On the Influence of Momentum Acceleration on Online Learning

Kun Yuan
Bicheng Ying
Ali H. Sayed
Department of Electrical Engineering
University of California
Los Angeles, CA 90095, USA

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 case 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 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 stochastic (online) setting when adaptation becomes necessary and when the true gradient vectors are not known beforehand. The analysis also suggests a method to retain some of the advantages of the momentum construction by employing a decaying momentum parameter, as opposed to a decaying step-size. In this way, the enhanced convergence rate during the initial stages of adaptation is preserved without the often-observed degradation in MSD performance.

Keywords: online learning, stochastic gradient, momentum acceleration, heavy-ball method, Nesterov 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

$$\min_{w \in \mathbb{R}^M} J(w) \triangleq \mathbb{E}_{\theta}[Q(w; \theta)],$$

where $\theta$ 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, $\theta$, 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, $Q(w; \theta)$, may be available at various observations $\theta_i$, where $i$ refers
to the sample index. We refer to these types of problems, where $J(w)$ is unknown but defined implicitly 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 $J$ 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, $\nabla_w Q(w; \theta_i)$, evaluated at successive samples of the streaming data $\theta_i$ over the iteration index $i$, say, as:

$$w_i = w_{i-1} - \mu \nabla_w Q(w_{i-1}; \theta_i), \quad i \geq 0.$$  \hspace{1cm} (2)

where $\mu > 0$ is a step-size parameter. Note that we are denoting the successive iterates by $w_i$ and using the boldface notation to refer to the fact that they are random quantities in view of the randomness in the measurements $\{\theta_i\}$. 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 Rates

Stochastic-gradient algorithms can be implemented with decaying step-sizes, such as $\mu(i) = \tau/i$ for some constant $\tau$, or with constant step-sizes, $\mu > 0$. The former generally ensure asymptotic convergence to the true minimizer of (1), denoted by $w^\phi$, at a convergence rate that is on the order of $O(1/i)$. 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 algorithms lose 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 of the true minimizer, $w^\phi$, whose radius is on the order of $O(\mu)$. This is still a desirable conclusion because the value of $\mu$ is controlled by the designer and can be chosen sufficiently small.

In other words, a well-known tradeoff 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. Therefore, 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 constant step-size implementations.

1.2 Acceleration Methods

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

$$w_i = w_{i-1} - \mu \nabla_w J(w_{i-1}), \quad i \geq 0,$$  \hspace{1cm} (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 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 $\nu$-strongly convex and has $\delta$-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 cannot be attained by standard gradient descent method. 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 $w^*$ occurs at the rate:

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

In contrast, 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

$$\|w_i - w^*\|^2 \leq \left(\frac{\delta - \nu}{\delta + \nu}\right)^2 \|w_{i-1} - w^*\|^2.$$  \hspace{1cm} (5)

It can be verified that

$$\frac{\sqrt{\delta} - \sqrt{\nu}}{\sqrt{\delta} + \sqrt{\nu}} < \frac{\delta - \nu}{\delta + \nu}$$  \hspace{1cm} (6)

when $\delta > \nu$. This inequality confirms that the momentum algorithm can achieve a faster rate in deterministic optimization and that 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, 1998; Bellanger, 2001; Wiegerinck et al., 1994; Hu et al., 2009; 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 their advantages from deterministic optimization do not necessarily carry over to the stochastic setting due to the presence of stochastic gradient noise (which is the difference between the actual gradient vector and its approximation). Specifically, we will show that any advantage they bring forth 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, $\mu_m$ for the momentum (heavy-ball or Nesterov) methods and $\mu$ for the standard stochastic gradient algorithms, are sufficiently small and satisfy the relation

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

where $\beta \in [0, 1)$ is the coefficient of the momentum term, then it will hold that

$$\mathbb{E} \|w_{m,i} - w_i\|^2 = O(\mu^{3/2}), \quad i = 0, 1, 2, \ldots$$

where $w_{m,i}$ and $w_i$ 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

$$\mathbb{E} \|w_{m,i} - w_i\|^2 = O(\mu^2), \quad i = 0, 1, 2, \ldots$$

What is important to note is that, we will show that these results hold for every $i$, and not only asymptotically. Therefore, when $\mu$ 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 theory under the situation where the cost function is convex and differentiable. However, as our numerical simulation over multi-layer neural network (see Sec. 7.4) shows, the equivalence between standard and momentum stochastic gradient methods can still hold in non-convex or 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 detailed 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; Baes, 2009; 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; Baes, 2009; Devolder et al., 2014) 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; 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 to some extent. However, 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 are limited analyses for this scenario.

It is worth noting that all the literature reviewed in this section focuses either on Heavy-ball momentum technique or on Nesterov’s acceleration. In our approach, we shall develop a framework that unifies both techniques and enables uniform performance analysis for momentum acceleration under stochastic noise.

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 framework, 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., \( w^o \triangleq \arg\min_w J(w) \). 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

\[
0 < \nu I_M \leq \nabla^2 J(w) \leq \delta I_M,
\]

for some positive parameters \( \nu \leq \delta \). Condition (10) is equivalent to requiring \( J(w) \) to be \( \nu \)-strongly convex and for its gradient vector to be \( \delta \)-Lipschitz, respectively.

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

\[
s_i(w_{i-1}) \triangleq \nabla Q(w_{i-1}; \theta_i) - \nabla_w E[Q(w_{i-1}; \theta_i)].
\]
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 (12) and (13) 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 $\mathcal{F}_{i-1}$ represent the filtration generated by the random process $w_j$ for $j \leq i - 1$ (basically, the collection of past history until time $i - 1$):

$$\mathcal{F}_{i-1} \triangleq \text{filtration}\{w_{-1}, w_0, w_1, \ldots, w_{i-1}\}.$$  

**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 $w \in \mathcal{F}_{i-1}$:

$$
\begin{align*}
\mathbb{E}[s_i(w)|\mathcal{F}_{i-1}] &= 0 \quad (12) \\
\mathbb{E}[\|s_i(w)\|^2|\mathcal{F}_{i-1}] &\leq \gamma^2\|w^o - w\|^2 + \sigma_s^2 \quad (13)
\end{align*}
$$

almost surely, for some nonnegative constants $\gamma^2$ and $\sigma_s^2$.

Condition (12) 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 (13) requires the size of the gradient noise (i.e., its mean-square value) to diminish as the iterate $w$ 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 $\tilde{w}_i = w^o - w_i$. Then, for sufficiently small step-sizes $\mu$, it holds for each iteration $i = 0, 1, 2, \ldots$ that

$$
\mathbb{E}\|\tilde{w}_i\|^2 \leq (1 - \mu\nu)\mathbb{E}\|\tilde{w}_{i-1}\|^2 + \mu^2\sigma_s^2, \quad (14)
$$

and, furthermore,

$$
\limsup_{i \to \infty} \mathbb{E}\|\tilde{w}_i\|^2 = O(\mu). \quad (15)
$$

The above lemma characterizes the stability of the second-order moment of the error vector, $\mathbb{E}\|\tilde{w}_i\|^2$. One useful conclusion that follows from (14) is that, again for sufficiently small step-sizes,

$$
\mathbb{E}\|\tilde{w}_i\|^2 \leq D, \quad \forall i \geq 0 \quad (16)
$$

for some constant $D$ — see equation (99). We will exploit this fact in the sequel.
We can also examine the stability of the fourth-order error moment, $\mathbb{E}\|\tilde{w}_i\|^4$, 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 (13) 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 $w \in \mathcal{F}_{i-1}$:

$$
\mathbb{E}[s_i(w)|\mathcal{F}_{i-1}] = 0 \quad (17)
$$

$$
\mathbb{E}[\|s_i(w)\|^4|\mathcal{F}_{i-1}] \leq \gamma_4^4\|w^0 - w\|^4 + \sigma_{s,4}^4 \quad (18)
$$

almost surely, for some nonnegative constants $\gamma_{s,4}^4$ 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-sizes, it holds that

$$
\mathbb{E}\|\tilde{w}_i\|^4 \leq (1 - \mu\nu)^{i+1}\mathbb{E}\|\tilde{w}_{i-1}\|^4 + c\mu^2, \quad (19)
$$

for some constant $c$, and, furthermore,

$$
\limsup_{i \to \infty} \mathbb{E}\|\tilde{w}_i\|^4 \leq O(\mu^2). \quad (20)
$$

**Proof** See Appendix A.

## 3. Momentum Acceleration

In this section, we present a generalized momentum stochastic gradient method, which is a unified framework that captures both the heavy-ball and Nesterov’s acceleration methods. 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 $\beta_1, \beta_2 \in [0, 1)$:

$$
\psi_{i-1} = w_{i-1} + \beta_1(w_{i-1} - w_{i-2}), \quad (21)
$$

$$
w_i = \psi_{i-1} - \mu_m \nabla_w Q(\psi_{i-1}; \theta_i) + \beta_2(\psi_{i-1} - \psi_{i-2}), \quad (22)
$$

with initial conditions

$$
w_{-2} = \psi_{-2} = \text{initial states}, \quad (23)
$$
\[ w_1 = w_2 - \mu_m \nabla_w Q(w_2; \theta_1), \]  

where \( \mu_m \) 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

\[
\beta_1 + \beta_2 = \beta, \quad \beta_1 \beta_2 = 0,
\]

for some \( \beta \in (0, 1) \) that is not too close to 1, i.e., there exists small constant \( \epsilon > 0 \) such that

\[
\beta \leq 1 - \epsilon.
\]

### 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 (21)-(22) can be transformed into a first-order recursion by defining extended state vectors. We introduce the transformation matrices:

\[
V = \begin{bmatrix}
I_M & -\beta_1 I_M \\
I_M & -\beta_2 I_M
\end{bmatrix}, \quad V^{-1} = \frac{1}{1 - \beta} \begin{bmatrix}
I_M & -\beta_1 I_M \\
I_M & -\beta_2 I_M
\end{bmatrix}.
\]

Recall \( \tilde{w}_i = w^0 - w_i \) and define the transformed error vectors, each of size \( 2M \times 1 \):

\[
\begin{bmatrix}
\tilde{w}_i \\
\tilde{w}_{i-1}
\end{bmatrix} = V^{-1} \begin{bmatrix}
\tilde{w}_i \\
\tilde{w}_{i-1}
\end{bmatrix} = \frac{1}{1 - \beta} \begin{bmatrix}
\tilde{w}_i - \beta \tilde{w}_{i-1} \\
\tilde{w}_{i-1} - \beta \tilde{w}_{i-2}
\end{bmatrix}.
\]

**Lemma 3 (Extended recursion)** Under Assumption [1] and condition (25), the momentum stochastic gradient recursion (21)-(22) can be transformed into the following extended recursion:

\[
\begin{bmatrix}
\tilde{w}_i \\
\tilde{w}_{i-1}
\end{bmatrix} = \begin{bmatrix}
I_M - \frac{\mu_m \beta_1}{1 - \beta} H_{i-1} & \frac{\mu_m \beta_2}{1 - \beta} H_{i-1} \\
-\frac{\mu_m \beta_1}{1 - \beta} H_{i-1} & \beta I_M + \frac{\mu_m \beta_2}{1 - \beta} H_{i-1}
\end{bmatrix} \begin{bmatrix}
\tilde{w}_{i-1} \\
\tilde{w}_{i-2}
\end{bmatrix} + \frac{\mu_m}{1 - \beta} \begin{bmatrix}
\mathbf{s}_i(\psi_{i-1}) \\
\mathbf{s}_i(\psi_{i-2})
\end{bmatrix},
\]

\[29\]

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 (21)-(22), we will use this same terminology to refer to both the heavy-ball and Nesterov’s acceleration methods.
where $s_i(\psi_{i-1})$ is defined according to (11) and
\[
\beta' \triangleq \beta_1 + \beta_2, \quad (30)
\]
\[
H_{i-1} \triangleq \int_0^1 \nabla w^2 J(w^o - t\tilde{\psi}_{i-1}) dt, \quad (31)
\]
where $\tilde{\psi}_{i-1} = w^o - \psi_{i-1}$.

**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 $\hat{w}_i$ and $\check{w}_i$ and, subsequently, of the error vector $\tilde{w}_i$ 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 (21)–(22) 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 and 2 hold and recall conditions (25) and (26). Consider the momentum stochastic gradient method (21)–(22) and the extended recursion (29). Then, for sufficiently small step-sizes $\mu_m$, it holds that the mean-square values of the transformed error vectors evolve according to the following recursive inequality:
\[
\begin{bmatrix}
\mathbb{E}\|\hat{w}_i\|^2 \\
\mathbb{E}\|\check{w}_i\|^2
\end{bmatrix} \preceq
\begin{bmatrix}
a & b \\
c & d
\end{bmatrix}
\begin{bmatrix}
\mathbb{E}\|\hat{w}_{i-1}\|^2 \\
\mathbb{E}\|\check{w}_{i-1}\|^2
\end{bmatrix} +
\begin{bmatrix}
e \\
f
\end{bmatrix}, \quad (32)
\]
where
\[
\begin{align*}
a &= 1 - \frac{\mu_m}{1-\beta^2} + O(\mu_m^2), & b &= O(\mu_m), \\
d &= \beta + O(\mu_m^2), & c &= O(\mu_m^2), \\
e &= O(\mu_m^2), & f &= O(\mu_m^2)
\end{align*} \quad (33)
\]
and the coefficient matrix appearing in (32) is stable, namely,
\[
\rho \left( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) < 1. \quad (34)
\]
Furthermore, it follows from (32) that
\[
\limsup_{i \to \infty} \mathbb{E}\|\hat{w}_i\|^2 = O(\mu_m) \quad \text{and} \quad \limsup_{i \to \infty} \mathbb{E}\|\check{w}_i\|^2 = O(\mu_m^2), \quad (35)
\]
and, consequently,
\[
\limsup_{i \to \infty} \mathbb{E}\|\tilde{w}_i\|^2 = O(\mu_m). \quad (36)
\]
Proof See Appendix C.

Although $\mathbb{E}\|\tilde{w}_i\|^2 = O(\mu_m^2)$ in result (35) 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

$$\mathbb{E}\|\tilde{w}_i\|^2 = O(\mu_m^2), \quad \forall i = 0, 1, 2, 3, \ldots$$

(37)

where $\tilde{w}_i$ is defined in (29).

Proof See Appendix D.

### 3.3 Stability of Fourth-Order Error Moment

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

**Theorem 6 (Fourth-order stability)** Let Assumptions 7 and 8 hold and recall conditions (25) and (26). Then, for sufficiently small step-sizes $\mu_m$, it holds that

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

(38)

$$\limsup_{i \to \infty} \mathbb{E}\|\tilde{w}_i\|^4 = O(\mu_m^4),$$

(39)

$$\limsup_{i \to \infty} \mathbb{E}\|\hat{w}_i\|^4 = O(\mu_m^2).$$

(40)

Proof The derivation follows by extending the argument from Appendix C in a manner similar to the proof of Theorem 9.2 in (Sayed, 2014a).

Again, result (39) is only shown to hold asymptotically in the statement of the theorem. In fact, $\mathbb{E}\|\tilde{w}_i\|^4$ 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

$$\mathbb{E}\|\tilde{w}_i\|^4 = O(\mu_m^2), \quad \forall i = 0, 1, 2, 3, \ldots$$

(41)

Furthermore, in the limit,

$$\limsup_{i \to \infty} \mathbb{E}\|\tilde{w}_i\|^4 = O(\mu_m^4).$$

(42)

Proof See Appendix E.
4. Equivalence in the Quadratic Case

In Section 3 we showed the momentum stochastic gradient algorithm (21)–(22) 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 $\mathbb{E} \| \tilde{\omega}_i \|^2$? 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

$$ J(w) = \frac{1}{2} \mathbb{E} \left( \langle d(i) - u_i^T w \rangle^2, \right. \quad (43) $$

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

$$ d(i) = u_i^T w^o + v(i), \quad (44) $$

for some unknown $w^o$, and where $v(i)$ is a zero-mean white noise process with power $\sigma_v^2 = \mathbb{E} v^2(i)$ and assumed independent of $u_j$ for all $i, j$. If we multiply (44) by $u_i$ 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 (43) also satisfies these same equations. Therefore, minimizing the quadratic risk (43) also recovers 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 (43), the true gradient vector at any location $w$ is

$$ \nabla_w J(w) = R_u w - r_{du} = -R_u (w^o - w), \quad (45) $$

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

$$ \nabla_w Q(w; d(i), u_i) = -u_i(d(i) - u_i^T w). \quad (46) $$

Here the loss function is defined by

$$ Q(w; d(i), u_i) \triangleq \frac{1}{2} \mathbb{E} \left( d(i) - u_i^T w \right)^2. \quad (47) $$
The resulting LMS (stochastic-gradient) recursion is given by
\[ w_i = w_{i-1} + \mu u_i(d(i) - u_i^T w_{i-1}) \]  
(48)

and the corresponding gradient noise process is
\[ s_i(w) = (R_u - u_iu_i^T)(w^o - w) - u_i\nu(i). \]  
(49)

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 (48), and recalling that \( \tilde{w}_i = w^o - w_i \), we obtain the error recursion that corresponds to the LMS implementation:
\[ \tilde{w}_i = (I_M - \mu R_u)\tilde{w}_{i-1} + \mu s_i(w_{i-1}), \]  
(50)

where \( \mu \) 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 \( \{w_i, \tilde{w}_i\} \) for LMS by \( \{x_i, \tilde{x}_i\} \) and keep the notation \( \{w_i, \tilde{w}_i\} \) for momentum LMS, i.e., for the LMS implementation (50) we shall write instead
\[ \tilde{x}_i = (I_M - \mu R_u)\tilde{x}_{i-1} + \mu s_i(x_{i-1}). \]  
(51)

On the other hand, we conclude from (21)–(22) that the momentum LMS recursion will be given by:
\[
\begin{align*}
\psi_{i-1} &= w_{i-1} + \beta_1(w_{i-1} - w_{i-2}), \\
\tilde{w}_i &= \psi_{i-1} + \mu u_i(d(i) - u_i^T \psi_{i-1}) + \beta_2(\psi_{i-1} - \psi_{i-2}).
\end{align*}
\]  
(52) (53)

Using the transformed recursion (29), we can transform the resulting relation for \( \tilde{w}_i \) into:
\[
\begin{bmatrix}
\tilde{w}_i \\
\tilde{x}_i
\end{bmatrix} = \begin{bmatrix}
I_M - \frac{\mu_m}{1 - \beta} R_u & \frac{\mu_m}{1 - \beta} R_u \\
-\frac{\mu_m}{1 - \beta} R_u & \beta I_M + \frac{\mu_m}{1 - \beta} R_u
\end{bmatrix}
\begin{bmatrix}
\tilde{w}_{i-1} \\
\tilde{x}_{i-1}
\end{bmatrix} + \frac{\mu_m}{1 - \beta}
\begin{bmatrix}
s_i(\psi_{i-1}) \\
s_i(\tilde{x}_{i-1})
\end{bmatrix},
\]  
(54)

where the Hessian matrix, \( H_{i-1} \), is independent of the weight iterates and given by \( R_u \) for quadratic risks. It follows from the first row that
\[
\tilde{w}_i = \left( I_M - \frac{\mu_m}{1 - \beta} R_u \right) \tilde{w}_{i-1} + \frac{\mu_m}{1 - \beta} R_u \tilde{w}_{i-1} + \frac{\mu_m}{1 - \beta} s_i(\psi_{i-1}).
\]  
(55)

Next, we assume the step-sizes \( \{\mu, \mu_m\} \) and the momentum parameter are selected to satisfy
\[
\mu = \frac{\mu_m}{1 - \beta}.
\]  
(56)

Since \( \beta \in [0, 1) \), this means that \( \mu_m < \mu \). Then, recursion (55) becomes
\[
\tilde{w}_i = (I_M - \mu R_u)\tilde{w}_{i-1} + \mu \beta' R_u \tilde{w}_{i-1} + \mu s_i(\psi_{i-1}).
\]  
(57)

Comparing (57) with the LMS recursion (51), we find that both relations are quite similar, except that the momentum recursion has an extra driving term dependent on \( \tilde{w}_{i-1} \). However, recall from (28) that \( \tilde{w}_{i-1} = (\tilde{w}_{i-2} - \tilde{w}_{i-1})/(1 - \beta) \), which is the difference between
two consecutive points generated by momentum LMS. Intuitively, it is not hard to see that \( \hat{w}_{i-1} \) is in the order of \( O(\mu) \), which makes \( \mu \beta \) in the order of \( O(\mu^2) \). When the step-size \( \mu \) is very small, this \( O(\mu^2) \) term can be ignored. Consequently, the above recursions for \( \hat{w}_i \) and \( \tilde{x}_i \) should evolve close to each other, which would imply that \( w_i \) and \( x_i \) will 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 (48) and (52)–(53). Assume both algorithms start from the same initial states, namely, \( \psi_2 = w_2 = x_1 \). Suppose conditions (25) and (26) hold, and that the step-sizes \( \{\mu, \mu_m\} \) satisfy (56). Then, it holds for sufficiently small \( \mu \) that

\[
E\|w_i - x_i\|^2 = O(\mu^2), \quad \forall i = 0, 1, 2, 3, \ldots
\]

**Proof** See Appendix F. 

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

\[
\hat{w}_i = \left( I_M - \frac{\mu_m}{1 - \beta} H_{i-1} \right) \hat{w}_{i-1} + \frac{\mu_m \beta}{1 - \beta} H_{i-1} \hat{w}_{i-1} + \frac{\mu_m}{1 - \beta} s_i(\psi_{i-1}), \quad (59)
\]

where \( H_{i-1} \) 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, \( \tilde{\psi}_{i-1} \), as well. Likewise, for the standard stochastic gradient iteration (2), we obtain that the error recursion in the general case is given by:

\[
\tilde{x}_i = (I_M - \mu R_{i-1}) \tilde{x}_{i-1} + \mu s_i(x_{i-1}), \quad (60)
\]

where we are introducing the matrix

\[
R_{i-1} = \int_0^1 \nabla^2_w J(w^o - r \tilde{x}_{i-1}) dr
\]

and \( \tilde{x}_i = w^o - x_i \). Note that \( H_{i-1} \) and \( R_{i-1} \) are different matrices. In contrast, in the quadratic case, they are both equal to \( R_u \). 

14
Under the assumed condition (56) relating \(\{\mu, \mu_m\}\), if we subtract (60) from (59) we obtain:

\[
\hat{w}_i - \tilde{x}_i = (I_M - \mu H_{i-1})(\tilde{w}_{i-1} - \tilde{x}_{i-1}) + \mu(R_{i-1} - H_{i-1})\tilde{x}_{i-1} \\
+ \mu \beta' H_{i-1} \tilde{w}_{i-1} + \mu [s_i(\psi_{i-1}) - s_i(x_{i-1})].
\]  

(62)

In the quadratic case, the second term on the right-hand side is zero since \(R_{i-1} = H_{i-1} = R_u\). It is the presence of this term that makes the analysis more demanding in the general case.

To examine how close \(\hat{w}_i\) gets to \(\tilde{x}_i\) for each iteration, we start by noting that

\[
E\|\hat{w}_i - \tilde{x}_i\|^2 = E \left\| (I_M - \mu H_{i-1})(\tilde{w}_{i-1} - \tilde{x}_{i-1}) + \mu(R_{i-1} - H_{i-1})\tilde{x}_{i-1} + \mu \beta' H_{i-1} \tilde{w}_{i-1} + \mu [s_i(\psi_{i-1}) - s_i(x_{i-1})] \right\|^2 \\
+ \mu^2 E\|s_i(\psi_{i-1}) - s_i(x_{i-1})\|^2.
\]

(63)

Now, applying a similar derivation to the one used to arrive at (122) in Appendix C and the inequality \(\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2\), we can conclude from (63) that

\[
E\|\hat{w}_i - \tilde{x}_i\|^2 \leq (1 - \mu \nu) E\|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^2 + \frac{{2\mu \beta^2 \delta^2}}{\nu} E\|\tilde{w}_{i-1}\|^2 \\
+ \frac{2\mu}{\nu} E\|R_{i-1} - H_{i-1}\|\tilde{x}_{i-1}\|^2 + \mu^2 E\|s_i(\psi_{i-1}) - s_i(x_{i-1})\|^2.
\]

(64)

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

\[
E\|R_{i-1} - H_{i-1}\|\tilde{x}_{i-1}\|^2 \leq E\|R_{i-1} - H_{i-1}\|^2 E\|\tilde{x}_{i-1}\|^2 \leq \sqrt{E\|R_{i-1} - H_{i-1}\|^2} E\|\tilde{x}_{i-1}\|^4.
\]

(65)

In the above inequality, the term \(E\|\tilde{x}_{i-1}\|^4\) can be bounded by using the result of Lemma 2. Therefore, we focus on bounding \(E\|R_{i-1} - H_{i-1}\|^4\) 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 G for the verification.

**Assumption 4** Consider the iterates \(\psi_{i-1}\) and \(x_{i-1}\) that are generated by the momentum recursion (21) and the stochastic gradient recursion (2). It is assumed that the gradient noise process satisfies:

\[
E\|s_i(\psi_{i-1}) - s_i(x_{i-1})\|^2 |\mathcal{F}_{i-1} \right\| \leq \xi_1 \|\psi_{i-1} - x_{i-1}\|^2,
\]

(66)

\[
E\|s_i(\psi_{i-1}) - s_i(x_{i-1})\|^4 |\mathcal{F}_{i-1} \right\| \leq \xi_2 \|\psi_{i-1} - x_{i-1}\|^4.
\]

(67)

for some constants \(\xi_1\) and \(\xi_2\).

**Assumption 5** The Hessian of the risk function \(J(w)\) in (1) is Lipschitz continuous, i.e., for any two variables \(w_1, w_2 \in \operatorname{dom} J(w)\), it holds that

\[
\|\nabla^2_w J(w_1) - \nabla^2_w J(w_2)\| \leq \kappa \|w_1 - w_2\|.
\]

(68)

for some constant \(\kappa \geq 0\).
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 (21)–(22) and assume they start from the same initial states, namely, \( \psi_{-2} = w_{-2} = x_{-1} \). We continue to assume conditions (25), (26), and (56). Under Assumptions 1, 3, 5 and for sufficiently small step-sizes \( \mu \), the following two results hold:

\[
\begin{align*}
\mathbb{E}\|\psi_i - \bar{x}_i\|^4 &= O(\mu), \quad \forall i = 0, 1, 2, 3, \ldots, \\
\limsup_{i \to \infty} \mathbb{E}\|\psi_i - \bar{x}_i\|^4 &= O(\mu^2).
\end{align*}
\]

**Proof** See Appendix II.

Although sufficient for our purposes, we remark that the bound (69) for \( \mathbb{E}\|\psi_i - \bar{x}_i\|^4 \) is not tight. The reason is that in the derivation in Appendix II we employed a looser bound for the term \( \mathbb{E}\|\!(\!R_{i-1} - H_{i-1}\!\!)\!\bar{x}_{i-1}\|^4 \) in order to avoid the appearance of higher-order powers, such as \( \mathbb{E}\|\!R_{i-1} - H_{i-1}\!\|^8 \) and \( \mathbb{E}\|\bar{x}_{i-1}\|^8 \). To avoid this possibility, we employed the following bound (using (10) to bound \( \|R_{i-1}\|^4 \) and \( \|H_{i-1}\|^4 \) and the inequality \( \|a + b\|^4 \leq 8\|a\|^4 + 8\|b\|^4 \)):

\[
\begin{align*}
\mathbb{E}\|(R_{i-1} - H_{i-1})\bar{x}_{i-1}\|^4 &\leq \mathbb{E}\|R_{i-1} - H_{i-1}\|^4\|\bar{x}_{i-1}\|^4 \\
&\leq 8\mathbb{E}\{\|R_{i-1}\|^4 + \|H_{i-1}\|^4\}\|\bar{x}_{i-1}\|^4 \leq 16\delta^4\mathbb{E}\|\bar{x}_{i-1}\|^4.
\end{align*}
\]

(71)

Based on Lemma 9, we can now bound \( \mathbb{E}\|R_{i-1} - H_{i-1}\|^4 \), 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, and 5, and for sufficiently small step-sizes \( \mu \), the following two results hold:

\[
\begin{align*}
\mathbb{E}\|R_{i-1} - H_{i-1}\|^4 &\leq C\mu, \quad \forall i = 0, 1, 2, 3, \ldots, \\
\limsup_{i \to \infty} \mathbb{E}\|R_{i-1} - H_{i-1}\|^4 &= O(\mu^2).
\end{align*}
\]

(72) (73)

**Proof** See Appendix II.

With the upper bounds of \( \mathbb{E}\|R_{i-1} - H_{i-1}\|^4 \) and \( \mathbb{E}\|\bar{x}_{i-1}\|^4 \) established in Lemma 10 and Lemma 2 respectively, we are able to bound \( \|(R_{i-1} - H_{i-1})\bar{x}_{i-1}\| \) in (65), 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 (21)–(22) and assume they start from the same initial states, namely, \( \psi_{-2} = w_{-2} = x_{-1} \). Suppose conditions (25), (26), and (56) hold. Under Assumptions 1, 3, 4, and 5 and for sufficiently small step-size \( \mu \), it holds that

\[
\mathbb{E}\|\bar{w}_i - \bar{x}_i\|^2 = O(\mu^{3/2}), \quad \forall i = 0, 1, 2, 3, \ldots
\]

(74)
Furthermore, in the limit,

$$\lim_{i \to \infty} \mathbb{E}\|\tilde{w}_i - \tilde{x}_i\|^2 = O(\mu^2).$$

(75)

**Proof** See Appendix J.

### 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 say that the two implementations are equivalent.

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 implementation 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 $\mathbf{w}_i - \mathbf{w}_{i-1}$ and $\mathbf{\psi}_i - \mathbf{\psi}_{i-1}$ 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 $\mathbf{w}_i$ and $\mathbf{\psi}_i$ 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

The above observations 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. 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.

To be more specific, the proposed construction takes the form

$$\psi_{i-1} = w_{i-1} + \beta_1(i)(w_{i-1} - w_{i-2}),$$  \hspace{1cm} (76)
$$w_i = \psi_{i-1} - \mu \nabla_w Q(\psi_{i-1}; \theta_i) + \beta_2(i)(\psi_{i-1} - \psi_{i-2}),$$  \hspace{1cm} (77)

with the same initial conditions as in (23)–(24). Similar to condition (25), \(\beta_1(i)\) and \(\beta_2(i)\) also need to satisfy

$$\beta_1(i) + \beta_2(i) = \beta(i), \quad \beta_1(i)\beta_2(i) = 0.$$  \hspace{1cm} (78)

The efficacy of (76)–(77) 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,

$$\beta(i) = \begin{cases} 
\beta_0 & \text{if } i \in [1, T] \\
\beta_0/T^\alpha & \text{if } i \in [T + 1, 2T] \\
\beta_0/(2T)^\alpha & \text{if } i \in [2T + 1, 3T] \\
\beta_0/(3T)^\alpha & \text{if } i \in [3T + 1, 4T] \\
\cdots
\end{cases}$$  \hspace{1cm} (79)

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 \(\alpha = 0.4\).

6. Diagonal Step-size Matrices

Sometimes it is advantageous to employ separate step-size for the individual entries of the weight vectors. 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:

$$x_i = x_{i-1} - D \nabla_w Q(x_{i-1}; \theta_i), \quad i \geq 0,$$  \hspace{1cm} (80)

where \(D = \text{diag}\{\mu_1, \mu_2, \ldots, \mu_M\}\). Here, we continue to use the letter “\(x\)” to refer to the variable iterates for the standard stochastic gradient descent iteration, while we reserve the letter “\(w\)” for the momentum recursion. We let \(\mu_{\text{max}} = \max\{\mu_1, \ldots, \mu_M\}\). Similarly, recursions (21) and (22) can be extended in the following manner:

$$\psi_{i-1} = w_{i-1} + B_1(w_{i-1} - w_{i-2}),$$  \hspace{1cm} (81)
$$w_i = \psi_{i-1} - D m \nabla_w Q(\psi_{i-1}; \theta_i) + B_2(\psi_{i-1} - \psi_{i-2}),$$  \hspace{1cm} (82)
Figure 1: \( \beta(i) \) changes with iteration \( i \) according to (79), where \( \beta_0 = 0.5 \), \( T = 20 \) and \( \alpha = 0.4 \). The reference curve is \( f(i) = 0.5/i^{0.4} \).

with initial conditions

\[
\begin{align*}
\mathbf{w}_{-2} &= \mathbf{\psi}_{-2} = \text{initial states}, \\
\mathbf{w}_{-1} &= \mathbf{w}_{-2} - D_m \nabla_{\mathbf{w}} Q(\mathbf{w}_{-2}; \mathbf{\theta}_{-1}),
\end{align*}
\]

where \( B_1 = \text{diag}\{\beta_1^1, \ldots, \beta_M^1\} \) and \( B_2 = \text{diag}\{\beta_1^2, \ldots, \beta_M^2\} \) are momentum coefficient matrices, while \( D_m \) is a diagonal step-size matrix for momentum stochastic gradient method. In a manner similar to (25), we also assume that

\[
0 \leq B_k < I_M, \quad k = 1, 2, \quad B_1 + B_2 = B, \quad B_1 B_2 = 0.
\]

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

\[
B \leq (1 - \epsilon)I_M, \quad \text{for some constant } \epsilon > 0.
\]

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 (85) and (86). Then, for the momentum stochastic gradient method (81)–(82), it holds under sufficiently small step-size \( \mu_{\text{max}} \) that

\[
\limsup_{i \to \infty} E \| \tilde{\mathbf{w}}_i \|^2 = O(\mu_{\text{max}}).
\]

**Theorem 3B (Equivalence for quadratic costs).** Consider recursions (48) and (52)–(53) 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, \( \psi_{-2} = w_{-2} = x_{-1} \). Suppose further that conditions (85) and (86) hold, and that the step-sizes matrices \( \{D, D_m\} \) satisfy a relation similar to (86), namely,

\[
D = (I - B)^{-1}D_m.
\]

Then, it holds under sufficiently small \( \mu_{\text{max}} \), that

\[
\mathbb{E}\|\bar{w}_i - \bar{x}_i\|^2 = O(\mu_{\text{max}}^2), \quad \forall i = 0, 1, 2, 3, \ldots
\]

(89)

**Theorem 4B (Equivalence for general costs).** Consider the stochastic gradient recursion (50) and the momentum stochastic gradient recursions (31)–(32) to solve the general problem (1). Assume they start from the same initial states, namely, \( \psi_{-2} = w_{-2} = x_{-1} \). Suppose conditions (55), (56), and (58) hold. Under Assumptions 1, 3, 4, and 5, and for sufficiently small step-sizes, it holds that

\[
\mathbb{E}\|\bar{w}_i - \bar{x}_i\|^2 = O(\mu_{\text{max}}^{3/2}), \quad \forall i = 0, 1, 2, 3, \ldots
\]

(90)

Furthermore, in the limit,

\[
\limsup_{i \to \infty} \mathbb{E}\|\bar{w}_i - \bar{x}_i\|^2 = O(\mu_{\text{max}}^2).
\]

(91)

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 (76)–(77) and verify its advantages in the stochastic context.

7.1 Least Mean-Squares Error Designs

We apply the standard LMS algorithm to (43). To do so, we generate data according to the linear regression model (44), where \( \omega^0 \in \mathbb{R}^10 \) is chosen randomly, and \( u_i \in \mathbb{R}^{10} \) is i.i.d and follows \( u_i \sim \mathcal{N}(0, \Lambda) \) where \( \Lambda \in \mathbb{R}^{10 \times 10} \) is randomly-generated diagonal matrix with positive diagonal entries. Besides, \( v(i) \) is also i.i.d and follows \( v(i) \sim \mathcal{N}(0, \sigma_v^2 I_{10}) \), where \( \sigma_v^2 = 0.01 \). All results are averaged over 300 random trials. For each trial we generated 800 samples of \( u_i, v(i) \) and \( d(i) \).

We first compare the standard and momentum LMS algorithms using \( \mu = \mu_m = 0.01 \). The momentum parameter \( \beta \) is set to 0.7. These two methods are illustrated 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 \( \mu_m = \mu(1 - \beta) = 0.003 \) 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 \( \mu_m = 0.01 \) and make \( \beta(i) \) decrease in a stair-wise fashion: when \( i \in [1, 50] \), \( \beta(i) = 0.7 \); when \( i \in [51, 100] \), \( \beta(i) = 0.7/(50^{0.3}) \); when \( i \in [101, 150] \), \( \beta(i) = 0.7/(100^{0.3}) \); \ldots; when \( i \in [751, 800] \), \( \beta(i) = 0.7/(750^{0.3}) \). 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.
7.2 Regularized Logistic Regression

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

\[ J(w) = \frac{\rho}{2} \|w\|^2 + E\{ \ln [1 + \exp(-\gamma(i)h_i^T w)] \} \quad (92) \]

where the approximate gradient vector is chosen as

\[ \nabla_w Q(w; h_i, \gamma(i)) = \rho w - \frac{\exp(-\gamma(i)h_i^T w)}{1 + \exp(-\gamma(i)h_i^T w)} \gamma(i)h_i \quad (93) \]

In the simulation, we generate 20000 samples \((h_i, \gamma(i))\). Among these training points, 10000 feature vectors \(h_i\) correspond to label \(\gamma(i) = 1\) and each \(h_i \sim \mathcal{N}(1.5 \times 1_{10}, R_h)\) for some diagonal covariance \(R_h\). The remaining 10000 feature vectors \(h_i\) correspond to label \(\gamma(i) = -1\) and each \(h_i \sim \mathcal{N}(-1.5 \times 1_{10}, R_h)\). 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.

Similar to the least-mean-squares error problem, we first compare the standard and momentum stochastic methods using \(\mu = \mu_m = 0.01\). The momentum parameter \(\beta\) is set to 0.5. These two methods are illustrated in Fig. 3 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 \(\mu_m = \mu(1 - \beta) = 0.005\) 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 \(\mu_m = 0.01\) and make \(\beta(i)\) decrease in a stair-wise manner: when \(i \in \)
[1, 200], \( \beta(i) = 0.5 \); when \( i \in [200, 400] \), \( \beta(i) = 0.5/(200^{0.3}) \); when \( i \in [400, 600] \), \( \beta(i) = 0.5/(400^{0.3}) \); \ldots; when \( i \in [1800, 2000] \), \( \beta(i) = 0.5/(1800^{0.3}) \). 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.

![Figure 3: Convergence behaviors of standard and momentum stochastic gradient methods applied to the logistic regression problem](image)

Next, we test the standard and momentum stochastic methods for regularized logistic regression problem over a benchmark data set — the Adult Data Set. 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 \( \mu = 0.01 \), \( \rho = 0.1 \), and \( \beta = 0.5 \). To check the equivalence of the algorithms, we set \( \mu_m = (1 - \beta)\mu = 0.005 \). In Fig. 4, 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. 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.

### 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., \( \mathbb{E}\|\mathbf{w}_i - \mathbf{x}_i\|^2 \), between the standard and momentum stochastic gradient methods.

---

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

3. To smooth the performance curve, we applied the weighted average technique from equation (74) of [Ying and Sayed, 2015](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/).
Figure 4: Performance accuracy of the standard and momentum stochastic gradient methods applied to logistic regression classification on the adult data test set.

For the least-mean-squares error problem, the selection of $u_i$, $v(i)$, $d(i)$ and $\beta$ is the same as in the simulation generated earlier in Subsection 7.1. For some specific step-size $\mu$, $x_i$ is the iterate generated through LMS recursion (48) with step-size $\mu$, and $w_i$ is the iterate generated momentum LMS recursion (52)–(53) with step-size $\mu_m = \mu(1 - \beta)$. Now we introduce the maximum difference:

$$d_{\text{max}}(\mu) = \max_i E \| w_i - x_i \|^2$$

and the difference at steady state

$$d_{\text{ss}}(\mu) = \limsup_{i \to \infty} E \| w_i - x_i \|^2.$$

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

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

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

where $d_{\text{max}}(\mu)(\text{dB}) = 10 \log d_{\text{max}}(\mu)$. Relation (94) can be confirmed with red circle line in Fig. 5. In this simulation, we choose 8 different step-size values $\{\mu_k\}_{k=1}^8$, and it can be verified that each data pair $(\log \mu_k, d_{\text{max}}(\mu_k)(\text{dB}))$ satisfies relation (94). For example, in the red circle solid line, at $\mu_1 = 10^{-2}$ we read $d_{\text{max}}(\mu_1)(\text{dB}) = -32\text{dB}$; while at $\mu_2 = 10^{-4}$ we read $d_{\text{max}}(\mu_2)(\text{dB}) = -72\text{dB}$. It can be verified that

$$d_{\text{max}}(\mu_1)(\text{dB}) - d_{\text{max}}(\mu_2)(\text{dB}) = 20(\log \mu_1 - \log \mu_2) = 40. \tag{95}$$
Using a similar argument, the blue square solid line can also implies that \( d_{ss} = O(\mu^2) \).

Figure 5 also reveals the order of \( d_{\text{max}} \) and \( d_{ss} \), with magenta and green dash lines respectively, for the regularized logistic regression problem from Subsection 7.2. With the same argument as above, \( d_{ss}(\mu) \) can be confirmed on the order of \( O(\mu^2) \). Now we check the order of \( d_{\text{max}}(\mu) \). The fact that \( d_{\text{max}}(\mu) = c\mu^{3/2} \) holds if and only if

\[
d_{\text{max}}(\mu)(\text{dB}) = 15\log \mu + 10\log c.
\]

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

\[
d_{\text{max}}(\mu_1)(\text{dB}) - d_{\text{max}}(\mu_2)(\text{dB}) = 15(\log \mu_1 - \log \mu_2) = 30.
\]

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

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

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

Figure 5: \( d_{\text{max}} \) and \( d_{ss} \) as a function of the step-size \( \mu \). MS stands for mean-square-error and LR stands for logistic regression.

### 7.4 Handwritten Digit Recognition

In this subsection we illustrate the conclusions in this work by re-examining the well-known problem of training a neural network to recognize handwritten digits. We will employ the MNSIT database which is a classical benchmark dataset of images for segmented

---

4. [http://yann.lecun.com/exdb/mnist/](http://yann.lecun.com/exdb/mnist/)
handwritten digits, each with $28 \times 28$ pixels. There are 60,000 training examples and 10,000 testing examples.

In this experiment we train the data through a two-layer fully-connected neural network. We employ the cross-entropy measure with $\ell_2$ regularization as a cost objective, and the ReLU as an activation function. The hidden layer has 100 units, the coefficient of the $\ell_2$ regularization term is set to 0.001, and the initial value $w_{-1}$ is generated by Gaussian distribution with 0.05 standard deviation. To accelerate training, we employ mini-batch stochastic-gradient learning with batch size equal to 100. First, we apply momentum stochastic-gradient algorithm to train this neural network. The momentum parameter is set at $\beta = 0.8$, and the initial step-size $\mu_\text{m}$ is set to 0.04. To achieve better accuracy, we follow a common technique (e.g., [Szegedy et al., 2015]) and reduce $\mu_\text{m}$ to $0.9\mu_\text{m}$ after every 500 iterations. With the above settings, we attain an accuracy of 97% (see Figure 6), which is the best performance we can achieve with momentum stochastic gradient method during the experiment.

However, what is interesting, and somewhat surprising, is that the above 97% accuracy can also be achieved with stochastic gradient descent itself. According to the step-size relation $\mu = \mu_\text{m}/(1-\beta)$, we set the initial step-size $\mu$ of SGD to 0.2. Similar to momentum method, we also reduce $\mu$ to $0.9\mu$ after every 500 iterations for SGD, and hence the relation $\mu = \mu_\text{m}/(1-\beta)$ still holds for each iteration. From Figure 6 we can observe that the accuracy performance curves of both scenarios, with and without momentum, are indistinguishable even when the overall risk is not necessarily convex or differentiable. Both scenarios attain an accuracy of 97%, which is better than the reference accuracy for this handwritten digit recognition task. This result indicates that the performance of momentum SGD can still be achieved by standard SGD by properly adjusting the step-size according to $\mu = \mu_\text{m}/(1-\beta)$.

In fact, apart from the test accuracy for both scenarios being similar, we can further verify a stronger conclusion that the converged weight matrix, i.e., the parameter $w$ for standard and momentum SGD are also almost identical. In Figure 7 we present visual representations for the 20 weight matrices from the first layer for both standard and momentum SGD training. From this figure, we observe that the images are almost indistinguishable by the human eye. The numerical result shows that the average difference for each entry between these two weights is around $2.536 \times 10^{-5}$, which amounts to a 0.4% relative error.

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 $\mu$. When larger values for $\mu$ 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 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?

5. See, for example, [http://yann.lecun.com/exdb/mnist/](http://yann.lecun.com/exdb/mnist/)
Figure 6: Classification accuracy of the standard and momentum stochastic gradient methods applied to a two-layer fully-connected neural network on the MNIST test data set.

Figure 7: Comparison between weights generated through standard and momentum SGD.

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 (43). We suppose $M = 5$, $u_i \sim N(0, 0.5I_5)$ and $d(i) = u_i^T w^o + v(i)$ where $v(i) \sim N(0, 0.01)$. 

26
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 $u_i$, $v(i)$ and $d(i)$. In Fig. 8, 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 8: Convergence comparison between standard and momentum LMS algorithms when $\mu = \mu_m = 0.4$ and $\beta = 0.5$.](image)

9. Conclusion

In this paper we carefully analyzed the behavior (convergence and performance) of momentum stochastic gradient methods in the constant step-size and slow adaptation regime. The results establish that the momentum method is essentially 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, $\beta$. 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 suggests 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. In this way, the enhanced convergence rate during the initial stages of adaptation is retained without the often-observed degradation in MSD performance.
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 Sec. 5.3.

References

Michel Baes. Estimate sequence methods: extensions and approximations. IFOR Internal report, ETH, Zürich, Switzerland, 2009.

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.

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.

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. Available as arXiv preprint arXiv: 1602.05419, Feb. 2016.

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. 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 Advances in Neural Information Processing Systems, pages 781–789, 2009.

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.

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.

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.

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.

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, 2014a.

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

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.

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.

B. Ying and A. H. Sayed. Performance limits of online stochastic sub-gradient learning. submitted for publication. Available as arXiv preprint arXiv:1511.07902, Oct. 2015.

S. Zareba, A. Gonczarek, J. M. Tomczak, and J. Światek. Accelerated learning for restricted Boltzmann machine with momentum term. In Proc. International Convergence 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.

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.

Appendix A. Proof of Lemma 2

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

$$E\|\bar{w}_i\|^4 \leq (1 - \mu \nu)E\|\bar{w}_{i-1}\|^4 + a_1 \mu^2 E\|\bar{w}_{i-1}\|^2 + a_2 \mu^4,$$

for some constants $a_1$ and $a_2$. If we iterate (14) we find that

$$E\|\bar{w}_{i-1}\|^2 \leq (1 - \mu \nu)^{i+1}E\|\bar{w}_{-1}\|^2 + a_3 \mu,$$

for some constant $a_3$. If we iterate (14) we find that

$$E\|\bar{w}_{i-1}\|^2 \leq (1 - \mu \nu)^{i+1}E\|\bar{w}_{-1}\|^2 + a_3 \mu.$$
where $a_3 = \sigma^2_t / \nu$. Substituting inequality (99) into (98), we find that it holds for each iteration $i = 0, 1, 2, \ldots$

$$\mathbb{E} \| \mathbf{w}_i \|^4 \leq (1 - \mu \nu) \mathbb{E} \| \mathbf{w}_{i-1} \|^4 + a_2 \mu^4 + a_1 a_3 \mu^3 + a_4 \mu^2 (1 - \mu \nu)^{i+1},$$

(100)

where $a_4 = a_1 \mathbb{E} \| \mathbf{w}_{i-1} \|^2$. Iterating the above inequality we get

$$\mathbb{E} \| \mathbf{w}_i \|^4 \leq (1 - \mu \nu)^{i+1} \mathbb{E} \| \mathbf{w}_{i-1} \|^4 + a_2 \mu^4 \sum_{s=0}^{i} (1 - \mu \nu)^s + a_1 a_3 \mu^3 \sum_{s=0}^{i} (1 - \mu \nu)^s + a_4 \mu^2 (i + 1)(1 - \mu \nu)^{i+1},$$

(101)

where $a_5 = a_4 / \nu$ and $a_6 = a_1 a_3 / \nu$. Let

$$z(i) \triangleq (i + 1)(1 - \mu \nu)^{i+1}, \ \forall i = 0, 1, 2, \ldots$$

(102)

Since $1 - \mu \nu < 1$, it holds that the sequence $z(i)$ converges to 0. Since convergent sequences are bounded, it follows that

$$z(i) \leq D, \ \forall i = 0, 1, 2, \ldots$$

(103)

for some constant $D$. In this way, expression (101) leads to the desired result (10). Furthermore, we conclude from this relation that (20) holds.

Appendix B. Proof of Lemma 3

We substitute the expression for the gradient noise from (11), evaluated at $\psi_{i-1}$, into (22) to get:

$$\mathbf{w}_i = \psi_{i-1} - \mu_m \nabla_w J(\psi_{i-1}) + \beta_2 (\psi_{i-1} - \psi_{i-2}) - \mu_m s_i(\psi_{i-1}).$$

(104)

Let again $\mathbf{w}_i = w^o - \mathbf{w}_i$ and $\tilde{\psi}_i = w^o - \psi_i$. Subtracting both sides of (104) from $w^o$ gives:

$$\mathbf{w}_i = \tilde{\psi}_{i-1} + \mu_m \nabla_w J(\psi_{i-1}) - \beta_2 (\psi_{i-1} - \psi_{i-2}) + \mu_m s_i(\psi_{i-1}).$$

(105)

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

$$\nabla J_w(\psi_{i-1}) = - \left( \int_0^1 \nabla^2 J_w(w^o - t \tilde{\psi}_{i-1}) dt \right) \tilde{\psi}_{i-1} \triangleq - H_{i-1} \tilde{\psi}_{i-1}.$$  

(106)

and express the momentum term in the form

$$\psi_{i-1} - \psi_{i-2} = \psi_{i-1} - w^o + w^o - \psi_{i-2} = -\tilde{\psi}_{i-1} + \tilde{\psi}_{i-2}. $$

(107)

Then, expression (105) can be rewritten as

$$\mathbf{w}_i = (I_M + \beta_2 I_M - \mu_m H_{i-1}) \tilde{\psi}_{i-1} - \beta_2 \tilde{\psi}_{i-2} + \mu_m s_i(\psi_{i-1}).$$

(108)
On the other hand, expression (21) gives
\[ \tilde{\psi}_{i-1} = \tilde{w}_{i-1} + \beta_1 (\tilde{w}_{i-1} - \tilde{w}_{i-2}). \] (109)

Substituting (109) into (108), we have
\[ \tilde{w}_i = J_{i-1} \tilde{w}_{i-1} + K_{i-1} \tilde{w}_{i-2} + L \tilde{w}_{i-3} + \mu_m s_i(\psi_{i-1}), \] (110)
where boldface quantities denote random variables:

\[
\begin{align*}
J_{i-1} &= (1 + \beta_1)(1 + \beta_2)I_M - \mu_m(1 + \beta_1)H_{i-1} \overset{(111)}{=} (1 + \beta)I_M - \mu_m(1 + \beta_1)H_{i-1} \\
K_{i-1} &= -(\beta_1 + \beta_2)I_M + \mu_m\beta_1 H_{i-1} = -\beta I_M + \mu_m\beta_1 H_{i-1} \quad \text{(112)} \\
L &= \beta_1\beta_2 = 0 \quad \text{(113)}
\end{align*}
\]

It follows that we can write the extended relation:
\[
\begin{bmatrix}
\tilde{w}_i \\
\tilde{w}_{i-1}
\end{bmatrix} = 
\begin{bmatrix}
J_{i-1} & K_{i-1} \\
I_M & 0
\end{bmatrix}
\begin{bmatrix}
\tilde{w}_{i-1} \\
\tilde{w}_{i-2}
\end{bmatrix} + \mu_m \begin{bmatrix} s_i(\psi_{i-1}) \\ 0 \end{bmatrix},
\] (114)
where we are denoting the coefficient matrix by \( B_{i-1} \), which can be written as the difference
\[
B_{i-1} \overset{\Delta}{=} P - M_{i-1},
\] (115)
with
\[
P = \begin{bmatrix}
(1 + \beta)I_M & -\beta I_M \\
I_M & 0
\end{bmatrix}, \quad M_{i-1} = \begin{bmatrix}
\mu_m(1 + \beta_1) H_{i-1} & -\mu_m\beta_1 H_{i-1} \\
0 & 0
\end{bmatrix}. \] (116)

The eigenvalue decomposition of \( P \) can be easily seen to be given by \( P = VDV^{-1} \), where
\[
V = \begin{bmatrix}
I_M & -\beta I_M \\
I_M & -I_M
\end{bmatrix}, \quad V^{-1} = \frac{1}{1 - \beta} \begin{bmatrix}
I_M & -\beta I_M \\
I_M & -I_M
\end{bmatrix}, \quad D = \begin{bmatrix}
I_M & 0 \\
0 & \beta I_M
\end{bmatrix}. \] (117)

Therefore, we have
\[
B_{i-1} = V(D - V^{-1} M_{i-1} V) V^{-1} = V \begin{bmatrix}
I_M - \frac{\mu_m}{1 - \beta} H_{i-1} & \frac{\mu_m\beta'}{1 - \beta} H_{i-1} \\
-\frac{\mu_m}{1 - \beta} H_{i-1} & \beta I_M + \frac{\mu_m\beta'}{1 - \beta} H_{i-1}
\end{bmatrix} V^{-1}, \] (118)
where
\[
\beta' \overset{\Delta}{=} \beta_1 + \beta - \beta_1 = \beta_1 + \beta_2. \] (119)

Multiplying both sides of (113) by \( V^{-1} \) from the left and recalling definition (28), we obtain
\[
\begin{bmatrix}
\hat{w}_i \\
\hat{w}_i
\end{bmatrix} = \begin{bmatrix}
I_M - \frac{\mu_m}{1 - \beta} H_{i-1} & \frac{\mu_m\beta'}{1 - \beta} H_{i-1} \\
-\frac{\mu_m}{1 - \beta} H_{i-1} & \beta I_M + \frac{\mu_m\beta'}{1 - \beta} H_{i-1}
\end{bmatrix} \begin{bmatrix}
\tilde{w}_{i-1} \\
\tilde{w}_{i-1}
\end{bmatrix} + \frac{\mu_m}{1 - \beta} \begin{bmatrix} s_i(\psi_{i-1}) \\ s_i(\psi_{i-1}) \end{bmatrix}. \] (120)
Appendix C. Proof of Theorem 4

From the first row of recursion (29), we have

\[ \tilde{w}_i = \left( I_M - \frac{\mu_m H_{i-1}}{1 - \beta} \right) \tilde{w}_{i-1} + \frac{\mu_m \beta' H_{i-1}}{1 - \beta} \tilde{w}_{i-1} + \frac{\mu_m}{1 - \beta} s_i(\psi_{i-1}). \tag{121} \]

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

\[
\mathbb{E}[\|\tilde{w}_i\|^2 | \mathcal{F}_{i-1}] = \left\| \left( I_M - \frac{\mu_m}{1 - \beta} H_{i-1} \right) \tilde{w}_{i-1} + \frac{\mu_m \beta'}{1 - \beta} H_{i-1} \tilde{w}_{i-1} \right\|^2 + \frac{\mu_m^2}{(1 - \beta)^2} \mathbb{E}[\|s_i(\psi_{i-1})\|^2 | \mathcal{F}_{i-1}]
\]

\[
\leq \left\| (1 - t) \frac{1}{1 - t} \left( I_M - \frac{\mu_m}{1 - \beta} H_{i-1} \right) \tilde{w}_{i-1} + t \frac{1}{1 - \beta} \frac{\mu_m \beta'}{1 - \beta} H_{i-1} \tilde{w}_{i-1} \right\|^2 + \frac{\mu_m^2}{(1 - \beta)^2} (\gamma^2 \|\tilde{\psi}_{i-1}\|^2 + \sigma_s^2)
\]

\[
\leq \frac{1}{1 - t} \left\| \left( I_M - \frac{\mu_m t}{1 - \beta} H_{i-1} \right) \tilde{w}_{i-1} \right\|^2 + \frac{\mu_m^2 t^2}{(1 - \beta)^2} \|\tilde{w}_{i-1}\|^2 + \frac{\mu_m^2}{(1 - \beta)^2} (\gamma^2 \|\tilde{\psi}_{i-1}\|^2 + \sigma_s^2)
\]

\[
\leq \left( 1 - \frac{\mu_m t}{1 - \beta} \right) \|\tilde{w}_{i-1}\|^2 + \frac{\mu_m \beta'^2 \sigma^2}{\nu (1 - \beta)} \|\tilde{w}_{i-1}\|^2 + \frac{\mu_m^2}{(1 - \beta)^2} (\gamma^2 \|\tilde{\psi}_{i-1}\|^2 + \sigma_s^2). \tag{122}
\]

where (a) holds because of equation (13) in Assumption (2), (b) holds because \( \nu I \leq H_{i-1} \leq \delta I \) under Assumption (1), and (c) holds because we selected \( t = \frac{\mu_m t}{1 - \beta} \). Taking expectation again, we remove the conditioning to find:

\[
\mathbb{E}[\|\tilde{w}_i\|^2] \leq \left( 1 - \frac{\mu_m t}{1 - \beta} \right) \mathbb{E}[\|\tilde{w}_{i-1}\|^2] + \frac{\mu_m \beta'^2 \sigma^2}{\nu (1 - \beta)} \mathbb{E}[\|\tilde{w}_{i-1}\|^2] + \frac{\mu_m^2}{(1 - \beta)^2} (\gamma^2 \mathbb{E}[\|\tilde{\psi}_{i-1}\|^2] + \sigma_s^2). \tag{123}
\]

Furthermore, squaring (109) and using the inequality \( \|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2 \) we get

\[
\|\tilde{\psi}_{i-1}\|^2 \leq 2(1 + \beta_1)^2 \|\tilde{w}_{i-1}\|^2 + 2\beta_1^2 \|\tilde{w}_{i-2}\|^2 \leq 2(1 + \beta_1)^2 (\|\tilde{w}_{i-1}\|^2 + \|\tilde{w}_{i-2}\|^2)
\]

\[
= 2(1 + \beta_1)^2 \left\| \frac{\tilde{w}_{i-1}}{\tilde{w}_{i-2}} \right\|^2 \mathbb{E}[\|\tilde{w}_{i-1}\|^2] + \mathbb{E}[\|\tilde{w}_{i-2}\|^2],
\]

\[
\leq 2(1 + \beta_1)^2 \mathbb{E}[\|\tilde{w}_{i-1}\|^2] + \mathbb{E}[\|\tilde{w}_{i-2}\|^2]. \tag{124}
\]

where we introduced \( \nu^2 = \|V\|^2 \). Therefore, under expectation, we conclude that it also holds:

\[
\mathbb{E}[\|\tilde{\psi}_{i-1}\|^2] \leq 2(1 + \beta_1)^2 \mathbb{E}[\|\tilde{w}_{i-1}\|^2] + \mathbb{E}[\|\tilde{w}_{i-2}\|^2]. \tag{125}
\]

Substituting (125) into (123), we get

\[
\mathbb{E}[\|\tilde{w}_i\|^2] \leq \left( 1 - \frac{\mu_m t}{1 - \beta} + \frac{2(1 + \beta_1)^2 \gamma^2 \nu^2}{\mu_m^2} \right) \mathbb{E}[\|\tilde{w}_{i-1}\|^2] + \frac{\mu_m^2 \sigma_s^2}{(1 - \beta)^2}. \tag{126}
\]
It then follows from (32) that
\[ \mu \text{ small step-size value} \]
the stability of the 2
Since the spectral radius of a matrix is upper bounded by its 1-norm, we have that
\[ \limsup \leq \leq \leq \leq \parallel \H_{i-1} \parallel \leq \leq \mu \parallel \beta \parallel \parallel \H_{i-1} \parallel \leq \leq \mu \parallel \beta \parallel \parallel \H_{i-1} \parallel \leq \leq \mu \parallel \beta \parallel \parallel \H_{i-1} \parallel \leq \leq \mu \parallel \beta \parallel \parallel \H_{i-1} \parallel . \]  

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

\[ E\| \tilde{w}_i \|^2 \]
\[ \leq E \left\| \beta \tilde{w}_{i-1} + \left( \frac{\mu_m \beta H_{i-1}}{1 - \beta} \right) \tilde{w}_{i-1} + \left( \beta I_M + \frac{\mu_m \beta I}{1 - \beta} H_{i-1} \right) \tilde{w}_{i-1} \right\|^2 + \frac{\mu_m^2}{(1 - \beta)^2} (\gamma^2 E\| \tilde{\psi}_{i-1} \|^2 + \sigma_s^2) \]
\[ \leq \beta E\| \tilde{w}_{i-1} \|^2 + \frac{1}{1 - \beta} E \left\| \frac{\mu_m \beta H_{i-1}}{1 - \beta} \tilde{w}_{i-1} - \frac{\mu_m \beta I_{i-1} \tilde{w}_{i-1}}{1 - \beta} \right\|^2 + \frac{\mu_m^2}{(1 - \beta)^2} (\gamma^2 E\| \tilde{\psi}_{i-1} \|^2 + \sigma_s^2) \]
\[ = \left( \beta + \frac{2 \mu_m^2 \beta^2 \sigma_s^2}{(1 - \beta)^3} \right) E\| \tilde{w}_{i-1} \|^2 + \frac{2 \mu_m^2 \gamma^2 (1 + \beta_1)^2 v^2}{(1 - \beta)^2} E\| \tilde{w}_{i-1} \|^2 + \frac{\mu_m^2 \sigma_s^2}{(1 - \beta)^2} \]  

where (a) holds since \( E\| \beta x + y \|^2 = E\| \beta x \| (1 - \beta) \leq E\| x \|^2 + \frac{1}{1 - \beta} E\| y \|^2 \). Substituting (124) into (128), it follows that:

\[ E\| \tilde{w}_i \|^2 \leq \left( \beta + \frac{2 \mu_m^2 \beta^2 \sigma_s^2}{(1 - \beta)^3} + \frac{2 \mu_m^2 \gamma^2 (1 + \beta_1)^2 v^2}{(1 - \beta)^2} \right) E\| \tilde{w}_{i-1} \|^2 + \frac{\mu_m^2 \sigma_s^2}{(1 - \beta)^2} \]  

Combing relations (126) and (129) leads to the desired result (32)–(33). Let us now examine the stability of the 2 \times 2 coefficient matrix:

\[ \Gamma \triangleq \begin{bmatrix} a & b \\ c & d \end{bmatrix} \]  

Since the spectral radius of a matrix is upper bounded by its 1-norm, we have that
\[ \rho(\Gamma) \leq \max \left\{ 1 - \frac{\mu_m \nu}{1 - \beta} + O(\mu_m^2), \beta + O(\mu_m) \right\} . \]  

Since \( \beta < 1 \) and is not close to one (in view of condition (25)), we conclude that a sufficiently small step-size value \( \mu_m \) exists that ensures \( \rho(\Gamma) < 1 \), in which case \( \Gamma \) will be a stable matrix. It then follows from (32) that

\[ \limsup_{i \to \infty} \left[ \frac{E\| \tilde{w}_i \|^2}{E\| \tilde{w}_{i-1} \|^2} \right] \leq (I_2 - \Gamma)^{-1} \begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} O(1/\mu_m) & O(1) \\ O(\mu_m) & O(\mu_m^2) \end{bmatrix} \begin{bmatrix} O(\mu_m^2) \\ O(\mu_m^2) \end{bmatrix} = O(\mu_m) \]
so that
\[ \limsup_{i \to \infty} E \| \hat{w}_i \|^2 = O(\mu_m), \quad \limsup_{i \to \infty} E \| \hat{w}_i \|^2 = O(\mu_m^2). \] (132)
and
\[ \limsup_{i \to \infty} E \left\| \begin{bmatrix} \hat{w}_i \\ \bar{w}_{i-1} \end{bmatrix} \right\|^2 = \limsup_{i \to \infty} E \left\| V \begin{bmatrix} \hat{w}_i \\ \bar{w}_i \end{bmatrix} \right\|^2 \leq v^2 \left( \limsup_{i \to \infty} E \left\| \begin{bmatrix} \hat{w}_i \\ \bar{w}_i \end{bmatrix} \right\|^2 \right) = v^2 \left( \limsup_{i \to \infty} (E \| \hat{w}_i \|^2 + E \| \bar{w}_i \|^2) \right) = O(\mu_m), \quad (133) \]
from which we conclude that (36) holds.

**Appendix D. Proof of Corollary 5**

To simplify the notation, we refer to (32) and introduce the quantities:
\[ z_i = \begin{bmatrix} E \| \hat{w}_i \|^2 \\ E \| \bar{w}_i \|^2 \end{bmatrix}, \quad \Gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad r = \begin{bmatrix} e \\ f \end{bmatrix}. \] (134)
Then, relation (32) can be rewritten as
\[ z_i \preceq \Gamma z_{i-1} + r. \] (135)
It follows that, in terms of the 1--norm,
\[ \| z_i \|_1 \leq \| \Gamma \|_1 \| z_{i-1} \|_1 + \| r \|_1, \] (136)
where
\[ \| \Gamma \|_1 = \max \left\{ 1 - \frac{\mu_m \nu}{1 - \beta} + c_1 \mu_m^2, \beta + c_2 \mu_m \right\} \] (137)
for some constant \( c_1 \) and \( c_2 \). We recall from (26) that \( \beta < 1 - \epsilon \), for some \( \epsilon > 0 \). Now we can choose \( \mu_m \) sufficiently small to satisfy
\[ c_1 \mu_m < \frac{\nu}{2(1 - \beta)}, \quad \left( c_2 + \frac{\nu}{2(1 - \beta)} \right) \mu_m < \epsilon, \] (138)
which implies that
\[ \| \Gamma \|_1 \leq 1 - \frac{\mu_m \nu}{1 - \beta} + c_1 \mu_m^2 \leq 1 - \frac{\mu_m \nu}{2(1 - \beta)} \Delta \rho_1 < 1 \] (139)
Then, from (136) we have
\[ \| z_i \|_1 \leq \rho_1 \| z_{i-1} \|_1 + \| r \|_1. \] (140)
Accordingly, using 1.4 we also find that $E \parallel z_i \parallel 1 \leq r_{i+1} \parallel z_i \parallel 1 + \frac{\parallel r \parallel 1}{1 - r_1}$. (141)

Since $\parallel r \parallel 1 = O(\mu_m^2)$ and $1 - r_1 = O(\mu_m)$, we get $\parallel r \parallel 1/(1 - r_1) = O(\mu_m)$. Moreover, since $\parallel z_i \parallel 1 = O(1)$, we conclude that $\parallel z_i \parallel 1 \leq D$ for $i = 0, 1, 2, \ldots$, where $D$ is some constant.

Accordingly, using

$$E\|\hat{w}_i\|^2 \leq E\|\hat{w}_i\|^2 + E\|\hat{w}_i\|^2 = \|z_i\|^1$$

(142)

we also find that $E\|\hat{w}_i\|^2 \leq D, \ i = 0, 1, 2, \ldots$

On the other hand, we know from the second row of (32), or from inequality (129) that

$$E\|\hat{w}_i\|^2 \leq (\beta + c_1 \mu_m^2)E\|\hat{w}_{i-1}\|^2 + c_2 \mu_m E\|\hat{w}_{i-1}\|^2 + c_3 \mu_m^2$$

(143)

for some constants $c_1, c_2$ and $c_3$. Using $\beta \leq 1 - \epsilon$ we have

$$\alpha \equiv \beta + c_1 \mu_m^2 \leq 1 - \epsilon + c_1 \mu_m^2 = 1 - \frac{1}{2} \epsilon + \left(c_1 \mu_m^2 - \frac{1}{2} \epsilon\right).$$

(144)

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

$$\alpha \leq 1 - \frac{1}{2} \epsilon < 1$$

(145)

Let $\rho_2 = 1 - \epsilon/2$. It follows that

$$E\|\hat{w}_i\|^2 \leq \rho_2 E\|\hat{w}_{i-1}\|^2 + c_2 \mu_m E\|\hat{w}_{i-1}\|^2 + c_3 \mu_m^2 \leq \rho_2 E\|\hat{w}_{i-1}\|^2 + c_2 \mu_m^2 D + c_3 \mu_m^2.$$ (146)

Let $c_4 = c_2 D + c_3$. Then, recursion (146) can be simplified to

$$E\|\hat{w}_i\|^2 \leq \rho_2 E\|\hat{w}_{i-1}\|^2 + c_4 \mu_m^2.$$ (147)

Iterating (147) gives

$$E\|\hat{w}_i\|^2 \leq \rho_2^{j+1}E\|\hat{w}_{-1}\|^2 + c_4 \mu_m^2 \sum_{j=0}^{i} \rho_2^j \leq \rho_2^{i+1}E\|\hat{w}_{-1}\|^2 + \frac{c_4 \mu_m^2}{1 - \rho_2}$$

$$\leq \left(\sum_{j=0}^{\infty} \rho_2^j\right) E\|\hat{w}_{-1}\|^2 + \frac{c_4 \mu_m^2}{1 - \rho_2} \leq \frac{1}{1 - \rho_2} E\|\hat{w}_{-1}\|^2 + \frac{c_4 \mu_m^2}{1 - \rho_2}. $$ (148)

To assess the term that depends on the initial state, $E\|\hat{w}_{-1}\|^2$, let us consider the boundary conditions (23) - (24). Then, from (28) it holds that

$$\hat{w}_{-1} = \frac{\hat{w}_{-1} - \hat{w}_{-2}}{1 - \beta} = \frac{w_{-2} - w_{-1}}{1 - \beta} = \mu_m \nabla w Q(w_{-2}; \theta_{-1})$$

(149)

so that $E\|\hat{w}_{-1}\|^2 = \mu_m^2$, where

$$c = E\|\nabla w Q(w_{-2}; \theta_{-1})\|^2/(1 - \beta)^2.$$ (150)

Substituting this conclusion into (148) we arrive at conclusion (37).
Appendix E. Proof of Corollary 7

Recall (127) that

\[ \mathbf{w}_i = -\frac{\mu_m}{1 - \beta} \mathbf{H}_{i-1} \mathbf{w}_{i-1} + \left( \beta \mathbf{I}_M + \frac{\mu_m \beta'}{1 - \beta} \mathbf{H}_{i-1} \right) \mathbf{w}_{i-1} + \frac{\mu_m}{1 - \beta} \mathbf{s}_i(\psi_{i-1}). \]  

(151)

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

\[ \|a + b\|^4 \leq \|a\|^4 + 3\|b\|^4 + 8\|a\|^2\|b\|^2 + 4\|a\|^2(a^Tb) \]

(152)

we get

\[
\mathbb{E}[\|\mathbf{w}_i\|^4 | \mathcal{F}_{i-1}] \\
= \left\| -\frac{\mu_m \mathbf{H}_{i-1}}{1 - \beta} \mathbf{w}_{i-1} + \left( \beta \mathbf{I}_M + \frac{\mu_m \beta'}{1 - \beta} \mathbf{H}_{i-1} \right) \mathbf{w}_{i-1} \right\|^4 + \frac{3\mu_m^4 \mathbb{E}[\|\mathbf{s}_i(\psi_{i-1})\|^4 | \mathcal{F}_{i-1}]}{(1 - \beta)^4} \\
+ \frac{8\mu_m^2}{(1 - \beta)^2} \left\| -\frac{\mu_m \mathbf{H}_{i-1}}{1 - \beta} \mathbf{w}_{i-1} + \left( \beta \mathbf{I}_M + \frac{\mu_m \beta'}{1 - \beta} \mathbf{H}_{i-1} \right) \mathbf{w}_{i-1} \right\|^2 \mathbb{E}[\|\mathbf{s}_i(\psi_{i-1})\|^2 | \mathcal{F}_{i-1}] \\
\leq \left\| -\frac{\mu_m \mathbf{H}_{i-1}}{1 - \beta} \mathbf{w}_{i-1} + \left( \beta \mathbf{I}_M + \frac{\mu_m \beta'}{1 - \beta} \mathbf{H}_{i-1} \right) \mathbf{w}_{i-1} \right\|^4 + \frac{3\mu_m^4 \gamma_4 \|\mathbf{w}_{i-1}\|^4 + \sigma_s^4}{(1 - \beta)^4} \\
+ \frac{8\mu_m^2}{(1 - \beta)^2} \left\| -\frac{\mu_m \mathbf{H}_{i-1}}{1 - \beta} \mathbf{w}_{i-1} + \left( \beta \mathbf{I}_M + \frac{\mu_m \beta'}{1 - \beta} \mathbf{H}_{i-1} \right) \mathbf{w}_{i-1} \right\|^2 \left( \gamma^2 \|\mathbf{w}_{i-1}\|^2 + \sigma_s^2 \right). 
\]

(153)

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 (10), and the inequality \( \|a + b\|^4 \leq 8\|a\|^4 + 8\|b\|^4 \), we have

\[
\left\| -\frac{\mu_m \mathbf{H}_{i-1}}{1 - \beta} \mathbf{w}_{i-1} + \left( \beta \mathbf{I}_M + \frac{\mu_m \beta'}{1 - \beta} \mathbf{H}_{i-1} \right) \mathbf{w}_{i-1} \right\|^4 \\
= \left\| \beta \mathbf{w}_{i-1} + (1 - \beta) \left( \frac{\mu_m}{(1 - \beta)^2} \mathbf{H}_{i-1} \mathbf{w}_{i-1} - \frac{\mu_m}{(1 - \beta)^2} \mathbf{H}_{i-1} \mathbf{w}_{i-1} \right) \right\|^4 \\
\leq \beta \|\mathbf{w}_{i-1}\|^4 + (1 - \beta) \left\| \frac{\mu_m \beta'}{(1 - \beta)^2} \mathbf{H}_{i-1} \mathbf{w}_{i-1} - \frac{\mu_m}{(1 - \beta)^2} \mathbf{H}_{i-1} \mathbf{w}_{i-1} \right\|^4 \\
\leq \beta \|\mathbf{w}_{i-1}\|^4 + \frac{8\mu_m^4 \beta^4 \|\mathbf{w}_{i-1}\|^4 + \frac{8\mu_m^4 \beta^4}{(1 - \beta)^2} \|\mathbf{w}_{i-1}\|^4 + \frac{8\mu_m^4 \beta^4}{(1 - \beta)^2} \|\mathbf{w}_{i-1}\|^4 \]

(154)

\[
= (\beta + p_1 \mu_m^4) \|\mathbf{w}_{i-1}\|^4 + p_2 \mu_m^4 \|\mathbf{w}_{i-1}\|^4, \\
\]

where

\[
p_1 = \frac{8\beta^4}{(1 - \beta)^2}, \quad p_2 = \frac{8\delta^4}{(1 - \beta)^2}
\]

are constants. Likewise, from a similar argument we get

\[
\left\| -\frac{\mu_m \mathbf{H}_{i-1}}{1 - \beta} \mathbf{w}_{i-1} + \left( \beta \mathbf{I}_M + \frac{\mu_m \beta'}{1 - \beta} \mathbf{H}_{i-1} \right) \mathbf{w}_{i-1} \right\|^2 \leq (\beta + p_3 \mu_m^2) \|\mathbf{w}_{i-1}\|^2 + p_4 \mu_m^2 \|\mathbf{w}_{i-1}\|^2, \\
\]

(155)
where

\[ p_3 = \frac{2\beta^2 \delta^2}{(1 - \beta)^3}, \quad p_4 = \frac{2\delta^2}{(1 - \beta)^3} \]

are constants. Moreover, from (124) we have

\[ \|\tilde{\psi}_{i-1}\|^2 \overset{p_5}{\leq} \|\tilde{w}_{i-1}\|^2 + \|w_{i-1}\|^2, \tag{156} \]

where \( p_5 = 2(1 + \beta_1)^2 \nu^2 \), which also implies that

\[ \|\tilde{\psi}_{i-1}\|^4 \overset{p_6}{\leq} 4(1 + \beta_1)^4 \nu^4 (\|\tilde{w}_{i-1}\|^2 + \|w_{i-1}\|^2)^2 \]

\[ \overset{p_6}{\leq} 8(1 + \beta_1)^4 \nu^4 (\|\tilde{w}_{i-1}\|^4 + \|w_{i-1}\|^4) = p_6 (\|\tilde{w}_{i-1}\|^4 + \|w_{i-1}\|^4), \tag{157} \]

where \( p_6 = 8(1 + \beta_1)^4 \nu^4 \). Substituting (154), (155), (156) and (157) into (153), we get

\[
E[\|\tilde{w}_i\|^4 | \mathcal{F}_{i-1}] \\
\leq (\beta + p_7 \mu_m^4) \|\tilde{w}_{i-1}\|^4 + p_8 \mu_m^2 \|\tilde{w}_{i-1}\|^4 + \frac{3p_6 \gamma_4 \mu_m^4 \|\tilde{w}_{i-1}\|^4}{(1 - \beta)^4} + \frac{3p_6 \gamma_4 \mu_m^4 \|w_{i-1}\|^4}{1 - \beta} \\
+ \frac{8p_5 \gamma_2 \mu_m^2}{(1 - \beta)^2} \left( (\beta + p_3 \mu_m^2) \|\tilde{w}_{i-1}\|^2 + p_4 \mu_m^2 \|w_{i-1}\|^2 \right)^2 (\|\tilde{w}_{i-1}\|^2 + \|w_{i-1}\|^2) \\
+ \frac{8p_2 \mu_m^2 \sigma_s^2}{(1 - \beta)^2} \left( (\beta + p_3 \mu_m^2) \|\tilde{w}_{i-1}\|^2 + p_4 \mu_m^2 \|w_{i-1}\|^2 \right) + \frac{3\mu_s^4 \sigma_s^4}{(1 - \beta)^4} \\
= (\beta + p_7 \mu_m^4) \|\tilde{w}_{i-1}\|^4 + p_8 \mu_m^2 \|\tilde{w}_{i-1}\|^4 + \frac{3\sigma_s^4 \mu_m^4}{(1 - \beta)^4} + p_9 \mu_m^2 \|w_{i-1}\|^4 + p_{10} \mu_m^4 \|\tilde{w}_{i-1}\|^4 \\
+ p_{11} \mu_m^2 \|\tilde{w}_{i-1}\|^2 \|\tilde{w}_{i-1}\|^2 + p_{12} \mu_m^2 \|w_{i-1}\|^2 + p_{13} \mu_m^4 \|\tilde{w}_{i-1}\|^2, \tag{158} \]

where

\[ p_7 = p_1 + \frac{3p_6 \gamma_4}{(1 - \beta)^4}, \quad p_8 = p_2 + \frac{3p_6 \gamma_4}{(1 - \beta)^4}, \quad p_{10} = \frac{8p_4 \gamma_2}{(1 - \beta)^2}, \quad p_{13} = \frac{8p_4 \sigma_s^2}{(1 - \beta)^2}, \]

and \( p_9, p_{11}, p_{12} \) are some positive constants such that

\[ \frac{8p_5 \gamma_2 \mu_m^2}{(1 - \beta)^2} (\beta + p_3 \mu_m^2) \leq p_9 \mu_m^2, \tag{159} \]

\[ \frac{8p_5 \gamma_2 \mu_m^2}{(1 - \beta)^2} (\beta + p_3 \mu_m^2) + \frac{8p_4 \gamma_2 \mu_m^4}{(1 - \beta)^2} \leq p_{11} \mu_m^2, \tag{160} \]

\[ \frac{8p_2 \mu_m^2 \sigma_s^2}{(1 - \beta)^2} (\beta + p_3 \mu_m^2) \leq p_{12} \mu_m^2. \tag{161} \]

Since

\[ p_{11} \mu_m^2 \|\tilde{w}_{i-1}\|^2 \|\tilde{w}_{i-1}\|^2 \leq \frac{p_{11}}{2} \mu_m^2 (\|\tilde{w}_{i-1}\|^4 + \|\tilde{w}_{i-1}\|^4), \tag{162} \]

then inequality (158) reduces to

\[
E[\|\tilde{w}_i\|^4 | \mathcal{F}_{i-1}] \leq a \|\tilde{w}_{i-1}\|^4 + b \|\tilde{w}_{i-1}\|^4 + c \|w_{i-1}\|^2 + d \|\tilde{w}_{i-1}\|^2 + e, \tag{163} \]

38
where
\[ a = \beta + O(\mu_m^2), \quad b = O(\mu_m^2), \quad c = O(\mu_m^2), \quad d = O(\mu_m^4), \quad e = O(\mu_m^4) \] (164)

Now taking expectation of both sides of (162), we have
\[ \mathbb{E}[\|\hat{w}_i\|^4] \leq a\mathbb{E}[\|\hat{w}_{i-1}\|^4] + b\mathbb{E}[\|\hat{w}_{i-1}\|^4] + c\mathbb{E}[\|\hat{w}_{i-1}\|^2] + d\mathbb{E}[\|\hat{w}_{i-1}\|^2] + e. \] (165)

Now, Recall that in (38) we established
\[ \limsup_{i \to \infty} \mathbb{E}[\|\hat{w}_i\|^4] = O(\mu_m^2). \] (166)

Let \( C = \limsup_{i \to \infty} \mathbb{E}[\|\hat{w}_i\|^4] \). The above equation states that \( C \) exists and \( C = O(\mu_m^2) \). From the definition of the limit superior operation, we know that
\[ \limsup_{i \to \infty} \mathbb{E}[\|\hat{w}_i\|^4] = \lim \left( \sup_{n \geq i} \mathbb{E}[\|\hat{w}_n\|^4] \right). \] (167)

Let \( z(i) = \sup_{n \geq i} \mathbb{E}[\|\hat{w}_n\|^4] \). Then, we have \( \lim_{i \to \infty} z(i) = C \), which implies that \( z(i) \) is a bounded sequence, i.e., \( z(i) \leq D_1 \) for all \( i \) and for some constant \( D_1 \). In particular, it holds that \( \sup_{n \geq 0} \mathbb{E}[\|\hat{w}_n\|^4] \leq D_1 \). Similarly from (35) we obtain \( \sup_{n \geq 0} \mathbb{E}[\|\hat{w}_n\|^2] \leq D_2 \) and \( \sup_{n \geq 0} \mathbb{E}[\|\hat{w}_n\|^2] \leq D_3 \) for some constant \( D_2 \) and \( D_3 \). Combining these facts with (37), namely, \( \mathbb{E}[\|\hat{w}_i\|^2] = O(\mu_m^2) \) for \( i = 0, 1, 2, \ldots \), and substituting into (165) we conclude that
\[ \mathbb{E}[\|\hat{w}_1\|^4] \leq (\beta + q_1 \mu_m^2) \mathbb{E}[\|\hat{w}_{1-1}\|^4] + q_2 \mu_m^2, \] (168)

for some constants \( q_1 \) and \( q_2 \). Following the same arguments used in (144)–(148), there must exist a constant \( \rho < 1 \) such that
\[ \mathbb{E}[\|\hat{w}_i\|^4] \leq \rho^{i+1} \mathbb{E}[\|\hat{w}_{i-1}\|^4] + q_2 \mu_m^2 \sum_{j=0}^{i} \rho^j \leq \rho^{i+1} \mathbb{E}[\|\hat{w}_{i-1}\|^4] + \frac{q_2 \mu_m^2}{1 - \rho} \leq \frac{1}{1 - \rho} \mathbb{E}[\|\hat{w}_{i-1}\|^4] + \frac{q_2 \mu_m^2}{1 - \rho}. \] (169)

Recall from (149) that \( \mathbb{E}[\|\hat{w}_{i-1}\|^4] = c \mu_m^4 \), where \( c = \mathbb{E}[\|\nabla w Q(w_{-2}; \theta_{-1})\|^4] \). Substituting into (169) we reach that \( \mathbb{E}[\|\hat{w}_i\|^4] = O(\mu_m^2), \forall i \geq 0 \).

On the other hand, from (35) and (166), the recursion (165) will reduce to
\[ \limsup_{i \to \infty} \mathbb{E}[\|\hat{w}_i\|^4] \leq \rho \limsup_{i \to \infty} \mathbb{E}[\|\hat{w}_{i-1}\|^4] + q \mu_m^4, \] (170)

where \( \rho \) is the same as in (169) and \( q \) is some constant. It then follows that
\[ \limsup_{i \to \infty} \mathbb{E}[\|\hat{w}_i\|^4] \leq \frac{q}{1 - \rho} \mu_m^4 = O(\mu_m^4). \] (171)
Appendix F. Proof of Theorem 8

Subtracting (51) and (57) we get
\[ \hat{w}_i - \bar{x}_i = (I_M - \mu R_u)(\hat{w}_{i-1} - \bar{x}_{i-1}) + \mu (s_i(\psi_{i-1}) - s_i(x_{i-1})) + \mu \beta' R_u \hat{w}_{i-1}, \]  
where
\[ s_i(\psi_{i-1}) = (R_u - u_iu_i^T)\tilde{\psi}_{i-1} - u_i\nu(i), \quad s_i(x_{i-1}) = (R_u - u_iu_i^T)\bar{x}_{i-1} - u_i\nu(i). \]  
Substituting into (172) gives
\[ \hat{w}_i - \bar{x}_i = (I_M - \mu R_u)(\hat{w}_{i-1} - \bar{x}_{i-1}) + \mu(R_u - u_iu_i^T)(\tilde{\psi}_{i-1} - \bar{x}_{i-1}) + \mu \beta' R_u \hat{w}_{i-1}. \]  
Now note that in the quadratic case, the Hessian matrix of \( J(w) \) is equal to \( R_u \). It follows that condition (10) is satisfied with the identifications \( \nu = \lambda_{\text{min}}(R_u), \delta = \lambda_{\text{max}}(R_u) \). Let \( t \in (0, 1) \). By squaring (174) and taking expectations, and applying Jensen’s inequality, we obtain
\[
\begin{align*}
\mathbb{E}\|\hat{w}_i - \bar{x}_i\|^2 &\leq \mathbb{E}\|(I_M - \mu R_u)(\hat{w}_{i-1} - \bar{x}_{i-1}) + \mu \beta' R_u \hat{w}_{i-1}\|^2 + \mu^2 \mathbb{E}\|R_u - u_iu_i^T\|^2 \mathbb{E}\|\tilde{\psi}_{i-1} - \bar{x}_{i-1}\|^2 \\
&\leq \frac{1}{1-t}(1 - \mu \nu)^2 \mathbb{E}\|\hat{w}_{i-1} - \bar{x}_{i-1}\|^2 + \frac{1}{t} \mu^2 \beta'^2 \delta^2 \mathbb{E}\|\hat{w}_{i-1}\|^2 + b_1 \mu^2 \mathbb{E}\|\tilde{\psi}_{i-1} - \bar{x}_{i-1}\|^2 \\
&\leq (1 - \mu \nu)\mathbb{E}\|\hat{w}_{i-1} - \bar{x}_{i-1}\|^2 + \frac{\mu \beta'^2 \delta^2}{\nu} \mathbb{E}\|\hat{w}_{i-1}\|^2 + b_1 \mu^2 \mathbb{E}\|\tilde{\psi}_{i-1} - \bar{x}_{i-1}\|^2,
\end{align*}
\]  
where (a) holds because of Jensen’s inequality and we let \( b_1 = \mathbb{E}\|R_u - u_iu_i^T\|^2 \), and (b) holds by choosing \( t = \mu \nu \). To bound the last term in the above relation, we use (109) to note that
\[ \tilde{\psi}_i - \bar{x}_i = \hat{w}_i + \beta_1(\hat{w}_i - \bar{w}_{i-1}) - \bar{x}_i = (\hat{w}_i - \bar{x}_i) - \beta_1(\bar{w}_{i-1} - \bar{w}_i) \]  
On the other hand, from (28) we have
\[ \hat{w}_i - \bar{x}_i = \frac{1}{1-\beta}(\bar{w}_i - \bar{x}_i) - \frac{\beta}{1-\beta}(\bar{w}_{i-1} - \bar{x}_i) = (\hat{w}_i - \bar{x}_i) - \frac{\beta}{1-\beta}(\bar{w}_{i-1} - \bar{w}_i). \]  
so that
\[ \tilde{\psi}_i - \bar{x}_i = \hat{w}_i - \bar{x}_i + \frac{\beta - \beta_1 + \beta \beta_1}{1-\beta}(\bar{w}_{i-1} - \bar{w}_i) = \hat{w}_i - \bar{x}_i + \frac{\beta'}{1-\beta}(\bar{w}_{i-1} - \bar{w}_i) = \hat{w}_i - \bar{x}_i - \beta' \hat{w}_i. \]  
where we used the definition for \( \beta' \) from (30) and the definition for \( \hat{w}_i \) from (28). Therefore, from Jensen’s inequality again, we get
\[ \mathbb{E}\|\tilde{\psi}_i - \bar{x}_i\|^2 \leq 2 \mathbb{E}\|\hat{w}_i - \bar{x}_i\|^2 + 2 \beta'^2 \mathbb{E}\|\hat{w}_i\|^2. \]  

Substituting into (175) gives
\[
\mathbb{E}||\tilde{w}_i - \bar{x}_i||^2 \leq (1 - \mu\nu + 2b_1\mu^2)\mathbb{E}||\tilde{w}_{i-1} - \bar{x}_{i-1}||^2 + \left(\frac{\mu^2\beta^2}{\nu} + 2b_1\beta^2\mu^2\right)\mathbb{E}||\tilde{w}_{i-1}||^2.
\] (180)

Recall from Corollary 5 that \(\mathbb{E}||\tilde{w}_i||^2 \leq c\mu^2\) for each iteration \(i = 0, 1, 2, 3, \ldots\). This fact, together with the above inequality, leads to
\[
\mathbb{E}||\tilde{w}_i - \bar{x}_i||^2 \leq (1 - \mu\nu + 2b_1\mu^2)\mathbb{E}||\tilde{w}_{i-1} - \bar{x}_{i-1}||^2 + b_2\mu^2,
\] (181)
for some constant \(b_2\). Note that
\[
\alpha \triangleq 1 - \mu\nu + 2b_1\mu^2 = 1 - \frac{1}{2}\mu\nu + \left(2b_1\mu^2 - \frac{1}{2}\mu\nu\right).
\] (182)

It is clear that we can choose a sufficiently small \(\mu\) for the last term between brackets to become negative, in which case
\[
\alpha \leq 1 - \frac{1}{2}\mu\nu < 1.
\] (183)

Let \(\rho = 1 - \mu\nu/2\). It then holds that
\[
\mathbb{E}||\tilde{w}_i - \bar{x}_i||^2 \leq \rho\mathbb{E}||\tilde{w}_{i-1} - \bar{x}_{i-1}||^2 + b_2\mu^3.
\] (184)

Recall from the first equation in (177) that for \(i = -1\):
\[
\tilde{w}_{-1} - \bar{x}_{-1} = \frac{1}{1-\beta}(\tilde{w}_{-1} - \bar{x}_{-1}) - \frac{\beta}{1-\beta}(\tilde{w}_{-2} - \bar{x}_{-1}).
\] (185)

Now, using the assumption that the momentum and standard recursions started from the same initial states, \(\bar{w}_{-2} = \bar{x}_{-1}\) and \(\bar{w}_{-1} = \bar{w}_{-2} = \mu_m\nabla w Q(\bar{w}_{-2}; d(-1), \bar{u}_{-1})\), and recall \(\mu = \mu_m/(1 - \beta)\), then we have
\[
\tilde{w}_{-1} - \bar{x}_{-1} = \frac{1}{1-\beta}(\tilde{w}_{-1} - \bar{w}_{-2}) = \mu \nabla w Q(\bar{w}_{-2}; d(-1), \bar{u}_{-1}).
\] (186)

Therefore, it holds that
\[
\mathbb{E}||\tilde{w}_{-1} - \bar{x}_{-1}||^2 = c\mu^2,
\] (187)
where \(c = \mathbb{E}||\nabla w Q(\bar{w}_{-2}; d(-1), \bar{u}_{-1})||^2\). On the other hand, it follows by iterating (184) that
\[
\mathbb{E}||\tilde{w}_i - \bar{x}_i||^2 \leq \rho^{i+1}\mathbb{E}||\tilde{w}_{i-1} - \bar{x}_{i-1}||^2 + \sum_{j=0}^{i} \rho^j b_2\mu^3 \leq c\rho^{i+1}\mu^2 + \frac{b_2\mu^3}{1-\rho}
\]
\[
\leq c\rho^{i+1}\mu^2 + \frac{2b_2\mu^3}{\nu} = c\rho^{i+1}\mu^2 + \frac{2b_2\mu^2}{\nu} \leq b_3\mu^2.
\] (188)

for some constant \(b_3\). Furthermore, from (177) we have
\[
\mathbb{E}||\tilde{w}_i - \bar{x}_i||^2 \leq 2\mathbb{E}||\tilde{w}_i - \bar{x}_i||^2 + \frac{2\beta^2}{(1-\beta)^2}\mathbb{E}||\tilde{w}_i||^2.
\] (189)

Since we have shown that \(\mathbb{E}||\tilde{w}_i - \bar{x}_i||^2 \leq b_3\mu^2\) in (188) and \(\mathbb{E}||\tilde{w}_i||^2 \leq c\mu^2\) in (181) for each iteration \(i = 0, 1, 2, 3, \ldots\), and it is obvious that \(\mathbb{E}||\tilde{w}_i - \bar{x}_i||^2 = \mathbb{E}||\tilde{w}_i - \bar{x}_i||^2\), we then arrive at the desired conclusion (58).

41
Appendix G. Verifying Assumptions 4 and 5

Least-mean-squares problem. Consider first the mean-squares cost \[ (13). \] Since in this case \( H_{i-1} = R_{i-1} = R_u \), we find that Assumption 5 holds automatically. With regards to Assumption 4 at any iteration \( i \), we have

\[
\mathbf{s}_i(w_{i-1}) - \mathbf{s}_i(x_{i-1}) = (R_u - \mathbf{u}_i \mathbf{u}_i^T)(\bar{w}_{i-1} - \bar{x}_{i-1}).
\] (190)

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

\[
\mathbb{E}[\|\mathbf{s}_i(w_{i-1}) - \mathbf{s}_i(x_{i-1})\|^2 | \mathcal{F}_{i-1}] \leq \xi_1 \|\bar{w}_{i-1} - \bar{x}_{i-1}\|^2,
\] (191)

where \( \xi_1 = \mathbb{E}\|R_u - \mathbf{u}_i \mathbf{u}_i^T\|^2 \). Similarly,

\[
\mathbb{E}[\|\mathbf{s}_i(w_{i-1}) - \mathbf{s}_i(x_{i-1})\|^4 | \mathcal{F}_{i-1}] \leq \xi_2 \|\bar{w}_{i-1} - \bar{x}_{i-1}\|^4,
\] (192)

where \( \xi_2 = \mathbb{E}\|R_u - \mathbf{u}_i \mathbf{u}_i^T\|^4 \). Therefore, Assumption 4 holds.

Regularized logistic regression. Consider next the regularized logistic regression risk

\[
J(w) \triangleq \frac{\rho}{2}\|w\|^2 + \mathbb{E}\left\{ \ln \left[ 1 + \exp(-\gamma(i)h_i^T w) \right] \right\},
\] (193)

where \( h_i \in \mathbb{R}^M \) is a streaming sequence of independent feature vectors with \( R_h = \mathbb{E}h_i h_i^T > 0 \), and \( \gamma(i) \in \{-1,+1\} \) is a streaming sequence of class labels. We assume the random processes \( \{\gamma(i), h_i\} \) are wide-sense stationary. Moreover, \( \rho > 0 \) is a regularization parameter. We first verify the feasibility of Assumption 4. Note that the approximate gradient vector is given by:

\[
\hat{\nabla}_w J(w) = \rho w - \frac{\exp(-\gamma(i)h_i^T w)}{1 + \exp(-\gamma(i)h_i^T w)} \gamma(i)h_i
\] (194)

and, hence,

\[
\hat{\nabla}_w J(\psi_{i-1}) - \hat{\nabla}_w J(x_{i-1}) \\
\leq \rho \|\psi_{i-1} - x_{i-1}\| + \|h_i\| \left\| \frac{\exp(-\gamma(i)h_i^T \psi_{i-1})}{1 + \exp(-\gamma(i)h_i^T \psi_{i-1})} - \frac{\exp(-\gamma(i)h_i^T x_{i-1})}{1 + \exp(-\gamma(i)h_i^T x_{i-1})} \right\|
\] (195)

Note that

\[
\left\| \frac{\exp(-\gamma(i)h_i^T \psi_{i-1})}{1 + \exp(-\gamma(i)h_i^T \psi_{i-1})} - \frac{\exp(-\gamma(i)h_i^T x_{i-1})}{1 + \exp(-\gamma(i)h_i^T x_{i-1})} \right\|
\equiv \left\| \frac{\exp(-\gamma(i)h_i^T x_{i-1}) - \exp(-\gamma(i)h_i^T \psi_{i-1})}{1 + \exp(-\gamma(i)h_i^T x_{i-1})} \frac{1}{1 + \exp(-\gamma(i)h_i^T \psi_{i-1})} \right\|
\leq \left\| \frac{\exp(-\gamma(i)h_i^T x_{i-1}) - \exp(-\gamma(i)h_i^T \psi_{i-1})}{\exp(-\gamma(i)h_i^T x_{i-1}) + \exp(-\gamma(i)h_i^T \psi_{i-1})} \right\|
\equiv \left\| \frac{\exp(\gamma(i)h_i^T x_{i-1} - \psi_{i-1}) - \exp(\gamma(i)h_i^T 2x_{i-1} - \psi_{i-1})}{\exp(\gamma(i)h_i^T x_{i-1} - \psi_{i-1}) + \exp(\gamma(i)h_i^T 2x_{i-1} - \psi_{i-1})} \right\|
\]
\[ \frac{1}{2} \| \eta_i \| \| \psi_{i-1} - x_{i-1} \|. \] (196)

where in the last inequality we used the property \(| \tanh(y) | \leq |y|, \forall y \in \mathbb{R}\). Substituting (196) into (195), we get

\[ \| \nabla w J(\psi_{i-1}) - \nabla w J(x_{i-1}) \| \leq \eta_{1,i} \| \psi_{i-1} - x_{i-1} \|, \] (197)

where \( \eta_{1,i} = \rho + \| h_i \|^2/2 \) is a random variable.

On the other hand, it is shown in Eq. (2.20) of [Save1, 2014a] that the Hessian matrix \( \nabla^2 w J(w) \) is upper bounded by \( \delta I_M \), where \( \delta = (\rho + \lambda_{\max}(R_h)) \). We conclude from Lemma E.3 in the same reference that \( \nabla w J(w) \) is Lipschitz continuous with modulus \( \delta \), i.e.,

\[ \| \nabla w J(\psi_{i-1}) - \nabla w J(x_{i-1}) \| \leq \delta \| \psi_{i-1} - x_{i-1} \|. \] (198)

Combining these results we get

\[ \| s_i(\psi_{i-1}) - s_i(x_{i-1}) \| = \left\| \left( \nabla J(\psi_{i-1}) - \nabla J(x_{i-1}) \right) - \left( \nabla J(\psi_{i-1}) - \nabla J(x_{i-1}) \right) \right\| \leq \eta_i \| \psi_{i-1} - x_{i-1} \|. \] (199)

where \( \eta_i = \eta_{1,i} + \delta \) is a random variable. Since the \( \{ h_i \} \) are independent feature vectors and \( \eta_i \) is only related to \( h_i \), it follows that

\[ \mathbb{E}[\| s_i(\psi_{i-1}) - s_i(x_{i-1}) \|^2 | \mathcal{F}_{i-1}] \leq \xi_i \| \psi_{i-1} - x_{i-1} \|^2, \] (200)

\[ \mathbb{E}[\| s_i(\psi_{i-1}) - s_i(x_{i-1}) \|^4 | \mathcal{F}_{i-1}] \leq \xi_2 \| \psi_{i-1} - x_{i-1} \|^4, \] (201)

where \( \xi_1 = \mathbb{E} \eta_i^2 \) and \( \xi_2 = \mathbb{E} \eta_i^4 \).

Next we check the feasibility of Assumption 5. For simplicity, we write \( \gamma \) instead of \( \gamma(i) \). It can be verified that for the cost function \( J(w) \) in (193):

\[ \nabla^2 w J(w) = \rho I_M + \mathbb{E} \left\{ h_i h_i^T \left( \frac{\exp(-\gamma h_i^T w)}{1 + \exp(-\gamma h_i^T w)} \right) \right\}. \] (202)

Now, for any two variables \( w_1 \) and \( w_2 \) we have

\[ \| \nabla^2 w J(w_1) - \nabla^2 w J(w_2) \| \]

\[ \leq \mathbb{E} \left\{ h_i h_i^T \left( \frac{\exp(-\gamma h_i^T w_1)}{1 + \exp(-\gamma h_i^T w_1)} - \frac{\exp(-\gamma h_i^T w_2)}{1 + \exp(-\gamma h_i^T w_2)} \right) \right\} \]

\[ \leq \mathbb{E} \left\{ \frac{\exp(-\gamma h_i^T w_1)}{1 + \exp(-\gamma h_i^T w_1)} - \frac{\exp(-\gamma h_i^T w_2)}{1 + \exp(-\gamma h_i^T w_2)} \right\} \] (203)

Let \( x_1 = -\gamma h_i^T w_1 \) and \( x_2 = -\gamma h_i^T w_2 \). Then,

\[ \left\| \frac{\exp(-\gamma h_i^T w_1)}{1 + \exp(-\gamma h_i^T w_1)} - \frac{\exp(-\gamma h_i^T w_2)}{1 + \exp(-\gamma h_i^T w_2)} \right\| \]
where (a) holds because of the following two facts:

\[
\exp(x_1)[1 + \exp(x_2)]^2 - \exp(x_2)[1 + \exp(x_1)]^2 \\
= \exp(x_1) + \exp(x_1 + 2x_2) - \exp(x_2) - \exp(x_2 + 2x_1) \\
= \exp(x_1 + 2x_2)[\exp(-x_2) + \exp(x_2) - \exp(-x_1) - \exp(x_1)],
\]

and

\[
[1 + \exp(x_1)]^2[1 + \exp(x_2)]^2 \\
= (1 + 2\exp(x_1) + \exp(2x_1))(1 + 2\exp(x_2) + \exp(2x_2)) \\
\geq 2\exp(x_1) + 2\exp(x_2) + 2\exp(x_1 + 2x_2) + 2\exp(x_2 + 2x_1) \\
= 2\exp(x_1 + 2x_2)[\exp(-x_2) + \exp(x_2) + \exp(-x_1) + \exp(x_1)].
\]

In addition, (b) holds if we extract \(\exp(-\frac{x_1 + x_2}{2})\) and \(\exp(\frac{x_1 + x_2}{2})\) from both the denominator and numerator of the first and second terms respectively.

Using the definitions for \(x_1\) and \(x_2\), this last expression gives

\[
\|\tanh[(x_2 - x_1)/2]\| = \|\tanh \left(\frac{1}{2} \gamma h_i^T(w_2 - w_1)\right)\| \leq \frac{1}{2} \|h_i\||w_2 - w_1|. 
\]

Substituting (207) into (203), we obtain \(\|\nabla_{w_1}^2 J(w_1) - \nabla_{w_2}^2 J(w_2)\| \leq \kappa\|w_1 - w_2\|\), where \(\kappa = \mathbb{E}[\|h_i h_i^T\|\|h_i\|]/2\). Therefore, Assumption 5 holds.

Appendix H. Proof of Lemma 9

We refer to relation (92) and apply inequality (152) to reach

\[
\mathbb{E}[\|\bar{w}_i - \bar{x}_i\|^4 | F_{i-1}] \\
= \| (I_M - \mu H_{i-1})(\bar{w}_{i-1} - \bar{x}_{i-1}) + \mu (R_{i-1} - H_{i-1})\bar{x}_{i-1} + \mu \beta' H_{i-1}\bar{w}_{i-1} \|^4
\]

44
where (a) holds because the facts that for any $a, b, c \in \mathbb{R}^m$,

$$
\|a + b + c\|^4 = \|(1-t)\frac{1}{1-t}a + t\frac{1}{t}(b + c)\|^4
\leq (1-t)\frac{1}{1-t}\|a\|^4 + t\frac{1}{t}\|(b + c)\|^4 = \frac{1}{(1-t)^3}\|a\|^4 + \frac{4}{(1-t)^3}\|b + c\|^4
\leq \frac{1}{(1-t)^3}\|a\|^4 + \frac{8}{t^3}\|b\|^4 + \frac{8}{t^3}\|c\|^4,
$$

(209)

and

$$
\|a + b + c\|^2 = \|(1-t)\frac{1}{1-t}a + t\frac{1}{t}(b + c)\|^2
\leq (1-t)\|(1-t)\frac{1}{1-t}a\|^2 + t\frac{1}{t}\|(b + c)\|^2 = \frac{1}{1-t}\|a\|^2 + \frac{1}{t}\|b + c\|^2
\leq \frac{1}{1-t}\|a\|^2 + \frac{2}{t}\|b\|^2 + \frac{2}{t}\|c\|^2,
$$

(210)

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

To further simplify inequality (208), we first note that

$$
\|R_{i-1} - H_{i-1}\|^2 \leq 2(\|R_{i-1}\|^2 + \|H_{i-1}\|^2) \leq 4\delta^2,
$$

(211)

$$
\|R_{i-1} - H_{i-1}\|^4 \leq 8(\|R_{i-1}\|^4 + \|H_{i-1}\|^4) \leq 16\delta^4.
$$

(212)

As a result, we have

$$
\frac{8\mu}{\nu^3}\|(R_{i-1} - H_{i-1})\bar{x}_{i-1}\|^4 \leq q_1\mu\|\bar{x}_{i-1}\|^4, \quad \frac{2\mu}{\nu}\|(R_{i-1} - H_{i-1})\bar{x}_{i-1}\|^2 \leq q_2\mu\|\bar{x}_{i-1}\|^2,
$$

(213)

where $q_1 = 128\delta^4/\nu^3$ and $q_2 = 8\delta^2/\nu$. On the other hand, from conditions (66)–(67), we have

$$
3\mu^4\mathbb{E}[\|s_i(\psi_{i-1}) - s_i(x_{i-1})\|^4 | \mathcal{F}_{i-1}] \leq q_3\mu^4\|\psi_{i-1} - \bar{x}_{i-1}\|^2,
$$

(214)
\[
8\mu^2 \mathbb{E}[\|s_i(\psi_{i-1}) - s_i(x_{i-1})\|^2 | \mathcal{F}_{i-1}] \leq q_4 \mu^2 \|\tilde{\psi}_{i-1} - \tilde{x}_{i-1}\|^2,
\]
where \(q_3 = 3\xi_2\) and \(q_4 = 8\xi_1\). In addition, from (178) we get
\[
\|\tilde{\psi}_i - \tilde{x}_i\|^2 \leq 2\|\tilde{w}_i - \tilde{x}_i\|^2 + 2\beta^2\|\tilde{w}_i\|^2, \quad \|\tilde{\psi}_i - \tilde{x}_i\|^4 \leq 8\|\tilde{w}_i - \tilde{x}_i\|^4 + 8\beta^4\|\tilde{w}_i\|^4.
\]
Combining (214) - (216), we have
\[
3\mu^4 \mathbb{E}[\|s_i(\psi_{i-1}) - s_i(x_{i-1})\|^4 | \mathcal{F}_{i-1}] \leq q_5 \mu^4 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^4 + q_6 \mu^4 \|\tilde{w}_{i-1}\|^4,
\]
\[
8\mu^2 \mathbb{E}[\|s_i(\psi_{i-1}) - s_i(x_{i-1})\|^2 | \mathcal{F}_{i-1}] \leq q_7 \mu^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^2 + q_8 \mu^2 \|\tilde{w}_{i-1}\|^2,
\]
for some constants \(q_5, q_6, q_7, q_8\). In this way, relation (208) becomes
\[
\mathbb{E}[\|\tilde{w}_i - \tilde{x}_i\|^4 | \mathcal{F}_{i-1}]
\]
\[
\leq (1 - \mu\nu)\|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^4 + q_1 \mu\|\tilde{x}_{i-1}\|^4 + q_2 \mu\|\tilde{w}_{i-1}\|^4 + \left(1 - \mu\nu\right)\|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^2
\]
\[
+ q_{10} \mu \|\tilde{w}_{i-1}\|^2 + q_{12} \mu \|\tilde{x}_{i-1}\|^2 \left[q_7 \mu^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^2 + q_8 \mu^2 \|\tilde{w}_{i-1}\|^2\right]
\]
\[
+ q_5 \mu^4 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^4 + q_6 \mu^4 \|\tilde{w}_{i-1}\|^4,
\]
where \(q_9 = 8\beta^4\delta^4/\nu^3\) and \(q_{10} = 2\mu\beta^2\delta^2/\nu\). Now note that
\[
\left[(1 - \mu\nu)\|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^2 + q_1 \mu\|\tilde{x}_{i-1}\|^2 + q_2 \mu\|\tilde{w}_{i-1}\|^2 \left[q_7 \mu^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^2 + q_8 \mu^2 \|\tilde{w}_{i-1}\|^2\right]
\]
\[
+ q_7 q_{10} \mu^3 \|\tilde{w}_{i-1}\|^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^2 + q_8 q_{10} \mu^3 \|\tilde{w}_{i-1}\|^4
\]
\[
+ q_2 q_7 \mu^3 \|\tilde{x}_{i-1}\|^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^2 + q_2 q_8 \mu^3 \|\tilde{x}_{i-1}\|^4 \|\tilde{w}_{i-1}\|^2\right]
\]
\[
\leq q_{11} \mu^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^4 + q_{12} \mu^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^4 + q_{13} \mu^2 \|\tilde{w}_{i-1}\|^4 + q_{14} \mu^2 \|\tilde{w}_{i-1}\|^4
\]
\[
+ q_{13} \mu^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^4 + q_{14} \mu^2 \|\tilde{w}_{i-1}\|^4 + q_{15} \mu^2 \|\tilde{x}_{i-1}\|^4
\]
\[
+ q_{15} \mu^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^4 + q_{16} \mu^2 \|\tilde{x}_{i-1}\|^4 + q_{16} \mu^2 \|\tilde{x}_{i-1}\|^4
\]
\[
\leq q_{17} \mu^2 \|\tilde{w}_{i-1} - \tilde{x}_{i-1}\|^4 + q_{18} \mu^2 \|\tilde{w}_{i-1}\|^4 + q_{19} \mu^2 \|\tilde{x}_{i-1}\|^4,
\]
where inequality (a) holds because of the fact that for any \(a > 0\) and \(b > 0\), it holds that \(ab < 2ab \leq a^2 + b^2\). Besides, in (220) \(q_{11}\) and \(q_{12}\) are some positive constants such that
\[
q_7 \mu^2 (1 - \mu\nu) \leq q_{11} \mu^2, \quad q_8 \mu^2 (1 - \mu\nu) \leq q_{12} \mu^2,
\]
and
\[
q_{13} = q_7 q_{10}, \quad q_{14} = q_8 q_{10}, \quad q_{15} = q_2 q_7, \quad q_{16} = q_2 q_8,
\]
where \(q_{17}, q_{18}\) and \(q_{19}\) are some positive constants such that
\[
q_{11} \mu^2 + q_{12} \mu^2 + q_{13} \mu^2 \leq q_{17} \mu^2,
\]
\[
q_{12} \mu^2 + q_{13} \mu^2 + q_{14} \mu^2 \leq q_{18} \mu^2,
\]
\[
q_{15} + q_{16} = q_{19}.
\]
Now, substituting (220) into (219), we reach

$$
\mathbb{E}[\|\hat{\omega}_i - \bar{x}_i\|^4 | \mathcal{F}_{i-1}]
\leq (1-\mu \nu + q_{17} \mu^2 + q_{19} \mu^4) \|\omega_{i-1} - \bar{x}_{i-1}\|^4 + (q_1 \mu + q_{19} \mu^3) \|\bar{x}_{i-1}\|^4 + (q_9 \mu + q_{19} \mu^4 + q_{19} \mu^3) \|\hat{\omega}_{i-1}\|^4
\leq (1 - \frac{1}{2} \mu \nu) \|\hat{\omega}_{i-1} - \bar{x}_{i-1}\|^4 + q_{20} \mu \|\bar{x}_{i-1}\|^4 + q_{21} \mu \|\hat{\omega}_{i-1}\|^4.
$$

(224)

where $q_{20}$ and $q_{21}$ are some positive constants such that

$$
q_1 \mu + q_{19} \mu^3 \leq q_{20} \mu, \quad q_9 \mu + q_{19} \mu^4 \leq q_{21} \mu.
$$

and $\mu$ is sufficiently small such that $1 - \mu \nu + q_{17} \mu^2 + q_{19} \mu^4 \leq 1 - \frac{1}{2} \mu \nu$.

Taking expectations of both sides of (224), we have

$$
\mathbb{E}[\|\hat{\omega}_i - \bar{x}_i\|^4] \leq (1 - \frac{1}{2} \mu \nu) \mathbb{E}[\|\hat{\omega}_{i-1} - \bar{x}_{i-1}\|^4 + q_{20} \mu \mathbb{E}[\|\bar{x}_{i-1}\|^4 + q_{21} \mu \mathbb{E}[\|\hat{\omega}_{i-1}\|^4].
$$

(225)

Recall from Corollary 7 that $\mathbb{E}[\|\hat{\omega}_i\|^4] = O(\mu^2)$ for $i \geq 0$, then there exists some constant $q_{22}$ such that

$$
\mathbb{E}[\|\hat{\omega}_i - \bar{x}_i\|^4] \leq (1 - \frac{1}{2} \mu \nu) \mathbb{E}[\|\hat{\omega}_{i-1} - \bar{x}_{i-1}\|^4 + q_{20} \mu \mathbb{E}[\|\bar{x}_{i-1}\|^4 + q_{21} \mu \mathbb{E}[\|\hat{\omega}_{i-1}\|^4 + q_{22} \mu^3.
$$

(226)

Now recall from (19) that

$$
\mathbb{E}[\|\bar{x}_i\|^4] \leq (1 - \mu \nu)^{i+1} \mathbb{E}[\|\bar{x}_{i-1}\|^4 + c \mu^2,
$$

(227)

where $c$ is some constant. Let $\rho = 1 - \frac{1}{2} \mu \nu$, then it holds that

$$
\mathbb{E}[\|\bar{x}_i\|^4] \leq \rho^{i+1} \mathbb{E}[\|\bar{x}_{i-1}\|^4 + c \mu^2,
$$

(228)

Substituting (228) into (225), we get

$$
\mathbb{E}[\|\hat{\omega}_i - \bar{x}_i\|^4] \leq \rho \mathbb{E}[\|\hat{\omega}_{i-1} - \bar{x}_{i-1}\|^4 + q_{23} \rho^i \mu + q_{24} \mu^3,
$$

(229)

where $q_{23} = q_{20} \mathbb{E}[\|\bar{x}_{i-1}\|^4$ and $q_{24} = q_{22} + q_{20} c$. If we iterate this recursion we obtain

$$
\mathbb{E}[\|\hat{\omega}_i - \bar{x}_i\|^4]
\leq \rho^i \mathbb{E}[\|\hat{\omega}_{i-1} - \bar{x}_{i-1}\|^4 + (i + 1) q_{23} \rho^i \mu + q_{24} \mu^3 \sum_{s=0}^{i} \rho^s
\leq \rho^i \mathbb{E}[\|\hat{\omega}_{i-1} - \bar{x}_{i-1}\|^4 + (i + 1) q_{23} \rho^i \mu + \frac{q_{24} \mu^3}{1 - \rho}
\leq \rho^i \mathbb{E}[\|\hat{\omega}_{i-1} - \bar{x}_{i-1}\|^4 + (i + 1) q_{23} \rho^i \mu + q_{25} \mu^2,
$$

(230)

where $q_{25} = 2q_{24}/\nu$. Recall from (186) that $\hat{\omega}_{i-1} - \bar{x}_{i-1} = \mu \nabla w Q(\omega_{i-2}; \theta_{i-1})$. Then it holds that

$$
\mathbb{E}[\|\hat{\omega}_i - \bar{x}_i\|^4] = c \mu^4,
$$

(231)
where $c = \mathbb{E}\|\nabla_w Q(w_{t-2}; \theta_{t-1})\|^4$. Besides, recall from (103) that $(i+1)\rho^i \leq C$ for each iteration $i$. Now substitute the above two facts into relation (230), to conclude that for $i = 0, 1, 2, \ldots$

$$E\|\hat{w}_i - \tilde{x}_i\|^4 \leq c\rho^{i+1} \mu + q_{25} \mu^2 + q_{23} C \mu = O(\mu).$$

(232)

In addition, as $i \to \infty$, in (230) we have

$$q_{23}(i+1)\rho^i \mu \to 0$$

and we conclude that

$$\limsup_{i \to \infty} E\|\hat{w}_i - \tilde{x}_i\|^4 = q_{25} \mu^2 = O(\mu^2).$$

(233)

Furthermore, from (216) and from the fact that $E\|\hat{\psi}_i\|^4 = O(\mu^2)$, we conclude from (232) and (233) that

$$E\|\hat{\psi}_i - \tilde{x}_i\|^4 = O(\mu), \forall i = 0, 1, 2, \ldots$$

(234)

and

$$\limsup_{i \to \infty} E\|\hat{\psi}_i - \tilde{x}_i\|^4 = O(\mu^2).$$

(235)

**Appendix I. Proof of Lemma 10**

Under Assumption 5, we have

$$\|H_{i-1} - R_{i-1}\| \leq \left\| \int_0^1 \nabla^2 J(w^0 - r\hat{\psi}_{i-1})dr - \int_0^1 \nabla^2 J(w^0 - r\tilde{x}_{i-1})dr \right\|$$

$$\leq \int_0^1 \|\nabla^2 J(w^0 - r\hat{\psi}_{i-1}) - \nabla^2 J(w^0 - r\tilde{x}_{i-1})\|dr$$

$$\leq \int_0^1 \kappa r\|\hat{\psi}_{i-1} - \tilde{x}_{i-1}\|dr = \frac{\kappa}{2}\|\hat{\psi}_{i-1} - \tilde{x}_{i-1}\|. $$

(236)

As a result, it holds that

$$E\|R_{i-1} - H_{i-1}\|^4 \leq \bar{\kappa} E\|\hat{\psi}_{i-1} - \tilde{x}_{i-1}\|^4, $$

(237)

where $\bar{\kappa} = \kappa^4/16$. Combining this inequality with (234)–(235), we reach the desired bounds

$$E\|R_{i-1} - H_{i-1}\|^4 \leq C\mu, \forall i = 0, 1, 2, \ldots$$

(238)

and

$$\limsup_{i \to \infty} E\|R_{i-1} - H_{i-1}\|^4 = O(\mu^2).$$

(239)
Appendix J. Proof of Theorem 11

For (66) in Assumption 4, if we take expectation over $\mathcal{F}_{i-1}$ of both sides, it holds that

$$\mathbb{E}\left\| s_i(\psi_{i-1}) - s_i(x_{i-1}) \right\|^2 \leq \xi_1 \mathbb{E}\left\| \psi_{i-1} - x_{i-1} \right\|^2. \quad (240)$$

Combining the above fact and inequalities (64)–(65), we get

$$\mathbb{E}\left\| \tilde{w}_i - \tilde{x}_i \right\|^2 \leq (1 - \mu \nu) \mathbb{E}\left\| \tilde{w}_{i-1} - \tilde{x}_{i-1} \right\|^2 + r_1 \mu \mathbb{E}\left\| \tilde{w}_{i-1} \right\|^2 + 2\mu \sqrt{\mathbb{E}\left\| R_{i-1} - H_{i-1} \right\|^4 \mathbb{E}\left\| \tilde{x}_{i-1} \right\|^4} + \xi_1 \mu^2 \mathbb{E}\left\| \tilde{\psi}_{i-1} - \tilde{x}_{i-1} \right\|^2, \quad (241)$$

for some constant $r_1$. Likewise, from (179) we have

$$\mathbb{E}\left\| \tilde{\psi}_i - \tilde{x}_i \right\|^2 \leq 2 \mathbb{E}\left\| \tilde{w}_i - \tilde{x}_i \right\|^2 + r_2 \mathbb{E}\left\| \tilde{w}_i \right\|^2, \quad (242)$$

for some constant $r_2$. Substituting this second inequality along with (238) into (241) gives

$$\mathbb{E}\left\| \tilde{w}_i - \tilde{x}_i \right\|^2 \leq (1 - \mu \nu + 2\xi_1 \mu^2) \mathbb{E}\left\| \tilde{w}_{i-1} - \tilde{x}_{i-1} \right\|^2 + (r_1 \mu + r_2 \xi_1 \mu^2) \mathbb{E}\left\| \tilde{w}_{i-1} \right\|^2 + r_3 \mu^{3/2} \sqrt{\mathbb{E}\left\| \tilde{x}_{i-1} \right\|^4}, \quad (243)$$

where $r_3$ is some constant. Similar to the argument used in (182)–(184), a sufficiently small step-size $\mu$ exists such that

$$1 - \mu \nu + 2\xi_1 \mu^2 < 1 - \frac{\mu \nu}{2}, \quad (244)$$

Using bound (37) that $\mathbb{E}\left\| \tilde{w}_i \right\|^2 = O(\mu^2)$, expression (243) can be simplified to

$$\mathbb{E}\left\| \tilde{w}_i - \tilde{x}_i \right\|^2 \leq \rho \mathbb{E}\left\| \tilde{w}_{i-1} - \tilde{x}_{i-1} \right\|^2 + r_3 \mu^{3/2} \sqrt{\mathbb{E}\left\| \tilde{x}_{i-1} \right\|^4} + r_4 \mu^3, \quad (245)$$

where $r_4$ is some constant. Next, using (19) we have

$$\sqrt{\mathbb{E}\left\| \tilde{x}_i \right\|^4} \leq \sqrt{(1 - \mu \nu)^{i+1} \mathbb{E}\left\| \tilde{x}_{i-1} \right\|^4 + c \mu^2} \leq r_5 \sqrt{(1 - \mu \nu)^{i+1} + r_6 \mu} \leq r_5 \sqrt{\rho}^{i+1} + r_6 \mu. \quad (246)$$

where $r_5 = \sqrt{\mathbb{E}\left\| \tilde{x}_{i-1} \right\|^4}$ and $r_6 = \sqrt{c}$, and where we applied the inequality $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ for positive $a$ and $b$. Substituting (246) into (245), we obtain

$$\mathbb{E}\left\| \tilde{w}_i - \tilde{x}_i \right\|^2 \leq \rho \mathbb{E}\left\| \tilde{w}_{i-1} - \tilde{x}_{i-1} \right\|^2 + r_3 r_5 \mu^{5/2} + r_4 \mu^3 + r_3 r_5 \mu^{3/2} \sqrt{\rho}^i \leq \rho \mathbb{E}\left\| \tilde{w}_{i-1} - \tilde{x}_{i-1} \right\|^2 + O(\mu^{5/2}) + r_3 r_5 \mu^{3/2} \sqrt{\rho}^i. \quad (247)$$

Since $\rho = 1 - \mu \nu / 2 < 1$, it holds that $\rho < \sqrt{\rho}$, and hence (a) holds in (247). Iterating (247) gives

$$\mathbb{E}\left\| \tilde{w}_i - \tilde{x}_i \right\|^2 \leq (\sqrt{\rho})^{i+1} \mathbb{E}\left\| \tilde{w}_{i-1} - \tilde{x}_{i-1} \right\|^2 + r_7 \mu^{5/2} \sum_{s=0}^{i} (\sqrt{\rho})^s + r_3 r_5 (i + 1) \mu^{3/2} \sqrt{\rho}^i$$

49
\[
\leq (\sqrt{\rho})^{i+1}E\|\hat{w}_{i-1} - \tilde{x}_{i-1}\|^2 + \frac{r_7\mu^{5/2}}{1 - \sqrt{\rho}} + r_8\mu^{3/2}, \quad (248)
\]

where \(r_7\) and \(r_8\) are some constants and where we applied the fact that \((i + 1)(\sqrt{\rho})^i \leq D\). Next we check the order of the term \(r_7\mu^{5/2}/(1 - \sqrt{\rho})\). Since \(\mu\) is sufficiently small, we can approximate \(\sqrt{\rho} = \sqrt{1 - \frac{1}{2}\mu\nu} = 1 - \frac{1}{4}\mu\nu\). As a result, it holds that

\[
\frac{r_7\mu^{5/2}}{1 - \sqrt{\rho}} = O(\mu^{3/2}) \quad (249)
\]

Now recall that \(E\|\hat{w}_{i-1} - \tilde{x}_{i-1}\|^2 = c\mu^2\), finally we can show that

\[
E\|\hat{w}_{i} - \tilde{x}_{i}\|^2 = O(\mu^{3/2}), \quad \forall i = 0, 1, 2, \ldots \quad (250)
\]

This proves the first part of Theorem 11. Furthermore, using (20) and (239), we get

\[
\limsup_{i \rightarrow \infty} \sqrt{E\|R_{i-1} - H_{i-1}\|^4E\|\tilde{x}_{i-1}\|^4} = O(\mu^2). \quad (251)
\]

Using again (242), it then follows from (241) that for sufficiently small step-sizes:

\[
\limsup_{i \rightarrow \infty} E\|\hat{w}_{i} - \tilde{x}_{i}\|^2 \leq (1 - \mu\nu + 2\xi_1\mu^2) \limsup_{i \rightarrow \infty} E\|\hat{w}_{i-1} - \tilde{x}_{i-1}\|^2 + O(\mu^3)
\]

\[
\leq \left(1 - \frac{\mu\nu}{2}\right) \limsup_{i \rightarrow \infty} E\|\hat{w}_{i-1} - \tilde{x}_{i-1}\|^2 + O(\mu^3), \quad (252)
\]

which leads to

\[
\limsup_{i \rightarrow \infty} E\|\hat{w}_{i} - \tilde{x}_{i}\|^2 = O(\mu^2). \quad (253)
\]

This proves the second part of the Theorem 11.