# Does Momentum Change the Implicit Regularization on Separable Data?

**Bohan Wang**[*]
University of Science & Technology of China
bhwangfy@gmail.com

**Qi Meng**[†]
Microsoft Research Asia
meq@microsoft.com

**Huishuai Zhang**
Microsoft Research Asia
huzhang@microsoft.com

**Ruoyu Sun**
The Chinese University of Hong Kong, Shenzhen, China
sunruoyu@cuhk.edu.cn

**Wei Chen**[†]
Chinese Academy of Sciences
chenwei2022@ict.ac.cn

**Zhi-Ming Ma**
Chinese Academy of Sciences
mazm@amt.ac.cn

**Tie-Yan Liu**
Microsoft Research Asia
tyliu@microsoft.com

## Abstract

The momentum acceleration technique is widely adopted in many optimization algorithms. However, there is no theoretical answer on how the momentum affects the generalization performance of the optimization algorithms. This paper studies this problem by analyzing the implicit regularization of momentum-based optimization. We prove that on the linear classification problem with separable data and exponential-tailed loss, gradient descent with momentum (GDM) converges to the $L^2$ max-margin solution, which is the same as vanilla gradient descent. That means gradient descent with momentum acceleration still converges to a low-complexity model, which guarantees their generalization. We then analyze the stochastic and adaptive variants of GDM (i.e., SGDM and deterministic Adam) and show they also converge to the $L^2$ max-margin solution. Technically, the implicit regularization of SGDM is established based on a novel convergence analysis of SGDM under a general noise condition called affine noise variance condition. To the best of our knowledge, we are the first to derive SGDM's convergence under such an assumption. Numerical experiments are conducted to support our theoretical results.

## 1 Introduction

It is widely believed that the optimizers have the implicit regularization in terms of selecting output parameters among all the minima on the landscape [26, 17, 45]. Parallel to the analysis of *coordinate descent* ([31, 40]), [34] shows that *gradient descent* would converge to the $L^2$ max-margin solution for the linear classification task with exponential-tailed loss, which mirrors its good generalization property in practice. Since then, many efforts have been taken on analyzing the implicit regularization of various local-search optimizers, including stochastic gradient descent [24], steepest descent [8], AdaGrad [29] and optimizers for homogeneous neural networks [21, 14, 43].

However, though the momentum acceleration technique is widely adopted in the optimization algorithms in both convex and non-convex learning tasks [37, 41, 38], the understanding of how

[*] The work was done when Bohan Wang was an intern at Microsoft Research Asia.

[†] Corresponding Authors

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Table 1: The algorithms investigated in this paper (GDM and Adam) along with algorithms (GD) already investigated in the existing literature. We also compare the learning rates required to obtain the characterization of implicit regularization. As for stochastic Adam with $\beta_1 \neq 0$, we leave its implicit regularization as future work.

| Method | With Random Sampling | Learning Rate | Corresponding Literature |
|---|---|---|---|
| GD | × | Constant | [34] |
| | ✓ | Constant | [24] |
| GDM | × | Constant | This Work |
| | ✓ | Constant | This Work |
| Adam | × | Constant | This Work |
| | ✓ | Decaying $\left(\frac{1}{\sqrt{t}}\right)$ | $\beta_1 = 0$ in this Work |

the momentum would affect the generalization performance of the optimization algorithms is still unclear, as the historical gradients in the momentum may significantly change the searching direction of the optimization dynamics. A natural question is:

*Can we theoretically analyze the implicit regularization of momentum-based optimizers?*

In this paper, we take the first step to analyze the convergence of momentum-based optimizers and unveil their implicit regularization. Our study starts from the classification problem with the linear model and exponential-tailed loss using Gradient Descent with Momentum (GDM) optimizer. Then the variants of GDM such as Stochastic Gradient Descent with Momentum (SGDM) and deterministic Adam are also analyzed. We consider the optimizers with constant learning rate and constant momentum hyper-parameters, which are widely adopted in practice, e.g., the default setting in popular machine learning frameworks [27] and in experiments [46]. Our main results are summarized in Theorem 1.

**Theorem 1** (informal)**.** *With linearly separable dataset $\boldsymbol{S}$, linear model and exponential-tailed loss:*

- *For GDM with a constant learning rate, the parameter norm diverges to infinity, with its direction converging to the $L^2$ max-margin solution. The same conclusion holds for SGDM with a constant learning rate.*
- *For deterministic Adam with a constant learning rate and stochastic RMSProp (i.e., Adam without momentum) with a decaying learning rate, the same conclusion holds.*

Theorem 1 states that GDM and its variants converge to the $L^2$ max-margin solution, which is the same as their without-momentum versions, indicating that momentum does not affect the convergent direction. Therefore, the good generalization behavior of the output parameters of these optimizers is well validated as the margin of a classifier is positively correlated with its generalization error [16] and is supported by existing experimental observations (e.g., [34, 23, 43]).

Our contributions are significant in terms of the following aspects:

- We establish the implicit regularization of the momentum-based optimizers, an open problem since the initial work [34]. The momentum-based optimizers are widely used in practice, and our theoretical characterization deepens the understanding of their generalization property, which is important on its own.
- Technically, we design a two-stage framework to analyze the momentum-based optimizers, which generalizes the proof techniques in [34] and [24]. The first stage shows the convergence of the loss. Specifically, we derive the first convergence analysis of SGDM under affine noise variance condition [2], generalizing the results under bounded noise variance condition in literature [47, 20]. The second stage shows the convergence of the parameters. We propose an easy-to-check condition of whether the difference between learned parameters and the scaled max-margin solution is bounded. This condition can be generalized to implicit regularization analyses of other momentum-based optimizers.

- We further verify our theory through numerical experiments.

The technical difficulty to analyzing the implicit regularization of the momentum-based optimizers mainly come from the historical information of training trajectory contained by the momentum term. Specifically, the foundation of the analysis is to derive the convergence of the loss. It is not trivial because, for momentum-based optimizers, the loss may not monotonously decrease due to the mismatch between the direction of the gradient and the momentum. Furthermore, even if the loss is proved to converge, the converging direction of the parameters is still hard to characterize as it's difficult to track the training trajectory due to the momentum term.

**Organization of This Paper.** Section 2 collects further related works on the implicit regularization of the first-order optimizers and the convergence of momentum-based optimizers. Section 3 shows basic settings and assumptions which will be used throughout this paper. Section 4 studies the implicit regularization of GDM, while Section 5 and Section 6 explore respectively the implicit regularization of SGDM and Adam. Discussions of these results are put in Section 7. Detailed proofs and experiments can be found in the appendix.

## 2 Further related works

**Implicit Regularization of First-order Optimization Methods.** [34] prove that gradient descent on linear classification problem with exponential-tailed loss converges to the direction of the max $L^2$ margin solution of the corresponding hard-margin Support Vector Machine. [24] extend the results in [34] to the stochastic case, proving that the convergent direction of SGD is the same as GD almost surely. [29] go beyond the vanilla gradient descent methods and consider the AdaGrad optimizer instead. They prove that the convergent direction of AdaGrad has a dependency on the optimizing trajectory, which varies according to the initialization. [15] propose a primal-dual analysis framework for the linear classification models and prove a faster convergent rate of the margin by increasing the learning rate according to the loss. Based on [15], [12] design another algorithm with an even faster convergent rate of margin by applying the Nesterov's Acceleration Method on the dual space. However, the corresponding form of the algorithm on the primal space is no longer a Nesterov's Acceleration Method nor GDM, which is significantly different from our settings.

On the other hand, another line of work is trying to extend the linear case result to deep neural networks. [13, 9] study the deep linear network and [34] study the two-layer neural network with ReLU activation. [21] propose a framework to analyze the asymptotic direction of GD on homogeneous neural networks, proving that given there exists a time the network achieves $100\%$ training accuracy, GD will converge to some KKT point of the $L^2$ max-margin problem. [43] extend the framework of [21] to adaptive optimizers and prove RMSProp and Adam without momentum have the same convergent direction as GD, while AdaGrad does not. The results [21, 43] indicate that results in the linear model can be extended to deep homogeneous neural networks and suggest that the linear model is an appropriate starting point to study the implicit bias.

Except for the exponential-tailed loss, there are also works on the implicit bias with squared loss. Interested readers can refer to [30, 19, 1] etc. for details.

**Convergence of Momentum-Based Optimization Methods.** For convex optimization problems, the convergence rate of Nesterov's Acceleration Method [25] has been proved in various approaches (e.g., [25, 35, 44]). In contrast, although GDM (Polyak's Heavy-Ball Method) was proposed in [28] before the Nesterov's Acceleration Method, the convergence of GDM on convex loss with Lipschitz gradient was not solved until [7] provides an ergodic convergent result for GDM, i.e., the convergent result for the running average of the iterates. [36] provide a non-ergodic analysis when the training loss is coercive (the training loss goes to infinity whenever parameter norm goes to infinity), convex, and globally smooth. However, all existing results cannot be directly applied to exponential-tailed loss, which is non-coercive.

There are also works on the convergence of SGDM under various settings. [47] prove SGDM converges to a bounded region assuming both bounded gradient norm and bounded gradient variance. The bounded gradient norm assumption is further removed by [48, 20]. Nevertheless, a converging-to-stationary-point analysis is required in the implicit regularization analysis. Thus their results can not be directly applied. [39] analyze a particular case when the momentum parameter increases over iterations, which, however, does not agree with the practice where the momentum parameter is fixed.

As for (stochastic) Adam, its convergence analysis is still an open problem, and the current analyses are restricted to specific settings (e.g., bounded gradient, dynamical momentum hyperparameters). We recommend interested readers to refer to [5, 4, 33, 10] for details.

## 3 Preliminaries

This paper focuses on the linear model with the exponential-tailed loss. We mainly investigate binary classification. However, the methodology can be easily extended to the multi-class classification problem (please refer to Appendix F.3 for details).

**Problem setting.** The dataset used for training is denoted as $\boldsymbol{S}=(\boldsymbol{x}_i,\boldsymbol{y}_i)_{i=1}^N$, where $\boldsymbol{x}_i\in\mathbb{R}^d$ is the $i$-th input feature, and $\boldsymbol{y}_i\in\mathbb{R}$ is the $i$-th label ($i=1,2,\cdots,N$). We will use the linear model to fit the label: for any feature $\boldsymbol{x}\in\mathbb{R}^d$ and parameter $\boldsymbol{w}\in\mathbb{R}^d$, the prediction is given by $\langle\boldsymbol{w},\boldsymbol{x}\rangle$.

For binary classification, given any data $\boldsymbol{z}_i=(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\boldsymbol{S}$, the individual loss for parameter $\boldsymbol{w}$ is given as $\ell(\boldsymbol{y}_i\langle\boldsymbol{w},\boldsymbol{x}_i\rangle)$. As only $\boldsymbol{y}_i\boldsymbol{x}_i$ is used in the loss, we then ensemble the feature and label together and assume $\boldsymbol{y}_i=1$ ($\forall i\in\{1,\cdots,N\}$) without the loss of generality. We then drop $\boldsymbol{y}_i$ for brevity and redefine $\boldsymbol{S}=(\boldsymbol{x}_i)_{i=1}^N$. The spectral norm of the data matrix $(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_N)$ is defined as $\sigma_{max}$. We use $\tilde{\ell}(\boldsymbol{w},\boldsymbol{x})\triangleq\ell(\langle\boldsymbol{w},\boldsymbol{x}\rangle)$ for brevity.

The optimization target is defined as the averaged loss: $\mathcal{L}(\boldsymbol{w})=\frac{\sum_{i=1}^N\tilde{\ell}(\boldsymbol{w},\boldsymbol{x}_i)}{N}$.

**Optimizer.** Here we will introduce the update rules of GDM, SGDM and deterministic Adam. GDM's update rule is

$$\boldsymbol{m}(0)=\boldsymbol{0},\boldsymbol{m}(t)=\beta\boldsymbol{m}(t-1)+(1-\beta)\nabla\mathcal{L}(\boldsymbol{w}(t)),\boldsymbol{w}(t+1)=\boldsymbol{w}(t)-\eta\boldsymbol{m}(t). \tag{1}$$

SGDM can be viewed as a stochastic version of GDM by randomly choosing a subset of the dataset to update. Specifically, SGDM changes the update of $\boldsymbol{m}(t)$ into

$$\boldsymbol{m}(t)=\beta\boldsymbol{m}(t-1)+(1-\beta)\nabla\mathcal{L}_{\xi(t)}(\boldsymbol{w}(t)), \tag{2}$$

where $\xi(t)$ is the gradient noise of the $t$-th step, independent of the randomness of the previous steps and satisfying $\mathbb{E}_{\xi(t)}\nabla\mathcal{L}_{\xi(t)}(\boldsymbol{w})=\nabla\mathcal{L}(\boldsymbol{w})$, $\forall\boldsymbol{w}$. In this paper, $\xi(t)$ comes from the random sampling of a subset $\boldsymbol{B}(t)$ of $\boldsymbol{S}$ with size $b$ (either with replacement (**abbreviated as "w/. r"**) or without replacement (i.e., with random shuffling, **abbreviated as "w/o. r"**)), and $\mathcal{L}_{\xi(t)}$ is defined as $\mathcal{L}_{\xi(t)}(\boldsymbol{w})=\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w})=\frac{\sum_{\boldsymbol{x}\in\boldsymbol{B}(t)}\tilde{\ell}(\boldsymbol{w},\boldsymbol{x})}{b}$. We also define $\mathcal{F}_t$ as the sub-sigma algebra over the gradient noise collecting randomness before the $t$-th step. Therefore, $\boldsymbol{w}(t)$ is adapted with respect to the sigma algebra flow $\mathcal{F}_t$.

The Adam optimizer can be viewed as a variant of SGDM in which the preconditioner is adopted, whose form is characterized as follows:

$$\begin{gathered}\boldsymbol{m}(0)=\boldsymbol{0},\boldsymbol{m}(t)=\beta_1\boldsymbol{m}(t-1)+(1-\beta_1)\nabla\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)),\\ \boldsymbol{\nu}(0)=0,\boldsymbol{\nu}(t)=\beta_2\boldsymbol{\nu}(t-1)+(1-\beta_2)\nabla\mathcal{L}(\boldsymbol{w}(t))\odot\nabla\mathcal{L}(\boldsymbol{w}(t))\\ \hat{\boldsymbol{m}}(t)=\frac{1}{1-\beta_1^t}\boldsymbol{m}(t),\hat{\boldsymbol{\nu}}(t)=\frac{1}{1-\beta_2^t}\boldsymbol{\nu}(t),\\ \textit{(Update Rule)}:\ \boldsymbol{w}(t)=\boldsymbol{w}(t-1)-\eta\frac{1}{\sqrt{\hat{\boldsymbol{\nu}}(t-1)+\varepsilon\mathbb{1}_d}}\odot\hat{\boldsymbol{m}}(t-1),\end{gathered} \tag{3}$$

where $\frac{1}{\sqrt{\hat{\boldsymbol{\nu}}(t-1)+\varepsilon\mathbb{1}_d}}$ is called the preconditioner.

**Assumptions:** The analysis of this paper are based on three common assumptions in existing literature (first proposed by [34]). They are respectively on the separability of the dataset, the individual loss behavior at the tail, and the smoothness of the individual loss. We list them as follows:

**Assumption 1** (Linearly Separable Dataset)**.** *There exists one parameter* $\boldsymbol{w}\in\mathbb{R}^d$*, such that*

$$\langle\boldsymbol{w},\boldsymbol{x}_i\rangle>0,\ \forall i\in[N].$$

**Assumption 2** (Exponential-tailed Loss)**.** *The individual loss* $\ell$ *is exponential-tailed, i.e.,*

- *Differentiable and monotonically decreasing to zero, with its derivative converging to zero at positive infinity and to non-zero at negative infinity, i.e.,* $\lim_{x\to\infty}\ell(x)=\lim_{x\to\infty}\ell'(x)=0$, $\overline{\lim}_{x\to-\infty}\ell'(x)<0$*, and* $\ell'(x)<0$, $\forall x\in\mathbb{R}$*;*

- *Close to exponential loss when $x$ is large enough, i.e., there exist positive constants $c, a, \mu_+, \mu_-, x_+$, and $x_-$, such that,*

$$\forall x > x_+ : -\ell'(x) \leq c(1 + e^{-\mu_+ x})e^{-ax}, \tag{4}$$

$$\forall x > x_- : -\ell'(x) \geq c(1 - e^{-\mu_- x})e^{-ax}. \tag{5}$$

**Assumption 3** (Smooth Loss)**.** *Either of the following assumptions holds regarding the case:*

***(D):*** *(Without Stochasticity) The individual loss $\ell$ is locally smooth, i.e., for any $s_0 \in \mathbb{R}$, there exists a positive real $H_{s_0}$, such that $\forall x, y \geq s_0$, $|\ell'(x) - \ell'(y)| \leq H_{s_0}|x - y|$.*

***(S):*** *(With Stochasticity) The individual loss $\ell$ is globally smooth, i.e., there exists a positive real $H$, such that $\forall x, y \in \mathbb{R}$, $|\ell'(x) - \ell'(y)| \leq H|x - y|$.*

We provide explanations of these three assumptions, respectively. Based on Assumption 1, we can formally define the margin and the maximum margin solution of an optimization problem:

**Definition 1.** *Let the margin $\hat{\gamma}(\boldsymbol{w})$ of parameter $\boldsymbol{w}$ defined as the lowest score of the prediction of $\boldsymbol{w}$ over the dataset $\boldsymbol{S}$, i.e., $\hat{\gamma}(\boldsymbol{w}) = \min_{\boldsymbol{x} \in \boldsymbol{S}} \langle \boldsymbol{w}, \boldsymbol{x} \rangle$. We then define the maximum margin solution $\hat{\boldsymbol{w}}$ and the $L^2$ max margin $\gamma$ of the dataset S as follows:*

$$\hat{\boldsymbol{w}} \triangleq \arg \min_{\hat{\gamma}(\boldsymbol{w}) \geq 1} \|\boldsymbol{w}\|^2, \ \gamma \triangleq \frac{1}{\|\hat{\boldsymbol{w}}\|}$$

Since $\|\cdot\|^2$ is strongly convex and set $\{\boldsymbol{w} : \hat{\gamma}(\boldsymbol{w}) \geq 1\}$ is convex, $\hat{\boldsymbol{w}}$ is uniquely defined.

Assumption 2 constraints the loss to be exponential-tailed, which is satisfied by many popular choices of $\ell$, including the exponential loss ($\ell_{exp}(x) = e^{-x}$) and the logistic loss ($\ell_{log}(x) = \log(1 + e^{-x})$). Also, as $c$ and $a$ can be respectively absorbed by resetting the learning rate and data as $\eta = c\eta$ and $\boldsymbol{x}_i = a\boldsymbol{x}_i$, without loss of generality, **in this paper we only analyze the case that** $c = a = 1$.

The globally smooth assumption (Assumption 3. (S)) is strictly stronger than the locally smooth assumption (Assumption 3. (D)). One can easily verify that both the exponential loss and the logistic loss meet Assumption 3. (**D**), and the logistic loss also meets Assumption 3. (**S**).

## 4 The implicit regularization of GDM

In this section, we analyze the implicit regularization of GDM with a two-stage framework [3]. Later, we will use this framework to investigate SGDM further and deterministic Adam. The formal theorem of the implicit regularization of GDM is as follows:

**Theorem 2.** *Let Assumptions 1, 2, and 3. (**D**) hold. Let $\beta \in [0, 1)$ and $\eta < 2\frac{N}{\sigma_{max}^2 H_{\ell^{-1}(N\mathcal{L}(\boldsymbol{w}_1))}}$ ($\ell^{-1}$ is the inverse function of $\ell$). Then, for almost every data set S, with arbitrary initialization point $\boldsymbol{w}(1)$, GDM (Eq. equation 1) satisfies that $\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}\| = \mathcal{O}(1/\log(t))$ and $\mathcal{L}(\boldsymbol{w}(t)) = \mathcal{O}(\frac{1}{t})$.*

Theorem 2 shows that the implicit regularization of GDM agrees with GD in linear classification with exponential-tailed loss (c.f. [34] for results on GD). This consistency can be verified by existing and our experiments (c.f. Section 7 for detailed discussions).

**Remark 1** (On the hyperparameter setting)**.** *Firstly, the learning rate upper bound $2\frac{N}{\sigma_{max}^2 H_{\ell^{-1}(N\mathcal{L}(\boldsymbol{w}_1))}}$ agrees with that of GD exactly [34], indicating our analysis is tight. Secondly, Theorem 2 adopts a constant momentum hyper-parameter, which agrees with the practical use (e.g., $\beta$ is fixed to be $0.9$ [46]). Also, Theorem 2 puts no restriction on the range of $\beta$, which allows wider choices of hyper-parameter tuning.*

We then present a proof sketch of Theorem 2, which is divided into two parts: we first prove that the sum of squared gradients is bounded, which indicates both the loss and the norm of gradient converge to 0 and the parameter diverges to infinity; these properties will then be applied to show the difference between $\boldsymbol{w}(t)$ and $\ln(t)\hat{\boldsymbol{w}}$ is bounded, and therefore, the direction of $\hat{\boldsymbol{w}}$ dominates as $t \to \infty$.

[3] It should be noticed that the proof sketches in Sections 4, 5, and 6 only hold for almost every dataset (means except a zero-measure set in $\mathbb{R}^{d \times N}$), as we want the presentation more simple and straightforward. However, the proof can be extended to every dataset with a more careful analysis (please refer to Appendix F.1 for details).

**Stage I: Loss Dynamics.** The goal of this stage is to characterize the dynamics of the loss and prove the convergence of GDM. The core of this stage is to select a proper potential function $\xi(t)$, which is required to correlate with the training loss $\mathcal{L}$ and be non-increasing along the optimization trajectory. For GD, since $\mathcal{L}$ is non-increasing with a properly chosen learning rate, we can pick $\xi(t) = \mathcal{L}(t)$. However, as the update of GDM does not align with the direction of the negative gradient, training loss $\mathcal{L}(t)$ in GDM is no longer monotonously decreasing, and the potential function requires special construction. Inspired by [36], we choose the following $\xi(t)$:

**Lemma 1.** *Let all conditions in Theorem 2 hold. Define* $\xi(t) \overset{\triangle}{=} \mathcal{L}(\boldsymbol{w}(t)) + \frac{1-\beta}{2\beta\eta}\|\boldsymbol{w}(t) - \boldsymbol{w}(t-1)\|^2$.[4] *Define* $C_1$ *as a positive real with* $C_1 \triangleq \frac{\sigma_{max}^2 H_{\ell-1}(N\mathcal{L}(\boldsymbol{w}_1))}{2N}\eta$*. We then have*

$$\xi(t) \geq \xi(t+1) + \frac{1-C_1}{\eta}\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|^2. \tag{6}$$

**Remark 2.** *Although this potential function is obtained by [36] by directly examining Taylor's expansion at* $\boldsymbol{w}(t)$*, the proof here is non-trivial as we only require the loss to be locally smooth instead of globally smooth in [36]. We need to prove that the smoothness parameter along the trajectory is upper bounded. We defer the detailed proof to Appendix C.1.1.*

By Lemma 1, we have that $\xi(t)$ is monotonously decreasing by gap $\frac{1-C_1}{\eta}\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|^2$. As $\xi(1) = \mathcal{L}(\boldsymbol{w}(1))$ is a finite number, we have $\sum_{t=1}^{\infty}\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|^2 < \infty$. By that $(1-\beta)\eta\nabla\mathcal{L}(\boldsymbol{w}(t)) = (\boldsymbol{w}(t+1) - \boldsymbol{w}(t)) - \beta(\boldsymbol{w}(t) - \boldsymbol{w}(t-1))$, it immediately follows that $\sum_{t=1}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2 < \infty$.

**Stage II. Parameter Dynamics.** The goal of this stage is to characterize the dynamics of the parameter and show that GDM asymptotically converges (in direction) to the max-margin solution $\hat{\boldsymbol{w}}$. To see this, we define a residual term $\boldsymbol{r}(t) \overset{\triangle}{=} \boldsymbol{w}(t) - \ln(t)\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}}$ with some constant vector $\tilde{\boldsymbol{w}}$ (specified in Appendix C.1.2). If we can show the norm of $\boldsymbol{r}(t)$ is bounded over the iterations, we complete the proof as $\ln(t)\hat{\boldsymbol{w}}$ will then dominates the dynamics of $\boldsymbol{w}(t)$.

For simplicity, we use the continuous dynamics approximation of GDM [36] to demonstrate why $\boldsymbol{r}(t)$ is bounded:

$$\frac{\beta}{1-\beta}\frac{\mathrm{d}^2\boldsymbol{w}(t)}{\mathrm{d}t^2} + \frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t} + \nabla\mathcal{L}(\boldsymbol{w}(t)) = 0. \tag{7}$$

We start by directly examining the evolution of $\|\boldsymbol{r}(t)\|$, i.e.,

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{r}(T)\|^2 - \frac{1}{2}\|\boldsymbol{r}(1)\|^2 &= \int_1^T \frac{1}{2}\frac{\mathrm{d}\|\boldsymbol{r}(s)\|^2}{\mathrm{d}s}\mathrm{d}s \\ =\int_1^T \left\langle \boldsymbol{r}(s), -\nabla\mathcal{L}(\boldsymbol{w}(s)) - \frac{1}{s}\hat{\boldsymbol{w}}\right\rangle \mathrm{d}s &+ \int_1^T \frac{\beta}{1-\beta}\left\langle \boldsymbol{r}(s), -\frac{\mathrm{d}^2\boldsymbol{w}(s)}{\mathrm{d}s^2}\right\rangle \mathrm{d}s,\end{aligned}$$

which by integration by part leads to

$$\mathrm{RHS} = \int_1^T \left\langle \boldsymbol{r}(s), -\nabla\mathcal{L}(\boldsymbol{w}(s)) - \frac{1}{s}\hat{\boldsymbol{w}}\right\rangle \mathrm{d}s + \frac{\beta}{1-\beta}\left(\left\langle \boldsymbol{r}(T), -\frac{\mathrm{d}\boldsymbol{w}(T)}{\mathrm{d}t}\right\rangle - \left\langle \boldsymbol{r}(1), -\frac{\mathrm{d}\boldsymbol{w}(1)}{\mathrm{d}t}\right\rangle + \int_1^T \left\langle \frac{\mathrm{d}\boldsymbol{r}(s)}{\mathrm{d}s}, \frac{\mathrm{d}\boldsymbol{w}(s)}{\mathrm{d}s}\right\rangle \mathrm{d}s\right),$$

We then check the terms one by one:

- $\int_1^T \left\langle \boldsymbol{r}(s), -\nabla\mathcal{L}(\boldsymbol{w}(s)) - \frac{1}{s}\hat{\boldsymbol{w}}\right\rangle \mathrm{d}s$: This term also occurs in the analysis of GD [34], the analysis of which can be generalized as it does not depend on the form of $\boldsymbol{w}(s)$;
- $\langle \boldsymbol{r}(T), -\frac{\mathrm{d}\boldsymbol{w}(T)}{\mathrm{d}t}\rangle$: as shown in Stage I, $\frac{\mathrm{d}\boldsymbol{w}(T)}{\mathrm{d}t} \to 0$ (i.e., $\boldsymbol{w}(T) - \boldsymbol{w}(T-1) \to 0$ in the discrete case) as $T \to \infty$. Thus, this term is $\boldsymbol{o}(\|\boldsymbol{r}(T)\|)$;
- $\int_1^T \left\langle \frac{\mathrm{d}\boldsymbol{r}(s)}{\mathrm{d}s}, \frac{\mathrm{d}\boldsymbol{w}(s)}{\mathrm{d}s}\right\rangle \mathrm{d}s$: finite due to $\frac{\mathrm{d}\boldsymbol{r}(s)}{\mathrm{d}s} = \frac{\mathrm{d}\boldsymbol{w}(s)}{\mathrm{d}s} - \frac{1}{s}\hat{\boldsymbol{w}}$, $\int_1^{\infty}\|\frac{\mathrm{d}\boldsymbol{w}(s)}{\mathrm{d}s}\|^2\mathrm{d}s$ is finite [5](i.e., $\sum_{t=1}^{\infty}\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|^2 < \infty$ in the discrete case) by Stage I, and the mean-value inequality.

[4] Please note that $\xi(t) \geq 0$ regardless of $t$.

[5] The derivation of this property is in the same manner as in the discrete case: we choose the potential function as $\xi(t) = \mathcal{L}(w(t)) + \frac{\beta}{2(1-\beta)}\|\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t}\|^2$. The derivative of $\xi(t)$ is $\frac{\mathrm{d}\xi(t)}{\mathrm{d}t} = \langle\nabla\mathcal{L}(\boldsymbol{w}(t)), \frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t}\rangle + \frac{\beta}{1-\beta}\langle\frac{\mathrm{d}^2\boldsymbol{w}(t)}{\mathrm{d}t^2}, \frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t}\rangle = -\|\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t}\|^2$. Therefore, $\xi(0) - \xi(T) = \int_0^T\|\frac{\mathrm{d}\boldsymbol{w}(t)}{\mathrm{d}t}\|^2\mathrm{d}t$. Then, taking $T \to \infty$ yields the conclusion.

Putting them together, we show that $\|\boldsymbol{r}(T)\|^2 + \boldsymbol{o}(\|\boldsymbol{r}(T)\|)$ is upper bounded over the iterations, which immediately leads to that $\|\boldsymbol{r}(T)\|$ is bounded. Applying similar methodology to the discrete update rule, we have the following lemma (the proof can be found in Appendix C.1.2).

**Lemma 2.** *Define potential function $g : \mathbb{Z}^+ \to \mathbb{R}$ as*

$$g(t) \triangleq \frac{1}{2}\|\boldsymbol{r}(t)\|^2 + \frac{\beta}{1-\beta}\langle \boldsymbol{r}(t), \boldsymbol{w}(t) - \boldsymbol{w}(t-1)\rangle.$$

*$g(t)$ is upper bounded, which further indicates $\|\boldsymbol{r}(t)\|$ is upper bounded.*

**Remark 3.** *Our technique for analyzing GDM here is essentially more complex and elaborate than that for GD in [34] due to the historical information of gradients GDM. The approach in [34] cannot be directly applied. It is worth mentioning that we provide a more easy-to-check condition for whether $\boldsymbol{r}(t)$ is bounded, i.e., "is $g(t)$ upper-bounded?". This condition can be generalized for other momentum-based implicit regularization analyses. E.g., SGDM and Adam later in this paper.*

## 5 Tackle the difficulty brought by random sampling

In this section, we analyze the implicit regularization of SGDM. Parallel to GDM, we establish the following implicit regularization result for SGDM:

**Theorem 3.** *Let Assumption 1, 2, and 3. (**S**) hold. Let $\beta \in [0,1)$ and $\eta < \frac{1}{\frac{H\beta\sigma_{max}^3}{\sqrt{Nb}\gamma(1-\beta)} + \frac{H\sigma_{max}^4}{2b\gamma^2}}$. Then, with arbitrary initialization point $\boldsymbol{w}(1)$, SGDM (w/. r) satisfies $\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}\| = \mathcal{O}(1/\log(t))$ and $\mathcal{L}(\boldsymbol{w}(t)) = \mathcal{O}(\frac{1}{t})$, almost surely (a.s.).*

Similar to the GDM case, Theorem 3 shows that the implicit regularization of SGDM under this setting is consistent with SGD (c.f. [24] for the implicit regularization of SGD). This matches the observations in practice (c.f. Section 7 for details), and is later supported by our experiments (e.g., Figure 1). We add two remarks on the learning rate upper bound and extension to SGDM (w/. r).

**Remark 4** (On the learning rate)**.** *Firstly, our learning rate upper bound $1/(\frac{H\beta\sigma_{max}^3}{\sqrt{Nb}\gamma(1-\beta)} + \frac{H\sigma_{max}^4}{2b\gamma^2})$ exactly matches that of SGD $2\frac{b\gamma^2}{H\sigma_{max}^4}$ [24] when $\beta = 0$, and matches that of SGD in terms of the order of $\sigma_{max}$, $H$, and $b$ when $\beta \neq 0$. This indicates our analysis is tight. Secondly, as the bound is monotonously increasing with respect to batch size b, Theorem 3 also sheds light on the learning rate tuning, i.e., the larger the batch size is, the larger the learning rate is.*

**Remark 5.** *(On SGDM (w/o. r)) Theorem 3 can be similarly extended to SGDM (w/o. r). We defer the detailed description of the corresponding theorem together with the proof to Appendix F.2.*

Next, we show the proof sketch for Theorem 3. The proof also contains two stages, where Stage II is similar to that for GDM. However, we highlight that **Stage I for SGDM is not a trivial extension of that for GDM**. The methodology used to construct GDM's potential function fails for SGDM: roughly speaking, the change of $\xi(t)$ (defined in Lemma 1) can be spitted into the "decrease term" and the "error term". The "decrease term" comes from the alignment between directions of the update of GDM and the negative gradient, and contributes to the decrease of $\xi(t)$. However, a gap remains between the directions of the update of GDM and the negative gradient. The "error term" measures this gap and potentially contributes to the increase of $\xi(t)$. For GDM, both the terms are in the order of $\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2$ while the "decrease term" has a larger coefficient, leading to the decrease of $\xi(t)$. For SGDM, the "decrease term" is in the order of $(\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|)^2$, while the "error term" is in the order of $\mathbb{E}(\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2)$, bigger than the "decrease term" in an unbounded magnitude. Therefore, $\xi(t)$ may no longer decrease. We defer a detailed discussion to Appendix C.2.3.

On the other hand, we find the stochastic gradient $\nabla\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w})$ is special, in the sense that it satisfies the strong growth condition [32] as follows.

**Lemma 3.** *Let all the assumptions in Theorem 3 hold. Then, for every $t$, $\nabla\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w})$ satisfies the strong growth condition, i.e., $\forall\boldsymbol{w}$, $\mathbb{E}_{\boldsymbol{B}(t)}\|\nabla\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w})\|^2 \le C_2\|\nabla\mathcal{L}(\boldsymbol{w})\|^2$, where $C_2 \triangleq \frac{N\sigma_{max}^2}{\gamma^2 b}$.*

The proof is deferred to Appendix B.2. SGD has been known to converge faster when the strong growth condition holds compared to the case when the variance of the stochastic gradient norm is bounded [32, 42]. However, to the best of our knowledge, convergence of SGDM under the strong growth condition has not been established. We then bridge this gap by proposing the following theorem.

**Theorem 4.** *Consider SGDM defined in Eq. (2) with general $\mathcal{L}$ and $\xi(t)$. Let $\nabla \mathcal{L}_{\xi(t)}(\boldsymbol{w})$ satisfy the following properties:*

- *(Affine noise variance condition). $\forall \boldsymbol{w}$, $\mathbb{E}_{\xi(t)} \|\nabla \mathcal{L}_{\xi(t)}(\boldsymbol{w})\|^2 \le \sigma_1 \|\nabla \mathcal{L}(\boldsymbol{w})\|^2 + \sigma_0$.*
- *(Bounded smoothness condition). $\forall \boldsymbol{w}_1, \boldsymbol{w}_2$, $\|\nabla \mathcal{L}(\boldsymbol{w}_1) - \nabla \mathcal{L}(\boldsymbol{w}_2)\| \le L \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$.*

*Then, with learning rate $\eta \le \frac{1}{2\frac{L\beta\sqrt{\sigma_1}}{1-\beta} + L\sigma_1}$, we have*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{w}(t))\|^2 \le \mathcal{O}\left(\frac{1}{T\eta}\right) + \mathcal{O}(\eta\sigma_0).$$

To the best of our knowledge, Theorem 4 is **the first to establish the convergence of SGDM under the affine noise variance condition [6]**, which generalizes the strong growth condition and the bounded gradient variance condition. Note that the affine noise variance condition with $\sigma_0 = 0$ is exact the strong growth condition, in which case Theorem 4 shows that $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla \mathcal{L}(\boldsymbol{w}(t))\|^2$ decays with rate $\frac{1}{T}$. Combining Lemma 3 and Theorem 4 completes the proof of Stage I for SGDM.

We then briefly state the proof idea of Theorem 4. Inspired by SGD's simple update rule, we rearrange the update rule of SGDM such that only the gradient information of the current step is contained, i.e.,

$$\frac{\boldsymbol{w}(t+1) - \beta \boldsymbol{w}(t)}{1-\beta} = \frac{\boldsymbol{w}(t) - \beta \boldsymbol{w}(t-1)}{1-\beta} - \eta \nabla \mathcal{L}_{\xi(t)}(\boldsymbol{w}(t)).$$

By defining $\boldsymbol{u}(t) \triangleq \frac{\boldsymbol{w}(t)-\beta\boldsymbol{w}(t-1)}{1-\beta} = \boldsymbol{w}(t) + \frac{\beta}{1-\beta}(\boldsymbol{w}(t) - \boldsymbol{w}(t-1))$, we have that $\boldsymbol{u}(t)$ is close to $\boldsymbol{w}(t)$ (differs by order of one-step update $\boldsymbol{w}(t) - \boldsymbol{w}(t-1)$), and the update rule of $\boldsymbol{u}(t)$ only contains the current-step-gradient information $\nabla \mathcal{L}(\boldsymbol{w}(t))$. We then select potential function as $\mathcal{L}(\boldsymbol{u}(t))$, and a simple Taylor's expansion directly leads to:

$$\mathbb{E}[\mathcal{L}(\boldsymbol{u}(t+1))|\mathcal{F}_t] \approx \mathcal{L}(\boldsymbol{u}(t)) - \eta\langle \nabla \mathcal{L}(\boldsymbol{w}(t)), \nabla \mathcal{L}(\boldsymbol{u}(t))\rangle \approx \mathcal{L}(\boldsymbol{u}(t)) - \eta \|\nabla \mathcal{L}(\boldsymbol{w}(t))\|^2,$$

i.e., $\mathcal{L}(\boldsymbol{u}(t))$ is a proper potential function. Summing the above equation across iterations completes the proof.

# 6 Analyze the effect of preconditioners

## 6.1 Implicit regularization of deterministic Adam

This section presents the implicit regularization of deterministic Adam, i.e., Adam without random sampling.

**Theorem 5.** *Let Assumption 1, 2, and 3. (**D**) hold. Let $1 > \beta_2 > \beta_1^4 \ge 0$, and the learning rate $\eta$ is a small enough constant (The upper bound of learning rate is complex, and we defer it to Appendix D.1). Then, with arbitrary initialization point $\boldsymbol{w}(1)$, deterministic Adam (Eq. 3) satisfies that $\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}\| = \mathcal{O}(1/\log(t))$ and $\mathcal{L}(\boldsymbol{w}(t)) = \mathcal{O}(\frac{1}{t})$.*

**Remark 6** (On the $\beta_1$ and $\beta_2$ range)**.** *Almost all existing literature assume a time-decaying hyper-parameter choice of $\beta_1$ or $\beta_2$ (c.f., [18, 3]). On the other hand, our result proves that deterministic Adam converges with constant settings of $\beta_1$ and $\beta_2$, which agrees with the practical use.*

**Remark 7** ((Discussion on the results in [34]))**.** *[34] observe that, on a synthetic dataset, the direction of the output parameter by Adam still does not converge to the max-margin direction after $2 \times 10^6$ iterations (but is getting closer). At the same time, GD seems to converge to the max-margin direction. However, this example does not contradict Theorem 5, as we study the asymptotic behavior. In the example, it can be observed that the angle gap keeps decreasing after $10^5$ iteration, which stands with our result. Furthermore, in Figure 3 and 4 of our paper, we reproduce the experiment and find that the angle gap still keeps decreasing after extending the training time. Secondly, this synthetic dataset is ill-posed, which has large singular value and makes the training rather slow, and it can be* *observed that the training loss of Adam does not vary much until* $10^4$ *iteration. On the well-posed dataset adopted by (Figure 1, [34]), we observe that Adam will converge to the max-margin solution rapidly (please refer to Figure 1 for details).*

We simply introduce the proof idea here and put the full proof in Appendix D. The proof of deterministic Adam is very similar to that of GDM with minor changes in potential functions. Specifically, $\xi(t)$ in Lemma 1 is changed to $\xi(t) \triangleq \mathcal{L}(\boldsymbol{w}(t)) + \frac{1}{2}\frac{1-\beta_1^{t-1}}{\eta(1-\beta_1)}\|\sqrt[4]{\varepsilon\mathbb{1}_d + \hat{\boldsymbol{\nu}}(t-1)} \odot (\boldsymbol{w}(t) - \boldsymbol{w}(t-1))\|^2$, and $g(t)$ in Lemma 2 is changed to

$$g(t) \triangleq \left\langle \boldsymbol{r}(t), (1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d + \hat{\boldsymbol{\nu}}(t-1)} \odot (\boldsymbol{w}(t) - \boldsymbol{w}(t-1)) \right\rangle \frac{\beta_1}{1-\beta_1} + \frac{\sqrt{\varepsilon}}{2}\|\boldsymbol{r}(t)\|^2.$$

The rest of the proof then flows similarly to the GDM case.

### 6.2 What if random sampling is added?

We have obtained the implicit regularization for GDM, SGDM, and deterministic Adam. One may wonder whether the implicit regularization of stochastic Adam can be obtained. Unfortunately, the gap can not be closed yet. This is because an implicit regularization analysis requires the knowledge of the loss dynamics, little of which, however, has been ever known even for stochastic RMSProp (i.e., Adam with $\beta_1 = 0$ in Eq. (3)) with constant learning rates. Specifically, the main difficulty lies in bounding the change of conditioner $\frac{1}{\sqrt{\varepsilon\mathbb{1}_d + \hat{\boldsymbol{\nu}}(t)}}$ across iterations, which is required to make the drift term $\langle \nabla\mathcal{L}(\boldsymbol{w}(t)), \mathbb{E}(\boldsymbol{w}(t+1) - \boldsymbol{w}(t))\rangle$ (derived by Taylor's expansion of the epoch start from $Kt$) negative to ensure a non-increasing loss.

On the other hand, if we adopt decaying learning rates $\eta_t = \frac{\eta_1}{\sqrt{t}}$, [33] shows $\beta_2$ close enough to 1, the following equation holds for stochastic RMSProp (w/o. r) (recall that $K \triangleq \frac{N}{b}$ is the epoch size)

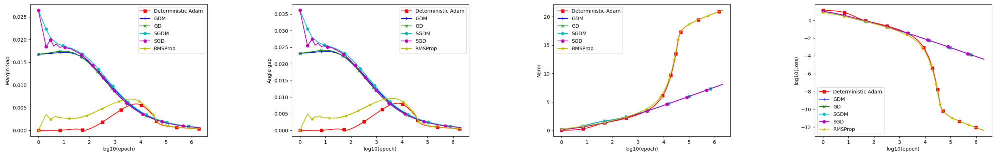
$$\sum_{t=0}^{T} \frac{1}{\sqrt{t+1}} \|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\| = \mathcal{O}(\ln T). \tag{8}$$

Based on this result, we have the following theorem for stochastic RMSProp (w/o. r):

**Theorem 6.** *Let Assumptions 1, 2, and 3. (S) hold. Let $\beta_2$ be close enough to 1. Then, with arbitrary initialization point $\boldsymbol{w}(1)$ and decaying learning rate $\eta_t = \frac{\eta_1}{\sqrt{t}}$, stochastic RMSProp satisfies $\|\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|} - \frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}\| = \mathcal{O}(1/\log(t))$ and $\mathcal{L}(\boldsymbol{w}(t)) = \mathcal{O}(\frac{1}{\sqrt{t}})$.*

**Remark 8** (On the decaying learning rate)**.** *The decaying learning rate is a "stronger" setting compared to the constant learning rate, both in the sense that GDM, SGDM, and deterministic Adam can be shown to converge to the max-margin solution following the same routine as Theorems 2, 3, and 5, and in the sense that we usually adopt constant learning rate in practice.*

The proof is on the grounds of a novel characterization of the loss convergence rate derived from Eq. (8), and readers can find the details in Appendix E.



(a) Margin Gap (b) Angle gap (c) Norm (d) Loss

Figure 1: Comparison of the implicit regularization of the optimizers. We use the synthetic dataset in [34] with learning rate $\frac{1}{\sigma_{max}^2}$. Figure (b) shows (1). all the optimizers converge to the max margin solution, and (2). the asymptotic behaviors with & without momentum are similar. The experimental observation support our theoretical results.

## 7 Discussions

**Consistency with the Experimental Results.** We conduct experiments to verify our theoretical findings. Specifically, we (1). run GD, GDM, SGD, SGDM, and Adam on a synthetic dataset to observe their implicit regularization; (2). run GD and Adam on ill-posed dataset proposed in [34] to verify Theorem 5; (3). run SGD and SGDM on neural networks to classify the MNIST dataset and compare their implicit regularization. The experimental observations stand with our theoretical results. Furthermore, it is worth mentioning that experimental phenomenons that adding momentum will not change the implicit regularization have also been observed by existing literature [34, 23, 43].

**Influence of hyperparameters on convergence rates.** Our results can be further extended to provide a precise characterization of the influence of the hyperparameters $\eta$ and $\beta$ on the convergence rate of (S)GDM. Specifically, in Appendix F.4, we show that the asymptotic convergence rate of (S)GDM is $C\frac{1}{\eta}\frac{1}{t}$, where $C$ is some constant independent of $\beta$ and $\eta$. Therefore, increase $\eta$ can lead to a faster convergence rate (with learning rate requirements in Theorems 2 and 3 satisfied). However, changing $\beta$ does not affect the convergence rate, which is also observed in our experiments (e.g., Figure 1). Furthermore, to the best of our knowledge, there still lacks the theoretical justification of the acceleration effect of the momentum except the strongly convex case. On the other hand, we still can not provide a precise characterization of the exact convergence behavior of Adam in this case due to the ever-changing preconditioner. As this work mainly focuses on the linear model, whether momentum and preconditioner can accelerate the training of non-linear model with the non-convex landscape is an exciting future direction and deserves further investigation.

**Generalization Behavior of Adam.** Our results characterize the implicit regularization behavior of optimizers when the time is large enough, which can be changed by early stopping, especially for Adam. This is because the proof of Theorem 5 relies on that when the training time is large enough, the gradient is small and the adaptive learning rate $\frac{1}{\sqrt{\boldsymbol{\nu}(t)+\varepsilon\mathbb{1}_d}}$ is dominated by $\frac{1}{\varepsilon}\mathbb{1}_d$, and then Adam behaves like GDM and thus converges to the max margin solution. This can also be verified by Figure 1, as the angle gap of deterministic Adam and RMSProp only start to decrease late in the training.

**Gap Between The Linear Model and Deep Neural Networks.** While our results only hold for the linear classification problem, extending the results to the deep neural networks is possible. Specifically, existing literature [21, 43] provide a framework for deriving implicit regularization for deep homogeneous neural networks. However, the approach in [21, 43] can not be trivially applied to the momentum-based optimizers, as their proofs require the specific gradient-based updates to lower bound a smoothed margin (c.f., Theorem 4.1, [21]). It remains an exciting work to see how our results can be expanded to GDM and Adam for deep neural networks.

## 8 Conclusion

This paper studies the implicit regularization of momentum-based optimizers in linear classification with exponential-tailed loss. Our results indicate that for SGD and the deterministic version of Adam, adding momentum will not influence the implicit regularization, and the direction of the parameter converges to the $L^2$ max-margin solution. Our theoretical results stand with existing experimental observations, and developed techniques such as the potential functions may inspire the analyses on other momentum-based optimizers. Motivated by the results and techniques for linear cases in this paper, it has the potential to extend them to the homogeneous neural network in the future. Another topic left for future work is to derive the implicit regularization of constant learning rate stochastic Adam. As discussed in Section 6.2, this topic is non-trivial, and it requires new techniques and assumptions to be developed.

## 9 Acknowledgement

This research is supported in part by CAS Project for Young Scientists in Basic Research under Grant No. YSBR-034 and Innovation Project of ICT CAS under Grants No. E261090.

# References

[1] A. Ali, E. Dobriban, and R. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pages 233–244. PMLR, 2020.

[2] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[3] X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2018.

[4] S. De, A. Mukherjee, and E. Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

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

[6] M. Faw, I. Tziotis, C. Caramanis, A. Mokhtari, S. Shakkottai, and R. Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. *arXiv preprint arXiv:2202.05791*, 2022.

[7] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

[8] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[9] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pages 9461–9471, 2018.

[10] Z. Guo, Y. Xu, W. Yin, R. Jin, and T. Yang. A novel convergence analysis for algorithms of the adam family. *arXiv preprint arXiv:2112.03459*, 2021.

[11] S. Jelassi and Y. Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pages 9965–10040. PMLR, 2022.

[12] Z. Ji, N. Srebro, and M. Telgarsky. Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR, 2021.

[13] Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[14] Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. *arXiv preprint arXiv:2006.06657*, 2020.

[15] Z. Ji and M. Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR, 2021.

[16] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[17] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

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

[19] J. Lin and L. Rosasco. Optimal rates for multi-pass stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

[20] Y. Liu, Y. Gao, and W. Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33, 2020.

[21] K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[23] M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[24] M. S. Nacson, N. Srebro, and D. Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019.

[25] Y. Nesterov. A method for solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics. Doklady*, volume 27, pages 367–372, 1983.

[26] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR*, 2015.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[28] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[29] Q. Qian and X. Qian. The implicit bias of adagrad on separable data. In *Advances in Neural Information Processing Systems*, pages 7761–7769, 2019.

[30] L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1630–1638, 2015.

[31] R. E. Schapire and Y. Freund. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

[32] M. Schmidt and N. L. Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

[33] N. Shi, D. Li, M. Hong, and R. Sun. Rmsprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021.

[34] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[35] W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

[36] T. Sun, P. Yin, D. Li, C. Huang, L. Guan, and H. Jiang. Non-ergodic convergence analysis of heavy-ball algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5033–5040, 2019.

[37] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[38] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[39] W. Tao, S. Long, G. Wu, and Q. Tao. The role of momentum parameters in the optimal convergence of adaptive polyak's heavy-ball methods. *arXiv preprint arXiv:2102.07314*, 2021.

[40] M. Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pages 307–315. PMLR, 2013.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[42] S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

[43] B. Wang, Q. Meng, W. Chen, and T.-Y. Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *International Conference on Machine Learning*, pages 10849–10858. PMLR, 2021.

[44] A. C. Wilson, B. Recht, and M. I. Jordan. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

[45] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in neural information processing systems*, pages 4148–4158, 2017.

[46] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[47] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang. A unified analysis of stochastic momentum methods for deep learning. In *IJCAI International Joint Conference on Artificial Intelligence*, 2018.

[48] H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

[49] D. Zou, Y. Cao, Y. Li, and Q. Gu. Understanding the generalization of adam in learning neural networks with proper regularization. *arXiv preprint arXiv:2108.11371*, 2021.

# Supplementary Materials for<br>“ Does Momentum Change the Implicit Regularization on Separable Data?”

## A Additional Related Works

[11] and [49] also investigate the generalization behavior of GDM and Adam but focus on a different setting from ours. Specifically, they work on a two-layer convolutional neural network with fixed and untrained second layer and cubic activation, while we work on the linear classifier. Also, except linear separability, they further pose additional specific requirements for the dataset (e.g., all data shares a same (scaled) patch). Based on the settings, [11] shows that the output hypothesis by GDM provably generalizes better than GD, while [49] proves that the output hypothesis by GD generalizes better than Adam. These results, however, do not contradict our findings due to the difference in the settings. We believe both the works and this paper have its own merits to unveil the mystery of the generalization of momentum-based optimizers.

## B Preparations

This section collect definitions and lemmas which will be used throughout the proofs.

### B.1 Characterization of the max-margin solution

This section collects several commonly-used characterization of the max-margin solution from [24] and [34].

To start with, we define support vectors and support set, which are two common terms in margin analysis. Recall that in the main text, we assume that without the loss of generality, $\boldsymbol{y}_i = 1$, $\forall i \in \{1, \cdots, N\}$.

**Definition 2** (Support vectors and support set)**.** *For any $i \in [N]$, $\boldsymbol{x}_i$ is called a support vector of the dataset $\boldsymbol{S}$, if*

$$\langle \boldsymbol{x}_i, \hat{\boldsymbol{w}} \rangle = 1.$$

*Correspondingly, $\boldsymbol{x}_i$ is called a non-support vector if $\langle \boldsymbol{x}_i, \hat{\boldsymbol{w}} \rangle > 1$. The support set of $\boldsymbol{S}$ is then defined as*

$$\boldsymbol{S}_s = \{\boldsymbol{x} \in \boldsymbol{S} : \langle \boldsymbol{x}, \hat{\boldsymbol{w}} \rangle = 1\}.$$

The following lemma delivers $\hat{\boldsymbol{w}}$ as an linear combination of support vectors.

**Lemma 4** (Lemma 12, [34])**.** *For almost every datasets $S$, there exists a unique vector $\boldsymbol{v} = (\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N)$, such that $\hat{\boldsymbol{w}}$ can be represented as*

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{N} \boldsymbol{v}_i \boldsymbol{x}_i, \tag{9}$$

*where $\boldsymbol{v}$ satisfies $\boldsymbol{v}_i = 0$ if $\boldsymbol{x}_i \notin \boldsymbol{S}_s$, and $\boldsymbol{v}_i > 0$ if $\boldsymbol{x}_i \in \boldsymbol{S}_s$. Furthermore, the size of $\tilde{\boldsymbol{S}}_s$ is at most $d$.*

By Lemma 4, we further have the following corollary:

**Corollary 1.** *For almost every datasets $S$, the unique $\boldsymbol{v}$ given by Lemma 4 further satisfies that for any positive constant $C_3$, there exists a non-zero vector $\tilde{\boldsymbol{w}}$, such that, $\boldsymbol{x}_i \in \boldsymbol{S}_s$, we have*

$$C_3 e^{-\langle \boldsymbol{x}_i, \tilde{\boldsymbol{w}} \rangle} = \boldsymbol{v}_i. \tag{10}$$

*Proof.* For almost every datasets $\boldsymbol{S}$, any subsets with size $d$ of $\boldsymbol{S}$ is linearly independent. Since $\tilde{\boldsymbol{S}}_s$ has size no larger than $d$ (by Lemma 4), and Eq. (10) is equivalent to linear equations, the proof is completed. □

For the stochastic case, we will also need the following lemma when we calculate the form of parameter at time $t$.

**Lemma 5** (Lemma 5, [24]). *Let $\boldsymbol{B}(s)$ be the random subset used in SGDM (w/. r). Almost surely, there exists a vector $\check{\boldsymbol{w}}$*

$$\frac{N}{b}\sum_{s=1}^{t-1}\frac{1}{s}\sum_{\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s}\boldsymbol{v}_i\boldsymbol{x}_i=\ln\left(\frac{bt}{N}\right)\hat{\boldsymbol{w}}+\boldsymbol{n}(t)+\check{\boldsymbol{w}},$$

*where $\boldsymbol{n}(t)$ satisfies $\|\boldsymbol{n}(t)\|=\boldsymbol{o}(t^{-0.5+\varepsilon})$ for any $\varepsilon>0$, and $\|\boldsymbol{n}(t+1)-\boldsymbol{n}(t)\|=O(t^{-1})$. As for SGDM (w/o. r), the a.s. condition can be removed.*

### B.2 Preparations of the optimization analysis

This section collects technical lemmas which will be used in latter proofs. We begin with a lemma bounding the smooth constants if the loss is bounded.

**Lemma 6.** *If loss $\ell$ satisfies (D) in Assumption 3, then for any $\boldsymbol{w}_0$, if $\mathcal{L}(\boldsymbol{w})\le\mathcal{L}(\boldsymbol{w}_0)$, then we have $\mathcal{L}$ is $\sigma_{max}^2H_{s_0}$ smooth at point $\boldsymbol{w}$, where $s_0=\ell^{-1}(N\mathcal{L}(\boldsymbol{w}_0))$. Furthermore, $\mathcal{L}$ is globally $\sigma_{max}^2H_{s_0}$ smooth over the set $\{\boldsymbol{w}:\mathcal{L}(\boldsymbol{w})\le\mathcal{L}(\boldsymbol{w}_0)\}$.*

*Proof.* Since $\ell$ is positive, we have $\forall i\in[N]$,

$$\frac{\tilde{\ell}(\boldsymbol{w},\boldsymbol{x}_i)}{N}<\frac{\sum_{j=1}^N\tilde{\ell}(\boldsymbol{w},\boldsymbol{x}_j)}{N}=\mathcal{L}(\boldsymbol{w})\le\mathcal{L}(\boldsymbol{w}_0),$$

which leads to $\tilde{\ell}(\boldsymbol{w},\boldsymbol{x}_i)<N\mathcal{L}(\boldsymbol{w}_0)$, and $\ell$ is $H_{s_0}$ smooth at $\langle\boldsymbol{w},\boldsymbol{x}_i\rangle$.

Furthermore, since $\nabla_{\boldsymbol{w}}\tilde{\ell}(\boldsymbol{w},\boldsymbol{x}_i)=\nabla_{\boldsymbol{w}}\ell(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle)=\ell'(\langle\boldsymbol{w},\boldsymbol{x}_i\rangle)\boldsymbol{x}_i$, for any two parameters $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ close enough to $\boldsymbol{w}$,

$$\begin{aligned}\|\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_1)-\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}_2)\|=&\left\|\sum_{\boldsymbol{x}\in\boldsymbol{S}}(\ell'(\langle\boldsymbol{w}_1,\boldsymbol{x}\rangle)-\ell'(\langle\boldsymbol{w}_2,\boldsymbol{x}\rangle))\boldsymbol{x}\right\|\\ \le&\sigma_{max}\sqrt{\sum_{\boldsymbol{x}\in\boldsymbol{S}}(\ell'(\langle\boldsymbol{w}_1,\boldsymbol{x}\rangle)-\ell'(\langle\boldsymbol{w}_2,\boldsymbol{x}\rangle))^2}\le\sigma_{max}H_{s_0}\sqrt{\sum_{\boldsymbol{x}\in\boldsymbol{S}}(\langle\boldsymbol{w}_1-\boldsymbol{w}_2,\boldsymbol{x}\rangle)^2}\\ \le&\sigma_{max}^2H_{s_0}\|\boldsymbol{w}_1-\boldsymbol{w}_2\|.\end{aligned}$$

Now if $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ both belong to $\{\boldsymbol{w}:\mathcal{L}(\boldsymbol{w})\le\mathcal{L}(\boldsymbol{w}_0)\}$, we have for any $\boldsymbol{x}_i\in\boldsymbol{S}$, $\langle\boldsymbol{w}_1,\boldsymbol{x}_i\rangle>\ell^{-1}(N\mathcal{L}(\boldsymbol{w}_0))$, and $\langle\boldsymbol{w}_2,\boldsymbol{x}_i\rangle>\ell^{-1}(N\mathcal{L}(\boldsymbol{w}_0))$. Following the same routine as the locally smooth proof, we complete the second argument.

The proof is completed. □

Based on Assumption 2, we also have the following lemma characterizing the relationship between loss $\ell$ and its derivative $\ell'$ when $x$ is large enough.

**Lemma 7.** *Let loss $\ell$ satisfy Assumption 2. Then, there exists an large enough $x_0$ and a positive real $K$, such that, $\forall x>x_0$, we have*

$$-\frac{1}{4}\ell'(x)\le\ell(x)\le-4\ell'(x).$$

*Proof.* By Assumption 2, there exists a large enough $x_0$, such that $\forall x>x_0$, we have

$$\frac{1}{2}e^{-x}\le-\ell'(x)\le2e^{-x}.\tag{11}$$

On the other hand, as $\lim_{t\to\infty}\ell(x)=0$, we have

$$\ell(x)=\int_{s=x}^{\infty}-\ell'(s)\mathrm{d}s,$$

which by Eq. (11) leads to

$$\frac{1}{2}e^{-x}=\frac{1}{2}\int_x^\infty e^{-s}\mathrm{d}s\le\ell(x)\le2\int_x^\infty e^{-s}\mathrm{d}s=2e^{-x}.$$

The proof is completed. □

By Lemma 7, we immediately get the following corollary:

**Corollary 2.** *Let loss $\ell$ satisfy Assumption 2. Then, there exist positive reals $C_g$ and $C_l$, such that, for any $\boldsymbol{w} \in \mathbb{R}^d$ satisfying either $\|\nabla \mathcal{L}(\boldsymbol{w})\| \le C_g$ or $\mathcal{L}(\boldsymbol{w}) \le C_l$, we have*

$$\frac{\gamma}{4}\mathcal{L}(\boldsymbol{w}) \le \|\nabla \mathcal{L}(\boldsymbol{w})\| \le 4\mathcal{L}(\boldsymbol{w}).$$

*Proof.* We start with the case $\|\nabla \mathcal{L}(\boldsymbol{w})\| \le C_g$. By simple calculation, we have

$$\|\nabla \mathcal{L}(\boldsymbol{w})\| = \frac{1}{N}\left\|\sum_{i=1}^{N} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)\boldsymbol{x}_i\right\| \le -\frac{\sigma_{max}}{N}\sum_{i=1}^{N} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle), \tag{12}$$

and

$$\|\nabla \mathcal{L}(\boldsymbol{w})\|\|\hat{\boldsymbol{w}}\| \ge \langle \nabla \mathcal{L}(\boldsymbol{w}), \hat{\boldsymbol{w}}\rangle \ge -\frac{1}{N}\sum_{i=1}^{N} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle). \tag{13}$$

By Assumption 2, we have there exists a constant $C'_g$, s.t., any $x$ with $-\ell'(x) > C'_g$ satisfies $x > x_0$. Let $C_g = \frac{C'_g\gamma}{N}$. We then have if $\|\nabla \mathcal{L}(\boldsymbol{w})\| \le C_g$, then $\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle > x_0$ $(\forall i)$, and thus $4\ell(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle) \ge -\ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle) \ge \frac{1}{4}\ell(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)$. Combing Eqs. (12) and (13), we then have

$$4\mathcal{L}(\boldsymbol{w}) = 4\sum_{i=1}^{N} \ell(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle) \ge \|\nabla \mathcal{L}(\boldsymbol{w})\| \ge \frac{\gamma}{4N}\sum_{i=1}^{N} \ell(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle) = \frac{\gamma}{4}\mathcal{L}(\boldsymbol{w}).$$

Similarly, as for the case $\mathcal{L}(\boldsymbol{w}) \le C_l$, we have there exists a constant $C'_l$, s.t., any $x$ with $\ell(x) < C'_l$ satisfies $x > x_0$. Let $C_l = \frac{C_l\sigma_{max}}{N}$ and the rest of the proof follows the same routine as the first case. The proof is completed. □

The following lemma bridges the second moment of $\nabla \mathcal{L}_{\boldsymbol{B}(t)}$ with its squared first moment.

**Lemma 8** (Lemma 3, restated)**.** *Let the dataset $\boldsymbol{S}$ satisfies the separable assumption 1. Let $\boldsymbol{B}$ be a random subset of $\boldsymbol{S}$ with size $b$ sampled independently and uniformly without replacement. Then, at any point $\mathbf{w}$, we have*

$$\|\nabla \mathcal{L}(\boldsymbol{w})\|^2 \le \mathbb{E}_{\boldsymbol{B}}\left[\|\nabla \mathcal{L}_{\boldsymbol{B}}(\boldsymbol{w})\|^2\right] \le \frac{N\sigma_{max}^2}{\gamma^2 b}\|\nabla \mathcal{L}(\boldsymbol{w})\|^2.$$

*Proof.* To start with, notice that

$$\|\nabla \mathcal{L}(\boldsymbol{w})\| = \|\mathbb{E}_{\boldsymbol{B}}\nabla \mathcal{L}_{\boldsymbol{B}}(\boldsymbol{w})\| \le \mathbb{E}_{\boldsymbol{B}}\|\nabla \mathcal{L}_{\boldsymbol{B}}(\boldsymbol{w})\|.$$

Therefore, the first inequality can be directly obtained by Cauchy-Schwartz's inequality. To prove the second inequality, we first calculate the explicit form of $\nabla \mathcal{L}_{\boldsymbol{B}}(\boldsymbol{w})$.

$$\|\nabla \mathcal{L}_{\boldsymbol{B}}(\boldsymbol{w})\|^2 = \frac{1}{b^2}\left\|\sum_{\boldsymbol{x}\in\boldsymbol{B}} \nabla\tilde{\ell}(\boldsymbol{w}, \boldsymbol{x})\right\|^2 = \frac{1}{b^2}\left\|\sum_{\boldsymbol{x}\in\boldsymbol{B}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}\right\|^2 \le \frac{\sigma_{max}^2}{b^2}\sum_{\boldsymbol{x}\in\boldsymbol{B}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)^2.$$

Therefore,

$$\mathbb{E}_{\boldsymbol{B}}\|\nabla \mathcal{L}_{\boldsymbol{B}}(\boldsymbol{w})\|^2 \le \frac{\sigma_{max}^2}{Nb}\sum_{\boldsymbol{x}\in\boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)^2 \le \frac{\sigma_{max}^2}{Nb}\left(\sum_{\boldsymbol{x}\in\boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\right)^2. \tag{14}$$

On the other hand,

$$\begin{aligned}\|\nabla \mathcal{L}(\boldsymbol{w})\| &= \frac{1}{N}\left\|\sum_{\boldsymbol{x}\in\boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}\right\| \\ &\ge \frac{1}{N}\left\langle \sum_{\boldsymbol{x}\in\boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}, -\frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}\right\rangle \overset{(\star)}{\ge} \frac{\gamma}{N}\sum_{\boldsymbol{x}\in\boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\end{aligned}$$

where Eq. ($\star$) is due to $\forall \boldsymbol{x} \in \boldsymbol{S}$, $\langle \boldsymbol{x}, -\hat{\boldsymbol{w}} \rangle \geq 1$ and $\ell' < 0$.

Therefore,

$$\|\nabla \mathcal{L}(\boldsymbol{w})\|^2 \geq \frac{\gamma^2}{N^2} \left( \sum_{\boldsymbol{x} \in \boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x} \rangle) \right)^2 . \tag{15}$$

The proof is completed by putting Eqs. (14) and (15) together. □

In the following lemma, we show the updates of GDM, Adam, and SGDM are all non-zero.

**Lemma 9.** *Regardless of GDM, Adam, or SGDM, the updates of all steps are non-zero, i.e.,*

$$\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\| > 0, \forall t > 1.$$

*Proof.* We start with the alternative forms of the update rule of GDM, Adam, and SGDM using the gradients along the trajectory respectively. For GDM, by Eq. (1), the update rule can be written as

$$\boldsymbol{w}(t+1) - \boldsymbol{w}(t) = -\eta(1-\beta) \left( \sum_{s=1}^{t} \beta^{t-s} \nabla \mathcal{L}(\boldsymbol{w}(s)) \right). \tag{16}$$

Similarly, the update rule of SGDM can be written as

$$\boldsymbol{w}(t+1) - \boldsymbol{w}(t) = -\eta(1-\beta) \left( \sum_{s=1}^{t} \beta^{t-s} \nabla \mathcal{L}_{\boldsymbol{B}(s)}(\boldsymbol{w}(s)) \right), \tag{17}$$

while the update rule of Adam can be given as

$$\boldsymbol{w}(t+1) - \boldsymbol{w}(t) = -\eta \frac{\sum_{s=1}^{t} \frac{1-\beta_1}{1-\beta_1^s} \beta_1^{t-s} \nabla \mathcal{L}(\boldsymbol{w}(s))}{\sqrt{\varepsilon \mathbf{1}_d + \sum_{s=1}^{t} \frac{1-\beta_2}{1-\beta_2^s} \beta_2^{t-s} (\nabla \mathcal{L}(\boldsymbol{w}(s)))^2}}. \tag{18}$$

On the other hand, by the definition of empirical risk $\mathcal{L}$, the gradient of $\mathcal{L}$ at point $\boldsymbol{w}$ can be given as

$$\nabla \mathcal{L}(\boldsymbol{w}) = \frac{\sum_{i=1}^{N} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \boldsymbol{x}_i}{N}. \tag{19}$$

By Eq. (19) and Eq. (16), we further have for GDM,

$$\boldsymbol{w}(t+1) - \boldsymbol{w}(t) = -\eta(1-\beta) \left( \sum_{s=1}^{t} \beta^{t-s} \frac{\sum_{i=1}^{N} \ell'(\langle \boldsymbol{w}(s), \boldsymbol{x}_i \rangle) \boldsymbol{x}_i}{N} \right). \tag{20}$$

By Assumption 1, there exists a non-zero parameter $\hat{\boldsymbol{w}}$, such that, $\langle \hat{\boldsymbol{w}}, \boldsymbol{x}_i \rangle > 0$, $\forall i$. Therefore, by executing inner product between Eq. (20) and $\hat{\boldsymbol{w}}$, we have

$$\begin{aligned} &\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\| \|\hat{\boldsymbol{w}}\| \geq \langle \boldsymbol{w}(t+1) - \boldsymbol{w}(t), \hat{\boldsymbol{w}} \rangle \\ = &- (1-\beta)\eta \left( \sum_{s=1}^{t} \beta^{t-s} \frac{\sum_{i=1}^{N} \ell'(\langle \boldsymbol{w}(s), \boldsymbol{x}_i \rangle) \langle \boldsymbol{x}_i, \hat{\boldsymbol{w}} \rangle}{N} \right) \overset{(*)}{>} 0, \end{aligned}$$

where Eq. $(*)$ is due to $\ell' < 0$. This complete the proof for GDM.

Similarly, for SGDM, we have

$$\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\| \|\hat{\boldsymbol{w}}\| \geq -\eta(1-\beta) \left( \sum_{s=1}^{t} \beta^{t-s} \frac{\sum_{(\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{B}} \ell'(\langle \boldsymbol{w}(s), \boldsymbol{y}\boldsymbol{x} \rangle) \langle \boldsymbol{y}\boldsymbol{x}, \hat{\boldsymbol{w}} \rangle}{b} \right) > 0,$$

which completes the proof of SGDM.

For Adam, we have

$$
\begin{aligned}
&\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\| \left\|\hat{\boldsymbol{w}} \odot \sqrt{\varepsilon \mathbf{1}_d+\sum_{s=1}^{t} \frac{1-\beta_2}{1-\beta_2^s} \beta_2^{t-s}(\nabla \mathcal{L}(\boldsymbol{w}(s)))^2}\right\| \\
\geq & -\left\langle \hat{\boldsymbol{w}} \odot \sqrt{\varepsilon \mathbf{1}_d+\sum_{s=1}^{t} \frac{1-\beta_2}{1-\beta_2^s} \beta_2^{t-s}(\nabla \mathcal{L}(\boldsymbol{w}(s)))^2}, \eta \frac{\sum_{s=1}^{t} \frac{1-\beta_1}{1-\beta_1^s} \beta_1^{t-s} \nabla \mathcal{L}(\boldsymbol{w}(s))}{\sqrt{\varepsilon \mathbf{1}_d+\sum_{s=1}^{t} \frac{1-\beta_2}{1-\beta_2^s} \beta_2^{t-s}(\nabla \mathcal{L}(\boldsymbol{w}(s)))^2}}\right\rangle \\
= & \left\langle \hat{\boldsymbol{w}}, \eta \sum_{s=1}^{t} \frac{1-\beta_1}{1-\beta_1^s} \beta_1^{t-s} \nabla \mathcal{L}(\boldsymbol{w}(s)\right\rangle \\
= & -\eta\left(\sum_{s=1}^{t} \frac{1-\beta_1}{1-\beta_1^s} \beta_1^{t-s} \frac{\sum_{i=1}^{N} \ell^{\prime}(\langle\boldsymbol{w}(s), \boldsymbol{x}_i\rangle)\langle\boldsymbol{x}_i, \hat{\boldsymbol{w}}\rangle}{N}\right)>0,
\end{aligned}
$$

which completes the proof of Adam.

The proof is completed. □

## C Implicit regularization of GD/SGD with momentum

This section collects the proof of the implicit regularization of gradient descent with momentum and stochastic gradient descent with momentum. The analyses of this section hold for almost every dataset, and the "almost every" constraint is further moved in Section F.1.

### C.1 Implicit regularization of GD with Momentum

This section collects the proof of Theorem 2.

#### C.1.1 Proof of the sum of squared gradients converges

To begin with, we will prove the sum of squared norm of gradients along the trajectory is finite for gradient descent with momentum. To see this, we first define the continuous-time update rule as

$$\boldsymbol{w}(t+\alpha)-\boldsymbol{w}(t)=\alpha(\boldsymbol{w}(t+1)-\boldsymbol{w}(t)), \forall t \in \mathbb{Z}^{+}, \forall \alpha \in[0,1].$$

We then prove a generalized case of Lemma 1 for any $\boldsymbol{w}(t+\alpha)$.

**Lemma 10** (Lemma 1, extended)**.** *Let all conditions in Theorem 2 hold. We then have*

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{w}(t))+\frac{\beta}{2\eta(1-\beta)}\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|^2 \geq & \mathcal{L}(\boldsymbol{w}(t+\alpha))+\frac{\beta}{2\eta(1-\beta)} \alpha^2\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2 \\
& +\frac{(1-C_1)\alpha^2}{\eta}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2, \qquad (21)
\end{aligned}
$$

*where $C_1$ is a positive real such that $\eta=2\frac{N}{H_{s_0}\sigma_{max}^2}C_1$ and $s_0 \overset{\triangle}{=} \ell^{-1}(N\mathcal{L}(\boldsymbol{w}_1))$.*

*Proof of Lemma 10.* We prove this lemma by reduction to absurdity.

Concretely, let $t^*$ be the smallest positive integer time such that there exists an $\alpha \in [0,1]$, such that Eq. (21) doesn't hold. Let $\alpha^*=\inf\{\alpha\in[0,1]: Eq.\ (21)\ doesn't\ hold\ for\ (t^*,\alpha)\}$. By continuity, Eq. (21) holds for $(t^*,\alpha^*)$.

We further divide the proof into two cases depending on the value of $\alpha^*$.

**Case 1:** $\alpha^*=0$: For any $t^*>t\geq 1$, we have Eq. (21) holds for $(t,1)$. Specifically, we have

$$\mathcal{L}(\boldsymbol{w}(t))+\frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|^2 \geq \mathcal{L}(\boldsymbol{w}(t+1))+\frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2,$$

which further leads to

$$\mathcal{L}(\boldsymbol{w}(1)) = \mathcal{L}(\boldsymbol{w}(1)) + \frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(1)-\boldsymbol{w}(0)\|^2 \geq \mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2.$$

Since $\frac{\beta}{2\eta}\|\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2$ is non-negative, we have

$$\mathcal{L}(\boldsymbol{w}(1)) \geq \mathcal{L}(\boldsymbol{w}(t^*)).$$

By Lemma 6, we have $\mathcal{L}$ is $H_{s_0}$ smooth at $\boldsymbol{w}(t^*)$. Therefore, by Taylor's expansion for $\mathcal{L}$ at point $\boldsymbol{w}(t^*)$, we have for small enough $\alpha > 0$

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}(t^*+\alpha))\\
\leq&\mathcal{L}(\boldsymbol{w}(t^*)) + \langle \nabla\mathcal{L}(\boldsymbol{w}(t^*)), \boldsymbol{w}(t^*+\alpha)-\boldsymbol{w}(t^*)\rangle + \frac{H_{s_0}\sigma_{max}^2}{2N}\|\boldsymbol{w}(t^*+\alpha)-\boldsymbol{w}(t^*)\|^2\\
=&\mathcal{L}(\boldsymbol{w}(t^*)) + \alpha\langle \nabla\mathcal{L}(\boldsymbol{w}(t^*)), \boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\rangle + \frac{H_{s_0}\alpha^2\sigma_{max}^2}{2N}\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
\overset{(*)}{=}&\mathcal{L}(\boldsymbol{w}(t^*)) + \alpha\left\langle \frac{1}{(1-\beta)\eta}(\beta(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))-(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))), \boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\right\rangle\\
&+\frac{H_{s_0}\alpha^2\sigma_{max}^2}{2N}\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
=&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\alpha\beta}{(1-\beta)\eta}\langle(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1), \boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\rangle + \left(\frac{H_{s_0}\alpha^2\sigma_{max}}{2N}-\frac{\alpha}{(1-\beta)\eta}\right)\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
\overset{(**)}{\leq}&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\alpha\beta}{2(1-\beta)\eta}\|(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2 + \frac{\alpha\beta}{2(1-\beta)\eta}\|(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
&+\left(\frac{H_{s_0}\alpha^2\sigma_{max}^2}{2N}-\frac{\alpha}{(1-\beta)\eta}\right)\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
=&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\alpha\beta}{2(1-\beta)\eta}\|(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2 + \left(\frac{\alpha\beta}{2(1-\beta)\eta}-\frac{\alpha}{(1-\beta)\eta}+\frac{H_{s_0}\alpha^2\sigma_{max}^2}{2N}\right)\|(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
=&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\beta}{2(1-\beta)\eta}\|(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2 - \frac{(1-\alpha)\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2\\
&+\left(\frac{\alpha\beta}{2(1-\beta)\eta}-\frac{\alpha}{(1-\beta)\eta}+\frac{H_{s_0}\alpha^2\sigma_{max}^2}{2N}\right)\|(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
\overset{(\diamond)}{\leq}&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2 - \frac{\beta}{2(1-\beta)\eta}\alpha^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
&-\frac{(1-C_1)\alpha^2}{\eta}\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2,
\end{aligned} \tag{22}$$

where Eq. $(*)$ is due to a simple rearrangement of the update rule of gradient descent with momentum (Eq. (1)), i.e.,

$$\nabla\mathcal{L}(\boldsymbol{w}(t)) = \frac{1}{(1-\beta)\eta}(\beta(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))-(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))), \forall t \geq 1, \tag{23}$$

Inequality $(**)$ is due to Cauchy Schwarz's inequality and arithmetic-geometric average inequality, and Inequality $(\diamond)$ is due to

$$\begin{aligned}
&-\frac{(1-\alpha)\beta}{2(1-\beta)\eta}\|(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2 + \left(\frac{\alpha\beta}{2(1-\beta)\eta}-\frac{\alpha}{(1-\beta)\eta}+\frac{H_{s_0}\alpha^2\sigma_{max}^2}{2N}\right)\|(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
=&-\frac{(1-\alpha)\beta}{2(1-\beta)\eta}\|(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2 + \mathcal{O}(\alpha)\\
\leq&-\frac{\beta}{2(1-\beta)\eta}\alpha^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2 - \frac{(1-C_1)\alpha^2}{\eta}\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2.
\end{aligned}$$

Here the inequality is due to that $-\frac{(1-\alpha)\beta}{2\eta}\|(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2$ tend to $-\frac{\beta}{2\eta}\|(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1)\|^2$ as $\alpha$ tend to zero, which is a negative constant by Lemma 9, and $-\frac{\beta}{2(1-\beta)\eta}\alpha^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2 - \frac{(1-C_1)\alpha^2}{\eta}\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2$ is $\mathcal{O}(\alpha^2)$.

Eq. (22) indicates Eq. (21) holds at $(t^*, \alpha)$ for $\alpha > 0$ is small enough, which contradicts to $\alpha^* = 0$.

**Case 2:** $\alpha^* \neq 0$: Same as **Case 1**, we have for any $1 \leq t < t^*$,

$$\mathcal{L}(\boldsymbol{w}(t)) + \frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t) - \boldsymbol{w}(t-1)\|^2 \geq \mathcal{L}(\boldsymbol{w}(t+1)) + \frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|^2,$$

which further leads to

$$\mathcal{L}(\boldsymbol{w}(1)) \geq \mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\beta}{2\eta}\|\boldsymbol{w}(t^*) - \boldsymbol{w}(t^*-1)\|^2. \tag{24}$$

On the other hand, by the definition of $\alpha^*$, we have for any $0 \leq \alpha < \alpha^*$, we have Eq. (21) holds for $(t^*, \alpha)$, which by continuity further leads to Eq. (21) holds for $(t^*, \alpha^*)$. Therefore, $\alpha^* < 1$, otherwise, Eq. (21) holds for $(t^*, \alpha)$, $\forall \alpha \in [0, 1]$ which contradicts the definition of $t^*$.

Combining Eq. (21) with $(t^*, \alpha)$ and Eq. (24), we further have

$$\mathcal{L}(\boldsymbol{w}(1)) \geq \mathcal{L}(\boldsymbol{w}(t^*+\alpha)) + \frac{\beta}{2(1-\beta)\eta}\alpha^2\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|^2 + \frac{1-C_1}{2C_1}\alpha^2\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|^2,$$

Consequently, for any $\alpha \in [0, \alpha^*]$

$$\mathcal{L}(\boldsymbol{w}(1)) \geq \mathcal{L}(\boldsymbol{w}(t^*+\alpha)),$$

and by Lemma 6, we then have $\mathcal{L}$ is $\frac{H_{s_0}\sigma_{max}^2}{N}$ smooth at $\boldsymbol{w}(t^*+\alpha)$, which further by Taylor's expansion leads to

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}(t^*+\alpha^*)) \\
\leq&\mathcal{L}(\boldsymbol{w}(t^*)) + \langle \nabla\mathcal{L}(\boldsymbol{w}(t^*)), \boldsymbol{w}(t^*+\alpha^*) - \boldsymbol{w}(t^*)\rangle + \frac{H_{s_0}\sigma_{max}^2}{2N}\|\boldsymbol{w}(t^*+\alpha^*) - \boldsymbol{w}(t^*)\|^2 \\
\overset{(\circ)}{\leq}&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\alpha^*\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*) - \boldsymbol{w}(t^*-1)\|^2 + \frac{\alpha^*\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*+1) - \boldsymbol{w}(t^*)\|^2 \\
&+\left(\frac{H_{s_0}(\alpha^*)^2\sigma_{max}^2}{2N} - \frac{\alpha^*}{(1-\beta)\eta}\right)\|\boldsymbol{w}(t^*+1) - \boldsymbol{w}(t^*)\|^2 \\
=&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\alpha^*\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*) - \boldsymbol{w}(t^*-1)\|^2 + \left(\frac{H_{s_0}(\alpha^*)^2\sigma_{max}^2}{2N} - \frac{\alpha^*(2-\beta)}{2(1-\beta)\eta}\right)\|\boldsymbol{w}(t^*+1) - \boldsymbol{w}(t^*)\|^2 \\
\overset{(\bullet)}{=}&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\alpha^*\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*) - \boldsymbol{w}(t^*-1)\|^2 + \left(\frac{C_1(\alpha^*)^2}{\eta} - \frac{\alpha^*(2-\beta)}{2(1-\beta)\eta}\right)\|\boldsymbol{w}(t^*+1) - \boldsymbol{w}(t^*)\|^2 \\
\overset{(*)}{<}&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*) - \boldsymbol{w}(t^*-1)\|^2 - \frac{(\alpha^*)^2\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t^*+1) - \boldsymbol{w}(t^*)\|^2 \\
&-\frac{(1-C_1)(\alpha^*)^2}{\eta}\|\boldsymbol{w}(t^*+1) - \boldsymbol{w}(t^*)\|^2
\end{aligned}$$

where Eq. $(\circ)$ follows the same routine as **Case 1**, Eq. $(\bullet)$ is due to the definition of $\eta$ and $C_1$, and Eq. $(*)$ is due to $\alpha^* < 1$, and $\|\boldsymbol{w}(t^*+1) - \boldsymbol{w}(t^*)\|^2 > 0$ (given by Lemma 9).

By the continuity of $\mathcal{L}$, for any small enough $\delta > 0$, Eq. (21) holds for $(t^*, \alpha^* + \delta)$, which contradicts to the definition of $\alpha^*$.

The proof is completed. □

By Lemma 1, one can easily obtain the sum of the squared norms of the updates across the trajectory converges.

**Corollary 3.** *Let all conditions in Theorem 2 hold. We have*

$$\sum_{t=1}^{\infty}\|\boldsymbol{w}(t+1) - \boldsymbol{w}(t)\|^2 < \infty. \tag{25}$$

*Consequentially, we have*

$$\|\boldsymbol{w}(t)\| = \mathcal{O}(\sqrt{t}).$$

*Proof.* By Lemma 1, we have

$$\mathcal{L}(\boldsymbol{w}(t))+\frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|^2-\left(\mathcal{L}(\boldsymbol{w}(t+1))+\frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\right)$$
$$\geq\frac{1-C_1}{\eta}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2,$$

which by summing over $t$ further leads to

$$\mathcal{L}(\boldsymbol{w}(1))\geq\mathcal{L}(\boldsymbol{w}(1))-\left(\mathcal{L}(\boldsymbol{w}(t+1))+\frac{\beta}{2(1-\beta)\eta}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\right)\geq\frac{1-C_1}{\eta}\sum_{s=1}^{t}\|\boldsymbol{w}(s+1)-\boldsymbol{w}(s)\|^2.$$

Taking $t\to\infty$ leads to

$$\sum_{s=1}^{\infty}\|\boldsymbol{w}(s+1)-\boldsymbol{w}(s)\|^2<\infty.$$

By triangle inequality, we further have

$$\begin{aligned}\|\boldsymbol{w}(t)\|&\leq\sum_{s=1}^{t}\|\boldsymbol{w}(s+1)-\boldsymbol{w}(s)\|+\|\boldsymbol{w}(1)\|\\&\overset{(\star)}{\leq}\sqrt{t\left(\sum_{s=1}^{t}\|\boldsymbol{w}(s+1)-\boldsymbol{w}(s)\|^2\right)}+\|\boldsymbol{w}(1)\|=\mathcal{O}(\sqrt{t}),\end{aligned}$$

where Eq. $(\star)$ is due to Cauchy-Schwartz's inequality.

The proof is completed. □

By the negative derivative of the loss and the separable data, we can finally prove the sum of squared gradient converges.

**Corollary 4.** *Let all conditions in Theorem 2 hold. We have,* $\sum_{t=1}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2<\infty$.

*Proof.* By Eq. (20), we have

$$\begin{aligned}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2=&\eta^2\left\|(1-\beta)\sum_{s=1}^{t}\beta^{t-s}\frac{\sum_{i=1}^{N}\ell'(\langle\boldsymbol{w}(s),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i}{N}\right\|^2\\=&\eta^2(1-\beta)^2\left\|\sum_{s=1}^{t}\beta^{t-s}\frac{\sum_{i=1}^{N}\ell'(\langle\boldsymbol{w}(s),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i}{N}\right\|^2\frac{\|\hat{\boldsymbol{w}}\|^2}{\|\hat{\boldsymbol{w}}\|^2}\\\overset{(*)}{\geq}&\eta^2\gamma^2(1-\beta)\left\langle\hat{\boldsymbol{w}},\sum_{s=1}^{t}\beta^{t-s}\frac{\sum_{i=1}^{N}\ell'(\langle\boldsymbol{w}(s),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i}{N}\right\rangle^2\\\overset{(**)}{\geq}&\eta^2\gamma^2(1-\beta)^2\left(\sum_{s=1}^{t}\beta^{t-s}\frac{\sum_{i=1}^{N}\ell'(\langle\boldsymbol{w}(s),\boldsymbol{x}_i\rangle)}{N}\right)^2\\\geq&\eta^2\gamma^2(1-\beta)^2\left(\frac{\sum_{i=1}^{N}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)}{N}\right)^2\\\overset{(\bullet)}{\geq}&\frac{\eta^2\gamma^2(1-\beta)^2}{\sigma_{max}^2}\left\|\frac{\sum_{i=1}^{N}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i}{N}\right\|^2\\=&\frac{\eta^2\gamma^2(1-\beta)^2}{\sigma_{max}^2}\left\|\nabla\mathcal{L}(\boldsymbol{w}(t))\right\|^2,\end{aligned}\tag{26}$$

where Inequality $(*)$ is due to Cauchy-Schwartz's inequality, Inequality $(**)$ is due to $\ell'(s)<0$, $\forall s\in\mathbb{R}$ and $\langle\hat{\boldsymbol{w}},\boldsymbol{x}_i\rangle\geq\gamma$, $\forall i\in[N]$, and Inequality $(\bullet)$ is due to the definition of $\sigma_{max}$. By combining Eq. (25) and Eq. (26), we complete the proof. □

By the exponential-tailed assumption of the loss (Assumption 2), we further have the following corollary.

**Corollary 5.** *Let all conditions in Theorem 2 hold. Then,* $\lim_{t\to\infty} \|\nabla\mathcal{L}(\boldsymbol{w}(t))\| = 0$, *and*

$$\lim_{t\to\infty} \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle = \infty, \forall i.$$

*Consequently, there exists an large enough time* $t_0$*, such that,* $\forall t > t_0$, $\forall i$*, we have* $\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle > 0$*, and*

$$-\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) \le (1 + e^{-\mu_+ \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle},$$
$$-\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) \ge (1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}.$$

### C.1.2 Parameter dynamics

To prove Theorem 2, we only need to show $\boldsymbol{w}(t) - \ln(t)\hat{\boldsymbol{w}}$ $(t \ge 1)$ has bounded norm for any iteration $t > 0$. Letting $C_3 = \frac{\eta}{N}$ in Corollary 1, we obtain an constant vector $\tilde{\boldsymbol{w}}$ satisfying Eq. (10). Define

$$\boldsymbol{r}(t) \overset{\triangle}{=} \boldsymbol{w}(t) - \ln(t)\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}}. \tag{27}$$

As $\tilde{\boldsymbol{w}}$ is a constant vector, that $\boldsymbol{w}(t) - \ln(t)\hat{\boldsymbol{w}}$ $(t \ge 1)$ has bounded norm is equivalent to $\boldsymbol{r}(t)$ has bounded norm. As discussed in the main body of the paper, we then propose an equivalent proposition of $\|\boldsymbol{r}(t)\|$ is bounded, and further prove this proposition is fulfilled. Specifically, we have

**Lemma 11.** *Let all conditions in Theorem 2 hold. Then,* $\|\boldsymbol{r}(t)\|$ *is bounded if and only if the function* $g(t)$ *is upper bounded, where* $g : \mathbb{Z}^+ \to \mathbb{R}$ *is defined as*

$$g(t) \overset{\triangle}{=} \frac{1}{2}\|\boldsymbol{r}(t)\|^2 + \frac{\beta}{1-\beta}\langle \boldsymbol{r}(t), \boldsymbol{w}(t) - \boldsymbol{w}(t-1)\rangle - \frac{\beta}{1-\beta}\sum_{\tau=2}^{t}\langle \boldsymbol{r}(\tau) - \boldsymbol{r}(\tau-1), \boldsymbol{w}(\tau) - \boldsymbol{w}(\tau-1)\rangle. \tag{28}$$

*Furthermore, for almost every dataset, we have* $\sum_{t=1}^{\infty}(g(t+1) - g(t))$ *is upper bounded.*

As the proof is rather complex, we separate it into two sub-lemmas. We first prove $\|\boldsymbol{r}(t)\|$ is bounded if and only if function $g(t)$ is upper bounded.

**Lemma 12** (First argument in Lemma 11)**.** *Let all conditions in Theorem 2 hold. Then,* $\|\boldsymbol{r}(t)\|$ *is bounded if and only if function* $g(t)$ *is upper bounded.*

*Proof.* We start the proof by showing that $A_1(t) \overset{\triangle}{=} \sum_{\tau=2}^{t}\langle \boldsymbol{r}(\tau) - \boldsymbol{r}(\tau-1), \boldsymbol{w}(\tau) - \boldsymbol{w}(\tau-1)\rangle$ has bounded absolute value.

By the definition of $\boldsymbol{r}(t)$ , we have

$$\boldsymbol{r}(t) - \boldsymbol{r}(t-1) = \boldsymbol{w}(t) - \boldsymbol{w}(t-1) - \ln\left(\frac{t}{t-1}\right)\hat{\boldsymbol{w}},$$

which further indicates

$$A_1(t) = \sum_{\tau=2}^{t}\left\langle \boldsymbol{w}(\tau) - \boldsymbol{w}(\tau-1) - \ln\left(\frac{\tau}{\tau-1}\right)\hat{\boldsymbol{w}}, \boldsymbol{w}(\tau) - \boldsymbol{w}(\tau-1)\right\rangle.$$

Therefore, the absolute value of $A_1(t)$ can be bounded as

$$\begin{aligned}|A_1(t)| =&\left|\sum_{\tau=2}^{t}\left\langle \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)-\ln\left(\frac{\tau}{\tau-1}\right)\hat{\boldsymbol{w}}, \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\right\rangle\right|\\
\leq&\sum_{\tau=2}^{t}\left|\left\langle \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)-\ln\left(\frac{\tau}{\tau-1}\right)\hat{\boldsymbol{w}}, \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\right\rangle\right|\\
\leq&\sum_{\tau=2}^{t}\|\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\|^2+\sum_{\tau=2}^{t}\left|\left\langle \ln\left(\frac{\tau}{\tau-1}\right)\hat{\boldsymbol{w}}, \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\right\rangle\right|\\
\leq&\sum_{\tau=2}^{t}\|\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\|^2+\sum_{\tau=2}^{t}\left\|\ln\left(\frac{\tau}{\tau-1}\right)\hat{\boldsymbol{w}}\right\|\|\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\|\\
\overset{(\star)}{\leq}&\frac{3}{2}\sum_{\tau=2}^{t}\|\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\|^2+\frac{1}{2}\sum_{\tau=2}^{t}\left\|\ln\left(\frac{\tau}{\tau-1}\right)\hat{\boldsymbol{w}}\right\|^2\\
\overset{(\circ)}{<}&\infty,\end{aligned}$$

where Inequality $(\star)$ is due to the Inequality of arithmetic and geometric means, and Inequality $(\circ)$ is due to Corollary 3 and $\ln\frac{\tau}{\tau-1}=\boldsymbol{O}(\frac{1}{\tau})$.

Therefore, $g(t)$ is upper bounded is then equivalent to $\frac{1}{2}\|\boldsymbol{r}(t)\|^2+\frac{\beta}{1-\beta}\langle\boldsymbol{r}(t),\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\rangle$ is upper bounded. Now if $\frac{1}{2}\|\boldsymbol{r}(t)\|^2+\frac{\beta}{1-\beta}\langle\boldsymbol{r}(t),\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\rangle$ is upper bounded, we will prove $\|\boldsymbol{r}(t)\|$ is bounded by reduction to absurdity.

Suppose that $\|\boldsymbol{r}(t)\|$ has unbounded norm. By Corollary 3, we have $\lim_{t\to\infty}\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|=0$, and there exists a large enough time $T$, such that $\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|<1$ for any $t\geq T$. On the other hand, since $\boldsymbol{r}(t)$ is unbounded from above, there exists an increasing time sequence $k_i>T$, $i\in\mathbb{Z}^+$, such that

$$\lim_{i\to\infty}\|\boldsymbol{r}(k_i)\|=\infty.$$

Therefore, we have

$$\begin{aligned}&\varliminf_{i\to\infty}\frac{1}{2}\|\boldsymbol{r}(k_i)\|^2+\frac{\beta}{1-\beta}\langle\boldsymbol{r}(k_i),\boldsymbol{w}(k_i)-\boldsymbol{w}(k_i-1)\rangle\\
\geq&\varliminf_{i\to\infty}\frac{1}{2}\|\boldsymbol{r}(k_i)\|^2-\frac{\beta}{1-\beta}\|\boldsymbol{r}(k_i)\|\|\boldsymbol{w}(k_i)-\boldsymbol{w}(k_i-1)\|\\
\geq&\varliminf_{i\to\infty}\frac{1}{2}\|\boldsymbol{r}(k_i)\|^2-\frac{\beta}{1-\beta}\|\boldsymbol{r}(k_i)\|=\infty,\end{aligned}$$

which leads to contradictory, and completes the proof of necessity.

On the other hand, if $\|\boldsymbol{r}(t)\|$ is upper bounded, since $\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|$ is also upper bounded, we have $\frac{1}{2}\|\boldsymbol{r}(t)\|^2+\frac{\beta}{1-\beta}\langle\boldsymbol{r}(t),\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\rangle$ is upper bounded, which completes the proof of sufficiency.

The proof is completed. □

Therefore, the last piece of this puzzle is to prove $g(t)$ is upper bounded $\forall t>0$.

**Lemma 13** (Second argument in Lemma 11)**.** *Let all conditions in Theorem 2 hold. Then, for almost every dataset, we have that $g(t)$ is upper bounded.*

*Proof.* We start the proof by calculating $g(t+1)-g(t)$. For any $t \geq 2$, we have

$$\begin{aligned} g(t+1)-g(t) =& \frac{1}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2 + \langle \boldsymbol{r}(t), \boldsymbol{r}(t+1)-\boldsymbol{r}(t)\rangle + \frac{\beta}{1-\beta}\langle \boldsymbol{r}(t+1), \boldsymbol{w}(t+1)-\boldsymbol{w}(t)\rangle \\ &-\frac{\beta}{1-\beta}\langle \boldsymbol{r}(t), \boldsymbol{w}(t)-\boldsymbol{w}(t-1)\rangle - \frac{\beta}{1-\beta}\langle \boldsymbol{r}(t+1)-\boldsymbol{r}(t), \boldsymbol{w}(t+1)-\boldsymbol{w}(t)\rangle \\ =& \frac{1}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2 + \langle \boldsymbol{r}(t), \boldsymbol{r}(t+1)-\boldsymbol{r}(t)\rangle + \frac{\beta}{1-\beta}\langle \boldsymbol{r}(t), \boldsymbol{w}(t+1)+\boldsymbol{w}(t-1)-2\boldsymbol{w}(t)\rangle. \end{aligned}$$

On the other hand, by simply rearranging the update rule Eq. (1), we have

$$\frac{\beta}{1-\beta}(\boldsymbol{w}(t+1)+\boldsymbol{w}(t-1)-2\boldsymbol{w}(t)) = -\eta\nabla\mathcal{L}(\boldsymbol{w}(t)) - (\boldsymbol{w}(t+1)-\boldsymbol{w}(t)), \tag{29}$$

which further indicates

$$\begin{aligned} & g(t+1)-g(t) \\ =& \frac{1}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2 + \langle \boldsymbol{r}(t), \boldsymbol{r}(t+1)-\boldsymbol{r}(t)\rangle + \langle \boldsymbol{r}(t), -\eta\nabla\mathcal{L}(\boldsymbol{w}(t)) - (\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\rangle \\ =& \frac{1}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2 + \left\langle \boldsymbol{r}(t), -\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}} - \eta\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle. \end{aligned}$$

Denote $A_2(t) = \|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2$, and $A_3(t) = \left\langle \boldsymbol{r}(t), -\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}} - \eta\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle$. We then prove respectively $\sum_{t=1}^{\infty} A_2(t)$ and $\sum_{t=1}^{\infty} A_3(t)$ are upper bounded.

First of all, by definition of $\boldsymbol{r}(t)$ Eq.(27), we have

$$\begin{aligned} \sum_{t=1}^{\infty} A_2(t) =& \sum_{t=1}^{\infty}\left(\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2 + \ln\left(\frac{t+1}{t}\right)^2\|\hat{\boldsymbol{w}}\|^2 - 2\ln\left(\frac{t+1}{t}\right)\langle \boldsymbol{w}(t+1)-\boldsymbol{w}(t), \hat{\boldsymbol{w}}\rangle\right) \\ \leq& 2\sum_{t=1}^{\infty}\left(\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2 + \ln\left(\frac{t+1}{t}\right)^2\|\hat{\boldsymbol{w}}\|^2\right) \overset{(\bullet)}{<} \infty, \end{aligned} \tag{30}$$

where Eq. $(\bullet)$ is due to Lemma 3 and $\ln\left(\frac{t+1}{t}\right) = \mathcal{O}(\frac{1}{t})$.

Then we only need to prove $\sum_{t=1}^{\infty} A_3(t) < \infty$.

To begin with, by adding one additional term $\frac{1}{t}\hat{\boldsymbol{w}}$ into $A_3$, we have

$$A_3(t) = \left\langle \boldsymbol{r}(t), \frac{1}{t}\hat{\boldsymbol{w}} - \ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}}\right\rangle + \left\langle \boldsymbol{r}(t), -\frac{1}{t}\hat{\boldsymbol{w}} - \eta\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle.$$

On the one hand, by Corollary 3, $\|\boldsymbol{w}(t)\| = \mathcal{O}(\sqrt{t})$, which further leads to

$$\|\boldsymbol{r}(t)\| = \|\boldsymbol{w}(t)\| + \ln(t)\|\hat{\boldsymbol{w}}\| + \|\hat{\boldsymbol{w}}\| = \mathcal{O}(\sqrt{t})$$

By $\frac{1}{t} - \ln\frac{t+1}{t} = \mathcal{O}\left(\frac{1}{t^2}\right)$, we have

$$\left\langle \boldsymbol{r}(t), \frac{1}{t}\hat{\boldsymbol{w}} - \ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}}\right\rangle = \mathcal{O}\left(\frac{1}{t^{\frac{3}{2}}}\right). \tag{31}$$

On the other hand, by direct calculation of the gradient, we have

$$\begin{aligned} & \left\langle \boldsymbol{r}(t), -\frac{1}{t}\hat{\boldsymbol{w}} - \eta\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle \\ =& \left\langle \boldsymbol{r}(t), -\frac{1}{t}\hat{\boldsymbol{w}} - \frac{\eta}{N}\sum_{i=1}^{N}\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle)\boldsymbol{x}_i\right\rangle \\ \overset{(\star)}{=}& \frac{1}{N}\left\langle \boldsymbol{r}(t), -\frac{1}{t}\eta\sum_{\boldsymbol{x}_i\in\boldsymbol{S}_s} e^{-\langle\tilde{\boldsymbol{w}}, \boldsymbol{x}_i\rangle}\boldsymbol{x}_i - \eta\sum_{i=1}^{N}\ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle)\boldsymbol{x}_i\right\rangle \\ =& \frac{1}{N}\left\langle \boldsymbol{r}(t), -\eta\sum_{\boldsymbol{x}_i\in\boldsymbol{S}_s}\left(\frac{1}{t}e^{-\langle\tilde{\boldsymbol{w}}, \boldsymbol{x}_i\rangle} + \ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle)\right)\boldsymbol{x}_i\right\rangle - \frac{1}{N}\left\langle \boldsymbol{r}(t), \eta\sum_{\boldsymbol{x}_i\notin\boldsymbol{S}_s}\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle)\boldsymbol{x}_i\right\rangle, \end{aligned}$$

where Eq. ($\star$) is due to the definition of $\tilde{\boldsymbol{w}}$ (Eq. (10) with $C_3 = \eta/N$).

Denote

$$A_4(t) = -\left\langle \boldsymbol{r}(t), \eta \sum_{\boldsymbol{x}_i \notin \boldsymbol{S}_s} \ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle)\boldsymbol{x}_i \right\rangle,$$

and

$$A_5(t) = \left\langle \boldsymbol{r}(t), -\eta \sum_{\boldsymbol{x}_i \in \boldsymbol{S}_s} \left( \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) \right) \boldsymbol{x}_i \right\rangle.$$

We then analysis these two terms respectively. As for $A_4(t)$, due to $\ell' < 0$, we have

$$A_4(t) \le -\eta \left\langle \boldsymbol{r}(t), \sum_{\boldsymbol{x}_i \notin \boldsymbol{S}_s, \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle > 0} \ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle)\boldsymbol{x}_i \right\rangle.$$

By Corollary 5, we further have $\forall t > t_0$

$$-\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) \le (1 + e^{-\mu_+ \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \le 2e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle},$$

which further indicates

$$\begin{aligned}
A_4(t) \le & -\eta \sum_{\boldsymbol{x}_i \notin \boldsymbol{S}_s, \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle > 0} \ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\le & \eta \sum_{\boldsymbol{x}_i \notin \boldsymbol{S}_s, \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle > 0} 2e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
= & \eta \sum_{\boldsymbol{x}_i \notin \boldsymbol{S}_s, \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle > 0} 2e^{-\langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\le & \eta \left( \max_i e^{\langle -\tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) \sum_{\boldsymbol{x}_i \notin \boldsymbol{S}_s, \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle > 0} 2e^{-\langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\overset{(\circ)}{\le} & \eta \frac{\left( \max_i e^{\langle -\tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right)}{t^{\theta}} \sum_{\boldsymbol{x}_i \notin \boldsymbol{S}_s, \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle > 0} 2e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\overset{(\diamond)}{\le} & \eta \frac{\left( \max_i e^{\langle -\tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right)}{t^{\theta}} 2N,
\end{aligned}$$

where $\theta$ in Eq. ($\circ$) is defined as

$$\theta = \min_{\boldsymbol{x}_i \notin \boldsymbol{S}_s} \langle \boldsymbol{x}_i, \hat{\boldsymbol{w}} \rangle > 1. \tag{32}$$

As $\sum_{t=1}^{\infty} \frac{1}{t^{\theta}} < \infty$, we have

$$\sum_{t=1}^{\infty} A_4(t) < \infty\text{[6]}. \tag{33}$$

For each term $\left\langle \boldsymbol{r}(t), -\eta \left( \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \right) \boldsymbol{x}_i \right\rangle$ ($\boldsymbol{x}_i \notin \boldsymbol{S}_s$) in $A_5(t)$, we divide the analysis into two parts depending on the sign of $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle$.

**Case 1:** $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \ge 0$**.** By Corollary 5, we have

$$\begin{aligned}
& \left\langle \boldsymbol{r}(t), -\eta \left( \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \right) \boldsymbol{x}_i \right\rangle \\
= & -\eta \left( \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\le & \eta \left( -\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + (1 + e^{-\mu_+ \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\overset{(\diamond)}{=} & \eta \left( -\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + (1 + e^{-\mu_+ \langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle,
\end{aligned}$$

[6] In this paper, for a real series $\{r_i\}_{i=1}^{\infty}$, we use $\sum_{i=1}^{\infty} r_i < \infty$ representing $\sum_{i=1}^{T} r_i$ is uniformly upper bounded for any $T$.

where Eq. ($\diamond$) is due to the definition of $\boldsymbol{r}(t)$ (Eq. (27).

Since $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \geq 0$, we further have

$$\begin{aligned}
&\left\langle \boldsymbol{r}(t), -\eta \left( \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \right) \boldsymbol{x}_i \right\rangle \\
\leq & \eta \left( -\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + (1 + e^{-\mu_+ \langle \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\overset{(\square)}{=} & \eta \left( -\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \frac{1}{t} (1 + t^{-\mu_+} e^{-\mu_+ \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{r}(t) + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
= & \eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + (1 + t^{-\mu_+} e^{-\mu_+ \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle ,
\end{aligned}$$

where Eq. ($\square$) is due to $\langle \hat{\boldsymbol{w}}, \boldsymbol{x}_i \rangle = 1$, $\forall \boldsymbol{x}_i \in \boldsymbol{S}_s$.

Specifically,

$$\begin{aligned}
& -1 + (1 + t^{-\mu_+} e^{-\mu_+ \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \\
= & -1 + e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} + t^{-\mu_+} e^{-\mu_+ \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \\
\leq & t^{-\mu_+} e^{-\mu_+ \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + (1 + t^{-\mu_+} e^{-\mu_+ \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq & \eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( t^{-\mu_+} e^{-\mu_+ \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq & \frac{\eta}{(1-\beta)e} \frac{1}{t^{1+\mu_+}} e^{-(1+\mu_+) \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} = \mathcal{O}\left( \frac{1}{t^{1+\mu_+}} \right).
\end{aligned}$$

**Case 2:** $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < 0$**.** Similar to **Case 1.**, in this case we have

$$\begin{aligned}
&\left\langle \boldsymbol{r}(t), -\eta \left( \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \right) \boldsymbol{x}_i \right\rangle \\
\leq & \eta \left( -\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \left( 1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
= & \eta \left( -\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \left( 1 - e^{-\mu_- \langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
= & \eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle .
\end{aligned}$$

Specifically, if $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \geq -t^{-0.5\mu_-}$,

$$\begin{aligned}
& \left| \eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + (1 - e^{-\mu_- \langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \right| \\
= & \left| \eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + (1 - t^{-\mu_-} e^{-\mu_- \langle \boldsymbol{r}(t) + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \right| \\
\leq & \eta \frac{1}{t^{1+0.5\mu_-}} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left| -1 + \left( 1 - t^{-\mu_-} e^{-\mu_- \langle \boldsymbol{r}(t) + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right| \\
\overset{(\dagger)}{=} & \mathcal{O}\left( \frac{1}{t^{1+0.5\mu_-}} \right),
\end{aligned}$$

where Eq. ($\dagger$) is due to if $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \geq -t^{-0.5\mu_-}$,

$$\lim_{t \to \infty} \left| -1 + \left( 1 - t^{-\mu_-} e^{-\mu_- \langle \boldsymbol{r}(t) + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right| = 0.$$

If $-2 \leq \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < -t^{-0.5\mu_-}$, we have

$$\begin{aligned}
&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
=&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{1}{t^{\mu_-}} e^{-\mu_- \langle \boldsymbol{r}(t) + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{2\mu_-}}{t^{\mu_-}} e^{-\mu_- \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle .
\end{aligned}$$

Therefore, when $t$ is large enough, $1 - \frac{e^{2\mu_-}}{t^{\mu_-}} e^{-\mu_- \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} > 0$, which by $e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \geq 1 - \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle$ leads to

$$\begin{aligned}
&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{2\mu_-}}{t^{\mu_-}} e^{-\mu_- \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{2\mu_-}}{t^{\mu_-}} e^{-\mu_- \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) (1 - \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle) \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{2\mu_-}}{t^{\mu_-}} e^{-\mu_- \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) \left( 1 + \frac{1}{t^{0.5\mu_-}} \right) \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
=&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( \frac{1}{t^{0.5\mu_-}} + \boldsymbol{o} \left( \frac{1}{t^{0.5\mu_-}} \right) \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < 0.
\end{aligned}$$

If $-2 > \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle$,

$$\begin{aligned}
&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{r}(t) + \ln t \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
=&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle .
\end{aligned}$$

For large enough $t$, $1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} > \frac{1}{2}$, and

$$\begin{aligned}
&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \right) e^{2} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq&\eta \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \frac{e^2}{2} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < 0.
\end{aligned}$$

Therefore, in **Case 2.**, for large enough $t$, we have

$$\left\langle \boldsymbol{r}(t), -\eta \left( \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \ell'(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) \right) \boldsymbol{x}_i \right\rangle \leq \mathcal{O} \left( \frac{1}{t^{1+0.5\mu_-}} \right).$$

Combining **Case 1.** and **Case 2.**, we conclude that

$$A_5(t) \leq \mathcal{O} \left( \frac{1}{t^{1+0.5\mu_+}} \right),$$

which further yields

$$\sum_{t=1}^{\infty} A_5(t) < \infty. \tag{34}$$

Combining Eq. (33) and Eq. (34), we conclude that $\sum_{t=1}^{\infty} A_3(t) < \infty$, which together with Eq. (30) yields $\sum_{t=2}^{\infty} g(t+1) - g(t) < \infty$, and completes the proof. □

We are now ready to prove Theorem 2.

*Proof of Theorem 2.* By Lemma 13, we have $g(t)$ is upper bounded. Therefore, by Lemma 11, we have $\|\boldsymbol{r}(t)\|$ is bounded, which further indicates $\|\boldsymbol{w}(t)-\ln(t)\hat{\boldsymbol{w}}\|$ is bounded.

Therefore, the direction of $\boldsymbol{w}(t)$ can be calculated as

$$\begin{aligned}\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}=&\frac{\ln(t)\hat{\boldsymbol{w}}}{\|\boldsymbol{w}(t)\|}+\frac{\boldsymbol{w}(t)-\ln(t)\hat{\boldsymbol{w}}}{\|\boldsymbol{w}(t)\|}=\frac{\ln(t)\hat{\boldsymbol{w}}}{\|\ln(t)\hat{\boldsymbol{w}}+\boldsymbol{w}(t)-\ln(t)\hat{\boldsymbol{w}}\|}+\frac{\boldsymbol{w}(t)-\ln(t)\hat{\boldsymbol{w}}}{\|\boldsymbol{w}(t)\|}\\=&\frac{\hat{\boldsymbol{w}}}{\left\|\hat{\boldsymbol{w}}+\frac{\boldsymbol{w}(t)-\ln(t)\hat{\boldsymbol{w}}}{\ln t}\right\|}+\frac{\boldsymbol{w}(t)-\ln(t)\hat{\boldsymbol{w}}}{\|\boldsymbol{w}(t)\|}\to\frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}\ (as\, t\to\infty).\end{aligned}$$

The proof is completed.

□

### C.2 Implicit regularization of SGDM

This section collects the proof of Theorem 3. Following the same framework as Appendix C.1, we will first prove that the sum of the squared gradient norms along the trajectory is finite. One may expect $\mathcal{L}(\boldsymbol{w}(t))+\frac{\beta}{2\eta}\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|^2$ is a Lyapunov function of SGDM. However, due to the randomness of the update rule of SGDM, $\mathcal{L}(\boldsymbol{w}(t))+\frac{\beta}{2\eta}\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|^2$ may no longer decrease (we will show this in the end of Appendix C.2, please see Appendix C.2.3 for explanation).

#### C.2.1 Loss dynamics

Recall that in the main text, we define $\boldsymbol{u}(t)$ as

$$\boldsymbol{u}(t)=\frac{\boldsymbol{w}(t)-\beta\boldsymbol{w}(t-1)}{1-\beta},\tag{35}$$

where the update of $\boldsymbol{u}(t)$ is given by $\boldsymbol{u}(t+1)=\boldsymbol{u}(t)-\eta\nabla\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))$. We then prove Theorem 4.

*Proof of Theorem 4.* By the bounded smoothness assumption, we have that

$$\mathcal{L}(\boldsymbol{u}(t+1))\le\mathcal{L}(\boldsymbol{u}(t))+\langle\boldsymbol{u}(t+1)-\boldsymbol{u}(t),\nabla\mathcal{L}(\boldsymbol{u}(t))\rangle+\frac{L}{2}\|\boldsymbol{u}(t+1)-\boldsymbol{u}(t)\|^2,$$

which by Eq. (35) leads to

$$\mathcal{L}(\boldsymbol{u}(t+1))\le\mathcal{L}(\boldsymbol{u}(t))-\eta\langle\nabla\mathcal{L}_{\xi(t)}(\boldsymbol{w}(t)),\nabla\mathcal{L}(\boldsymbol{u}(t))\rangle+\frac{L\eta^2}{2}\|\nabla\mathcal{L}_{\xi(t)}(\boldsymbol{w}(t))\|^2.\tag{36}$$

Taking the expectation of Eq. (36) with respect to $\boldsymbol{w}(t+1)$ conditioning on $\mathcal{F}_t$ (recall that $\mathcal{F}_t$ is the sub-sigma algebra over the mini-batch sampling, such that $\forall t\in\mathbb{N}$, $\boldsymbol{w}(t)$ is adapted with respect to the sigma algebra flow $\mathcal{F}_t$), we have

$$\begin{aligned}&\mathbb{E}[\mathcal{L}(\boldsymbol{u}(t+1))|\mathcal{F}_t]\\\overset{(\star)}{=}&\mathbb{E}_{\xi(t)}[\mathcal{L}(\boldsymbol{u}(t+1))]\\\le&\mathbb{E}_{\xi(t)}\left[\mathcal{L}(\boldsymbol{u}(t))-\eta\langle\nabla\mathcal{L}_{\xi(t)}(\boldsymbol{w}(t)),\nabla\mathcal{L}(\boldsymbol{u}(t))\rangle+\frac{L\eta^2}{2}\|\nabla\mathcal{L}_{\xi(t)}(\boldsymbol{w}(t))\|^2\right]\\\overset{(\circ)}{=}&\mathcal{L}(\boldsymbol{u}(t))-\eta\langle\nabla\mathcal{L}(\boldsymbol{w}(t)),\nabla\mathcal{L}(\boldsymbol{u}(t))\rangle+\frac{L\eta^2}{2}\mathbb{E}_{\xi(t)}\left[\|\nabla\mathcal{L}_{\xi(t)}(\boldsymbol{w}(t))\|^2\right]\\\overset{(\bullet)}{\le}&\mathcal{L}(\boldsymbol{u}(t))-\eta\langle\nabla\mathcal{L}(\boldsymbol{w}(t)),\nabla\mathcal{L}(\boldsymbol{u}(t))\rangle+\frac{L\eta^2}{2}(\sigma_1\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\sigma_0),\end{aligned}\tag{37}$$

where Eq. $(\star)$ is due to that $\boldsymbol{w}(t+1)$ is uniquely determined by $\boldsymbol{B}(t)$ given $\{\boldsymbol{w}(s)\}_{s=1}^t$, Eq. $(\circ)$ is due to $\boldsymbol{u}(t)$ is uniquely determined by $\{\boldsymbol{w}(s)\}_{s=1}^t$, and Inequality. $(\bullet)$ is due to the affine variance assumption .

Therefore, we have

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\boldsymbol{u}(t+1))|\mathcal{F}_t]\\
\le&\mathcal{L}(\boldsymbol{u}(t))-\eta\langle\nabla\mathcal{L}(\boldsymbol{w}(t)),\nabla\mathcal{L}(\boldsymbol{u}(t))\rangle+\frac{L\eta^2}{2}(\sigma_1\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\sigma_0)\\
=&\mathcal{L}(\boldsymbol{u}(t))-\eta\langle\nabla\mathcal{L}(\boldsymbol{w}(t)),\nabla\mathcal{L}(\boldsymbol{w}(t))\rangle+\eta\langle\nabla\mathcal{L}(\boldsymbol{w}(t)),\nabla\mathcal{L}(\boldsymbol{w}(t))-\nabla\mathcal{L}(\boldsymbol{u}(t))\rangle\\
&+\frac{L\eta^2}{2}(\sigma_1\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\sigma_0)\\
=&\mathcal{L}(\boldsymbol{u}(t))-\eta\left(1-\frac{L\eta\sigma_1}{2}\right)\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\langle\eta\nabla\mathcal{L}(\boldsymbol{w}(t)),\nabla\mathcal{L}(\boldsymbol{w}(t))-\nabla\mathcal{L}(\boldsymbol{u}(t))\rangle+\frac{L\eta^2\sigma_0}{2}\\
\le&\mathcal{L}(\boldsymbol{u}(t))-\eta\left(1-\frac{L\eta\sigma_1}{2}\right)\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\frac{1}{2\lambda}\left\|\eta\nabla\mathcal{L}(\boldsymbol{w}(t))\right\|^2+\frac{\lambda}{2}\left\|\nabla\mathcal{L}(\boldsymbol{w}(t))-\nabla\mathcal{L}(\boldsymbol{u}(t))\right\|^2\\
&+\frac{L\eta^2\sigma_0}{2}\\
=&\mathcal{L}(\boldsymbol{u}(t))-\eta\left(1-\left(\frac{1}{2\lambda}+\frac{L\sigma_1}{2}\right)\eta\right)\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\frac{\lambda}{2}\left\|\nabla\mathcal{L}(\boldsymbol{w}(t))-\nabla\mathcal{L}(\boldsymbol{u}(t))\right\|^2+\frac{L\eta^2\sigma_0}{2},
\end{aligned}$$

where $\lambda$ is a positive constant that will be specified latter.

Reusing the bounded smoothness assumption leads to

$$\begin{aligned}
&\|\nabla\mathcal{L}(\boldsymbol{w}(t))-\nabla\mathcal{L}(\boldsymbol{u}(t))\|^2\\
\le&L^2\|\boldsymbol{w}(t)-\boldsymbol{u}(t)\|^2\overset{(\square)}{=}\frac{\beta^2L^2}{(1-\beta)^2}\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|^2\\
=&\beta^2L^2\left\|\sum_{s=1}^{t-1}\eta\beta^{t-1-s}\nabla\mathcal{L}_{\xi(s)}(\boldsymbol{w}(s))\right\|^2\overset{(\diamond)}{\le}\beta^2L^2\eta^2\left(\sum_{s=1}^{t-1}\beta^{t-1-s}\left\|\nabla\mathcal{L}_{\xi(s)}(\boldsymbol{w}(s))\right\|\right)^2\\
\overset{(\clubsuit)}{\le}&\beta^2L^2\eta^2\left(\sum_{s=1}^{t-1}\beta^{t-1-s}\left\|\nabla\mathcal{L}_{\xi(s)}(\boldsymbol{w}(s))\right\|^2\right)\left(\sum_{s=1}^{t-1}\beta^{t-1-s}\right)\\
\le&\frac{\beta^2L^2\eta^2}{1-\beta}\left(\sum_{s=1}^{t-1}\beta^{t-1-s}\left\|\nabla\mathcal{L}_{\xi(s)}(\boldsymbol{w}(s))\right\|^2\right),
\end{aligned}\tag{38}$$

where Inequality $(\square)$ is due to $\beta(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))=(1-\beta)(\boldsymbol{u}(t)-\boldsymbol{w}(t))$ by Eq. (35), Inequality $(\diamond)$ is due to triangular inequality, and Inequality $(\clubsuit)$ is due to Cauchy-Schwartz Inequality.

Combining Eqs. (37) and (38), we have

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\boldsymbol{u}(t+1))|\mathcal{F}_t]\\
\le&\mathcal{L}(\boldsymbol{u}(t))-\eta\left(1-\left(\frac{1}{2\lambda}+\frac{L\sigma_1}{2}\right)\eta\right)\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\frac{\lambda\beta^2L^2\eta^2}{2(1-\beta)}\left(\sum_{s=1}^{t-1}\beta^{t-1-s}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(s)}(\boldsymbol{w}(s))\right\|^2\right)\\
&+\frac{L\eta^2\sigma_0}{2},
\end{aligned}$$

which by taking expectation with respect to $\mathcal{F}_t$ leads to

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\boldsymbol{u}(t+1))]\\
\le&\mathbb{E}\mathcal{L}(\boldsymbol{u}(t))-\eta\left(1-\left(\frac{1}{2\lambda}+\frac{L\sigma_1}{2}\right)\eta\right)\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\mathbb{E}\frac{\lambda\beta^2L^2\eta^2}{2(1-\beta)}\left(\sum_{s=1}^{t-1}\beta^{t-1-s}\left\|\nabla\mathcal{L}_{\xi(s)}(\boldsymbol{w}(s))\right\|^2\right)\\
&+\frac{L\eta^2\sigma_0}{2}\\
\le&\mathbb{E}\mathcal{L}(\boldsymbol{u}(t))-\eta\left(1-\left(\frac{1}{2\lambda}+\frac{L\sigma_1}{2}\right)\eta\right)\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2+\frac{\lambda\beta^2L^2\eta^2}{2(1-\beta)}\sigma_1\left(\sum_{s=1}^{t-1}\beta^{t-1-s}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{w}(s))\right\|^2\right)\\
&+\frac{\lambda\beta^2L^2\eta^2}{2(1-\beta)^2}\sigma_0+\frac{L\eta^2\sigma_0}{2},
\end{aligned}$$

where the last inequality is due to the affine variance assumption. Letting $\lambda = \frac{1-\beta}{L\beta\sqrt{\sigma_1}}$ then leads to

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\boldsymbol{u}(t+1))] \\
\le&\mathbb{E}\mathcal{L}(\boldsymbol{u}(t)) - \eta\left(1-\left(\frac{L\beta\sqrt{\sigma_1}}{2(1-\beta)}+\frac{L\sigma_1}{2}\right)\eta\right)\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2 + \frac{L\beta\eta^2}{2}\sqrt{\sigma_1}\left(\sum_{s=1}^{t-1}\beta^{t-1-s}\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{w}(s))\right\|^2\right) \\
&+\frac{\beta L\eta^2}{2(1-\beta)\sqrt{\sigma_1}}\sigma_0 + \frac{L\eta^2\sigma_0}{2}.
\end{aligned}$$

By the learning rate upper bound $\eta \le \frac{1}{2\frac{L\beta\sqrt{\sigma_1}}{1-\beta}+L\sigma_1}$, summing the above inequality over $t$ then leads to

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}(\boldsymbol{u}(T+1))] \\
\le&\mathcal{L}(\boldsymbol{u}(1)) - \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2 + T\left(\frac{\beta L\eta^2}{2(1-\beta)\sqrt{\sigma_1}}\sigma_0+\frac{L\eta^2\sigma_0}{2}\right).
\end{aligned}$$

The proof is completed. □

As a direct corollary of Theorem 4 and Lemma 3, we have the following corollary.

**Corollary 6.** *Let all conditions in Theorem 3 hold. Then, we have*

$$\sum_{t=1}^{\infty}\mathbb{E}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2 < \infty. \tag{39}$$

*Consequently,*

$$\sum_{t=1}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2 < \infty$$

*and*

$$\langle \boldsymbol{w}(t), \boldsymbol{x}\rangle \to \infty, \forall \boldsymbol{x} \in \tilde{\boldsymbol{S}}$$

*hold almost surely.*

*Proof.* Eq. (39) directly follows from Theorem 4 and Lemma 3. The rest of claims follows immediately by Fubini's Theorem and Assumption 2.

The proof is completed. □

### C.2.2 Parameter dynamics

Similar to the case of GDM, we define $\tilde{\boldsymbol{w}}$ as the solution of Eq. (10) with $C_3 = \frac{\eta}{(1-\beta)N}$. We also let $\boldsymbol{n}(t)$ be given by Lemma 5, and define $\boldsymbol{r}(t)$ in this case as

$$\boldsymbol{r}(t) \overset{\triangle}{=} \boldsymbol{w}(t) - \ln(t)\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}} - \boldsymbol{n}(t). \tag{40}$$

As $\tilde{\boldsymbol{w}}$ is a constant vector, and $\|\boldsymbol{n}(t)\| \to 0$ as $t \to \infty$, we have $\boldsymbol{w}(t) - \ln(t)\hat{\boldsymbol{w}}$ has bounded norm if and only if $\|\boldsymbol{r}(t)\|$ is upper bounded. Similar to the GDM case, we have the following equivalent condition of that $\|\boldsymbol{r}(t)\|$ is bounded.

**Lemma 14.** *Let all conditions in Theorem 3 hold. Then, $\|\boldsymbol{r}(t)\|$ is bounded almost surely if and only if function $g(t)$ is upper bounded almost surely, where $g : \mathbb{Z}^+ \to \mathbb{R}$ is defined as*

$$g(t) \overset{\triangle}{=} \frac{1}{2}\|\boldsymbol{r}(t)\|^2 + \frac{\beta}{1-\beta}\langle \boldsymbol{r}(t), \boldsymbol{w}(t)-\boldsymbol{w}(t-1)\rangle - \frac{\beta}{1-\beta}\sum_{\tau=2}^{t}\langle \boldsymbol{r}(\tau)-\boldsymbol{r}(\tau-1), \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\rangle. \tag{41}$$

*Proof.* To begin with, we prove that almost surely $|\sum_{\tau=2}^{t}\langle \boldsymbol{r}(\tau)-\boldsymbol{r}(\tau-1), \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\rangle|$ is upper bounded for any $t$. By Corollary 6, we have almost surly

$$\sum_{t=1}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2 < \infty.$$

On the other hand, for any $\boldsymbol{w}$, we have

$$
\begin{aligned}
&\|\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w})\| = \frac{1}{b}\left\| \sum_{\boldsymbol{x}\in \boldsymbol{B}(t)} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}\right\| \\
\leq & -\frac{\sigma_{max}}{b}\sum_{\boldsymbol{x}\in \boldsymbol{B}(t)} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) < -\frac{\sigma_{max}}{b}\sum_{\boldsymbol{x}\in \boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle) \\
\leq & -\frac{\sigma_{max}}{b}\sum_{\boldsymbol{x}\in \boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\langle \hat{\boldsymbol{w}}, \boldsymbol{x}\rangle \leq \frac{N\sigma_{max}}{b}\left\| \frac{1}{N}\sum_{\boldsymbol{x}\in \boldsymbol{S}} \ell'(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)\boldsymbol{x}\right\| \|\hat{\boldsymbol{w}}\| \\
= & \frac{N\sigma_{max}}{b\gamma}\left\|\nabla \mathcal{L}(\boldsymbol{w})\right\|.
\end{aligned}
$$

Therefore, we have almost surely,

$$
\sum_{t=1}^{\infty} \left\|\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))\right\|^2 < \infty,
$$

which further leads to almost surely

$$
\begin{aligned}
&\sum_{t=1}^{\infty}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2 \leq \eta^2(1-\beta)^2\sum_{t=1}^{\infty}\left\|\sum_{s=1}^{t}\beta^{t-s}\nabla \mathcal{L}_{\boldsymbol{B}(s)}(\boldsymbol{w}(s))\right\|^2 \\
\leq & \eta^2(1-\beta)^2\sum_{t=1}^{\infty}\left(\sum_{s=1}^{t}\beta^{t-s}\left\|\nabla \mathcal{L}_{\boldsymbol{B}(s)}(\boldsymbol{w}(s))\right\|\right)^2 \\
\leq & \eta^2(1-\beta)^2\sum_{t=1}^{\infty}\left(\sum_{s=1}^{t}\beta^{t-s}\left\|\nabla \mathcal{L}_{\boldsymbol{B}(s)}(\boldsymbol{w}(s))\right\|^2\right)\left(\sum_{s=1}^{t}\beta^{t-s}\right) \\
\leq & \eta^2\sum_{s=1}^{\infty}\left\|\nabla \mathcal{L}_{\boldsymbol{B}(s)}(\boldsymbol{w}(s))\right\|^2 < \infty.
\end{aligned}
$$

By the definition of $\boldsymbol{r}(t)$ (Eq. (40)), we further have

$$
\begin{aligned}
&\left|\sum_{\tau=2}^{t}\langle \boldsymbol{r}(\tau)-\boldsymbol{r}(\tau-1), \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\rangle\right| \\
\leq & \sum_{\tau=2}^{t}\left|\langle \boldsymbol{r}(\tau)-\boldsymbol{r}(\tau-1), \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\rangle\right| \\
= & \sum_{\tau=2}^{t}\left|\left\langle \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)-\ln\left(\frac{\tau+1}{\tau}\right)-(\boldsymbol{n}(\tau)-\boldsymbol{n}(\tau-1)), \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\right\rangle\right| \\
\leq & \sum_{\tau=2}^{t}\|\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\|^2 + \sum_{\tau=2}^{t}\left|\left\langle -\ln\left(\frac{\tau+1}{\tau}\right)-(\boldsymbol{n}(\tau)-\boldsymbol{n}(\tau-1)), \boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\right\rangle\right| \\
\leq & \frac{3}{2}\sum_{\tau=2}^{t}\|\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\|^2 + \frac{1}{2}\sum_{\tau=2}^{t}\left\|-\ln\left(\frac{\tau+1}{\tau}\right)-(\boldsymbol{n}(\tau)-\boldsymbol{n}(\tau-1))\right\|^2 \\
\overset{(\star)}{\leq} & \frac{3}{2}\sum_{\tau=2}^{t}\|\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\|^2 + \frac{1}{2}\sum_{\tau=2}^{t}\mathcal{O}\left(\frac{1}{\tau}\right)^2 < \infty,
\end{aligned}
$$

where Inequality $(\star)$ is due to $\|\boldsymbol{n}(\tau)-\boldsymbol{n}(\tau-1)\| = \mathcal{O}(\frac{1}{\tau})$ and $\ln\frac{\tau+1}{\tau} = \mathcal{O}(\frac{1}{\tau})$.

Therefore, $g(t)$ is upper bounded almost surely is equivalent to $\frac{1}{2}\|\boldsymbol{r}(t)\|^2 + \frac{\beta}{1-\beta}\langle \boldsymbol{r}(t), \boldsymbol{w}(t)-\boldsymbol{w}(t-1)\rangle$ is upper bounded, which can be shown to be equivalent with $\|\boldsymbol{r}(t)\|$ is bounded following the same routine as Lemma 11.

The proof is completed. □

As the case of GDM, we only need to prove $g(t)$ is upper bounded to complete the proof of Theorem 3.

**Lemma 15.** *Let all conditions in Theorem 3 hold. Then, for almost every dataset, we have $g(t)$ is upper bounded.*

*Proof.* Following the same routine as Lemma 11, we have

$$\begin{aligned}&g(t+1)-g(t)\\=&\frac{1}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2+\langle\boldsymbol{r}(t),\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\rangle+\left\langle\boldsymbol{r}(t),-\eta\nabla\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))-(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\rangle,\end{aligned}$$

where $\sum_{t=1}^{\infty}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2$ is upper bounded.

On the other hand, by the definition of $\boldsymbol{r}(t)$ (Eq. (40)), we have

$$\begin{aligned}&\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\\=&\boldsymbol{w}(t+1)-\boldsymbol{w}(t)-\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}}-\boldsymbol{n}(t+1)+\boldsymbol{n}(t),\end{aligned}$$

while by Lemma 5,

$$\frac{N}{b}\frac{1}{t}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(t)\cap\boldsymbol{S}_s}\boldsymbol{v}_i\boldsymbol{x}_i=\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}}+\boldsymbol{n}(t+1)-\boldsymbol{n}(t).$$

Combining the above two equations, we further have

$$\boldsymbol{r}(t+1)-\boldsymbol{r}(t)=\boldsymbol{w}(t+1)-\boldsymbol{w}(t)-\frac{N}{bt}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(t)\cap\boldsymbol{S}_s}\boldsymbol{v}_i\boldsymbol{x}_i,$$

which further indicates

$$g(t+1)-g(t)=\frac{1}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2+\left\langle\boldsymbol{r}(t),-\eta\nabla\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))-\frac{N}{bt}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(t)\cap\boldsymbol{S}_s}\boldsymbol{v}_i\boldsymbol{x}_i\right\rangle.$$

Therefore, we only need to prove $\sum_{t=1}^{\infty}\langle\boldsymbol{r}(t),-\eta\nabla\mathcal{L}(\boldsymbol{w}(t))-\frac{N}{bt}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(t)\cap\boldsymbol{S}_s}\boldsymbol{v}_i\boldsymbol{x}_i\rangle<\infty$. By directly applying the form of $\nabla\mathcal{L}(\boldsymbol{w}(t))$, we have

$$\begin{aligned}&\left\langle\boldsymbol{r}(t),-\eta\nabla\mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))-\frac{N}{bt}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s}\boldsymbol{v}_i\boldsymbol{x}_i\right\rangle\\=&\left\langle\boldsymbol{r}(t),-\frac{\eta}{(1-\beta)b}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i-\frac{N}{bt}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s}\boldsymbol{v}_i\boldsymbol{x}_i\right\rangle\\=&\frac{\eta}{(1-\beta)b}\left\langle\boldsymbol{r}(t),-\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i-\frac{N(1-\beta)}{\eta}\frac{1}{t}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s}\boldsymbol{v}_i\boldsymbol{x}_i\right\rangle\\=&\frac{\eta}{(1-\beta)b}\left\langle\boldsymbol{r}(t),-\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i-\frac{1}{t}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s}e^{-\langle\tilde{\boldsymbol{w}},\boldsymbol{x}_i\rangle}\boldsymbol{x}_i\right\rangle\\=&\frac{\eta}{(1-\beta)b}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s}\left(-\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)-\frac{1}{t}e^{-\langle\tilde{\boldsymbol{w}},\boldsymbol{x}_i\rangle}\right)\langle\boldsymbol{r}(t),\boldsymbol{x}_i\rangle\\+&\frac{\eta}{(1-\beta)b}\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s^c}\langle\boldsymbol{r}(t),-\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i\rangle.\end{aligned}$$

Let $A_6(t)=\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s}\left(-\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)-\frac{1}{t}e^{-\langle\tilde{\boldsymbol{w}},\boldsymbol{x}_i\rangle}\right)\langle\boldsymbol{r}(t),\boldsymbol{x}_i\rangle$, and $A_7(t)=\sum_{i:\boldsymbol{x}_i\in\boldsymbol{B}(s)\cap\boldsymbol{S}_s^c}\langle\boldsymbol{r}(t),-\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}_i\rangle)\boldsymbol{x}_i\rangle$. We will investigate these two terms respectively.

As $\langle \boldsymbol{w}(t), \boldsymbol{x}\rangle \to \infty$, $\forall \boldsymbol{x} \in \boldsymbol{S}$, a.s., we have a.s., there exists a large enough time $t_0$, s.t., $\forall t \geq t_0$, $\forall \boldsymbol{x} \in \boldsymbol{S}$,

$$\begin{aligned}
-\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}\rangle) &\leq (1 + e^{-\mu_+ \langle \boldsymbol{w}(t), \boldsymbol{x}\rangle}) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}\rangle}, \\
-\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle) &\geq (1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}\rangle}) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}\rangle}, \\
\langle \boldsymbol{x}, \boldsymbol{w}(t)\rangle &> 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
A_7(t) \leq & \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s^c} -\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle) \langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \mathbb{1}_{\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \geq 0} \\
\leq & \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s^c} (1 + e^{-\mu_+ \langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle}) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \mathbb{1}_{\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \geq 0} \\
\leq & 2 \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s^c} e^{-\langle \boldsymbol{r}(t) + \ln(t)\hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i\rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \mathbb{1}_{\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \geq 0} \\
\overset{(\star)}{\leq} & 2 \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s^c} \frac{1}{t^\theta} e^{-\langle \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i\rangle} e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \mathbb{1}_{\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \geq 0} \\
\overset{(\dagger)}{\leq} & \frac{2}{e} \frac{1}{t^\theta} \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s^c} e^{-\langle \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i\rangle} \mathbb{1}_{\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \geq 0} \\
\overset{(\circ)}{=} & \mathcal{O}\left(\frac{1}{t^\theta}\right),
\end{aligned}$$

where Inequality. $(\star)$ is due the definition of $\theta$ (Eq. (32)), Inequality. $(\dagger)$ is due to $e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \leq e^{-1}$ , and Eq. $(\circ)$ is due to $\lim_{t\to\infty} e^{-\langle \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i\rangle} = e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i\rangle}$. Thus,

$$\sum_{t=1}^{\infty} A_7(t) < \infty.$$

On the other hand, $A_6(t)$ can be rewritten as

$$\begin{aligned}
A_6(t) = & \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s} \left( -\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle) - \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i\rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \mathbb{1}_{\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \geq 0} \\
& + \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s} \left( -\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i\rangle) - \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i\rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle \mathbb{1}_{\langle \boldsymbol{r}(t), \boldsymbol{x}_i\rangle < 0}.
\end{aligned}$$

If $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \geq 0$, we have for $\varepsilon < 0.5$,

$$
\begin{aligned}
&\left(-\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) - \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}\right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq &\left(\left(1 + e^{-\mu_+ \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}\right) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} - \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}\right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
= &\left(\left(1 + e^{-\mu_+ \langle \boldsymbol{r}(t) + \ln(t) \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle}\right) e^{-\langle \boldsymbol{r}(t) + \ln(t) \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} - \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}\right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
= &\left(\left(1 + e^{-\mu_+ \langle \boldsymbol{r}(t) + \ln(t) \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle}\right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} - 1\right) \frac{1}{t} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \\
\leq &\left(\left(1 + \frac{1}{t^{\mu_+}} e^{-\mu_+ \langle \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle}\right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} - 1\right) \frac{1}{t} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \\
\overset{(\bullet)}{=} &\left(\left(1 + \mathcal{O}\left(\frac{1}{t^{\mu_+}}\right)\right) \left(1 + \mathcal{O}\left(\frac{1}{t^{0.5-\varepsilon}}\right)\right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} - 1\right) \frac{1}{t} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \\
= &\left(e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} - 1\right) \frac{1}{t} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} + \frac{1}{t} \mathcal{O}\left(\frac{1}{t^{\min\{\mu_+, 0.5-\varepsilon\}}}\right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \\
\leq &\frac{1}{t} \mathcal{O}\left(\frac{1}{t^{\min\{\mu_+, 0.5-\varepsilon\}}}\right) e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \\
\overset{(\diamond)}{=} &\mathcal{O}\left(\frac{1}{t^{\min\{1+\mu_+, 1.5-\varepsilon\}}}\right),
\end{aligned}
$$

where Eq. $(\bullet)$ is due to $\boldsymbol{n}(t) = \mathcal{O}(\frac{1}{t^{0.5-\varepsilon}})$, and Eq. $(\diamond)$ is due to $e^{-\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle} \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \leq \frac{1}{e}$.
On the other hand, if $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < 0$, we have

$$
\begin{aligned}
&\left(-\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) - \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}\right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq &\left(\left(1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}\right) e^{-\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} - \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}\right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
= &\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left(-1 + \left(1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}\right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle}\right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle
\end{aligned}
$$

Specifically, if $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \geq -t^{-0.5 \min\{\mu_-, 0.5\}}$,

$$
\begin{aligned}
&\left|\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left(-1 + \left(1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}\right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle}\right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle\right| \\
\leq &\frac{1}{t^{1+0.5\min\{\mu_-, 0.5\}}} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left|-1 + \left(1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle}\right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle}\right| \\
\overset{(\square)}{=} &\mathcal{O}\left(\frac{1}{t^{1+0.5\min\{\mu_-, 0.5\}}}\right),
\end{aligned}
$$

where Eq. $(\square)$ is due to that as $\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle \to \infty$ and $t^{-0.5\min\{\mu_-, 0.5\}} \to 0$ as $t \to \infty$, there exists a large enough time $T$, s.t., $\forall t > T$, under the circumstance $0 > \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \geq -t^{-0.5\min\{\mu_-, 0.5\}}$, $e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} < 1$ and $e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} < 1$.

If $-2 \leq \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < -t^{-0.5\min\{\mu_-, 0.5\}}$, then, for large enough $t$, $|\langle \boldsymbol{x}_i, \boldsymbol{n}(t) \rangle| < 2$, $1 - \frac{e^{\mu_-(-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle + 4)}}{t^{\mu_-}} > 0$, and

$$
\begin{aligned}
&\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
=& \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{r}(t) + \ln(t) \hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}} + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
=& \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{-\mu_- \langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle}}{t^{\mu_-}} e^{-\mu_- \langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq& \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{\mu_-(-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle + 4)}}{t^{\mu_-}} \right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq& \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{\mu_-(-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle + 4)}}{t^{\mu_-}} \right) (1 - \langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle) \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq& \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{\mu_-(-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle + 4)}}{t^{\mu_-}} \right) \left( 1 + t^{-0.5\min\{\mu_-, 0.5\}} - \langle \boldsymbol{n}(t), \boldsymbol{x}_i \rangle \right) \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
=& \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - \frac{e^{\mu_-(-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle + 4)}}{t^{\mu_-}} \right) \left( 1 + t^{-0.5\min\{\mu_-, 0.5\}} + \boldsymbol{o}\left( t^{-0.5\min\{\mu_-, 0.5\}} \right) \right) \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
=& \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + 1 + t^{-0.5\min\{\mu_-, 0.5\}} + \boldsymbol{o}\left( t^{-0.5\min\{\mu_-, 0.5\}} \right) \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < 0.
\end{aligned}
$$

If $-2 > \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle$, then for large enough time $t$, $e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} \geq e^{\frac{3}{2}}$, $1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \geq e^{-\frac{1}{2}}$, and

$$
\begin{aligned}
&\frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \left( -1 + \left( 1 - e^{-\mu_- \langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle} \right) e^{-\langle \boldsymbol{r}(t) + \boldsymbol{n}(t), \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
\leq& \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} (-1 + e) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < 0.
\end{aligned}
$$

Conclusively, if $\langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle < 0$, for large enough $t$, we have

$$
\left( -\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) - \frac{1}{t} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \leq \mathcal{O}\left( \frac{1}{t^{1+0.5\min\{\mu_-, 0.5\}}} \right),
$$

which further indicates, for large enough $t$, we have

$$
A_6(t) \leq \max\left\{ \mathcal{O}\left( \frac{1}{t^{1+0.5\min\{\mu_-, 0.5\}}} \right), \mathcal{O}\left( \frac{1}{t^{\min\{1+\mu_+, 1.5-\varepsilon\}}} \right) \right\},
$$

which indicates

$$
\sum_{t=1}^{\infty} A_6(t) < \infty.
$$

Therefore,

$$
\begin{aligned}
&\sum_{t=1}^{\infty} (g(t+1) - g(t)) \\
=& \sum_{t=1}^{\infty} \left( \frac{1}{2} \|\boldsymbol{r}(t+1) - \boldsymbol{r}(t)\|^2 + \left\langle \boldsymbol{r}(t), -\eta \nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)) - \frac{N}{bt} \sum_{i: \boldsymbol{x}_i \in \boldsymbol{B}(t) \cap \boldsymbol{S}_s} \boldsymbol{v}_i \boldsymbol{x}_i \right\rangle \right) \\
=& \sum_{t=1}^{\infty} \left( \frac{1}{2} \|\boldsymbol{r}(t+1) - \boldsymbol{r}(t)\|^2 + \eta A_6(t) + \eta A_7(t) \right) \\
<& \infty.
\end{aligned}
$$

The proof is completed. □

### C.2.3 Explanation for proper lyapunov function

Based on the success of applying Lyapunov function $\mathcal{L}(\boldsymbol{w}(t)) + \frac{\beta}{2\eta}\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\|^2$ to analyze gradient descent with momentum, it is natural to try to extend this routine to analyze stochastic gradient descent with momentum. However, in this section, we will show such Lyapunov function is not proper to analyze SGDM as this will put constraints on the range of the momentum rate $\beta$. Specifically, at any step $t$, since the loss $\mathcal{L}$ is $\frac{H\sigma_{max}^2}{N}$ smooth at $\boldsymbol{w}(t)$, we can expand the loss $\mathcal{L}$ in the same way as the GDM case:

$$\mathcal{L}(\boldsymbol{w}(t+1)) \leq \mathcal{L}(\boldsymbol{w}(t)) + \langle \boldsymbol{w}(t+1)-\boldsymbol{w}(t), \nabla\mathcal{L}(\boldsymbol{w}(t))\rangle + \frac{H\sigma_{max}^2}{2N}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2.$$

By taking expectation with respect to $\boldsymbol{w}(t+1)$ conditioning on $\{\boldsymbol{w}(s)\}_{s=1}^t$ for both sides, we further obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}(t+1))\,|\mathcal{F}_t\right] \\
\leq&\mathcal{L}(\boldsymbol{w}(t)) + \langle \mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right], \nabla\mathcal{L}(\boldsymbol{w}(t))\rangle + \frac{H\sigma_{max}^2}{2N}\mathbb{E}\left[\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\,|\mathcal{F}_t\right] \\
\overset{(\star)}{=}&\mathcal{L}(\boldsymbol{w}(t)) + \frac{1}{(1-\beta)\eta}\left\langle \mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right], \beta\left(\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\right) - \mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\rangle \\
&+\frac{H\sigma_{max}^2}{2N}\mathbb{E}\left[\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\,|\mathcal{F}_t\right] \\
=&\mathcal{L}(\boldsymbol{w}(t)) + \frac{\beta}{(1-\beta)\eta}\left\langle \left(\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\right), \mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\rangle \\
&+\frac{H\sigma_{max}^2}{2N}\mathbb{E}\left[\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\,|\mathcal{F}_t\right] - \frac{1}{(1-\beta)\eta}\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2 \\
\leq&\mathcal{L}(\boldsymbol{w}(t)) + \frac{\beta}{2(1-\beta)\eta}\left\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\right\|^2 + \frac{\beta}{2(1-\beta)\eta}\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2 \\
&+\frac{H\sigma_{max}^2}{2N}\mathbb{E}\left[\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\,|\mathcal{F}_t\right] - \frac{1}{(1-\beta)\eta}\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2,
\end{aligned}$$

where Eq. $(\star)$ is becasue $\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right] = -(1-\beta)\eta\nabla\mathcal{L}(\boldsymbol{w}(t)) + \beta\left(\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\right)$ due to the definition of SGDM (Eq. (2)). Rearranging the above inequality and taking expectations of both sides with respect to $\{\boldsymbol{w}(s)\}_{s=1}^t$ leads to

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}(t+1))\right] + \frac{2-\beta}{2(1-\beta)\eta}\mathbb{E}\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2 \\
&-\frac{H\sigma_{max}^2}{2N}\mathbb{E}\left[\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\right] \\
\leq&\mathbb{E}\mathcal{L}(\boldsymbol{w}(t)) + \frac{\beta}{2(1-\beta)\eta}\mathbb{E}\left\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\right\|^2. \qquad (42)
\end{aligned}$$

On the other hand, we wish to obtain some positive constant $\alpha$ from Eq. (42), such that (at least),

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}(\boldsymbol{w}(t+1))\right] + \alpha\mathbb{E}\left\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\right\|^2. \\
\leq&\mathbb{E}\mathcal{L}(\boldsymbol{w}(t)) + \alpha\mathbb{E}\left\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\right\|^2, \qquad (43)
\end{aligned}$$

which requires to lower bound $\mathbb{E}\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2$ by $\mathbb{E}\parallel \boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2$. However, in general cases, $\mathbb{E}\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2$ is only upper bounded by $\mathbb{E}\parallel \boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2$ (Holder's Inequality), although in our case, $\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2$ can be bounded as

$$\begin{aligned}
&\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2 \\
=&\left\|-(1-\beta)\eta\nabla\mathcal{L}(\boldsymbol{w}(t)) + \beta\left(\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\right)\right\|^2 \\
=&\left\|-(1-\beta)\eta\nabla\mathcal{L}(\boldsymbol{w}(t))\right\|^2 + \left\|\beta\left(\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\right)\right\|^2 + 2\beta(1-\beta)\eta\langle \boldsymbol{w}(t)-\boldsymbol{w}(t-1), -\nabla\mathcal{L}(\boldsymbol{w}(t))\rangle,
\end{aligned}$$

while by the separability of the dataset and that the loss is non-increasing, $\mathbb{E}\left[\|\,\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\,|\mathcal{F}_t\right]$ can be bounded as

$$\begin{aligned}&\mathbb{E}\left[\|\,\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2\,|\mathcal{F}_t\right]\\ \le&\frac{N\sigma_{max}^2}{b\gamma^2}\left\|\mathbb{E}\left[\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\,|\mathcal{F}_t\right]\right\|^2.\end{aligned} \tag{44}$$

By Eqs. (42) and (44), we have that to ensure Eq. (43), it is required that

$$\frac{2-\beta}{2(1-\beta)\eta}\frac{b\gamma^2}{N\sigma_{max}^2}-\frac{H\sigma_{max}^2}{2N}\ge\frac{\beta}{2(1-\beta)\eta},$$

which puts additional constraint on $\beta$ as

$$\beta\le\frac{2b\gamma^2-H\eta\sigma_{max}^4}{b\gamma^2+N\sigma_{max}^2-H\sigma_{max}^4\eta}.$$

Specifically, the upper bound becomes close to 0 when $N$ becomes large, and constrains $\beta$ in a small range.

# D Implicit regularization of deterministic Adam

This section collects the proof of the convergent direction of Adam, i.e., Theorem 5. The methodology of this section bears great similarity with GDM, although the preconditioner of Adam requires specific treatment for analysis. The proof is still divided into two stages: (1). we first prove the sum of squared gradients along the trajectory is finite. Additionally, we prove the convergent rate of loss is $\mathcal{O}(\frac{1}{t})$; (2). we prove $\boldsymbol{w}(t)-\ln(t)\hat{\boldsymbol{w}}$ has bounded norm. Before we present these two stages of proof, we will first give the required range of $\eta$ for which Theorem 3 holds. The analyses of this section hold for almost every dataset, and the "almost every" constraint is further moved in Section F.1.

## D.1 Choice of learning rate

Let $H_{s_0}$ be the smooth parameter over $[s_0,\infty)$ given by Assumption 3. (D). Let $\beta_2=(c\beta_1)^4$ $(c>1)$. The "sufficiently small learning rate" in Theorem 3 means

$$\eta\le\frac{\sqrt{\varepsilon}\inf_{t\ge2}\left(\frac{1-\beta_1^t}{1-\beta_1}-\frac{1-\beta_1^{t-1}}{c(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}}\right)}{H_{\ell^{-1}((1-c\beta_1)^{-1}N\mathcal{L}(\boldsymbol{w}(1)))}}.$$

To ensure $\eta$ is well-defined, we need to prove

$$\inf_{t\ge2}\left(\frac{1-\beta_1^t}{1-\beta_1}-\frac{1-\beta_1^{t-1}}{c(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}}\right)>0,$$

and we introduce the following technical lemma:

**Lemma 16.** *Define $f_t(x)=\frac{1-x^t}{x(1-x^{t-1})}$, $\forall t\in\mathbb{Z},t\ge2$. We have $f_t(x)$ is decreasing with respect to $x$. Furthermore, for any $x\in[0,1)$, we have*

$$f(x)\ge\sqrt[4]{f(x^4)}. \tag{45}$$

*Proof.* First of all, by definition,

$$f(x)=\frac{1-x^t}{x-x^t}=1+\frac{1-x}{x-x^t}=1+\frac{1-x}{x(1-x^{t-1})}=1+\frac{1}{x(1+x+\cdots+x^{t-2})}$$

is monotonously decreasing as $0\le x<1$. Secondly, Eq. (45) is equivalent to

$$\begin{aligned}&\frac{(1-x^t)^4}{\beta_1^4(1-x^{t-1})^4}\ge\frac{(1-x^{4t})}{x^4(1-x^{4(t-1)})}\\ \Longleftrightarrow&\frac{(1-x^t)^3}{(1-x^{t-1})^3}\ge\frac{(1+x^t)(1+x^{2t})}{(1+x^{t-1})(1+x^{2(t-1)})}.\end{aligned}$$

The left side of the above inequality is no smaller than 1, while the right side is no larger than 1, which completes the proof. □

We are now ready to prove $\eta$ is well-defined. First of all, for every $t$, we have

$$\begin{aligned}
&\frac{1-\beta_1^t}{1-\beta_1} - \frac{1-\beta_1^{t-1}}{c(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}} \\
=&\frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1}\left(\frac{1-\beta_1^t}{\beta_1(1-\beta_1^{t-1})} - \frac{1-(c\beta_1)^t}{(c\beta_1)(1-(c\beta_1)^{t-1})}\right) \\
\overset{(\star)}{=}&\frac{\beta_1(1-\beta_1^{t-1})}{1-\beta_1}\left(f_t(\beta_1)-f_t(c\beta_1)\right) > 0,
\end{aligned} \tag{46}$$

where Eq. $(\star)$ is by Lemma 16 and $c\beta_1 = \sqrt[4]{\beta_2} < 1$.

On the other hand, we have

$$\lim_{t\to\infty}\left(\frac{1-\beta_1^t}{1-\beta_1} - \frac{1-\beta_1^{t-1}}{c(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}}\right) = \left(1-\frac{1}{c}\right)\frac{1}{1-\beta_1}. \tag{47}$$

By Eq. (46) and Eq. (47), we obtain $\frac{1-\beta_1^t}{1-\beta_1} - \frac{1-\beta_1^{t-1}}{c(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}}$ is lower bounded by some positive constant across $t$, and $\eta$ is well defined.

### D.2 Sum of gradients along the trajectory is bounded

We start with the following lemma, which indicates $\mathcal{L}(\boldsymbol{w}(t)) + \|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\|^2$ is a proper Lyapunov function for Adam.

**Lemma 17.** *Let all conditions in Theorem 5 hold. Then, for any* $t \geq 1$*,*

$$\begin{aligned}
&\mathcal{L}(t+1) + \frac{1}{2}\frac{1-\beta_1^t}{\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2 \\
\leq&\mathcal{L}(\boldsymbol{w}(t)) + \left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\frac{1-\beta_1^{t-1}}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}}.
\end{aligned} \tag{48}$$

*Proof.* We start with the case $t=1$. To begin with, we have $\mathcal{L}$ is $H_{\ell^{-1}(N\mathcal{L}(\boldsymbol{w}(1)))}$ smooth around $\boldsymbol{w}(1)$. By definition $H_x$ is non-increasing with respect to $x$, and since $\ell^{-1}$ is also non-increasing, we have

$$H_{\ell^{-1}(N\mathcal{L}(\boldsymbol{w}(1)))} \leq H_{\ell^{-1}(\frac{1}{1-c\beta_1}N\mathcal{L}(\boldsymbol{w}(1)))},$$

which further indicates when $\alpha$ is small enough,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}(1+\alpha)) \overset{(\star)}{\leq}&\mathcal{L}(\boldsymbol{w}(1)) + \alpha\langle\nabla\mathcal{L}(\boldsymbol{w}(1)),\boldsymbol{w}(2)-\boldsymbol{w}(1)\rangle + \frac{L}{2}\alpha^2\|\boldsymbol{w}(2)-\boldsymbol{w}(1)\| \\
=&\mathcal{L}(\boldsymbol{w}(1)) - \alpha\left\langle\nabla\mathcal{L}(\boldsymbol{w}(1)),\eta\frac{1}{\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(1)}}\odot\nabla\mathcal{L}(\boldsymbol{w}(1))\right\rangle + \boldsymbol{o}(\alpha^2) \\
\leq&\mathcal{L}(\boldsymbol{w}(1)) - \frac{1}{2\eta}\alpha^2\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(1)}\odot(\boldsymbol{w}(2)-\boldsymbol{w}(t))\right\|^2,
\end{aligned}$$

where in Eq. $(\star)$ we denote $L \overset{\triangle}{=} H_{\ell^{-1}(\frac{1}{1-c\beta_1}N\mathcal{L}(\boldsymbol{w}(1)))}$, and the last inequality is due to $\frac{1}{2\eta}\alpha^2\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(1)}\odot(\boldsymbol{w}(2)-\boldsymbol{w}(t))\right\|^2 = \boldsymbol{o}(\alpha^2)$, and $\left\langle\nabla\mathcal{L}(\boldsymbol{w}(1)),\eta\frac{1}{\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(1)}}\odot\nabla\mathcal{L}(\boldsymbol{w}(1))\right\rangle$ is positive.

Now if there exists an $\alpha \in (0,1)$, such that Eq. (48) fails, we denote $\alpha^* = \inf\{\alpha : Eq.(48)\ fails\ for\ 1+\alpha\}$. We have $\alpha^* > 0$, and the equality in Eq. (48) holds for $1+\alpha^*$. Therefore, we have for any $\alpha \in (0,\alpha^*)$,

$$\mathcal{L}(\boldsymbol{w}(1+\alpha)) \leq \mathcal{L}(\boldsymbol{w}(1+\alpha)) + \frac{1}{2\eta}\alpha^2\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(1)}\odot(\boldsymbol{w}(2)-\boldsymbol{w}(t))\right\|^2 \leq \mathcal{L}(\boldsymbol{w}(1)),$$

which by Lemma 6 leads to $\mathcal{L}$ is $H_{\ell^{-1}(N\mathcal{L}(\boldsymbol{w}(1)))}$ smooth (thus $L$ smooth) over the set $\{\boldsymbol{w}(1+\alpha) : \alpha \in [0, \alpha^*]\}$, and

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}(1+\alpha^*))\\
\le&\mathcal{L}(\boldsymbol{w}(1)) + \alpha^* \langle \nabla \mathcal{L}(\boldsymbol{w}(1)), \boldsymbol{w}(2) - \boldsymbol{w}(1) \rangle + \frac{L}{2}(\alpha^*)^2 \|\boldsymbol{w}(2) - \boldsymbol{w}(1)\|^2\\
=&\mathcal{L}(\boldsymbol{w}(1)) - \alpha^* \left\langle \frac{1}{\eta}\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)} \odot (\boldsymbol{w}(2) - \boldsymbol{w}(1)), \boldsymbol{w}(2) - \boldsymbol{w}(1) \right\rangle + \frac{L}{2}(\alpha^*)^2 \|\boldsymbol{w}(2) - \boldsymbol{w}(1)\|^2\\
=&\mathcal{L}(\boldsymbol{w}(1)) - \alpha^* \frac{1}{\eta} \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)} \odot (\boldsymbol{w}(2) - \boldsymbol{w}(1)) \right\|^2 + \frac{L}{2}(\alpha^*)^2 \left\| \frac{1}{\sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)}} \odot \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)} \odot (\boldsymbol{w}(2) - \boldsymbol{w}(1)) \right\|^2\\
\le&\mathcal{L}(\boldsymbol{w}(1)) - \alpha^* \frac{1}{\eta} \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)} \odot (\boldsymbol{w}(2) - \boldsymbol{w}(1)) \right\|^2 + \frac{L}{2\sqrt{\varepsilon}}(\alpha^*)^2 \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)} \odot (\boldsymbol{w}(2) - \boldsymbol{w}(1)) \right\|^2\\
<&\mathcal{L}(\boldsymbol{w}(1)) - (\alpha^*)^2 \frac{1}{\eta} \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)} \odot (\boldsymbol{w}(2) - \boldsymbol{w}(1)) \right\|^2 + \frac{L}{2\sqrt{\varepsilon}}(\alpha^*)^2 \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)} \odot (\boldsymbol{w}(2) - \boldsymbol{w}(1)) \right\|^2\\
\le&\mathcal{L}(\boldsymbol{w}(1)) - (\alpha^*)^2 \frac{1}{2\eta} \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(1)} \odot (\boldsymbol{w}(2) - \boldsymbol{w}(1)) \right\|^2,
\end{aligned} \tag{49}$$

where the second-to-last inequality is due to $\|\boldsymbol{w}(2) - \boldsymbol{w}(1)\| > 0$ (by Lemma 9) and $\alpha^* > (\alpha^*)^2$, while the last inequality is due to

$$\begin{aligned}
\eta \le& \frac{\sqrt{\varepsilon}\inf_{t\ge 2}\left(\frac{1-\beta_1^t}{1-\beta_1} - \frac{1-\beta_1^{t-1}}{c(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}}\right)}{L} \le \frac{\sqrt{\varepsilon}\left(\frac{1-\beta_1^2}{1-\beta_1} - \frac{1-\beta_1}{c(1-\beta_1)}\frac{1-(c\beta_1)^2}{1-(c\beta_1)}\right)}{L}\\
=& \frac{\sqrt{\varepsilon}\left(1+\beta_1 - \frac{1+(c\beta_1)}{c}\right)}{L} = \frac{\sqrt{\varepsilon}\left(1 - \frac{1}{c}\right)}{L} < \frac{\sqrt{\varepsilon}}{L}.
\end{aligned}$$

Eq. (49) contradicts the fact that the equality in Eq. (48) holds for $1+\alpha^*$, which completes the proof of $t=1$.

If $t \ge 2$, following the similar routine as $t=1$, we also prove Eq. (48) by reduction to absurdity. If there exist $t$ and $\alpha$ such that Eq. (48) fails. Denote $t^*$ as the smallest time such that there exists an $\alpha \in [0,1)$ such that Eq. (48) fails for $t^*$ and $\alpha$. By Lemma 9, $\left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t^*-1)} \odot (\boldsymbol{w}(t^*) - \boldsymbol{w}(t^*-1)) \right\|^2$ is positive, and strict inequality in Eq. (48) holds for $t$ and $\alpha = 0$, which by continuity leads to

$$1 > \alpha^* \overset{\triangle}{=} \inf\{\alpha \in [0,1] : Eq.(48)\ fails\ for\ 1+\alpha\} > 0.$$

Then, for any $\alpha \in [0, \alpha^*]$, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}(t^*+\alpha))\\
\le&\mathcal{L}(\boldsymbol{w}(t^*+\alpha)) + \frac{1}{2}\alpha^2 \frac{1-\beta_1^{t^*}}{\eta(1-\beta_1)} \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t^*)} \odot (\boldsymbol{w}(t^*+1) - \boldsymbol{w}(t^*)) \right\|^2\\
\le&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{1-\beta_1^{t^*-1}}{2c\eta(1-\beta_1)} \frac{1-(c\beta_1)^{t^*}}{1-(c\beta_1)^{t^*-1}} \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t^*-1)} \odot (\boldsymbol{w}(t^*) - \boldsymbol{w}(t^*-1)) \right\|^2.
\end{aligned}$$

On the other hand, for any time $2 \le s \le t^*-1$, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}(s+1)) + \frac{1}{2}\frac{1-\beta_1^s}{\eta(1-\beta_1)} \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(s)} \odot (\boldsymbol{w}(s+1) - \boldsymbol{w}(s)) \right\|^2\\
\le&\mathcal{L}(\boldsymbol{w}(s)) + \frac{\beta_1(1-\beta_1^{s-1})}{2c\eta(1-\beta_1)} \frac{1-(c\beta_1)^s}{1-(c\beta_1)^{s-1}} \left\| \sqrt[4]{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(s-1)} \odot (\boldsymbol{w}(s) - \boldsymbol{w}(s-1)) \right\|^2.
\end{aligned} \tag{50}$$

By Eq. (47), we have

$$\frac{1-\beta_1^s}{\eta(1-\beta_1)} > \frac{1-\beta_1^s}{c\eta(1-\beta_1)} = \frac{1-\beta_1^s}{c\eta(1-\beta_1)} \frac{1-(c\beta_1)^{s+1}}{1-(c\beta_1)^s} \frac{1-(c\beta_1)^s}{1-(c\beta_1)^{s+1}},$$

which by $\frac{(1-\beta_1^{s-1})}{(1-\beta_1^s)}$ further leads to

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}(s)) + \frac{1-\beta_1^{s-1}}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^s}{1-(c\beta_1)^{s-1}}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(s-1)}\odot(\boldsymbol{w}(s)-\boldsymbol{w}(s-1))\right\|^2 \\
\geq&\mathcal{L}(\boldsymbol{w}(s+1)) + \frac{1}{2}\frac{1-\beta_1^s}{\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(s)}\odot(\boldsymbol{w}(s+1)-\boldsymbol{w}(s))\right\|^2 \\
>&\mathcal{L}(\boldsymbol{w}(s+1)) + \frac{1-\beta_1^s}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^{s+1}}{1-(c\beta_1)^s}\frac{1-(c\beta_1)^s}{1-(c\beta_1)^{s+1}}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(s)}\odot(\boldsymbol{w}(s+1)-\boldsymbol{w}(s))\right\|^2 \\
>&\frac{1-(c\beta_1)^s}{1-(c\beta_1)^{s+1}}\left(\mathcal{L}(\boldsymbol{w}(s+1)) + \frac{1-\beta_1^s}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^{s+1}}{1-(c\beta_1)^s}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(s)}\odot(\boldsymbol{w}(s+1)-\boldsymbol{w}(s))\right\|^2\right).
\end{aligned} \tag{51}$$

On the other hand, for $s=1$, we have

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}(1)) \geq&\mathcal{L}(\boldsymbol{w}(2)) + \frac{1}{2}\frac{1-\beta_1}{\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(1)}\odot(\boldsymbol{w}(2)-\boldsymbol{w}(1))\right\|^2 \\
\geq&\frac{1-(c\beta_1)}{1-(c\beta_1)^2}\left(\mathcal{L}(\boldsymbol{w}(2)) + \frac{1-\beta_1}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^2}{1-(c\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(1)}\odot(\boldsymbol{w}(2)-\boldsymbol{w}(1))\right\|^2\right).
\end{aligned} \tag{52}$$

Combining Eqs. (50), (51), and (52), we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}(t^*+\alpha)) \\
\leq&\mathcal{L}(\boldsymbol{w}(t^*)) + \frac{1-\beta_1^{t^*-1}}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^{t^*}}{1-(c\beta_1)^{t^*-1}}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\right\|^2 \\
<&\frac{1-(c\beta_1)^{t^*}}{1-(c\beta_2)^{t^*-1}}\left(\mathcal{L}(\boldsymbol{w}(t^*-1)) + \frac{1-\beta_1^{t^*-2}}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^{t^*-1}}{1-(c\beta_1)^{t^*-2}}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-2)}\odot(\boldsymbol{w}(t^*-1)-\boldsymbol{w}(t^*-2))\right\|^2\right) \\
<&\cdots \\
<&\frac{1-(c\beta_1)^{t^*}}{1-(c\beta_1)^2}\left(\mathcal{L}(\boldsymbol{w}(2)) + \frac{1-\beta_1}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^2}{1-(c\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(1)}\odot(\boldsymbol{w}(2)-\boldsymbol{w}(1))\right\|^2\right) \\
\leq&\frac{1-(c\beta_1)^{t^*}}{1-c\beta_1}\mathcal{L}(\boldsymbol{w}(1)) < \frac{1}{1-c\beta_1}\mathcal{L}(\boldsymbol{w}(1)).
\end{aligned}$$

Therefore, by Lemma 6, $\mathcal{L}$ is $H_{\ell^{-1}(\frac{1}{1-c\beta_1}N\mathcal{L}(\boldsymbol{w}(1)))}$ smooth (thus $L$ smooth) over the set $\{\boldsymbol{w}(t^*+\alpha):\alpha\in[0,\alpha^*]\}$, which further leads to

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{w}(t^*+\alpha^*))\\
\le&\mathcal{L}(\boldsymbol{w}(t^*))+\alpha^*\langle\nabla\mathcal{L}(\boldsymbol{w}(t^*)),\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\rangle+\frac{L}{2}(\alpha^*)^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
\overset{(\bullet)}{=}&-\frac{\alpha^*}{\eta(1-\beta_1)}\Big\langle\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*),(1-\beta_1^{t^*})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\\
&-\beta_1(1-\beta_1^{t^*-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\Big\rangle\\
&+\mathcal{L}(\boldsymbol{w}(t^*))+\frac{L}{2}(\alpha^*)^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2\\
=&\mathcal{L}(\boldsymbol{w}(t^*))+\frac{L}{2}(\alpha^*)^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2-\frac{\alpha^*(1-\beta_1^{t^*})}{\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2\\
&+\beta_1\frac{\alpha^*(1-\beta_1^{t^*-1})}{\eta(1-\beta_1)}\Big\langle\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*),\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\Big\rangle\\
=&\mathcal{L}(\boldsymbol{w}(t^*))+\frac{L}{2}(\alpha^*)^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2-\frac{\alpha^*(1-\beta_1^{t^*})}{\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2\\
&+\beta_1\frac{\alpha^*(1-\beta_1^{t^*-1})}{\eta(1-\beta_1)}\left\langle\frac{\sqrt[8]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}}{\sqrt[8]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}}\odot\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)),\right.\\
&\qquad\qquad\qquad\qquad\left.\frac{\sqrt[8]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}}{\sqrt[8]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}}\odot\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\right\rangle\\
\le&\mathcal{L}(\boldsymbol{w}(t^*))+\frac{L}{2}(\alpha^*)^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2-\frac{\alpha^*(1-\beta_1^{t^*})}{\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2\\
&+\beta_1\frac{(\alpha^*)^2(1-\beta_1^{t^*-1})}{2\eta(1-\beta_1)}\left\|\frac{\sqrt[8]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}}{\sqrt[8]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}}\odot\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2\\
&+\beta_1\frac{(1-\beta_1^{t^*-1})}{2\eta(1-\beta_1)}\left\|\frac{\sqrt[8]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}}{\sqrt[8]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}}\odot\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\right\|^2\\
\overset{(\diamond)}{\le}&\mathcal{L}(\boldsymbol{w}(t^*))+\frac{L}{2}(\alpha^*)^2\|\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*)\|^2-\frac{\alpha^*(1-\beta_1^{t^*})}{\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2\\
&+\beta_1\frac{(\alpha^*)^2(1-\beta_1^{t^*-1})}{2\eta(1-\beta_1)}\frac{1-(c\beta_1)^{t^*}}{c\beta_1(1-(c\beta_1)^{t^*-1})}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2\\
&+\beta_1\frac{(1-\beta_1^{t^*-1})}{2\eta(1-\beta_1)}\frac{1-(c\beta_1)^{t^*}}{c\beta_1(1-(c\beta_1)^{t^*-1})}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\right\|^2\\
\le&\mathcal{L}(\boldsymbol{w}(t^*))+\frac{L}{2\sqrt{\varepsilon}}(\alpha^*)^2\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\|^2\\
&-\frac{\alpha^*(1-\beta_1^{t^*})}{\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2\\
&+\beta_1\frac{(\alpha^*)^2(1-\beta_1^{t^*})}{2\eta(1-\beta_1)}\frac{1-(c\beta_1)^{t^*}}{c\beta_1(1-(c\beta_1)^{t^*-1})}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2\\
&+\beta_1\frac{(1-\beta_1^{t^*})}{2\eta(1-\beta_1)}\frac{1-(c\beta_1)^{t^*}}{c\beta_1(1-(c\beta_1)^{t^*-1})}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\right\|^2\\
\overset{(\square)}{<}&\mathcal{L}(\boldsymbol{w}(t^*))-\frac{(\alpha^*)^2(1-\beta_1^{t^*})}{2\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2
\end{aligned}
$$

$$+\frac{(1-\beta_1^{t^*-1})}{2\eta(1-\beta_1)}\frac{1-(c\beta_1)^{t^*}}{c(1-(c\beta_1)^{t^*-1})}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\right\|^2,$$

where Eq. (●) is due to an alternative form of the Adam's update rule:

$$\begin{aligned}&(1-\beta_1^{t^*})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))-\beta_1(1-\beta_1^{t^*-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}\odot(\boldsymbol{w}(t^*)-\boldsymbol{w}(t^*-1))\\=&-\eta(1-\beta_1)\nabla\mathcal{L}(\boldsymbol{w}(t^*)),\end{aligned}\tag{53}$$

Inequality ($\diamond$) is due to

$$\begin{aligned}&\frac{\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}}{\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}}=\sqrt[4]{\frac{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}}=\sqrt[4]{\frac{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}{\varepsilon\mathbb{1}_d+\frac{\beta_2\boldsymbol{\nu}(t^*-1)+(1-\beta_2)\nabla\mathcal{L}(\boldsymbol{w}(t^*))^2}{1-\beta_2^{t^*}}}}\\&\le\sqrt[4]{\frac{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}{\varepsilon\mathbb{1}_d+\frac{\beta_2\boldsymbol{\nu}(t^*-1)}{1-\beta_2^{t^*}}}}=\sqrt[4]{\frac{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}{\varepsilon\mathbb{1}_d+\frac{\beta_2(1-\beta_2^{t^*-1})\hat{\boldsymbol{\nu}}(t^*-1)}{1-\beta_2^{t^*}}}}\le\sqrt[4]{\frac{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*-1)}{\frac{\beta_2(1-\beta_2^{t^*-1})\hat{\boldsymbol{\nu}}(t^*-1)}{1-\beta_2^{t^*}}\varepsilon\mathbb{1}_d+\frac{\beta_2(1-\beta_2^{t^*-1})\hat{\boldsymbol{\nu}}(t^*-1)}{1-\beta_2^{t^*}}}}\\&=\sqrt[4]{\frac{1-\beta_2^{t^*}}{\beta_2(1-\beta_2^{t^*-1})}}\mathbb{1}_d\ (\textit{all the computings are component-wisely}),\end{aligned}$$

and $f(c\beta_1)\ge\sqrt[4]{f((c\beta_1)^4)}$, and Inequality ($\Box$) is due to

$$\frac{L}{2\sqrt{\varepsilon}}\le\frac{\inf_{t\ge2}\left(\frac{1-\beta_1^t}{1-\beta_1}-\frac{1-\beta_1^{t-1}}{c(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}}\right)}{2\eta}\le\frac{\left(\frac{1-\beta_1^{t^*}}{1-\beta_1}-\frac{1-\beta_1^{t^*-1}}{c(1-\beta_1)}\frac{1-(c\beta_1)^{t^*}}{1-(c\beta_1)^{t^*-1}}\right)}{2\eta},$$

$\alpha^*>(\alpha^*)^2$, and $\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t^*)}\odot(\boldsymbol{w}(t^*+1)-\boldsymbol{w}(t^*))\right\|^2>0.$

This contradicts to that the equality in Eq. (48) holds for $t^*+\alpha^*$.

The proof is completed. □

As $\lim_{t\to\infty}\beta_1^t=0$ and $\lim_{t\to\infty}(c\beta_1)^t=0$, we have the following corollary based on Lemma 1.

**Corollary 7.** *Let all assumptions in Theorem 5 hold. Then, for large enough t, we have*

$$\begin{aligned}&\mathcal{L}(\boldsymbol{w}(t+1))+\frac{1}{2\sqrt[4]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\\\le&\mathcal{L}(\boldsymbol{w}(t))+\frac{1}{2\sqrt[2]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2.\end{aligned}\tag{54}$$

*Consequently, we have*

$$\sum_{t=1}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2<\infty.\tag{55}$$

The proof of Corollary 7 relies on the following classical lemma on the equivalence between the convergence of two non-negative sequence. The proof is omitted here and can be found in [43].

**Lemma 18** (c.f. Lemma 27, [43])**.** *Let $\{a_i\}_{i=1}^{\infty}$ be a series of non-negative reals, and $\varepsilon$ be a positive real. Then, $\sum_{i=1}^{\infty}a_i<\infty$ is equivalent to $\sum_{i=1}^{\infty}\frac{a_i}{\sqrt{\varepsilon+\sum_{s=1}^{i}a_s}}<\infty$.*

*Proof of Corollary 7.* We have

$$\begin{aligned}\lim_{t\to\infty}\frac{1-\beta_1^{t-1}}{2c\eta(1-\beta_1)}\frac{1-(c\beta_1)^t}{1-(c\beta_1)^{t-1}}=\frac{1}{2c\eta(1-\beta_1)}<\frac{1}{2\sqrt[2]{c}\eta(1-\beta_1)},\\\lim_{t\to\infty}\frac{1-\beta_1^t}{2\eta(1-\beta_1)}=\frac{1}{2\eta(1-\beta_1)}>\frac{1}{2\sqrt[4]{c}\eta(1-\beta_1)},\end{aligned}$$

which completes the proof of Eq. (54). Rearranging Eq. (54) leads to

$$\frac{\sqrt[4]{c}-1}{2\sqrt[2]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\le\frac{1}{2\sqrt[4]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2$$
$$+\mathcal{L}(\boldsymbol{w}(t))-\left(\mathcal{L}(\boldsymbol{w}(t+1))+\frac{1}{2\sqrt[4]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\right),$$

which by iteration further leads to that for a large enough time $T_1$

$$\begin{aligned}&\sum_{t=T_1}^{T_2}\frac{\sqrt[4]{c}-1}{2\sqrt[2]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\\
\le&\mathcal{L}(\boldsymbol{w}(T_1))+\frac{1}{2\sqrt[4]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(T_1-1)}\odot(\boldsymbol{w}(T_1)-\boldsymbol{w}(T_1-1))\right\|^2\\
-&\mathcal{L}(\boldsymbol{w}(T_2+1))+\frac{1}{2\sqrt[4]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(T_2)}\odot(\boldsymbol{w}(T_2+1)-\boldsymbol{w}(T_2+1))\right\|^2\\
<&\mathcal{L}(\boldsymbol{w}(T_1))+\frac{1}{2\sqrt[4]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(T_1-1)}\odot(\boldsymbol{w}(T_1)-\boldsymbol{w}(T_1-1))\right\|^2.\end{aligned}$$

Consequently, we obtain

$$\sum_{t=1}^{\infty}\frac{\sqrt[4]{c}-1}{2\sqrt[2]{c}\eta(1-\beta_1)}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2<\infty. \tag{56}$$

On the other hand, for any $t$, we have

$$\begin{aligned}&\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot\hat{\boldsymbol{w}}\right\|\\
\ge&\left\langle\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t)),\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot\hat{\boldsymbol{w}}\right\rangle=\left\langle\sqrt[2]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t)),\hat{\boldsymbol{w}}\right\rangle\\
=&\langle-\eta\hat{\boldsymbol{m}}(t),\hat{\boldsymbol{w}}\rangle=-\frac{\eta(1-\beta_1)}{1-\beta_1^t}\left\langle\sum_{s=1}^{t}\beta_1^{t-s}\nabla\mathcal{L}(\boldsymbol{w}(s)),\hat{\boldsymbol{w}}\right\rangle\\
=&-\frac{\eta(1-\beta_1)}{1-\beta_1^t}\frac{1}{N}\left\langle\sum_{s=1}^{t}\beta_1^{t-s}\sum_{\boldsymbol{x}_i\in\boldsymbol{S}}\ell'(\langle\boldsymbol{x}_i,\boldsymbol{w}(s)\rangle)\boldsymbol{x}_i,\hat{\boldsymbol{w}}\right\rangle\ge-\frac{\eta(1-\beta_1)}{1-\beta_1^t}\frac{1}{N}\sum_{s=1}^{t}\beta_1^{t-s}\sum_{\boldsymbol{x}_i\in\boldsymbol{S}}\ell'(\langle\boldsymbol{x}_i,\boldsymbol{w}(s)\rangle)\\
\ge&-\frac{\eta(1-\beta_1)}{1-\beta_1^t}\frac{1}{N}\sum_{\boldsymbol{x}_i\in\boldsymbol{S}}\ell'(\langle\boldsymbol{x}_i,\boldsymbol{w}(t)\rangle)\ge\frac{\eta(1-\beta_1)}{1-\beta_1^t}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|,\end{aligned}$$

which by Eq. (56) indicates

$$\sum_{t=1}^{\infty}\left(\frac{\eta(1-\beta_1)}{1-\beta_1^t}\right)^2\frac{\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}{\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot\hat{\boldsymbol{w}}\right\|^2}<\infty.$$

As $\lim_{t\to\infty}\left(\frac{\eta(1-\beta_1)}{1-\beta_1^t}\right)^2=\eta^2(1-\beta_1)^2$, we then obtain

$$\begin{aligned}&\sum_{t=1}^{\infty}\frac{\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}{\sqrt[2]{\varepsilon+\sum_{s=1}^{t}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}}\le\sum_{t=1}^{\infty}\frac{\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}{\sqrt[2]{\varepsilon+\sum_{s=1}^{t}(1-\beta)\beta^{t-s}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}}\\
\le&\sqrt{\frac{1}{1-\beta}}\sum_{t=1}^{\infty}\frac{\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}{\sqrt[2]{\varepsilon+\frac{\sum_{s=1}^{t}(1-\beta)\beta^{t-s}\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}{1-\beta^t}}}\le d\sqrt{\frac{1}{1-\beta}}\sum_{t=1}^{\infty}\frac{\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}{\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\frac{\sum_{s=1}^{t}(1-\beta)\beta^{t-s}\nabla\mathcal{L}(\boldsymbol{w}(t))^2}{1-\beta^t}}\right\|^2}\\
=&d\sqrt{\frac{1}{1-\beta}}\sum_{t=1}^{\infty}\frac{\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}{\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\right\|^2}\le d\|\hat{\boldsymbol{w}}\|_\infty^2\sqrt{\frac{1}{1-\beta}}\sum_{t=1}^{\infty}\frac{\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2}{\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot\hat{\boldsymbol{w}}\right\|^2}<\infty,\end{aligned}$$

which by Lemma 18 completes the proof. □

Based on Corollary 7, we can further prove Lemma 19, characterizing the convergent rate of loss $\mathcal{L}$ directly.

**Lemma 19.** *Let all conditions in Theorem 5 hold. Then,* $\mathcal{L}(\boldsymbol{w}(t)) = \Theta\left(t^{-1}\right), \|\boldsymbol{w}(t)\| = \Theta(\ln(t))$*, and* $\|\boldsymbol{w}(t) - \boldsymbol{w}(t-1)\| = \Theta(t^{-1})$.

*Proof of Lemma 19.* To begin with, Eq. (53) indicates

$$
\begin{aligned}
&\|\eta(1-\beta_1)\nabla\mathcal{L}(\boldsymbol{w}(t))\|^2\\
=&\left\|(1-\beta_1^t)\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))-\beta_1(1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\\
\le&\left\|(1-\beta_1^t)\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|+\left\|\beta_1(1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\\
\le&\left(\left\|\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|+\left\|\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|\right)^2\\
\le&2\left(\left\|\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2+\left\|\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\right) \qquad (57)
\end{aligned}
$$

On the other hand, by Corollary 7,

$$\sum_{s=1}^{\infty}\|\nabla\mathcal{L}(\boldsymbol{w}(s))\|^2<\infty,$$

which following the same routine as Corollary 6 leads to

$$\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle\to\infty,\forall\boldsymbol{x}\in\tilde{\boldsymbol{S}}.$$

Therefore, by Lemma 7, there exists a large enough time $T_1$, such that $\forall t\ge T_1$,

$$\frac{1}{K}\ell(\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle)\le-\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle)\le K\ell(\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle),\forall\boldsymbol{x}\in\boldsymbol{S},$$

which by the separable assumption further leads to

$$
\begin{aligned}
&\frac{\gamma}{K}\mathcal{L}(\boldsymbol{w}(t))\le-\frac{\gamma}{N}\sum_{\boldsymbol{x}\in\boldsymbol{S}}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle)\le\frac{1}{N}\left\langle-\sum_{\boldsymbol{x}\in\boldsymbol{S}}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle)\boldsymbol{x},\gamma\hat{\boldsymbol{w}}\right\rangle\\
\le&\frac{1}{N}\left\|\sum_{\boldsymbol{x}\in\boldsymbol{S}}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle)\boldsymbol{x}\right\|\|\gamma\hat{\boldsymbol{w}}\|=\|\nabla\mathcal{L}(\boldsymbol{w}(t))\|\\
\le&-\frac{1}{N}\sum_{\boldsymbol{x}\in\boldsymbol{S}}\ell'(\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle)\le K\mathcal{L}(\boldsymbol{w}(t)). \qquad (58)
\end{aligned}
$$

Combining Eq. (26) and the above inequality, we have

$$
\begin{aligned}
\left(\frac{\eta(1-\beta_1)\gamma}{K}\right)^2\mathcal{L}(\boldsymbol{w}(t))^2\le2\Bigg(&\left\|\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\\
\text{‘}\qquad&+\left\|\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\Bigg). \qquad (59)
\end{aligned}
$$

On the other hand, by Eq. (56), we have

$$\sum_{t=1}^{\infty}\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2<\infty.$$

Therefore, there exists large enough time $T_2$, such that $\forall t>T_2$,

$$\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2<1,$$

and thus,

$$\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^4<\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2. \qquad (60)$$

Combining Eq. (59) and Eq. (60), there exists a positive real constant $C$, such that

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}(t))^2 \leq & C\left(\left\|\sqrt{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)} \odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\right. \\ & \left.+\left\|\sqrt{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\right), \\ \left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)} \odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^4 \leq & C\left(\left\|\sqrt{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)} \odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\right. \\ & \left.+\left\|\sqrt{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\right).\end{aligned}$$

Rearranging Eq. (54) leads to

$$\begin{aligned}& \frac{\sqrt[4]{c}-1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\left(\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\right. \\ & \qquad\left.+\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\right) \\ \leq & \mathcal{L}(\boldsymbol{w}(t))+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2 \\ & \qquad-\left(\mathcal{L}(\boldsymbol{w}(t+1))+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)} \odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\right),\end{aligned}$$

which further indicates

$$\begin{aligned}& \left(\mathcal{L}(\boldsymbol{w}(t))+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\right)^2 \\ \leq & 2\left(\mathcal{L}(\boldsymbol{w}(t))^2+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^4\right) \\ \leq & 2 C\left(1+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\right)\left(\left\|\sqrt{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)} \odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\right. \\ & \qquad\left.+\left\|\sqrt{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2\right) \\ \leq & 2 C\left(1+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\right) \frac{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}{\sqrt[4]{c}-1}\left(\mathcal{L}(\boldsymbol{w}(t))+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\right. \\ & \cdot\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2-(\mathcal{L}(\boldsymbol{w}(t+1)) \\ & \left.\left.+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)} \odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\|^2\right)\right).\end{aligned}$$

Denote $\xi(t)$ as

$$\xi(t) \triangleq \mathcal{L}(\boldsymbol{w}(t))+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2.$$

We then have

$$\xi(t)^2 \leq 2 C\left(1+\frac{\sqrt[4]{c}+1}{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}\right) \frac{4 \sqrt[2]{c} \eta\left(1-\beta_1\right)}{\sqrt[4]{c}-1}(\xi(t)-\xi(t+1)),$$

which leads to

$$\begin{gathered}\xi(t)=\mathcal{O}\left(\frac{1}{t}\right),\ \textit{i.e.},\ \mathcal{L}(\boldsymbol{w}(t))=\mathcal{O}\left(\frac{1}{t}\right), \\ \textit{and}\ \left\|\sqrt[4]{\varepsilon \mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)} \odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|^2=\mathcal{O}\left(\frac{1}{t}\right).\end{gathered}$$

Due to Eq. (58), we further have $\|\nabla\mathcal{L}(\boldsymbol{w}(t))\| = \mathcal{O}(t^{-1})$, which indicates

$$\begin{aligned}
\|\boldsymbol{w}(t)\| \leq & \|\boldsymbol{w}(1)\| + \sum_{s=1}^{t}\|\boldsymbol{w}(s+1)-\boldsymbol{w}(s)\| = \|\boldsymbol{w}(1)\| + \eta\sum_{s=1}^{t}\left\|\frac{\hat{\boldsymbol{m}}(s)}{\sqrt{\hat{\boldsymbol{\nu}}(s)+\varepsilon\mathbb{1}_d}}\right\| \\
\leq & \|\boldsymbol{w}(1)\| + \frac{\eta}{\sqrt{\varepsilon}}\sum_{s=1}^{t}\|\hat{\boldsymbol{m}}(s)\| = \|\boldsymbol{w}(1)\| + \frac{\eta}{\sqrt{\varepsilon}}\sum_{s=1}^{t}\frac{1}{(1-\beta^s)}\left\|\sum_{i=1}^{s}\beta^{s-i}\nabla\mathcal{L}(\boldsymbol{w}(i))\right\| \\
\leq & \|\boldsymbol{w}(1)\| + \frac{\eta}{\sqrt{\varepsilon}(1-\beta)}\sum_{s=1}^{t}\sum_{i=1}^{s}\beta^{s-i}\left\|\nabla\mathcal{L}(\boldsymbol{w}(i))\right\| \\
\leq & \|\boldsymbol{w}(1)\| + \frac{\eta}{\sqrt{\varepsilon}(1-\beta)^2}\sum_{s=1}^{t}\|\nabla\mathcal{L}(\boldsymbol{w}(s))\| = \mathcal{O}(\ln(t)).
\end{aligned}$$

Therefore, for any $\boldsymbol{x} \in \boldsymbol{S}$, we have $\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle = \mathcal{O}(\ln(t))$, which by $\ell$ is exponential-tailed leads to $\ell(\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle) = \Omega(t^{-1})$, and thus $\mathcal{L}(\boldsymbol{w}(t)) = \Theta(t^{-1})$. Also, since $\mathcal{L}(\boldsymbol{w}(t)) = \mathcal{O}(t^{-1})$, we have $\langle\boldsymbol{w}(t),\boldsymbol{x}\rangle = \Omega(\ln(t))$, which further leads to $\|\boldsymbol{w}(t)\| = \Omega(\ln(t))$, and thus $\|\boldsymbol{w}(t)\| = \Theta(\ln(t))$.

Finally, we have

$$\begin{aligned}
&\gamma\sum_{s=1}^{t}\beta^{t-s}\|\nabla\mathcal{L}(\boldsymbol{w}(s))\| = \frac{\gamma}{N}\sum_{s=1}^{t}\beta^{t-s}\left\|\sum_{\boldsymbol{x}\in\boldsymbol{S}}\ell'(\langle\boldsymbol{w}(s),\boldsymbol{x}\rangle)\boldsymbol{x}\right\| \\
\leq & -\frac{\gamma}{N}\sum_{s=1}^{t}\beta^{t-s}\sum_{\boldsymbol{x}\in\boldsymbol{S}}\ell'(\langle\boldsymbol{w}(s),\boldsymbol{x}\rangle) \leq -\frac{1}{N}\sum_{s=1}^{t}\beta^{t-s}\left\langle\sum_{\boldsymbol{x}\in\boldsymbol{S}}\ell'(\langle\boldsymbol{w}(s),\boldsymbol{x}\rangle)\boldsymbol{x},\gamma\hat{\boldsymbol{w}}\right\rangle \\
\leq & \left\|\sum_{s=1}^{t}\beta^{t-s}\nabla\mathcal{L}(\boldsymbol{w}(s))\right\|\|\gamma\hat{\boldsymbol{w}}\| = \|\boldsymbol{m}(t)\| \leq \sum_{s=1}^{t}\beta^{t-s}\|\nabla\mathcal{L}(\boldsymbol{w}(s))\|,
\end{aligned}$$

which leads to $\|\boldsymbol{m}(t)\| = \Theta(t^{-1})$. Similarly, we have $\boldsymbol{\nu}(t) = \mathcal{O}(t^{-2})$, component-wisely. As $\lim_{t\to\infty}\beta_1^t = 0$ and $\lim_{t\to\infty}\beta_2^t = 0$, we have

$$\|\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\| = \left\|\frac{\hat{\boldsymbol{m}}(t)}{\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}}\right\| = \Theta(t^{-1}).$$

The proof is completed. □

### D.3 Parameter dynamics

By Lemma 4, there exists a solution $\tilde{\boldsymbol{w}}$ as the solution of Eq. (10) with $C_3 = \frac{\eta}{(1-\beta)\sqrt{\varepsilon}}$. Define $\boldsymbol{r}(t)$ as

$$\boldsymbol{r}(t) \overset{\triangle}{=} \boldsymbol{w}(t) - \ln(t)\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}}, \tag{61}$$

and we only need to prove $\|\boldsymbol{r}(t)\|$ is bounded over time. We then prove $\boldsymbol{r}(t)$ has bounded norm. Specifically, we will prove the following lemma:

**Lemma 20.** *Let all conditions in Theorem 5 hold. Then, $\|\boldsymbol{r}(t)\|$ is bounded if and only if $g(t)$ is upper bounded, where $g(t)$ is defined as follows.*

$$\begin{aligned}
g(t) \overset{\triangle}{=} & \left\langle\boldsymbol{r}(t),(1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\rangle \\
& \cdot\frac{\beta_1}{1-\beta_1} + \frac{\sqrt{\varepsilon}}{2}\|\boldsymbol{r}(t)\|^2 - \frac{\beta_1}{1-\beta_1}\sum_{\tau=2}^{t}\langle\boldsymbol{r}(\tau)-\boldsymbol{r}(\tau-1), \\
& (1-\beta_1^{\tau-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(\tau-1)}\odot(\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1))\rangle.
\end{aligned}$$

*Furthermore, we have $\sum_{s=1}^{\infty}(g(t+1)-g(t))$ is upper bounded.*

Similar to GDM, the proof of Lemma 20 is divided into two parts, each focus on one claim of it. We start with the first claim.

**Lemma 21.** *Let all conditions in Theorem 5 hold. Then, $\|\boldsymbol{r}(t)\|$ is bounded if and only if $g(t)$ is upper bounded.*

*Proof.* Following the same routine as Lemma 11 and Lemma 14, we only need to prove

$$\lim_{t\to\infty}\left\|(1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|=0, \tag{62}$$

and

$$\sum_{\tau=2}^{\infty}\left|\langle\boldsymbol{r}(\tau)-\boldsymbol{r}(\tau-1),(1-\beta_1^{\tau-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(\tau-1)}\odot(\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1))\rangle\right|<\infty. \tag{63}$$

As for Eq. (62), by Lemma 19, we have

$$\begin{aligned}&\left\|(1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\|\\=&\mathcal{O}(t^{-1})=\boldsymbol{o}(1).\end{aligned}$$

As for Eq. (62), we have

$$\begin{aligned}&\left|\left\langle\boldsymbol{r}(\tau)-\boldsymbol{r}(\tau-1),(1-\beta_1^{\tau-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(\tau-1)}\odot(\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1))\right\rangle\right|\\=&\left|\left\langle\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)-\ln\frac{\tau}{\tau-1}\hat{\boldsymbol{w}},(1-\beta_1^{\tau-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(\tau-1)}\odot(\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1))\right\rangle\right|\\\le&\left|\left\langle\ln\frac{\tau}{\tau-1}\hat{\boldsymbol{w}},(1-\beta_1^{\tau-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(\tau-1)}\odot(\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1))\right\rangle\right|\\&+(1-\beta_1^{\tau-1})\left\|\sqrt[4]{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(\tau-1)}\odot(\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1))\right\|^2\\\overset{(\star)}{=}&\mathcal{O}(\tau^{-2}),\end{aligned}$$

where Eq. $(\star)$ is due to Lemma 19 and $\ln(\frac{\tau}{\tau-1})=\Theta(\tau^{-1})$.

The proof is completed. □

We conclude the proof of Theorem 5 by showing $g(t)$ is upper bounded.

**Lemma 22.** *Let all conditions in Theorem 5 hold. Then, $g(t)$ is upper bounded.*

*Proof.* $g(t)$ is upper bounded is equivalent to $\sum_{t=1}^{\infty}g(t+1)-g(t)<\infty$. We then prove this lemma by calculating $g(t+1)-g(t)$ directly.

$$\begin{aligned}&g(t+1)-g(t)\\=&\frac{\sqrt{\varepsilon}}{2}\|\boldsymbol{r}(t+1)\|^2+\frac{\beta_1}{1-\beta_1}\left\langle\boldsymbol{r}(t+1),(1-\beta_1^{t})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\rangle\\&-\left(\frac{\sqrt{\varepsilon}}{2}\|\boldsymbol{r}(t)\|^2+\frac{\beta_1}{1-\beta_1}\left\langle\boldsymbol{r}(t),(1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\rangle\right)\\&-\frac{\beta_1}{1-\beta_1}\langle\boldsymbol{r}(t+1)-\boldsymbol{r}(t),(1-\beta_1^{t})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\rangle\\=&\frac{\sqrt{\varepsilon}}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2+\frac{\beta_1}{1-\beta_1}\left\langle\boldsymbol{r}(t),(1-\beta_1^{t})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right.\\&\left.-(1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right\rangle+\sqrt{\varepsilon}\langle\boldsymbol{r}(t+1)-\boldsymbol{r}(t),\boldsymbol{r}(t)\rangle\\\overset{(\star)}{=}&\left\langle\boldsymbol{r}(t),-(1-\beta_1^{t})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))-\frac{\eta}{1-\beta}\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle\\&+\frac{\sqrt{\varepsilon}}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2+\sqrt{\varepsilon}\langle\boldsymbol{r}(t+1)-\boldsymbol{r}(t),\boldsymbol{r}(t)\rangle,\end{aligned}$$

where Eq. $(\star)$ is due to a simple rearranging of the update rule of Adam, i.e.,

$$\frac{\beta_1}{1-\beta_1}\left((1-\beta_1^t)\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))-(1-\beta_1^{t-1})\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t-1)}\odot(\boldsymbol{w}(t)-\boldsymbol{w}(t-1))\right)$$
$$=-\frac{\eta}{1-\beta_1}\nabla\mathcal{L}(\boldsymbol{w}(t))-(1-\beta_1^t)\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\,.$$

On the one hand, as $\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|=\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)-\ln\frac{t+1}{t}\hat{\boldsymbol{w}}\|=\mathcal{O}(t^{-1})$,

$$\sum_{t=1}^{\infty}\frac{\sqrt{\varepsilon}}{2}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2<\infty.$$

On the other hand,

$$\begin{aligned}
&\left\langle\boldsymbol{r}(t),-(1-\beta_1^t)\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))-\frac{\eta}{1-\beta}\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle\\
&+\sqrt{\varepsilon}\langle\boldsymbol{r}(t+1)-\boldsymbol{r}(t),\boldsymbol{r}(t)\rangle\\
=&\left\langle\boldsymbol{r}(t),-(1-\beta_1^t)\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))-\frac{\eta}{1-\beta}\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle\\
&+\sqrt{\varepsilon}\left\langle\boldsymbol{w}(t+1)-\boldsymbol{w}(t)-\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}},\boldsymbol{r}(t)\right\rangle\\
=&\left\langle\boldsymbol{r}(t),-(1-\beta_1^t)\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t)}\odot(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))+\sqrt{\varepsilon}(\boldsymbol{w}(t+1)-\boldsymbol{w}(t))\right\rangle\\
&+\left\langle\boldsymbol{r}(t),-\sqrt{\varepsilon}\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}}-\frac{\eta}{1-\beta}\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle\\
\overset{(\bullet)}{=}&\mathcal{O}(\beta_1^t+t^{-2})+\left\langle\boldsymbol{r}(t),-\sqrt{\varepsilon}\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}}-\frac{\eta}{1-\beta}\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle,
\end{aligned}$$

where Eq. $(\bullet)$ is due to $\hat{\boldsymbol{\nu}}(t)=\mathcal{O}(t^{-2})$.

Furthermore, following exactly the same routine as Lemma 13, we have

$$\sum_{t=1}^{\infty}\left\langle\boldsymbol{r}(t),-\sqrt{\varepsilon}\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}}-\frac{\eta}{1-\beta}\nabla\mathcal{L}(\boldsymbol{w}(t))\right\rangle<\infty.$$

The proof is completed. □

# E Implicit regularization of RMSProp (w/o. r) with decaying learning rate

This section collects the proof of Theorem 6. To begin with, we formally define RMSProp (w/o. r) as follows to facilitate latter analysis: for each $t\in\{0,1,2\cdots\}$, divide the sample set $S$ into $K$ subsets $\{\boldsymbol{B}(Kt+1),\cdots,\boldsymbol{B}(K(t+1))\}$ uniformly and i.i.d., and let

$$\begin{aligned}
&\boldsymbol{\nu}(0)=0,\boldsymbol{\nu}(\tau)=\beta_2\boldsymbol{\nu}(\tau-1)+(1-\beta_2)\left(\nabla\mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w}(\tau-1))\right)^2,\\
&\textit{(RMSProp (w/o. r))}:\ \boldsymbol{w}(\tau)=\boldsymbol{w}(\tau-1)-\eta_\tau\frac{\nabla\mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w}(\tau-1))}{\sqrt{\boldsymbol{\nu}(\tau)+\varepsilon\mathbb{1}_d}}.
\end{aligned}\tag{64}$$

Here the $\mathcal{L}_{\boldsymbol{B}(\tau)}$ is the individual loss average over $\boldsymbol{B}(\tau)$, i.e., $\mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w})=\frac{\sum_{(\boldsymbol{x},\boldsymbol{y})\in\boldsymbol{B}(\tau)}\ell(-\boldsymbol{y}\langle\boldsymbol{w},\boldsymbol{x}\rangle)}{b}$, where $b=\frac{N}{K}$ is the batch size. With Eq. (64), we restate the loss convergence result in [33] as the following proposition.

**Proposition 1** (Corollary 4.1 in [33], restated)**.** *Suppose $\ell$ is non-negative and $L$-smooth. Furthermore, assume that there exists a constant $D$, s.t., $\forall t\geq 0$, $\forall\boldsymbol{w}\in\mathbb{R}^d$*

$$\sum_{\tau=Kt+1}^{K(t+1)}\|\nabla\mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w})\|^2\leq D\|N\nabla\mathcal{L}(\boldsymbol{w})\|^2,\tag{65}$$

*and*

$$T_2\left(\beta_2\right) \triangleq \sqrt{\frac{10dK}{\beta_2^K}}dKD\left((1-\beta_2)\frac{\left(\frac{4K^2}{\beta_2^K}-1\right)}{2}+\left(\frac{1}{\sqrt{\beta_2^K}}-1\right)\right)\le\frac{\sqrt{2}-1}{2\sqrt{2}}.$$

*Then, RMSProp (w/o. r) with decaying learning rate $\eta_\tau=\frac{\eta_1}{\sqrt{\tau}}$ satisfies*

$$\sum_{t=0}^{T}\frac{1}{\sqrt{t+1}}\|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\|=\mathcal{O}(\ln(T)).$$

Combining Assumptions 1 and 2, we immediately get the following corollary:

**Corollary 8.** *Let Assumptions 1, 2, and 3. (S) hold. Let*

$$T_2\left(\beta_2\right) \triangleq \sqrt{\frac{10dK}{\beta_2^K}}\frac{dK}{b^2\gamma^2}\left((1-\beta_2)\frac{\left(\frac{4K^2}{\beta_2^K}-1\right)}{2}+\left(\frac{1}{\sqrt{\beta_2^K}}-1\right)\right)\le\frac{\sqrt{2}-1}{2\sqrt{2}}.$$

*Then, there exists a positive constant $C_1$ and $C_3$ independent of random sampling, s.t.,*

$$\sum_{t=0}^{T}\frac{1}{\sqrt{t+1}}\|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\|\le C_1\ln(T),\forall T\ge 0,$$

*and*

$$\mathcal{L}(\boldsymbol{w}(K(t+1)))\le\mathcal{L}(\boldsymbol{w}(Kt))-\frac{C_3}{\sqrt{t+1}}\|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\|+\frac{C_1C_3}{t+1}.$$

Applying relationship between $\ell$ and $\ell'$ of the exponentially-tailed loss, we can further obtain the loss convergent rate.

**Lemma 23.** *Let all the conditions in Theorem 6 hold. Then, we have*

$$\mathcal{L}(\boldsymbol{w}(Kt))=\mathcal{O}\left(\frac{1}{\sqrt{t}}\right).$$

*Proof.* To begin with, we show that there exists an increasing positive integer sequence $\{t_i\}_{i=1}^{\infty}$, and $\|\nabla\mathcal{L}(\boldsymbol{w}(Kt_i))\|\le 2C_1\frac{1}{\sqrt{t_i+1}}$ by reduction to absurdity. Otherwise, suppose there exists a positive integer $T_1$, such that $\|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\|>2C_1\frac{1}{\sqrt{t+1}}$, $\forall t\ge T_1$. Therefore, for $T\ge T_1$, we have

$$\sum_{t=0}^{T}\frac{1}{\sqrt{t+1}}\|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\|\ge\sum_{t=T_1}^{T}\frac{1}{\sqrt{t+1}}\|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\|$$
$$\ge 2C_1\sum_{t=T_1}^{T}\frac{1}{t+1}\ge 2C_1\ln\frac{T+2}{T_1+1}.$$

Let $T$ be large enough, we have $2C_1\ln\frac{T+2}{T_1+1}>C_1\ln T$, which contradicts Corollary 8.

Denote $\boldsymbol{T}=\{t>0:\|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\|\le 2C_1\frac{1}{\sqrt{t+1}}\}$. By the above discussion, $\boldsymbol{T}$ contains an increasing positive integer sequence. We then prove that if $s\in\boldsymbol{T}$ and $s>s^{\boldsymbol{T}}$, where $s^{\boldsymbol{T}}$ is defined as

$$s^{\boldsymbol{T}}\triangleq\max\left\{\left(\frac{2C_1}{C_g}\right)^2-1,(C_3\gamma)^2-1,\left(\frac{9C_1}{\gamma C_l}\right)^2-1,4,\frac{4}{C_3\gamma(\sqrt{\frac{3}{2}}-1)}\ln\frac{18\sqrt{\frac{3}{2}}}{\gamma}-2\right\}$$

($C_l$ and $C_g$ is defined in Corollary 2), there exists $s<r\le 2s$, and $r\in\boldsymbol{T}$. We slightly abuse the notation and let $r=\inf\{t:t\in\boldsymbol{T},t>s\}$. If $r=s+1$, this claim trivially holds. Otherwise, as $s\in\boldsymbol{T}$, we have

$$\|\nabla\mathcal{L}(\boldsymbol{w}(Ks))\|\le 2C_1\frac{1}{\sqrt{s+1}}\le C_g,$$

which by Corollary 2 further leads to

$$\mathcal{L}(\boldsymbol{w}(Ks)) \le \frac{4}{\gamma}\|\nabla\mathcal{L}(\boldsymbol{w}(Ks))\|.$$

Therefore, by Corollary 8, we have

$$\mathcal{L}(\boldsymbol{w}(K(s+1))) \le \mathcal{L}(\boldsymbol{w}(Ks)) + \frac{C_1C_3}{s+1} \le \frac{8C_1}{\gamma}\frac{1}{\sqrt{s+1}} + \frac{C_1C_3}{s+1} \le \frac{9C_1}{\gamma}\frac{1}{\sqrt{s+1}},$$

which by $s > s^{\boldsymbol{T}}$ further leads to

$$\mathcal{L}(\boldsymbol{w}(K(s+1))) \le C_l,$$

and thus $\mathcal{L}(\boldsymbol{w}(K(s+1))) \le \frac{4}{\gamma}\|\nabla\mathcal{L}(\boldsymbol{w}(K(s+1)))\|$. As $(s+1) \notin \boldsymbol{T}$, we have

$$\|\nabla\mathcal{L}(\boldsymbol{w}(K(s+1)))\| > 2C_1\frac{1}{\sqrt{s+2}},$$

which leads to

$$\mathcal{L}(\boldsymbol{w}(K(s+2))) \le \mathcal{L}(\boldsymbol{w}(K(s+1))) - \frac{C_3}{2\sqrt{s+2}}\|\nabla\mathcal{L}(\boldsymbol{w}(K(s+1)))\| \le \left(1 - \frac{C_3\gamma}{8\sqrt{s+2}}\right)\mathcal{L}(\boldsymbol{w}(K(s+1))),$$

and thus $\mathcal{L}(\boldsymbol{w}(K(s+2))) \le C_l$. By the inductive method, we have for any $j \in \{s+2, \cdots, r\}$,

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}(K(j))) &\le \Pi_{i=s+2}^{j}\left(1 - \frac{C_3\gamma}{8\sqrt{i}}\right)\mathcal{L}(\boldsymbol{w}(K(s+1))) \le e^{-\sum_{i=s+2}^{j}\frac{C_3\gamma}{8\sqrt{i}}}\mathcal{L}(\boldsymbol{w}(K(s+1)))\\ &\le e^{-\frac{C_3\gamma(\sqrt{j+1}-\sqrt{s+2})}{4}}\mathcal{L}(\boldsymbol{w}(K(s+1))) \le e^{-\frac{C_3\gamma(\sqrt{j+1}-\sqrt{s+2})}{4}}\frac{9C_1}{\gamma\sqrt{s+1}}.\end{aligned}$$

If $r > 2s$, applying $j = 2s$ into the above equation leads to

$$\mathcal{L}(\boldsymbol{w}(K(2s))) \le e^{-\frac{C_3\gamma(\sqrt{2s+1}-\sqrt{s+2})}{4}}\frac{9C_1}{\gamma\sqrt{s+1}} \le C_l,$$

and

$$\|\nabla\mathcal{L}(\boldsymbol{w}(K(2s)))\| \le e^{-\frac{C_3\gamma(\sqrt{2s+1}-\sqrt{s+2})}{4}}\frac{36C_1}{\gamma\sqrt{s+1}} \le \frac{2C_1}{\sqrt{2s+1}},$$

which leads to $2s \in \boldsymbol{T}$, and contradicts the definition of $r$. Therefore, we have $r \le 2s$.

As $\boldsymbol{T}$ contains an increasing integer sequence, there exists an $s_0$, s.t., $s_0 \in \boldsymbol{T}$ and $s_0 > s^{\boldsymbol{T}}$. Let $t$ be any positive integer larger than $s_0$ and let $t'$ be the largest integer smaller than $t$ and belongs to $\boldsymbol{T}$. We have $t' \ge s_0$, and $t \le 2t'$ by the above discussion. Therefore, we have

$$\mathcal{L}(Kt) \le \frac{9C_1}{\gamma}\frac{1}{\sqrt{t'+1}} \le \frac{9\sqrt{2}C_1}{\gamma}\frac{1}{\sqrt{t+2}}.$$

The proof is completed. □

As a corollary, we can obtain an asymptotic estimation of $\nabla\mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w}(\tau))$.

**Corollary 9.** *Let all the conditions in Theorem 6 hold. Then, we have*

$$\|\nabla\mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w}(\tau))\| = \mathcal{O}\left(\frac{1}{\sqrt{\tau}}\right).$$

*Proof.* Let $\tau > K(s^{\boldsymbol{T}}+1)$, where $s^{\boldsymbol{T}}$ is defined as Lemma 23. Let $t = \lceil\frac{\tau}{K}\rceil > s^{\boldsymbol{T}}$ and $s = \tau - Kt$. Then, we have

$$\|\nabla\mathcal{L}(Kt)\| \le C_l,$$

and

$$\|\nabla\mathcal{L}(\boldsymbol{w}(Kt))\| \le 4\mathcal{L}(\boldsymbol{w}(Kt)) \le \frac{36\sqrt{2}C_1}{\gamma}\frac{1}{\sqrt{t+2}}.$$

On the other hand, we have

$$\|\nabla \mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w}(Kt))\| \le \frac{1}{\gamma}\frac{N}{b}\|\nabla \mathcal{L}(\boldsymbol{w}(Kt))\| \le \frac{N}{b\gamma}\frac{36\sqrt{2}C_1}{\gamma}\frac{1}{\sqrt{t+2}}.$$

As $\mathcal{L}_{\boldsymbol{B}(\tau)}$ is $H$ smooth, we further have

$$\begin{aligned}\|\nabla \mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w}(Kt)) - \nabla \mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w}(\tau))\| \le & H\left\|\sum_{i=0}^{s-1}\eta_{Kt+i+1}\frac{\nabla \mathcal{L}_{\boldsymbol{B}(Kt+i+1)}(\boldsymbol{w}(Kt+i))}{\sqrt{\hat{\boldsymbol{\nu}}(Kt+i+1)+\varepsilon\mathbb{1}_d}}\right\| \\ \le & \frac{KH}{\sqrt{1-\beta_2}\sqrt{Kt+1}}.\end{aligned}$$

Combining the above two equations, we have

$$\begin{aligned}\|\nabla \mathcal{L}_{\boldsymbol{B}(\tau)}(\boldsymbol{w}(\tau))\| \le & \frac{KH}{\sqrt{1-\beta_2}\sqrt{Kt+1}} + \frac{N}{b\gamma}\frac{36\sqrt{2}C_1}{\gamma}\frac{1}{\sqrt{t+2}} \\ \le & \frac{KH}{\sqrt{1-\beta_2}\sqrt{\tau-K}} + \frac{N}{b\gamma}\frac{36\sqrt{2}C_1}{\gamma}\frac{\sqrt{K}}{\sqrt{\tau+K}}.\end{aligned}$$

The proof is completed. □

Using Corollary 9, we are now able to prove Theorem 6.

*Proof of Theorem 6 (for almost every dataset).* To begin with, define $\boldsymbol{r}(t) \triangleq \boldsymbol{w}(t) - \frac{1}{2}\ln\left(\frac{\eta}{K}t\right)\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}} - \frac{1}{2}\boldsymbol{n}(t)$, where $\tilde{\boldsymbol{w}}$ is the solution of Eq. (10) with $C_3 = \frac{2}{N\sqrt{\varepsilon}}$, and $\boldsymbol{n}(t)$ is given by Lemma 5. As in the case of GDM, SGDM, and Adam (w/s), $\boldsymbol{w}(t) - \frac{1}{2}\ln\left(\frac{\eta}{K}t\right)$ has bounded norm if and only if $\boldsymbol{r}(t)$ has bounded norm. Also, it is a sufficient condition to ensure $\|\boldsymbol{r}(t)\|$ is bounded that both $A \triangleq \sum_{t=1}^{\infty}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2 < \infty$ and $B \triangleq \sum_{t=1}^{\infty}\langle \boldsymbol{r}(t+1)-\boldsymbol{r}(t), \boldsymbol{r}(t)\rangle < \infty$.

As for $A$, we have

$$\begin{aligned}A =& \sum_{t=1}^{\infty}\|\boldsymbol{r}(t)-\boldsymbol{r}(t-1)\|^2 \\ =& \sum_{t=1}^{\infty}\left\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)-\frac{1}{2}\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}}-\frac{1}{2}\boldsymbol{n}(t+1)+\frac{1}{2}\boldsymbol{n}(t)\right\|^2 \\ \le & 3\left(\sum_{t=1}^{\infty}\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\|^2 + \frac{1}{2}\sum_{t=1}^{\infty}\ln\left(\frac{t+1}{t}\right)\|\hat{\boldsymbol{w}}\|^2 + \sum_{t=1}^{\infty}\frac{1}{4}\|-\boldsymbol{n}(t+1)+\boldsymbol{n}(t)\|^2\right) \\ \overset{(*)}{<} & \infty,\end{aligned}$$

where Inequality $(*)$ is due to

$$\|\boldsymbol{w}(t+1)-\boldsymbol{w}(t)\| = \frac{\eta_1}{\sqrt{t+1}}\left\|\frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon\mathbb{1}_d+\hat{\boldsymbol{\nu}}(t+1)}}\right\| \le \frac{\eta_1}{\sqrt{t+1}\sqrt{\varepsilon}}\left\|\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))\right\| = \mathcal{O}\left(\frac{1}{t}\right),$$

$\ln\left(\frac{t+1}{t}\right) = \mathcal{O}\left(\frac{1}{t}\right)$, and $\|-\boldsymbol{n}(t+1)+\boldsymbol{n}(t)\| = \mathcal{O}\left(\frac{1}{t}\right)$.

As for $B$, we have

$$\begin{aligned}
B =& \sum_{t=1}^{\infty} \left\langle \boldsymbol{w}(t+1) - \boldsymbol{w}(t) - \frac{1}{2}\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}} - \frac{1}{2}\boldsymbol{n}(t+1) + \frac{1}{2}\boldsymbol{n}(t), \boldsymbol{r}(t) \right\rangle \\
=& \sum_{t=1}^{\infty} \left\langle \boldsymbol{w}(t+1) - \boldsymbol{w}(t) - \frac{N}{2bt} \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(t) \cap \boldsymbol{S}_s} \boldsymbol{v}_i \boldsymbol{x}_i, \boldsymbol{r}(t) \right\rangle \\
=& \sum_{t=1}^{\infty} \left\langle -\frac{1}{\sqrt{t+1}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}} - \frac{N}{2bt} \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(t) \cap \boldsymbol{S}_s} \boldsymbol{v}_i \boldsymbol{x}_i, \boldsymbol{r}(t) \right\rangle \\
=& \sum_{t=1}^{\infty} \left\langle -\frac{1}{\sqrt{t+1}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}} + \frac{1}{\sqrt{t}} \frac{\nabla \mathcal{L}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}}, \boldsymbol{r}(t) \right\rangle \\
&+ \sum_{t=1}^{\infty} \left\langle -\frac{1}{\sqrt{t}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}} + \frac{1}{\sqrt{t}\sqrt{\varepsilon}} \nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)), \boldsymbol{r}(t) \right\rangle \\
&+ \sum_{t=1}^{\infty} \left\langle -\frac{1}{\sqrt{t}\sqrt{\varepsilon}} \nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)) - \frac{N}{2bt} \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(t) \cap \boldsymbol{S}_s} \boldsymbol{v}_i \boldsymbol{x}_i, \boldsymbol{r}(t) \right\rangle. \qquad (66)
\end{aligned}$$

On the other hand, as

$$\|\boldsymbol{w}(t)\| \le \|\boldsymbol{w}(1)\| + \sum_{s=1}^{t-1} \frac{1}{\sqrt{s+1}} \left\| \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(s))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(s+1)}} \right\| = \mathcal{O}(\ln(t)),$$

we have $\|\boldsymbol{r}(t)\| = \mathcal{O}(\ln(t))$. Also, we have $\|\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))^2\| = \mathcal{O}(\frac{1}{t})$, and thus $\left\| \frac{1}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t)}} - \frac{1}{\sqrt{\varepsilon \mathbb{1}_d}} \right\| = \mathcal{O}(\frac{1}{t})$. Combining these estimations, we have

$$\left\langle -\frac{1}{\sqrt{t+1}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}} + \frac{1}{\sqrt{t}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}}, \boldsymbol{r}(t) \right\rangle = \mathcal{O}\left(\frac{\ln t}{t^{\frac{5}{2}}}\right).$$

and

$$\left\langle -\frac{1}{\sqrt{t}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}} + \frac{1}{\sqrt{t}\sqrt{\varepsilon}} \nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)), \boldsymbol{r}(t) \right\rangle = \mathcal{O}\left(\frac{\ln t}{t^2}\right),$$

Therefore,

$$\sum_{t=1}^{\infty} \left\langle -\frac{1}{\sqrt{t+1}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}} + \frac{1}{\sqrt{t}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}}, \boldsymbol{r}(t) \right\rangle < \infty,$$

$$\sum_{t=1}^{\infty} \left\langle -\frac{1}{\sqrt{t}} \frac{\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t))}{\sqrt{\varepsilon \mathbb{1}_d + \hat{\boldsymbol{\nu}}(t+1)}} + \frac{1}{\sqrt{t}\sqrt{\varepsilon}} \nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)), \boldsymbol{r}(t) \right\rangle < \infty.$$

As for the last term in Eq. (66), we have

$$\begin{aligned}
&\left\langle -\frac{1}{\sqrt{t}\sqrt{\varepsilon}} \nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)) - \frac{N}{2bt} \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(t) \cap \boldsymbol{S}_s} \boldsymbol{v}_i \boldsymbol{x}_i, \boldsymbol{r}(t) \right\rangle \\
=& \frac{1}{b\sqrt{t}\sqrt{\varepsilon}} \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s} \left( -\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) - \frac{1}{\sqrt{t}} e^{-\langle \tilde{\boldsymbol{w}}, \boldsymbol{x}_i \rangle} \right) \langle \boldsymbol{r}(t), \boldsymbol{x}_i \rangle \\
&+ \frac{1}{b\sqrt{t}\sqrt{\varepsilon}} \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(s) \cap \boldsymbol{S}_s^c} \langle \boldsymbol{r}(t), -\ell'(\langle \boldsymbol{w}(t), \boldsymbol{x}_i \rangle) \boldsymbol{x}_i \rangle.
\end{aligned}$$

Then, following the same routine as Lemma 15, we have

$$\sum_{t=1}^{\infty} \left\langle -\frac{1}{\sqrt{t}\sqrt{\varepsilon}} \nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)) - \frac{N}{2bt} \sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(t) \cap \boldsymbol{S}_s} \boldsymbol{v}_i \boldsymbol{x}_i, \boldsymbol{r}(t) \right\rangle < \infty.$$

The proof is completed. □

# F Applications & Extensions

## F.1 Deriving the conclusion for every dataset

In Sections 4, 5, and 6, we only derive the implicit regularization for almost every dataset, but not all the separable datasets. In this section, we show that the "almost every" condition can be removed as the following theorem.

**Theorem 7.** *We have the following conclusions:*

- *For GDM, let all the conditions in Theorem 2 hold. Then, GDM converges to the $L^2$ max-margin solution;*
- *For SGDM sampling with replacement, let all the conditions in Theorem 3 hold. Then, SGDM (w/. r) converges to the $L^2$ max-margin solution (the same for SGDM (w/o. r), except a different learning rate upper bound);*
- *For deterministic Adam, let all the conditions in Theorem 5 hold. Then, Adam (w/s) converges to the $L^2$ max-margin solution;*
- *For RMSProp (w/o. r), let all the conditions in Theorem 6 hold. Then, RMSProp (w/o. r) converges to the $L^2$ max-margin solution.*

It can be easily observe that to prove Theorem 7, the analysis of Stage I of every optimizer still works. Therefore, we only need to change Stage II. As the analyses of Stage II are highly overlapped for different optimizers, we only provide a proof of GDM.

To begin with, we present some notations and results on the structure of the separable dataset from [34].

Let $\bar{\boldsymbol{S}}_0 = \{1, \cdots, N\}$, and $\bar{\mathbf{P}}_0 = \mathbb{I}_{d\times d}$. We then recursively define the index sets $\boldsymbol{S}_m^+$, $\boldsymbol{S}_m^=$, $\boldsymbol{S}_m$, and $\bar{\boldsymbol{S}}_m$:

$$\boldsymbol{S}_m^+ = \left\{ i \in \bar{\boldsymbol{S}}_{m-1} \mid \left\langle \hat{\boldsymbol{w}}_m, \bar{\mathbf{P}}_{m-1}\boldsymbol{x}_i \right\rangle > 1 \right\},$$
$$\boldsymbol{S}_m^= = \left\{ i \in \bar{\boldsymbol{S}}_{m-1} \mid \left\langle \hat{\boldsymbol{w}}_m, \bar{\mathbf{P}}_{m-1}\boldsymbol{x}_i \right\rangle = 1 \right\} = \bar{\boldsymbol{S}}_{m-1}/\boldsymbol{S}_m^+,$$
$$\boldsymbol{S}_m = \left\{ i \in \boldsymbol{S}_m^= \mid \exists \boldsymbol{\alpha} \in \mathbb{R}^N_{\geq 0} : \hat{\boldsymbol{w}}_m = \sum_{k=1}^N \alpha_k \bar{\mathbf{P}}_{m-1}\boldsymbol{x}_k, \alpha_i > 0, \forall j \notin \mathcal{S}_m^= : \alpha_j = 0 \right\},$$
$$\bar{\boldsymbol{S}}_m = \boldsymbol{S}_m^= / \boldsymbol{S}_m.$$

where $\bar{\mathbf{P}}_m = \bar{\mathbf{P}}_{m-1}\left(\mathbf{I}_d - \boldsymbol{S}_{\boldsymbol{S}_m}\boldsymbol{S}^{\dagger}_{\boldsymbol{S}_m}\right)$ (we also denote $\mathbf{P}_m = \mathbb{I}_{d\times d} - \bar{\mathbf{P}}_m$), and $\hat{\boldsymbol{w}}_m$ is defined as the max-margin solution of dataset $\bar{\mathbf{P}}_{m-1}\boldsymbol{S}_{\bar{\boldsymbol{S}}_{m-1}}$ (that is, the transferred data $x_i$ with index in $\bar{\boldsymbol{S}}_{m-1}$ projected through matrix $\bar{\mathbf{P}}_{m-1}$):

$$\hat{\boldsymbol{w}}_m = \underset{\boldsymbol{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\boldsymbol{w}\|^2, \text{ s.t. } \langle \boldsymbol{w}, \bar{\mathbf{P}}_{m-1}\boldsymbol{x}_i \rangle \geq 1, \forall i \in \bar{\boldsymbol{S}}_{m-1}. \tag{67}$$

The existence of the $\boldsymbol{\alpha}$ is guaranteed by the KKT condition of Eq. (67). The above procedure will produce a sequence $\hat{\boldsymbol{w}}_1$, $\hat{\boldsymbol{w}}_2$, $\cdots$, and will stop at $\hat{\boldsymbol{w}}_M$ if $\bar{\boldsymbol{S}}_M$ is empty (if the sequence is infinite, we let $M = \infty$). For every $i \leq M$, we have $\hat{\boldsymbol{w}}_i$ is non-zero, and $\boldsymbol{S}_i$ is non-empty, which leads to $|\bar{\boldsymbol{S}}_i| < |\bar{\boldsymbol{S}}_{i+1}|$, and $M \leq N$.

The following lemma characterize the structure of the dataset.

**Lemma 24** (Lemma 17, [34])**.** $\forall \boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{S}_1|}_{>0}$*, we can find a unique* $\tilde{\boldsymbol{w}}_1$*, such that*

$$\sum_{i\in\boldsymbol{S}_1} \boldsymbol{x}_i \beta_i \exp\left(-\langle \boldsymbol{x}_i, \tilde{\boldsymbol{w}}_1\rangle\right) = \hat{\boldsymbol{w}}_1,$$

*and* $\tilde{\boldsymbol{w}}_1 \in \operatorname{Col}(\boldsymbol{S}_{\boldsymbol{S}_1})$.

With Lemma 24, we then define $\tilde{\boldsymbol{w}}_m$ as the solution of

$$\frac{\eta}{1-\beta}\sum_{i\in\mathcal{S}_m}\exp\left(-\sum_{k=1}^{m}\langle\tilde{\boldsymbol{w}}_k,\boldsymbol{x}_i\rangle\right)\bar{\mathbf{P}}_{m-1}\boldsymbol{x}_i=\hat{\boldsymbol{w}}_m,$$

with $\mathbf{P}_{m-1}\tilde{\boldsymbol{w}}_m=0$ and $\bar{\mathbf{P}}_m\tilde{\boldsymbol{w}}_m=0$. We also define $\tilde{\boldsymbol{w}}=\sum_{m=1}^{M}\tilde{\boldsymbol{w}}_m$.

We then have the following lemma:

**Lemma 25** (Lemma 18, [34])**.** *$\forall m>k\geq 1$, the equations*

$$\sum_{i\in\boldsymbol{S}_m}\exp\left(-\langle\tilde{\boldsymbol{w}},\boldsymbol{x}_i\rangle\right)\mathbf{P}_{m-1}\boldsymbol{x}_i=\sum_{k=1}^{m-1}\left[\sum_{i\in\mathcal{S}_k}\exp\left(-\langle\tilde{\boldsymbol{w}},\boldsymbol{x}_i\right)\boldsymbol{x}_i\boldsymbol{x}_i^\top\right]\check{\boldsymbol{w}}_{k,m}$$

*under the constraints $\mathbf{P}_{k-1}\check{\boldsymbol{w}}_{k,m}=0$ and $\bar{\mathbf{P}}_k\check{\boldsymbol{w}}_{k,m}=0$ have the unique solution $\check{\boldsymbol{w}}_{k,m}$.*

We then denote

$$\boldsymbol{r}(t)=\boldsymbol{w}(t)-\hat{\boldsymbol{w}}_1\log(t)-\tilde{\boldsymbol{w}}-\tau(t),$$

where $\tau(t)=\left(\sum_{m=2}^{M}\hat{\boldsymbol{w}}_m\log^{\circ m}(t)+\sum_{m=1}^{M}\sum_{k=1}^{m-1}\frac{\check{\boldsymbol{w}}_{k,m}}{\prod_{r=k}^{m-1}\log^{\circ r}(t)}\right)$

Similar to Eq. 28, we define

$$g(t)\overset{\triangle}{=}\frac{1}{2}\|\boldsymbol{r}(t)\|^2+\frac{\beta}{1-\beta}\langle\boldsymbol{r}(t),\boldsymbol{w}(t)-\boldsymbol{w}(t-1)\rangle-\frac{\beta}{1-\beta}\sum_{\tau=2}^{t}\langle\boldsymbol{r}(\tau)-\boldsymbol{r}(\tau-1),\boldsymbol{w}(\tau)-\boldsymbol{w}(\tau-1)\rangle.$$

We then have the following lemma parallel to Lemma 11:

**Lemma 26.** *Let all conditions in Theorem 2 hold. We have $\sup_t g(t)$ is finite. Furthermore, $\sup_t\|\boldsymbol{r}(t)\|$ is finite if and only if $\sup_t g(t)$ is finite, and consequently $\sup_t\|\boldsymbol{r}(t)\|$ is finite.*

*Proof.* The proof of the second argument follows the same routine as the proof of the first argument in Lemma 11, and we omit it here.

As for the first argument, we have

$$\sum_{t=1}^{\infty}(g(t+1)-g(t))=\frac{1}{2}\sum_{t=1}^{\infty}\|\boldsymbol{r}(t+1)-\boldsymbol{r}(t)\|^2$$
$$+\sum_{t=1}^{\infty}\left\langle\boldsymbol{r}(t),-\frac{\eta}{1-\beta}\nabla\mathcal{L}(\boldsymbol{w}(t))-\ln\frac{t+1}{t}\hat{\boldsymbol{w}}_1-(\tau(t+1)-\tau(t))\right\rangle,$$

where the first term can be shown to be finite similar to Lemma 11, while the second term is finite by Lemma 14 in [34].

The proof is completed. □

## F.2 Implicit regularization of SGDM (w/o. r)

This section provides formal description of the implicit regularization of SGDM (w/o. r) and its corresponding proof. To begin with, we would like to provide a formal definition of SGDM (w/o. r). SGDM (w/o. r) differs from SGDM by applying sampling without replacement to obtain $\boldsymbol{B}(t)$ in Eq. (2). Specifically, let $K=\frac{N}{b}$. For any $T\geq 0$, we call time series $\{KT+1,\cdots,KT+K\}$ the $(T+1)$-th epoch, and during the $T+1$-th epoch, the dataset $\boldsymbol{S}$ is randomly uniformly divided into $K$ parts $\{\boldsymbol{B}(KT+1),\cdots,\boldsymbol{B}(KT+K)\}$, with $\bigcup_{t=KT+1}^{KT+K}\boldsymbol{B}(t)=\boldsymbol{S}$. The implicit regularization of SGDM (w/o. r) is then stated as the following theorem:

**Theorem 8.** *Let Assumptions 1, 2, and 3. (S) hold. Let learning rate $\eta$ be small enough, and $\beta\in[0,1)$.Then, for almost every dataset $\boldsymbol{S}$, SGDM (w/o. r) satisfies $\boldsymbol{w}(t)-\ln(t)\hat{\boldsymbol{w}}$ is bounded as $t\to\infty$, and $\lim_{t\to\infty}\frac{\boldsymbol{w}(t)}{\|\boldsymbol{w}(t)\|}=\frac{\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|}$.*

The without-replacement sampling method leads to the direction of every trajectory of mini-SGDM converge to the max-margin solution, compared to the same conclusion holds for SGDM a.s.. We prove the theorem following the same framework of GDM, by proceeding with two stages.

**Stage I.** The following lemma proves $\mathcal{L}(\boldsymbol{u}(t))$ is an Lyapunov function for SGDM (w/o. r) and without the a.s. condition.

**Lemma 27.** *Let all conditions in Theorem 8 hold. Then, we have*

$$\mathcal{L}(\boldsymbol{u}(t+1)) \le \mathcal{L}(\boldsymbol{u}(1)) - \Omega(\eta)\sum_{s=1}^{t}\|\nabla\mathcal{L}(\boldsymbol{w}(s))\|^2.$$

*Proof.* By the Taylor Expansion of $\mathcal{L}(\boldsymbol{u}(t+1))$ at $\boldsymbol{u}(t)$, we have

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{u}(KT+T+1))\\
\le&\mathcal{L}(\boldsymbol{u}(KT+1)) - \eta\left\langle \nabla\mathcal{L}(\boldsymbol{u}(KT+1)), \sum_{t=1}^{K}\nabla\mathcal{L}_{\boldsymbol{B}(t+KT)}(\boldsymbol{w}(t+KT))\right\rangle\\
+&\frac{H\eta^2}{2}\left\|\sum_{t=1}^{K}\mathcal{L}_{\boldsymbol{B}(t+KT)}(\boldsymbol{w}(t+KT))\right\|^2.
\end{aligned}\tag{68}$$

On the other hand, for any $t\in\{2,\cdots,K\}$, we have

$$\begin{aligned}
&\boldsymbol{w}(KT+t)-\boldsymbol{w}(KT+1) = \eta\sum_{s=1}^{t}\left(\sum_{\ell=1}^{KT+s}\beta^{KT+s-\ell}\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))\right)\\
=&\eta\sum_{s=1}^{t}\left(\sum_{\ell=KT+1}^{KT+s}\beta^{KT+s-\ell}\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))\right)+\eta\sum_{s=1}^{t}\left(\sum_{\ell=1}^{KT}\beta^{KT+s-\ell}\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))\right)\\
=&\eta\sum_{\ell=1}^{t}\frac{1-\beta^{t-\ell+1}}{1-\beta}\nabla\mathcal{L}_{\boldsymbol{B}(KT+\ell)}(\boldsymbol{w}(KT+\ell))+\eta\frac{\beta(1-\beta^t)}{1-\beta}\sum_{\ell=1}^{KT}\beta^{KT-\ell}\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))\\
=&\eta\sum_{\ell=1}^{t}\frac{1-\beta^{t-\ell+1}}{1-\beta}\nabla\mathcal{L}_{\boldsymbol{B}(KT+\ell)}(\boldsymbol{w}(KT+\ell))-\eta\sum_{\ell=1}^{t}\frac{1-\beta^{t-\ell+1}}{1-\beta}\nabla\mathcal{L}_{\boldsymbol{B}(KT+\ell)}(\boldsymbol{w}(KT+1))\\
+&\eta\frac{\beta(1-\beta^t)}{1-\beta}\sum_{\ell=1}^{KT}\beta^{KT-\ell}\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))+\eta\sum_{\ell=1}^{t}\frac{1-\beta^{t-\ell+1}}{1-\beta}\nabla\mathcal{L}_{\boldsymbol{B}(KT+\ell)}(\boldsymbol{w}(KT+1)),
\end{aligned}$$

which by $\eta$ is small enough further indicates

$$\begin{aligned}
&\|\boldsymbol{w}(KT+t)-\boldsymbol{w}(KT+1)\|\\
\le&\eta\left\|\sum_{\ell=1}^{t}\frac{1-\beta^{t-\ell+1}}{1-\beta}\nabla\mathcal{L}_{\boldsymbol{B}(KT+\ell)}(\boldsymbol{w}(KT+\ell))-\sum_{\ell=1}^{t}\frac{1-\beta^{t-\ell+1}}{1-\beta}\nabla\mathcal{L}_{\boldsymbol{B}(KT+\ell)}(\boldsymbol{w}(KT+1))\right\|\\
+&\eta\left\|\frac{\beta(1-\beta^t)}{1-\beta}\sum_{\ell=1}^{KT}\beta^{KT-\ell}\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))\right\|+\eta\left\|\sum_{\ell=1}^{t}\frac{1-\beta^{t-\ell+1}}{1-\beta}\nabla\mathcal{L}_{\boldsymbol{B}(KT+\ell)}(\boldsymbol{w}(KT+1))\right\|\\
=&\mathcal{O}(\eta)\sum_{\ell=2}^{t}\|\boldsymbol{w}(KT+\ell)-\boldsymbol{w}(KT+1)\|+\mathcal{O}(\eta)\left(\sum_{\ell=1}^{KT}\beta^{KT-\ell}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))\right\|\right)\\
&+\mathcal{O}(\eta)\left\|\nabla\mathcal{L}(\boldsymbol{w}(KT+1))\right\|.
\end{aligned}$$

Applying the same analysis to $\|\boldsymbol{w}(KT+t-1)-\boldsymbol{w}(KT+1)\|$ recursively, we finally obtain

$$\begin{aligned}
&\|\boldsymbol{w}(KT+t)-\boldsymbol{w}(KT+1)\|\\
\le&\mathcal{O}(\eta)\left(\sum_{\ell=1}^{KT}\beta^{KT-\ell}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))\right\|\right)+\mathcal{O}(\eta)\left\|\nabla\mathcal{L}(\boldsymbol{w}(KT+1))\right\|.
\end{aligned}\tag{69}$$

Applying Eq. (69) to the $\left\|\nabla\mathcal{L}_{\boldsymbol{B}(\ell)}(\boldsymbol{w}(\ell))\right\|$ in Eq. (69) ($\forall \ell \in [1, KT]$) iterative and choosing $\eta$ to be small enough, we further have

$$
\begin{aligned}
&\|\boldsymbol{w}(KT+t)-\boldsymbol{w}(KT+1)\| \\
\leq&\mathcal{O}(\eta)\left(\sum_{\ell=0}^{T-1}\sqrt{\beta^{K(T-\ell)}}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(K\ell+1)}(\boldsymbol{w}(K\ell+1))\right\|\right)+\mathcal{O}(\eta)\left\|\nabla\mathcal{L}(\boldsymbol{w}(KT+1))\right\| \\
=&\mathcal{O}(\eta)\left(\sum_{\ell=0}^{T}\sqrt{\beta^{K(T-\ell)}}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(K\ell+1)}(\boldsymbol{w}(K\ell+1))\right\|\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sum_{t=1}^{K}\nabla\mathcal{L}_{\boldsymbol{B}(t+KT)}(\boldsymbol{w}(t+KT)) \\
=&\sum_{t=1}^{K}\nabla\mathcal{L}_{\boldsymbol{B}(t+KT)}(\boldsymbol{w}(t))+\mathcal{O}\left(\eta\left(\sum_{\ell=0}^{T}\sqrt{\beta^{K(T-\ell)}}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(K\ell+1)}(\boldsymbol{w}(K\ell+1))\right\|\right)\right) \\
=&K\nabla\mathcal{L}(\boldsymbol{w}(t))+\mathcal{O}\left(\eta\left(\sum_{\ell=0}^{T}\sqrt{\beta^{K(T-\ell)}}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(K\ell+1)}(\boldsymbol{w}(K\ell+1))\right\|\right)\right). \qquad (70)
\end{aligned}
$$

Similarly, one can obtain

$$
\begin{aligned}
&\nabla\mathcal{L}(\boldsymbol{u}(KT+1)) \\
=&\nabla\mathcal{L}(\boldsymbol{w}(KT+1))+\mathcal{O}\left(\|\boldsymbol{w}(KT+1)-\boldsymbol{w}(KT)\|\right) \\
=&\nabla\mathcal{L}(\boldsymbol{w}(KT+1))+\mathcal{O}\left(\eta\left(\sum_{\ell=0}^{T}\sqrt{\beta^{K(T-\ell)}}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(K\ell+1)}(\boldsymbol{w}(K\ell+1))\right\|\right)\right). \qquad (71)
\end{aligned}
$$

Applying Eq. (70) and Eq. (71) back to the Taylor Expansion (Eq. (68)), we have

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{u}(KT+T+1)) \\
\leq&\mathcal{L}(\boldsymbol{u}(KT+1))-\Omega(\eta)\left\langle\nabla\mathcal{L}(\boldsymbol{w}(KT+1)),\nabla\mathcal{L}(\boldsymbol{w}(KT+1))\right\rangle \\
&+\mathcal{O}\left(\eta^2\left(\sum_{\ell=0}^{T}\sqrt{\beta^{K(T-\ell)}}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(K\ell+1)}(\boldsymbol{w}(K\ell+1))\right\|\right)^2\right) \\
\leq&\mathcal{L}(\boldsymbol{u}(KT+1))-\Omega(\eta)\left\langle\nabla\mathcal{L}(\boldsymbol{w}(KT+1)),\nabla\mathcal{L}(\boldsymbol{w}(KT+1))\right\rangle \\
&+\mathcal{O}\left(\eta^2\left(\sum_{\ell=0}^{T}\sqrt{\beta^{K(T-\ell)}}\left\|\nabla\mathcal{L}_{\boldsymbol{B}(K\ell+1)}(\boldsymbol{w}(K\ell+1))\right\|^2\right)\right).
\end{aligned}
$$

Summing the above inequality over $T$ and setting $\eta$ small enough leads to the conclusion.

The proof is completed. □

### F.3 Extension to the multi-class classification problem

As mentioned in Section 3, despite all the previous analyses are aimed at the binary classification problem, they can be naturally extended to the analyses multi-class classification problem. Specifically, in the linear multi-class classification problem, for any $(\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}^{d_X}\times\{1,\cdots,C\}$ in the sample set $\boldsymbol{S}$, the (individual) logistic loss with parameter $\boldsymbol{W}\in\mathbb{R}^{C\times d_X}$ is denoted as

$$
\ell(\boldsymbol{y},\boldsymbol{W}\boldsymbol{x})=\ln\frac{e^{\boldsymbol{W}_{\boldsymbol{y},}\boldsymbol{x}}}{\sum_{i=1}^{C}e^{\boldsymbol{W}_{i,}\boldsymbol{x}}}.
$$

Correspondingly, dataset $\boldsymbol{S}$ is separable if there exists a parameter $\boldsymbol{W}$, such that $\forall(\boldsymbol{x},\boldsymbol{y})\in\boldsymbol{S}$, we have $\boldsymbol{W}_{\boldsymbol{y},}\boldsymbol{x}>\boldsymbol{W}_{i,}\boldsymbol{x}, \forall i\neq\boldsymbol{y}$. The multi-class $L^2$ max-margin problem is then defined as

$$
\min\|\boldsymbol{W}\|_F,\ \mathit{subject\ to}:\ \boldsymbol{W}_{\boldsymbol{y},}\boldsymbol{x}\geq\boldsymbol{W}_{i,}\boldsymbol{x}+1,\forall(\boldsymbol{x},\boldsymbol{y})\in\boldsymbol{S},i\neq\boldsymbol{y},
$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Denote $\hat{\boldsymbol{W}}$ as the $L^2$ max-margin solution, we have SGDM and Adam  (w/s) still converges to the direction of $\hat{\boldsymbol{W}}$.

**Theorem 9.** *For linear multi-class classification problem using logistic loss and almost every separable data, with a small enough learning rate, and $1 > \beta_2 > \beta_1^4 \geq 0$ (for Adam  (w/s)), SGDM and Adam  (w/s) converge to the multi-class $L^2$ max-margin solution (a.s. for SGDM SGDM (w/. r)).*

Here we use several notations and lemmas from [34]. We define $\boldsymbol{w} = \operatorname{vec}(\boldsymbol{W})$, $\hat{\boldsymbol{w}} = \operatorname{vec}(\hat{\boldsymbol{W}})$, $\boldsymbol{e}_i \in \boldsymbol{R}^C$ ($i \in \{1, \cdots, C\}$) satisfying $(\boldsymbol{e}_i)_j = \delta_{ij}$, and $\boldsymbol{A}_i = \boldsymbol{e}_i \otimes \boldsymbol{I}_{d_X}$, where $\boldsymbol{I}_{d_X}$ is the identity matrix with dimension $d_X$. We still consider the normalized data, i.e., $\|\boldsymbol{x}\| \leq 1, \forall (\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{S}$. Then, the individual loss of sample $(\boldsymbol{x}, \boldsymbol{y})$ can be then represented as

$$\ell(\boldsymbol{y}, \boldsymbol{W}\boldsymbol{x}) = \ln \frac{e^{\langle \boldsymbol{w}, \boldsymbol{A}_{\boldsymbol{y}}\boldsymbol{x}\rangle}}{\sum_{i=1}^C e^{\langle \boldsymbol{w}, \boldsymbol{A}_i \boldsymbol{x}\rangle}}.$$

Furthermore, the gradient of training error at $\boldsymbol{W}$ has the form

$$\nabla \mathcal{L}(\boldsymbol{w}) = \frac{1}{N} \sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{S}} \sum_{i=1}^{C} \frac{1}{\sum_{j=1}^C e^{\langle \boldsymbol{w}, (\boldsymbol{A}_j - \boldsymbol{A}_i)\boldsymbol{x}\rangle}} (\boldsymbol{A}_i - \boldsymbol{A}_{\boldsymbol{y}})\boldsymbol{x}.$$

and the Hessian matrix of $\mathcal{L}$ can be represented as

$$\mathcal{H}\mathcal{L}(\boldsymbol{w}) = \frac{1}{N} \sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{S}} \sum_{i=1}^{C} \frac{\sum_{j=1}^C e^{\langle \boldsymbol{w}, (\boldsymbol{A}_j - \boldsymbol{A}_i)\boldsymbol{x}\rangle}}{\left(\sum_{j=1}^C e^{\langle \boldsymbol{w}, (\boldsymbol{A}_j - \boldsymbol{A}_i)\boldsymbol{x}\rangle}\right)^2} (\boldsymbol{A}_i - \boldsymbol{A}_{\boldsymbol{y}})\boldsymbol{x}((\boldsymbol{A}_j - \boldsymbol{A}_i)\boldsymbol{x})^\top,$$

one can then easily verify all absolute value of the eigenvalues of $\mathcal{H}\mathcal{L}(\boldsymbol{w})$ is no larger than 2, which indicates $\mathcal{L}$ is 2-globally smooth.

On the other hand, the separable assumption leads to $\langle \hat{\boldsymbol{w}}, (\boldsymbol{A}_{\boldsymbol{y}} - \boldsymbol{A}_i)\boldsymbol{x}\rangle > 0, \forall \boldsymbol{y} \neq i$, which further indicates

$$\langle \nabla \mathcal{L}(\boldsymbol{w}), \hat{\boldsymbol{w}}\rangle > 0.$$

Let $\gamma = \frac{1}{\|\hat{\boldsymbol{w}}\|}$, following the similar routine as the binary case, we have for a random subset of $\boldsymbol{S}$ sampled uniformly without replacement with size $b$, we have

$$\|\nabla \mathcal{L}(\boldsymbol{w})\|^2 \leq \mathbb{E}_{\boldsymbol{B}(t)} \|\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w})\|^2 \leq \frac{2N}{\gamma b^2} \|\nabla \mathcal{L}(\boldsymbol{w})\|^2. \tag{72}$$

Similarly, we have for any positive real series $\{a_t\}_{t=t_1}^{t_2}$,

$$\gamma \sum_{t=t_1}^{t_2} a(t) \|\nabla \mathcal{L}(\boldsymbol{w}(t))\| \leq \left\| \sum_{t=t_1}^{t_2} a(t) \nabla \mathcal{L}(\boldsymbol{w}(t)) \right\| \leq \sum_{t=t_1}^{t_2} a(t) \|\nabla \mathcal{L}(\boldsymbol{w}(t))\|. \tag{73}$$

The proofs of Stage I can then be obtained with Lyapunov functions unchanged and by replacing the corresponding lemmas using Eq. (72) and Eq. (73).

As for the proofs of Stage II, the Lyapunov functions are still the same, while we only need to prove the sum of $\langle \boldsymbol{r}(t), -\frac{\eta}{1-\beta}\nabla \mathcal{L}(\boldsymbol{w}(t)) - \ln \frac{t+1}{t}\hat{\boldsymbol{w}}\rangle$ (for GDM, $\langle \boldsymbol{r}(t), -\frac{\eta}{1-\beta}\nabla \mathcal{L}_{\boldsymbol{B}(t)}(\boldsymbol{w}(t)) - \frac{N}{bt}\sum_{i:\boldsymbol{x}_i \in \boldsymbol{B}(t)\cap \boldsymbol{S}_s} \boldsymbol{v}_i \boldsymbol{x}_i\rangle$ for SGDM, $\langle \boldsymbol{r}(t), -\sqrt{\varepsilon}\ln\left(\frac{t+1}{t}\right)\hat{\boldsymbol{w}} - \frac{\eta}{1-\beta}\nabla \mathcal{L}(\boldsymbol{w}(t))\rangle$ for Adam). For the multi-class case using GDM, We present the following lemma from [34], while the other two cases can be proved similarly:

**Lemma 28** (Part of the proof of Lemma 20, [34])**.** *If $\langle \boldsymbol{w}(t), (\boldsymbol{A}_{\boldsymbol{y}} - \boldsymbol{A}_i)\boldsymbol{x}\rangle \to \infty$ as $t \to \infty$, $\forall (\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{S}$ and $\forall i \neq \boldsymbol{y}$, we have the sum of $\langle \boldsymbol{r}(t), -\frac{\eta}{1-\beta}\nabla \mathcal{L}(\boldsymbol{w}(t)) - \ln \frac{t+1}{t}\hat{\boldsymbol{w}}\rangle$ is upper bounded.*

The proof of Theorem 9 is then completed.

### F.4 Precisely characterize the convergence rate

Theorem 2 and Theorem 3 can be further extended to precisely characterize the asymptotic convergence rate for (S)GDM as the following theorem.

**Theorem 10.** *Let all the conditions in Theorem 2 (Theorem 3) hold. Assume the linear span of support vectors contains the whole dataset. Then, we have*

$$\lim_{t\to\infty} t\mathcal{L}(\boldsymbol{w}(t)) = C\frac{1}{\eta},$$

*where $C$ is a constant independent of learning rate $\eta$, momentum hyperparameter $\beta$, and mini-batch size $b$.*

The proof follows exactly the same routine as (Corollary 1, [23]) and we omit it here.

## G Experiments

This section collects several experiments supporting our theoretical results.

### G.1 Experiments on linear model

In the following sections, we conduct experiments on the well=posed and ill-posed datasets in (Figures 1 and 3, [34]). Note that while we try our best to recover the result in [34] by using the same hyperparameters and the same dataset, there is still some difference between the result of [34] and us, as the code has not been open source and we are unable to know the initialization method and the random seed.

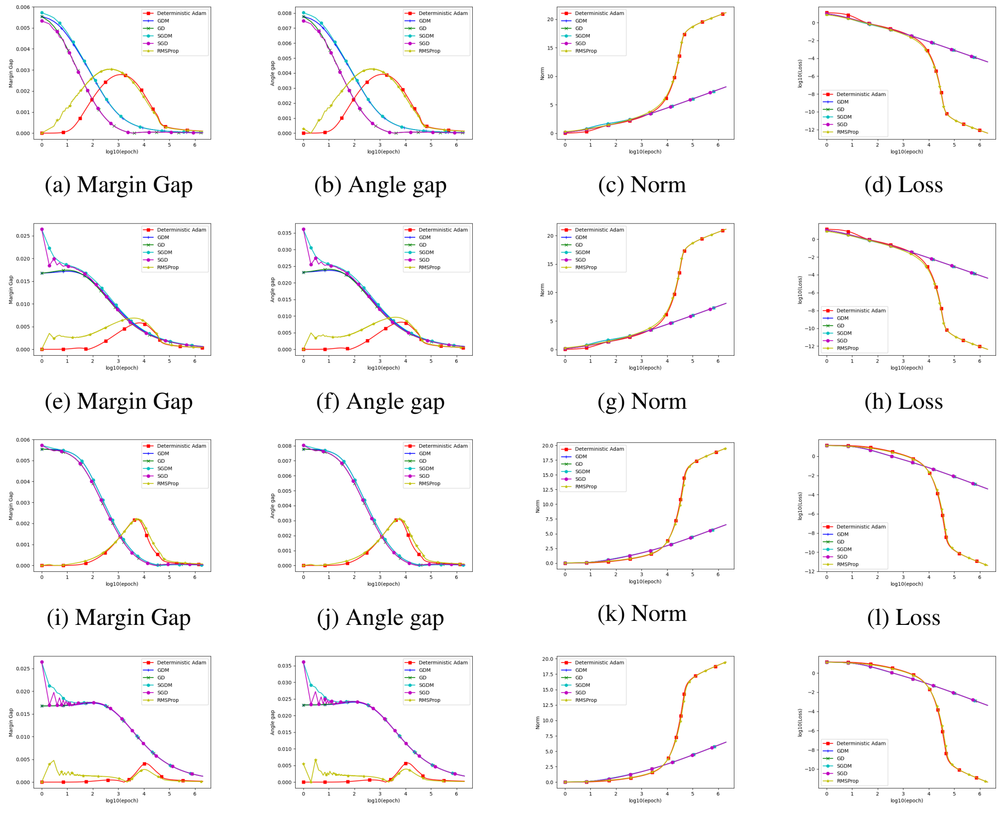
#### G.1.1 Comparing the training behavior of (S)GD, (S)GDM, deterministic Adam and RMSProp

The experiments in this section is designed to verify our theoretical results, i.e., Theorems 2, 3, 5, and 6. We use exact the same setting as (Figure 1, [34]) (including the same separable synthetic dataset and the same learning rate $1/\sigma_{max}^2$), and run (S)GD, (S)GDM, deterministic Adam and stochastic RMSProp over it different random seeds (for random samples despite the support sets $\{((1.5, 0.5), 1), ((0.5, 1.5), 1), ((-1.5, -0.5), -1), ((-1.5, -0.5), -1)\}$), and random mini-batches. We also report the results with a smaller learning rate $0.1/\sigma_{max}^2$. Both the angle between the output parameter and max-margin solution, the gap between current margin and max margin, parameter norm and the training loss are plotted in Figure 2. The observations can be summarized as follows:

- With proper learning rates, all of (S)GD, (S)GDM, deterministic Adam and stochastic RMSProp converge to the max-margin solution, which supports our theoretical results;
- (Similarity between (S)GD and (S)GDM). The asymptotic training behaviors of GD, SGD, GDM and SGDM are highly similar, which supports our Theorem 10.
- (Different behavior of Adaptive Optimizers). While deterministic Adam and stochastic RMSProp both converge to the max-margin solution, their training behaviors are different from those of (S)GD(M). Specifically, their angle gap from the max margin solution first increases then decreases. This phenomenon, however, can be explained by the proof of Theorems 5 and 6, the key insight of which is that when the training time is large enough, the gradient is small and the adaptive learning rate $1/\sqrt{\boldsymbol{\nu}(t) + \varepsilon\mathbb{1}_d}$ is dominated by $1/\varepsilon\mathbb{1}_d$, and thus Adam behaves like GDM and converges to the max margin solution.

#### G.1.2 Adam on ill-posed dataset

In (Figure 3, [34]), an ill-posed synthetic dataset is proposed to support the argument "Adam does not converge to max-margin solution". Specifically, such a dataset is derived by multiplying 20 to the second coordinate of all of the data in the dataset adopted in Section G.1.1, and thus both make the dataset (almost) degenerates to an one dimension line and lead to large singular value $\sigma_{max}$. However, their experiment does not contradict our implicit regularization result for Adam (i.e., Theorem 5), as our results hold asymptotically and it can be observed that at the end of their



(a) Margin Gap (b) Angle gap (c) Norm (d) Loss

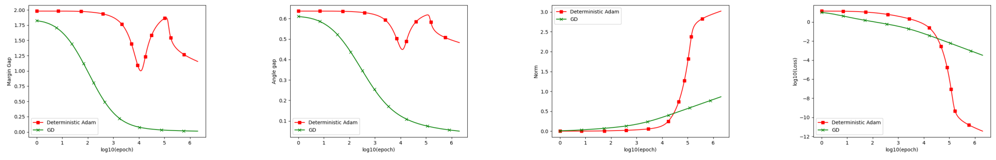
(e) Margin Gap (f) Angle gap (g) Norm (h) Loss

(i) Margin Gap (j) Angle gap (k) Norm (l) Loss

(m) Margin Gap (n) Angle gap (o) Norm (p) Loss

Figure 2: Comparison of (S)GD, (S)GDM, deterministic Adam, and stochastic RMSProp on the synthetic dataset in [34]. The first line use learning rate $1/\sigma_{max}^2$ and random seed 1. The second line use learning rate $1/\sigma_{max}^2$ and random seed 2. The third line use learning rate $0.1/\sigma_{max}^2$ and random seed 1. The forth line use learning rate $0.1/\sigma_{max}^2$ and random seed 2.

experiment, the angle gap is still decreasing. Furthermore, we reconduct the experiment of (Figure 3, [34]) in Figure 3. Specifically, we use the exact setting as [34] (the same ill-posed dataset, the same learning rate $1/\sigma_{max}^2$ both for GD and Adam, and $2 \times 10^6$ training iterations). It can be observed that the convergence of loss is rather slow at the first $10^4$ iteration. Also, the angle gap keeps decreasing at the end of experiment. We increase the training iterations to $2 \times 10^7$ in Figure 3 and find the angle gap is still decreasing. This stands with our result that Adam asymptotically converges to the max-margin solution.



(a) Margin Gap (b) Angle gap (c) Norm (d) Loss

Figure 3: We reproduce the experiment in (Figure 3, [34]). It can be observed that the loss does not vary much until $10^4$ iteration.

## G.2 Evidence from deep neural networks

We conduct an experiment on the MNIST dataset using the four layer convolutional networks used in [21, 43] (first proposed by [22]) to verify whether SGD and SGDM still behave similarly in

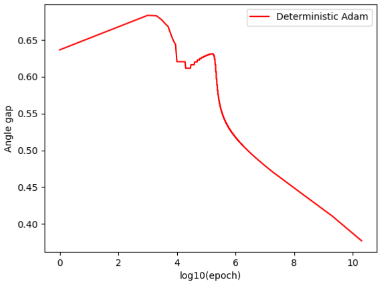
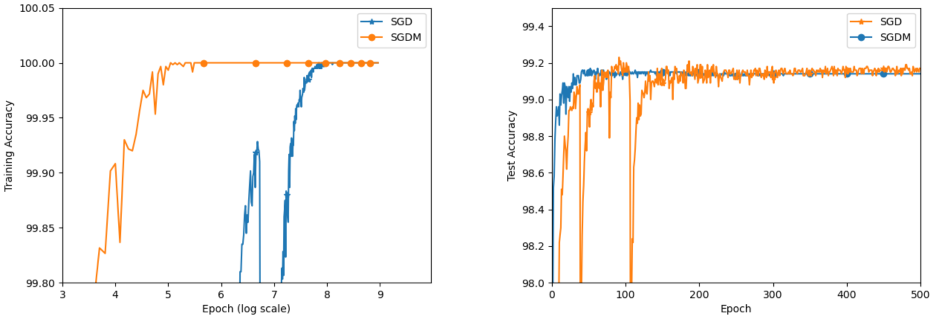


Figure 4: We extend the training time in Figure 2 and find that the angle gap keeps decreasing.

(homogeneous) deep neural networks. The learning rates of the optimizers are all set to be the default in Pytorch. The results can be seen in Figure 5. It can be observed that (1). SGDM achieves similar test accuracy compared to SGD while (2). SGDM converges faster than SGD.



(a) Comparison of Training Accuracy

(b) Comparison of Test Accuracy

Figure 5: Comparison of SGD and SGDM on MNIST with a four-layer CNN.