An adaptive Gradient-type Method for Composite Optimization Problems with Gradient Dominance Condition and Generalized Smoothness

Fedor Stonyakin
V. I. Vernadsky Crimean Federal University,
Russia, 295007, Republic of Crimea, Simferopol, Academician Vernadsky Avenue, 4
Moscow Institute of Physics and Technology (National Research University),
Russia, 141701, Moscow Region, Dolgoprudny, Institutskiy per., 9
E-mail: fedyor@mail.ru

Olesya Kuznetsova
Moscow Institute of Physics and Technology (National Research University),
Russia, 141701, Moscow Region, Dolgoprudny, Institutskiy per., 9
E-mail: kuznetsova.oa@phystech.edu

Sergei Artamonov
National Research University "Higher School of Economics",
20 Myasnitskaya, 101000 Moscow, Russia
E-mail: sartamonov@hse.ru

Inna Baran
V. I. Vernadsky Crimean Federal University,
Russia, 295007, Republic of Crimea, Simferopol, Academician Vernadsky Avenue, 4
E-mail: matemain@mail.ru

Ibraim Fitaev
V. I. Vernadsky Crimean Federal University,
Russia, 295007, Republic of Crimea, Simferopol, Academician Vernadsky Avenue, 4
E-mail: fitaev.i@yandex.ru

Abstract. We consider an interesting class of composite optimization problems with a gradient dominance condition and introduce corresponding analogue of the recently proposed concept of an inexact oracle. This concept is applied to some classes of smooth functional.

1 The research in Theorem 1 was partially supported by Russian Foundation for Basic Research grant (project 18-31-20005 mol-a-ved). The research of F. Stonyakin in Theorem 2 was partially supported by the grant of the President of Russian Federation for young candidates of sciences (project MK-15.2020.1). The research of F. Stonyakin and I. Fitaev in numerical experiments was supported by Russian Science Foundation (project 18-71-00048).
Theoretical estimate for the rate of convergence for proximal gradient method with constant step is obtained. Adaptive version of the proximal gradient method is proposed for the considered class of problems.

1. Introduction

It is well known that the condition of strong convexity of the smooth objective functional of a large-scale optimization problem allows us to guarantee a higher rate of convergence in comparison with the usual convex case. The condition of strong convexity of functionals allows to obtain estimates of the rate of convergence of gradient-type methods that are uniform with respect to dimension of the problem. This is important in the context of large-scale applications. However, many problems lead to non-convex cases. Therefore, there is a natural interest in distinguishing the most general class of functions for which efficient numerical methods that guarantee a convergence rate close to linear can be constructed. One approach of this type is to replace convexity with the well-known condition of gradient dominance (hereinafter, the \((PL)\)-condition) \cite{1, 2}.

Further, in applications there are often arise problems with inexact data and weakened assumptions on smoothness of the functionals. One of the approaches to obtain estimates for problems with generalized smoothness or inexact data is the concept of \((\delta, L, \mu)\)-oracle \cite{3} (see also \cite{4}). It is said that the function \(f\) admits \((\delta, L, \mu)\)-oracle \((f_\delta(y), g_\delta(y)) \in \mathbb{R} \times \mathbb{R}^d\) at the point \(y \in Q \subset \mathbb{R}^d\), if for all \(x, y \in Q\) it holds
\[
f_\delta(x) + \langle g_\delta(x), y-x \rangle + \frac{\mu}{2} \|y-x\|^2 \leq f(y) \leq f_\delta(x) + \langle g_\delta(x), y-x \rangle + \frac{L}{2} \|y-x\|^2 + \delta, \quad \forall x, y \in Q, \tag{1}\]
for some fixed \(\delta > 0\). Hereinafter, the norm \(\| \cdot \|\) means the usual Euclidean norm in a finite-dimensional space \(\mathbb{R}^d\). The right inequality in (1) can be considered as a relaxation of the Lipschitz condition for the gradient of the convex function \(f\) in the case of presence of an error. For example, it is clear that \(f_\delta(x) \in [f(x) - \delta, f(x)] \quad \forall x \in Q\), i.e. this approach allows us to describe the case when, instead of the exact value of the objective functional at the current point, its approximation \(f_\delta(x)\) is considered.

Further, in the article \cite{5} a modification of the concept of the \((\delta, L, \mu)\)-oracle was considered with replacement of the left inequality in (1) by the gradient dominance condition. At the same time, a gradient type method with adaptive tuning of the smoothness constant \(L\) is proposed in \cite{5}. However, the method in \cite{5} is essentially based on explicit choice of the step. This complicates the applicability of the developed approach to structural (composite) optimization problems. And such problems naturally arise in various applications. It is worth noting here, for example, the \(\ell_1\) regularized LASSO regression problem with a sparse matrix (see e.g. \cite{6, 7}). This problem is now being widely studied and new techniques are being developed (see e.g. \cite{8, 9}).

In this paper, we consider the general class of composite optimization problems \cite{6} of the form
\[
F(x) = f(x) + g(x) \rightarrow \min_{x \in \mathbb{R}^d}, \tag{2}\]
where \(f(x)\) is a differentiable function with \(L\)-Lipschitz-continuous gradient, and \(g(x)\) is a convex function of a simple structure (possibly, nonsmooth). In \cite{2} the linear convergence rate of the proximal gradient method with constant step was shown for a certain class of problems. The assumption was that for some \(\mu > 0\) a functional \(F(x)\) satisfies the inequality that generalizes the Polyak-Lojasiewicz condition to the case of composite problems:
\[
\frac{1}{2} D_g(x, L) \geq \mu (F(x) - F^*), \tag{3}\]
where
\[ D_g(x, \alpha) \equiv -2\alpha \min_{y \in \mathbb{R}^d} \left[ \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2 + g(y) - g(x) \right] \quad (4) \]

As an example of a function satisfying (3), we note \( F(x) = f(Ax) + g(x) \), where \( f(x) \) is strongly convex, \( g(x) \) is an indicator function of some polyhedron \( \chi \) and \( A \) is a matrix of a linear operator \[2\).

Analogously with (1) we consider the class of problems with the following generalization of smoothness
\[ f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2 + \delta, \quad (5) \]
for all \( x,y \in Q \) with fixed \( \delta > 0 \).

For \( \delta = 0 \) this inequality is true for any function \( f \) with an \( L \)-Lipschitz-continuous gradient. If \( \delta > 0 \) then (9) holds, for example, for functions \( f \) with Hölder-continuous gradient \( (\nu \in [0; 1]) \) (see, e.g. \[3\])
\[ \|\nabla f(x) - \nabla f(y)\| \leq L_\nu \|x - y\| \quad \forall x, y \in \mathbb{R}^d \]
and fixed \( L_\nu < +\infty \). In this situation we can put
\[ L = L_\nu \left[ \frac{1 - \nu}{1 + \nu} \cdot \frac{L_\nu}{2\delta} \right]^{\frac{1 + \nu}{2\nu}} \]

Let us consider another example: \( f(x) = \varphi(x) + g(x) \), where \( \varphi \) has a \( L_\varphi \)-Lipschitz-continuous gradient and \( g(x) \) is differentiable \( M \)-Lipschitz-continuous function for some \( M > 0 \). In this case, for each \( x,y \in \mathbb{R}^n \) we have
\[ g(y) \leq g(x) + \nabla g(x), y - x \rangle + 2M \|y - x\| \leq g(x) + \langle \nabla g(x), y - x \rangle + \frac{M^2}{\delta} \|y - x\|^2 + \delta \]
for fixed \( \delta > 0 \). So, (5) holds for \( L = L_\varphi + \frac{2M^2}{\delta} \).

In this paper estimates of the rate of convergence of the gradient method with a constant step are obtained for the problems (2) under the assumptions (3)–(5). However both the method itself and the estimates of its rate of convergence make significant use of the global constant \( L > 0 \). Therefore, by analogy with \[4\]–[6] we also propose a method with adaptive tuning of the \( L \) and \( \delta \) parameters. For such a setting it is important to calculate the values of the function at each iteration of the method. We consider the assumption that at each step not the exact value of \( f \) is available, but only some approximation \( f_\delta(x) \in [f(x) - \delta, f(x)] \) \( (\delta > 0) \). The estimates of the rate of convergence are obtained for this method. At the end of the work, we present a numerical comparison of the considered methods for the usual regression problem.

The contributions of this paper can be summarized as follows.

- An adaptive version of the proximal gradient method on a certain class of problems is presented. Moreover, we emphasize the possibility of using inexact values of the objective functional for the choice of a step. This result is new even in the case of smooth mappings.
- However, the class of problems with generalized smoothness is also considered. Estimates of the rate of convergence of both the adaptive method and its non-adaptive version are obtained on this class.

2. Gradient-type Methods for Composite Optimization Problems with Considered Inexact Oracle Concept Modification

We start with theoretical estimate for non-adaptive proximal gradient method for the considered class of problems.
Theorem 1 Let the differentiable functional \( f(x) \) in the problem (2) satisfies the condition of generalized smoothness (5). Assume also that \( g(x) \) is a convex functional of simple structure, and the objective functional \( F(x) \) satisfies the proximal Polyak-Lojasiewicz condition (3), and the set of solutions of the optimization problem (2) is not empty. Then, after \( k \) iterations of the proximal gradient method

\[
x_{k+1} = \arg \min_{y \in \mathbb{R}^d} \left[ \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2} \| x_k - y \|^2 + g(y) - g(x) \right]
\]

the following inequality is true:

\[
F(x_k) - F^* \leq \left( 1 - \frac{\mu}{L} \right)^k [F(x_0) - F^*] + \frac{L}{\mu} \left( 1 - \left( 1 - \frac{\mu}{L} \right)^k \right) \delta.
\]

In fact, the proved statement guarantees a linear rate of convergence of the method up to a value determined by the parameter \( \delta \).

Proof. Similarly to ([2], Theorem 5) we have:

\[
F(x_{k+1}) = f(x_{k+1}) + g(x_k) - g(x_{k+1})
\]

\[
\leq F(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \| x_{k+1} - x_k \|^2 + \delta + g(x_{k+1}) - g(x_k)
\]

\[
\leq F(x_k) - \frac{1}{2L} \mathcal{D}_g(x_k, L) + \delta \leq F(x_k) - \frac{\mu}{L} (F(x_k) - F^*) + \delta.
\]

So, we have

\[
F(x_{k+1}) - F^* \leq \left( 1 - \frac{\mu}{L} \right)^k [F(x_0) - F^*] + 2\delta.
\]

Finally, taking into account that

\[
\sum_{l=0}^{k} \left( 1 - \frac{\mu}{L} \right)^l = \frac{L}{\mu} \left( 1 - \left( 1 - \frac{\mu}{L} \right)^{k+1} \right),
\]

we obtain the statement of the theorem.

In the previous result, we considered a method with constant values of the smoothness constants \( L \) and corresponding value of \( \delta \). Based on the work [5], we propose a modified proximal gradient method for the problem (2) with adaptive tuning of values \( L \) and \( \delta \) (see Algorithm 1). Assume that \( F_\delta(x) \in [F(x) - \delta, F(x)] \) for each \( x \in \mathbb{R}^d \).

To prove the result of its rate of convergence, instead of (3) we will consider its amplification: for each \( x \in Q \)

\[
\frac{1}{2} \mathcal{D}_g(x, \bar{L}) \geq \mu (F(x) - F^*), \quad \forall \bar{L} : \mu \leq \bar{L}.
\]

We apply the result of [2] (Lemma 1): for each differentiable function \( f \) and any convex function \( g \), for \( \mu_2 > \mu_1 > 0 \) we have

\[
\mathcal{D}_g(x, \mu_2) \geq \mathcal{D}_g(x, \mu_1), \quad \forall x \in \mathbb{R}^d.
\]

Therefore, (9) immediately follows from the following inequality

\[
\frac{1}{2} \mathcal{D}_g(x, \mu) \geq \mu (F(x) - F^*), \quad \forall x \in \mathbb{R}^d.
\]
Algorithm 1. Adaptive proximal gradient method for functions satisfying the condition of gradient dominance.

\( x_0 \) - starting point, \( L_0 : \mu \leq L_0 \leq L, \delta_0 > 0 \)

For \( k \leftarrow 0 \) To \( N - 1 \) do

\[
\delta_{k+1} = \frac{\delta_k}{2};
\]

\[
L_{k+1} = \begin{cases} \frac{L_k}{2} & \text{if } L_k > 2\mu \\ L_k & \text{if } L_k \leq 2\mu \end{cases} \quad \text{and} \quad L_{k+1} = L_k \text{ if } L_k \leq 2\mu;
\]

\[
x_{k+1} = \arg \min_{y \in \mathbb{R}^d} \left[ \langle \nabla f(x_k), y - x_k \rangle + \frac{L_{k+1}}{2} \| y - x_k \|^2 + g(y) - g(x_k) \right];
\]

While \( F_\delta(x_{k+1}) \leq F_\delta(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_{k+1}}{2} \| x_{k+1} - x_k \|^2 + \delta_{k+1} + g(x_{k+1}) - g(x_k) \) do

\[
L_{k+1} = 2L_{k+1}; \quad \delta_{k+1} = 2\delta_{k+1};
\]

\[
x_{k+1} = \arg \min_{y \in \mathbb{R}^d} \left[ \langle \nabla f(x_k), y - x_k \rangle + \frac{L_{k+1}}{2} \| y - x_k \|^2 + g(y) - g(x_k) \right];
\]

endWhile

endFor

Theorem 2 Let the differentiable function \( f(x) \) in the problem (2) satisfies the condition

\[
f_\delta(y) \leq f(y) \leq f_\delta(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| y - x \|^2 + \delta, \tag{10}\]

where \( f_\delta(x) \in [f(x) - \delta, f(x)] \) for all \( x \in \mathbb{R}^d \) with fixed \( \delta > 0 \). Assume that the additional replacement of the gradient domination condition by a stronger version (9) is done. Then:

(i) For each \( k = 1, 2, \ldots, N \) after \( k \) iterations of the adaptive Algorithm 1, the function \( F \) satisfies the following inequality:

\[
F(x_k) - F^* \leq \sum_{s=1}^{k} \left( 1 - \frac{\mu}{L_s} \right) [F(x_0) - F^*] + \delta_k + \delta \tag{11}
\]

\[
+ \sum_{l=0}^{k-1} \prod_{m=l+1}^{k} \left( 1 - \frac{\mu}{L_m} \right) (\delta_l + \tilde{\delta})
\]

(ii) After \( N \) iterations the auxiliary problem of minimizing a convex function is solved no more than

\[
2N + \max \left\{ \log_2 \frac{2L}{L_0}, \log_2 \frac{2\delta}{\delta_0} \right\} \tag{12}
\]

times, thus it has the same computational complexity as a non-adaptive analogue.

Proof.

(i) Following the scheme of the proof of Theorem 1 and taking into account the condition for passing to the next iteration we have:

\[
F_\delta(x_{k+1}) = f_\delta(x_{k+1}) + g(x_k) - g(x_k) + g(x_{k+1})
\leq F_\delta(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \| x_{k+1} - x_k \|^2 + \delta + g(x_{k+1}) - g(x_k) \tag{13}
\leq F_\delta(x_k) - \frac{1}{2L} Dg(x_k, L) + \delta \leq F_\delta(x_k) - \frac{\mu}{L} (F(x_k) - F^*) + \delta.
\]
So, from $F_\delta(x_{k+1}) - F_\delta(x_k) \geq F(x_{k+1}) - F(x_k) - \delta$ we have

$$F(x_{k+1}) \leq F(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_{k+1}}{2}\|x_{k+1} - x_k\|^2 + \delta_{k+1} + \tilde{\delta} + g(x_{k+1}) - g(x_k)$$

Together with (9) it gives an opportunity to continue the estimate

$$F(x_{k+1}) \leq F(x_k) - \frac{\mu}{L_{k+1}} (F(x_k) - F^*) + \delta_{k+1} + \delta,$$

from which, revealing the recurrence relation, we obtain the desired inequality (11).

(ii) Denote by $i_k$ the number of times when the auxiliary minimization problem is solved on the $k$-th iteration of the algorithm (9). Then, since at the beginning of each iterations we divide $\delta_k$ by two, but probably leave $L_k$ unchanged

$$i_k = 2 + \log \frac{\delta_{k+1}}{\delta_k} \leq 2 + \log \frac{L_{k+1}}{L_k}.$$

Summing over all iterations

$$\sum_{k=0}^{N-1} i_k = \sum_{k=0}^{N-1} \log \frac{\delta_{k+1}}{\delta_k} + 2N \leq \sum_{k=0}^{N-1} \log \frac{L_{k+1}}{L_k} + 2N.$$

It is clear that at least one of the inequalities holds $\delta_N \leq 2\delta$, $L_N \leq L$, so the resulting estimate is

$$\sum_{k=0}^{N-1} i_k \leq 2N + \max \left\{ \log \frac{2L}{L_0}, \log \frac{2\delta}{\delta_0} \right\}$$

It is worth noting that in such an adaptive scheme, the algorithm is tuned to the local value of the constants $L$ and $\delta$ at specific points $x \in \mathbb{R}^d$. It may turn out that for some values of $k$ the corresponding value $L_k$ will be significantly smaller than the Lipschitz constant of the gradient $f$. Moreover, this behavior can be observed experimentally, which leads to an increase in the rate of convergence in the transition to the adaptive version. This shows one more advantage of the adaptive approach compared to the non-adaptive one.

3. Numerical Experiments

Let a given matrix $A \in \mathbb{R}^{1000 \times 1000}$ and a vector $b \in \mathbb{R}^{1000}$. The main diagonal of the matrix $A$ is filled with random integers from the interval $[1, 20]$. Also 10 randomly selected non-diagonal elements of this matrix are integers from the interval $[1, 20]$. Consider the problem of solving a matrix equation. This problem, in the case of solvability, is equivalent to the problem of minimizing a convex functional $f(x) = 0.5\|Ax - b\|_2^2$. The indicated function is $\mu$-strongly convex and has a $L$-Lipschitz-continuous gradient, where $\mu$ is the smallest positive eigenvalue of the matrix $A^T A$, $L$ is the largest eigenvalue of $A^T A$ ($A^T$ is the matrix transposed to $A$). The starting point is $x^0 = (0.1, \ldots, 0.1)$. Assume that for adaptive method at each point we have only inexact value $f_\delta(x) = f(x) - \delta$, where $\delta \leq \delta = 0.01$. We run the adaptive and non-adaptive methods described above in section 1 for the indicated problem. In the comparison Table 1 (see also Figure 1), depending on the number of iterations, the estimate of the quality of the solution found is given in accordance with the obtained in Theorems 1 and 2 theoretical estimates $(F(x_0) - F^*)$ is used instead of $F(x_0) - F^*$. This is quite natural, because, due to the random selection of parameters, we do not know where the exact solution of the problem is located. As we can see,
Table 1: Results of Algorithm 1.

| K  | Adaptive     |               | Non-adaptive |               |
|----|--------------|---------------|--------------|---------------|
|    | Time, s      | Estimate      | Time, s      | Estimate      |
| 1000| 105.0        | 266.64791     | 31.0         | 4483.59369    |
| 2000| 199.8        | 5.23964       | 62.2         | 1892.32799    |
| 3000| 294.0        | 0.5986        | 93.4         | 914.77683     |
| 4000| 388.2        | 0.07701       | 124.4        | 477.62367     |
| 5000| 482.4        | 0.01821       | 156.0        | 261.195       |
| 6000| 576.8        | 0.01166       | 187.2        | 146.68122     |
| 7000| 671.2        | 0.01091       | 218.6        | 83.58596      |
| 8000| 764.6        | 0.01083       | 250.2        | 48.00781      |
| 9000| 858.2        | 0.01075       | 281.6        | 27.68953      |
| 10000| 951.6       | 0.01082       | 313.4        | 16.00631      |

Table 2: Results of Algorithm 1.

| K   | Adaptive     |               | Non-adaptive |               |
|-----|--------------|---------------|--------------|---------------|
|     | F(x)         | Time, s       | F(x)         | Time, s       |
| 1000| 105.0        | 0.0791838216  | 11.7527626238| 31.0         |
| 2000| 199.8        | 0.0019284084  | 1.7007326643 | 62.2         |
| 3000| 294.0        | 0.0003489     | 0.5670845121 | 93.4         |
| 4000| 388.2        | 0.000153289   | 0.2204354268 | 124.4        |
| 5000| 482.4        | 0.0001510414  | 0.0973512542 | 156.0        |
| 6000| 576.8        | 0.0001507847  | 0.0476449185 | 187.2        |
| 7000| 671.2        | 0.0001507515  | 0.0250337712 | 218.6        |
| 8000| 764.6        | 0.0001507519  | 0.0137397343 | 250.2        |
| 9000| 858.2        | 0.0001507515  | 0.0044101557 | 281.6        |
| 10000| 951.6       | 0.0001507515  | 313.4        |               |

for approximately the same time and number of iterations, an estimate for the adaptive method (Theorem 2) allows us to guarantee a significantly better quality of the solution. In the Table 2 (see also Figure 2) the values of the objective function are compared. The comparison tables show the average results of five experiments with a random choice of the matrix and the vector. All calculations were performed using CPython 3.7.7 software on a computer with a 3-core AMD Athlon II X3 450 processor with a clock frequency of 3.2 GHz for each core. The computer’s RAM was 8 GB.
Conclusion

Some analogue of the concept of \((\delta, L, \mu)\)-oracle of Devolder-Gliner-Nesterov for composite optimization problems in which strong convexity is replaced by the proximal analogue of the Polyak-Lojasiewicz gradient domination condition is proposed in the paper. Estimates of the convergence rate of gradient type methods are obtained.

Both non-adaptive and adaptive versions of the proximal gradient method are considered. For the adaptive method, some refinement of the proximal gradient dominance condition has been introduced. At the same time, adaptive Algorithm 1 completely implies adaptive step tuning to the value of the local smoothness constants \(L_k\) as well as partial tuning to another parameter \(\delta\). It is proved that adaptive tuning does not increase the average cost of the iteration of the method, but experimentally it can lead to a significant increase in the quality of the solution, as demonstrated by experiments.
An important difference from gradient methods for the standard version of the \((\delta, L, \mu)\)-oracle is the ability to evaluate the quality of the solution found only by the value of the objective functional at the initial point \(F(x^0)\), without using an estimate of the distance from \(x^0\) to the exact solution \(F(x^0) - F^*\). Some numerical experiments were performed for the problem of solving the matrix equation, which demonstrated the best efficiency of the adaptive algorithm compared to non-adaptive one.

References

[1] B. T. Polyak: Gradient methods for minimizing functionals. In: Journal of Computational Mathematics and Mathematical Physics. 3(4), 643–653 (1963).

[2] Karimi H., Nutini J., Schmidt M.: Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 795–811. Springer (2016).

[3] Devolder O.: Exactness, inexactness and stochasticity in first-order methods for large-scale convex optimization: PhD thesis. 2013. 320 p.

[4] Stonyakin F. S., Dvinskikh D., Dvurechensky P., Kroshnin A., Kuznetsova O., Agafonov A., Gasnikov A., Tyurin A., Uribe C. A., Pasechnyuk D., Artamonov S.: Gradient methods for problems with inexact model of the objective. In: International Conference on Mathematical Optimization Theory and Operations Research. Lecture Notes in Computer Science. 11548, 97–114 (2019).

[5] Stonyakin F.S.: Adaptation to inexactness for some gradient-type optimization methods. Trudy Instituta Matematiki i Mekhaniki UrO RAN 25(4), 210–225 (2019).

[6] Nesterov Yu.: Gradient methods for minimizing composite functions. Math. Program. 140(1), 125–161 (2013).

[7] A. V. Gasnikov. Modern Numerical Optimization Methods. Universal Gradient Descent. MIPT, Moscow, 2018 (in Russian).

[8] Zhang H., Wang J., Sun Z., Zurada J.M., Pal N.R.: Feature selection for neural networks using group lasso regularization. IEEE Transactions on Knowledge and Data Engineering (2019).

[9] Yang S., Wen J., Zhan X., Kifer D.: ET-Lasso: A New Efficient Tuning of Lasso-type Regularization for High-Dimensional Data. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 607–616 (2019).