Beyond Worst-Case Analysis in Stochastic Approximation: Moment Estimation Improves Instance Complexity

Jingzhao Zhang [1] **Hongzhou Lin** [2] **Subhro Das** [3] **Suvrit Sra** [4] **Ali Jadbabaie** [4]

Abstract

We study oracle complexity of gradient based methods for stochastic approximation problems. Though in many settings optimal algorithms and tight lower bounds are known for such problems, these optimal algorithms do not achieve the best performance when used in practice. We address this theory-practice gap by focusing on *instance-dependent complexity* instead of worst case complexity. In particular, we first summarize known instance-dependent complexity results and categorize them into three levels. We identify the domination relation between different levels and propose a fourth instance-dependent bound that dominates existing ones. We then provide a sufficient condition according to which an adaptive algorithm with moment estimation can achieve the proposed bound without knowledge of noise levels. Our proposed algorithm and its analysis provide a theoretical justification for the success of moment estimation as it achieves improved instance complexity.

1. Introduction

Stochastic approximation (SA) methods, introduced and analyzed in the seminal works (Robbins & Monro, 1951; Fabian et al., 1968; Polyak, 1987; Kushner & Yin, 2003; Moulines & Bach, 2011), are central to modern machine learning algorithms. Due to their wide applicability and importance, they have been extensively studied. It is now well established that an algorithm as simple as stochastic gradient descent achieves minimax optimal rates in many settings (Nesterov, 2013; Agarwal et al., 2009; Fang et al., 2018a; Arjevani et al., 2019; Lin et al., 2015; Allen-Zhu, 2016; Ghadimi & Lan, 2012).

[1]IIIS, Tsinghua University [2]Amazon [3]MIT-IBM Watson AI Lab, IBM Research [4]Massachusetts Institute of Technology. Correspondence to: Jingzhao Zhang <jingzhaoz@mail.tsinghua.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

These well-studied stochastic methods are, however, optimal in the sense of a worst case measure of complexity:

$$\inf_{A_\theta \in \mathcal{A}} \sup_{x_1 \in \mathbb{R}^d, f \in \mathcal{F}} T_\epsilon(A_\theta[f, x_1], f), \tag{1}$$

where $T_\epsilon(A_\theta[f, x_1], f)$ denotes the number of iterations to obtain an ϵ-solution initialising at x_1. In the most common setting of convex problems, the ϵ-solution is measured by the function suboptimality gap:

$$T_\epsilon(\{x_t\}_{t\in\mathbb{N}}, f) := \inf\{t \in \mathbb{N} | f(x_t) - f(x^*) \leq \epsilon\}. \tag{2}$$

The worst case complexity measure is naturally achieved in certain extreme cases within a function class, which may be quite rare in practice. Empirical observations (Defazio & Bottou, 2018; Reddi et al., 2019) suggests that a gap does exist between worst case analysis and empirical performance.

To reduce the mistmatch between theory and practice, it is necessary to go beyond worst case scenarios, and to include more domain-specific nuances. One alternative is to consider average case complexity analysis. Such an approach is well established in theoretical computer science, and is used to explain the superior empirical performance of QuickSort (Hoare, 1962), quickhull (Preparata & Shamos, 2012), simplex methods (Borgwardt, 2012). In contrast, the use of average-case complexity analysis in optimization scenarios is less mature and in its infancy (Pedregosa & Scieur, 2020; Lacotte & Pilanci, 2020; Paquette et al., 2021).

Another approach to move away from the worst-case is to conduct smoothed analysis (Spielman, 2005), in which the complexity is instance-dependent (Fagin et al., 2003; Afshani et al., 2017). Here, instance could be referring to a specific sample, or more broadly to an easily parametrizable subclass of samples. The goal is to build a middle ground between worst-case and average-case analyses, such that we can eventually interpolate between them (Spielman, 2005). This idea has been recently extended to the study of complexity in Q-learning (Khamaru et al., 2021; Pananjady & Wainwright, 2020), and further analyzed in Markovian linear stochastic approximation setting (Mou et al., 2021).

Following this line of work, we propose an *instance-dependent analysis* in the general convex setup, extending

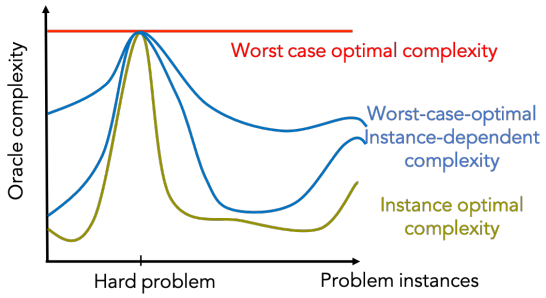

Figure 1. Different notions of complexity and optimality.

the linear setting and worst-case optimal rates in (Mou et al., 2021). We define instances by the subclass of functions with non-stationary noise, and make the following contributions:

- We categorize previous instance complexities into three categories based on their dependence on the iteration-wise noise levels as well as the information available to the algorithm. We further identify the gap between different types of instance complexities.
- We propose a new instance dependent rate, named *dynamic error bound*, that dominates (see Definition 2.5) previous ones when additional information of the problem instance is available.
- We study an algorithm based on moment estimation that achieves the *dynamic error bound* without knowing additional information when the total variation in noise level is bounded. We hence provide the first theoretical justification for why moment estimation speeds up gradient methods empirically.

We test the noise variation and empirical performance of our moment estimation algorithm on policy optimization and neural network training. We conclude by discussing numerous open problems that this work raises.

2. Problem Setup and Instance Complexity

In this work, we focus on stochastic approximation problems within the convex smooth function class $\mathcal{F}_{L,R}$:

Assumption 2.1. A continuously differentiable function f is in $\mathcal{F}_{L,R}$ if f is convex, $L-$smooth, and bounded below,

$$\begin{aligned} &f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \\ &\|x_1 - x^*\| \leq R, \\ &\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \end{aligned}$$

where x^* denotes a global optimal solution.

We follow the classical stochastic approximation setup (*à la* Robbins-Monro), where **the noise is additive**. More precisely, the stochastic gradient oracle g is given as the sum of the actual gradient ∇f and random noise ξ:

$$g(x) = \nabla f(x) + \xi. \tag{3}$$

This setup holds whenever the error comes from an exogenous source. In pursuit of our goal to go beyond worst case analysis, we follow the recent line of work on instance-complexity for stochastic approximation and reinforcement learning (Khamaru et al., 2021; Pananjady & Wainwright, 2020; Mou et al., 2021), and define the $\{\sigma_k\}_{k\in\mathbb{N}}$-instance as the following subclass of gradient oracles:

Definition 2.2 ($\{\sigma_k\}_{k\in\mathbb{N}}$-instance)**.** At the k^{th} iteration, the gradient oracle g_k returns a stochastic gradient

$$g_k(x) = \nabla f(x) + \xi_k,$$

where ξ_k is zero mean and has bounded second moment σ_k:

$$\mathbb{E}[\xi_k] = 0, \quad \mathbb{E}[\|\xi_k\|^2] = \sigma_k^2 \leq M^2.$$

For simplicity we assume that the noise is nontrivial:

Assumption 2.3 (Nontrivial noise)**.**

$$\sigma_k^2 > L^2R^2/T, \text{ for all } k \leq T.$$

Indeed, when the noise level is too small, g_k is essentially the full gradient ∇f, leading to a deterministic style convergence. Assumption 2.3 is not necessary but saves us from discussing such trivial cases, which deviate from the main purpose of studying instance-dependent bounds.

Unlike standard analysis where uniform variance bound is imposed on all the stochastic gradients, we allow the noise level to be non-stationary, i.e., can vary from iteration to iteration. To recover the worst case analysis, it suffices to set the noise at a maximum level, $\forall k, \sigma_k = M$. By allowing the noise to be non-stationary, we create a fine-grained middle ground where we can investigate how the variation of noise σ_k influence the convergence rate of an optimization algorithm.

As in most convex problems, we study the level of sub-optimality achieved after querying the oracle T times to upper bound $\mathbb{E}[f(x_T) - \min_x f(x)]$. In the following subsection, we derive *theoretical convergence rates under a* $\{\sigma_k\}_{k\in\mathbb{N}}$*-instance, named as the instance complexity*.

2.1. Different levels of instance complexity

In this subsection, we categorize rates in stochastic approximation into four levels based on their dependence on the problem instance defined by $\{\sigma_i\}_{i\in\mathbb{N}}$. **Our proposed categorization later enables us to identify a nontrivial gap between vanilla stochastic gradient descent and its adaptive variant using moment estimation**. Although we focus

	Worst	Agnostic	Adaptive	Dynamic
Error bound	$\frac{2RM}{\sqrt{T}}$	$(R^2 + \frac{1}{T}\sum_k \sigma_k^2)/\sqrt{T}$	$2R\left(\frac{1}{T}\sum_{k=1}^{T}\sigma_k^2\right)^{1/2}/\sqrt{T}$	$2R\left(\frac{1}{T}\sum_{k=1}^{T}\frac{1}{\sigma_k}\right)^{-1}/\sqrt{T}$
η_k	$R/\sqrt{TM^2}$	$1/\sqrt{T}$	$R/\sqrt{\sum_{k=1}^{T}\sigma_t^2}$ or $R/\sqrt{2\sum_{\tau\le k}\|g_k\|^2}$	$R/(\sigma_k\sqrt{T})$
Can be achieved via	Fixed step, known R, M	Fixed step, unknown R, M	Fixed step, known $R, \{\sigma_k\}_k$ or Adapt. step, unknown $\{\sigma_k\}_k$	Adaptive step, known $R, \{\sigma_k\}_k$

Table 1. Comparisons of four types of instance-dependent theoretical convergence rates under the noise level $\{\sigma_k\}_k$. The parameter R^2 denotes the initial distance $\|x_1 - x^*\|^2$ and the parameter M is an upper bound of σ_k. The convergence rate improves from the left to the right monotonically, with different choices of stepsizes η_k.

on analysis for convex functions, similar categorization may also hold for analyses of nonconvex functions.

We begin by summarizing known results that are dependent on iteration-wise noise levels $\{\sigma_k\}_k$. Previous results on instance-dependent bounds usually do not consider the exact same problem setup, and hence not directly comparable. To have a unified framework, we view these results under the smooth-convex setting. Recall the following folklore result:

Theorem 2.4. *When $f \in \mathcal{F}_{L,R}$ (convex, L-smooth), the stochastic gradient descent algorithm of the form*

$$x_{k+1} = x_k - \eta_k g(x_k),$$

with predetermined stepsizes $\eta_k \le \frac{1}{L}$, satisfies the suboptimality bound

$$\mathbb{E}[f(\overline{x}_T) - f^*] \le \frac{R^2 + \sum_{k=1}^{T}\eta_k^2\sigma_k^2}{\sum_{k=1}^{T}\eta_k}, \tag{4}$$

where $\overline{x}_T = (\sum_{k=1}^{T}\eta_k x_k)/(\sum_{k=1}^{T}\eta_k)$ is the weighted average of iterates.

The above theorem already shows that different step size choices can induce a variety of instance complexities, and some choices might give faster convergence. However, better step size choice might require more information about the problem instance. We will now discuss how different step size choice requires different level of information and classify previous work based on their step size strategy.

1. Worst case bounds. In classical stochastic optimization literature (e.g. (Rakhlin et al., 2011; Ghadimi & Lan, 2012; Ghadimi et al., 2013; Fang et al., 2018a)), only an upper bound on the variances is given. In other words, we do not have access to the entire sequence of σ_k but only an upper bound $M \ge \sigma_k, \forall k$. Using the stepsize $\eta_k = R/\sqrt{TM^2}$, this choice results in the worst case bound:

$$\mathbb{E}[f(\overline{x}_T) - f^*] \le 2RM/\sqrt{T} := \epsilon_{\text{worst}}. \tag{5}$$

2. Agnostic bounds. Another type of bound appears in the analysis of adaptive methods, and involves an additive bound on the initialization ($\|x_0 - x^*\|, f(x_0) - f(x^*)$) and noise (see Theorem 2.1 of (Ward et al., 2018), or (Zhou et al., 2018)) instead of a multiplicative one as in other bounds.

Though these bounds are usually achieved by adaptive methods, to distinguish them from other instance-dependent bounds, we call these bounds "*agnostic bounds*" because they can be obtained using a step size without knowing the maximum noise level, the smoothness constant or the domain diameter of the problem. In smooth convex case, these bounds can be achieved by setting $\eta_k = \frac{1}{\sqrt{T}}$:

$$\mathbb{E}[f(\overline{x}_T) - f^*] \le (R^2 + \tfrac{1}{T}\sum_{k=1}^{T}\sigma_k^2)/\sqrt{T} := \epsilon_{\text{agnostic}}. \tag{6}$$

3. Adaptive bounds. One theoretically justified advantage of adaptive gradient methods is that if the smoothness and level of suboptimality are known, then adaptive methods can automatically adjust the learning rate for the noise level in stochastic gradients. Results of this type can be found for example in (Duchi et al., 2011; Levy et al., 2018; Ward et al., 2018; Zhou et al., 2018). In particular, the setup in (Levy et al., 2018) (Thm 2.1 and equation (11)) subsumes our smooth-convex setup and is directly comparable. In particular, by setting $\eta_k = R(2\sum_{\tau\le k}\|g_k\|^2)^{-1/2}$, one gets the following rate:

$$\mathbb{E}[f(\overline{x}_T) - f^*] \le \frac{2R}{\sqrt{T}}\left(\tfrac{1}{T}\sum_{k=1}^{T}\sigma_k^2\right)^{\frac{1}{2}} := \epsilon_{\text{adaptive}}. \tag{7}$$

	Worst	Agnostic	Adaptive	Dynamic
Error bound	$T^{-\frac{1}{2}}$	$T^{-\frac{1}{2}}$	$T^{-((1+\alpha)/2)}$	$T^{-((1+2\alpha)/2)}$

Table 2. Comparison of convergence rate under multiple shapes of the noise level σ_k, where constant terms are omitted. The parameter $\alpha \in [0, \frac{1}{2}]$ is the exponent controlling the ratio between $\max \sigma_k$ and $\min \sigma_k$, indicating the significance in noise changes.

Notice that results in (Mou et al., 2021) also fall in this category. Though the bounds in (Mou et al., 2021) may not be directly comparable as they are in the strongly convex regime, these bounds, as well as the adaptive bound in (7) can be achieved by using a fixed step size if the noise levels σ_k are known in advance by setting $\eta_k = R\left(\sum_{\tau=1}^{T} \sigma_\tau^2\right)^{-1/2}$. This is the best attainable rate using a fixed step size by optimizing the error bound in Theorem 2.4.

This observation naturally leads to the bounds in the next part, which are achieved by an iteration dependent step size when σ_k are known. From the following result, we will see that the adaptive bounds above are **worst-case optimal** but **not instance-level optimal** (see Figure 1).

4. Dynamic bounds. A natural question is whether we can further improve adaptive bounds by appropriately selecting the stepsizes. This is indeed possible by setting a stepsize inversely proportional to σ_k. More concretely, when setting $\eta_k = R/(\sigma_k\sqrt{T})$, we get

$$\mathbb{E}[f(\overline{x}_T) - f^*] \leq 2R\left(\frac{1}{T}\sum_{k=1}^{T}\frac{1}{\sigma_k}\right)^{-1}/\sqrt{T} := \epsilon_{\text{dynamic}}. \tag{8}$$

We name these bounds "*dynamic bounds*", analogous to "dynamic regret" in online learning (Jadbabaie et al., 2015; Yang et al., 2016), indicating that the optimal policy can have iteration dependent action (i.e. iteration-wise step sizes) rather than a fixed action (i.e. a fixed step size).

Such bounds are less studied, as previous works mainly focus on the dependence of T instead of the fine-grained dependency on the instance $\{\sigma_k\}_k$. In our setting, it is important to notice that different dependency on $\{\sigma_k\}_k$ could lead to non-trivial differences in the convergence rate. For instance, the dynamic bound depends on the harmonic sum of σ_k, whereas the adaptive bound depends on its 2-norm. To make comparison between different error bounds, we adopt the following notation.

Definition 2.5 (Dominating bounds)**.** We say an error bound dominates another, denoted as $\epsilon_1 \preceq \epsilon_2$, if there exists an absolute constant $c > 0$ such that for any $R, \{\sigma_k\}_k, T$, we have that the instantiated value $\epsilon_1(R, T, \{\sigma_k\}_k) \leq c\epsilon_2(R, T, \{\sigma_k\}_k)$.

With the above definition, we could provide the following order of instance complexities.

Lemma 2.6. *The four different types of bounds satisfy the following ordering:*

$$\epsilon_{dynamic} \preceq \epsilon_{adaptive} \preceq \min\{\epsilon_{worst}, \epsilon_{agnostic}\}.$$

The above result is expected. We see in Table 1 that the dominance between different bounds is ordered the same way the amount of information available to the algorithm.

Another straightforward consequence of the lemma is that both $\epsilon_{\text{adaptive}}$ and $\epsilon_{\text{dynamic}}$ are worst case optimal. In other words, they all fall into the same rate when $\sigma_k = M$ is constant. Similarly, when all the ratios σ_i/σ_j are bounded by a constant, then the ratio of these rates are bounded, i.e. differ only by a constant ratio.

A more interesting question to ask is when does the adaptive rate $\epsilon_{\text{adaptive}}$ or the dynamic rate $\epsilon_{\text{dynamic}}$ non trivially improve the worst case rate. This can happen when the variation in the noise level $\{\sigma_k\}_k$ is comparable to the total iteration number T. In fact, **we will later show that gradient methods with moment estimation can achieve such bounds without knowing the noise level in advance, and hence improve over vanilla SGD.** To make the discussion concrete, we first use a synthetic example to illustrate the phenomenon.

2.2. Complexity comparison via synthetic example

In this section, we use an example with synthetic noise to illustrate the instance complexity of different types.

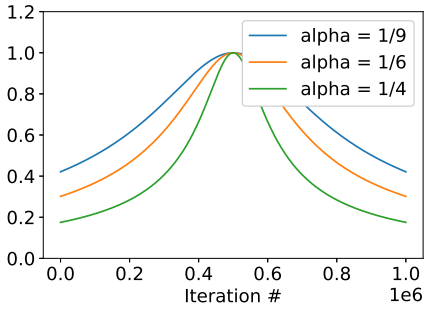

Figure 2. Mountain shape noise for different values of α.

Example 2.7. We consider a noise sequence σ_k of a mountain shape, illustrated in Figure 2. The noise level increases smoothly in the first half of iterations, then decreases in the

second half, parametrized by a positive constant $\alpha \in (0,1)$:

$$\sigma_k = \frac{1}{\sqrt{1+T^{2\alpha}\left(\frac{2k}{T}-1\right)^2}}.$$

In particular, the maximum of σ_k is 1, attained in the middle of the iterations; and the minimum of it is $T^{-\alpha}$. Substituting the values of σ_k into the rates summarized in Table 1, we obtain the rates shown in Table 2.

For this specific $\{\sigma_k\}$-instance, adapting the noise level is beneficial. As we can see, the adaptive bound improves the worst case bound by $T^{\alpha/2}$, and, the dynamic bound improves further the adaptive bound by $T^{\alpha/2}$. Such a non-trivial gap is unobservable under worst case analysis.

Though dynamic error bounds can achieve faster convergence than standard worst case rates (in terms of T), we note that achieving better dependence on T is not the main purpose of our work. *Instead, we want to show via different exponents on T that the gap between different types of instance complexity is significant and cannot be considered as off by an absolute constant.*

We showed that using a step size that is inversely proportional to the noise level can generate a better instance-dependent bound. Such a construction relies on the knowledge of σ_k. However, in practice, we usually do not have access to the exact noise level. We show in the next section that combining the idea of moment estimation with SGD can achieve the dynamic error bound even when the $\{\sigma_k\}_k$ are unknown.

3. Moment Estimation Achieves Dynamic Error Bounds

The main difficulty to achieve the dynamic bounds lies in the estimation of σ_k. A naive first idea is to draw multiple samples of the gradient at each iteration and perform an empirical estimate of the variance. Later use the estimate $\widehat{\sigma}_k$ to update the stepsizes. In order to ensure nonasymptotic concentration of the estimator, one convenient way is to impose a bounded higher moment assumption.

Assumption 3.1 (Higher moment boundedness.)**.** We assume that the fourth moment of g_k is bounded, namely, $\mathbb{E}[\|\xi_k\|^4] = \mathbb{E}[\|g_k - \nabla f\|^4] \le M^4$ for all k.

This assumption allows us to guarantee the estimation of σ_k^2 is non-asymptotically concentrated with high probability. In particular, given N samples of the gradients $g_{k,1}, g_{k,2}, \cdots g_{k,N}$ at the k-th iteration, the standard variance estimator

$$Y_k = \frac{\sum_{i=1}^N \|g_{k,i}-\overline{g}_k\|^2}{N-1}, \text{ where } \overline{g}_k = \frac{1}{N}\sum_{i=1}^N g_{k,i}, \quad (9)$$

is an unbiased, i.e. $\mathbb{E}[Y_k] = \sigma_k^2$. Moreover, its variance is bounded: $\mathrm{Var}[Y_k^2] = \frac{1}{N}\left(\mathbb{E}[\|\xi_k\|^4] - \frac{N-3}{N-1}\sigma_k^4\right) \le \frac{M^4}{N}$.

Algorithm 1 Moment Estimation SGD (x_1, T, c, m)

1: Initialize $\hat{\sigma}_1 = \frac{\|g_{1,1}-g_{1,2}\|^2}{2}$, where $g_{1,1}, g_{1,2}$ are two independent stochastic gradients at x_1.
2: **for** $t = 1, 2, ..., T/2$ **do**
3: Query two stochastic gradients $g_{t,1}, g_{t,2}$ at x_t.
4: Update

$$\overline{g_t} = \frac{g_{t,1}+g_{t,2}}{2} \text{ and } \eta_t = \frac{c}{\hat{\sigma}_t + m}$$
$$x_{t+1} = x_t - \eta_t \overline{g_t}$$
$$\hat{\sigma}_{t+1}^2 = \beta\hat{\sigma}_t^2 + (1-\beta)\frac{\|g_{t,1}-g_{t,2}\|^2}{2}$$

5: **end for**
6: **Return** x_I where I is the random variable such that $\mathbb{P}(I=i) \propto \eta_i$.

In other words, with high probability, the empirical estimator Y_k concentrates around $\sigma_k^2 \pm \frac{M^2}{\sqrt{N}}$. However, if $\sigma_k = O(T^{-\alpha})$, we would need $T^{4\alpha}$ samples to make sure the approximation is in the same magnitude of σ_k. This leads to an extremely inefficient sample complexity, barely implementable in practice.

In the real-world setting, often the environment changes in a smooth way. This provides another possibility if we are able to exploit the continuity in the noise changes. In particular, we can combine the previous gradients with the actual ones to perform the variance estimation, i.e. conducting a moving average estimator. In order to provide theoretical guarantee, we impose the following condition on the total variation of noise level σ_k^2.

Assumption 3.2 (Total variation boundedness.)**.** We assume that the total variation of σ_k^2 is bounded, i.e., $\sum_k |\sigma_k^2 - \sigma_{k+1}^2| \le D^2$, where D is a constant independent of T. To facilitate discussion, we assume $D^2 = \Omega(M^2)$.

The bounded total variation assumption provides us the necessary smoothness on σ_k. This assumption is commonly used in the dynamic, online learning literature (Besbes et al., 2014; Jadbabaie et al., 2015; Mokhtari et al., 2016). A key consequence of this assumption is that it avoids infinite oscillations, such as the pathological setting where $\sigma_{2k} = \frac{1}{T^\alpha}$ and $\sigma_{2k+1} = 1$, in which case the total variation scales with the number of iterations T. The specific constant in $D^2 = \Omega(M^2)$ depends on the shape of the noise. When σ_k is increasing in the first half and decreasing in the second half, as in Example 2.7, the total variation is bounded by $D^2 \le 2M^2$. More generally, if the noise can be decomposed into K piece-wise monotone fragments, then the bound $D^2 \le KM^2$ holds.

	Worst	Adaptive	Dynamic	Moment Est.
$\alpha = 1/9$	$T^{-\frac{1}{2}}$	$T^{-5/9}$	$T^{-11/18}$	$T^{-11/18}$
$\alpha = 1/8$	$T^{-\frac{1}{2}}$	$T^{-9/16}$	$T^{-11/18}$	$T^{-5/8}$

Table 3. Comparison of convergence rate under the mountain shape noise σ_k, where constant terms are omitted. The parameter $\alpha = 1/9$ is the exponent controlling the ratio between $\max \sigma_k$ and $\min \sigma_k$, indicating the significance in noise changes.

3.1. Design and analysis of our proposed algorithm

With the bounded total variation in hand, we are now ready to exploit the continuity of σ_k. Without loss of generality, we use the empirical variance estimator (9) with two samples per iteration. We construct an online estimator using an exponential moving average:

$$\hat{\sigma}_{t+1}^2 = \beta \hat{\sigma}_t^2 + (1-\beta)\frac{\|g_{t,1} - g_{t,2}\|^2}{2}, \qquad \text{(ExpMvAvg)}$$

where $g_{t,1}$ and $g_{t,2}$ are two samples of the gradient at t-th iteration. Note that *t is the index for algorithm iteration, whereas k is the index for number of oracle calls*. The oracle call at iteration t corresponding to the $(2t)_{th}$ and $(2t+1)_{th}$ call to the stochastic oracle.

The moving average is very similar to the one used in the modern adaptive methods such as RMSProp (Tieleman & Hinton, 2012), Adam (Kingma & Ba, 2014), etc. One major difference compared to the current adaptive methods is that we are estimating the variance of gradients, whereas RMSProp/Adam directly accumulate the gradient square in a coordinate-wise manner.

The above estimator combined with dynamic stepsizes lead to our design of Algorithm 1. To avoid the explosion of stepsize when $\hat{\sigma}_t$ is underestimated, we include a constant correction term m in the denominator of η_t. Such correction term is also commonly used in the practical implementation of adaptive methods (Duchi et al., 2011; Kingma & Ba, 2014). In reinforcement learning, this term is sometimes referred to as the exploration bonus (Strehl & Littman, 2008; Azar et al., 2017).

Algorithm 1 gives us the following convergence guarantee.

Theorem 3.3 (Main result)**.** *Under Assumptions 3.1,3.2, with probability at least $1/2$, the iterates generated by Algorithm 1 using parameters $\beta = 1 - 2T^{-2/3}$, $m = 4\sqrt{D^2 + M^2}T^{-\frac{1}{9}}\ln(T)^{\frac{1}{2}}$, $c = \frac{R}{\sqrt{T}}$ satisfy*

$$f(\overline{x}_T) - f^* \le \frac{64R}{\sqrt{T}} \cdot \left(\frac{1}{T}\sum_{k=1}^{T}\frac{1}{\sigma_k + m}\right)^{-1} := \epsilon_{ours}.$$

Corollary 3.4. *Our result directly implies a $1-\delta$ high probability convergence rate, by restarting it $2\log(1/\delta)$ times. An additional $\log(1/\delta)$ dependency will be introduced in the complexity, as in standard high probability results (Nemirovski et al., 2009; Jin et al., 2017; Fang et al., 2018b).*

The main challenge in proving the theorem is to effectively bound the estimation error $|\hat{\sigma}_t^2 - \sigma_{2t}^2/2 - \sigma_{2t+1}^2/2|$. As $\hat{\sigma}_t^2$ is an exponential average, the past errors accumulates into the current estimator, which requires a careful bound using concentration and the total variation of σ_k. In particular, the decay parameter β plays a critical role, determining the contribution of past gradients in the current estimator. A consequence of using past gradients in the estimate is that the online estimator $\hat{\sigma}_k$ is no longer unbiased. The proposed choice of β and m carefully balances the bias error and the variance error, leading to a sublinear regret, see Appendix D.

Understanding the convergence rate Due to the correction constant m, the obtained convergence rate inversely depends on $\sum_{k=1}^{T}\frac{1}{\sigma_k+m}$, instead of the $\sum_{k=1}^{T}\frac{1}{\sigma_k}$ dependency in the dynamic bound. This additional term makes the comparison less straightforward, especially when some of the σ_k are small. We provide several scenarios to facilitate the comparison.

Corollary 3.5. *If the ratio $M/(\min_k \sigma_k) \le T^{\frac{1}{9}}$, then the Moment Estimation SGD method converges in the same order as the dynamic error bound $\epsilon_{dynamic}$.*

This result is remarkable since our proposed method does not require any knowledge of σ_k values, and yet it achieves the dynamic rate. In other words, the exponential moving average estimator successfully adapts to the variation in noise, allowing faster convergence than adaptive/worst bounds. In particular when taking $\alpha = 1/9$ in Example 2.7, we get the following error bounds in Table 3. The comparison between different error bounds suggest that moment estimation can non-trivially improve the convergence rates achieved by conventional analysis for adaptive step sizes.

Corollary 3.6. *Let $\sigma_{avg}^2 = \sum \sigma_k^2/T$ be the average second moment. If $M/\sigma_{avg} \le T^{\frac{1}{9}}$, then adaptive method is no slower than the adaptive error bound (7).*

The condition in Corollary 3.6 is strictly weaker than the condition in Corollary 3.5, which means even though an adaptive method may not match the dynamic bound, it can still be non-trivially better than the adaptive bound. This case happens for instance when $\alpha > \frac{1}{9}$ in Table 3, where the proposed method is $\mathcal{O}(T^{\frac{1}{9}})$ faster than the adaptive bound. Indeed, $\mathcal{O}(T^{\frac{1}{9}})$ is the maximum improvement one can expect according to our current analysis.

We now proceed to evaluate different algorithms and bounds with both synthetic and deep learning experiments.

4. Experiments

In this section, we present three sets of experiments: synthetic least squares, policy optimization for mujoco tasks and neural network training on Cifar10 dataset. Our synthetic experiments verify the theory prediction, whereas our deep learning experiments show that the studied moment-estimation algorithm is practical and performs as well as common baselines.

4.1. Synthetic experiments

In the synthetic experiment, we generate a random linear regression dataset using the `sklearn` library. We design the stochastic oracle as full gradient with injected Gaussian noise, whose coordinate-wise standard deviation σ is shown in the top row figures of Fig 3. We then run the three algorithms discussed in this work: the fixed best step-size, the dynamic step size and the moment estimation algorithm in Alg 1. We fine-tune the step sizes for each algorithm by grid-searching. We repeat the experiment for 10 runs and show the average training trajectory in the second row in Fig 3. From Figure 3, we see that moment estimation algorithm can achieve comparable performance against the dynamic version and outperforms the fixed step size SGD.

4.2. Policy optimization

In this set of experiments, we simulated the proximal policy optimization algorithm in the nonstationary environment described in (Dulac-Arnold et al., 2019). We consider two Mujoco tasks: one in which a walker gets reward for walking; another in which a cart-pole gets reward for swinging up the pole. The original implementation[1] update the parameters with ADAM (Kingma & Ba, 2014) and gradient clipping, which as we will show later, closely connects to noise-dependent step size. To highlight the comparison between fixed step size SA and adaptive step size SA, we replaced the update with the constant baseline and the moment estimation algorithm.

During training, we estimate the variance level by sampling a batch of 5 thousand state, action pairs. For each algorithm, we fine-tune the step sizes by grid-searching among 10^k, where k is an integer. The estimated variance of noise and average reward over 5 runs are plotted in Figure 4. The result shows that the noise level is indeed nonstationary during training, where the SA algorithm can benefit from an adaptive step size. We note that the adaptive step size method learns much faster than fixed stepsize in the cart-pole task over the first few iterations. The advantage might result from the low noise level in initial epochs of the policy gradient due to the simplicity of swinging up a pole.

[1] https://github.com/google-research/realworldrl_suite

4.3. Neural network training

We will now discuss the application of our proposed method to neural network training. In order to apply our result in such settings, the challenge is not merley about extending the results to nonsmooth scenarios or non-convex cases. Rather, the main challenge we face is that noise levels in stochastic optimization for neural network training are step-size dependent. The source of nonstationary noise in this setting is entirely *endogenous*, i.e., it is determined by the iterative output x_t. In such settings, it is unclear how baselines could be defined, or improvement could be quantified.

To this end, we show in Figure 5 that the noise trajectories are different for training ResNet18 on Cifar10 with different algorithms and hyperparameters. Even though our theoretical guarantees do not directly apply to this setting, we can apply our adaptive step-size algorithm and still exploit the variations in noise.

Directly applying Algorithm 1 may be cumbersome on standard image classification pipelines due to requiring an unbiased within-batch variance estimator. Instead, we notice from Figure 5 that the noise level dominates the gradient norm in neural network training, and hence we can simply use the empirical second moment $\mathbb{E}[\|g_k\|^2]$ as a substitute for the empirical variance $\mathbb{E}[\|g_k - \nabla f(x_k)\|^2]$. This gives us the following update,

$$x_{k+1} = x_k - \eta_k g_k \quad \text{with} \quad \eta_t = \tfrac{c}{\hat{m}_k + m},$$
$$\text{and } \hat{m}_{k+1}^2 = \beta \hat{m}_k^2 + (1-\beta)\frac{\|g_k\|^2}{2}.$$

We should point out that interestingly, the above update is *exactly the same* as in the RMSProp algorithm (Tieleman & Hinton, 2012) if the step sizes were coordinate-wise. Since the moment estimation algorithm is very similar to popular optimizers for training Cifar10, we do not expect it to significantly outperform the well tuned baselines. Instead, we highlight that our analysis provides theoretical evidence for the popularity of moment estimation techniques in practice.

Such observations point to the following interesting question: "*why is fixed stepsize SGD minimax optimal (Arjevani et al., 2019), yet adaptive methods such as RMSProp and ADAM outperforms fixed step size SGD in many real world settings?*" Alongside with many recent works (Reddi et al., 2019; Wilson et al., 2017; Zhang et al., 2019; Ward et al., 2018; Luo et al., 2019; Liu et al., 2020), we believe that our more fine-grained analysis provides a new perspective and motivates new avenues for proving the effectiveness of adaptive algorithms such as ADAM and RMSProp.

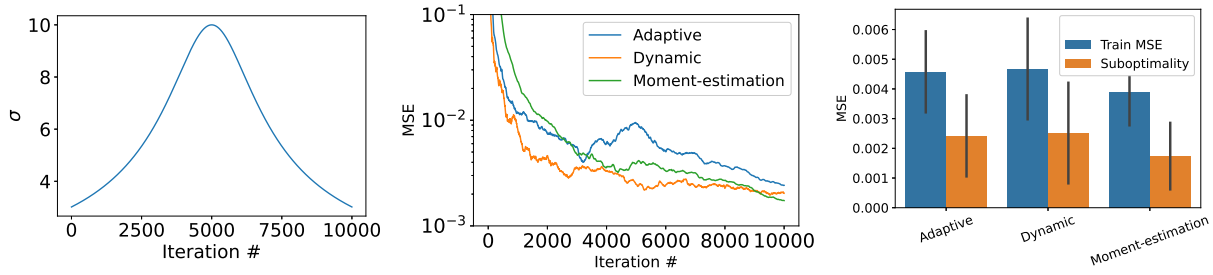

Figure 3. Results for synthetic least square problems. The three plots from left to right corresponds to the shape of noise, the average MSE out of 10 random runs and the numerical values of suboptimality.

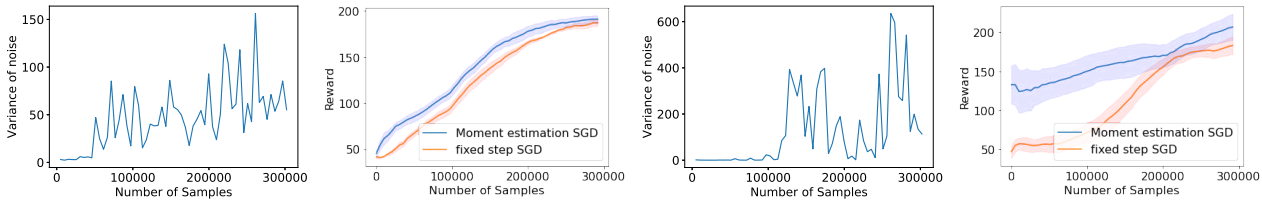

Figure 4. The left two figures plot the noise level and the average reward during training a controller for the walker task. The right two plots correspond to the cartpole swing-up task.

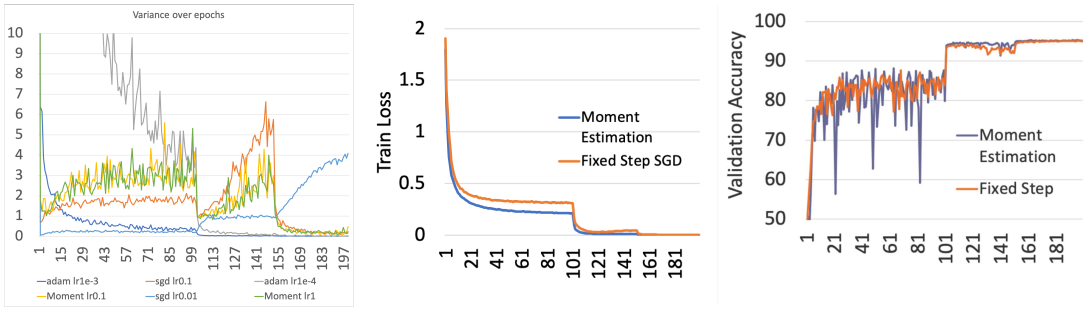

Figure 5. The left figure plots the noise level while training a ResNet model on Cifar10 for 200 epochs. It shows that noise levels are very much algorithm dependent. The second left figure shows that the variance in stochastic gradient almost equals the variance. The right two figures show that in Cifar10 training, the moment estimation algorithm performs as well as the fine-tuned SGD baseline.

5. Conclusions and Discussions

In this work, we provided a new perspective for instance-dependent complexity of stochastic approximation methods. We first categorized existing instance-dependent error bounds into different levels based on dominance relations. We then proposed a new dynamic error bound that dominates known ones. Simple algorithms that achieves this bound requires knowing the exact noise levels and is not implementable. To address this issue, we showed that when noise levels have bounded total variation, moment estimation can achieve the desired rate. Our results are validated by both synthetic and real-world experiments. We believe the instance complexity we developed shed new insights to the following interesting question: "*why is fixed stepsize SGD minimax optimal (Arjevani et al., 2019), yet adaptive methods such as RMSProp and ADAM outperforms fixed step size SGD in many real world settings?*"

Many important instance-complexity problems are still open. First, in traditional complexity theory, instance dependent lower bounds can sometimes be tight up to constants (Afshani et al., 2017). However, determining the lower bounds for stochastic approximation instances requires a reformulation of the complexity definition, such as what information is available. For example, in the dynamic bounds setup, a different step size choice,

$$\eta_k = R/(\sigma_k^2\sqrt{\textstyle\sum_{i=1}^{T}\frac{1}{\sigma_i^2}}),$$

may lead to a better error bound. Yet, this kind of step size is at the same time dynamic and non-causal (depends on information from future iterates). Therefore, information dependence needs to be properly integrated in the lower bound of instance complexity.

Second, beyond settings such as smooth-convex problems, we currently do not know of any faster instance-dependent bounds. Even if better bounds can be achieved when noise levels are known, it could still be unclear whether it can be achieved by practical and implementable algorithms. Another question is whether in our setup where the function is smooth and convex the $T^{1/9}$ factor can be improved. We believe understanding these problems can provide more insight into practical algorithm performances and lead to invention of new gradient based algorithms.

Last but not least, an apparent limitation of the stochastic approximation setting is the assumption of an exogenous noise. It is usually not satisfied in the standard empirical risk minimization framework, where the noise is iterate dependent, i.e. x-dependent. Note that the iterates generated by one algorithm is mostly very different from the iterates generated by another one, how to appropriately quantify the state-dependency such that we can derive non-vacuous instance complexity result becomes challenging. Though simplified, we believe our work provides a solid first step towards this ultimate goal.

Acknowledgments

SS acknowledges support from an NSF CAREER award (number 1846088), and NSF CCF-2112665 (TILOS AI Research Institute). JZ acknowledges support from a IIIS young scholar fellowship.

References

Afshani, P., Barbay, J., and Chan, T. M. Instance-optimal geometric algorithms. *Journal of the ACM (JACM)*, 64 (1):1–38, 2017.

Agarwal, A., Wainwright, M. J., Bartlett, P. L., and Ravikumar, P. K. Information-theoretic lower bounds on the oracle complexity of convex optimization. 2009.

Allen-Zhu, Z. Katyusha: The first direct acceleration of stochastic gradient methods. March 2016.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. 2017.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pp. 199–207, 2014.

Borgwardt, K. H. *The Simplex Method: a probabilistic analysis*, volume 1. Springer Science & Business Media, 2012.

Defazio, A. and Bottou, L. On the ineffectiveness of variance reduced optimization for deep learning. *arXiv preprint arXiv:1812.04529*, 2018.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Dulac-Arnold, G., Mankowitz, D., and Hester, T. Challenges of real-world reinforcement learning. *arXiv preprint arXiv:1904.12901*, 2019.

Fabian, V. et al. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, 39(4):1327–1332, 1968.

Fagin, R., Lotem, A., and Naor, M. Optimal aggregation algorithms for middleware. *Journal of computer and system sciences*, 66(4):614–656, 2003.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018a.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator, 2018b.

Ghadimi, S. and Lan, G. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, 2013.

Hoare, C. A. Quicksort. *The Computer Journal*, 5(1):10–16, 1962.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., and Sridharan, K. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pp. 398–406, 2015.

Jin, C., Netrapalli, P., and Jordan, M. I. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv preprint arXiv:1711.10456*, 2017.

Khamaru, K., Xia, E., Wainwright, M. J., and Jordan, M. I. Instance-optimality in optimal value estimation: Adaptivity via variance-reduced q-learning. *arXiv preprint arXiv:2106.14352*, 2021.

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

Kushner, H. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Lacotte, J. and Pilanci, M. Optimal randomized first-order methods for least-squares problems. In *International Conference on Machine Learning*, pp. 5587–5597. PMLR, 2020.

Levy, K. Y., Yurtsever, A., and Cevher, V. Online adaptive methods, universality and acceleration. 2018.

Lin, H., Mairal, J., and Harchaoui, Z. A universal catalyst for first-order optimization. In *Advances in Neural Information Processing Systems*, pp. 3384–3392, 2015.

Liu, M., Zhang, W., Orabona, F., and Yang, T. Adam$\hat{}+$: A stochastic method with adaptive variance reduction. *arXiv preprint arXiv:2011.11985*, 2020.

Luo, L., Xiong, Y., Liu, Y., and Sun, X. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

Mokhtari, A., Shahrampour, S., Jadbabaie, A., and Ribeiro, A. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 7195–7201. IEEE, 2016.

Mou, W., Pananjady, A., Wainwright, M. J., and Bartlett, P. L. Optimal and instance-dependent guarantees for markovian linear stochastic approximation. *arXiv preprint arXiv:2112.12770*, 2021.

Moulines, E. and Bach, F. R. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. 2011.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Pananjady, A. and Wainwright, M. J. Instance-dependent -bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67 (1):566–585, 2020.

Paquette, C., Lee, K., Pedregosa, F., and Paquette, E. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. *arXiv preprint arXiv:2102.04396*, 2021.

Pedregosa, F. and Scieur, D. Average-case acceleration through spectral density estimation. *arXiv preprint arXiv:2002.04756*, 2020.

Polyak, B. T. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

Preparata, F. P. and Shamos, M. I. *Computational geometry: an introduction*. Springer Science & Business Media, 2012.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

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

Robbins, H. and Monro, S. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

Spielman, D. A. The smoothed analysis of algorithms. In *International Symposium on Fundamentals of Computation Theory*, pp. 17–18. Springer, 2005.

Strehl, A. L. and Littman, M. L. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Tieleman, T. and Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2):26–31, 2012.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *arXiv preprint arXiv:1806.01811*, 2018.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pp. 4148–4158, 2017.

Yang, T., Zhang, L., Jin, R., and Yi, J. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning*, pp. 449–457. PMLR, 2016.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Analysis of gradient clipping and adaptive scaling with a relaxed smoothness condition. *arXiv preprint arXiv:1905.11881*, 2019.

Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

A. Proof of Theorem 2.4

Proof. Using the fact that $\frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \le \langle \nabla f(y) - \nabla f(x), y - x\rangle$, we have

$$\begin{aligned}\mathbb{E}[\|x_{k+1} - x^*\|^2|\mathcal{F}_k] &= \|x_k - x^*\|^2 - 2\eta_k\langle \nabla f(x_k), x_k - x^*\rangle + \eta_k^2\|\nabla f(x_k)^2\| + \eta_k^2\sigma_k^2 \\ &\le \|x_k - x^*\|^2 - \eta_k(2 - L\eta_k)\langle \nabla f(x_k), x_k - x^*\rangle + \eta_k^2\sigma_k^2 \\ &\le \|x_k - x^*\|^2 - \eta_k\langle \nabla f(x_k), x_k - x^*\rangle + \eta_k^2\sigma_k^2\end{aligned}$$

As $f(x_k) - f^* \le \langle \nabla f(x_k), x_k - x^*\rangle$. We have

$$\eta_k(f(x_k) - f^*) \le \|x_k - x^*\|^2 - \mathbb{E}[\|x_{k+1} - x^*\|^2|\mathcal{F}_k] + \eta_k^2\sigma_k^2$$

Taking expectation and telescoping yields

$$\mathbb{E}\left[\sum_{k=1}^T \eta_k(f(x_k) - f^*)\right] \le \|x_1 - x^*\|^2 - \mathbb{E}[\|x_{T+1} - x^*\|^2] + \sum_{k=1}^T \eta_k^2\sigma_k^2$$

□

B. Proof of Lemma 2.6

The only nontrivial one is $\epsilon_{\text{dynamic}} \preceq \epsilon_{\text{adaptive}}$. This follows from Jensen's inequality $\mathbb{E}[X]^{-2} \le \mathbb{E}[X^{-2}]$, and we have

$$(\tfrac{1}{T}\textstyle\sum_k \frac{1}{\sigma_k})^{-2} \le \frac{1}{T}\sum_k \sigma_k^2,$$

.

C. Summation series in Example 2.7

In Example 2.7, we have

$$\sigma_k = \frac{1}{\sqrt{1 + T^\alpha\left(\frac{2k}{T} - 1\right)^2}}$$

- The second moment of σ_k is given by

$$\begin{aligned}\frac{1}{T}\sum_{k=1}^T \sigma_k^2 = \frac{1}{T}\sum_{k=1}^T \frac{1}{1 + T^\alpha\left(\frac{2k}{T} - 1\right)^2} \quad &\sim \quad \frac{1}{T}\int_0^T \frac{1}{1 + T^\alpha\left(\frac{2x}{T} - 1\right)^2}dx \\ &\overset{u:=\frac{2x}{T}-1}{=} \frac{1}{T}\int_{-1}^1 \frac{1}{1 + T^\alpha u^2}\frac{du}{\frac{2}{T}} \\ &= \int_0^1 \frac{1}{1 + T^\alpha u^2}du \\ &= \left[\frac{1}{T^{\frac{\alpha}{2}}}arctan(x)\right]_0^{T^{\frac{\alpha}{2}}} \sim \frac{1}{T^{\frac{\alpha}{2}}}\end{aligned}$$

  This implies that the adaptive bound 7 is of order $O(T^{-\left(\frac{1}{2}+\frac{\alpha}{4}\right)})$.

- The harmonic sum of σ_k is given by

$$\begin{aligned}\frac{1}{T}\sum_{k=1}^T \frac{1}{\sigma_k} = \frac{1}{T}\sum_{k=1}^T \sqrt{1 + T^\alpha\left(\frac{2k}{T} - 1\right)^2} &= O\left(\frac{1}{T}\sum_{k=1}^T 1 + T^{\frac{\alpha}{2}}\left|\frac{2k}{T} - 1\right|\right) \\ &= O\left(1 + \frac{1}{T}T^{\frac{\alpha}{2}}\frac{2\sum_{k=0}^{\frac{T}{2}}(T - 2k)}{T}\right) \\ &= O\left(1 + \frac{1}{T}T^{\frac{\alpha}{2}}\frac{T^2}{T}\right) = O\left(T^{\frac{\alpha}{2}}\right)\end{aligned}$$

This implies that the dynamic bound8 is of order $O(T^{-\left(\frac{1}{2}+\frac{\alpha}{2}\right)})$.

D. Key Lemma: Total estimation error of the variance estimator

Lemma D.1. *Under Assumptions 3.2, taking $\beta = 1 - 2T^{-2/3}$, the total estimation error of the $\hat{\sigma}_k^2$ based on* (ExpMvAvg) *is bounded by:*

$$\mathbb{E}\left[\sum_{t=1}^{T/2} |\hat{\sigma}_t^2 - \sigma_{2t}^2/2 - \sigma_{2t+1}^2/2|\right] \leq 2(D^2 + M^2)T^{2/3}\ln(T^{2/3})$$

Proof. On a high level, we decouple the error in a bias term and a variance term. We use the total variation assumption to bound the bias term, and use the exponential moving average to reduce variance. Then we pick β to balance the two terms.

For simplicity, we denote $\text{var}_t^2 = \sigma_{2t}^2/2 + \sigma_{2t+1}^2/2$.

From triangle inequality, we have

$$\sum_{t=0}^{T/2} \mathbb{E}\left[|\hat{\sigma}_t^2 - \text{var}_t^2|\right] \leq \sum_{t=1}^{T/2} \underbrace{\mathbb{E}\left[|\hat{\sigma}_t^2 - \mathbb{E}[\hat{\sigma}_t^2]|\right]}_{\text{Variance term}} + \sum_{t=1}^{T/2} \underbrace{\left|\mathbb{E}[\hat{\sigma}_t^2] - \text{var}_t^2\right|}_{\text{Bias term}} \tag{10}$$

We first bound the bias term. By definition of $\hat{\sigma}_t$, we have

$$\begin{aligned}\mathbb{E}[\hat{\sigma}_t^2] - \text{var}_t^2 &= \beta\mathbb{E}[\hat{\sigma}_{t-1}^2] + (1-\beta)\text{var}_{t-1}^2 - \text{var}_t^2 \\ &= \beta(\mathbb{E}[\hat{\sigma}_{t-1}^2] - \text{var}_{t-1}^2) + (\text{var}_{t-1}^2 - \text{var}_t^2)\end{aligned}$$

Hence by recursion,

$$\mathbb{E}[\hat{\sigma}_t^2] - \text{var}_t^2 = \beta^{t-1}\underbrace{(\mathbb{E}[\hat{\sigma}_1^2] - \text{var}_1^2)}_{=0} + \beta^{t-2}(\sigma_1^2 - \text{var}_2^2) + \cdots + (\text{var}_{t-1}^2 - \text{var}_t^2)$$

Therefore, the bias term could be bounded by

$$\begin{aligned}\sum_{t=1}^{T/2} \left|\mathbb{E}[\hat{\sigma}_t^2] - \text{var}_t^2\right| &\leq \sum_{t=1}^{T/2}\sum_{j=1}^{t-1} \beta^{t-1-j}\left|\text{var}_j^2 - \text{var}_{j+1}^2\right| \\ &\leq \sum_{k=1}^{T}\sum_{j=1}^{k-1} \beta^{k-1-j}\left|\sigma_k^2 - \sigma_{k+1}^2\right| \\ &= \sum_{k=1}^{T-1}\left|\sigma_k^2 - \sigma_{k+1}^2\right| \sum_{j=0}^{T-1-k} \beta^j \\ &\leq \frac{1}{1-\beta}\sum_{k=1}^{T-1}\left|\sigma_k^2 - \sigma_{k+1}^2\right| \\ &\leq \frac{D^2}{1-\beta} \quad \text{(From Assumption (3.2))}\end{aligned}$$

The first inequality follows by traingle inequality. The third inequality uses the geometric sum over β. To bound the variance term, we remark that

$$\hat{\sigma}_t^2 = (1-\beta)y_{t-1}^2 + (1-\beta)\beta y_{t-2}^2 + \cdots + (1-\beta)\beta^{t-2}y_1^2 + \beta^{t-1}y_0^2.$$

where we denote

$$y_t = \frac{\|g_{t,1} - g_{t,2}\|^2}{2}.$$

Hence from independence of the gradients, we have

$$\begin{aligned}\mathbb{E}\left[|\hat{\sigma}_t^2 - \mathbb{E}[\hat{\sigma}_t^2]|\right] &\leq \sqrt{\text{Var}[\hat{\sigma}_t^2]} \\ &= \sqrt{\text{Var}[(1-\beta)y_{t-1}^2] + \text{Var}[(1-\beta)\beta y_{t-2}^2] + \cdots + \text{Var}[(1-\beta)\beta^{t-2}y_1^2] + \text{Var}[\beta^{t-1}y_0^2]} \\ &\leq \sqrt{(1-\beta)^2 + (1-\beta)^2\beta^2 + \cdots + (1-\beta)^2\beta^{2(t-2)} + \beta^{2(t-1)}}M^2,\end{aligned}$$

where M^2 is an upperbound on the variance. The first inequality follows by Jensen's inequality. The second equality uses independence of y_i given $g_1, ..., g_{i-1}$. The last inequality follows by assumption 3.2.

We distinguish two cases, when t is small, we simply bound the coefficient by 1, i.e.

$$\sqrt{(1-\beta)^2 + (1-\beta)^2\beta^2 + \cdots + (1-\beta)^2\beta^{2(t-2)} + \beta^{2(t-1)}} \leq 1$$

When t is large such that $t \geq 1 + \gamma$, with $\gamma = \frac{1}{2(1-\beta)}\ln(\frac{1}{1-\beta})$, we have $\beta^{2(t-1)} \leq 1 - \beta$, thus

$$\begin{aligned}&\sqrt{(1-\beta)^2 + (1-\beta)^2\beta^2 + \cdots + (1-\beta)^2\beta^{2(t-2)} + \beta^{2(t-1)}} \\ \leq &\sqrt{\frac{(1-\beta)^2}{1-\beta^2} + \beta^{2(t-1)}} \\ \leq &\sqrt{\frac{(1-\beta)^2}{1-\beta^2} + (1-\beta)} \\ \leq &\sqrt{2(1-\beta)}\end{aligned}$$

The second inequality follows by $t \geq 1 + \gamma$, with $\gamma = \frac{1}{2(1-\beta)}\ln(\frac{1}{1-\beta})$. Therefore, when $t \geq 1 + \gamma$,

$$\mathbb{E}\left[|\hat{\sigma}_t^2 - \mathbb{E}[\hat{\sigma}_t^2]|\right] \leq \sqrt{2(1-\beta)}M$$

Therefore, substitute in the above equation into the

$$\begin{aligned}\sum_{t=1}^{T/2}\mathbb{E}\left[|\hat{\sigma}_t^2 - \mathbb{E}[\hat{\sigma}_t^2]|\right] &= \sum_{t=1}^{\gamma}\mathbb{E}\left[|\hat{\sigma}_t^2 - \mathbb{E}[\hat{\sigma}_t^2]|\right] + \sum_{t=\gamma+1}^{T/2}\mathbb{E}\left[|\hat{\sigma}_t^2 - \mathbb{E}[\hat{\sigma}_t^2]|\right] \\ &\leq (\gamma + (T-\gamma)\sqrt{2(1-\beta)})M^2\end{aligned}$$

Summing up the variance term and the bias term yields,

$$\sum_{t=0}^{T/2}\mathbb{E}\left[|\hat{\sigma}_t^2 - \text{var}_t^2|\right] \leq \frac{D^2}{1-\beta} + (\gamma + (T-\gamma)\sqrt{2(1-\beta)})M^2 \tag{11}$$

Taking $\beta = 1 - T^{-2/3}/2$ yields,

$$\sum_{t=0}^{T/2}\mathbb{E}\left[|\hat{\sigma}_t^2 - \text{var}_t^2|\right] \leq 2(D^2 + M^2)T^{2/3}\ln(T^{2/3}) \tag{12}$$

□

E. Proof of Theorem 3.3

On a high level, the difference between the adaptive stepsize and the idealized harmonic stepsize mainly depends on the estimation error $|\hat{\sigma}_k^2 - \sigma_k^2|$, which has a sublinear regret according to Lemma D. Then we carefully integrate this regret bound to control the derivation from the idealized algorithm, reaching the conclusion.

Proof. By the update rule of x_{t+1}, we have,

$$\|x_{t+1} - x^*\|^2 = \|x_t - \eta_t \overline{g_t} - x^*\|^2 = \|x_t - x^*\|^2 - 2\eta_t \langle \overline{g_t}, x_t - x^* \rangle + \eta_t^2 \|\overline{g_t}\|^2.$$

Noting that the stepsize η_t is independent of $\overline{g_t}$ conditioned on $\overline{g_{t-1}}$. Recall that for simplicity, we denote $\text{var}_t^2 = \sigma_{2t}^2/2 + \sigma_{2t+1}^2/2$. Taking expectation with respect to g_t conditional on the past iterates lead to

$$\begin{aligned}
2\eta_t(f(x_t) - f^*) &\le 2\eta_t \langle \nabla f(x_t), x_t - x^* \rangle \\
&= \mathbb{E}[2\eta_t \langle \overline{g_t}, x_t - x^* \rangle | x_t, \cdots, x_1] \\
&= -\mathbb{E}[\|x_{t+1} - x^*\|^2 | x_t, \cdots, x_1] + \|x_t - x^*\|^2 + \eta_t^2(\|\nabla f(x_t)\|^2 + \text{var}_t^2) \\
&\le -\mathbb{E}[\|x_{t+1} - x^*\|^2 | x_t, \cdots, x_1] + \|x_t - x^*\|^2 + \eta_t^2 \sigma_t^2 + L\eta_t^2 \langle \nabla f(x_t), x_t - x^* \rangle.
\end{aligned}$$

Recall that $R = \|x_1 - x^*\|$, taking expectation and sum over iterations k, we get

$$\mathbb{E}[(\textstyle\sum_{t=1}^{T/2} \eta_t)(f(\overline{x}_T) - f^*)] \le R^2 + \mathbb{E}[\sum_{t=1}^{T/2} \eta_t^2 \text{var}_t^2].$$

Hence by Markov's inequality, with probability at least $3/4$,

$$(\textstyle\sum_{t=1}^{T/2} \eta_t)(f(\overline{x}_T) - f^*) \le 4\mathbb{E}[2(\sum_{t=1}^{T/2} \eta_t)(f(\overline{x}_T) - f^*)] \le 4(R^2 + \mathbb{E}[\sum_{t=1}^{T/2} \eta_t^2 \text{var}_t^2]). \tag{13}$$

Now we can upper bound the right hand side, indeed

$$\begin{aligned}
\sum_{t=1}^{T/2} \mathbb{E}[\eta_t^2 \text{var}_t]^2 &= c^2 \sum_{t=1}^{T/2} \mathbb{E}\left[\frac{\text{var}_t^2}{(\hat{\sigma}_t + m)^2}\right] \\
&\le c^2 \left(\sum_{t=1}^{T/2} \mathbb{E}\left[\frac{\text{var}_t^2 - \hat{\sigma}_t^2}{(\hat{\sigma}_t + m)^2}\right] + \sum_{t=1}^{T/2} \mathbb{E}\left[\frac{\hat{\sigma}_t^2}{(\hat{\sigma}_t + m)^2}\right] \right) \\
&\le c^2 \left(\frac{1}{m^2} \sum_{t=1}^{T} \mathbb{E}\left[|\text{var}_t^2 - \hat{\sigma}_t^2|\right] + T \right) \\
&\le c^2 \left(\frac{(M^2 + D^2) T^{2/3} \ln(T^{2/3})}{m^2} + T \right) \le 3c^2 T
\end{aligned} \tag{14}$$

The last inequality follows by the choice on m. Hence, from Eq. (13), we have with probability at least $3/4$,

$$(\textstyle\sum_{t=1}^{T} \eta_t)(f(\overline{x}_T) - f^*) \le 4(R^2 + 3c^2 T) \tag{15}$$

Next, by denoting $(x)_+ = \max(x, 0)$, we lower bound the left hand side,

$$\begin{aligned}
\frac{1}{c}\sum \eta_t &= \sum \frac{1}{\hat{\sigma}_t + m} \\
&= \sum \frac{1}{\text{var}_t + m} + \sum \left(\frac{1}{\hat{\sigma}_t + m} - \frac{1}{\text{var}_t + m} \right) \\
&\geq \sum \frac{1}{\text{var}_t + m} - \sum \frac{(\hat{\sigma}_t - \text{var}_t)_+}{(\text{var}_t + m)(\hat{\sigma}_t + m)} \\
&\geq \sum \frac{1}{\text{var}_t + m} - \sum \frac{(\hat{\sigma}_t - \text{var}_t)_+}{\sqrt{\text{var}_t + m} \cdot m^{3/2}} \\
&\geq \sum \frac{1}{\text{var}_t + m} - \sum \frac{1}{2} \left(\frac{(\hat{\sigma}_t - \text{var}_t)_+^2}{m^3} + \frac{1}{\text{var}_t + m} \right) \\
&= \frac{1}{2} \sum \frac{1}{\text{var}_t + m} - \frac{1}{2m^3} \sum (\hat{\sigma}_t - \text{var}_t)_+^2 \qquad (16) \\
& \qquad (17)
\end{aligned}$$

Note that

$$\begin{aligned}
\frac{1}{\text{var}_t + m} &= \frac{1}{\sqrt{\sigma_{2t}^2 + \sigma_{2t+1}^2}/2 + m} \geq \frac{1}{\sigma_{2t} + \sigma_{2t+1} + m} \\
&\geq \frac{1}{4}\left(\frac{1}{\sigma_{2t} + m} + \frac{1}{\sigma_{2t+1} + m}\right) - \frac{1}{m^2}(\sigma_{2t} - \sigma_{2t+1}) \\
&\geq \frac{1}{4}\left(\frac{1}{\sigma_{2t} + m} + \frac{1}{\sigma_{2t+1} + m}\right) - \frac{1}{m^2}\left|\sigma_{2t}^2 - \sigma_{2t+1}\right|)
\end{aligned}$$

Therefore,

$$\frac{1}{c}\sum \eta_t \geq \frac{1}{2}\sum_k \frac{1}{\sigma_l + m} - \frac{1}{2m^3}\sum (\hat{\sigma}_t - \text{var}_t)_+^2 - \frac{\sqrt{T}}{m^2}D \qquad (18)$$

Finally, by Markov's inequality, with probability $3/4$

$$\sum (\hat{\sigma}_t - \text{var}_t)_+^2 \leq 4\mathbb{E}[\sum (\hat{\sigma}_k - \text{var}_t)_+^2] \leq 4\mathbb{E}[\sum (\hat{\sigma}_k - \text{var}_t)^2] \leq 8(D^2 + M^2)T^{2/3}\ln(T^{2/3}).$$

Following the choice of $m = 4\sqrt{D^2 + M^2}T^{-\frac{1}{9}}\ln(T)^{\frac{1}{2}}$, we have

$$\frac{1}{2m^3}\sum_{k=1}^{T}(\hat{\sigma}_k - \text{var}_t)_+^2 \leq \frac{T}{4(M+m)} \leq \frac{1}{4}\sum_{k=1}^{T}\frac{1}{\sigma_k + m}$$

$$\frac{\sqrt{T}}{m^2}D \leq T/8(M+m) \leq \frac{1}{8}\sum_{k=1}^{T}\frac{1}{\sigma_k + m}$$

Consequently, together with (15) and (18), we know that with probability at least $1 - \frac{1}{4} - \frac{1}{4} = 1/2$,

$$f(\overline{x}_T) - f^* \leq \frac{4(R^2 + 3c^2T)}{\sum_k \frac{c}{8(\sigma_k + m)}} \leq \frac{2R}{\sqrt{T}} \cdot \frac{64}{\sum_k \frac{1}{(\sigma_k + m)}}, \qquad (19)$$

where the last inequality follows by setting $c = \frac{R}{\sqrt{T}}$.

□