Remove that Square Root: A New Efficient Scale-Invariant Version of AdaGrad

Sayantan Choudhury[*]
MBZUAI & Johns Hopkins University

Nazarii Tupitsa
MBZUAI & Innopolis University

Nicolas Loizou
Johns Hopkins University

Samuel Horváth
MBZUAI

Martin Takáč
MBZUAI

Eduard Gorbunov
MBZUAI

Abstract

Adaptive methods are extremely popular in machine learning as they make learning rate tuning less expensive. This paper introduces a novel optimization algorithm named KATE, which presents a scale-invariant adaptation of the well-known AdaGrad algorithm. We prove the scale-invariance of KATE for the case of Generalized Linear Models. Moreover, for general smooth non-convex problems, we establish a convergence rate of $\mathcal{O}(\log T/\sqrt{T})$ for KATE, matching the best-known ones for AdaGrad and Adam. We also compare KATE to other state-of-the-art adaptive algorithms Adam and AdaGrad in numerical experiments with different problems, including complex machine learning tasks like image classification and text classification on real data. The results indicate that KATE consistently outperforms AdaGrad and matches/surpasses the performance of Adam in all considered scenarios.

1 Introduction

In this work, we consider the following unconstrained optimization problem:

$$\min_{w\in\mathbb{R}^d} f(w), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a L-smooth and generally non-convex function. In particular, we are interested in the situations when the objective has either expectation $f(w) = \mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(w)]$ or finite-sum $f(w) = \frac{1}{n}\sum_{i=1}^n f_i(w)$ form. Such minimization problems are crucial in machine learning, where w corresponds to the model parameters. Solving these problems with stochastic gradient-based optimizers has gained much interest owing to their wider applicability and low computational cost. Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951) and similar algorithms require the knowledge of parameters like L for convergence and are very sensitive to the choice of the stepsize in general. Therefore, SGD requires hyperparameter tuning, which can be computationally expensive. To address these issues, it is common practice to use adaptive variants of stochastic gradient-based methods that can converge without knowing the function's structure.

There exist many adaptive algorithms such as AdaGrad (Duchi et al., 2011), Adam (Kingma and Ba, 2014), AMSGrad (Reddi et al., 2019), D-Adaptation (Defazio and Mishchenko, 2023), Prodigy (Mishchenko and Defazio, 2023), AI-SARAH (Shi et al., 2023) and their variants. These adaptive techniques are capable of updating their step sizes on the fly. For instance, the AdaGrad method determines its step sizes using a cumulative sum of the coordinate-wise squared (stochastic) gradient

[*]Part of this work was done when S. Choudhury was an intern at MBZUAI, UAE.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

of all the previous iterates:

$$\text{AdaGrad: } w_{t+1} = w_t - \frac{\beta g_t}{\sqrt{\operatorname{diag}\left(\Delta I + \sum_{\tau=1}^{t} g_\tau g_\tau^\top\right)}}, \tag{2}$$

where g_t represents an unbiased estimator of $\nabla f(w_t)$, i.e., $\mathbb{E}[g_t \mid w_t] = \nabla f(w_t)$, $\operatorname{diag}(M) \in \mathbb{R}^d$ is a vector of diagonal elements of matrix $M \in \mathbb{R}^{d\times d}$, $\Delta > 0$, and the division by vector is done component-wise. Ward et al. (2020) has shown that this method achieves a convergence rate of $\mathcal{O}\left(\log T/\sqrt{T}\right)$ for smooth functions, similar to SGD, without prior knowledge of the functions' parameters. However, the performance of AdaGrad deteriorates when applied to data that may exhibit poor scaling or ill-conditioning. In this work, we propose a novel algorithm, KATE, to address the issues of poor data scaling. KATE is also a stochastic adaptive algorithm that can achieve a convergence rate of $\mathcal{O}\left(\log T/\sqrt{T}\right)$ for smooth non-convex functions in terms of $\min_{t\in[T]} \mathbb{E}\left[\|\nabla f(w_t)\|\right]^2$.

1.1 Related Work

A significant amount of research has been done on adaptive methods over the years, including AdaGrad (Duchi et al., 2011; McMahan and Streeter, 2010), AMSGrad (Reddi et al., 2019), RMSProp (Tieleman and Hinton, 2012), AI-SARAH (Shi et al., 2023), and Adam (Kingma and Ba, 2014). However, all these works assume that the optimization problem is contained in a bounded set. To address this issue, Li and Orabona (2019) proposes a variant of the AdaGrad algorithm, which does not use the gradient of the last iterate (this makes the step sizes of t-th iteration conditionally independent of g_t) for computing the step sizes and proves convergence for the unbounded domain.

Each of these works considers a vector of step sizes for each coefficient. Duchi et al. (2011) and McMahan and Streeter (2010) simultaneously proposed the original AdaGrad algorithm. However, McMahan and Streeter (2010) was the first to consider the vanilla scalar form of AdaGrad, known as

$$\text{AdaGradNorm: } w_{t+1} = w_t - \frac{\beta g_t}{\sqrt{\Delta + \sum_{\tau=0}^{t} \|g_\tau\|^2}}. \tag{3}$$

Later, Ward et al. (2020) analyzed AdaGradNorm for minimizing smooth non-convex functions. In a follow-up study, Xie et al. (2020) proves a linear convergence of AdaGradNorm for strongly convex functions. Recently, Liu et al. (2022) analyzed AdaGradNorm for solving smooth convex functions without the bounded domain assumption. Moreover, Liu et al. (2022) extends the convergence guarantees of AdaGradNorm to quasar-convex functions [2] using the function value gap. Orabona et al. (2015) introduce the notion of scale-invariance, which is a special case of affine invariance (Nesterov and Nemirovskii, 1994; Nesterov, 2018; d'Aspremont et al., 2018), propose a scale-invariant version of AdaGrad for online convex optimization for generalized linear models, and prove $\mathcal{O}(\sqrt{T})$ regret bounds in this setup.

Recently, Defazio and Mishchenko (2023) introduced the D-Adaptation method, which has gathered considerable attention due to its promising empirical performances. In order to choose the adaptive step size optimally, one requires knowledge of the initial distance from the solution, i.e., $D := \|w_0 - w_*\|$ where $w_* \in \operatorname{argmin}_{w\in\mathbb{R}^d} f(w)$. The D-Adaptation method works by maintaining an estimate of D and the stepsize choice in this case is $d_t/\sqrt{\sum_{\tau=0}^{t}\|g_\tau\|^2}$ for the t-th iteration (here d_t is an estimate of D). Mishchenko and Defazio (2023) further modifies the algorithm in a follow-up work and introduces Prodigy (with stepsize choice $d_t^2/\sqrt{\sum_{\tau=0}^{t} d_\tau^2\|g_\tau\|^2}$) to improve the convergence speed.

Another exciting line of work on adaptive methods is Polyak stepsizes. Polyak (1969) first proposed Polyak stepsizes for subgradient methods, and recently, the stochastic version (also known as SPS) was introduced by Oberman and Prazeres (2019); Loizou et al. (2021); Abdukhakimov et al. (2024, 2023); Li et al. (2023) and Gower et al. (2021). For a finite sum problem $\min_{w\in\mathbb{R}^d} f(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, Loizou et al. (2021) uses $\frac{f_i(w_t) - f_i^*}{c\|\nabla f_i(w_t)\|^2}$ as their stepsize choices (here $f_i^* := \min_{w\in\mathbb{R}^d} f_i(w)$), while Oberman and Prazeres (2019) uses $\frac{2(f(w_t)-f^*)}{\mathbb{E}[\|\nabla f_i(w_t)\|^2]}$ for k-th iteration. However, these methods are impractical when f^* or f_i^* is unknown. Following its introduction,

[2] f satisfy $f^* \geq f(w) + \frac{1}{\zeta}\langle f(w), w^* - w\rangle$ for some $\zeta \in (0,1]$ where $w^* \in \operatorname{argmin}_w f(w)$.

Table 1: Summary of convergence guarantees for closely-related adaptive algorithms to solve *smooth non-convex stochastic* optimization problems. Convergence rates are given in terms of $\min_{t\in[T]} \mathbb{E}\left[\|\nabla f(w_t)\|\right]^2$. We highlight KATE's *scale-invariance* property for problems of type (4).

Algorithm	Convergence rate	Scale invariance
AdaGradNorm (Ward et al., 2020)	$\mathcal{O}\left(\log T/\sqrt{T}\right)$	✗
AdaGrad (Défossez et al., 2020)	$\mathcal{O}\left(\log T/\sqrt{T}\right)$	✗
Adam (Défossez et al., 2020)	$\mathcal{O}\left(\log T/\sqrt{T}\right)$	✗
KATE (this work)	$\mathcal{O}\left(\log T/\sqrt{T}\right)$	✓

several variants of the SPS algorithm emerged (Li et al., 2023; D'Orazio et al., 2021). Lately, Orvieto et al. (2022) tackled the issues with unknown f_i^* and developed a truly adaptive variant. In practice, the SPS method shows excellent empirical performance on overparameterized deep learning models (which satisfy the interpolation condition i.e. $f_i^* = 0,\ \forall i \in [n]$) (Loizou et al., 2021).

1.2 Main Contribution

Our main contributions are summarized below.

• **KATE: new scale-invariant version of AdaGrad.** We propose a new method called KATE that can be seen as a version of AdaGrad, which does not use a square root in the denominator of the stepsize. To compensate for this change, we introduce a new sequence defining the numerator of the stepsize. We prove that KATE is scale-invariant for generalized linear models: if the starting point is zero, then the loss values (and training and test accuracies in the case of classification) at points generated by KATE are independent of the data scaling (Proposition 2.1), meaning that the speed of convergence of KATE is the same as for the best scaling of the data.

• **Convergence for smooth non-convex problems.** We prove that for smooth non-convex problems with noise having bounded variance KATE has $\mathcal{O}\left(\log(T)/\sqrt{T}\right)$ convergence rate (Theorem 3.4), matching the best-known rates for AdaGrad and Adam (Défossez et al., 2020).

• **Numerical experiments.** We empirically illustrate the scale-invariance of KATE on the logistic regression task and test its performance on logistic regression (see Section 4.1), image classification, and text classification problems (see Section 4.2). In all the considered scenarios, KATE outperforms AdaGrad and works either better or comparable to Adam.

1.3 Notation

We denote the set $\{1, 2, \cdots, n\}$ as $[n]$. For a vector $a \in \mathbb{R}^d$, $a[k]$ is the k-th coordinate of a and a^2 represents the element-wise suqare of a, i.e., $a^2[k] = (a[k])^2$. For two vectors a and b, $\frac{a}{b}$ stands for element-wise division of a and b, i.e., k-th coordinate of $\frac{a}{b}$ is $\frac{a[k]}{b[k]}$. Given a function $h : \mathbb{R}^d \to \mathbb{R}$, we use $\nabla h \in \mathbb{R}^d$ to denote its gradient and $\nabla_k h$ to indicate the k-th component of ∇h. Throughout the paper $\|\cdot\|$ represents the ℓ_2-norm and $f_* = \inf_{w\in\mathbb{R}^d} f(w)$. Moreover, we use $\|w\|_A$ for a positive-definite matrix A to define $\|w\|_A := \sqrt{w^\top A w}$. Furthermore, $\mathbb{E}[\cdot]$ denotes the total expectation while $\mathbb{E}_t[\cdot]$ denotes the conditional expectation conditioned on all iterates up to step t i.e. $w_0, w_1, \ldots, w_t$.

2 Motivation and Algorithm Design

We focus on solving the minimization problem (1) using a variant of AdaGrad. We aim to design an algorithm that performs well, irrespective of how poorly the data is scaled.

Generalized linear models. Here, we consider the parameter estimation problem in generalized linear models (GLMs) (Nelder and Wedderburn, 1972; Agresti, 2015) using maximum likelihood estimation. GLMs are an extension of linear models and encompass several other valuable models, such as logistic (Hosmer Jr et al., 2013) and Poisson regression (Frome, 1983), as special cases. The parameter estimation to fit GLM on dataset $\{x_i, y_i\}_{i=1}^n$ (where $x_i \in \mathbb{R}^d$ are feature vectors and y_i are response variables) can be reformulated as

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} \varphi_i \left(x_i^\top w\right) \tag{4}$$

for differentiable functions $\varphi_i : \mathbb{R} \to \mathbb{R}$ (Shalev-Shwartz and Ben-David, 2014; Nguyen et al., 2017b; Takáč et al., 2013; He et al., 2018; Chezhegov et al., 2024). For example, the linear regression on data $\{x_i, y_i\}_{i=1}^n$ is equivalent to solving (4) with $\varphi_i(z) = (z - y_i)^2$. Next, the choice of φ_i for logistic regression is $\varphi_i(z) = \log\left(1 + \exp\left(-y_i z\right)\right)$.

Scale-invariance. Now consider the instances of fitting GLMs on two datasets $\{x_i, y_i\}_{i=1}^n$ and $\{V x_i, y_i\}_{i=1}^n$, where $V \in \mathbb{R}^{d \times d}$ is a diagonal matrix with positive entries. Note that the second dataset is a scaled version of the first one where the k-th component of feature vectors x_i are multiplied by a scalar V_{kk}. Then, the minimization problems corresponding to datasets $\{x_i, y_i\}_{i=1}^n$ and $\{V x_i, y_i\}_{i=1}^n$ are (4) and

$$\min_{w \in \mathbb{R}^d} f^V(w) \quad := \quad \frac{1}{n} \sum\nolimits_{i=1}^{n} \varphi_i \left(x_i^\top V w\right), \tag{5}$$

respectively, for functions φ_i. In this work, *we want to design an algorithm with equivalent performance for the problems* (4) *and* (5). If we can do that, the new algorithm's performance will not deteriorate for poorly scaled data, i.e., the method will be scale-invariant (Orabona et al., 2015), which is a special case of affine-invariance, see (Nesterov and Nemirovskii, 1994; Nesterov, 2018; d'Aspremont et al., 2018). To develop such an algorithm, we replace the denominator of AdaGrad step size with its square (remove the square root from the denominator), i.e., $\forall k \in [d]$

$$w_{t+1}[k] \quad = \quad w_t[k] - \frac{\beta m_t[k]}{\sum_{\tau=0}^{t} g_\tau^2[k]} g_t[k] \tag{6}$$

for some $m_t \in \mathbb{R}^d$.[3] The following proposition shows that this method (6) satisfies a scale-invariance property with respect to functional value.

Proposition 2.1 (Scale invariance)**.** Suppose we solve problems (4) and (5) using algorithm (6). Then, the iterates $\hat{w}_t$ and $\hat{w}_t^V$ corresponding to (4) and (5) follow: $\forall k \in [d]$

$$\hat{w}_{t+1}[k] \quad = \quad \hat{w}_t[k] - \frac{\beta m_t[k]}{\sum_{\tau=0}^{t} g_\tau^2[k]} g_t[k], \tag{7}$$

$$\hat{w}_{t+1}^V[k] \quad = \quad \hat{w}_t^V[k] - \frac{\beta m_t[k]}{\sum_{\tau=0}^{t} \left(g_\tau^V[k]\right)^2} g_t^V[k] \tag{8}$$

with $g_\tau = \varphi_{i_\tau}'(x_{i_\tau}^\top \hat{w}_\tau) x_{i_\tau}$ and $g_\tau^V = \varphi_{i_\tau}'(x_{i_\tau}^\top V \hat{w}_\tau) V x_{i_\tau}$ for i_τ chosen uniformly from $[n]$, $\tau = 0, 1, \ldots, t$, $t \geq 0$. Moreover, updates (7) and (8) satisfy

$$\hat{w}_t = V \hat{w}_t^V, \quad V g_t = g_t^V, \quad f\left(\hat{w}_t\right) = f^V\left(\hat{w}_t^V\right) \tag{9}$$

for all $t \geq 0$ when $\hat{w}_0 = \hat{w}_0^V = 0 \in \mathbb{R}^d$. Furthermore we have

$$\left\|g_t^V\right\|_{V^{-2}}^2 \quad = \quad \left\|g_t\right\|^2. \tag{10}$$

The Proposition 2.1 highlights that the update rule of the form (6) satisfies a scale-invariance property for GLMs. In contrast, AdaGrad does not satisfy (9) and (10). In Appendix C, we illustrate numerically the scale-invariance of KATE and the lack of the scale-invariance of AdaGrad. We also emphasize that AdaGrad with $\Delta = 0$ is known to be a scale-free method[4].

[3] Sequence $\{m_t\}_{t \geq 0}$ can depend on the problem but is assumed to be scale-invariant.

[4] The algorithm is called scale-free if for any $c > 0$, it generates the same sequence of points for functions f and cf given the same initialization and hyperparameters. To the best of our knowledge, this definition is

Algorithm 1 KATE

Require: Initial point $w_0 \in \mathbb{R}^d$, step size $\beta > 0, \eta \in \mathbb{R}^d_+$ and $b_{-1}, m_{-1} = 0$.

1: **for** $t = 0, 1, ..., T$ **do**

2: Compute $g_t \in \mathbb{R}^d$ such that $\mathbb{E}[g_t] = \nabla f(w_t)$.

3: $b_t^2 = b_{t-1}^2 + g_t^2$

4: $m_t^2 = m_{t-1}^2 + \eta g_t^2 + \frac{g_t^2}{b_t^2}$

5: $w_{t+1} = w_t - \frac{\beta m_t}{b_t^2} g_t$

Design of KATE. In order to construct an algorithm following the update rule (6), one may choose $m_t[k] = 1\ \forall k \in [d]$. However, the step size from (6) in this case may decrease very fast, and the resulting method does not necessarily converge. Therefore, we need a more aggressive choice of m_t, which grows with t. It motivates the construction of our algorithm KATE (Algorithm 1),[5] where we choose $m_t[k] = \sqrt{\eta[k] b_t^2[k] + \sum_{\tau=0}^{t} \frac{g_\tau^2[k]}{b_\tau^2[k]}}$. Note that the term $\sum_{\tau=0}^{t} \frac{g_\tau^2[k]}{b_\tau^2[k]}$ is scale-invariant for GLMs (follows from Proposition 2.1). To make m_t scale-invariant, we choose $\eta \in \mathbb{R}^d$ in the following way:

- $\underline{\eta \to 0}$: When η is very small, m_t is also approximately scale-invariant for GLMs.
- $\underline{\eta = 1/(\nabla f(w_0))^2}$: In this case $\eta b_t^2 = b_t^2/(\nabla f(w_0))^2$ is scale-invariant for GLMs (follows from Proposition 2.1) as well as m_t.

KATE can be rewritten in the following coordinate form

$$w_{t+1}[k] = w_t[k] - \nu_t[k] g_t[k], \qquad \forall k \in [d], \tag{11}$$

where g_t is an unbiased estimator of $\nabla f(w_t)$ and the per-coefficient step size $\nu_t[k]$ is defined as

$$\nu_t[k] := \frac{\beta \sqrt{\eta[k] b_t^2[k] + \sum_{\tau=0}^{t} \frac{g_\tau^2[k]}{b_\tau^2[k]}}}{b_t^2[k]}. \tag{12}$$

Note that the numerator of the steps $\nu_t[k]$ is increasing with iterations t. However, one of the crucial properties of this step size choice is that the steps always decrease with t, which we rely on in our convergence analysis.

Lemma 2.2 (Decreasing step size)**.** For $\nu_t[k]$ defined in (11) we have

$$\nu_{t+1}[k] \le \nu_t[k], \qquad \forall k \in [d]. \tag{13}$$

Comparison with the scale-invariant version of AdaGrad by Orabona et al. (2015). In the special case of GLMs, Orabona et al. (2015) propose a different version of AdaGrad. The method is proposed for the case of online convex optimization, and in the case of standard optimization with GLMs (4), it has the following form

$$w_0 := 0, \quad w_{t+1} := -\beta \frac{\sum_{\tau=0}^{t} \nabla f_{i_\tau}(w_\tau)}{a_t^2 \sqrt{d} \sqrt{\gamma^2 + \sum_{\tau=0}^{t} \left(\nabla f_{i_\tau}(w_\tau)/a_\tau\right)^2}}, \quad a_t := \max_{\tau=0,\ldots,t} |x_{i_\tau}|, \tag{14}$$

where $\{i_\tau\}_{\tau=0}^t$ are arbitrary indices from $[n]$ (e.g., selected uniformly at random), functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are defined as $f_i(w) := \varphi_i(x_i^\top w)$ for $i \in [n]$, and γ is such that $f_i(w)$ is γ-Lipschitz for $i \in [n]$. In this setup, the update rule of KATE with $w_0 = 0$ can be written as follows:

$$w_{t+1} := -\beta \sum_{\tau=0}^{t} \frac{m_\tau}{b_\tau^2} \nabla f_{i_\tau}(w_\tau), \quad m_t := \sqrt{\eta \sum_{\tau=0}^{t} (\nabla f_{i_\tau}(w_\tau))^2 + \sum_{\tau=0}^{t} (\nabla f_{i_\tau}(w_\tau))^2/b_\tau^2},$$

introduced by Cesa-Bianchi et al. (2005, 2007) in the context of learning with expert advice and extended to the context of generic online convex optimization by Orabona and Pál (2015, 2018). We emphasize that scale-freeness and scale-invariance are completely different concepts.

[5]Note that, for $m_t = b_t \forall t$ we get the AdaGrad algorithm.

where $b_t := \sqrt{\sum_{\tau=0}^{t}(\nabla f_{i_\tau}(w_\tau))^2}$, $\{i_\tau\}_{\tau=0}^{t}$ are sampled from $[n]$ uniformly at random. Although both methods can be seen as variations of AdaGrad due to the terms $\sum_{\tau=0}^{t}\left(\nabla f_{i_\tau}(w_\tau/a_\tau)\right)^2$ and $\sum_{\tau=0}^{t}\left(\nabla f_{i_\tau}(w_\tau)\right)^2$ respectively, the scale-invariance is achieved quite differently in these methods. The method from (14) uses the feature vectors explicitly in the update rule to ensure scale-invariance: indeed, the square root in the definition of w_{t+1} is independent of scaling, and a_t^2 in the denominator ensures that $\hat{w}_{t+1} = V\hat{w}_{t+1}^V$ if we define them similarly to KATE (see equations (7)-(8)). In contrast, KATE achieves the scale-invariance by removing the square root from the denominator (as explained earlier). Moreover, unlike the method from (14), KATE does not use the feature vectors explicitly in its update rule (only in the gradients of f_{i_τ}) and, thus, can be used for general stochastic optimization (not necessarily for the case of GLMs).

3 Convergence Analysis

In this section, we present and discuss the convergence guarantees of KATE. In the first subsection, we list the assumptions made about the problem.

3.1 Assumptions

In all our theoretical results, we assume that f is smooth as defined below.

Assumption 3.1 (L-smooth)**.** Function f is L-smooth, i.e. for all $w, w' \in \mathbb{R}^d$

$$f(w') \leq f(w) + \langle \nabla f(w), w' - w\rangle + \frac{L}{2}\|w - w'\|^2. \tag{15}$$

This assumption is standard in the literature of adaptive methods (Li and Orabona, 2019; Ward et al., 2020; Liu et al., 2022; Nguyen et al., 2018, 2021, 2017a; Beznosikov and Takáč, 2021). Moreover, we assume that at any iteration t of KATE, we can access g_t — a noisy and unbiased estimate of $\nabla f(w_t)$. We also make the following assumption on the noise of the gradient estimate g_t.

Assumption 3.2 (Bounded Variance)**.** For fixed constant $\sigma > 0$, the variance of the stochastic gradient g_t (unbiased estimate of $\nabla f(w_t)$) at any time t satisfies

$$\mathbb{E}_t\left[\|g_t - \nabla f(w_t)\|^2\right] \leq \sigma^2. \tag{BV}$$

Bounded variance is a common assumption to study the convergence of stochastic gradient-based methods. Several assumptions on stochastic gradients are used in the literature to explore the adaptive methods. Ward et al. (2020) used the BV, while Liu et al. (2022) assumed the sub-Weibull noise, i.e. $\mathbb{E}\left[\exp\left(\|g_t - \nabla f(w_t)\|/\sigma\right)^{1/\theta}\right] \leq \exp(1)$ for some $\theta > 0$, to prove the convergence of AdaGradNorm. Li and Orabona (2019) assumes sub-Gaussian ($\theta = 1/2$ in sub-Weibull condition) noise to study a variant of AdaGrad. However, sub-Gaussian noise is strictly stronger than BV. Recently, Faw et al. (2022) analyzed AdaGradNorm under a more relaxed condition known as affine variance $\left(\text{i.e. } \mathbb{E}_t\left[\|g_t - \nabla f(w_t)\|^2\right] \leq \sigma_0^2 + \sigma_1^2\|\nabla f(w_t)\|^2\right)$.

3.2 Main Results

In this section, we cover the main convergence guarantees of KATE for both deterministic and stochastic setups.

Deterministic setting. We first present our results for the deterministic setting. In this setting, we consider the gradient estimate to have no noise (i.e. $\sigma^2 = 0$) and $g_t = \nabla f(w_t)$. The main result in this setting is summarized below.

Theorem 3.3. Suppose f satisfy Assumption 3.1 and $g_t = \nabla f(w_t)$. Moreover, $\beta > 0$ and $\eta[k] > 0$ are chosen such that $\nu_0[k] \leq \frac{1}{L}$ for all $k \in [d]$. Then the iterates of KATE satisfies

$$\min_{t \leq T} \|\nabla f(w_t)\|^2 \quad \leq \quad \frac{\left(\frac{2(f(w_0)-f_*)}{\sqrt{\eta_0}\beta} + \sum_{k=1}^{d} b_0[k]\right)^2}{T+1},$$

where $\eta_0 := \min_{k \in [d]} \eta[k]$.

Discussion on Theorem 3.3. Theorem 3.3 establishes an $\mathcal{O}\left(1/T\right)$ convergence rate for KATE, which is optimal for finding a first-order stationary point of a non-convex problem (Carmon et al., 2020). However, this result is not parameter-free. To prove the convergence, we assume that $\nu_0[k] \leq \frac{1}{L}$, $\forall k \in [d]$ in Theorem 3.3, which is equivalent to $\beta\sqrt{1 + \eta_0 \left(\nabla_k f(w_0)\right)^2} \leq (\nabla_k f(w_0))^2/L$, $\forall k \in [d]$. Note that the later condition holds for sufficiently small (dependent on L) values of $\beta, \eta_0 > 0$.

However, it is possible to derive a parameter-free version of Theorem 3.3. Indeed, Lemma 2.2 implies that the step sizes are decreasing. Therefore, we can break down the analysis of KATE into two phases: Phase I when $\nu_0[k] > 1/L$ and Phase II when $\nu_0[k] \leq 1/L$, when the current analysis works, and then follow the proof techniques of Ward et al. (2020) and Xie et al. (2020). We leave this extension as a possible future direction of our work.

Stochastic setting. Next, we present the convergence guarantees for KATE in the stochastic case, when we can access an unbiased gradient estimate g_t with non-zero noise.

Theorem 3.4. Suppose f satisfy Assumption 3.1 and g_t is an unbiased estimator of $\nabla f(w_t)$ such that BV holds. Moreover, we assume $\|\nabla f(w_t)\|^2 \leq \gamma^2$ for all t. Then the iterates of KATE satisfy

$$\min_{t \leq T} \mathbb{E}\left[\|\nabla f(w_t)\|\right] \leq \left(\frac{\|g_0\|}{T} + \frac{2(\gamma+\sigma)}{\sqrt{T}}\right)^{1/2} \sqrt{\frac{2\mathcal{C}_f}{\beta\sqrt{\eta_0}}},$$

where $\eta_0 := \min_{k \in [d]} \eta[k]$ and

$$\begin{aligned}\mathcal{C}_f \quad := \quad & f(w_0) - f_* + 2\beta\sigma\sum_{k=1}^{d}\sqrt{\eta[k]}\log\left(\frac{e(\sigma^2+\gamma^2)T}{g_0^2[k]}\right) \\ & + \sum_{k=1}^{d}\left(\frac{\beta^2\eta[k]L}{2} + \frac{\beta^2 L}{2g_0^2[k]}\right)\log\left(\frac{e(\sigma^2+\gamma^2)T}{g_0^2[k]}\right).\end{aligned}$$

Comparison with prior work. Theorem 3.4 shows an $\mathcal{O}(\log^{1/2} T/T^{1/4})$ convergence rate for KATE with respect to the metric $\min_{t \leq T} \mathbb{E}\left[\|\nabla f(w_t)\|\right]$ for the stochastic setting. Note that, in the stochastic setting, KATE achieves a slower rate than Theorem 3.3 due to noise accumulation. Up to the logarithmic factor, this rate is optimal (Arjevani et al., 2023). Similar rates for the same metric follow from the results[6] of (Défossez et al., 2020) for AdaGrad and Adam.

Finally, Li and Orabona (2019) considers a variant of AdaGrad closely related to KATE:

$$w_{t+1} = w_t - \frac{\beta g_t}{\left(\mathrm{diag}\left(\Delta I + \sum_{\tau=1}^{t-1} g_\tau g_\tau^\top\right)\right)^{\frac{1}{2}+\varepsilon}}, \tag{16}$$

for some $\varepsilon \in [0, 1/2)$ and $\Delta > 0$. It differs from AdaGrad in two key aspects: the denominator of the stepsize does not contain the last stochastic gradient, and also, instead of the square root of the sum of squared gradients, this sum is taken in the power of $1/2 + \varepsilon$. However, the results from Li and Orabona (2019) do not imply convergence for the case of $\varepsilon = 1/2$, which is expected since, in this case, the stepsize converges to zero too quickly in general. To compensate for such a rapid decrease, in KATE, we introduce an increasing sequence m_t in the numerator of the stepsize.

[6] Défossez et al. (2020) derive $\mathcal{O}(\log T/\sqrt{T})$ convergence rates for AdaGrad and Adam in terms of $\min_{t \leq T} \mathbb{E}\left[\|\nabla f(w_t)\|^2\right]$ which is not smaller than $\min_{t \leq T} \left(\mathbb{E}\left[\|\nabla f(w_t)\|\right]\right)^2$.

Proof technique. Compared to the AdaGrad, KATE uses more aggressive steps (the larger numerator of KATE due to the extra term $\sum_{\tau=0}^{t} g_\tau^2[k]/b_\tau^2[k]$). Therefore, we expect KATE to have better empirical performance. However, introducing $\sum_{\tau=0}^{t} g_\tau^2[k]/b_\tau^2[k]$ in the numerator raises additional technical difficulties in the proof technique. Fortunately, as we rigorously show, the KATE steps $\nu_t[k]$ retain some of the critical properties of AdaGrad steps. For instance, they (i) are lower bounded by AdaGrad steps up to a constant, (ii) decrease with iteration t (Lemma 2.2), and (iii) have closed-form upper bounds for $\sum_{t=0}^{T} \nu_t^2[k] g_t^2[k]$. These are indeed the primary building blocks of our proof technique.

4 Numerical Experiments

In this section, we implement KATE in several machine learning tasks to evaluate its performance. To ensure transparency and facilitate reproducibility, we provide an access to the source code for all of our experiments at https://github.com/nazya/KATE.

4.1 Logistic Regression

In this section, we consider the logistic regression model

$$\min_{w \in \mathbb{R}^d} f(w) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i x_i^\top w\right)\right), \tag{17}$$

to elaborate on the scale-invariance and robustness of KATE for various initializations. For the experiments of this Section 4.1, we used Mac mini (M1, 2020), RAM 8 GB and storage 256 GB. Each of these plots took about 20 minutes to run.

4.1.1 Robustness of KATE

To conduct this experiment, we set the total number of samples to 1000 (i.e. $n = 1000$). Here, we simulate the independent vectors $x_i \in \mathbb{R}^{20}$ such that each entry is from $\mathcal{N}(0,1)$. Moreover, we generate a diagonal matrix $V \in \mathbb{R}^{20\times 20}$ such that $\log V_{kk} \overset{\text{iid}}{\sim} \mathrm{Unif}(-10, 10)$, $\forall k \in [20]$. Similarly, we generate $w^* \in \mathbb{R}^{20}$ with each component from $\mathcal{N}(0,1)$ and set the labels

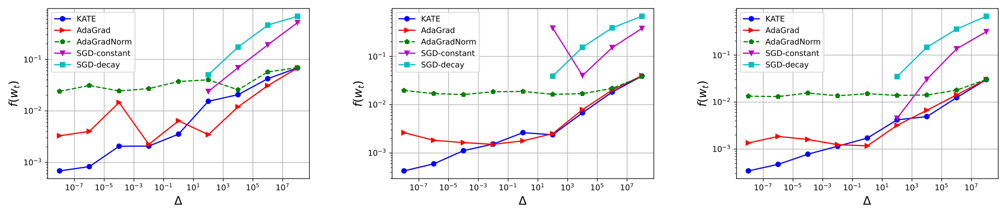
$$y_i = \begin{cases} 1, & x_i^\top V w^* \geq 0, \\ -1, & x_i^\top V w^* < 0, \end{cases} \qquad \forall i \in [n].$$

We compare KATE's performance with four other algorithms: AdaGrad, AdaGradNorm, SGD-decay and SGD-constant, similar to the section 5.1 of Ward et al. (2020). For each algorithm, we initialize with $w_0 = 0 \in \mathbb{R}^{20}$ and independently draw a sample of mini-batch size 10 to update the weight vector w_t. We compare the algorithms • AdaGrad with stepsize $\frac{\beta}{\sqrt{\Delta + \sum_{\tau=0}^{t} g_\tau^2}}$, • AdaGradNorm with step size $\frac{\beta}{\sqrt{\Delta + \sum_{\tau=0}^{t} \|g_\tau\|^2}}$, • SGD-decay with stepsize $\beta/\Delta\sqrt{t+1}$, and • SGD-constant with step size β/Δ. Similarly, for KATE we use stepsize $\frac{\beta m_t}{b_t^2}$ where $m_t^2 = \eta b_t^2 + \sum_{\tau=0}^{t} g_\tau^2/b_\tau^2$ and $b_t^2 = \Delta + \sum_{\tau=0}^{t} g_\tau^2$. Here, we choose $\beta = f(w_0) - f(w^*)$ and vary Δ in $\{10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1, 10^2, 10^4, 10^6, 10^8\}$.

In Figures 1a, 1b, and 1c, we plot the functional value $f(w_t)$ (on the y-axis) after $10^4, 5\times 10^4$, and 10^5 iterations, respectively. In theory, the convergence of SGD requires the knowledge of smoothness constant L. Therefore, when the Δ is small (hence the stepsize is large), SGD-decay and SGD-constant diverge. However, the adaptive algorithms KATE, AdaGrad, and AdaGradNorm can auto-tune themselves and converge for a wide range of Δs (even when the Δ is too small). As we observe in Figure 1, when the Δ is small, KATE outperforms all other algorithms. For instance, when $\Delta = 10^{-8}$, KATE achieves a functional value of 10^{-3} after only 10^4 iterations (see Figure 1a), while other algorithms fail to achieve this even after 10^5 iterations (see Figure 1c). Furthermore, KATE performs as well as AdaGrad and better than other algorithms when the Δ is large. *In particular, this experiment highlights that* KATE *is robust to initialization* Δ.

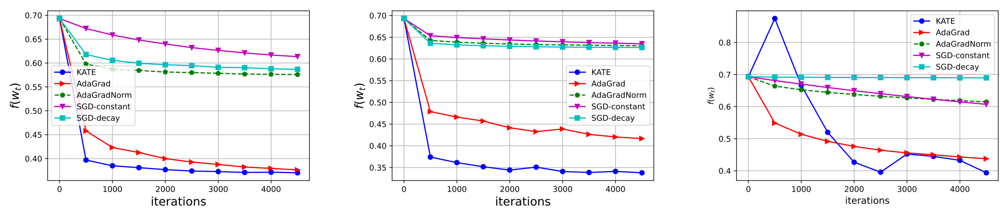
4.1.2 Peformance of KATE on Real Data

In this section, we examine KATE's performance on real data. We test KATE on three datasets: heart, australian, and splice from the LIBSVM library (Chang and Lin, 2011). The response variables y_i of

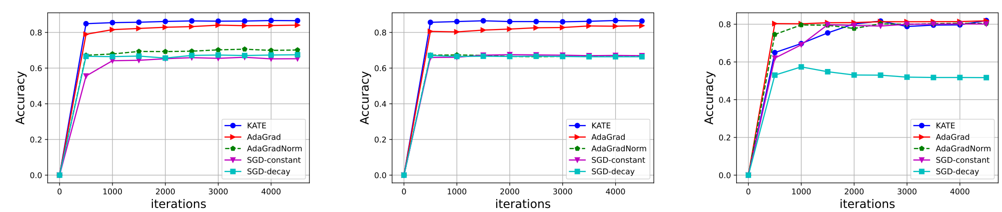


(a) Plot of $f(w_t)$ (b) Plot of Accuracy (c) Plot of Gradient Norm

Figure 1: Comparison of KATE with AdaGrad, AdaGradNorm, SGD-decay and SGD-constant for different values of Δ (on x-axis for logistic regression model. Figure 1a, 1b and 1c plots the functional value $f(w_t)$ (on y-axis) after 10^4, 5×10^4, and 10^5 iterations respectively.



(a) Dataset: heart (b) Dataset: australian (c) Dataset: splice



(d) Dataset: heart (e) Dataset: australian (f) Dataset: splice

Figure 2: Comparison of KATE with AdaGrad, AdaGradNorm, SGD-decay and SGD-constant on datasets heart, australian, and splice from LIBSVM. Figures 2a, 2b and 2c plot the functional value $f(w_t)$, while 2d, 2e and 2f plot the accuracy on y-axis for $5,000$ iterations.

each of these datasets contain two classes, and we use them for binary classification tasks using a logistic regression model (17). We take $\eta = {}^1/(\nabla f(w_0))^2$ for KATE and tune β in all the experiments. For tuning β, we do a grid search on the list $\{10^{-10}, 10^{-8}, 10^{-6}, 10^{-4}, 10^{-2}, 1\}$. Similarly, we tune stepsizes for other algorithms. We take 5 trials for each of these algorithms and plot the mean of their trajectories.

We plot the functional value $f(w_t)$ (i.e. loss function) in Figures 2a, 2b and 2c, whereas Figures 2d, 2e and 2f plot the corresponding accuracy of the weight vector w_t on the y-axis for $5,000$ iterations. We observe that KATE performs superior to all other algorithms, even on real datasets.

4.2 Training of Neural Networks

In this section, we compare the performance of KATE, AdaGrad and Adam on two tasks, i.e. training ResNet18 (He et al., 2016) on the CIFAR10 dataset (Krizhevsky and Hinton, 2009) and BERT (Devlin et al., 2018) fine-tuning on the emotions dataset (Saravia et al., 2018) from the Hugging Face Hub. We use internal cluster with the following hardware: AMD EPYC 7552 48-Core Processor, 512GiB RAM, NVIDIA A100 40GB GPU, 200gb user storage space.

General comparison. We choose standard parameters for Adam ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) that are default values in PyTorch and select the learning rate of 10^{-5} for all considered methods. We run KATE with different values of $\eta \in \{0, 10^{-1}, 10^{-2}\}$. For the image classification task, we normalize the images (similar to Horváth and Richtárik (2020)) and use a mini-batch size of 500. For the BERT fine-tuning, we use a mini-batch size 160 for all methods.

Figures 3-8 report the evolution of top-1 accuracy and cross-entropy loss (on the y-axis) calculated on the test data. For the image classification task, we observe that KATE with different choices of

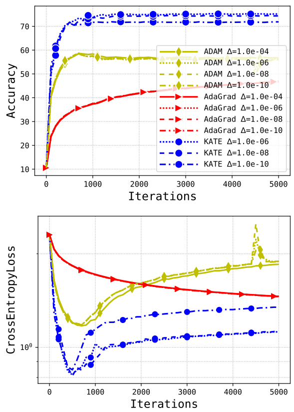


Figure 3: CIFAR10: $\eta = 0$

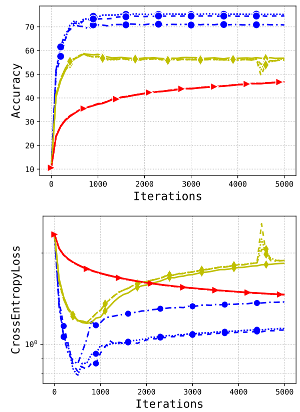


Figure 4: CIFAR10: $\eta = 0.001$

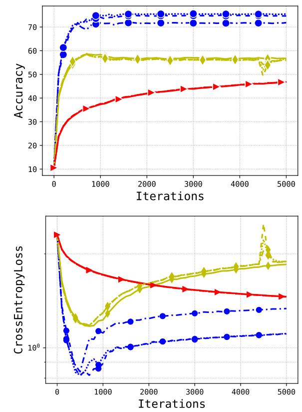


Figure 5: CIFAR10: $\eta = 0.1$

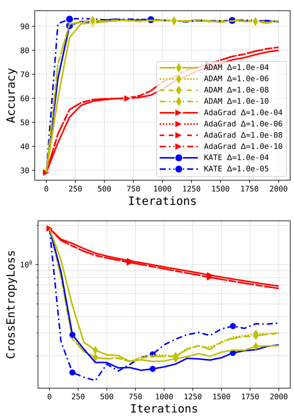


Figure 6: Emotion: $\eta = 0$

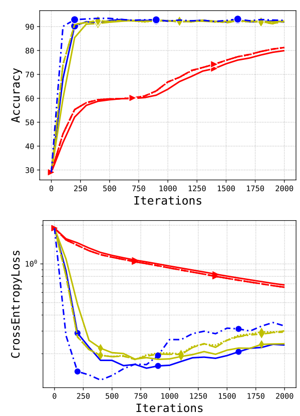


Figure 7: Emotion: $\eta = 0.001$

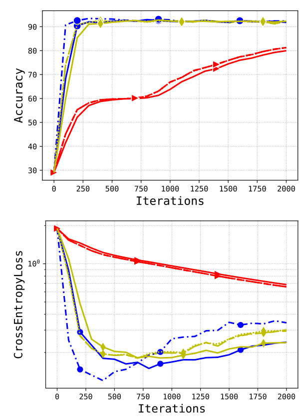


Figure 8: Emotion: $\eta = 0.1$

η outperforms Adam and AdaGrad. Finally, we also observe that KATE performs comparably to Adam on the BERT fine-tuning task and is better than AdaGrad. These preliminary results highlight the potential of KATE to be applied for training neural networks for different tasks. For BERT each run takes about 35 minutes, and 25 minutes for ResNet.

Hyper-parameters tuning. Next, we compare baselines presented in Saravia et al. (2018) for emotions classification and Zhang et al. (2019) for image classification. These papers provide efficient setups for learning rates and learning rate schedulers that are reasonable to compare with. Saravia et al. (2018) performs a search of efficient learning rate and uses a linear learning rate scheduler with warmup for Adam optimizer. A different learning rate (1e-5), Δ=1e-5 and the same scheduler applied for KATE lead to the same performance, see Figure 9. We would like to point out that it is challenging to find a reference for hyper-parameters for a certain setup. Thus, to fairly compare with Saravia et al. (2018) we use distilroberta-base model. Zhang et al. (2019) did a grid search for an efficient learning rate and used a multi-step scheduler for Adam optimizer, decaying the learning rate by a factor of 5 at the 60th, 120th, and 160th epochs. Zhang et al. (2019) refers to DeVries and Taylor (2017) for the code implementing special techniques, namely data augmentation and cutout to achieve higher accuracy. A different learning rate (1e-3), the same scheduler and Δ=1e-3 applied for KATE demonstrates comparable performance, see Figure 10. For BERT each run takes about 20 minutes, while 100 minutes for ResNet.

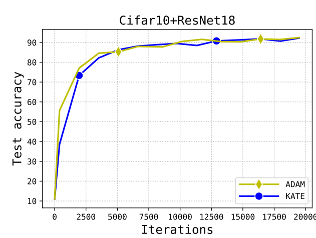


Figure 9: Cifar10: $\eta = 0.001$

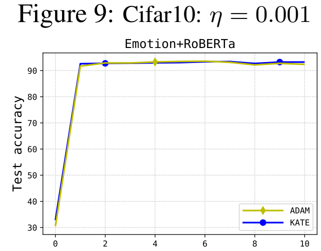


Figure 10: Emotion: $\eta = 0.001$

Acknowledgments and Disclosure of Funding

We thank Francesco Orabona and Dmitry Kamzolov for the pointers to the related works that we missed while preparing the first version of this paper. We also thank anonymous reviewers for their useful feedback and suggestions.

References

Abdukhakimov, F., Xiang, C., Kamzolov, D., Gower, R., and Takáč, M. (2023). Sania: Polyak-type optimization framework leads to scale invariant stochastic algorithms. *arXiv preprint arXiv:2312.17369*.

Abdukhakimov, F., Xiang, C., Kamzolov, D., and Takáč, M. (2024). Stochastic gradient descent with preconditioned polyak step-size. *Computational Mathematics and Mathematical Physics*, 64(4):621–634.

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2023). Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214.

Beznosikov, A. and Takáč, M. (2021). Random-reshuffled sarah does not need a full gradient computations. In *Optimization for Machine Learning Workshop @ NeurIPS 2021*.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2020). Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. (2005). Improved second-order bounds for prediction with expert advice. In *International Conference on Computational Learning Theory*, pages 217–232. Springer.

Cesa-Bianchi, N., Mansour, Y., and Stoltz, G. (2007). Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352.

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

Chezhegov, S., Skorik, S., Khachaturov, N., Shalagin, D., Avetisyan, A., Beznosikov, A., Takáč, M., Kholodov, Y., and Gasnikov, A. (2024). Local methods with adaptivity via scaling. *arXiv preprint arXiv:2406.00846*.

d'Aspremont, A., Guzman, C., and Jaggi, M. (2018). Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405.

Defazio, A. and Mishchenko, K. (2023). Learning-rate-free learning by D-adaptation. *arXiv preprint arXiv:2301.07733*.

Défossez, A., Bottou, L., Bach, F., and Usunier, N. (2020). A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DeVries, T. and Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

D'Orazio, R., Loizou, N., Laradji, I., and Mitliagkas, I. (2021). Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic polyak stepsize. *arXiv preprint arXiv:2110.15412*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. (2022). The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pages 313–355. PMLR.

Frome, E. L. (1983). The analysis of rates using poisson regression models. *Biometrics*, pages 665–674.

Gower, R. M., Defazio, A., and Rabbat, M. (2021). Stochastic polyak stepsize with a moving target. *arXiv preprint arXiv:2106.11851*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, X., Tappenden, R., and Takac, M. (2018). Dual free adaptive minibatch sdca for empirical risk minimization. *Frontiers in Applied Mathematics and Statistics*, 4:33.

Horváth, S. and Richtárik, P. (2020). A better alternative to error feedback for communication-efficient distributed learning. *arXiv preprint arXiv:2006.11077*.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images.

Li, S., Swartworth, W. J., Takáč, M., Needell, D., and Gower, R. M. (2023). Sp2: A second order stochastic polyak method. *ICLR*.

Li, X. and Orabona, F. (2019). On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pages 983–992. PMLR.

Liu, Z., Nguyen, T. D., Ene, A., and Nguyen, H. L. (2022). On the convergence of adagrad on $\mathbb{R}^d$: Beyond convexity, non-asymptotic rate and acceleration. *arXiv preprint arXiv:2209.14827*.

Loizou, N., Vaswani, S., Laradji, I. H., and Lacoste-Julien, S. (2021). Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR.

McMahan, H. B. and Streeter, M. (2010). Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*.

Mishchenko, K. and Defazio, A. (2023). Prodigy: An expeditiously adaptive parameter-free learner. *arXiv preprint arXiv:2306.06101*.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nesterov, Y. and Nemirovskii, A. (1994). *Interior-point polynomial algorithms in convex programming*. SIAM.

Nguyen, L., Liu, J., Scheinberg, K., and Takáč, M. (2017a). Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *In 34th International Conference on Machine Learning, ICML 2017*.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017b). Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*.

Nguyen, L. M., Nguyen, P. H., van Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. (2018). Sgd and hogwild! convergence without the bounded gradients assumption. In *In 34th International Conference on Machine Learning, ICML 2018*.

Nguyen, L. M., Scheinberg, K., and Takáč, M. (2021). Inexact sarah algorithm for stochastic optimization. *Optimization Methods and Software*, 36(1):237–258.

Oberman, A. M. and Prazeres, M. (2019). Stochastic gradient descent with polyak's learning rate. *arXiv preprint arXiv:1903.08688*.

Orabona, F., Crammer, K., and Cesa-Bianchi, N. (2015). A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99:411–435.

Orabona, F. and Pál, D. (2015). Scale-free algorithms for online linear optimization. In *International Conference on Algorithmic Learning Theory*, pages 287–301. Springer.

Orabona, F. and Pál, D. (2018). Scale-free online learning. *Theoretical Computer Science*, 716:50–69.

Orvieto, A., Lacoste-Julien, S., and Loizou, N. (2022). Dynamics of sgd with stochastic polyak stepsizes: Truly adaptive variants and convergence to exact solution. *Advances in Neural Information Processing Systems*, 35:26943–26954.

Polyak, B. T. (1969). Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29.

Reddi, S. J., Kale, S., and Kumar, S. (2019). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., and Chen, Y.-S. (2018). CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shi, Z., Sadiev, A., Loizou, N., Richtárik, P., and Takáč, M. (2023). AI-SARAH: Adaptive and implicit stochastic recursive gradient methods. *Transactions on Machine Learning Research*.

Takáč, M., Bijral, A., Richtárik, P., and Srebro, N. (2013). Mini-batch primal and dual methods for svms. In *In 30th International Conference on Machine Learning, ICML 2013*.

Tieleman, T. and Hinton, G. (2012). Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *COURSERA Neural Networks Mach. Learn*, 17.

Ward, R., Wu, X., and Bottou, L. (2020). Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21(1):9047–9076.

Xie, Y., Wu, X., and Ward, R. (2020). Linear convergence of adaptive stochastic gradient descent. In *International conference on artificial intelligence and statistics*, pages 1475–1485. PMLR.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. (2019). Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems*, 32.

Supplementary Material

Contents

A Technical Lemmas

Lemma A.1 (AM-GM)**.** For $\lambda > 0$ we have

$$ab \leq \frac{\lambda}{2}a^2 + \frac{1}{2\lambda}b^2. \tag{18}$$

Lemma A.2 (Cauchy-Schwarz Inequality)**.** For $a_1, \cdots, a_n, b_1, \cdots, b_n \in \mathbb{R}$ we have

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \quad \geq \quad \left(\sum_{i=1}^{n} a_i b_i\right)^2. \tag{19}$$

Lemma A.3 (Holder's Inequality)**.** Suppose X, Y are two random variables and $p, q > 1$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\mathbb{E}\left(|XY|\right) \leq \left(\mathbb{E}\left(|X|^p\right)\right)^{\frac{1}{p}} \left(\mathbb{E}\left(|Y|^q\right)\right)^{\frac{1}{q}}. \tag{20}$$

Lemma A.4 (Jensen's Inequality)**.** For a convex function $g : \mathbb{R}^d \to \mathbb{R}$ and a random variable X such that $\mathbb{E}(\Psi(X))$ and $\Psi\left(\mathbb{E}(X)\right)$ are finite, we have

$$\Psi\left(\mathbb{E}(X)\right) \leq \mathbb{E}(\Psi(X)). \tag{21}$$

Lemma A.5. For $a_1, a_2, \cdots, a_n \geq 0$ and $b_1, b_2, \cdots, b_n > 0$ we have

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{b_i}} \geq \frac{\sum_{i=1}^{n} a_i}{\sqrt{\sum_{i=1}^{n} b_i}}. \tag{22}$$

Proof. Expanding the LHS of (22) we get

$$\begin{aligned}
\left(\sum_{i=1}^{n} \frac{a_i}{\sqrt{b_i}}\right)^2 \quad &= \quad \sum_{i=1}^{n} \frac{a_i^2}{b_i} + 2\sum_{i \neq j} \frac{a_i a_j}{\sqrt{b_i b_j}} \\
&\geq \quad \sum_{i=1}^{n} \frac{a_i^2}{b_i}.
\end{aligned} \tag{23}$$

The last inequality follows from $\frac{a_i}{\sqrt{b_i}} \geq 0$ for all $i \in [n]$. Now, using Cauchy-Schwarz Inequality (19), we have

$$\left(\sum_{i=1}^{n} \frac{a_i^2}{b_i}\right)\left(\sum_{i=1}^{n} b_i\right) \quad \geq \quad \left(\sum_{i=1}^{n} a_i\right)^2. \tag{24}$$

Then combining (23) and (24), we get

$$\left(\sum_{i=1}^{n} \frac{a_i}{\sqrt{b_i}}\right)^2 \left(\sum_{i=1}^{n} b_i\right) \quad \geq \quad \left(\sum_{i=1}^{n} a_i\right)^2.$$

Finally dividing both sides by $\sum_{i=1}^{n} b_i$ and taking square root we get the desired result. $\square$

Lemma A.6. For $k \in [d]$ and $t \geq 1$ we have

$$\mathbb{E}_t\left[\left(\frac{\beta\sqrt{\eta[k]}}{\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}} - \nu_t[k]\right)\nabla_k f(w_t) g_t[k]\right] \leq \frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{2\sqrt{b_{t-1}^2[k] + (\nabla f(w_t))^2 + \sigma^2}} + 2\beta\sqrt{\eta[k]}\sigma\mathbb{E}_t\left[\frac{g_t^2[k]}{b_t^2[k]}\right] \tag{25}$$

Proof. Note that, using $\nu_t[k] \geq \frac{\beta\sqrt{\eta[k]}}{b_t[k]}$ we have

$$\begin{aligned}
&\frac{\beta\sqrt{\eta[k]}}{\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}} - \nu_t[k] \\
\leq\ & \beta\sqrt{\eta[k]}\left(\frac{1}{\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}} - \frac{1}{b_t[k]}\right) \\
=\ & \beta\sqrt{\eta[k]}\left(\frac{b_t^2[k] - b_{t-1}^2[k] - (\nabla_k f(w_t))^2 - \sigma^2}{b_t[k]\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\left(b_t[k] + \sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\right)}\right) \\
=\ & \beta\sqrt{\eta[k]}\left(\frac{g_t^2[k] - (\nabla_k f(w_t))^2 - \sigma^2}{b_t[k]\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\left(b_t[k] + \sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\right)}\right) \\
=\ & \beta\sqrt{\eta[k]}\left(\frac{\left(g_t[k] + \nabla_k f(w_t)\right)\left(g_t[k] - \nabla_k f(w_t)\right) - \sigma^2}{b_t[k]\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\left(b_t[k] + \sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\right)}\right) \\
\leq\ & \frac{\beta\sqrt{\eta[k]}\left|\left(g_t[k] + \nabla_k f(w_t)\right)\left(g_t[k] - \nabla_k f(w_t)\right)\right|}{b_t[k]\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\left(b_t[k] + \sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\right)} \\
&+ \frac{\beta\sqrt{\eta[k]}\sigma^2}{b_t[k]\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\left(b_t[k] + \sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}\right)} \\
\leq\ & \frac{\beta\sqrt{\eta[k]}\left|g_t[k] - \nabla_k f(w_t)\right|}{b_t[k]\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}} + \frac{\beta\sqrt{\eta[k]}\sigma}{b_t[k]\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}}.
\end{aligned} \tag{26}$$

Note that the second last inequality follows from the use of triangle inequality in the following way

$$\begin{aligned}
\left(g_t[k] + \nabla_k f(w_t)\right)\left(g_t[k] - \nabla_k f(w_t)\right) - \sigma^2 &\leq \left|\left(g_t[k] + \nabla_k f(w_t)\right)\left(g_t[k] - \nabla_k f(w_t)\right) - \sigma^2\right| \\
&\leq \left|\left(g_t[k] + \nabla_k f(w_t)\right)\left(g_t[k] - \nabla_k f(w_t)\right)\right| + \sigma^2,
\end{aligned}$$

while the last inequality follows from

$$\begin{aligned}
b_t[k] + \sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2} &\geq |g_t[k]| + |\nabla_k f(w_t)| \geq |g_t[k] + \nabla_k f(w_t)|, \\
b_t[k] + \sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2} &\geq \sigma.
\end{aligned}$$

Then from (26) we have

$$
\begin{aligned}
&\mathbb{E}_t\left[\left(\frac{\beta\sqrt{\eta[k]}}{\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}-\nu_t[k]\right)\nabla_k f(w_t) g_t[k]\right] \\
\leq\ & \underbrace{\beta\sqrt{\eta[k]}\mathbb{E}_t\left[\frac{\left|g_t[k]-\nabla_k f(w_t)\right|\left|\nabla_k f(w_t)\right|\left|g_t[k]\right|}{b_t[k]\sqrt{b_{t-1}^2[k]+(\nabla_k f(w_t))^2+\sigma^2}}\right]}_{\text{term I}} \\
&\underbrace{+\beta\sqrt{\eta[k]}\mathbb{E}_t\left[\frac{\sigma\left|\nabla_k f(w_t)\right|\left|g_t[k]\right|}{b_t[k]\sqrt{b_{t-1}^2[k]+(\nabla_k f(w_t))^2+\sigma^2}}\right]}_{\text{term II}}. \qquad (27)
\end{aligned}
$$

For term I in (27), we use Lemma A.1 with

$$
\begin{aligned}
\lambda &= \frac{2\sigma^2}{\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}, \\
a &= \frac{|g_t[k]|}{b_t[k]}, \\
b &= \frac{\left|g_t[k]-\nabla_k f(w_t)\right|\left|\nabla_k f(w_t)\right|}{\sqrt{b_{t-1}^2[k]+(\nabla_k f(w_t))^2+\sigma^2}},
\end{aligned}
$$

to get

$$
\begin{aligned}
&\beta\sqrt{\eta[k]}\mathbb{E}_t\left[\frac{\left|g_t[k]-\nabla_k f(w_t)\right|\left|\nabla_k f(w_t)\right|\left|g_t[k]\right|}{b_t[k]\sqrt{b_{t-1}^2[k]+(\nabla_k f(w_t))^2+\sigma^2}}\right] \\
\leq\ & \frac{\beta\sqrt{\eta[k]}\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}{4\sigma^2}\frac{\left(\nabla_k f(w_t)\right)^2\mathbb{E}_t\left[g_t[k]-\nabla_k f(w_t)\right]^2}{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2} \\
&+\frac{\beta\sqrt{\eta[k]}\sigma^2}{\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}\mathbb{E}_t\left[\frac{g_t^2[k]}{b_t^2[k]}\right] \\
\leq\ & \frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{4\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}+\beta\sqrt{\eta[k]}\sigma\mathbb{E}_t\left[\frac{g_t^2[k]}{b_t^2[k]}\right]. \qquad (28)
\end{aligned}
$$

The last inequality follows from BV. Similarly, we again use Lemma A.1 with

$$
\begin{aligned}
\lambda &= \frac{2}{\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}, \\
a &= \frac{\sigma\left|g_t[k]\right|}{b_t[k]}, \\
b &= \frac{\left|\nabla_k f(w_t)\right|}{\sqrt{b_t^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}
\end{aligned}
$$

and $\sqrt{b_t^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}\geq\sigma$ to get

$$
\begin{aligned}
\beta\sqrt{\eta[k]}\mathbb{E}_t\left[\frac{\sigma\left|\nabla_k f(w_t)\right|\left|g_t[k]\right|}{b_t[k]\sqrt{b_{t-1}^2[k]+(\nabla_k f(w_t))^2+\sigma^2}}\right] &\leq \beta\sqrt{\eta[k]}\sigma\mathbb{E}_t\left[\frac{g_t^2[k]}{b_t^2[k]}\right] \\
&\quad+\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{4\sqrt{b_{t-1}^2[k]+\left(\nabla f(w_t)\right)^2+\sigma^2}}. \qquad (29)
\end{aligned}
$$

Therefore using (28) and (29) in (28) we get

$$\mathbb{E}_t\left[\left(\frac{\beta\sqrt{\eta[k]}}{\sqrt{b_{t-1}^2[k]+(\nabla_k f(w_t))^2+\sigma^2}}-\nu_t[k]\right)\nabla_k f(w_t) g_t[k]\right] \leq 2\beta\sqrt{\eta[k]}\sigma\mathbb{E}_t\left[\frac{g_t^2[k]}{b_t^2[k]}\right] + \frac{\beta\sqrt{\eta[k]}\,(\nabla_k f(w_t))^2}{2\sqrt{b_{t-1}^2[k]+(\nabla f(w_t))^2+\sigma^2}}.$$

This completes the proof of this Lemma. □

Lemma A.7.

$$\sum_{t=0}^{T}\frac{g_t^2[k]}{b_t^2[k]} \leq \log\left(\frac{b_T^2[k]}{b_0^2[k]}\right)+1 \tag{30}$$

Proof. Using $b_t^2[k]=\sum_{\tau=0}^{t} g_\tau^2[k]$ we have

$$\begin{aligned}
\sum_{t=0}^{T}\frac{g_t^2[k]}{b_t^2[k]} &= 1+\sum_{t=1}^{T}\frac{g_t^2[k]}{b_t^2[k]} \\
&= 1+\sum_{t=1}^{T}\frac{b_t^2[k]-b_{t-1}^2[k]}{b_t^2[k]} \\
&= 1+\sum_{t=1}^{T}\frac{1}{b_t^2[k]}\int_{b_{t-1}^2[k]}^{b_t^2[k]} dz \\
&\leq 1+\sum_{t=1}^{T}\int_{b_{t-1}^2[k]}^{b_t^2[k]}\frac{dz}{z} \\
&= 1+\int_{b_0^2[k]}^{b_T^2[k]}\frac{dz}{z} \\
&= 1+\log\left(\frac{b_T^2[k]}{b_0^2[k]}\right).
\end{aligned}$$

The inequality follows from the fact $\frac{1}{b_t^2[k]} \leq \frac{1}{z}$ when $b_{t-1}^2[k] \leq z \leq b_t^2[k]$. This completes the proof of the Lemma. □

B Proof of Main Results

B.1 Proof of Proposition 2.1

Proposition B.1 (Scale invariance)**.** Suppose we solve problems (4) and (5) using algorithm (6). Then, the iterates $\hat{w}_t$ and $\hat{w}_t^V$ corresponding to (4) and (5) follow: $\forall k \in [d]$

$$\hat{w}_{t+1}[k] = \hat{w}_t[k] - \frac{\beta m_t[k]}{\sum_{\tau=0}^{t} g_\tau^2[k]} g_t[k], \tag{31}$$

$$\hat{w}_{t+1}^V[k] = \hat{w}_t^V[k] - \frac{\beta m_t[k]}{\sum_{\tau=0}^{t} \left(g_\tau^V[k]\right)^2} g_t^V[k] \tag{32}$$

with $g_\tau = \varphi_{i_\tau}'(x_{i_\tau}^\top \hat{w}_\tau) x_{i_\tau}$ and $g_\tau^V = \varphi_{i_\tau}'(x_{i_\tau}^\top V \hat{w}_\tau) V x_{i_\tau}$ for i_τ chosen uniformly from $[n]$, $\tau = 0, 1, \ldots, t$, $t \geq 0$. Moreover, updates (31) and (32) satisfy

$$\hat{w}_t = V \hat{w}_t^V, \quad V g_t = g_t^V, \quad f\left(\hat{w}_t\right) = f^V\left(\hat{w}_t^V\right)$$

for all $t \geq 0$ when $\hat{w}_0 = \hat{w}_0^V = 0 \in \mathbb{R}^d$. Furthermore we have

$$\left\|g_t^V\right\|_{V^{-2}}^2 = \left\|g_t\right\|^2. \tag{33}$$

Proof. First, we will show $\hat{w}_t = V \hat{w}_t^V$ and $V g_t = g_t^V$ using induction. Note that for $\tau = 1$ and $k \in [d]$, we get

$$\hat{w}_1[k] = \frac{-\beta m_0[k] \varphi_{i_0}'(0) x_{i_0}[k]}{\left(\varphi_{i_0}'(0) x_{i_0}[k]\right)^2} = \frac{-\beta m_0[k]}{\varphi_{i_0}'(0) x_{i_0}[k]},$$

$$\hat{w}_1^V[k] = \frac{-\beta m_0[k] \varphi_{i_0}'(0) V_{kk} x_{i_0}[k]}{\left(\varphi_{i_0}'(0) V_{kk} x_{i_0}[k]\right)^2} = \frac{-\beta m_0[k]}{\varphi_{i_0}'(0) V_{kk} x_{i_0}[k]}.$$

as $\hat{w}_0 = \hat{w}_0^V = 0$. Therefore, we have $\forall k \in [d], \hat{w}_1[k] = V_{kk} \hat{w}_1^V[k]$. This can be equivalently written as $\hat{w}_1 = V \hat{w}_1^V$, as V is a diagonal matrix. Then it is easy to check

$$V g_1 = \varphi_{i_1}'\left(x_{i_1}^\top \hat{w}_1\right) V x_{i_1} = \varphi_{i_1}'\left(x_{i_1}^\top V \hat{w}_1^V\right) V x_{i_1} = g_1^V, \tag{34}$$

where the second equality follows from $\hat{w}_1 = V \hat{w}_1^V$. Now, we assume the proposition holds for $\tau = 1, \cdots, t$. Then, we need to prove this proposition for $\tau = t + 1$. Note that, from (7) we have

$$\hat{w}_{t+1}[k] = \hat{w}_t[k] - \frac{\beta m_t[k]}{\sum_{\tau=0}^{t} g_\tau^2[k]} g_t[k] = V_{kk} \hat{w}_t^V[k] - \frac{\beta m_t[k] V_{kk}^2}{\sum_{\tau=0}^{t} \left(g_\tau^V[k]\right)^2} \frac{g_t^V[k]}{V_{kk}} = V_{kk} \hat{w}_{t+1}^V[k].$$

Here, the second last equality follows from $\hat{w}_\tau = V \hat{w}_\tau^V$ and $V g_\tau = g_\tau^V \quad \forall \tau \in [t]$, while the last equality holds due to (32). Therefore, we have $\hat{w}_{t+1} = V \hat{w}_{t+1}^V$. Then similar to (34) we get $V g_{t+1} = g_{t+1}^V$ using $\hat{w}_{t+1} = V \hat{w}_{t+1}^V$. Again, using $\hat{w}_t = V \hat{w}_t^V$, we can rewrite $f(\hat{w}_t)$ as follow

$$f(\hat{w}_t) = \frac{1}{n} \sum_{i=1}^{n} \varphi_i\left(x_i^\top \hat{w}_t\right) = \frac{1}{n} \sum_{i=1}^{n} \varphi_i\left(x_i^\top V \hat{w}_t^V\right) = f^V\left(\hat{w}_t^V\right).$$

The last equality follows from (5). This proves $f(\hat{w}_t) = f^V\left(\hat{w}_t^V\right)$. Finally using $V g_t = g_t^V$ we get

$$\left\|g_t^V\right\|_{V^{-2}}^2 = \left(g_t^V\right)^\top V^{-2} g_t^V = g_t^\top V V^{-2} V g_t = \left\|g_t\right\|^2.$$

This completes the proof of Proposition 2.1. $\square$

B.2 Proof of Lemma 2.2

Lemma B.2 (Decreasing step size)**.** For $\nu_t[k]$ defined in (11) we have

$$\nu_{t+1}[k] \leq \nu_t[k] \qquad \forall k \in [d].$$

Proof. We want to show that $\nu_{t+1}[k] \leq \nu_t[k]$. Taking square and rearranging the terms (13) is equivalent to proving

$$b_t^4[k] m_{t+1}^2[k] \leq b_{t+1}^4[k] m_t^2[k]. \tag{35}$$

Using the expansion of $m_{t+1}^2[k], b_{t+1}^2[k]$, LHS of (35) can be expanded as follow

$$b_t^4[k] m_{t+1}^2[k] \quad = \quad b_t^4[k] \left(m_t^2[k] + \eta[k] g_{t+1}^2[k] + \frac{g_{t+1}^2[k]}{b_t^2[k] + g_{t+1}^2[k]} \right). \tag{36}$$

Similarly, the RHS of (35) can be expanded to

$$\begin{aligned} b_{t+1}^4[k] m_t^2[k] \quad &= \quad m_t^2[k] \left(b_t^2[k] + g_{t+1}^2[k] \right)^2 \\ &= \quad m_t^2[k] b_t^4[k] + m_t^2[k] g_{t+1}^4[k] + 2 m_t^2[k] g_{t+1}^2[k] b_t^2[k]. \end{aligned} \tag{37}$$

Therefore using (36) and (37), inequality (35) is equivalent to

$$\begin{aligned} b_t^4[k] \left(m_t^2[k] + \eta[k] g_{t+1}^2[k] + \frac{g_{t+1}^2[k]}{b_t^2[k] + g_{t+1}^2[k]} \right) \quad \leq \quad & m_t^2[k] b_t^4[k] + m_t^2[k] g_{t+1}^4[k] \\ & + 2 m_t^2[k] g_{t+1}^2[k] b_t^2[k]. \end{aligned} \tag{38}$$

Now subtracting $m_t^2[k] b_t^4[k]$ from both sides of (38) and then multiplying both sides by $b_t^2[k] + g_{t+1}^2[k]$, (38) is equivalent to

$$\begin{aligned} \eta[k] g_{t+1}^2[k] b_t^6[k] + \eta[k] g_{t+1}^4[k] b_t^4[k] + g_{t+1}^2[k] b_t^4[k] \quad \leq \quad & m_t^2[k] g_{t+1}^4[k] b_t^2[k] + 2 m_t^2[k] g_{t+1}^2[k] b_t^4[k] \\ & + m_t^2[k] g_{t+1}^6[k] + 2 m_t^2[k] g_{t+1}^4[k] b_t^2[k]. \end{aligned} \tag{39}$$

Therefore, proving (13) is equivalent to proving (39). Note that, from the expansion $m_t^2[k] = \eta[k] b_t^2[k] + \sum_{\tau=0}^{t} \frac{g_t^2[k]}{b_t^2[k]}$, we have $m_t^2[k] \geq \frac{g_0^2[k]}{b_0^2[k]} = 1$ and $m_t^2[k] \geq \eta[k] b_t^2[k]$. Then using $m_t^2[k] \geq 1$ we get

$$g_{t+1}^4[k] b_t^2[k] \leq m_t^2[k] g_{t+1}^4[k] b_t^2[k]. \tag{40}$$

Again, using $m_t^2[k] \geq \eta[k] b_t^2[k]$, we have

$$\eta[k] g_{t+1}^2[k] b_t^6[k] + \eta[k] g_{t+1}^4[k] b_t^4[k] \quad \leq \quad m_t^2[k] g_{t+1}^2[k] b_t^4[k] + m_t^2[k] g_{t+1}^4[k] b_t^2[k]. \tag{41}$$

Then adding (40) and (41) we get

$$\eta[k] g_{t+1}^2[k] b_t^6[k] + \eta[k] g_{t+1}^4[k] b_t^4[k] + g_{t+1}^2[k] b_t^4[k] \quad \leq \quad m_t^2[k] g_{t+1}^4[k] b_t^2[k] + 2 m_t^2[k] g_{t+1}^2[k] b_t^4[k]. \tag{42}$$

Therefore, (39) is true due to (42) and $m_t^2[k] g_{t+1}^6[k] + 2 m_t^2[k] g_{t+1}^4[k] b_t^2[k] \geq 0$. This completes the proof of the Lemma. $\square$

B.3 Proof of Theorem 3.3

Theorem B.3. Suppose f is L-smooth, $g_t = \nabla f(w_t)$ and η, β are chosen such that $\nu_0[k] \leq \frac{1}{L}$ for all $k \in [d]$. Then for (11) we have

$$\min_{t \leq T} \|\nabla f(w_t)\|^2 \leq \frac{1}{T+1}\left(\sum_{k=1}^{d} b_0[k] + \frac{2(f(w_0) - f_*)}{\sqrt{\eta}\beta}\right)^2.$$

Proof. Suppose $g_t = \nabla f(w_t)$. Then using the smoothness of f we get

$$\begin{aligned} f(w_{T+1}) &\leq f(w_T) + \langle g_T, w_{T+1} - w_T \rangle + \frac{L}{2}\|w_{T+1} - w_T\|^2 \\ &= f(w_T) + \sum_{k=1}^{d} g_T[k]\left(w_{T+1}[k] - w_T[k]\right) + \frac{L}{2}\sum_{k=1}^{d}\left(w_{T+1}[k] - w_T[k]\right)^2 \\ &= f(w_T) - \sum_{k=1}^{d} \nu_T[k] g_T^2[k] + \frac{L}{2}\sum_{k=1}^{d} \nu_T^2[k] g_T^2[k] \\ &= f(w_T) - \sum_{k=1}^{d} \nu_T[k]\left(1 - \nu_T[k]\frac{L}{2}\right) g_T^2[k]. \end{aligned}$$

Then using this bound recursively we get

$$f(w_{T+1}) \leq f(w_0) - \sum_{t=0}^{T}\sum_{k=1}^{d} \nu_t[k]\left(1 - \nu_t[k]\frac{L}{2}\right) g_t^2[k].$$

Note that, we initialized KATE such that $\nu_0[k] \leq \frac{1}{L} \forall k \in [d]$. Therefore using Lemma 2.2 we have $\nu_t[k] \leq \frac{1}{L}$, which is equivalent to $1 - \nu_t[k]\frac{L}{2} \geq \frac{1}{2}$ for all $k \in [d]$. Hence from (43) we have

$$f(w_{T+1}) \leq f(w_0) - \sum_{t=0}^{T}\sum_{k=1}^{d} \frac{\nu_t[k]}{2} g_t^2[k].$$

Then rearranging the terms and using $f(w_{T+1}) \geq f_*$ we get

$$\sum_{t=0}^{T}\sum_{k=1}^{d} \frac{\nu_t[k]}{2} g_t^2[k] \leq f(w_0) - f_*. \tag{43}$$

Then from (43) and $m_t[k] \geq \sqrt{\eta_0} b_t[k]$ we get

$$\sum_{t=0}^{T}\sum_{k=1}^{d} \frac{g_t^2[k]}{b_t[k]} \leq \frac{2(f(w_0) - f_*)}{\sqrt{\eta_0}\beta}. \tag{44}$$

Now from the definition of $b_t^2[k]$, we have $b_t^2[k] = b_{t-1}^2[k] + g_t^2[k]$. This can be rearranged to get

$$\begin{aligned} b_T[k] &= b_{T-1}[k] + \frac{g_T^2[k]}{b_T[k] + b_{T-1}[k]} \\ &\leq b_{T-1}[k] + \frac{g_T^2[k]}{b_T[k]} \qquad (45) \\ &\leq b_0[k] + \sum_{t=0}^{T} \frac{g_t^2[k]}{b_t[k]}. \qquad (46) \end{aligned}$$

Here the last inequality (46) follows from recursive use of (45). Then, taking squares on both sides and summing over $k \in [d]$ we get

$$\begin{aligned} \sum_{k=1}^{d} b_T^2[k] &\leq \sum_{k=1}^{d}\left(b_0[k] + \sum_{t=0}^{T}\frac{g_t^2[k]}{b_t[k]}\right)^2 \\ &\leq \left(\sum_{k=1}^{d} b_0[k] + \sum_{t=0}^{T}\sum_{k=1}^{d}\frac{g_t^2[k]}{b_t[k]}\right)^2 \\ &\leq \left(\sum_{k=1}^{d} b_0[k] + \frac{2(f(w_0) - f_*)}{\sqrt{\eta_0}\beta}\right)^2. \qquad (47) \end{aligned}$$

The second inequality follows from $b_0[k] + \sum_{t=0}^{T} \frac{g_t^2[k]}{b_t[k]} \geq 0$ for all $k \in [d]$ and the last inequality from (44). Now note that $\sum_{t=0}^{T} \|g_t\|^2 = \sum_{t=0}^{T} \sum_{k=1}^{d} g_t^2[k] = \sum_{k=1}^{d} b_t^2[k]$. Therefore dividing both sides of (47) by $T+1$, we get

$$\min_{t \leq T} \|\nabla f(w_t)\|^2 \leq \frac{1}{T+1} \left(\sum_{k=1}^{d} b_0[k] + \frac{2(f(w_0) - f_*)}{\sqrt{\eta_0}\beta} \right)^2 .$$

This completes the proof of the theorem. □

B.4 Proof of Theorem 3.4

Theorem B.4. Suppose f is a L-smooth function and g_t is an unbiased estimator of $\nabla f(w_t)$ such that BV holds. Moreover, we assume $\|\nabla f(w_t)\|^2 \leq \gamma^2$ for all t. Then KATE satisfies

$$\min_{t\leq T} \mathbb{E}\|\nabla f(w_t)\| \quad \leq \quad \left(\frac{\|g_0\|}{T} + \frac{2(\gamma+\sigma)}{\sqrt{T}}\right)^{1/2} \sqrt{\frac{2\mathcal{C}_f}{\beta\sqrt{\eta_0}}}$$

where

$$\mathcal{C}_f = f(w_0) - f_* + \sum_{k=1}^{d}\left(2\beta\sqrt{\eta[k]}\sigma + \frac{\beta^2\eta[k]L}{2} + \frac{\beta^2 L}{2g_0^2[k]}\right)\left(\log\left(\frac{(\sigma^2+\gamma^2)T}{g_0^2[k]}\right)+1\right).$$

Proof. Using smoothness, we have

$$\begin{aligned}
f(w_{t+1}) &\leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t\rangle + \frac{L}{2}\|w_{t+1}-w_t\|^2 \\
&= f(w_t) + \sum_{k=1}^{d} \nabla_k f(w_t)\left(w_{t+1}[k] - w_t[k]\right) + \frac{L}{2}\sum_{k=1}^{d}\left(w_{t+1}[k]-w_t[k]\right)^2 \\
&= f(w_t) - \sum_{k=1}^{d}\nu_t[k]\nabla_k f(w_t) g_t[k] + \frac{L}{2}\sum_{k=1}^{d}\nu_t^2[k]g_t^2[k].
\end{aligned}$$

Then, taking the expectation conditioned on w_t, we have

$$\begin{aligned}
\mathbb{E}_t\left[f(w_{t+1})\right] &\leq f(w_t) - \sum_{k=1}^{d}\mathbb{E}_t\left[\nu_t[k]\nabla_k f(w_t)g_t[k]\right] + \frac{L}{2}\sum_{k=1}^{d}\mathbb{E}_t\left[\nu_t^2[k]g_t^2[k]\right] \\
&= f(w_t) - \sum_{k=1}^{d}\mathbb{E}_t\left[\nu_t[k]\nabla_k f(w_t)g_t[k]\right] + \frac{L}{2}\sum_{k=1}^{d}\mathbb{E}_t\left[\nu_t^2[k]g_t^2[k]\right] \\
&\quad - \sum_{k=1}^{d}\frac{\beta\sqrt{\eta[k]}}{\sqrt{b_{t-1}^2[k] + \left(\nabla_k f(w_t)\right)^2 + \sigma^2}}\mathbb{E}_t\left[\nabla_k f(w_t)\left(\nabla_k f(w_t) - g_t[k]\right)\right] \\
&= f(w_t) + \sum_{k=1}^{d}\mathbb{E}_t\left[\left(\frac{\beta\sqrt{\eta[k]}}{\sqrt{b_{t-1}^2[k] + \left(\nabla_k f(w_t)\right)^2+\sigma^2}} - \nu_t[k]\right)\nabla_k f(w_t) g_t[k]\right] \\
&\quad + \frac{L}{2}\sum_{k=1}^{d}\mathbb{E}_t\left[\nu_t^2[k]g_t^2[k]\right] - \sum_{k=1}^{d}\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}.
\end{aligned}$$

The second last equality follows from $\mathbb{E}_t\left[\nabla_k f(w_t)\left(\nabla_k f(w_t) - g_t[k]\right)\right] = \nabla_k f(w_t)\left(\nabla_k f(w_t) - \mathbb{E}_t\left[g_t[k]\right]\right) = 0$. Now we use (25) to get

$$\begin{aligned}
\mathbb{E}_t\left[f(w_{t+1})\right] &\leq f(w_t) + \sum_{k=1}^{d} 2\beta\sqrt{\eta[k]}\sigma\mathbb{E}_t\left[\frac{g_t^2[k]}{b_t^2[k]}\right] + \frac{L}{2}\sum_{k=1}^{d}\mathbb{E}_t\left[\nu_t^2[k]g_t^2[k]\right] \\
&\quad - \sum_{k=1}^{d}\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{2\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}.
\end{aligned}$$

Then rearranging the terms we have

$$\begin{aligned}
\sum_{k=1}^{d}\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{2\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}} &\leq f(w_t) - \mathbb{E}_t\left[f(w_{t+1})\right] + \sum_{k=1}^{d}2\beta\sqrt{\eta[k]}\sigma\mathbb{E}_t\left[\frac{g_t^2[k]}{b_t^2[k]}\right] \\
&\quad + \frac{L}{2}\sum_{k=1}^{d}\mathbb{E}_t\left[\nu_t^2[k]g_t^2[k]\right].
\end{aligned}$$

Now we take the total expectations to derive

$$\sum_{k=1}^{d}\mathbb{E}\left[\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{2\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}\right] \leq \mathbb{E}\left[f(w_t)\right]-\mathbb{E}\left[f(w_{t+1})\right]+\sum_{k=1}^{d}2\beta\sqrt{\eta[k]}\sigma\mathbb{E}\left[\frac{g_t^2[k]}{b_t^2[k]}\right] +\frac{L}{2}\sum_{k=1}^{d}\mathbb{E}\left[\nu_t^2[k]g_t^2[k]\right].$$

The above inequality holds for any t. Therefore summing up from $t=0$ to $t=T$ and using $f(w_{T+1})\geq f_*$ we get

$$\sum_{t=0}^{T}\sum_{k=1}^{d}\mathbb{E}\left[\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{2\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}\right] \leq f(w_0)-f_*+\sum_{t=0}^{T}\sum_{k=1}^{d}2\beta\sqrt{\eta[k]}\sigma\mathbb{E}\left[\frac{g_t^2[k]}{b_t^2[k]}\right] +\frac{L}{2}\sum_{t=0}^{T}\sum_{k=1}^{d}\mathbb{E}\left[\nu_t^2[k]g_t^2[k]\right]. \tag{48}$$

Note that, using the expansion of $\nu_t^2[k]$ we have

$$\begin{aligned}
\nu_t^2[k] &= \frac{\beta^2\eta[k]b_t^2[k]+\beta^2\sum_{j=0}^{t}\frac{g_j^2[k]}{b_j^2[k]}}{b_t^4[k]}\\
&= \frac{\beta^2\eta[k]}{b_t^2[k]}+\frac{\beta^2}{b_t^4[k]}\sum_{j=0}^{t}\frac{g_j^2[k]}{b_j^2[k]}\\
&\leq \frac{\beta^2\eta[k]}{b_t^2[k]}+\frac{\beta^2}{b_t^4[k]b_0^2[k]}\sum_{j=0}^{t}g_j^2[k] \qquad (49)\\
&= \frac{\beta^2\eta[k]}{b_t^2[k]}+\frac{\beta^2}{b_t^2[k]g_0^2[k]}. \qquad (50)
\end{aligned}$$

Here (49) follows from $b_j^2[k]\geq b_0^2[k]$ and (50) from $b_t^2[k]=\sum_{j=0}^{t}g_j^2[k]$. Then using (50) in (48) we derive

$$\begin{aligned}
\sum_{t=0}^{T}\sum_{k=1}^{d}\mathbb{E}\left[\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{2\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}\right] &\leq f(w_0)-f_*+\sum_{t=0}^{T}\sum_{k=1}^{d}\left(2\beta\sqrt{\eta[k]}\sigma+\frac{\beta^2\eta[k]L}{2}+\frac{\beta^2L}{2g_0^2[k]}\right)\mathbb{E}\left[\frac{g_t^2[k]}{b_t^2[k]}\right]\\
&\leq f(w_0)-f_*\\
&\quad+\sum_{k=1}^{d}\left(2\beta\sqrt{\eta[k]}\sigma+\frac{\beta^2\eta[k]L}{2}+\frac{\beta^2L}{2g_0^2[k]}\right)\mathbb{E}\left[\log\left(\frac{b_T^2[k]}{b_0^2[k]}\right)+1\right].
\end{aligned}$$

Here the last inequality follows from (30). Now using Jensen's Inequality (21) with $\Psi(z)=\log(z)$ we have

$$\begin{aligned}
\sum_{t=0}^{T}\sum_{k=1}^{d}\mathbb{E}\left[\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{2\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}\right] &\leq f(w_0)-f_*\\
&\quad+\sum_{k=1}^{d}\left(2\beta\sqrt{\eta[k]}\sigma+\frac{\beta^2\eta[k]L}{2}+\frac{\beta^2L}{2g_0^2[k]}\right)\left(\log\left(\frac{\mathbb{E}\left[b_T^2[k]\right]}{b_0^2[k]}\right)+1\right).
\end{aligned}$$

Now note that $\mathbb{E}\left[b_T^2[k]\right]=\sum_{t=0}^{T}\mathbb{E}\left[g_t^2[k]\right]=\sum_{t=0}^{T}\mathbb{E}\left[g_t[k]-\nabla_k f(w_t)\right]^2+\left(\nabla_k f(w_t)\right)^2\leq(\sigma^2+\gamma^2)T$. Therefore, we have the bound

$$\sum_{t=0}^{T}\sum_{k=1}^{d}\mathbb{E}\left[\frac{\beta\sqrt{\eta[k]}\left(\nabla_k f(w_t)\right)^2}{2\sqrt{b_{t-1}^2[k]+\left(\nabla_k f(w_t)\right)^2+\sigma^2}}\right] \leq f(w_0)-f_*+2\beta\sigma\sum_{k=1}^{d}\sqrt{\eta[k]}\log\left(\frac{e(\sigma^2+\gamma^2)T}{b_0^2[k]}\right) +\sum_{k=1}^{d}\left(\frac{\beta^2\eta[k]L}{2}+\frac{\beta^2L}{2g_0^2[k]}\right)\log\left(\frac{e(\sigma^2+\gamma^2)T}{b_0^2[k]}\right). \tag{51}$$

Here the RHS is exactly $\mathcal{C}_f$. Using (22) we have

$$\sum_{k=1}^{d} \frac{(\nabla_k f(w_t))^2}{\sqrt{b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}} \geq \frac{\sum_{k=1}^{d} (\nabla_k f(w_t))^2}{\sqrt{\sum_{k=1}^{d} b_{t-1}^2[k] + (\nabla_k f(w_t))^2 + \sigma^2}}$$

$$= \frac{\|\nabla f(w_t)\|^2}{\sqrt{\|b_{t-1}\|^2 + \|\nabla f(w_t)\|^2 + d\sigma^2}}. \tag{52}$$

Therefore using (52) in (51) we arrive at

$$\sum_{t=0}^{T} \mathbb{E}\left[\frac{\|\nabla f(w_t)\|^2}{\sqrt{\|b_{t-1}\|^2 + \|\nabla f(w_t)\|^2 + d\sigma^2}}\right] \leq \frac{2\mathcal{C}_f}{\beta\sqrt{\eta_0}}. \tag{53}$$

Now we use Holder's Inequality (20) $\frac{\mathbb{E}(XY)}{\left(\mathbb{E}|Y|^3\right)^{\frac{1}{3}}} \leq \left(\mathbb{E}|X|^{\frac{3}{2}}\right)^{\frac{2}{3}}$ with

$$X = \left(\frac{\|\nabla f(w_t)\|^2}{\sqrt{\|b_{t-1}\|^2 + \|\nabla f(w_t)\|^2 + d\sigma^2}}\right)^{\frac{2}{3}} \quad \text{and} \quad Y = \left(\sqrt{\|b_{t-1}\|^2 + \|\nabla f(w_t)\|^2 + d\sigma^2}\right)^{\frac{2}{3}}$$

to get a lower bound on LHS of (53):

$$\mathbb{E}\left[\frac{\|\nabla f(w_t)\|^2}{\sqrt{\|b_{t-1}\|^2 + \|\nabla f(w_t)\|^2 + d\sigma^2}}\right] \geq \frac{\mathbb{E}\left[\|\nabla f(w_t)\|^{\frac{4}{3}}\right]^{\frac{3}{2}}}{\sqrt{\mathbb{E}\left(\|b_{t-1}\|^2 + \|\nabla f(w_t)\|^2 + d\sigma^2\right)}}$$

$$\geq \frac{\mathbb{E}\left[\|\nabla f(w_t)\|^{\frac{4}{3}}\right]^{\frac{3}{2}}}{\sqrt{\|b_0\|^2 + 2t(\gamma^2 + d\sigma^2)}}. \tag{54}$$

Therefore from (53) and (54) we get

$$\frac{T}{\sqrt{\|b_0\|^2 + 2T(\gamma^2 + d\sigma^2)}} \min_{t \leq T} \mathbb{E}\left[\|\nabla f(w_t)\|^{\frac{4}{3}}\right]^{\frac{3}{2}} \leq \frac{2\mathcal{C}_f}{\beta\sqrt{\eta_0}}.$$

Then multiplying both sides by $\frac{\|b_0\| + \sqrt{2T}(\gamma + \sqrt{d}\sigma)}{T}$ we have

$$\min_{t \leq T} \mathbb{E}\left[\|\nabla f(w_t)\|^{\frac{4}{3}}\right]^{\frac{3}{2}} \leq \left(\frac{\|b_0\|}{T} + \frac{2(\gamma + \sigma)}{\sqrt{T}}\right) \frac{2\mathcal{C}_f}{\beta\sqrt{\eta_0}}.$$

Here we use $\mathbb{E}\left[\|\nabla f(w_t)\|\right]^{\frac{4}{3}} \leq \mathbb{E}\left[\|\nabla f(w_t)\|^{\frac{4}{3}}\right]$ (follows from Jensen's Inequality (21) with $\Psi(z) = z^{4/3}$) in the above equation to get

$$\min_{t \leq T} \mathbb{E}\left[\|\nabla f(w_t)\|\right]^2 \leq \left(\frac{\|b_0\|}{T} + \frac{2(\gamma + \sigma)}{\sqrt{T}}\right) \frac{2\mathcal{C}_f}{\beta\sqrt{\eta_0}}.$$

This completes the proof of the Theorem. $\square$

C Additional Experiments: Scale-Invariance Verification

In this experiment, we implement KATE on problems (4) (for unscaled data) and (5) (for scaled data) with

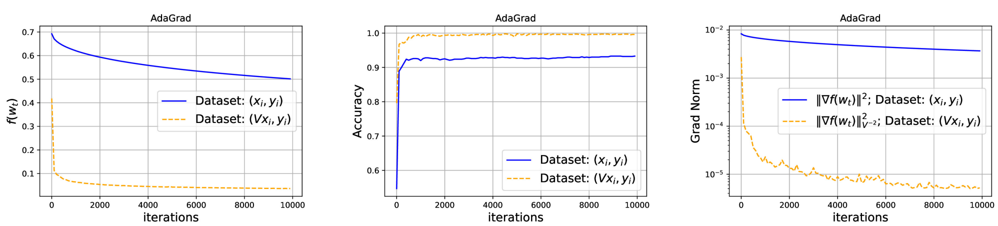
$$\varphi_i(z) = \log\left(1 + \exp\left(-y_i z\right)\right).$$

We generate the data similar to Section 4.1.1. We run KATE for $10,000$ iterations with mini-batch size 10, $\eta = {}^{1}/_{(\nabla f(w_0))^2}$ and plot functional value $f(w_t)$ and accuracy in Figures 11a and 11b. We use the proportion of correctly classified samples to compute accuracy, i.e. $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\left\{y_i x_i^\top w_t \geq 0\right\}}$.

In plots 11a and 11b, the functional value and accuracy of KATE coincide, which aligns with our theoretical findings (Proposition 2.1). Figure 11c plots $\|\nabla f(w_t)\|^2$ and $\|\nabla f(w_t)\|^2_{V^{-2}}$ for unscaled and scaled data respectively. Here, (10) explains the identical values taken by the corresponding gradient norms of KATE iterates for the scaled and unscaled data. Similarly, in Figure 12, we compare the performance of AdaGrad on scaled and un-scaled data. This figure illustrates the lack of the scale-invariance for AdaGrad.

(a) Plot of $f(w_t)$ (b) Plot of Accuracy (c) Plot of Gradient Norm

Figure 11: Comparison of KATE on scaled and un-scaled data. Figures 11a, and 11b plot the functional value $f(w_t)$ and accuracy on scaled and unscaled data, respectively. Figure 11c plots $\|\nabla f(w_t)\|^2$ and $\|\nabla f(w_t)\|^2_{V^{-2}}$ for unscaled and scaled data respectively.



(a) Plot of $f(w_t)$ (b) Plot of Accuracy (c) Plot of Gradient Norm

Figure 12: Comparison of AdaGrad on scaled and un-scaled data. Figures 12a, and 12b plot the functional value $f(w_t)$ and accuracy on scaled and unscaled data, respectively. Figure 12c plots $\|\nabla f(w_t)\|^2$ and $\|\nabla f(w_t)\|^2_{V^{-2}}$ for unscaled and scaled data respectively.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the new method and its scale invariance property are introduced in Section 2, main convergence results are provided in Section 3, and the numerical results are provided in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: see Section 3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: see Section 3 and the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: see Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: see Section 4.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: see Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: the results are consistent for different runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have added all the details in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: the paper follows NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: the paper is mostly theoretical and does not have a direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: we do not release data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: see Section 4.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: not applicable.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.