Journal of Machine Learning Research 17 (2016) 1-66 Submitted 3/16; Revised 8/16; Published 10/16

On the Influence of Momentum Acceleration on Online Learning

Kun Yuan KUNYUAN@UCLA.EDU
Bicheng Ying YBC@UCLA.EDU
Ali H. Sayed SAYED@UCLA.EDU
Department of Electrical Engineering
University of California
Los Angeles, CA 90095, USA

Editor: Leon Bottou

Abstract

The article examines in some detail the convergence rate and mean-square-error performance of momentum stochastic gradient methods in the constant step-size and slow adaptation regime. The results establish that momentum methods are equivalent to the standard stochastic gradient method with a re-scaled (larger) step-size value. The size of the re-scaling is determined by the value of the momentum parameter. The equivalence result is established for all time instants and not only in steady-state. The analysis is carried out for general strongly convex and smooth risk functions, and is not limited to quadratic risks. One notable conclusion is that the well-known benefits of momentum constructions for deterministic optimization problems do not necessarily carry over to the adaptive online setting when small constant step-sizes are used to enable continuous adaptation and learning in the presence of persistent gradient noise. From simulations, the equivalence between momentum and standard stochastic gradient methods is also observed for non-differentiable and non-convex problems.

Keywords: Online Learning, Stochastic Gradient, Momentum Acceleration, Heavy-ball Method, Nesterov's Method, Mean-Square-Error Analysis, Convergence Rate

1. Introduction

Stochastic optimization focuses on the problem of optimizing the expectation of a loss function, written as

$$\min_{w \in \mathbb{R}^M} J(w) \stackrel{\Delta}{=} \mathbb{E}_{\boldsymbol{\theta}}[Q(w;\boldsymbol{\theta})], \tag{1}$$

where $\boldsymbol{\theta}$ is a random variable whose distribution is generally unknown and $J(w)$ is a convex function (usually strongly-convex due to regularization). If the probability distribution of the data, $\boldsymbol{\theta}$, is known beforehand, then one can evaluate $J(w)$ and seek its minimizer by means of a variety of gradient-descent or Newton-type methods (Polyak, 1987; Bertsekas, 1999; Nesterov, 2004). We refer to these types of problems, where $J(w)$ is known, as *deterministic* optimization problems. On the other hand, when the probability distribution of the data is unknown, then the risk function $J(w)$ is unknown as well; only instances of the loss function, $Q(w;\boldsymbol{\theta})$, may be available at various observations $\boldsymbol{\theta}_i$, where i refers to the sample index. We refer to these types of problems, where $J(w)$ is unknown but defined implicity as the expectation of some known loss form, as *stochastic* optimization problems. This article deals with this second type of problems, which are prevalent in online adaptation and learning contexts (Widrow and Stearns, 1985; Haykin, 2008; Sayed, 2008; Theodoridis, 2015).

When $J(w)$ is differentiable, one of the most popular techniques to seek minimizers for (1) is to employ the *stochastic* gradient method. This algorithm is based on employing instantaneous approximations for the true (unavailable) gradient vectors, $\nabla_w J(w)$, by using the gradients of the loss function, $\nabla_w Q(w;\boldsymbol{\theta}_i)$, evaluated at successive samples of the streaming data $\boldsymbol{\theta}_i$ over the iteration index i, say, as:

$$\boldsymbol{w}_i \;=\; \boldsymbol{w}_{i-1} \;-\; \mu\nabla_w\, Q(\boldsymbol{w}_{i-1};\boldsymbol{\theta}_i), \quad i\geq 0. \tag{2}$$

where $\mu > 0$ is a step-size parameter. Note that we are denoting the successive iterates by $\boldsymbol{w}_i$ and using the boldface notation to refer to the fact that they are random quantities in view of the randomness in the measurements $\{\boldsymbol{\theta}_i\}$. Due to their simplicity, robustness to noise and uncertainty, and scalability to big data, such stochastic gradient methods have become popular in large-scale optimization, machine learning, and data mining applications (Zhang, 2004; Bottou, 2010; Gemulla et al., 2011; Sutskever et al., 2013; Kahou et al., 2013; Cevher et al., 2014; Szegedy et al., 2015; Zareba et al., 2015).

1.1 Convergence Rate

Stochastic-gradient algorithms can be implemented with decaying step-sizes, such as $\mu(i) = \tau/i$ for some constant τ, or with constant step-sizes, $\mu > 0$. The former generally ensure asymptotic convergence to the true minimizer of (1), denoted by w^o, at a convergence rate that is on the order of $O(1/i)$ for strongly-convex risk functions. This guarantee, however, comes at the expense of turning off adaptation and learning as time progresses since the step-size value approaches zero in the limit, as $i \to \infty$. As a result, the algorithm loses the ability to track concept drifts. In comparison, constant step-sizes keep adaptation and learning alive and infuse a desirable tracking mechanism into the operation of the algorithm: even if the minimizers drift with time, the algorithm will generally be able to adjust and track their locations. Moreover, convergence can now occur at the considerably faster exponential rate, $O(\alpha^i)$, for some $\alpha \in (0,1)$. These favorable properties come at the expense of a small deterioration in the limiting accuracy of the iterates since almost-sure convergence is not guaranteed any longer. Instead, the algorithm converges in the mean-square-error sense towards a small neighborhood around the true minimizer, w^o, whose radius is on the order of $O(\mu)$. This is still a desirable conclusion because the value of μ is controlled by the designer and can be chosen sufficiently small.

A well-known tradeoff therefore develops between convergence rate and mean-square-error (MSE) performance. The asymptotic MSE performance level approaches $O(\mu)$ while the convergence rate is given by $\alpha = 1 - O(\mu)$ (Polyak, 1987; Sayed, 2014a). It is nowadays well-recognized that the small $O(\mu)$ degradation in performance is acceptable in most large-scale learning and adaptation problems (Bousquet and Bottou, 2008; Bottou, 2010; Sayed, 2014b). This is because, in general, there are always modeling errors in formulating optimization problems of the form (1); the cost function may not reflect perfectly the scenario and data under study. As such, insisting on attaining asymptotic convergence to the true minimizer may not be necessarily the best course of action or may not be worth the effort. It is often more advantageous to tolerate a small steady-state error that is negligible in most cases, but is nevertheless attained at a faster exponential rate of convergence than the slower rate of $O(1/i)$. Furthermore, the data models in many applications are more complex than assumed, with possibly local minima. In these cases, constant step-size implementations can help reduce the risk of being trapped at local solutions.

For these various reasons, and since our emphasis is on algorithms that are able to learn continuously, we shall focus on small *constant* step-size implementations. In these cases, gradient noise is always present, as opposed to decaying step-size implementations where the gradient noise terms get annihilated with time. The analysis in the paper will establish analytically, and illustrate by simulations, that, for sufficiently small step-sizes, any benefit from a momentum stochastic-construction can be attained by adjusting the step-size parameter for the original stochastic-gradient implementation. We emphasize here the qualification "small" for the step-size. The reason we focus on small step-sizes (which correspond to the slow adaptation regime) is because, in the stochastic context, mean-square-error stability and convergence require small step-sizes.

1.2 Acceleration Methods

In the *deterministic* optimization case, when the true gradient vectors of the smooth risk function $J(w)$ are available, the iterative algorithm for seeking the minimizer of $J(w)$ becomes the following gradient-descent recursion

$$w_i = w_{i-1} - \mu \nabla_w J(w_{i-1}), \quad i \geq 0, \tag{3}$$

There have been many ingenious methods proposed in the literature to enhance the convergence of these methods for both cases of convex and strongly-convex risks, $J(w)$. Two of the most notable and successful techniques are the heavy-ball method (Polyak, 1964, 1987; Qian, 1999) and Nesterov's acceleration method (Nesterov, 1983, 2004, 2005) (the recursions for these algorithms are described in Section 3.1). The two methods are different but they both rely on the concept of adding a momentum term to the recursion. When the risk function $J(w)$ is ν-strongly convex and has δ-Lipschitz continuous gradients, both methods succeed in accelerating the gradient descent algorithm to attain a faster exponential convergence rate (Polyak, 1987) (Nesterov, 2004), and this rate is proven to be optimal for problems with smooth $J(w)$ and cannot be attained by standard gradient descent methods. Specifically, it is shown in (Polyak, 1987) (Nesterov, 2004) that for heavy-ball and Nesterov's acceleration methods, the convergence of the iterates w_i towards w^o occurs at the rate:

$$\|w_i - w^o\|^2 \leq \left(\frac{\sqrt{\delta} - \sqrt{\nu}}{\sqrt{\delta} + \sqrt{\nu}}\right)^2 \|w_{i-1} - w^o\|^2, \tag{4}$$

In comparison, in Theorem 2.1.15 of (Nesterov, 2005) and Theorem 4 in Section 1.4 of (Polyak, 1987), the fastest rate for gradient descent method is shown to be

$$\|w_i - w^o\|^2 \le \left(\frac{\delta-\nu}{\delta+\nu}\right)^2 \|w_{i-1} - w^o\|^2. \tag{5}$$

It can be verified that

$$\frac{\sqrt{\delta}-\sqrt{\nu}}{\sqrt{\delta}+\sqrt{\nu}} < \frac{\delta-\nu}{\delta+\nu} \tag{6}$$

when $\delta > \nu$. This inequality confirms that the momentum algorithm can achieve a faster rate in deterministic optimization and, moreover, this faster rate cannot be attained by standard gradient descent.

Motivated by these useful acceleration properties in the *deterministic* context, momentum terms have been subsequently introduced into *stochastic* optimization algorithms as well (Polyak, 1987; Proakis, 1974; Sharma et al., 1998; Shynk and Roy, June 1988; Roy and Shynk, 1990; Tugay and Tanik, 1989; Bellanger, 2001; Wiegerinck et al., 1994; Hu et al., 2009; Xiao, 2010; Lan, 2012; Ghadimi and Lan, 2012; Zhong and Kwok, 2014) and applied, for example, to problems involving the tracking of chirped sinusoidal signals (Ting et al., 2000) or deep learning (Sutskever et al., 2013; Kahou et al., 2013; Szegedy et al., 2015; Zareba et al., 2015). However, the analysis in this paper will show that the advantages of the momentum technique for deterministic optimization do not necessarily carry over to the *adaptive* online setting due to the presence of stochastic gradient noise (which is the difference between the actual gradient vector and its approximation). Specifically, for sufficiently small step-sizes and for a momentum paramter not too close to one, we will show that any advantage brought forth by the momentum term can be achieved by staying with the original stochastic-gradient algorithm and adjusting its step-size to a larger value. For instance, for optimization problem (1), we will show that if the step-sizes, μ_m for the momentum (heavy-ball or Nesterov) methods and μ for the standard stochastic gradient algorithms, are sufficiently small and satisfy the relation

$$\mu = \frac{\mu_m}{1-\beta} \tag{7}$$

where β, a positive constant that is not too close to 1, is the momentum parameter, then it will hold that

$$\mathbb{E}\|\boldsymbol{w}_{m,i} - \boldsymbol{w}_i\|^2 = O(\mu^{3/2}),\ i = 0, 1, 2, \ldots \tag{8}$$

where $\boldsymbol{w}_{m,i}$ and $\boldsymbol{w}_i$ denote the iterates generated at time i by the momentum and standard implementations, respectively. In the special case when $J(w)$ is quadratic in w, as happens in mean-square-error design problems, we can tighten (8) to

$$\mathbb{E}\|\boldsymbol{w}_{m,i} - \boldsymbol{w}_i\|^2 = O(\mu^2),\ i = 0, 1, 2, \ldots \tag{9}$$

What is important to note is that, we will show that these results hold *for every* i, and not only asymptotically. Therefore, when μ is sufficiently small, property (8) establishes that the stochastic gradient method and the momentum versions are fundamentally equivalent since their iterates evolve close to each other at all times. We establish this equivalence result under the situation where the risk function is convex and differentiable. However, as our numerical simulations over a multi-layer fully connected neural network and a second convolutional neural network (see Section 7.4) show, the equivalence between standard and momentum stochastic gradient methods are also observed in non-convex and non-differentiable scenarios.

1.3 Related Works in the Literature

There are useful results in the literature that deal with special instances of the general framework developed in this work. These earlier results focus mainly on the mean-square-error case when $J(w)$ is quadratic in w, in which case the stochastic gradient algorithm reduces to the famed least-mean-squares (LMS) algorithm. We will not be limiting our analysis to this case so that our results will be applicable to a broader class of learning problems beyond mean-square-error estimation (e.g., logistic regression would be covered by our results as well). As the analysis and derivations will reveal, the treatment of the general $J(w)$ case is demanding because the Hessian matrix of $J(w)$ is now $w-$dependent, whereas it is a constant matrix in the quadratic case.

Some of the earlier investigations in the literature led to the following observations. It was noted in (Polyak, 1987) that, for quadratic costs, stochastic gradient implementations with a momentum term do not necessarily perform well. This work remarks that although the heavy-ball method can lead to faster convergence in the early stages of learning, it nevertheless converges to a region with worse mean-square-error in comparison to standard stochastic-gradient (or LMS) iteration. A similar phenomenon is also observed in (Proakis, 1974; Sharma et al., 1998). However, in the works (Proakis, 1974; Polyak, 1987; Sharma et al., 1998), no claim is made or established about the equivalence between momentum and standard methods.

Heavy-ball LMS was further studied in the useful works (Roy and Shynk, 1990) and (Tugay and Tanik, 1989). The reference (Roy and Shynk, 1990) claimed that no significant gain is achieved in convergence speed if both the heavy-ball and standard LMS algorithms approach the same *steady-state* MSE performance. Reference (Tugay and Tanik, 1989) observed that when the step-sizes satisfy relation (7), then heavy-ball LMS is "equivalent" to standard LMS. However, they assumed Gaussian measurement noise in their data model, and the notion of "equivalence" in this work is only referring to the fact that the algorithms have similar starting convergence rates and similar steady-state MSE levels. There was no analysis in (Tugay and Tanik, 1989) of the behavior of the algorithms during all stages of learning – see also (Bellanger, 2001). Another useful work is (Wiegerinck et al., 1994), which considered the heavy-ball stochastic gradient method for general risk, $J(w)$. By assuming a sufficiently small step-size, and by transforming the error difference recursion into a differential equation, the work concluded that heavy-ball can be equivalent to the standard stochastic gradient method asymptotically (i.e., for i large enough). No results were provided for the earlier stages of learning.

All of these previous works were limited to examining the heavy-ball momentum technique; none of them considered other forms of acceleration such as Nesterov's technique although this latter technique is nowadays widely applied to stochastic gradient learning, including deep learning (Sutskever et al., 2013; Kahou et al., 2013; Szegedy et al., 2015; Zareba et al., 2015). The performance of Nesterov's acceleration with *deterministic* and *bounded* gradient error was examined in (d'Aspremont, 2008; Devolder et al., 2014; Lessard et al., 2016). The source of the inaccuracy in the gradient vector in these works is either because the gradient was assessed by solving an auxiliary "simpler" optimization problem or because of numerical approximations. Compared to the standard gradient descent implementation, the works by (d'Aspremont, 2008; Lessard et al., 2016) claimed that Nesterov's acceleration is not robust to the errors in gradient. The work by (Devolder et al., 2014) also observed that the superiority of Nesterov's acceleration is no longer absolute when inexact gradients are used, and they further proved that the performance of Nesterov's acceleration may be even worse than gradient descent due to error accumulation. These works assumed bounded errors in the gradient vectors and focused on the context of deterministic optimization. None of the works examined the stochastic setting where the gradient error is random in nature and where the assumption of bounded errors are generally unsuitable. We may add that there have also been analyses of Nesterov's acceleration for *stochastic* optimization problems albeit for *decaying* step-sizes in more recent literature (Hu et al., 2009; Xiao, 2010; Lan, 2012; Ghadimi and Lan, 2012; Zhong and Kwok, 2014). These works proved that Nesterov's acceleration can improve the convergence rate of stochastic gradient descent at the initial stages when deterministic risk components dominate; while at the asymptotic stages when the stochastic gradient noise dominates, the momentum correction cannot accelerate convergence any more. Another useful study is (Flammarion and Bach, 2015), in which the authors showed that momentum and averaging methods for stochastic optimization are equivalent to the same second-order difference equations but with different step-sizes. However, (Flammarion and Bach, 2015) does not study the equivalence between standard and momentum stochastic gradient methods, and they focus on quadratic problems and also employ decaying step-sizes.

Finally, we note that there are other forms of stochastic gradient algorithms for empirical risk minimization problems where momentum acceleration has been shown to be useful. Among them, we list recent algorithms like SAG (Roux et al., 2012), SVRG (Johnson and Zhang, 2013) and SAGA (Defazio et al., 2014). In these algorithms, the variance of the stochastic gradient noise diminishes to zero and the deterministic component of the risk becomes dominant in the asymptotic regime. In these situations, momentum acceleration helps improve the convergence rate, as noted by (Nitanda, 2014) and (Zhu, 2016). Another family of algorithms to solve empirical risk minimization problems are stochastic dual coordinate ascent (SDCA) algorithms. It is proved in (Shalev-Shwartz, 2015; Johnson and Zhang, 2013) that SDCA can be viewed as a variance-reduced stochastic algorithm, and hence momentum acceleration can also improve its convergence for the same reason noted by (Shalev-Shwartz and Zhang, 2014).

In this paper, we are studying online training algorithms where data can stream in continuously as opposed to running multiple passes over a finite amount of data. In this case, the analysis will help clarify the limitations of momentum acceleration in the slow adaptation regime. We are particularly interested in the constant step-size case, which enables continuous adaptation and learning and is regularly used, e.g., in deep learning implementations. There is a non-trivial difference between the decaying and constant step-size situations. This is because gradient noise is always present in the constant step-size case, while it is annihilated in the decaying step-size case. The presence of the gradient noise interferes with the dynamics of the algorithms in a non-trivial way, which is what our analysis discovers. There are limited analyses for the constant step-sizes scenario.

1.4 Outline of Paper

The outline of the paper is as follows. In Section 2, we introduce some basic assumptions and review the stochastic gradient method and its convergence properties. In Section 3 we embed the heavy-ball and Nesterov's acceleration methods into a unified momentum algorithm, and subsequently establish the mean-square stability and fourth-order stability of the error moments. Next, we analyze the equivalence between momentum and standard LMS algorithms in Section 4 and then extend the results to general risk functions in Section 5. In Section 6 we extend the equivalence results into a more general setting with diagonal step-size matrices. We illustrate our results in Section 7, and in Section 8 we comment on the stability ranges of standard and momentum stochastic gradient methods.

2. Stochastic Gradient Algorithms

In this section we review the stochastic gradient method and its convergence properties. We denote the minimizer for problem (1) by w^o, i.e.,

$$w^o \stackrel{\triangle}{=} \arg\min_{w} \ J(w). \tag{10}$$

We introduce the following assumption on $J(w)$, which essentially amounts to assuming that $J(w)$ is strongly-convex with Lipschitz gradient. These conditions are satisfied by many problems of interest, especially when regularization is employed (e.g., mean-square-error risks, logistic risks, etc.). Under the strong-convexity condition, the minimizer w^o is unique.

Assumption 1 (Conditions on risk function) *The cost function $J(w)$ is twice differentiable and its Hessian matrix satisfies*

$$0 < \nu I_M \le \nabla^2 J(w) \le \delta I_M, \tag{11}$$

for some positive parameters $\nu \le \delta$. Condition (11) is equivalent to requiring $J(w)$ to be ν-strongly convex and for its gradient vector to be δ-Lipschitz, respectively (Boyd and Vandenberghe, 2004; Sayed, 2014a).

■

The stochastic-gradient algorithm for seeking w^o takes the form (2), with initial condition $\boldsymbol{w}_{-1}$. The difference between the true gradient vector and its approximation is designated *gradient noise* and is denoted by:

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) \stackrel{\triangle}{=} \nabla_w Q(\boldsymbol{w}_{i-1};\boldsymbol{\theta}_i) - \nabla_w \mathbb{E}[Q(\boldsymbol{w}_{i-1};\boldsymbol{\theta}_i)]. \tag{12}$$

In order to examine the convergence of the standard and momentum stochastic gradient methods, it is necessary to introduce some assumptions on the stochastic gradient noise.

Assumptions (13) and (14) below are satisfied by important cases of interest, as shown in (Sayed, 2014a) and (Sayed, 2014b), such as logistic regression and mean-square-error risks. Let the symbol $\boldsymbol{\mathcal{F}}_{i-1}$ represent the filtration generated by the random process $\boldsymbol{w}_j$ for $j \leq i-1$ (basically, the collection of past history until time $i-1$):

$$\boldsymbol{\mathcal{F}}_{i-1} \stackrel{\triangle}{=} \text{filtration}\{\boldsymbol{w}_{-1}, \boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i-1}\}.$$

Assumption 2 (Conditions on gradient noise) *It is assumed that the first and second-order conditional moments of the gradient noise process satisfy the following conditions for any* $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$*:*

$$\mathbb{E}[\boldsymbol{s}_i(\boldsymbol{w})|\boldsymbol{\mathcal{F}}_{i-1}] = 0 \tag{13}$$

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{w})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \leq \gamma^2\|w^o - \boldsymbol{w}\|^2 + \sigma_s^2 \tag{14}$$

almost surely, for some nonnegative constants γ^2 *and* σ_s^2*.*

■

Condition (13) essentially requires the gradient noise process to have zero mean, which amounts to requiring the approximate gradient to correspond to an unbiased construction for the true gradient. This is a reasonable requirement. Condition (14) requires the size of the gradient noise (i.e., its mean-square value) to diminish as the iterate $\boldsymbol{w}$ gets closer to the solution w^o. This is again a reasonable requirement since it amounts to expecting the gradient noise to get reduced as the algorithm approaches the minimizer. Under Assumptions 1 and 2, the following conclusion is proven in Lemma 3.1 of (Sayed, 2014a).

Lemma 1 (Second-order stability) *Let Assumptions 1 and 2 hold, and consider the stochastic gradient recursion* (2)*. Introduce the error vector* $\widetilde{\boldsymbol{w}}_i = w^o - \boldsymbol{w}_i$*. Then, for any step-sizes* μ *satisfying*

$$\mu < \frac{2\nu}{\delta^2 + \gamma^2}, \tag{15}$$

it holds for each iteration $i = 0, 1, 2, \ldots$ *that*

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \leq (1-\mu\nu)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \mu^2\sigma_s^2, \tag{16}$$

and, furthermore,

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \leq \frac{\sigma_s^2\mu}{\nu} = O(\mu). \tag{17}$$

■

We can also examine the the stability of the fourth-order error moment, $\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4$, which will be used later in Section 5 to establish the equivalence between the standard and momentum stochastic implementations. For this case, we tighten the assumption on the gradient noise by replacing the bound in (14) on its second-error moment by a similar bound involving its fourth-order moment. Again, this assumption is satisfied by problems of interest, such as mean-square-error and logistic risks (Sayed, 2014a,b).

Assumption 3 (Conditions on gradient noise) *It is assumed that the first and fourth-order conditional moments of the gradient noise process satisfy the following conditions for any* $\boldsymbol{w} \in \boldsymbol{\mathcal{F}}_{i-1}$*:*

$$\mathbb{E}[\boldsymbol{s}_i(\boldsymbol{w})|\boldsymbol{\mathcal{F}}_{i-1}] = 0 \tag{18}$$

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{w})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \leq \gamma_4^4\|w^o - \boldsymbol{w}\|^4 + \sigma_{s,4}^4 \tag{19}$$

almost surely, for some nonnegative constants γ_4^4 *and* $\sigma_{s,4}^4$*.* ■

It is straightforward to check that if Assumption 3 holds, then Assumption 2 will also hold. The following conclusion is a modified version of Lemma 3.2 of (Sayed, 2014a).

Lemma 2 (Fourth-order stability) *Let the conditions under Assumptions 1 and 3 hold, and consider the stochastic gradient iteration* (2)*. For sufficiently small step-size* μ*, it holds that*

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 \leq \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^4 + A\sigma_s^2(i+1)\rho^{i+1}\mu^2 + \frac{B\sigma_s^4\mu^2}{\nu^2} \tag{20}$$

where $\rho \stackrel{\Delta}{=} 1 - \mu\nu$*, and* A *and* B *are some constants. Furthermore,*

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 \leq \frac{B\sigma_s^4\mu^2}{\nu^2} = O\left(\mu^2\right) \tag{21}$$

Proof See Appendix A. ■

3. Momentum Acceleration

In this section, we present a generalized momentum stochastic gradient method, which captures both the heavy-ball and Nesterov's acceleration methods as special cases. Subsequently, we derive results for its convergence property.

3.1 Momentum Stochastic Gradient Method

Consider the following general form of a stochastic-gradient implementation, with two momentum parameters $\beta_1, \beta_2 \in [0, 1)$:

$$\boldsymbol{\psi}_{i-1} = \boldsymbol{w}_{i-1} + \beta_1(\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}), \tag{22}$$

$$\boldsymbol{w}_i = \boldsymbol{\psi}_{i-1} - \mu_m \nabla_w Q(\boldsymbol{\psi}_{i-1}; \boldsymbol{\theta}_i) + \beta_2(\boldsymbol{\psi}_{i-1} - \boldsymbol{\psi}_{i-2}), \tag{23}$$

with initial conditions

$$\boldsymbol{w}_{-2} = \boldsymbol{\psi}_{-2} = \text{initial states}, \tag{24}$$

$$\boldsymbol{w}_{-1} = \boldsymbol{w}_{-2} - \mu_m \nabla_w Q(\boldsymbol{w}_{-2}; \boldsymbol{\theta}_{-1}), \tag{25}$$

where μ_m is some constant step-size. We refer to this formulation as the momentum stochastic gradient method. [1]

When $\beta_1 = 0$ and $\beta_2 = \beta$ we recover the heavy-ball algorithm (Polyak, 1964, 1987), and when $\beta_2 = 0$ and $\beta_1 = \beta$, we recover Nesterov's algorithm (Nesterov, 2004). We note that Nesterov's method has several useful variations that fit different scenarios, such as situations involving smooth but not strongly-convex risks (Nesterov, 1983, 2004) or non-smooth risks (Nesterov, 2005; Beck and Teboulle, 2009). However, for the case when $J(w)$ is strongly convex and has Lipschitz continuous gradients, the Nesterov construction reduces to what is presented above, with a constant momentum parameter. This type of construction has also been studied in (Lessard et al., 2016; Dieuleveut et al., 2016) and applied in deep learning implementations (Sutskever et al., 2013; Kahou et al., 2013; Szegedy et al., 2015; Zareba et al., 2015).

In order to capture both the heavy-ball and Nesterov's acceleration methods in a unified treatment, we will assume that

$$\beta_1 + \beta_2 = \beta, \quad \beta_1 \beta_2 = 0, \tag{26}$$

for some fixed constant $\beta \in [0, 1)$. Next we introduce a condition on the momentum parameter.

Assumption 4 *The momentum parameter β is a constant that is not too close to 1, i.e., there exists a small fixed constant $\epsilon > 0$ such that $\beta \leq 1 - \epsilon$.* ■

Assumption 4 is quite common in studies on adaptive signal processing and neural networks — see, e.g., (Tugay and Tanik, 1989; Roy and Shynk, 1990; Bellanger, 2001; Wiegerinck et al., 1994; Attoh-Okine, 1999). Also, in recent deep learning applications it is common to set $\beta = 0.9$, which satisfies Assumption 4 (Krizhevsky et al., 2012; Szegedy et al., 2015; Zhang and LeCun, 2015). Under (26), the work (Flammarion and Bach, 2015) also considers recursions related to (22)–(23) for the special case of quadratic risks.

3.2 Mean-Square Error Stability

In preparation for studying the performance of the momentum stochastic gradient method, we first show in the next result how recursions (22)-(23) can be transformed into a first-order recursion by defining extended state vectors. We introduce the transformation matrices:

$$V = \begin{bmatrix} I_M & -\beta I_M \\ I_M & -I_M \end{bmatrix}, \; V^{-1} = \frac{1}{1-\beta} \begin{bmatrix} I_M & -\beta I_M \\ I_M & -I_M \end{bmatrix}. \tag{27}$$

Recall $\widetilde{\boldsymbol{w}}_i = w^o - \boldsymbol{w}_i$ and define the transformed error vectors, each of size $2M \times 1$:

$$\begin{bmatrix} \widehat{\boldsymbol{w}}_i \\ \check{\boldsymbol{w}}_i \end{bmatrix} \overset{\Delta}{=} V^{-1} \begin{bmatrix} \widetilde{\boldsymbol{w}}_i \\ \widetilde{\boldsymbol{w}}_{i-1} \end{bmatrix} = \frac{1}{1-\beta} \begin{bmatrix} \widetilde{\boldsymbol{w}}_i - \beta \widetilde{\boldsymbol{w}}_{i-1} \\ \widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{w}}_{i-1} \end{bmatrix}. \tag{28}$$

1. Traditionally, the terminology of a "momentum method" has been used more frequently for the heavy-ball method, which corresponds to the special case $\beta_1 = 0$ and $\beta_2 = \beta$. Given the unified description (22)–(23), we will use this same terminology to refer to both the heavy-ball and Nesterov's acceleration methods.

Lemma 3 (Extended recursion) *Under Assumption 1 and condition* (26)*, the momentum stochastic gradient recursion* (22)–(23) *can be transformed into the following extended recursion:*

$$\begin{bmatrix} \widehat{\boldsymbol{w}}_i \\ \check{\boldsymbol{w}}_i \end{bmatrix} = \begin{bmatrix} I_M - \frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1} & \frac{\mu_m \beta'}{1-\beta}\boldsymbol{H}_{i-1} \\ -\frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1} & \beta I_M + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{w}}_{i-1} \\ \check{\boldsymbol{w}}_{i-1} \end{bmatrix} + \frac{\mu_m}{1-\beta}\begin{bmatrix} \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) \\ \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) \end{bmatrix}, \tag{29}$$

where $\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})$ *is defined according to* (12) *and*

$$\beta' \triangleq \beta\beta_1 + \beta_2, \tag{30}$$

$$\boldsymbol{H}_{i-1} \triangleq \int_0^1 \nabla_w^2 J(w^o - t\widetilde{\boldsymbol{\psi}}_{i-1})dt, \tag{31}$$

where $\widetilde{\boldsymbol{\psi}}_{i-1} = w^o - \boldsymbol{\psi}_{i-1}$.

Proof See Appendix B. ■

The transformed recursion (29) is important for at least two reasons. First, it is a first-order recursion, which facilitates the convergence analysis of $\widehat{\boldsymbol{w}}_i$ and $\check{\boldsymbol{w}}_i$ and, subsequently, of the error vector $\widetilde{\boldsymbol{w}}_i$ in view of (28) — see next theorem. Second, as we will explain later, the first row of (29) turns out to be closely related to the standard stochastic gradient iteration; this relation will play a critical role in establishing the claimed equivalence between momentum and standard stochastic gradient methods.

The following statement establishes the convergence property of the momentum stochastic gradient algorithm. It shows that recursions (22)–(23) converge exponentially fast to a small neighborhood around w^o with a steady-state error variance that is on the order of $O(\mu_m)$. Note that in the following theorem the notation $a \preceq b$, for two vectors a and b, signifies element-wise comparisons.

Theorem 4 (Mean-square stability) *Let Assumptions 1 , 2 and 4 hold and recall conditions* (26)*. Consider the momentum stochastic gradient method* (22)–(23) *and the extended recursion* (29)*. Then, when step-sizes* μ_m *satisfies*

$$\mu_m \le \frac{(1-\beta)^2\nu}{32\gamma^2\nu^2 + 4\delta^2}, \tag{32}$$

it holds that the mean-square values of the transformed error vectors evolve according to the following recursive inequality:

$$\begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \end{bmatrix} \preceq \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}, \tag{33}$$

where

$$\begin{array}{lll} a = 1 - \frac{\mu_m\nu}{1-\beta} + O(\mu_m^2), & b = \frac{\mu_m\beta'^2\delta^2}{\nu(1-\beta)} + O(\mu_m^2), & c = \frac{2\mu_m^2\delta^2}{(1-\beta)^3} + \frac{2\mu_m^2\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2} \\ d = \beta + O(\mu_m^2), & e = \frac{\mu_m^2\sigma_s^2}{(1-\beta)^2}, & f = \frac{\mu_m^2\sigma_s^2}{(1-\beta)^2} \end{array} \tag{34}$$

and the coefficient matrix appearing in (33) is stable, namely,

$$\rho\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) < 1. \tag{35}$$

Furthermore, if μ_m is sufficiently small it follows from (33) *that*

$$\limsup_{i\to\infty} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 = O\left(\frac{\mu_m \sigma_s^2}{(1-\beta)\nu}\right), \qquad \limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 = O\left(\frac{\mu_m^2 \sigma_s^2}{(1-\beta)^3}\right), \tag{36}$$

and, consequently,

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 = O\left(\frac{\mu_m \sigma_s^2}{(1-\beta)\nu}\right). \tag{37}$$

Proof See Appendix C. ■

Although $\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 = O(\mu_m^2)$ in result (36) is shown to hold asymptotically in the statement of the theorem, it can actually be strengthened and shown to hold for *all* time instants. This fact is crucial for our later proof of the equivalence between standard and momentum stochastic gradient methods.

Corollary 5 (Uniform mean-square bound) *Under the same conditions as Theorem 4, it holds for sufficiently small step-sizes that*

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 = O\left(\frac{(\delta^2+\gamma^2)\rho_1^{i+1}\mu_m^2}{(1-\beta)^4} + \frac{\sigma_s^2 \mu_m^2}{(1-\beta)^3}\right), \forall i = 0, 1, 2, \cdots \tag{38}$$

where $\rho_1 \overset{\Delta}{=} 1 - \frac{\mu_m \nu}{2(1-\beta)}$, and $\check{\boldsymbol{w}}_i$ is defined in (29)*.*

Proof See Appendix D. ■

Corollary 5 has two implications. First, since β, δ, γ, σ_s^2 are all constants, and $\rho_1 < 1, \alpha < 1$, we conclude that

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 = O(\mu_m^2), \quad \forall i = 0, 1, 2, \cdots \tag{39}$$

Besides, since $\rho_1^i \to 0$ as $i \to \infty$, according to (38) we also achieve

$$\limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 = O\left(\frac{\sigma_s^2 \mu_m^2}{(1-\beta)^3}\right), \tag{40}$$

which is consistent with (36).

3.3 Stability of Fourth-Order Error Moment

In a manner similar to the treatment in Section 2, we can also establish the convergence of the fourth-order moments of the error vectors, $\mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^4$ and $\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4$.

Theorem 6 (Fourth-order stability) *Let Assumptions 1, 3 and 4 hold and recall conditions (26). Then, for sufficiently small step-sizes μ_m, it holds that*

$$\limsup_{i\to\infty} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^4 = O(\mu_m^2), \tag{41}$$

$$\limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 = O\left(\mu_m^4\right), \tag{42}$$

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 = O\left(\mu_m^2\right). \tag{43}$$

Proof See Appendix E. ■

Again, result (42) is only shown to hold asymptotically in the statement of the theorem. In fact, $\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4$ can also be shown to be bounded for all time instants, as the following corollary states.

Corollary 7 (Uniform forth-moment bound) *Under the same conditions as Theorem 6, it holds for sufficiently small step-sizes that*

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 = O\left(\frac{\gamma^2\rho_2^{i+1}}{(1-\beta)^3}\mu_m^2 + \left[\frac{\sigma_s^2(\delta^2+\gamma^2)(i+1)\rho_2^{i+1}}{(1-\beta)^7} + \frac{(\gamma^2+\nu^2)\sigma_s^4+\nu^2\sigma_{s,4}^4}{(1-\beta)^6\nu^2}\right]\mu_m^4\right) \tag{44}$$

where $\rho_2 \stackrel{\Delta}{=} 1 - \frac{\mu_m\nu}{4(1-\beta)} \in (0,1)$.

Proof See Appendix F. ■

Corollary 7 also has two implications. First, since β, δ, γ, σ_s and $\sigma_{s,4}$ are constants, we conclude that

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 = O(\mu_m^2), \quad \forall i = 0,1,2,\cdots \tag{45}$$

Besides, since $\rho_2^i \to 0$ and $i\rho_2^i \to 0$ as $i \to \infty$, we will achieve the following fact according to (44)

$$\limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 = O\left(\frac{(\gamma^2+\nu^2)\sigma_s^4+\nu^2\sigma_{s,4}^4}{(1-\beta)^6\nu^2}\mu_m^4\right) = O(\mu_m^4), \tag{46}$$

which is consistent with (42).

4. Equivalence in the Quadratic Case

In Section 3 we showed the momentum stochastic gradient algorithm (22)–(23) converges exponentially for sufficiently small step-sizes. But some important questions remain. Does the momentum implementation converge faster than the standard stochastic gradient method (2)? Does the momentum implementation lead to superior steady-state mean-square-deviation (MSD) performance, measured in terms of the limiting value of $\mathbb{E}\,\|\widetilde{\boldsymbol{w}}_i\|^2$? Is the momentum method generally superior to the standard method when considering both the convergence rate and MSD performance? In this and the next sections, we answer these questions in some detail. Before treating the case of general risk functions, $J(w)$, we examine first the special case when $J(w)$ is quadratic in w to illustrate the main conclusions that will follow.

4.1 Quadratic Risks

We consider mean-square-error risks of the form

$$J(w) = \frac{1}{2}\mathbb{E}\left(\boldsymbol{d}(i) - \boldsymbol{u}_i^\mathsf{T} w\right)^2, \tag{47}$$

where $\boldsymbol{d}(i)$ denotes a streaming sequence of zero-mean random variables with variance $\sigma_d^2 = \mathbb{E}\boldsymbol{d}^2(i)$, and $\boldsymbol{u}_i \in \mathbb{R}^M$ denotes a streaming sequence of independent zero-mean random vectors with covariance matrix $R_u = \mathbb{E}\boldsymbol{u}_i\boldsymbol{u}_i^\mathsf{T} > 0$. The cross covariance vector between $\boldsymbol{d}(i)$ and $\boldsymbol{u}_i$ is denoted by $r_{du} = \mathbb{E}\boldsymbol{d}(i)\boldsymbol{u}_i$. The data $\{\boldsymbol{d}(i), \boldsymbol{u}_i\}$ are assumed to be wide-sense stationary and related via a linear regression model of the form:

$$\boldsymbol{d}(i) = \boldsymbol{u}_i^\mathsf{T} w^o + \boldsymbol{v}(i), \tag{48}$$

for some unknown w^o, and where $\boldsymbol{v}(i)$ is a zero-mean white noise process with power $\sigma_v^2 = \mathbb{E}\boldsymbol{v}^2(i)$ and assumed independent of $\boldsymbol{u}_j$ for all i, j. If we multiply (48) by $\boldsymbol{u}_i$ from the left and take expectations, we find that the model parameter w^o satisfies the normal equations $R_u w^o = r_{du}$. The unique solution that minimizes (47) also satisfies these same equations. Therefore, minimizing the quadratic risk (47) enables us to recover the desired w^o. This observation explains why mean-square-error costs are popular in the context of regression models.

4.2 Adaptation Methods

For the least-mean-squares problem (47), the true gradient vector at any location $\boldsymbol{w}$ is

$$\nabla_w J(\boldsymbol{w}) = R_u\boldsymbol{w} - r_{du} \;=\; -R_u(w^o - \boldsymbol{w}), \tag{49}$$

while the approximate gradient vector constructed from an instantaneous sample realization is:

$$\nabla_w Q(\boldsymbol{w}; \boldsymbol{d}(i), \boldsymbol{u}_i) = -\boldsymbol{u}_i(\boldsymbol{d}(i) - \boldsymbol{u}_i^\mathsf{T}\boldsymbol{w}). \tag{50}$$

Here the loss function is defined by

$$Q(w; \boldsymbol{d}(i), \boldsymbol{u}_i) \;\stackrel{\Delta}{=}\; \frac{1}{2}\mathbb{E}\left(\boldsymbol{d}(i) - \boldsymbol{u}_i^\mathsf{T} w\right)^2 \tag{51}$$

The resulting LMS (stochastic-gradient) recursion is given by

$$\boldsymbol{w}_i = \boldsymbol{w}_{i-1} + \mu\boldsymbol{u}_i(\boldsymbol{d}(i) - \boldsymbol{u}_i^\mathsf{T}\boldsymbol{w}_{i-1}) \tag{52}$$

and the corresponding gradient noise process is

$$\boldsymbol{s}_i(\boldsymbol{w}) = (R_u - \boldsymbol{u}_i\boldsymbol{u}_i^\mathsf{T})(w^o - \boldsymbol{w}) - \boldsymbol{u}_i\boldsymbol{v}(i). \tag{53}$$

It can be verified that this noise process satisfies Assumption 2 — see Example 3.3 in (Sayed, 2014a). Subtracting w^o from both sides of (52), and recalling that $\widetilde{\boldsymbol{w}}_i = w^o - \boldsymbol{w}_i$, we obtain the error recursion that corresponds to the LMS implementation:

$$\widetilde{\boldsymbol{w}}_i = (I_M - \mu R_u)\widetilde{\boldsymbol{w}}_{i-1} + \mu\boldsymbol{s}_i(\boldsymbol{w}_{i-1}), \tag{54}$$

where μ is some constant step-size. In order to distinguish the variables for LMS from the variables for the momentum LMS version described below, we replace the notation $\{\boldsymbol{w}_i, \widetilde{\boldsymbol{w}}_i\}$ for LMS by $\{\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i\}$ and keep the notation $\{\boldsymbol{w}_i, \widetilde{\boldsymbol{w}}_i\}$ for momentum LMS, i.e., for the LMS implementation (54) we shall write instead

$$\widetilde{\boldsymbol{x}}_i = (I_M - \mu R_u)\widetilde{\boldsymbol{x}}_{i-1} + \mu\boldsymbol{s}_i(\boldsymbol{x}_{i-1}). \tag{55}$$

On the other hand, we conclude from (22)–(23) that the momentum LMS recursion will be given by:

$$\begin{aligned} \boldsymbol{\psi}_{i-1} &= \boldsymbol{w}_{i-1} + \beta_1(\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}), &(56)\\ \boldsymbol{w}_i &= \boldsymbol{\psi}_{i-1} + \mu_m\boldsymbol{u}_i(\boldsymbol{d}(i) - \boldsymbol{u}_i^\mathsf{T}\boldsymbol{\psi}_{i-1}) + \beta_2(\boldsymbol{\psi}_{i-1} - \boldsymbol{\psi}_{i-2}), &(57) \end{aligned}$$

Using the transformed recursion (29), we can transform the resulting relation for $\widetilde{\boldsymbol{w}}_i$ into:

$$\begin{bmatrix} \widehat{\boldsymbol{w}}_i \\ \check{\boldsymbol{w}}_i \end{bmatrix} = \begin{bmatrix} I_M - \frac{\mu_m}{1-\beta}R_u & \frac{\mu_m\beta'}{1-\beta}R_u \\ -\frac{\mu_m}{1-\beta}R_u & \beta I_M + \frac{\mu_m\beta'}{1-\beta}R_u \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{w}}_{i-1} \\ \check{\boldsymbol{w}}_{i-1} \end{bmatrix} + \frac{\mu_m}{1-\beta}\begin{bmatrix} \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) \\ \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) \end{bmatrix}, \tag{58}$$

where the Hessian matrix, $\boldsymbol{H}_{i-1}$, is independent of the weight iterates and given by R_u for quadratic risks. It follows from the first row that

$$\widehat{\boldsymbol{w}}_i = \left(I_M - \frac{\mu_m}{1-\beta}R_u\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'}{1-\beta}R_u\check{\boldsymbol{w}}_{i-1} + \frac{\mu_m}{1-\beta}\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{59}$$

Next, we assume the step-sizes $\{\mu, \mu_m\}$ and the momentum parameter are selected to satisfy

$$\mu = \frac{\mu_m}{1-\beta}. \tag{60}$$

Since $\beta \in [0, 1)$, this means that $\mu_m < \mu$. Then, recursion (59) becomes

$$\widehat{\boldsymbol{w}}_i = (I_M - \mu R_u)\widehat{\boldsymbol{w}}_{i-1} + \mu\beta' R_u\check{\boldsymbol{w}}_{i-1} + \mu\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{61}$$

Comparing (61) with the LMS recursion (55), we find that both relations are quite similar, except that the momentum recursion has an extra driving term dependent on $\check{\boldsymbol{w}}_{i-1}$. However, recall from (28) that $\check{\boldsymbol{w}}_{i-1} = (\widetilde{\boldsymbol{w}}_{i-2} - \widetilde{\boldsymbol{w}}_{i-1})/(1-\beta)$, which is the difference between two consecutive points generated by momentum LMS. Intuitively, it is not hard to see that $\check{\boldsymbol{w}}_{i-1}$ is in the order of $O(\mu)$, which makes $\mu\beta' R_u\check{\boldsymbol{w}}_{i-1}$ in the order of $O(\mu^2)$. When the step-size μ is very small, this $O(\mu^2)$ term can be ignored. Consequently, the above recursions for $\widehat{\boldsymbol{w}}_i$ and $\widetilde{\boldsymbol{x}}_i$ should evolve close to each other, which would help to prove that $\boldsymbol{w}_i$ and $\boldsymbol{x}_i$ will also evolve close to each other as well. This conclusion can be established formally as follows, which proves the equivalence between the momentum and standard LMS methods.

Theorem 8 (Equivalence for LMS) *Consider the LMS and momentum LMS recursions (52) and (56)–(57). Let Assumptions 1, 2 and 4 hold. Assume both algorithms start from the same initial states, namely, $\boldsymbol{\psi}_{-2} = \boldsymbol{w}_{-2} = \boldsymbol{x}_{-1}$. Suppose conditions (26) holds, and that the step-sizes $\{\mu, \mu_m\}$ satisfy (60). Then, it holds for sufficiently small μ that for $\forall\ i = 0, 1, 2, 3, \cdots$*

$$\mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2 = O\left(\left[\frac{\delta^2+\gamma^2}{(1-\beta)^2}\rho_1^{i+1} + \frac{\delta^2\sigma_s^2}{\nu^2(1-\beta)}\right]\mu^2 + \frac{\delta^2(\delta^2+\gamma^2)(i+1)\rho_1^{i+1}}{\nu(1-\beta)^2}\mu^3\right). \tag{62}$$

where $\rho_1 = 1 - \frac{\mu\nu}{2} \in (0,1)$.

Proof See Appendix G. ∎

Similar to Corollary 5 and 7, Theorem 8 also has two implications. First, it holds that

$$\mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2 = O(\mu^2), \quad \forall i = 0, 1, 2, \cdots \tag{63}$$

Besides, since $\rho_1^i \to 0$ and $i\rho_1^i$ as $i \to \infty$, we also conclude

$$\limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2 = O\left(\frac{\delta^2\sigma_s^2\mu^2}{\nu^2(1-\beta)}\right). \tag{64}$$

Theorem 8 establishes that the standard and momentum LMS algorithms are fundamentally equivalent since their iterates evolve close to each other at all times for sufficiently small step-sizes. More interpretation of this result is discussed in Section 5.2.

5. Equivalence in the General Case

We now extend the analysis from quadratic risks to more general risks (such as logistic risks). The analysis in this case is more demanding because the Hessian matrix of $J(w)$ is now $w-$dependent, but the same equivalence conclusion will continue to hold as we proceed to show.

5.1 Equivalence in the General Case

Note from the momentum recursion (29) that

$$\widehat{\boldsymbol{w}}_i = \left(I_M - \frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1} + \frac{\mu_m}{1-\beta}\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}), \tag{65}$$

where $\boldsymbol{H}_{i-1}$ is defined by (31). In the quadratic case, this matrix was constant and equal to the covariance matrix, R_u. Here, however, it is time-variant and depends on the error vector, $\widetilde{\boldsymbol{\psi}}_{i-1}$, as well. Likewise, for the standard stochastic gradient iteration (2), we obtain that the error recursion in the general case is given by:

$$\widetilde{\boldsymbol{x}}_i = (I_M - \mu\boldsymbol{R}_{i-1})\widetilde{\boldsymbol{x}}_{i-1} + \mu\boldsymbol{s}_i(\boldsymbol{x}_{i-1}), \tag{66}$$

where we are introducing the matrix

$$\boldsymbol{R}_{i-1} = \int_0^1 \nabla_w^2 J(w^o - r\widetilde{\boldsymbol{x}}_{i-1})dr \tag{67}$$

and $\widetilde{\boldsymbol{x}}_i = w^o - \boldsymbol{x}_i$. Note that $\boldsymbol{H}_{i-1}$ and $\boldsymbol{R}_{i-1}$ are different matrices. In contrast, in the quadratic case, they are both equal to R_u.

Under the assumed condition (60) relating $\{\mu, \mu_m\}$, if we subtract (66) from (65) we obtain:

$$\begin{aligned}\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i = &(I_M - \mu \boldsymbol{H}_{i-1})(\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}) + \mu(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1} \\ &+ \mu\beta' \boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1} + \mu[\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})]. \end{aligned} \tag{68}$$

In the quadratic case, the second term on the right-hand side is zero since $\boldsymbol{R}_{i-1} = \boldsymbol{H}_{i-1} = R_u$. It is the presence of this term that makes the analysis more demanding in the general case.

To examine how close $\widehat{\boldsymbol{w}}_i$ gets to $\widetilde{\boldsymbol{x}}_i$ for each iteration, we start by noting that

$$\begin{aligned}\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 = &\mathbb{E}\left\|(I_M - \mu \boldsymbol{H}_{i-1})(\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}) + \mu(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1} + \mu\beta'\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1}\right\|^2 \\ &+ \mu^2\mathbb{E}\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2. \end{aligned} \tag{69}$$

Now, applying a similar derivation to the one used to arrive at (137) in Appendix C, and the inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, we can conclude from (69) that

$$\begin{aligned}\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le &(1-\mu\nu)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + \frac{2\mu\beta'^2\delta^2}{\nu}\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \\ &+ \frac{2\mu}{\nu}\mathbb{E}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^2 + \mu^2\mathbb{E}\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2. \end{aligned} \tag{70}$$

Using the Cauchy-Schwartz inequality we can bound the cross term as

$$\mathbb{E}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^2 \le \mathbb{E}(\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^2\|\widetilde{\boldsymbol{x}}_{i-1}\|^2) \le \sqrt{\mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4\,\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4}. \tag{71}$$

In the above inequality, the term $\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4$ can be bounded by using the result of Lemma 2. Therefore, we focus on bounding $\mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4$ next. To do so, we need to introduce the following smoothness assumptions on the second and fourth-order moments of the gradient noise process and on the Hessian matrix of the risk function. These assumptions hold automatically for important cases of interest, such as least-mean-squares and logistic regression problems — see Appendix H for the verification.

Assumption 5 *Consider the iterates $\boldsymbol{\psi}_{i-1}$ and $\boldsymbol{x}_{i-1}$ that are generated by the momentum recursion* (22) *and the stochastic gradient recursion* (2)*. It is assumed that the gradient noise process satisfies:*

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \le \xi_1\|\boldsymbol{\psi}_{i-1} - \boldsymbol{x}_{i-1}\|^2, \tag{72}$$

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \le \xi_2\|\boldsymbol{\psi}_{i-1} - \boldsymbol{x}_{i-1}\|^4. \tag{73}$$

for some constants ξ_1 and ξ_2. ■

Assumption 6 *The Hessian of the risk function $J(w)$ in* (1) *is Lipschitz continuous, i.e., for any two variables $w_1, w_2 \in \mathrm{dom}\, J(w)$, it holds that*

$$\|\nabla_w^2 J(w_1) - \nabla_w^2 J(w_2)\| \le \kappa\|w_1 - w_2\|. \tag{74}$$

for some constant $\kappa \ge 0$. ■

Using these assumptions, we can now establish two auxiliary results in preparation for the main equivalence theorem in the general case.

Lemma 9 (Uniform bound) *Consider the standard and momentum stochastic gradient recursions* (2) *and* (22)-(23) *and assume they start from the same initial states, namely,* $\boldsymbol{\psi}_{-2} = \boldsymbol{w}_{-2} = \boldsymbol{x}_{-1}$. *We continue to assume conditions (26), and (60). Under Assumptions 1, 3, 4, 5 and for sufficiently small step-sizes* μ*, the following result holds:*

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{x}}_i\|^4 = O\left(\frac{\delta^4 (i+1)\rho_2^{i+1}\mu}{\nu^3} + \frac{\delta^4\sigma_s^4}{\nu^6}\mu^2\right), \tag{75}$$

where $\rho_2 = 1 - \mu\nu/4$.

Proof See Appendix I. ■

Although sufficient for our purposes, we remark that the bound (75) for $\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{x}}_i\|^4$ is not tight. The reason is that in the derivation in Appendix I we employed a looser bound for the term $\mathbb{E}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^4$ in order to avoid the appearance of higher-order powers, such as $\mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^8$ and $\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^8$. To avoid this possibility, we employed the following bound (using (11) to bound $\|\boldsymbol{R}_{i-1}\|^4$ and $\|\boldsymbol{H}_{i-1}\|^4$ and the inequality $\|a+b\|^4 \le 8\|a\|^4 + 8\|b\|^4$):

$$\begin{aligned}\mathbb{E}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^4 &\le \mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4\|\widetilde{\boldsymbol{x}}_{i-1}\|^4 \\ &\le 8\mathbb{E}\big\{(\|\boldsymbol{R}_{i-1}\|^4 + \|\boldsymbol{H}_{i-1}\|^4)\|\widetilde{\boldsymbol{x}}_{i-1}\|^4\big\} \le 16\delta^4\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4. \end{aligned} \tag{76}$$

Based on Lemma 9, we can now bound $\mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4$, which is what the following lemma states.

Lemma 10 (Bound on Hessian difference) *Consider the same setting of Lemma 9. Under Assumptions 1, 3, 4, 6 and for sufficiently small step-sizes* μ*, the following two result holds:*

$$\mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4 = O\left(\frac{\delta^4 i\rho_2^{i}\mu}{\nu^3} + \frac{\delta^4\sigma_s^4}{\nu^6}\mu^2\right), \tag{77}$$

where $\rho_2 = 1 - \mu\nu/4$.

Proof See Appendix J. ■

With the upper bounds of $\mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4$ and $\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4$ established in Lemma 10 and Lemma 2 respectively, we are able to bound $\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|$ in (71), which in turn helps to establish the main equivalence result.

Theorem 11 (Equivalence for general risks) *Consider the standard and momentum stochastic gradient recursions* (2) *and* (22)–(23) *and assume they start from the same initial states, namely,* $\boldsymbol{\psi}_{-2} = \boldsymbol{w}_{-2} = \boldsymbol{x}_{-1}$. *Suppose conditions (26) and (60) hold. Under*

Assumptions 1, 3, 4, 5, and 6, and for sufficiently small step-size μ, it holds that

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 = O\left(\frac{\delta^2\sigma_s^2 i^2 \tau_2^{i+1}\mu^{3/2}}{(1-\beta)\nu^{5/2}} + \left[\frac{(\delta^2+\gamma^2)\rho_1^{i+1}}{(1-\beta)^2} + \frac{\delta^2\sigma_s^4}{(1-\beta)\nu^2}\right]\mu^2\right), \quad \forall i = 0,1,2,3,\ldots \tag{78}$$

where $\rho_1 = 1 - \frac{\mu\nu}{2} \in (0,1)$ *and* $\tau_2 \overset{\Delta}{=} \sqrt{1-\mu\nu/4} \in (0,1)$.

Proof See Appendix K. ■

Similar to Corollary 5, 7 and Theorem 8, Theorem 11 implies that

$$\mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2 = O(\mu^{3/2}), \quad \forall i = 0,1,2,\cdots \tag{79}$$

Besides, since $\rho_1^i \to 0$ and $i^2\tau_2^i \to 0$ as $i \to \infty$, we will also conclude

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 = O\left(\frac{\delta^2\sigma_s^4\mu^2}{(1-\beta)\nu^2}\right) = O(\mu^2). \tag{80}$$

Remark When we refer to "sufficiently small step-sizes" in Theorems 8 and 11, we mean that step-sizes are smaller than the stability bound, and are also small enough to ensure a desirable level of mean-square-error based on the performance expressions.

5.2 Interpretation of Equivalence Result

The result of Theorem 11 shows that, for sufficiently small step-sizes, the trajectories of momentum and standard stochastic gradient methods remain within $O(\mu^{3/2})$ from each other for *every* i (for quadratic cases the trajectories will remain within $O(\mu^2)$ as stated in Theorem 8). This means that these trajectories evolve together for all practical purposes and, hence, we shall say that the two implementations are "equivalent" (meaning that their trajectories remain close to each other in the mean-square-error sense).

A second useful insight from Theorem 8 is that the momentum method is essentially equivalent to running a standard stochastic gradient method with a larger step-size (since $\mu > \mu_m$). This interpretation explains why the momentum method is observed to converge faster during the transient phase albeit towards a worse MSD level in steady-state than the standard method. This is because, as is well-known in the adaptive filtering literature (Sayed, 2008, 2014a) that larger step-sizes for stochastic gradient method do indeed lead to faster convergence but worse limiting performance.

In addition, Theorem 11 enables us to compute the steady-state MSD performance of the momentum stochastic gradient method. It is guaranteed by Theorem 11 that momentum method is equivalent to standard stochastic gradient method with larger step-size, $\mu = \mu_m/(1-\beta)$. Therefore, once we compute the MSD performance of the standard stochastic gradient, according to (Haykin, 2008; Sayed, 2008, 2014a), we will also know the MSD performance for the momentum method.

Another consequence of the equivalence result is that any benefits that would be expected from a momentum stochastic gradient descent can be attained by simply using a standard stochastic gradient implementation with a larger step-size; this is achieved without the additional computational or memory burden that the momentum method entails.

Besides the theoretical analysis given above, there is an intuitive explanation as to why the momentum variant leads to worse steady-state performance. While the momentum terms $\boldsymbol{w}_i - \boldsymbol{w}_{i-1}$ and $\boldsymbol{\psi}_i - \boldsymbol{\psi}_{i-1}$ can smooth the convergence trajectories, and hence accelerate the convergence rate, they nevertheless introduce additional noise into the evolution of the algorithm because all iterates $\boldsymbol{w}_i$ and $\boldsymbol{\psi}_i$ are distorted by perturbations. This fact illustrates the essential difference between stochastic methods with constant step-sizes, and stochastic or deterministic methods with decaying step-sizes: in the former case, the presence of gradient noise essentially eliminates the benefits of the momentum term.

5.3 Stochastic Gradient Method with Diminishing Momentum

(Tygert, 2016; Yuan et al., 2016) suggest one useful technique to retain the advantages of the momentum implementation by employing a *diminishing* momentum parameter, $\beta(i)$, and by ensuring $\beta(i) \to 0$ in order not to degrade the limiting performance of the implementation. By doing so, the momentum term will help accelerate the convergence rate during the transient phase because it will smooth the trajectory (Nedić and Bertsekas, 2001; Xiao, 2010; Lan, 2012). On the other hand, momentum will not cause degradation in MSD performance because the momentum effect would have died before the algorithm reaches state-state.

According to (Tygert, 2016; Yuan et al., 2016), we adapt the momentum stochastic method into the following algorithm

$$\boldsymbol{\psi}_{i-1} = \boldsymbol{w}_{i-1} + \beta_1(i)(\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}), \tag{81}$$

$$\boldsymbol{w}_i = \boldsymbol{\psi}_{i-1} - \mu \nabla_w Q(\boldsymbol{\psi}_{i-1}; \boldsymbol{\theta}_i) + \beta_2(i)(\boldsymbol{\psi}_{i-1} - \boldsymbol{\psi}_{i-2}), \tag{82}$$

with the same initial conditions as in (24)–(25). Similar to condition (26), $\beta_1(i)$ and $\beta_2(i)$ also need to satisfy

$$\beta_1(i) + \beta_2(i) = \beta(i), \quad \beta_1(i)\beta_2(i) = 0, \tag{83}$$

The efficacy of (81)–(82) will depend on how the momentum decay, $\beta(i)$, is selected. A satisfactory sequence $\{\beta(i)\}$ should decay slowly during the initial stages of adaptation so that the momentum term can induce an acceleration effect. However, the sequence $\{\beta(i)\}$ should also decrease drastically prior to steady-state so that the vanishing momentum term will not introduce additional gradient noise and degrade performance. One strategy, which is also employed in the numerical experiments in Section 7, is to design $\beta(i)$ to decrease in a stair-wise fashion, namely,

$$\beta(i) = \begin{cases} \beta_0 & \text{if } i \in [1, T], \\ \beta_0 / T^\alpha & \text{if } i \in [T+1, 2T], \\ \beta_0 / (2T)^\alpha & \text{if } i \in [2T+1, 3T], \\ \beta_0 / (3T)^\alpha & \text{if } i \in [3T+1, 4T], \\ \cdots \end{cases} \tag{84}$$

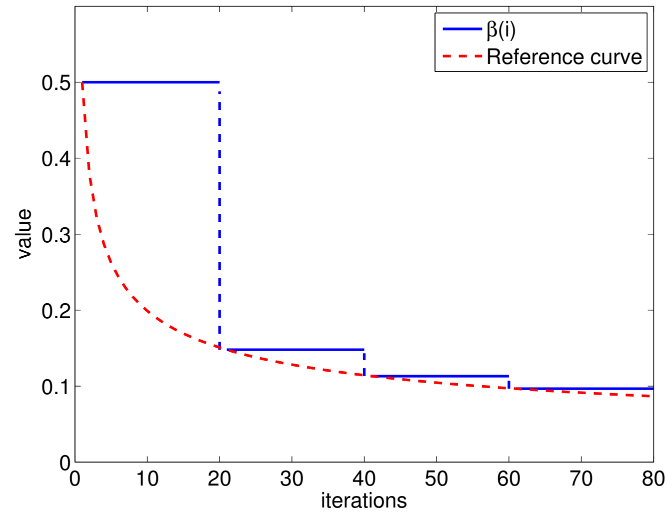

Figure 1: $\beta(i)$ changes with iteration i according to (84), where $\beta_0 = 0.5$, $T = 20$ and $\alpha = 0.4$. The reference curve is $f(i) = 0.5/i^{0.4}$.

where the constants $\beta_0 \in [0, 1)$, $\alpha \in (0, 1)$ and $T > 0$ determines the width of the stair steps. Fig. 1 illustrates how $\beta(i)$ varies when $T = 20$, $\beta_0 = 0.5$ and $\alpha = 0.4$.

Algorithm (81)–(82) works well when $\beta(i)$ decreases according to (84) (see Section 7). However, with Theorems 8 and 11, we find that this algorithm is essentially equivalent to the standard stochastic gradient method with decaying step-size, i.e.,

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - \mu_s(i)\nabla_w Q(\boldsymbol{x}_{i-1}; \theta_i), \tag{85}$$

where

$$\mu_s(i) = \frac{\mu}{1 - \beta(i)} \tag{86}$$

will decrease from $\mu/[1 - \beta(0)]$ to μ. In another words, the stochastic algorithm with decaying momentum is still not helpful.

6. Diagonal Step-size Matrices

Sometimes it is advantageous to employ separate step-size for the individual entries of the weight vectors, see (Duchi et al., 2011). In this section we comment on how the results from the previous sections extend to this scenario. First, we note that recursion (2) can be generalized to the following form, with a diagonal matrix serving as the step-size parameter:

$$\boldsymbol{x}_i = \boldsymbol{x}_{i-1} - D\nabla_w Q(\boldsymbol{x}_{i-1}; \boldsymbol{\theta}_i),\ i \ge 0, \tag{87}$$

where $D = \mathrm{diag}\{\mu_1, \mu_2, \ldots, \mu_M\}$. Here, we continue to use the letter "$\boldsymbol{x}$" to refer to the variable iterates for the standard stochastic gradient descent iteration, while we reserve the letter "$\boldsymbol{w}$" for the momentum recursion. We let $\mu_{\max} = \max\{\mu_1, \ldots, \mu_M\}$. Similarly, recursions (22) and (23) can be extended in the following manner:

$$\boldsymbol{\psi}_{i-1} = \boldsymbol{w}_{i-1} + B_1(\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}), \tag{88}$$

$$\boldsymbol{w}_i = \boldsymbol{\psi}_{i-1} - D_m \nabla_w Q(\boldsymbol{\psi}_{i-1}; \boldsymbol{\theta}_i) + B_2(\boldsymbol{\psi}_{i-1} - \boldsymbol{\psi}_{i-2}), \tag{89}$$

with initial conditions

$$\boldsymbol{w}_{-2} = \boldsymbol{\psi}_{-2} = \text{initial states}, \tag{90}$$

$$\boldsymbol{w}_{-1} = \boldsymbol{w}_{-2} - D_m \nabla_w Q(\boldsymbol{w}_{-2}; \boldsymbol{\theta}_{-1}), \tag{91}$$

where $B_1 = \mathrm{diag}\{\beta_1^1, \ldots, \beta_M^1\}$ and $B_2 = \mathrm{diag}\{\beta_1^2, \ldots, \beta_M^2\}$ are momentum coefficient matrices, while D_m is a diagonal step-size matrix for momentum stochastic gradient method. In a manner similar to (26), we also assume that

$$0 \le B_k < I_M, \; k = 1, 2, \qquad B_1 + B_2 = B, \qquad B_1 B_2 = 0. \tag{92}$$

where $B = \mathrm{diag}\{\beta_1, \ldots, \beta_M\}$ and $0 < B < I_M$. In addition, we further assume that B is not too close to I_M, i.e.

$$B \le (1 - \epsilon) I_M, \text{ for some constant } \epsilon > 0. \tag{93}$$

The following results extend Theorems 1, 3, and 4 and they can be established following similar derivations.

Theorem 1B (Mean-square stability). *Let Assumptions 1 and 2 hold and recall conditions (92) and* (93). *Then, for the momentum stochastic gradient method* (88)–(89)*, it holds under sufficiently small step-size* $\mu_{\max}$ *that*

$$\limsup_{i \to \infty} \mathbb{E} \|\widetilde{\boldsymbol{w}}_i\|^2 = O(\mu_{\max}). \tag{94}$$

■

Theorem 3B (Equivalence for quadratic costs). *Consider recursions* (52) *and* (56)–(57) *with* $\{\mu, \mu_m, \beta_1, \beta_2\}$ *replaced by* $\{D, D_m, B_1, B_2\}$. *Assume they start from the same initial states, namely,* $\boldsymbol{\psi}_{-2} = \boldsymbol{w}_{-2} = \boldsymbol{x}_{-1}$. *Suppose further that conditions* (92) *and* (93) *hold, and that the step-sizes matrices* $\{D, D_m\}$ *satisfy a relation similar to* (60)*, namely,*

$$D = (I - B)^{-1} D_m. \tag{95}$$

Then, it holds under sufficiently small $\mu_{\max}$*, that*

$$\mathbb{E} \|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 = O(\mu_{\max}^2), \quad \forall i = 0, 1, 2, 3, \ldots \tag{96}$$

■

Theorem 4B (Equivalence for general costs). *Consider the stochastic gradient recursion* (87) *and the momentum stochastic gradient recursions* (88)–(89) *to solve the general problem* (1). *Assume they start from the same initial states, namely,* $\boldsymbol{\psi}_{-2} = \boldsymbol{w}_{-2} = \boldsymbol{x}_{-1}$.

Suppose conditions (92), (93), and (95) hold. Under Assumptions 1, 3, 5, and 6, and for sufficiently small step-sizes, it holds that

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 = O(\mu_{\max}^{3/2}), \quad \forall i = 0, 1, 2, 3, \ldots \tag{97}$$

Furthermore, in the limit,

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 = O(\mu_{\max}^2). \tag{98}$$

■

7. Experimental Results

In this section we illustrate the main conclusions by means of computer simulations for both cases of mean-square-error designs and logistic regression designs. We also run simulations for algorithm (81)–(82) and verify its advantages in the stochastic context.

7.1 Least Mean-Squares Error Designs

We apply the standard LMS algorithm to (47). To do so, we generate data according to the linear regression model (48), where $w^o \in \mathbb{R}^{10}$ is chosen randomly, and $\boldsymbol{u}_i \in \mathbb{R}^{10}$ is i.i.d and follows $\boldsymbol{u}_i \sim \mathcal{N}(0, \Lambda)$ where $\Lambda \in \mathbb{R}^{10\times 10}$ is randomly-generated diagonal matrix with positive diagonal entries. Besides, $\boldsymbol{v}(i)$ is also i.i.d and follows $\boldsymbol{v}(i) \sim \mathcal{N}(0, \sigma_s^2 I_{10})$, where $\sigma_s^2 = 0.01$. All results are averaged over 300 random trials. For each trial we generated 800 samples of $\boldsymbol{u}_i$, $\boldsymbol{v}(i)$ and $\boldsymbol{d}(i)$.

We first compare the standard and momentum LMS algorithms using $\mu = \mu_m = 0.003$. The momentum parameter β is set as 0.9. Furthermore, we employ the heavy-ball option for the momentum LMS, i.e., $\beta_1 = 0, \beta_2 = \beta$. Both the standard and momentum LMS methods are illustrated in the left plot in Fig. 2 with blue and red curves, respectively. It is seen that the momentum LMS converges faster, but the MSD performance is much worse. Next we set $\mu_m = \mu(1-\beta) = 0.0003$ and illustrate this case with the magenta curve. It is observed that the magenta and blue curves are almost indistinguishable, which confirms the equivalence predicted by Theorem 8 for all time instants. We also illustrate an implementation with a decaying momentum parameter $\beta(i)$ by the green curve. In this simulation, we set $\mu_m = 0.003$ and make $\beta(i)$ decrease in a stair-wise fashion: when $i \in [1, 100]$, $\beta(i) = 0.9$; when $i \in [101, 200]$, $\beta(i) = 0.9/(100^{0.3})$; $\ldots$; when $i \in [2401, 2500]$, $\beta(i) = 0.9/(2400^{0.3})$. With this decaying $\beta(i)$, it is seen that the momentum LMS method recovers its faster convergence rate and attains the same steady-state MSD performance as the LMS implementation. Finally, we also implemented the standard LMS with initial step-size $\mu = 0.003$ and then decrease it gradually according to $\mu_s(i) = \mu/[1-\beta(i)]$. As implied by Theorem 8, it is observed that the green and black curves are also almost indistinguishable, which confirms that the LMS algorithm with decaying momentum is still equivalent to the standard LMS with appropriately chosen decaying step-sizes. We also compared the standard and momentum LMS algorithms when $\mu = \mu_m = 0.003$ and β is set as $0.5, 0.6, 0.7, 0.8$, and the same performance as the left plot in Fig. 2 is observed. To save space, we show the right plot in Fig. 2 in which $\beta = 0.5$ and omit the figures when β is set as $0.6, 0.7, 0.8$.

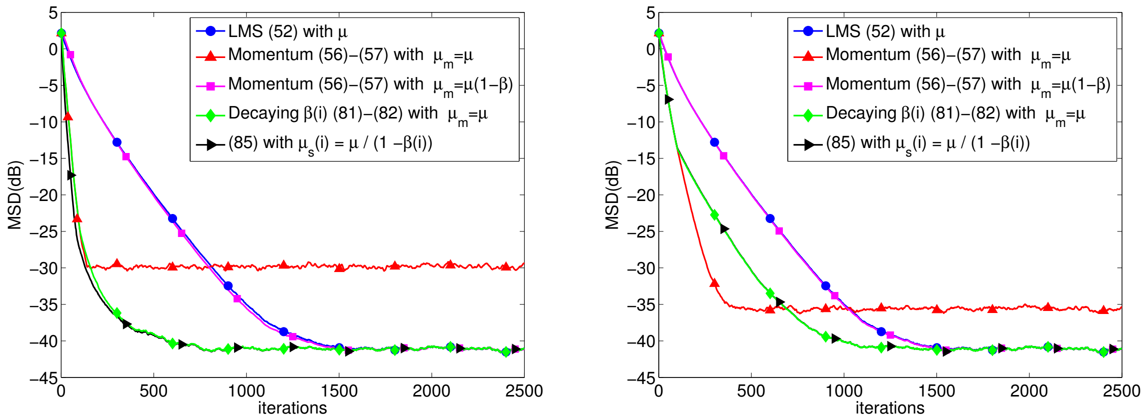

Figure 2: Convergence behavior of standard and momentum LMS (heavy-ball LMS) algorithms applied to the mean-square-error design problem (47) with $\beta = 0.9$ in the left plot and $\beta = 0.5$ in the right plot. Mean-square-deviation (MSD) means $\mathbb{E}\|w^o - \boldsymbol{w}_i\|^2$.

Next we employ the Nesterov's acceleration option for the momentum LMS method, and compare it with standard LMS. The experimental settings are exactly the same as the above except that $\beta_1 = \beta$ and $\beta_2 = 0$. Both the standard and momentum LMS methods are illustrated in Fig. 3. As implied by Theorem 8, it is observed that Nesterov's acceleration applied to LMS is equivalent to standard LMS with rescaled step-size. Besides, by comparing Figs. 2 and 3, it is also observed that both momentum options, the heavy-ball and the Nesterov's acceleration, have the same performance. To save space, in the following experiments in Section 7.2–7.4 we just show the performance of momentum method with the option of heavy-ball.

7.2 Regularized Logistic Regression

We next consider a regularized logistic regression risk of the form:

$$J(w) \;\stackrel{\Delta}{=}\; \frac{\rho}{2}\|w\|^2 + \mathbb{E}\Big\{\ln\left[1+\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}} w)\right]\Big\} \tag{99}$$

where the approximate gradient vector is chosen as

$$\nabla_w Q(\boldsymbol{w};\boldsymbol{h}_i,\boldsymbol{\gamma}(i)) = \rho\boldsymbol{w} - \frac{\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w})}{1+\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}\boldsymbol{w})}\boldsymbol{\gamma}(i)\boldsymbol{h}_i \tag{100}$$

In the simulation, we generate 20000 samples $(\boldsymbol{h}_i, \boldsymbol{\gamma}(i))$. Among these training points, 10000 feature vectors $\boldsymbol{h}_i$ correspond to label $\boldsymbol{\gamma}(i) = 1$ and each $\boldsymbol{h}_i \sim \mathcal{N}(1.5 \times \mathbb{1}_{10}, R_h)$ for some diagonal covariance R_h. The remaining 10000 feature vectors $\boldsymbol{h}_i$ correspond to label $\boldsymbol{\gamma}(i) = -1$ and each $\boldsymbol{h}_i \sim \mathcal{N}(-1.5 \times \mathbb{1}_{10}, R_h)$. We set $\rho = 0.1$. The optimal solution w^o is computed via the classic gradient descent method. All simulation results shown below are averaged over 300 trials.

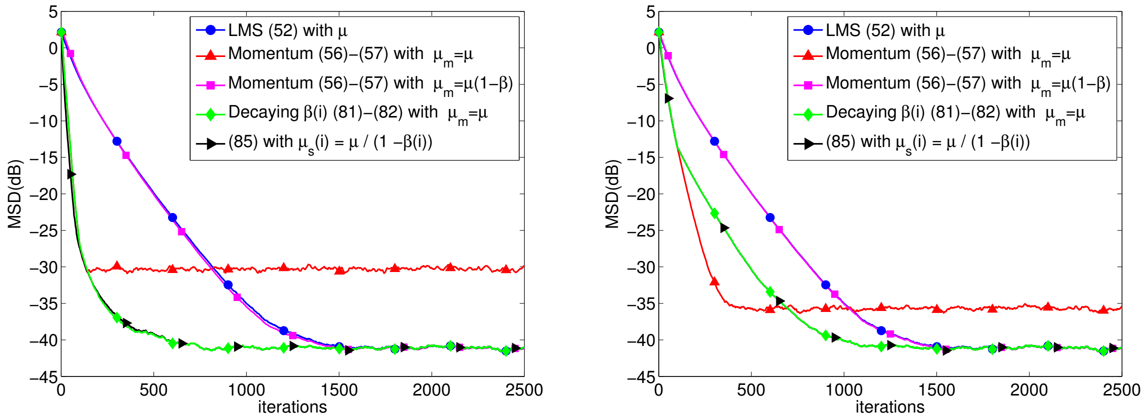

Figure 3: Convergence behavior of standard and momentum LMS (Nesterov's acceleration LMS) algorithms applied to the mean-square-error design problem (47) with $\beta = 0.9$ in the left plot and $\beta = 0.5$ in the right plot.

Similar to the least-mean-squares error problem, we first compare the standard and momentum stochastic methods using $\mu = \mu_m = 0.005$. The momentum parameter β is set to 0.9. These two methods are illustrated in Fig. 4 with blue and red curves, respectively. It is seen that the momentum method converges faster, but the MSD performance is much worse. Next we set $\mu_m = \mu(1-\beta) = 0.0005$ and illustrate this case with the magenta curve. It is observed that the magenta and blue curves are indistinguishable, which confirms the equivalence predicted by Theorem 11 for all time instants. Again we illustrate an implementation with a decaying momentum parameter $\beta(i)$ by the green curve. In this simulation, we set $\mu_m = 0.005$ and make $\beta(i)$ decrease in a stair-wise manner: when $i \in [1, 200]$, $\beta(i) = 0.9$; when $i \in [201, 400]$, $\beta(i) = 0.9/(200^{0.3})$; when $i \in [401, 600]$, $\beta(i) = 0.9/(400^{0.3})$; …; when $i \in [1801, 2000]$, $\beta(i) = 0.9/(1800^{0.3})$. With this decaying $\beta(i)$, it is seen that the momentum method recovers its faster convergence rate and attains the same steady-state MSD performance as the stochastic-gradient implementation. Finally, we implemented the standard stochastic gradient descent with initial step-size $\mu_m = \mu = 0.005$ and then decrease it gradually according to $\mu_s(i) = \mu/[1-\beta(i)]$. As implied by Theorem 11, it is observed that the green and black curves are almost indistinguishable, which confirms that the algorithm with decaying momentum is still equivalent to the standard stochastic gradient descent with appropriately chosen decaying step-sizes.

Next, we test the standard and momentum stochastic methods for regularized logistic regression problem over a benchmark data set — the Adult Data Set[2]. The aim of this dataset is to predict whether a person earns over \$50K a year based on census data such as age, workclass, education, race, etc. The set is divided into 6414 training data and 26147 test data, and each feature vector has 123 entries. In the simulation, we set $\mu = 0.1$, $\rho = 0.1$, and $\beta = 0.9$. To check the equivalence of the algorithms, we set $\mu_m = (1-\beta)\mu = 0.01$. In Fig. 5,

2. Source: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/` or `http://archive.ics.uci.edu/ml/datasets/Adult`

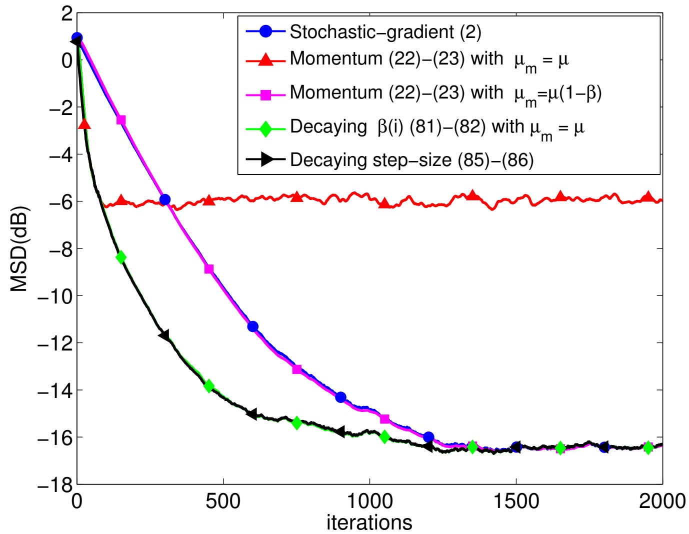

Figure 4: Convergence behaviors of standard and momentum stochastic gradient methods applied to the logistic regression problem (99).

the curve shows how the accuracy performance, i.e., the percentage of correct prediction, over the test dataset evolved as the algorithm received more training data[3]. The horizontal x-axis indicates the number of training data used. It is observed that the momentum and standard stochastic gradient methods cannot be distinguished, which confirms their equivalence when training the Adult Data Set.

For the experiments shown in this section, Section 7.3 and 7.4, we also tested the cases when β is set as 0.5, 0.6, 0.7, 0.8. Since the experimental results with different β are similar, we just plot the situation when $\beta = 0.9$, a setting which is usually employed in practice (Szegedy et al., 2015; Krizhevsky et al., 2012; Zhang and LeCun, 2015).

7.3 Further Verification of Theorems 8 and 11

In this section we further illustrate the conclusions of Theorems 8 and 11 by checking the behavior of the iterate difference, i.e., $\mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2$, between the standard and momentum stochastic gradient methods.

For the least-mean-squares error problem, the selection of $\boldsymbol{u}_i$, $\boldsymbol{v}(i)$, $\boldsymbol{d}(i)$ and β is the same as in the simulation generated earlier in Subsection 7.1. For some specific step-size μ, $\boldsymbol{x}_i$ is the iterate generated through LMS recursion (52) with step-size μ, and $\boldsymbol{w}_i$ is the iterate generated momentum LMS recursion (56)–(57) with step-size $\mu_m = \mu(1-\beta)$. Now we introduce the maximum difference:

$$d_{\max}(\mu) = \max_i \mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2 \tag{101}$$

3. To smooth the performance curve, we applied the weighted average technique from equation (74) of (Ying and Sayed, 2015, 2016).

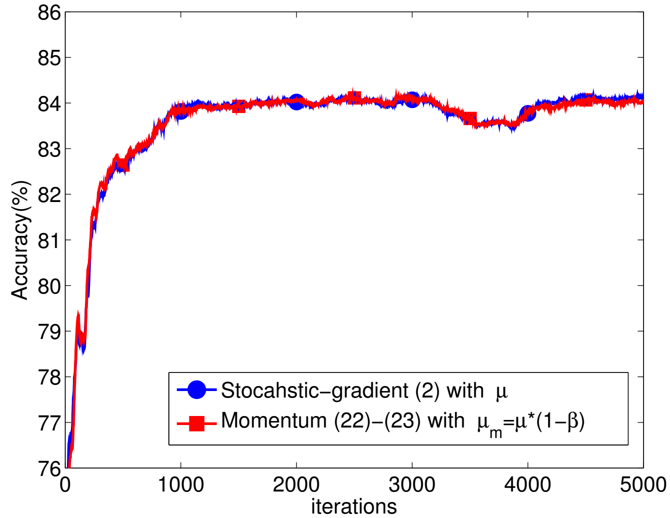

Figure 5: Performance accuracy of the standard and momentum stochastic gradient methods applied to logistic regression classification on the adult data test set.

and the difference at steady state

$$d_{\rm ss}(\mu) = \limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2. \tag{102}$$

Note that both $d_{\max}(\mu)$ and $d_{\rm ss}(\mu)$ are related with μ and we will examine how they vary according to different step-sizes. Obviously, since $\mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2 \le d_{\max}(\mu)$, if $d_{\max}(\mu)$ is illustrated to be on the order of $O(\mu^2)$, then it follows that $\mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2 = O(\mu^2)$ for $i \ge 0$. Similarly, if we can illustrate $d_{\rm ss}(\mu) = O(\mu^2)$, then it follows that $\limsup_{i\to\infty} \mathbb{E}\|\boldsymbol{w}_i - \boldsymbol{x}_i\|^2 = O(\mu^2)$.

Note that the fact $d_{\max}(\mu) = c\mu^2$ for some constant c holds if and only if

$$d_{\max}(\mu)(\text{dB}) = 20 \log \mu + 10 \log c, \tag{103}$$

where $d_{\max}(\mu)(\text{dB}) = 10 \log d_{\max}(\mu)$. Relation (103) can be confirmed with red circle line in Fig. 6. In this simulation, we choose 8 different step-size values $\{\mu_k\}_{k=1}^8$, and it can be verified that each data pair $\Big(\log \mu_k,\ d_{\max}(\mu_k)(\text{dB})\Big)$ satisfies relation (103). For example, in the red circle solid line, at $\mu_1 = 10^{-2}$ we read $d_{\max}(\mu_1)(\text{dB}) = -32\text{dB}$; while at $\mu_2 = 10^{-4}$ we read $d_{\max}(\mu_2)(\text{dB}) = -72\text{dB}$. It can be verified that

$$d_{\max}(\mu_1)(\text{dB}) - d_{\max}(\mu_2)(\text{dB}) = 20(\log \mu_1 - \log \mu_2) = 40. \tag{104}$$

Using a similar argument, the blue square solid line can also implies that $d_{\rm ss} = O(\mu^2)$.

Figure 6 also reveals the order of $d_{\max}$ and $d_{\rm ss}$, with magenta and green dash lines respectively, for the regularized logistic regression problem from Subsection 7.2. With the same argument as above, $d_{\rm ss}(\mu)$ can be confirmed on the order of $O(\mu^2)$. Now we check the order of $d_{\max}(\mu)$. The fact that $d_{\max}(\mu) = c\mu^{3/2}$ holds if and only if

$$d_{\max}(\mu)(\text{dB}) = 15 \log \mu + 10 \log c. \tag{105}$$

According to the above relation, at $\mu_1 = 10^{-2}$ and $\mu_2 = 10^{-4}$ we should have

$$d_{\max}(\mu_1)(\text{dB}) - d_{\max}(\mu_2)(\text{dB}) = 15(\log \mu_1 - \log \mu_2) = 30. \tag{106}$$

However, in the triangle magenta dash line we read $d_{\max}(\mu_1) = -30\text{dB}$ while $d_{\max}(\mu_2) = -66\text{dB}$ and hence

$$30\text{dB} < d_{\max}(\mu_1)(\text{dB}) - d_{\max}(\mu_2)(\text{dB}) < 40\text{dB}$$

Therefore, the order of $d_{\max}$ should be between $O(\mu^{3/2})$ and $O(\mu^2)$, which still confirms Theorem 11.

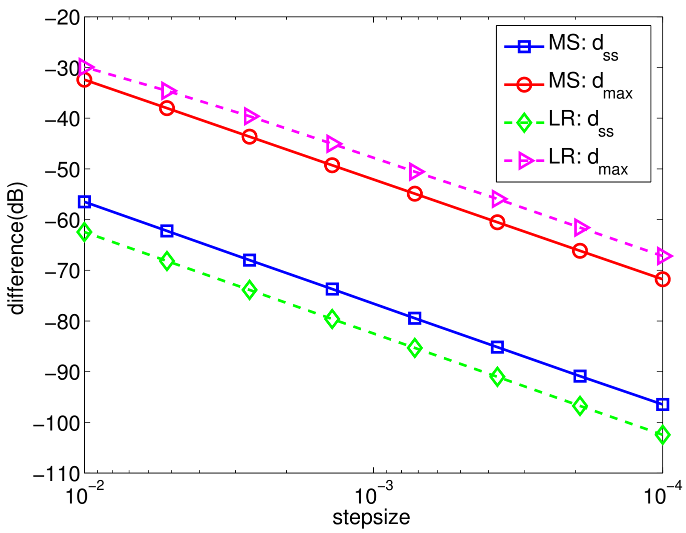

Figure 6: $d_{\max}$ and d_{ss} as a function of the step-size μ. *MS* stands for mean-square-error and *LR* stands for logistic regression.

7.4 Visual Recognition

In this subsection we illustrate the conclusions of this work by re-examining the problem of training a neural network to recognize objects from images. We employ the CIFAR-10 database[4], which is a classical benchmark dataset of images for visual recognition. The CIFAR-10 dataset consists of 60000 color images in 10 classes, each with 32×32 pixels. There are 50000 training images and 10000 test images. Similar to (Sutskever et al., 2013), and since the focus of this paper is on optimization, we only report training errors in our experiment.

To help illustrate that the conclusions also hold for non-differentiable and non-convex problems, in this experiment we train the data with two different neural network structures: (a) a 6-layer fully connected neural network and (b) a 4-layer convolutional neural network, both with ReLU activation functions. For each neural network, we will compare the performance of the momentum and standard stochastic gradient methods.

4. `https://www.cs.toronto.edu/~kriz/cifar.html`

6-Layer Fully Connected Nuerual Network. For this neural network structure, we employ the softmax measure with ℓ_2 regularization as a cost objective, and the ReLU as an activation function. Each hidden layer has 100 units, the coefficient of the ℓ_2 regularization term is set to 0.001, and the initial value w_{-1} is generated by a Gaussian distribution with 0.05 standard deviation. We employ mini-batch stochastic-gradient learning with batch size equal to 100. First, we apply a momentum backpropagation (i.e., momentum stochastic gradient) algorithm to train the 6-layer neural network. The momentum parameter is set to $\beta = 0.9$, and the initial step-size μ_m is set to 0.01. To achieve better accuracy, we follow a common technique (e.g., (Szegedy et al., 2015)) and reduce μ_m to $0.95\mu_m$ after every epoch. With the above settings, we attain an accuracy of about 90% in 80 epochs.

However, what is interesting, and somewhat surprising, is that the same 90% accuracy can also be achieved with the standard backpropagation (i.e., stochastic gradient descent) algorithm in 80 epochs. According to the step-size relation $\mu = \mu_m/(1-\beta)$, we set the initial step-size μ of SGD to 0.1. Similar to the momentum method, we also reduce μ to 0.95μ after every epoch for SGD, and hence the relation $\mu = \mu_m/(1-\beta)$ still holds for each iteration. From Figure 7, we observe that the accuracy performance curves for both scenarios, with and without momentum, are overlapping even when the overall risk is not necessarily convex or differentiable.

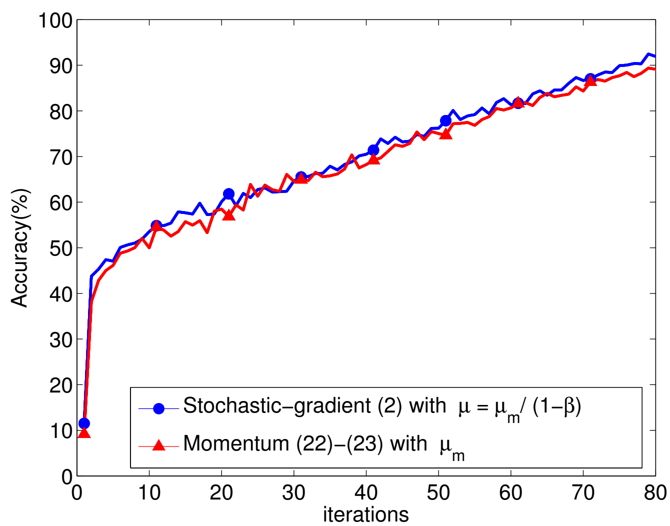

Figure 7: Classification accuracy of the standard and momentum stochastic gradient methods applied to a 6-layer fully-connected neural network on the CIFAR-10 test data set.

4-Layer Convolutional Neural Network. In a second experiment, we consider a 4-layer convolutional neural network. We employ the same objective and activation functions. This network has the structure:

$$\big(\text{conv} - \text{ReLU} - \text{pool}\big) \times 2 - \big(\text{ affine} - \text{ReLU }\big) - \text{affine}$$

In the first convolutional layer, we use filters of size $7 \times 7 \times 3$, stride value 1, zero padding 3, and the number of these filters is 32. In the second convolutional layer, we use filters of size

$7 \times 7 \times 32$, stride value 1, zero padding 3, and the number of filters is still 32. We implement MAX operation in all pooling layers, and the pooling filters are of size 2×2, stride value 2 and zero padding 0. The hidden layer has 500 units. The coefficient of the ℓ_2 regularization term is set to 0.001, and the initial value w_{-1} is generated by a Gaussian distribution with 0.001 standard deviation. We employ mini-batch stochastic-gradient learning with batch size equal to 50, and the step-size decreases by 5% after each epoch.

First, we apply the momentum backpropagation algorithm to train the neural network. The momentum parameter is set at $\beta = 0.9$, and we performed experiments with step-sizes $\mu_m \in \{0.01, 0.005, 0.001, 0.0005, 0.0001\}$ and find that $\mu_m = 0.001$ gives the highest training accuracy after 10 epochs. In Fig. 8 we draw the momentum stochastic gradient method with red curve when $\mu_m = 0.001$ and $\beta = 0.9$. The curve reaches an accuracy of 94%. Next we set the step-size of the standard backpropagation $\mu = \mu_m/(1-\beta) = 0.01$, and illustrate its convergence performance with the blue curve. It is also observed that the two curves are indistinguishable. The numerical results shown in Figs. 7 and 8 imply that the performance of momentum SGD can still be achieved by standard SGD by properly adjusting the step-size according to $\mu = \mu_m/(1-\beta)$.

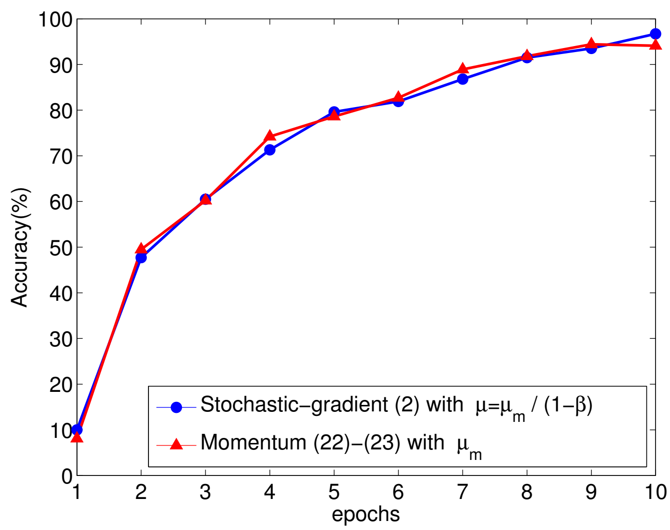

Figure 8: Classification accuracy of the standard and momentum stochastic gradient methods applied to a 4-layer convolutional neural network on the CIFAR-10 training data set.

8. Comparison for Larger Step-sizes

According to Theorem 11, the equivalence results between the standard and momentum stochastic gradient methods hold for sufficiently small step-sizes μ. When larger values for μ are used, the $O(\mu^{3/2})$ term is not negligible any longer so that the momentum and gradient-descent implementations are not equivalent anymore under these conditions. While in practical implementations small step-sizes are widely employed in order to ensure satisfactory steady-state MSD performance, one may still wonder how both algorithms would compare to each other under larger step-sizes. For example, it is known that the larger the step-size value is, the more likely it is that the stochastic-gradient algorithm will become unstable. Does the addition of momentum help enlarge the stability range and allow for proper adaptation and learning over a wider range of step-sizes?

Unfortunately, the answer to the above question is generally negative. In fact, we can construct a simple numerical example in which the momentum can hurt the stability range. This example considers the case of quadratic risks, namely problems of the form (47). We suppose $M = 5$, $\boldsymbol{u}_i \sim \mathcal{N}(0, 0.5I_5)$ and $\boldsymbol{d}(i) = \boldsymbol{u}_i^{\mathsf{T}} w^o + \boldsymbol{v}(i)$ where $\boldsymbol{v}(i) \sim \mathcal{N}(0, 0.01)$. We compare the convergence of standard LMS and Nesterov's acceleration method with fixed parameter $\beta_2 = 0$ and $\beta = 0.5$. Both algorithms are set with the same step-size $\mu = \mu_m = 0.4$, which is a relatively large step-size. All results are averaged over 1000 random trials. For each trial we generated 200 samples of $\boldsymbol{u}_i$, $\boldsymbol{v}(i)$ and $\boldsymbol{d}(i)$. In Fig. 9, it shows that standard LMS converges at $\mu = 0.4$ while momentum LMS diverges, which indicates that momentum LMS has narrower stability range than standard LMS.

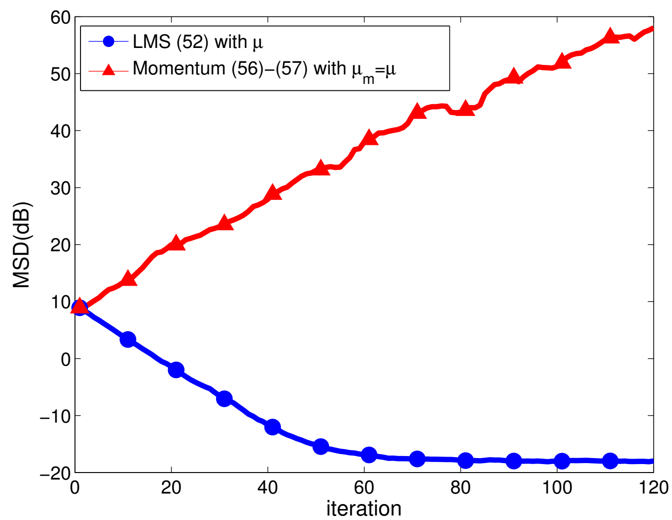

Figure 9: Convergence comparison between standard and momentum LMS algorithms when $\mu = \mu_m = 0.4$ and $\beta = 0.5$.

9. Conclusion

In this paper we analyzed the convergence and performance behavior of momentum stochastic gradient methods in the constant step-size and slow adaptation regime. The results establish that the momentum method is equivalent to employing the standard stochastic gradient method with a re-scaled (larger) step-size value. The size of the re-scaling is determined by the momentum parameter, β. The analysis was carried out under general conditions and was not limited to quadratic risks, but is also applicable to broader choices of the risk function. Overall, the conclusions indicate that the well-known benefits of momentum constructions in the deterministic optimization scenario do not necessarily carry over to the stochastic setting when adaptation becomes necessary and gradient noise is present. The analysis also comments on a way to retain some of the advantages of the momentum construction by employing a decaying momentum parameter: one that starts at a constant level and decays to zero over time. adaptation is retained without the often-observed degradation in MSD performance.

Acknowledgments

This work was supported in part by NSF grants CIF-1524250 and ECCS-1407712, by DARPA project N66001-14-2-4029, and by a Visiting Professorship from the Leverhulme Trust, United Kingdom. The authors would like to thank PhD student Chung-Kai Yu for contributing to Section 5.3, and undergraduate student Gabrielle Robertson for contributing to the simulation in Section 7.4.

Appendix A. Proof of Lemma 2

It is shown in Eq. (3.76) of (Sayed, 2014a) that $\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4$ evolves as follows:

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 \le (1-\mu\nu)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + a_1\mu^2\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + a_2\mu^4, \tag{107}$$

where the constants a_1 and a_2 are defined as

$$a_1 \;\stackrel{\Delta}{=}\; 16\sigma_s^2, \quad a_2 \;\stackrel{\Delta}{=}\; 3\sigma_{s,4}^4. \tag{108}$$

If we iterate (16) we find that

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le (1-\mu\nu)^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^2 + a_3\mu, \tag{109}$$

where a_3 is defined as

$$a_3 \;\stackrel{\Delta}{=}\; \frac{\sigma_s^2}{\nu}. \tag{110}$$

Substituting inequality (109) into (107), we find that it holds for each iteration $i = 0, 1, 2, \ldots$

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 \le& (1-\mu\nu)\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + a_2\mu^4 + a_1a_3\mu^3 + a_4\mu^2(1-\mu\nu)^i,\\ =& \rho\mathbb{E}\|\widetilde{\boldsymbol{w}}_{i-1}\|^4 + a_2\mu^4 + a_1a_3\mu^3 + a_4\mu^2\rho^i\end{aligned} \tag{111}$$

where

$$\rho \;\stackrel{\Delta}{=}\; 1-\mu\nu, \quad a_4 \;\stackrel{\Delta}{=}\; a_1\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^2. \tag{112}$$

Iterating the inequality (111) we get

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 &\le \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^4 + a_2\mu^4\sum_{s=0}^{i}\rho^s + a_1a_3\mu^3\sum_{s=0}^{i}\rho^s + a_4\mu^2(i+1)\rho^i\\ &\le \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^4 + \frac{a_2\mu^4}{1-\rho} + \frac{a_1a_3\mu^3}{1-\rho} + a_4\mu^2(i+1)\rho^i\end{aligned}$$

$$
\begin{aligned}
&\le \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^4 + a_5\mu^3 + a_6\mu^2 + a_4\mu^2(i+1)\rho^i \\
&\overset{(a)}{\le} \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^4 + 2a_6\mu^2 + a_4\mu^2(i+1)\rho^i,
\end{aligned} \tag{113}
$$

where

$$
a_5 \overset{\Delta}{=} \frac{a_2}{\nu}, \quad a_6 \overset{\Delta}{=} \frac{a_1 a_3}{\nu}, \tag{114}
$$

and (a) holds because for sufficiently small μ such that $a_6\mu^2 > a_5\mu^3$, we have

$$
a_5\mu^3 + a_6\mu^2 = 2a_6\mu^2 - (a_6\mu^2 - a_5\mu^3) \le 2a_6\mu^2. \tag{115}
$$

Substituting (108), (110), (112) and (114) into (113), we get

$$
\begin{aligned}
\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 &\le \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^4 + A_1\sigma_s^2(i+1)\rho^i\mu^2 + \frac{A_2\sigma_s^4\mu^2}{\nu^2} \\
&= \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^4 + \frac{A_1}{\rho}\sigma_s^2(i+1)\rho^{i+1}\mu^2 + \frac{A_2\sigma_s^4\mu^2}{\nu^2}
\end{aligned} \tag{116}
$$

for some constants A_1 and A_2. When μ is sufficiently small, there must exist some constant A_3 such that

$$
\frac{A_1}{\rho} = \frac{A_1}{1-\mu\nu} \le A_3. \tag{117}
$$

Therefore, (116) becomes

$$
\mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 \le \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{w}}_{-1}\|^4 + A_3\sigma_s^2(i+1)\rho^{i+1}\mu^2 + \frac{A_2\sigma_s^4\mu^2}{\nu^2}. \tag{118}
$$

Appendix B. Proof of Lemma 3

We substitute the expression for the gradient noise from (12), evaluated at $\boldsymbol{\psi}_{i-1}$, into (23) to get:

$$
\boldsymbol{w}_i = \boldsymbol{\psi}_{i-1} - \mu_m\nabla_w J(\boldsymbol{\psi}_{i-1}) + \beta_2(\boldsymbol{\psi}_{i-1} - \boldsymbol{\psi}_{i-2}) - \mu_m\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{119}
$$

Let again $\widetilde{\boldsymbol{w}}_i = w^o - \boldsymbol{w}_i$ and $\widetilde{\boldsymbol{\psi}}_i = w^o - \boldsymbol{\psi}_i$. Subtracting both sides of (119) from w^o gives:

$$
\widetilde{\boldsymbol{w}}_i = \widetilde{\boldsymbol{\psi}}_{i-1} + \mu_m\nabla_w J(\boldsymbol{\psi}_{i-1}) - \beta_2(\boldsymbol{\psi}_{i-1} - \boldsymbol{\psi}_{i-2}) + \mu_m\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{120}
$$

We now appeal to the mean-value theorem (relation (D.9) in (Sayed, 2014a)) to write

$$
\nabla J_w(\boldsymbol{\psi}_{i-1}) = -\Big(\int_0^1 \nabla_w^2 J_w(w^o - t\widetilde{\boldsymbol{\psi}}_{i-1})dt\Big)\widetilde{\boldsymbol{\psi}}_{i-1} \overset{\Delta}{=} -\boldsymbol{H}_{i-1}\widetilde{\boldsymbol{\psi}}_{i-1}. \tag{121}
$$

and express the momentum term in the form

$$
\boldsymbol{\psi}_{i-1} - \boldsymbol{\psi}_{i-2} = \boldsymbol{\psi}_{i-1} - w^o + w^o - \boldsymbol{\psi}_{i-2} = -\widetilde{\boldsymbol{\psi}}_{i-1} + \widetilde{\boldsymbol{\psi}}_{i-2}. \tag{122}
$$

Then, expression (120) can be rewritten as

$$\widetilde{\boldsymbol{w}}_i = (I_M + \beta_2 I_M - \mu_m \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{\psi}}_{i-1} - \beta_2 \widetilde{\boldsymbol{\psi}}_{i-2} + \mu_m \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{123}$$

On the other hand, expression (22) gives

$$\widetilde{\boldsymbol{\psi}}_{i-1} = \widetilde{\boldsymbol{w}}_{i-1} + \beta_1(\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{w}}_{i-2}). \tag{124}$$

Substituting (124) into (123), we have

$$\widetilde{\boldsymbol{w}}_i = \boldsymbol{J}_{i-1}\widetilde{\boldsymbol{w}}_{i-1} + \boldsymbol{K}_{i-1}\widetilde{\boldsymbol{w}}_{i-2} + L\widetilde{\boldsymbol{w}}_{i-3} + \mu_m \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}), \tag{125}$$

where boldface quantities denote random variables:

$$\boldsymbol{J}_{i-1} = (1+\beta_1)(1+\beta_2)I_M - \mu_m(1+\beta_1)\boldsymbol{H}_{i-1} \overset{(26)}{=} (1+\beta)I_M - \mu_m(1+\beta_1)\boldsymbol{H}_{i-1} \tag{126}$$

$$\boldsymbol{K}_{i-1} = -(\beta_1 + \beta_2 + 2\beta_1\beta_2)I_M + \mu_m\beta_1\boldsymbol{H}_{i-1} = -\beta I_M + \mu_m\beta_1\boldsymbol{H}_{i-1} \tag{127}$$

$$L = \beta_1\beta_2 = 0 \tag{128}$$

It follows that we can write the extended relation:

$$\begin{bmatrix} \widetilde{\boldsymbol{w}}_i \\ \widetilde{\boldsymbol{w}}_{i-1} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{J}_{i-1} & \boldsymbol{K}_{i-1} \\ I_M & 0 \end{bmatrix}}_{\triangleq\, \boldsymbol{B}_{i-1}} \begin{bmatrix} \widetilde{\boldsymbol{w}}_{i-1} \\ \widetilde{\boldsymbol{w}}_{i-2} \end{bmatrix} + \mu_m \begin{bmatrix} \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) \\ 0 \end{bmatrix}. \tag{129}$$

where we are denoting the coefficient matrix by $\boldsymbol{B}_{i-1}$, which can be written as the difference

$$\boldsymbol{B}_{i-1} \triangleq P - \boldsymbol{M}_{i-1}, \tag{130}$$

with

$$P = \begin{bmatrix} (1+\beta)I_M & -\beta I_M \\ I_M & 0 \end{bmatrix}, \quad \boldsymbol{M}_{i-1} = \begin{bmatrix} \mu_m(1+\beta_1)\boldsymbol{H}_{i-1} & -\mu_m\beta_1\boldsymbol{H}_{i-1} \\ 0 & 0 \end{bmatrix}. \tag{131}$$

The eigenvalue decomposition of P can be easily seen to be given by $P = VDV^{-1}$, where

$$V = \begin{bmatrix} I_M & -\beta I_M \\ I_M & -I_M \end{bmatrix}, \quad V^{-1} = \frac{1}{1-\beta}\begin{bmatrix} I_M & -\beta I_M \\ I_M & -I_M \end{bmatrix}, \quad D = \begin{bmatrix} I_M & 0 \\ 0 & \beta I_M \end{bmatrix}. \tag{132}$$

Therefore, we have

$$\boldsymbol{B}_{i-1} = V(D - V^{-1}\boldsymbol{M}_{i-1}V)V^{-1} = V\begin{bmatrix} I_M - \frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1} & \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1} \\ -\frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1} & \beta I_M + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1} \end{bmatrix}V^{-1}, \tag{133}$$

where

$$\beta' \triangleq \beta\beta_1 + \beta - \beta_1 = \beta\beta_1 + \beta_2. \tag{134}$$

Multiplying both sides of (129) by V^{-1} from the left and recalling definition (28), we obtain

$$\begin{bmatrix} \widehat{\boldsymbol{w}}_i \\ \check{\boldsymbol{w}}_i \end{bmatrix} = \begin{bmatrix} I_M - \frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1} & \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1} \\ -\frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1} & \beta I_M + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1} \end{bmatrix}\begin{bmatrix} \widehat{\boldsymbol{w}}_{i-1} \\ \check{\boldsymbol{w}}_{i-1} \end{bmatrix} + \frac{\mu_m}{1-\beta}\begin{bmatrix} \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) \\ \boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) \end{bmatrix}. \tag{135}$$

Appendix C. Proof of Theorem 4

From the first row of recursion (29) we have

$$\widehat{\boldsymbol{w}}_i \;\; = \;\; \left(I_M - \frac{\mu_m \boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m \beta' \boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1} + \frac{\mu_m}{1-\beta}\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{136}$$

Let $t \in (0,1)$. Squaring both sides and taking expectations conditioned on $\boldsymbol{\mathcal{F}}_{i-1}$, and using Jensen's inequality, we obtain under Assumptions 1 and 2:

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{w}}_i\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \\
=& \left\|\left(I_M - \frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1}\right\|^2 + \frac{\mu_m^2}{(1-\beta)^2}\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \\
\overset{(a)}{\le}& \left\|(1-t)\frac{1}{1-t}\left(I_M - \frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1}\right)\widehat{\boldsymbol{w}}_{i-1} + t\frac{1}{t}\frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1}\right\|^2 + \frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2 + \sigma_s^2) \\
\le& \frac{1}{1-t}\left\|\left(I_M - \frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1}\right)\widehat{\boldsymbol{w}}_{i-1}\right\|^2 + \frac{1}{t}\left\|\frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1}\right\|^2 + \frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2 + \sigma_s^2) \\
\overset{(b)}{\le}& \frac{1}{1-t}\left(1 - \frac{\mu_m\nu}{1-\beta}\right)^2\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + \frac{1}{t}\frac{\mu_m^2\beta'^2\delta^2}{(1-\beta)^2}\|\check{\boldsymbol{w}}_{i-1}\|^2 + \frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2 + \sigma_s^2) \\
\overset{(c)}{=}& \left(1 - \frac{\mu_m\nu}{1-\beta}\right)\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + \frac{\mu_m\beta'^2\delta^2}{\nu(1-\beta)}\|\check{\boldsymbol{w}}_{i-1}\|^2 + \frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2 + \sigma_s^2).
\end{aligned} \tag{137}$$

where (a) holds because of equation (14) in Assumption (2), (b) holds because $\nu I \le \boldsymbol{H}_{i-1} \le \delta I$ under Assumption (1), and (c) holds because we selected $t = \frac{\mu_m\nu}{1-\beta}$. Taking expectation again, we remove the conditioning to find:

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 \le \left(1 - \frac{\mu_m\nu}{1-\beta}\right)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + \frac{\mu_m\beta'^2\delta^2}{\nu(1-\beta)}\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + \frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2 + \sigma_s^2). \tag{138}$$

Furthermore, squaring (124) and using the inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ we get

$$\begin{aligned}
\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2 &\le 2(1+\beta_1)^2\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + 2\beta_1^2\|\widetilde{\boldsymbol{w}}_{i-2}\|^2 \le 2(1+\beta_1)^2(\|\widetilde{\boldsymbol{w}}_{i-1}\|^2 + \|\widetilde{\boldsymbol{w}}_{i-2}\|^2) \\
&= 2(1+\beta_1)^2\left\|\begin{bmatrix}\widetilde{\boldsymbol{w}}_{i-1}\\ \widetilde{\boldsymbol{w}}_{i-2}\end{bmatrix}\right\|^2 = 2(1+\beta_1)^2\left\|VV^{-1}\begin{bmatrix}\widetilde{\boldsymbol{w}}_{i-1}\\ \widetilde{\boldsymbol{w}}_{i-2}\end{bmatrix}\right\|^2 \\
&\le 2(1+\beta_1)^2\|V\|^2\left\|\begin{bmatrix}\widehat{\boldsymbol{w}}_{i-1}\\ \check{\boldsymbol{w}}_{i-1}\end{bmatrix}\right\|^2.
\end{aligned} \tag{139}$$

It is known that there exists some constant $d > 0$ such that $\|V\|^2 \le d\|V\|_F^2$. From expression (27) for V we have

$$\|V\|_F^2 = 3\|I_M\|_F^2 + \beta^2\|I_M\|_F^2 \le 4\|I_M\|_F^2 = 4M.$$

Let $v^2 \overset{\Delta}{=} 4dM$, so that $\|V\|^2 \le v^2$. Therefore, under expectation, we conclude that it also holds:

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2 \le 2(1+\beta_1)^2 v^2(\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2). \tag{140}$$

Substituting (140) into (138), we get

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 \le \Big(1-\frac{\mu_m\nu}{1-\beta}+\frac{2(1+\beta_1)^2\gamma^2v^2}{(1-\beta)^2}\mu_m^2\Big)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2+\frac{\mu_m^2\sigma_s^2}{(1-\beta)^2} + \Big(\frac{\mu_m\beta'^2\delta^2}{\nu(1-\beta)}+\frac{2(1+\beta_1)^2\gamma^2v^2}{(1-\beta)^2}\mu_m^2\Big)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2. \tag{141}$$

Now, let us consider the second row of (29), namely,

$$\check{\boldsymbol{w}}_i = -\frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1}\widehat{\boldsymbol{w}}_{i-1}+\left(\beta I_M+\frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\right)\check{\boldsymbol{w}}_{i-1}+\frac{\mu_m}{1-\beta}\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{142}$$

As before, squaring and taking expectations of both sides, and using Jensen's inequality, we obtain under Assumptions 1 and 2:

$$\begin{aligned}
&\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2\\
&\le\ \mathbb{E}\left\|\beta\check{\boldsymbol{w}}_{i-1}+\left(\frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}-\frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\widehat{\boldsymbol{w}}_{i-1}\right)\right\|^2+\frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2+\sigma_s^2)\\
&\overset{(a)}{\le}\ \beta\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2+\frac{1}{1-\beta}\mathbb{E}\left\|\frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}-\frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\widehat{\boldsymbol{w}}_{i-1}\right\|^2+\frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2+\sigma_s^2)\\
&\le\ \beta\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2+\frac{2\mu_m^2\beta'^2\delta^2}{(1-\beta)^3}\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2+\frac{2\mu_m^2\delta^2}{(1-\beta)^3}\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2+\frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2+\sigma_s^2)\\
&=\ \Big(\beta+\frac{2\mu_m^2\beta'^2\delta^2}{(1-\beta)^3}\Big)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2+\frac{2\mu_m^2\delta^2}{(1-\beta)^3}\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2+\frac{\mu_m^2}{(1-\beta)^2}(\gamma^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2+\sigma_s^2),
\end{aligned} \tag{143}$$

where (a) holds since $\mathbb{E}\|\beta\boldsymbol{x}+\boldsymbol{y}\|^2=\mathbb{E}\|\beta\boldsymbol{x}+(1-\beta)\frac{1}{1-\beta}\boldsymbol{y}\|^2\le\beta\mathbb{E}\|\boldsymbol{x}\|^2+\frac{1}{1-\beta}\mathbb{E}\|\boldsymbol{y}\|^2$. Substituting (139) into (143), it follows that:

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le\Big(\beta+\frac{2\mu_m^2\beta'^2\delta^2}{(1-\beta)^3}+\frac{2\mu_m^2\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2}\Big)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2+\frac{\mu_m^2\sigma_s^2}{(1-\beta)^2} + \Big(\frac{2\mu_m^2\delta^2}{(1-\beta)^3}+\frac{2\mu_m^2\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2}\Big)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \tag{144}$$

Combining relations (141) and (144) leads to the desired result (33)–(34). Let us now examine the stability of the 2×2 coefficient matrix:

$$\Gamma \triangleq \begin{bmatrix} a & b\\ c & d\end{bmatrix}, \tag{145}$$

where

$$a = 1-\frac{\mu_m\nu}{1-\beta}+\frac{2(1+\beta_1)^2\gamma^2v^2}{(1-\beta)^2}\mu_m^2,\quad b=\frac{\mu_m\beta'^2\delta^2}{\nu(1-\beta)}+\frac{2(1+\beta_1)^2\gamma^2v^2}{(1-\beta)^2}\mu_m^2,$$
$$c=\frac{2\mu_m^2\delta^2}{(1-\beta)^3}+\frac{2\mu_m^2\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2},\qquad d=\beta+\frac{2\mu_m^2\beta'^2\delta^2}{(1-\beta)^3}+\frac{2\mu_m^2\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2}. \tag{146}$$

When μ_m is sufficiently small, a, b, c, d are all positive. Since the spectral radius of a matrix is upper bounded by its 1-norm, we have that

$$\rho(\Gamma) \le \max\{a + c,\ b + d\}. \tag{147}$$

From (146), we further have

$$\begin{aligned}
a + c &\le 1 - \frac{\mu_m \nu}{1-\beta} + \frac{2(1+\beta_1)^2\gamma^2 v^2}{(1-\beta)^2}\mu_m^2 + \frac{2\mu_m^2\delta^2}{(1-\beta)^3} + \frac{2\mu_m^2\gamma^2(1+\beta_1)^2 v^2}{(1-\beta)^2} \\
&= 1 - \frac{\mu_m \nu}{1-\beta} + \frac{4(1-\beta)(1+\beta_1)^2\gamma^2\nu^2 + 2\delta^2}{(1-\beta)^3}\mu_m^2 \\
&\le 1 - \frac{\mu_m \nu}{1-\beta} + \frac{16\gamma^2\nu^2 + 2\delta^2}{(1-\beta)^3}\mu_m^2,
\end{aligned} \tag{148}$$

where the last inequality holds because $1 - \beta < 1$ and $1 + \beta_1 < 2$. Similarly, we also have

$$b + d \le \beta + \frac{\delta^2 \mu_m}{\nu(1-\beta)} + \frac{16\gamma^2\nu^2 + 2\delta^2}{(1-\beta)^3}\mu_m^2. \tag{149}$$

Combining (147)–(149), we reach

$$\rho(\Gamma) \le \max\left\{1 - \frac{\mu_m\nu}{1-\beta} + \frac{16\gamma^2\nu^2 + 2\delta^2}{(1-\beta)^3}\mu_m^2,\ \ \beta + \frac{\delta^2\mu_m}{\nu(1-\beta)} + \frac{16\gamma^2\nu^2 + 2\delta^2}{(1-\beta)^3}\mu_m^2\right\}. \tag{150}$$

If the step-size μ_m is small enough to satisfy the following conditions

$$\begin{cases}
\frac{\mu_m\nu}{2(1-\beta)} > \frac{16\gamma^2\nu^2+2\delta^2}{(1-\beta)^3}\mu_m^2, \\
\frac{\delta^2\mu_m}{\nu(1-\beta)} > \frac{16\gamma^2\nu^2+2\delta^2}{(1-\beta)^3}\mu_m^2, \\
1-\beta > \frac{2\delta^2\mu_m}{\nu(1-\beta)},
\end{cases} \tag{151}$$

which is also equivalent to

$$\mu_m < \min\left\{\frac{(1-\beta)^2\nu}{32\gamma^2\nu^2 + 4\delta^2}, \frac{(1-\beta)^2\delta^2}{16\gamma^2\nu^3 + 2\delta^2\nu}, \frac{\nu(1-\beta)^2}{2\delta^2}\right\} = \frac{(1-\beta)^2\nu}{32\gamma^2\nu^2 + 4\delta^2}, \tag{152}$$

then it holds that

$$\rho(\Gamma) < \max\left\{1 - \frac{\mu_m\nu}{2(1-\beta)},\ \ \beta + \frac{2\delta^2\mu_m}{\nu(1-\beta)}\right\} \le 1, \tag{153}$$

in which case Γ will be a stable matrix.

When Γ is stable, it then follows from (33) that

$$\limsup_{i\to\infty} \begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \end{bmatrix} \le (I_2 - \Gamma)^{-1} \begin{bmatrix} e \\ f \end{bmatrix}. \tag{154}$$

Notice that

$$(I_2 - \Gamma)^{-1} = \begin{bmatrix} 1-a & -b \\ -c & 1-d \end{bmatrix}^{-1} = \frac{1}{(1-a)(1-d) - bc} \begin{bmatrix} 1-d & b \\ c & 1-a \end{bmatrix}$$

$$\overset{(146)}{=} \frac{1}{\mu_m\nu + p_1\mu_m^2 + p_2\mu_m^3 + p_3\mu_m^4} \begin{bmatrix} 1-\beta+p_4\mu_m^2 & \frac{\mu_m\beta'^2\delta^2}{\nu(1-\beta)} + p_5\mu_m^2 \\ \frac{2\mu_m^2\delta^2}{(1-\beta)^3} + \frac{2\mu_m^2\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2} & \frac{\mu_m\nu}{1-\beta} + p_6\mu_m^2 \end{bmatrix} \tag{155}$$

where

$$\begin{aligned} p_1 &\triangleq -\frac{2(1+\beta_1)^2\gamma^2\nu^2}{1-\beta} < 0, \quad & p_4 &\triangleq -\frac{2\beta'^2\delta^2}{(1-\beta)^3} - \frac{2\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2} < 0, \\ p_5 &\triangleq \frac{2(1+\beta_1)^2\gamma^2\nu^2}{(1-\beta)^2} > 0, & p_6 &\triangleq -\frac{2(1+\beta_1)^2\gamma^2\nu^2}{(1-\beta)^2} < 0. \end{aligned} \tag{156}$$

For simplicity, we omit the expression of p_2 and p_3 here. Notice that

$$\mu_m\nu + p_1\mu_m^2 + p_2\mu_m^3 + p_3\mu_m^4 = \frac{\mu_m\nu}{2} + \left(\frac{\mu_m\nu}{2} + p_1\mu_m^2 + p_2\mu_m^3 + p_3\mu_m^4\right). \tag{157}$$

Although $p_1 < 0$, it still holds that $\frac{\mu_m\nu}{2} + p_1\mu_m^2 + p_2\mu_m^3 + p_3\mu_m^4 > 0$ when μ_m is sufficiently small, which implies that

$$\mu_m\nu + p_1\mu_m^2 + p_2\mu_m^3 + p_3\mu_m^4 > \frac{\mu_m\nu}{2}. \tag{158}$$

Similarly, it holds that

$$\frac{\mu_m\beta'\delta^2}{\nu(1-\beta)} + p_5\mu_m^2 = \frac{2\mu_m\beta'\delta^2}{\nu(1-\beta)} - \left(\frac{\mu_m\beta'\delta^2}{\nu(1-\beta)} - p_5\mu_m^2\right) \le \frac{2\mu_m\beta'\delta^2}{\nu(1-\beta)}, \tag{159}$$

where the last inequality holds because $\frac{2\mu_m\beta'\delta^2}{\nu(1-\beta)} - p_5\mu_m^2 > 0$ for sufficiently small step-size. Furthermore, since $p_4 < 0$ and $p_6 < 0$, we also have

$$1-\beta+p_4\mu_m^2 < 1-\beta, \quad \frac{\mu_m\nu}{1-\beta} + p_6\mu_m^2 < \frac{\mu_m\nu}{1-\beta}. \tag{160}$$

Substitute (158), (159) and (160) into (155), we have

$$(I_2-\Gamma)^{-1} \le \begin{bmatrix} \frac{2(1-\beta)}{\mu_m\nu} & \frac{4\beta'^2\delta^2}{\nu^2(1-\beta)} \\ \frac{4\mu_m\delta^2}{(1-\beta)^3\nu} + \frac{4\mu_m\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2\nu} & \frac{2}{1-\beta} \end{bmatrix} \tag{161}$$

Combining (154) and (161), we have

$$\begin{aligned} \limsup_{i\to\infty} \begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \end{bmatrix} &\le (I_2-\Gamma)^{-1} \begin{bmatrix} e \\ f \end{bmatrix} \\ &\le \begin{bmatrix} \frac{2(1-\beta)}{\mu_m\nu} & \frac{4\beta'^2\delta^2}{\nu^2(1-\beta)} \\ \frac{4\mu_m\delta^2}{(1-\beta)^3\nu} + \frac{4\mu_m\gamma^2(1+\beta_1)^2v^2}{(1-\beta)^2\nu} & \frac{2}{1-\beta} \end{bmatrix} \begin{bmatrix} \frac{\mu_m^2\sigma_s^2}{(1-\beta)^2} \\ \frac{\mu_m^2\sigma_s^2}{(1-\beta)^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{2\mu_m\sigma_s^2}{(1-\beta)\nu} + \frac{4\beta'^2\delta^2\sigma_s^2\mu_m^2}{(1-\beta)^3\nu^2} \\ \frac{2\mu_m^2\sigma_s^2}{(1-\beta)^3} + \frac{4\mu_m^3\delta^2\sigma_s^2}{(1-\beta)^5\nu} + \frac{4\mu_m^3\gamma^2(1+\beta_1)^2v^2\sigma_s^2}{(1-\beta)^4\nu} \end{bmatrix} \le \begin{bmatrix} \frac{3\mu_m\sigma_s^2}{(1-\beta)} \\ \frac{3\mu_m^2\sigma_s^2}{(1-\beta)^3} \end{bmatrix} \end{aligned} \tag{162}$$

where in the last inequality we choose sufficiently small μ_m such that

$$\frac{4\beta'^2\delta^2\sigma_s^2\mu_m^2}{(1-\beta)^3\nu^2} < \frac{\mu_m\sigma_s^2}{(1-\beta)\nu}, \quad \frac{4\mu_m^3\delta^2\sigma_s^2}{(1-\beta)^5\nu} + \frac{4\mu_m^3\gamma^2(1+\beta_1)^2v^2\sigma_s^2}{(1-\beta)^4\nu} < \frac{\mu_m^2\sigma_s^2}{(1-\beta)^3} \tag{163}$$

Therefore, we have the following result

$$\limsup_{i\to\infty} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 = O\left(\frac{\mu_m\sigma_s^2}{(1-\beta)\nu}\right), \quad \limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 = O\left(\frac{\mu_m^2\sigma_s^2}{(1-\beta)^3}\right). \tag{164}$$

and

$$\begin{aligned}
\limsup_{i\to\infty} \mathbb{E}\left\|\begin{bmatrix} \widetilde{\boldsymbol{w}}_i \\ \widetilde{\boldsymbol{w}}_{i-1} \end{bmatrix}\right\|^2 &= \limsup_{i\to\infty} \mathbb{E}\left\|V\begin{bmatrix} \widehat{\boldsymbol{w}}_i \\ \check{\boldsymbol{w}}_i \end{bmatrix}\right\|^2 \\
&\le v^2\left(\limsup_{i\to\infty} \mathbb{E}\left\|\begin{bmatrix} \widehat{\boldsymbol{w}}_i \\ \check{\boldsymbol{w}}_i \end{bmatrix}\right\|^2\right) \\
&= v^2\left(\limsup_{i\to\infty}(\mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 + \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2)\right) = O\left(\frac{\mu_m\sigma_s^2}{(1-\beta)\nu}\right),
\end{aligned} \tag{165}$$

from which we conclude that (37) holds.

Appendix D. Proof of Corollary 5

To simplify the notation, we refer to (33) and introduce the quantities:

$$z_i = \begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \end{bmatrix}, \ \Gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \ r = \begin{bmatrix} e \\ f \end{bmatrix}. \tag{166}$$

Then, relation (33) can be rewritten as

$$z_i \preceq \Gamma z_{i-1} + r. \tag{167}$$

It follows that, in terms of the 1−norm,

$$\|z_i\|_1 \le \|\Gamma\|_1\|z_{i-1}\|_1 + \|r\|_1, \tag{168}$$

where

$$\|\Gamma\|_1 = \max\left\{1 - \frac{\mu_m\nu}{1-\beta} + B_1\mu_m^2, \beta + B_2\mu_m\right\} \tag{169}$$

for some constant B_1 and B_2. Now we can choose μ_m sufficiently small to satisfy

$$B_1\mu_m^2 < \frac{\nu\mu_m}{2(1-\beta)}, \qquad \left(B_2 + \frac{\nu}{2(1-\beta)}\right)\mu_m < 1-\beta, \tag{170}$$

which implies that

$$\|\Gamma\|_1 \le 1 - \frac{\mu_m\nu}{1-\beta} + B_1\mu_m^2 \le 1 - \frac{\mu_m\nu}{2(1-\beta)} \stackrel{\Delta}{=} \rho_1 < 1 \tag{171}$$

Then, from (168) we have

$$\|z_i\|_1 \le \rho_1 \|z_{i-1}\|_1 + \|r\|_1. \tag{172}$$

Iterating (172) gives

$$\|z_i\|_1 \le \rho_1^{i+1} \|z_{-1}\|_1 + \frac{\|r\|_1}{1-\rho_1}. \tag{173}$$

Recall the expressions of e and f from (34), we have $\|r\|_1 \le \frac{B_3 \mu_m^2 \sigma_s^2}{(1-\beta)^2}$ for some constant B_3. Since $1-\rho_1 = \frac{\mu_m \nu}{2(1-\beta)}$, we get $\|r\|_1/(1-\rho_1) \le \frac{2B_3\mu_m\sigma_s^2}{(1-\beta)\nu}$. From (173), we have

$$\|z_i\|_1 \le \rho_1^{i+1} \|z_{-1}\|_1 + \frac{2B_3\mu_m\sigma_s^2}{(1-\beta)\nu}. \tag{174}$$

Accordingly, using

$$\|z_i\|_1 = \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 + \mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \tag{175}$$

we also find that

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^2 \le \rho_1^{i+1} \|z_{-1}\|_1 + \frac{2B_3\mu_m\sigma_s^2}{(1-\beta)\nu}. \tag{176}$$

On the other hand, we know from the second row of (33) that

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le \big(\beta + c_1\mu_m^2\big)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + c_2\mu_m^2 \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + c_3\mu_m^2 \tag{177}$$

for constants

$$c_1 \;\stackrel{\Delta}{=}\; \frac{2\beta'^2\delta^2}{(1-\beta)^3} + \frac{2\gamma^2(1+\beta_1)^2 v^2}{(1-\beta)^2}, \quad c_2 \;\stackrel{\Delta}{=}\; \frac{2\delta^2}{(1-\beta)^3} + \frac{2\gamma^2(1+\beta)^2 v^2}{(1-\beta)^2}, \quad c_3 \;\stackrel{\Delta}{=}\; \frac{\sigma_s^2}{(1-\beta)^2}. \tag{178}$$

To simplify the notation, with the facts that $\beta' < 1, \beta < 1$ and $\beta_1 < 1$, we have

$$c_1 \le \frac{B_4(\delta^2+\gamma^2)}{(1-\beta)^3} \;\stackrel{\Delta}{=}\; c_4, \quad c_2 \le \frac{B_4(\delta^2+\gamma^2)}{(1-\beta)^3} = c_4 \tag{179}$$

for some constant B_4. Substituting (179) into (177) we get

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le \big(\beta + c_4\mu_m^2\big)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + c_4\mu_m^2 \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + c_3\mu_m^2. \tag{180}$$

Now we substitute (176) into (180), and reach

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le \big(\beta + c_4\mu_m^2\big)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + c_4\rho_1^i \|z_{-1}\|_1 \mu_m^2 + \frac{2B_3c_4\sigma_s^2}{(1-\beta)\nu}\mu_m^3 + c_3\mu_m^2. \tag{181}$$

When μ_m is sufficiently small such that

$$\frac{2B_3c_4\sigma_s^2}{(1-\beta)\nu}\mu_m^3 \le c_3\mu_m^2, \tag{182}$$

(181) becomes

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le \big(\beta + c_4\mu_m^2\big)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + c_4\rho_1^i\|z_{-1}\|_1\mu_m^2 + 2c_3\mu_m^2. \tag{183}$$

Notice that

$$\beta + c_4\mu_m^2 = 1 - (1-\beta) + c_4\mu_m^2 = 1 - \frac{1-\beta}{2} + \left(c_1\mu_m^2 - \frac{1-\beta}{2}\right). \tag{184}$$

It is clear that we can choose a sufficiently small μ_m for the last term between brackets to become negative, in which case

$$\beta + c_4\mu_m^2 \le 1 - \frac{1-\beta}{2} = \frac{1+\beta}{2} \;\stackrel{\Delta}{=}\; \alpha < 1 \tag{185}$$

It follows that

$$\begin{aligned}\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le&\; \alpha\,\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + \big(c_4\|z_{-1}\|_1\rho_1^i\big)\,\mu_m^2 + 2c_3\mu_m^2\\ \le&\; \alpha^{i+1}\mathbb{E}\|\check{\boldsymbol{w}}_{-1}\|^2 + c_4\|z_{-1}\|_1\mu_m^2\rho_1^i\sum_{s=0}^{i}\left(\frac{\alpha}{\rho_1}\right)^s + \frac{2c_3\mu_m^2}{1-\alpha}.\end{aligned} \tag{186}$$

Recall that $\rho_1 = 1 - \frac{\mu_m\nu}{2(1-\beta)}$ and $\alpha = 1-(1-\beta)/2$. Therefore, it holds that $\alpha/\rho_1 < 1$ for sufficiently small μ_m. As a result, we have

$$\sum_{s=0}^{i}\left(\frac{\alpha}{\rho_1}\right)^s \le \frac{1}{1-\frac{\alpha}{\rho_1}} = \frac{\rho_1}{\rho_1-\alpha} = \frac{2(1-\beta)-\mu_m\nu}{(1-\beta)^2-\mu_m\nu} \le \frac{B_5}{1-\beta} \tag{187}$$

for some constant B_5 when μ_m is sufficiently small. Substituting (187) into (186), we get

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le\; \alpha^{i+1}\mathbb{E}\|\check{\boldsymbol{w}}_{-1}\|^2 + \frac{B_5c_4\|z_{-1}\|_1\rho_1^i}{1-\beta}\mu_m^2 + \frac{4c_3\mu_m^2}{1-\beta}. \tag{188}$$

To assess the term that depends on the initial state, $\mathbb{E}\|\check{\boldsymbol{w}}_{-1}\|^2$, let us consider the boundary conditions (24)–(25), Then, from (28) it holds that

$$\check{\boldsymbol{w}}_{-1} = \frac{\widetilde{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{w}}_{-2}}{1-\beta} = \frac{\boldsymbol{w}_{-2} - \boldsymbol{w}_{-1}}{1-\beta} = \frac{\mu_m\nabla_w Q(\boldsymbol{w}_{-2};\boldsymbol{\theta}_{-1})}{1-\beta} \tag{189}$$

so that $\mathbb{E}\|\check{\boldsymbol{w}}_{-1}\|^2 = c_5\mu_m^2$, where

$$c_5 \;\stackrel{\Delta}{=}\; \mathbb{E}\|\nabla_w Q(\boldsymbol{w}_{-2};\boldsymbol{\theta}_{-1})\|^2/(1-\beta)^2. \tag{190}$$

Substituting this conclusion into (188), and recalling the expression of c_3, c_4 and c_5, we arrive at

$$\begin{aligned}\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le&\; \frac{B_6\alpha^{i+1}\mu_m^2}{(1-\beta)^2} + \frac{B_7(\delta^2+\gamma^2)\rho_1^i}{(1-\beta)^4}\mu_m^2 + \frac{B_8\mu_m^2\sigma_s^2}{(1-\beta)^3}\\ \stackrel{(a)}{\le}&\; \frac{B_6\rho_1^{i+1}\mu_m^2}{(1-\beta)^2} + \frac{B_9(\delta^2+\gamma^2)\rho_1^{i+1}}{(1-\beta)^4}\mu_m^2 + \frac{B_8\mu_m^2\sigma_s^2}{(1-\beta)^3}\end{aligned}$$

$$\overset{(b)}{\leq} \frac{B_{10}(\delta^2+\gamma^2)\rho_1^{i+1}}{(1-\beta)^4}\mu_m^2 + \frac{B_8\mu_m^2\sigma_s^2}{(1-\beta)^3}, \tag{191}$$

where (a) holds because $\alpha \leq \rho_1$ when μ_m is sufficiently small, and there must exist some constant B_9 such that $B_7/\rho_1 < B_9$; (b) holds because there must exist some constant B_{10} such that

$$B_6(1-\beta)^2 + B_9(\delta^2+\gamma^2) \leq B_{10}(\delta^2+\gamma^2). \tag{192}$$

Appendix E. Proof of Theorem 6

The argument below is motivated by the derivation of Theorem 9.2 in (Sayed, 2014a). Here, however, we extend the arguments and expand the details in order to clearly identify the constants inside the $O(\mu)$ notation, which was not necessary in (Sayed, 2014a). The derivation becomes more demanding, as the arguments show.

From the first row of recursion (29) we have

$$\widehat{\boldsymbol{w}}_i = \left(I_M - \frac{\mu_m \boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1} + \frac{\mu_m}{1-\beta}\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{193}$$

Now applying the following inequality, for any two vectors $\{a,b\}$:

$$\|a+b\|^4 \leq \|a\|^4 + 3\|b\|^4 + 8\|a\|^2\|b\|^2 + 4\|a\|^2(a^{\mathsf{T}}b) \tag{194}$$

we get

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \\
&= \left\|\left(I_M - \frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}\right\|^4 + \frac{3\mu_m^4}{(1-\beta)^4}\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \\
&\quad + \frac{8\mu_m^2}{(1-\beta)^2}\left\|\left(I_M - \frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}\right\|^2 \mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \\
&\leq \left\|\left(I_M - \frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}\right\|^4 + \frac{3\mu_m^4(\gamma_4^4\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^4+\sigma_{s,4}^4)}{(1-\beta)^4} \\
&\quad + \frac{8\mu_m^2}{(1-\beta)^2}\left\|\left(I_M - \frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}\right\|^2(\gamma^2\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2+\sigma_s^2).
\end{aligned} \tag{195}$$

We next bound each of the terms that appear on the right-hand side. Using Jensen's inequality, the lower and upper bounds on the Hessian matrix from (11), we have

$$\begin{aligned}
&\left\|\left(I_M - \frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1} + \frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}\right\|^4 \\
&= \left\|(1-t)\frac{1}{1-t}\left(I_M - \frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1} + t\frac{1}{t}\frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}\right\|^4
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{1}{(1-t)^3}\left(1-\frac{\mu_m\nu}{1-\beta}\right)^4\|\widehat{\boldsymbol{w}}_{i-1}\|^4+\frac{1}{t^3}\frac{\mu_m^4\beta'^4\delta^4}{(1-\beta)^4}\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&\overset{(a)}{=} \left(1-\frac{\mu_m\nu}{1-\beta}\right)\|\widehat{\boldsymbol{w}}_{i-1}\|^4+\frac{\mu_m\beta'^4\delta^4}{(1-\beta)\nu^3}\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&= (1-q_1\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^4+q_2\mu_m\|\check{\boldsymbol{w}}_{i-1}\|^4.
\end{aligned}
\tag{196}
$$

where (a) holds because we set $t=\mu_m\nu/(1-\beta)$, and q_1, q_2 are defined as

$$
q_1 \;\overset{\Delta}{=}\; \frac{\nu}{1-\beta}, \qquad q_2 \;\overset{\Delta}{=}\; \frac{\beta'^4\delta^4}{(1-\beta)\nu^3}.
\tag{197}
$$

Next we check the terms $\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}]$ and $\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}]$. From (139) we have

$$
\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2 \quad \le B_1(\|\widehat{\boldsymbol{w}}_{i-1}\|^2+\|\check{\boldsymbol{w}}_{i-1}\|^2),
\tag{198}
$$

where $B_1=2(1+\beta_1)^2v^2$, which also implies that

$$
\begin{aligned}
\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^4 \;&\le\; B_1^2(\|\widehat{\boldsymbol{w}}_{i-1}\|^2+\|\check{\boldsymbol{w}}_{i-1}\|^2)^2\\
&\le\; 2B_1^2(\|\widehat{\boldsymbol{w}}_{i-1}\|^4+\|\check{\boldsymbol{w}}_{i-1}\|^4) \;=\; B_2(\|\widehat{\boldsymbol{w}}_{i-1}\|^4+\|\check{\boldsymbol{w}}_{i-1}\|^4),
\end{aligned}
\tag{199}
$$

where $B_2=2B_1^2$. Furthermore, recall in (137) that

$$
\begin{aligned}
&\left\|\left(I_M-\frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\right)\widehat{\boldsymbol{w}}_{i-1}+\frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1}\right\|^2\\
&\le \left(1-\frac{\mu_m\nu}{1-\beta}\right)\|\widehat{\boldsymbol{w}}_{i-1}\|^2+\frac{\mu_m\beta'^2\delta^2}{\nu(1-\beta)}\|\check{\boldsymbol{w}}_{i-1}\|^2\\
&= (1-q_1\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^2+q_3\mu_m\|\check{\boldsymbol{w}}_{i-1}\|^2,
\end{aligned}
\tag{200}
$$

where we define

$$
q_3 \;\overset{\Delta}{=}\; \frac{\beta'^2\delta^2}{\nu(1-\beta)}.
\tag{201}
$$

Now substituting (196), (198), (199) and (200) into (195), we get

$$
\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}]\\
&\le (1-q_1\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^4+q_2\mu_m\|\check{\boldsymbol{w}}_{i-1}\|^4+\frac{3B_2\gamma_4^4\mu_m^4}{(1-\beta)^4}\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^4+\|\check{\boldsymbol{w}}_{i-1}\|^4\right)+\frac{3\sigma_{s,4}^4\mu_m^4}{(1-\beta)^4}\\
&\quad+\frac{8\mu_m^2}{(1-\beta)^2}\left[(1-q_1\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^2+q_3\mu_m\|\check{\boldsymbol{w}}_{i-1}\|^2\right]\left[\gamma^2B_1\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^2+\|\check{\boldsymbol{w}}_{i-1}\|^2\right)+\sigma_s^2\right]\\
&= (1-q_1\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^4+q_2\mu_m\|\check{\boldsymbol{w}}_{i-1}\|^4+q_4\mu_m^4\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^4+\|\check{\boldsymbol{w}}_{i-1}\|^4\right)+q_5\mu_m^4\\
&\quad+q_6\mu_m^2\left[(1-q_1\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^2+q_3\mu_m\|\check{\boldsymbol{w}}_{i-1}\|^2\right]\left[\gamma^2B_1\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^2+\|\check{\boldsymbol{w}}_{i-1}\|^2\right)+\sigma_s^2\right],\\
&= (1-q_1\mu_m+q_4\mu_m^4)\|\widehat{\boldsymbol{w}}_{i-1}\|^4+(q_2\mu_m+q_4\mu_m^4)\|\check{\boldsymbol{w}}_{i-1}\|^4+q_5\mu_m^4\\
&\quad+q_6\gamma^2B_1(1-q_1\mu_m)\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^4+q_6q_3\gamma^2B_1\mu_m^3\|\check{\boldsymbol{w}}_{i-1}\|^4
\end{aligned}
$$

$$
\begin{aligned}
&\quad + q_6\gamma^2 B_1\mu_m^2(1-q_1\mu_m+q_3\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^2\|\check{\boldsymbol{w}}_{i-1}\|^2\\
&\quad + q_6\sigma_s^2\mu_m^2(1-q_1\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + q_6q_3\sigma_s^2\mu_m^3\|\check{\boldsymbol{w}}_{i-1}\|^2\\
&\overset{(a)}{\le} (1-q_1\mu_m+q_4\mu_m^4)\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + (q_2\mu_m+q_4\mu_m^4)\|\check{\boldsymbol{w}}_{i-1}\|^4 + q_5\mu_m^4\\
&\quad + q_6\gamma^2 B_1(1-q_1\mu_m)\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + q_6q_3\gamma^2 B_1\mu_m^3\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&\quad + q_6\gamma^2 B_1\mu_m^2(1-q_1\mu_m+q_3\mu_m)\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^4+\|\check{\boldsymbol{w}}_{i-1}\|^4\right)\\
&\quad + q_6\sigma_s^2\mu_m^2(1-q_1\mu_m)\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + q_6q_3\sigma_s^2\mu_m^3\|\check{\boldsymbol{w}}_{i-1}\|^2,
\end{aligned} \tag{202}
$$

where we define

$$
q_4 \overset{\Delta}{=} \frac{3B_2\gamma_4^4}{(1-\beta)^4},\quad q_5 \overset{\Delta}{=} \frac{3\sigma_{s,4}^4}{(1-\beta)^4},\quad q_6 \overset{\Delta}{=} \frac{8}{(1-\beta)^2}, \tag{203}
$$

and (a) holds because for any two variables $a, b > 0$ we have

$$
ab < 2ab \le a^2 + b^2. \tag{204}
$$

When μ_m is chosen sufficiently small, from (202) we reach

$$
\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}]\\
&\le \left(1-\frac{q_1\mu_m}{2}\right)\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + 2q_2\mu_m\|\check{\boldsymbol{w}}_{i-1}\|^4 + q_6\sigma_s^2\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + q_6q_3\sigma_s^2\mu_m^3\|\check{\boldsymbol{w}}_{i-1}\|^2 + q_5\mu_m^4\\
&= \left(1-\frac{q_1\mu_m}{2}\right)\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + 2q_2\mu_m\|\check{\boldsymbol{w}}_{i-1}\|^4 + q_7\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + q_8\mu_m^3\|\check{\boldsymbol{w}}_{i-1}\|^2 + q_5\mu_m^4,
\end{aligned} \tag{205}
$$

where we define

$$
q_7 \overset{\Delta}{=} q_6\sigma_s^2,\qquad q_8 \overset{\Delta}{=} q_6q_3\sigma_s^2. \tag{206}
$$

On the other hand, recall from (142) that

$$
\check{\boldsymbol{w}}_i = -\frac{\mu_m}{1-\beta}\boldsymbol{H}_{i-1}\widehat{\boldsymbol{w}}_{i-1} + \left(\beta I_M + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\right)\check{\boldsymbol{w}}_{i-1} + \frac{\mu_m}{1-\beta}\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}). \tag{207}
$$

Now we also apply inequality (194) to the above equation and get

$$
\begin{aligned}
&\mathbb{E}[\|\check{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}]\\
&\quad= \left\|-\frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\widehat{\boldsymbol{w}}_{i-1} + \left(\beta I_M + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\right)\check{\boldsymbol{w}}_{i-1}\right\|^4 + \frac{3\mu_m^4\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}\right]}{(1-\beta)^4}\\
&\qquad + \frac{8\mu_m^2}{(1-\beta)^2}\left\|\frac{-\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\widehat{\boldsymbol{w}}_{i-1} + \left(\beta I_M + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\right)\check{\boldsymbol{w}}_{i-1}\right\|^2\mathbb{E}\left[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\right]\\
&\quad\le \left\|-\frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\widehat{\boldsymbol{w}}_{i-1} + \left(\beta I_M + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\right)\check{\boldsymbol{w}}_{i-1}\right\|^4 + \frac{3\mu_m^4(\gamma_4^4\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^4+\sigma_{s,4}^4)}{(1-\beta)^4}\\
&\qquad + \frac{8\mu_m^2}{(1-\beta)^2}\left\|\frac{-\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\widehat{\boldsymbol{w}}_{i-1} + \left(\beta I_M + \frac{\mu_m\beta'}{1-\beta}\boldsymbol{H}_{i-1}\right)\check{\boldsymbol{w}}_{i-1}\right\|^2(\gamma^2\|\widetilde{\boldsymbol{\psi}}_{i-1}\|^2+\sigma_s^2).
\end{aligned} \tag{208}
$$

We next bound each of the terms that appear on the right-hand side. Using Jensen's inequality, the lower and upper bounds on the Hessian matrix from (11), and the inequality $\|a+b\|^4 \le 8\|a\|^4 + 8\|b\|^4$, we have

$$\begin{aligned}
&\left\| -\frac{\mu_m \boldsymbol{H}_{i-1}}{1-\beta}\widehat{\boldsymbol{w}}_{i-1} + \left(\beta I_M + \frac{\mu_m \beta'}{1-\beta}\boldsymbol{H}_{i-1}\right)\check{\boldsymbol{w}}_{i-1}\right\|^4 \\
&\qquad = \left\|\beta\check{\boldsymbol{w}}_{i-1} + (1-\beta)\left(\frac{\mu_m\beta'}{(1-\beta)^2}\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1} - \frac{\mu_m}{(1-\beta)^2}\boldsymbol{H}_{i-1}\widehat{\boldsymbol{w}}_{i-1}\right)\right\|^4 \\
&\qquad \le \beta\|\check{\boldsymbol{w}}_{i-1}\|^4 + (1-\beta)\left\|\frac{\mu_m\beta'}{(1-\beta)^2}\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1} - \frac{\mu_m}{(1-\beta)^2}\boldsymbol{H}_{i-1}\widehat{\boldsymbol{w}}_{i-1}\right\|^4 \\
&\qquad \le \beta\|\check{\boldsymbol{w}}_{i-1}\|^4 + \frac{8\mu_m^4\beta'^4\delta^4}{(1-\beta)^7}\|\check{\boldsymbol{w}}_{i-1}\|^4 + \frac{8\mu_m^4\delta^4}{(1-\beta)^7}\|\widehat{\boldsymbol{w}}_{i-1}\|^4 \\
&\qquad = (\beta + p_1\mu_m^4)\|\check{\boldsymbol{w}}_{i-1}\|^4 + p_2\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^4,
\end{aligned} \tag{209}$$

where we define

$$p_1 \triangleq \frac{8\beta'^4\delta^4}{(1-\beta)^7}, \qquad p_2 \triangleq \frac{8\delta^4}{(1-\beta)^7}. \tag{210}$$

Moreover, recall in (143) that

$$\begin{aligned}
&\left\|\beta\check{\boldsymbol{w}}_{i-1} + \left(\frac{\mu_m\beta'\boldsymbol{H}_{i-1}}{1-\beta}\check{\boldsymbol{w}}_{i-1} - \frac{\mu_m\boldsymbol{H}_{i-1}}{1-\beta}\widehat{\boldsymbol{w}}_{i-1}\right)\right\|^2 \\
&\qquad\qquad \le \left(\beta + \frac{2\mu_m^2\beta'^2\delta^2}{(1-\beta)^3}\right)\|\check{\boldsymbol{w}}_{i-1}\|^2 + \frac{2\mu_m^2\delta^2}{(1-\beta)^3}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\
&\qquad\qquad = (\beta + p_3\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^2 + p_4\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^2,
\end{aligned} \tag{211}$$

where we define

$$p_3 \triangleq \frac{2\beta'^2\delta^2}{(1-\beta)^3}, \quad p_4 \triangleq \frac{2\delta^2}{(1-\beta)^3}. \tag{212}$$

Now substituting (209), (211), (198) and (199) into (208), we have

$$\begin{aligned}
&\mathbb{E}[\|\check{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \\
&\le (\beta + p_1\mu_m^4)\|\check{\boldsymbol{w}}_{i-1}\|^4 + p_2\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + \frac{3B_2\gamma_4^4\mu_m^4}{(1-\beta)^4}\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + \|\check{\boldsymbol{w}}_{i-1}\|^4\right) + \frac{3\sigma_{s,4}^4\mu_m^4}{(1-\beta)^4} \\
&\qquad + \frac{8\mu_m^2}{(1-\beta)^2}\left[(\beta+p_3\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^2 + p_4\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^2\right]\left[\gamma^2 B_1\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + \|\check{\boldsymbol{w}}_{i-1}\|^2\right) + \sigma_s^2\right] \\
&= (\beta + p_1\mu_m^4)\|\check{\boldsymbol{w}}_{i-1}\|^4 + p_2\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + p_5\mu_m^4\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + \|\check{\boldsymbol{w}}_{i-1}\|^4\right) + p_6\mu_m^4 \\
&\qquad + p_7\mu_m^2\left[(\beta+p_3\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^2 + p_4\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^2\right]\left[\gamma^2 B_1\left(\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + \|\check{\boldsymbol{w}}_{i-1}\|^2\right) + \sigma_s^2\right] \\
&= [\beta + (p_1+p_5)\mu_m^4]\|\check{\boldsymbol{w}}_{i-1}\|^4 + (p_2+p_5)\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + p_6\mu_m^4 \\
&\qquad + p_7\gamma^2 B_1\mu_m^2(\beta+p_3\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^4 + p_4p_7\gamma^2 B_1\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^4 \\
&\qquad + p_7\gamma^2 B_1\mu_m^2(\beta + p_3\mu_m^2 + p_4\mu_m^2)\|\widehat{\boldsymbol{w}}_{i-1}\|^2\|\check{\boldsymbol{w}}_{i-1}\|^2
\end{aligned}$$

$$
\begin{aligned}
&\quad + p_7\sigma_s^2\mu_m^2(\beta + p_3\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^2 + p_4 p_7\sigma_s^2\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\
&\le [\beta + (p_1+p_5)\mu_m^4]\|\check{\boldsymbol{w}}_{i-1}\|^4 + (p_2+p_5)\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + p_6\mu_m^4 \\
&\quad + p_7\gamma^2 B_1\mu_m^2(\beta + p_3\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^4 + p_4 p_7\gamma^2 B_1\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^4 \\
&\quad + p_7\gamma^2 B_1\mu_m^2(\beta + p_3\mu_m^2 + p_4\mu_m^2)\left(\|\check{\boldsymbol{w}}_{i-1}\|^4 + \|\widehat{\boldsymbol{w}}_{i-1}\|^4\right) \\
&\quad + p_7\sigma_s^2\mu_m^2(\beta + p_3\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^2 + p_4 p_7\sigma_s^2\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^2,
\end{aligned} \tag{213}
$$

where we define

$$
p_5 \;\stackrel{\Delta}{=}\; \frac{3B_2\gamma_4^4}{(1-\beta)^4}, \quad p_6 \;\stackrel{\Delta}{=}\; \frac{3\sigma_{s,4}^4}{(1-\beta)^4}, \quad p_7 \;\stackrel{\Delta}{=}\; \frac{8}{(1-\beta)^2}. \tag{214}
$$

When μ_m is sufficiently small, we obtain from (213):

$$
\begin{aligned}
&\mathbb{E}[\|\check{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \\
&\le (\beta + 2p_7\gamma^2 B_1\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^4 + 2p_7\gamma^2 B_1\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + p_4p_7\sigma_s^2\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\
&\quad + 2p_7\beta\sigma_s^2\mu_m^2\|\check{\boldsymbol{w}}_{i-1}\|^2 + p_6\mu_m^4 \\
&= (\beta + p_8\mu_m^2)\|\check{\boldsymbol{w}}_{i-1}\|^4 + p_8\mu_m^2\|\widehat{\boldsymbol{w}}_{i-1}\|^4 + p_9\mu_m^4\|\widehat{\boldsymbol{w}}_{i-1}\|^2 + p_{10}\mu_m^2\|\check{\boldsymbol{w}}_{i-1}\|^2 + p_6\mu_m^4
\end{aligned} \tag{215}
$$

where we define

$$
p_8 \;\stackrel{\Delta}{=}\; 2p_7\gamma^2 B_1, \quad p_9 \;\stackrel{\Delta}{=}\; p_4p_7\sigma_s^2, \quad p_{10} \;\stackrel{\Delta}{=}\; 2p_7\beta\sigma_s^2. \tag{216}
$$

Combining (205) and (215), we have

$$
\begin{bmatrix} \mathbb{E}[\|\widehat{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \\ \mathbb{E}[\|\check{\boldsymbol{w}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \end{bmatrix} \le \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \|\widehat{\boldsymbol{w}}_{i-1}\|^4 \\ \|\check{\boldsymbol{w}}_{i-1}\|^4 \end{bmatrix} + \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} \begin{bmatrix} \|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\ \|\check{\boldsymbol{w}}_{i-1}\|^2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}, \tag{217}
$$

where the constants are

$$
\begin{aligned}
&a \;\stackrel{\Delta}{=}\; 1 - \frac{q_1}{2}\mu_m, \quad && b \;\stackrel{\Delta}{=}\; 2q_2\mu_m, \quad && a' \;\stackrel{\Delta}{=}\; q_6\sigma_s^2\mu_m^2, \quad && b' \;\stackrel{\Delta}{=}\; q_3q_6\sigma_s^2\mu_m^3, \\
&c \;\stackrel{\Delta}{=}\; p_8\mu_m^2, && d \;\stackrel{\Delta}{=}\; \beta + p_8\mu_m^2, && c' \;\stackrel{\Delta}{=}\; p_9\mu_m^4, && d' \;\stackrel{\Delta}{=}\; p_{10}\mu_m^2, \\
&e \;\stackrel{\Delta}{=}\; q_5\mu_m^4, && f \;\stackrel{\Delta}{=}\; p_6\mu_m^4.
\end{aligned} \tag{218}
$$

Taking expectations again over $\boldsymbol{\mathcal{F}}_{i-1}$ for both sides of the inequality (217), we have

$$
\begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^4 \\ \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \end{bmatrix} \le \underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{\Gamma} \begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^4 \\ \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4 \end{bmatrix} + \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} \begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}, \tag{219}
$$

Recall from Theorem 4 that

$$
\limsup_{i\to\infty} \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 = O(\mu_m), \quad \limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 = O(\mu_m^2), \tag{220}
$$

then it holds that

$$
\limsup_{i\to\infty} \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} \begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} O(\mu_m^3) \\ O(\mu_m^4) \end{bmatrix}. \tag{221}
$$

When μ_m is sufficiently small, it can be verified that Γ is stable. Therefore, it holds that

$$\begin{aligned}\begin{bmatrix} \limsup_{i\to\infty} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^4 \\ \limsup_{i\to\infty} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \end{bmatrix} &= (I-\Gamma)^{-1}\left(\limsup_{i\to\infty}\begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}\begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix}\right) \\ &= \begin{bmatrix} O(\mu_m^2) \\ O(\mu_m^4) \end{bmatrix}. \end{aligned} \tag{222}$$

Furthermore,

$$\limsup_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^4 \le 2v^4 \limsup_{i\to\infty}(\mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^4 + \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4) = O\left(\mu_m^2\right). \tag{223}$$

Appendix F. Proof of Corollary 7

Recall from (219) that

$$\underbrace{\begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^4 \\ \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \end{bmatrix}}_{y_i} \le \underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{\Gamma_1}\underbrace{\begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^4 \\ \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4 \end{bmatrix}}_{y_{i-1}} + \underbrace{\begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}}_{\Gamma_2}\underbrace{\begin{bmatrix} \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}\|^2 \\ \mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \end{bmatrix}}_{z_{i-1}} + \underbrace{\begin{bmatrix} e \\ f \end{bmatrix}}_{r}. \tag{224}$$

We then have

$$\|y_i\|_1 \le \|\Gamma_1\|_1\|y_{i-1}\|_1 + \|\Gamma_2\|_1\|z_{i-1}\|_1 + \|r\|_1. \tag{225}$$

Notice that

$$\|\Gamma_1\|_1 = \max\left\{1 - \frac{q_1\mu_m}{2} + p_8\mu_m^2, \beta + 2q_2\mu_m + p_8\mu_m^2\right\}, \tag{226}$$

we can always choose μ_m small enough such that

$$\|\Gamma_1\|_1 \le 1 - \frac{q_1\mu_m}{4} = 1 - \frac{\mu_m\nu}{4(1-\beta)} \stackrel{\Delta}{=} \rho_2. \tag{227}$$

Similarly, we can also choose μ_m small enough such that

$$\|\Gamma_2\|_1 = \max\left\{q_6\sigma_s^2\mu_m^2 + p_9\mu_m^4, q_3q_6\sigma_s^2\mu_m^3 + p_{10}\mu_m^2\right\} \le (q_6\sigma_s^2 + p_{10})\mu_m^2. \tag{228}$$

Also recall (174) that

$$\|z_i\|_1 \le \rho_1^{i+1}\|z_{-1}\|_1 + s_1\mu_m \stackrel{(a)}{\le} \rho_2^{i+1}\|z_{-1}\|_1 + s_1\mu_m \tag{229}$$

where we define

$$s_1 \stackrel{\Delta}{=} \frac{2E_1\sigma_s^2}{(1-\beta)\nu}, \tag{230}$$

for some constant E_1, and (a) holds because $\rho_1 = 1 - \frac{\mu_m\nu}{2(1-\beta)} \le \rho_2$. Substituting (227), (228) and (229) into (225), we have

$$\|y_i\|_1 \le \rho_2\|y_{i-1}\|_1 + (q_6\sigma_s^2 + p_{10})(\rho_2^i\|z_{-1}\|_1 + s_1\mu_m)\mu_m^2 + 2p_6\mu_m^4$$

$$
\begin{aligned}
&= \rho_2\|y_{i-1}\|_1 + s_2\rho_2^i\mu_m^2 + s_3\mu_m^3 + 2p_6\mu_m^4 \\
&\overset{(a)}{\le} \rho_2\|y_{i-1}\|_1 + s_2\rho_2^i\mu_m^2 + 2s_3\mu_m^3, \qquad (231)
\end{aligned}
$$

where (a) holds when μ_m is sufficiently small, and the constants s_2 and s_3 are defined as

$$
s_2 \;\triangleq\; (q_6\sigma_s^2 + p_{10})\|z_{-1}\|_1, \quad s_3 \;\triangleq\; (q_6\sigma_s^2 + p_{10})s_1. \qquad (232)
$$

Iterating (231), we reach

$$
\|y_i\|_1 \le \rho_2^{i+1}\|y_{-1}\|_1 + s_2(i+1)\rho_2^i\mu_m^2 + \frac{2s_3\mu_m^3}{1-\rho_2}. \qquad (233)
$$

Since $\|y_i\|_1 = \mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^4 + \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4$, we have

$$
\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{w}}_i\|^4 &\le \rho_2^{i+1}\|y_{-1}\|_1 + s_2(i+1)\rho_2^i\mu_m^2 + \frac{2s_3\mu_m^3}{1-\rho_2} \\
&= \rho_2^{i+1}\|y_{-1}\|_1 + s_2(i+1)\rho_2^i\mu_m^2 + \frac{8(1-\beta)s_3\mu_m^2}{\nu} \\
&= \rho_2^{i+1}\|y_{-1}\|_1 + s_2(i+1)\rho_2^i\mu_m^2 + s_4\mu_m^2, \qquad (234)
\end{aligned}
$$

where s_4 is defined as

$$
s_4 \;\triangleq\; \frac{8(1-\beta)s_3}{\nu} \qquad (235)
$$

Now we substitute (234) and (176) into the second row of (219) and reach

$$
\begin{aligned}
\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \le{}& (\beta + p_8\mu_m^2)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4 + p_8\mu_m^2\left[\rho_2^i\|y_{-1}\|_1 + s_2 i\rho_2^{i-1}\mu_m^2 + s_4\mu_m^2\right] \\
&+ p_{10}\mu_m^2\|\check{\boldsymbol{w}}_{i-1}\|^2 + p_9\mu_m^4\left[\rho_2^i\|z_{-1}\|_1 + s_1\mu_m\right] + p_6\mu_m^4. \qquad (236)
\end{aligned}
$$

Using the bounds for $\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2$ from Corollary 5, the above inequality becomes

$$
\begin{aligned}
\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \le{}& (\beta + p_8\mu_m^2)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4 + p_8\mu_m^2\left[\rho_2^i\|y_{-1}\|_1 + s_2 i\rho_2^{i-1}\mu_m^2 + s_4\mu_m^2\right] \\
&+ p_{10}E_2\mu_m^2\left(\frac{(\delta^2+\gamma^2)\rho_1^i\mu_m^2}{(1-\beta)^4} + \frac{\sigma_s^2\mu_m^2}{(1-\beta)^3}\right) \\
&+ p_9\mu_m^4\left[\rho_2^i\|z_{-1}\|_1 + s_1\mu_m\right] + p_6\mu_m^4 \\
\overset{(a)}{\le}{}& (1 - \frac{1-\beta}{2})\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4 + s_5\rho_2^i\mu_m^2 + s_6 i\rho_2^{i-1}\mu_m^4 + s_7\mu_m^4 \\
&+ s_8\rho_1^i\mu_m^4 + s_9\mu_m^4 + s_{10}\rho_2^i\mu_m^4 + p_6\mu_m^4 + s_1p_9\mu_m^5 \\
\overset{(b)}{\le}{}& \alpha\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4 + s_5\rho_2^i\mu_m^2 + s_6 i\rho_2^{i-1}\mu_m^4 + s_8\rho_1^i\mu_m^4 + s_{10}\rho_2^i\mu_m^4 \\
&+ (s_7 + s_9 + 2p_6)\mu_m^4 \\
\overset{(c)}{\le}{}& \alpha\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4 + s_5\rho_2^i\mu_m^2 + s_6 i\rho_2^{i-1}\mu_m^4 + s_8\rho_2^i\mu_m^4 + s_{10}\rho_2^i\mu_m^4 \\
&+ (s_7 + s_9 + 2p_6)\mu_m^4 \qquad (237)
\end{aligned}
$$

where E_2 is some constant. The inequality (a) holds because step-size μ_m is chosen small enough such that $(1-\beta)/2 > p_8\mu_m^2$, (b) holds because $\alpha = 1-(1-\beta)/2$ and μ_m is chosen such that $p_6\mu_m^4 > s_1 p_9 \mu_m^5$, and (c) holds because $\rho_1 < \rho_2$. Moreover, the other constants are defined as

$$\begin{aligned} s_5 &\triangleq p_8\|y_{-1}\|_1, & s_6 &\triangleq p_8 s_2, & s_7 &\triangleq p_8 s_4 \\ s_8 &\triangleq \frac{p_{10}E_2(\delta^2+\gamma^2)}{(1-\beta)^4}, & s_9 &\triangleq \frac{p_{10}E_2\sigma_s^2}{(1-\beta)^3}, & s_{10} &\triangleq p_9\|z_{-1}\|_1. \end{aligned} \tag{238}$$

Now we continue iterating (237) and reach

$$\begin{aligned} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \le{}& \alpha^{i+1}\mathbb{E}\|\check{\boldsymbol{w}}_{-1}\|^4 + s_5\mu_m^2\rho_2^i\sum_{k=0}^{i}\left(\frac{\alpha}{\rho_2}\right)^k + s_6\mu_m^4\rho_2^{i-1}\sum_{k=0}^{i}(i-k)\left(\frac{\alpha}{\rho_2}\right)^k \\ &+ s_8\mu_m^4\rho_2^i\sum_{k=0}^{i}\left(\frac{\alpha}{\rho_2}\right)^k + s_{10}\mu_m^4\rho_2^i\sum_{k=0}^{i}\left(\frac{\alpha}{\rho_2}\right)^k + \frac{2(s_7+s_9+2p_6)\mu_m^4}{1-\beta}. \end{aligned} \tag{239}$$

Recall $\rho_2 = 1 - \frac{\mu_m\nu}{4(1-\beta)}$, and we can choose μ_m small enough such that $\rho_2 > \alpha = 1-\frac{1-\beta}{2}$. In this situation, we have

$$\frac{\alpha}{\rho_2} < 1, \quad \text{and} \quad \sum_{k=0}^{i}\left(\frac{\alpha}{\rho_2}\right)^k < \sum_{k=0}^{\infty}\left(\frac{\alpha}{\rho_2}\right)^k = \frac{\rho_2}{\rho_2-\alpha} \le \frac{E_3}{1-\beta}, \tag{240}$$

where E_3 is some constant. Meanwhile, we also have

$$\sum_{k=0}^{i}(i-k)\left(\frac{\alpha}{\rho_2}\right)^k \le i\sum_{k=0}^{i}\left(\frac{\alpha}{\rho_2}\right)^k \le i\sum_{k=0}^{\infty}\left(\frac{\alpha}{\rho_2}\right)^k \le \frac{iE_3}{1-\beta}. \tag{241}$$

We substitute (240) and (241) into (239), and reach

$$\begin{aligned} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \le{}& \rho_2^{i+1}\mathbb{E}\|\check{\boldsymbol{w}}_{-1}\|^4 + \frac{E_3 s_5\mu_m^2\rho_2^i}{1-\beta} + \frac{iE_3 s_6\mu_m^4\rho_2^{i-1}}{1-\beta} \\ &+ \frac{E_3 s_8\mu_m^4\rho_2^i}{1-\beta} + \frac{E_3 s_{10}\mu_m^4\rho_2^i}{1-\beta} + \frac{2(s_7+s_9+2p_6)\mu_m^4}{1-\beta}. \end{aligned} \tag{242}$$

Recall from (189) that $\mathbb{E}\|\check{\boldsymbol{w}}_{-1}\|^4 = E_4\mu_m^4$, where $E_4 = \mathbb{E}\|\nabla_w Q(\boldsymbol{w}_{-2};\boldsymbol{\theta}_{-1})\|^4$. Substituting into (242) we reach that

$$\begin{aligned} \mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \le{}& E_4\rho_2^{i+1}\mu_m^4 + \frac{E_3 s_5\mu_m^2\rho_2^i}{1-\beta} + \frac{iE_3 s_6\mu_m^4\rho_2^{i-1}}{1-\beta} \\ &+ \frac{E_3 s_8\mu_m^4\rho_2^i}{1-\beta} + \frac{E_3 s_{10}\mu_m^4\rho_2^i}{1-\beta} + \frac{2(s_7+s_9+2p_6)\mu_m^4}{1-\beta}. \end{aligned} \tag{243}$$

Substituting (238) into (243) and recall $\alpha < \rho_2$, we finally reach

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 = O\Bigg(\frac{\gamma^2\rho_2^i}{(1-\beta)^3}\mu_m^2 + \rho_2^{i+1}\mu_m^4 + \frac{\gamma^2\sigma_s^2 i\rho_2^{i-1}}{(1-\beta)^5}\mu_m^4 + \frac{\sigma_s^2(\delta^2+\gamma^2)\rho_2^i}{(1-\beta)^7}\mu_m^4 + \frac{\delta^2\sigma_s^2\rho_2^i}{(1-\beta)^6}\mu_m^4$$

$$+\frac{\gamma^2\sigma_s^4}{(1-\beta)^5\nu^2}\mu_m^4+\frac{\sigma_s^4}{(1-\beta)^6}\mu_m^4+\frac{\sigma_{s,4}^4}{(1-\beta)^5}\mu_m^4\Bigg). \tag{244}$$

Since there must exist some constants E_5, E_6 and E_7 such that

$$\begin{aligned}\frac{\gamma^2\rho_2^i}{(1-\beta)^3}\mu_m^2 &\le \frac{E_5\gamma^2\rho_2^{i+1}}{(1-\beta)^3}\mu_m^2,\\ \rho_2^{i+1}\mu_m^4+\frac{\gamma^2\sigma_s^2 i\rho_2^{i-1}}{(1-\beta)^5}\mu_m^4+\frac{\sigma_s^2(\delta^2+\gamma^2)\rho_2^i}{(1-\beta)^7}\mu_m^4+\frac{\delta^2\sigma_s^2\rho_2^i}{(1-\beta)^6}\mu_m^4 &\le \frac{E_6\sigma_s^2(\delta^2+\gamma^2)(i+1)\rho_2^{i+1}}{(1-\beta)^7}\mu_m^4,\\ \frac{\gamma^2\sigma_s^4}{(1-\beta)^5\nu^2}\mu_m^4+\frac{\sigma_s^4}{(1-\beta)^6}\mu_m^4+\frac{\sigma_{s,4}^4}{(1-\beta)^5}\mu_m^4 &\le \frac{E_7[(\gamma^2+\nu^2)\sigma_s^4+\sigma_{s,4}^4\nu^2]}{(1-\beta)^6\nu^2}\mu_m^4,\end{aligned} \tag{245}$$

we finally reach the conclusion in (44).

Appendix G. Proof of Theorem 8

Subtracting (55) and (61) we get

$$\widehat{\boldsymbol{w}}_i-\widetilde{\boldsymbol{x}}_i=(I_M-\mu R_u)(\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1})+\mu(\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1}))+\mu\beta' R_u\check{\boldsymbol{w}}_{i-1}, \tag{246}$$

where

$$\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})=(R_u-\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}})\widetilde{\boldsymbol{\psi}}_{i-1}-\boldsymbol{u}_i\boldsymbol{v}(i),\quad \boldsymbol{s}_i(\boldsymbol{x}_{i-1})=(R_u-\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}})\widetilde{\boldsymbol{x}}_{i-1}-\boldsymbol{u}_i\boldsymbol{v}(i). \tag{247}$$

Substituting into (246) gives

$$\widehat{\boldsymbol{w}}_i-\widetilde{\boldsymbol{x}}_i=(I_M-\mu R_u)(\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1})+\mu(R_u-\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}})(\widetilde{\boldsymbol{\psi}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1})+\mu\beta' R_u\check{\boldsymbol{w}}_{i-1}. \tag{248}$$

Now note that in the quadratic case, the Hessian matrix of $J(w)$ is equal to R_u. It follows that condition (11) is satisfied with the identifications $\nu=\lambda_{\min}(R_u),\delta=\lambda_{\max}(R_u)$. Let $t\in(0,1)$. By squaring (248) and taking expectations, and applying Jensen's inequality, we obtain

$$\begin{aligned}&\mathbb{E}\|\widehat{\boldsymbol{w}}_i-\widetilde{\boldsymbol{x}}_i\|^2\\ &\le \mathbb{E}\left\|(I_M-\mu R_u)(\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1})+\mu\beta' R_u\check{\boldsymbol{w}}_{i-1}\right\|^2+\mu^2\mathbb{E}\|R_u-\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}}\|^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2\\ &\overset{(a)}{\le} \frac{1}{1-t}(1-\mu\nu)^2\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2+\frac{1}{t}\mu^2\beta'^2\delta^2\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2+B_1\mu^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2\\ &\overset{(b)}{\le} (1-\mu\nu)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2+\frac{\mu\beta'^2\delta^2}{\nu}\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2+B_1\mu^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2,\end{aligned} \tag{249}$$

where (a) holds because of Jensen's inequality and we let $B_1=\mathbb{E}\|R_u-\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}}\|^2$, and (b) holds by choosing $t=\mu\nu$. To bound the last term in the above relation, we use (124) to note that

$$\widetilde{\boldsymbol{\psi}}_i-\widetilde{\boldsymbol{x}}_i=\widetilde{\boldsymbol{w}}_i+\beta_1(\widetilde{\boldsymbol{w}}_i-\widetilde{\boldsymbol{w}}_{i-1})-\widetilde{\boldsymbol{x}}_i=(\widetilde{\boldsymbol{w}}_i-\widetilde{\boldsymbol{x}}_i)-\beta_1(\widetilde{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{w}}_i) \tag{250}$$

On the other hand, from (28) we have

$$\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i = \frac{1}{1-\beta}(\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i) - \frac{\beta}{1-\beta}(\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_i) = (\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i) - \frac{\beta}{1-\beta}(\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{w}}_i). \quad (251)$$

so that

$$\widetilde{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{x}}_i = \widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i + \frac{\beta-\beta_1+\beta\beta_1}{1-\beta}(\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{w}}_i) = \widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i + \frac{\beta'}{1-\beta}(\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{w}}_i) = \widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i - \beta'\check{\boldsymbol{w}}_i. \quad (252)$$

where we used the definition for β' from (30) and the definition for $\check{\boldsymbol{w}}_i$ from (28). Therefore, from Jensen's inequality again, we get

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le 2\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 + 2\beta'^2\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2. \quad (253)$$

Substituting into (249) gives

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 &\le (1-\mu\nu+2B_1\mu^2)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + \left(\frac{\mu\beta'^2\delta^2}{\nu} + 2B_1\beta'^2\mu^2\right)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \\
&\overset{(a)}{\le} \left(1-\frac{\mu\nu}{2}\right)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + \frac{2\mu\beta'^2\delta^2}{\nu}\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \\
&\le \left(1-\frac{\mu\nu}{2}\right)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + \frac{B_2\mu\delta^2}{\nu}\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2, \quad (254)
\end{aligned}$$

where $B_2 = 2\beta'^2$ and the inequality (a) holds when μ is chosen small enough such that

$$\frac{\mu\nu}{2} > 2B_1\mu^2 \quad \text{and} \quad \frac{\mu\beta'^2\delta^2}{\nu} > 2B_1\beta'^2\mu^2. \quad (255)$$

Recall from Corollary 5 that

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le C_1\left(\frac{(\delta^2+\gamma^2)\rho_1^{i+1}\mu_m^2}{(1-\beta)^4} + \frac{\sigma_s^2\mu_m^2}{(1-\beta)^3}\right) \quad (256)$$

for each iteration $i = 0, 1, 2, 3, \ldots$, where C_1 is some constant. Recall $\mu = \mu_m/(1-\beta)$, we then have

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le C_1\left(\frac{(\delta^2+\gamma^2)\rho_1^{i+1}\mu^2}{(1-\beta)^2} + \frac{\sigma_s^2\mu^2}{1-\beta}\right) \quad (257)$$

This fact, together with inequality (254), leads to

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le \left(1-\frac{\mu\nu}{2}\right)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + B_2C_1\left(\frac{\delta^2(\delta^2+\gamma^2)\rho_1^i\mu^3}{\nu(1-\beta)^2} + \frac{\delta^2\sigma_s^2\mu^3}{\nu(1-\beta)}\right). \quad (258)$$

Recall from Corollary 5 that $\rho_1 = 1 - \frac{\mu_m\nu}{2(1-\beta)} = 1 - \frac{\mu\nu}{2}$, then (258) becomes

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le \rho_1\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + B_2C_1\left(\frac{\delta^2(\delta^2+\gamma^2)\rho_1^i\mu^3}{\nu(1-\beta)^2} + \frac{\delta^2\sigma_s^2\mu^3}{\nu(1-\beta)}\right). \quad (259)$$

For brevity, we denote

$$e_1 \triangleq \frac{B_2 C_1(\delta^2+\gamma^2)\delta^2}{\nu(1-\beta)^2}, \quad e_2 \triangleq \frac{B_2 C_1 \delta^2 \sigma_s^2}{\nu(1-\beta)}. \tag{260}$$

Inequality (259) will become

$$\begin{aligned} \mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 &\le \rho_1 \mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + e_1\rho_1^i\mu^3 + e_2\mu^3 \\ &\le \rho_1^{i+1}\mathbb{E}\|\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1}\|^2 + e_1(i+1)\rho_1^i\mu^3 + \frac{e_2\mu^3}{1-\rho_1}. \end{aligned} \tag{261}$$

Recall from the first equation in (251) that for $i=-1$:

$$\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1} = \frac{1}{1-\beta}(\widetilde{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1}) - \frac{\beta}{1-\beta}(\widetilde{\boldsymbol{w}}_{-2} - \widetilde{\boldsymbol{x}}_{-1}). \tag{262}$$

Now, using the assumption that the momentum and standard recursions started from the same initial states, $\boldsymbol{w}_{-2} = \boldsymbol{x}_{-1}$ and $\boldsymbol{w}_{-1} = \boldsymbol{w}_{-2} - \mu_m \nabla_w Q(\boldsymbol{w}_{-2}; \boldsymbol{d}(-1), \boldsymbol{u}_{-1})$, and recall $\mu = \mu_m/(1-\beta)$, then we have

$$\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1} = \frac{1}{1-\beta}(\widetilde{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{w}}_{-2}) = \mu \nabla_w Q(\boldsymbol{w}_{-2}; \boldsymbol{d}(-1), \boldsymbol{u}_{-1}). \tag{263}$$

Therefore, it holds that

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1}\|^2 = B_4\mu^2, \tag{264}$$

where $B_4 = \mathbb{E}\|\nabla_w Q(\boldsymbol{w}_{-2}; \boldsymbol{d}(-1), \boldsymbol{u}_{-1})\|^2$. Substituting (264) and (260) into (261), we reach

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le B_4\rho_1^{i+1}\mu^2 + \frac{B_2C_1(\delta^2+\gamma^2)\delta^2(i+1)\rho_1^i}{\nu(1-\beta)^2}\mu^3 + \frac{4B_2C_1\delta^2\sigma_s^2}{\nu^2(1-\beta)}\mu^2. \tag{265}$$

Furthermore, using (251) and $\check{\boldsymbol{w}}_i = \frac{\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{w}}_{i-1}}{1-\beta}$ from (28) we have

$$\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i = (\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i) + \beta\check{\boldsymbol{w}}_i, \tag{266}$$

which implies that

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le 2\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 + 2\beta^2\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le 2\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 + 2\mathbb{E}\|\check{\boldsymbol{w}}_i\|^{\cdot} \tag{267}$$

Substituting (257) and (265) into (267), we have

$$\begin{aligned} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le\ & B_4\rho_1^{i+1}\mu^2 + \frac{B_2C_1(\delta^2+\gamma^2)\delta^2(i+1)\rho_1^i}{\nu(1-\beta)^2}\mu^3 + \frac{4B_2C_1\delta^2\sigma_s^2}{\nu^2(1-\beta)}\mu^2 \\ & + 2C_1\left(\frac{(\delta^2+\gamma^2)\rho_1^{i+1}\mu^2}{(1-\beta)^2} + \frac{\sigma_s^2\mu^2}{1-\beta}\right) \\ =\ & O\left(\frac{\delta^2+\gamma^2}{(1-\beta)^2}\rho_1^{i+1}\mu^2 + \frac{\delta^2(\delta^2+\gamma^2)(i+1)\rho_1^{i+1}}{\nu(1-\beta)^2}\mu^3 + \frac{\delta^2\sigma_s^2\mu^2}{\nu^2(1-\beta)}\right). \end{aligned} \tag{268}$$

Appendix H. Verifying Assumptions 5 and 6

Least-mean-squares problem. Consider first the mean-squares cost (47). Since in this case $\boldsymbol{H}_{i-1} = \boldsymbol{R}_{i-1} = R_u$, we find that Assumption 6 holds automatically. With regards to Assumption 5, at any iteration i, we have

$$\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1}) = (R_u - \boldsymbol{u}_i\boldsymbol{u}_i^\mathsf{T})(\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}). \tag{269}$$

so that, under the assumption of independent and stationary regression vectors,

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2 | \boldsymbol{\mathcal{F}}_{i-1}] \le \ \xi_1 \, \|\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2, \tag{270}$$

where $\xi_1 = \mathbb{E}\|R_u - \boldsymbol{u}_i\boldsymbol{u}_i^\mathsf{T}\|^2$. Similarly,

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{w}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^4 | \boldsymbol{\mathcal{F}}_{i-1}] \le \ \xi_2 \, \|\widetilde{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^4, \tag{271}$$

where $\xi_2 = \mathbb{E}\|R_u - \boldsymbol{u}_i\boldsymbol{u}_i^\mathsf{T}\|^4$. Therefore, Assumption 5 holds.

Regularized logistic regression. Consider next the regularized logistic regression risk

$$J(w) \ \stackrel{\Delta}{=} \ \frac{\rho}{2}\|w\|^2 + \mathbb{E}\Big\{ \ln\big[1 + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T} w)\big]\Big\}, \tag{272}$$

where $\boldsymbol{h}_i \in \mathbb{R}^M$ is a streaming sequence of independent feature vectors with $R_h = \mathbb{E}\boldsymbol{h}_i\boldsymbol{h}_i^\mathsf{T} > 0$, and $\boldsymbol{\gamma}(i) \in \{-1, +1\}$ is a streaming sequence of class labels. We assume the random processes $\{\boldsymbol{\gamma}(i), \boldsymbol{h}_i\}$ are wide-sense stationary. Moreover, $\rho > 0$ is a regularization parameter. We first verify the feasibility of Assumption 5. Note that the approximate gradient vector is given by:

$$\widehat{\nabla_w J}(\boldsymbol{w}) = \rho \boldsymbol{w} - \frac{\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{w})}{1 + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{w})}\boldsymbol{\gamma}(i)\boldsymbol{h}_i \tag{273}$$

and, hence,

$$\begin{aligned}
&\widehat{\nabla_w J}(\boldsymbol{\psi}_{i-1}) - \widehat{\nabla_w J}(\boldsymbol{x}_{i-1}) \\
&\quad \le \rho\|\boldsymbol{\psi}_{i-1} - \boldsymbol{x}_{i-1}\| + \|\boldsymbol{h}_i\| \left\| \frac{\exp(-\boldsymbol{\gamma}_i\boldsymbol{h}_i^\mathsf{T}\boldsymbol{\psi}_{i-1})}{1 + \exp(-\boldsymbol{\gamma}_i\boldsymbol{h}_i^\mathsf{T}\boldsymbol{\psi}_{i-1})} - \frac{\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{x}_{i-1})}{1 + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{x}_{i-1})} \right\|
\end{aligned} \tag{274}$$

Note that

$$\begin{aligned}
&\left\| \frac{\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{\psi}_{i-1})}{1 + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{\psi}_{i-1})} - \frac{\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{x}_{i-1})}{1 + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{x}_{i-1})} \right\| \\
&\qquad = \left\| \frac{\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{\psi}_{i-1}) - \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{x}_{i-1})}{[1 + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{\psi}_{i-1})][1 + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{x}_{i-1})]} \right\| \\
&\qquad \le \left\| \frac{\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{\psi}_{i-1}) - \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{x}_{i-1})}{\exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{\psi}_{i-1}) + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\boldsymbol{x}_{i-1})} \right\| \\
&\qquad = \left\| \frac{\exp(\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\frac{\boldsymbol{x}_{i-1} - \boldsymbol{\psi}_{i-1}}{2}) - \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\frac{\boldsymbol{x}_{i-1} - \boldsymbol{\psi}_{i-1}}{2})}{\exp(\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\frac{\boldsymbol{x}_{i-1} - \boldsymbol{\psi}_{i-1}}{2}) + \exp(-\boldsymbol{\gamma}(i)\boldsymbol{h}_i^\mathsf{T}\frac{\boldsymbol{x}_{i-1} - \boldsymbol{\psi}_{i-1}}{2})} \right\|
\end{aligned}$$

$$= \left|\tanh\Big(\boldsymbol{\gamma}(i)\boldsymbol{h}_i^{\mathsf{T}}(\boldsymbol{x}_{i-1}-\boldsymbol{\psi}_{i-1})/2\Big)\right| \le \frac{1}{2}\|\boldsymbol{h}_i\|\|\boldsymbol{\psi}_{i-1}-\boldsymbol{x}_{i-1}\|. \tag{275}$$

where in the last inequality we used the property $|\tanh(y)| \le |y|,\ \forall y \in \mathbb{R}$. Substituting (275) into (274), we get

$$\|\widehat{\nabla_w J}(\boldsymbol{\psi}_{i-1}) - \widehat{\nabla_w J}(\boldsymbol{x}_{i-1})\| \le \boldsymbol{\eta}_{1,i}\|\boldsymbol{\psi}_{i-1}-\boldsymbol{x}_{i-1}\|, \tag{276}$$

where $\boldsymbol{\eta}_{1,i} = \rho + \|\boldsymbol{h}_i\|^2/2$ is a random variable.

On the other hand, it is shown in Eq. (2.20) of (Sayed, 2014a) that the Hessian matrix $\nabla_w^2 J(w)$ is upper bounded by δI_M, where $\delta = \big(\rho + \lambda_{\max}(R_h)\big)$. We conclude from Lemma E.3 in the same reference that $\nabla_w J(w)$ is Lipschitz continuous with modulus δ, i.e.,

$$\|\nabla_w J(\boldsymbol{\psi}_{i-1}) - \nabla_w J(\boldsymbol{x}_{i-1})\| \le \delta\|\boldsymbol{\psi}_{i-1}-\boldsymbol{x}_{i-1}\|. \tag{277}$$

Combining these results we get

$$\begin{aligned}\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1})\| &= \Big\|[\widehat{\nabla J}(\boldsymbol{\psi}_{i-1})-\widehat{\nabla J}(\boldsymbol{x}_{i-1})]-[\nabla J(\boldsymbol{\psi}_{i-1})-\nabla J(\boldsymbol{x}_{i-1})]\Big\| \\ &\le \boldsymbol{\eta}_i\|\boldsymbol{\psi}_{i-1}-\boldsymbol{x}_{i-1}\|.\end{aligned} \tag{278}$$

where $\boldsymbol{\eta}_i = \boldsymbol{\eta}_{1,i} + \delta$ is a random variable. Since the $\{\boldsymbol{h}_i\}$ are independent feature vectors and $\boldsymbol{\eta}_i$ is only related to $\boldsymbol{h}_i$, it follows that

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \le \xi_1\|\boldsymbol{\psi}_{i-1}-\boldsymbol{x}_{i-1}\|^2, \tag{279}$$

$$\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \le \xi_2\|\boldsymbol{\psi}_{i-1}-\boldsymbol{x}_{i-1}\|^4, \tag{280}$$

where $\xi_1 = \mathbb{E}\boldsymbol{\eta}_i{}^2$ and $\xi_2 = \mathbb{E}\boldsymbol{\eta}_i{}^4$.

Next we check the feasibility of Assumption 6. For simplicity, we write $\boldsymbol{\gamma}$ instead of $\boldsymbol{\gamma}(i)$. It can be verified that for the cost function $J(w)$ in (272):

$$\nabla_w^2 J(w) = \rho I_M + \mathbb{E}\Big\{\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\Big(\frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w)]^2}\Big)\Big\}. \tag{281}$$

Now, for any two variables w_1 and w_2 we have

$$\begin{aligned}&\|\nabla_w^2 J(w_1) - \nabla_w^2 J(w_2)\| \\ &\quad= \left\|\mathbb{E}\Big\{\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\Big(\frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1)]^2} - \frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2)]^2}\Big)\Big\}\right\| \\ &\quad\le \mathbb{E}\left\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\Big(\frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1)]^2} - \frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2)]^2}\Big)\right\| \\ &\quad\le \mathbb{E}\left\{\left\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\right\|\left\|\frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1)]^2} - \frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2)]^2}\right\|\right\}\end{aligned} \tag{282}$$

Let $\boldsymbol{x}_1 = -\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1$ and $\boldsymbol{x}_2 = -\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2$. Then,

$$\left\|\frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_1)]^2} - \frac{\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2)}{[1+\exp(-\boldsymbol{\gamma}\boldsymbol{h}_i^{\mathsf{T}}w_2)]^2}\right\|$$

$$
\begin{aligned}
&= \left\| \frac{\exp(\boldsymbol{x}_1)}{[1+\exp(\boldsymbol{x}_1)]^2} - \frac{\exp(\boldsymbol{x}_2)}{[1+\exp(\boldsymbol{x}_2)]^2} \right\| \\
&= \left\| \frac{\exp(\boldsymbol{x}_1)[1+\exp(\boldsymbol{x}_2)]^2 - \exp(\boldsymbol{x}_2)[1+\exp(\boldsymbol{x}_1)]^2}{[1+\exp(\boldsymbol{x}_1)]^2[1+\exp(\boldsymbol{x}_2)]^2} \right\| \\
&\overset{(a)}{\le} \left\| \frac{\exp(-\boldsymbol{x}_2) - \exp(-\boldsymbol{x}_1) + \exp(\boldsymbol{x}_2) - \exp(\boldsymbol{x}_1)}{2\big(\exp(-\boldsymbol{x}_2) + \exp(-\boldsymbol{x}_1) + \exp(\boldsymbol{x}_2) + \exp(\boldsymbol{x}_1)\big)} \right\| \\
&\le \left\| \frac{\exp(-\boldsymbol{x}_2) - \exp(-\boldsymbol{x}_1)}{2\big(\exp(-\boldsymbol{x}_2) + \exp(-\boldsymbol{x}_1) + \exp(\boldsymbol{x}_2) + \exp(\boldsymbol{x}_1)\big)} \right\| + \left\| \frac{\exp(\boldsymbol{x}_2) - \exp(\boldsymbol{x}_1)}{2\big(\exp(-\boldsymbol{x}_2) + \exp(-\boldsymbol{x}_1) + \exp(\boldsymbol{x}_2) + \exp(\boldsymbol{x}_1)\big)} \right\| \\
&\le \left\| \frac{\exp(-\boldsymbol{x}_2) - \exp(-\boldsymbol{x}_1)}{2\big(\exp(-\boldsymbol{x}_2) + \exp(-\boldsymbol{x}_1)\big)} \right\| + \left\| \frac{\exp(\boldsymbol{x}_2) - \exp(\boldsymbol{x}_1)}{2\big(\exp(\boldsymbol{x}_2) + \exp(\boldsymbol{x}_1)\big)} \right\| \\
&\overset{(b)}{=} \frac{1}{2} \left\| \frac{\exp(-\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{2}) - \exp(\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{2})}{\exp(-\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{2}) + \exp(\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{2})} \right\| + \frac{1}{2} \left\| \frac{\exp(\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{2}) - \exp(-\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{2})}{\exp(\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{2}) + \exp(-\frac{\boldsymbol{x}_2-\boldsymbol{x}_1}{2})} \right\| \\
&= |\tanh[(\boldsymbol{x}_2 - \boldsymbol{x}_1)/2]|,
\end{aligned} \tag{283}
$$

where (a) holds because of the following two facts:

$$
\begin{aligned}
&\exp(\boldsymbol{x}_1)[1+\exp(\boldsymbol{x}_2)]^2 - \exp(\boldsymbol{x}_2)[1+\exp(\boldsymbol{x}_1)]^2 \\
&\qquad = \exp(\boldsymbol{x}_1) + \exp(\boldsymbol{x}_1 + 2\boldsymbol{x}_2) - \exp(\boldsymbol{x}_2) - \exp(\boldsymbol{x}_2 + 2\boldsymbol{x}_1) \\
&\qquad = \exp(\boldsymbol{x}_1 + \boldsymbol{x}_2)[\exp(-\boldsymbol{x}_2) + \exp(\boldsymbol{x}_2) - \exp(-\boldsymbol{x}_1) - \exp(\boldsymbol{x}_1)],
\end{aligned} \tag{284}
$$

and

$$
\begin{aligned}
&[1+\exp(\boldsymbol{x}_1)]^2[1+\exp(\boldsymbol{x}_2)]^2 \\
&\qquad = \big(1 + 2\exp(\boldsymbol{x}_1) + \exp(2\boldsymbol{x}_1)\big)\big(1 + 2\exp(\boldsymbol{x}_2) + \exp(2\boldsymbol{x}_2)\big) \\
&\qquad \ge 2\exp(\boldsymbol{x}_1) + 2\exp(\boldsymbol{x}_2) + 2\exp(\boldsymbol{x}_1 + 2\boldsymbol{x}_2) + 2\exp(\boldsymbol{x}_2 + 2\boldsymbol{x}_1) \\
&\qquad = 2\exp(\boldsymbol{x}_1 + \boldsymbol{x}_2)[\exp(-\boldsymbol{x}_2) + \exp(\boldsymbol{x}_2) + \exp(-\boldsymbol{x}_1) + \exp(\boldsymbol{x}_1)].
\end{aligned} \tag{285}
$$

In addition, (b) holds if we extract $\exp(-\frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2})$ and $\exp(\frac{\boldsymbol{x}_1+\boldsymbol{x}_2}{2})$ from both the denominator and numerator of the first and second terms respectively.

Using the definitions for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, this last expression gives

$$
|\tanh[(\boldsymbol{x}_2 - \boldsymbol{x}_1)/2]| = \left|\tanh\left(\frac{1}{2}\gamma \boldsymbol{h}_i^{\mathsf{T}}(w_2 - w_1)\right)\right| \le \frac{1}{2}\|\boldsymbol{h}_i\|\|w_2 - w_1\|. \tag{286}
$$

Substituting (286) into (282), we obtain $\|\nabla_w^2 J(w_1) - \nabla_w^2 J(w_2)\| \le \kappa\|w_1 - w_2\|$, where $\kappa = \mathbb{E}\|\boldsymbol{h}_i\boldsymbol{h}_i^{\mathsf{T}}\|\|\boldsymbol{h}_i\|/2$. Therefore, Assumption 6 holds.

Appendix I. Proof of Lemma 9

Referring to relation (68) and apply the inequality (194), we reach

$$
\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^4 | \boldsymbol{\mathcal{F}}_{i-1}] \\
&= \|(\boldsymbol{I}_M - \mu\boldsymbol{H}_{i-1})(\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}) + \mu(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1} + \mu\beta'\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1}\|^4
\end{aligned}
$$

$$
\begin{aligned}
&+ 3\mu^4\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] + 8\mu^2\|(I_M - \mu\boldsymbol{H}_{i-1})(\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}) \\
&+ \mu(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1} + \mu\beta'\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1}\|^2\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \\
\overset{(a)}{\le}& \frac{1}{(1-t)^3}\|(I_M - \mu\boldsymbol{H}_{i-1})(\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1})\|^4 + \frac{8\mu^4}{t^3}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^4 + \frac{8\mu^4\beta'^4}{t^3}\|\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1}\|^4 \\
&+ 3\mu^4\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] + 8\mu^2\left(\frac{1}{1-t}\|(I_M - \mu\boldsymbol{H}_{i-1})(\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1})\|^2\right. \\
&\left. + \frac{2\mu^2}{t}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^2 + \frac{2\mu^2\beta'^2}{t}\|\boldsymbol{H}_{i-1}\check{\boldsymbol{w}}_{i-1}\|^2\right)\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \\
\overset{(b)}{\le}& (1 - \mu\nu)\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^4 + \frac{8\mu}{\nu^3}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^4 + \frac{8\mu\beta'^4\delta^4}{\nu^3}\|\check{\boldsymbol{w}}_{i-1}\|^4 \\
&+ 3\mu^4\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] + 8\mu^2\Big((1 - \mu\nu)\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 \\
&+ \frac{2\mu}{\nu}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^2 + \frac{2\mu\beta'^2\delta^2}{\nu}\|\check{\boldsymbol{w}}_{i-1}\|^2\Big)\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1}) - \boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}], \quad (287)
\end{aligned}
$$

where (a) holds because the facts that for any $a, b, c \in \mathbb{R}^m$,

$$
\begin{aligned}
\|a + b + c\|^4 &= \|(1-t)\frac{1}{1-t}a + t\frac{1}{t}(b + c)\|^4 \\
&\le (1-t)\|\frac{1}{1-t}a\|^4 + t\|\frac{1}{t}(b+c)\|^4 = \frac{1}{(1-t)^3}\|a\|^4 + \frac{1}{t^3}\|b + c\|^4 \\
&\le \frac{1}{(1-t)^3}\|a\|^4 + \frac{8}{t^3}\|b\|^4 + \frac{8}{t^3}\|c\|^4, \quad (288)
\end{aligned}
$$

and

$$
\begin{aligned}
\|a + b + c\|^2 &= \|(1-t)\frac{1}{1-t}a + t\frac{1}{t}(b + c)\|^2 \\
&\le (1-t)\|\frac{1}{1-t}a\|^2 + t\|\frac{1}{t}(b+c)\|^2 = \frac{1}{1-t}\|a\|^2 + \frac{1}{t}\|b + c\|^2 \\
&\le \frac{1}{1-t}\|a\|^2 + \frac{2}{t}\|b\|^2 + \frac{2}{t}\|c\|^2, \quad (289)
\end{aligned}
$$

In addition, (b) holds by choosing $t = \mu\nu$.

To further simplify inequality (287), we first note that

$$
\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^2 \le 2(\|\boldsymbol{R}_{i-1}\|^2 + \|\boldsymbol{H}_{i-1}\|^2) \le 4\delta^2, \quad (290)
$$

$$
\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4 \le 8(\|\boldsymbol{R}_{i-1}\|^4 + \|\boldsymbol{H}_{i-1}\|^4) \le 16\delta^4. \quad (291)
$$

As a result, we have

$$
\frac{8\mu}{\nu^3}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^4 \le a_1\mu\|\widetilde{\boldsymbol{x}}_{i-1}\|^4, \quad \frac{2\mu}{\nu}\|(\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1})\widetilde{\boldsymbol{x}}_{i-1}\|^2 \le a_2\mu\|\widetilde{\boldsymbol{x}}_{i-1}\|^2, \quad (292)
$$

where we define

$$
a_1 \triangleq 128\delta^4/\nu^3, \quad a_2 \triangleq 8\delta^2/\nu. \quad (293)
$$

On the other hand, from conditions (72)–(73), we have

$$3\mu^4\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \le 3\xi_2\mu^4\|\widetilde{\boldsymbol{\psi}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4, \tag{294}$$
$$8\mu^2\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \le 8\xi_1\mu^2\|\widetilde{\boldsymbol{\psi}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2, \tag{295}$$

In addition, from (252) we get

$$\|\widetilde{\boldsymbol{\psi}}_i-\widetilde{\boldsymbol{x}}_i\|^2 \le 2\|\widehat{\boldsymbol{w}}_i-\widetilde{\boldsymbol{x}}_i\|^2+2\beta'^2\|\check{\boldsymbol{w}}_i\|^2, \quad \|\widetilde{\boldsymbol{\psi}}_i-\widetilde{\boldsymbol{x}}_i\|^4 \le 8\|\widehat{\boldsymbol{w}}_i-\widetilde{\boldsymbol{x}}_i\|^4+8\beta'^4\|\check{\boldsymbol{w}}_i\|^4. \tag{296}$$

Combining (294)–(296), we have

$$3\mu^4\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^4|\boldsymbol{\mathcal{F}}_{i-1}] \;\le\; 24\xi_2\mu^4\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+24\xi_2\mu^4\|\check{\boldsymbol{w}}_{i-1}\|^4, \tag{297}$$
$$8\mu^2\mathbb{E}[\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}] \;\le\; 16\xi_1\mu^2\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2+16\xi_1\mu^2\|\check{\boldsymbol{w}}_{i-1}\|^2, \tag{298}$$

In this way, relation (287) becomes

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{w}}_i-\widetilde{\boldsymbol{x}}_i\|^4|\boldsymbol{\mathcal{F}}_{i-1}]\\
&\le (1-\mu\nu)\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_1\mu\|\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_3\mu\|\check{\boldsymbol{w}}_{i-1}\|^4+\big[(1-\mu\nu)\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2\\
&\quad+a_4\mu\|\check{\boldsymbol{w}}_{i-1}\|^2+a_2\mu\|\widetilde{\boldsymbol{x}}_{i-1}\|^2\big]\big[a_5\mu^2\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2+a_5\mu^2\|\check{\boldsymbol{w}}_{i-1}\|^2\big]\\
&\quad+a_6\mu^4\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_6\mu^4\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&\le (1-\mu\nu)\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_1\mu\|\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_3\mu\|\check{\boldsymbol{w}}_{i-1}\|^4+a_5(1-\mu\nu)\mu^2\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4\\
&\quad+a_5\mu^2(1-\mu\nu+a_4\mu)\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2\|\check{\boldsymbol{w}}_{i-1}\|^2+a_4a_5\mu^3\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&\quad+a_2a_5\mu^3\|\widetilde{\boldsymbol{x}}_{i-1}\|^2\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2+a_2a_5\mu^3\|\widetilde{\boldsymbol{x}}_{i-1}\|^2\|\check{\boldsymbol{w}}_{i-1}\|^2\\
&\quad+a_6\mu^4\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_6\mu^4\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&\overset{(a)}{\le} (1-\mu\nu)\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_1\mu\|\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_3\mu\|\check{\boldsymbol{w}}_{i-1}\|^4+a_5(1-\mu\nu)\mu^2\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4\\
&\quad+a_5\mu^2(1-\mu\nu+a_4\mu)\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_5\mu^2(1-\mu\nu+a_4\mu)\|\check{\boldsymbol{w}}_{i-1}\|^4+a_4a_5\mu^3\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&\quad+a_2a_5\mu^3\|\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_2a_5\mu^3\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_2a_5\mu^3\|\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_2a_5\mu^3\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&\quad+a_6\mu^4\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+a_6\mu^4\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&\le (1-\frac{\mu\nu}{2})\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+2a_1\mu\|\widetilde{\boldsymbol{x}}_{i-1}\|^4+2a_3\mu\|\check{\boldsymbol{w}}_{i-1}\|^4.
\end{aligned} \tag{299}$$

where we define

$$a_3 \;\triangleq\; \frac{8\beta'^4\delta^4}{\nu^4}, \quad a_4 \;\triangleq\; \frac{2\beta'^2\delta^2}{\nu}, \quad a_5 \;\triangleq\; 16\xi_1, \quad a_6 \;\triangleq\; 24\xi_2. \tag{300}$$

Taking expectations over $\boldsymbol{\mathcal{F}}_{i-1}$ for both sides of (299), we have

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{w}}_i-\widetilde{\boldsymbol{x}}_i\|^4 &\le (1-\frac{\mu\nu}{2})\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+2a_1\mu\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4+2a_3\mu\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4\\
&= \rho_1\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^4+2a_1\mu\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4+2a_3\mu\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^4.
\end{aligned} \tag{301}$$

Now recall from (20) that

$$\mathbb{E}\|\widetilde{\boldsymbol{x}}_i\|^4 \le \rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{x}}_{-1}\|^4+A_1\sigma_s^2(i+1)\rho^{i+1}\mu^2+\frac{A_2\sigma_s^4\mu^2}{\nu^2}, \tag{302}$$

where $\rho = 1-\mu\nu$ and A_1 and A_2 are some constants. On the other hand, recall from (44) that

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4 \le \frac{B_1\gamma^2\rho_2^{i+1}}{1-\beta}\mu^2 + \frac{B_1\sigma_s^2(\delta^2+\gamma^2)(i+1)\rho_2^{i+1}}{(1-\beta)^3}\mu^4 + \frac{B_1[(\gamma^2+\nu^2)\sigma_s^4+\sigma_{s,4}^4\nu^2]}{(1-\beta)^2\nu^2}\mu^4, \tag{303}$$

where $\rho_2 = 1-\mu\nu/4$. Besides, we denote $\rho_1 = 1-\mu\nu/2$ and clearly $\rho < \rho_1 < \rho_2$. Substituting (302) and (303) into (301) we reach

$$\begin{aligned}\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^4 &\le \rho_1\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^4 + a_7\rho^i\mu + a_8 i\rho^i\mu^3 + a_9\mu^3 + a_{10}\rho_2^i\mu^3 + a_{11}i\rho_2^i\mu^5 + a_{12}\mu^5\\ &\overset{(a)}{\le} \rho_1\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^4 + a_7\rho_1^i\mu + a_8 i\rho_2^i\mu^3 + a_{10}\rho_2^i\mu^3 + a_{11}i\rho_2^i\mu^5 + 2a_9\mu^3,\end{aligned} \tag{304}$$

where the constants are defined as

$$\begin{aligned} a_7 &\triangleq 2a_1\mathbb{E}\|\widetilde{\boldsymbol{x}}_{-1}\|^4, \quad a_8 \triangleq 2A_1a_1\sigma_s^2, \quad a_9 \triangleq \frac{2A_2a_1\sigma_s^4}{\nu^2}, \quad a_{10} \triangleq \frac{2B_1a_3\gamma^2}{1-\beta}\\ a_{11} &\triangleq \frac{2B_1a_3\sigma_s^2(\delta^2+\gamma^2)}{(1-\beta)^3}, \quad a_{12} \triangleq \frac{2B_1a_3[(\gamma^2+\nu^2)\sigma_s^4+\sigma_{s,4}^4\nu^2]}{(1-\beta)^2\nu^2}. \end{aligned} \tag{305}$$

The inequality (a) holds because μ is chosen small enough such that $a_9\mu^3 > a_{12}\mu^5$. Now we continue iterating (304) and get

$$\begin{aligned}&\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^4\\ &\le \rho_1^{i+1}\mathbb{E}\|\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1}\|^4 + a_7(i+1)\rho_1^i\mu + a_8\rho_2^i\mu^3\sum_{k=0}^{i}(i-k)\left(\frac{\rho_1}{\rho_2}\right)^k\\ &\quad + a_{10}\rho_2^i\mu^3\sum_{k=0}^{i}\left(\frac{\rho_1}{\rho_2}\right)^k + a_{11}\rho_2^i\mu^5\sum_{k=0}^{i}(i-k)\left(\frac{\rho_1}{\rho_2}\right)^k + \frac{2a_9\mu^3}{1-\rho_2}.\end{aligned} \tag{306}$$

Note that $\rho_1 < \rho_2$, we then have

$$\sum_{k=0}^{i}\left(\frac{\rho_1}{\rho_2}\right)^k \le \sum_{k=0}^{\infty}\left(\frac{\rho_1}{\rho_2}\right)^k \le \frac{\rho_2}{\rho_2-\rho_1} = \frac{4-\mu\nu}{\mu\nu} \le \frac{B_2}{\mu\nu}, \tag{307}$$

where B_2 is some constant. Meanwhile, it also holds that

$$\sum_{k=0}^{i}(i-k)\left(\frac{\rho_1}{\rho_2}\right)^k \le i\sum_{k=0}^{i}\left(\frac{\rho_1}{\rho_2}\right)^k \le \frac{iB_2}{\mu\nu}. \tag{308}$$

Substituting (307) and (308) into (306), we get

$$\begin{aligned}\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^4 \le \rho_1^{i+1}\mathbb{E}\|\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1}\|^4 + a_7(i+1)\rho_1^i\mu + \frac{a_8B_2i\rho_2^i\mu^2}{\nu}\\ + \frac{a_{10}B_2\rho_2^i\mu^2}{\nu} + \frac{a_{11}i\rho_2^i\mu^4}{\nu} + \frac{4a_9\mu^2}{\nu}.\end{aligned} \tag{309}$$

Recall from (263) that $\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1} = \mu \nabla_w Q(\boldsymbol{w}_{-2}; \boldsymbol{\theta}_{-1})$. Then it holds that

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1}\|^4 = B_3 \mu^4, \tag{310}$$

where $B_3 = \mathbb{E}\|\nabla_w Q(\boldsymbol{w}_{-2}; \boldsymbol{\theta}_{-1})\|^4$. With this fact, expressions (309) becomes

$$\begin{aligned} &\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^4 \\ &\le B_3 \rho_1^{i+1}\mu^4 + a_7(i+1)\rho_1^i \mu + \frac{a_8 B_2 i \rho_2^i \mu^2}{\nu} + \frac{a_{10} B_2 \rho_2^i \mu^2}{\nu} + \frac{a_{11} i \rho_2^i \mu^4}{\nu} + \frac{4 a_9 \mu^2}{\nu} \\ &\le B_3 \rho_2^{i+1}\mu^4 + a_7(i+1)\rho_2^i \mu + \frac{a_8 B_2 i \rho_2^i \mu^2}{\nu} + \frac{a_{10} B_2 \rho_2^i \mu^2}{\nu} + \frac{a_{11} i \rho_2^i \mu^4}{\nu} + \frac{4 a_9 \mu^2}{\nu}. \end{aligned} \tag{311}$$

Furthermore, recall from (296) that

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{x}}_i\|^4 \le 8\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^4 + 8\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4. \tag{312}$$

and recall the upper bound of $\mathbb{E}\|\check{\boldsymbol{w}}_i\|^4$ in (303). With the definition of all constants we finally reach

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{x}}_i\|^4 = O\left(\frac{\delta^4 i \rho_2^i \mu}{\nu^3} + \frac{\delta^4(\sigma_s^2 + \gamma^2) i \rho_2^i \mu^2}{(1-\beta)\nu^5} + \frac{\delta^4(\delta^2 + \gamma^2)\sigma_s^2 i \rho_2^i \mu^4}{(1-\beta)^3 \nu} + \frac{\delta^4 \sigma_s^4}{\nu^6}\mu^2 \right). \tag{313}$$

To further simplify the notation, we notice that when μ is sufficiently small it holds that

$$\begin{aligned} &\frac{\delta^4 i \rho_2^i \mu}{\nu^3} + \frac{\delta^4(\sigma_s^2 + \gamma^2) i \rho_2^i \mu^2}{(1-\beta)\nu^5} + \frac{\delta^4(\delta^2 + \gamma^2)\sigma_s^2 i \rho_2^i \mu^4}{(1-\beta)^3 \nu} \\ &= i\rho_2^i \mu \left(\frac{\delta^4}{\nu^3} + \frac{\delta^4(\sigma_s^2 + \gamma^2)\mu}{(1-\beta)\nu^5} + \frac{\delta^4(\delta^2 + \gamma^2)\sigma_s^2 \mu^3}{(1-\beta)^3 \nu} \right) \\ &\le \frac{2\delta^4 i \rho_2^i \mu}{\nu^3} \le \frac{B_4 \delta^4 (i+1)\rho_2^{i+1}\mu}{\nu^3} \end{aligned} \tag{314}$$

for some constant B_4. As a result, we obtain

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{x}}_i\|^4 = O\left(\frac{\delta^4 (i+1)\rho_2^{i+1}\mu}{\nu^3} + \frac{\delta^4 \sigma_s^4}{\nu^6}\mu^2 \right). \tag{315}$$

Appendix J. Proof of Lemma 10

Under Assumption 6, we have

$$\begin{aligned} &\|\boldsymbol{H}_{i-1} - \boldsymbol{R}_{i-1}\| \\ &= \left\| \int_0^1 \nabla_w^2 J(w^o - r\widetilde{\boldsymbol{\psi}}_{i-1}) dr - \int_0^1 \nabla_w^2 J(w^o - r\widetilde{\boldsymbol{x}}_{i-1}) dr \right\| \\ &\le \int_0^1 \|\nabla_w^2 J(w^o - r\widetilde{\boldsymbol{\psi}}_{i-1}) - \nabla_w^2 J(w^o - r\widetilde{\boldsymbol{x}}_{i-1})\| dr \\ &\le \int_0^1 \kappa r \|\widetilde{\boldsymbol{\psi}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\| dr = \frac{\kappa}{2}\|\widetilde{\boldsymbol{\psi}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|. \end{aligned} \tag{316}$$

As a result, it holds that

$$\mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4 \le \bar{\kappa}\, \mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^4, \tag{317}$$

where $\bar{\kappa} = \kappa^4/16$. Using (75), we reach the desired bounds shown in (77).

Appendix K. Proof of Theorem 11

For (72) in Assumption 5, if we take expectation over $\boldsymbol{\mathcal{F}}_{i-1}$ of both sides, it holds that

$$\mathbb{E}\|\boldsymbol{s}_i(\boldsymbol{\psi}_{i-1})-\boldsymbol{s}_i(\boldsymbol{x}_{i-1})\|^2 \le \xi_1\mathbb{E}\|\boldsymbol{\psi}_{i-1}-\boldsymbol{x}_{i-1}\|^2. \tag{318}$$

Combining the above fact and inequalities (70)–(71), we get

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \\
&\quad\le (1-\mu\nu)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + r_1\mu\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 \\
&\qquad + r_2\mu\sqrt{\mathbb{E}\|\boldsymbol{R}_{i-1} - \boldsymbol{H}_{i-1}\|^4\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4} + \xi_1\mu^2\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2,
\end{aligned} \tag{319}$$

where the constants are defined as

$$r_1 \;\triangleq\; \frac{2\beta'^2\delta^2}{\nu}, \quad r_2 \;\triangleq\; \frac{2}{\nu}. \tag{320}$$

Likewise, from (253) we have

$$\mathbb{E}\|\widetilde{\boldsymbol{\psi}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le 2\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 + 2\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2. \tag{321}$$

Substituting the above inequality along with (77) into (319) gives

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \\
&\le (1-\mu\nu+2\xi_1\mu^2)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2+(r_1\mu+2\xi_1\mu^2)\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + \left[\sqrt{i}r_3\rho_2^{i/2}\mu^{3/2} + r_4\mu^2\right]\sqrt{\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4} \\
&\le \left(1-\frac{\mu\nu}{2}\right)\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2+2r_1\mu\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + \left[r_3\sqrt{i}\rho_2^{i/2}\mu^{3/2} + r_4\mu^2\right]\sqrt{\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4} \\
&= \rho_1\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1}-\widetilde{\boldsymbol{x}}_{i-1}\|^2+2r_1\mu\mathbb{E}\|\check{\boldsymbol{w}}_{i-1}\|^2 + \left[r_3\sqrt{i}\rho_2^{i/2}\mu^{3/2} + r_4\mu^2\right]\sqrt{\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4}
\end{aligned} \tag{322}$$

where the constants are defined as

$$r_3 \;\triangleq\; \frac{r_2\delta^2}{\nu^{3/2}}, \quad r_4 \;\triangleq\; \frac{r_2\delta^2\sigma_s^2}{\nu^3}. \tag{323}$$

Recall the upper bound of $\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2$ in (38) that

$$\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2 \le \frac{C_2(\delta^2+\gamma^2)\rho_1^i\mu^2}{(1-\beta)^2} + \frac{C_1\sigma_s^2\mu^2}{1-\beta} \tag{324}$$

where $\alpha = 1-\epsilon/2 < \rho_1$, and C_1 and C_2 are some constants. Substituting (324) into (322), we have

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \\
&\quad\le \rho_1\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + (r_5\rho_1^i\mu^3 + r_6\mu^3) + \left[2r_3\sqrt{i}\rho_2^{i/2}\mu^{3/2} + r_6\mu^2\right]\sqrt{\mathbb{E}\|\widetilde{\boldsymbol{x}}_{i-1}\|^4},
\end{aligned} \tag{325}$$

where the constants are defined as

$$r_5 \;\triangleq\; \frac{2C_2r_1(\delta^2+\gamma^2)}{(1-\beta)^2}, \quad r_6 \;\triangleq\; \frac{2C_1r_1\sigma_s^2}{1-\beta}. \tag{326}$$

Next, using (20) we have

$$\sqrt{\mathbb{E}\|\widetilde{\boldsymbol{x}}_i\|^4} \le \sqrt{\rho^{i+1}\mathbb{E}\|\widetilde{\boldsymbol{x}}_{-1}\|^4 + A_3\sigma_s^2(i+1)\rho^{i+1}\mu^2 + \frac{A_2\sigma_s^4\mu^2}{\nu^2}}$$
$$\le C_3\rho^{(i+1)/2} + C_4\sigma_s\sqrt{i+1}\rho^{(i+1)/2}\mu + C_5\frac{\sigma_s^2\mu}{\nu}$$
$$\le C_3\rho_2^{(i+1)/2} + C_4\sigma_s\sqrt{i+1}\rho_2^{(i+1)/2}\mu + C_5\frac{\sigma_s^2\mu}{\nu}. \quad (327)$$

Substituting (327) into (325), we reach

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \\
&\le \rho_1\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + (r_5\rho_1^i\mu^3 + r_6\mu^3) + r_7\sqrt{i}\rho_2^i\mu^{3/2} + r_8 i\rho_2^i\mu^{5/2} + r_9\sqrt{i}\rho_2^{i/2}\mu^{5/2} \\
&\quad + r_{10}\mu^2\rho_2^{i/2} + r_{11}\sqrt{i}\rho_2^{i/2}\mu^3 + r_{12}\mu^3, \qquad (328)
\end{aligned}$$

where the constants are defined as

$$r_7 \triangleq 2C_3r_3, \quad r_8 \triangleq 2C_4r_3\sigma_s, \quad r_9 \triangleq \frac{2C_5r_3\sigma_s^2}{\nu}$$
$$r_{10} \triangleq C_3r_6, \quad r_{11} \triangleq C_4r_6\sigma_s, \quad r_{12} \triangleq \frac{C_5r_6\sigma_s^2}{\nu}. \quad (329)$$

Now we denote

$$\tau_1 \triangleq \rho_1^{1/2}, \quad \tau_2 \triangleq \rho_2^{1/2}. \quad (330)$$

Clearly, we have

$$\rho_1 < \tau_1, \quad \rho_2 < \tau_2, \quad \tau_1 < \tau_2. \quad (331)$$

With the above relation, expressions (328) becomes

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \\
&\le \tau_2\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + (r_5\tau_2^i\mu^3 + r_6\mu^3) + r_7\sqrt{i}\tau_2^i\mu^{3/2} + r_8 i\tau_2^i\mu^{5/2} + r_9\sqrt{i}\tau_2^i\mu^{5/2} \\
&\quad + r_{10}\mu^2\tau_2^i + r_{11}\sqrt{i}\tau_2^i\mu^3 + r_{12}\mu^3 \\
&\le \tau_2\mathbb{E}\|\widehat{\boldsymbol{w}}_{i-1} - \widetilde{\boldsymbol{x}}_{i-1}\|^2 + 2r_{10}\tau_2^i\mu^2 + 2r_7\sqrt{i}\tau_2^i\mu^{3/2} + r_8 i\tau_2^i\mu^{5/2} + (r_6 + r_{12})\mu^3 \\
&\le \tau_2^{i+1}\mathbb{E}\|\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1}\|^2 + 2r_{10}(i+1)\tau_2^i\mu^2 + 2r_7\tau_2^i\mu^{3/2}\left(\sum_{k=0}^{i}\sqrt{i-k}\right) \\
&\quad + r_8\tau_2^i\mu^{5/2}i(i+1) + \frac{(r_6 + r_{12})\mu^3}{1-\tau_2}. \qquad (332)
\end{aligned}$$

Note that $\tau_2 = \sqrt{\rho_2} = \sqrt{1-\mu\nu/4}$. When μ is sufficiently small, we have $\tau_2 = 1 - \mu\nu/8$ and hence $1 - \tau_2 = \mu\nu/2$. With this fact and recall that $\mathbb{E}\|\widehat{\boldsymbol{w}}_{-1} - \widetilde{\boldsymbol{x}}_{-1}\|^2 = C_6\mu^2$, finally we can show that

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \le C_6\tau_2^{i+1}\mu^2 + 2r_{10}(i+1)\tau_2^i\mu^2 + 2r_7\tau_2^i\mu^{3/2}\left(\sum_{k=0}^{i}\sqrt{i-k}\right)$$

$$+ r_8\tau_2^i\mu^{5/2}i(i+1) + \frac{8(r_6+r_{12})\mu^2}{\nu}. \tag{333}$$

Substituting the definitions of all constants, we get

$$\begin{aligned}
&\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \\
&\leq C_7\left(\frac{\delta^2 s_1(i)\tau_2^i\mu^{3/2}}{\nu^{5/2}} + \tau_2^{i+1}\mu^2 + \frac{\sigma_s^2\delta^2(i+1)\tau_2^i\mu^2}{(1-\beta)\nu} + \frac{\sigma_s\delta^2 s_2(i)\tau_2^i\mu^{5/2}}{\nu^{5/2}} + \frac{\delta^2\sigma_s^4\mu^2}{(1-\beta)\nu^2}\right) \\
&\leq C_8\left(\frac{\delta^2\sigma_s^2 s_2(i)\tau_2^{i+1}\mu^{3/2}}{(1-\beta)\nu^{5/2}} + \frac{\delta^2\sigma_s^4\mu^2}{(1-\beta)\nu^2}\right).
\end{aligned} \tag{334}$$

where

$$s_1(i) \triangleq \sum_{k=0}^{i}\sqrt{i-k}, \quad s_2(i) \triangleq i(i+1) \tag{335}$$

Furthermore, it holds that

$$\mathbb{E}\|\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 \leq 2\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 + 2\beta^2\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2. \tag{336}$$

Using the upper bound for $\mathbb{E}\|\check{\boldsymbol{w}}_i\|^2$ in (324), we then have

$$\mathbb{E}\|\widehat{\boldsymbol{w}}_i - \widetilde{\boldsymbol{x}}_i\|^2 = O\left(\frac{\delta^2\sigma_s^2 s_2(i)\tau_2^{i+1}\mu^{3/2}}{(1-\beta)\nu^{5/2}} + \frac{(\delta^2+\gamma^2)\rho_1^{i+1}\mu^2}{(1-\beta)^2} + \frac{\delta^2\sigma_s^4\mu^2}{(1-\beta)\nu^2}\right). \tag{337}$$

References

N. O. Attoh-Okine. Analysis of learning rate and momentum term in backpropagation neural network algorithm trained to predict pavement performance. *Advances in Engineering Software*, 30(4):291–302, 1999.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

M. Bellanger. *Adaptive Digital Filters and Signal Analysis.* 2nd Edition, Marcel Dekker, 2001.

D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, 1999.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proc. International Conference on Computational Statistics*, pages 177–186. Springer, Paris, France, 2010.

O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Proc. Advances in Neural Information Processing Systems*, pages 161–168, Vancouver, Canada, 2008.

S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, 2004.

V. Cevher, S. Becker, and M. Schmidt. Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics. *IEEE Signal Processing Magazine*, 31(5):32–43, 2014.

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

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Proc. Advances in Neural Information Processing Systems*, pages 1646–1654, Montreal, Canada, 2014.

O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *arXiv: 1602.05419*, Feb. 2016.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(2):2121–2159, 2011.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. *Journal of Machine Learning Research*, 40(1):1–38, 2015.

R. Gemulla, E. Nijkamp, P. J. Haas, and Y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proc. International Conference on Knowledge Discovery and Data Mining*, pages 69–77, Alberta, Canada, 2011.

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

S. Haykin. *Adaptive Filter Theory.* Fourth Edition, Prentice-Hall, NJ, 2008.

C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. In *Proc. Advances in Neural Information Processing Systems*, pages 781–789, Vancouver, Canada, 2009.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proc. Advances in Neural Information Processing Systems* (NIPS), pages 315–323, Lake Tahoe, Navada, 2013.

S. Kahou, C. Pal, X. Bouthillier, P. Froumenty, and et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proc. International Conference on Multimodal Interaction*, pages 543–550, Sydney, Australia, 2013.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

A. Nedić and D. P. Bertsekas. Convergence rate of incremental subgradient algorithms. In S. Uryasey and M. Pardalos P, editors, *Stochastic Optimization: Algorithms and Applications*, volume 54, pages 223–264. Springer, 2001.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer, 2004.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

A. Nitanda. Stochastic proximal gradient descent with acceleration techniques. In *Proc. Advances in Neural Information Processing Systems*, pages 1574–1582, Montreal, Canada, 2014.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

B. T. Polyak. *Introduction to Optimization*. Optimization Software, NY, 1987.

J. G. Proakis. Channel identification for high speed digital communications. *IEEE Transactions on Automatic Control*, 19(6):916–922, 1974.

N. Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999.

N. L. Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Proc. Advances in Neural Information Processing Systems* (NIPS), pages 2663–2671, Lake Tahoe, Navada, 2012.

S. Roy and J. J. Shynk. Analysis of the momentum LMS algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(12):2088–2098, 1990.

A. H. Sayed. *Adaptive Filters*. Wiley, NY, 2008.

A. H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*, 7(4-5):311–801, Jul. 2014a.

A. H. Sayed. Adaptive networks. *Proceedings of the IEEE*, 102(4):460–497, 2014b.

S. Shalev-Shwartz. SDCA without duality. *arXiv:1502.06177*, Feb. 2015.

S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *Proc. International Conference on Machine Learning*, pages 64–72, Beijing, China, 2014.

R. Sharma, W. A. Sethares, and J. A. Bucklew. Analysis of momentum adaptive filtering algorithms. *IEEE Transactions on Signal Processing*, 46(5):1430–1434, 1998.

J. J. Shynk and S. Roy. The LMS algorithm with momentum updating. In *Proc. IEEE International Symposium on Circuits and Systems*, pages 2651–2654, Espoo, Finland, June 1988.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proc. International Conference on Machine Learning*, pages 1139–1147, Atlanta, USA, 2013.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelo, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, Boston, USA, June 2015.

S. Theodoridis. *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, NY, 2015.

L. K. Ting, C. F. N. Cowan, and R. F. Woods. Tracking performance of momentum LMS algorithm for a chirped sinusoidal signal. In *Proc. European Signal Processing Conference*, pages 1–4, Tampere, Finland, 2000.

M. A. Tugay and Y. Tanik. Properties of the momentum LMS algorithm. *Signal Processing*, 18(2):117–127, 1989.

M. Tygert. Poor starting points in machine learning. *arXiv:1602.02823*, Feb. 2016.

B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice-Hall, NJ, 1985.

W. Wiegerinck, A. Komoda, and T. Heskes. Stochastic dynamics of learning with momentum in neural networks. *Journal of Physics A: Mathematical and General*, 27(13): 4425–4438, 1994.

L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

B. Ying and A. H. Sayed. Performance limits of online stochastic sub-gradient learning. *arXiv:1511.07902*, Oct. 2015.

B. Ying and A. H. Sayed. Performance limits of single-agent and multi-agent sub-gradient stochastic learning. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 4905–4909, Shanghai, China, March 2016.

K. Yuan, B. Ying, and A. H. Sayed. On the influence of momentum acceleration on online learning. In *Proc. International Conference on Acoustics, Speech and Signal Processing*, pages 4915–4919, Shanghai, China, March 2016.

S. Zareba, A. Gonczarek, J. M. Tomczak, and J. Światek. Accelerated learning for restricted Boltzmann machine with momentum term. In *Proc. International Conference on Systems Engineering*, pages 187–192, Coventry, UK, 2015.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proc. International Conference on Machine Learning*, page 116, Alberta, Canada, 2004.

X. Zhang and Y. LeCun. Text understanding from scratch. *arXiv:1502.01710*, Feb. 2015.

W. Zhong and J. T. Kwok. Accelerated stochastic gradient method for composite regularization. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 1086–1094, Reykjavik, Iceland, 2014.

Z. Zhu. Katyusha: Accelerated variance reduction for faster SGD. *arXiv:1603.05953*, Mar. 2016.