

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

SIGN-SGD VIA PARAMETER-FREE OPTIMIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Large language models have achieved major advances across domains, yet training them remains extremely resource-intensive. We revisit SIGN-SGD, which serves both as a memory-efficient optimizer for single-node training and as a gradient compression mechanism for distributed learning. This paper addresses a central limitation: the effective stepsize cannot be determined *a priori* because it relies on unknown, problem-specific quantities. We present a parameter-free SIGN-SGD that removes manual stepsize selection. We analyze the deterministic single-node case, and extend the method to stochastic single-node training and multi-node settings. We also incorporate the momentum technique into our algorithms and propose a memory-efficient variant that stores only gradient signs instead of full gradients. We evaluate our methods on pre-training LLaMA models (130M and 350M) and fine-tuning a Swin Transformer (28M). Across considered tasks, the proposed methods match the performance of tuned SIGN-SGD and ADAMW (grid-searched stepsizes with a cosine schedule), while avoiding tuning overhead. Employing parameter-free training yields approximately $1.5 \times$ end-to-end speedup compared to runs with grid-searched stepsizes.

1 INTRODUCTION

Models and datasets continue to scale rapidly (Vaswani, 2017; Hoffmann et al., 2022; Alzubaidi et al., 2021). This growth drives steep increases in compute requirements, memory footprint, and wall-clock training time, consequently raising hardware costs. These pressures motivate the development of methods that accelerate training and reduce resource usage without sacrificing accuracy. A significant breakthrough arose not from designing advanced learning algorithms, but primarily from the manner in which these algorithms can be applied: distributed learning (Konečný et al., 2016; McMahan et al., 2017; Verbraeken et al., 2020). However, distributing training across M nodes does not yield an M -fold speedup in practice, as inter-device communication remains a significant bottleneck.

The reduction of the number of transmitted packages through compression is one of the key techniques to address this issue (Seide et al., 2014; Alistarh et al., 2018). Among others, the SIGN-SGD method stands out (Bernstein et al., 2018). Solving the classic optimization problem $\min_{x \in \mathbb{R}^d} f(x)$, it utilizes

Algorithm 1 SIGN-SGD

```

1: Input: Start point  $x^0 \in \mathbb{R}^d$ , number of iterations  $T$ 
2: Parameter: Stepsize  $\gamma > 0$ 
3: for  $t = 0, \dots, T - 1$  do
4:    $x^{t+1} = x^t - \gamma \text{sign}(\nabla f(x^t))$ 
5: end for

```

an intuitive heuristic that takes the sign of each gradient coordinate (Algorithm 1). In the distributed setup, aggregation is performed by a majority vote on the transmitted signs of the gradients.

Additionally, SIGN-SGD is rapidly gaining popularity, even for single-node training. In contrast to methods such as ADAM (Kingma, 2014) and ADAMW (Loshchilov, 2017), which require substantial memory for storing statistics, SIGN-SGD is free from this constraint. Moreover, sign-based approaches offer both theoretical and practical advantages over traditional SGD (Robbins & Monro, 1951), demonstrating superior convergence (Balles & Hennig, 2018; Balles et al., 2020) and empirical performance (Kunstner et al., 2023; Zhao et al., 2024; Zmushko et al., 2024) in training large models.

Although SIGN-SGD is effectively used both for compression in distributed learning and as a memory-efficient method in a single-node regime, achieving its full potential requires selecting an appropriate stepsize. The optimal choice depends on problem-specific quantities that are unknown

054 in practice, necessitating costly manual tuning. To address this issue, we introduce parameter-free
 055 SIGN-SGD algorithm that employ automatic stepsize selection schemes.
 056

057 2 BRIEF LITERATURE REVIEW AND CONTRIBUTIONS

059 To situate the problem and motivate our algorithms, this section reviews the literature and distills the
 060 open challenges that guide our contributions.
 061

- 062 • We begin by revisiting SIGN-SGD and identifying the theoretically desirable stepsize that would
 063 enable effective training without manual tuning.
- 064 • Next, we survey parameter-free optimization methods, highlighting their advantages and limitations.
- 065 • We conclude by stating the contributions of this work and explaining how they address the gaps.

066 2.1 RELATED WORK

069 • **Sign-SGD.** In the original paper on SIGN-SGD (Bernstein et al., 2018), the authors explored
 070 convergence in the paradigm of finding a near-stationary point, i.e., such $x \in \mathbb{R}^d$, that $\|\nabla f(x)\| \leq \varepsilon$,
 071 where ε represents the accuracy of the solution. Moreover, to achieve convergence with respect to
 072 the variance term, the authors utilized mini-batches. Both this convergence criterion and the use
 073 of mini-batches are essential components of the analysis. As shown in (Karimireddy et al., 2019),
 074 SIGN-SGD may fail to converge when considered the regret minimization. Moreover, to achieve
 075 convergence with respect to the variance term, the authors of (Karimireddy et al., 2019) utilized
 076 mini-batches. Meanwhile, Safaryan & Richtárik (2021) proposed a relaxation of the SIGN-SGD
 077 method and showed that at least half of the coordinates in the sign of the stochastic gradient align
 078 with those of the exact gradient, thereby enabling convergence with respect to the variance term. A
 079 number of works have also emerged around SIGN-SGD, extending it with momentum (Sun et al.,
 080 2023), providing high-probability convergence bounds (Kornilov et al., 2025), and studying it in
 081 the context of differential privacy (Jin & Dai, 2025). *Nevertheless, the possibility of selecting a
 082 stepsize independent of problem properties while achieving optimal convergence rate has been largely
 083 overlooked.*

084 Let us provide the basic estimate of SIGN-SGD convergence with the exact gradient oracles. This
 085 can be simply derived from Theorem 1 in (Bernstein et al., 2018):

$$086 \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{\Delta^*}{\gamma T} + \frac{\gamma L_\infty}{2},$$

087 where L_∞ is the smoothness constant of the objective f with respect to l_∞ -norm, and $\Delta^* = f(x^0) - f(x^*)$ represents the initial distance to the solution. Putting

$$088 \gamma = \frac{\sqrt{\Delta^*}}{\sqrt{L_\infty T}}, \text{ we obtain optimal } \mathcal{O}\left(\frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}}\right) \text{ convergence rate.} \quad (1)$$

089 This stepsize poses challenges, as it depends on the problem's hyperparameters. To address this issue,
 090 we turn to various techniques that facilitate the provision of an adaptive stepsize.

091 • **Parameter-free approaches.** In the non-smooth setting, considering regret minimization, classic
 092 gradient methods (Robbins & Monro, 1951; Moulines & Bach, 2011; Stich, 2019; Lan, 2020) require

$$093 \gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{T}} \text{ to have } \mathcal{O}\left(\frac{\|x^0 - x^*\|_2 M}{\sqrt{T}}\right) \text{ convergence rate.} \quad (2)$$

094 This estimate is (worst-case) optimal in its complexity class (Nemirovskij & Yudin, 1983). We let M
 095 denote the Lipschitz constant ($|f(x) - f(y)| \leq M \|x - y\|_2$ for all $x, y \in \mathbb{R}^d$). The parameter-free
 096 setting aims to adapt the stepsize automatically, without prior knowledge of the initial distance
 097 $\|x^0 - x^*\|_2$ or the Lipschitz constant M .

098 For the first time, the idea of an automatic stepsize setting was proposed to achieve adaptation to
 099 constant M . It was embodied in methods such as ADAGRAD (Duchi et al., 2011), ADAM (Kingma,
 100 2014), RMSPROP (Tieleman & Hinton, 2012), ADADELTA (Zeiler, 2012), and ADAPTIVE SGD
 101 (Gupta et al., 2017; Attia & Koren, 2023). In these methods, computed gradients were utilized

108 to adjust the stepsize based on the properties of M . *However, these methods required additional*
 109 *memory and computations, and they lacked adaptivity to the initial distance.* Attempts to modify γ in
 110 equation 2 led to approaches within the general online stochastic learning setting (Orabona, 2019),
 111 such as coin betting and reward-doubling techniques (Streeter & McMahan, 2012; Orabona, 2013;
 112 McMahan & Orabona, 2014; Orabona & Pál, 2016; Cutkosky & Orabona, 2018; Cutkosky, 2019),
 113 which can also be classified as parameter-free algorithms. *Nevertheless, these approaches assumed*
 114 *that the stochastic oracles have some (loose) bound.*

115 Further studies suggested more intricate solutions in parameter-free convex stochastic optimization.
 116 These methods achieved asymptotic convergence rates comparable to classic approaches
 117 while adapting to essential hyperparameters. The starting point was the work (Carmon & Hinder,
 118 2022) which provided adaptivity to the initial distance $\|x^0 - x^*\|_2$ through estimators of the form
 119 $\max_{t \leq T} \|x^0 - x^t\|_2$. To find such estimators, the authors employed an additional grid search
 120 procedure *which increased the required number of steps only in double-logarithmic time*. The primary
 121 objective of this work was to derive high-probability convergence estimates in the stochastic convex
 122 non-smooth setup. Several studies that did not utilize the additional search procedure were built upon,
 123 including (Khaled et al., 2023), (Ivgi et al., 2023) and (Kreisler et al., 2024).

124 The work (Defazio & Mishchenko, 2023) provided another approach for sensitivity to the initial
 125 distance. The authors iteratively constructed a sequence upper bounded by $\|x^0 - x^*\|_2$ and approx-
 126 imated it accordingly. *However, they considered only exact gradient oracles, which represents a*
 127 *significant limitation.* Later, in (Mishchenko & Defazio, 2023), the authors introduced a damping
 128 factor in the denominator to improve convergence in the logarithmic factor’s square root. *Never-*
 129 *theless, theoretical analysis depended on the knowledge of the Lipschitz constant, which is not a*
 130 *parameter-free approach.* We note that the use of the classic ADAGRAD-NORM stepsize (Duchi
 131 et al., 2011; Streeter & McMahan, 2010; Ward et al., 2020), possibly with additional factors in the
 132 denominators, remains standard for adaptation to M .

133 The orthogonal approach was presented in the work (Mishkin et al., 2024). The authors considered a
 134 smooth setup and proposed the use of local approximations of the smoothness constant L to achieve
 135 adaptivity. However, the authors employed the stepsize $\gamma^t = \frac{\|x^{t+1}(\gamma^t) - x^t\|_2}{\|\nabla f(x^{t+1}(\gamma^t)) - \nabla f(x^t)\|_2}$ at the t -th
 136 iteration, where γ^t was determined by exponential search in the manner (Carmon & Hinder, 2022) or
 137 by Newton’s method. *Both variants are inefficient.*

138 In light of the literature, we present the main directions of this study. Our goal is to provide the
 139 parameter-free SIGN-SGD method that achieves a convergence rate comparable to the optimal
 140 stepsize tuning 1.

143 2.2 CONTRIBUTIONS

144 We propose a novel mechanism for estimators compared to existing approaches. Instead of the classic
 145 $\|x^0 - x^*\|$ and M hyperparameters in equation 2, we aim to gain the tolerance to $f(x^0) - f(x^*)$ and
 146 L_∞ from equation 1. We now outline our contributions.

147 **• Parameter-free SIGN-SGD.** We introduce a parameter-free SIGN-SGD method. The core
 148 idea involves per-iteration step-size adaptation. Every iteration, we choose estimators of L_∞ and
 149 $f(x^0) - f(x^*)$ using the current gradient information. This design is practical, as it requires no
 150 additional hyperparameter search or restarts. As a starting point, we analyze the exact gradients setup.

151 **• Stochastic and distributed settings.** We study our algorithm in the distributed setting and the
 152 case of stochastic gradient oracles. A lack of stochastic analysis presents a significant drawback in
 153 parameter-free optimization. Our work addresses this limitation.

154 **• Practical extensions.** We extend our approach in two important directions.

- 155 • We incorporate momentum to improve practical performance.
- 156 • We provide a memory-efficient parameter-free version. It stores only the sign of the gradient
 157 from the previous step while remaining an adaptivity to the problem properties.

158 **• Theoretical analysis.** We provide a comprehensive theoretical analysis of the proposed methods
 159 and establish convergence guarantees. In our setup, we consider a convex and smooth objective.

160 **• Experimental validation.** We demonstrate that our methods are competitive in practical tasks,
 161 including LLM and ViT training. An Adam-style momentum variant further improves performance

162 across both language and vision benchmarks. Empirically, parameter-free training matches or is
 163 slightly below tuned SIGN-SGD and AdamW with cosine schedules, while achieving appreciably
 164 better overall training time.
 165

166 3 ALGORITHMS AND CONVERGENCE ANALYSIS

167 • **Notation.** We begin with the following notation: $\mathbb{E}[\cdot]$ denotes the expected value of a random
 168 variable, $\|x\|_2 = \sqrt{\langle x, x \rangle}$ represents the Euclidean norm of the vector $x \in \mathbb{R}^d$, $\|x\|_1 = \sum_{i=1}^d |x_i|$
 169 refers to the ℓ_1 -norm of the vector x , and $\|x\|_\infty = \max_{i \in [d]} |x_i|$ defines the ℓ_∞ -norm of the vector x .
 170

171 • **Assumptions.** We present the assumptions regarding the objective function f

172 **Assumption 3.1.** The function f is L_∞ -smooth, i.e., it satisfies $\|\nabla f(x) - \nabla f(y)\|_1 \leq L_\infty \|x - y\|_\infty$
 173 for any $x, y \in \mathbb{R}^d$.
 174

175 **Assumption 3.2.** The function f is convex, i.e., it satisfies $f(x) \leq f(y) + \langle \nabla f(x), x - y \rangle$ for any $x, y \in \mathbb{R}^d$.
 176

177 Although neural networks are inherently non-convex, theoretical analysis under convexity assumptions remains relevant. Recent studies suggest that deep neural networks often exhibit properties similar to convexity in certain regions, making insights from convex analysis applicable (Kleinberg et al., 2018; Zhou et al., 2019; Liu et al., 2022). Moreover, convex optimization serves as a theoretical foundation for the design of optimization algorithms. For example, momentum (Nesterov et al., 2018) and AdaGrad (Duchi et al., 2011) were initially developed and analyzed for convex problems.

178 **Assumption 3.3.** The function f has a (maybe not unique) finite minimum, i.e., $f(x^*) = \inf_{x \in \mathbb{R}^d} f(x) > -\infty$.
 179

180 Now we move to the base point of our analysis: the algorithms with exact gradient oracles.
 181

182 3.1 EXACT GRADIENTS SETTING

183 We begin with an additional assumption regarding the gradient oracles.
 184

185 **Assumption 3.4.** At any point $x \in \mathbb{R}^d$, we have access to the exact gradient, i.e., we can compute
 186 the full gradient value $\nabla f(x)$.
 187

188 We now present the main algorithm of this paper named ALIAS (Automatic Local per-Iteration
 189 Approximation of the StepSize, Algorithm 2). At each iteration, it utilizes the stepsize selection in a
 190 specific manner to gain adaptivity to the global parameters of the problem. Below, we provide an
 191 explanation of the algorithm and offer some intuition why the presented stepsize facilitates adaptivity.
 192

193 Algorithm 2 ALIAS

1: **Input:** Starting point $x^0 \in \mathbb{R}^d$, initial L_∞ -approximation $\eta^{-1} = 0$, initial Δ^* -approximation
 2: **for** $t = 0, \dots, T-1$ **do**
 3: Compute gradient $\nabla f(x^t)$
 4: $\eta^t = \eta^{t-1} + \frac{\|\nabla f(x^t) - \nabla f(x^{t-1})\|_1}{\|x^t - x^{t-1}\|_\infty}$; $\lambda^t = \frac{1}{\sqrt{\eta^t}}$
 5: **if** $t \neq 0$ **then**
 6: $\tilde{d}^t = \sum_{i=0}^{t-1} \gamma^i \langle \nabla f(x^{i+1}), \text{sign}(\nabla f(x^i)) \rangle$
 7: $d^t = \max(d^{t-1}, \tilde{d}^t)$
 8: **end if**
 9: Option I: $\gamma^t = \lambda^t \sqrt{d^t}$
 10: Option II: $\gamma^t = \lambda^t \sqrt{f(x^0) - \tilde{f}}$
 11: $x^{t+1} = x^t - \gamma^t \text{sign}(\nabla f(x^t))$
 12: **end for**

critical. As shown in (Boyd et al., 2003), the condition $f(x^*) = 0$ arises in problems such as finding a point in the intersection of convex sets, completing positive semi-definite matrices, or solving systems of convex inequalities. Moreover, a lower bound \tilde{f} on $f(x^*)$ is often known or readily available. For instance, $\tilde{f} = 0$ serves as a valid estimate in the empirical risk minimization setting. Taking this into account, we present the second option for our method, where we use $f(x^0) - \tilde{f}$ with $\tilde{f} \leq f(x^*)$ (Line 10) instead of the sequence $\{d^t\}_{t=0}^{T-1}$.

As for the denominator, at the t -th iteration, we approximate the local Lipschitz constant L_∞ between x^t and x^{t-1} . We accumulate it in the manner of ADAGRAD-NORM by adding it to the sum of previous approximations:

$$\eta^t = \eta^{t-1} + \frac{\|\nabla f(x^t) - \nabla f(x^{t-1})\|_1}{\|x^t - x^{t-1}\|_\infty}.$$

In the stepsize, the corresponding to the denominator coefficient appears as:

$$\lambda^t = \frac{1}{\sqrt{\eta^t}} = \frac{1}{\sqrt{\sum_{i=0}^{t-1} \frac{\|\nabla f(x^{i+1}) - \nabla f(x^i)\|_1}{\|x^{i+1} - x^i\|_\infty}}}.$$

This stepsize facilitates iterative adaptation to the objective landscape. We are now prepared to present the main theoretical results of this section.

Theorem 3.5. Suppose Assumptions 3.1, 3.2, 3.3, 3.4 hold. We denote $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, $L_\infty^0 = \frac{\|\nabla f(x^1) - \nabla f(x^0)\|_1}{\|x^1 - x^0\|_\infty}$. Then Algorithm 2 with $d^0 < \Delta^*$ to reach ε -accuracy needs

$$\tilde{\mathcal{O}}\left(\frac{(\Delta^*)^2 (L_\infty)^3}{d^0 (L_\infty^0)^2 \varepsilon^2}\right) \text{ and } \tilde{\mathcal{O}}\left(\frac{\Delta^* (L_\infty)^3}{(L_\infty^0)^2 \varepsilon^2}\right) \text{ iterations with Options I and II, respectively.}$$

Remark 3.6. Under conditions of Theorem 3.5, Algorithm 2 with $\lambda^t = \frac{1}{\sqrt{L_\infty + \sum_{i=0}^{t-1} \frac{\|\nabla f(x^{i+1}) - \nabla f(x^i)\|_1}{\|x^{i+1} - x^i\|_\infty}}}$ to reach ε -accuracy, where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, needs

$$\tilde{\mathcal{O}}\left(\frac{(\Delta^*)^2 L_\infty}{d^0 \varepsilon^2}\right) \text{ and } \tilde{\mathcal{O}}\left(\frac{\Delta^* L_\infty}{\varepsilon^2}\right) \text{ iterations with Options I and II, respectively.}$$

Discussion of the results. Since we provide convergence guarantees for finding near-stationary points for a convex objective, we first examine the relationship between convergence rates in convex and non-convex settings. For instance, we compare gradient descent rates using the gradient norm as the convergence criterion. While the behavior of gradient norm minimization is well understood in the non-convex setting (Arjevani et al., 2023), it is specific in the context of convex optimization. Notably, Allen-Zhu (2018) showed that vanilla gradient descent – without acceleration or additional techniques – achieves the same $\mathcal{O}(1/\varepsilon^2)$ rate for finding near-stationary points in both convex and non-convex settings. However, as previously noted, SIGN-SGD does not admit convergence guarantees beyond any criterion except the gradient norm, even in the convex case. Consequently, convergence analysis for sign-based methods must be framed in terms of finding the near-stationary point. Thus, our convex rate is not superior to that of the non-convex case. Moreover, the bound in Theorem 3.5 includes an additional factor of $(L_\infty/L_\infty^0)^2$ compared to Remark 3.6. However, the algorithm analyzed in Remark 3.6 is not parameter-free: it requires prior knowledge of L_∞ . In Appendix A, we present empirical results for varying values of L_∞ , which demonstrate that this additive factor has negligible impact on the practical convergence of Algorithm 2.

So far, we propose an algorithm and provide the theoretical analysis behind it. However, the analysis assumes access to exact gradient oracles – an unrealistic assumption in practice. We now extend the analysis to more realistic scenarios involving stochastic oracles.

3.2 STOCHASTIC GRADIENTS SETTING

We begin with the assumption regarding the gradient oracles.

270 **Assumption 3.7.** At any point $x \in \mathbb{R}^d$ we have access to the stochastic gradient, i.e., we can compute
 271 $g_\xi(x) = \nabla f(x, \xi)$ – the stochastic gradient value with respect to the randomness in the choice of
 272 samples ξ . Additionally, the variance of these stochastic estimators is coordinate-wise bounded, i.e.,
 273 $\mathbb{E} ([g_\xi(x)]_i - [\nabla f(x)]_i)^2 \leq \sigma_i^2$. Furthermore, this implies that $\mathbb{E} \|g_\xi(x) - \nabla f(x)\|_1 \leq \|\sigma\|_1$.
 274

275 It is a classic assumption in stochastic optimization (Bernstein et al., 2018). Furthermore, the
 276 batch gradient g_ξ typically exhibits smoothness (Liu et al., 2023). Thus, we introduce an additional
 277 assumption.

278 **Assumption 3.8.** The stochastic function f_ξ is L_∞ -smooth according to the realization ξ , i.e., it
 279 satisfies $\|g_\xi(x) - g_\xi(y)\|_1 \leq L_\infty \|x - y\|_\infty$ for any $x, y \in \mathbb{R}^d, \xi$.
 280

281 The stochastic formulation of the problem (Assumption 3.7) necessitates modifications of Algorithm
 282 2. This algorithm assumes access to the exact gradients, and the estimation of the local smoothness
 283 constant relies on computing full gradients. Thus, our goal is to modify Line 4 in Algorithm 2.
 284 Utilizing Assumption 3.8, we can construct a local approximation of L_∞ on the t -th iteration via
 285 stochastic gradients with respect to the stochastic realization ξ^t . Namely,

$$286 \lambda^t = \frac{1}{\sqrt{\sum_{i=0}^{t-1} \frac{\|g_{\xi^{i+1}}^{i+1} - g_{\xi^{i+1}}^i\|_1}{\|x^{i+1} - x^i\|_\infty}}},$$

290 where $g_{\xi^t}^t$ is the stochastic gradient computed at the t -th iteration based on the stochastic realization
 291 ξ^t . We query the oracle twice per iteration, utilizing the current and subsequent stochastic realizations.
 292 Another change in Algorithm 2 involves performing a step in Line 11 regarding $\text{sign}(g_{\xi^t}^t)$. In the
 293 subsequent theoretical analysis, we focus solely on Option II in Algorithm 2. **We provide a formal**
 294 **description of the stochastic method, Algorithm 7, in Appendix F.2.** There, we present both the
 295 practical and theoretical versions.

296 We now present the convergence results.

297 **Theorem 3.9.** Suppose Assumptions 3.8, 3.2, 3.3, 3.7 hold. Then Algorithm 2 with Option II to reach

298 ε -accuracy, where $\varepsilon \geq \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{i=0}^{t-1} \gamma^i} \|\nabla f(x^t)\|_1 \right]$ and $L_\infty^{t, \xi^{t+1}} = \frac{\|g_{\xi^{t+1}}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty}$, needs

302 $\tilde{\mathcal{O}} \left(\frac{\Delta^* (L_\infty)^3}{\varepsilon^2} \left(\mathbb{E} \left(\frac{1}{L_\infty^{0, \xi^1}} \right)^2 \right) + \|\sigma\|_1^2 L_\infty \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t, \xi^{t+1}}} \right) \right)$ iterations.

306 **Remark 3.10.** Under the conditions of Theorem 3.9, Algorithm 2 with $\lambda^t = \frac{1}{\sqrt{L_\infty + \sum_{i=0}^{t-1} \frac{\|g_{\xi^{i+1}}^{i+1} - g_{\xi^i}^i\|_1}{\|x^{i+1} - x^i\|_\infty}}}$,

309 Option II and mini-batch of the size $t+1$ at t -th iteration, to reach ε -accuracy needs

310 $\tilde{\mathcal{O}} \left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \frac{\|\sigma\|_1^2 L_\infty}{\varepsilon^2} \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t, \xi^{t+1}}} \right) \right)$ iterations,

314 where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, $L_\infty^{t, \xi^{t+1}} = \frac{\|g_{\xi^{t+1}}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty}$.

317 **Discussion of the results.** With Assumption 3.8, a more stringent version of Assumption 3.1,
 318 we approximate the smoothness constant via stochastic gradients. The key point is to measure the
 319 gradient at the current point while considering the stochastic realization from the next iteration.
 320 Since x^t , ξ^t , and ξ^{t+1} are independent, we can provide a theoretical analysis. Thus, we surpass
 321 works such as (Defazio & Mishchenko, 2023; Mishchenko & Defazio, 2023; Mishkin et al., 2024),
 322 which employed a similar idea of the adaptation to the Lipschitz constant but lacked a stochastic
 323 analysis. Notably, the result of Theorem 3.9 achieves convergence only to a neighborhood, the size of
 which is determined by the variance. This rate fully aligns with the original SIGN-SGD convergence

(Bernstein et al., 2018). To address it in theory, we introduce increasing mini-batches analogously to (Bernstein et al., 2018) in Remark 3.10. We note that mini-batching enables convergence guarantees concerning the variance term; however, the method remains parameter-free even without it. In our experiments, we do not employ mini-batching.

We develop an analysis not only for the stochastic setting, but also for the distributed one. A full description of Algorithm 2 in the distributed setup, along with theoretical statements and proofs, is presented in Appendix F.3.

Above, we present an algorithm that can be easily applied to practical tasks. It does not require multiple restarts or additional search procedures. However, Algorithm 2 lacks the main advantage of the original SIGN-SGD method. Indeed, performing a step on the t -th iteration requires storing the entire gradient $\nabla f(x^{t-1})$ instead of just its sign. To address this limitation, we propose a memory-efficient modification in the next section.

3.3 MEMORY-EFFICIENT ALIAS

In Algorithm 2, memory efficiency is sacrificed to achieve a parameter-free stepsize. Indeed,

$$\gamma^t = \lambda^t \sqrt{d^t} = \sqrt{\frac{d^t}{\eta^t}} = \sqrt{\frac{\sum_{i=0}^{t-1} \gamma^i \langle \nabla f(x^{i+1}), \text{sign}(\nabla f(x^i)) \rangle}{\sum_{i=0}^{t-1} \frac{\|\nabla f(x^{i+1}) - \nabla f(x^i)\|_1}{\|x^{i+1} - x^i\|_\infty}}}.$$

To compute d^t , it is sufficient to store only $\text{sign}(\nabla f(x^{t-1}))$, incurring no additional memory costs. Regarding λ^t , we calculate $\|\nabla f(x^t) - \nabla f(x^{t-1})\|_1$ and $\|x^{t+1} - x^t\|_\infty$ at each step. The last term does not present an issue since $\|x^t - x^{t-1}\|_\infty = \|\gamma^{t-1} \text{sign}(\nabla f(x^{t-1}))\|_\infty$. However, to find $\|\nabla f(x^t) - \nabla f(x^{t-1})\|_1$, it is necessary to store the entire gradient $\nabla f(x^{t-1})$.

We address this concern by modifying λ^t in Algorithm 2:

$$\eta^t = \eta^{t-1} + \frac{\|\nabla f(x^t) - \nabla f(x^{t-1})\|_\infty}{\|x^t - x^{t-1}\|_1} \text{ followed by } \lambda^t = \frac{1}{\sqrt{\sum_{i=0}^{t-1} \frac{\|\nabla f(x^{i+1}) - \nabla f(x^i)\|_\infty}{\|x^{i+1} - x^i\|_1}}}. \quad (3)$$

To approximate the smoothness constant, we interchange the l_∞ -norm and l_1 -norm in the expression, leveraging their duality relationship. Thus, we approximate the constant L_1 , not L_∞ , as indicated in Algorithm 2. Theoretically, this approach still requires memorizing $\nabla f(x^{t-1})$. For this reason, we consider a practical option by the approximation $\|\nabla f(x^t) - \nabla f(x^{t-1})\|_\infty \approx \max(|\max_j [\nabla f(x^t)]_j - \min_j [\nabla f(x^{t-1})]_j|, |\max_j [\nabla f(x^{t-1})]_j - \min_j [\nabla f(x^t)]_j|)$. It necessitates storing only two additional constants: $\max_j [\nabla f(x^{t-1})]_j$ and $\min_j [\nabla f(x^{t-1})]_j$. In the theoretical analysis, we provide convergence guarantees only for the λ^t choice, as in equation 3. However, we additionally validate the methods empirically with the approximation of the l_∞ -norm and provide an ablation study that shows a small deviation of the approximate solution from the exact one. More precisely, this approximation provides an upper bound on the initial l_∞ -norm, while remaining close to it (see Section 4 and Appendix A).

We present a theoretical analysis of a memory-efficient approach, utilizing an additional assumption on the L_1 -smoothness.

Assumption 3.11. The function f is L_1 -smooth, i.e., it satisfies $\|\nabla f(x) - \nabla f(y)\|_\infty \leq L_1 \|x - y\|_1$ for any $x, y \in \mathbb{R}^d$.

Now we present the convergence guarantees of Algorithm 2 with λ^t as in equation 3.

Theorem 3.12. Suppose Assumptions 3.11, 3.2, 3.3, 3.4 hold. We denote $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, $L_1^0 = \frac{\|\nabla f(x^1) - \nabla f(x^0)\|_\infty}{\|x^1 - x^0\|_1}$. Then Algorithm 2 with $d^0 < \Delta^*$ and $d \cdot \lambda^t$ as in equation 3, to reach ε -accuracy needs

$$\tilde{\mathcal{O}}\left(\frac{(\Delta^*)^2 (L_1)^3 d^2}{d^0 (L_1^0)^2 \varepsilon^2}\right) \text{ and } \tilde{\mathcal{O}}\left(\frac{\Delta^* (L_1)^3 d^2}{(L_1^0)^2 \varepsilon^2}\right) \text{ iterations with Options I and II, respectively.}$$

378 The result of Theorem 3.12 deteriorates the rate established in Theorem 3.5. Indeed, we can derive
 379 the $L_\infty \leq dL_1$ inequality. However, the proposed approach offers significant advantages in terms of
 380 memory efficiency.

381 Nevertheless, the theoretical convergence rates presented in this section are not optimal. In the
 382 stochastic case, we aim to achieve $\tilde{\mathcal{O}}\left(\frac{\Delta^* L + \|\sigma\|_1^2}{\varepsilon^2}\right)$ rate. This issue is discussed in detail in Appendix
 383 B, where we present an algorithm that attains this rate.

386 3.4 ALIAS WITH MOMENTUM

387 In previous sections, we presented
 388 methods that do not utilize the
 389 momentum parameter (Polyak, 1987;
 390 Nesterov et al., 2018). However,
 391 many modern optimizers, such as
 392 ADAM (Kingma, 2014), PRODIGY
 393 (Mishchenko & Defazio, 2023),
 394 MUON (Jordan et al., 2024), and
 395 MARS (Yuan et al., 2024), employ
 396 this technique. We address this gap
 397 in the current section and present
 398 Algorithm 3, which incorporates the
 399 momentum parameter into Algorithm
 400 2 in a manner similar to (Mishchenko
 401 & Defazio, 2023). Specifically, we
 402 include exponential moving averages
 403 of the first and second statistics, as in
 404 ADAM to aggregate past gradients and provide coordinate-wise
 405 normalization that mitigates sharp directions and gradient noise.

4 EXPERIMENTS

406 In this section, we present empirical results for the LLM pre-training task. In Appendix A, we
 407 validate our approach on vision tasks, specifically by fine-tuning the SWIN Transformer architecture
 408 (Liu et al., 2021). Our code is open-sourced¹.

409 **Language model pre-training.** Following the protocol of (Lialin et al., 2023), we train a LLaMA-
 410 based architecture (Touvron et al., 2023) with 130M parameters on the C4 dataset (Raffel et al., 2020).
 411 A detailed description of the experimental setup is provided in Appendix A.1. We compare several
 412 optimization methods: SIGN-SGD with a tuned constant learning rate (lr), and three methods using a
 413 tuned learning rate with a cosine scheduler (cosine sc) – namely, SIGN-SGD, STEEPEST DESCENT,
 414 and NORMALIZED SGD. All of these methods are compared against ALIAS (Algorithm 2), which is
 415 used without any tuning. Additionally, we evaluate all methods with weight decay (wd). We provide
 416 final validation loss and perplexity in Table 1.

417 Table 1: SIGN methods on LLAMA pre-training.

Algorithm	Validation Loss (↓)	Perplexity (↓)
SIGN-SGD (lr)	3.041	20.923
SIGN-SGD (lr, cosine sc)	2.992	19.923
STEEPEST DESCENT (lr, cosine sc)	3.035	20.791
NORMALIZED SGD (lr, cosine sc)	3.135	22.982
ALIAS (ours)	3.017	20.422
SIGN-SGD (wd, lr)	3.041	20.923
SIGN-SGD (wd, lr, cosine sc)	2.980	19.693
STEEPEST DESCENT (wd, lr, cosine sc)	3.022	20.537
NORMALIZED SGD (wd, lr, cosine sc)	3.006	20.169
ALIAS (wd) (ours)	3.006	20.169

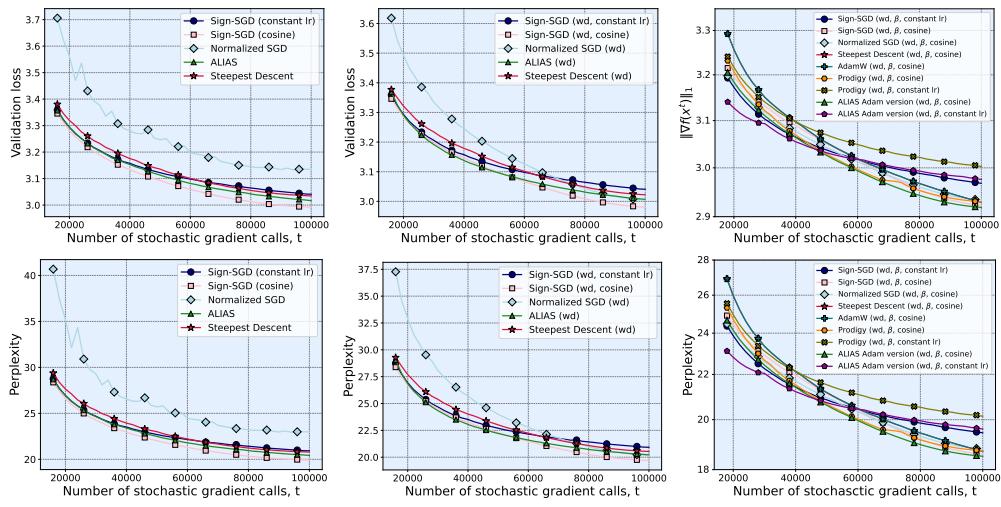
418 Table 2: SIGN-SGD methods with added momentum parameter (β), ADAMW and PRODIGY
 419 on LLAMA pre-training.

Algorithm	Validation Loss (↓)	Perplexity (↓)
SIGN-SGD (wd, β , lr)	2.968	19.459
SIGN-SGD (wd, β , lr, cosine sc)	2.923	18.596
STEEPEST DESC. (wd, β , lr, cosine sc)	2.932	18.765
NORM. SGD (wd, β , lr, cosine sc)	2.934	18.803
ADAMW (wd, β , lr, cosine sc)	2.929	18.698
PRODIGY (wd, β)	3.003	20.145
PRODIGY (wd, β , cosine sc)	2.930	18.727
ALIAS Adam version (wd, β) (ours)	2.976	19.609
ALIAS Adam version (wd, β, cosine sc) (ours)	2.918	18.504

420 In Table 2, we present the results for methods incorporating momentum (β) (all methods with weight
 421 decay). ALIAS Adam version utilizes sign descent with momentum and an additional scaling factor
 422 (see Algorithm 3 for details). We consider two options for this method: with and without a cosine

423 ¹https://anonymous.4open.science/r/PF_SignSGD

432 scheduler. We provide a comparison with ADAMW (Loshchilov, 2017) and PRODIGY (Mishchenko
 433 & Defazio, 2023). We test PRODIGY with and without a learning rate scheduler. We present the
 434 pre-training dynamic in Figure 1. These results coincides with those in Tables 1, 2.
 435



452 Figure 1: Comparison of SIGN-SGD methods in LLAMA pre-training. The left column shows
 453 results without weight decay, the central column presents results with weight decay (wd), and the
 454 right column displays results with weight decay (wd) and momentum parameter (β).
 455

456 We highlight that our basic ALIAS achieves performance only slightly inferior to that of SIGN-SGD
 457 with a tuned cosine scheduler. The Adam-based version of ALIAS outperforms all competing
 458 methods, including tuned ADAMW and the state-of-the-art parameter-free optimizer PRODIGY with a
 459 tuned cosine scheduler. These results are particularly competitive given that our approach eliminates
 460 the need for learning rate tuning – a significant practical advantage. This feature enhances the
 461 method’s usability, making it appealing for large-scale applications.
 462

463 **Memory-efficient version of Algorithm 2.** We proceed with testing the memory-efficient ap-
 464 proach, presented in Section 3.3. Recall that we approximate $\|\nabla f(x^t) - \nabla f(x^{t-1})\|_\infty \approx$
 465 $\max(|\max_j [\nabla f(x^t)]_j - \min_j [\nabla f(x^{t-1})]_j|, |\max_j [\nabla f(x^{t-1})]_j - \min_j [\nabla f(x^t)]_j|)$. We compare
 466 the performance of ALIAS with λ^t as in equation 3, considering exact and approximated l_∞ -norm
 467 (me), SIGN-SGD with a constant (tuned) stepsize, and SIGN-SGD with a (tuned) cosine scheduler.
 468 The results of the 130M LLAMA-based model pre-training are presented in Table 3. We provide an
 469 ablation comparing exact and approximated values of l_∞ -norms during training in Appendix A.
 470

471 The results indicate a slight performance degra-
 472 dation of the memory-efficient version of
 473 ALIAS compared to SIGN-SGD with a cosine
 474 scheduler baseline, as well as relative to the orig-
 475 inal ALIAS method. However, it is crucial to
 476 emphasize that this variant is a parameter-free
 477 algorithm that retains only the sign of the
 478 gradient from the previous iteration. Despite these
 479 simplifications, its performance remains competitive with significantly more memory-intensive
 480 methods. We report performance metrics, memory footprint, and runtime efficiency in Appendix A, along
 481 with detailed training configurations for full reproducibility.
 482

5 CONCLUSION

483 In this work, we present a novel parameter-free SIGN-SGD that eliminates manual stepsize selection.
 484 The method is analyzed in deterministic, stochastic, and distributed settings. Additionally, we
 485 introduce a memory-efficient variant that stores only gradient signs while maintaining adaptivity. We
 486 also explore a momentum-adapted version that demonstrates strong performance in practice.
 487

486 REFERENCES
487

488 Dan Alistarh, Torsten Hoefer, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric
489 Renggli. The convergence of sparsified gradient methods. *Advances in Neural Information
490 Processing Systems*, 31, 2018.

491 Zeyuan Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex
492 sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

493

494 Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma,
495 José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep
496 learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big
497 Data*, 8:1–74, 2021.

498 Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth.
499 Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):
500 165–214, 2023.

501

502 Amit Attia and Tomer Koren. Sgd with adagrad stepsizes: Full adaptivity with high probability to
503 unknown parameters, unbounded gradients and affine variance. In *International Conference on
504 Machine Learning*, pp. 1147–1171. PMLR, 2023.

505 Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic
506 gradients. In *International Conference on Machine Learning*, pp. 404–413. PMLR, 2018.

507

508 Lukas Balles, Fabian Pedregosa, and Nicolas Le Roux. The geometry of sign gradient descent. *arXiv
509 preprint arXiv:2002.08056*, 2020.

510 Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd:
511 Compressed optimisation for non-convex problems. In *International Conference on Machine
512 Learning*, pp. 560–569. PMLR, 2018.

513

514 Stephen Boyd, Lin Xiao, and Almir Mutapcic. Subgradient methods. *lecture notes of EE392o,
515 Stanford University, Autumn Quarter*, 2004(01), 2003.

516 Yair Carmon and Oliver Hinder. Making sgd parameter-free. In *Conference on Learning Theory*, pp.
517 2360–2389. PMLR, 2022.

518

519 Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM
520 transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

521

522 Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on
523 Learning Theory*, pp. 874–894. PMLR, 2019.

524

525 Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in
526 banach spaces. In *Conference On Learning Theory*, pp. 1493–1529. PMLR, 2018.

527

528 George E Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry,
529 Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, et al. Bench-
marking neural network training algorithms. *arXiv preprint arXiv:2306.07179*, 2023.

530

531 Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *Interna-
532 tional Conference on Machine Learning*, pp. 7449–7479. PMLR, 2023.

533

534 Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale
535 hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*,
pp. 248–255. Ieee, 2009.

536

537 John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and
538 stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

539

Vineet Gupta, Tomer Koren, and Yoram Singer. A unified approach to adaptive regularization in
online and stochastic optimization. *arXiv preprint arXiv:1706.06569*, 2017.

540 Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza
 541 Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al.
 542 Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

543

544 Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd's best friend: A parameter-free dynamic step
 545 size schedule. In *International Conference on Machine Learning*, pp. 14465–14499. PMLR, 2023.

546

547 Richeng Jin and Huaiyu Dai. Noisy signsgd is more differentially private than you (might) think. In
 548 *Forty-second International Conference on Machine Learning*, 2025.

549

550 Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy
 551 Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL <https://kellerjordan.github.io/posts/muon/>.

552

553 Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback
 554 fixes signsgd and other gradient compression schemes. In *International Conference on Machine
 555 Learning*, pp. 3252–3261. PMLR, 2019.

556

557 Priya Kasimbeg, Vincent Roulet, Naman Agarwal, Sourabh Medapati, Fabian Pedregosa, Atish
 558 Agarwala, and George E Dahl. How far away are truly hyperparameter-free learning algorithms?
 559 *arXiv preprint arXiv:2505.24005*, 2025.

560

561 Ahmed Khaled, Konstantin Mishchenko, and Chi Jin. Dowg unleashed: An efficient universal
 562 parameter-free gradient descent method. *Advances in Neural Information Processing Systems*, 36:
 563 6748–6769, 2023.

564

565 Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*,
 566 2014.

567

568 Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local
 569 minima? In *International conference on machine learning*, pp. 2698–2707. PMLR, 2018.

570

571 Jakub Konečný, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization:
 572 Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

573

574 Nikita Kornilov, Philip Zmushko, Andrei Semenov, Mark Ikonnikov, Alexander Gasnikov, and
 575 Alexander Beznosikov. Sign operator for coping with heavy-tailed noise in non-convex optimiza-
 576 tion: High probability bounds under (l_0, l_1) -smoothness. *arXiv preprint arXiv:2502.07923*,
 577 2025.

578

579 Itai Kreisler, Maor Ivgi, Oliver Hinder, and Yair Carmon. Accelerated parameter-free stochastic
 580 optimization. *arXiv preprint arXiv:2404.00666*, 2024.

581

582 Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the
 583 main factor behind the gap between sgd and adam on transformers, but sign descent might be.
 584 *arXiv preprint arXiv:2304.13960*, 2023.

585

586 Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1.
 587 Springer, 2020.

588

589 Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

590

591 Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. Relora: High-
 592 rank training through low-rank updates. In *The Twelfth International Conference on Learning
 593 Representations*, 2023.

594

595 Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized
 596 non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:
 597 85–116, 2022.

598

599 Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo.
 600 Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the
 601 IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

594 Zijian Liu, Srikanth Jagabathula, and Zhengyuan Zhou. Near-optimal non-convex stochastic opti-
 595 mization under generalized smoothness. *arXiv preprint arXiv:2302.06032*, 2023.

596

597 I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

598

599 Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas.
 600 Communication-efficient learning of deep networks from decentralized data. In *Artificial intelli-
 601 gence and statistics*, pp. 1273–1282. PMLR, 2017.

602 H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert
 603 spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pp.
 604 1020–1039. PMLR, 2014.

605

606 Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free
 607 learner. *arXiv preprint arXiv:2306.06101*, 2023.

608

609 Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert Gower. Directional
 610 smoothness and gradient methods: Convergence and adaptivity. *Advances in Neural Information
 611 Processing Systems*, 37:14810–14848, 2024.

612

613 Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms
 for machine learning. *Advances in neural information processing systems*, 24, 2011.

614

615 Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method
 616 efficiency in optimization. 1983.

617

618 Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

619

620 Francesco Orabona. Dimension-free exponentiated gradient. *Advances in Neural Information
 621 Processing Systems*, 26, 2013.

622

623 Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*,
 624 2019.

625

626 Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in
 627 Neural Information Processing Systems*, 29, 2016.

628

629 Boris T Polyak. Introduction to optimization. 1987.

630

631 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi
 632 Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text
 633 transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

634

635 Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical
 636 statistics*, pp. 400–407, 1951.

637

638 Mher Safaryan and Peter Richtárik. Stochastic sign descent methods: New algorithms and better
 639 theory. In *International Conference on Machine Learning*, pp. 9224–9234. PMLR, 2021.

640

641 Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, and Robert M Gower. Momo:
 642 Momentum models for adaptive learning rates. *arXiv preprint arXiv:2305.07583*, 2023.

643

644 Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its
 645 application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014, pp.
 646 1058–1062. Singapore, 2014.

647

648 Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

649

650 Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint
 651 arXiv:1907.04232*, 2019.

652

653 Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint
 654 arXiv:1002.4862*, 2010.

648 Matthew Streeter and H Brendan McMahan. No-regret algorithms for unconstrained online convex
 649 optimization. *arXiv preprint arXiv:1211.2260*, 2012.

650

651 Tao Sun, Qingsong Wang, Dongsheng Li, and Bao Wang. Momentum ensures convergence of signsgd
 652 under weaker assumptions. In *International Conference on Machine Learning*, pp. 33077–33099.
 653 PMLR, 2023.

654 Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop, coursera: Neural networks for machine
 655 learning. *University of Toronto, Technical Report*, 6, 2012.

656

657 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
 658 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and
 659 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

660 A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

661

662 Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S
 663 Rellermeyer. A survey on distributed machine learning. *Acm computing surveys (csur)*, 53(2):
 664 1–33, 2020.

665 Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex
 666 landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

667

668 Huizhuo Yuan, Yifeng Liu, Shuang Wu, Xun Zhou, and Quanquan Gu. Mars: Unleashing the power
 669 of variance reduction for training large models. *arXiv preprint arXiv:2411.10438*, 2024.

670 Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*,
 671 2012.

672

673 Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing
 674 what makes a good optimizer for language models. *arXiv preprint arXiv:2407.07972*, 2024.

675 Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global
 676 minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

677

678 Philip Zmushko, Aleksandr Beznosikov, Martin Takáč, and Samuel Horváth. Frugal: Memory-
 679 efficient optimization by reducing state overhead for scalable training. *arXiv preprint
 680 arXiv:2411.07837*, 2024.

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702	CONTENTS	
703		
704		
705	1 Introduction	1
706		
707	2 Brief Literature Review and Contributions	2
708	2.1 Related work	2
709	2.2 Contributions	3
710		
711		
712	3 Algorithms and Convergence Analysis	4
713	3.1 Exact gradients setting	4
714	3.2 Stochastic gradients setting	5
715	3.3 Memory-efficient ALIAS	7
716	3.4 ALIAS with momentum	8
717		
718		
719		
720	4 Experiments	8
721		
722	5 Conclusion	9
723		
724		
725	A Additional Experiments	16
726	A.1 LLaMA pre-training	16
727	A.1.1 Experimental setup.	16
728	A.1.2 Additional results	16
729	A.1.3 Comparison with parameter-free approaches	18
730	A.1.4 Experiments on big model	18
731	A.1.5 Compute resources.	19
732	A.2 Tiny ImageNet classification with Swin Transformer Fine-Tuning	20
733	A.2.1 Experimental setup	20
734	A.2.2 Performance on Image Classification	20
735	A.2.3 Compute Resources	20
736	A.3 AlgoPerf benchmark	21
737		
738		
739		
740		
741		
742		
743	B SIGN-SGD with Additional Stepsize Search Procedure	21
744	B.1 SOS SIGN-SGD experiments	23
745	B.1.1 Logistic regression.	23
746	B.1.2 Non-convex problem	24
747		
748		
749	C Additional Notation and General Inequalities	24
750		
751		
752	D Lemmas for SOS SIGN-SGD	25
753		
754	E Main Proofs and Details for SOS SIGN-SGD	28
755	E.1 Exact gradient setting	28

756	E.2 Stochastic gradient setting	31
757	E.3 Distributed setting	36
758		
759		
760	F Proofs for ALIAS	39
761	F.1 Exact gradient setting	39
762	F.2 Stochastic gradient setting	43
763	F.3 Distributed setting	51
764		
765	F.4 Memory-efficient ALIAS	55
766		
767		
768	G Steepest Descent	58
769		
770		
771		
772		
773		
774		
775		
776		
777		
778		
779		
780		
781		
782		
783		
784		
785		
786		
787		
788		
789		
790		
791		
792		
793		
794		
795		
796		
797		
798		
799		
800		
801		
802		
803		
804		
805		
806		
807		
808		
809		

810 A ADDITIONAL EXPERIMENTS
811

812 This section supplements our experimental validation by examining the internal mechanisms of
813 parameter-free sign-based optimizers across LLaMA pre-training and Tiny ImageNet classification.
814 We analyze how step-size dynamics naturally emerge without manual scheduling, investigate memory
815 consumption and computational time compared to established optimizers, and demonstrate robustness
816 to hyperparameter choices.

817
818 A.1 LLAMA PRE-TRAINING
819

820 A.1.1 EXPERIMENTAL SETUP.

821 Our experiments use a LLaMA-based architecture (Touvron et al., 2023) equipped with RMSNorm
822 and SwiGLU (Shazeer, 2020) activations, trained on the C4 dataset (Raffel et al., 2020). The training
823 consists of 100k steps. We use batch size of 512 sequences and sequence length of 256, as in Lialin
824 et al. (2023), and T5 tokenizer with the dictionary size of 32k since it was originally trained on C4.
825

826 For all experiments, the respective optimization method is applied to the main model parameters,
827 while the LM Head layer is optimized with AdamW. This design follows prior work Zhao et al. (2024)
828 which showed that the LM Head layer requires more fine-grained learning rate adjustment.
829

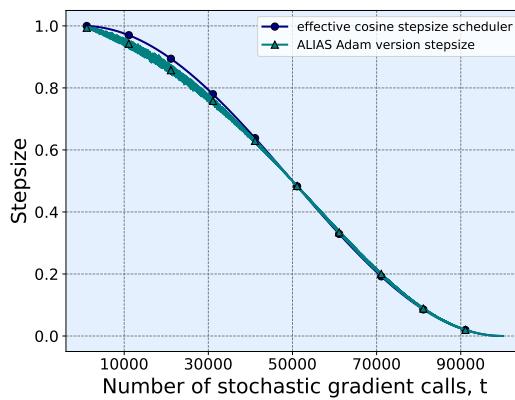
830 The learning rate was selected through a grid search with multiplicative step of $10^{\frac{1}{4}}$. We employ a
831 cosine learning rate schedule with a warmup of 10% of the total steps and decay to 10% of the peak
832 learning rate. For ALIAS Adam version (Algorithm 3), we choose stepsize $\gamma^t = 10^{-3}$.

833 The weight decay value was selected from $[0, 0.01, 0.1]$ through validation. We also applied gradient
834 clipping with threshold of 1.0 for all methods except STEEPEST DESCENT and NORMALIZED SGD.
835 All methods with momentum utilize the Nesterov acceleration scheme with a momentum value of
836 0.9. For AdamW we use the standard hyperparameters: $\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 1e - 8$.

837 A.1.2 ADDITIONAL RESULTS
838

839 In this section, we explore key aspects of our method. We analyze the stepsize derived from our
840 approach and compare it to the effective learning rate induced by the cosine scheduler. Next, we ex-
841 amine the memory and computational efficiency of all considered optimizers. We present an ablation
842 study on the approximation used in the stepsize of the memory-efficient variant, demonstrating its
843 close alignment with exact computation. We provide empirical evidence for the robustness of ALIAS
844 to an additional constant L_∞ term (see Remark 3.6). Finally, we discuss the question regarding the
845 performance dependence on the choice of the initial value d^0 and the level of gradient noise.
846

847 **Study on the stepsize.** A question arises regarding how $\gamma^t \sqrt{\frac{(d^t)^2}{1 + \frac{v^{t+1} - (m^{t+1})^2}{(m^{t+1})^2}}}$ performs compared
848 to the effective cosine scheduler when γ^t remains constant. This pairing is presented in Figure 2.



863 Figure 2: Comparison of ALIAS Adam version stepsize with constant γ^t with effective cosine
864 stepsize scheduler.

864 One can state that the cosine nature of the stepsize is automatically obtained. This feature highlights
 865 the distinctiveness of our parameter-free approach.
 866

867 **Study on the time and memory consumption.** In Table 4, we present details of memory require-
 868 ments and time consumption per-iteration.
 869

870 Table 4: Comparison of memory and time consumption.
 871

Algorithm	Memory consumption (gb)	Time consumption per-iteration (s)
SIGN-SGD	0.41	0.004
STEEPEST DESCENT	0.41	0.01
NORMALIZED SGD	0.41	0.01
ADAMW	1.5	0.007
PRODIGY	3.5	0.05
ALIAS (ours)	1.22	0.01
ALIAS Adam version (ours)	1.91	0.03
memory-efficient ALIAS (ours)	0.41	0.007

882 Table 4 shows a higher time per-iteration for ALIAS Adam version and PRODIGY, which we adopt
 883 from the work (Mishchenko & Defazio, 2023). We attribute this to the suboptimal implementation of
 884 these algorithms, in contrast to others that have been utilized for an extended period. Simultaneously,
 885 our algorithms are comparable to ADAMW in terms of required memory, while PRODIGY occupies
 886 more GPU resources because it stores a vector of initial model parameters. Note that the memory-
 887 efficient version of ALIAS is superior to ADAMW and comparable to the basic SIGN-SGD.
 888

889 **Study on the memory-efficient ALIAS.**

890 We now analyze the memory-efficient variant
 891 of ALIAS, focusing on the accuracy of
 892 the approximated l_∞ -norm used in its
 893 update rule. Figure 3 shows the dynamics
 894 of $\|\nabla f(x^t) - \nabla f(x^{t-1})\|_\infty$ across iterations,
 895 along with the deviation range of its approxi-
 896 mation (see Section 3.3 for details on the ap-
 897 proximation scheme). The ablation study re-
 898 veals that the approximate norm deviates from
 899 the exact value by approximately 50% on aver-
 900 age. Notably, the approximation consistently
 901 exceeds the true norm – as expected,
 902 since it constitutes an upper bound by
 903 design. This leads to smaller effective step-
 904 sizes, which explains the slightly degraded per-
 905 formance of the memory-efficient variant compared
 906 to the basic ALIAS algorithm (Algorithm 2).
 907

908 **Study on the robustness to L_∞ .** In Table 5,
 909 we provide empirical evidence supporting the
 910 claim made in Section 3.1 that the modification
 911 of ALIAS (Algorithm 2) is robust concerning
 912 the L_∞ parameter. Hence, although the ver-
 913 sion of the algorithm considered in Remark 3.6
 914 requires prior knowledge of L_∞ , this additive
 915 factor has negligible impact on the practical con-
 916 vergence of Algorithm 2.
 917

918 **Performance dependence on d^0 choice.** In this paragraph, we investigate the robustness of our
 919 ALIAS Adam version concerning the choice of the initial distance d^0 . To this end, we compare the
 920 performance of Algorithm 3 on LLAMA pre-training using $d^0 = 1$ and $d^0 = 10^{-3}$. In both cases,
 921 we obtain the same validation metric: validation loss = 2.918. Based on these results, we conclude
 922 that our method is insensitive to the choice of d^0 .

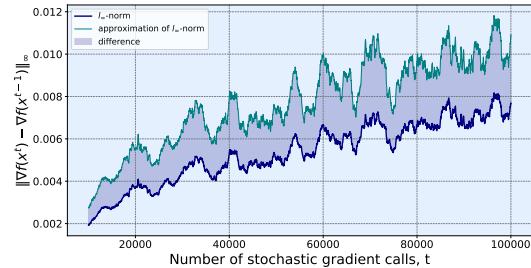


Figure 3: Ablation study on approximated l_∞ -norm deviation from the exact one in the memory-efficient version of ALIAS.

Table 5: Robustness to L_∞ .

L_∞ value	Validation loss (\downarrow)
0	3.006
50	3.006
100	3.007
500	3.005
1000	3.006

918 **Performance dependence on gradient noise.** We investigate the dependence of the performance
 919 of our ALIAS procedure (Algorithm 2) on the level of gradient noise. To simulate different noise
 920 levels, we vary the batch size. Indeed, decreasing the batch size increases the stochasticity and the
 921 variance of the gradient estimate, thereby leading to a higher level of gradient noise. While in previous
 922 experiments we used a batch size of 512 sequences, here we use 256, 128, and 64 sequences. Then
 923 we compare the validation loss on these runs. Table 6 provides a pairwise comparison of ALIAS and
 924 SIGN-SGD across these batch sizes.

925 Table 6: SIGN-SGD and ALIAS with different batch sizes on LLAMA pre-training.
 926

Batch Size (# of Sequences)	Algorithm	Validation Loss (\downarrow)
512	SIGN-SGD	2.980
512	ALIAS (ours)	3.006
256	SIGN-SGD	2.986
256	ALIAS (ours)	3.013
128	SIGN-SGD	2.992
128	ALIAS (ours)	3.021
64	SIGN-SGD	2.999
64	ALIAS (ours)	3.029

937
 938 The experimental results demonstrate that, when the batch size is reduced – thereby increasing the
 939 level of gradient noise – both SIGN-SGD and ALIAS exhibit a comparable decline in performance.
 940 This suggests that ALIAS is not disproportionately affected by the increased stochasticity in gradient
 941 estimates, underscoring its robustness to gradient noise.
 942

943 A.1.3 COMPARISON WITH PARAMETER-FREE APPROACHES
 944

945 In this section, we present an experimental comparison of our ALIAS Adam version algorithm
 946 with competing parameter-free optimization methods. For this additional evaluation, we selected the
 947 following approaches: DOG (Ivgi et al., 2023), D-ADAPTATION (Defazio & Mishchenko, 2023), and
 948 MOMO (Schaipp et al., 2023). These methods are chosen based on their performance reported in
 949 the work (Kasimbeg et al., 2025) on the ALGOPERF benchmark (Dahl et al., 2023). Our validation
 950 results for pre-training the LLAMA-based architecture are summarized in Table 7.
 951

952 Table 7: Parameter-free methods on LLAMA pre-training.
 953

Algorithm	Validation Loss (\downarrow)	Perplexity (\downarrow)
DOG	2.939	18.897
D-ADAPTATION (with Adam)	2.927	18.672
MOMO (with Adam)	2.925	18.634
PRODIGY	2.930	18.727
ALIAS Adam version (wd) (ours)	2.918	18.504

961
 962 These results complement our comparison against sign-based methods and ADAMW. They demon-
 963 strate that our approach achieves stronger performance than prior parameter-free methods.
 964

965 A.1.4 EXPERIMENTS ON BIG MODEL
 966

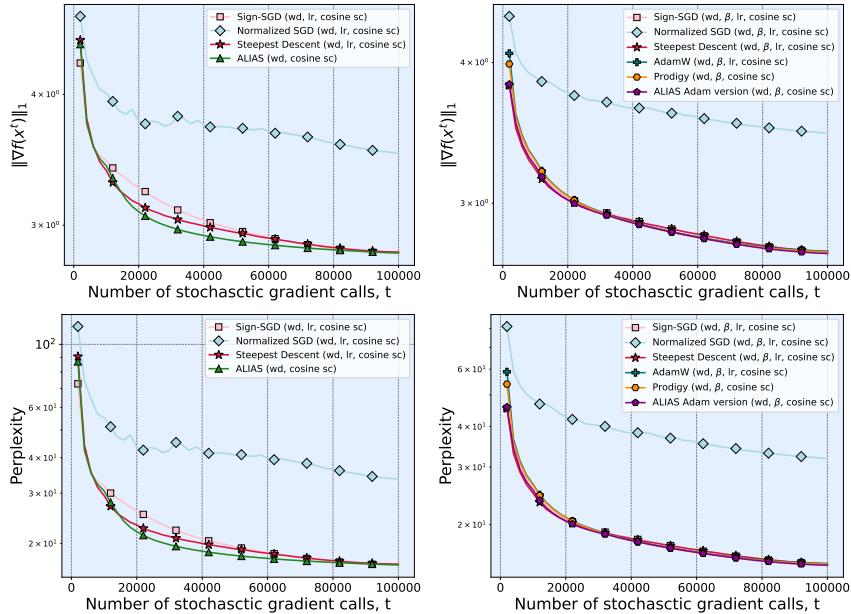
967 We evaluate the methods on LLAMA with 350M parameters. The training setup remains consistent
 968 with the previous experiment. However, the number of layers in the model increases, leading to a
 969 total parameter count that rises from 130M to 350M. This experiment is essential to demonstrate the
 970 sustainability of our approaches to increasing dimensionality. We conduct experiments comparing
 971 methods with and without the momentum parameter β along with weight decay. The results are
 972 presented Tables 8, 9, and Figure 4.

Table 8: SIGN-SGD methods on LLAMA pre-training.

Algorithm	Validation Loss (\downarrow)	Perplexity (\downarrow)
SIGN-SGD (wd, lr, cosine sc)	2.819	16.760
STEEPEST DESCENT (wd, lr, cosine sc)	2.828	16.912
NORMALIZED SGD (wd, lr, cosine sc)	3.510	33.448
ALIAS (wd) (ours)	2.821	16.793

Table 9: SIGN-SGD methods with added momentum parameter (β), ADAMW (wd) and PRODIGY on LLAMA pre-training.

Algorithm	Validation Loss (\downarrow)	Perplexity (\downarrow)
SIGN-SGD (wd, β , lr, cosine sc)	2.717	15.135
STEEPEST DESCENT (wd, β , lr, cosine sc)	2.711	15.044
NORMALIZED SGD (wd, β , lr, cosine sc)	3.460	31.817
ADAMW (wd, β , lr, cosine sc)	2.719	15.165
PRODIGY (wd, β , cosine sc)	2.715	15.105
ALIAS Adam version (wd, β, cosine sc) (ours)	2.707	14.984

Figure 4: Comparison of SIGN-SGD methods on 350M parameters LLAMA pre-training. Left column is results for methods with weight decay and without momentum parameter β , right column – methods with momentum β .

The results are consistent with those obtained for the smaller model. Among the momentum-based methods, ALIAS Adam version demonstrates the best performance, while among the methods without momentum, ALIAS exhibits comparable performance to other solutions.

A.1.5 COMPUTE RESOURCES.

We conducted all experiments described in Section A.1 on the cluster equipped with 4×NVIDIA A100 GPUs. A complete run of 100,000 steps took approximately 12 hours using a full node.

1026
1027

A.2 TINY IMAGENET CLASSIFICATION WITH SWIN TRANSFORMER FINE-TUNING

1028
1029

A.2.1 EXPERIMENTAL SETUP

1030
1031

Our image classification experiments on the Tiny ImageNet dataset (Le & Yang, 2015) employed the Tiny Swin Transformer architecture (Liu et al., 2021). This lightweight variant of the Swin Transformer is characterized by its hierarchical design and the use of shifted windows for efficient self-attention computation. The specific configuration utilized involved non-overlapping 4×4 input patches and a 7×7 window size for local self-attention.

1032
10331034
1035
1036
1037

We initialized the model using pretrained weights from ImageNet-1K (Deng et al., 2009), specifically the `swin_T_patch4_window7_224` checkpoint provided in the official Swin Transformer repository². The model was then fine-tuned on Tiny ImageNet.

1038
1039
1040
1041

The Tiny ImageNet dataset comprises 200 classes with images of 64×64 resolution. To meet the model’s input requirements, all images were upsampled to 224×224 . A standard ImageNet-style data augmentation pipeline was implemented, including random resized cropping and horizontal flipping.

1042
1043
1044
1045
1046
1047
1048

Training spanned 50 epochs, with a batch size of 256. The learning rate was determined via a grid search, employing a multiplicative step of $10^{\frac{1}{4}}$. A cosine learning rate schedule was adopted, featuring a linear warm-up phase for the initial 10% of total training steps, followed by decay to 10% of the peak learning rate. Weight decay was selected from $\{0, 0.01, 0.1\}$ based on validation performance. All optimization methods incorporated gradient clipping with a threshold of 1.0. When momentum was applied, Nesterov acceleration with a coefficient of 0.99 was used. For AdamW, the standard configuration of $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$ was maintained.

1049
1050
1051
1052
1053
1054

A.2.2 PERFORMANCE ON IMAGE CLASSIFICATION

Further results and training curves for the Tiny Swin Transformer on the Tiny ImageNet classification task are presented in Figure 5 and Table 10. We provide plots for the same methods with the incorporated momentum parameter as for the LLAMA pre-training task.

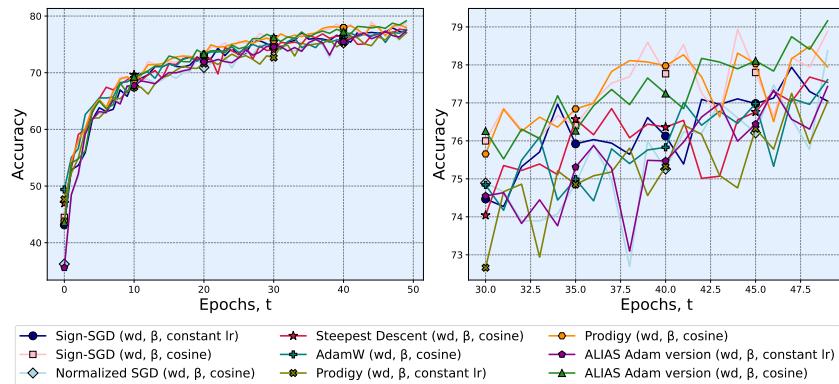
1065
1066
1067

Figure 5: SIGN-SGD methods with added momentum parameter (β), ADAMW (wd) and PRODIGY on SWIN fine-tuning. Left plot represents full process of training, right plot demonstrates accuracy on last 20 epoch.

1071
1072
1073
1074

The results demonstrate the superiority of our algorithms over both tuned sign-based methods and advanced optimizers, such as PRODIGY and ADAMW.

1075
1076

A.2.3 COMPUTE RESOURCES

1077
1078
1079

We conducted all experiments described in Section A.2 using a single NVIDIA A100 GPU. A complete run of 50 epochs required approximately 3 hours.

²<https://github.com/microsoft/Swin-Transformer/blob/main/MODELHUB.md>

1080 Table 10: Final accuracy of SIGN-SGD methods with added momentum parameter (β), ADAMW
 1081 (wd) and PRODIGY on SWIN fine-tuning.

Algorithm	Final accuracy (\uparrow)
SIGN-SGD (wd, β , lr)	77.045
SIGN-SGD (wd, β , lr, cosine sc)	78.885
NORMALIZED SGD (wd, β , lr, cosine sc)	78.375
STEEPEST DESCNET (wd, β , lr, cosine sc)	77.547
ADAMW (wd, β , lr, cosine sc)	77.612
PRODIGY (wd, β)	77.035
PRODIGY (wd, β , cosine sc)	77.944
ALIAS Adam version (wd, β) (ours)	77.433
ALIAS Adam version (wd, β , cosine sc) (ours)	79.161

A.3 ALGOPERF BENCHMARK

In this section, we evaluate our method on some tasks from the ALGOPERF benchmark (Dahl et al., 2023). To test our approach across different modalities, we chose the MRI reconstruction and molecular property prediction (MPP) tasks. We preserve the original setups from the benchmark implementation³. Specifically, for the MRI reconstruction task, we use the fastMRI dataset and a U-Net model; for molecular property prediction, we utilize the OGBG dataset with a GNN model. We validate only our ALIAS Adam version algorithm. The results for the other methods are taken from Table 4 in (Kasimbeg et al., 2025), which reports comparisons between parameter-free optimizers and tuned ADAMW. For our method, we fix $\gamma^t = 10^{-3}$. The results are presented in Table 11.

Table 11: Parameter-free methods and ADAMW on MRI reconstruction and molecular property prediction tasks.

Algorithm	MRI, SSIM (\uparrow)	MPP, mAP (\uparrow)
ADAMW	0.723	0.254
DoG	0.714	0.231
D-ADAPTATION (with Adam)	0.722	0.221
MOMO (with Adam)	0.723	0.221
PRODIGY	0.723	0.212
ALIAS Adam version (wd) (ours)	0.724	0.242

The results demonstrate that our approach improves upon the metrics of prior parameter-free methods on the evaluated tasks. We also note that, for the MRI reconstruction task, the performance of our method surpasses that of the tuned ADAMW.

B SIGN-SGD WITH ADDITIONAL STEPSIZE SEARCH PROCEDURE

In this section, we present an algorithm that achieves near-optimal convergence rates for SIGN-SGD – $\tilde{\mathcal{O}}\left(\frac{\Delta^* L_\infty}{\varepsilon^2}\right)$ in the deterministic case, and $\mathcal{O}\left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \|\sigma\|_1^2\right)$ in the stochastic case. The method does not utilize prior knowledge about the parameters of the problem and incorporates an additional automatic stepsize search procedure.

Exact gradients. To design the necessary algorithm, we should provide a stepsize γ in Algorithm 1 that yields an estimate as in equation 1. Let us begin with the description of the approximation of the stepsize 1 that we utilize. We establish that the desired value is $\gamma = \frac{\mathfrak{N}_T}{\mathfrak{D}_T}$, where $\mathfrak{N}_T = \tilde{\Delta}_T = f(x^0) - \min_{0 \leq t \leq T} f(x^t)$ is the numerator and $\mathfrak{D}_T = \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1$

³https://github.com/mlcommons/algorithmic-efficiency/blob/main/docs/GETTING_STARTED.md

is the denominator. The intuition behind this choice is that due to L_∞ -smoothness, we have $\mathfrak{D}_T \sim L_\infty \sum_{t=0}^{T-1} \|x^{t+1} - x^t\|_\infty = \gamma L_\infty \sum_{t=0}^{T-1} \|\text{sign}(\nabla f(x^t))\|_\infty = \gamma L_\infty T$; then γ has $\frac{\sqrt{\Delta_T}}{\sqrt{L_\infty T}}$ magnitude. However, we face a more complex situation compared to the regret minimization paradigm: in our case, $\tilde{\Delta}_T$ can be non-negative (in regret minimization, the analog of Δ_T is the norm of the points' difference $\|x^0 - x^T\|$ (Carmon & Hinder, 2022) which is always positive). To address this, we add an extra step to the SIGN-SGD algorithm. Define $e = \text{sign}(\nabla f(x^{-1}))$. Let τ be a small parameter. The update is:

$$f(x^0) = \min \{f(x^{-1} + \tau e), f(x^{-1} - \tau e)\}, \quad (4)$$

The rationale behind selecting the step is as follows. Due to the smoothness of the objective function, there exists a small neighborhood around any point within which moving in any direction decreases the objective value. The exception arises when x^{-1} is the minimum itself. In this case, the sign descent algorithm would not take any steps, and we return this point as the solution. Since the neighborhood size τ depends on L_∞ , we iteratively decrease τ until it is sufficiently small. The choice of τ and the guarantee $f(x^0) < f(x^{-1})$ are discussed in Lemma D.4. In this manner, we ensure that $\mathfrak{N}_T = \tilde{\Delta}_T = f(x^{-1}) - \min_{-1 \leq t \leq T} f(x^t) > 0$. To prevent the denominator from being zero, we introduce a small constant ζ , which represents the minimum gradient norm encountered during learning. This leads to $\mathfrak{D}_T = \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 + \zeta$ (see Lemma D.2 for details). However, determining these values necessitates completing all T iterations. To address this, we employ the BISECTION procedure from (Carmon & Hinder, 2022), which is outlined in Algorithm 4.

Our goal is to have $\gamma = \phi(\gamma) = \frac{\mathfrak{N}_T(\gamma)}{\mathfrak{D}_T(\gamma)}$. To find such γ , we take an initial interval $[\gamma_{\text{lo}}, \gamma_{\text{hi}}]$ and, iteratively narrowing it, obtain a small enough interval $[\gamma_{\text{lo}}^*, \gamma_{\text{hi}}^*]$ that contains the $\gamma - \phi(\gamma) = 0$ point. To perform this, we firstly have to make sure that the initial interval contains the desired point. For this purpose, we require $\gamma_{\text{hi}} > \phi(\gamma_{\text{hi}})$ and $\gamma_{\text{lo}} < \phi(\gamma_{\text{lo}})$. We designate the group of these two requirements as the bisection start condition (Lines 3, 5). Note that we can always satisfy the first condition, as shown in Lemma D.2. Regarding the second requirement, we can choose a sufficiently small initial γ_{lo} value. Even if γ_{lo} is still greater than $\phi(\gamma_{\text{lo}})$, we can select this γ_{lo} value as the desired stepsize without performing the BISECTION procedure, thereby obtaining optimal convergence guarantees. This is demonstrated in **Step 2** of the proof of Theorem B.1 (Theorem E.2). This enables us to avoid early infinite termination (non-compliance with the first condition) and prevents convergence from being compromised by early non-infinite termination (non-compliance with the second condition). Additionally, we ensure that, by entering the procedure with the desired point between γ_{lo} and γ_{hi} , it remains invariant throughout the procedure. Indeed, at each iteration we compute γ_{mid} as the geometric average of the bounds and perform T iterations of the SIGN-SGD method with this

Algorithm 4 BISECTION procedure

```

1: Input: Optimal stepsize value  $\phi(\gamma)$ , lower stepsize bound  $\gamma_{\text{lo}}$ ,  

   upper stepsize bound  $\gamma_{\text{hi}}$ ,  $x^{-1} \in \mathbb{R}^d$ , number of iterations  $T$ 
2:  $\phi(\gamma)$  (it is always in the form  $\phi(\gamma) = \frac{\mathfrak{N}_T(\gamma)}{\mathfrak{D}_T(\gamma)}$ )
3: if  $\gamma_{\text{hi}} \leq \phi(\gamma_{\text{hi}})$  then return  $\infty$  // Early infinite termination
4: end if
5: if  $\gamma_{\text{lo}} > \phi(\gamma_{\text{lo}})$  then return  $\gamma_{\text{lo}}^* = \gamma_{\text{lo}}$  // Early non-infinite termination
6: end if
7: while  $\gamma_{\text{hi}} > 2\gamma_{\text{lo}}$  do
8:    $\gamma_{\text{mid}} = \sqrt{\gamma_{\text{lo}}\gamma_{\text{hi}}}$ 
9:    $\mathfrak{N}_T(\gamma_{\text{mid}}), \mathfrak{D}_T(\gamma_{\text{mid}}) \leftarrow \text{SIGN-SGD}(x^{-1}, T, \gamma_{\text{mid}})$  // First
   step in Sign-SGD is made by equation 4
10:  if  $\gamma_{\text{mid}} \leq \phi(\gamma_{\text{mid}})$  then
11:     $\gamma_{\text{lo}} = \gamma_{\text{mid}}$ 
12:  else
13:     $\gamma_{\text{hi}} = \gamma_{\text{mid}}$ 
14:  end if // Bisection invariants:  $\gamma_{\text{lo}} < \phi(\gamma_{\text{lo}})$ ,  $\gamma_{\text{hi}} > \phi(\gamma_{\text{hi}})$ 
15: end while // Bisection stop condition:  $\gamma_{\text{hi}} \leq 2\gamma_{\text{lo}}$ 
16: if  $\mathfrak{N}_T(\gamma_{\text{hi}}) \leq \mathfrak{N}_T(\gamma_{\text{lo}}) \frac{\phi(\gamma_{\text{hi}})}{\gamma_{\text{hi}}}$  then return  $\gamma_{\text{hi}}^* = \gamma_{\text{hi}}$  //  $\gamma_{\text{hi}}$  return
   condition
17: else return  $\gamma_{\text{lo}}^* = \gamma_{\text{lo}}$  //  $\gamma_{\text{lo}}$  return condition
18: end if

```

Algorithm 5 SOS SIGN-SGD

```

1: Input: Initial stepsize bound  $\gamma_s$ , initial bound step  $k$ , start point
    $x^{-1} \in \mathbb{R}^d$ , number of iterations  $T$ 
2:  $\gamma_0 = \text{BISECTION}(\phi(\gamma), \gamma_s, 2^{2^k} \gamma_s, T)$ 
3:  $x^T = \text{SIGN-SGD}(x^{-1}, T, \gamma_0)$ 

```

22

1188 stepsize to find $\phi(\gamma_{\text{mid}})$ (Lines 8, 9). It remains for us to choose such a part of the segment $([\gamma_{\text{lo}}, \gamma_{\text{mid}}]$
 1189 or $[\gamma_{\text{mid}}, \gamma_{\text{hi}}]$) in which $\phi(\gamma_{\text{mid}})$ lies (Lines 10 - 14). We perform this bisection, until γ_{hi} exceeds γ_{lo}
 1190 by more than 2 times (Line 7). In the end, by utilizing return conditions, the procedure returns γ_{lo}^* or
 1191 γ_{hi}^* (Lines 16 - 18). They satisfy the specific bounds explored in Lemma D.3.

1192 Using this procedure, we present a description of the SOS (Search of the Optimal Stepsize) SIGN-
 1193 SGD (Algorithm 5). Before we pass to the convergence rate, we discuss the number of iterations
 1194 required by Algorithm 4. Since we calculate the average geometric at each iteration, we need
 1195 $\log \log \frac{\gamma_{\text{hi}}}{\gamma_{\text{lo}}}$ steps, where γ_{lo} and γ_{hi} are the boundaries of the initial segment. Thus, according to
 1196

1197 Algorithm 5, it requires $\log \log \frac{2^{2^k} \gamma_s}{\gamma_s} = k$ iterations. We establish a lower bound on k by requiring
 1198 that the initial γ_{hi} is greater than $\phi(\gamma_{\text{hi}})$. According to Lemma D.2, γ_{hi} should be at least $\frac{\Delta^*}{\|\nabla f(x^0)\|_1}$.
 1199 In this way, $k = \log \log \frac{\Delta^*}{\gamma_s \|\nabla f(x^0)\|_1}$. Therefore, allowing Algorithm 5 to perform T iterations, the
 1200 total number of iterations (considering Algorithm 4 performance time) is $T \log \log \frac{\Delta^*}{\gamma_s \|\nabla f(x^0)\|_1}$. We
 1201 regard this additional double-logarithmic factor as negligible, as it aligns with the results in (Carmon
 1202 & Hinder, 2022). We now present the main theoretical result of this section.

1204 **Theorem B.1.** *Suppose Assumptions 3.1, 3.2, 3.3, 3.4 hold. Then for Algorithm 5 after obtaining the
 1205 stepsize γ_0 the following estimate is valid:*

$$1206 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{3 \|\nabla f(x^0)\|_1}{T}.$$

1210 Moreover, taking into account the complexity of Algorithm 4 in relation to the initial stepsize bound
 1211 γ_s , to reach ε -accuracy, where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, Algorithm 5 needs
 1212

$$1213 \quad \tilde{\mathcal{O}}\left(\frac{\Delta^* L_\infty}{\varepsilon^2}\right) \text{ iterations.}$$

1216 **Discussion of the results.** We obtain the near-optimal convergence rate 1. Our method retains a
 1217 dependency on the initial approximation. Indeed, we should take γ_s to be less than $\frac{\Delta^*}{L_\infty T}$, according to
 1218 **Step 2** in the proof of Theorem B.1 (Theorem E.2). An analogous requirement was established in the
 1219 work (Carmon & Hinder, 2022) and we do not consider this to be an issue. Nevertheless, despite the
 1220 theoretical optimality of the proposed approach, its practical application is not promising. Launching
 1221 multiple training sessions on large models does not appear effective. In this context, Algorithm 2
 1222 remains our main contribution. While proposing Algorithm 5, we demonstrate how to obtain the
 1223 near-optimal rate for SIGN-SGD in parameter-free optimization.

1225 **Stochastic gradients and distributed settings.** The description and analysis of Algorithm 5 in
 1226 stochastic and distributed setups can be found in Appendix E.2, E.3.

1228 B.1 SOS SIGN-SGD EXPERIMENTS

1229 B.1.1 LOGISTIC REGRESSION.

1232 We present toy experiments on logistic regression. We provide a comparison of SIGN-SGD with the
 1233 theoretical stepsize $\frac{1}{\sqrt{T}}$ (Algorithm 1), SOS SIGN-SGD (Algorithm 5), ALIAS (Algorithm 2) and
 1234 STEEPEST DESCENT (Algorithms 8, 9). We validate the criteria $\|\nabla f(x^t)\|_1$ on four datasets sourced
 1235 from the LIBSVM library (Chang & Lin, 2011): a9a, w8a, ijcnn1 and skin-nonskin. The
 1236 results are presented in Figure 6.

1237 The plots show that even on the convex problems, SOS SIGN-SGD performs worse than ALIAS.
 1238 This was expected, however, testing this method on a real non-convex problem, such as training LLMs,
 1239 lacks justification. Additionally, it is noteworthy that STEEPEST DESCENT performs worse compared
 1240 to SIGN-SGD, highlighting the limited practical applicability of this approach. Consequently, we
 1241 provide analysis for STEEPEST DESCENT only with incorporated Algorithm 4 in Appendix G. We do
 not focus on the analysis and development of efficient parameter-free methods based on this approach.

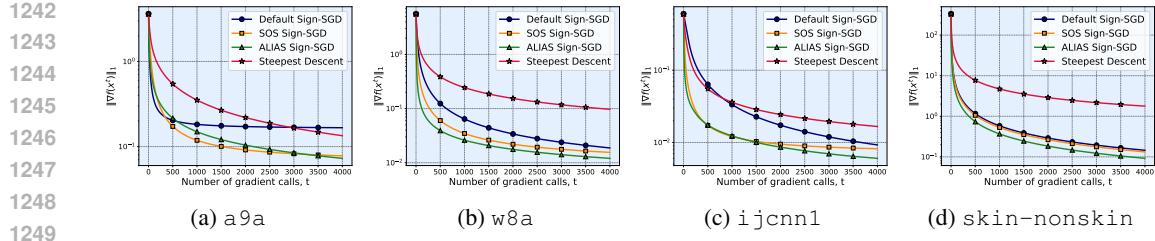


Figure 6: SIGN-SGD methods on logistic regression.

B.1.2 NON-CONVEX PROBLEM

We provide the comparison of SIGN-SGD with theoretical stepsize $\frac{1}{\sqrt{T}}$ (Algorithm 1), SOS SIGN-SGD (Algorithm 5), ALIAS (Algorithm 2) and STEEPEST DESCENT (Algorithms 8, 9). We validate criteria $\|\nabla f(x^t)\|_1$ on four datasets, sourced from the LIBSVM library (Chang & Lin, 2011): a9a, w8a, ijcnn1 and skin-nonskin. In the main part we presented the results for the convex problem. Now we consider the non-convex objective, namely the non-linear least squares loss:

$$f(x) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \frac{1}{1 + \exp(-a_i^T x)} \right)^2. \quad (5)$$

There we denote $a_i \in \mathbb{R}^{1 \times d}$ as the sample and $y_i \in \{0, 1\}$ as the target. The results are presented in Figure 7.

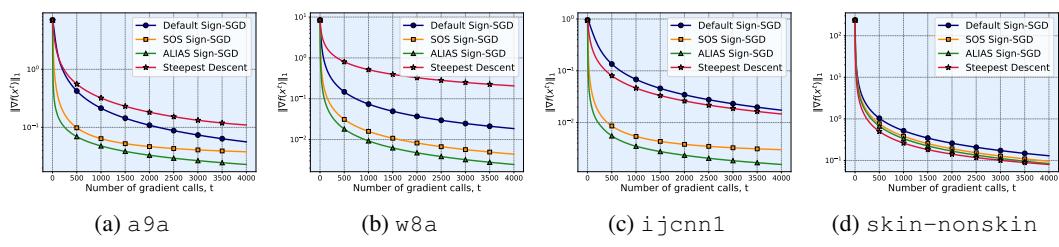


Figure 7: Comparison of SIGN-SGD methods on problem equation 5.

The plots demonstrate the superiority of our methods, SOS SIGN-SGD and ALIAS, over classical SIGN-SGD with the vanilla step size choice $\frac{1}{\sqrt{T}}$. This highlights the importance of adapting the stepsize to the problem structure and hyperparameters. However, SOS SIGN-SGD still underperforms compared to ALIAS.

C ADDITIONAL NOTATION AND GENERAL INEQUALITIES

Notation. Here we present the full list of notation, used in our paper.

- We denote d as the dimension of the problem; T as the total number of iterations in the algorithms; x^{-1} as the starting point in the SOS SIGN-SGD algorithm, x^0 as the starting point in the ALIAS algorithm; x^t as the point at t -th iteration in the algorithms; x^* as the optimal solution of the problem; $\tilde{\Delta}_T = f(x^{-1}) - \min_{-1 \leq t \leq T} f(x^t)$; $\Delta^* = f(x^{-1}) - f(x^*)$ for the SOS SIGN-SGD method, $\Delta^* = f(x^0) - f(x^*)$ for the ALIAS method.
- We denote $\nabla f(x^t)$ as the honest full gradient of the objective function at the point x^t ; g^t (or $g_{\xi^t}^t$) as the stochastic gradient of the objective function at the point x^t , according to the stochastic realization ξ^t (we add lower index only when we use different stochastic realizations in the method); g_j^t (or g_{j,ξ^t}^t) as the stochastic gradient of the objective function at the point x^t on j -th device in the distributed setup, according to the stochastic realization ξ^t .
- For vectors $x, y \in \mathbb{R}^d$ we denote $\text{sign}(x)$ as the vector of the dimension d , where the i -th coordinate

1296 defines as
 1297

1298 $[\text{sign}(x)]_i = \text{sign}(x_i) = \begin{cases} 1, & \text{if } x_i > 0 \\ 0, & \text{if } x_i = 0 \\ -1, & \text{if } x_i < 0 \end{cases}$
 1299
 1300
 1301

1302 $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$ is the scalar product; $\|x\|_1 = \sum_{i=1}^d |x_i|$ is l_1 -norm; $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$ is l_2 -norm;
 1303
 1304 $\|x\|_\infty = \max_{i \in \overline{1, d}} |x_i|$ is l_∞ -norm.
 1305

1306 • For a random vector $\xi \in \mathbb{R}^d$ and fixed vector $\psi \in \mathbb{R}^d$ we denote $\mathbb{E}[\xi]$ is the expected value with
 1307 respect to a random vector ξ and $\mathbb{E}[\xi|\psi]$ as the expected value with the respect to a random vector ξ ,
 1308 conditioned on the fixed vector ψ .

1309
 1310
 1311 **General inequalities.** Suppose $x, y \in \mathbb{R}^d$, $a, b \in \mathbb{R}$, $f(\cdot)$ is under Assumption 3.1 and $\xi, \psi \in \mathbb{R}_+$
 1312 are random variables. Then,

1313 $\|\nabla f(x) - \nabla f(y)\|_1 \leq L_\infty \|x - y\|_\infty$ (Lip)
 1314
 1315 $\|x + y\|_1 \leq \|x\|_1 + \|y\|_1 \text{ or } \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ (CS)
 1316
 1317 $\langle x, y \rangle \leq \|x\|_1 \|y\|_\infty$ (Conj)
 1318
 1319 $\mathbb{E}[\xi\psi] \leq (\mathbb{E}[\xi]^p)^{\frac{1}{p}} (\mathbb{E}[\psi]^q)^{\frac{1}{q}}, \text{ where } \frac{1}{p} + \frac{1}{q} = 1$ (HöL)

1320 D LEMMAS FOR SOS SIGN-SGD

1321
 1322 **Lemma D.1** (Quadratic inequality). Let $x \in \mathbb{R}_+$ be a variable and $u, v \in \mathbb{R}_+$ be constants. Then
 1323 $x^2 - ux - v \leq 0$ implies $x \leq u + \sqrt{v}$. Additionally, $x^2 + ux - v \leq 0$ implies $x \leq \sqrt{v}$.
 1324

1325 *Proof.* Since u, v are non-negative constants, the plain algebra involves $x_{\text{s.p.}} = \frac{u \pm \sqrt{u^2 + 4v}}{2}$ being
 1326 stationary points of $x^2 - 2ux - v \leq 0$ inequality. Since x is the positive variable, the boundary
 1327 $x \leq x_{\text{s.p.}+}$ is the appropriate area of the solution. It remains for us to say that
 1328

1329 $x \leq \frac{1}{2}u + \frac{1}{2}\sqrt{u^2 + 4v} \stackrel{CS}{\leq} u + \sqrt{v},$
 1330
 1331

1332 which finishes the proof of the first statement. Proceeding analogically for the second part, we obtain
 1333 $x \leq -\frac{1}{2}u + \frac{1}{2}\sqrt{u^2 + 4v} \leq -\frac{1}{2}u + \frac{1}{2}u + \sqrt{v} = \sqrt{v}$. \square

1334
 1335 **Lemma D.2** (Bisection entry). Let $\gamma_{\max} = \frac{\Delta^*}{\|\text{Grad}(f(x^0))\|_1}$ (or $\gamma_{\max} = \frac{\Delta^*}{\frac{1}{M} \sum_{j=1}^M \|\text{Grad}(f(x^0))\|_1}$ for dis-
 1336
 1337 tributed setting), where $\Delta^* = f(x^{-1}) - f(x^*)$ and the gradient oracle $\text{Grad}(f(\cdot))$ can be specified
 1338 as $\nabla f(\cdot)$ or $g(\cdot)$ or $g_j(\cdot)$, that depends on the algorithm setting (exact gradient, stochastic gradient
 1339 or gradient on the i -th node in distributed setting). Then we can always entry the bisection procedure
 1340 without infinite early terminations taking $\gamma_{hi} \geq \gamma_{\max}$.
 1341

1342
 1343 *Proof.* We can entry the BISECTION procedure, when $\gamma_{hi} \geq \phi(\gamma_{hi})$. Thus, to proof the lemma
 1344 statement we can show that $\gamma_{hi} < \phi(\gamma_{hi})$ is impossible, when $\gamma_{hi} \geq \gamma_{\max} = \frac{\Delta^*}{\|\text{Grad}(f(x^0))\|_1}$. Using
 1345 $\tilde{\Delta}_T = f(x^{-1}) - \min_{-1 \leq t \leq T} f(x^t)$ notation, we consider

1346
 1347
 1348
 1349
$$\frac{\tilde{\Delta}_T(\gamma_{hi})}{\mathfrak{D}_T(\gamma_{hi})} = \frac{\mathfrak{N}_T(\gamma_{hi})}{\mathfrak{D}_T(\gamma_{hi})} = \phi(\gamma_{hi}) > \gamma_{hi} \geq \gamma_{\max} = \frac{\Delta^*}{\|\text{Grad}(f(x^0))\|_1}. \quad (6)$$

Let us look at the numerators of the fractions in the obtained inequality. According to Assumption 3.3, $f(x^*) \leq \min_{-1 \leq t \leq T} f(x^t)$. In that way,

$$\tilde{\Delta}_T(\gamma_{hi}) \leq \Delta^*. \quad (7)$$

Now we consider denominators in equation 6. $\mathfrak{D}_T(\gamma_{hi})$ has the following form in any setting: $\sum_{t=0}^{T-1} \|\text{Grad}(f(x^{t+1}(\gamma_{hi})) - \text{Grad}(f(x^t(\gamma_{hi})))\|_1 + \zeta(\gamma_{hi})$, where $\zeta(\gamma)$ is defined as the minimum of gradients norm over the training: $\zeta(\gamma) = \min_{0 \leq t \leq T} \|\text{Grad}(f(x^t(\gamma_{hi})))\|_1$. Using equation CS, we obtain

$$\begin{aligned} \|\text{Grad}(f(x^0))\|_1 &\stackrel{(i)}{\leq} \sum_{t=0}^{\bar{t}-1} \|\text{Grad}(f(x^{t+1}(\gamma_{hi})) - \text{Grad}(f(x^t(\gamma_{hi})))\|_1 + \|\text{Grad}(f(x^{\bar{t}}(\gamma_{hi})))\|_1 \\ &\leq \sum_{t=0}^{T-1} \|\text{Grad}(f(x^{t+1}(\gamma_{hi})) - \text{Grad}(f(x^t(\gamma_{hi})))\|_1 + \|\text{Grad}(f(x^{\bar{t}}(\gamma_{hi})))\|_1 \\ &\stackrel{(ii)}{=} \sum_{t=0}^{T-1} \|\text{Grad}(f(x^{t+1}(\gamma_{hi})) - \text{Grad}(f(x^t(\gamma_{hi})))\|_1 \\ &\quad + \min_{0 \leq t \leq T} \|\text{Grad}(f(x^t(\gamma_{hi})))\|_1 \\ &\stackrel{(iii)}{=} \sum_{t=0}^{T-1} \|\text{Grad}(f(x^{t+1}(\gamma_{hi})) - \text{Grad}(f(x^t(\gamma_{hi})))\|_1 + \zeta(\gamma_{hi}) \\ &= \mathfrak{D}_T(\gamma_{hi}), \end{aligned} \quad (8)$$

where inequality (i) holds for any $1 \leq \bar{t} \leq T$ and in (ii) we choose $\bar{t} = \arg \min_{0 \leq t \leq T} \|\text{Grad}(f(x^t(\gamma_{hi})))\|_1$. One can note that we omit the case when the norm of the oracle reaches its minimum at iteration $t = 0$ in ζ definition, when use it in (iii). However, it is a trivial case and it satisfies

$$\|\text{Grad}(f(x^0))\|_1 \leq \zeta(\gamma_{hi}) \leq \sum_{t=0}^{T-1} \|\text{Grad}(f(x^{t+1}(\gamma_{hi})) - \text{Grad}(f(x^t(\gamma_{hi})))\|_1 + \zeta(\gamma_{hi}) = \mathfrak{D}_T(\gamma_{hi}).$$

In that way, combining it with equation 8 and equation 7, we obtain

$$\frac{\tilde{\Delta}_T(\gamma_{hi})}{\mathfrak{D}_T(\gamma_{hi})} \leq \frac{\Delta^*}{\|\text{Grad}(f(x^0))\|_1},$$

which contradicts to equation 6. Thus, we can entry the Algorithm 4 without infinite early termination if take initial γ_{hi} at least $\frac{\Delta^*}{\|\text{Grad}(f(x^0))\|_1}$. Note that for the distributed case we can obtain

$$\frac{1}{M} \sum_{j=1}^M \|\text{Grad}(f(x^0))\|_1 \leq \mathfrak{D}_T(\gamma_{hi}) \text{ in the same way as in equation 8.} \quad \square$$

Lemma D.3 (Bisection invariants). *If The BISECTION procedure (Algorithm 4) has no early termination at all, it returns γ_0 such that*

$$\frac{\mathfrak{N}_T(\gamma_0)}{2\mathfrak{D}_T(\gamma_{hi}^*)} \leq \gamma_0 \leq \frac{\mathfrak{N}_T(\gamma_{lo}^*)}{\mathfrak{D}_T(\gamma_0)}, \quad (9)$$

where γ_{lo}^* and γ_{hi}^* are values, from which γ_0 is chosen in the end of Algorithm 4. Moreover,

$$\mathfrak{N}_T(\gamma_0) \leq \mathfrak{N}_T(\gamma_{lo}^*), \quad (10)$$

$$\mathfrak{D}_T(\gamma_0) \leq \mathfrak{D}_T(\gamma_{hi}^*). \quad (11)$$

Proof. Consider the case procedure returns $\gamma_0 = \gamma_{lo}^*$. Then

$$\frac{\mathfrak{N}_T(\gamma_{lo}^*)}{2\mathfrak{D}_T(\gamma_{hi}^*)} = \frac{\mathfrak{N}_T(\gamma_{lo}^*)}{2\mathfrak{N}_T(\gamma_{hi}^*)} \cdot \frac{\mathfrak{N}_T(\gamma_{hi}^*)}{\mathfrak{D}_T(\gamma_{hi}^*)} = \frac{\mathfrak{N}_T(\gamma_{lo}^*)}{2\mathfrak{N}_T(\gamma_{hi}^*)} \phi(\gamma_{hi}^*) \stackrel{(i)}{\leq} \frac{1}{2} \gamma_{hi}^* \stackrel{(ii)}{\leq} \gamma_{lo}^*$$

$$1404 \quad \leqslant \quad \phi(\gamma_{\text{lo}}^*) \stackrel{(iii)}{=} \frac{\mathfrak{N}_T(\gamma_{\text{lo}}^*)}{\mathfrak{D}_T(\gamma_{\text{lo}}^*)}, \quad (12)$$

$$1405$$

$$1406$$

1407 where (i) is correct due to the γ_{lo} return condition, (ii) – bisection stop condition, (iii) – bisection
1408 invariant. Consider the case when procedure returns $\gamma_0 = \gamma_{\text{hi}}^*$. Then

$$1409 \quad \frac{\mathfrak{N}_T(\gamma_{\text{hi}}^*)}{2\mathfrak{D}_T(\gamma_{\text{hi}}^*)} = \frac{1}{2}\phi(\gamma_{\text{hi}}^*) \stackrel{(i)}{\leqslant} \frac{1}{2}\gamma_{\text{hi}}^* \leqslant \gamma_{\text{hi}}^* \stackrel{(ii)}{\leqslant} \frac{\mathfrak{N}_T(\gamma_{\text{hi}}^*)}{\mathfrak{D}_T(\gamma_{\text{hi}}^*)}, \quad (13)$$

$$1410$$

$$1411$$

1412 where (i) is correct due to the bisection invariant and (ii) – the return condition. Combining
1413 equation 12 with equation 13, we obtain the first claim of the lemma whether Algorithm 4 returns
1414 $\gamma_0 = \gamma_{\text{lo}}^*$ or $\gamma_0 = \gamma_{\text{hi}}^*$. It remains to notice that equation 12 is followed by $\mathfrak{D}_T(\gamma_{\text{lo}}^*) \leqslant \mathfrak{D}_T(\gamma_{\text{hi}}^*)$ when
1415 $\gamma_0 = \gamma_{\text{lo}}^*$, and, consequently, $\mathfrak{D}_T(\gamma_0) \leqslant \mathfrak{D}_T(\gamma_{\text{hi}}^*)$ since $\mathfrak{D}_T(\gamma_{\text{hi}}^*) \leqslant \mathfrak{D}_T(\gamma_{\text{lo}}^*)$ is trivial. Analogically,
1416 equation 13 is followed by $\mathfrak{N}_T(\gamma_{\text{hi}}^*) \leqslant \mathfrak{N}_T(\gamma_{\text{lo}}^*)$ when $\gamma_0 = \gamma_{\text{hi}}^*$, and, consequently, $\mathfrak{N}_T(\gamma_0) \leqslant \mathfrak{N}_T(\gamma_{\text{lo}}^*)$. This finishes the proof. \square

$$1417$$

$$1418$$

$$1419$$

Lemma D.4 (Extra step). *Suppose Assumptions 3.1, 3.2, 3.3 hold. Then, considering update of the following form:*

$$1420 \quad f(x^0) = \min \{f(x^{-1} + \tau e), f(x^{-1} - \tau e)\},$$

$$1421$$

1422 where e is the random vector from the unit basis, and we can guarantee $f(x^0) < f(x^{-1})$, when
1423 $\tau < \frac{\|\nabla f(x^{-1})\|_1}{L_\infty}$. Moreover, Algorithm 4, starting with $\tau = \tau_s$ and performing $\tau = \frac{\tau}{2}$, needs at least
1424 $\log \left(\frac{\tau_s L_\infty}{\|\nabla f(x^{-1})\|_1} \right)$ extra iterations to find efficient value of τ .

$$1425$$

$$1426$$

$$1427$$

Proof. We choose $f(x^0) = \min \{f(x^{-1} + \tau e), f(x^{-1} - \tau e)\}$. We use convexity to show

$$1428 \quad f(x^{-1} + \tau e) \leqslant f(x^{-1}) + \langle \nabla f(x^{-1} + \tau e), \tau e \rangle$$

$$1429 \quad = f(x^{-1}) + \tau \langle \nabla f(x^{-1}), e \rangle + \tau \langle \nabla f(x^{-1} + \tau e) - \nabla f(x^{-1}), e \rangle$$

$$1430 \quad \stackrel{\text{Conj}}{\leqslant} f(x^{-1}) + \tau \langle \nabla f(x^{-1}), e \rangle + \tau \|\nabla f(x^{-1} + \tau e) - \nabla f(x^{-1})\|_1 \|e\|_\infty$$

$$1431 \quad \stackrel{\text{Lip}}{\leqslant} f(x^{-1}) + \tau \langle \nabla f(x^{-1}), e \rangle + \tau^2 L_\infty \|e\|_\infty^2,$$

$$1432 \quad f(x^{-1} - \tau e) \leqslant f(x^{-1}) - \langle \nabla f(x^{-1} - \tau e), \tau e \rangle$$

$$1433 \quad = f(x^{-1}) - \tau \langle \nabla f(x^{-1}), e \rangle - \tau \langle \nabla f(x^{-1} - \tau e) - \nabla f(x^{-1}), e \rangle$$

$$1434 \quad \stackrel{\text{Conj}}{\leqslant} f(x^{-1}) - \tau \langle \nabla f(x^{-1}), e \rangle + \tau \|\nabla f(x^{-1} - \tau e) - \nabla f(x^{-1})\|_1 \|e\|_\infty$$

$$1435 \quad \stackrel{\text{Lip}}{\leqslant} f(x^{-1}) - \tau \langle \nabla f(x^{-1}), e \rangle + \tau^2 L_\infty \|e\|_\infty^2.$$

$$1436$$

$$1437$$

$$1438$$

$$1439$$

$$1440$$

Utilizing $e = \text{sign}(\nabla f(x^{-1}))$, we take expectation and obtain

$$1441 \quad f(x^0) \leqslant f(x^{-1}) - \tau |\langle \nabla f(x^{-1}), e \rangle| + \tau^2 L_\infty \|e\|_\infty^2$$

$$1442 \quad = f(x^{-1}) - \tau \left| \sum_{i=1}^d [\nabla f(x^{-1})]_i \right| + \tau^2 L_\infty \|\text{sign}(\nabla f(x^{-1}))\|_\infty^2$$

$$1443 \quad \leqslant f(x^{-1}) - \tau \|\nabla f(x^{-1})\|_1 + \tau^2 L_\infty$$

$$1444 \quad = f(x^{-1}) - \tau (\|\nabla f(x^{-1})\|_1 - \tau L_\infty).$$

$$1445$$

$$1446$$

$$1447$$

$$1448$$

$$1449$$

1450 In that way, if we have $\tau < \frac{\|\nabla f(x^{-1})\|_1}{L_\infty}$, we derive

$$1451$$

$$1452 \quad f(x^0) < f(x^{-1}).$$

$$1453$$

1454 Since in the algorithm we start with $\tau = \tau_s$ and divide it by 2, after l divisions, we have

$$1455 \quad \frac{\tau_s}{2^l} < \frac{\|\nabla f(x^{-1})\|_1}{L_\infty}.$$

$$1456$$

$$1457$$

Thus, we need at least $l = \log \left(\frac{\tau_s L_\infty}{\|\nabla f(x^{-1})\|_1} \right)$ iterations. \square

1458 **E MAIN PROOFS AND DETAILS FOR SOS SIGN-SGD**
 1459

1460 **E.1 EXACT GRADIENT SETTING**
 1461

1462 **Lemma E.1** (Descent lemma). *For Algorithm 5 under Assumptions 3.1, 3.2, 3.3, 3.4, the following
 1463 estimate is valid:*

$$1464 \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{f(x^{-1}) - f(x^T)}{\gamma_0} + \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1.$$

1468 *Proof.* Starting from the convexity of the objective,

$$1469 \begin{aligned} f(x^{t+1}) - f(x^t) &\leq \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle = -\gamma^t \langle \nabla f(x^{t+1}), \text{sign}(\nabla f(x^t)) \rangle \\ 1470 &= -\gamma^t \langle \nabla f(x^t), \text{sign}(\nabla f(x^t)) \rangle \\ 1471 &\quad -\gamma^t \langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign}(\nabla f(x^t)) \rangle \\ 1472 &\stackrel{\text{Conj}}{\leq} -\gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \|\text{sign}(\nabla f(x^t))\|_\infty \\ 1473 &\leq -\gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1. \end{aligned}$$

1477 Now we express the gradient norm and sum over all iterations to obtain

$$1478 \begin{aligned} \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 &\leq \sum_{t=0}^{T-1} [f(x^t) - f(x^{t+1})] + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \\ 1479 &= f(x^0) - f(x^T) + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1. \end{aligned}$$

1484 Using Lemma D.4 to consider the extra step, we get

$$1486 \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 \leq f(x^{-1}) - f(x^T) + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1.$$

1489 Since Algorithm 5 performs all the steps with the constant rate γ_0 which we define later, we can
 1490 rewrite the result in the following form:

$$1492 \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{f(x^{-1}) - f(x^T)}{\gamma_0} + \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1,$$

1495 which ends the proof of the lemma. \square

1496 **Theorem E.2 (Theorem B.1).** *Suppose Assumptions 3.1, 3.2, 3.3, 3.4 hold. Then for Algorithm 5
 1497 after obtaining the stepsize γ_0 , the following estimate is valid:*

$$1499 \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{3 \|\nabla f(x^0)\|_1}{T}.$$

1502 Moreover, taking into account the complexity of Algorithm 4 in relation to the initial stepsize bound

1503 γ_s , to reach ε -accuracy, where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, Algorithm 5 needs

$$1506 \mathcal{O} \left(\frac{\Delta^* L_\infty}{\varepsilon^2} \log \log \frac{\Delta^*}{\gamma_s \|\nabla f(x^0)\|_1} \right) \text{ iterations.}$$

1508 *Proof.* Let us start with the result of Lemma E.1:

$$1511 \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{f(x^{-1}) - f(x^T)}{\gamma_0} + \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1$$

$$1512 \quad \leq \quad \frac{\tilde{\Delta}_T}{\gamma_0} + \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1, \quad (14)$$

1515 where $\tilde{\Delta}_T = f(x^{-1}) - \min_{-1 \leq t \leq T} f(x^t)$. Now, we accurately estimate the last term in equation 14,

1517 which is additionally denoted as $F_T = \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1$. Thus,

$$1520 \quad F_T = \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \stackrel{Lip}{\leq} L_\infty \sum_{t=0}^{T-1} \|x^{t+1} - x^t\|_\infty \\ 1521 \\ 1523 \quad = L_\infty \sum_{t=0}^{T-1} \gamma^t \|\text{sign}(\nabla f(x^t))\|_\infty \leq L_\infty \sum_{t=0}^{T-1} \gamma^t. \quad (15)$$

1526 Now let us choose $\phi(\gamma)$, which we push into the BISECTION procedure (Algorithm 4): $\phi(\gamma) =$
1527 $\frac{\mathfrak{N}(\gamma)}{\mathfrak{D}(\gamma)} = \frac{\tilde{\Delta}_T(\gamma)}{F_T(\gamma) + \zeta(\gamma)}$, where $\tilde{\Delta}_T = f(x^{-1}) - \min_{-1 \leq t \leq T} f(x^t)$ and $\zeta = \min_{0 \leq t \leq T} \|\nabla f(x^t)\|_1$. In that way,
1528 we obtain some γ_0 , which can be equal to γ_{lo}^* or γ_{hi}^* (see Lemma D.2, Lemma D.3) and use it as a
1529 constant stepsize for our method. Thus, equation 15 transforms into

$$1531 \quad F_T(\gamma_0) \leq \gamma_0 L_\infty T. \quad (16)$$

1532 Mention that, according to Lemma D.2, we can always entry to the procedure without infinite early
1533 termination. In that way, we have two situations: when we have no early terminations at all and we
1534 are under the activity of Lemma D.3, and when we have early termination with initial γ_{lo}^* . We divide
1535 the following proof into two steps, where we separately show the convergence guarantees in this two
1536 situations.

1537 Step 1: no early terminations.

1538 Since we have only two cases: $\gamma_0 = \gamma_{\text{lo}}^*$ or $\gamma_0 = \gamma_{\text{hi}}^*$, let us consider them separately.

- 1539 • $\gamma_0 = \gamma_{\text{hi}}^*$: equation 16 transforms into

$$1540 \quad F_T(\gamma_{\text{hi}}^*) \leq \gamma_{\text{hi}}^* L_\infty T \stackrel{\text{Lemma D.3,9}}{\leq} \frac{\mathfrak{N}_T(\gamma_{\text{lo}}^*)}{\mathfrak{D}_T(\gamma_{\text{hi}}^*)} L_\infty T \stackrel{(i)}{=} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{F_T(\gamma_{\text{hi}}^*) + \zeta(\gamma_{\text{hi}}^*)} L_\infty T,$$

1543 where (i) is correct due to the $\phi(\gamma)$ choice. Solving this quadratic inequality with respect to $F_T(\gamma_{\text{hi}}^*)$
1544 (Lemma D.1), we obtain

$$1545 \quad F_T(\gamma_{\text{hi}}^*) \leq \sqrt{\tilde{\Delta}_T(\gamma_{\text{lo}}^*) L_\infty T} \leq \sqrt{\Delta^* L_\infty T}, \quad (17)$$

1548 where $\Delta^* = f(x^{-1}) - f(x^*)$. Plugging it into equation 14, we obtain

$$1549 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{hi}}^*)}{\gamma_{\text{hi}}^*} + \frac{1}{T} F_T(\gamma_{\text{hi}}^*) \\ 1550 \\ 1552 \quad \stackrel{\text{Lemma D.3,9}}{\leq} \frac{1}{T} \frac{2\mathfrak{D}_T(\gamma_{\text{hi}}^*)}{\mathfrak{N}_T(\gamma_{\text{hi}}^*)} \tilde{\Delta}_T(\gamma_{\text{hi}}^*) + \frac{1}{T} F_T(\gamma_{\text{hi}}^*) \\ 1553 \\ 1555 \quad = \frac{2 [F_T(\gamma_{\text{hi}}^*) + \zeta(\gamma_{\text{hi}}^*)] \tilde{\Delta}_T(\gamma_{\text{hi}}^*)}{\tilde{\Delta}_T(\gamma_{\text{hi}}^*)} + \frac{1}{T} F_T(\gamma_{\text{hi}}^*) \\ 1556 \\ 1558 \quad = \frac{3}{T} F_T(\gamma_{\text{hi}}^*) + \frac{2\zeta(\gamma_{\text{hi}}^*)}{T} \\ 1559 \\ 1560 \quad \stackrel{17}{\leq} \frac{3\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{2\|\nabla f(x^0)\|_1}{T}. \quad (18)$$

1561 In that way, equation 18 is the final estimate when BISECTION procedure returns γ_{hi}^* .

- 1563 • $\gamma_0 = \gamma_{\text{lo}}^*$: equation 16 transforms into

$$1564 \quad F_T(\gamma_{\text{lo}}^*) \leq \gamma_{\text{lo}}^* L_\infty T \stackrel{\text{Lemma D.3,9}}{\leq} \frac{\mathfrak{N}_T(\gamma_{\text{lo}}^*)}{\mathfrak{D}_T(\gamma_{\text{lo}}^*)} L_\infty T \stackrel{(i)}{=} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{F_T(\gamma_{\text{lo}}^*) + \zeta(\gamma_{\text{lo}}^*)} L_\infty T,$$

1566 where (i) is correct due to the $\phi(\gamma)$ choice. Solving this quadratic inequality with respect to $F_T(\gamma_{\text{lo}}^*)$
1567 (Lemma D.1), we obtain

$$1569 \quad F_T(\gamma_{\text{lo}}^*) \leq \sqrt{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)L_\infty T} \leq \sqrt{\Delta^*L_\infty T}. \quad (19)$$

1570 Now we make an additional distinction and consider the estimates separately: one case when
1571 $\gamma_{\text{lo}}^* > \sqrt{\frac{\Delta^*}{L_\infty T}}$, and another when $\gamma_{\text{lo}}^* \leq \sqrt{\frac{\Delta^*}{L_\infty T}}$. We can do this without any limitations, since
1573 combining the intervals considered for γ_{lo}^* returns all possible values.

1574 $\circ \quad \gamma_{\text{lo}}^* > \sqrt{\frac{\Delta^*}{L_\infty T}}$: we straightforwardly move to the equation 14 estimation:

$$\begin{aligned} 1576 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 &\leq \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\gamma_{\text{lo}}^*} + \frac{1}{T} F_T(\gamma_{\text{lo}}^*) \\ 1577 &\leq \frac{\sqrt{L_\infty}}{\sqrt{\Delta^*T}} \tilde{\Delta}_T(\gamma_{\text{lo}}^*) + \frac{1}{T} F_T(\gamma_{\text{lo}}^*) \\ 1578 &\stackrel{19}{\leq} \frac{\sqrt{\Delta^*L_\infty}}{\sqrt{T}} + \frac{\sqrt{\Delta^*L_\infty}}{\sqrt{T}} = 2 \frac{\sqrt{\Delta^*L_\infty}}{\sqrt{T}}. \end{aligned} \quad (20)$$

1584 $\circ \quad \gamma_{\text{lo}}^* \leq \sqrt{\frac{\Delta^*}{L_\infty T}}$: in this case, we start from the estimate that is followed by equation 16:

$$1586 \quad F_T(\gamma_{\text{hi}}^*) \leq \gamma_{\text{hi}}^* L_\infty T \stackrel{(i)}{\leq} 2\gamma_{\text{lo}}^* L_\infty T \leq 2\sqrt{L_\infty \Delta^* T}, \quad (21)$$

1588 where (i) is done due to the bisection stop condition. Now we proceed with estimation of equation 14:

$$\begin{aligned} 1590 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 &\leq \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\gamma_{\text{lo}}^*} + \frac{1}{T} F_T(\gamma_{\text{lo}}^*) \\ 1591 &\stackrel{\text{Lemma D.3,9}}{\leq} \frac{1}{T} \frac{2\mathfrak{D}_T(\gamma_{\text{hi}}^*)}{\mathfrak{N}_T(\gamma_{\text{lo}}^*)} \tilde{\Delta}_T(\gamma_{\text{lo}}^*) + \frac{1}{T} F_T(\gamma_{\text{lo}}^*) \\ 1592 &\stackrel{\text{Lemma D.3,11}}{=} \frac{2}{T} \frac{[F_T(\gamma_{\text{hi}}^*) + \zeta(\gamma_{\text{hi}}^*)] \tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)} + \frac{F_T(\gamma_{\text{hi}}^*) + \zeta(\gamma_{\text{hi}}^*)}{T} \\ 1593 &= \frac{3F_T(\gamma_{\text{hi}}^*)}{T} + \frac{3\zeta(\gamma_{\text{hi}}^*)}{T} \\ 1594 &\stackrel{21}{\leq} 6 \frac{\sqrt{\Delta^*L_\infty}}{\sqrt{T}} + \frac{3\zeta(\gamma_{\text{hi}}^*)}{T} \\ 1595 &\leq 6 \frac{\sqrt{\Delta^*L_\infty}}{\sqrt{T}} + \frac{3\|\nabla f(x^0)\|_1}{T}. \end{aligned} \quad (22)$$

1604 Combining equation 20 and equation 22, we get that equation 22 is the final estimate when BISECTION
1605 procedure returns γ_{lo}^* .

1606 In the end, equation 18 and equation 22 give us the estimate in the case when BISECTION procedure
1607 does not have early terminations at all and outputs any value of γ_0 :

$$1609 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq 6 \frac{\sqrt{\Delta^*L_\infty}}{\sqrt{T}} + \frac{3\|\nabla f(x^0)\|_1}{T}. \quad (23)$$

1612 Step 2: early termination with γ_{lo} .

1613 Now we consider the scenario when with initial γ_{lo} , there is $\gamma_{\text{lo}} \geq \phi(\gamma_{\text{lo}})$ and algorithm early returns
1614 γ_{lo}^* . To dissect this, we should choose an initial $\gamma_{\text{lo}} = \gamma_{\text{lo}}^* \leq \frac{\Delta^*}{L_\infty T}$. Thus, equation 16 transforms into

$$1615 \quad F_T(\gamma_{\text{lo}}^*) \leq \gamma_{\text{lo}} L_\infty T \leq \sqrt{L_\infty \Delta^* T}. \quad (24)$$

1616 In that way, equation 14 turns into

$$1618 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\gamma_{\text{lo}}^*} + \frac{1}{T} F_T(\gamma_{\text{lo}}^*)$$

$$\begin{aligned}
1620 &\leq \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\phi(\gamma_{\text{lo}}^*)} + \frac{1}{T} F_T(\gamma_{\text{lo}}^*) \\
1621 &= \frac{1}{T} \frac{[F_T(\gamma_{\text{lo}}^*) + \zeta(\gamma_{\text{lo}}^*)] \tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)} + \frac{1}{T} F_T(\gamma_{\text{lo}}^*) \\
1622 &= \frac{2F_T(\gamma_{\text{lo}}^*)}{T} + \frac{\zeta(\gamma_{\text{lo}}^*)}{T} \stackrel{24}{\leq} 2 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{\|\nabla f(x^0)\|_1}{T}. \quad (25) \\
1623 \\
1624 \\
1625 \\
1626 \\
1627
\end{aligned}$$

Hence, equation 25 delivers the estimate, when Algorithm 4 makes an early termination.

Combining equation 23 with equation 25, we finally obtain the estimate for all possible cases of the BISECTION procedure return:

$$1631 \quad 1632 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t(\gamma_0))\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{3 \|\nabla f(x^0)\|_1}{T}.$$

Expressing the number of iterations and using $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$ as a criterion, we obtain that

algorithm needs $\mathcal{O}\left(\frac{\Delta^* L_\infty}{\varepsilon^2}\right)$ iterations to reach ε -accuracy. Note that we drop the term $\frac{\|\nabla f(x^0)\|_1}{T}$, since it is asymptotically smaller than the main one. However, we firstly need to find the step γ_0 with the bisection procedure which takes $T \log \log \left(\frac{\gamma_\varepsilon 2^{2^k}}{\gamma_\varepsilon}\right) = \mathcal{O}\left(\frac{\Delta^* L_\infty}{\varepsilon^2} k\right)$ iterations, where 2^k denotes the length of the initial interval for the stepsize. We have already discussed in the main part that, according to Lemma D.2, k should be at least $k = \log \log \frac{\Delta^*}{\gamma_s \|\nabla f(x^0)\|_1}$. Thus, $\mathcal{O}\left(\frac{\Delta^* L_\infty}{\varepsilon^2} \log \log \frac{\Delta^*}{\gamma_s \|\nabla f(x^0)\|_1}\right)$ is the final iteration complexity. \square

E.2 STOCHASTIC GRADIENT SETTING

Let us start with the description of the stepsize choice for stochastic version of Algorithm 5. The main purpose of the BISECTION procedure (Algorithm 4) is to find stepsize γ close enough to the $\phi(\gamma)$ desired value utilizing small number of sign descent launches. Recall we establish

$$1651 \quad \phi(\gamma) = \frac{\tilde{\Delta}_T(\gamma)}{\sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 + \zeta(\gamma)}$$

for the exact gradient case. The numerator can remain unchanged. However, since we lack access to exact gradients, we cannot use the original denominator. Instead, we employ stochastic oracles: $\mathfrak{D}_T(\gamma) = \sum_{t=0}^{T-1} \|g(x^{t+1}) - g(x^t)\|_1 + \zeta(\gamma)$. Other details remain the same, and we can straightforwardly pass to the convergence results.

Lemma E.3 (Descent lemma). *For Algorithm 5 under Assumptions 3.1, 3.2, 3.3, 3.7, the following estimate is valid:*

$$1661 \quad \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{f(x^{-1}) - f(x^T)}{\gamma_0} + \sum_{t=0}^{T-1} \|g^{t+1} - g^t\|_1 + 3\delta^t + \delta^{t+1},$$

$$1664 \quad \text{where } \delta^t = \sum_{t=0}^{T-1} \|\nabla f(x^t) - g^t\|_1.$$

Proof. Starting from the convexity of the objective,

$$\begin{aligned}
1668 \quad f(x^{t+1}) - f(x^t) &\leq \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle = -\gamma^t \langle \nabla f(x^{t+1}), \text{sign}(g^t) \rangle \\
1669 &= -\gamma^t \langle g^t, \text{sign}(g^t) \rangle - \gamma^t \langle \nabla f(x^{t+1}) - g^t, \text{sign}(g^t) \rangle \\
1670 &= -\gamma^t \|g^t\|_1 - \gamma^t \langle \nabla f(x^t) - g^t, \text{sign}(g^t) \rangle \\
1671 &\quad - \gamma^t \langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign}(g^t) \rangle \\
1672 &\stackrel{\text{Conj}}{\leq} -\gamma^t \|\nabla f(x^t)\|_1
\end{aligned}$$

$$\begin{aligned}
& + \gamma^t \|\nabla f(x^t) - g^t\|_1 + \gamma^t \|\nabla f(x^t) - g^t\|_1 \|\text{sign}(g^t)\|_\infty \\
& + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \|\text{sign}(g^t)\|_\infty \\
\leq & -\gamma^t \|\nabla f(x^t)\|_1 + 3\gamma^t \|\nabla f(x^t) - g^t\|_1 + \gamma^t \|\nabla f(x^{t+1}) - g^{t+1}\|_1 \\
& + \gamma^t \|g^{t+1} - g^t\|_1.
\end{aligned}$$

Now we rearrange terms and summarize over all iterations to obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 & \leq \sum_{t=0}^{T-1} [f(x^t) - f(x^{t+1})] + \sum_{t=0}^{T-1} \gamma^t \|g^{t+1} - g^t\|_1 \\
& + 3 \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t) - g^t\|_1 + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^{t+1}) - g^{t+1}\|_1.
\end{aligned}$$

Since Algorithm 5 performs all the steps with the constant rate γ_0 , which we define later, we can rewrite the result in the following form:

$$\begin{aligned}
\sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 & \leq \sum_{t=0}^{T-1} \frac{[f(x^t) - f(x^{t+1})]}{\gamma_0} + \sum_{t=0}^{T-1} \|g^{t+1} - g^t\|_1 \\
& + 3 \sum_{t=0}^{T-1} \|\nabla f(x^t) - g^t\|_1 + \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - g^{t+1}\|_1.
\end{aligned}$$

In the obtained estimate the last two terms consist from differences between the honest and stochastic gradient at the t -th and $(t+1)$ -th moments. One of our goals is to estimate them, however, we want perform analogically to Theorem E.2 and continue the proof with the $\sum_{t=0}^{T-1} \|g^{t+1} - g^t\|_1$ term estimate. In order to simplify our following writing we give additional notation and denote $\delta^t = \sum_{t=0}^{T-1} \|\nabla f(x^t) - g^t\|_1$. In that way, additionally considering the extra step (Lemma D.4), we derive

$$\sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{f(x^{-1}) - f(x^T)}{\gamma_0} + \sum_{t=0}^{T-1} \|g^{t+1} - g^t\|_1 + 3\delta^t + \delta^{t+1},$$

which ends the proof of the lemma. \square

Theorem E.4. *Suppose Assumptions 3.1, 3.2, 3.3, 3.7 hold. Then for Algorithm 5 using at t -th iteration mini-batches of sizes $t+1$, after obtaining the stepsize γ_0 , the following estimate is valid:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(x^t)\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + 10\|\sigma\|_1 + \frac{3\mathbb{E} \|g^0\|_1}{T}.$$

Moreover, taking into account the complexity of Algorithm 4 in relation to the initial stepsize bound γ_s , to reach ε -accuracy, where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, Algorithm 5 needs

$$\mathcal{O} \left(\left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \|\sigma\|_1^2 \right) \log \log \frac{\Delta^*}{\gamma_s \|g^0\|_1} \right) \text{ iterations.}$$

Proof. Let us start with the result of Lemma E.3. We transform it due to the fact that Algorithm 5 performs all the steps with the constant rate γ_0 , which we define later:

$$\begin{aligned}
\sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 & \leq \frac{f(x^{-1}) - f(x^T)}{\gamma_0} + \sum_{t=0}^{T-1} \|g^{t+1} - g^t\|_1 + 3\delta^t + \delta^{t+1} \\
& \leq \frac{\tilde{\Delta}_T}{\gamma_0} + \sum_{t=0}^{T-1} \|g^{t+1} - g^t\|_1 + 3\delta^t + \delta^{t+1},
\end{aligned} \tag{26}$$

1728 where $\tilde{\Delta}_T = f(x^{-1}) - \min_{-1 \leq t \leq T} f(x^t)$. Now, we focus on estimating $G_T = \sum_{t=0}^{T-1} \|g^{t+1} - g^t\|_1$ term
 1729 in equation 26. Thus,
 1730

$$\begin{aligned}
 1732 \quad G_T &= \sum_{t=0}^{T-1} \|g^{t+1} - g^t\|_1 \leq \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - g^{t+1}\|_1 + \sum_{t=0}^{T-1} \|\nabla f(x^t) - g^t\|_1 \\
 1733 &\quad + \sum_{t=0}^{T-1} \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \\
 1734 &\stackrel{Lip}{\leq} \delta^t + \delta^{t+1} + L_\infty \sum_{t=0}^{T-1} \|x^{t+1} - x^t\|_\infty \\
 1735 &= \delta^t + \delta^{t+1} + L_\infty \sum_{t=0}^{T-1} \gamma^t \|\text{sign}(\nabla f(x^t))\|_\infty \\
 1736 &\leq \delta^t + \delta^{t+1} + L_\infty \sum_{t=0}^{T-1} \gamma^t. \tag{27}
 \end{aligned}$$

1747 Now let us choose $\phi(\gamma)$, which we push to the BISECTION procedure (Algorithm 4): $\phi(\gamma) = \frac{\mathfrak{N}(\gamma)}{\mathfrak{D}(\gamma)} =$
 1748 $\frac{\tilde{\Delta}_T(\gamma)}{G_T(\gamma) + \zeta(\gamma)}$, where $\tilde{\Delta}_T = f(x^{-1}) - \min_{-1 \leq t \leq T} f(x^t)$ and $\zeta = \min_{0 \leq t \leq T} \|g^t\|_1$. In that way, we obtain
 1749 some γ_0 , which can be equal to γ_{lo}^* or γ_{hi}^* (see Lemma D.2, Lemma D.3) and use it as a constant
 1750 stepsize for our method. Thus, equation 27 transforms into
 1751

$$G_T(\gamma_0) \leq \delta^t + \delta^{t+1} + \gamma_0 L_\infty T. \tag{28}$$

1752 Mention that, according to Lemma D.2, we can always entry to the procedure without infinite early
 1753 termination. In that way we have two situations: when we have no early terminations at all and we
 1754 are under the activity of Lemma D.3, and when we have an early termination with the initial γ_{lo}^* . We
 1755 divide the following proof into two steps, where we separately show the convergence guarantees in
 1756 these two situations.
 1757

Step 1: no early terminations.

1758 Since we have only two cases: $\gamma_0 = \gamma_{\text{lo}}^*$ or $\gamma_0 = \gamma_{\text{hi}}^*$, let us consider them separately.
 1759

- 1760 • $\gamma_0 = \gamma_{\text{hi}}^*$: equation 28 transforms into

$$\begin{aligned}
 1762 \quad G_T(\gamma_{\text{hi}}^*) &\leq \delta^t + \delta^{t+1} + \gamma_{\text{hi}}^* L_\infty T \stackrel{\text{Lemma D.3.9}}{\leq} \delta^t + \delta^{t+1} + \frac{\mathfrak{N}_T(\gamma_{\text{lo}}^*)}{\mathfrak{D}_T(\gamma_{\text{hi}}^*)} L_\infty T \\
 1763 &\stackrel{(i)}{=} \delta^t + \delta^{t+1} + \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{G_T(\gamma_{\text{hi}}^*) + \zeta(\gamma_{\text{hi}}^*)} L_\infty T \leq \delta^t + \delta^{t+1} + \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{G_T(\gamma_{\text{hi}}^*)} L_\infty T,
 \end{aligned}$$

1764 where (i) is correct due to the $\phi(\gamma)$ choice. Solving this quadratic inequality with respect to $G_T(\gamma_{\text{hi}}^*)$
 1765 (Lemma D.1), we obtain
 1766

$$G_T(\gamma_{\text{hi}}^*) \leq \delta^t + \delta^{t+1} + \sqrt{\tilde{\Delta}_T(\gamma_{\text{lo}}^*) L_\infty T} \leq \delta^t + \delta^{t+1} + \sqrt{\Delta^* L_\infty T}, \tag{29}$$

1767 where $\Delta^* = f(x^{-1}) - f(x^*)$. Plugging it into equation 26, we obtain
 1768

$$\begin{aligned}
 1769 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 &\leq \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{hi}}^*)}{\gamma_{\text{hi}}^*} + \frac{1}{T} G_T(\gamma_{\text{hi}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\
 1770 &\stackrel{\text{Lemma D.3.9}}{\leq} \frac{1}{T} \frac{2\mathfrak{D}_T(\gamma_{\text{hi}}^*)}{\mathfrak{N}_T(\gamma_{\text{hi}}^*)} \tilde{\Delta}_T(\gamma_{\text{hi}}^*) + \frac{1}{T} G_T(\gamma_{\text{hi}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\
 1771 &= \frac{2}{T} \frac{[G_T(\gamma_{\text{hi}}^*) + \zeta(\gamma_{\text{hi}}^*)] \tilde{\Delta}_T(\gamma_{\text{hi}}^*)}{\tilde{\Delta}_T(\gamma_{\text{hi}}^*)} + \frac{1}{T} G_T(\gamma_{\text{hi}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\
 1772 &= \frac{3}{T} G_T(\gamma_{\text{hi}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) + \frac{2\zeta(\gamma_{\text{hi}}^*)}{T}
 \end{aligned}$$

$$\begin{aligned} & \stackrel{29}{\leqslant} 3 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (6\delta^t + 4\delta^{t+1}) + \frac{2\|g^0\|_1}{T}. \end{aligned} \quad (30)$$

In that way, equation 30 is the final estimate when BISECTION procedure returns γ_{hi}^* .

• $\gamma_0 = \gamma_{\text{lo}}^*$: equation 28 transforms into

$$\begin{aligned} G_T(\gamma_{\text{lo}}^*) & \leqslant \delta^t + \delta^{t+1} + \gamma_{\text{lo}}^* L_\infty T \stackrel{\text{Lemma D.3.9}}{\leqslant} \delta^t + \delta^{t+1} + \frac{\mathfrak{N}_T(\gamma_{\text{lo}}^*)}{\mathfrak{D}_T(\gamma_{\text{lo}}^*)} L_\infty T \\ & \stackrel{(i)}{=} \delta^t + \delta^{t+1} + \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{G_T(\gamma_{\text{lo}}^*) + \zeta(\gamma_{\text{lo}}^*)} L_\infty T \leqslant \delta^t + \delta^{t+1} + \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{G_T(\gamma_{\text{lo}}^*)} L_\infty T, \end{aligned}$$

where (i) is correct due to $\phi(\gamma)$ choice. Solving this quadratic inequality with respect to $G_T(\gamma_{\text{lo}}^*)$ (Lemma D.1), we obtain

$$G_T(\gamma_{\text{lo}}^*) \leqslant \delta^t + \delta^{t+1} + \sqrt{\tilde{\Delta}_T(\gamma_{\text{lo}}^*) L_\infty T} \leqslant \delta^t + \delta^{t+1} + \sqrt{\Delta^* L_\infty T}. \quad (31)$$

Now we make an additional distinction and consider the estimates separately: one case when $\gamma_{\text{lo}}^* > \sqrt{\frac{\Delta^*}{L_\infty T}}$ and another when $\gamma_{\text{lo}}^* \leqslant \sqrt{\frac{\Delta^*}{L_\infty T}}$. We can do this without any limitations, since combining the intervals considered for γ_{lo}^* returns all possible values.

◦ $\gamma_{\text{lo}}^* > \sqrt{\frac{\Delta^*}{L_\infty T}}$: we straightforwardly move to the equation 26 estimation:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 & \leqslant \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\gamma_{\text{lo}}^*} + \frac{1}{T} G_T(\gamma_{\text{lo}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\ & \leqslant \frac{\sqrt{L_\infty}}{\sqrt{\Delta^* T}} \tilde{\Delta}_T(\gamma_{\text{lo}}^*) + \frac{1}{T} G_T(\gamma_{\text{lo}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\ & \stackrel{31}{\leqslant} \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (4\delta^t + 2\delta^{t+1}) \\ & = 2 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (4\delta^t + 2\delta^{t+1}). \end{aligned} \quad (32)$$

◦ $\gamma_{\text{lo}}^* \leqslant \sqrt{\frac{\Delta^*}{L_\infty T}}$: in this case we start from the estimate that is followed by equation 28:

$$G_T(\gamma_{\text{hi}}^*) \leqslant \delta^t + \delta^{t+1} + \gamma_{\text{hi}}^* L_\infty T \stackrel{(i)}{\leqslant} \delta^t + \delta^{t+1} + 2\gamma_{\text{lo}}^* L_\infty T \leqslant \delta^t + \delta^{t+1} + 2\sqrt{\Delta^* L_\infty T} \quad (33)$$

where (i) is done due to bisection stop condition. Now we proceed to the equation 26 estimation:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 & \leqslant \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\gamma_{\text{lo}}^*} + \frac{1}{T} G_T(\gamma_{\text{lo}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\ & \stackrel{\text{Lemma D.3.9}}{\leqslant} \frac{1}{T} \frac{2\mathfrak{D}_T(\gamma_{\text{hi}}^*)}{\mathfrak{N}_T(\gamma_{\text{lo}}^*)} \tilde{\Delta}_T(\gamma_{\text{lo}}^*) + \frac{1}{T} G_T(\gamma_{\text{lo}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\ & \stackrel{\text{Lemma D.3.11}}{\leqslant} \frac{2}{T} \frac{[G_T(\gamma_{\text{hi}}^*) + \zeta(\gamma_{\text{hi}}^*)] \tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)} + \frac{G_T(\gamma_{\text{hi}}^*) + \zeta(\gamma_{\text{hi}}^*)}{T} \\ & \quad + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\ & = \frac{3G_T(\gamma_{\text{hi}}^*)}{T} + \frac{1}{T} (3\delta^t + \delta^{t+1}) + \frac{3\zeta(\gamma_{\text{hi}}^*)}{T} \\ & \stackrel{33}{\leqslant} 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (6\delta^t + 4\delta^{t+1}) + \frac{3\zeta(\gamma_{\text{hi}}^*)}{T} \\ & \leqslant 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (6\delta^t + 4\delta^{t+1}) + \frac{3\|g^0\|_1}{T}. \end{aligned} \quad (34)$$

Combining equation 32 and equation 34, we get that equation 34 is the final estimate when BISECTION procedure returns γ_{lo}^* .

In the end, equation 30 and equation 34 give us the estimate in the case when BISECTION procedure does not have early terminations at all and outputs any value of γ_0 :

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (6\delta^t + 4\delta^{t+1}) + \frac{3\|g^0\|_1}{T}. \quad (35)$$

Step 2: early termination with γ_{lo} .

Now we consider the scenario when, with the initial γ_{lo} , there is $\gamma_{\text{lo}} \geq \phi(\gamma_{\text{lo}})$ and algorithm early returns γ_{lo}^* . To consider this, we should choose the initial $\gamma_{\text{lo}} = \gamma_{\text{lo}}^* \leq \frac{\Delta^*}{L_\infty T}$. Thus, equation 28 transforms into

$$G_T(\gamma_{\text{lo}}^*) \leq \delta^t + \delta^{t+1} + \gamma_{\text{lo}} L_\infty T \leq \delta^t + \delta^{t+1} + \sqrt{L_\infty \Delta^* T}. \quad (36)$$

In that way, equation 26 transforms into

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 &\leq \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\gamma_{\text{lo}}^*} + \frac{1}{T} G_T(\gamma_{\text{lo}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\ &\leq \frac{1}{T} \frac{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\phi(\gamma_{\text{lo}}^*)} + \frac{1}{T} G_T(\gamma_{\text{lo}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\ &= \frac{1}{T} \frac{[G_T(\gamma_{\text{lo}}^*) + \zeta(\gamma_{\text{lo}}^*)] \tilde{\Delta}_T(\gamma_{\text{lo}}^*)}{\tilde{\Delta}_T(\gamma_{\text{lo}}^*)} + \frac{1}{T} G_T(\gamma_{\text{lo}}^*) + \frac{1}{T} (3\delta^t + \delta^{t+1}) \\ &= \frac{2G_T(\gamma_{\text{lo}}^*)}{T} + \frac{1}{T} (3\delta^t + \delta^{t+1}) + \frac{\zeta(\gamma_{\text{lo}}^*)}{T} \\ &\stackrel{36}{\leq} 2 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (5\delta^t + 3\delta^{t+1}) + \frac{\|g^0\|_1}{T}. \end{aligned} \quad (37)$$

In that way, equation 37 delivers the estimate, when Algorithm 4 makes an early termination.

Combining equation 35 with equation 37, we finally obtain the estimate for all possible cases of the BISECTION procedure return:

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t(\gamma_0))\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (6\delta^t + 4\delta^{t+1}) + \frac{3\|g^0\|_1}{T}. \quad (38)$$

Now it is time to take expectation and give estimate to δ^t . One can note, using the law of total expectation ($\mathbb{E}[\xi] = \mathbb{E}[\mathbb{E}[\xi|\psi]]$),

$$\begin{aligned} \mathbb{E}\|\nabla f(x^t) - g^t\|_1 &= \sum_{i=1}^d \mathbb{E}|\nabla f(x^t)_i - [g^t]_i| \stackrel{(Jen)}{\leq} \sum_{i=1}^d \sqrt{\mathbb{E}((\nabla f(x^t)_i - [g^t]_i)^2)} \\ &= \sum_{i=1}^d \sqrt{\mathbb{E}[(\nabla f(x^t)_i - [g^t]_i)^2 | x^t]} \leq \sum_{i=1}^d \sigma_i^t. \end{aligned}$$

In that way, we obtain important bound:

$$\mathbb{E}\|\nabla f(x^t) - g^t\|_1 \leq \|\sigma\|_1. \quad (39)$$

Then,

$$\begin{aligned} \mathbb{E}\delta^t &= \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(x^t) - g^t\|_1 \leq \sum_{t=0}^{T-1} \|\sigma\|_1 \leq \|\sigma\|_1 T, \\ \mathbb{E}\delta^{t+1} &= \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(x^{t+1}) - g^{t+1}\|_1 \leq \sum_{t=0}^{T-1} \|\sigma\|_1 = \|\sigma\|_1 T. \end{aligned}$$

Substituting it to equation 38, we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(x^t)\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + 10\|\sigma\|_1 + \frac{3\mathbb{E}\|g^0\|_1}{T}.$$

1890 Expressing the number of iterations and using $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$ as a criterion, we obtain
 1891 that algorithm needs $\mathcal{O}\left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \|\sigma\|_1^2\right)$ iterations to reach ε -accuracy. Note the we drop the term
 1892 $\frac{3\mathbb{E}\|g^0\|_1}{T}$, since it is asymptotically smaller than the main one. However we firstly need to find step
 1893 γ_0 with bisection procedure that takes $T \log \log \left(\frac{\gamma_\varepsilon 2^{2^k}}{\gamma_\varepsilon}\right) = \mathcal{O}\left(\left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \|\sigma\|_1^2\right) k\right)$ iterations,
 1894 where 2^{2^k} denotes the length of the initial interval for the stepsize. We have already discussed
 1895 in the main part that, according to Lemma D.2, k should be at least $k = \log \log \frac{\Delta^*}{\gamma_s \|g^0\|_1}$. Thus,
 1896 $\mathcal{O}\left(\left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \|\sigma\|_1^2\right) \log \log \frac{\Delta^*}{\gamma_s \|g^0\|_1}\right)$ is the final iteration complexity. \square
 1897

1900 *Remark E.5.* Under conditions of Theorem E.4 Algorithm 5 with mini-batch of the size $t+1$ at t -th
 1901 iteration to reach ε -accuracy needs

$$\mathcal{O}\left(\frac{\Delta^* L_\infty + \|\sigma\|_1^2}{\varepsilon^2} \log \log \frac{\Delta^*}{\gamma_s \|g^0\|_1}\right) \text{ iterations.}$$

1902 *Proof.* The proof of the remark repeats the proof of Theorem 3.9 except for the estimate on
 1903 $\mathbb{E}\|\nabla f(x^t) - g^t\|_1^2$ term. Since we now use mini-batches, we can bound
 1904

$$\mathbb{E}\|\nabla f(x^t) - g^t\|_1^2 \leq \frac{\|\sigma\|_1}{\sqrt{t+1}},$$

1905 instead of equation 39. In that way,

$$\frac{1}{T} \mathbb{E} \delta^t = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(x^t) - g^t\|_1 \leq \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\sigma\|_1}{t+1} \leq 2 \frac{\|\sigma\|_1}{\sqrt{T}},$$

1906 which ends the proof of the remark. \square

E.3 DISTRIBUTED SETTING

1907 To begin with, we present the modification of the classic SIGN-SGD algorithm (Algorithm 1) that
 1908 aligns with the distributed learning. We consider SIGN-SGD with majority vote (Algorithm 6),
 1909 similarly to (Bernstein et al., 2018). We present the assumption that we utilize in distributed regime.

1910 **Assumption E.6.** In the multi-node regime of learning each node $j = 1, M$ at any point $x \in \mathbb{R}^d$ has
 1911 an access to the stochastic gradient, i.e., it can compute $g_j(x) = \nabla f(x, \xi_j)$ – the stochastic gradient
 1912 value with respect to the randomness in the choice of samples ξ_j . Additionally, this stochastic
 1913 estimators is unbiased, i.e., $\mathbb{E}[g_j(x)] = \nabla f(x)$, and its variance is coordinate-wise bounded, i.e.,
 1914 $\mathbb{E}([g_j(x)]_i - [\nabla f(x)]_i)^2 \leq \sigma_i^2$. \square

Algorithm 6 SIGN-SGD with majority vote

- 1: **Input:** Start point $x^0 \in \mathbb{R}^d$, number of iterations T
- 2: **Parameter:** Step size $\gamma > 0$
- 3: **for** $t = 0, \dots, T-1$ **do**
- 4: **for** all nodes $j = 1, \dots, M$ in parallel **do**
- 5: Compute stochastic gradient $g_j(x^t) = \nabla f(x^t, \xi_j)$
- 6: Send sign($g_j(x^t)$) to the server
- 7: **end for**
- 8: $x^{t+1} = x^t - \gamma \text{sign}\left(\sum_{j=1}^M \text{sign}(g_j(x^t))\right)$
- 9: **end for**

1915 Proceeding analogically to the stochastic one-node regime, we establish $\mathfrak{N}_T(\gamma)$ and $\mathfrak{D}_T(\gamma)$
 1916 that we use in $\phi(\gamma)$ in the BISECTION procedure: $\mathfrak{N}_T(\gamma) = \tilde{\Delta}_T(\gamma), \mathfrak{D}_T(\gamma) =$

1944 $\sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M (\|g_j(x^{t+1}) - g_j(x^t)\|_1 + \zeta(\gamma))$. Let us emphasize how this affects Algorithms
 1945 4, 5. Firstly, we now need to call the SIGN-SGD with majority vote method (Algorithm 6) instead of
 1946 SIGN-SGD (Algorithm 1). Secondly, to obtain $\mathfrak{D}_T(\gamma)$ in the bisection procedure, each node j counts
 1947 $\sum_{t=0}^{T-1} \|g_j(x^{t+1}) - g_j(x^t)\|_1$ using locally stored gradients, and sends the complete sum to the server
 1948 in the end. Note that this requirement has no effect on extra memory and communication complexity,
 1949 since each device requires only $\mathcal{O}(d)$ extra memory and performs only one extra communication
 1950 during the whole learning. Now we present the theoretical result for the distributed setting.

1951 **Lemma E.7** (Theorem 2 (a) from (Bernstein et al., 2018)). *Suppose Assumption E.6 holds. Then, at*
 1952 *any point $x \in \mathbb{R}^d$, the following estimate is valid:*

$$1954 \quad 1955 \quad 1956 \quad |[\nabla f(x)]_i| \mathbb{P} \left(\text{sign} \left(\sum_{j=1}^M \text{sign} ([g_j(x)]_i) \right) \neq \text{sign} ([\nabla f(x)]_i) \right) \leq \sigma_i.$$

1957 **Lemma E.8** (Descent lemma). *For Algorithm 5 under Assumptions 3.1, 3.2, 3.3, E.6, the following*
 1958 *estimate is valid:*

$$1959 \quad 1960 \quad 1961 \quad \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{f(x^{-1}) - f(x^T)}{\gamma_0} + \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|g_j^{t+1} - g_j^t\|_1 + 2\tilde{\delta}^T + \delta^t + \delta^{t+1},$$

1962 where $\delta^t = \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_j^t\|_1$

1963 and $\tilde{\delta}^T = \sum_{t=0}^{T-1} \sum_{i=1}^d |[\nabla f(x^t)]_i| \mathbb{I} \left(\text{sign} \left(\sum_{j=1}^M \text{sign} ([g_j^t]_i) \right) \neq \text{sign} ([\nabla f(x^t)]_i) \right)$.

1968 *Proof.* Starting from the convexity of the objective,

$$1969 \quad 1970 \quad 1971 \quad f(x^{t+1}) - f(x^t) \leq \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle = -\gamma^t \left\langle \nabla f(x^{t+1}), \text{sign} \left(\sum_{j=1}^M \text{sign}(g_j^t) \right) \right\rangle$$

$$1972 \quad 1973 \quad 1974 \quad = -\gamma^t \left\langle \nabla f(x^t), \text{sign} \left(\sum_{j=1}^M \text{sign}(g_j^t) \right) \right\rangle$$

$$1975 \quad 1976 \quad 1977 \quad -\gamma^t \left\langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign} \left(\sum_{j=1}^M \text{sign}(g_j^t) \right) \right\rangle$$

$$1978 \quad 1979 \quad 1980 \quad = -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \sum_{i=1}^d |[\nabla f(x^t)]_i|$$

$$1981 \quad 1982 \quad 1983 \quad \cdot \mathbb{I} \left(\text{sign} \left(\sum_{j=1}^M \text{sign} ([g_j^t]_i) \right) \neq \text{sign} ([\nabla f(x^t)]_i) \right)$$

$$1984 \quad 1985 \quad 1986 \quad -\gamma^t \left\langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign} \left(\sum_{j=1}^M \text{sign}(g_j^t) \right) \right\rangle$$

$$1987 \quad 1988 \quad 1989 \quad \stackrel{\text{Conj},(i)}{\leq} -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t$$

$$1990 \quad 1991 \quad 1992 \quad + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \left\| \text{sign} \left(\sum_{j=1}^M \text{sign} (g_j^t) \right) \right\|_\infty$$

$$1993 \quad 1994 \quad 1995 \quad \leq -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1$$

$$1996 \quad 1997 \quad 1998 \quad = -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1$$

$$\begin{aligned}
& \stackrel{CS}{\leq} -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t + \gamma^t \frac{1}{M} \sum_{j=1}^M \|g_j^{t+1} - g_j^t\|_1 \\
& + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_j^{t+1}\|_1 + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_j^t\|_1, \quad (40)
\end{aligned}$$

where in (i) we denote $\tilde{\delta}^t = \sum_{i=1}^d |\nabla f(x^t)|_i \mathbb{I} \left(\text{sign} \left(\sum_{j=1}^M \text{sign} \left([g_j^t]_i \right) \right) \neq \text{sign} \left([\nabla f(x^t)]_i \right) \right)$. Now

we rearrange terms and summarize over all iterations to obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 & \leq \sum_{t=0}^{T-1} [f(x^t) - f(x^{t+1})] + 2 \sum_{t=0}^{T-1} \gamma^t \tilde{\delta}^t + \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \gamma^t \|g_j^{t+1} - g_j^t\|_1 \\
& + \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \gamma^t \|\nabla f(x^t) - g_j^t\|_1 + \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \gamma^t \|\nabla f(x^{t+1}) - g_j^{t+1}\|_1.
\end{aligned}$$

Since Algorithm 5 performs all the steps with the constant rate γ_0 , which we define later, denoting

$\tilde{\delta}^T = \sum_{t=0}^{T-1} \tilde{\delta}^t$, we can rewrite the result in the following form:

$$\begin{aligned}
\sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 & \leq \sum_{t=0}^{T-1} \frac{[f(x^t) - f(x^{t+1})]}{\gamma_0} + 2\tilde{\delta}^T + \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|g_j^{t+1} - g_j^t\|_1 \\
& + \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_j^t\|_1 + \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_j^{t+1}\|_1.
\end{aligned}$$

In the obtained estimate the last two terms consist from differences between the honest and stochastic gradient at the t -th and $(t+1)$ -th moments. One of our goals is to estimate them, however, we want

to perform analogically to Theorem E.4 and continue the proof with the $\sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|g_j^{t+1} - g_j^t\|_1$

term estimate. To simplify the subsequent notation, we introduce the following definition: let

$\delta^t = \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_j^t\|_1$. In that way, the following inequality finishes the proof of the

lemma:

$$\sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{f(x^{-1}) - f(x^T)}{\gamma_0} + \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|g_j^{t+1} - g_j^t\|_1 + 2\tilde{\delta}^T + \delta^t + \delta^{t+1}.$$

□

Theorem E.9. Suppose Assumptions 3.1, 3.2, 3.3, E.6 hold. Then for Algorithm 5 using at t -th iteration mini-batches of sizes $t+1$, after obtaining the stepsize γ_0 , the following estimate is valid:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(x^t)\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + 10 \|\sigma\|_+ \frac{\frac{3}{M} \sum_{j=1}^M \mathbb{E} \|g_j^0\|_1}{T}.$$

Moreover, taking into account the complexity of Algorithm 4 in relation to the initial stepsize bound

γ_s , to reach ε -accuracy, where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, Algorithm 5 needs

$$\mathcal{O} \left(\left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \|\sigma\|_1^2 \right) \log \log \frac{\Delta^*}{\gamma_s \sum_{j=1}^M \|g_j^0\|_1} \right) \text{ iterations.}$$

2052 *Proof.* Let us mention that the result of Lemma E.8 almost matches the starting point of Theorem
 2053 E.4 equation 26. If we now denote $G_T = \sum_{t=0}^{T-1} \frac{1}{M} \sum_{j=1}^M \|g_j^{t+1} - g_j^t\|_1$, the only difference is that there
 2054 we have an additional $2\tilde{\delta}^T$ term. However, we do not estimate it yet and it does not require any
 2055 transformations. Thus, we can proceed in a manner completely analogous to the proof of Theorem
 2056 E.4 and obtain an analog of the estimate in equation 38:
 2057

$$2059 \quad 2060 \quad 2061 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t(\gamma_0))\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{T} (2\tilde{\delta}^T + 4\delta^t + 4\delta^{t+1}) + \frac{\frac{3}{M} \sum_{j=1}^M \|g_j^0\|_1}{T}, \quad (41)$$

2063 where $\Delta^* = f(x^{-1}) - f(x^*)$. Now we take expectation and use Lemma E.7 to obtain
 2064

$$2065 \quad 2066 \quad \mathbb{E} \tilde{\delta}^t = \sum_{i=1}^d |[\nabla f(x^t)]_i| \mathbb{P} \left(\text{sign} \left(\sum_{j=1}^M \text{sign} ([g_j^t]_i) \right) \neq \text{sign} ([\nabla f(x^t)]_i) \right) \\ 2067 \quad \leq \sum_{i=1}^d \sigma_i^t = \|\sigma\|_1. \quad (42)$$

2071 For $\mathbb{E} \delta^t$, under Assumption E.6, we have the estimate as equation 39:
 2072

$$2073 \quad \mathbb{E} \|\nabla f(x^t) - g_j^t\|_1 \leq \|\sigma\|_1.$$

2074 Thus, substituting both of these estimates to equation 41, we obtain the final convergence result:
 2075

$$2076 \quad 2077 \quad 2078 \quad \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(x^t)\|_1 \leq 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + \frac{1}{M} \sum_{j=1}^M \frac{1}{T} \sum_{t=0}^{T-1} 10 \|\sigma\|_1 + \frac{\frac{3}{M} \sum_{j=1}^M \mathbb{E} \|g_j^0\|_1}{T} \\ 2079 \\ 2080 \quad = 6 \frac{\sqrt{\Delta^* L_\infty}}{\sqrt{T}} + 10 \|\sigma\|_1 + \frac{\frac{3}{M} \sum_{j=1}^M \mathbb{E} \|g_j^0\|_1}{T}.$$

2083 Since we obtain the same convergence estimate as in Theorem E.4, we can analogically establish the
 2084 $\mathcal{O} \left(\left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \|\sigma\|_1^2 \right) \log \log \frac{\Delta^*}{\gamma_s \frac{1}{M} \sum_{j=1}^M \|g_j^0\|_1} \right)$ iteration complexity. \square
 2085
 2086
 2087

2088 *Remark E.10.* Under conditions of Theorem E.9 Algorithm 5 with mini-batches of the size $t+1$ at
 2089 t -th iteration to reach ε -accuracy needs
 2090

$$2091 \quad 2092 \quad 2093 \quad 2094 \quad \mathcal{O} \left(\frac{\Delta^* L_\infty + \|\sigma\|_1^2}{\varepsilon^2} \log \log \frac{\Delta^*}{\gamma_s \frac{1}{M} \sum_{j=1}^M \|g_j^0\|_1} \right) \text{ iterations.}$$

2096 *Proof.* Proof repeats the proofs of Remark E.5. \square
 2097

2098 F PROOFS FOR ALIAS

2100 F.1 EXACT GRADIENT SETTING

2102 **Lemma F.1** (Approximating sequence). *Let the initial Δ^* -approximation d^0 be $0 < d^0 < \Delta^*$, where
 2103 $\Delta^* = f(x^0) - f(x^*)$. Then for Algorithm 2 under Assumptions 3.1, 3.2, 3.3, 3.4, the following
 2104 estimate is valid:*

$$2105 \quad \Delta^* \geq d^n \quad \forall n \in [0, T-1].$$

2106 *Proof.* Starting from the convexity of the objective,

$$2107 \quad f(x^{t+1}) - f(x^t) \leq \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle = -\gamma^t \langle \nabla f(x^{t+1}), \text{sign}(\nabla f(x^t)) \rangle. \quad (43)$$

2108 Now we summarize both sides over the first n iterations:

$$2109 \quad -\Delta^* = f(x^*) - f(x^0) \stackrel{(i)}{\leq} f(x^n) - f(x^0) = \sum_{t=0}^{n-1} f(x^{t+1}) - f(x^t) \\ 2110 \\ 2111 \quad \stackrel{43}{\leq} -\sum_{t=0}^{n-1} \gamma^t \langle \nabla f(x^{t+1}), \text{sign}(\nabla f(x^t)) \rangle,$$

2112 where (i) is correct due to Assumption 3.3. Changing the sign of the inequality,

$$2113 \quad \tilde{d}^n = \sum_{t=0}^{n-1} \gamma^t \langle \nabla f(x^{t+1}), \text{sign}(\nabla f(x^t)) \rangle \leq \Delta^*.$$

2114 Since our algorithm performs $d^n = \max(d^{n-1}, \tilde{d}^n)$ and we initialize our sequence with $d^0 < \Delta^*$,
2115 we obtain the required statement. \square

2116 **Lemma F.2** (Descent lemma). *For Algorithm 2 under Assumptions 3.1, 3.2, 3.3, 3.4, the following
2117 estimate is valid:*

$$2118 \quad \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 \leq \Delta^* + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^t,$$

2119 where $L_\infty^t = \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty}$.

2120 *Proof.*

$$2121 \quad \begin{aligned} f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle = f(x^t) - \gamma^t \langle \nabla f(x^{t+1}), \text{sign}(\nabla f(x^t)) \rangle \\ 2122 &= f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 - \gamma^t \langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign}(\nabla f(x^t)) \rangle \\ 2123 &\stackrel{\text{Conj}}{\leq} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \|\text{sign}(\nabla f(x^t))\|_\infty \\ 2124 &\leq f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \\ 2125 &\stackrel{(i)}{=} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty} \|x^{t+1} - x^t\|_\infty \\ 2126 &= f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + (\gamma^t)^2 \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty}, \end{aligned}$$

2127 where in (i) we assume $\|x^{t+1} - x^t\|_\infty \neq 0$. Indeed, $\|x^{t+1} - x^t\|_\infty = 0$ follows from the equality
2128 $\text{sign}(\nabla f(x^t)) = 0$, which means that we find the optimum and do not need to find another point x^{t+1} .

2129 Now we denote $L_\infty^t = \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty}$. Summing over all iterations, we obtain

$$2130 \quad \begin{aligned} \sum_{t=0}^{T-1} \gamma^t \|f(x^t)\|_1 &\leq \sum_{t=0}^{T-1} [f(x^t) - f(x^{t+1})] + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^t \\ 2131 &= f(x^0) - f(x^*) + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^t \leq \Delta^* + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^t, \end{aligned}$$

2132 which ends the proof of the lemma. \square

2133 **Theorem F.3 (Theorem 3.5).** *Suppose Assumptions 3.1, 3.2, 3.3, 3.4 hold. We denote $\varepsilon \geq$
2134 $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, $L_\infty^t = \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty}$. Then Algorithm 2 with Option I, $d^0 < \Delta^*$
2135 to reach ε -accuracy needs*

$$2136 \quad \tilde{\mathcal{O}} \left(\frac{(\Delta^*)^2 (L_\infty^0)^3}{d^0 (L_\infty^0)^2 \varepsilon^2} \right) \text{ iterations.}$$

2160 *Algorithm 2 with Option II to reach ε -accuracy needs*
 2161

$$2162 \quad 2163 \quad \tilde{\mathcal{O}}\left(\frac{\Delta^*(L_\infty)^3}{(L_\infty^0)^2 \varepsilon^2}\right) \text{ iterations.}$$

2164
 2165 *Proof.* Let us start with the result of Lemma F.2:
 2166

$$2167 \quad \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 \leq \Delta^* + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^t. \quad (44)$$

2169 Now we use our γ^t choice. Let us firstly estimate the denominator that is exactly $\lambda^t =$
 2170 $\frac{1}{\sqrt{\sum_{i=0}^{t-1} \frac{\|\nabla f(x^{i+1}) - \nabla f(x^i)\|_1}{\|x^{i+1} - x^i\|_\infty}}} = \frac{1}{\sqrt{\sum_{i=0}^{t-1} L_\infty^i}}$ and is the same for both Options I and II. Let us estimate
 2171 the following term.
 2172
 2173

$$2174 \quad 2175 \quad \sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^t = \sum_{t=0}^{T-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i}.$$

2179 We mention, that each L_∞^i is bounded from the definition of smoothness (see Assumption 3.1), i.e.,
 2180 $L_\infty^i \leq L_\infty$. We consider the sequence $\{L_\infty^i\}_{i=0}^{T-1}$. Since each term in this sequence is bounded,
 2181 there exists r such that $\sum_{i=0}^{r-2} L_\infty^i \leq L_\infty^{r-1}$ and for each $t \geq r-1$ such that $\sum_{i=0}^t L_\infty^i \geq L_\infty^{t+1}$. In that
 2182 way, we divide the sum into two parts:
 2183
 2184

$$2185 \quad 2186 \quad \sum_{t=0}^{T-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i} = \sum_{t=0}^{r-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i} + \sum_{t=r}^{T-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i}. \quad (45)$$

2186 Considering the first sum in equation 45, we mention, that we can estimate the denominator as
 2187 $\sum_{i=0}^{t-1} L_\infty^i \geq L_\infty^0$. As for the numerator. Thus,
 2188

$$2189 \quad 2190 \quad \sum_{t=0}^{r-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i} \leq \frac{1}{L_\infty^0} \left(\sum_{t=0}^{r-2} L_\infty^t + L_\infty^{r-1} \right) \leq \frac{2L_\infty^{r-1}}{L_\infty^0} \leq \frac{2L_\infty}{L_\infty^0}. \quad (46)$$

2191 Considering the second sum in equation 45, we have
 2192

$$2193 \quad 2194 \quad \sum_{t=r}^{T-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i} = \sum_{t=r}^{T-1} \frac{L_\infty^t}{\frac{1}{2} \sum_{i=0}^{t-1} L_\infty^i + \frac{1}{2} \sum_{i=0}^{t-1} L_\infty^i}.$$

2195 Estimating any of the sums in the denominator, we claim, that $\sum_{i=0}^{t-1} L_\infty^i \geq L_\infty^t$, since $t-1 \geq r-1$.
 2196 In that way,
 2197

$$2198 \quad 2199 \quad \sum_{t=r}^{T-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i} \leq \sum_{t=r}^{T-1} \frac{2L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i} \leq 2 \sum_{t=0}^{T-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i}. \quad (47)$$

2214
2215 Next we denote $s^t = \sum_{i=0}^t L_\infty^i$ and have
2216

$$2217 \quad 2218 \quad 2219 \quad 2220 \quad L_\infty^t \frac{1}{\sum_{i=0}^t L_\infty^i} = (s^t - s^{t-1}) \frac{1}{\sum_{i=0}^t L_\infty^i} = \int_{s^{t-1}}^{s^t} \frac{1}{\sum_{i=0}^t L_\infty^i} dx \stackrel{(i)}{\leq} \int_{s^{t-1}}^{s^t} \frac{1}{x} dx, \quad (48)$$

2221 where (i) was done due to $\frac{1}{x}$ is a non-increasing function on $(0, +\infty)$. Summing over t , we obtain
2222

$$2223 \quad 2224 \quad 2225 \quad 2226 \quad 2 \sum_{t=1}^T \frac{L_\infty^t}{\sum_{i=0}^t L_\infty^i} \leq 2 \int_{s^0}^{s^T} \frac{1}{x} dx = 2 \log(s^T) - 2 \log(s^0) = 2 \log \left(\frac{\sum_{t=1}^T L_\infty^t}{L_\infty^0} \right) \leq 2 \log \left(\frac{L_\infty T}{L_\infty^0} \right).$$

2227 Combining this estimate with equation 47,
2228

$$2229 \quad 2230 \quad 2231 \quad 2232 \quad 2233 \quad \sum_{t=r}^{T-1} \frac{L_\infty^t}{\sum_{i=0}^{t-1} L_\infty^i} \leq 2 \sum_{t=1}^T \frac{L_\infty^t}{\sum_{i=0}^t L_\infty^i} + 2 \leq 2 \left(\log \left(\frac{L_\infty T}{L_\infty^0} \right) + 1 \right) \leq 4 \log \left(\frac{L_\infty T}{L_\infty^0} \right). \quad (49)$$

2234 Substituting equation 46 and equation 49 into equation 45, we obtain
2235

$$2236 \quad 2237 \quad 2238 \quad 2239 \quad \sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^t \leq 2 \frac{L_\infty}{L_\infty^0} + 4 \log \left(\frac{L_\infty T}{L_\infty^0} \right). \quad (50)$$

2240 We additionally note, that if $r > T - 1$, only first term remains in this estimate, consequently our
2241 bound equation 50 is correct.
2242

2243 In this way, utilizing Option I from Algorithm 2, equation 44 together with equation 50 yields

$$2244 \quad 2245 \quad 2246 \quad \sqrt{d^0} \lambda^{T-1} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \stackrel{(i)}{\leq} \sum_{t=0}^{T-1} \sqrt{d^t} \lambda^t \|\nabla f(x^t)\|_1 \leq \Delta^* + \sum_{t=0}^{T-1} d^t (\lambda^t)^2 L_\infty^t \\ 2247 \quad 2248 \quad 2249 \quad \stackrel{\text{Lemma F.1}}{\leq} \Delta^* + \Delta^* \sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^t, \\ 2250 \quad 2251 \quad 2252 \quad \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{\Delta^*}{\sqrt{d^0} \lambda^{T-1}} + \frac{\Delta^*}{\sqrt{d^0} \lambda^{T-1}} \sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^t \\ 2253 \quad 2254 \quad 2255 \quad \stackrel{50}{\leq} \frac{\Delta^*}{\sqrt{d^0} \lambda^{T-1}} + 4 \frac{\Delta^*}{\sqrt{d^0} \lambda^{T-1}} \log \left(\frac{L_\infty T}{L_\infty^0} \right) + 2 \frac{\Delta^* L_\infty}{\sqrt{d^0} \lambda^{T-1} L_\infty^0} \\ 2256 \quad 2257 \quad 2258 \quad \leq 7 \frac{\Delta^* L_\infty}{\sqrt{d^0} \lambda^{T-1} L_\infty^0} \log \left(\frac{L_\infty T}{L_\infty^0} \right), \quad (51)$$

2259 where (i) was done due to the fact that d^0 is minimal from all $\{d^t\}_{t=0}^{T-1}$ (Line 7 from Algorithm 2)

2260 and the definition of λ^t . Utilizing $\frac{1}{\lambda^{T-1}} = \sqrt{\sum_{t=0}^{T-2} L_\infty^t} \leq \sqrt{L_\infty T}$, we obtain the final estimate:
2261

$$2262 \quad 2263 \quad 2264 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{7 \Delta^* (L_\infty)^{\frac{3}{2}}}{\sqrt{d^0 T L_\infty^0}} \log \left(\frac{L_\infty T}{L_\infty^0} \right).$$

2265 Expressing the number of iterations and using $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$ as a criterion, we obtain that
2266 the algorithm needs $\tilde{\mathcal{O}} \left(\frac{(\Delta^*)^2 (L_\infty)^3}{d^0 (L_\infty^0)^2 \varepsilon^2} \right)$ iterations to reach ε -accuracy.
2267

Considering Option II from Algorithm 2, we can proceed absolutely analogical, however, using $f(x^0) - \tilde{f} \geq \Delta^*$ instead of Lemma F.1. In that way,

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 &\leq \frac{\Delta^* \sqrt{L_\infty}}{\sqrt{(f(x^0) - \tilde{f})T}} + \frac{4(f(x^0) - \tilde{f})\sqrt{L_\infty}}{\sqrt{(f(x^0) - \tilde{f})T}} \log\left(\frac{L_\infty T}{L_\infty^0}\right) \\ &\quad + \frac{2(f(x^0) - \tilde{f})(L_\infty)^{\frac{3}{2}}}{\sqrt{(f(x^0) - \tilde{f})T L_\infty^0}} \\ &\leq \frac{7\sqrt{(f(x^0) - \tilde{f})(L_\infty)^{\frac{3}{2}}}}{\sqrt{T} L_\infty^0} \log\left(\frac{L_\infty T}{L_\infty^0}\right). \end{aligned}$$

Expressing the number of iterations, using $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$ as a criterion, and utilizing \tilde{f} is an approximation of $f(x^*)$, we obtain that the algorithm needs $\tilde{\mathcal{O}}\left(\frac{\Delta^*(L_\infty)^3}{(L_\infty^0)^2 \varepsilon^2}\right)$ iterations to reach ε -accuracy. \square

Remark F.4 (Remark 3.6). Under conditions of Theorem 3.5 Algorithm 2 with $\lambda^t = \sqrt{L_\infty + \sum_{i=0}^{t-1} \frac{\|\nabla f(x^{i+1}) - \nabla f(x^i)\|_1}{\|x^{i+1} - x^i\|_\infty}}$ and Option II to reach ε -accuracy needs

$$\tilde{\mathcal{O}}\left(\frac{\Delta^* L_\infty}{\varepsilon^2}\right) \text{ iterations,}$$

where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$.

Proof. The proof of the remark repeats the proof of Theorem 3.5 except for the estimate on $\sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^t$ term. Let us derive it. We use definition $L_\infty^t = \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty}$.

$$\sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^t = \sum_{t=0}^{T-1} \frac{L_\infty^t}{L_\infty + \sum_{i=0}^{t-1} L_\infty^i} \leq \sum_{t=0}^{T-1} \frac{L_\infty^t}{\sum_{i=0}^t L_\infty^i}.$$

Continuing analogically to equation 48 - equation 49, we get

$$\sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^t \leq 2 \log\left(\frac{L_\infty T}{L_\infty^0}\right).$$

Substituting this bound into equation 51 instead of equation 50, we ends the proof of the remark. \square

F.2 STOCHASTIC GRADIENT SETTING

In this section we denote $g_{\xi^t}^t$ the stochastic gradient at the t -th iteration (point x^t), according to the stochastic realization ξ^t at the t -th iteration.

Before proceeding to the theoretical analysis of the algorithm, we present its formal description, Algorithm 7, specifying which option for the sequence d^t we use in practice and which one we analyze theoretically.

Algorithm 7 ALIAS stochastic version

```

2322 1: Input: Starting point  $x^0 \in \mathbb{R}^d$ , initial  $L_\infty$ -approximation  $\eta^{-1} = 0$ , initial  $\Delta^*$ -approximation  $d^0$ 
2323    $\in \mathbb{R}_+$ , lower bound  $\tilde{f}$  on  $f(x^*)$ , number of iterations  $T$ 
2324 2: for  $t = 0, \dots, T-1$  do
2325   3:   Compute gradients  $g_{\xi^t}^t, g_{\xi^t}^{t-1}$ 
2326   4:    $\eta^t = \eta^{t-1} + \frac{\|g_{\xi^t}^t - g_{\xi^t}^{t-1}\|_1}{\|x^t - x^{t-1}\|_\infty}; \lambda^t = \frac{1}{\sqrt{\eta^t}}$ 
2327   5:   if  $t \neq 0$  then
2328     6:      $\tilde{d}^t = \sum_{i=0}^{t-1} \gamma^i \langle g_{\xi^{i+1}}^i, \text{sign}(g_{\xi^{i+1}}^i) \rangle$ 
2329     7:      $d^t = \max(d^{t-1}, \tilde{d}^t)$ 
2330   8:   end if
2331   9:   Option I (Practical):  $\gamma^t = \lambda^t \sqrt{d^t}$ 
2332 10:   Option II (Theoretical):  $\gamma^t = \lambda^t \sqrt{f(x^0) - \tilde{f}}$ 
2333 11:    $x^{t+1} = x^t - \gamma^t \text{sign}(g_{\xi^t}^t)$ 
2334 12: end for

```

2340
2341 In the practical version of the algorithm, we use the stochastic gradient at the previous point with the
2342 current stochastic realization to update d^t . We use the same stochastic samples, similar to the update
2343 of the smoothness constant approximation, to reduce noise from the stochastic gradients.

2344 We now proceed to the convergence analysis.

2345 **Lemma F.5** (Descent lemma). *For Algorithm 2 under Assumptions 3.8, 3.2, 3.3, 3.7, the following
2346 estimate is valid:*

$$\begin{aligned}
2347 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \right] &\leq \Delta^* \mathbb{E} \left[\frac{1}{\sum_{t=0}^{T-1} \gamma^t} \right] + 2 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
2348 &+ \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
2349 &+ \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] + \mathbb{E} \left[\frac{\sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^{t, \xi^{t+1}}}{\sum_{t=0}^{T-1} \gamma^t} \right],
\end{aligned}$$

$$2351 \text{where } L_\infty^{t, \xi^t} = \frac{\|g_{\xi^t}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty}.$$

2352 *Proof.*

$$\begin{aligned}
2353 f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle = f(x^t) - \gamma^t \langle \nabla f(x^{t+1}), \text{sign}(g_{\xi^t}^t) \rangle \\
2354 &= f(x^t) - \gamma^t \|g_{\xi^t}^t\|_1 - \gamma^t \langle \nabla f(x^{t+1}) - g_{\xi^t}^t, \text{sign}(g_{\xi^t}^t) \rangle \\
2355 &\stackrel{\text{Conj}}{\leq} f(x^t) - \gamma^t \|g_{\xi^t}^t\|_1 + \gamma^t \|\nabla f(x^{t+1}) - g_{\xi^t}^t\|_1 \|\text{sign}(g_{\xi^t}^t)\|_\infty \\
2356 &\stackrel{\text{CS}}{\leq} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1 \\
2357 &\quad + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \|\text{sign}(g_{\xi^t}^t)\|_\infty \\
2358 &\stackrel{\text{CS}}{\leq} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1 + \gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1 \\
2359 &\quad + \gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1 + \gamma^t \|g_{\xi^{t+1}}^{t+1} - g_{\xi^t}^t\|_1 \|\text{sign}(g_{\xi^t}^t)\|_\infty
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(i)}{=} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1 + \gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1 \\
&\quad + \gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1 + \gamma^t \frac{\|g_{\xi^{t+1}}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty} \|x^{t+1} - x^t\|_\infty \\
&= f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1 + \gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1 \\
&\quad + \gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1 + (\gamma^t)^2 \frac{\|g_{\xi^{t+1}}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty},
\end{aligned}$$

where in (i) we assume $\|x^{t+1} - x^t\|_\infty \neq 0$. Indeed, $\|x^{t+1} - x^t\|_\infty = 0$ follows from the equality $\text{sign}(g_{\xi^t}^t) = 0$, which means $\|\text{sign}(g_{\xi^t}^t)\|_\infty = 0$ and at the t -th iteration this term equals zero. Thus, we can omit these iterations and consider this term only when it is non-zero, without any limitations. Now we denote $L_\infty^{t,\xi^t} = \frac{\|g_{\xi^t}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty}$. Summing over all iterations, we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 &\leq \sum_{t=0}^{T-1} f(x^t) - f(x^{t+1}) + 2 \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1 \\
&\quad + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1 + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1 \\
&\quad + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^{t,\xi^{t+1}} \\
&= f(x^0) - f(x^T) + 2 \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1 \\
&\quad + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1 + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1 \\
&\quad + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^{t,\xi^{t+1}} \\
&\leq \Delta^* + 2 \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1 + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1 \\
&\quad + \sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1 + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^{t,\xi^{t+1}}.
\end{aligned}$$

We divide both sides of inequality on $\sum_{t=0}^{T-1} \gamma^t$.

$$\begin{aligned}
\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 &\leq \frac{\Delta^*}{\sum_{t=0}^{T-1} \gamma^t} + 2 \sum_{t=0}^{T-1} \frac{\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \\
&\quad + \sum_{t=0}^{T-1} \frac{\gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} + \sum_{t=0}^{T-1} \frac{\gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \\
&\quad + \sum_{t=0}^{T-1} \frac{(\gamma^t)^2 L_\infty^{t,\xi^{t+1}}}{\sum_{t=0}^{T-1} \gamma^t}.
\end{aligned}$$

2430 Taking expectation, we obtain the final result of the lemma:
 2431

$$\begin{aligned}
 2432 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \right] &\leq \mathbb{E} \left[\frac{\Delta^*}{\sum_{t=0}^{T-1} \gamma^t} \right] + 2 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
 2433 &\quad + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
 2434 &\quad + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{(\gamma^t)^2 L_{\infty}^{t, \xi^{t+1}}}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
 2435 &= \Delta^* \mathbb{E} \left[\frac{1}{\sum_{t=0}^{T-1} \gamma^t} \right] + 2 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
 2436 &\quad + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
 2437 &\quad + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] + \mathbb{E} \left[\frac{\sum_{t=0}^{T-1} (\gamma^t)^2 L_{\infty}^{t, \xi^{t+1}}}{\sum_{t=0}^{T-1} \gamma^t} \right].
 \end{aligned}$$

□

2459
 2460 **Theorem F.6 (Theorem 3.9).** Suppose Assumptions 3.8, 3.2, 3.3, 3.7 hold. Then Algorithm 2 with
 2461 Option II to reach ε -accuracy, where $\varepsilon \geq \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \right]$ needs
 2462

$$\tilde{\mathcal{O}} \left(\frac{\Delta^* (L_{\infty})^3}{\varepsilon^2} \left(\mathbb{E} \left(\frac{1}{L_{\infty}^{0, \xi^1}} \right)^2 \right) + \|\sigma\|_1^2 L_{\infty} \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_{\infty}^{t, \xi^{t+1}}} \right) \right) \text{ iterations,}$$

2463 where $L_{\infty}^{t, \xi^{t+1}} = \frac{\|g_{\xi^{t+1}}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_{\infty}}$.
 2464

2465 *Proof.* Let us start with the result of Lemma F.5:
 2466

$$\begin{aligned}
 2467 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \right] &\leq \Delta^* \mathbb{E} \left[\frac{1}{\sum_{t=0}^{T-1} \gamma^t} \right] + 2 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^t) - g_{\xi^t}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
 2468 &\quad + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
 2469 &\quad + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] + \mathbb{E} \left[\frac{\sum_{t=0}^{T-1} (\gamma^t)^2 L_{\infty}^{t, \xi^{t+1}}}{\sum_{t=0}^{T-1} \gamma^t} \right].
 \end{aligned}$$

2484 Using equation Höl with $p = q = 2$, we rewrite it in the following form:
2485

$$\begin{aligned}
2486 \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \right] &\leq \Delta^* \mathbb{E} \left[\frac{1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
2487 &+ 2 \sum_{t=0}^{T-1} \left(\mathbb{E} \|\nabla f(x^t) - g_{\xi^t}^t\|_1^2 \right)^{\frac{1}{2}} \left(\mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \right]^2 \right)^{\frac{1}{2}} \\
2488 &+ \sum_{t=0}^{T-1} \left(\mathbb{E} \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1^2 \right)^{\frac{1}{2}} \left(\mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \right]^2 \right)^{\frac{1}{2}} \\
2489 &+ \sum_{t=0}^{T-1} \left(\mathbb{E} \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1^2 \right)^{\frac{1}{2}} \left(\mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \right]^2 \right)^{\frac{1}{2}} \\
2490 &+ \left(\mathbb{E} \left[\sum_{t=0}^{T-1} (\gamma^t)^2 L_{\infty}^{t, \xi^{t+1}} \right]^2 \right)^{\frac{1}{2}} \left(\mathbb{E} \left[\frac{1}{\sum_{t=0}^{T-1} \gamma^t} \right]^2 \right)^{\frac{1}{2}}. \quad (52)
\end{aligned}$$

2510 Now we use our choice of γ^t . Let us firstly estimate the denominator that is exactly $\lambda^t =$
2511 $\frac{1}{\sqrt{\sum_{i=0}^{t-1} \frac{\|g_{\xi^{i+1}}^{i+1} - g_{\xi^i}^i\|_1}{\|x^{i+1} - x^i\|_{\infty}}}} = \frac{1}{\sqrt{\sum_{i=0}^{t-1} L_{\infty}^{i, \xi^{i+1}}}}$. Let us estimate the following term.

$$\begin{aligned}
2512 \sum_{t=0}^{T-1} (\lambda^t)^2 L_{\infty}^{t, \xi^{t+1}} &= \sum_{t=0}^{T-1} \frac{L_{\infty}^{t, \xi^{t+1}}}{\sum_{i=0}^{t-1} L_{\infty}^{i, \xi^{i+1}}}.
\end{aligned}$$

2513 We mention, that each $L_{\infty}^{i, \xi^{i+1}}$ is bounded from the definition of smoothness (see Assumption 3.8),
2514 i.e., $L_{\infty}^{i, \xi^{i+1}} \leq L_{\infty}$. We consider the sequence $\{L_{\infty}^{i, \xi^{i+1}}\}_{i=0}^{T-1}$. Since each term in this sequence
2515 is bounded, there exists r such that $\sum_{i=0}^{r-2} L_{\infty}^{i, \xi^{i+1}} \leq L_{\infty}^{r-1, \xi^r}$ and for each $t \geq r-1$ such that
2516 $\sum_{i=0}^t L_{\infty}^{i, \xi^{i+1}} \geq L_{\infty}^{t+1, \xi^{t+2}}$. In that way, we divide the sum into two parts:

$$\sum_{t=0}^{T-1} \frac{L_{\infty}^{t, \xi^{t+1}}}{\sum_{i=0}^{t-1} L_{\infty}^{i, \xi^{i+1}}} = \sum_{t=0}^{r-1} \frac{L_{\infty}^{t, \xi^{t+1}}}{\sum_{i=0}^{t-1} L_{\infty}^{i, \xi^{i+1}}} + \sum_{t=r}^{T-1} \frac{L_{\infty}^{t, \xi^{t+1}}}{\sum_{i=0}^{t-1} L_{\infty}^{i, \xi^{i+1}}}. \quad (53)$$

2517 Considering the first sum in equation 53, we mention, that we can estimate the denominator as
2518 $\sum_{i=0}^{t-1} L_{\infty}^{i, \xi^{i+1}} \geq L_{\infty}^{0, \xi^1}$. As for the numerator. Thus,

$$\sum_{t=0}^{r-1} \frac{L_{\infty}^{t, \xi^{t+1}}}{\sum_{i=0}^{t-1} L_{\infty}^{i, \xi^{i+1}}} \leq \frac{1}{L_{\infty}^{0, \xi^1}} \left(\sum_{t=0}^{r-2} L_{\infty}^{t, \xi^{t+1}} + L_{\infty}^{r-1, \xi^r} \right) \leq \frac{2L_{\infty}^{r-1, \xi^r}}{L_{\infty}^{0, \xi^1}} \leq \frac{2L_{\infty}}{L_{\infty}^{0, \xi^1}}. \quad (54)$$

2538 Considering the second sum in equation 53, we have
 2539

2540

$$2541 \sum_{t=r}^{T-1} \frac{L_{\infty}^{t,\xi^{t+1}}}{\sum_{i=0}^{t-1} L_{\infty}^{i,\xi^{i+1}}} = \sum_{t=r}^{T-1} \frac{L_{\infty}^{t,\xi^{t+1}}}{\frac{1}{2} \sum_{i=0}^{t-1} L_{\infty}^{i,\xi^{i+1}} + \frac{1}{2} \sum_{i=0}^{t-1} L_{\infty}^{i,\xi^{i+1}}}.$$

2545

2546 Estimating any of the sums in the denominator, we claim, that $\sum_{i=0}^{t-1} L_{\infty}^{i,\xi^{i+1}} \geq L_{\infty}^{t,\xi^{t+1}}$, since $t-1 \geq r-1$. In that way,
 2547

2548

$$2549 \sum_{t=r}^{T-1} \frac{L_{\infty}^{t,\xi^{t+1}}}{\sum_{i=0}^{t-1} L_{\infty}^{i,\xi^{i+1}}} \leq \sum_{t=r}^{T-1} \frac{2L_{\infty}^{t,\xi^{t+1}}}{\sum_{i=0}^t L_{\infty}^{i,\xi^{i+1}}} \leq 2 \sum_{t=0}^{T-1} \frac{L_{\infty}^{t,\xi^{t+1}}}{\sum_{i=0}^t L_{\infty}^{i,\xi^{i+1}}}. \quad (55)$$

2554

2555 Next we denote $s^t = \sum_{i=0}^t L_{\infty}^{i,\xi^{i+1}}$ and have
 2556

$$2557 L_{\infty}^{t,\xi^{t+1}} \frac{1}{\sum_{i=0}^t L_{\infty}^{i,\xi^{i+1}}} = (s^t - s^{t-1}) \frac{1}{\sum_{i=0}^t L_{\infty}^{i,\xi^{i+1}}} = \int_{s^{t-1}}^{s^t} \frac{1}{\sum_{i=0}^t L_{\infty}^{i,\xi^{i+1}}} dx \stackrel{(i)}{\leq} \int_{s^{t-1}}^{s^t} \frac{1}{x} dx, \quad (56)$$

2562 where (i) was done due to $\frac{1}{x}$ is a non-increasing function on $(0, +\infty)$. Summing over t , we obtain
 2563

$$2564 2 \sum_{t=1}^T \frac{L_{\infty}^{t,\xi^{t+1}}}{\sum_{i=0}^t L_{\infty}^{i,\xi^{i+1}}} \leq 2 \int_{s^0}^{s^T} \frac{1}{x} dx = 2 \log(s^T) - 2 \log(s^0) = 2 \log \left(\frac{\sum_{t=1}^T L_{\infty}^{t,\xi^{t+1}}}{L_{\infty}^{0,\xi^1}} \right) \leq 2 \log \left(\frac{L_{\infty} T}{L_{\infty}^{0,\xi^1}} \right).$$

2568

2569 Combining this estimate with equation 55,

2570

$$2571 \sum_{t=r}^{T-1} \frac{L_{\infty}^{t,\xi^{t+1}}}{\sum_{i=0}^{t-1} L_{\infty}^{i,\xi^{i+1}}} \leq 2 \sum_{t=1}^T \frac{L_{\infty}^{t,\xi^{t+1}}}{\sum_{i=0}^t L_{\infty}^{i,\xi^{i+1}}} + 2 \leq 2 \left(\log \left(\frac{L_{\infty} T}{L_{\infty}^{0,\xi^1}} \right) + 1 \right) \leq 4 \log \left(\frac{L_{\infty} T}{L_{\infty}^{0,\xi^1}} \right). \quad (57)$$

2575

2576 Substituting equation 54 and equation 57 into equation 53, we obtain
 2577

2578

$$2579 \sum_{t=0}^{T-1} (\lambda^t)^2 L_{\infty}^{t,\xi^{t+1}} \leq 2 \frac{L_{\infty}}{L_{\infty}^{0,\xi^1}} + 4 \log \left(\frac{L_{\infty} T}{L_{\infty}^{0,\xi^1}} \right). \quad (58)$$

2581

2582 We additionally note, that if $r > T-1$, only first term remains in this estimate, consequently our
 2583 bound equation 58 is correct. Next, we estimate
 2584

$$2585 \frac{1}{\sum_{t=0}^{T-1} \lambda^t} = \frac{1}{\sum_{t=0}^{T-1} \frac{1}{\sqrt{L_{\infty} + \sum_{i=0}^{t-1} L_{\infty}^{i,\xi^{i+1}}}}} \leq \frac{\sqrt{L_{\infty}}}{\sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}}} \leq \frac{\sqrt{L_{\infty}}}{\sqrt{T}}. \quad (59)$$

2588

2589 Now we estimate the second, third and forth terms in equation 52. In the same manner, as in
 2590 equation 39, we can estimate
 2591

$$\mathbb{E} \left\| \nabla f(x^t) - g_{\xi^t}^t \right\|_1^2 \leq \|\sigma\|_1^2,$$

$$\begin{aligned}
2592 \quad & \mathbb{E} \left\| \nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1} \right\|_1^2 \leq \|\sigma\|_1^2, \\
2593 \quad & \mathbb{E} \left\| \nabla f(x^t) - g_{\xi^{t+1}}^t \right\|_1^2 \leq \|\sigma\|_1^2,
\end{aligned} \tag{60}$$

2596 where the last inequality is correct due to the fact that that stochastic realization ξ^{t+1} is independent
2597 from the point x^t . Thus, using equation 59,

$$\begin{aligned}
2599 \quad & \sum_{t=0}^{T-1} \left(\mathbb{E} \left\| \nabla f(x^t) - g_{\xi^t}^t \right\|_1^2 \right)^{\frac{1}{2}} \cdot \left(\mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \right]^2 \right)^{\frac{1}{2}} \\
2600 \quad & \leq \frac{\sqrt{L_\infty} \|\sigma\|_1}{\sqrt{T}} \sum_{t=0}^{T-1} \left(\mathbb{E} \frac{1}{\sum_{i=0}^{t-1} L_\infty^{i, \xi^{i+1}}} \right)^{\frac{1}{2}} \\
2601 \quad & \leq \frac{\sqrt{L_\infty} \|\sigma\|_1}{\sqrt{T}} \left(\mathbb{E} \min_{0 \leq t \leq T-1} L_\infty^{t, \xi^{t+1}} \right)^{\frac{1}{2}} \sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}} \\
2602 \quad & \leq 2\sqrt{L_\infty} \|\sigma\|_1 \left(\mathbb{E} \min_{0 \leq t \leq T-1} L_\infty^{t, \xi^{t+1}} \right)^{\frac{1}{2}}.
\end{aligned}$$

2616 It is clear that we can bound the rest two terms in the same manner. Now, substituting this estimate
2617 along with equation 58 and equation 59 into equation 52, we obtain
2618

$$\begin{aligned}
2619 \quad & \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \left\| \nabla f(x^t) \right\|_1 \right] \leq \frac{\Delta^* \sqrt{L_\infty}}{\sqrt{(f(x^0) - \tilde{f})T}} \\
2620 \quad & + 8\sqrt{L_\infty} \|\sigma\|_1 \left(\mathbb{E} \min_{0 \leq t \leq T-1} L_\infty^{t, \xi^{t+1}} \right)^{\frac{1}{2}} \\
2621 \quad & + 8 \frac{(f(x^0) - \tilde{f}) \sqrt{L_\infty}}{\sqrt{(f(x^0) - \tilde{f})T}} \left(\mathbb{E} \log^2 \left(\frac{L_\infty T}{L_\infty^{0, \xi^1}} \right) \right)^{\frac{1}{2}} \\
2622 \quad & + 4 \frac{(f(x^0) - \tilde{f}) \sqrt{L_\infty}}{\sqrt{(f(x^0) - \tilde{f})T}} \left(\mathbb{E} \left(\frac{L_\infty}{L_\infty^{0, \xi^1}} \right)^2 \right)^{\frac{1}{2}}.
\end{aligned} \tag{61}$$

2634 Now we use $\Delta^* \leq f(x^0) - \tilde{f}$ to obtain the final estimate:
2635

$$\begin{aligned}
2636 \quad & \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \left\| \nabla f(x^t) \right\|_1 \right] \leq 13 \frac{\sqrt{(f(x^0) - \tilde{f})} (L_\infty)^{\frac{3}{2}}}{T} \left(\mathbb{E} \left(\frac{1}{L_\infty^{0, \xi^1}} \right)^2 \right)^{\frac{1}{2}} \\
2637 \quad & \cdot \left(\mathbb{E} \log^2 \left(\frac{L_\infty T}{L_\infty^{0, \xi^1}} \right) \right)^{\frac{1}{2}} \\
2638 \quad & + 8\|\sigma\|_1 \left(\sqrt{L_\infty} \left(\mathbb{E} \min_{0 \leq t \leq T-1} L_\infty^{t, \xi^{t+1}} \right)^{\frac{1}{2}} \right).
\end{aligned}$$

2646
2647 Expressing the number of iterations and using $\varepsilon \geq \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{i=0}^{T-1} \gamma^i} \|\nabla f(x^t)\|_1 \right]$ as a criterion, we
2648
2649 obtain that the algorithm needs $\tilde{\mathcal{O}} \left(\frac{\Delta^* (L_\infty)^3}{\varepsilon^2} \left(\mathbb{E} \left(\frac{1}{L_\infty^{0,\xi^1}} \right)^2 \right) + \|\sigma\|_1^2 L_\infty \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t,\xi^{t+1}}} \right) \right)$ it-
2650
2651 iterations to reach ε -accuracy. \square
2652

2653
2654 *Remark F.7 (Remark 3.10).* Under conditions of Theorem 3.9 Algorithm 2 with $\lambda^t =$
2655 $\frac{1}{\sqrt{L_\infty + \sum_{i=0}^{t-1} \frac{\|g_{\xi^{i+1}}^{t+1} - g_{\xi^i}^t\|_1}{\|x^{i+1} - x^i\|_\infty}}}$, Option II and mini-batch of the size $t+1$ at t -th iteration to reach ε -accuracy
2656 needs
2657

$$\tilde{\mathcal{O}} \left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \frac{\|\sigma\|_1^2 L_\infty}{\varepsilon^2} \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t,\xi^{t+1}}} \right) \right) \text{ iterations,}$$

2663 where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, $L_\infty^{t,\xi^{t+1}} = \frac{\|g_{\xi^{t+1}}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty}$.
2664
2665

2666 *Proof.* The proof of the remark repeats the proof of Theorem 3.9 except for the estimate on
2667 $\sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^{t,\xi^{t+1}}$ term and $\mathbb{E} \|\nabla f(x^t) - g_{\xi^t}^t\|_1^2$ term. Let us derive them. We use definition
2668
2669 $L_\infty^{t,\xi^{t+1}} = \frac{\|g_{\xi^{t+1}}^{t+1} - g_{\xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty}$.
2670
2671

$$\sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^{t,\xi^{t+1}} = \sum_{t=0}^{T-1} \frac{L_\infty^{t,\xi^{t+1}}}{L_\infty + \sum_{i=0}^{t-1} L_\infty^{i,\xi^{i+1}}} \leq \sum_{t=0}^{T-1} \frac{L_\infty^{t,\xi^{t+1}}}{\sum_{i=0}^t L_\infty^{i,\xi^{i+1}}}.$$

2677 Continuing analogically to equation 56 - equation 57, we get
2678

$$\sum_{t=0}^{T-1} (\lambda^t)^2 L_\infty^t \leq 2 \log \left(\frac{L_\infty T}{L_\infty^{0,\xi^1}} \right).$$

2684 We substitute this bound into equation 61 instead of equation 58. Next, since we now use mini-batches,
2685 we can bound

$$\begin{aligned} \mathbb{E} \|\nabla f(x^t) - g_{\xi^t}^t\|_1^2 &\leq \frac{\|\sigma\|_1^2}{t+1}, \\ \mathbb{E} \|\nabla f(x^{t+1}) - g_{\xi^{t+1}}^{t+1}\|_1^2 &\leq \frac{\|\sigma\|_1^2}{t+2}, \\ \mathbb{E} \|\nabla f(x^t) - g_{\xi^{t+1}}^t\|_1^2 &\leq \frac{\|\sigma\|_1^2}{t+1}, \end{aligned}$$

2693 instead of equation 60. In that way,
2694

$$\sum_{t=0}^{T-1} \left(\mathbb{E} \|\nabla f(x^t) - g_{\xi^t}^t\|_1^2 \right)^{\frac{1}{2}} \cdot \left(\mathbb{E} \left[\frac{\gamma^t}{\sum_{i=0}^{T-1} \gamma^i} \right]^2 \right)^{\frac{1}{2}}$$

$$\begin{aligned}
& \leq \frac{\sqrt{L_\infty} \|\sigma\|_1}{\sqrt{T}} \sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}} \left(\mathbb{E} \frac{1}{\sum_{i=0}^{t-1} L_\infty^{i, \xi^{i+1}}} \right)^{\frac{1}{2}} \\
& \leq \frac{\sqrt{L_\infty} \|\sigma\|_1}{\sqrt{T}} \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t, \xi^{t+1}}} \right)^{\frac{1}{2}} \sum_{t=0}^{T-1} \frac{1}{t+1} \\
& \leq 2 \frac{\sqrt{L_\infty} \|\sigma\|_1}{\sqrt{T}} \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t, \xi^{t+1}}} \right)^{\frac{1}{2}} \log(T),
\end{aligned}$$

which ends the proof of the remark. \square

F.3 DISTRIBUTED SETTING

We remind, that in distributed setting we consider Assumption E.6. We present the theoretical result with the following approximation of L_∞ in Algorithm 2:

$$\lambda^t = \frac{1}{\sqrt{\sum_{i=0}^{t-1} \frac{1}{M} \sum_{j=1}^M \frac{\|g_{j, \xi^{i+1}}^{i+1} - g_{j, \xi^{i+1}}^i\|_1}{\|x^{i+1} - x^i\|_\infty}}}.$$

In this section, we denote g_{j, ξ^t}^t the stochastic gradient from the j -th device, computed at the t -th iteration, according to the stochastic realization ξ^t .

Lemma F.8 (Descent lemma). *For Algorithm 2 under Assumptions 3.8, 3.2, 3.3, E.6, the following estimate is valid:*

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E} [\gamma^t \|\nabla f(x^t)\|_1] & \leq \Delta^* \mathbb{E} \left[\frac{1}{\sum_{t=0}^{T-1} \gamma^t} \right] + 2 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \tilde{\delta}^t}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
& + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_{j, \xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
& + \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_{j, \xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\
& + \mathbb{E} \left[\frac{\sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^{t, \xi^{t+1}}}{\sum_{t=0}^{T-1} \gamma^t} \right],
\end{aligned}$$

where $\tilde{\delta}^t = \sum_{i=1}^d |\nabla f(x^t)|_i \mathbb{I} \left(\text{sign} \left(\sum_{j=1}^M \text{sign} \left([g_{j, \xi^t}^t]_i \right) \right) \neq \text{sign} ([\nabla f(x^t)]_i) \right)$

and $L_\infty^{t, \xi^t} = \frac{1}{M} \sum_{j=1}^M \frac{\|g_{j, \xi^t}^{t+1} - g_{j, \xi^t}^t\|_1}{\|x^{t+1} - x^t\|_\infty}$.

2754 *Proof.*

$$\begin{aligned}
2756 \quad f(x^{t+1}) - f(x^t) &\leq \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle \\
2757 &= -\gamma^t \left\langle \nabla f(x^{t+1}), \text{sign} \left(\sum_{j=1}^M \text{sign}(g_{j,\xi^t}^t) \right) \right\rangle \\
2758 &= -\gamma^t \left\langle \nabla f(x^t), \text{sign} \left(\sum_{j=1}^M \text{sign}(g_{j,\xi^t}^t) \right) \right\rangle \\
2759 &\quad -\gamma^t \left\langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign} \left(\sum_{j=1}^M \text{sign}(g_{j,\xi^t}^t) \right) \right\rangle \\
2760 &= -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \sum_{i=1}^d |[\nabla f(x^t)]_i| \\
2761 &\quad \cdot \mathbb{I} \left(\text{sign} \left(\sum_{j=1}^M \text{sign}([g_{j,\xi^t}^t]_i) \right) \neq \text{sign}([\nabla f(x^t)]_i) \right) \\
2762 &\quad -\gamma^t \left\langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign} \left(\sum_{j=1}^M \text{sign}(g_{j,\xi^t}^t) \right) \right\rangle \\
2763 &\stackrel{Conj, (i)}{\leq} -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t \\
2764 &\quad + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \left\| \text{sign} \left(\sum_{j=1}^M \text{sign}(g_{j,\xi^t}^t) \right) \right\|_\infty \\
2765 &= -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t \\
2766 &\quad + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \left\| \text{sign} \left(\sum_{j=1}^M \text{sign}(g_{j,\xi^t}^t) \right) \right\|_\infty \\
2767 &\stackrel{CS}{\leq} -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_{j,\xi^{t+1}}^t\|_1 \\
2768 &\quad + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_{j,\xi^{t+1}}^{t+1}\|_1 \\
2769 &\quad + \gamma^t \frac{1}{M} \sum_{j=1}^M \|g_{j,\xi^{t+1}}^{t+1} - g_{j,\xi^{t+1}}^t\|_1 \left\| \text{sign} \left(\sum_{j=1}^M \text{sign}(g_{j,\xi^t}^t) \right) \right\|_\infty \\
2770 &\stackrel{(ii)}{=} -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_{j,\xi^{t+1}}^t\|_1 \\
2771 &\quad + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_{j,\xi^{t+1}}^{t+1}\|_1 \\
2772 &\quad + \gamma^t \frac{1}{M} \sum_{j=1}^M \frac{\|g_{j,\xi^{t+1}}^{t+1} - g_{j,\xi^{t+1}}^t\|_1}{\|x^{t+1} - x^t\|_\infty} \|x^{t+1} - x^t\|_\infty \\
2773 &= -\gamma^t \|\nabla f(x^t)\|_1 + 2\gamma^t \tilde{\delta}^t + \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_{j,\xi^{t+1}}^t\|_1
\end{aligned}$$

2808
 2809 $+\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_{j,\xi^{t+1}}^{t+1}\|_1$
 2810
 2811 $+(\gamma^t)^2 \frac{1}{M} \sum_{j=1}^M \frac{\|g_{j,\xi^{t+1}}^{t+1} - g_{j,\xi^{t+1}}^t\|_1}{\|x^{t+1} - x^t\|_\infty},$
 2812
 2813
 2814
 2815 where in (i) we denote $\tilde{\delta}^t = \sum_{i=1}^d |\nabla f(x^t)|_i \mathbb{I} \left(\text{sign} \left(\sum_{j=1}^M \text{sign} \left(\left[g_{j,\xi^t}^t \right]_i \right) \right) \neq \text{sign} ([\nabla f(x^t)]_i) \right)$
 2816
 2817 and in (ii) we assume $\|x^{t+1} - x^t\|_\infty \neq 0$ (analogically to Lemma F.5). Defining $L_\infty^{t,\xi^{t+1}} =$
 2818
 2819 $\frac{1}{M} \sum_{j=1}^M \frac{\|g_{j,\xi^{t+1}}^{t+1} - g_{j,\xi^{t+1}}^t\|_1}{\|x^{t+1} - x^t\|_\infty}$ and summing over all iterations gives us
 2820
 2821
 2822 $\sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 \leq \Delta^* + 2 \sum_{t=0}^{T-1} \gamma^t \tilde{\delta}^t + \sum_{t=0}^{T-1} \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_{j,\xi^{t+1}}^t\|_1$
 2823
 2824 $+ \sum_{t=0}^{T-1} \gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_{j,\xi^{t+1}}^{t+1}\|_1 + \sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^{t,\xi^t},$
 2825
 2826
 2827
 2828 $\sum_{t=0}^{T-1} \frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \leq \frac{\Delta^*}{\sum_{t=0}^{T-1} \gamma^t} + 2 \sum_{t=0}^{T-1} \frac{\gamma^t \tilde{\delta}^t}{\sum_{t=0}^{T-1} \gamma^t} + \sum_{t=0}^{T-1} \frac{\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_{j,\xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t}$
 2829
 2830 $+ \sum_{t=0}^{T-1} \frac{\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_{j,\xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} + \sum_{t=0}^{T-1} \frac{(\gamma^t)^2 L_\infty^{t,\xi^t}}{\sum_{t=0}^{T-1} \gamma^t}.$
 2831
 2832
 2833
 2834
 2835
 2836
 2837

2838 Taking expectation, we derive the result of the lemma:
 2839

2840
 2841 $\sum_{t=0}^{T-1} \mathbb{E} [\gamma^t \|\nabla f(x^t)\|_1] \leq \Delta^* \mathbb{E} \left[\frac{1}{\sum_{t=0}^{T-1} \gamma^t} \right] + 2 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \tilde{\delta}^t}{\sum_{t=0}^{T-1} \gamma^t} \right]$
 2842
 2843
 2844
 2845 $+ \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_{j,\xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right]$
 2846
 2847
 2848
 2849
 2850 $+ \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_{j,\xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right]$
 2851
 2852
 2853
 2854
 2855
 2856
 2857
 2858
 2859
 2860
 2861

□

2862 **Theorem F.9.** Suppose Assumptions 3.8, 3.2, 3.3, E.6 hold. Then Algorithm 2 with Option II to reach
 2863

2864 ε -accuracy, where $\varepsilon \geq \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \right]$ needs
 2865

2866 $\tilde{\mathcal{O}} \left(\frac{\Delta^* (L_\infty)^3}{\varepsilon^2} \left(\mathbb{E} \left(\frac{1}{L_\infty^{0,\xi^1}} \right)^2 \right) + \|\sigma\|_1^2 L_\infty \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t,\xi^{t+1}}} \right) \right)$ iterations,
 2867

2868 where $L_\infty^{t,\xi^{t+1}} = \frac{1}{M} \sum_{j=1}^M \frac{\|g_{j,\xi^{t+1}}^{t+1} - g_{j,\xi^{t+1}}^t\|_1}{\|x^{t+1} - x^t\|_\infty}$.
 2869

2870 *Proof.* Let us start with the result of Lemma F.8:

2871
$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E} [\gamma^t \|\nabla f(x^t)\|_1] &\leq \Delta^* \mathbb{E} \left[\frac{1}{\sum_{t=0}^{T-1} \gamma^t} \right] + 2 \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \tilde{\delta}^t}{\sum_{t=0}^{T-1} \gamma^t} \right] \\ &+ \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^t) - g_{j,\xi^{t+1}}^t\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\ &+ \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \frac{1}{M} \sum_{j=1}^M \|\nabla f(x^{t+1}) - g_{j,\xi^{t+1}}^{t+1}\|_1}{\sum_{t=0}^{T-1} \gamma^t} \right] \\ &+ \mathbb{E} \left[\frac{\sum_{t=0}^{T-1} (\gamma^t)^2 L_\infty^{t,\xi^{t+1}}}{\sum_{t=0}^{T-1} \gamma^t} \right]. \end{aligned}$$

2872 Note that we have already estimated all terms in Theorem F.6 except $\sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \tilde{\delta}^t}{\sum_{t=0}^{T-1} \gamma^t} \right]$. However,
 2873 using Lemma E.7 together with equation Höl, we can do the same thing and obtain
 2874

2875
$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t \tilde{\delta}^t}{\sum_{t=0}^{T-1} \gamma^t} \right] &\leq \sum_{t=0}^{T-1} \left(\mathbb{E} [\tilde{\delta}]^2 \right)^{\frac{1}{2}} \left(\mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \right]^2 \right)^{\frac{1}{2}} \\ &\leq 2\sqrt{L_\infty} \|\sigma\|_1 \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t,\xi^{t+1}}} \right)^{\frac{1}{2}}. \end{aligned}$$

2876 In that way, we get the same estimate as in Theorem F.6:
 2877

2878
$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \right] &\leq 13 \frac{\sqrt{(f(x^0) - \tilde{f})(L_\infty)^{\frac{3}{2}}}}{T} \left(\mathbb{E} \left(\frac{1}{L_\infty^{0,\xi^1}} \right)^2 \right)^{\frac{1}{2}} \\ &\quad \cdot \left(\mathbb{E} \log^2 \left(\frac{L_\infty T}{L_\infty^{0,\xi^1}} \right) \right)^{\frac{1}{2}} \end{aligned}$$

$$+8\|\sigma\|_1 \left(\sqrt{L_\infty} \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t,\xi^{t+1}}} \right)^{\frac{1}{2}} \right).$$

Expressing the number of iterations and using $\varepsilon \geq \sum_{t=0}^{T-1} \mathbb{E} \left[\frac{\gamma^t}{\sum_{t=0}^{T-1} \gamma^t} \|\nabla f(x^t)\|_1 \right]$ as a criterion, we

obtain that the algorithm needs $\tilde{\mathcal{O}} \left(\frac{\Delta^*(L_\infty)^3}{\varepsilon^2} \left(\mathbb{E} \left(\frac{1}{L_\infty^{0,\xi^1}} \right)^2 \right) + \|\sigma\|_1^2 L_\infty \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t,\xi^{t+1}}} \right) \right)$ iterations to reach ε -accuracy. \square

Remark F.10. Under conditions of Theorem F.9 Algorithm 2 with $\lambda^t = \frac{1}{\sqrt{L_\infty + \sum_{i=0}^{t-1} \frac{1}{M} \sum_{j=1}^M \frac{\|g_{j,\xi^{i+1}} - g_{j,\xi^i}\|_1}{\|x^{i+1} - x^i\|_\infty}}}$, Option II and mini-batch of the size $t+1$ at t -th iteration to reach ε -accuracy needs

$$\tilde{\mathcal{O}} \left(\frac{\Delta^* L_\infty}{\varepsilon^2} + \frac{\|\sigma\|_1^2 L_\infty}{\varepsilon^2} \left(\mathbb{E} \frac{1}{\min_{0 \leq t \leq T-1} L_\infty^{t,\xi^{t+1}}} \right) \right) \text{ iterations,}$$

where $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, $L_\infty^{t,\xi^{t+1}} = \frac{1}{M} \sum_{j=1}^M \frac{\|g_{j,\xi^{t+1}} - g_{j,\xi^t}\|_1}{\|x^{t+1} - x^t\|_\infty}$.

Proof. Proof repeats the proof of Remark 3.10. \square

F.4 MEMORY-EFFICIENT ALIAS

Lemma F.11 (Descent lemma). *For Algorithm 2 under Assumptions 3.11, 3.2, 3.3, 3.4, the following estimate is valid:*

$$\sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 \leq \Delta^* + \sum_{t=0}^{T-1} (\gamma^t)^2 d^2 L_1^t,$$

where $L_1^t = \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_\infty}{\|x^{t+1} - x^t\|_1}$.

Proof.

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle = f(x^t) - \gamma^t \langle \nabla f(x^{t+1}), \text{sign}(\nabla f(x^t)) \rangle \\ &= f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 - \gamma^t \langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign}(\nabla f(x^t)) \rangle \\ &\stackrel{\text{Conj}}{\leq} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_\infty \|\text{sign}(\nabla f(x^t))\|_1 \\ &\leq f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t d \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_\infty \\ &\stackrel{(i)}{=} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t d \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_\infty}{\|x^{t+1} - x^t\|_1} \|x^{t+1} - x^t\|_1 \\ &= f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + (\gamma^t)^2 d^2 \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_\infty}{\|x^{t+1} - x^t\|_1}, \end{aligned}$$

where in (i) we assume $\|x^{t+1} - x^t\|_1 \neq 0$. Indeed, $\|x^{t+1} - x^t\|_1 = 0$ follows from the equality $\text{sign}(\nabla f(x^t)) = 0$, which means that we find the optimum and do not need to find another point x^{t+1} .

Now we denote $L_1^t = \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_\infty}{\|x^{t+1} - x^t\|_1}$. Summing over all iterations, we obtain

$$\sum_{t=0}^{T-1} \gamma^t \|f(x^t)\|_1 \leq \sum_{t=0}^{T-1} [f(x^t) - f(x^{t+1})] + \sum_{t=0}^{T-1} (\gamma^t)^2 d^2 L_1^t$$

$$= f(x^0) - f(x^*) + \sum_{t=0}^{T-1} (\gamma^t)^2 d^2 L_1^t \leq \Delta^* + \sum_{t=0}^{T-1} (\gamma^t)^2 d^2 L_\infty^t,$$

which ends the proof of the lemma. \square

Theorem F.12 (Theorem 3.12). Suppose Assumptions 3.11, 3.2, 3.3, 3.4 hold. We denote $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$, $L_1^0 = \frac{\|\nabla f(x^1) - \nabla f(x^0)\|_\infty}{\|x^1 - x^0\|_1}$. Then Algorithm 2 with $d^0 < \Delta^*$ and $d \cdot \lambda^t$ as in equation 3 to reach ε -accuracy needs

$$\tilde{\mathcal{O}}\left(\frac{(\Delta^*)^2 (L_1)^3 d^2}{d^0 (L_1^0)^2 \varepsilon^2}\right) \text{ and } \tilde{\mathcal{O}}\left(\frac{\Delta^* (L_1)^3 d^2}{(L_1^0)^2 \varepsilon^2}\right) \text{ iterations with Options I and II, respectively.}$$

Proof. Let us start with the result of Lemma F.11:

$$\sum_{t=0}^{T-1} \gamma^t \|\nabla f(x^t)\|_1 \leq \Delta^* + \sum_{t=0}^{T-1} (\gamma^t)^2 d^2 L_1^t. \quad (62)$$

Now we use our γ^t choice. Let us firstly estimate the denominator that is exactly $\lambda^t = \frac{1}{d \sqrt{\sum_{i=0}^{t-1} \frac{\|\nabla f(x^{i+1}) - \nabla f(x^i)\|_\infty}{\|x^{i+1} - x^i\|_1}}} = \frac{1}{d \sqrt{\sum_{i=0}^{t-1} L_1^i}}$ and is the same for both Options I and II. Let us estimate the following term.

$$\sum_{t=0}^{T-1} (\lambda^t)^2 d^2 L_1^t = \sum_{t=0}^{T-1} \frac{L_1^t}{\sum_{i=0}^{t-1} L_1^i}.$$

We mention, that each L_1^i is bounded from the definition of smoothness (see Assumption 3.11), i.e., $L_1^i \leq L_1$. We consider the sequence $\{L_1^i\}_{i=0}^{T-1}$. Since each term in this sequence is bounded, there exists r such that $\sum_{i=0}^{r-2} L_1^i \leq L_1^{r-1}$ and for each $t \geq r-1$ such that $\sum_{i=0}^t L_1^i \geq L_1^{t+1}$. In that way, we divide the sum into two parts:

$$\sum_{t=0}^{T-1} \frac{L_1^t}{\sum_{i=0}^{t-1} L_1^i} = \sum_{t=0}^{r-1} \frac{L_1^t}{\sum_{i=0}^{t-1} L_1^i} + \sum_{t=r}^{T-1} \frac{L_1^t}{\sum_{i=0}^{t-1} L_1^i}. \quad (63)$$

Considering the first sum in equation 63, we mention, that we can estimate the denominator as $\sum_{i=0}^{t-1} L_1^i \geq L_1^0$. As for the numerator. Thus,

$$\sum_{t=0}^{r-1} \frac{L_1^t}{\sum_{i=0}^{t-1} L_1^i} \leq \frac{1}{L_1^0} \left(\sum_{t=0}^{r-2} L_1^t + L_1^{r-1} \right) \leq \frac{2L_1^{r-1}}{L_1^0} \leq \frac{2L_1}{L_1^0}. \quad (64)$$

Considering the second sum in equation 63, we have

$$\sum_{t=r}^{T-1} \frac{L_1^t}{\sum_{i=0}^{t-1} L_1^i} = \sum_{t=r}^{T-1} \frac{L_1^t}{\frac{1}{2} \sum_{i=0}^{t-1} L_1^i + \frac{1}{2} \sum_{i=0}^{t-1} L_1^i}.$$

3024
 3025 Estimating any of the sums in the denominator, we claim, that $\sum_{i=0}^{t-1} L_1^i \geq L_1^t$, since $t-1 \geq r-1$. In
 3026 that way,
 3027

3028
 3029
$$\sum_{t=r}^{T-1} \frac{L_1^t}{\sum_{i=0}^{t-1} L_1^i} \leq \sum_{t=r}^{T-1} \frac{2L_1^t}{\sum_{i=0}^t L_1^i} \leq 2 \sum_{t=0}^{T-1} \frac{L_1^t}{\sum_{i=0}^t L_1^i}. \quad (65)$$

 3030
 3031
 3032

3033
 3034 Next we denote $s^t = \sum_{i=0}^t L_1^i$ and have
 3035

3036
 3037
$$L_1^t \frac{1}{\sum_{i=0}^t L_1^i} = (s^t - s^{t-1}) \frac{1}{\sum_{i=0}^t L_1^i} = \int_{s^{t-1}}^{s^t} \frac{1}{\sum_{i=0}^t L_1^i} dx \stackrel{(i)}{\leq} \int_{s^{t-1}}^{s^t} \frac{1}{x} dx, \quad (66)$$

 3038
 3039
 3040

3041 where (i) was done due to $\frac{1}{x}$ is a non-increasing function on $(0, +\infty)$. Summing over t , we obtain
 3042

3043
 3044
$$2 \sum_{t=1}^T \frac{L_1^t}{\sum_{i=0}^t L_1^i} \leq 2 \int_{s^0}^{s^T} \frac{1}{x} dx = 2 \log(s^T) - 2 \log(s^0) = 2 \log \left(\frac{\sum_{t=1}^T L_1^t}{L_1^0} \right) \leq 2 \log \left(\frac{L_1 T}{L_1^0} \right).$$

 3045
 3046
 3047

3048 Combining this estimate with equation 65,

3049
 3050
 3051
$$\sum_{t=r}^{T-1} \frac{L_1^t}{\sum_{i=0}^{t-1} L_1^i} \leq 2 \sum_{t=1}^T \frac{L_1^t}{\sum_{i=0}^t L_1^i} + 2 \leq 2 \left(\log \left(\frac{L_1 T}{L_1^0} \right) + 1 \right) \leq 4 \log \left(\frac{L_1 T}{L_1^0} \right). \quad (67)$$

 3052
 3053
 3054

3055 Substituting equation 64 and equation 67 into equation 63, we obtain
 3056

3057
 3058
$$\sum_{t=0}^{T-1} (\lambda^t)^2 d^2 L_1^t \leq 2 \frac{L_1}{L_1^0} + 4 \log \left(\frac{L_1 T}{L_1^0} \right). \quad (68)$$

 3059
 3060

3061 We additionally note, that if $r > T - 1$, only first term remains in this estimate, consequently our
 3062 bound equation 68 is correct.
 3063

3064 In this way, utilizing Option I from Algorithm 2, equation 62 together with equation 68 yields

3065
 3066
$$\sqrt{d^0} \lambda^{T-1} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \stackrel{(i)}{\leq} \sum_{t=0}^{T-1} \sqrt{d^t} \lambda^t \|\nabla f(x^t)\|_1 \leq \Delta^* + \sum_{t=0}^{T-1} d^t (\lambda^t)^2 d^2 L_1^t$$

 3067
 3068
 3069
$$\stackrel{\text{Lemma F.1}}{\leq} \Delta^* + \Delta^* \sum_{t=0}^{T-1} (\lambda^t)^2 d^2 L_1^t,$$

 3070
 3071
 3072
$$\sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{\Delta^*}{\sqrt{d^0} \lambda^{T-1}} + \frac{\Delta^*}{\sqrt{d^0} \lambda^{T-1}} \sum_{t=0}^{T-1} (\lambda^t)^2 d^2 L_1^t$$

 3073
 3074
 3075
$$\stackrel{68}{\leq} \frac{\Delta^*}{\sqrt{d^0} \lambda^{T-1}} + 4 \frac{\Delta^*}{\sqrt{d^0} \lambda^{T-1}} \log \left(\frac{L_1 T}{L_1^0} \right) + 2 \frac{\Delta^* L_1}{\sqrt{d^0} \lambda^{T-1} L_1^0}$$

 3076
 3077
$$\leq 7 \frac{\Delta^* L_1}{\sqrt{d^0} \lambda^{T-1} L_1^0} \log \left(\frac{L_1 T}{L_1^0} \right), \quad (69)$$

3078 where (i) was done due to the fact that d^0 is minimal from all $\{d^t\}_{t=0}^{T-1}$ (Line 7 from Algorithm 2)

3079
3080 and the definition of λ^t . Utilizing $\frac{1}{\lambda^{T-1}} = d \sqrt{\sum_{t=0}^{T-2} L_1^t} \leq d\sqrt{L_1 T}$, we obtain the final estimate:

$$3082 \quad 3083 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{7\Delta^*(L_1)^{\frac{3}{2}} d}{\sqrt{d^0 T L_1^0}} \log\left(\frac{L_1 T}{L_1^0}\right).$$

3084
3085 Expressing the number of iterations and using $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$ as a criterion, we obtain that

3086 the algorithm needs $\tilde{\mathcal{O}}\left(\frac{(\Delta^*)^2 (L_1)^3 d^2}{d^0 (L_1^0)^2 \varepsilon^2}\right)$ iterations to reach ε -accuracy.

3087
3088 Considering Option II from Algorithm 2, we can proceed absolutely analogical, however, using
3089 $f(x^0) - \tilde{f} \geq \Delta^*$ instead of Lemma F.1. In that way,

$$3090 \quad 3091 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1 \leq \frac{\Delta^* \sqrt{L_1 d}}{\sqrt{(f(x^0) - \tilde{f}) T}} + \frac{4(f(x^0) - \tilde{f}) \sqrt{L_1 d}}{\sqrt{(f(x^0) - \tilde{f}) T}} \log\left(\frac{L_1 T}{L_1^0}\right)$$

$$3092 \quad 3093 \quad + \frac{2(f(x^0) - \tilde{f}) (L_1)^{\frac{3}{2}} d}{\sqrt{(f(x^0) - \tilde{f}) T L_1^0}}$$

$$3094 \quad 3095 \quad \leq \frac{7\sqrt{(f(x^0) - \tilde{f})} (L_1)^{\frac{3}{2}} d}{\sqrt{T L_1^0}} \log\left(\frac{L_1 T}{L_1^0}\right).$$

3096
3097 Expressing the number of iterations, using $\varepsilon \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1$ as a criterion, and utilizing \tilde{f} is
3098 an approximation of $f(x^*)$, we obtain that the algorithm needs $\tilde{\mathcal{O}}\left(\frac{\Delta^* (L_1)^3 d^2}{(L_1^0)^2 \varepsilon^2}\right)$ iterations to reach
3099 ε -accuracy. \square

3100
3101 The proofs under stochastic and distributed settings for the memory-efficient version of ALIAS can
3102 be obtained analogously to Theorems F.6, F.9, and F.12.

3111 G STEEPEST DESCENT

3112 There is one more approach for sign descent. Classically, we perform the step in the direction of the
3113 gradient. However, we do not take into account the length of the gradient in any way in the step. The
3114 approach, called steepest descent, is supposed to utilize this information and provide the steps in the
3115 direction $\|\nabla f(x^t)\|_1 \text{sign}(\nabla f(x^t))$ at the t -th iteration. We provide the formal description of this
3116 approach (Algorithm 9).

3117 Algorithm 8 STEEPEST DESCENT

3118 1: **Input:** Initial point $x^0 \in \mathbb{R}^d$, number of iterations T
 3119 2: **Parameter:** Stepsize $c > 0$
 3120 3: **for** $t = 0, \dots, T - 1$ **do**
 3121 4: $x^{t+1} = x^t - c \|\nabla f(x^t)\|_1 \text{sign}(\nabla f(x^t))$
 3122 5: **end for**

3123 We present the analysis of SOS STEEPEST DESCENT. We start with the descent lemma.

3124 **Lemma G.1** (Descent lemma). *For Algorithm 9 under Assumptions 3.1, 3.2, 3.3, 3.4, the following
3125 estimate is valid:*

$$3126 \quad 3127 \quad -\Delta^* \leq -c_0 \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2 \left(1 - c_0 \tilde{L}_\infty\right),$$

3132 **Algorithm 9** SOS STEEPEST DESCENT

3133 1: **Input:** Initial stepsize bound c_s , initial bound step k , initial point $x^0 \in \mathbb{R}^d$, number of iterations
 3134 T
 3135 2: $c_0 = \text{BISECTION} \left(\phi(c), \frac{c_s}{2^{2k}}, c_s, T \right)$ in Algorithm 4 we utilize Algorithm 8
 3136 instead of Algorithm 1
 3137 3: $x^T = \text{STEEPEST DESCENT}(x^0, T, c_0)$

3139
 3140 where $\tilde{L}_\infty = \max_{0 \leq t \leq T-1} L_\infty^t$ and $L_\infty^t = \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty}$.
 3141
 3142

3143 *Proof.* Starting from the convexity of the objective,

$$\begin{aligned}
 3144 \quad f(x^{t+1}) &\leq f(x^t) + \langle \nabla f(x^{t+1}), x^{t+1} - x^t \rangle = f(x^t) - \gamma^t \langle \nabla f(x^{t+1}), \text{sign}(\nabla f(x^t)) \rangle \\
 3145 &= f(x^t) - \gamma^t \langle \nabla f(x^t), \text{sign}(\nabla f(x^t)) \rangle \\
 3146 &\quad - \gamma^t \langle \nabla f(x^{t+1}) - \nabla f(x^t), \text{sign}(\nabla f(x^t)) \rangle \\
 3147 &\stackrel{\text{Conj}}{\leq} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \|\text{sign}(\nabla f(x^t))\|_\infty \\
 3148 &\leq f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1 \\
 3149 &\stackrel{(i)}{=} f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + \gamma^t \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty} \|x^{t+1} - x^t\|_\infty,
 \end{aligned}$$

3150 where in (i) we assume $\|x^{t+1} - x^t\|_\infty \neq 0$. Indeed, $\|x^{t+1} - x^t\|_\infty = 0$ follows from
 3151 $\text{sign}(\nabla f(x^t)) = 0$, which means we find the optimum and do need to search the point x^{t+1} .
 3152

3153 Now we denote $L_\infty^t = \frac{\|\nabla f(x^{t+1}) - \nabla f(x^t)\|_1}{\|x^{t+1} - x^t\|_\infty}$. Continue estimate,
 3154

$$\begin{aligned}
 3155 \quad f(x^{t+1}) &\leq f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + (\gamma^t)^2 L_\infty^t \|\text{sign}(\nabla f(x^t))\|_\infty \\
 3156 &\leq f(x^t) - \gamma^t \|\nabla f(x^t)\|_1 + (\gamma^t)^2 L_\infty^t.
 \end{aligned}$$

3157 Now we choose $\gamma^t = c_0 \|\nabla f(x^t)\|_1$, where we find the constant c_0 using BISECTION procedure
 3158 (Algorithm 4). Thus,

$$\begin{aligned}
 3159 \quad f(x^{t+1}) &\leq f(x^t) - c_0 \|\nabla f(x^t)\|_1^2 + c_0^2 \|\nabla f(x^t)\|_1^2 L_\infty^t \\
 3160 &= f(x^t) - c_0 \|\nabla f(x^t)\|_1^2 (1 - c_0 L_\infty^t).
 \end{aligned}$$

3161 Summing over all iterations and utilizing $\tilde{L}_\infty = \max_{0 \leq t \leq T-1} L_\infty^t$ notation, we have
 3162

$$-\Delta^* = f(x^*) - f(x^0) \leq f(x^T) - f(x^0) \leq -c_0 \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2 (1 - c_0 \tilde{L}_\infty),$$

3163 which ends the proof of the lemma. \square

3164 Now we present the purposes of Algorithm 4. Let us take an arbitrary point $x^{-1} \in \mathbb{R}^d$. We denote
 3165 $L_\infty^{-1} = \frac{\|\nabla f(x^0) - \nabla f(x^{-1})\|_1}{\|x^0 - x^{-1}\|_\infty}$ and $\tilde{L}_\infty^{-1} = \max_{-1 \leq t \leq T-1} L_\infty^t$. It is obvious that it implies
 3166

$$\begin{aligned}
 3167 \quad L_\infty^{-1} &\leq \tilde{L}_\infty^{-1} \leq L_\infty, \\
 3168 \quad \tilde{L}_\infty &\leq \tilde{L}_\infty^{-1}.
 \end{aligned} \tag{70}$$

3169 Let us put $\phi(c) = \frac{1}{\tilde{L}_\infty^{-1}(c)}$ in the BISECTION procedure. The following lemma shows guarantees of
 3170 $\phi(c_{hi}) \leq c_{hi}$ and $\phi(c_{lo}) \geq c_{lo}$.
 3171

3172 **Lemma G.2** (Bisection entry). *Let $c_{\max} = \frac{1}{\tilde{L}_\infty^{-1}}$. Thus, with the initial $c_{hi} = c_{\max}$, Algorithm
 3173 4 always avoids an early infinite termination. Moreover, with the initial $c_{lo} = \frac{1}{2^{2k}} c_{hi}$, where
 3174 $k \geq \log \log \frac{L_\infty}{L_\infty^{-1}}$, Algorithm 4 always avoids early non-infinite termination.*

3186 *Proof.* Let us start with c_{hi} . The choice of c_{max} implies
 3187

$$3188 \quad c_{\text{hi}} = c_{\text{max}} = \frac{1}{L_{\infty}^{-1}} \stackrel{70}{\geq} \frac{1}{\tilde{L}_{\infty}^{-1}(c_{\text{hi}})} = \phi(c_{\text{hi}}),$$

3190 which means we avoid early infinite termination. As for c_{lo} :
 3191

$$3192 \quad c_{\text{lo}} = \frac{1}{2^{2k}} c_{\text{hi}} \leq \frac{1}{\frac{L_{\infty}^{-1}}{L_{\infty}^{-1}}} \cdot \frac{1}{L_{\infty}^{-1}} = \frac{1}{L_{\infty}} \stackrel{70}{\leq} \frac{1}{\tilde{L}_{\infty}^{-1}(c_{\text{lo}})} = \phi(c_{\text{lo}}),$$

3195 which means we avoid early non-infinite termination. \square
 3196

3197 Since we always entry to the BISECTION procedure, we are under the performing of Lemma D.3.
 3198 Now we are ready to prove the final convergence guarantees for SOS STEEPEST DESCENT.

3199 **Theorem G.3.** *Suppose Assumptions 3.1, 3.2, 3.3, 3.4 hold. Then for Algorithm 9 after obtaining the
 3200 stepsize c_0 , the following estimate is valid:*

$$3201 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2 \leq 8 \frac{\Delta^* L_{\infty}}{T}.$$

3205 *Moreover, taking into account the complexity of Algorithm 4 in relation to the initial stepsize bound
 3206 c_s , to reach ε -accuracy, where $\varepsilon^2 \geq \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2$, Algorithm 9 needs*

$$3208 \quad \mathcal{O}\left(\frac{\Delta^* L_{\infty}}{\varepsilon^2} \log \log \frac{L_{\infty}}{L_{\infty}^{-1}}\right) \text{ iterations.}$$

3211 *Proof.* Firstly, we recall the result of Lemma G.1:
 3212

$$3213 \quad -\Delta^* \leq -c_0 \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2 \left(1 - c_0 \tilde{L}_{\infty}\right).$$

3216 We have already mentioned that we can always avoid early terminations of Algorithm 4, due to
 3217 Lemma G.2, and thus, $\frac{1}{2\tilde{L}_{\infty}^{-1}(c_{\text{hi}}^*)} \leq c_0 \leq \frac{1}{\tilde{L}_{\infty}^{-1}(c_0)}$. Tuning $c_0 = \frac{c_0}{2}$, we obtain
 3218

$$3219 \quad -\Delta^* \leq -c_0 \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2 \left(1 - \frac{1}{2\tilde{L}_{\infty}^{-1}(c_0)} \tilde{L}_{\infty}(c_0)\right)$$

$$3222 \quad \stackrel{70}{\leq} -c_0 \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2 \left(1 - \frac{1}{2}\right).$$

3225 Expressing gradient norms, we obtain

$$3226 \quad \frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2 \leq \frac{2\Delta^*}{c_0 T} \leq \frac{8\Delta^* \tilde{L}_{\infty}^{-1}(c_{\text{hi}}^*)}{T} \stackrel{70}{\leq} \frac{8\Delta^* L_{\infty}}{T}.$$

3229 Assuming $\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(x^t)\|_1^2 \leq \varepsilon^2$ as a criterion, we easily obtain the estimate on the number of
 3230 iterations required — $\mathcal{O}\left(\frac{\Delta^* L_{\infty}}{\varepsilon^2}\right)$. Mention that the total number of iterations (together with the
 3232 Algorithm 4 performance) — $\mathcal{O}\left(\frac{\Delta^* L_{\infty}}{\varepsilon^2} \log \log \frac{L_{\infty}}{L_{\infty}^{-1}}\right)$. \square
 3233

3236 THE USE OF LARGE LANGUAGE MODELS (LLMs)

3237 In this work, large language models (LLMs) were used exclusively for spelling edits.
 3238