Journal of Machine Learning Research 25 (2024) 1-56 Submitted 1/22; Revised 10/23; Published 1/24

On the Generalization of Stochastic Gradient Descent with Momentum

Ali Ramezani-Kebrya ALI@UIO.NO
Department of Informatics, University of Oslo and Visual Intelligence Centre
Integreat, Norwegian Centre for Knowledge-driven Machine Learning
Gaustadalléen 23B, Ole-Johan Dahls hus, 0373 Oslo, Norway

Kimon Antonakopoulos KIMON.ANTONAKOPOULOS@EPFL.CH
Laboratory for Information and Inference Systems (LIONS), EPFL
EPFL STI IEL LIONS, Station 11, CH-1015 Lausanne, Switzerland

Volkan Cevher VOLKAN.CEVHER@EPFL.CH
Laboratory for Information and Inference Systems (LIONS), EPFL
EPFL STI IEL LIONS, Station 11, CH-1015 Lausanne, Switzerland

Ashish Khisti AKHISTI@ECE.UTORONTO.CA
Department of Electrical and Computer Engineering, University of Toronto
40 St. George Street, Toronto, ON M5S 2E4, Canada

Ben Liang LIANG@ECE.UTORONTO.CA
Department of Electrical and Computer Engineering, University of Toronto
40 St. George Street, Toronto, ON M5S 2E4, Canada

Editor: Francesco Orabona

Abstract

While momentum-based accelerated variants of stochastic gradient descent (SGD) are widely used when training machine learning models, there is little theoretical understanding on the generalization error of such methods. In this work, we first show that there exists a convex loss function for which the stability gap for multiple epochs of SGD with standard heavy-ball momentum (SGDM) becomes unbounded. Then, for smooth Lipschitz loss functions, we analyze a modified momentum-based update rule, *i.e.,* SGD with early momentum (SGDEM) under a broad range of step-sizes, and show that it can train machine learning models for multiple epochs with a guarantee for generalization. Finally, for the special case of strongly convex loss functions, we find a range of momentum such that multiple epochs of standard SGDM, as a special form of SGDEM, also generalizes. Extending our results on generalization, we also develop an upper bound on the expected true risk, in terms of the number of training steps, sample size, and momentum. Our experimental evaluations verify the consistency between the numerical results and our theoretical bounds. SGDEM improves the generalization error of SGDM when training ResNet-18 on ImageNet in practical distributed settings.

Keywords: Uniform stability, generalization error, heavy-ball momentum, stochastic gradient descent, non-convex

1 Introduction

Stochastic gradient descent (SGD) and its variants are the most popular algorithms for training deep neural networks due to their support of efficient parallel implementations and excellent generalization performance (Krizhevsky et al., 2012; Wilson et al., 2017). To accelerate the convergence of SGD, a momentum term is often added in the iterative update of the stochastic gradient (Goodfellow et al., 2016). This approach has a long history, with proven benefits in various settings. The heavy-ball momentum method was first introduced by Polyak (1964) where a weighted version of the previous update is added to the current gradient update. Polyak (1964) motivated his method by its resemblance to a heavy ball moving in a potential well defined by the objective function. Momentum methods have been used to accelerate empirical risk minimization when training neural networks (Rumelhart et al., 1986). In particular, momentum methods are used for training deep neural networks with complex and nonconvex loss landscapes (Sutskever et al., 2013). Intuitively, adding momentum accelerates convergence by circumventing sharp curvatures and long ravines of the sub-level sets of the objective function (Wilson et al., 2021). Ochs et al. (2015) present an illustrative example to show that the momentum can potentially avoid local minima.

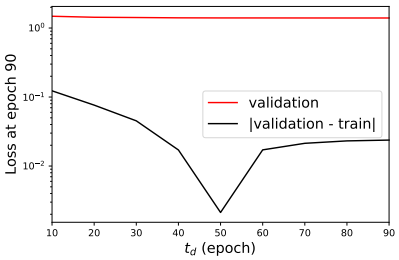

Figure 1: *Validation loss and generalization error of SGDEM when training ResNet-18 (He et al., 2016) on ImageNet (Deng et al., 2009) in a distributed setting with 4 GPUs under tuned step-size and global mini-batch size of 128. For each t_d, the momentum is set to $\mu_d = 0.9$ in the first t_d epochs and then zero for the next $90 - t_d$ epochs. SGDM is a special form of SGDEM with $t_d = 90$. The details are provided in Section 5 and Appendix L .*

Beyond convergence, the generalization of machine learning algorithms is a fundamental problem in learning theory. A classical framework used to study the generalization error in machine learning is PAC learning (Vapnik and Chervonenkis, 1971; Valiant, 1984). However, the associated bounds using this approach can be conservative. The connection between stability and generalization has been studied in the literature, which captures how the learning algorithm explores a hypothesis class (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Hardt et al., 2016). According to the definition of Bousquet and Elisseeff (2002), uniform stability requires the algorithm to generate almost the same predictions for all datasets that are different in only one example. Recently, this notion of uniform stability is leveraged to analyze the generalization error of SGD (Hardt et al., 2016). Hardt et al. (2016) have derived the stability bounds for SGD and analyzed its generalization for different loss functions. This is a substantial step forward, since SGD is widely used in many practical systems. However, the algorithms studied in these works do not include momentum.

In this work, we study SGD with momentum (SGDM). Although momentum methods are empirically observed to accelerate training in deep learning (Goodfellow et al., 2016), their effect on the generalization error is not well understood. Even though momentum is not studied in (Hardt et al., 2016), it is conjectured therein that momentum might speed up training but adversely impact generalization.

By providing a counter example, we show that the stability gap for multiple epochs of SGDM can become unbounded even for convex loss functions. This motivates us to consider a modified momentum-based update rule, called SGD with *early momentum* (SGDEM) where a momentum term is added in the earlier training steps. We show that SGDEM is guaranteed to generalize for smooth Lipschitz loss functions and any momentum. To the best of our knowledge, stability and generalization of SGDEM have not been considered in the existing literature. As Fig. 1 shows in a practical and distributed setting on ImageNet, while validation loss remains unaffected, the minimum generalization error happens if the momentum is applied for 50 epochs, which indicates that tuning momentum is useful to achieve the best generalization error.

We study the generalization error and true risk of SGDEM. In order to find an upper bound on the expected generalization error of SGDEM, we use the framework of uniform stability (Bousquet and Elisseeff, 2002; Hardt et al., 2016).

1.1 Main Contributions

In Section 3, we show that there exists a convex loss function for which the stability gap for multiple epochs of SGDM becomes unbounded. We introduce SGDEM and show that it is guaranteed to generalize for smooth Lipschitz loss functions. We obtain a bound on the generalization error of SGDEM that decreases inversely with the size of the training set. Our results show that the number of iterations can grow as n^l for a small $l > 1$ where n is the sample size, which explains why complicated models such as deep neural networks can be trained for multiple epochs of SGDEM while their generalization errors are limited. We also establish an explicit convergence rate for SGDEM and smooth Lipschitz loss functions under a broad range of hyperparameters including a general step-size rule that covers popular step-sizes in the optimization literature. Our convergence and generalization bounds capture the inherent trade-off between optimization and generalization.

In Section 4, we focus on the special case of strongly convex loss functions. In this case, we show that one can obtain a bound on the generalization error of standard SGDM, which suggests that this special form of SGDEM suffices for generalization. Our bound is independent of the number of training iterations and decreases inversely with the size of the training set. Finally, we establish an upper bound on the expected true risk of SGDM as a function of various problem parameters.

Our generalization bounds for both strongly convex and smooth Lipschitz loss functions tend to zero as the number of samples increases. In addition, our results confirm that using a momentum parameter, $\mu \approx 1$, for the entire training improves

optimization error under certain settings. However, it adversely affects the generalization error bounds. Hence, it is crucial to establish an appropriate balance between the optimization error associated with the empirical risk and the generalization error.

Finally, our experimental results show that SGDEM outperforms both vanilla SGD and SGDM in terms of test error on CIFAR10 and generalization error on ImageNet.

1.2 Related Work

Studies on the generalization of momentum methods are scarce in the literature. As explained above, momentum is not considered in (Bousquet and Elisseeff, 2002; Hardt et al., 2016). While the generalization error of SGDM is studied in (Ong, 2017) and (Chen et al.), their analysis is limited to the special case of quadratic loss functions. In this work, we show that unlike SGDM, multiple epochs of SGDEM is guaranteed to generalize for smooth Lipschitz loss functions. A similar hybrid method has been shown to generalize better than both vanilla SGD and Adaptive Moment Estimation (Adam) in deep learning practice (Keskar and Socher). However, it remains unclear why such hybrid methods generalize better. Our work sheds theoretical light on this question.

Convergence of first-order methods with momentum has been studied in (Polyak, 1964; Ochs et al., 2014, 2015; Ghadimi et al., 2015; Lessard et al., 2016; Yan et al., 2018; Wilson et al., 2021; Gadat et al., 2018; Orvieto et al., 2020; Can et al., 2019). Most of these works consider the deterministic setting for gradient update (Polyak, 1964; Ochs et al., 2014, 2015; Ghadimi et al., 2015; Lessard et al., 2016; Wilson et al., 2021). Only a few works have analyzed convergence in the stochastic setting (Yan et al., 2018; Gadat et al., 2018; Orvieto et al., 2020; Can et al., 2019). In (Yan et al., 2018), a unified convergence analysis of SGDM has been studied for both convex and nonconvex loss functions with bounded variance. Gadat et al. (2018) have studied the almost sure convergence results of the stochastic heavy-ball method with nonconvex coercive loss functions and provided a complexity analysis for the case of quadratic strongly convex. In (Orvieto et al., 2020), differential equation-based analysis is used to study convergence of SGDM. Can et al. (2019) have obtained linear convergence rates for SGDM under a particular momentum for the special case of quadratic loss functions. In this paper, we introduce *early momentum* for the class of smooth Lipschitz loss functions, which requires unique convergence analysis as shown in Section 3.

We further note that Lessard et al. (2016) have provided a specific loss function for which the heavy-ball method does not converge. This loss function does not contradict our convergence analysis. The loss function in (Lessard et al., 2016) has been carefully constructed and does not satisfy the assumptions considered in this paper.

In addition to Polyak's heavy-ball momentum method, Nesterov (1983) has proposed an accelerated gradient descent, which converges as $O(1/k^2)$ in a deterministic

and convex setting where k is the number of iterations. Convergence rates are obtained for Nesterov's accelerated gradient method in various settings (Su et al., 2014; Laborde and Oberman, 2020; Assran and Rabbat, 2020). However, the Netstrov momentum does not seem to improve the rate of convergence for stochastic gradient settings (Goodfellow et al., 2016, Section 8.3.3). Therefore, in this work we focus on the heavy-ball momentum.

High-probability bounds on the generalization error of uniformly stable algorithms over the random choice of the dataset have been recently established in (Feldman and Vondrak, 2018, 2019; Bousquet et al., 2020; Klochkov and Zhivotovskiy, 2021). Momentum-based methods have not been considered in these works. In this paper, we establish generalization errors for momentum-based methods with a focus on the randomness of the algorithm.

Our results complement the recent results of Attia and Koren (2021), which show exponential growth in uniform stability bounds of accelerated gradient descent methods. We focus on stochastic gradient descent with heavy-ball momentum for possibly nonconvex problems under nonconstant step-sizes, while Attia and Koren (2021) focus on convex problems with full-batch Nesterov's accelerated gradient under a fixed step-size.

Notation: We use $\mathbb{E}[\cdot]$ to denote the expectation and $\|\cdot\|$ to represent the Euclidean norm of a vector. We use lower-case bold font to denote vectors. We use sans-serif font to denote random quantities. Sets and scalars are represented by calligraphic and standard fonts, respectively.

2 Problem and Assumptions

We consider a general supervised learning problem, where $\mathcal{S} = \{\mathsf{z}_1, \ldots, \mathsf{z}_n\}$ denotes the set of samples of size n drawn i.i.d. from some space $\mathcal{Z}$ with an unknown distribution D. We assume a learning model described by parameter vector $\mathbf{w} \in \Omega$. Let $\ell(\mathbf{w}; \mathbf{z})$ denote the loss of the model described by parameter $\mathbf{w}$ on example $\mathbf{z} \in \mathcal{Z}$.

The ultimate goal of learning is to minimize the true or population risk given by

$$R(\mathbf{w}) := \mathbb{E}_{\mathbf{z}\sim D}[\ell(\mathbf{w}; \mathbf{z})]. \tag{2.1}$$

Since the distribution D is unknown, we approximate this objective by the empirical risk during training, *i.e.*, $R_{\mathcal{S}}(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}; \mathbf{z}_i)$. We assume $\mathbf{w} = A(\mathcal{S})$ for some potentially randomized algorithm A.

2.1 Generalization Error and Stability

In order to find an upper bound on the true risk of algorithm A, in this work, we consider the generalization error, which is the expected difference of empirical and true risk:

$$\epsilon_g := \mathbb{E}_{\mathcal{S},A}[R(A(\mathcal{S})) - R_{\mathcal{S}}(A(\mathcal{S}))]. \tag{2.2}$$

In order to find an upper bound on the generalization error of algorithm A, we consider the uniform stability property.

Definition 1 *Let $\mathcal{S}$ and $\mathcal{S}'$ denote two datasets from space $\mathcal{Z}^n$ such that $\mathcal{S}$ and $\mathcal{S}'$ differ in at most one example. Algorithm A is ϵ_s-uniformly stable if for all datasets $\mathcal{S}$ and $\mathcal{S}'$, we have*

$$\sup_{\mathbf{z}} \mathbb{E}_A[\ell(A(\mathcal{S}); \mathbf{z}) - \ell(A(\mathcal{S}'); \mathbf{z})] \leq \epsilon_s. \tag{2.3}$$

It is known that uniform stability implies generalization in expectation:

Theorem 2 (Hardt et al. 2016) *If A is an ϵ_s-uniformly stable algorithm, then the generalization error of A is upper bounded by ϵ_s.*

Theorem 2 suggests that it is enough to control the uniform stability of an algorithm to bound the generalization error.

2.2 SGDM

The update rule for SGDM is given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \mu(\mathbf{w}_t - \mathbf{w}_{t-1}) - \alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}) \tag{SGDM}$$

where $\alpha_t > 0$ is the step-size, $\mu \in (0, 1]$ is the momentum parameter, $\mathsf{i}_t \in \{1, \ldots, n\}$ is a selected index drawn uniformly at random at each iteration, and $\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})$ is the loss evaluated on sample $\mathbf{z}_{\mathsf{i}_t}$.[1] In SGDM, we run the update iteratively for T steps and let $\mathbf{w}_T$ denote the final output.

In the case where the parameter space Ω is a compact and convex set, we consider the update rule for projected SGDM:

$$\mathbf{w}_{t+1} = \mathbf{P}\Big(\mathbf{w}_t + \mu(\mathbf{w}_t - \mathbf{w}_{t-1}) - \alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\Big) \tag{P-SGDM}$$

where $\mathbf{P}$ denotes the Euclidean projection onto Ω. The key quantity of interest in this paper is the generalization error given by

$$\epsilon_g = \mathbb{E}_{\mathcal{S},A}[R(\mathbf{w}_T) - R_{\mathcal{S}}(\mathbf{w}_T)] = \mathbb{E}_{\mathcal{S},\mathsf{i}_0,\ldots,\mathsf{i}_{T-1}}[R(\mathbf{w}_T) - R_{\mathcal{S}}(\mathbf{w}_T)]$$

since the randomness in A arises from the choice of $\mathsf{i}_0, \ldots, \mathsf{i}_{T-1}$.

1. Another variant to select i_t is to permutate $\{1, \ldots, n\}$ randomly once and then select the examples repeatedly in a cyclic manner. Our stability analysis in Sections 3.2 and 4 holds under both variants, i.e., uniformly at random with replacement and random permutation.

2.3 Assumptions on Loss Function

Let $\mathbf{z} \in \mathcal{Z}$. In our analysis, we will assume that the loss function $\ell(\cdot; \mathbf{z})$ satisfies the following properties, which are used also in (Hardt et al., 2016).

Assumption 1 (Lipschitzness & smoothness) *Let $\mathbf{z} \in \mathcal{Z}$. The loss function $\ell(\cdot; \mathbf{z})$ satisfies the following properties: 1) L-Lipschitzness: There exists some $L > 0$ such that $|\ell(\mathbf{u}; \mathbf{z}) - \ell(\mathbf{v}; \mathbf{z})| \leq L\|\mathbf{u} - \mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in \Omega$; 2) β-smoothness: There exists some $\beta > 0$ such that $\|\nabla \ell(\mathbf{u}; \mathbf{z}) - \nabla \ell(\mathbf{v}; \mathbf{z})\| \leq \beta\|\mathbf{u} - \mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in \Omega$.*

Our assumptions hold for neural networks with smooth activation functions such as smooth approximations of ReLU including softplus or Gaussian error Linear Units (GeLU) (Dugas et al., 2000; Hendrycks and Gimpel). We note that softplus and GeLU typically match and exceed performance compared to ReLU (Clevert et al., 2016; Xu et al.).

3 Smooth Lipschitz Loss

We first show that there exists a convex loss function for which the stability gap for multiple epochs of SGDM becomes unbounded. For the case of smooth Lipschitz loss functions, we introduce SGDEM and show that machine learning models can be trained for multiple epochs of SGDEM while their generalization errors are bounded.

In SGDEM, the momentum μ is set to some constant $\mu_d \in (0, 1]$ in the first t_d steps and then zero for $t = t_d + 1, \ldots, T$. Thus, the update rule for SGDEM is given by

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \mu_d \mathbb{1}(t \leq t_d)(\mathbf{w}_t - \mathbf{w}_{t-1}) - \alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{i_t}) \tag{SGDEM}$$

where $\mathbb{1}$ denotes the indicator function, and the projected version can be similarly defined based on P-SGDM.

3.1 Uniform Stability Bounds for SGDM and SGDEM

Example 1 *Let $w \in [-1, 1]$ denote a parameter. Consider the one-dimensional and convex loss function $\ell(w; \mathbf{z}) = L_{\mathbf{z}} w + c_{\mathbf{z}}$ where $L_{\mathbf{z}} \in \{L, -L\}$ depending on $\mathbf{z} \in \mathcal{Z}$ and $c_{\mathbf{z}} \geq 0$ is constant w.r.t. w. For a specific choice of $\mathcal{S}$ with*

$$R_{\mathcal{S}}(w) = \frac{1}{n} \sum_{i=1}^{n} (Lw + c_i) = Lw + \sum_{i=1}^{n} \frac{c_i}{n},$$

the optimal parameter minimizing the empirical risk is $w^*_{\mathcal{S}} = -1$.

Both SGDM and SGDEM can find the optimal solution of our convex empirical risk minimization problem. We first establish a lower bound on the stability gap

when SGDM is run for multiple epochs, which shows that the gap can be unbounded.[2] In our analysis of the stability, our goal is to track the divergence of two different iterative sequences of update rules with the same starting point.

Theorem 3 *Let $\mathbf{z} \in \mathcal{Z}$. Suppose that the SGDM update is executed for T steps with constant step-size $\alpha > 0$ and $\mu \in (0, 1]$ on Example 1. There exist datasets $\mathcal{S}$ and $\mathcal{S}'$ such that $\mathbb{E}_A[\ell(A(\mathcal{S}); \mathbf{z}) - \ell(A(\mathcal{S}'); \mathbf{z})]$ is lower bounded by $\Omega(\frac{T}{n}(1 + k\mu^k))$ with $k = \lceil \log(T) \rceil$.*[3]

Proof We consider two neighbouring datasets $\mathcal{S}$ and $\mathcal{S}'$ with

$$R_{\mathcal{S}}(w) = \frac{1}{n}\sum_{i=1}^{n}(Lw + c_i) = Lw + \bar{c}$$

and

$$R_{\mathcal{S}'}(w) = -\frac{1}{n}Lw + \frac{c'_k}{n} + \frac{1}{n}\sum_{i=1, i\neq k}^{n}(Lw + c_i) = \frac{n-2}{n}Lw + \frac{c'_k - c_k}{n} + \bar{c}$$

where $\bar{c} = \frac{1}{n}\sum_{i=1}^{n} c_i$. Suppose we select an index i_t uniformly at random from $\{1, \ldots, n\}$. Then we have $\mathbb{E}_{\mathsf{i}_t}[\nabla\ell(w; \mathbf{z}_{\mathsf{i}_t})] = \nabla R_{\mathcal{S}}(w) = L$ and $\mathbb{E}_{\mathsf{i}_t}[\nabla\ell(w; \mathbf{z}'_{\mathsf{i}_t})] = \nabla R_{\mathcal{S}'}(w) = (n-2)L/n$, which holds for all $w \in \Omega$. Let w_T and w'_T denote the outputs of SGDM on $\mathcal{S}$ and $\mathcal{S}'$, respectively. Suppose $\mathsf{w}_0 = \mathsf{w}'_0$. We can follow the steps of SGDM on $\mathcal{S}$ and obtain[4]

$$\mathbb{E}_{\mathsf{i}_0, \ldots, \mathsf{i}_{T-1}}[\mathsf{w}_T] = -(T + (T-1)\mu + (T-2)\mu^2 + \cdots + \mu^{T-1})\alpha L + \mathbb{E}[\mathsf{w}_0].$$

Similarly, we have

$$\mathbb{E}_{\mathsf{i}_0, \ldots, \mathsf{i}_{T-1}}[\mathsf{w}'_T] = -(T + (T-1)\mu + (T-2)\mu^2 + \cdots + \mu^{T-1})\frac{(n-2)\alpha L}{n} + \mathbb{E}[\mathsf{w}_0].$$

Hence, we have

$$\mathbb{E}_A[\mathsf{w}'_T - \mathsf{w}_T] = \frac{2\alpha L}{n}(T + (T-1)\mu + (T-2)\mu^2 + \cdots + \mu^{T-1}).$$

Let $\mathbf{z} \in \mathcal{Z}$. Using Jensen's inequality, we can show that

$$\begin{aligned}\mathbb{E}_A[|\ell(\mathsf{w}_T; \mathbf{z}) - \ell(\mathsf{w}'_T; \mathbf{z})|] &\geq |\mathbb{E}_A[\ell(\mathsf{w}_T; \mathbf{z}) - \ell(\mathsf{w}'_T; \mathbf{z})]| \\ &= \frac{2\alpha L^2}{n}\sum_{j=0}^{T-1}(T-j)\mu^j.\end{aligned}$$

2. If we set $T = n^l$ for some $l > 1$, the stability gap will become unbounded as $n \to \infty$ regardless of μ. If we set $T = kn$ with $k > 1$, there will still be a nonvanishing gap regardless of μ (the gap does not blow up but it does not vanish).
3. We note that the lower bound in Theorem 3 matches an upper bound for SGD with smooth and convex losses, and early-stopped SGDM will be stable by setting T sublinearly in n.
4. We assume $\mathsf{w}_0, \alpha, \mu$ are set such that the updated parameter remains in the parameter space.

Hence, $\mathbb{E}_A[|\ell(\mathsf{w}_T;\mathbf{z})-\ell(\mathsf{w}'_T;\mathbf{z})|]$ is lower bounded by $\Omega(T/n(1+k\mu^k))$ where $k=o(T)$. ■

Note that the stability lower bound increases monotonically with μ. For the same example, we can show that the stability gap for multiple epochs of SGDEM goes to zero.

Theorem 4 *Let $\mu_d \in (0,1)$. For Example 1 and datasets described in Theorem 3, the stability gap for SGDEM and SGDM goes to zero as $n\to\infty$ as long as $\sum_{j=1}^T \alpha_j = o(n)$.*

Furthermore, under $T=n^l$ with some $l>1$ and $t_d\big(\sum_{j=1}^T \alpha_j^2\big)^{1/3} = o(T^{\frac{2}{3l}})$, the stability gap for SGDEM goes to zero for any $\mu_d \in (0,1]$.

Proof Let w_T and w'_T denote the outputs of SGDEM on $\mathcal{S}$ and $\mathcal{S}'$, respectively. Following the steps of SGDEM on $\mathcal{S}$, we have

$$\begin{aligned}\mathbb{E}_{\mathsf{i}_0,\ldots,\mathsf{i}_{T-1}}[\mathsf{w}_T] = &-L(\alpha_1+\cdots+\alpha_T) - \alpha_1 L(\mu_d+\cdots+\mu_d^{t_d-1}) - \alpha_2 L^2(\mu_d+\cdots+\mu_d^{t_d-2})\\ &-\cdots-\alpha_{t_d-1}L\mu_d + \mathbb{E}[\mathsf{w}_0].\end{aligned}$$

Similarly, we have

$$\begin{aligned}E_{\mathsf{i}_0,\ldots,\mathsf{i}_{T-1}}[\mathsf{w}'_T] = &-((n-2)/n)L(\alpha_1+\cdots+\alpha_T) - ((n-2)/n)\alpha_1 L(\mu_d+\cdots+\mu_d^{t_d-1})\\ &-((n-2)/n)\alpha_2 L^2(\mu_d+\cdots+\mu_d^{t_d-2}) - \cdots - ((n-2)/n)\alpha_{t_d-1}L\mu_d + \mathbb{E}[\mathsf{w}_0].\end{aligned}$$

Let $\mathbf{z}\in\mathcal{Z}$. For this particular example, we have $\mathbb{E}_A[|\ell(\mathsf{w}_T;\mathbf{z})-\ell(\mathsf{w}'_T;\mathbf{z})|] = L\mathbb{E}_A|\mathsf{w}_T-\mathsf{w}'_T|$.

Lemma 5 $\mathsf{w}'_T - \mathsf{w}_T \geq 0$ *everywhere.*

Proof Let w_T and w'_T denote a realization of w_T and w'_T, respectively. Let w_0 denote a realization of w_0, and fix $(\mathsf{i}_0,\ldots,\mathsf{i}_{T-1})$ where the differing index between $\mathcal{S}$ and $\mathcal{S}'$ happens at steps in $\mathcal{J}=\{j_1,\ldots,j_m\}$. Then we have

$$\begin{aligned}w_T = &-L(\alpha_1+\cdots+\alpha_T) - \alpha_1 L(\mu_d+\cdots+\mu_d^{t_d-1}) - \alpha_2 L^2(\mu_d+\cdots+\mu_d^{t_d-2})\\ &-\cdots-\alpha_{t_d-1}L\mu_d + w_0\end{aligned}$$

and

$$\begin{aligned}w'_T = &-L\sum_{j\in\mathcal{T}/\mathcal{J}}\alpha_j + L\sum_{j\in\mathcal{J}}\alpha_j - L\sum_{j\in\mathcal{T}_d/\mathcal{J}_d}\alpha_j(\mu_d+\cdots+\mu_d^{t_d-j})\\ &+L\sum_{j\in\mathcal{J}_d}\alpha_j(\mu_d+\cdots+\mu_d^{t_d-j}) + w_0\end{aligned}$$

where $\mathcal{T} = \{1, \dots, T\}$, $\mathcal{T}_d = \{1, \dots, t_d\}$, and $\mathcal{J}_d = \{j \in \mathcal{J} : \ j \leq t_d\}$. This shows $w'_T \geq w_T$. We note that the set of realizations with $w'_T < w_T$ is empty. ■

This lemma shows $\mathbb{E}_A|\mathsf{w}_T - \mathsf{w}'_T| = \mathbb{E}_A[\mathsf{w}'_T] - \mathbb{E}_A[\mathsf{w}_T]$. We first note that

$$\begin{aligned}
\mathbb{E}_A[|\ell(\mathsf{w}_T; \mathbf{z}) - \ell(\mathsf{w}'_T; \mathbf{z})|] &= \frac{2L^2}{n} \sum_{j=1}^{T} \alpha_j + \frac{2L^2}{n} \sum_{j=1}^{t_d} \alpha_j (\mu_d + \cdots + \mu_d^{t_d - j}) \\
&\leq \frac{2L^2}{n} \sum_{j=1}^{T} \alpha_j + \frac{2L^2 \mu_d}{n(1-\mu_d)} \sum_{j=1}^{t_d} \alpha_j \\
&\leq \frac{2L^2 \sum_{j=1}^{T} \alpha_j}{n} \Big(1 + \frac{\mu_d}{1-\mu_d}\Big).
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathbb{E}_A[|\ell(\mathsf{w}_T; \mathbf{z}) - \ell(\mathsf{w}'_T; \mathbf{z})|] &= \frac{2L^2}{n} \sum_{j=1}^{T} \alpha_j + \frac{2L^2}{n} \sum_{j=1}^{t_d} \alpha_j (\mu_d + \cdots + \mu_d^{t_d - j}) \\
&\leq \frac{2L^2}{n} \sum_{j=1}^{T} \alpha_j + \frac{2L^2}{n} \sum_{j=1}^{t_d} \alpha_j (t_d - j) \\
&\leq \frac{2L^2}{n} \sum_{j=1}^{T} \alpha_j + \frac{2L^2}{n} \sqrt{\frac{(t_d^2 - 1)(2t_d - 1)}{6} \sum_{j=1}^{t_d} \alpha_j^2}
\end{aligned}$$

where the last inequality holds due to Cauchy-Schwarz. This completes the proof. ■

For constant step-size, we can show an $\Omega(\frac{T}{n})$ lower bound on the stability gap for SGDM even when we set momentum to zero. However, this does not explain what happens if we use another step-size. To highlight the importance of early momentum on bounding the stability gap, in Appendix A, we show that the stability gap for multiple epochs of SGDM may become unbounded for any step-size schedule. This includes $\alpha_1 = 1$ and $\alpha_j = 0$ for $j > 1$, *i.e.,* the gradient term is added only in the first iteration. We also establish an $\Omega(\frac{T}{n})$ lower bound for SGDM on Example 1 even with a *time-decaying* step-size, which shows that it is important to control both step-size and momentum to establish uniform stability.[5] In Section 4, we show that SGDM with fixed step-size is stable for strongly convex loss functions.This demonstrates the role of the loss function.

Corollary 6 *For Example 1 with datasets described in Theorem 3 and for time-decaying step-size, the stability gap for SGDM is lower bounded by $\Omega(\frac{T}{n})$ for any momentum $\mu > 0$.*

5. Time-decaying step-size is required to establish uniform stability for multiple epochs of SGD without momentum in the convex case (Hardt et al., 2016, Theorem 3.8).

Proof It follows immediately from the proof of Theorem 4. ■

Remark 7 *As shown in Section 3.2, unlike SGDM, SGDEM is stable and thus guaranteed to generalize for smooth Lipschitz loss functions and any momentum. We remark that, since uniform stability is only a sufficient condition for generalization, our result here does not necessarily imply that SGDM does not generalize. Our results highlight that step-size schedule, momentum, and the structure of the loss play roles in establishing uniform stability. In Section 5, we show an empirical example that SGDM does not generalize in a nonconvex problem.*

3.2 Generalization Analysis of SGDEM

Since generalization is predicated on the convergence of a learning algorithm, we first show that SGDEM is guaranteed to converge to a local minimum for general and possibly nonconvex problems. Then, we show that SGDEM is guaranteed to generalize for any μ_d, when t_d is chosen appropriately. Our analysis captures the inherent trade-off between optimization and generalization.

Let $q \in [\frac{1}{2}, 1)$. We establish an explicit convergence rate for SGDEM and a general step-size $\alpha_t = \alpha_0/t^q$, which includes as special cases popular choices of step-sizes in the optimization literature. Together with the generalization bounds, our analyses characterize the optimization error in terms of the expected norm of gradients of empirical risk and generalization error for stochastic gradient descent with heavy-ball momentum under a broad range of hyperparameters and smooth Lipschitz loss functions. To the best of our knowledge, this is the first work providing such results.

Theorem 8 *Let $q \in [\frac{1}{2}, 1)$. Suppose that ℓ satisfies Assumption 1 and that the SGDEM update is executed for T steps with step-size $\alpha_t = \alpha_0/t^q$ and any $1 \leq t_d \leq T$. Then we have*

$$\min_{1\leq t\leq T} \mathbb{E}_A[\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2] \leq \frac{2(R_{\mathcal{S}}(\mathbf{w}_0) - \inf_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}))}{\sum_{t=1}^T \alpha_t} + \frac{\max\left\{\frac{\beta L^2}{2(1-\mu_d)^2}, \frac{\beta\mu_d L^2}{(1-\mu_d)^3}, \frac{\beta L^2}{2}\right\} \sum_{t=1}^T \alpha_t^2}{\sum_{t=1}^T \alpha_t}. \tag{3.1}$$

In particular, SGDEM achieves the rate of $\mathcal{O}(T^{q-1})$ for any t_d.

Proof See Appendix B. ■

Remark 9 *Convergence of SGDEM with constant step-size and another time-dependent step-size are provided in Appendix C and Appendix D, respectively. In Appendix E, we provide a sufficient condition for the optimization bound to become a*

monotonically decreasing function of t_d. In Appendix F, we study the convergence bound for a special form of SGDEM and show the benefit of using momentum. We also provide a simple sufficient condition for the non-vanishing term in the convergence bound to become a monotonically decreasing function of μ_d.

We first show that for $\alpha_t = \alpha_0/t$ and carefully designed t_d, SGDEM updates satisfy uniform stability, and the number of stochastic gradient steps can grow as n^l for a small $l > 1$ while the generalization error is limited. We note that SGDEM is guaranteed to generalize for any μ_d. Then we establish an upper bound on the generalization error of SGDEM for a general step-size $\alpha_t = \alpha_0/t^q$ with $q \in [\frac{1}{2}, 1)$.

Theorem 10 *Suppose that ℓ satisfies Assumption 1 and that the SGDEM update is executed for T steps with step-size $\alpha_t = \alpha_0/t$ and some constant $\mu_d \in (0, 1]$ in the first t_d steps. Then, for any $1 \leq \tilde{t} \leq t_d \leq T$, SGDEM satisfies ϵ_s-uniform stability with*

$$\epsilon_s \leq \frac{2\alpha_0 L^2}{n} T^u \tilde{h}(\mu_d, t_d) + \frac{\tilde{t}M}{n} + \frac{2L^2}{\beta(n-1)} \Big(\frac{T}{\tilde{t}}\Big)^u \tag{3.2}$$

where $\tilde{h}(\mu_d, t_d) = \exp(2\mu_d t_d)(E_1(2\mu_d \tilde{t}) - E_1(2\mu_d t_d))$, $E_1(x) := \int_x^\infty \frac{\exp(-t)}{t}\,\mathrm{d}t$, $u = (1 - \frac{1}{n})\alpha_0\beta$, and $M = \sup_{\mathbf{w},\mathbf{z}} \ell(\mathbf{w}; \mathbf{z})$.

Proof Let $\mathcal{S}$ and $\mathcal{S}'$ be two sets of samples of size n that differ in at most one example. Let $\mathbf{w}_T$ and $\mathbf{w}'_T$ denote the outputs of SGDM on $\mathcal{S}$ and $\mathcal{S}'$, respectively. We consider the updates $\mathbf{w}_{t+1} = G_t(\mathbf{w}_t) + \mu_t(\mathbf{w}_t - \mathbf{w}_{t-1})$ and $\mathbf{w}'_{t+1} = G'_t(\mathbf{w}'_t) + \mu_t(\mathbf{w}'_t - \mathbf{w}'_{t-1})$ where $\mu_t := \mu_d \mathbb{1}(t \leq t_d)$ with $G_t(\mathbf{w}_t) = \mathbf{w}_t - \alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})$ and $G'_t(\mathbf{w}'_t) = \mathbf{w}'_t - \alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}'_t; \mathbf{z}'_{\mathsf{i}_t})$, respectively, for $t = 1, \ldots, T$. We denote $\delta_t := \|\mathbf{w}_t - \mathbf{w}'_t\|$. Suppose $\mathbf{w}_0 = \mathbf{w}'_0$, *i.e.,* $\delta_0 = 0$.

First, as a preliminary step, we observe that the expected loss difference under $\mathbf{w}_T$ and $\mathbf{w}'_T$ for every $\mathbf{z} \in Z$ and every $\tilde{t} \in \{1, \ldots, T\}$ is bounded by

$$\mathbb{E}[|\ell(\mathbf{w}_T; \mathbf{z}) - \ell(\mathbf{w}'_T; \mathbf{z})|] \leq \frac{\tilde{t}M}{n} + L\mathbb{E}[\delta_T | \delta_{\tilde{t}} = 0]. \tag{3.3}$$

This follows from the argument for a similar claim in (Hardt et al., 2016) and applying it to our expression of SGDEM parameter update.

Now, let us define $\Delta_{t,\tilde{t}} := \mathbb{E}[\delta_t | \delta_{\tilde{t}} = 0]$. Our goal is to find an upper bound on $\Delta_{T,\tilde{t}}$ and then minimize it over $\tilde{t}$.

At step t, with probability $1 - 1/n$, the example is the same in both $\mathcal{S}$ and $\mathcal{S}'$. Hence, we have

$$\begin{aligned} \delta_{t+1} &= \|(1 + \mu_t)(\mathbf{w}_t - \mathbf{w}'_t) - \mu_t(\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}) - \alpha_t \phi_1\| \\ &\leq (1 + \mu_t)\|\mathbf{w}_t - \mathbf{w}'_t\| + \mu_t \|\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}\| + \alpha_t \|\phi_1\| \\ &\leq (1 + \mu_t + \alpha_t \beta)\delta_t + \mu_t \delta_{t-1} \end{aligned} \tag{3.4}$$

where $\phi_1 = \nabla_{\mathbf{w}}\ell(\mathbf{w}_t;\mathbf{z}_{\mathrm{i}_t}) - \nabla_{\mathbf{w}}\ell(\mathbf{w}'_t;\mathbf{z}_{\mathrm{i}_t})$. Note that the last inequality in (3.4) holds due to the β-smooth property. With probability $1/n$, the selected example is different in $\mathcal{S}$ and $\mathcal{S}'$. In this case, we have

$$\begin{aligned}\delta_{t+1} &= \|(1+\mu_t)(\mathbf{w}_t - \mathbf{w}'_t) - \mu_t(\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}) - \alpha_t\phi_2\| \\ &\le (1+\mu_t)\delta_t + \mu_t\delta_{t-1} + \alpha_t\|\nabla_{\mathbf{w}}\ell(\mathbf{w}_t;\mathbf{z}_{\mathrm{i}_t})\| + \alpha_t\|\nabla_{\mathbf{w}}\ell(\mathbf{w}'_t;\mathbf{z}'_{\mathrm{i}_t})\| \\ &\le (1+\mu_t)\delta_t + \mu_t\delta_{t-1} + 2\alpha_t L\end{aligned} \tag{3.5}$$

where $\phi_2 = \nabla_{\mathbf{w}}\ell(\mathbf{w}_t;\mathbf{z}_{\mathrm{i}_t}) - \nabla_{\mathbf{w}}\ell(\mathbf{w}'_t;\mathbf{z}'_{\mathrm{i}_t})$.

After taking expectation, for every $t \ge \tilde{t}$, we have

$$\Delta_{t+1,\tilde{t}} \le \big(1 + \mu_t + (1-1/n)\alpha_t\beta\big)\Delta_{t,\tilde{t}} + \mu_t\Delta_{t-1,\tilde{t}} + 2\alpha_t L/n.$$

Let us consider the recursion

$$\tilde{\Delta}_{t+1,\tilde{t}} = \big(1 + \mu_t + (1-1/n)\alpha_t\beta\big)\tilde{\Delta}_{t,\tilde{t}} + \mu_t\Delta_{t-1,\tilde{t}} + 2\alpha_t L/n.$$

Note that we have $\tilde{\Delta}_{t+1,\tilde{t}} \ge \tilde{\Delta}_{t,\tilde{t}}$. Then, we have the following inequality:

$$\tilde{\Delta}_{t+1,\tilde{t}} \le \big(1 + 2\mu_t + (1-1/n)\alpha_t\beta\big)\tilde{\Delta}_{t,\tilde{t}} + \frac{2\alpha_t L}{n}.$$

Noting that $\tilde{\Delta}_{t,\tilde{t}} \ge \Delta_{t,\tilde{t}}$ for all $t \ge \tilde{t}$, we have $\mathbb{E}[\Delta_{T,\tilde{t}}] \le S_3 + S_4$ where

$$S_3 = \sum_{t=\tilde{t}+1}^{t_d}\prod_{p=t+1}^{T}\Big(1 + 2\mu_p + (1-\frac{1}{n})\frac{\alpha_0\beta}{p}\Big)\frac{2\alpha_0 L}{nt}$$

and

$$S_4 = \sum_{t=t_d+1}^{T}\prod_{p=t+1}^{T}\Big(1 + 2\mu_p + (1-\frac{1}{n})\frac{\alpha_0\beta}{p}\Big)\frac{2\alpha_0 L}{nt}.$$

Substituting $\mu_p = \mu_d$ for $p = 1,\dots,t_d$, we can find an upper bound on S_3 as follows:

$$\begin{aligned}S_3 &= \sum_{t=\tilde{t}+1}^{t_d}\prod_{p=t+1}^{T}\Big(1 + 2\mu_p + (1-\frac{1}{n})\frac{\alpha_0\beta}{p}\Big)\frac{2\alpha_0 L}{nt} \\ &\le \sum_{t=\tilde{t}+1}^{t_d}\prod_{p=t+1}^{T}\exp\Big(2\mu_p + (1-\frac{1}{n})\frac{\alpha_0\beta}{p}\Big)\frac{2\alpha_0 L}{nt} \\ &\le \sum_{t=\tilde{t}+1}^{t_d}\exp\Big(2\mu_d(t_d - t) + (1-\frac{1}{n})\alpha_0\beta\ln\big(\frac{T}{t}\big)\Big)\frac{2\alpha_0 L}{nt} \\ &\le \frac{2\alpha_0 L}{n}T^{(1-\frac{1}{n})\alpha_0\beta}\exp(2\mu_d t_d)\int_{\tilde{t}}^{t_d} h_1(t)t^{-(1-\frac{1}{n})\alpha_0\beta}\,\mathrm{d}t \\ &\le \frac{2\alpha_0 L}{n}T^{(1-\frac{1}{n})\alpha_0\beta}\exp(2\mu_d t_d)\int_{\tilde{t}}^{t_d} h_1(t)\,\mathrm{d}t \\ &= \frac{2\alpha_0 L}{n}T^{(1-\frac{1}{n})\alpha_0\beta}\exp(2\mu_d t_d)\big(E_1(2\mu_d\tilde{t}) - E_1(2\mu_d t_d)\big)\end{aligned}$$

where $h_1(t) = \frac{\exp(-2\mu_d t)}{t}$ and the exponential integral function E_1 is defined as

$$E_1(x) := \int_x^\infty \frac{\exp(-t)}{t}\,\mathrm{d}t. \tag{3.6}$$

Note that the following inequalities hold for the exponential integral function for $t > 0$ (Abramowitz and Stegun, 1972):

$$\tfrac{1}{2}\exp(-t)\ln\left(1+\tfrac{2}{t}\right) < E_1(t) < \exp(-t)\ln\left(1+\tfrac{1}{t}\right). \tag{3.7}$$

Applying both upper bound and lower bound in (3.7), we have

$$S_3 \leq \frac{2\alpha_0 L}{n} T^{(1-\frac{1}{n})\alpha_0\beta} h(\mu_d, t_d) \tag{3.8}$$

where $h(\mu_d, t_d) = \exp\left(2\mu_d(t_d - \tilde{t})\right)\ln\left(1+\frac{1}{2\mu_d \tilde{t}}\right) - \frac{1}{2}\ln\left(1+\frac{1}{\mu_d t_d}\right)$.

We can also find an upper bound on S_4 as follows:

$$\begin{aligned} S_4 &= \sum_{t=t_d+1}^{T}\prod_{p=t+1}^{T}\Big(1+(1-\frac{1}{n})\frac{\alpha_0\beta}{p}\Big)\frac{2\alpha_0 L}{nt} \\ &\leq \frac{2L}{\beta(n-1)}\Big(\frac{T}{t_d}\Big)^{(1-\frac{1}{n})\alpha_0\beta} \\ &\leq \frac{2L}{\beta(n-1)}\Big(\frac{T}{\tilde{t}}\Big)^{(1-\frac{1}{n})\alpha_0\beta}. \end{aligned} \tag{3.9}$$

Replacing $\Delta_{T,\tilde{t}}$ with its upper bound in Eq. (3.3), we obtain Eq. (3.2). ■

Theorem 10 suggests that the stability bound decreases inversely with the size of the training set. It increases as the momentum parameter μ_d increases. By setting $t_d = \tilde{t}$ in Theorem 10 and comparing with (Hardt et al., 2016, Theorem 3.12), we note that we slightly improve the exponent of T. We can also establish a simpler but looser bound by noting $\tilde{h}(\mu_d, t_d) < h(\mu_d, t_d) = \exp\left(2\mu_d(t_d - \tilde{t})\right)\ln\left(1+\frac{1}{2\mu_d\tilde{t}}\right) - \frac{1}{2}\ln\left(1+\frac{1}{\mu_d t_d}\right)$.

Remark 11 *We can show that our stability bound in Theorem 10 holds for the projected SGDEM since Euclidean projection does not increase the distance between projected points.*

Corollary 12 *For SGDEM with the step-size $\alpha_t = \alpha_0/t$, suppose we set $t_d = \tilde{t}^* + K$ where $\tilde{t}^* = \left(\frac{2\alpha_0 L^2}{M}\right)^{\frac{1}{u+1}} T^{\frac{u}{u+1}}$ for some constant K. Provided that $\alpha_0\beta < 1$, i.e., $u < 1$, the generalization error of SGDEM for T steps with $\alpha_t = \alpha_0/t$ is upper bounded by $\mathcal{O}\Big(\frac{\exp(\mu_d)T^u}{n}\Big)$, and the number of stochastic gradient steps can grow as n^l for a small $l > 1$ while still allowing $\epsilon_s \to 0$ as $n \to \infty$.*

Proof Note that we can minimize the expression $\frac{\tilde{t}M}{n} + \frac{2L^2}{\beta(n-1)}\Big(\frac{T}{\tilde{t}}\Big)^u$ in Eq. (3.2) by optimizing $\tilde{t}$, where the optimal $\tilde{t}$ is given by $\tilde{t}^*$ as defined in the theorem statement. After substituting the optimal $\tilde{t}^*$ into Eq. (3.2) and setting $t_d = \tilde{t}^* + K$ for some constant K, we obtain

$$\epsilon_s \leq \frac{2\alpha_0 L^2}{n} T^u \chi_1 + \frac{1 + \frac{1}{\alpha_0 \beta}}{n-1} (2\alpha_0 L^2)^{\frac{1}{u+1}} (MT)^{\frac{u}{u+1}} \tag{3.10}$$

where $\chi_1 = \exp(2\mu_d K) \ln\big(1 + \frac{1}{2\mu_d \tilde{t}^*}\big) - \frac{1}{2}\ln\big(1 + \frac{1}{\mu_d(\tilde{t}^* + K)}\big)$.

Note that by substituting $\tilde{t} = T$ into Eq. (3.3), for any training algorithm, we have

$$\begin{aligned} \mathbb{E}[|\ell(\mathbf{w}_T; \mathbf{z}) - \ell(\mathbf{w}'_T; \mathbf{z})|] &\leq \frac{TM}{n} + L\mathbb{E}[\delta_T | \delta_T = 0] \\ &= \frac{TM}{n}. \end{aligned} \tag{3.11}$$

Combining the above bounds, an upper bound on the generalization error of SGDEM is given by $\mathcal{O}\Big(\min\Big\{\frac{\exp(\mu_d)T^u}{n}, \frac{TM}{n}\Big\}\Big)$. Here, we consider a nontrivial case where u is small enough, i.e., the generalization error is bounded by the first term. ■

We obtain $\tilde{t}^*$ in Corollary 12 by optimizing over the second and third terms of the upper bound in Theorem 10.

High-probability bounds. In Appendix G, we establish high-probability bounds for generalization error of SGDEM along the lines of (Feldman and Vondrak, 2018).

To complete our generalization analysis, in the following, we further show that SGDEM updates may not satisfy uniform stability depending on how t_d is set.

Corollary 13 *Suppose, in Theorem 10, we set $t_d = \rho T$ and $\tilde{t} = \rho T - K$ for some $0 < \rho \leq 1$ and $K < \rho T$. Then SGDEM updates do not satisfy uniform stability for multiple epochs $T = \kappa n$ and the asymptotic upper bound on the penalty of generalization error is given by $\rho\kappa M$, i.e.,*

$$\lim_{n \to \infty : T = \kappa n} \epsilon_g \leq \rho\kappa M.$$

Proof Substituting $t_d = \rho T$ and $\tilde{t} = \rho T - K$ into Eq. (3.2), we obtain

$$\epsilon_s \leq \frac{2\alpha_0 L^2}{n} T^u \chi_2 + \frac{(\rho T - K)M}{n} + \frac{2L^2}{\beta(n-1)}\Big(\frac{T}{\rho T - K}\Big)^u \tag{3.12}$$

where

$$\chi_2 = \exp\big(2\mu_d K\big) \ln\big(1 + \frac{1}{2\mu_d(\rho T - K)}\big) - \frac{1}{2}\ln\big(1 + \frac{1}{\mu_d \rho T}\big).$$

We can derive the asymptotic penalty by substituting $T = \kappa n$ into the upper bound (3.12), letting $n \to \infty$, and using Theorem 2. ■

Corollary 13 suggests that increasing t_d worsens the generalization penalty when t_d is linear in T. Furthermore, increasing T improves the convergence bound. However, the stability upper bound increases as T increases, which is expected.

Let $q \in [\frac{1}{2}, 1)$. In the following, we establish an upper bound on the generalization error of SGDEM for a general step-size $\alpha_t = \alpha_0/t^q$, which includes as special cases popular choices of step-sizes whose convergence are studied in the optimization literature (Bubeck, 2015).

Theorem 14 *Let $q \in [\frac{1}{2}, 1)$. Suppose that ℓ satisfies Assumption 1 and that the SGDEM update is executed for T steps with step-size $\alpha_t = \alpha_0/t^q$ and some constant $\mu_d \in (0, 1]$ in the first t_d steps. Then, for any $1 \leq \tilde{t} \leq t_d \leq T$, SGDEM satisfies ϵ_s-uniform stability with*

$$\epsilon_s \leq \frac{\alpha_0 L^2 \sqrt{\pi}}{n\sqrt{2\mu_d}(1-q)} \exp(u_q T^{1-q}) \breve{h}(\mu_d, t_d) + \frac{\tilde{t} M}{n} + \frac{2L^2}{\beta(n-1)} \exp\left(u_q (T^{1-q} - \tilde{t}^{1-q})\right) \tag{3.13}$$

where $\breve{h}(\mu_d, t_d) = \exp(2\mu_d t_d + u_q^2/(8\mu_d))\big(\Phi(\sqrt{2\mu_d}(t_d^{1-q} + \frac{u_q}{4\mu_d})) - \Phi(\sqrt{2\mu_d}(\tilde{t}^{1-q} + \frac{u_q}{4\mu_d}))\big)$, $\Phi(x) = \operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,\mathrm{d}t$, $u_q = (1 - \frac{1}{n})\frac{\alpha_0 \beta}{1-q}$, *and* $M = \sup_{\mathbf{w},\mathbf{z}} \ell(\mathbf{w}; \mathbf{z})$.

Proof Similar to the proof of Theorem 10, we have the following inequality:

$$\tilde{\Delta}_{t+1,\tilde{t}} \leq \big(1 + 2\mu_t + (1 - 1/n)\alpha_t \beta\big)\tilde{\Delta}_{t,\tilde{t}} + \frac{2\alpha_t L}{n}.$$

Noting that $\tilde{\Delta}_{t,\tilde{t}} \geq \Delta_{t,\tilde{t}}$ for all $t \geq \tilde{t}$, we have $\mathbb{E}[\Delta_{T,\tilde{t}}] \leq S_3 + S_4$ where

$$S_3 = \sum_{t=\tilde{t}+1}^{t_d} \prod_{p=t+1}^{T} \Big(1 + 2\mu_p + (1 - \frac{1}{n})\frac{\alpha_0 \beta}{p^q}\Big)\frac{2\alpha_0 L}{n t^q}$$

and

$$S_4 = \sum_{t=t_d+1}^{T} \prod_{p=t+1}^{T} \Big(1 + 2\mu_p + (1 - \frac{1}{n})\frac{\alpha_0 \beta}{p^q}\Big)\frac{2\alpha_0 L}{n t^q}.$$

Substituting $\mu_p = \mu_d$ for $p = 1, \ldots, t_d$, we can find an upper bound on S_3 as follows:

$$\begin{aligned}
S_3 &= \sum_{t=\tilde{t}+1}^{t_d} \prod_{p=t+1}^{T} \Big(1 + 2\mu_p + (1 - \frac{1}{n})\frac{\alpha_0\beta}{p^q}\Big)\frac{2\alpha_0 L}{nt^q} \\
&\leq \sum_{t=\tilde{t}+1}^{t_d} \prod_{p=t+1}^{T} \exp\Big(2\mu_p + (1 - \frac{1}{n})\frac{\alpha_0\beta}{p^q}\Big)\frac{2\alpha_0 L}{nt^q} \\
&\leq \sum_{t=\tilde{t}+1}^{t_d} \exp\Big(2\mu_d(t_d - t) + u_q(T^{1-q} - t^{1-q})\Big)\frac{2\alpha_0 L}{nt^q} \\
&\leq \frac{2\alpha_0 L}{n}\exp(u_q T^{1-q} + 2\mu_d t_d)\int_{\tilde{t}}^{t_d} \frac{\exp(-2\mu_d t - u_q t^{1-q})}{t^q}\,\mathrm{d}t \\
&\leq \frac{2\alpha_0 L}{n}\exp(u_q T^{1-q} + 2\mu_d t_d)\int_{\tilde{t}}^{t_d} \frac{\exp(-2\mu_d t^{2(1-q)} - u_q t^{1-q})}{t^q}\,\mathrm{d}t \\
&\leq \frac{2\alpha_0 L}{n}\exp(u_q T^{1-q} + 2\mu_d t_d + u_q^2/(8\mu_d))\int_{\tilde{t}}^{t_d} \frac{\exp\big(-2\mu_d(t^{1-q} + u_q/(4\mu_d))^2\big)}{t^q}\,\mathrm{d}t \\
&= \frac{\alpha_0 L\sqrt{\pi}}{n\sqrt{2\mu_d}(1-q)}\exp(u_q T^{1-q} + 2\mu_d t_d + u_q^2/(8\mu_d))\cdot \\
&\quad \cdot \big(\Phi(\sqrt{2\mu_d}(t_d^{1-q} + \frac{u_q}{4\mu_d})) - \Phi(\sqrt{2\mu_d}(\tilde{t}^{1-q} + \frac{u_q}{4\mu_d}))\big)
\end{aligned}$$

where the fifth step holds since $2(1-q) < 1$ and the last inequality follows (Gradshteyn and Ryzhik, 2014, Eq. 3.321).

We can also find an upper bound on S_4 as follows:

$$\begin{aligned}
S_4 &= \sum_{t=t_d+1}^{T} \prod_{p=t+1}^{T} \Big(1 + (1 - \frac{1}{n})\frac{\alpha_0\beta}{p^q}\Big)\frac{2\alpha_0 L}{nt^q} \\
&\leq \frac{2\alpha_0 L}{n} \sum_{t=t_d+1}^{T} \exp\Big(u_q(T^{1-q} - t^{1-q})\Big)t^{-q} \\
&\leq \frac{2L}{\beta(n-1)}\exp\Big(u_q(T^{1-q} - t_d^{1-q})\Big) \\
&\leq \frac{2L}{\beta(n-1)}\exp\Big(u_q(T^{1-q} - \tilde{t}^{1-q})\Big).
\end{aligned} \tag{3.14}$$

Replacing $\Delta_{T,\tilde{t}}$ with its upper bound in Eq. (3.3), we obtain Eq. (3.13).

By its definition, we have $\Phi(x) \leq 1$. We also note that $1 - \exp(-x^2) \leq \Phi(x)$ for $x > 0$ following the upper bound developed for $1 -$ erf in (Chiani et al., 2003). Applying both lower bound and upper bound on Φ in Eq. (3.13) and after rearranging the terms, we have

$$\epsilon_s \leq \Big(\frac{\alpha_0 L^2\sqrt{\pi}}{n\sqrt{2\mu_d}(1-q)}\exp\big(2\mu_d(t_d - \tilde{t}^{2(1-q)})\big) + \frac{2L^2}{\beta(n-1)}\Big)\exp\Big(u_q(T^{1-q} - \tilde{t}^{1-q})\Big) + \frac{\tilde{t}M}{n}. \tag{3.15}$$

■

We now find a simpler expression for the generalization bound in Theorem 14 by substituting t_d and optimizing over $\tilde{t}$.

Corollary 15 *Let $q \in [\frac{1}{2}, 1)$. For SGDEM with a general step-size $\alpha_t = \alpha_0/t^q$, suppose we set $t_d = \tilde{t}^{*2(1-q)} + K$ for some constant K where $\tilde{t}^*$ satisfies Eq. (3.16). Then the generalization error of SGDEM for T steps with $\alpha_t = \alpha_0/t^q$ is upper bounded by $\mathcal{O}\Big(\min\Big\{\frac{\exp(uT^{1-q}/(u+1)+\mu_d)}{n}, \frac{TM}{n}\Big\}\Big)$.*

Proof Note that we can minimize:

$$\min_{1\le\tilde{t}\le t_d} \frac{\tilde{t}M}{n} + \Big(\frac{\alpha_0 L^2\sqrt{\pi}}{n\sqrt{2\mu_d}(1-q)}\exp\left(2\mu_d K\right) + \frac{2L^2}{\beta(n-1)}\Big)\exp\Big(u_q(T^{1-q} - \tilde{t}^{1-q})\Big)$$

by optimizing $\tilde{t}$ after setting $t_d = \tilde{t} + K$ where the objective is the upper bound in Eq. (3.13). We note that an optimal $\tilde{t}^*$ satisfies

$$M\exp(u_q\tilde{t}^{*1-q})\tilde{t}^{*q} = \Big(\frac{u_q\alpha_0 L^2\sqrt{\pi}}{\sqrt{2\mu_d}} + 2L^2\alpha_0\Big)\exp(u_q T^{1-q}). \tag{3.16}$$

Note Eq. (3.16) does not have an analytic solution but can be solve numerically. Instead, we consider a suboptimal solution by taking ln on both sides of Eq. (3.16) and applying the well-known inequality $\ln(x+1) \le x$, $\forall x \ge -1$, which leads to:

$$\tilde{t}^{1-q} = \frac{\ln\Big((\frac{u\alpha_0 L^2\sqrt{\pi}}{\sqrt{2\mu_d}} + L^2\alpha_0)/M\Big)}{u+1} + \frac{u}{u+1}T^{1-q}. \tag{3.17}$$

Substituting Eq. (3.17) into Eq. (3.13) and combining with the upper bound in Eq. (3.11) complete the proof. ■

As an important special case the problem considered in Theorem 14, we provide an uppe-bound on the generalization error of SGDEM with the larger step size $\alpha_t = \alpha_0/\sqrt{t}$, which is a common choice in the optimization literature (Bubeck, 2015). See Appendix K for the exact expression of $\tilde{t}$.

Corollary 16 *For SGDEM with the step-size $\alpha_t = \alpha_0/\sqrt{t}$, suppose we set $t_d = \tilde{t}^* + K$ for some constant K under an optimized $\tilde{t}^*$, which satisfies Eq. (3.16) with $q = \frac{1}{2}$. Then the generalization error of SGDEM for T steps with $\alpha_t = \alpha_0/\sqrt{t}$ is upper bounded by $\mathcal{O}\Big(\min\Big\{\frac{\exp(u\sqrt{T}/(u+1)+\mu_d)}{n}, \frac{TM}{n}\Big\}\Big)$.*

The bounds in Corollaries 15 and 16 complement the results of Attia and Koren (2021) by showing exponential growth in uniform stability bounds for stochastic gradient descent with heavy-ball momentum for possibly nonconvex problems under nonconstant step-sizes. The bound in Corollary 16 shows that as long as $T = o(\log(n)^{1/(1-q)})$, SGDEM is guaranteed to generalize.

Remark 17 (Stability of SGDEM does not follow that of SGD) *As shown in Theorem 4 and Corollary 12, t_d can grow with T, i.e., two iterative sequences of rules with the same starting points on two neighboring datasets can be possibly arbitrarily far after applying momentum for t_d iterations. Then even a contraction map does not make the algorithm stable. In other words, the stability of SGDEM does not directly follow the stability of SGD.*

4 Strongly Convex Loss

While we have discussed in the previous section the generalization of SGDEM for smooth Lipschitz loss functions, in this section, we focus on the important class of strongly convex loss functions. We show that it suffices to consider the case $t_d = T$, *i.e.,* where SGDEM becomes SGDM, to achieve generalization.

Assumption 2 (Strong convexity) *Let $\mathbf{z} \in \mathcal{Z}$ and $\mathbf{u}, \mathbf{v} \in \Omega$. The loss function $\ell(\cdot; \mathbf{z})$ is γ-strongly convex: there exists $\gamma > 0$ such that*

$$\ell(\mathbf{u}; \mathbf{z}) \geq \ell(\mathbf{v}; \mathbf{z}) + \nabla_{\mathbf{w}} \ell(\mathbf{v}; \mathbf{z})^\top (\mathbf{u} - \mathbf{v}) + \frac{\gamma}{2} \|\mathbf{u} - \mathbf{v}\|^2.$$

An example for γ-strongly convex loss function is Tikhonov regularization, where the empirical risk is given by $R_{\mathcal{S}}(\mathbf{w}) = \sum_{i=1}^n \ell(\mathbf{w}; \mathbf{z}_i) + \frac{\gamma}{2}\|\mathbf{w}\|^2$ with a convex $\ell(\cdot; \mathbf{z})$ for all $\mathbf{z}$. In the following, we assume that $\ell(\mathbf{w}; \mathbf{z})$ is a γ-strongly convex function of $\mathbf{w}$ for all $\mathbf{z} \in \mathcal{Z}$.

To satisfy the L-Lipschitz property of the loss function, we further assume that the parameter space Ω is a compact and convex set. Since Ω is compact, the SGDM update has to involve projection.

We present a bound on the generalization of P-SGDM for γ-strongly convex loss.

Theorem 18 *Suppose that ℓ satisfies Assumptions 1 and 2 and that P-SGDM is executed for T steps with constant step-size α and momentum μ. Provided that $\frac{\alpha\beta\gamma}{\beta+\gamma} - \frac{1}{2} \leq \mu < \frac{\alpha\beta\gamma}{3(\beta+\gamma)}$ and $\alpha \leq \frac{2}{\beta+\gamma}$, P-SGDM satisfies ϵ_s-uniform stability where*

$$\epsilon_s \leq \frac{2\alpha L^2(\beta+\gamma)}{n\big(\alpha\beta\gamma - 3\mu(\beta+\gamma)\big)}. \tag{4.1}$$

Proof sketch: the update rule in the strongly convex case is a contraction, which is not the case in the convex case. In particular, the contraction term due to γ-strong convexity can be leveraged to control the additional expansion term due to

momentum. The overall update remains a contraction assuming the momentum is not too large. See Appendix H for the complete proof.

Theorem 18 implies that the stability bound decreases inversely with the size of the training set. It increases as the momentum parameter μ increases. These properties are also verified in our experimental evaluation.[6]

The theoretically advocated momentum parameters in (Polyak, 1964; Nesterov, 1983) are based on *convergence* analysis of gradient descent with momentum, and do not account for *generalization*. Depending on the condition number of the problem, these values may not satisfy the range of momentum in Theorem 18. These values are not necessarily optimal for P-SGDM, in terms of our objective of true risk. Our goal in Theorem 18 is to show nontrivial cases that P-SGDM satisfies uniform stability.

Remark 19 *Compared with the stability bound in (Hardt et al., 2016) for SGD, both bounds are in $O(1/n)$. Our bound in Theorem 18 holds for $\alpha \leq \frac{2}{\beta+\gamma}$, which is slightly less restrictive than the range of step-size in (Hardt et al., 2016, Theorem 3.9). By substituting $\mu = 0$ in Eq. (4.1), we note that the constant term of our bound, $\frac{\beta+\gamma}{\beta\gamma}$, is slightly larger than that of (Hardt et al., 2016, Theorem 3.9), which is $1/\gamma$. Compared with (Chen et al.), our bound is independent of T and our work analyzes the case of strongly convex loss.*

Classical generalization bounds using Rademacher complexity, which measures the rate of uniform convergence, are obtained for linear predictors with various norm constraints (Shalev-Shwartz and Ben-David, 2014). Those classical generalization bounds are typically $\mathcal{O}(1/\sqrt{n})$. For linear predictors with a Lipschitz loss and a strongly convex regularizer, by bounding Rademacher complexity, it has been shown that with high probability, the generalization error is bounded by $\mathcal{O}(1/\sqrt{n})$ for all parameters in a certain bounded set (Kakade et al., 2008). The fast rates by Sridharan et al. (2008) for regularized linear prediction are built based on the notion of localized Rademacher complexity (Bartlett et al., 2002), which requires an additional boundedness on the dual norm of data mapping. Our high-probability generalization bound for general smooth Lipschitz loss functions in Appendix G is $\mathcal{O}(1/n)$.

Our stability analysis captures how the learning algorithm explores the hypothesis class, in particular, how the generalization gap depends on the momentum. More broadly, unlike stability, uniform convergence is not necessary for learning (based on the learnability definition in (Shalev-Shwartz et al., 2010)) in the general learning setting (Shalev-Shwartz et al., 2010).

Finally, in the case of strongly convex loss, we can further consider the minimization of the true risk as defined in Eq. (2.1), since we are able to derive an upper bound on the optimization error (shown in Appendix I). In Appendix J, we study

6. Our purpose in this work is *not* to show the superiority of SGDM or SGDEM, in terms of the *stability bound*, over SGD. Given the known advantage of SGDM in terms of speeding up training, our purpose is to further analyze the stability/generalization properties of SGDM and SGDEM.

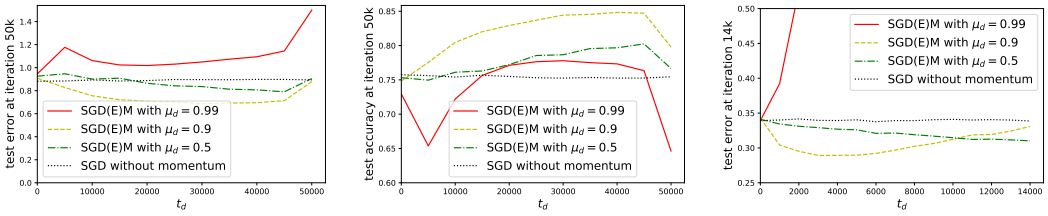

Figure 2: *Test error (left) and test accuracy (middle) of ResNet-20 on CIFAR10. Test error of a feedforward fully connected neural network for notMNIST dataset (right).*

how the uniform stability results in an upper bound on the true risk of P-SGDM. We first show that stability results similar to Theorem 18 hold even if the average parameter $\hat{\mathbf{w}}_T$ is considered as the output of algorithm A. We then decompose the expected true risk into a stability error term and an optimization one. We also compare the final results with SGD with no momentum, and we show that one can achieve tighter bounds by using P-SGDM than vanilla SGD.

5 Experimental Evaluation

5.1 Nonconvex Loss

In this section, we validate the insights obtained in our theoretical results using experimental evaluation. Our main goal is to study how adding momentum affects the generalization and convergence of SGD.

We first investigate the performance of SGDEM when applied to both CIFAR10 (Krizhevsky) and notMINIST datasets for nonconvex loss functions. We set T to 50000 and 14000 for CIFAR10 and notMNIST experiments, respectively. For each value of μ_d, we add momentum for 0-10 epochs. For each pair of (μ_d, t_d), we repeat the experiments 10 times with random initializations. SGDM can be viewed as a special form of SGDEM when the momentum is added for the entire training (*i.e.,* $t_d = T$). For 10 epochs and without data augmentation, we train ResNet-20 on CIFAR10 and a feedforward fully connected neural network with 1000 hidden nodes on notMINIST. For the feedforward fully connected neural network, we use ReLU activation functions, a cross-entropy loss function, and a softmax output layer with Xavier initialization to initialize the weights (Glorot and Bengio, 2010).[7] We set the step-size $\alpha = 0.01$. The minibatch size is set to 10. We use 10 (SGDEM) realizations to evaluate the average performance. We compare the test performance of SGD without momentum with that of SGDEM under $\mu_d = 0.5$, $\mu_d = 0.9$, and $\mu_d = 0.99$.

Outperforming both SGD and SGDM. We show the test error and test accuracy versus t_d under SGDEM for CIFAR10 dataset in Fig. 2 (left and middle). We observe that adding momentum for the entire training (*i.e.,* $t_d = T$ or SGDM) is

7. We observe similar results for smooth activation functions.

Table 1: Ablation studies where we optimize performance of SGD and SGDM by obtaining the minimum test error over step-sizes $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$ and momentum parameters $\mu \in \{0, 0.5, 0.9, 0.99\}$. We do not tune t_d and used the fixed $t_d = 3000$. The rest of the setup is similar to Fig. 2 (right).

	SGD	**SGDEM**	**SGDM**
Test error	0.3596 ± 0.0142	**0.2892** ± 0.0042	0.3194 ± 0.0030

useful when the momentum parameter is small. For different μ_d values, we notice there exists an optimal t_d in Fig. 2 (left) when test error is minimized. We plot the test error versus t_d for notMNIST dataset in Fig. 2 (right). The test accuracy is shown in Appendix L. We observe an overshooting phenomenon for $\mu_d = 0.99$, which is consistent with our convergence analysis in Theorem 25. We observe similar phenomenon when we train a feedforward fully connected neural network with 1000 hidden nodes on MINIST dataset. In terms of test accuracy, we observe that it is not helpful to use a momentum parameter, $\mu_d \approx 1$, for the entire training. In an online framework with high dimensional parameters, early momentum is particularly useful since we can minimize memory utilization as SGDEM does not require $\mathbf{w}_{t-1}$ for the entire iterative updates.

To see whether SGDM is able to match performance of SGDEM by further tuning hyperparameters, in Table 1, we show the results of an ablation study where we minimize the test error of SGD and SGDM by optimizing over step-sizes $\alpha \in \{0.1, 0.01, 0.001, 0.0001\}$ and momentum parameters $\mu \in \{0, 0.5, 0.9, 0.99\}$. For SGDEM, we do *not* tune t_d and used the fixed $t_d = 3000$. The rest of the setup is similar to Fig. 2 (right). We repeat the experiments for 5 times to report confidence intervals. These results show that even under tuned hyperparameters, SGDEM outperforms both SGDM and SGD in terms of test error.

Distributed training on ImageNet. Fig. 1 shows validation loss and generalization error of SGDEM at epoch 90 when training ResNet-18 on ImageNet in a practical data-parallel setting with 4 GPUs under *tuned step-sizes* for SGD and SGDM. We observe that the minimum generalization error happens if the momentum is applied for 50 epochs. In Fig. 3, we plot validation accuracy and generalization gap of SGDEM at epoch 90. Similar to the loss results, we observe that the minimum generalization error happens if the momentum is applied for 50 epochs. Our accuracy results are on par with existing results (He et al., 2016).

Details of ImageNet experiments. The global minibatch size and weight decay are set to 128 and 5×10^{-5}, respectively. For each t_d, the momentum is set to $\mu_d = 0.9$ in the first t_d epochs and then zero for the next $90 - t_d$ epochs. We use a cluster with 4 NVIDIA 2080 Ti GPUs with the following CPU details: Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz; 48 cores; GPU2GPU bandwidth: unidirectional 10GB/s and bidirectional 15GB/s.

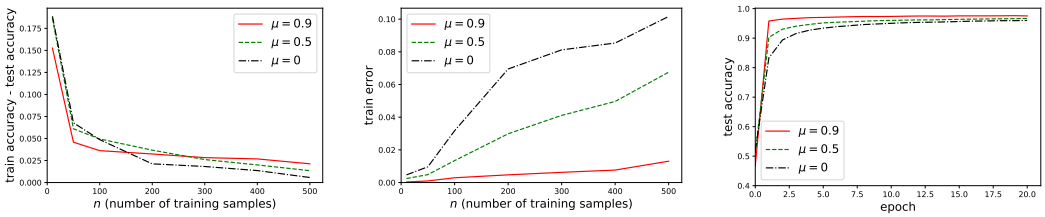

Figure 4: *Generalization error (left) and training error (middle) of logistic regression (cross entropy loss) for notMNIST dataset with $T = 1000$ iterations . Test accuracy of logistic regression for notMNIST dataset with $n = 500$ (right).*

5.2 Strongly Convex Loss

We now study the performance of SGDEM for a smooth and strongly convex loss function. We train a logistic regression model with the weight decay regularization on notMNIST and MNIST. The setup's details are provided in Appendix L. We plot the test error and test accuracy versus t_d under SGDEM for notMNIST and MNIST in Appendix L and observe that, unlike the case of nonconvex loss functions, it does not hurt to add momentum for the entire training. We then focus on SGDM and compare the optimization and generalization performance of vanilla SGD with that of SGDM under $\mu = 0.5$ and $\mu = 0.9$, which are common momentum values used in practice (Goodfellow et al., 2016, Section 8.3.2).

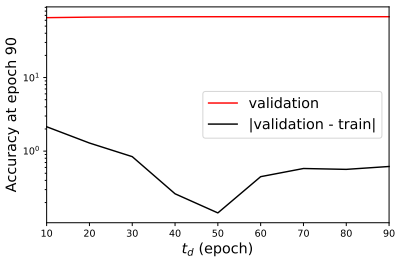

Figure 3: *Validation accuracy and generalization gap of SGDEM when training ResNet-18 on ImageNet in a distributed setting with 4 GPUs under tuned step-size and global minibatch size of 128. For each t_d, the momentum is set to $\mu_d = 0.9$ in the first t_d epochs and then zero for the next $90 - t_d$ epochs. SGDM is a special form of SGDEM with $t_d = 90$.*

Hurting generalization error and improving training error. In Fig. 4 (left) and (middle), we plot generalization and training error versus n with fixed T and observe that generalization error decreases as n increases for all values of μ, which is suggested by our stability upper bound in Theorem 18. In addition, for sufficiently large n, we observe that the generalization error increases with μ, consistent with Theorem 18. On the other hand, training error increases as n increases with fixed T, which is expected. We can observe that adding momentum reduces training error as it improves the convergence rate.

Negligible improvement of test accuracy. In Fig. 4 (right), we plot test accuracy versus T with fixed n (See Appendix L for training error, training accuracy, and test error). As the number of epochs increases, we note that the benefit of momentum on the test accuracy becomes negligible. This happens because adding

momentum results in a higher generalization error thus penalizing the gain in training error.

6 Conclusions and Future Work

We study the generalization error of SGDEM under mild technical conditions. We show that there exists a convex loss function for which the stability gap for multiple epochs of SGDM becomes unbounded and investigate a modified momentum-based update rule, *i.e.,* SGDEM. We establish a bound on the generalization error of SGDEM for the class of smooth Lipschitz loss functions. Our results confirm that deep neural networks can be trained for multiple epochs of SGDEM while their generalization errors are bounded. We also study the convergence of SGDEM in terms of a bound on the expected norm of the gradient. Then, for the case of strongly convex loss functions, we establish an upper bound on the generalization error, which decreases with the size of the training set, and increases as the momentum parameter is increased. We establish an upper bound on the expected difference between the true risk of P-SGDM and the global minimum of the empirical risk. Finally, we present experimental evaluation and show that the numerical results are consistent with our theoretical bounds and SGDEM is an effective algorithm for nonconvex problems.

Beyond uniform stability analysis, which is sufficient for generalization, developing necessary conditions for generalization of various learning algorithms remains an open problem. In particular, "on-average" stability is a more relaxed notion and depends on the data-generating distribution (Shalev-Shwartz et al., 2010). Finally, practical methods for adaptive training, such as Adam, use a variation of the heavy-ball momentum (Kingma and Ba, 2015). Adapting our analysis for such extensions is also an interesting area of future work.

Acknowledgments and Disclosure of Funding

This work was supported by 1) the Research Council of Norway through its Centres of Excellence scheme, Integreat - Norwegian Centre for knowledge-driven machine learning, project number 332645; 2) the Research Council of Norway through its Centre for Research-based Innovation funding scheme (Visual Intelligence under grant no. 309439), and Consortium Partners; 3) Hasler Foundation Program: Hasler Responsible AI (project number 21043); 4) the Swiss National Science Foundation (SNSF) under grant number 200021_205011; 5) the Natural Sciences and Engineering Research Council of Canada under a Discovery Grant.

Content of the appendix. The appendix is organized as follows:

- In Appendix A, we show that the stability gap for multiple epochs of SGDM may become unbounded for any step-size schedule.
- Theorem 8 (convergence of SGDEM) is proved in Appendix B.
- Convergence of SGDEM with constant step-size is provided in Appendix C.
- Convergence guarantees for SGDEM with other time-dependent step-sizes are provided in Appendix D.
- A sufficient condition for the upper bound in Theorem 25 to become monotonically decreasing is provided in Appendix E.
- Benefit of using momentum in terms of convergence is elaborated in Appendix F.
- High-probability generalization bounds for SGDEM are established in Appendix G.
- Theorem 18 (stability of strongly convex problems) is proved in Appendix H.
- Convergence bound for strongly convex loss is provided in Appendix I.
- An upper bound on true risk of strongly convex loss is established in Appendix J.
- Generalization error of SGDEM with $\alpha_t = \alpha_0/\sqrt{t}$ is discussed in Appendix K.
- Additional experimental details are included in Appendix L.

In our analysis of the stability of SGDM, we will consider the following two properties of the growth of the update rule. Let Ω denote the model parameter space. Consider a general update rule G which maps $\mathbf{w} \in \Omega$ to another point $G(\mathbf{w}) \in \Omega$. Our goal is to track the divergence of two different iterative sequences of update rules with the same starting point.

Definition 20 *An update rule G is η-expansive if*

$$\sup_{\mathbf{v},\mathbf{w}\in\Omega} \frac{\|G(\mathbf{v}) - G(\mathbf{w})\|}{\|\mathbf{v} - \mathbf{w}\|} \leq \eta.$$

Definition 21 *An update rule G is σ-bounded if*

$$\sup_{\mathbf{w}\in\Omega} \|\mathbf{w} - G(\mathbf{w})\| \leq \sigma.$$

Appendix A. Importance of early momentum on bounding the stability gap

To highlight the importance of early momentum on bounding the stability gap, in this section, we show that the stability gap for multiple epochs of SGDM may become unbounded for any step-size schedule. This includes $\alpha_1 = 1$ and $\alpha_j = 0$ for $j > 1$, *i.e.,* the gradient term is added only in the first iteration. We also establish a $\Omega(\frac{T}{n})$ lower bound for SGDM on Example 1 even with a *time-decaying* step-size, which shows that it is important to control both step-size and momentum to establish uniform stability.

Theorem 22 *For Example 1 with datasets described in Theorem 3 and for any step-size schedule, there exists a momentum such that the stability gap for SGDM is lower bounded by* $\Omega(\frac{T}{n})$.

In addition, if $\alpha_j \geq \alpha_{\min}$ *for* $j = 1, 2, \ldots, \eta T$ *where* $\alpha_{\min}$ *and* $\eta < 1$ *are some constants that do not depend on* T*, then the stability gap for SGDM is lower bounded by* $\Omega(\frac{T}{n})$ *for any momentum* $\mu > 0$.

Proof Following the proofs of Theorem 3 for SGDM with $\alpha_1 > 0$, we have

$$\mathbb{E}_A[|\ell(\mathsf{w}_T; \mathbf{z}) - \ell(\mathsf{w}'_T; \mathbf{z})|] \geq \frac{2\alpha_1 L^2}{n} \sum_{j=0}^{T-1} \mu^j.$$

Substituting $\mu = 1$, $\mathbb{E}_A[|\ell(\mathsf{w}_T; \mathbf{z}) - \ell(\mathsf{w}'_T; \mathbf{z})|]$ is lower bounded by $\Omega(\frac{T}{n})$.

For the second part of the theorem, suppose $\alpha_j \geq \alpha_{\min}$ for $j = 1, 2, \ldots, \eta T$. Then we have

$$\mathbb{E}_A[|\ell(\mathsf{w}_T; \mathbf{z}) - \ell(\mathsf{w}'_T; \mathbf{z})|] \geq \frac{2\alpha_{\min} L^2}{n} \sum_{j=0}^{\eta T-1} (T - j)\mu^j$$

and $\mathbb{E}_A[|\ell(\mathsf{w}_T; \mathbf{z}) - \ell(\mathsf{w}'_T; \mathbf{z})|]$ is lower bounded by $\Omega(\frac{T}{n})$. ■

Corollary 23 *For Example 1 with datasets described in Theorem 3 and for time-decaying step-size, the stability gap for SGDM is lower bounded by* $\Omega(\frac{T}{n})$ *for any momentum* $\mu > 0$.

Proof It follows immediately from the proof of Theorem 4. ■

Appendix B. Proof of Theorem 8 (convergence of SGDEM)

The following lemmas are useful for our proofs:

Lemma 24 *For any integers t, p, and q, such that $0 \leq t \leq T-1$, $p < T$, and $q < T - t$, and for any sequences $a_0, a_1, \ldots$, and $b_0, b_1, \ldots$, we have*

$$\sum_{i=p}^{T} a_i \sum_{j=q}^{i-t} b_j = \sum_{i=q}^{T-t} b_i \sum_{j=i+t}^{T} a_j.$$

Proof We prove by induction. It clearly holds for $T = 1$. Suppose it holds for all $k < T$. Then, we have

$$\begin{aligned}
\sum_{i=p}^{k+1} a_i \sum_{j=q}^{i-t} b_j &= \sum_{i=p}^{k} a_i \sum_{j=q}^{i-t} b_j + a_{k+1} \sum_{j=q}^{k+1-t} b_j \\
&= \sum_{i=q}^{k-t} b_i \sum_{j=i+t}^{k} a_j + a_{k+1} \sum_{j=q}^{k+1-t} b_j \\
&= \sum_{i=q}^{k-t} b_i \sum_{j=i+t}^{k} a_j + a_{k+1} \sum_{j=q}^{k-t} b_j + a_{k+1} b_{k+1-t} \qquad \text{(B.1)} \\
&= \sum_{i=q}^{k-t} b_i \sum_{j=i+t}^{k+1} a_j + a_{k+1} b_{k+1-t} \\
&= \sum_{i=q}^{k+1-t} b_i \sum_{j=i+t}^{k+1} a_j.
\end{aligned}$$

■

As two special cases of Lemma 24, we obtain (Li and Orabona, 2020, Lemma 4) by substituting $q = 1,\ t = 0,\ p = 1$ and $q = 0,\ t = 1,\ p = 1$:

$$\begin{aligned}
\sum_{i=1}^{T} a_i \sum_{j=1}^{i} b_j &= \sum_{i=1}^{T} b_i \sum_{j=i}^{T} a_j, \\
\sum_{i=1}^{T} a_i \sum_{j=0}^{i-1} b_j &= \sum_{i=1}^{T-1} b_i \sum_{j=i+1}^{T} a_j.
\end{aligned}$$

To facilitate the convergence analysis, we define $\mathbf{q}_t := \mathbf{w}_t - \mathbf{w}_{t-1}$ with $\mathbf{q}_0 = 0$ and $\mathbf{q}_1 = 0$. It is not difficult to show that $\mathbf{q}_{t+1} = \mu_d \mathbf{q}_t - \alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathrm{i}_t})$. Since the empirical risk $R_{\mathcal{S}}$ is a β-smooth function, we have

$$R_{\mathcal{S}}(\mathbf{w}_{t+1}) \leq R_{\mathcal{S}}(\mathbf{w}_t) + \nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \mathbf{q}_{t+1} + \frac{\beta}{2} \|\mathbf{q}_{t+1}\|^2. \qquad \text{(B.2)}$$

Based on the definition of $\mathbf{q}_t$, the inner-product term in Eq. (B.2) is bounded:

$$\begin{aligned}
\nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \mathbf{q}_{t+1} &= \mu_d \nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \mathbf{q}_t - \alpha_t \nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathrm{i}_t}) \\
&= \mu_d \nabla R_{\mathcal{S}}(\mathbf{w}_{t-1})^\top \mathbf{q}_t + \mu_d \left(\nabla R_{\mathcal{S}}(\mathbf{w}_t) - \nabla R_{\mathcal{S}}(\mathbf{w}_{t-1})\right)^\top \mathbf{q}_t \\
&\quad - \alpha_t \nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathrm{i}_t}) \\
&\leq \mu_d \nabla R_{\mathcal{S}}(\mathbf{w}_{t-1})^\top \mathbf{q}_t + \mu_d \|\nabla R_{\mathcal{S}}(\mathbf{w}_t) - \nabla R_{\mathcal{S}}(\mathbf{w}_{t-1})\| \|\mathbf{q}_t\| \\
&\quad - \alpha_t \nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathrm{i}_t}) \\
&\leq \mu_d \nabla R_{\mathcal{S}}(\mathbf{w}_{t-1})^\top \mathbf{q}_t + \mu_d \beta \|\mathbf{q}_t\|^2 - \alpha_t \nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathrm{i}_t})
\end{aligned} \tag{B.3}$$

where the last inequality holds due to smoothness. Unraveling the recursion Eq. (B.3), we have

$$\nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \mathbf{q}_{t+1} \leq \beta \sum_{i=0}^{t-1} \mu_d^{t-i} \|\mathbf{q}_{i+1}\|^2 - \sum_{i=1}^{t} \mu_d^{t-i} \alpha_i \nabla R_{\mathcal{S}}(\mathbf{w}_i)^\top \nabla \ell(\mathbf{w}_i; \mathbf{z}_{\mathrm{i}_i})$$

For simplicity of analysis, we first suppose that the momentum is applied in T steps. Then we modify the bound considering it is set to zero after t_d steps. Substituting this bound in Eq. (B.2) and summing for $t = 1, \ldots, T$, we have

$$\begin{aligned}
R_{\mathcal{S}}(\mathbf{w}_{T+1}) \leq R_{\mathcal{S}}(\mathbf{w}_0) &+ \beta \sum_{t=1}^{T} \sum_{i=1}^{t-1} \mu_d^{t-i} \|\mathbf{q}_{i+1}\|^2 + \frac{\beta}{2} \sum_{t=1}^{T} \|\mathbf{q}_{t+1}\|^2 \\
&- \sum_{t=1}^{T} \sum_{i=1}^{t} \mu_d^{t-i} \alpha_i \nabla R_{\mathcal{S}}(\mathbf{w}_i)^\top \nabla \ell(\mathbf{w}_i; \mathbf{z}_{\mathrm{i}_i}).
\end{aligned} \tag{B.4}$$

Using Lemma 24, we expand the the second and fourth terms in the upper bound Eq. (B.4):

$$\begin{aligned}
\beta \sum_{t=1}^{T} \sum_{i=1}^{t-1} \mu_d^{t-i} \|\mathbf{q}_{i+1}\|^2 &= \beta \sum_{t=1}^{T-1} \|\mathbf{q}_{t+1}\|^2 \mu_d^{-t} \sum_{i=t+1}^{T} \mu_d^i \\
&= \beta \sum_{t=1}^{T-1} \frac{\mu_d - \mu_d^{T-t+1}}{1 - \mu_d} \|\mathbf{q}_{t+1}\|^2 \\
&\leq \frac{\beta \mu_d}{1 - \mu_d} \sum_{t=1}^{T-1} \|\mathbf{q}_{t+1}\|^2.
\end{aligned} \tag{B.5}$$

Furthermore, we have

$$
\begin{aligned}
-\sum_{t=1}^{T}\sum_{i=1}^{t}\mu_d^{t-i}\alpha_i\nabla R_{\mathcal{S}}(\mathbf{w}_i)^\top\nabla\ell(\mathbf{w}_i;\mathbf{z}_{\mathsf{i}_i}) &= -\sum_{t=1}^{T}\sum_{i=1}^{t}\mu_d^{t-i}\alpha_i\nabla R_{\mathcal{S}}(\mathbf{w}_i)^\top\nabla R_{\mathcal{S}}(\mathbf{w}_i)\\
&\quad+\sum_{t=1}^{T}\sum_{i=1}^{t}\mu_d^{t-i}\alpha_i\nabla R_{\mathcal{S}}(\mathbf{w}_i)^\top(\nabla R_{\mathcal{S}}(\mathbf{w}_i)-\nabla\ell(\mathbf{w}_i;\mathbf{z}_{\mathsf{i}_i}))\\
&= -\sum_{t=1}^{T}\mu_d^{-t}\alpha_t\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2\sum_{i=t}^{T}\mu_d^i\\
&\quad+\sum_{t=1}^{T}\mu_d^{-t}\alpha_t\nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top(\nabla R_{\mathcal{S}}(\mathbf{w}_t)-\nabla\ell(\mathbf{w}_t;\mathbf{z}_{\mathsf{i}_t}))\sum_{i=t}^{T}\mu_d^i\\
&= -\sum_{t=1}^{T}\frac{1-\mu_d^{T-t+1}}{1-\mu_d}\alpha_t\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2\\
&\quad+\sum_{t=1}^{T}\frac{1-\mu_d^{T-t+1}}{1-\mu_d}\alpha_t\nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top(\nabla R_{\mathcal{S}}(\mathbf{w}_t)-\nabla\ell(\mathbf{w}_t;\mathbf{z}_{\mathsf{i}_t}))
\end{aligned}
\tag{B.6}
$$

Substituting Eqs. (B.5) and (B.6) into Eq. (B.4) and rearranging the terms, we obtain

$$
\begin{aligned}
\sum_{t=1}^{T}\frac{1-\mu_d^{T-t+1}}{1-\mu_d}\alpha_t\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2 &\le R_{\mathcal{S}}(\mathbf{w}_0)-R_{\mathcal{S}}(\mathbf{w}_{T+1})+\frac{\beta}{2}\sum_{t=1}^{T}\|\mathbf{q}_{t+1}\|^2\\
&\quad+\sum_{t=1}^{T}\frac{1-\mu_d^{T-t+1}}{1-\mu_d}\alpha_t\nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top(\nabla R_{\mathcal{S}}(\mathbf{w}_t)-\nabla\ell(\mathbf{w}_t;\mathbf{z}_{\mathsf{i}_t}))\\
&\quad+\frac{\beta\mu_d}{1-\mu_d}\sum_{t=1}^{T-1}\|\mathbf{q}_{t+1}\|^2.
\end{aligned}
\tag{B.7}
$$

We now find an upper bound on $\sum_{t=1}^{T} \|\mathbf{q}_{t+1}\|^2$:

$$\begin{aligned}
\sum_{t=1}^{T} \|\mathbf{q}_{t+1}\|^2 &= \sum_{t=1}^{T} \|\mu_d \mathbf{q}_t - \alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 \\
&\leq \sum_{t=1}^{T} \mu_d \|\mathbf{q}_t\|^2 + \sum_{t=1}^{T} \frac{1}{1-\mu_d} \|\alpha_t \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 \\
&\leq \sum_{t=1}^{T+1} \mu_d \|\mathbf{q}_t\|^2 + \sum_{t=1}^{T} \frac{1}{1-\mu_d} \|\alpha_t \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 \\
&\leq \sum_{t=1}^{T} \mu_d \|\mathbf{q}_{t+1}\|^2 + \sum_{t=1}^{T} \frac{1}{1-\mu_d} \|\alpha_t \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 \\
&\leq \sum_{t=1}^{T} \frac{1}{(1-\mu_d)^2} \|\alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 \\
&\leq \frac{L^2}{(1-\mu_d)^2} \sum_{t=1}^{T} \alpha_t^2.
\end{aligned}$$

where the second and last lines hold due to Jensen's inequality and L-Lipschitz property, respectively.

Applying this upper bound in Eq. (B.7), taking expectation over $\mathsf{i}_0, \ldots, \mathsf{i}_{t_d}$, we have:

$$\begin{aligned}
\mathbb{E}_A\Big[\sum_{t=1}^{t_d} \frac{1-\mu_d^{t_d-t+1}}{1-\mu_d} \alpha_t \|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2\Big] \leq R_{\mathcal{S}}(\mathbf{w}_0) - \inf_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}) + \frac{\beta L^2}{2(1-\mu_d)^2} \sum_{t=1}^{t_d} \alpha_t^2 \\
+ \frac{\beta \mu_d L^2}{(1-\mu_d)^3} \sum_{t=1}^{t_d-1} \alpha_t^2.
\end{aligned} \tag{B.8}$$

For $t = t_d + 1, \ldots, T$, using smoothness property, we have:

$$\begin{aligned}
R_{\mathcal{S}}(\mathbf{w}_{t+1}) &\leq R_{\mathcal{S}}(\mathbf{w}_t) - \alpha_t \nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}) + \frac{\beta}{2} \|\nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 \\
&\leq R_{\mathcal{S}}(\mathbf{w}_t) - \alpha_t \nabla R_{\mathcal{S}}(\mathbf{w}_t)^\top \nabla \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}) + \frac{\beta L^2}{2}.
\end{aligned}$$

Using a similar argument, we can find the following upper bound when the momentum is set to zero:

$$\mathbb{E}_A[\sum_{t=t_d+1}^{T} \alpha_t \|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2] \leq R_{\mathcal{S}}(\mathbf{w}_0) - \inf_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}) + \frac{\beta L^2}{2} \sum_{t=t_d+1}^{T} \alpha_t^2. \tag{B.9}$$

Adding Eqs. (B.8) and (B.9), we have

$$\begin{aligned}
&\mathbb{E}_A\Big[\sum_{t=1}^{t_d}\frac{1-\mu_d^{t_d-t+1}}{1-\mu_d}\alpha_t\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2+\sum_{t=t_d+1}^{T}\alpha_t\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2\Big]\\
&\leq 2\big(R_{\mathcal{S}}(\mathbf{w}_0)-\inf_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{w})\big)+\frac{\beta L^2}{2(1-\mu_d)^2}\sum_{t=1}^{t_d}\alpha_t^2+\frac{\beta\mu_d L^2}{(1-\mu_d)^3}\sum_{t=1}^{t_d-1}\alpha_t^2+\frac{\beta L^2}{2}\sum_{t=t_d+1}^{T}\alpha_t^2\\
&\leq 2\big(R_{\mathcal{S}}(\mathbf{w}_0)-\inf_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{w})\big)+\max\Big\{\frac{\beta L^2}{2(1-\mu_d)^2},\frac{\beta\mu_d L^2}{(1-\mu_d)^3},\frac{\beta L^2}{2}\Big\}\sum_{t=1}^{T}\alpha_t^2.
\end{aligned}$$

On the other hand, we find a lower bound on the left hand side:

$$\begin{aligned}
&\mathbb{E}_A\Big[\sum_{t=1}^{t_d}\frac{1-\mu_d^{t_d-t+1}}{1-\mu_d}\alpha_t\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2+\sum_{t=t_d+1}^{T}\alpha_t\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2\Big]\\
&\geq \min\Big\{\min_{1\leq t\leq t_d}\frac{1-\mu_d^{t_d-t+1}}{1-\mu_d},1\Big\}\mathbb{E}_A\big[\sum_{t=1}^{T}\alpha_t\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2\big]\\
&\geq \sum_{t=1}^{T}\alpha_t\min_{1\leq t\leq T}\mathbb{E}_A\big[\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2\big].
\end{aligned}$$

Combining the above upper bound and lower bound, we have

$$\begin{aligned}
\min_{1\leq t\leq T}\mathbb{E}_A\big[\|\nabla R_{\mathcal{S}}(\mathbf{w}_t)\|^2\big]\leq\;&\frac{2\big(R_{\mathcal{S}}(\mathbf{w}_0)-\inf_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{w})\big)}{\sum_{t=1}^{T}\alpha_t}\\
&+\frac{\max\Big\{\frac{\beta L^2}{2(1-\mu_d)^2},\frac{\beta\mu_d L^2}{(1-\mu_d)^3},\frac{\beta L^2}{2}\Big\}\sum_{t=1}^{T}\alpha_t^2}{\sum_{t=1}^{T}\alpha_t}
\end{aligned}\tag{B.10}$$

which completes the proof. In particular, by substituting step-size $\alpha_t=\alpha_0/t^q$, we achieve convergence with the rate of $\mathcal{O}(T^{q-1})$ for SGDEM with any t_d.

Appendix C. Convergence of SGDEM with constant step-size

Theorem 25 *Suppose that ℓ satisfies Assumption 1 and that the SGDEM update is executed for T steps with constant step-size $\alpha<2(1-\mu_d)$ and momentum $\mu_d\in(0,1)$ in the first t_d steps. Then, for any $\mathcal{S}$ and $0<t_d\leq T$, we have*

$$\min_{t=0,\ldots,T}\epsilon(t)\leq\frac{W+J_2}{W_1}\tag{C.1}$$

where $\epsilon(t):=\mathbb{E}_A[\|\nabla_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{w}_t)\|^2]$, $J_2=(t_d+1)\big(\frac{\beta}{2}\big(\frac{\alpha L}{1-\mu_d}\big)^2+\frac{1}{2}\big(\frac{\alpha\beta L\mu_d}{(1-\mu_d)^2}\big)^2\big)+(T-t_d)\frac{\beta}{2}\alpha^2L^2$, $W_1=(t_d+1)\big(\frac{\alpha}{1-\mu_d}-\frac{\alpha^2}{2(1-\mu_d)^2}\big)+(T-t_d)\alpha$, and $W=\mathbb{E}_A[R_{\mathcal{S}}(\mathbf{w}_0)-R_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}^*)]$ with $\mathbf{w}_{\mathcal{S}}^*=\arg\min_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{w})$.

Proof We analyze the convergence of SGDEM for a smooth Lipschitz loss function with constant step-size. To facilitate the convergence analysis, for $t \in [t_d]$, we define $\mathbf{p}_t := \frac{\mu_d}{1-\mu_d}(\mathbf{w}_t - \mathbf{w}_{t-1})$ with $\mathbf{p}_0 = 0$. Substituting this into the SGDEM update, the parameter recursion is given by

$$\mathbf{w}_{t+1} + \mathbf{p}_{t+1} = \mathbf{w}_t + \mathbf{p}_t - \frac{\alpha}{1-\mu_d}\nabla_{\mathbf{w}}\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}). \tag{C.2}$$

We also define $\mathbf{x}_t := \mathbf{w}_t + \mathbf{p}_t$. Note that for a β-smooth function f and for all $\mathbf{u},\ \mathbf{v} \in \Psi$, we have

$$f(\mathbf{u}) \leq f(\mathbf{v}) + \nabla f(\mathbf{v})^\top(\mathbf{u} - \mathbf{v}) + \frac{\beta}{2}\|\mathbf{u} - \mathbf{v}\|^2. \tag{C.3}$$

We note that $\mathbf{x}_t = \mathbf{w}_t$ for $t > t_d$. Let $t \in [t_d]$. Since the empirical risk $R_{\mathcal{S}}$ is a β-smooth function, we have

$$\begin{aligned} R_{\mathcal{S}}(\mathbf{x}_{t+1}) &\leq R_{\mathcal{S}}(\mathbf{x}_t) + \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{\beta\alpha^2}{2(1-\mu_d)^2}\|\nabla_{\mathbf{w}}\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 \\ &\leq R_{\mathcal{S}}(\mathbf{x}_t) + \frac{\beta}{2}\Big(\frac{\alpha L}{1-\mu_d}\Big)^2 - \frac{\alpha}{1-\mu_d}\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t)^\top \nabla_{\mathbf{w}}\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}) \end{aligned} \tag{C.4}$$

where we use the fact that $\|\nabla_{\mathbf{w}}\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\| \leq L$, due to the L-Lipschitz property.

Upon taking the expectation w.r.t. i_t in Eq. (C.4) and defining $\mathsf{r}_t := R_{\mathcal{S}}(\mathbf{x}_{t+1}) - R_{\mathcal{S}}(\mathbf{x}_t)$, we have

$$\begin{aligned} \mathbb{E}_{\mathsf{i}_t}[\mathsf{r}_t] &\leq -\frac{\alpha}{1-\mu_d}\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t)^\top \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t) + \frac{\beta}{2}\Big(\frac{\alpha L}{1-\mu_d}\Big)^2 \\ &= -\frac{\alpha}{1-\mu_d}\big(\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\big)^\top \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t) - \frac{\alpha}{1-\mu_d}\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2 + \frac{\beta}{2}\Big(\frac{\alpha L}{1-\mu_d}\Big)^2 \\ &\leq \frac{1}{2}\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2 + \frac{\beta}{2}\Big(\frac{\alpha L}{1-\mu_d}\Big)^2 + \Big(\frac{\alpha^2}{2(1-\mu_t)^2} - \frac{\alpha}{1-\mu_d}\Big)\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2 \end{aligned} \tag{C.5}$$

where the last inequality is obtained using $2\mathbf{u}^\top\mathbf{v} \leq \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$. For $t > t_d$, we have

$$\begin{aligned} \mathbb{E}_{\mathsf{i}_t}[\mathsf{r}_t] &\leq -\alpha\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)^\top \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t) + \frac{\beta}{2}(\alpha L)^2 \\ &\leq \frac{\beta}{2}(\alpha L)^2 - \alpha\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2. \end{aligned} \tag{C.6}$$

In the following, we obtain an upper bound on $\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2$ in Eq. (C.5) for $t \in [t_d]$.

Since $R_{\mathcal{S}}$ is β-smooth, we have

$$\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2 \leq \beta^2\|\mathbf{x}_t - \mathbf{w}_t\|^2. \tag{C.7}$$

We also note that $\beta^2\|\mathbf{x}_t - \mathbf{w}_t\|^2 = \frac{\beta^2\mu_d^2}{(1-\mu_d)^2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2$.

For notational simplicity, we define $\mathbf{q}_t := \frac{1-\mu_d}{\mu_d}\mathbf{p}_t$ with $\mathbf{q}_0 = 0$. Rewriting the SGDEM update rule, the parameter recursion is given by

$$\mathbf{q}_{t+1} = \mu_d\mathbf{q}_t - \alpha\nabla_{\mathbf{w}}\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}). \tag{C.8}$$

Unraveling the recursion Eq. (C.8), we have

$$\begin{aligned}\mathbf{q}_t &= -\alpha\sum_{k=0}^{t-1}\mu_d^{t-1-k}\nabla_{\mathbf{w}}\ell(\mathbf{w}_k; \mathbf{z}_{\mathsf{i}_k}) \\ &= -\alpha\sum_{k=0}^{t-1}\mu_d^{k}\nabla_{\mathbf{w}}\ell(\mathbf{w}_{t-1-k}; \mathbf{z}_{\mathsf{i}_{t-1-k}}).\end{aligned} \tag{C.9}$$

We define $\Theta_{t-1} := \sum_{k=0}^{t-1}\mu_d^k = \frac{1-\mu_d^t}{1-\mu_d}$. Then we can find an upper bound on $\|\mathbf{q}_t\|$ as follows:

$$\begin{aligned}\|\mathbf{q}_t\| &= \|-\alpha\sum_{k=0}^{t-1}\mu_d^k\nabla_{\mathbf{w}}\ell(\mathbf{w}_{t-1-k}; \mathbf{z}_{\mathsf{i}_{t-1-k}})\| \\ &= \alpha\|\sum_{k=0}^{t-1}\mu_d^k\nabla_{\mathbf{w}}\ell(\mathbf{w}_{t-1-k}; \mathbf{z}_{\mathsf{i}_{t-1-k}})\| \\ &\le \alpha\sum_{k=0}^{t-1}\mu_d^k\|\nabla_{\mathbf{w}}\ell(\mathbf{w}_{t-1-k}; \mathbf{z}_{\mathsf{i}_{t-1-k}})\| \\ &\le \alpha\Theta_{t-1}L \\ &\le \frac{\alpha L}{1-\mu_d}.\end{aligned} \tag{C.10}$$

Substituting Eq. (C.10) into Eq. (C.7), we obtain the following upper bound on $\|\nabla_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{w}_t)\|^2$:

$$\|\nabla_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{w}_t)\|^2 \le \frac{\alpha^2\beta^2L^2\mu_d^2}{(1-\mu_d)^4}. \tag{C.11}$$

Substituting Eq. (C.11) into Eq. (C.5) and taking expectation over $\mathsf{i}_0, \ldots, \mathsf{i}_t$, we have

$$\mathbb{E}_A[\mathsf{r}_t] \le -\Big(\frac{\alpha}{1-\mu_d} - \frac{\alpha^2}{2(1-\mu_d)^2}\Big)\mathbb{E}_A[\|\nabla_{\mathbf{w}}R_{\mathcal{S}}(\mathbf{w}_t)\|^2] + \frac{\beta}{2}\Big(\frac{\alpha L}{1-\mu_d}\Big)^2 + \frac{1}{2}\Big(\frac{\alpha\beta L\mu_d}{(1-\mu_d)^2}\Big)^2. \tag{C.12}$$

Summing Eq. (C.12) for $t \in [t_d]$ and Eq. (C.6) for $t = t_d + 1, \ldots, T$, we have

$$\begin{aligned}J_1 &\le \mathbb{E}_A[R_{\mathcal{S}}(\mathbf{x}_0) - R_{\mathcal{S}}(\mathbf{x}_{t+1})] + J_2 \\ &\le \mathbb{E}_A[R_{\mathcal{S}}(\mathbf{w}_0) - R_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}^*)] + J_2\end{aligned} \tag{C.13}$$

where

$$J_1 = \Big(\frac{\alpha}{1-\mu_d} - \frac{\alpha^2}{2(1-\mu_d)^2}\Big)\sum_{t=0}^{t_d}\mathbb{E}_A[\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2] + \alpha\sum_{t=t_d+1}^{T}\mathbb{E}_A[\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2] \tag{C.14}$$

and

$$J_2 = (t_d+1)\Big(\frac{\beta}{2}\Big(\frac{\alpha L}{1-\mu_d}\Big)^2 + \frac{1}{2}\Big(\frac{\alpha\beta L\mu_d}{(1-\mu_d)^2}\Big)^2\Big) + (T-t_d)\frac{\beta}{2}\alpha^2L^2.$$

Noting $\alpha \leq 2(1-c)(1-\mu_d)$ for some $0 < c < 1$, we obtain the following lower bound on J_1 in Eq. (C.14):

$$\begin{aligned} J_1 &\geq \frac{(t_d+1)\alpha c}{1-\mu_d}\min_{t=0,\ldots,t_d}\epsilon(t) + (T-t_d)\alpha\min_{t=t_d+1,\ldots,T}\epsilon(t) \\ &\geq \chi_3\min_{t=0,\ldots,T}\epsilon(t) \end{aligned} \tag{C.15}$$

where $\chi_3 = \frac{(t_d+1)\alpha c}{1-\mu_d} + (T-t_d)\alpha$.

Substituting Eq. (C.15) into Eq. (C.13), we obtain Eq. (C.1), which completes the proof.

■

We cannot directly combine standard bounds for SGDM and SGD to analyze convergence of SGDEM because the standard analysis requires characterization of the empirical risk at $\mathbf{w}_{t_d}$. Instead our proof is inspired by the convergence proof for SGDM by carefully handling time-varying momentum. We now study the upper bound (C.1) as a function of t_d for a given μ_d. Note that the first term in the upper bound vanishes as $T \to \infty$.

Remark 26 *In Appendix E, we provide a sufficient condition for the upper bound* (C.1) *to become a monotonically decreasing function of t_d. In Theorem 25, $\frac{J_2}{W_1} \approx \frac{aT+bt_d}{cT+dt_d}$ for some a, b, c, d. We may provide a looser bound by establishing an upper bound on J_2 and a lower bound on W_1. However, such looser bound is not useful since we will not be able to recover standard bounds for SGD and SGDM. In order to provide a simpler expression and understand how adding momentum affects the convergence, in Appendix F we study the convergence bound for a special form of SGDEM and show the benefit of using momentum. We also provide a simple sufficient condition for the non-vanishing term in the convergence bound to become a monotonically decreasing function of μ_d.*

We also establish convergence guarantees for SGDEM with another time-dependent step-size in Appendix D.

Remark 27 *In Theorem 25, our focus is on guaranteeing convergence to a local minimum, which holds for any $t_d \leq T$. We note that optimizing the upper bound in* (C.1) over t_d will not provide much intuition on the optimal t_d in terms of training error since we cannot guarantee the actual suboptimality gap (optimization error) of nonconvex loss functions. In practice, we need to tune t_d when training, *e.g.,* neural networks. Our experimental results show that a nontrivial t_d can be optimal in terms of test error.

Appendix D. Convergence guarantees for SGDEM with time-dependent and time-decaying step-sizes

We establish convergence guarantees for SGDEM with time-dependent and time-decaying step-sizes as follows.

Theorem 28 *Suppose that ℓ satisfies Assumption 1 and that the SGDEM update is executed for T steps with momentum μ_d in the first t_d steps and time-dependent step-size $\alpha = \min\{2(1-c)(1-\mu_d), \frac{K}{\max\{\sqrt{t_d+1}, \sqrt{T-t_d}\}}\}$ for some $0 < c < 1$ and $0 < K$. Then, for any $\mathcal{S}$ and $0 < t_d \leq T$, we have*

$$\min_{t=0,\ldots,T} \epsilon(t) \leq \frac{\tilde{T}(W + \tilde{J}_2)}{\tilde{W}_1} \tag{D.1}$$

where

$$\begin{aligned}
\tilde{J}_2 &= \frac{\beta}{2}\left(\frac{KL}{1-\mu_d}\right)^2 + \frac{1}{2}\left(\frac{K\beta L\mu_d}{(1-\mu_d)^2}\right)^2 + \frac{\beta}{2}K^2L^2, \\
\tilde{W}_1 &= \frac{(t_d+1)c}{1-\mu_d} + T - t_d, \\
\tilde{T} &= \max\left\{\frac{1}{2(1-c)(1-\mu_d)}, \frac{\max\{\sqrt{t_d+1}, \sqrt{T-t_d}\}}{K}\right\}.
\end{aligned}$$

Theorem 29 *Suppose that ℓ satisfies Assumption 1 and that the SGDEM update is executed for T steps with time-decaying step-size $\alpha_t = \frac{\alpha_0}{t+1}$ for $t = 0, 1, \ldots, T$ with $\alpha_0 \leq 2(1-c)(1-\mu_d)$ for some $0 < c < 1$ and momentum $\mu_d > \exp(-1)$ in the first t_d steps. Then, for any $\mathcal{S}$ and $0 < t_d \leq T$, we have*

$$\min_{t=0,\ldots,T} \epsilon(t) \leq \frac{W + \hat{J}_2}{\hat{W}_1} \tag{D.2}$$

where

$$\begin{aligned}
\hat{J}_2 &= \beta\left(\frac{\alpha_0 L}{1-\mu_d}\right)^2 + \frac{\beta}{2}(\alpha_0 L)^2 \frac{1}{t_d+1} + \sum_{t=1}^{t_d}\left(\frac{\alpha_0 \bar{c}_t \beta L \mu_d}{1-\mu_d}\right)^2, \\
\hat{W}_1 &= \frac{\ln(t_d+1)\alpha_0 c}{1-\mu_d} + \ln\left(\frac{T}{t_d+2}\right)\alpha_0, \\
\bar{c}_t &= \min\left\{\frac{1}{1-\mu_d}, 1 + \ln(t), \mu_d^t(\mu_d^{-1} + I(t))\right\}, \text{ and } I(t) = \int_1^t \frac{\mu_d^{-u}}{u}\,\mathrm{d}u.
\end{aligned}$$

Proof Following the proof of Theorem 25 for $t \in [t_d]$, we have

$$\mathbb{E}_{\mathsf{i}_t}[\mathsf{r}_t] \leq \frac{1}{2}\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2 + \frac{\beta}{2}\Big(\frac{\alpha_t L}{1-\mu_d}\Big)^2 + \Big(\frac{\alpha_t^2}{2(1-\mu_d)^2} - \frac{\alpha_t}{1-\mu_d}\Big)\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2. \tag{D.3}$$

For $t > t_d$, we have $\mathbf{x}_t = \mathbf{w}_t$ and

$$\begin{aligned}\mathbb{E}_{\mathsf{i}_t}[\mathsf{r}_t] &\leq -\alpha_t \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)^\top \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t) + \frac{\beta}{2}(\alpha_t L)^2 \\ &\leq \frac{\beta}{2}(\alpha_t L)^2 - \alpha_t \|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2.\end{aligned} \tag{D.4}$$

In the following, we obtain an upper bound on $\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2$ in Eq. (D.3). Since $R_{\mathcal{S}}$ is β-smooth, we have

$$\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2 \leq \beta^2 \|\mathbf{x}_t - \mathbf{w}_t\|^2. \tag{D.5}$$

We also note that

$$\beta^2 \|\mathbf{x}_t - \mathbf{w}_t\|^2 = \frac{\beta^2 \mu_d^2}{(1-\mu_d)^2}\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2.$$

For notational simplicity, we define $\mathbf{q}_t := \frac{1-\mu_d}{\mu_d}\mathbf{p}_t$ with $\mathbf{q}_0 = 0$. Rewriting the (SGDEM) update rule, the parameter recursion is given by

$$\mathbf{q}_{t+1} = \mu_d \mathbf{q}_t - \alpha_t \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}). \tag{D.6}$$

Unraveling the recursion Eq. (D.6), we have

$$\mathbf{q}_t = -\alpha_0 \sum_{k=0}^{t-1} \frac{\mu_d^{t-1-k}}{k+1} \nabla_{\mathbf{w}} \ell(\mathbf{w}_k; \mathbf{z}_{\mathsf{i}_k}). \tag{D.7}$$

Lemma 30 *Provided that $\mu_d \geq \exp(-1)$, we have $\|\mathbf{q}_t\|^2 \leq \bar{c}_t^2 \alpha_0^2 L^2$ for $t \leq t_d$, where*

$$\bar{c}_t = \min\{\frac{1}{1-\mu_d}, 1 + \ln(t), \mu_d^t(\mu_d^{-1} + I(t))\} \text{ and } I(t) = \int_1^t \frac{\mu_d^{-u}}{u}\,\mathrm{d}u.$$

Proof *Following the proof of Theorem 25, an upper bound on $\|\mathbf{q}_t\|^2$ is given by*

$$\|\mathbf{q}_t\|^2 \leq \alpha_0^2 L^2 \tilde{S}^2$$

where

$$\tilde{S} = \sum_{k=0}^{t-1} \frac{\mu_d^{t-1-k}}{k+1}.$$

Note that

$$\tilde{S} \leq \sum_{k=0}^{t-1} \mu_d^k = \frac{1-\mu_d^t}{1-\mu_d} \text{ and } \tilde{S} \leq \sum_{k=1}^{t} 1/k \leq 1 + \int_1^t 1/u \, \mathrm{d}u = 1 + \ln(t).$$

Rewriting $\tilde{S}$ *as* $\tilde{S} = \mu_d^t \sum_{k=1}^t \frac{\mu_d^{-k}}{k}$ *and noting* $f(u) = \mu_d^{-u}/u$ *is convex and non-increasing for* $1 \leq u \leq t$ *due to the lower bound* $\mu_d \geq \exp(-1)$*. Therefore, we have* $\tilde{S} \leq \mu_d^t(\mu_d^{-1} + I(t))$*.* ■

Substituting the upper bound on $\|\mathbf{q}_t\|^2$ into Eq. (D.5), we obtain the following upper bound on $\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2$:

$$\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{x}_t) - \nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2 \leq \frac{\alpha_0^2 \bar{c}_t^2 \beta^2 L^2 \mu_d^2}{(1-\mu_d)^2}. \tag{D.8}$$

Substituting Eq. (D.8) into Eq. (D.3) and taking expectation over $\mathsf{i}_0, \ldots, \mathsf{i}_t$, we have

$$\mathbb{E}_A[\mathsf{r}_t] \leq -\Big(\frac{\alpha_t}{1-\mu_t} - \frac{\alpha_t^2}{2(1-\mu_d)^2}\Big)\mathbb{E}_A[\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2] + \frac{\beta}{2}\Big(\frac{\alpha_t L}{1-\mu_d}\Big)^2 + \Big(\frac{\alpha_0 \bar{c}_t \beta L \mu_d}{1-\mu_d}\Big)^2. \tag{D.9}$$

Summing Eq. (D.9) for $t \in [t_d]$ and Eq. (D.4) for $t = t_d + 1, \ldots, T$, we have

$$\begin{aligned} \hat{J}_1 &\leq \mathbb{E}_A[R_{\mathcal{S}}(\mathbf{x}_0) - R_{\mathcal{S}}(\mathbf{x}_{t+1})] + \hat{J}_2 \\ &\leq \mathbb{E}_A[R_{\mathcal{S}}(\mathbf{w}_0) - R_{\mathcal{S}}(\mathbf{w}_{\mathcal{S}}^*)] + \hat{J}_2 \end{aligned} \tag{D.10}$$

where

$$\hat{J}_1 = \sum_{t=0}^{t_d} \Big(\frac{\alpha_t}{1-\mu_d} - \frac{\alpha_t^2}{2(1-\mu_d)^2}\Big)\mathbb{E}_A[\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2] + \sum_{t=t_d+1}^{T} \alpha_t \mathbb{E}_A[\|\nabla_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w}_t)\|^2]. \tag{D.11}$$

Noting $\alpha_0 \leq 2(1-c)(1-\mu_d)$ for some $0 < c < 1$, we obtain the following lower bound on $\hat{J}_1$ in Eq. (D.11):

$$\begin{aligned} \hat{J}_1 &\geq \sum_{t=0}^{t_d} \frac{\alpha_t c}{1-\mu_d} \min_{t=0,\ldots,t_d} \epsilon(t) + \sum_{t=t_d+1}^{T} \alpha_t \min_{t=t_d+1,\cdots,T} \epsilon(t) \\ &\geq \hat{\chi}_3 \min_{t=0,\ldots,T} \epsilon(t) \end{aligned} \tag{D.12}$$

where

$$\hat{\chi}_3 = \frac{\ln(t_d+1)\alpha_0 c}{1-\mu_d} + \ln\big(\frac{T}{t_d+2}\big)\alpha_0. \tag{D.13}$$

Finally, we note that

$$\frac{\beta}{2}\Big(\frac{\alpha_0 L}{1-\mu_d}\Big)^2 \sum_{t=0}^{t_d} \frac{1}{(t+1)^2} \leq \beta\Big(\frac{\alpha_0 L}{1-\mu_d}\Big)^2 \quad \text{and} \quad \frac{\beta}{2}(\alpha_0 L)^2 \sum_{t=t_d+1}^{T} \frac{1}{(t+1)^2} \leq \frac{\beta}{2}(\alpha_0 L)^2 \frac{1}{t_d+1}.$$

■

Appendix E. Sufficient condition for the upper bound in Theorem 25

In the following corollary, we provide a simple sufficient condition for the upper bound (C.1) to become a monotonically decreasing function of t_d.

Corollary 31 *Suppose that ℓ satisfies Assumption 1 and that SGDEM is executed for finite T steps with constant step-size $\alpha < 2c(1-\mu_d)$ with some $c < \frac{1}{2-\mu_d}$ and momentum μ_d in the first t_d steps. Then the upper bound* (C.1) *is a monotonically decreasing function of t_d if the following condition is satisfied:*

$$W > \frac{(K_1 - K_2)(K_3 + TK_4)}{K_3 - K_4} - K_1 - TK_2 \tag{E.1}$$

where $K_1 = \frac{\beta}{2}\big(\frac{\alpha L}{1-\mu_d}\big)^2 + \frac{1}{2}\big(\frac{\alpha\beta L\mu_d}{(1-\mu_d)^2}\big)^2$, $K_2 = \frac{\beta}{2}\alpha^2 L^2$, $K_3 = \frac{\alpha}{1-\mu_d} - \frac{\alpha^2}{2(1-\mu_d)^2}$, and $K_4 = \alpha$.

Proof Note that we can express the upper bound (C.1) as

$$U(t_d) = \frac{W + K_1 + TK_2 + t_d(K_1 - K_2)}{K_3 + TK_4 + t_d(K_3 - K_4)}.$$

The proof follows by taking the first derivative of U w.r.t. t_d. ■

Corollary 31 implies that adding momentum for a longer time is particularly useful when our initial parameter is sufficiently far from a local minimum.

Appendix F. Understanding the role of momentum on convergence

In order to understand how adding momentum affects the convergence, we study the convergence bound for a special form of SGDEM and show the benefit of using momentum.

Corollary 32 *Suppose we set $t_d = T$ with constant step-size $\alpha < 2(1-\mu_d)$. Then, for any $\mathcal{S}$, we have*

$$\min_{t=0,\cdots,T} \epsilon(t) \leq \frac{W}{(T+1)\big(\frac{\alpha}{1-\mu_d} - \frac{\alpha^2}{2(1-\mu_d)^2}\big)} + \frac{\beta\alpha^2 L^2 + \frac{(\alpha\beta L\mu_d)^2}{(1-\mu_d)^2}}{2\alpha(1-\mu_d) - \alpha^2}. \tag{F.1}$$

Note that the upper bound (F.1) is a function of μ_d. The first term in the upper bound vanishes as $T \to \infty$. In the following corollary, we provide a simple sufficient condition for the non-vanishing term in the upper bound (F.1) to become a monotonically decreasing function of μ_d.

Corollary 33 *Suppose $\beta \leq \frac{2\mu_d - \mu_d^2}{(1-\mu_d)^2}$ and we set $t_d = T$ with $\alpha \leq 2c(1-\mu_d)$ for some $0 < c < 1$. Then the non-vanishing term in the upper bound* (C.1) *is a monotonically decreasing function of μ_d.*

Proof Noting $\alpha \leq 2c(1-\mu_d)$ for some $0 < c < 1$, we obtain the following lower bound on J_1 in Eq. (C.14):

$$J_1 \geq \chi_4 \min_{t=0,\ldots,T} \epsilon(t) \tag{F.2}$$

where

$$\chi_4 := \frac{(t_d+1)\alpha(1-c)}{1-\mu_d} + T - t_d.$$

Substituting Eq. (F.2) into Eq. (C.1), the non-vanishing term in the upper bound $\frac{2L^2\beta(1-\mu_d)}{\alpha(1-c)}c^2\Big(1+\beta\frac{\mu_d^2}{(1-\mu_d)^2}\Big)$ becomes a function of μ_d through $(1-\mu_d)\Big(1+\beta\frac{\mu_d^2}{(1-\mu_d)^2}\Big)$. We can prove the proposition by taking the first derivative w.r.t. μ_d. ■

Appendix G. High-probability generalization bounds

We consider the generalization error that depends on a random set of n samples $\mathcal{S}$ drawn i.i.d. from some space $\mathcal{Z}$ with an unknown distribution D:

$$\epsilon_g(\mathcal{S}) = \mathbb{E}_A[R(\mathbf{w}_T) - R_{\mathcal{S}}(\mathbf{w}_T)].$$

We establish high-probability bounds for generalization error of SGDEM along the lines of (Feldman and Vondrak, 2018).

Theorem 34 (High-probability generalization bound) *Let $0 < \delta < 1$. For the setting described in Corollary 12, with probability at least $1-\delta$ over $\mathcal{S} \sim D^n$, the generalization error of SGDEM $\epsilon_g(\mathcal{S})$ is bounded by $\mathcal{O}\Big(\sqrt{(\frac{\exp(\mu_d)T^u}{n} + \frac{1}{n})\log(\frac{1}{\delta})}\Big)$.*

Proof *The proof follows by the arguments in (Feldman and Vondrak, 2018, Theorem 1.2) and substituting the stability upper bound in Corollary 12.* ■

Appendix H. Proof of Theorem 18

We track the divergence of two different iterative sequences of update rules with the same starting point. We remark that our analysis is more involved than (Hardt

et al., 2016) as the presence of momentum term requires a more careful bound on the iterative expressions.

To keep the notation uncluttered, we first consider SGDM without projection and defer the discussion of projection to the end of this proof. Let $\mathcal{S} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ and $\mathcal{S}' = \{\mathbf{z}'_1, \dots, \mathbf{z}'_n\}$ be two samples of size n that differ in at most one example. Let $\mathbf{w}_T$ and $\mathbf{w}'_T$ denote the outputs of SGDM on $\mathcal{S}$ and $\mathcal{S}'$, respectively. We consider the updates $\mathbf{w}_{t+1} = G_t(\mathbf{w}_t) + \mu(\mathbf{w}_t - \mathbf{w}_{t-1})$ and $\mathbf{w}'_{t+1} = G'_t(\mathbf{w}'_t) + \mu(\mathbf{w}'_t - \mathbf{w}'_{t-1})$ with $G_t(\mathbf{w}_t) = \mathbf{w}_t - \alpha\nabla_{\mathbf{w}}\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})$ and $G'_t(\mathbf{w}'_t) = \mathbf{w}'_t - \alpha\nabla_{\mathbf{w}}\ell(\mathbf{w}'_t; \mathbf{z}'_{\mathsf{i}_t})$, respectively, for $t = 1, \cdots, T$. We denote $\delta_t := \|\mathbf{w}_t - \mathbf{w}'_t\|$. Suppose $\mathbf{w}_0 = \mathbf{w}'_0$, *i.e.,* $\delta_0 = 0$.

We first establish an upper bound on $\mathbb{E}_A[\delta_T]$. At step t, with probability $1 - 1/n$, the example is the same in both $\mathcal{S}$ and $\mathcal{S}'$, *i.e.,* $\mathbf{z}_{\mathsf{i}_t} = \mathbf{z}'_{\mathsf{i}_t}$, which implies $G_t = G'_t$. Then G_t becomes $(1 - \frac{\alpha\beta\gamma}{\beta+\gamma})$-expansive for $\alpha \leq \frac{2}{\beta+\gamma}$ (see, *e.g.,* (Hardt et al., 2016, Appendix A)). Hence, we have

$$\begin{aligned}
\delta_{t+1} &= \|\mu(\mathbf{w}_t - \mathbf{w}'_t) - \mu(\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}) + G_t(\mathbf{w}_t) - G_t(\mathbf{w}'_t)\| \\
&\leq \mu\|\mathbf{w}_t - \mathbf{w}'_t\| + \mu\|\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}\| + \|G_t(\mathbf{w}_t) - G_t(\mathbf{w}'_t)\| \\
&\leq \vartheta\delta_t + \mu\delta_{t-1}
\end{aligned} \tag{H.1}$$

where $\vartheta = 1 + \mu - \frac{\alpha\beta\gamma}{\beta+\gamma}$. With probability $1/n$, the selected example is different in $\mathcal{S}$ and $\mathcal{S}'$. In this case, we have

$$\begin{aligned}
\delta_{t+1} &= \|\mu(\mathbf{w}_t - \mathbf{w}'_t) - \mu(\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}) + G_t(\mathbf{w}_t) - G'_t(\mathbf{w}'_t)\| \\
&\leq \mu\|\mathbf{w}_t - \mathbf{w}'_t\| + \mu\|\mathbf{w}_{t-1} - \mathbf{w}'_{t-1}\| + \phi_3 \\
&\leq \vartheta\delta_t + \mu\delta_{t-1} + \|G_t(\mathbf{w}'_t) - G'_t(\mathbf{w}'_t)\| \\
&\leq \vartheta\delta_t + \mu\delta_{t-1} + \|\mathbf{w}'_t - G_t(\mathbf{w}'_t)\| + \|\mathbf{w}'_t - G'_t(\mathbf{w}'_t)\| \\
&\leq \vartheta\delta_t + \mu\delta_{t-1} + 2\alpha L
\end{aligned} \tag{H.2}$$

where $\phi_3 = \|G_t(\mathbf{w}_t) + G_t(\mathbf{w}'_t) - G_t(\mathbf{w}'_t) - G'_t(\mathbf{w}'_t)\|$. The last inequality in (H.2) holds due to the L-Lipschitz property. Combining Eqs. (H.1) and (H.2), we have

$$\begin{aligned}
\mathbb{E}_A[\delta_{t+1}] &\leq (1 - 1/n)\big(\vartheta\mathbb{E}_A[\delta_t] + \mu\mathbb{E}_A[\delta_{t-1}]\big) + 1/n\big(\vartheta\mathbb{E}_A[\delta_t] + \mu\mathbb{E}_A[\delta_{t-1}] + 2\alpha L\big) \\
&= \vartheta\mathbb{E}_A[\delta_t] + \mu\mathbb{E}_A[\delta_{t-1}] + \frac{2\alpha L}{n}.
\end{aligned} \tag{H.3}$$

Let us consider the recursion

$$\mathbb{E}_A[\tilde{\delta}_{t+1}] = \vartheta\mathbb{E}_A[\tilde{\delta}_t] + \mu\mathbb{E}_A[\tilde{\delta}_{t-1}] + \frac{2\alpha L}{n} \tag{H.4}$$

with $\tilde{\delta}_0 = \delta_0 = 0$. Upon inspecting Eq. (H.4) it is clear that

$$\mathbb{E}_A[\tilde{\delta}_t] \geq \vartheta\mathbb{E}_A[\tilde{\delta}_{t-1}], \qquad \forall t \geq 1, \tag{H.5}$$

as we simply drop the remainder of positive terms. Substituting Eq. (H.5) into Eq. (H.4), we have

$$\begin{aligned}\mathbb{E}_A[\tilde{\delta}_{t+1}] &\leq \Big(1+\mu+\frac{\mu}{\vartheta}-\frac{\alpha\beta\gamma}{\beta+\gamma}\Big)\mathbb{E}_A[\tilde{\delta}_t]+\frac{2\alpha L}{n} \\ &\leq (\vartheta+2\mu)\mathbb{E}_A[\tilde{\delta}_t]+\frac{2\alpha L}{n}\end{aligned} \tag{H.6}$$

where the second inequality holds due to $\mu \geq \frac{\alpha\beta\gamma}{\beta+\gamma}-\frac{1}{2}$.

Noting that $\mathbb{E}_A[\tilde{\delta}_t] \geq \mathbb{E}_A[\delta_t]$ for all t including T, we have

$$\mathbb{E}_A[\delta_T] \leq \frac{2\alpha L}{n}\sum_{t=1}^{T}(\vartheta+2\mu)^t \leq \frac{2\alpha L(\beta+\gamma)}{n\big(\alpha\beta\gamma-3\mu(\beta+\gamma)\big)}$$

where the second expression holds since $0 \leq \mu < \frac{\alpha\beta\gamma}{3(\beta+\gamma)}$.

Applying the L-Lipschitz property on $\ell(\cdot,\mathbf{z})$, it follows

$$\begin{aligned}\mathbb{E}_A[|\ell(\mathbf{w}_T;\mathbf{z})-\ell(\mathbf{w}'_T;\mathbf{z})|] &\leq L\mathbb{E}_A[\delta_T] \\ &\leq \frac{2\alpha L^2(\beta+\gamma)}{n\big(\alpha\beta\gamma-3\mu(\beta+\gamma)\big)}.\end{aligned} \tag{H.7}$$

Since this bound holds for all $\mathcal{S}$, $\mathcal{S}'$, and $\mathbf{z}$, we obtain an upper bound on the uniform stability and the proof is complete.

Our stability bound in above holds for the (P-SGDM) update because Euclidean projection onto a convex set does not increase the distance between projected points (Rockafellar, 1976). In particular, note that inequalities (H.1) and (H.2) still hold under P-SGDM.

Appendix I. Convergence bound for strongly convex loss

In this section, we develop an upper bound on the optimization error for the case of strongly convex loss, which is defined as

$$\epsilon_{\text{opt}} := \mathbb{E}_{\mathcal{S},A}[R_{\mathcal{S}}(\hat{\mathbf{w}}_T)-R_{\mathcal{S}}(\mathbf{w}^*_{\mathcal{S}})] \tag{I.1}$$

where $\hat{\mathbf{w}}_T$ denotes the average of T steps of the algorithm, *i.e.,* $\hat{\mathbf{w}}_T = \frac{1}{T+1}\sum_{t=0}^{T}\mathbf{w}_t$, $R_{\mathcal{S}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{w};\mathbf{z}_i)$, and $\mathbf{w}^*_{\mathcal{S}} = \arg\min_{\mathbf{w}} R_{\mathcal{S}}(\mathbf{w})$.

The optimization error quantifies the gap between the empirical risk of P-SGDM and the optimal empirical risk.

Theorem 35 *Suppose that ℓ satisfies Assumptions 1 and 2 and that P-SGDM is executed for T steps with constant step-size α and momentum μ. Then we have*[8]

$$\epsilon_{\text{opt}} \leq \frac{\mu W_0}{(1-\mu)T} + \frac{(1-\mu)W_1}{2\alpha T} - \frac{\gamma W_2}{2} - \frac{\mu\gamma W_3}{2(1-\mu)} + \frac{\alpha L^2}{2(1-\mu)} \tag{I.2}$$

where $W_0 = \mathbb{E}_{\mathcal{S},A}[R_{\mathcal{S}}(\mathbf{w}_0) - R_{\mathcal{S}}(\mathbf{w}_T)]$, $W_1 = \mathbb{E}_{\mathcal{S},A}[\|\mathbf{w}_0 - \mathbf{w}^*_{\mathcal{S}}\|^2]$, $W_2 = \mathbb{E}_{\mathcal{S},A}[\|\hat{\mathbf{w}}_T - \mathbf{w}^*_{\mathcal{S}}\|^2]$, *and* $W_3 = \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}_{\mathcal{S},A}[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2]$.

Proof Again, we first consider SGDM without projection and discuss the extension to projection at the end of this proof. To facilitate the convergence analysis, we define:

$$\mathbf{p}_t := \frac{\mu}{1-\mu}(\mathbf{w}_t - \mathbf{w}_{t-1}) \tag{I.3}$$

with $\mathbf{p}_0 = 0$. Substituting into SGDM, we have

$$\|\mathbf{s}_{t+1} - \mathbf{w}\|^2 = \|\mathbf{s}_t - \mathbf{w}\|^2 + \left(\frac{\alpha}{1-\mu}\right)^2\|\nabla_{\mathbf{w}}\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 - \frac{2\alpha}{1-\mu}(\mathbf{s}_t - \mathbf{w})^\top\nabla_{\mathbf{w}}\ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}) \tag{I.4}$$

where $\mathbf{s}_t = \mathbf{w}_t + \mathbf{p}_t$. Substituting $\mathbf{s}_t$, taking the expectation w.r.t. i_t, using the L-Lipschitz assumption, noting $R_{\mathcal{S}}$ is a γ-strongly convex function, summing for $t = 0, \ldots, T$, and rearranging terms, we have

$$\begin{aligned}\varsigma_T \leq{}& \frac{2\alpha\mu}{(1-\mu)^2}\mathbb{E}_A[R_{\mathcal{S}}(\mathbf{w}_0) - R_{\mathcal{S}}(\mathbf{w}_T)] - \frac{\alpha\gamma}{1-\mu}\sum_{t=0}^{T}\mathbb{E}_A[\|\mathbf{w}_t - \mathbf{w}\|^2] \\ &+ \frac{\alpha^2L^2(T+1)}{(1-\mu)^2} + \mathbb{E}_A[\|\mathbf{w}_0 - \mathbf{w}\|^2] - \frac{\alpha\mu\gamma}{(1-\mu)^2}\sum_{t=0}^{T}\mathbb{E}_A[\|\mathbf{w}_t - \mathbf{w}_{t-1}\|^2]\end{aligned} \tag{I.5}$$

where $\varsigma_T = \frac{2\alpha}{1-\mu}\sum_{t=0}^{T}\mathbb{E}_A[R_{\mathcal{S}}(\mathbf{w}_t) - R_{\mathcal{S}}(\mathbf{w})]$. Since $\|\cdot\|^2$ is a convex function, we have $\|\hat{\mathbf{w}}_T - \mathbf{w}\|^2 \leq \frac{1}{T+1}\sum_{t=0}^{T}\|\mathbf{w}_t - \mathbf{w}\|^2$ for all $\mathbf{w}_T$ and $\mathbf{w}$. Furthermore, due to the convexity of R_S, we have

$$R_{\mathcal{S}}(\hat{\mathbf{w}}_T) - R_{\mathcal{S}}(\mathbf{w}) \leq \frac{1}{T+1}\sum_{t=0}^{T}\left(R_{\mathcal{S}}(\mathbf{w}_t) - R_{\mathcal{S}}(\mathbf{w})\right). \tag{I.6}$$

Taking expectation over $\mathcal{S}$, applying the above inequalities, and substituting $\mathbf{w} = \mathbf{w}^*_{\mathcal{S}}$, we obtain Eq. (I.2).

8. Linear convergence results for SGD can be obtained under a stringent condition (Needell et al., 2014). Such a condition requires that the loss function is simultaneously minimized on each training example, and it does not apply to our setting. Different from (Yan et al., 2018; Ghadimi et al., 2015), we analyze the convergence of P-SGDM for a smooth and strongly convex loss function with constant step-size.

Our convergence bound in (I.2) can be extended to P-SGDM. Let use denote

$$\mathbf{y}_{t+1} := \mathbf{w}_t + \mu(\mathbf{w}_t - \mathbf{w}_{t-1}) - \alpha \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}).$$

Then, for any feasible $\mathbf{w} \in \Omega$, Eq. (I.4) holds for $\mathbf{y}_{t+1}$, *i.e.,*

$$\begin{aligned} \|\hat{\mathbf{y}}_{t+1} - \mathbf{w}\|^2 = \|\mathbf{s}_t - \mathbf{w}\|^2 + (\frac{\alpha}{1-\mu})^2 \|\nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t})\|^2 \\ - \frac{2\alpha}{1-\mu} (\mathbf{s}_t - \mathbf{w})^\top \nabla_{\mathbf{w}} \ell(\mathbf{w}_t; \mathbf{z}_{\mathsf{i}_t}) \end{aligned} \tag{I.7}$$

where $\hat{\mathbf{y}}_t = \mathbf{y}_t + \frac{\mu}{1-\mu}(\mathbf{y}_t - \mathbf{w}_{t-1})$.

Note that the LHS of Eq. (I.7) can be written as

$$\|\hat{\mathbf{y}}_{t+1} - \mathbf{w}\|^2 = \tfrac{1}{(1-\mu)^2} \|\mathbf{y}_{t+1} - \big(\mu \mathbf{w}_t + (1-\mu)\mathbf{w}\big)\|^2.$$

We note that $\tilde{\mathbf{w}}_t = \mu \mathbf{w}_t + (1-\mu)\mathbf{w} \in \Omega$ for any $\mathbf{w} \in \Omega$ and $\mathbf{w}_t \in \Omega$ since Ω is convex.

Now in projected SGDM, we have

$$\begin{aligned} \|\mathbf{w}_{t+1} - \tilde{\mathbf{w}}_t\|^2 &= \|\mathbf{P}(\mathbf{y}_{t+1}) - \tilde{\mathbf{w}}_t\|^2 \\ &\leq \|\mathbf{y}_{t+1} - \tilde{\mathbf{w}}_t\|^2 \end{aligned}$$

since projection a point onto Ω moves it closer to any point in Ω. This shows inequality (I.5) holds, and the convergence results do not change. ■

Theorem 35 bounds the optimization error, *i.e.,* the expected difference between the empirical risk achieved by SGDM and the global minimum. Upon setting $\mu = 0$ and $\gamma = 0$ in Eq. (I.2), we can recover the classical bound on optimization error for SGD (Nemirovskij and Yudin, 1983), (Hardt et al., 2016, Theorem 5.2). The first two terms in Eq. (I.2) vanish as T increases. The terms with negative sign improve the convergence due to the strongly convexity. The last term depends on the step-size, α, the momentum parameter μ, and the Lipschitz constant L. This term can be reduced by selecting α sufficiently small.

Appendix J. Upper bound on true risk

We now study how the uniform stability results in an upper bound on the true risk in the strongly convex case. We also compare the final results with SGD with no momentum and we show that one can achieve tighter bounds by using SGDM than vanilla SGD.

The expected true risk estimate under parameter $\hat{\mathbf{w}}_T$ can be decomposed into a stability error term and an optimization one. In Appendix I, we present an upper bound on the optimization error for strongly convex loss. The optimization error reflects the optimality gap when we optimize the empirical risk under some step-size

and momentum. By combining the result Appendix I and our stability error bound, and adjusting the hyper parameters, we minimize the upper bound on the expected true risk estimate.

In the following lemma, we show that stability results similar to Theorem 18 hold even if we consider the average parameter $\hat{\mathbf{w}}_T$ instead of $\mathbf{w}_T$. In other words, the same upper bound holds even if $\hat{\mathbf{w}}_T$ is considered as the output of algorithm A.

Lemma 36 *Suppose that that ℓ satisfies Assumptions 1 and 2 and P-SGDM is executed for T steps with step-size α and momentum μ. Provided that $\frac{\alpha\beta\gamma}{\beta+\gamma} - \frac{1}{2} \leq \mu < \frac{\alpha\beta\gamma}{3(\beta+\gamma)}$ and $\alpha \leq \frac{2}{\beta+\gamma}$, then the average of the first T steps of P-SGDM satisfies ϵ_s-uniform stability with Eq.* (4.1).

Proof Let us define $\hat{\mathbf{w}}_t = \frac{1}{t}\sum_{k=1}^{t} \mathbf{w}_k$ and $\hat{\delta}_t := \|\hat{\mathbf{w}}_t - \hat{\mathbf{w}}'_t\|$ where $\hat{\mathbf{w}}'_t$ is obtained as specified in the proof of Theorem 18. Following the proof of Theorem 18, we have

$$\mathbb{E}[\tilde{\delta}_{k+1}] \leq \Big(1 + 3\mu - \frac{\alpha\beta\gamma}{\beta+\gamma}\Big)\mathbb{E}[\tilde{\delta}_k] + \frac{2\alpha L}{n}. \tag{J.1}$$

for $k = 0, \ldots, T$. Defining $\overline{\delta}_t := \frac{\sum_{k=1}^{t} \tilde{\delta}_k}{t}$, we have $\hat{\delta}_T \leq \overline{\delta}_T$ by the triangle inequality. Summing Eq. (J.1) for $k = 0, \ldots, T$ and dividing by T, we have $\mathbb{E}[\hat{\delta}_T] \leq \mathbb{E}[\overline{\delta}_T] \leq \frac{2\alpha L(\beta+\gamma)}{n\big(\alpha\beta\gamma - 3\mu(\beta+\gamma)\big)}$. Applying the L-Lipschitz property on $\ell(\cdot, \mathbf{z})$, we have

$$\mathbb{E}[|\ell(\hat{\mathbf{w}}_T; \mathbf{z}) - \ell(\hat{\mathbf{w}}'_T; \mathbf{z})|] \leq \frac{2\alpha L^2(\beta+\gamma)}{n\big(\alpha\beta\gamma - 3\mu(\beta+\gamma)\big)}, \tag{J.2}$$

which holds for all $\mathcal{S}$, $\mathcal{S}'$, and $\mathbf{z}$. ■

Adding the stability error following Lemma 36, we have

$$\mathbb{E}_{\mathcal{S},A}[R(\hat{\mathbf{w}}_T)] \leq \mathbb{E}_{\mathcal{S},A}[R_{\mathcal{S}}(\hat{\mathbf{w}}_T)] + \epsilon_s \leq \mathbb{E}_{\mathcal{S},A}[R_{\mathcal{S}}(\mathbf{w}^*_{\mathcal{S}})] + \epsilon_{\text{opt}} + \epsilon_s \tag{J.3}$$

where $\epsilon_{\text{opt}} := \mathbb{E}_{\mathcal{S},A}[R_{\mathcal{S}}(\hat{\mathbf{w}}_T) - R_{\mathcal{S}}(\mathbf{w}^*_{\mathcal{S}})]$.

Note that there is a tradeoff between the optimization error and stability one. We can balance these errors to achieve reasonable expected true risk.

Theorem 37 *Suppose that ℓ satisfies Assumptions 1 and 2 and that P-SGDM is executed for T steps with constant step-size $\alpha = C/T^q$ for $q \in [\frac{1}{2}, 1)$ and momentum μ, satisfying the conditions in Theorem 18 with $\mu = o(\alpha\gamma)$. Then, the risk $\mathbb{E}[R(\hat{\mathbf{w}}_T)] - \mathbb{E}[R_{\mathcal{S}}(\mathbf{w}^*_{\mathcal{S}})]$ goes to zero as T and n increase with the rate:*

$$\mathbb{E}[R(\hat{\mathbf{w}}_T)] - \mathbb{E}[R_{\mathcal{S}}(\mathbf{w}^*_{\mathcal{S}})] = \mathcal{O}\Big(\max\Big\{\frac{1}{T^{q\mathbb{1}\{q\leq 1/2\}+(1-q)\mathbb{1}\{q>1/2\}}}, \frac{1}{n}\Big\}\Big). \quad \text{(Excess Risk)}$$

Proof By our convergence analysis in Theorem 35, we have

$$\epsilon_{\text{opt}} \leq \frac{\mu W_0}{(1-\mu)T} + \frac{(1-\mu)W_1}{2\alpha T} - \frac{\gamma W_2}{2} - \frac{\mu\gamma W_3}{2(1-\mu)} + \frac{\alpha L^2}{2(1-\mu)}.$$

By our stability analysis in Lemma 36, we have

$$\epsilon_s \leq \frac{2\alpha L^2(\beta+\gamma)}{n\big(\alpha\beta\gamma - 3\mu(\beta+\gamma)\big)}.$$

Adding the upper bounds of ϵ_{opt} and ϵ_s above, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{S},A}[R(\hat{\mathbf{w}}_T)] \leq \mathbb{E}_{\mathcal{S},A}[R_{\mathcal{S}}(\mathbf{w}^*_{\mathcal{S}})] + \frac{\mu W_0}{(1-\mu)T} + \frac{(1-\mu)W_1}{2\alpha T} - \frac{\gamma W_2}{2} \\ - \frac{\mu\gamma W_3}{2(1-\mu)} + \frac{\alpha L^2}{2(1-\mu)} + \frac{2\alpha L^2(\beta+\gamma)}{n\big(\alpha\beta\gamma - 3\mu(\beta+\gamma)\big)}.\end{aligned} \tag{J.4}$$

We note that the condition $\mu < \frac{\alpha\beta\gamma}{3(\beta+\gamma)}$ in Theorem 18 implies that $\mu < \alpha\gamma/3$. For sufficiently small $\mu = o(\alpha\gamma)$, the last term in the upper bound becomes independent of α and we have

$$\frac{2\alpha L^2(\beta+\gamma)}{n\big(\alpha\beta\gamma - 3\mu(\beta+\gamma)\big)} = \mathcal{O}(1/n).$$

Then for any $\alpha = C/T^q$ for $q \in [\frac{1}{2}, 1)$, T and n, the upper bound on the risk goes to zero as T and n increase with the rate in Eq. (Excess Risk). ■

Theorem 37 provides a bound on the expected true risk of P-SGDM in terms of the global minimum of the empirical risk.

Appendix K. Generalization error of SGDEM with $\alpha_t = \alpha_0/\sqrt{t}$

We establish an upper bound on the generalization error of SGDEM with the larger step size $\alpha_t = \alpha_0/\sqrt{t}$, which is a common choice in the optimization literature (Bubeck, 2015).

Theorem 38 *Suppose that ℓ satisfies Assumption 1 and that the SGDEM update is executed for T steps with step-size $\alpha_t = \alpha_0/\sqrt{t}$ and some constant $\mu_d \in (0,1]$ in the first t_d steps. Then, for any $1 \leq \tilde{t} \leq t_d \leq T$, SGDEM satisfies ϵ_s-uniform stability with*

$$\epsilon_s \leq \frac{2\alpha_0 L^2\sqrt{\pi}}{n\sqrt{2\mu_d}} \exp(u\sqrt{T})\breve{h}(\mu_d, t_d) + \frac{\tilde{t}M}{n} + \frac{2L^2}{\beta(n-1)} \exp\Big(u\big(\sqrt{T} - \sqrt{\tilde{t}}\big)\Big) \tag{K.1}$$

where $\breve{h}(\mu_d, t_d) = \exp(2\mu_d t_d + u^2/(8\mu_d))\big(\Phi(\sqrt{2\mu_d}(\sqrt{t_d} + \frac{u}{4\mu_d})) - \Phi(\sqrt{2\mu_d}(\sqrt{\tilde{t}} + \frac{u}{4\mu_d}))\big)$, $\Phi(x) = \operatorname{erf}(x) := \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)\,\mathrm{d}t$, $u = (1 - \frac{1}{n})\alpha_0\beta$, *and* $M = \sup_{\mathbf{w},\mathbf{z}} \ell(\mathbf{w};\mathbf{z})$.

Proof Similar to the proof of Theorem 10, we have the following inequality:

$$\tilde{\Delta}_{t+1,\tilde{t}} \leq (1 + 2\mu_t + (1 - 1/n)\alpha_t\beta)\tilde{\Delta}_{t,\tilde{t}} + \frac{2\alpha_t L}{n}.$$

Noting that $\tilde{\Delta}_{t,\tilde{t}} \geq \Delta_{t,\tilde{t}}$ for all $t \geq \tilde{t}$, we have $\mathbb{E}[\Delta_{T,\tilde{t}}] \leq S_3 + S_4$ where

$$S_3 = \sum_{t=\tilde{t}+1}^{t_d} \prod_{p=t+1}^{T} \Big(1 + 2\mu_p + (1 - \frac{1}{n})\frac{\alpha_0\beta}{\sqrt{p}}\Big)\frac{2\alpha_0 L}{n\sqrt{t}}$$

and

$$S_4 = \sum_{t=t_d+1}^{T} \prod_{p=t+1}^{T} \Big(1 + 2\mu_p + (1 - \frac{1}{n})\frac{\alpha_0\beta}{\sqrt{p}}\Big)\frac{2\alpha_0 L}{n\sqrt{t}}.$$

Substituting $\mu_p = \mu_d$ for $p = 1, \ldots, t_d$, we can find an upper bound on S_3 as follows:

$$\begin{aligned}
S_3 &= \sum_{t=\tilde{t}+1}^{t_d} \prod_{p=t+1}^{T} \Big(1 + 2\mu_p + (1 - \frac{1}{n})\frac{\alpha_0\beta}{\sqrt{p}}\Big)\frac{2\alpha_0 L}{n\sqrt{t}} \\
&\leq \sum_{t=\tilde{t}+1}^{t_d} \prod_{p=t+1}^{T} \exp\Big(2\mu_p + (1 - \frac{1}{n})\frac{\alpha_0\beta}{\sqrt{p}}\Big)\frac{2\alpha_0 L}{n\sqrt{t}} \\
&\leq \sum_{t=\tilde{t}+1}^{t_d} \exp\Big(2\mu_d(t_d - t) + (1 - \frac{1}{n})\alpha_0\beta(\sqrt{T} - \sqrt{t})\Big)\frac{2\alpha_0 L}{n\sqrt{t}} \\
&\leq \frac{2\alpha_0 L}{n}\exp(u\sqrt{T} + 2\mu_d t_d)\int_{\tilde{t}}^{t_d} \frac{\exp(-2\mu_d t - u\sqrt{t})}{\sqrt{t}}\,\mathrm{d}t \\
&\leq \frac{2\alpha_0 L}{n}\exp(u\sqrt{T} + 2\mu_d t_d + u^2/(8\mu_d))\int_{\tilde{t}}^{t_d} \frac{\exp\big(-2\mu_d(\sqrt{t} + u/(4\mu_d))^2\big)}{\sqrt{t}}\,\mathrm{d}t \\
&= \frac{2\alpha_0 L\sqrt{\pi}}{n\sqrt{2\mu_d}}\exp(u\sqrt{T} + 2\mu_d t_d + u^2/(8\mu_d))(\Phi(\sqrt{2\mu_d}(\sqrt{t_d} + \frac{u}{4\mu_d})) - \Phi(\sqrt{2\mu_d}(\sqrt{\tilde{t}} + \frac{u}{4\mu_d})))
\end{aligned}$$

where the last line follows (Gradshteyn and Ryzhik, 2014, Eq. 3.321).

We can also find an upper bound on S_4 as follows:

$$\begin{aligned}
S_4 &= \sum_{t=t_d+1}^{T} \prod_{p=t+1}^{T} \Big(1 + (1 - \frac{1}{n})\frac{\alpha_0\beta}{\sqrt{p}}\Big)\frac{2\alpha_0 L}{n\sqrt{t}} \\
&\leq \frac{2L}{\beta(n-1)}\exp\Big(u(\sqrt{T} - \sqrt{t_d})\Big) \qquad \text{(K.2)} \\
&\leq \frac{2L}{\beta(n-1)}\exp\Big(u(\sqrt{T} - \sqrt{\tilde{t}})\Big).
\end{aligned}$$

Replacing $\Delta_{T,\tilde{t}}$ with its upper bound in Eq. (3.3), we obtain Eq. (K.1).

By its definition, we have $\Phi(x) \leq 1$. We also note that $1 - \exp(-x^2) \leq \Phi(x)$ for $x > 0$ following the upper bound developed for $1 -$ erf in (Chiani et al., 2003). Applying both lower bound and upper bound on Φ in Eq. (K.1) and after rearranging the terms, we have

$$\epsilon_s \leq \Big(\frac{2\alpha_0 L^2\sqrt{\pi}}{n\sqrt{2\mu_d}} \exp\left(2\mu_d(t_d - \tilde{t})\right) + \frac{2L^2}{\beta(n-1)}\Big) \exp\Big(u(\sqrt{T} - \sqrt{\tilde{t}})\Big) + \frac{\tilde{t}M}{n}. \quad \text{(K.3)}$$

■

Corollary 39 *Suppose, in Theorem 38, we set $t_d = \tilde{t}^* + K$for some constant K where $\tilde{t}^*$ satisfies:*

$$M \exp(u\sqrt{\tilde{t}^*})\sqrt{\tilde{t}^*} = \Big(\frac{u\alpha_0 L^2\sqrt{\pi}}{\sqrt{2\mu_d}} + L^2\alpha_0\Big) \exp(u\sqrt{T}). \quad \text{(K.4)}$$

Then the generalization error of SGDEM for T steps with $\alpha_t = \alpha_0/\sqrt{t}$ is upper bounded by $\mathcal{O}\Big(\frac{\exp(u\sqrt{T}/(u+1)+\mu_d)}{n}\Big)$.

Proof Note that we can minimize:

$$\min_{1\leq\tilde{t}\leq t_d} \frac{\tilde{t}M}{n} + \Big(\frac{2\alpha_0 L^2\sqrt{\pi}}{n\sqrt{2\mu_d}} \exp\left(2\mu_d K\right) + \frac{2L^2}{\beta(n-1)}\Big) \exp\Big(u(\sqrt{T} - \sqrt{\tilde{t}})\Big)$$

by optimizing $\tilde{t}$ after setting $t_d = \tilde{t} + K$ where the objective is the upper bound in Eq. (K.1). We note that an optimal $\tilde{t}^*$ satisfies Eq. (K.4), which does not have an analytic solution but can be solve numerically. Instead, we consider a suboptimal solution by taking ln from both sides of Eq. (K.4) and applying the well-known inequality $\ln(x+1) \leq x$, $\forall x \geq -1$, which leads to:

$$\sqrt{\tilde{t}} = \frac{\ln\Big((\frac{u\alpha_0 L^2\sqrt{\pi}}{\sqrt{2\mu_d}} + L^2\alpha_0)/M\Big)}{u+1} + \frac{u\sqrt{T}}{u+1}. \quad \text{(K.5)}$$

Substituting Eq. (K.5) into Eq. (K.1) completes the proof. ■

Appendix L. Additional experiments

In Fig. 5, we plot the test accuracy versus t_d of SGDEM and SGDM (which is a special case of SGDEM with $t_d = T$) for the notMNIST dataset for different μ_d values. We observe dramatic decrease in the test accuracy for $\mu_d = 0.99$, which is consistent with our convergence analysis in Theorem 25.

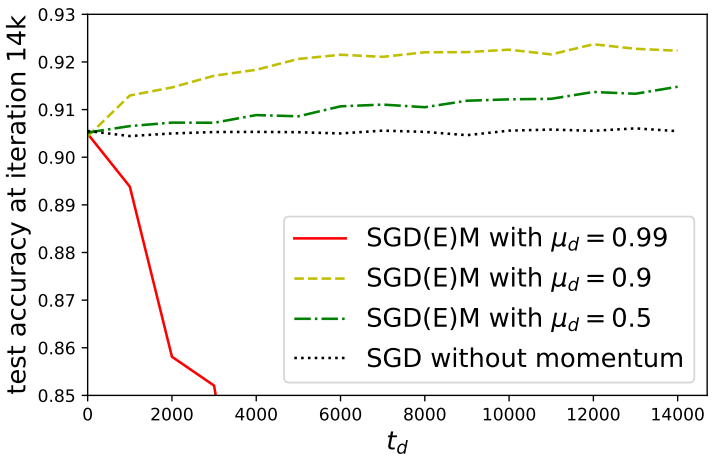

Figure 5: *Test accuracy of a feedforward fully connected neural network for notMNIST dataset.*

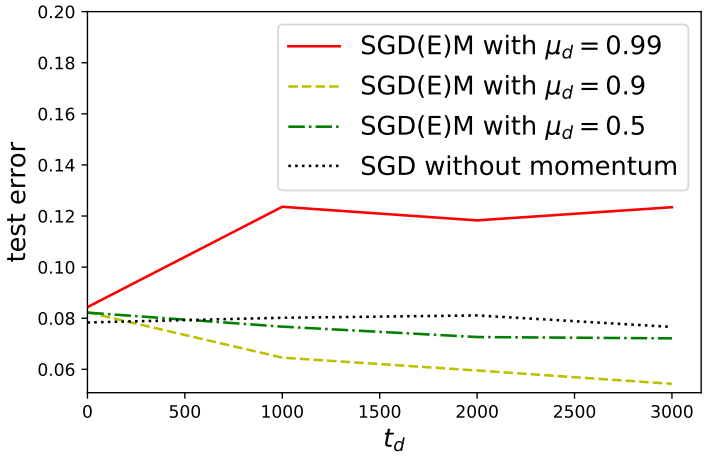

Figure 6: *Test error of logistic regression for notMNIST dataset.*

We now study the performance of SGDEM for a smooth and strongly convex loss function. We train a logistic regression model with the weight decay regularization using SGDEM for binary classification on the two-class notMNIST and MNIST datasets that contain the images from letter classes "C" and "J", and digit classes "2" and "9", respectively. We set the step-size $\alpha = 0.01$. The weight decay coefficient and the minibatch size are set to 0.001 and 10, respectively. We use 100 SGDEM realizations to evaluate the average performance.

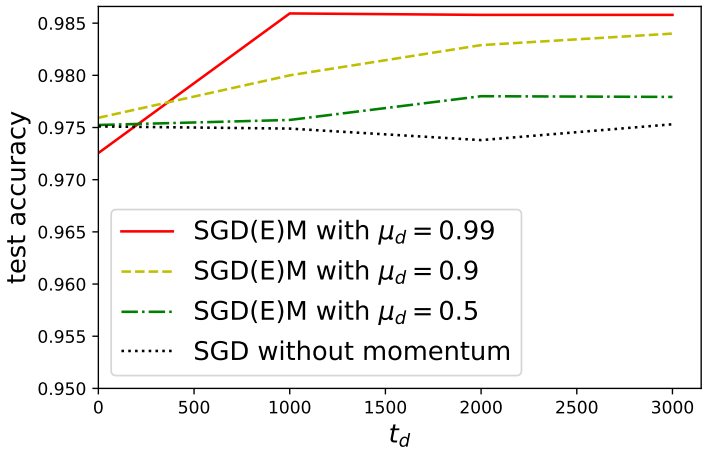

Figure 7: *Test accuracy of logistic regression for notMNIST dataset.*

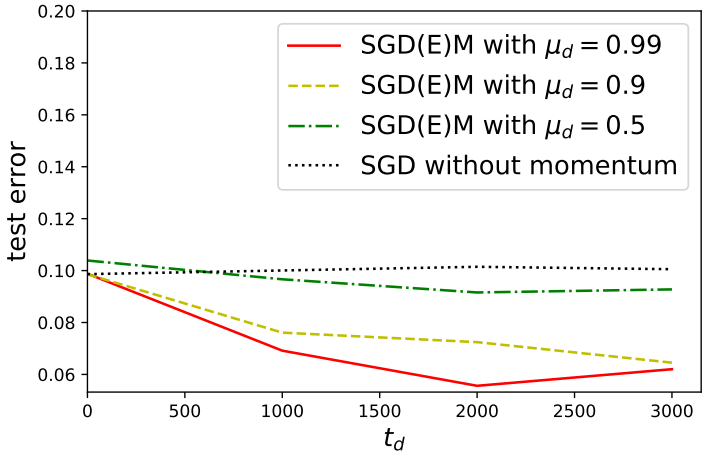

Figure 8: *Test error of logistic regression for MNIST dataset.*

We plot the test error and test accuracy versus t_d under SGDEM for the notMNIST dataset in Figs. 6 and 7, respectively. We show the same performance measures for the MNIST dataset in Figs. 8 and 9 respectively. We observe that, unlike the case of nonconvex loss functions, it does not hurt to add momentum for the entire training. In the following, we focus on SGDM with the classical momentum update rule for a smooth and strongly convex loss function for the notMNIST dataset.

In the following, we focus on SGDM with the classical momentum update rule for a smooth and strongly convex loss function on notMINIST.

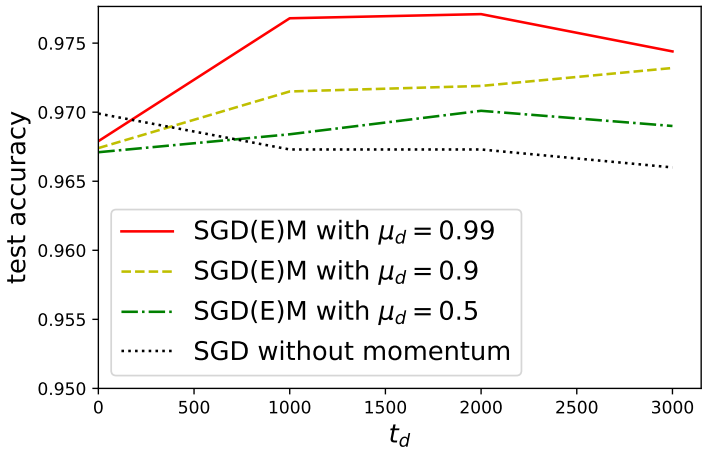

Figure 9: *Test accuracy of logistic regression for MNIST dataset.*

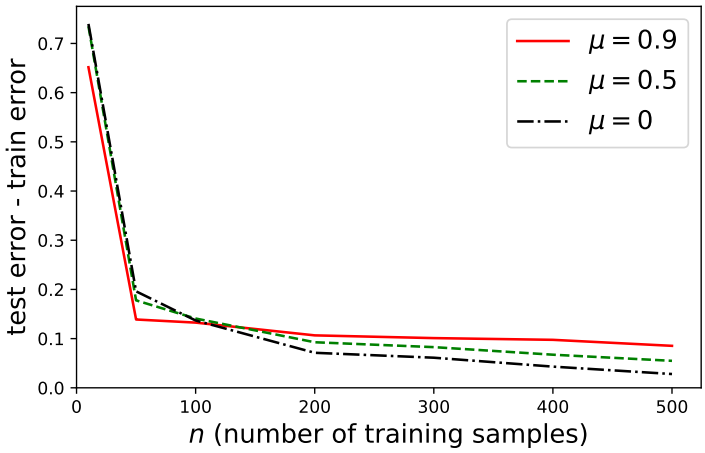

Figure 10: *Generalization error (cross entropy) of logistic regression for notMNIST dataset with $T = 1000$ iterations.*

We compare the training and generalization performance of SGD without momentum with that of SGDM under $\mu = 0.5$ and $\mu = 0.9$, which are common momentum values used in practice (Goodfellow et al., 2016, Section 8.3.2).

We show in Fig. 10 the generalization error (w.r.t. cross entropy) versus the number of training samples, n, under SGDM with fixed $T = 1000$ iterations for $\mu = 0, 0.5, 0.9$. In Fig. 11, we plot the training accuracy as a function of the number of training samples for the same dataset. First, we observe that the generalization error decreases as n increases for all values of μ, which is also suggested by our

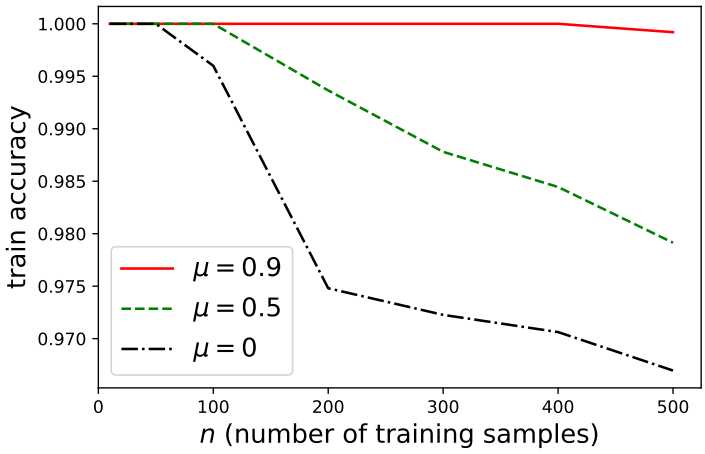

Figure 11: *Training accuracy of logistic regression for notMNIST dataset with $T = 1000$ iterations.*

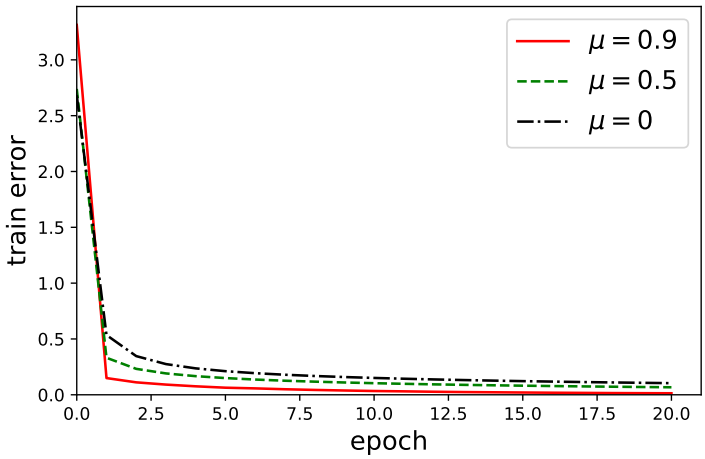

Figure 12: *Training error (cross entropy) of logistic regression for notMNIST dataset with $n = 500$.*

stability upper bound in Theorem 18. In addition, for sufficiently large n, we observe that the generalization error increases with μ, consistent with Theorem 18. The training accuracy also improves by adding momentum as illustrated in Fig. 11.

In order to study the optimization error of SGDM, we show in Figs. 12 and 13, the training error and test error, respectively, versus the number of epochs, under SGDM trained with $n = 500$ samples. We plot the classification accuracy for training dataset in Fig. 14. We observe that the training error decreases as the number of epochs increases for all values of μ, which is consistent with the convergence analysis

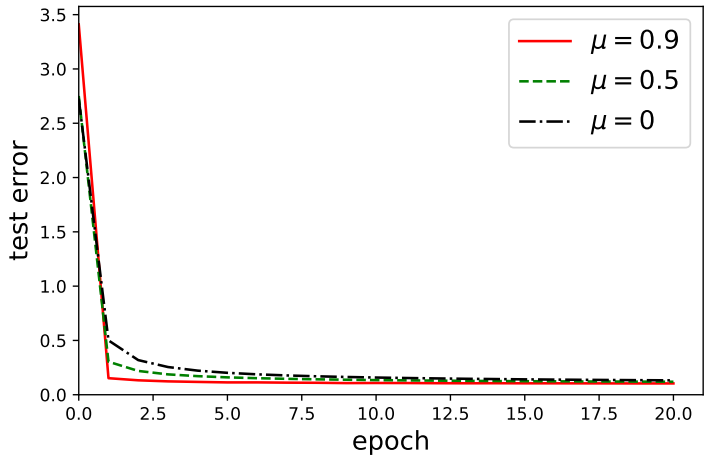

Figure 13: *Test error (cross entropy) of logistic regression for notMNIST dataset with* $n = 500$.

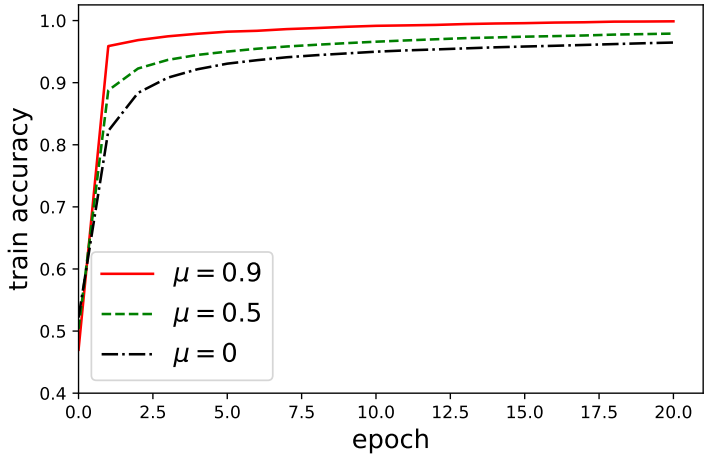

Figure 14: *Training accuracy of logistic regression for notMNIST dataset with* $n = 500$.

in Theorem 35. Furthermore, as expected, we see that adding momentum improves the training error and accuracy. However, as the number of epochs increases, we note that the benefit of momentum on the test error becomes negligible. This happens because adding momentum also results in a higher generalization error thus offsetting the gain in training error.

References

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York, USA: Dover, 1972.

Mahmoud Assran and Michael Rabbat. On the convergence of Nesterov's accelerated gradient method in stochastic settings. In *International Conference on Machine Learning (ICML)*, 2020.

Amit Attia and Tomer Koren. The instability of accelerated gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Localized rademacher complexities. In *Conference on Learning Theory (COLT)*, 2002.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research (JMLR)*, 2:499–526, 2002.

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory (COLT)*, 2020.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Bugra Can, Mert Gürbüzbalaban, and Lingjiong Zhu. Accelerated linear convergence of stochastic momentum methods in Wasserstein distances. In *International Conference on Machine Learning (ICML)*, 2019.

Yuansi Chen, Chi Jin, and Bin Yu. Stability and convergence trade-off of iterative optimization algorithms. arXiv preprint arXiv:1804.01619.

Marco Chiani, Davide Dardari, and Marvin K. Simon. New exponential bounds and approximations for the computation of error probability in fading channels. *IEEE Transactions on Wireless Communications*, 2(4):840–845, 2003.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations (ICLR)*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2000.

Vitaly Feldman and Jan Vondrak. Generalization bounds for uniformly stable algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Vitaly Feldman and Jan Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory (COLT)*, 2019.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *European Control Conference (ECC)*, 2015.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, USA: The MIT Press, 2016.

Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of Integrals, Series, and Products*. Academic press, 2014.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML)*, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

Sham M. Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2008.

Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from Adam to SGD. arXiv preprint arXiv:1712.07628.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Yegor Klochkov and Nikita Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

Maxime Laborde and Adam Oberman. A Lyapunov analysis for accelerated gradient methods: from deterministic to stochastic case. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

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

Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

Deanna Needell, Nati Srebro, and Rachel Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Arkadij Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Yurii Nesterov. A method of solving a convex programming problem with convergence $O(1/k^2)$. *Doklady Akademii Nauk*, 269(3):543–547, 1983.

Peter Ochs, Yunjin Chen, Thomas Brox, and Thomas Pock. iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2): 1388–1419, 2014.

Peter Ochs, Thomas Brox, and Thomas Pock. iPiasco: Inertial proximal algorithm for strongly convex optimization. *Journal of Mathematical Imaging and Vision*, 53(2):171–181, 2015.

Ming Yang Ong. Understanding generalization. 2017. Thesis available at `https://dspace.mit.edu/bitstream/handle/1721.1/113119/1016455698-MIT.pdf?sequence=1`.

Antonio Orvieto, Jonas Kohler, and Aurelien Lucchi. The role of memory in stochastic optimization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

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

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by backpropagating errors. *Nature*, 323(6088):533–536, 1986.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research (JMLR)*, 11:2635–2670, 2010.

Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2008.

Weijie Su, Stephen Boyd, and Emmanuel Candès. A differential equation for modeling Nesterov's acceleratedgradient method: Theory and insights. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning (ICML)*, 2013.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142, 1984.

Vladimir N. Vapnik and Alexey Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ashia C. Wilson, Ben Recht, and Michael I. Jordan. A Lyapunov analysis of momentum methods in optimization. *Journal of Machine Learning Research (JMLR)*, 22:34–1, 2021.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853v2*.

Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.