold:

1. *The Nemirovski's Conjugate Gradient method 2 makes*

$$K = \left\lceil \frac{2}{1-\alpha} \log \frac{1}{\varepsilon} \right\rceil$$

   restarts and

$$T = \left\lceil \frac{8}{\gamma} \sqrt{\frac{L}{\mu}} \frac{\sqrt{1+\alpha}}{1-\alpha} \right\rceil,$$

   iterations per each restart, where $\varepsilon = \frac{64}{\gamma^2 \mu}\delta_1^2$

2. *For some iteration $N \leq N^*$, at the N-th iteration of Nemirovski's Conjugate Gradient method 2, stopping criterion $\|g(x_N)\| \leq \frac{8}{\gamma}\delta_1$ is satisfied for the first time.*

Then for the output point $\widehat{x}$ ($\widehat{x} = x_N$ or $\widehat{x} = x_{N_*}$) of Nemirovski's Conjugate Gradient method 2, the following inequalities hold:

$$f(\widehat{x}) - f^* \leqslant \frac{64\delta_1^2}{\gamma^2 \mu},$$

We can see that restarts technique allows obtaining optimal convergence rate for considered non-convex case. Nevertheless, in this case we have additional parameter for tuning - frequency of restarts.

4 Auxiliary Low-Dimensional Subproblems

In this section, we assume that all auxiliary subproblems in Algorithms 1 and 2 are convex. Namely, subproblems in step 4 of Algorithm 1 and in step 3 of Algorithm 2 are convex.

4.1 ELLIPSOID METHODS FOR SESOP

To estimate the number of gradient calculations, we need to choose some procedure for optimization on subspace on the second string of algorithm 1. It is the three-dimensional problem, so we can use some methods for low-dimensional problems. Examples of such methods are ellipsoid method (see Nemirovski et al. (2010)), Vaidya method (see Vaidya (1996)) and Dichotomy methods for hypercube (see Gladin et al. (2020)). For all these methods there are results of method works with inexact gradient (see Gladin et al. (2020)). The Dichotomy method has worse estimate for number of calculations than other methods. Nevertheless, it demonstrates enough good performance for two dimensional case. Therefore, we consider it for CG method below. The Ellipsoid and Vaid's methods require $O\left(\log \frac{1}{\varepsilon}\right)$ number of gradient calculations. So in current work we chose Ellipsoid method (see 4) for subproblem.

For Ellipsoid Methods there is the following estimate (see Theorem 2 in Gladin et al. (2020)). If algorithm 4 was run on a ball $\mathcal{B} \subset \mathbb{R}^n$ in n-dimensional space of radius R, the constant B is such that $\max_x f(x) - \min_x f(x) \leq B$ then the Ellispoid method with δ-subgradient converges to solution with the following speed:

$$f(x_N) - f(x^*) \leq B \exp\left(-\frac{N}{2n^2}\right) + \delta \tag{15}$$

In our case n is equal to 3, dimension of subproblem. So, according to equation 15 when $\delta \leq \frac{\varepsilon}{2}$ to approach the quality ε we need

$$N \geq 18 \ln \frac{2B}{\varepsilon} \tag{16}$$

iterations of method 4.

In this section, we will estimate the work of SESOP algorithm in two modes:

- One has exact low-dimensional gradient (gradient for subproblem) but there is only inexact gradient for full problem
- One has only inexact gradient both in low-dimensional problem and full problem

Theorem 12. *Let inexact gradient required condition equation 8 with $\delta_1 \leq \frac{\varepsilon}{\frac{R}{\gamma}+10}$. Also, let us assume that we have the ball $\mathcal{B}_R^k \subset \mathbb{R}^3$ with radius R on each iteration such that $\tau_k \in \mathcal{B}_R^j$. If we can use exact gradient of function f_k than to approach quality ε on initial problem by SESOP method one requires not more than*

$$N = \left\lceil \sqrt{\frac{40LR^2}{\gamma^2 \varepsilon}} \right\rceil$$

of inexact gradient calculations with respect to x and not more than

$$M = \left\lceil 18N \ln \frac{12800LBC_N}{\varepsilon^4} \right\rceil$$

of exact gradient calculations with respect to τ where

$$B = \max_{k=\overline{1,N}} \max_{\tau \in \mathcal{B}_R^j} f_j(\tau) - f^*$$

$$C_N = 1 + \sqrt{\max_{k=\overline{1,N}} (\|D_k\| \|\tau_k\|)} + \sqrt{\| \max_{k=\overline{1,N}} \mathbf{d}_{k-1}^1 \|} + \max_{k=\overline{1,N}} \|\mathbf{d}_k^3\|.$$

Remark 13. *Note, that the SESOP in such implementation requires $O\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$ inexact gradient calculations with respect to x and $O\left(\sqrt{\frac{LR^2}{\varepsilon}} \ln \frac{1}{\varepsilon}\right)$ inexact gradient calculations with respect to τ.*

Remark 14. *The main theoretical advantage of SESOP with inexact gradient is that there is no additive part depends on $\max_k R_k$ as in early works. It approaches through solving additional low-dimensional subproblem. Nevertheless, it leads to requirements for high accuracy of solution of auxiliary problem.*

Further, let us consider the case of inexact gradient in internal problems. In this case, the quality of subproblem solution can not be better than inexactness of gradient.

Theorem 15. *Let us assume that we have the ball $\mathcal{B}_R^k \subset \mathbb{R}^3$ with radius R on each iteration such that $\tau_k \in \mathcal{B}_R^j$. Let inexact gradient require condition equation 8 with*

$$\delta_1 \leq \min\left\{\frac{\varepsilon}{\frac{R}{\gamma}+10}, \frac{\varepsilon^4}{6400A_kL}\right\} \tag{17}$$

where $A_N = C_N \max_{k=\overline{1,N}} \|D_k\|_2$ for C_k defined as in Theorem 12.

Then to approach quality ε on initial problem by SESOP method one requires not more than

$$N = \left\lceil \sqrt{\frac{40LR^2}{\gamma^2\varepsilon}} \right\rceil$$

of inexact gradient calculations with respect to x and not more than

$$M = \left\lceil 18N \ln \frac{12800LBC_N}{\varepsilon^4} \right\rceil$$

of inexact gradient calculations with respect to τ.

We can see that the number of gradient calculation is almost the same as in previous theorem in case of exact low-dimensional gradient but in this case we have significantly more strong conditions equation 17. This condition allows testing inequalities equation 37 and equation 38 and to approach quality equation 39 in subproblem. But in this case we can see that the inexactness gradient should be not more than $O(\varepsilon^4)$.

Note, that the advantages of SESOP method leads to requirements for extra low inexactness for both gradient inexactness and solution of auxiliary problem. Nevertheless, the required inexactness can be easily controlled through values $\mathbf{d}_k^J$ during the algorithm work.

4.2 Multidimensional Dichotomy

Ellipsoid method can be applied for problems in CG method too. Nevertheless, in Gladin et al. (2020) it was shown that there is another effective method for two-dimensional subproblem. It is generalization of one-dimensional dichotomy. Despite the little worse convergence rate in comparison with Ellipsoid method, it demonstrates better performance. So, we provide the result for CG method with two-dimensional dichotomy in the following theorem.

Theorem 16. *Let assumptions of Theorem 11 hold and all subproblems are convex. Besides, there is R such that $\hat{x}_k - x_k \in B_{R_x}$ for all k. Each point $\hat{x}_k$ is output of two-dimensional dichotomy algorithm (see Gladin et al. (2020)) after M steps, where M is given by:*

$$M = \left\lceil 16 \left(\ln \frac{CR_x}{\varepsilon^4} \right)^2 \right\rceil$$

Let one of the following alternatives hold:

1. *The Nemirovski's Conjugate Gradient method 2 makes*

$$K = \left\lceil \frac{2}{1-\alpha} \log \frac{1}{\varepsilon} \right\rceil$$

   restarts and

$$T = \left\lceil \frac{8}{\gamma} \sqrt{\frac{L}{\mu}} \frac{\sqrt{1+\alpha}}{1-\alpha} \right\rceil,$$

   iterations per each restart, where $\varepsilon = \frac{64}{\gamma^2\mu}\delta_1^2$

2. *For some iteration $N \leq N^*$, at the N-th iteration of Nemirovski's Conjugate Gradient method 2, stopping criterion $\|g(x_N)\| \leq \frac{8}{\gamma}\delta_1$ is satisfied for the first time.*

Then for the output point $\widehat{x}$ ($\widehat{x} = x_N$ or $\widehat{x} = x_{N_*}$) of Nemirovski's Conjugate Gradient method 2, the following inequalities hold:

$$f(\widehat{x}) - f^* \leqslant \frac{64\delta_1^2}{\gamma^2\mu}.$$

As the result the algorithm requires not more N of calculations of inexact gradient with respect to x and $MN = O(\ln^3(1/\varepsilon))$ of low-dimensional inexact gradient calculations.

5 Numerical Experiments

In this section, we present some preliminary numerical results. We compare SESOP method with Ellipsoid methods for auxiliary subproblem (see Algorithms 1 and 4), CG with restarts with different methods for subproblems (see Algorithms 3, 4 and Gladin et al. (2020))) and Similar Triangle method (see Gasnikov & Nesterov (2018)).

We consider the problem of logistic regression:

$$f(x) = (1/m)\sum_{j=1}^{m} \log(1 + \exp(-y_j\langle f_j, x\rangle)) + \mu\|x\|^2 \tag{18}$$

on synthetic data in dimensions $n = 100, m = 200$. In all cases we considered constant inexactness with different norm δ_1. Parameter for restarts of CG, Lipschitz constant and strong convexity parameter were found analytically. It was found that all method approach the close accuracy in our problem settings. So, in this work we demonstrate time comparison (see 1). The presented time is time required to approach quality $10\delta_1^2/\mu$

δ_1	SESOP	CG+Ellipsoids	CG+Dichotomy	STM
10^{-3}	1	1.4	0.9	1.7
10^{-5}	10.1	15.3	9.5	13.8
10^{-7}	35.3	60.9	36.8	42.1

Table 1: Time comparison (s) for problem equation 18

We can see that multidimensional dichotomy works better than Ellipsoid method for Nemirovski's Conjugate Gradient Method in all cases. At the same time, methods based on subspace optimization outperforms STM method. The result for CG and SESOP method are close enough.

Conclusion

In this paper, one considered generalization of convexity condition that is known as quasar-convexity or weakly-quasi-convexity. We propose modification of Nemirovski's Conjugate Gradient Method with a δ-additive noise in the gradient equation 2 for γ-quasar convex functions satisfying quadratic growth condition.

We estimate computational complexity for solving the internal auxiliary problem in SESOP and CG method. For this, we used well-known low-dimensional optimization methods – Ellipsoid Method and generalization of dichotomy. We prove that these methods do not significantly increase complexity for convex case. Besides, these methods are still the methods of the first order.

Moreover, we provide numerical experiments which demonstrate the effectiveness of the approaches proposed in this paper.

References

Mathieu Barre, Adrien Taylor, and Alexandre d'Aspremont. Complexity guarantees for polyak steps with momentum, 02 2020.

A. Beck. *First-Order Methods in Optimization. Society for Industrial and Applied Mathematics.* 01 2017.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 08 2021. doi: 10.1017/S0962492921000039.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8:231–357, 01 2015. doi: 10.1561/2200000050.

O. Devolder. Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization. *Ph.D. thesis, ICTEAM and CORE, Universit'e Catholique de Louvain*, 01 2013.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization 19(3)*, M 146(1):1171–1183, 01 2008.

Alexandre d'Aspremont, Damien Scieur, and Adrien Taylor. Acceleration methods. 5:1–245, 01 2021. doi: 10.1561/2400000036.

Alexander Gasnikov. Universal gradient descent. 11 2017.

Alexander Gasnikov and Yu Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58:48–64, 01 2018. doi: 10.1134/S0965542518010050.

Egor Gladin, Ilya Kuruzov, Fedor Stonyakin, Dmitry Pasechnyuk, Mohammad Alkousa, and Alexander Gasnikov. Solving strongly convex-concave composite saddle point problems with a small dimension of one of the variables, 10 2020.

I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. *MIT Press, `http://www.deeplearningbook.org`*, 01 2016.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19, 09 2016.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond, 06 2020.

Sergey Kabanikhin. Inverse and ill-posed problems: Theory and applications. 01 2011. doi: 10.1515/9783110224016.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition. volume 9851, pp. 795–811, 09 2016. ISBN 978-3-319-46127-4. doi: 10.1007/978-3-319-46128-1_50.

Ilya Kuruzov and Fedor Stonyakin. *Sequential Subspace Optimization for Quasar-Convex Optimization Problems with Inexact Gradient*, pp. 19–33. 12 2021. ISBN 978-3-030-92710-3. doi: 10.1007/978-3-030-92711-0_2.

W¿ Haskell L. Hien, R. Zhao. An inexact primal-dual smoothing framework for large-scale non-bilinear saddle point problems. *ournal of Optimization Theory and Applications. 10.1007/s10957-023-02351-9*, 01 2023.

G. Lan. First-order and stochastic optimization methods for machine learning. *Switzerland: Springer Series in the Data Sciences*, 01 2020.

Guy Narkiss and Michael Zibulevsky. Sequential subspace optimization method for large-scale unconstrained problems. 01 2005.

Arkadi Nemirovski, Shmuel Onn, and Uriel Rothblum. Accuracy certificates for computational problems with convex structure. *Math. Oper. Res.*, 35:52–78, 02 2010. doi: 10.1287/moor.1090.0427.

A.S. Nemirovsky and D.B. Yudin. Problem complexity and optimization method efficiency [in russian]. *Nauka, Moscow*, 11 1979a.

A.S. Nemirovsky and D.B. Yudin. Problem complexity and optimization method efficiency. *Moscow, Nauka*, 01 1979b.

Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 108:177–205, 08 2006. doi: 10.1007/s10107-006-0706-8.

F. Glineur O. Devolder and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming M*, M 146(1):37–75, 01 2014.

B.T. Polyak. Gradient methods for minimizing functionals. *Comput. Math. Math. Phys., 3:4*, pp. 864 — 878, 01 1963.

B.T. Polyak. Introduction to optimization. *Optimization Software*, 01 1987.

I. Kuruzov S. Guminov, A. Gasnikov. Accelerated methods for weakly-quasi-convex optimization problems. *Computational Management Science. 20. 10.1007/s10287-023-00468-w*, 01 2008.

Pravin Vaidya. New algorithm for minimizing convex functions over convex sets. *Math Program*, 73:291–341, 01 1996. doi: 10.1007/BF02592216.

F. Vasilyev. Optimization methods. *Moscow, Russia: FP*, 01 2002.

Artem Vasin, Alexander Gasnikov, and Vladimir Spokoiny. Stopping rules for accelerated gradient methods with additive noise in gradient, 02 2021.

Gasnikov A.V. Vorontsova E.A., Hildbrand R.F. and Stonyakin F.S. Convex optimization. *Moscow, MIPT*, pp. 251, 11 2021.

Jun-Kun Wang and Andre Wibisono. Continuized acceleration for quasar convex functions in non-convex optimization, 02 2023.

A CONVERGENCE OF CG METHOD

A.1 PROOF OF LEMMA 4

Proof. Note, that $q_T = q_{T-1} + \frac{1}{L} g(\hat{x}_T)$ for $T \geq 1$. So, we have the following expression for $\|q_T\|$:

$$\|q_T\|^2 = \|g(\hat{x}_T)\|^2 + \|q_{T-1}\|^2 + 2\langle g(\hat{x}_T), q_{T-1}\rangle.$$

Because of exact solution of auxiliary problem on each iteration, we have that

$$\nabla f(\hat{x}_T) \perp q_{T-1}.$$

At the same time, we have that $\|\nabla f(\hat{x}_T) - g(\hat{x}_T)\|^2 \leq \delta_1$. So, we have the following estimations:

$$\|q_T\|^2 \leq \sum_{k=0}^{T} \|g(\hat{x}_k)\|^2 + 2\delta_1 \sum_{k=0}^{T-1} \|q_k\|, \tag{19}$$

and

$$\|q_T\|^2 \geq \|q_{T-1}\|^2 - 2\delta_1 \|q_{T-1}\|. \tag{20}$$

From the lower bound equation 20, we have that

$$\|q_{T-1}\| \leq \delta_1 + \sqrt{\delta_1^2 + \|q_T\|^2},$$

or

$$\|q_{T-1}\| \leq 2\delta_1 + \|q_T\|.$$

Using the inequality above, we can obtain estimation for $\|q_j\|$ for all j:

$$\|q_j\| \leq 2\delta_1 (T - j) + \|q_T\|. \tag{21}$$

Using inequalities equation 19 and equation 21, we have the following estimation:

$$\|q_T\|^2 \leq \sum_{k=0}^{T} \|g(\hat{x}_k)\|^2 + 4\delta_1^2 \sum_{i=0}^{T-1} (T - j) + 2\delta_1 T \|q_T\|, \tag{22}$$

and from equation 22 we have the following quadratic inequality on $\|q_T\|$:

$$\|q_T\|^2 \leq \sum_{k=0}^{T} \|g(\hat{x}_k)\|^2 + 2\delta_1^2 T^2 + 2\delta_1 T \|q_T\|, \tag{23}$$

Using inequality equation 23 we obtain the estimation equation 9:

$$\|q_T\| \leq 3\delta_1 T + \left(\sum_{k=0}^{T} \|g(\hat{x}_k)\|^2 \right)^{\frac{1}{2}}.$$

□

A.2 PROOF OF THEOREM 5

Proof. Let us assume that $\varepsilon_T = f(x_T) - f^* \geq \beta \varepsilon_0 + c\delta_1$, where $\beta \in (0, 1)$ and c are some constants. Using estimation equation 31 for $s_0 = \frac{1}{2L}$ we obtain the following estimation:

$$\|g(\hat{x}_k)\|^2 \leq 4L(f(\hat{x}_k) - f(x_{k+1})) + 2\delta_1.$$

Because of exact solution of auxiliary problem, we have that

$$\|g(\hat{x}_k)\|^2 \leq 4L(\varepsilon_k - \varepsilon_{k+1}) + 2\delta_1. \tag{24}$$

Telescoping inequality above, we obtain the following inequality:

$$\sum_{k=0}^{T-1} \|g(\hat{x}_k)\|^2 \leq 4L(\varepsilon_0 - \varepsilon_T) \leq 4L(1 - \beta)\varepsilon_0. \tag{25}$$

On the other hand, from quasar-convexity we have the following estimation:

$$f(\hat{x}_k) - f^* \leq \frac{1}{\gamma}\langle \nabla f(\hat{x}_k), \hat{x}_k - x^* \rangle.$$

Because of exact solution of auxiliary problem. we have the following inequality:

$$f(\hat{x}_k) - f^* \leq \frac{1}{\gamma}\langle \nabla f(\hat{x}_k), x_0 - x^* \rangle.$$

Similarly to proof of results for SESOP, we obtain the final estimations:

$$f(\hat{x}_k) - f^* \leq \frac{1}{\gamma}\langle \nabla g(\hat{x}_k), x_0 - x^* \rangle + \delta_1 \frac{R}{\gamma}.$$

Note, that from equation 24 and condition $\|g(\hat{x}_k)\| \geq 2\delta_1$, we have that $f(x_{k+1}) \leq f(\hat{x}_k)$. By construction of $\hat{x}_k$, we have that $f(\hat{x}_k) \geq f(x_k)$. So, when $\varepsilon_T = f(x_T) - f^* \geq \beta\varepsilon_0 + c\delta_1$ the following inequality holds:

$$\beta\varepsilon_0 + c\delta_1 \leq \frac{1}{\gamma}\langle \nabla g(\hat{x}_k), x_0 - x^* \rangle + \delta_1 \frac{R}{\gamma}.$$

When one sum up this inequalities above, we obtain:

$$T\beta\varepsilon_0 + cT\delta_1 \leq \frac{1}{\gamma}\langle q_T, x_0 - x^* \rangle + \delta_1 \frac{RT}{\gamma},$$

$$-\|q_T\|\|x_0 - x_*\| \leq -T\gamma\beta\varepsilon_0 + \delta_1 T\left(R + c\gamma\right). \tag{26}$$

Firstly, from Lemma 4 and inequality equation 25, we have that:

$$\|q_T\| \leq 3\delta_1 T + \sqrt{4L(1-\beta)\varepsilon_0 - 4Lc\delta_1}. \tag{27}$$

On the other hand, from quadratic growth we can obtain the following estimation for $\|x_0 - x_*\|$:

$$\|x_0 - x_*\| \leq \sqrt{\frac{2\varepsilon_0}{\mu}}. \tag{28}$$

Uniting inequalities equation 26-equation 28 we obtain the following inequalities for T:

$$\left(3\delta_1 T + \sqrt{4L(1-\beta)\varepsilon_0}\right)\sqrt{\frac{2\varepsilon_0}{\mu}} \geq T\gamma\beta\varepsilon_0 + \delta_1 T\left(-R + c\gamma\right) \tag{29}$$

Let us rewrite equation 29 in the following form:

$$T\gamma\beta\varepsilon_0 + \delta_1 T\left(c\gamma - R - 3\sqrt{\frac{2\varepsilon_0}{\mu}}\right) \leq 2\varepsilon_0\sqrt{\frac{2(1-\beta)L}{\mu}}.$$

So, when $c \geq \frac{1}{\gamma}\left(R + 3\sqrt{\frac{2\varepsilon_0}{\mu}}\right)$, we have the following estimation on T:

$$T \leq \frac{2}{\gamma\beta}\sqrt{\frac{2(1-\beta)L}{\mu}}.$$

So, after $T = \left\lceil \frac{2}{\gamma\beta}\sqrt{\frac{2(1-\beta)L}{\mu}} \right\rceil$, we have that $\varepsilon_T \leq \beta\varepsilon_0 + c\delta_1$. Finally, note, that because of quadratic growth we have estimation $\frac{\mu R^2}{2} \geq \varepsilon_0$. So, we have that after T iterations we have a point x_T that meets the following estimation:

$$f(x_T) - f^* \leq \beta\varepsilon_0 + \frac{4}{\gamma}\sqrt{\frac{2\varepsilon_0}{\mu}}\delta_1.$$

□

B SOME RESULTS FOR SESOP METHOD

On the base of the last inequality and the right part of equation 32 for $\mathbf{y} := \mathbf{x}_k + s_0 g(\mathbf{x}_k)$ and $\mathbf{x} = \mathbf{x}_k$ we can conclude that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \left(s_0 + s_0^2 L\right) \|g(\mathbf{x}_k)\|^2 + \frac{1}{2L}\delta_1^2 \tag{30}$$

for each $s_0 \in \mathbb{R}$. Further,

$$-\left(s_0 + s_0^2 L\right) \|g(\mathbf{x}_k)\|^2 \leq f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) + \frac{1}{2L}\delta_1^2. \tag{31}$$

$$f(\mathbf{x}_{k+1}) = \min_{\mathbf{s}\in\mathbb{R}^3} f\left(\mathbf{x}_k + \sum_{i=0}^{2} s_i d_k^i\right) \leq f\left(\mathbf{x}_k + s_0 g(\mathbf{x}_k)\right). \tag{32}$$

Theorem 17. *Let the objective function f be L-smooth and γ-quasar-convex with respect to $\mathbf{x}^*$. Let τ_k be the step value obtained with the inexact solution of the auxiliary problem equation 6 on step 2 in Algorithm 1 on the k-th iteration. Namely, the following conditions for inexactness hold:*

(i) For the inexact gradient $g : \mathbb{R}^n \to \mathbb{R}^n$ there is some constant $\delta_1 \geq 0$ such that for all points $\mathbf{x} \in \mathbb{R}^n$ condition equation 8 holds.

(ii) The inexact solution τ_k meets the following condition:

$$|\left\langle \nabla f(\mathbf{x}_k), \mathbf{d}_{k-1}^2 \right\rangle| \leq k^2 \delta_2 \tag{33}$$

for some constant $\delta_2 \geq 0$ and each $k \in \mathbb{N}$. Note that $\mathbf{x}_k = \mathbf{x}_{k-1} + D_{k-1}\tau_{k-1}$.

(iii) The inexact solution τ_k meets the following condition for some constant $\delta_3 \geq 0$:

$$|\left\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}_0 \right\rangle| \leq \delta_3. \tag{34}$$

(iv) The problem from step 2 in Algorithm 1 is solved with accuracy $\delta_4 \geq 0$ on the function on each iteration, i.e. $f(\mathbf{x}_k) - \min_{\tau\in\mathbb{R}^n} f(\mathbf{x}_{k-1} + D_{k-1}\tau) \leq \delta_4$.

Then the sequence $\{\mathbf{x}_k\}$ generated by Algorithm 1 satisfies

$$f(\mathbf{x}_k) - f^* \leq \frac{8LR^2}{\gamma^2 k^2} + \left(\frac{R}{\gamma} + 10\right)\delta_1 + 4\sqrt{\delta_2} + \delta_3 + 5\sqrt{\frac{L\delta_4}{k}} \tag{35}$$

for each $k \geq 8$, where $R = \|\mathbf{x}^* - \mathbf{x}_0\|$.

Theorem 18. *If condition (iv) from Theorem 17 holds, then we can choose $\delta_2, \delta_3 \geq 0$ according to the following estimates:*

$$\delta_3 \leq \sqrt{2L\delta_4}\left(\sqrt{\max_k(\|D_k\|\|\tau_k\|)} + \sqrt{\|\max_k \mathbf{d}_{k-1}^1\|}\right)$$

and

$$\delta_2 \leq \frac{1}{k^2}\sqrt{2L \max_k \|\mathbf{d}_k^3\|\delta_4}.$$

C AUXILLARY LOW-DIMENSIONAL SUBPROBLEMS

C.1 PROOF OF THEOREM 12

In the first case we will estimate the number of gradient calculations with respect to x and to τ separately. In the second case we will estimate this gradients calculation in total. In the second case we can get the inexact gradient for subproblem through full gradient calculation:

$$\frac{d}{d\tau} f(x_k + D_k\tau) = D_k^\top \nabla f(x)\Big|_{x=x_k+D_k\tau}$$

for all k. In such scheme, any low-dimensional gradient calculation requires high-dimensional gradient calculation because we can calculate total number of calculations.

Theorem 18 gives correspondence between inaccuracies δ_2, δ_3 and δ_4:

$$\delta_3 \leq \sqrt{2L\delta_4}\left(\sqrt{\max_k(\|D_k\|\|\tau_k\|)} + \sqrt{\|\max_k \mathbf{d}_{k-1}^1\|}\right)$$

and

$$\delta_2 \leq \frac{1}{k^2}\sqrt{2L\max_k \|\mathbf{d}_k^3\|\delta_4}.$$

The condition

$$\max\left\{\frac{8LR^2}{\gamma^2 k^2}, \left(\frac{R}{\gamma}+10\right)\delta_1, 4\sqrt{\delta_2}, \delta_3, 5\sqrt{\frac{L\delta_4}{k}}\right\} \leq \frac{\varepsilon}{5}, \tag{36}$$

is sufficient to approach quality ε on function. So, according to equation 36 we have the following conditions for inexactness $\delta_2, \delta_3, \delta_4$:

$$\delta_2 \leq \frac{\varepsilon^2}{400} \tag{37}$$

$$\delta_3 \leq \frac{\varepsilon}{5} \tag{38}$$

$$\delta_4 \leq \frac{\varepsilon^2}{625L} \tag{39}$$

So we obtain the main statement about quality of subproblem solution and condition for iterations count and inexactness of gradient.

Lemma 19. *Let inexact gradient meets condition equation 8 with*

$$\delta_1 \leq \frac{\varepsilon}{\frac{R}{\gamma}+10}.$$

To obtain quality ε on function the SESOP method 1 with inexact gradient should have

$$T \geq \sqrt{\frac{40LR^2}{\gamma^2\varepsilon}},$$

iterations and on each iteration subproblem should be solved such that conditions equation 37, equation 38 and equation 39 are met.

Moreover, uniting the conditions equation 37-equation 39 with early obtained results of Theorem 18 we can obtain sufficient quality of subproblem for obtaining these results.

Lemma 20. *The following quality of subproblem on kth iteration is sufficient for conditions equation 37, equation 38 and equation 39 are met:*

$$\delta_4 \leq \min\left\{\frac{\varepsilon^4}{6400L\left(\sqrt{\max_k(\|D_k\|\|\tau_k\|)} + \sqrt{\|\max_k \mathbf{d}_{k-1}^1\|}\right)}, \frac{\varepsilon^2}{50L\max_k\|\mathbf{d}_k^3\|}, \frac{\varepsilon^2}{625L}\right\}.$$

In other words, we need to solve on each iteration auxiliary problem with accuracy $O\left(\frac{\varepsilon^4}{LC_k}\right)$ where constant $C_k = 1 + \sqrt{\max_k(\|D_k\|\|\tau_k\|)} + \sqrt{\|\max_k \mathbf{d}_{k-1}^1\|} + \max_k\|\mathbf{d}_k^3\|$ is defined by generated by algorithm sequence $\{x_k\}$.

This quality is the worst case when the subproblem procedure will be stopped. One supposes that for the most problems conditions equation 37, equation 38 and equation 39 will be met significantly early.

The statements above are true for the both modes, and they allow estimating a number of exact and inexact gradients in these cases. In the following theorem, we suppose that we have the ball $\mathcal{B}_R^k \subset \mathbb{R}^3$ with radius R on each iteration such that $\tau_k \in \mathcal{B}_R^j$. So for the first mode we have the following estimations. In this case, to approach quality

$$\delta_4 = \frac{\varepsilon^4}{6400 C_k L}$$

where $C_k = 1 + \sqrt{\max_k(\|D_k\|\|\tau_k\|)} + \sqrt{\|\max_k \mathbf{d}_{k-1}^1\|} + \max_k \|\mathbf{d}_k^3\|$ we need

$$18 \ln \frac{12800 L B_k C_k}{\varepsilon^4}$$

iterations of ellispoid method where $B_k = \max_{\tau \in \mathcal{B}_R^j} f_j(\tau) - f^*$.

C.2 Proof of Theorem 15

Proof. Let us estimate inexactness in internal problems. Early, we obtained the following equality:

$$\frac{d}{d\tau} f(x_k + D_k \tau) = D_k^\top \nabla f(x)\Big|_{x = x_k + D_k \tau}$$

So, let us consider ellipsoid method with inexact gradient in the following form:

$$g_k(\tau) = D_k^\top g(x_k + D_k \tau).$$

For such gradient we have the following estimation for inexactness:

$$\|g_k(\tau) - \nabla_\tau f_k(\tau)\| \leq \|D_k\|_2 \delta_1. \tag{40}$$

So to approach quality ε we need the more hard conditions for inexactness of gradient. To approach quality δ_4 by ellispoid method, we need that the following condition holds:

$$\delta_1 \leq \frac{\delta_4}{\|D_k\|_2}.$$

Under assumptions of Theorem 12, we have that δ_4 is given by

$$\delta_4 = \frac{\varepsilon^4}{6400 C_k L}$$

is sufficient accuracy where $C_k = 1 + \sqrt{\max_k(\|D_k\|\|\tau_k\|)} + \sqrt{\|\max_k \mathbf{d}_{k-1}^1\|} + \max_k \|\mathbf{d}_k^3\|$. So, for δ_1 we have the following condition:

$$\delta_1 \leq \frac{\varepsilon^4}{6400 C_k L \|D_k\|_2}.$$

for all k or

$$\delta_1 \leq \frac{\varepsilon^4}{6400 A_k L}.$$

where $A_k = C_k \max_k \|D_k\|_2$. □

C.3 Ellispoids Method

Algorithm 4 Ellipsoids Method with δ-subgradient.

Require: Number of iterations $N \geqslant 1$, $\delta \geqslant 0$, ball $\mathcal{B}_{\mathcal{R}} \supseteq Q_x$, its center c and radius $\mathcal{R}$.

$\mathcal{E}_0 := \mathcal{B}_{\mathcal{R}}, \quad H_0 := \mathcal{R}^2 I_n, \quad c_0 := c.$

for $k = 0, \ldots, N-1$ **do**

  if $c_k \in Q_x$ **then**

    $w_k := w \in \partial_\delta g(c_k)$,

    if $w_k = 0$ **then return** c_k,

  else

    $w_k := w$, where $w \neq 0$ is such that $Q_x \subset \{x \in \mathcal{E}_k : \langle w, x - c_k \rangle \leqslant 0\}$.

  $c_{k+1} := c_k - \frac{1}{n+1} \frac{H_k w_k}{\sqrt{w_k^T H_k w_k}}$,

$H_{k+1} := \frac{n^2}{n^2-1} \left(H_k - \frac{2}{n+1} \frac{H_k w_k w_k^T H_k}{w_k^T H_k w_k} \right)$,

$\mathcal{E}_{k+1} := \{x : (x - c_{k+1})^T H_{k+1}^{-1} (x - c_{k+1}) \leqslant 1\}$,

Ensure: $x^N = \arg \min_{x \in \{c_0, \ldots, c_N\} \cap Q_x} g(x)$.