IMPLICIT SPARSE REGULARIZATION IN MOMENTUM GRADIENT DESCENT: A SHIFTED EARLY STOPPING WINDOW INDUCED BY MOMENTUM AND DEPTH

Anonymous authors
Paper under double-blind review

ABSTRACT

In high-dimensional sparse regression, optimization algorithms are known to induce implicit sparsity through the interplay of overparameterization and early stopping. However, existing theory largely focuses on (stochastic) gradient descent (GD), leaving momentum-based methods — despite their widespread empirical success — poorly understood. This paper provides a theoretical characterization of implicit sparse regularization in momentum gradient descent (MGD). We show that for depth-N diagonal networks, MGD with early stopping achieves minimax optimal sparse recovery with sufficiently small initialization and learning rate. Moreover, increasing depth significantly enlarges both the admissible initialization scale and the effective early stopping window, making implicit sparsity more likely to emerge. More importantly, we uncover a distinct temporal effect of momentum: a larger momentum shortens the early stopping window, but it shifts the entire window toward earlier iterations. As a result, MGD induces sparsity earlier than GD. Experimental results strongly support our theoretical findings. The code is available at `https://anonymous.4open.science/r/Implicit-Sparse-Regularization-in-MGD`

1 INTRODUCTION

In highly overparameterized deep learning models, gradient-based methods often achieve remarkable generalization performance despite the model capacity far exceeding the number of training samples. This phenomenon cannot be fully explained by the classical statistical learning theory, such as VC dimension Vapnik & Chervonenkis (1971) or Rademacher complexity Koltchinskii (2000; 2006). In recent years, a growing body of work has suggested that implicit regularization induced by optimization algorithms plays a key role in explaining the generalization behavior of deep learning models Zhang et al. (2017; 2021). Specifically, although the hypothesis space is extremely rich, gradient-based methods tend to favor solutions with low-complexity structures, thereby enabling the good generalization Soudry et al. (2018); Vardi & Shamir (2021); Lyu & Zhu (2023); Gatmiry et al. (2024); Pesme et al. (2021); HaoChen et al. (2021); Gunasekar et al. (2018b).

In this context, understanding how the network architecture influences implicit regularization has become a fundamental problem. Existing studies on gradient descent (GD) reveal that its implicit bias is closely tied to the network depth. For the linear regression ($N = 1$), Ali et al. (2019; 2020) show that GD (and SGD) converges to the minimum ℓ_2-norm solution. For the two-layer diagonal linear network ($N = 2$), Woodworth et al. (2020) shows that the solution selected by GD can be characterized as the solution to a constrained norm minimization problem that interpolates between ℓ_1-norm and ℓ_2-norm, where the interpolation is controlled by the initialization scale. For the N-layer diagonal linear network ($N > 2$), Li et al. (2021); Gissin et al. (2020) indicate that increasing depth strengthens GD's preference for sparse solutions in regression problems. Moreover, increasing depth enlarges the admissible initialization scale and extends the effective early stopping window, making implicit sparsification more likely to emerge Li et al. (2021). These findings suggest that network depth fundamentally reshapes the implicit regularization of gradient-based optimization. Related perspectives on depth-dependent implicit regularization can also be found in Arora et al. (2018; 2019); Razin et al. (2021); Hariz et al. (2024). However, most existing studies focus primarily

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

on GD and SGD, while the implicit regularization behavior of momentum-based gradient methods, such as momentum gradient descent (MGD) Polyak (1964), remains far less understood.

Although momentum methods demonstrate significant advantages in optimization acceleration Sebbouh et al. (2021); Fu et al. (2023); Sun et al. (2019); Xie et al. (2024), the theoretical understanding of how momentum affects implicit regularization is still limited. Gunasekar et al. (2018a); Wang et al. (2022; 2023) show that for single-layer linear networks, momentum does not change the implicit bias of GD. In contrast, for deep models, the situation becomes more intricate. For diagonal linear networks with $N = 2$, Lyu et al. (2025) reveals a dual effect of momentum: on the one hand, momentum induces extra initialization mitigation effects similar to SGD that are beneficial for generalization of sparse regression; on the other hand, it also depends on the initialization of gradients explicitly, which may not be benign for generalization. However, these analyses do not extend to deeper networks ($N > 2$). Wang et al. (2023) shows that increasing depth strengthens the tendency of MGD toward sparse solutions, but they do not theoretically characterize the role of momentum.

Prior evidence indicates that network depth Eldan & Shamir (2016); Lu et al. (2017); Bengio & LeCun (2007); Delalleau & Bengio (2011); Cheng et al. (2016); Liang & Srikant (2017); Bubeck & Sellke (2021) and momentum Leclerc & Madry (2020); Jelassi & Li (2022); Ghosh et al. (2023); Sato & Iiduka (2024); Ramezani-Kebrya et al. (2024) are beneficial to generalization. Therefore, jointly studying the effects of network depth and momentum parameters on implicit regularization is both theoretically important and practically necessary. This leads to a fundamental open question:

> **How do network depth and momentum jointly influence the implicit regularization of MGD in deep neural networks ($N > 2$)?**

To address this question, we study high-dimensional sparse regression and characterize how depth and momentum affect the implicit regularization of MGD, with a focus on their roles in initialization scale and early stopping behavior. Our main contributions are summarized as follows:

- We prove that for a depth-N diagonal network, under sufficiently small initialization $\boldsymbol{\theta}_0$ and learning rate ($\eta = \mathcal{O}\left(\frac{1}{N^2\gamma}\right)$, where $\gamma = \frac{1+\mu}{2(1-\mu)^3}$ and μ is the momentum parameter) MGD with early stopping achieves minimax optimal sparse recovery.
- Our analysis reveals that increasing depth enlarges the working initialization scale and extends the effective early stopping window size, making this implicit sparse regularization effect more likely to take place.
- We further uncover a nontrivial role of momentum: increasing the momentum not only enlarges the scale of working initialization, but also shifts the early stopping window earlier and shortens its size, indicating that implicit sparsification occurs earlier during training compared with standard GD.
- Finally, extensive numerical experiments validate our theoretical findings.

2 PRELIMINARIES

In this section, we introduce the necessary background for our theoretical analysis, including the setup of high-dimensional sparse regression, the structure of diagonal linear networks, and MGD.

High-dimensional Sparse Regression. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ ($n \ll p$) be the design matrix, where n and p denote the number of samples and features, respectively. Suppose we observe n data points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ such that $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle + \xi_i$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ is a noise vector following a Gaussian distribution, and $\boldsymbol{\theta}^*$ is a p-dimensional sparse vector with π non-zero entries. The linear model can be expressed in the matrix-vector form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\xi}, \tag{1}$$

The goal of the sparse regression is to estimate the unknown, sparse vector $\boldsymbol{\theta}^*$ from the observations.

Definition 1. *Assume the columns of $\boldsymbol{X}$ are ℓ_2-normalized, i.e., $\|\boldsymbol{X}_{\cdot,i}\|_2 = 1$ for $i = 1, \ldots, p$. The coherence $\sigma = \sigma(\boldsymbol{X})$ is defined as $\sigma = \max_{i\neq j} \left|\boldsymbol{X}_{\cdot,i}^\top \boldsymbol{X}_{\cdot,j}\right|$. We say that $\boldsymbol{X}$ satisfies σ-incoherence.*

The coherence is a measure of the suitability of the measurement matrix in compressive sensing Foucart & Rauhut (2013). And the smaller the coherence, the better the recovery algorithms perform.

Notation. We first introduce the notation used throughout the paper. Let $S = i \in 1, \dots, p \mid \theta_i^* \neq 0$ denote the support of the target sparse vector $\boldsymbol{\theta}^*$, and let $S^c = 1, \dots, p \setminus S$ denote its complement. Define the largest and smallest magnitudes on the support as $\theta^*_{\max} = \max_{i \in S} |\theta_i^*|$ and $\theta^*_{\min} = \min_{i \in S} |\theta_i^*|$, respectively. Let $\mathbf{1}$ denote the all-ones vector, and let $\mathbf{1}_S$ denote the indicator vector of S, with entries equal to 1 on S and 0 elsewhere. The symbol $\odot$ denotes the Hadamard product. For any $\boldsymbol{\theta}_k$, we decompose it at iteration k into the signal component $\boldsymbol{s}_k = \mathbf{1}_S \odot \boldsymbol{\theta}_k$ and the error component $\boldsymbol{e}_k = \mathbf{1}_{S^c} \odot \boldsymbol{\theta}_k$. We use $\vee$ and $\wedge$ to denote the coordinate-wise maximum and minimum, respectively, and $\succcurlyeq$ to denote coordinate-wise inequalities. Finally, $a \lesssim b$ means that $a \leq Cb$ for some absolute constant $C > 0$ independent of the problem parameters.

Diagonal Linear Network. The diagonal linear network Pesme et al. (2021); Woodworth et al. (2020); Li et al. (2021); Gissin et al. (2020); Wang et al. (2023); Lyu et al. (2025) serve as an intriguing and informative simplification proxy model for DNNs. We adopt the depth-N diagonal linear network with $N > 2$. The model reparameterizes the linear predictor $\boldsymbol{\theta}$ as:

$$\boldsymbol{\theta} = \boldsymbol{u}^N - \boldsymbol{v}^N, \qquad \boldsymbol{u},\ \boldsymbol{v} \in \mathbb{R}^p, \tag{2}$$

where $\boldsymbol{\vartheta}^N$ denotes the element-wise N-th power of the vector $\boldsymbol{\vartheta}$. For the new parameters $\boldsymbol{u}$ and $\boldsymbol{v}$, we give the mean squared error loss:

$$L(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2n} \left\| \boldsymbol{X}(\boldsymbol{u}^N - \boldsymbol{v}^N) - \boldsymbol{y} \right\|^2 := \frac{1}{2n} \sum_{i=1}^{n} (f(\boldsymbol{\theta}, \boldsymbol{x}_i) - y_i)^2 := \frac{1}{2n} \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}). \tag{3}$$

Momentum Gradient Descent. Let $\boldsymbol{w}_k = (\boldsymbol{u}_k, \boldsymbol{v}_k)$, the gradient $\nabla L(\boldsymbol{w}_k) = \left(\frac{\partial L(\boldsymbol{w}_k)}{\partial \boldsymbol{u}_k}, \frac{\partial L(\boldsymbol{w}_k)}{\partial \boldsymbol{v}_k}\right)$, and the momentum vector $\boldsymbol{h}_k$. The iteration of MGD can be given by:

$$\begin{aligned} &\boldsymbol{w}_0 = (\alpha \mathbf{1}, \alpha \mathbf{1}), \quad \boldsymbol{h}_0 = \mathbf{0}, \\ &\boldsymbol{h}_{k+1} = \mu \boldsymbol{h}_k - \eta \nabla L(\boldsymbol{w}_k), \\ &\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \boldsymbol{h}_{k+1}, \qquad k = 0, 1, \dots \end{aligned} \tag{MGD}$$

where $\mu \in [0, 1)$ is momentum parameter, η is learning rate, $\boldsymbol{\alpha}$ is the initialization scale of $\boldsymbol{u}$ and $\boldsymbol{v}$.

3 Main Results

3.1 Modified MGD

The update rule of MGD is inherently a two-step iteration, which makes its dynamics substantially more difficult to analyze than those of standard GD. To facilitate the theoretical analysis, we first introduce a continuous-time approximation of MGD Ghosh et al. (2023).

Lemma 1. *Ghosh et al. (2023) Suppose the gradient of the full-batch loss function $\nabla L(\boldsymbol{w})$ is smooth and four-times continuously differentiable. Then, for a sufficiently large iteration k, the MGD are $\mathcal{O}(\eta^2)$-close to the flow of the continuous trajectory of the first-order ODE*

$$\dot{\boldsymbol{w}}(t) = -\frac{1}{1-\mu} \nabla \hat{L}(\boldsymbol{w}(t)), \quad t \in (0, T],$$

where the modified loss $\hat{L}(\boldsymbol{w}(t))$ is given by $\hat{L}(\boldsymbol{w}(t)) = L(\boldsymbol{w}(t)) + \frac{(1+\mu)\eta}{4(1-\mu)^2} \|\nabla L(\boldsymbol{w}(t))\|_2^2$.

Lemma 1 shows that the discrete two-step MGD approximates the gradient flow driven by the modified loss $\hat{L}$ i.e., the trajectory of MGD is equivalent to that of GD on $\hat{L}$ augmented by the regularizer induced by momentum. This implies that the first-order ODE has already captured the essential role of momentum. From Lemma 1, we note that $\nabla \hat{L}(\boldsymbol{w})$ involves the term $\nabla \|\nabla L(\boldsymbol{w})\|$, which in turn depends on the Hessian of the loss $\nabla^2 L$. To make the analysis tractable, we adopt two simplifications: (1) we adopt the Gauss–Newton approximation of the Hessian Sagun et al. (2018):

$$\nabla^2 L(\boldsymbol{\theta}) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla_{f_i(\boldsymbol{\theta})} \ell_i(\boldsymbol{\theta}) \nabla f_i(\boldsymbol{\theta}) \nabla f_i(\boldsymbol{\theta})^\top,$$

where $\nabla_{f_i(\boldsymbol{\theta})}\ell_i$ denotes the derivative of the per-sample loss ℓ_i with respect to the predictor $f_i(\boldsymbol{\theta})$, and $\nabla f_i(\boldsymbol{\theta})$ is the gradient of the predictor with respect to parameters $\boldsymbol{\theta}$; (2) we simplify the Hessian matrix to a block-diagonal matrix Li et al. (2019); Martens & Grosse (2015), i.e., $\nabla^2 L(\boldsymbol{\theta}) = [\nabla^2_{\boldsymbol{u}} L(\boldsymbol{\theta}), \mathbf{0}; \mathbf{0}, \nabla^2_{\boldsymbol{v}} L(\boldsymbol{\theta})]$. Based on Lemma 1, we then discretize the corresponding continuous-time dynamics using the forward Euler discretization, and employ the Gauss-Newton approximation to obtain the modified MGD (M-MGD)

$$\begin{aligned}
\boldsymbol{u}_{k+1}^{\mathrm{GN}} &= \boldsymbol{u}_k^{\mathrm{GN}} - \eta\tau\nabla_{\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}}) - \eta^2\gamma\nabla^2_{\mathrm{GN},\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}})\nabla_{\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}}) \\
\boldsymbol{v}_{k+1}^{\mathrm{GN}} &= \boldsymbol{v}_k^{\mathrm{GN}} - \eta\tau\nabla_{\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}}) - \eta^2\gamma\nabla^2_{\mathrm{GN},\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}})\nabla_{\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}}) \qquad \text{(M-MGD)}
\end{aligned}$$

where the momentum term τ and γ are defined as $\tau = \frac{1}{1-\mu}, \quad \gamma = \frac{1+\mu}{2(1-\mu)^3}$. The subsequent analysis is based on the modified one-step MGD iteration. Therefore, we use Theorem 1 to guarantee that M-MGD is a reliable approximation to the original standard MGD.

Theorem 1. *For a four times continuously differentiable loss function $L(\boldsymbol{u}, \boldsymbol{v}) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$, let $\boldsymbol{w} = (\boldsymbol{u}, \boldsymbol{v})^\top$ and $\boldsymbol{w}^{\mathrm{GN}} = (\boldsymbol{u}^{\mathrm{GN}}, \boldsymbol{v}^{\mathrm{GN}})^\top$ denote the trajectories generated by MGD and M-MGD, respectively. Assume both trajectories share the same initialization, i.e., $\boldsymbol{w}_0 = \boldsymbol{w}_0^{\mathrm{GN}} = (\alpha\mathbf{1}, \alpha\mathbf{1})^\top$. If $K = \mathcal{O}(\frac{1}{\eta})$, then at iteration K, we have*

$$\|\boldsymbol{w}_K - \boldsymbol{w}_K^{\mathrm{GN}}\|_\infty \leq \mathcal{O}(\eta).$$

Theorem 1 establishes that the Gauss-Newton approximation (M-MGD) is first-order accurate in η. Its proof and empirical validation are provided in Appendix A. Under the diagonal network model and the mean squared loss, we explicitly compute the Gauss-Newton approximation, which is formally stated in Corollary 1.

Corollary 1. *Under the setting of mean squared loss (3) and the depth-N diagonal networks (2) and using the Gauss-Newton decomposition approximation of the Hessian $\nabla^2 L(\boldsymbol{u}, \boldsymbol{v})$, discretizing the ODE defined in Lemma 1 yields the following update rule for MGD*

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k - \eta\left\{\tau\boldsymbol{I} + \eta\gamma\left[\mathrm{Diag}\left(\boldsymbol{u}_k^{N-1}\right)\frac{N^2}{n}\boldsymbol{X}^\top\boldsymbol{X}\mathrm{Diag}\left(\boldsymbol{u}_k^{N-1}\right)\right]\right\}\left(\frac{N}{n}\boldsymbol{X}^\top\boldsymbol{r}_k \odot \boldsymbol{u}_k^{N-1}\right) \quad (4)$$

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \eta\left\{\tau\boldsymbol{I} + \eta\gamma\left[\mathrm{Diag}\left(\boldsymbol{v}_k^{N-1}\right)\frac{N^2}{n}\boldsymbol{X}^\top\boldsymbol{X}\mathrm{Diag}\left(\boldsymbol{v}_k^{N-1}\right)\right]\right\}\left(\frac{N}{n}\boldsymbol{X}^\top\boldsymbol{r}_k \odot \boldsymbol{v}_k^{N-1}\right) \quad (5)$$

where $\mathrm{Diag}(\cdot)$ denotes the operator that maps a vector to the corresponding diagonal matrix and the residual term $\boldsymbol{r}_k$ is denoted as $\boldsymbol{r}_k = \boldsymbol{X}\boldsymbol{\theta}_k - \boldsymbol{y}$.

The detailed proof is provided in Appendix B. In the following analysis, at each iteration k of update (4) and (5), we compute $\boldsymbol{\theta}_k = \boldsymbol{u}_k^N - \boldsymbol{v}_k^N$ to study the evolution of the signal.

3.2 Convergence of MGD

In this section, we state the convergence of MGD in the ℓ_∞-norm.

Theorem 2. *Assume $\boldsymbol{\theta}^* \succ 0$ (all components are positive) and sparse parameter $\pi \geq 1$ Let the normalized design matrix $\frac{1}{\sqrt{n}}\boldsymbol{X}$ satisfies σ-incoherence with $\sigma \leq \frac{C_1}{\pi d}$, where C_1 is a sufficiently small constant, and $d := \frac{\theta^*_{\max}}{\theta^*_{\min}}$. Given an arbitrary precision $\epsilon > 0$, suppose initialization satisfies:*

$$0 < \alpha \leq \left(\frac{\epsilon}{p+1}\right)^{\frac{4}{N}} \vee \mathrm{Init}(\theta^*_{\max}, \theta^*_{\min}, \epsilon, N, \mu),$$

where

$$\begin{aligned}
\mathrm{Init}(\theta^*_{\max}, \theta^*_{\min}, \epsilon, N, \mu) := \left(\frac{1}{8\beta}\right)^{\frac{2N}{N-2}} &\vee \left(\frac{\left[\tau + \frac{25}{64}\eta\gamma N^2(\theta^*_{\max})^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}}(\theta^*_{\max})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\max}}{\epsilon}}\right)^{\frac{2N}{N-2}} \\
&\vee \left(\frac{\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}}(\theta^*_{\min})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\min}}{\epsilon}}\right)^{\frac{4N}{N-2}},
\end{aligned}$$

where $\beta = \frac{\tau+1/B}{1-(\tau+1/B)/\gamma}$ *(see Lemma 14). Let* $\zeta := \frac{1}{5}\theta^*_{\min} \vee \frac{2}{C_2}\left\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{\xi}\right\|_\infty \vee \frac{2}{C_2}\epsilon$. *For any time step* k *satisfying* $K_1 + K_3 \le k \le K_2 + K_4$ *where* $K_1 + K_3$ *and* $K_2 + K_4$ *are defined in equation (18) and equation (19) (see Appendix G), generated by MGD (4) with* $\eta \le \frac{\alpha^N}{N^2\gamma(\theta^*_{\max})^{\frac{3N}{N-1}}}$ *and* $\gamma - 3\tau \ge \max\left\{\frac{1}{B}, \frac{2N}{N-2}\right\}$ *satisfies the following bound:*

$$|\theta_{k,i} - \theta^*_i| \lesssim \begin{cases} \left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty \wedge \epsilon & \text{if } i \in S \text{ and } \theta^*_{\min} \lesssim \left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty \wedge \epsilon, \\ \left|\frac{1}{n}(\boldsymbol{X}^\top\boldsymbol{\xi})_i\right| \wedge k\mu\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\odot\mathbf{1}_S\right\|_\infty \wedge \epsilon & \text{if } i \in S \text{ and } \theta^*_{\min} \gtrsim \left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty \wedge \epsilon, \\ \alpha^{\frac{N}{4}} & \text{if } i \notin S. \end{cases}$$

Remark 1. (**How depth and momentum affect the initialization scale.**) The proof of Theorem 2 is deferred to Appendix G. Since $\theta_0 = \alpha^N$, the initialization condition in Theorem 2 can be rewritten as

$$\theta_0 \le \left(\frac{1}{8\beta}\right)^{\frac{2N}{N-2}} \vee \left(\frac{\left[\tau + \frac{25}{64}\eta\gamma N^2(\theta^*_{\max})^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}}(\theta^*_{\max})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\max}}{\epsilon}}\right)^{\frac{2N}{N-2}} \vee \left(\frac{\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}}(\theta^*_{\min})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\min}}{\epsilon}}\right)^{\frac{4N}{N-2}}. \tag{6}$$

First, we analyze how increasing the depth N enlarges the admissible initialization scale.

- For the first term of (6) ($\beta \ge 1$): it is evident that $\left(\frac{1}{8\beta}\right)^{\frac{2N}{N-2}}$ increases as N increases.
- For the second term of (6): when ϵ is sufficiently small, it can be simplified as:

$$\left[\tau + \frac{25}{64}\eta\gamma N^2(\theta^*_{\max})^{\frac{2N-2}{N}}\right]^2(\theta^*_{\max})^2 \Bigg/ \left(\log\frac{\theta^*_{\max}}{\epsilon}\right)^{\frac{2N}{N-2}}.$$

- Under the learning rate condition $\eta = \mathcal{O}(\frac{1}{N^2\gamma(\theta^*_{\max})^{\frac{3N-2}{N}}})$, it follows that

$$\left[\tau + \frac{25}{64}\eta\gamma N^2(\theta^*_{\max})^{\frac{2N-2}{N}}\right]^2 \lesssim \tau^2.$$

Since $\left(\log\frac{\theta^*_{\max}}{\epsilon}\right)^{\frac{2N}{N-2}}$ decreases as N increases, thus the second term becomes larger as N grows.

- For the third term in (6): we simplify and obtain:

$$\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right]^4(\theta^*_{\min})^4 \Bigg/ \left(\log\frac{\theta^*_{\min}}{\epsilon}\right)^{\frac{4N}{N-2}},$$

and the same argument shows that this term also increases with N.

Next, we analyze the effect of momentum μ on the initialization scale. Since $\beta \ge 1$ and $\frac{1}{8\beta} \le \frac{1}{8}$, we focus only on the last two terms of (6).

- For the second term of (6): under $\eta = \mathcal{O}(\frac{1}{N^2\gamma(\theta^*_{\max})^{\frac{3N-2}{N}}})$, the factor τ increases with μ. This shows that this term increases as μ increases.
- The third term of (6) follows a similar argument.

It is worth noting that, compared with GD, MGD exhibits a larger scale of the working initialization. For comparison, the initialization condition for GD established by Li et al. Li et al. (2021) is

$$\theta_0 \lesssim \left(\frac{1}{8}\right)^{\frac{2N}{N-2}} \vee \left(\frac{(\theta^*_{\max})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\max}}{\epsilon}}\right)^{\frac{2N}{N-2}} \vee \left(\frac{(\theta^*_{\min})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\min}}{\epsilon}}\right)^{\frac{4N}{N-2}}.$$

For MGD, we have $\tau = \frac{1}{1-\mu}$. When the momentum $\mu \to 1$, τ is amplified, thereby relaxing the requirement on the initialization scale. **Consequently, MGD permits significantly larger initialization while still ensuring convergence and implicit sparse regularization.**

For general signals, the corresponding ℓ_∞-norm convergence result for MGD is presented in Theorem 5 in Appendix H. Compared with Theorem 2, the main difference lies in the initialization requirement. In the general case, an additional condition is needed:

$$0 < \alpha \leq \left(\frac{\epsilon}{p+1}\right)^{\frac{4}{N}} \vee \mathrm{Init}(\theta^*_{\max}, \theta^*_{\min}, \epsilon, N, \mu) \vee \tilde{\mathrm{Init}}(\theta^*_{\min}, N),$$

where $\tilde{\mathrm{Init}}(\theta^*_{\min}, N) := (2 - 2^{\frac{N-2}{N}})^{\frac{1}{N-2}} (\theta^*_{\min})^{\frac{1}{N}} \vee 2^{\frac{3}{N}} (2^{\frac{1}{N}} - 1)^{\frac{1}{N-2}} (\theta^*_{\min})^{\frac{1}{N}}$.

3.3 Early Stopping Window Size

Early stopping is a widely used regularization technique in boosting Zhang & Yu (2004) and nonparametric regression Raskutti et al. (2014). Recent studies have further shown that early stopping plays a crucial role in inducing implicit sparse regularization and may even be necessary for sparse recovery in overparameterized models Vaskevicius et al. (2019); Zhao et al. (2022). Motivated by these observations, we investigate how the effective early stopping window depends on the depth N and the momentum parameter μ, and derive a lower bound on its size.

Theorem 3. *Let K_1, K_2, K_3 and K_4 be the numbers of iterations defined in Proposition 1 and Proposition 2. Assume that $\zeta \geq 1$, $\theta^*_{\max} \geq 1$ and the initialization $\alpha \leq \exp(-\frac{5}{3})$. If α and η are fixed and do not vary with N and μ, then both $K_2 - K_1$ and $K_4 - K_3$ have a lower bound, and this lower bound increasing as the depth N increases ($N > 2$).*

Remark 2. (**How depth affects the early stopping window**) The proof and detailed derivation of Theorem 3 are deferred to Appendix I. Due to the difficulty of exactly characterizing the optimal stopping interval for MGD, the quantity $K_{\mathrm{up}} - K_{\mathrm{low}}$ should be interpreted as a theoretically tractable lower bound on the maximal effective early stopping window. Importantly, we prove that this lower bound increases with N, indicating that deeper networks admit a broader time interval in which implicit sparsification can successfully emerge.

Theorem 4. *Under the same settings as Theorem 3, the size of the early stopping window $K_{\mathrm{up}} - K_{\mathrm{low}}$ decreases as the momentum parameter μ increases.*

Remark 3. (**How momentum affects the early stopping window**) The proof and detailed derivation of Theorem 4 are deferred to Appendix J. Under the learning rate condition $\eta = \mathcal{O}(\frac{1}{N^2 \gamma (\theta^*_{\max})^{\frac{3N-2}{N}}})$, we have:

$$K_{\mathrm{low}} = \frac{25}{N^2 \eta \tau (\theta^*_{\min})^{\frac{2N-2}{N}}} \log \frac{\theta^*_{\min}}{\epsilon} \mathbb{I}\{A\} + \frac{15}{4N(N-2)\eta\zeta\tau\alpha^{N-2}} + \frac{75}{4N^2\eta\tau\zeta^{\frac{2N-2}{N}}} \log \frac{|\theta^*_{\max} - \alpha^N|}{\epsilon},$$

$$K_{\mathrm{up}} = \frac{100}{\beta N(N-1)\eta\theta^*_{\min}} \left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) \mathbb{I}\{A\} + \frac{20}{\beta N(N-1)\eta\zeta} \left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}}\right),$$

where $\mathbb{I}\{A\}$ denotes the indicator function that equals 1 if condition A holds and 0 otherwise, and $A := \frac{1}{5}\theta^*_{\min} > \frac{2}{C_2} \left\| \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{\xi} \right\|_\infty \wedge \frac{2}{C_2}\epsilon$. From these formulas, we observe that K_{up} is inversely proportional to τ. Since β increases with the momentum when the momentum is large, K_{low} is also inversely proportional to μ. The above conclusion indicates that the left and right endpoints of the early stopping window for GD $K_{\mathrm{up-GD}}$ and $K_{\mathrm{low-GD}}$, are larger than those for MGD. Intuitively, the addition of momentum accelerates the convergence of both the signal and error components, leading to an earlier and narrower early stopping window for MGD compared to GD.

Remark 4. (**How initialization affects the early stopping window**) Similar to Theorems 3 and 4, we also investigate the effect of initialization scale on the early stopping window for MGD. For a fixed learning rate η, decreasing the initialization scale α enlarges the effective early stopping window. The detailed analysis is deferred to Theorem 6 in Appendix K. This result extends the observations in Li et al. (2021); Gidel et al. (2019) from GD to MGD, showing that vanishing initialization enlarges the separation between different phase transition times and further promotes implicit sparse regularization.

4 Simulation Study

In this section, we validate the above theoretical results through numerical experiments. The experimental setup is as follows: (1) each entry of the design matrix $\boldsymbol{X}$ is an independent and identically

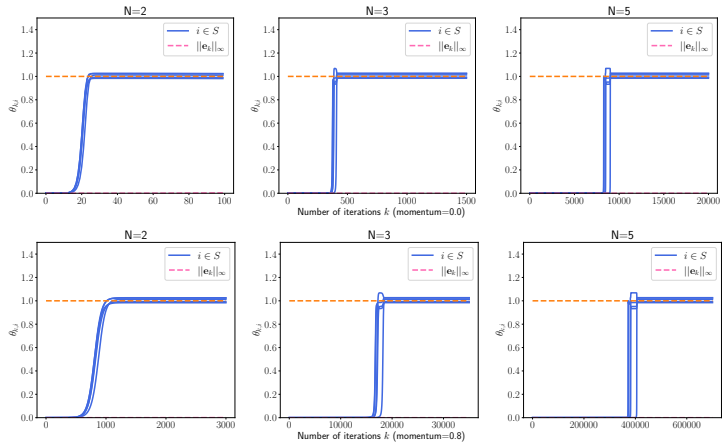


Figure 1: The evolution path of the components of $\boldsymbol{\theta}_k$ with $\alpha = 1e-6$ and $\eta = \frac{1}{N^2\gamma}$.

distributed (i.i.d.) Rademacher random variable (i.e., ± 1 with equal probability); (2) each entry of the noise vector $\boldsymbol{\xi}$ is i.i.d. following a normal distribution $\mathcal{N}(0, 0.5^2)$; (3) the ground-truth solution vector is set to $\boldsymbol{\theta}^* = \mathbf{1}_S$; (4) sample size $n = 500$, number of features per sample $p = 3000$, and sparsity level $\pi = 5$ (meaning that $\boldsymbol{\theta}^*$ has exactly 5 nonzero components).

4.1 Convergence of MGD

We first demonstrate that different choices of depth N and momentum parameter μ enable sparse recovery. To this end, we conduct experiments with various N and μ to illustrate the convergence behavior of MGD. The results are shown in Figure 1. Note that the learning rate is set to $\eta = \frac{1}{N^2\gamma}$; consequently, the number of iterations increases significantly with larger N and μ. Under a very small initialization scale, MGD requires many iterations to escape the near-zero region. Additional results for $\mu \in \{0.0, 0.2, 0.4, 0.6, 0.8, 0.9\}$ on the simulated data and experiments on the MNIST dataset are provided in Appendix L.1.

4.2 Initialization Scale

According to the theoretical analysis in Remark 1, increasing the depth N enlarges the admissible initialization scale, thereby allowing MGD to maintain the implicit sparsification effect even under larger initialization values. In our experiments, we set $\alpha^N = 2 \times 10^{-3}$. Under this initialization, MGD fails to converge to the optimal solution when $N = 2$. However, as N increases, sparse recovery gradually emerges. For instance, when $N = 3$, the error $\|e_k\|_\infty$ of MGD with momentum $\mu = 0.8$ is smaller than that of GD ($\mu = 0$), indicating that a larger μ leads to better recovery under a large initialization. The results are shown in Figure 2. Each row corresponds to a fixed μ with varying N, while each column corresponds to a fixed N with varying μ.

4.3 Early Stopping Window Size

We now examine the effect of the depth and the momentum on the early stopping window size. According to Theorem 3 and Theorem 4, we fix the initialization $\alpha = 5 \times 10^{-3}$ and $\eta = 0.01$, and conduct experiments for $N = \{2, 3, 4\}$. In each figure, every row corresponds to a fixed μ with varying N, while each column corresponds to a fixed N with varying μ. The results are shown in each row of Figure 3. Additional results for $\mu \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$ are provided in Appendix L.2. Based on these observations, we summarize two key findings: (1) Increasing the network depth enlarges the early stopping window, thereby making the implicit sparsification effect of MGD more likely to occur. (2) Adding momentum shortens the early stopping window but shifts it to occur earlier, so that MGD induces implicit sparsity earlier than GD.

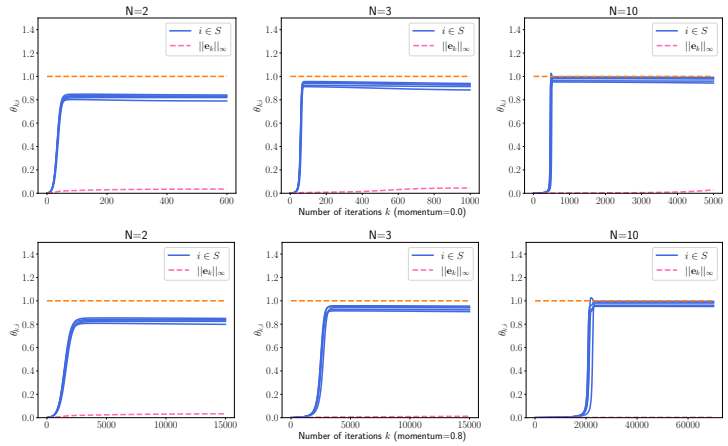


Figure 2: The effect of N and μ on the initialization with $\eta = \frac{1}{N^2\gamma}$.

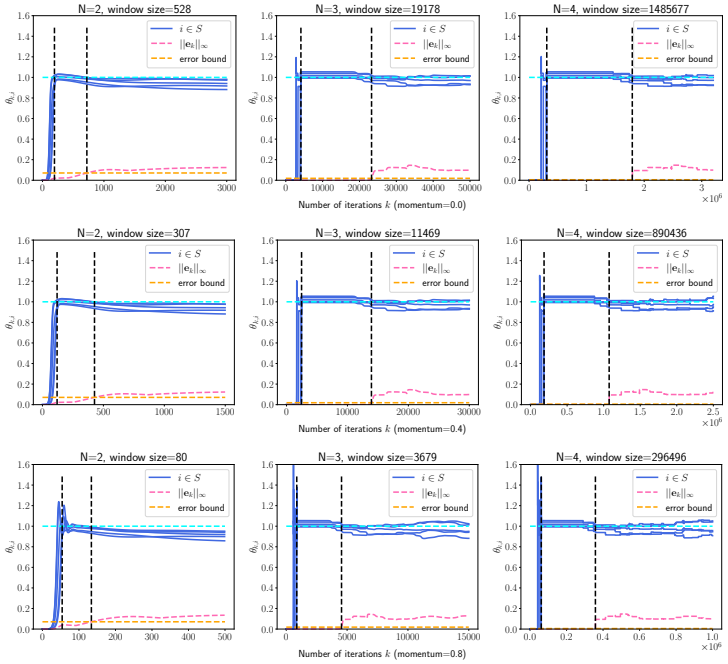


Figure 3: The blue solid line shows the evolution path of the components of $\boldsymbol{\theta}_k$ corresponding to the support set S. The black dashed line marks the early stopping window. The pink dashed line shows the absolute value of the largest component among those corresponding to S_c. The orange line represents the error bound in Theorem 2 ($|\theta_{k,i} - \theta_i^*| \lesssim \alpha^{\frac{N}{4}}$), which allows us to stop early.

To show how the early stopping window varies with N and μ, we plot the ℓ_2 error of the recovered signal. Each simulation is repeated 30 times, and the median ℓ_2 error is reported. The shaded area represents the pointwise region between the 25th and 75th percentiles. The results are shown in Figure 4. More results for $\mu \in \{0.0, 0.2, 0.4, 0.6, 0.8, 0.9\}$ are provided in Appendix L.2.

4.4 Incremental Learning

Under different choices of N and μ, MGD exhibits incremental learning dynamics Gissin et al. (2020); Wang et al. (2023). We also observe distinct learning stages in MGD. Given the same α^N and η, a larger N leads to clearer separation of the learning stages among weight components, and

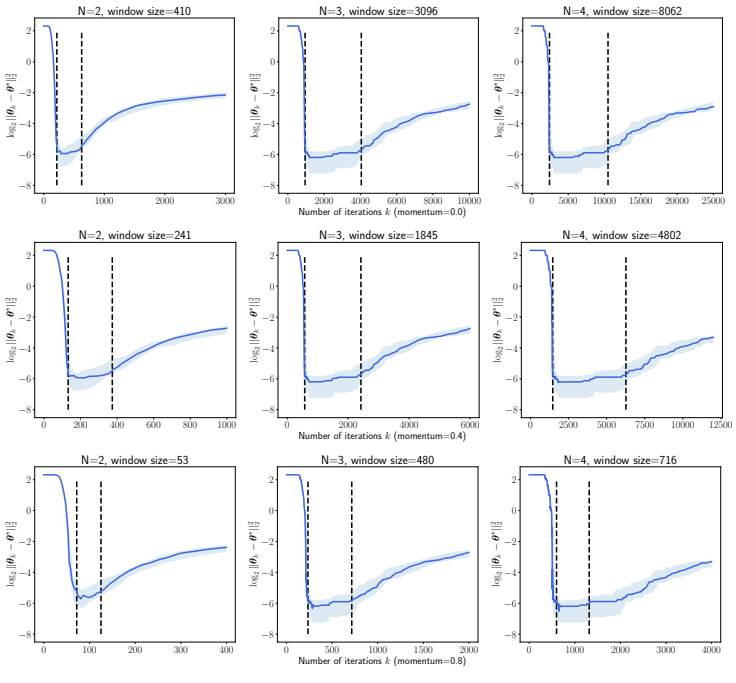


Figure 4: $\log$-ℓ_2 error of $N \in \{2, 3, 4\}$ and $\mu \in \{0.0, 0.4, 0.8\}$.

the convergence speed of each component increases with μ. Consequently, under larger μ, the early stopping time required to obtain a sparse solution occurs earlier than that of GD, and the effective window is narrower than that of GD, which is consistent with our theoretical findings in Theorem 3 and Theorem 4. The corresponding experimental results are provided in the Appendix L.3.

4.5 Kernel Regime

For diagonal linear networks, the scale of initialization determines whether the gradient dynamics operate in the kernel regime or the rich regime (the latter corresponding to the sparse recovery problem with small initialization studied in this paper) Gissin et al. (2020). To investigate the kernel regime in a more practical setting, we consider the overdetermined case where $n = 500$ and $p = 100$. The components of the ground-truth vector $\boldsymbol{\theta}^*$ are drawn independently from the distribution $\mathcal{N}(0, 25^2)$. For all tested values of N and μ, we set the same $\alpha^N = 1 \times 10^{-3}$ and $\eta = 1 \times 10^{-7}$. Experiments (see Figure 11 in Appendix L.4) show that for large initialization, adopting early stopping makes the gradient dynamics behave similarly to ridge regression. Moreover, as μ increases, MGD requires an earlier early stopping point to align with ridge regression.

5 Conclusions and Future Work

In this paper, we theoretically investigated the implicit sparse regularization of MGD in high-dimensional sparse regression. By studying depth-N diagonal linear networks, we established that MGD with early stopping achieves minimax optimal sparse recovery under sufficiently small initialization and learning rate. Moreover, we showed that increasing depth enlarges both the admissible initialization scale and the effective early stopping window, thereby making implicit sparsification more likely to occur. More importantly, we characterized the distinct temporal effect of momentum: larger momentum shifts the early stopping window toward earlier iterations while simultaneously shortening its duration. This indicates that MGD induces sparse structures earlier during training compared with GD. Future directions include providing a rigorous theoretical understanding of the incremental learning dynamics and kernel regimes observed empirically, and applying our theoretical insights to guide practical hyperparameter selection in real-world deep learning tasks, such as network depth, momentum coefficient, learning rate, and the design of early stopping strategies.

REFERENCES

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1370–1378. PMLR, 2019.

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International conference on machine learning*, pp. 233–244. PMLR, 2020.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International conference on machine learning*, pp. 244–253. PMLR, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards ai. *Large Scale Kernel Machines*, pp. 321–359, 2007. doi: 10.7551/mitpress/7496.003.0016.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. In *Proceedings of the 34th Advances in Neural Information Processing Systems*, pp. 28811–28822, 2021. doi: 10.1145/3578580.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pp. 7–10. ACM, 2016. doi: 10.1145/2988450.2988454.

Olivier Delalleau and Yoshua Bengio. Shallow vs. deep sum-product networks. In *Proceedings of the 24th Advances in Neural Information Processing Systems*, pp. 666–674, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/8e6b42f1644ecb1327dc03ab345e618b-Abstract.html.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Conference on Learning Theory*, volume 49, pp. 907–940. JMLR.org, 2016. doi: 10.48550/arXiv.1512.03965.

Simon Foucart and Holger Rauhut. An invitation to compressive sensing. In *A mathematical introduction to compressive sensing*, pp. 1–39, 2013. URL https://api.semanticscholar.org/CorpusID:118414184.

Jingwen Fu, Bohan Wang, Huishuai Zhang, Zhizheng Zhang, Wei Chen, and Nanning Zheng. When and why momentum accelerates SGD: an empirical study. *CoRR*, abs/2306.09000, 2023. doi: 10.48550/ARXIV.2306.09000. URL https://doi.org/10.48550/arXiv.2306.09000.

Khashayar Gatmiry, Zhiyuan Li, Sashank J. Reddi, and Stefanie Jegelka. Simplicity bias via global convergence of sharpness minimization. In *Proceedings of the 41th International Conference on Machine Learning*, pp. 15102–15129. OpenReview.net, 2024. doi: 10.48550/arXiv.2410.16401.

Avrajit Ghosh, He Lyu, Xitong Zhang, and Rongrong Wang. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=ZzdBhtEH9yB.

Gauthier Gidel, Francis R. Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3196–3206, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f39ae9ff3a81f499230c4126e01f421b-Abstract.html.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1lj0nNFwB.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018a.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1827–1836. PMLR, 2018b. doi: 10.48550/arXiv.1802.08246.

Jeff Z. HaoChen, Colin Wei, Jason D. Lee, and Tengyu Ma. Shape matters: Understanding the implicit bias of the noise covariance. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2315–2357. PMLR, 2021. doi: 10.48550/arXiv.2006.08680.

Kais Hariz, Hachem Kadri, Stéphane Ayache, Maher Moakher, and Thierry Artières. Implicit regularization in deep tucker factorization: Low-rankness via structured sparsity. In *International Conference on Artificial Intelligence and Statistics*, 2024. URL https://api.semanticscholar.org/CorpusID:269648701.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pp. 9965–10040. PMLR, 2022. URL https://proceedings.mlr.press/v162/jelassi22a.html.

V. Koltchinskii. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability II*, pp. 443–457, 2000.

V. Koltchinskii. Rademacher penalties and structural risk minimization. 47(5):1902–1914, 2006. ISSN 0018-9448. doi: 10.1109/18.930926. URL https://doi.org/10.1109/18.930926.

Guillaume Leclerc and Aleksander Madry. The two regimes of deep network training. *CoRR*, abs/2002.10376, 2020. URL https://arxiv.org/abs/2002.10376.

Dawei Li, Jin Wang, and Xuejie Zhang. Yun-hpcc at semeval-2019 task 3: Multi-step ensemble neural network for sentiment analysis in textual conversation. In *International Workshop on Semantic Evaluation*, 2019. URL https://api.semanticscholar.org/CorpusID:184482660.

Jiangyuan Li, Thanh Nguyen, Chinmay Hegde, and Ka Wai Wong. Implicit sparse regularization: The impact of depth and early stopping. *Advances in Neural Information Processing Systems*, 34: 28298–28309, 2021.

Shiyu Liang and R. Srikant. Why deep neural networks for function approximation? In *Proceedings of the 5th International Conference on Learning Representations*. OpenReview.net, 2017. doi: 10.48550/arXiv.1610.04161.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Proceedings of the 30th Advances in Neural Information Processing Systems*, pp. 6231–6239, 2017. doi: 10.48550/arXiv.1709.02540.

Bochen Lyu and Zhanxing Zhu. Implicit bias of (stochastic) gradient descent for rank-1 linear neural network. *Advances in Neural Information Processing Systems 36*, 2023. URL https://api.semanticscholar.org/CorpusID:268064140.

Bochen Lyu, He Wang, Zheng Wang, and Zhanxing Zhu. Effects of momentum in implicit bias of gradient flow for diagonal linear networks. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 19242–19250. AAAI Press, 2025. doi: 10.1609/AAAI.V39I18.34118. URL https://doi.org/10.1609/aaai.v39i18.34118.

James Martens and Roger Baker Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, 2015. URL https://api.semanticscholar.org/CorpusID:11480464.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Proceedings of the 34th Advances in Neural Information Processing Systems*, pp. 29218–29230, 2021. doi: 10.48550/arXiv.2106.09524.

Boris Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4:1–17, 1964. URL https://api.semanticscholar.org/CorpusID:120243018.

Ali Ramezani-Kebrya, Kimon Antonakopoulos, Volkan Cevher, Ashish Khisti, and Ben Liang. On the generalization of stochastic gradient descent with momentum. *Journal of Machine Learning Research*, 25(22):1–56, 2024. URL http://jmlr.org/papers/v25/22-0068.html.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15(1):335–366, January 2014. ISSN 1532-4435.

Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pp. 8913–8924. PMLR, 2021.

Levent Sagun, Utku Evci, V. Ugur Güney, Yann N. Dauphin, and Léon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJO1_M0Lf.

Naoki Sato and Hideaki Iiduka. Role of momentum in smoothing objective function in implicit graduated optimization. *ArXiv*, abs/2402.02325, 2024. URL https://api.semanticscholar.org/CorpusID:267412561.

Othmane Sebbouh, Robert Mansel Gower, Aaron Defazio, and IP Paris. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Annual Conference Computational Learning Theory*, 2021. URL https://api.semanticscholar.org/CorpusID:235310402.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19:2822–2878, 2018. doi: 10.5555/3291125.3309632.

Tao Sun, Dongsheng Li, Zhe Quan, Hao Jiang, Shengguo Li, and Yong Dou. Heavy-ball algorithms always escape saddle points. In *International Joint Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:198179405.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL https://doi.org/10.1137/1116025.

Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4224–4258. PMLR, 2021. doi: 10.48550/arXiv.2012.05156.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 2968–2979, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5cf21ce30208cfffaa832c6e44bb567d-Abstract.html.

Bohan Wang, Qi Meng, Huishuai Zhang, Ruoyu Sun, Wei Chen, Zhi-Ming Ma, and Tie-Yan Liu. Does momentum change the implicit regularization on separable data? In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ab3f6bbe121a8f7a0263a9b393000741-Abstract-Conference.html.

Li Wang, Zhiguo Fu, Yingcong Zhou, and Zili Yan. The implicit regularization of momentum gradient descent in overparametrized models. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:259681665.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9508–9520, 2024.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tong Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Ann. Statist.*, 33:1538–1579, 2004. doi: 10.1214/009053605000000255.

Peng Zhao, Yun Yang, and Qiao-Chu He. High-dimensional linear regression via implicit regularization. *Biometrika*, 109(4):1033–1046, 12 2022. ISSN 1464-3510. doi: 10.1093/biomet/asac010. URL https://doi.org/10.1093/biomet/asac010.

Appendix

A THE PROOF OF THEOREM 1

Controllability of approximation errors. We analyze the three-step approximation as follows: (a) The two-step MGD $\rightarrow$ first-order ODE (Lemma 1) incurs an error of order $\mathcal{O}(\eta^2)$, which is proved in Ghosh et al. (2023); (b) the forward Euler discretization of the continuous ODE has a global error of order $\mathcal{O}(\eta)$, which is a standard result in numerical analysis; (c) the Gauss-Newton approximation introduces an error of order $\mathcal{O}(\eta)$. The total error is $\mathcal{O}(\eta)$ which is controllable. The experimental results are shown in Figure 5.

The error analysis of the Gauss-Newton approximation derived from (c) is given below. Let the true Hessian be $\nabla^2 L(\boldsymbol{u}, \boldsymbol{v})$ and its Gauss-Newton approximation is $\nabla^2_{\mathrm{GN}} L(\boldsymbol{u}, \boldsymbol{v})$. Crucially, all analysis below is performed in the original parameter space $(\boldsymbol{u}, \boldsymbol{v})$ of the depth-N diagonal network. The first-order discretized dynamics in $(\boldsymbol{u}, \boldsymbol{v})$ space are:

$$\boldsymbol{u}_{k+1}^{\mathrm{lst}} = \boldsymbol{u}_k^{\mathrm{lst}} - \eta\tau\nabla_{\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{lst}}, \boldsymbol{v}_k^{\mathrm{lst}}) - \eta^2\gamma\nabla^2_{\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{lst}}, \boldsymbol{v}_k^{\mathrm{lst}})\nabla_{\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{lst}}, \boldsymbol{v}_k^{\mathrm{lst}}),$$
$$\boldsymbol{v}_{k+1}^{\mathrm{lst}} = \boldsymbol{v}_k^{\mathrm{lst}} - \eta\tau\nabla_{\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{lst}}, \boldsymbol{v}_k^{\mathrm{lst}}) - \eta^2\gamma\nabla^2_{\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{lst}}, \boldsymbol{v}_k^{\mathrm{lst}})\nabla_{\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{lst}}, \boldsymbol{v}_k^{\mathrm{lst}}).$$

where $\tau = \frac{1}{1-\mu}$ and $\gamma = \frac{1+\mu}{2(1-\mu)^3}$.

The Gauss-Newton approximated dynamics can be obtained by

$$\boldsymbol{u}_{k+1}^{\mathrm{GN}} = \boldsymbol{u}_k^{\mathrm{GN}} - \eta\tau\nabla_{\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}}) - \eta^2\gamma\nabla^2_{\mathrm{GN},\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}})\nabla_{\boldsymbol{u}} L(\boldsymbol{u}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}}),$$
$$\boldsymbol{v}_{k+1}^{\mathrm{GN}} = \boldsymbol{v}_k^{\mathrm{GN}} - \eta\tau\nabla_{\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}}) - \eta^2\gamma\nabla^2_{\mathrm{GN},\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}})\nabla_{\boldsymbol{v}} L(\boldsymbol{v}_k^{\mathrm{GN}}, \boldsymbol{v}_k^{\mathrm{GN}}).$$

To avoid cumbersome notation with the full $(\boldsymbol{u}, \boldsymbol{v})$-state, we write the combined state as $\boldsymbol{w}_k = (\boldsymbol{u}_k, \boldsymbol{v}_k) \in \mathbb{R}^{2p}$ and denote the two one-step maps as $\boldsymbol{w}_{k+1}^{\mathrm{1st}} = \Phi_{\mathrm{1st}}(\boldsymbol{w}_k^{\mathrm{1st}})$ and $\boldsymbol{w}_{k+1}^{\mathrm{GN}} = \Phi_{\mathrm{GN}}(\boldsymbol{w}_k^{\mathrm{GN}})$, with the same initial condition $\boldsymbol{w}_0^{\mathrm{1st}} = \boldsymbol{w}_0^{\mathrm{GN}} = (\alpha\mathbf{1}, \alpha\mathbf{1})$.

It is worth noting that $\Phi_{\mathrm{1st}}(\boldsymbol{w})$ is a finite composition of elementary functions, and is therefore continuously differentiable. Below, we examine the boundedness of the Jacobian of $\Phi_{\mathrm{1st}}(\boldsymbol{w})$. From Propositions 1 and 2, for every $k \in [K_{\mathrm{low}}, K_{\mathrm{up}}]$, we have $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|_\infty \lesssim \zeta$. Consequently, the parameters $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are bounded. We define the compact set of parameters satisfying this condition as $\Omega = \{(\boldsymbol{u}, \boldsymbol{v}) | \|\boldsymbol{u}\|_\infty, \|\boldsymbol{v}\|_\infty \leq c\}$. Therefore, the Jacobian matrix $J_{\Phi_{\mathrm{1st}}}$ of $\Phi_{\mathrm{1st}}(\boldsymbol{w})$ is bounded on the set Ω , i.e.,

$$||J_{\Phi_{\mathrm{1st}}}(\boldsymbol{w})||_\infty = ||\boldsymbol{I} - \eta\tau\nabla^2 L(\boldsymbol{w}) - \eta^2\gamma\frac{d}{dz}[\nabla^2 L(\boldsymbol{w})\nabla L(\boldsymbol{w})]||_\infty \leq 1 + C\eta.$$

For any $\boldsymbol{w} \in \Omega$, the difference between the first-order dynamics and the GN-approximated dynamics is

$$\Phi_{\mathrm{1st}}(\boldsymbol{w}) - \Phi_{\mathrm{GN}}(\boldsymbol{w}) = -\eta^2\gamma R(\boldsymbol{w})\nabla L(\boldsymbol{w}),$$

where $R(\boldsymbol{w}) = \nabla^2 L(\boldsymbol{w}) - \nabla^2_{\mathrm{GN}} L(\boldsymbol{w})$ is the Hessian residual in $(\boldsymbol{u}, \boldsymbol{v})$ space.

Within Ω, the residual $\boldsymbol{r} = \boldsymbol{X}(\boldsymbol{u}^N - \boldsymbol{v}^N) - \boldsymbol{y}$ is bounded. Thus, the Hessian residual $R(\boldsymbol{w})$ is bounded: $\|R(\boldsymbol{w})\| \leq c$. Therefore,

$$\delta(\boldsymbol{w}) := ||\Phi_{\mathrm{1st}}(\boldsymbol{w}) - \Phi_{\mathrm{GN}}(\boldsymbol{w})||_\infty \leq \delta_{\max},$$

uniformly for all $\boldsymbol{w} \in \Omega$, where $\delta_{\max} = \mathcal{O}\left(\eta^2\right)$.

Let $\boldsymbol{\Delta}_k := \boldsymbol{w}_k^{\mathrm{1st}} - \boldsymbol{w}_k^{\mathrm{GN}}$. Then

$$\boldsymbol{\Delta}_{k+1} = \Phi_{\mathrm{1st}}(\boldsymbol{w}_k^{\mathrm{1st}}) - \Phi_{\mathrm{GN}}(\boldsymbol{w}_k^{\mathrm{GN}}) = [\Phi_{\mathrm{1st}}(\boldsymbol{w}_k^{\mathrm{1st}}) - \Phi_{\mathrm{1st}}(\boldsymbol{w}_k^{\mathrm{GN}})] + [\Phi_{\mathrm{1st}}(\boldsymbol{w}_k^{\mathrm{GN}}) - \Phi_{\mathrm{GN}}(\boldsymbol{w}_k^{\mathrm{GN}})].$$

Taking the ℓ_∞-norm and applying the Lipschitz bound,

$$||\boldsymbol{\Delta}_{k+1}||_\infty \leq (1 + C\eta)||\Delta_k||_\infty + \delta_{\max}.$$

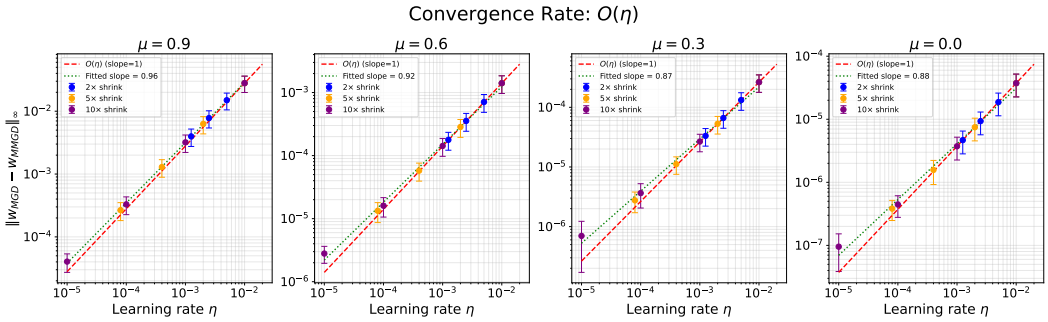


Figure 5: We verify Theorem 1 on a depth-$N = 3$ diagonal linear network. The data are generated according to $y = X\theta^\star + \xi$, with $n = 50$, $p = 200$, $\pi = 10$, and $\sigma = 0.01$. MGD and MMGD are initialized identically with $\alpha = 0.1$ and run for $K_{\text{up}} = \mathcal{O}(1/(\gamma\eta))$ iterations. We consider four momentum coefficients, $\mu \in \{0.9, 0.6, 0.3, 0\}$. For each μ, the learning rate η is shrunk from 0.01 along three geometric schedules with reduction factors of $(2\times, 5\times, 10\times)$, yielding four learning rates for each schedule. Each (η, μ) configuration is repeated over 10 random seeds. We report the mean and standard deviation of the final discrepancy $\|\boldsymbol{w}_{\text{MGD}} - \boldsymbol{w}_{\text{MMGD}}\|_\infty$.

This is a standard linear recursion. Unrolling from $k = 0$ to $K - 1$, with $\boldsymbol{\Delta}_0 = 0$,

$$\|\boldsymbol{\Delta}_K\|_\infty \le \delta_{\max} \sum_{j=0}^{K-1} (1 + C\eta)^j = \delta_{\max} \frac{(1 + C\eta)^K - 1}{C\eta} \le \delta_{\max} \frac{e^{CK\eta} - 1}{C\eta}.$$

Take $K = K_{\text{up}}$, the upper bound of the early-stopping window defined in Theorem 1. From equations (18) and (19) in the paper,$K_{\text{up}} = \mathcal{O}\left(\frac{1}{\eta}\right)$, Hence$K_{\text{up}}\eta = \mathcal{O}(1) =: T$, which is independent of η. Therefore $e^{CK_{\text{up}}\eta} - 1 = e^{CT} - 1 = \mathcal{O}(1)$.

Combining with the bound on $\delta_{\max}$,

$$||\Delta_{K_{\text{up}}}||_\infty \le \mathcal{O}(\eta)\,.$$

We have completed the error analysis of the Gauss-Newton approximation.

Figure 5 empirically confirms the predicted $\mathcal{O}(\eta)$ approximation error as shown in Theorem 1: the error decreases monotonically with η and exhibits fitted slopes consistently close to 1 across all momentum.

B THE PROOF OF COROLLARY 1

We consider the modified Moment Gradient Flow (MGF) as follows:

$$\dot{\boldsymbol{w}}(t) = -\frac{1}{1-\mu}\nabla\hat{L}(\boldsymbol{w}(t)), \quad t \in (0, T], \tag{7}$$

where μ is the momentum parameter. The modified loss is

$$\hat{L}(\boldsymbol{w}(t)) = L(\boldsymbol{w}(t)) + \frac{(1+\mu)\eta}{4(1-\mu)^2}\|\nabla L(\boldsymbol{w}(t))\|_2^2.$$

We can discretize (7) as follows,

$$\begin{aligned}\boldsymbol{w}_{k+1} =& \boldsymbol{w}_k - \frac{\eta}{1-\mu}\nabla L(\boldsymbol{w}_k) - \frac{(1+\mu)\eta^2}{4(1-\mu)^3}\nabla\|\nabla L(\boldsymbol{w}_k)\|_2^2 \\ =& \boldsymbol{w}_k - \frac{\eta}{1-\mu}\nabla L(\boldsymbol{w}_k) - \frac{(1+\mu)\eta^2}{2(1-\mu)^3}\nabla^2 L(\boldsymbol{w}_k)\nabla L(\boldsymbol{w}_k).\end{aligned} \tag{8}$$

where $\boldsymbol{w}_k = (\boldsymbol{u}_k, \boldsymbol{v}_k)$. We consider the following square loss function

$$L(\boldsymbol{\theta}) = \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}\|^2 = \frac{1}{2n}\sum_{i=1}^n (\boldsymbol{x}_i\boldsymbol{\theta} - y_i)^2 = \frac{1}{2n}\sum_{i=1}^n \ell(f_i(\boldsymbol{\theta}), y_i) = \frac{1}{2n}\sum_{i=1}^n \ell_i(\boldsymbol{\theta}).$$

where $\boldsymbol{\theta} = \boldsymbol{u}^N - \boldsymbol{v}^N$.

By using the Gauss-Newton decomposition, we have

$$\nabla^2 L(\boldsymbol{\theta}) \approx \frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}(f_i(\boldsymbol{\theta}))\nabla f_i(\boldsymbol{\theta})\nabla f_i(\boldsymbol{\theta})^\top,$$

In what follows, we focus on $\boldsymbol{\theta} = \boldsymbol{u}^N$; the case for $\boldsymbol{\theta} = \boldsymbol{u}^N - \boldsymbol{v}^N$ follows a similar argument. The first-order derivatives of $f_i(\boldsymbol{u})$ ($i \in [n]$) are calculated as follows

$$\nabla f_i(\boldsymbol{u}) = \begin{pmatrix} N \cdot x_i^1 \cdot u_1^{N-1} \\ N \cdot x_i^2 \cdot u_2^{N-1} \\ \vdots \\ N \cdot x_i^p \cdot u_p^{N-1} \end{pmatrix} = N \begin{pmatrix} u_1^{N-1} & & \\ & u_2^{N-1} & \\ & \ddots & \\ & & u_p^{N-1} \end{pmatrix} \begin{pmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^p \end{pmatrix} = N\mathrm{Diag}\left(\boldsymbol{u}^{N-1}\right)\boldsymbol{x}_i^\top,$$

where $\mathrm{Diag}(\cdot)$ denotes the operation of converting a vector into a diagonal matrix.

Based on the above, we approximate the Hessian as follows

$$\begin{aligned} J_1 =& \frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}(f_i(\boldsymbol{u}))\nabla f_i(\boldsymbol{u})\nabla f_i(\boldsymbol{u})^T \\ =& \frac{N^2}{n}\sum_{i=1}^{n} \mathrm{Diag}(\boldsymbol{u}^{N-1})\boldsymbol{x}_i^T\boldsymbol{x}_i\mathrm{Diag}(\boldsymbol{u}^{N-1}) \\ =& \mathrm{Diag}(\boldsymbol{u}^{N-1})\left\{\frac{N^2}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^T\boldsymbol{x}_i\right\}\mathrm{Diag}(\boldsymbol{u}^{N-1}) \\ =& \mathrm{Diag}(\boldsymbol{u})^{N-1}\frac{N^2}{n}\boldsymbol{X}^T\boldsymbol{X}\mathrm{Diag}(\boldsymbol{u})^{N-1}. \end{aligned}$$

Therefore, (8) can be calculated as

$$\begin{aligned} &\boldsymbol{u}_{k+1} = \boldsymbol{u}_k - \frac{\eta}{1-\mu}\nabla L(\boldsymbol{u}_k) - \frac{(1+\mu)\eta^2}{2(1-\mu)^3}\nabla^2 L(\boldsymbol{u}_k)\nabla L(\boldsymbol{u}_k) \\ =& \boldsymbol{u}_k - \left[\frac{\eta}{1-\mu}\boldsymbol{I} + \frac{(1+\mu)\eta^2}{2(1-\mu)^3}\nabla^2 L(\boldsymbol{u}_k)\right]\nabla L(\boldsymbol{u}_k) \\ =& \boldsymbol{u}_k - \eta\left\{\tau\boldsymbol{I} + \eta\gamma\left[\mathrm{Diag}\left(\boldsymbol{u}_k^{N-1}\right)\frac{N^2}{n}\boldsymbol{X}^T\boldsymbol{X}\mathrm{Diag}\left(\boldsymbol{u}_k^{N-1}\right)\right]\right\}\cdot\left(\frac{N}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{u}_k^N - \boldsymbol{y})\odot\boldsymbol{u}_k^{N-1}\right) \end{aligned} \tag{9}$$

where $\odot$ denotes the Hadamard product, and

$$\tau = \frac{1}{1-\mu} \in [1, +\infty), \quad \gamma = \frac{1+\mu}{2(1-\mu)^3} \in [\frac{1}{2}, +\infty).$$

The proof of Corollary 1 is completed.

Furthermore, we consider the decomposition of signal and noise,

$$\begin{aligned}
&\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\theta}_k-\boldsymbol{y})\\
=&\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\theta}_k-\boldsymbol{X}\boldsymbol{\theta}^*-\boldsymbol{X})\\
=&\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{X}(\boldsymbol{e}_k+\boldsymbol{s}_k-\boldsymbol{\theta}^*)-\boldsymbol{X})\\
=&\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\theta}_k-\boldsymbol{X}\boldsymbol{\theta}^*-\boldsymbol{X})\\
=&\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{X}(\mathbf{1}_S\boldsymbol{\theta}_k+\mathbf{1}_{S^c}\boldsymbol{\theta}_k^N-\boldsymbol{\theta}^*)-\boldsymbol{X})\\
=&\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{e}_k+\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{s}_k-\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}^*-\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\\
=&\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{e}_k-\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}+\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{s}_k-\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}^*+\boldsymbol{s}_k-\boldsymbol{\theta}^*-(\boldsymbol{s}_k-\boldsymbol{\theta}^*)\\
=&\boldsymbol{b}_k+\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}-I\right)(\boldsymbol{s}_k-\boldsymbol{\theta}^*)+\boldsymbol{s}_k-\boldsymbol{\theta}^*\\
=&\boldsymbol{p}_k+\boldsymbol{b}_k+\boldsymbol{s}_k-\boldsymbol{\theta}^*,
\end{aligned}$$

Let $\boldsymbol{\theta}_k=\boldsymbol{\theta}_k^+-\boldsymbol{\theta}_k^-$ where $\boldsymbol{\theta}_k^+:=\boldsymbol{u}^N$ and $\boldsymbol{\theta}_k^-:=\boldsymbol{v}^N$. Let S be the support set of $\boldsymbol{\theta}^*$, and let $S_+=\{i\mid\theta_i^*>0\}$ denote the set of indices where the components are positive, and $S_-=\{i\mid\theta_i^*<0\}$ denote the set of indices where the components are negative, and

$$\begin{aligned}
\boldsymbol{s}_k&:=\mathbf{1}_{S^+}\odot\boldsymbol{\theta}_k^+-\mathbf{1}_{S^-}\odot\boldsymbol{\theta}_k^-,\\
\boldsymbol{e}_k&:=\mathbf{1}_{S^c}\odot\boldsymbol{\theta}_k+\mathbf{1}_{S^-}\odot\boldsymbol{\theta}_k^+-\mathbf{1}_{S^+}\odot\boldsymbol{\theta}_k^-,\\
\boldsymbol{b}_k&:=\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{e}_k-\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi},\\
\boldsymbol{p}_k&:=\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{I}\right)(\boldsymbol{s}_k-\boldsymbol{\theta}^*).
\end{aligned}$$

Substituting $\boldsymbol{s}_k,\boldsymbol{e}_k,\boldsymbol{b}_k$ and $\boldsymbol{p}_k$ into equation (9) yields

$$\begin{aligned}
\boldsymbol{u}_{k+1}=&\boldsymbol{u}_k-\eta\left\{\tau\boldsymbol{I}+\eta\gamma\left[\mathrm{Diag}\left(\boldsymbol{u}_k^{N-1}\right)\frac{N^2}{n}\boldsymbol{X}^T\boldsymbol{X}\mathrm{Diag}\left(\boldsymbol{u}_k^{N-1}\right)\right]\right\}\\
&\cdot\left((\boldsymbol{p}_k+\boldsymbol{b}_k+\boldsymbol{s}_k-\boldsymbol{\theta}^*)\odot\boldsymbol{\theta}_k^{N-1}\right).
\end{aligned}$$

Similarly, it follows that

$$\begin{aligned}
\boldsymbol{v}_{k+1}=&\boldsymbol{v}_k+\eta\left\{\tau\boldsymbol{I}+\eta\gamma\left[\mathrm{Diag}\left(\boldsymbol{v}_k^{N-1}\right)\frac{N^2}{n}\boldsymbol{X}^T\boldsymbol{X}\mathrm{Diag}\left(\boldsymbol{v}_k^{N-1}\right)\right]\right\}\\
&\cdot\left((\boldsymbol{p}_k+\boldsymbol{b}_k+\boldsymbol{s}_k-\boldsymbol{\theta}^*)\odot\boldsymbol{\theta}_k^{N-1}\right).
\end{aligned}$$

To facilitate the subsequent theoretical analysis, we need to separate $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ from the Gauss–Newton term.

Let $\boldsymbol{G}:=\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}=\boldsymbol{I}+\boldsymbol{E}$, $\boldsymbol{\rho}_k:=\boldsymbol{s}_k-\boldsymbol{\theta}^*$ and $\boldsymbol{q}_k:=\boldsymbol{\rho}_k+\boldsymbol{p}_k+\boldsymbol{b}_k$. For the $\boldsymbol{u}$-branch, we define $\boldsymbol{H}_k^{\boldsymbol{u}}:=\tau\boldsymbol{I}+\eta\gamma N^2\,\mathrm{Diag}(\boldsymbol{u}_k^{2N-2})$, and $\boldsymbol{\delta}_k^{\boldsymbol{u}}:=\eta\gamma N^2(\boldsymbol{H}_k^u)^{-1}\boldsymbol{E}\left(\boldsymbol{q}_k\odot\boldsymbol{u}_k^{2N-2}\right)$. With the notation introduced, we obtain

$$\boldsymbol{p}_k=(\boldsymbol{G}-\boldsymbol{I})(\boldsymbol{s}_k-\boldsymbol{\theta}^*)=\boldsymbol{E}\boldsymbol{\rho}_k,\quad \boldsymbol{b}_k=\boldsymbol{G}\boldsymbol{e}_k-\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}.$$

Hence the residual gradient can be written as

$$\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{r}_k=\boldsymbol{\rho}_k+\boldsymbol{p}_k+\boldsymbol{b}_k:=\boldsymbol{q}_k.$$

From Corollary 1, the $\boldsymbol{u}$-update is

$$\boldsymbol{u}_{k+1}=\boldsymbol{u}_k-\eta\left[\tau\boldsymbol{I}+\eta\gamma N^2\boldsymbol{D}_k^{\boldsymbol{u}}\boldsymbol{G}\boldsymbol{D}_k^{\boldsymbol{u}}\right]\left(N\boldsymbol{q}_k\odot\boldsymbol{u}_k^{N-1}\right).$$

where $\boldsymbol{D}_k^{\boldsymbol{u}} := \mathrm{Diag}(\boldsymbol{u}_k^{N-1})$. Using $\boldsymbol{G} = \boldsymbol{I} + \boldsymbol{E}$, we have

$$\boldsymbol{D}_k^{\boldsymbol{u}}\boldsymbol{G}\boldsymbol{D}_k^{\boldsymbol{u}} = \boldsymbol{D}_k^{\boldsymbol{u}}(\boldsymbol{I}+\boldsymbol{E})\boldsymbol{D}_k^{\boldsymbol{u}} = \boldsymbol{D}_k^{\boldsymbol{u}}\boldsymbol{D}_k^{u} + \boldsymbol{D}_k^{u}\boldsymbol{E}\boldsymbol{D}_k^{\boldsymbol{u}} = \mathrm{Diag}(\boldsymbol{u}_k^{2N-2}) + \boldsymbol{D}_k^{\boldsymbol{u}}\boldsymbol{E}\boldsymbol{D}_k^{\boldsymbol{u}}.$$

Finally, we obtain

$$\begin{aligned}\boldsymbol{u}_{k+1} &= \boldsymbol{u}_k - N\eta\left[\tau\boldsymbol{I} + \eta\gamma N^2\,\mathrm{Diag}(\boldsymbol{u}_k^{2N-2})\right]\left(\boldsymbol{q}_k \odot \boldsymbol{u}_k^{N-1}\right) - \eta^2\gamma N^3 \boldsymbol{D}_k^{\boldsymbol{u}}\boldsymbol{E}\boldsymbol{D}_k^{\boldsymbol{u}}\left(\boldsymbol{q}_k \odot \boldsymbol{u}_k^{N-1}\right)\\ &= \boldsymbol{u}_k - N\eta H_k^{\boldsymbol{u}}\left(\boldsymbol{q}_k \odot \boldsymbol{u}_k^{N-1}\right) - \eta^2\gamma N^3 \boldsymbol{u}_k^{N-1} \odot \left(\boldsymbol{E}\left(\boldsymbol{q}_k \odot \boldsymbol{u}_k^{2N-2}\right)\right).\end{aligned}$$

Define $\boldsymbol{R}_k^{\boldsymbol{u}} := \eta\gamma N^2(\boldsymbol{H}_k^{\boldsymbol{u}})^{-1}$, and $\boldsymbol{\delta}_k^{\boldsymbol{u}} := \boldsymbol{R}_k^{\boldsymbol{u}}\boldsymbol{E}\big(\boldsymbol{q}_k \odot \boldsymbol{u}_k^{2N-2}\big)$. It is easy to know that $\boldsymbol{H}_k^{\boldsymbol{u}}\boldsymbol{\delta}_k^{\boldsymbol{u}} = \eta\gamma N^2\boldsymbol{E}\left(\boldsymbol{q}_k \odot \boldsymbol{u}_k^{2N-2}\right)$. Thus,

$$N\eta\boldsymbol{H}_k^{\boldsymbol{u}}\left(\boldsymbol{\delta}_k^{\boldsymbol{u}} \odot \boldsymbol{u}_k^{N-1}\right) = N\eta\left[\boldsymbol{H}_k^{\boldsymbol{u}}\boldsymbol{\delta}_k^{\boldsymbol{u}}\right] \odot \boldsymbol{u}_k^{N-1} = \eta^2\gamma N^3 \boldsymbol{u}_k^{N-1} \odot \left(\boldsymbol{E}\left(\boldsymbol{q}_k \odot \boldsymbol{u}_k^{2N-2}\right)\right).$$

Finally, we know

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k - N\eta\boldsymbol{H}_k^{\boldsymbol{u}}\left[(\boldsymbol{q}_k + \boldsymbol{\delta}_k^{\boldsymbol{u}}) \odot \boldsymbol{u}_k^{N-1}\right],$$

Therefore,

$$\begin{aligned}\boldsymbol{\delta}_k^{\boldsymbol{u}} &= \boldsymbol{R}_k^{\boldsymbol{u}}\boldsymbol{E}\left[(\boldsymbol{\rho}_k + \boldsymbol{p}_k + \boldsymbol{b}_k) \odot \boldsymbol{u}_k^{2N-2}\right]\\ &= \boldsymbol{R}_k^{\boldsymbol{u}}\boldsymbol{E}\left[(\boldsymbol{\rho}_k + \boldsymbol{p}_k) \odot \boldsymbol{u}_k^{2N-2}\right] + \boldsymbol{R}_k^{\boldsymbol{u}}\boldsymbol{E}\left[\boldsymbol{b}_k \odot \boldsymbol{u}_k^{2N-2}\right]\\ &= \boldsymbol{\delta}_k^{\boldsymbol{u},s} + \boldsymbol{\delta}_k^{\boldsymbol{u},\boldsymbol{b}}.\end{aligned}$$

Hence

$$\begin{aligned}\boldsymbol{q}_k + \boldsymbol{\delta}_k^{\boldsymbol{u}} &= \boldsymbol{\rho}_k + \boldsymbol{p}_k + \boldsymbol{b}_k + \boldsymbol{\delta}_k^{\boldsymbol{u},s} + \boldsymbol{\delta}_k^{\boldsymbol{u},b}\\ &= \boldsymbol{\rho}_k + (\boldsymbol{p}_k + \boldsymbol{\delta}_k^{\boldsymbol{u},s}) + \left(\boldsymbol{b}_k + \boldsymbol{\delta}_k^{\boldsymbol{u},b}\right)\\ &:= \boldsymbol{\rho}_k + \widetilde{\boldsymbol{p}}_k^{\boldsymbol{u}} + \widetilde{\boldsymbol{b}}_k^{\boldsymbol{u}}.\end{aligned}$$

Consequently,

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k - N\eta\boldsymbol{H}_k^{\boldsymbol{u}}\left[\left(\boldsymbol{\rho}_k + \widetilde{\boldsymbol{p}}_k^{\boldsymbol{u}} + \widetilde{\boldsymbol{b}}_k^{\boldsymbol{u}}\right) \odot \boldsymbol{u}_k^{N-1}\right]. \tag{10}$$

Similarly, it follows that

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + N\eta\boldsymbol{H}_k^{\boldsymbol{v}}\left[\left(\boldsymbol{\rho}_k + \widetilde{\boldsymbol{p}}_k^{\boldsymbol{v}} + \widetilde{\boldsymbol{b}}_k^{\boldsymbol{v}}\right) \odot \boldsymbol{v}_k^{N-1}\right]. \tag{11}$$

C CONSIDERING THE CASE OF ONE-DIMENSION

We consider the one-dimensional iteration sequence of MGD with a positive target and ignore perturbations, i.e., assume that $1/n\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{I}$.

$$\theta_{k+1} = \theta_k\left\{1 - N\eta\left[\tau + \eta\gamma N^2\theta_k^{\frac{2N-2}{N}}\right](\theta_k - \theta^*)\,\theta_k^{\frac{N-2}{N}}\right\}^N. \tag{12}$$

Lemma 2. *Let $\eta > 0$ be the step size and suppose the updates are given by (12). Then the following holds*

- *If $0 < \theta_0 \le \theta^*$, $\gamma - 2\tau \ge \frac{2N}{N-1}$ and $\eta \le \frac{1}{\gamma N(N-1)(\theta^*)^{\frac{2N-2}{N}}}$ then for any $k > 0$ we have $\theta_0 \le \theta_{k-1} \le \theta_k \le \theta^*$;*
- *If $\theta^* \le \theta_0 \le \nu\theta^*$, $\gamma - 2\nu\tau \ge \frac{\nu^2(N+1)}{N}$ and $\eta \le \frac{1}{\nu\gamma N^2(\theta^*)^{(2N-2)/N}}$ then for any $t \ge 0$ and $\nu > 1$ we have $\theta^* \le \theta_k \le \theta_{k-1} \le \nu\theta^*$.*

Proof. Note that if $\theta_0 \le \theta_k \le \theta^*$ then $\theta_k - \theta^* \le 0$. The sequence is monotonically increasing for any momentum $1 > \mu \ge 0$. Thus for the first part it is enough to show that for all $k > 0$, we have

$\theta_k \leq \theta \leq \theta^*$. Assume for a contradiction that exists k such that: $\theta_0 \leq \theta_k \leq \theta^*$ and $\theta_{k+1} \geq \theta^*$. Let $\delta = \frac{\theta_k^{\frac{2N-2}{N}}}{(\theta^*)^{\frac{2N-2}{N}}} \in (0, 1]$ and the condition on η, then the original expression can be simplified to

$$\begin{aligned}
\theta_k &\leq \theta^* \leq \theta_{k+1} \\
&= \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right] (\theta_k - \theta^*)\, \theta_k^{\frac{N-2}{N}}\right\}^N \\
&\leq \theta_k \left\{1 - \frac{1}{\gamma(N-1)} \left[\tau + \frac{N}{N-1} \frac{\theta_k^{\frac{2N-2}{N}}}{(\theta^*)^{\frac{2N-2}{N}}}\right] \cdot \left(\frac{\theta_k^{\frac{2N-2}{N}}}{(\theta^*)^{\frac{2N-2}{N}}} - \frac{\theta_k^{\frac{N-2}{N}}\theta^*}{(\theta^*)^{\frac{2N-2}{N}}}\right)\right\}^N \\
&\leq \theta_k \left\{1 - \frac{1}{\gamma(N-1)} \left[\tau + \frac{N}{N-1}\delta\right] (\delta - 1)\right\}^N.
\end{aligned}$$

Hence we can obtain that

$$\begin{aligned}
&\left(\frac{\theta^*}{\theta_k}\right)^{\frac{1}{N}} = \left(\frac{1}{\delta}\right)^{\frac{1}{2N-2}} \leq 1 - \frac{1}{\gamma(N-1)} \left[\tau + \frac{N}{N-1}\delta\right] (\delta - 1) \\
&\quad \Rightarrow \left(\frac{1}{\delta}\right)^{\frac{1}{2N-2}} + \frac{\tau}{\gamma(N-1)}(\delta - 1) + \frac{N}{\gamma(N-1)^2}\delta(\delta - 1) := f_1(\delta) \leq 1, \quad \delta \in (0, 1].
\end{aligned}$$

Calculate the derivative of $f_1(\delta)$ as

$$\nabla f_1(\delta) = -\frac{1}{2N-2}\left(\frac{1}{\delta}\right)^{\frac{1}{2N-2}+1} + \frac{\tau}{\gamma(N-1)} + \frac{N}{\gamma(N-1)^2}(\delta - 1) + \frac{N}{\gamma(N-1)^2}\delta.$$

and

$$\nabla^2 f_1(\delta) = \frac{1}{2N-2}\left(\frac{2N-1}{2N-2}\right)\left(\frac{1}{\delta}\right)^{\frac{1}{2N-2}+2} + \frac{2N}{\gamma(N-1)^2} \geq 0.$$

According to the condition of the momentum: $\gamma - 2\tau \geq \frac{2N}{N-1}$,we know that

$$f_1'(\delta) \leq f_1'(1) = -\frac{1}{2N-2} + \frac{N}{\gamma(N-1)^2} + \frac{\tau}{\gamma(N-1)} = \frac{(2\tau - \gamma)(N-1) + 2N}{2\gamma(N-1)^2} \leq 0.$$

Therefore $f_1(\delta) \geq f_1(1) = 1$. This gives us the desired contradiction and concludes our proof for the first part.

For the second part, Suppose that $\theta^* \leq \theta_k \leq \nu\theta^*$ and hence we can write $\theta_k = (1 + \rho)\theta^* \Rightarrow \theta_k - \theta^* = \rho\theta^*$ where $\rho \in (0, \nu - 1), \nu > 1$. If we take $\eta \leq \frac{1}{\nu\gamma N^2 (\theta^*)^{\frac{2N-2}{N}}}$, we have

$$\begin{aligned}
\theta_{t+1} =& (1+\rho)\theta^* \left\{1 - N\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right] (\theta_k - \theta^*)\, \theta_k^{\frac{N-2}{N}}\right\}^N \\
=& (1+\rho)\theta^* \left\{1 - N\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right] \rho\theta^* \theta_k^{\frac{N-2}{N}}\right\}^N \\
\geq& (1+\rho)\theta^* \left\{1 - N\eta \left[\tau + \eta\gamma N^2 \nu^{\frac{2N-2}{N}} (\theta^*)^{\frac{2N-2}{N}}\right] \rho\theta^* \nu^{\frac{N-2}{N}} (\theta^*)^{\frac{N-2}{N}}\right\}^N \\
\geq& (1+\rho)\theta^* \left\{1 - N\eta \left[\tau + \eta\gamma N^2 \nu^2 (\theta^*)^{\frac{2N-2}{N}}\right] \rho\nu (\theta^*)^{\frac{2N-2}{N}}\right\}^N \\
\geq& (1+\rho)\theta^* \left\{1 - \frac{1}{N\gamma}\left[\tau + \nu\right]\rho\right\}^N.
\end{aligned}$$

Then, let $f_2(\rho) = 1 - \frac{1}{N\gamma}\left[\tau + \nu\right]\rho$ and $f_3(\rho) = (1 + \rho)\left\{1 - \frac{1}{N\gamma}\left[\tau + \nu\right]\rho\right\}^N$. First, due to $\gamma - \frac{\nu-1}{N}\tau \geq \frac{\nu(\nu-1)}{N}$, we ensure that the value of $f_2(\rho)$ is non-negative for any $\rho \in (0, \nu - 1)$

$$f_2(\rho) = 1 - \frac{1}{N\gamma}\left[\tau + \nu\right]\rho \geq \frac{N\gamma - \left[\tau + \nu\right](\nu - 1)}{N\gamma} \geq 0.$$

On the other hand, we have

$$f_3(\rho) = (1+\rho)\left\{1 - \frac{1}{N\gamma}\left[\tau+\nu\right]\rho\right\}^N := f_2^N(\rho)(1+\rho).$$

Derivative with respect to $f_4(\rho)$, we have

$$\begin{aligned}
f_3'(\rho) =& N f_2(\rho)^{N-1}\left(-\frac{\tau+\nu}{N\gamma}\right)(1+\rho) + f_2(\rho)^N = f_2(\rho)^{N-1}\left(-\frac{\tau+\nu}{\gamma}(1+\rho) + (1-\frac{\tau+\nu}{N\gamma}\rho)\right)\\
\geq& f_2(\rho)^{N-1}\left(1 - \frac{\tau+\nu}{\gamma}\nu - \frac{\tau+\nu}{N\gamma}(\nu-1)\right) = f_2(\rho)^{N-1}\left\{1 - \frac{(\tau+\nu)\big[(N+1)\nu-1\big]}{N\gamma}\right\}\\
\geq& f_2(\rho)^{N-1}\left\{1 - \frac{\tau\big[(N+1)\nu\big] + \nu^2\big[(N+1)\big]}{N\gamma}\right\}.
\end{aligned}$$

Thus, when the momentum satisfies the condition $\gamma - \frac{(N+1)\nu}{N}\tau \geq \frac{\nu^2(N+1)}{N}$, we can conclude that $f_3'(\rho) \geq 0$. Therefore, $f_3(\rho) \geq f_3(0) = 1$. Finally, we consolidate conditions $\gamma - \frac{\nu-1}{N}\tau \geq \frac{\nu(\nu-1)}{N}$ and $\gamma - \frac{(N+1)\nu}{N}\tau \geq \frac{\nu^2(N+1)}{N}$: It is straightforward to show that $\frac{(N+1)\nu}{N} > \frac{\nu-1}{N}$ and $\frac{\nu^2(N+1)}{N} > \frac{\nu(\nu-1)}{N}$,

$$\gamma - \frac{\nu-1}{N}\tau > \gamma - \frac{(N+1)\nu}{N}\tau > \gamma - 2\nu\tau \geq \frac{\nu^2(N+1)}{N} > \frac{\nu(\nu-1)}{N}.$$

Therefore, the momentum parameter must satisfy the condition: $\gamma - 2\nu\tau \geq \frac{\nu^2(N+1)}{N}$. □

Lemma 3. *Consider the same setting as before. Let $\theta^* > 0$ and $|\theta_0 - \theta^*| \leq c\theta^*$, $\frac{1}{2} \leq c < 1$. Then the following holds.*

- *If $\theta_0 \leq \theta^*$ and $\eta \leq \frac{1}{\gamma N(N-1)(\theta^*)^{\frac{2N-2}{N}}}$ then for any*

$$k \geq \frac{1}{N^2\eta\left[\tau + \eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right]\left[(1-c)\theta^*\right]^{\frac{2N-2}{N}}},$$

  we have $0 \leq \theta^* - \theta_k \leq c|\theta_0 - \theta^*|$;

- *If $\theta^* \leq \theta_0 \leq (c+1)\theta^*$ and $\eta \leq \frac{1}{(c+1)N^2\gamma(\theta^*)^{(2N-2)/N}}$ (let $\nu = (1+c)$ in Lemma 2), then for any*

$$k \geq \frac{1}{N^2\eta(\theta^*)^{\frac{2N-2}{N}}\left(\tau + \eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right)},$$

  we have

$$0 \leq \theta_k - \theta^* \leq c|\theta_0 - \theta^*|.$$

Proof. Setting $|\theta_0 - \theta^*| = \rho\theta^*$, $\rho \in [0, c]$. For the first part, we have $\theta_0 = (1-\rho)\theta^*$. We want to know how many steps k are needed to halve the error, i.e.,

$$\theta_k\left\{1 - N\eta\left[\tau + \eta\gamma N^2\theta_k^{\frac{2N-2}{N}}\right](\theta_k - \theta^*)\,\theta_k^{\frac{N-2}{N}}\right\}^N \geq \theta^* - c|\theta_0 - \theta^*| = (1-c\rho)\theta^*.$$

The given theorem implies that $\theta_0 \leq \theta_k \leq \theta^*$, then $\theta_k \geq \theta_0 = (1-\rho)\theta^*$

$$\begin{aligned}
&\theta_k\left\{1 + N\eta\left[\tau + \eta\gamma N^2\theta_k^{\frac{2N-2}{N}}\right](\theta^* - \theta_k)\,\theta_k^{\frac{N-2}{N}}\right\}^N\\
\geq&\theta_k\left\{1 + N\eta\left[\tau + \eta\gamma N^2(1-\rho)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right]c\rho\theta^*\left[(1-\rho)\theta^*\right]^{\frac{N-2}{N}}\right\}^N.
\end{aligned}$$

Therefore, we have

$$\theta_k \left\{1 + N\eta \left[\tau + \eta\gamma N^2(1-\rho)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right] c\rho\theta^*[(1-\rho)\theta^*]^{\frac{N-2}{N}}\right\}^N \geq (1-c\rho)\theta^*$$
$$\theta_0 \left\{1 + N\eta \left[\tau + \eta\gamma N^2(1-\rho)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right] c\rho\theta^*[(1-\rho)\theta^*]^{\frac{N-2}{N}}\right\}^{Nk} \geq (1-c\rho)\theta^*$$
$$1 + N^2 k\eta \left[\tau + \eta\gamma N^2(1-\rho)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right] c\rho\theta^*[(1-\rho)\theta^*]^{\frac{N-2}{N}} \geq \frac{(1-c\rho)}{1-\rho}.$$

Then, we have

$$k \geq \frac{\left(\frac{1-c\rho}{1-\rho} - 1\right)\frac{1}{\rho}}{N^2\eta \left[\tau + \eta\gamma N(1-\rho)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right] c\theta^*[(1-\rho)\theta^*]^{\frac{N-2}{N}}}$$
$$= \frac{\frac{1-c}{1-\rho}}{N^2\eta \left[\tau + \eta\gamma N(1-\rho)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right] c\theta^*[(1-\rho)\theta^*]^{\frac{N-2}{N}}} := f_4(\rho).$$

It is easy to know that $f_4(\rho)$ is a positive increment function. Thus we can obtain

$$k \geq \frac{1}{N^2\eta \left[\tau + \eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right] [(1-c)\theta^*]^{\frac{2N-2}{N}}}$$

the error is halved.

To deal with the second part, we write $\theta_0 = \theta^*(1+\rho)$. We will use a similar approach as the one in the first part. If for some θ_k we have $\theta_k \leq (1+c\rho)\theta^*$ we would be done. If $\theta_k > \theta^*(1+c\rho)$ we have

$$\theta_{k+1} \leq \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right] \theta^* c\rho(\theta^*)^{\frac{N-2}{N}}\right\}^N.$$

Therefore

$$\theta_0 \left\{1 - N\eta \left[\tau + \eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right] \theta^* c\rho(\theta^*)^{\frac{N-2}{N}}\right\}^{Nk} \leq (1+c\rho)\theta^*.$$

We have

$$k \geq \frac{1}{N} \frac{\log \frac{(1+c\rho)\theta^*}{\theta_0}}{\log \left\{1 - N\eta \left[\tau + \eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right] \theta^* c\rho(\theta^*)^{\frac{N-2}{N}}\right\}}$$
$$\geq \frac{1}{N} \frac{\log \frac{(1+c\rho)}{1+\rho}}{\log \left\{1 - N\eta \left[\tau + \eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right] \theta^* c\rho(\theta^*)^{\frac{N-2}{N}}\right\}}.$$

and

$$k \geq \frac{1}{N^2\eta(\theta^*)^{\frac{2N-2}{N}} \left(\tau + \eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right)}$$
$$\geq \frac{\frac{(1-c)\rho}{1+c\rho}}{N^2\eta \left[\tau + \eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right] \theta^* c\rho(\theta^*)^{\frac{N-2}{N}}}.$$

where $\frac{1-c}{(1+c\rho)c} \leq \frac{1-c}{c} \leq 1$. □

Lemma 4. *Consider the same setting as before. If* $0 < \theta_0 \leq c\theta^*$, $\eta \leq \frac{\theta_0}{\frac{1}{2}N(2N-4)\gamma(\theta^*)^{\frac{3N-2}{N}}}$ *and* $\gamma - 2\tau \geq \frac{4N}{2N-4}$, *for any* $k \geq \frac{1+c}{N(N-2)(1-c)\eta\theta^*\tau\theta_0^{\frac{N-2}{N}}}$, *we will have* $c\theta^* \leq \theta_k \leq \theta^*$ *where* $c \in [\frac{1}{2}, \frac{2N-4}{2N-3}]$.

Proof. We need to find a lower-bound on time K which ensures that $\theta_K \geq c\theta^*$. At any time k, we have

$$\begin{aligned}\theta_{k+1} &= \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right](\theta_k - \theta^*)\,\theta_k^{\frac{N-2}{N}}\right\}^N \\ &\geq \theta_k \left\{1 - N^2\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right](\theta_k - \theta^*)\,\theta_k^{\frac{N-2}{N}}\right\}.\end{aligned}$$

We have

$$\sum_{k=0}^{K-1} \frac{\theta_{k+1} - \theta_k}{-N^2\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right](\theta_k - \theta^*)\,\theta_k^{\frac{2N-2}{N}}} \geq K.$$

Therefore, for k that is larger than the left hand side, we have $\theta_k \geq c\theta^*$.

$$\begin{aligned}&\sum_{k=1}^{K-1} \frac{\theta_{k+1} - \theta_k}{N^2\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right](\theta^* - \theta_k)\,\theta_k^{\frac{2N-2}{N}}} \\ \leq &\frac{1}{(1-c)\theta^* N^2\eta} \sum_{k=1}^{K-1} \frac{\theta_{k+1} - \theta_k}{\left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right]\theta_k^{\frac{2N-2}{N}}} \\ \leq &\frac{1}{(1-c)\theta^* N^2\eta\tau} \sum_{k=1}^{K-1} \int_{\theta_k}^{\theta_{k+1}} \frac{1}{\theta^{\frac{2N-2}{N}}} d\theta + \frac{1}{(1-c)\theta^* N^2\eta\tau} \sum_{k=1}^{K-1} \int_{\theta_k}^{\theta_{k+1}} \left(\frac{1}{\theta_k^{\frac{2N-2}{N}}} - \frac{1}{\theta^{\frac{2N-2}{N}}}\right) d\theta \\ \leq &\frac{1}{(1-c)\theta^* N^2\eta\tau} \left(\sum_{k=1}^{K-1} \int_{\theta_k}^{\theta_{k+1}} \frac{1}{\theta^{\frac{2N-2}{N}}} d\theta + \max_{0\leq k\leq K-1} \left(\frac{1}{\theta_k^{\frac{2N-2}{N}}} - \frac{1}{\theta_{k+1}^{\frac{2N-2}{N}}}\right)(\theta_K - \theta_0)\right) \\ \leq &\frac{1}{(1-c)\theta^* N^2\eta\tau} \int_{\theta_0}^{c\theta^*} \frac{1}{\theta^{\frac{2N-2}{N}}} d\theta + \frac{1}{(1-c)\theta^* N^2\eta\tau} \max_{0\leq k\leq K-1} \underbrace{\left(\frac{1}{\theta_k^{\frac{2N-2}{N}}} - \frac{1}{\theta_{k+1}^{\frac{2N-2}{N}}}\right)}_{J_1}(c\theta^* - \theta_0) \\ &+ \frac{1}{(1-c)\theta^* N^2\eta\tau} \frac{1}{(c\theta^*)^{\frac{2N-2}{N}}}(\theta_K - c\theta^*) := J. \end{aligned} \tag{13}$$

For the second term J_1 in (13), if we further require the step size satisfies $\eta \leq \frac{1}{\frac{1}{2}N(2N-3)\gamma(\theta^*)^{\frac{2N-2}{N}}}$ and $\gamma - 2\tau \geq \frac{4N}{2N-3}$, we have

$$\begin{aligned}\theta_{k+1}^{\frac{2N-2}{N}} =&\theta_k^{\frac{2N-2}{N}} \left\{1 - N\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right](\theta_k - \theta^*)\,\theta_k^{\frac{N-2}{N}}\right\}^{(2N-2)} \\ \leq&\theta_k^{\frac{2N-2}{N}} \left\{1 + N\eta \left[\tau + \eta\gamma N^2 (\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}\right\}^{(2N-2)} \\ \leq&\theta_k^{\frac{2N-2}{N}} \frac{1 + N\eta \left[\tau + \eta\gamma N^2 (\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}}{1 - (2N-3)N\eta \left[\tau + \eta\gamma N^2 (\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}}.\end{aligned}$$

Then, using the condition of $\eta \leq \frac{\theta_0}{\frac{1}{2}N(2N-4)\gamma(\theta^*)^{\frac{3N-2}{N}}}$, $\gamma - 2\tau \geq \frac{4N}{2N-4}$, we have

$$\begin{aligned}
&\frac{1}{\theta_k^{\frac{2N-2}{N}}} - \frac{1}{\theta_{k+1}^{\frac{2N-2}{N}}} \\
\leq &\frac{1}{\theta_k^{(2N-2)/N}} \left(1 - \frac{1-(2N-3)N\eta\left[\tau+\eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}}{1+N\eta\left[\tau+\eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}}\right) \\
= &\frac{1}{\theta_k^{\frac{2N-2}{N}}} \frac{(2N-4)N\eta\left[\tau+\eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}}{1+N\eta\left[\tau+\eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}} \\
\leq &\frac{(2N-4)N\eta\left[\tau+\eta\gamma N^2(\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}}{\theta_0^{\frac{2N-2}{N}}} \\
\leq &\frac{\frac{2}{\gamma}[\tau+\frac{2N\theta_0}{(2N-4)\theta^*}]}{\theta^*\theta_0^{\frac{N-2}{N}}} \leq \frac{\frac{2}{\gamma}[\tau+\frac{2cN}{(2N-4)}]}{\theta^*\theta_0^{\frac{N-2}{N}}} \leq \frac{1}{\theta^*\theta_0^{\frac{N-2}{N}}}.
\end{aligned} \tag{14}$$

Substituting equation (14) into equation (13), we obtain

$$\begin{aligned}
J \leq &\frac{1}{(1-c)N^2\eta\theta^*\tau}\left\{\int_{\theta_0}^{c\theta^*}\frac{1}{\theta^{\frac{2N-2}{N}}}d\theta + \frac{1}{\theta_0^{\frac{N-2}{N}}\theta^*}(c\theta^*-\theta_0) + \frac{1}{(c\theta^*)^{\frac{2N-2}{N}}}(\theta_K - c\theta^*)\right\} \\
= &\frac{1}{(1-c)N^2\eta\theta^*\tau}\left\{\frac{N}{N-2}\left(\frac{1}{\theta_0^{\frac{N-2}{N}}} - \frac{(1/c)^{\frac{N-2}{N}}}{(\theta^*)^{\frac{N-2}{N}}}\right) + \frac{c}{\theta_0^{\frac{N-2}{N}}} + \frac{c^{\frac{1-c}{\frac{2N-2}{N}}}}{(\theta^*)^{\frac{N-2}{N}}}\right\} \\
\leq &\frac{1+c}{N(N-2)(1-c)\eta\theta^*\tau\theta_0^{\frac{N-2}{N}}}.
\end{aligned}$$

Finally, we consolidate the conditions for η.

$$\begin{aligned}
&\theta_0 N(2N-3)\gamma(\theta^*)^{\frac{2N-2}{N}} - N(2N-4)\gamma(\theta^*)^{\frac{3N-2}{N}} \\
= &[\theta_0(2N-3)-(2N-4)\theta^*]N\gamma(\theta^*)^{\frac{2N-2}{N}} = [(\theta_0-\theta^*)(2N-3)+\theta^*]N\gamma(\theta^*)^{\frac{2N-2}{N}} \\
\leq &[(c-1)\theta^*(2N-3)+\theta^*]N\gamma(\theta^*)^{\frac{2N-2}{N}} = [1-(1-c)(2N-3)]N\gamma(\theta^*)^{\frac{3N-2}{N}} \leq 0.
\end{aligned}$$

□

Lemma 5. *Consider the same setting as before. Fix any $\epsilon > 0$.*

1. *If $\epsilon < |\theta^* - \theta_0| \leq c\theta^*$ and $\eta \leq \frac{1}{(c+1)N^2\gamma(\theta^*)^{(2N-2)/N}}$ then for any $k \geq \frac{\log_{1/c} e}{N^2\eta\left[\tau+\eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right][(1-c)\theta^*]^{\frac{2N-2}{N}}}\log\frac{|\theta^*-\theta_0|}{\epsilon}$ we have*

$$|\theta^* - \theta_k| \leq \epsilon.$$

2. *If $0 < \theta_0 \leq c\theta^*$ and $\eta \leq \frac{\theta_0}{N^2\gamma(\theta^*)^{(3N-2)/N}}$ then for any*

$$\begin{aligned}
k \geq &\frac{\log_{1/c} e}{N^2\eta\left[\tau+\eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right][(1-c)\theta^*]^{\frac{2N-2}{N}}}\log\frac{|\theta^*-\theta_0|}{\epsilon} \\
&+ \frac{1+c}{N(N-2)(1-c)\eta\theta^*\tau\theta_0^{(N-2)/N}},
\end{aligned}$$

we have

$$\theta^* - \epsilon \leq \theta_k \leq \theta^*.$$

Proof. By Lemma 3, after a sufficient number of iterations, it holds that $|\theta_k - \theta^*| \leq c|\theta_0 - \theta^*|$. Therefore, after m such iterations, we have: $|\theta_k - \theta^*| \leq c^m|\theta_0 - \theta^*|$. Hence, setting $c^m|\theta_0 - \theta^*| = \epsilon$, we derive that $m = \log_{1/c}\left(\frac{|\theta_0-\theta^*|}{\epsilon}\right)$. This means that we only need to apply Lemma 3 m times, where $m = \log_{1/c}\left(\frac{|\theta_0-\theta^*|}{\epsilon}\right)$.

$$\begin{aligned}&\frac{1}{N^2\eta\left[\tau+\eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right]\left[(1-c)\theta^*\right]^{\frac{2N-2}{N}}}\log_{1/c}\left(\frac{|\theta_0-\theta^*|}{\epsilon}\right)\\ =&\frac{\log_{1/c}e}{N^2\eta\left[\tau+\eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right]\left[(1-c)\theta^*\right]^{\frac{2N-2}{N}}}\log\frac{|\theta^*-\theta_0|}{\epsilon}\end{aligned}$$

For the second part, the process is divided into two stages. When $0 < \theta_0 \leq \frac{1}{2}\theta^*$, according to Lemma 4, after a sufficient number of iterations $k \geq \frac{1+c}{N(N-2)(1-c)\eta\theta^*\tau\theta_0^{\frac{N-2}{N}}}$, it holds that $c\theta^* \leq \theta_k \leq \theta^*$. Based on the first part, after further $k \geq \frac{\log_{1/c}e}{N^2\eta\left[\tau+\eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right]\left[(1-c)\theta^*\right]^{\frac{2N-2}{N}}}\log\frac{|\theta^*-\theta_0|}{\epsilon}$ iterations, we can achieve$|\theta^* - \theta_k| \leq \epsilon$. Thus, it is enough to choose k larger than or equal to

$$\begin{aligned}&\frac{\log_{1/c}e}{N^2\eta\left[\tau+\eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^*)^{\frac{2N-2}{N}}\right]\left[(1-c)\theta^*\right]^{\frac{2N-2}{N}}}\log\frac{|\theta^*-\theta_0|}{\epsilon}\\ &+\frac{1+c}{N(N-2)(1-c)\eta\theta^*\tau\theta_0^{\frac{N-2}{N}}}.\end{aligned}$$

□

D DEALING WITH BOUNDED ERRORS $\boldsymbol{b}_k$

In this section, we consider the following iterative form with bounded errors. We revisit the iteration with a positive target.

$$\begin{aligned}\boldsymbol{u}_{k+1} =&\boldsymbol{u}_k - \eta\left\{\tau\boldsymbol{I}+\eta\gamma\left[\mathrm{Diag}\left(\boldsymbol{u}_k^{N-1}\right)\frac{N^2}{n}\boldsymbol{X}^T\boldsymbol{X}\mathrm{Diag}\left(\boldsymbol{u}_k^{N-1}\right)\right]\right\}\\ &\cdot\left((\boldsymbol{p}_k+\boldsymbol{b}_k+\boldsymbol{s}_k-\boldsymbol{\theta}^*)\odot\boldsymbol{\theta}_k^{N-1}\right).\end{aligned}$$

$$\theta_{k+1}^+ =\theta_k^+\left\{1-N\eta\left[\tau+\eta\gamma N^2(\theta_k^+)^{\frac{2N-2}{N}}\right]\left(\theta_k^+-\theta^*+b_k\right)(\theta_k^+)^{\frac{N-2}{N}}\right\}^N, \tag{15}$$

where for any $k \geq 0$ we have $\|\boldsymbol{b}_k\| \leq B$. For a given B, we consider two extreme scenarios,

$$\theta_{k+1}^+ =\theta_k^+\left\{1-N\eta\left[\tau+\eta\gamma N^2(\theta_k^+)^{\frac{2N-2}{N}}\right]\left(\theta_k^+-(\theta^*+B)\right)(\theta_k^+)^{\frac{N-2}{N}}\right\}^N, \tag{16}$$

$$\theta_{k+1}^- =\theta_k^-\left\{1-N\eta\left[\tau+\eta\gamma N^2(\theta_k^-)^{\frac{2N-2}{N}}\right]\left(\theta_k^--(\theta^*-B)\right)(\theta_k^-)^{\frac{N-2}{N}}\right\}^N. \tag{17}$$

It is easy to see that when $B \geq \theta^*$, the iteration cannot converge to θ^*. Therefore, we restrict $B \leq \beta\theta^*$ with $\beta \in [0, 1)$.

Lemma 6. *Consider the sequences $(\theta_k^+)_{k\geq 0}, (\theta_k)_{k\geq 0}$ and $(\theta_k^-)_{k\geq 0}$ as defined above with*

$$0 < \theta_0^- = \theta_0^+ = \theta_0 \leq \theta^* + B$$

If $\eta \leq \frac{1}{2\gamma N^2(\theta^*)^{\frac{2N-2}{N}}}$ and $\gamma - \frac{N+1}{N}\tau \geq \frac{N+1}{N}$ then for all $k \geq 0$

$$0 \leq \theta_k^- \leq \theta_k \leq \theta_k^+ \leq \theta^* + B.$$

Proof. We will prove the claim by induction. The claim holds trivially for $k = 0$. If $\theta_k^+ \geq \theta_k$, we rewrite $\theta_k^+ - (\theta^* + B) = (\theta_k - \theta^* + b_k) + (\theta_k^+ - \theta_k - B - b_k) := p_k + q_k$ and $y_k = \theta_k^+ - \theta_k$. If we further require $\eta \leq \frac{1}{\gamma N^2 (\theta^* + B)^{\frac{2N-2}{N}}}$,

$$
\begin{aligned}
\theta_{k+1}^+ =& \theta_k^+ \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k^+)^{\frac{2N-2}{N}}\right] \left[\theta_k^+ - (\theta^* + B)\right] (\theta_k^+)^{\frac{N-2}{N}}\right\}^N \\
\geq& \theta_k^+ \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[\theta_k^+ - (\theta^* + B)\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[\theta_k - \theta^* + b_k\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& + \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[\theta_k - \theta^* + b_k\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
\geq& \theta_k^+ \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[p_k + y_k\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& + \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N.
\end{aligned}
$$

Due to the condition of η and $B \leq \beta\theta^*$, we know that

$$
\begin{aligned}
& 1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}} \\
\geq& 1 - \frac{N \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}}{\gamma N^2 (\theta^* + B)^{\frac{2N-2}{N}}} \\
\geq& 1 - \frac{(\tau+1) p_k}{\gamma N (\theta^* + B)} \geq 1 - \frac{2(\tau+1)B}{\gamma N (1/\beta + 1) B} \\
\geq& 1 - \frac{2(\tau+1)}{\gamma N (1/\beta + 1)} \geq 1 - \frac{\tau+1}{N\gamma}.
\end{aligned}
$$

Due to $\eta \leq \frac{1}{2\gamma N^2 (\theta^*)^{\frac{2N-2}{N}}}$ and $\gamma - \frac{N+1}{N}\tau \geq \frac{N+1}{N}$, we have

$$
\begin{aligned}
& \theta_k^+ \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}} - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] y_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
\geq& \theta_k^+ \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left\{1 - \frac{N^2\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] y_k (\theta_k)^{\frac{N-2}{N}}}{1 - \frac{\tau+1}{N\gamma}}\right\} \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
\geq& \theta_k^+ \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left\{1 - \frac{\frac{\tau+1}{2\gamma} y_k}{\left(1 - \frac{\tau+1}{N\gamma}\right)\theta^*}\right\} \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
=& \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left(\theta_k^+ - \frac{\frac{\tau+1}{2\gamma} y_k \theta^+}{\left(1 - \frac{\tau+1}{N\gamma}\right)\theta^*} - \theta_k\right) \\
=& y_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left(1 - \frac{N(\tau+1)(y_k + \theta_k)}{2(N\gamma - \tau - 1)\theta^*}\right) \\
\geq& y_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left(1 - \frac{N(\tau+1)}{N\gamma - \tau - 1}\right) \geq 0.
\end{aligned}
$$

Finally, it can be concluded that $\theta_{k+1}^{+} \geq \theta_{k+1}$.

For the second part, the claim holds trivially for $t = 0$. If $\theta_k^- \leq \theta_k$, we rewrite $\theta_k^- - (\theta^* - B) = (\theta_k - \theta^* + b_k) + (\theta_k^- - \theta_k + B - b_k) := p_k + n_k$ and $z_k = \theta_k^- - \theta_k \leq 0$. If we further require $\eta \leq \frac{1}{\gamma N^2 (\theta^* + B)^{\frac{2N-2}{N}}}$

$$
\begin{aligned}
\theta_{k+1}^- =& \theta_k^- \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k^-)^{\frac{2N-2}{N}}\right] \left[\theta_k^- - (\theta^* - B)\right] (\theta_k^-)^{\frac{N-2}{N}}\right\}^N \\
\leq& \theta_k^- \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[\theta_k^- - (\theta^* - B)\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[\theta_k - \theta^* + b_k\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& + \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[\theta_k - \theta^* + b_k\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
\leq& \theta_k^- \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[p_k + z_k\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& + \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N .
\end{aligned}
$$

It is the same as θ_k^+, we have

$$
\begin{aligned}
& \theta_k^- \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] \left[p_k + z_k\right] (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
\leq& \theta_k^- \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left\{1 - \frac{N^2\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] z_k (\theta_k)^{\frac{N-2}{N}}}{1 - \frac{\tau+1}{N\gamma}}\right\} \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
\leq& \theta_k^- \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left\{1 - \frac{\frac{\tau+1}{2\gamma} z_k}{\left(1 - \frac{\tau+1}{N\gamma}\right)\theta^*}\right\} \\
& - \theta_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \\
=& \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left(\theta_k^- - \frac{\frac{\tau+1}{2\gamma} z_k \theta^-}{\left(1 - \frac{\tau+1}{N\gamma}\right)\theta^*} - \theta_k\right) \\
=& z_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left(1 - \frac{N(\tau+1)(z_k + \theta_k)}{2(N\gamma - \tau - 1)\theta^*}\right) \\
\leq& z_k \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_k)^{\frac{2N-2}{N}}\right] p_k (\theta_k)^{\frac{N-2}{N}}\right\}^N \left(1 - \frac{N(\tau+1)}{N\gamma - \tau - 1}\right) \leq 0.
\end{aligned}
$$

The proof is completed. □

Lemma 7. *Consider the previous setting with $B \leq \beta\theta^*$, $\eta \leq \frac{1}{2N^2\gamma(\theta^*)^{(2N-2)/N}}$. Then the following holds*

1. *If $|\theta_k - \theta^*| > B$ then $|\theta_{k+1} - \theta^*| < |\theta_k - \theta^*|$.*
2. *If $|\theta_k - \theta^*| \leq B$ then $|\theta_{k+1} - \theta^*| \leq B$.*

Proof. When $\theta_k \geq \theta^* + B$, by Lemma 2, we know

$$\theta^* \leq \theta_{k+1} \leq \theta_k \leq \nu\theta^* \Rightarrow 0 \leq \theta_{k+1} - \theta^* \leq \theta_k - \theta^* \leq (\nu - 1)\theta^* \Rightarrow |\theta_{k+1} - \theta^*| \leq |\theta_k - \theta^*|$$

When $\theta_k \leq \theta^* - B$, by Lemma 2, we know

$$0 \leq \theta_k \leq \theta_{k+1} \leq \theta^* \Rightarrow -\theta^* \leq \theta_k - \theta^* \leq \theta_{k+1} - \theta^* \leq 0 \Rightarrow |\theta_{k+1} - \theta^*| \leq |\theta_k - \theta^*|$$

When $\theta^* \leq \theta_k \leq B + \theta^*$. By Lemma 6, we have $\theta^* \leq \theta_{k+1} \leq \theta_k \leq B + \theta^*$.
When $\theta^* - B \leq \theta_k \leq \theta^*$. By Lemma 6, we have $\theta^* - B \leq \theta_k \leq \theta_{k+1} \leq \theta^*$. □

Lemma 8. *Consider the setting as before and $c = \frac{1}{2}, \beta = \frac{1}{5}$ in Lemma 3. Then the following holds:*

1. *If $\frac{1}{2}(\theta^* - B) \leq \theta_0 \leq \theta^* - 5B$ then for any $k \geq \frac{25}{2N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right][\theta^*]^{\frac{2N-2}{N}}}$ we have*

$$|\theta^* - \theta_k| \leq \frac{1}{2}|\theta_0 - \theta^*|.$$

2. *If $\theta^* + 4B < \theta_0 < \frac{3}{2}\theta^*$ then for any $k \geq \frac{16}{N^2\eta\left[\tau + \frac{1}{4}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right][\theta^*]^{\frac{2N-2}{N}}}$ we have*

$$|\theta^* - \theta_k| \leq \frac{1}{2}|\theta_0 - \theta^*|.$$

Proof. We use Lemma 3 on θ_k^- twice and take $c = \frac{1}{2}$,

$$\begin{aligned}
&\frac{2}{N^2\eta\left[\tau + \eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^* - B)^{\frac{2N-2}{N}}\right][(1-c)(\theta^* - B)]^{\frac{2N-2}{N}}} \\
\leq &\frac{25}{2N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right][\theta^*]^{\frac{2N-2}{N}}} \leq k
\end{aligned}$$

We have

$$\theta^* - B - \theta_k^- \leq \frac{1}{4}|\theta_0 - (\theta^* - B)| \leq \frac{1}{4}|\theta_0 - \theta^*| + \frac{1}{4}B$$

Therefore, due to $\theta_0 \leq \theta^* - 5B$

$$\begin{cases}
\theta^* - \theta_k \leq \theta^* - \theta_k^- \leq \frac{1}{4}|\theta_0 - \theta^*| + \frac{5}{4}B \leq \frac{1}{2}|\theta_0 - \theta^*| \quad \text{when} \quad \theta^* \geq \theta_k \\
\theta_k - \theta^* \leq B \leq \frac{1}{5}|\theta_0 - \theta^*| \leq \frac{1}{2}|\theta_0 - \theta^*| \quad \text{when} \quad \theta^* \leq \theta_k
\end{cases}$$

The second part follows a very similar proof.

$$\begin{aligned}
&2\frac{2}{N^2\eta\left[\tau + \eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^* + B)^{\frac{2N-2}{N}}\right][(1-c)(\theta^* + B)]^{\frac{2N-2}{N}}} \\
\leq &\frac{16}{N^2\eta\left[\tau + \frac{1}{4}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right][\theta^*]^{\frac{2N-2}{N}}} \leq k
\end{aligned}$$

Due to $\theta^* + 4B < \theta_0$, we have

$$\begin{aligned}
&0 \leq \theta_k^+ - (\theta^* + B) \leq \frac{1}{4}|\theta_0 - (\theta^* + B)| = \frac{1}{4}(\theta_0 - \theta^* - B) = \frac{1}{4}|\theta_0 - \theta^*| - \frac{1}{4}B \\
\Rightarrow &\theta_k^+ - \theta^* \leq \frac{1}{4}|\theta_0 - \theta^*| + \frac{3}{4}B.
\end{aligned}$$

If $\theta_k \geq \theta_k^*$, thus

$$0 \leq \theta_k - \theta^* \leq \theta_k^+ - \theta^* \leq \frac{1}{4}|\theta_0 - \theta^*| + \frac{3}{4}B.$$

Since $\theta_0 \geq \theta^* + 4B$, we get $B \leq \frac{1}{4}|\theta_0 - \theta^*|$, and therefore

$$0 \leq \theta_k - \theta^* \leq \theta_k^+ - \theta^* \leq \frac{1}{4}|\theta_0 - \theta^*| + \frac{3}{16}|\theta_0 - \theta^*| \leq \frac{1}{2}|\theta_0 - \theta^*|.$$

If $\theta_k < \theta^*$, Lemma 6 still ensures $\theta_k^- \leq \theta_k \leq \theta_k^+$ and in particular

$$\theta_k \geq \theta^* - B.$$

Therefore,

$$|\theta_k - \theta^*| = \theta^* - \theta_k \leq B \leq \frac{1}{4}|\theta_0 - \theta^*| \leq \frac{1}{2}|\theta_0 - \theta^*|.$$

□

Lemma 9. *Consider the same setting as before. Fix any $\epsilon > 0$ and $c = \frac{1}{2}$, then the following holds*

1. *If $B + \epsilon < |\theta^* - \theta_0| \leq \frac{1}{5}\theta^*$ then for any $k \geq \frac{4\log_2 e}{N^2\eta\left[\tau + \frac{1}{4}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}} \log\frac{|\theta^* - \theta_0|}{\epsilon}$ iterations we have $|\theta^* - \theta_k| \leq B + \epsilon$.*
2. *If $0 < \theta_0 \leq \theta^* - B - \epsilon$ then for any*

$$k\frac{25\log_2 e}{4N^2\eta\left[\tau + \frac{25}{4}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right][\theta^*]^{\frac{2N-2}{N}}} \log\frac{|\theta^* - \theta_0|}{\epsilon} + \frac{15}{4N(N-2)\eta\theta^*\tau\theta_0^{(N-2)/N}}$$

we have $\theta^* - B - \epsilon \leq \theta_k \leq \theta^* + B$.

Proof. For the first part, the lower bound is obtained from Lemma 6 and Lemma 7. If $\theta_0 > \theta^* + B$. From the first part of Lemma 5, we see that

$$\begin{aligned}
&\frac{4\log_2 e}{N^2\eta\left[\tau + \frac{1}{4}\eta\gamma N(\theta^* + B)^{\frac{2N-2}{N}}\right](\theta^* + B)^{\frac{2N-2}{N}}} \log\frac{|\theta^* + B - \theta_0|}{\epsilon} \\
\leq &\frac{4\log_2 e}{N^2\eta\left[\tau + \frac{1}{4}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right](\theta^*)^{\frac{2N-2}{N}}} \log\frac{|\theta^* - \theta_0|}{\epsilon}
\end{aligned}$$

iterations are enough. For the second part, the upper bound is obtained from Lemma 6. For the lower bound, we simply apply the second part of Lemma 5 to the sequence $(\theta_k^-)_{t\geq 0}$ to get

$$\begin{aligned}
&\frac{\log_2 e \log\frac{|(\theta^* - B) - \theta_0|}{\epsilon}}{N^2\eta\left[\tau + \eta\gamma N(1-c)^{\frac{2N-2}{N}}(\theta^* - B)^{\frac{2N-2}{N}}\right][(1-c)(\theta^* - B)]^{\frac{2N-2}{N}}} \\
&+ \frac{3}{2N(N-2)(1-c)\eta(\theta^* - B)\tau\theta_0^{(N-2)/N}} \\
\leq &\frac{25\log_2 e}{4N^2\eta\left[\tau + \frac{25}{4}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right][\theta^*]^{\frac{2N-2}{N}}} \log\frac{|\theta^* - \theta_0|}{\epsilon} + \frac{15}{4N(N-2)\eta\theta^*\tau\theta_0^{(N-2)/N}}.
\end{aligned}$$

□

Lemma 10. *Suppose the error sequences $(\boldsymbol{b}_k)_{k\geq 0}$ and $(\boldsymbol{p}_k)_{k\geq 0}$ satisfy the following for any $k \geq 0$:*

$$\|\boldsymbol{b}_k \odot \mathbf{1}_S\| \leq B,$$
$$\|\boldsymbol{p}_k\|_\infty \leq \frac{1}{20}\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty.$$

Suppose that

$$20B < \|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty \leq \frac{1}{5}\theta^*_{\min}.$$

Then for $\eta \leq \frac{1}{2N^2\gamma(\theta^*_{\max})^{(2N-2)/N}}$ and any $k \geq \frac{25}{2N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N(\theta^*_{\max})^{\frac{2N-2}{N}}\right][\theta^*_{\max}]^{\frac{2N-2}{N}}}$ we have

$$\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty \leq \frac{1}{2}\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty.$$

Proof. It is easy to see that $\|\boldsymbol{b}_0\|_\infty + \|\boldsymbol{p}_k\|_\infty \leq \frac{1}{10}\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty$. For any i satisfying $|s_{0,i} - \theta_i^*| \leq \frac{1}{2}\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty$, Lemma 7 implies that $|s_{k,i} - \theta_i^*| \leq \frac{1}{2}\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty$ for all $k \geq 0$. For any i satisfying $|s_{0,i} - \theta_i^*| > \frac{1}{2}\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty$, Lemma 8 implies that when $k \geq \frac{25}{2N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N(\theta^*_{\max})^{\frac{2N-2}{N}}\right][\theta^*_{\max}]^{\frac{2N-2}{N}}}$, we have $|s_{0,i} - \theta_i^*| \leq \frac{1}{2}\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty$. □

E $\boldsymbol{\theta}^*$ HAS NEGATIVE COMPONENTS

Lemma 11. *Let $\theta_k = u_k^N - v_k^N$ and $|\theta^*| > 0$. The update rules for $(u_k)_{k\geq 0}$ and $(v_k)_{k\geq 0}$ are as follows,*

$$0 < u_0 = \alpha, u_{k+1} = u_k\left\{1 - N\eta\left[\tau + \eta\gamma N^2 u_k^{2N-2}\right](\theta_k - \theta^* + b_k)\, u_k^{N-2}\right\},$$
$$0 < v_0 = \alpha, v_{k+1} = v_k\left\{1 + N\eta\left[\tau + \eta\gamma N^2 v_k^{2N-2}\right](\theta_k - \theta^* + b_k)\, v_k^{N-2}\right\},$$

where $\alpha \leq \left(2 - 2\frac{N-2}{N}\right)^{\frac{1}{N-2}}|\theta^*|^{1/N}$. There exists $B > 0$ such that $|b_k| \leq B$ and $\eta \leq \frac{1}{N^2\gamma(\theta^*)^{\frac{2N-2}{N}}}$. Then the following holds ($\nu > 1$):

- *If $\theta^* > 0$ and $\nu\theta^* > u_k^N \geq \theta^*, c\theta^* \geq v_k^N$, then $v_k^N \leq \frac{1}{2}\alpha^N$.*
- *If $\theta^* < 0$ and $\nu\theta^* > v_k^N \geq |\theta^*|, c\theta^* \geq u_k^N$, then $u_k^N \leq \frac{1}{2}\alpha^N$.*

Proof. For the first part, from the updating equation, we obtain that

$$\frac{u_{k+1} - u_k}{u_k^{N-1}} = -N\eta\left[\tau + \eta\gamma N^2 u_k^{2N-2}\right](\theta_k - \theta^* + b_k).$$

And then,

$$\begin{aligned}
&\sum_{i=0}^{k} -N\eta\,(\theta_k - \theta^* + b_k) = \sum_{i=0}^{k} \frac{u_{i+1} - u_i}{\left[\tau + \eta\gamma N^2 u_i^{2N-2}\right] u_i^{N-1}} \\
\geq &\sum_{i=0}^{t}\int_{u_i}^{u_{i+1}} \frac{1}{\left[\tau + \eta\gamma N^2 u^{2N-2}\right] u^{N-1}}\mathrm{d}u = \int_{u_0}^{u_k} \frac{1}{\left[\tau + \eta\gamma N^2 u^{2N-2}\right] u^{N-1}}\mathrm{d}u \\
\geq &\int_{u_0}^{u_k} \frac{1}{\left[\tau + \nu\right] u^{N-1}} = \frac{2-N}{\left[\tau + \nu\right]}(u_k^{2-N} - u_0^{2-N}).
\end{aligned}$$

When $u_k^N \geq \theta^*$, we have that $u_k^{2-N} \leq (\theta^*)^{\frac{2-N}{N}}$. Therefore,

$$\sum_{i=1}^{k} -N\eta\,(\theta_k - \theta^* + b_k) \geq \frac{2-N}{\tau + \nu}(u_k^{2-N} - u_0^{2-N}).$$

Similarly, it can be obtained

$$\sum_{i=1}^{k} N\eta\,(\theta_k - \theta^* + b_k) \geq \frac{2-N}{\tau + \nu}(v_k^{2-N} - v_0^{2-N}).$$

Therefore, we have that

$$\begin{aligned}
&\frac{N-2}{\tau+\nu}\left((\theta^*)^{\frac{2-N}{N}}-\alpha^{2-N}\right)\geq\frac{2-N}{\tau+\nu}(v_k^{2-N}-\alpha^{2-N})\\
\implies&\left(\alpha^{2-N}-(\theta^*)^{\frac{2-N}{N}}\right)+\alpha^{2-N}\leq v_k^{2-N}\\
\implies& v_k\leq\left(\frac{1}{2\alpha^{2-N}-(\theta^*)^{\frac{2-N}{N}}}\right)^{\frac{1}{N-2}}\\
\implies& v_k\leq\left(\frac{1}{2-\alpha^{N-2}/(\theta^*)^{\frac{N-2}{N}}}\right)^{\frac{1}{N-2}}\alpha\\
\implies& v_k\leq\left(\frac{1}{2-(2-2^{\frac{N-2}{N}})}\right)^{\frac{1}{N-2}}\alpha\\
\implies& v_k\leq\frac{1}{2^{\frac{1}{N}}}\alpha.
\end{aligned}$$

For the second part, when $v_k^N\geq|\theta^*|$, we have that $v_k^{2-N}\leq|\theta^*|^{\frac{2-N}{N}}$. Therefore, we have that

$$\begin{aligned}
&\frac{N-2}{\tau+c}\left((\theta^*)^{\frac{2-N}{N}}-\alpha^{2-N}\right)\geq\frac{2-N}{\tau+c}(u_k^{2-N}-\alpha^{2-N})\\
\implies&\left(\alpha^{2-N}-|\theta^*|^{\frac{2-N}{N}}\right)+\alpha^{2-N}\leq u_k^{2-N}\\
\implies& u_k\leq\left(\frac{1}{2\alpha^{2-N}-|\theta^*|^{\frac{2-N}{N}}}\right)^{\frac{1}{N-2}}\\
\implies& u_k\leq\left(\frac{1}{2-\alpha^{N-2}/|\theta^*|^{\frac{N-2}{N}}}\right)^{\frac{1}{N-2}}\alpha\\
\implies& u_k\leq\left(\frac{1}{2-(2-2^{\frac{N-2}{N}})}\right)^{\frac{1}{N-2}}\alpha\\
\implies& u_k\leq\frac{1}{2^{\frac{1}{N}}}\alpha.
\end{aligned}$$

□

Lemma 12. *Let $\theta_k=\theta_k^+-\theta_k^-$ and $\theta^*\in\mathbb{R}$ be the target such that $|\theta^*|>0$. Suppose the sequences $(\theta_k^+)_{k\geq0}$ and $(\theta_k^-)_{k\geq0}$ evolve as follows*

$$\begin{aligned}
&0<\theta_0^+=\alpha^N\leq2^3(2^{\frac{1}{N}}-1)^{\frac{N}{N-2}}|\theta^*|,\\
&\theta_{k+1}^+=\theta_k^+\left\{1-N\eta\left[\tau+\eta\gamma N^2(\theta_k^+)^{\frac{2N-2}{N}}\right](\theta_k-\theta^*+b_k)(\theta_k^+)^{\frac{N-2}{N}}\right\}^N,\\
&0<\theta_0^-=\alpha^N\leq2^3(2^{\frac{1}{N}}-1)^{\frac{N}{N-2}}|\theta^*|,\\
&\theta_{k+1}^-=\theta_k^-\left\{1+N\eta\left[\tau+\eta\gamma N^2(\theta_k^-)^{\frac{2N-2}{N}}\right](\theta_k-\theta^*+b_k)(\theta_k^-)^{\frac{N-2}{N}}\right\}^N,
\end{aligned}$$

and that there exists $B>0$ such that $|b_k|\leq B$ and $\eta\leq\frac{1}{4\nu N^2\gamma(|\theta^*|+B)|\theta^*|^{\frac{N-2}{N}}}$, $(\nu>1)$. Then the following holds:

- *If $\theta^*>0$ then $\theta_k^-\leq\alpha^N\prod_{i=0}^{t-1}\left\{1+N\eta\left[\tau+\eta\gamma N^2(\theta_i^-)^{\frac{2N-2}{N}}\right](|b_i|)(\theta_i^-)^{\frac{N-2}{N}}\right\}^N$.*
- *If $\theta^*<0$ then $\theta_k^+\leq\alpha^N\prod_{i=0}^{t-1}\left\{1-N\eta\left[\tau+\eta\gamma N^2(\theta_i^+)^{\frac{2N-2}{N}}\right](|b_i|)(\theta_i^+)^{\frac{N-2}{N}}\right\}^N$.*

Proof. For the first part, assume $\theta^* > 0$ and fix any $t \geq 0$. Let $0 \leq t \leq k$ be the largest t such that $\nu\theta^* > \theta_t^+ > \theta^*$.

Case 1. If no such t exists, we have $\theta_t^+ \leq \theta^*$ for any $0 \leq t \leq k$. Then,

$$\theta_t - \theta^* + b_k = \theta_t^+ - \theta_t^- - \theta^* + b_t \leq |b_t|.$$

Therefore

$$\theta_t^- \leq \alpha^N \left\{1 + N\eta \left[\tau + \eta\gamma N^2 (\theta_t^-)^{\frac{2N-2}{N}}\right] (|b_t|) (\theta_t^-)^{\frac{N-2}{N}}\right\}^N.$$

The conclusion holds.

Case 2. If $t = k$ then by the first part of Lemma 11, we have $\theta_k^- \leq \alpha$. The conclusion holds.

Case 3. If t exists, we have $\nu\theta^* \geq \theta_t^+ \geq \theta^*$ for any $0 \leq t \leq k$. By the first part of Lemma 11, we have $\theta_t^- \leq \frac{1}{2}\alpha^N$. From the requirement of initialization, we have

$$\begin{aligned}
&\left\{1 + N\eta \left[\tau + \eta\gamma N^2 (\theta_t^-)^{\frac{2N-2}{N}}\right] (\theta_t - \theta^* + b_t) (\theta_t^-)^{\frac{N-2}{N}}\right\}^N \\
\leq &\left\{1 + \left[\tau + \frac{(\frac{1}{2}\alpha^N)^{\frac{2N-2}{N}}}{4\nu(\theta^* + B)(\theta^*)^{\frac{N-2}{N}}}\right] \frac{(\nu\theta^* + |b_t|)(\frac{1}{2}\alpha^N)^{\frac{N-2}{N}}}{4N\nu\gamma(\theta^* + B)(\theta^*)^{\frac{N-2}{N}}}\right\}^N \\
\leq &\left\{1 + \left[\tau + \frac{(2^2(2^{\frac{1}{N}} - 1)^{\frac{N}{N-2}}|\theta^*|)^{\frac{2N-2}{N}}}{4\nu(\theta^* + B)(\theta^*)^{\frac{N-2}{N}}}\right] \frac{(\nu\theta^* + |b_t|)(2^2(2^{\frac{1}{N}} - 1)^{\frac{N}{N-2}}|\theta^*|)^{\frac{N-2}{N}}}{4N\nu\gamma(\theta^* + B)(\theta^*)^{\frac{N-2}{N}}}\right\}^N \\
\leq &\left\{1 + \left[\tau + \frac{4^{\frac{2N-2}{N}}(2^{\frac{1}{N}} - 1)^{\frac{2N-2}{N-2}}}{4}\right] \frac{4^{\frac{N-2}{N}}(2^{\frac{1}{N}} - 1)}{4N\gamma}\right\}^N \\
\leq &\left\{1 + \left[\tau + 4^{\frac{N-2}{N}}(2^{\frac{1}{N}} - 1)^{\frac{2N-2}{N-2}}\right] \frac{2^{\frac{1}{N}} - 1}{N\gamma}\right\}^N \\
\leq &\left\{1 + \frac{(\tau + 4)\left(2^{\frac{1}{N}} - 1\right)}{N\gamma}\right\}^N \leq 2.
\end{aligned}$$

where the last inequality follows from the condition on the momentum parameter $\gamma - 2\tau \geq 4$. Therefore

$$\begin{aligned}
\theta_k^- =&\theta_t^- \prod_{i=t}^{k-1} \left\{1 + N\eta \left[\tau + \eta\gamma N^2 (\theta_i^-)^{\frac{2N-2}{N}}\right] \left(\theta_i^+ - \theta_i^- - \theta^* + b_i\right) (\theta_i^-)^{\frac{N-2}{N}}\right\}^N \\
=&\frac{1}{2}\alpha^N \cdot 2 \prod_{i=t+1}^{k-1} \left\{1 + N\eta \left[\tau + \eta\gamma N^2 (\theta_i^-)^{\frac{2N-2}{N}}\right] \left(\theta_i^+ - \theta_i^- - \theta^* + b_i\right) (\theta_i^-)^{\frac{N-2}{N}}\right\}^N \\
\leq&\alpha^N \prod_{i=t+1}^{k-1} \left\{1 + N\eta \left[\tau + \eta\gamma N^2 (\theta_i^-)^{\frac{2N-2}{N}}\right] (|b_i|) (\theta_i^-)^{\frac{N-2}{N}}\right\}^N.
\end{aligned}$$

This completes the proof for $\theta^* > 0$. For the case $\theta^* < 0$.

Case 1. If no such t exists, we have $\theta_t^- \leq |\theta^*|$ for any $0 \leq t \leq k$. Therefore,

$$\theta_t - \theta^* + b_k = \theta_t^+ - \theta_t^- - \theta^* + b_t \geq -\theta_t^- - \theta^* + b_t \geq |b_t|$$

The conclusion holds.

Case 2. If $t = k$ then by the first part of Lemma 11, we have $\theta_k^+ \leq \alpha$. The conclusion holds.

Case 3. If t exists, we have $\nu|\theta^*| \geq \theta_t^- \geq |\theta^*|$ for any $0 \leq t \leq k$. By the first part of Lemma 11, we have $\theta_t^+ \leq \frac{1}{2}\alpha^N$. From the requirement of initialization, we have

$$\left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_t^+)^{\frac{2N-2}{N}}\right] (\theta_t - \theta^* + b_t) (\theta_t^+)^{\frac{N-2}{N}}\right\}^N \leq 2.$$

Therefore

$$\begin{aligned}\theta_k^+ =&\theta_t^+ \prod_{i=t}^{k-1} \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_i^+)^{\frac{2N-2}{N}}\right] \left(\theta_i^+ - \theta_i^- - \theta^* + b_i\right) (\theta_i^+)^{\frac{N-2}{N}}\right\}^N \\ =&\frac{1}{2}\alpha^N \cdot 2 \prod_{i=t+1}^{k-1} \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_i^+)^{\frac{2N-2}{N}}\right] \left(\theta_i^+ - \theta_i^- - \theta^* + b_i\right) (\theta_i^+)^{\frac{N-2}{N}}\right\}^N \\ \leq&\alpha^N \prod_{i=t+1}^{k-1} \left\{1 - N\eta \left[\tau + \eta\gamma N^2 (\theta_i^+)^{\frac{2N-2}{N}}\right] (|b_i|) (\theta_i^+)^{\frac{N-2}{N}}\right\}^N.\end{aligned}$$

This completes the proof for $\theta^* < 0$. □

F THE PROOF OF PROPOSITIONS 1 AND 2

Before proving Proposition 1 and Proposition 2, we introduce two lemmas that characterize the error growth during the training process.

Lemma 13. *Consider the momentum gradient descent update rule 4. Assume there exists a constant $B > 0$ such that for all k, $\|\boldsymbol{b}_k\|_\infty + \|\boldsymbol{p}_k\|_\infty \leq B$, then we have*

$$\|\mathbf{1}_{S^c} \boldsymbol{e}_{k+1}\|_\infty \leq \|\mathbf{1}_{S^c} \boldsymbol{e}_k\|_\infty \left\{1 + N\eta \left[\tau + \eta\gamma N^2 \|\boldsymbol{e}_k\|_\infty^{2N-2}\right] \cdot (\|\boldsymbol{p}_k\|_\infty + \|\boldsymbol{b}_k\|_\infty) \|\boldsymbol{e}_k\|_\infty^{N-2}\right\}^N.$$

Proof. Taking the Hadamard product of both sides of the iteration with $\mathbf{1}_{S^c}$, we obtain

$$\begin{aligned}&\mathbf{1}_{S^c} \odot \boldsymbol{\theta}_{k+1} \\ =&\mathbf{1}_{S^c} \odot \mathbf{1}_{S^c} \odot \boldsymbol{\theta}_{k+1} \odot \mathbf{1}_{S^c} \odot \mathbf{1}_{S^c} \\ =&\mathbf{1}_{S^c} \odot \boldsymbol{\theta}_k \odot \left\{\mathbf{1}_{S^c} - N\eta \left[\tau \mathbf{1}_{S^c} + \eta N^2 \mathbf{1}_{S^c} \odot \boldsymbol{\theta}_k^{\frac{2N-2}{N}}\right] \odot \mathbf{1}_{S^c} \right. \\ &\left. \odot (p_k + b_k + s_k - \theta^*) \odot \mathbf{1}_{S^c} \odot \boldsymbol{\theta}_k^{\frac{N-2}{N}}\right\}^N \\ =&\mathbf{1}_{S^c} \odot e_k \odot \left\{\mathbf{1}_{S^c} - N\eta \left[\tau \mathbf{1}_{S^c} + \eta N^2 e_k^{\frac{2N-2}{N}}\right] \odot (p_k + b_k) \odot e_k^{\frac{N-2}{N}}\right\}^N.\end{aligned}$$

Therefore, we have

$$\|\mathbf{1}_{S^c} \odot e_{k+1}\|_\infty \leq \|\mathbf{1}_{S^c} \odot e_k\|_\infty \left\{1 + N\eta \left[\tau + \eta N^2 \|e_k\|_\infty^{\frac{2N-2}{N}}\right] \cdot (\|p_k\|_\infty + \|b_k\|_\infty) \|e_k\|_\infty^{\frac{N-2}{N}}\right\}^N.$$

□

Lemma 14. *For all $k > 0$, assume that the sequence $(b_k)_{k\geq 0}$ is bounded, i.e., there exists a positive constant B such that $|b_k| \leq B$. Consider the following error iteration sequence*

$$\theta_{k+1} = \theta_k \left\{1 + N\eta \left[\tau + \eta N^2 \theta_k^{\frac{2N-2}{N}}\right] \cdot b_k \theta_k^{\frac{N-2}{N}}\right\}^N,$$

and assume that the learning rate satisfies $\eta \leq \frac{1}{N^{2\gamma} B \theta_0^{\frac{2N-2}{2N}}}$. Then for any iteration

$$k < \frac{1}{2\beta N(N-1)B\eta} \left(\frac{1}{\theta_0^{\frac{N-2}{N}}} - \frac{1}{\theta_0^{\frac{N-2}{2N}}}\right),$$

we have $\theta_k \leq \sqrt{\theta_0}$ where $\beta = \frac{\beta'}{1 - \frac{\beta'}{\gamma}}$ and $\beta' = \tau + \frac{1}{B}$.

Proof. Recalling the error iteration sequence, we have

$$\begin{aligned}\theta_{k+1} =&\theta_k \left\{1 + N\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right] \cdot b_k \theta_k^{\frac{N-2}{N}}\right\}^N \\ \leq&\theta_k \left\{1 + N\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right] \cdot B \theta_k^{\frac{N-2}{N}}\right\}^N \\ \leq&\theta_k \left(1 + \frac{N^2\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right] \cdot B \theta_k^{\frac{N-2}{N}}}{1 - (N-1)N\eta \left[\tau + \eta\gamma N^2 \theta_k^{\frac{2N-2}{N}}\right] \cdot B \theta_k^{\frac{N-2}{N}}}\right) \\ \leq&\theta_k \left(1 + \beta N^2 \eta B \theta_k^{\frac{N-2}{N}}\right),\end{aligned}$$

where $\beta = \frac{\beta'}{1-\frac{\beta'}{\gamma}}$ and $\beta' = \tau + \frac{1}{B}$. When $\gamma - \tau > \frac{1}{B}$, we know that

$$\beta \geq 1 \Leftrightarrow 1 - \frac{\beta'}{\gamma} \leq \beta' \Leftrightarrow \gamma - \tau - \frac{1}{B} \leq \gamma(\tau + \frac{1}{B}) \Leftrightarrow \gamma(\tau + \frac{1}{B} - 1) \geq 0 \geq -\tau - \frac{1}{B}$$

Therefore, for the number of iterations K required for θ_K to reach a given value, the following lower bound holds:

$$\sum_{k=1}^{K-1} \frac{\theta_{k+1} - \theta_k}{\beta N^2 B\eta\, \theta_k^{\frac{2N-2}{N}}} \leq K.$$

Therefore

$$\begin{aligned}&\sum_{t=0}^{K-1} \frac{\theta_{k+1} - \theta_k}{\beta N^2 B\eta \theta_k^{\frac{2N-2}{N}}} \geq \sum_{k=0}^{K-1} \int_{\theta_k}^{\theta_{k+1}} \frac{1}{\beta N^2 B\eta \theta^{\frac{2N-2}{N}}}\, d\theta \\ &\geq \int_{\theta_0}^{\theta_K} \frac{1}{\beta N^2 B\eta \theta^{\frac{2N-2}{N}}}\, d\theta \geq \int_{\theta_0}^{\sqrt{\theta_0}} \frac{1}{\beta N^2 B\eta \theta^{\frac{2N-2}{N}}}\, d\theta - \int_{\theta_K}^{\sqrt{\theta_0}} \frac{1}{\beta N^2 B\eta \theta^{\frac{2N-2}{N}}}\, d\theta \\ &\geq \int_{\theta_0}^{\sqrt{\theta_0}} \frac{1}{\beta N^2 B\eta \theta^{\frac{2N-2}{N}}}\, d\theta - 1 \geq \frac{1}{2\beta N(N-1)B\eta} \left(\frac{1}{x_0^{\frac{N-2}{N}}} - \frac{1}{x_0^{\frac{N-2}{2N}}}\right) - 1.\end{aligned}$$

In summary, for any $k > 0$ satisfying $k \leq \frac{1}{2\beta(N-1)B\eta} \left(\frac{\frac{1}{N^2}}{\theta_0^{\frac{N-2}{2}}} - \frac{1}{\theta_0^{\frac{N-2}{2}}}\right) - 1$, we have $\theta_k \leq \sqrt{\theta_0}$.

This completes the proof. □

Lemma 15 (Weighted incoherence inequality)**.** *Let $\boldsymbol{z} \in \mathbb{R}^p$ and let $\boldsymbol{w} \in \mathbb{R}^p_+$. Define*

$$M_S(\boldsymbol{w}) := \max_{j \in S} \boldsymbol{w}_j, \qquad M_{S^c}(\boldsymbol{w}) := \max_{j \in S^c} \boldsymbol{w}_j.$$

Then

$$\|\boldsymbol{E}(\boldsymbol{z} \odot \boldsymbol{w})\|_\infty \leq \sigma \left[\pi M_S(\boldsymbol{w}) \|\boldsymbol{z}_S\|_\infty + p M_{S^c}(\boldsymbol{w}) \|\boldsymbol{z}_{S^c}\|_\infty\right].$$

Proof. For an arbitrary coordinate i, $[\boldsymbol{E}(\boldsymbol{z} \odot \boldsymbol{w})]_i = \sum_{j=1}^p \boldsymbol{E}_{ij} \boldsymbol{z}_j \boldsymbol{w}_j$. Since $\boldsymbol{E}_{ii} = 0$, the diagonal term vanishes. Splitting the remaining indices into S and S^c gives

$$[\boldsymbol{E}(\boldsymbol{z} \odot \boldsymbol{w})]_i = \sum_{j \in S} \boldsymbol{E}_{ij} \boldsymbol{z}_j \boldsymbol{w}_j + \sum_{j \in S^c} \boldsymbol{E}_{ij} \boldsymbol{z}_j \boldsymbol{w}_j.$$

Hence

$$\begin{aligned}|[\boldsymbol{E}(\boldsymbol{z} \odot \boldsymbol{w})]_i| &\leq \sum_{j \in S} |\boldsymbol{E}_{ij}||\boldsymbol{z}_j| \boldsymbol{w}_j + \sum_{j \in S^c} |\boldsymbol{E}_{ij}||\boldsymbol{z}_j| \boldsymbol{w}_j \\ &\leq \sum_{j \in S} \sigma \|\boldsymbol{z}_S\|_\infty M_S(\boldsymbol{w}) + \sum_{j \in S^c} \sigma \|\boldsymbol{z}_{S^c}\|_\infty M_{S^c}(\boldsymbol{w}) \\ &= \pi\sigma M_S(\boldsymbol{w}) \|\boldsymbol{z}_S\|_\infty + |S^c| \sigma M_{S^c}(\boldsymbol{w}) \|\boldsymbol{z}_{S^c}\|_\infty.\end{aligned}$$

Since $|S^c| \leq p$, we have

$$|[\boldsymbol{E}(\boldsymbol{z} \odot \boldsymbol{w})]_i| \leq \sigma \left[\pi M_S(\boldsymbol{w}) \|\boldsymbol{z}_S\|_\infty + p M_{S^c}(\boldsymbol{w}) \|\boldsymbol{z}_{S^c}\|_\infty\right].$$

Taking the maximum over i proves the claim. □

Proposition 1. *Under the setting of the nonnegative iteration (4), for condition number* $d = C_1 \frac{\theta^*_{\min}}{\theta^*_{\max}}$ *and any precision* $\epsilon > 0$*, where* C_1 *is a small positive constant. Assume that the two error sequences* $(\boldsymbol{b}_k)_{k\geq 0}$ *and* $(\boldsymbol{p}_k)_{k\geq 0}$ *satisfy the following conditions:*

$$\|\boldsymbol{b}_k\|_\infty \leq C_2\zeta - \alpha^{\frac{N}{4}},$$
$$\|\boldsymbol{p}_k\|_\infty \leq d\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty,$$

where C_2 *is a small positive constant and* $\zeta := \frac{1}{5}\theta^*_{\min} \vee \frac{200}{n}\left\|\mathbf{X}^\top \boldsymbol{\xi}\right\|_\infty \vee 200\epsilon$*. If the initial value* α *satisfies*

$$\alpha \leq \left(\frac{1}{8c}\right)^{\frac{2}{N-2}} \vee \left[\tau + \frac{25}{64}\eta N^2 \zeta^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}} \zeta^{\frac{N-2}{N}} \log\frac{\theta^*_{\max}}{\epsilon},$$

then for any $K \in [K_1, K_2]$*, where*

$$K_1 = \frac{15}{4N(N-2)\eta\zeta\tau\alpha^{N-2}} + \frac{75}{4N^2\eta\left[\tau + \frac{25}{64}\eta N^2\zeta^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}}} \log\frac{\theta^*_{\max} - \alpha^N}{\epsilon},$$

and

$$K_2 = \frac{20}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}}\right),$$

we have for any $k \in [0, K]$*,*

$$\|\boldsymbol{s}_K - \boldsymbol{\theta}^*\|_\infty \leq \zeta,$$
$$\|\boldsymbol{e}_k\|_\infty \leq \alpha^{\frac{N}{2}}.$$

Proof. By the assumptions on $(\boldsymbol{b}_k)_{k\geq 0}$ and $(\boldsymbol{p}_k)_{k\geq 0}$, we obtain that

$$\|\boldsymbol{b}_k\|_\infty \leq C_2\zeta - \alpha^{N/4},$$
$$\|\boldsymbol{p}_k\|_\infty \leq \frac{C_1}{\theta^*_{\max}/\zeta}\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty \leq \frac{C_1}{\theta^*_{\max}/\zeta}\theta^*_{\max} \leq C_1\zeta.$$

Choose C_1 and C_2 such that $C_1 + C_2 \leq \frac{1}{40}$. Therefore, we have

$$B \leq \|\boldsymbol{b}_k\|_\infty + \|\mathbf{p}_k\|_\infty + \alpha^{N/4} \leq (C_1 + C_2)\zeta \leq \frac{1}{40}\zeta.$$

For any $\theta^*_j \geq \frac{1}{2}\zeta$ that is satisfied, we have $B \leq \frac{1}{20}\theta^*_j$. Therefore, applying Lemma 9, it follows that when

$$k \geq \frac{15}{4N(N-2)\eta\zeta\tau\alpha^{N-2}} + \frac{75}{4N^2\eta\left[\tau + \frac{25}{64}\eta\gamma N\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{2N-2}{N}}} \log\frac{|\theta^*_{\max} - \alpha^N|}{\epsilon} =: K_1,$$

we have $|\theta_{j,t} - \theta^*_j| \leq \zeta$.

For any j such that $\theta^*_j \leq \frac{1}{2}\zeta$, applying Lemma 6, $\theta_{j,k} \in (0, \theta^*_j + \frac{1}{40}\zeta]$ maintaining $|\theta_{j,k} - \theta^*_j| \leq \zeta$ as required.

By Lemma 14, we have that $\|\boldsymbol{e}_k\|_\infty \leq \alpha^{\frac{N}{2}}$ up to

$$K_2 = \frac{20}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}}\right).$$

From our choice of initialization α,

$$\alpha \leq \left(\frac{1}{8\beta}\right)^{2/(N-2)} \vee \left(\frac{\left[\tau + \frac{25}{64}\eta\gamma N\zeta^{\frac{2N-2}{N}}\right]^{(N-2)/N}\zeta^{(N-2)/N}}{\log\frac{\theta^*_{\max}}{\epsilon}}\right)^{2/(N-2)}$$

We can see that the interval $[K_1, K_2]$ is nonempty. Since $\frac{4\beta x}{1-4\beta x} \leq 1$ $(0 < x \leq \frac{1}{8\beta})$, we have

$$
\begin{aligned}
&\alpha^{\frac{N-2}{2}} \frac{4\beta\alpha^{\frac{N-2}{2}}}{1-4\beta\alpha^{\frac{N-2}{2}}} \leq \alpha^{\frac{N-2}{2}} \leq \frac{16}{15} \frac{\left[\tau + \frac{25}{64}\eta\gamma N\zeta^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}} \zeta^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\max}}{\epsilon}} \\
\Rightarrow\ & 4\beta\alpha^{\frac{N-2}{2}} \left(\frac{15}{16}\alpha^{\frac{N-2}{2}} \log\frac{\theta^*_{\max}}{\epsilon} + \left[\tau + \frac{25}{64}\eta\gamma\zeta^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}} \zeta^{\frac{N-2}{N}}\right) \leq \left[\tau + \frac{25}{64}\eta\gamma N\zeta^{\frac{2N-2}{N}}\right]^{\frac{N-2}{N}} \zeta^{\frac{N-2}{N}} \\
\Rightarrow\ & \frac{\frac{15\beta}{2}\alpha^{N-2}\log\frac{\theta^*_{\max}}{\epsilon}}{\left[\tau + \frac{25}{64}\eta\gamma\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{N-2}{N}}} + 6\beta\alpha^{\frac{N-2}{2}} \leq 2 - 2\alpha^{\frac{N-2}{2}} \\
\Rightarrow\ & \frac{\frac{15\beta}{2}\alpha^{N-2}\log\frac{\theta^*_{\max}}{\epsilon}}{\left[\tau + \frac{25}{64}\eta\gamma\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{N-2}{N}}} + 6 \leq 4 - 4\alpha^{\frac{N-2}{2}} \\
\Rightarrow\ & \frac{15}{2\left[\tau + \frac{25}{64}\eta\gamma\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{N-2}{N}}} \log\frac{\theta^*_{\max}}{\epsilon} + \frac{6}{\beta\alpha^{N-2}} \leq \frac{4}{\beta}\left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}}\right) \\
\Rightarrow\ & \frac{75}{4N(N-1)\eta\left[\tau + \frac{25}{16}\eta\gamma N\zeta^{\frac{2N-2}{N}}\right]\zeta^{(2N-2)/2}} \log\frac{\theta^*_{\max}}{\epsilon} + \frac{15}{N(N-1)\eta\tau\zeta\alpha^{(N-2)}} \\
& \leq \frac{20}{2\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}}\right) \\
\Rightarrow\ & \frac{75}{4N(N-1)\eta\left[\tau + \frac{25}{16}\eta\gamma N\zeta^{\frac{2N-2}{N}}\right]\zeta^{(2N-2)/2}} \log\frac{\theta^*_{\max}}{\epsilon} + \frac{15}{4N(N-2)\eta\tau\zeta\alpha^{(N-2)}} \\
& \leq \frac{20}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{1}{\alpha^{(N-2)/2}}\right) \\
\Rightarrow\ & K_1 \leq K_2.
\end{aligned}
$$

□

Remark 5. *At the beginning of the iteration $(k = 0)$, we have $\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty \leq \mathcal{O}(\theta^*_{\max})$ and $\|\boldsymbol{e}_0\|_\infty \leq \alpha^N$. Furthermore, Proposition 1 shows that after a certain number of iterations, we obtain $\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty \leq \mathcal{O}(\theta^*_{\min})$ and $\|\boldsymbol{e}_k\|_\infty \leq \alpha^{\frac{N}{2}}$.*

Proposition 2. *Under the setting of the nonnegative iteration (4). Fix any $0 < \zeta \leq \theta^*_{\max}$ and assume that the two error sequences $(\boldsymbol{b}_k)$ and $(\boldsymbol{p}_k)$ for any $k \geq 0$ satisfy*

$$
\begin{aligned}
&B = \|\boldsymbol{b}_k\|_\infty + \|\boldsymbol{p}_k\|_\infty \leq \frac{1}{200}\theta^*_{\min}, \\
&\|\boldsymbol{b}_k \odot \mathbf{1}_i\|_\infty \leq B_i \leq \frac{1}{10}\theta^*_{\min}, \\
&\|\boldsymbol{p}_k\|_\infty \leq \frac{1}{20}\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty.
\end{aligned}
$$

If the initial value α, the initial signal $\boldsymbol{s}_0$ and the initial error $\boldsymbol{e}_0$ satisfy the following condition

$$
\begin{aligned}
&\alpha \leq \left(\frac{1}{4}\right)^{\frac{2}{N-2}} \vee \left(\frac{(\theta^*_{\min})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\min}}{\epsilon}}\right)^{\frac{4}{N-2}}, \\
&\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty \leq \frac{1}{5}\theta^*_{\min}, \\
&\|\mathbf{e}_0\| \leq \alpha^{\frac{N}{2}}.
\end{aligned}
$$

Let the step size satisfy $\eta \le \frac{\alpha^N}{8N^2(\theta^*_{\min})^{(3N-2)/N}}$. *Then for any* $K_3 \le t \le K_4$,

$$K_3 := \frac{25}{N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}} \log\frac{\theta^*_{\min}}{\epsilon}$$
$$K_4 = \frac{100}{\beta N(N-1)\eta\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right),$$

and any $i \in S$ *we have*

$$|s_{i,t} - \theta^*_i| \lesssim k\mu \max_{j\in S} B_j \wedge B_i \wedge \epsilon, \quad \|\mathbf{e}_k\|_\infty \le \alpha^{\frac{N}{4}}.$$

Proof. Choose $B = \frac{1}{200}\theta^*_{\min}$. By Lemma 14, when the number of iterations

$$k \le \frac{100}{\beta N(N-1)\eta\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) := K_4,$$

we can maintain $\|\mathbf{e}_k\|_\infty \le \alpha^{\frac{N}{4}}$.

Denote $B_i := (\boldsymbol{b}_k)_i$ and let $B := \max_{j\in S} B_j$. If the signal error is reduced from $\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty < \frac{1}{5}\theta^*_{\min}$ to $\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty < 20(B \vee \epsilon)$, then we apply Lemma 10,

$$\frac{1}{5}\theta^*_{\min}2^{-m} = 20(B\vee\epsilon) \Rightarrow m = \log_2\frac{\theta^*_{\min}}{100(B\vee\epsilon)}.$$

Thus, by applying Lemma 10 $\log_2\frac{\theta^*_{\min}}{100(BV\epsilon)}$ times,

$$k \ge \frac{25}{2N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right][\theta^*]^{\frac{2N-2}{N}}} \log_2\frac{\theta^*_{\min}}{100(B\wedge\epsilon)}$$

we obtain the desired result. For any π-sparse vector $\boldsymbol{v}$, if $\frac{1}{\sqrt{n}}\boldsymbol{X}$ is an ℓ_2-normalized matrix satisfying the σ-incoherence condition, we have $\|\boldsymbol{p}_k\|_\infty = \|(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{I})\boldsymbol{v}\|_\infty \le \pi\sigma\|\boldsymbol{v}\|_\infty$. Then, we know that $\|\boldsymbol{p}_k\|_\infty < \pi\sigma\cdot 20(B\vee\epsilon)$. Therefore, for any $i \in S$, we have $\|\boldsymbol{b}_k \odot \mathbf{1}\|_\infty + \|\boldsymbol{p}_k\|_\infty \le B_i + \pi\sigma 20(B\vee\epsilon)$. For each coordinate $i \in S$, by Lemma 9, we obtain:

$$|\theta_{i,k} - \theta^*_i| \le \left|\frac{1}{n}(\boldsymbol{X}^T\boldsymbol{\xi})_i\right| \vee \pi\sigma\left\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{\xi}\odot\mathbf{1}_S\right\|_\infty \vee \epsilon.$$

The number of iteration required at this point is

$$\frac{12}{N^2\eta\left[\tau + \frac{1}{4}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}}\log\frac{\theta^*_{\min}}{\epsilon}.$$

Therefore, the total number of iterations required to further reduce $\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty$ is

$$\begin{aligned}K_3 :=& \frac{25}{N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}}\log\frac{\theta^*_{\min}}{\epsilon}\\ \ge& \frac{25}{2N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N(\theta^*)^{\frac{2N-2}{N}}\right][\theta^*]^{\frac{2N-2}{N}}}\log_2\frac{\theta^*_{\min}}{100(B\wedge\epsilon)}\\ &+ \frac{12}{N^2\eta\left[\tau + \frac{1}{4}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}}\log\frac{\theta^*_{\min}}{\epsilon}.\end{aligned}$$

Similarly, we need to ensure that the interval $[K_3, K_4]$ is nonempty. According to the condition on the initial value

$$\alpha^{\frac{N-2}{4}} \le \frac{\left[\tau + \frac{4}{25}\eta\gamma N(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{(N-2)/N}}{\log\frac{\theta^*_{\min}}{\epsilon}} \vee \frac{1}{2\beta}$$

Since $\frac{\beta x}{1-\beta x} \le 1$ for $0 \le x \le \frac{1}{2\beta}$, we have

$$
\begin{aligned}
\alpha^{\frac{N-2}{4}} &\le \frac{\left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\min}}{\epsilon}} \vee \frac{1}{2\beta} \\
&\Rightarrow \alpha^{\frac{N-2}{4}} \frac{\beta\alpha^{\frac{N-2}{4}}}{1-\beta\alpha^{\frac{N-2}{4}}} \le \alpha^{\frac{N-2}{4}} \le \frac{\left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{N-2}{N}}}{\log\frac{\theta^*_{\min}}{\epsilon}} \\
&\Rightarrow \beta\alpha^{\frac{N-2}{4}} \left(\alpha^{\frac{N-2}{4}} \log\frac{\theta^*_{\min}}{\epsilon} + \left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{N-2}{N}}\right) \\
&\quad \le \left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{N-2}{N}} \\
&\Rightarrow \frac{\beta\alpha^{\frac{2N-4}{4}}}{\left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{N-2}{N}}} \log\frac{\theta^*_{\min}}{\epsilon} + \alpha^{\frac{N-2}{4}} \le 1 \\
&\Rightarrow \frac{\beta\alpha^{\frac{2N-4}{4}}}{\left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{N-2}{N}}} \log\frac{\theta^*_{\min}}{\epsilon} \le 1 - \alpha^{\frac{N-2}{4}} \\
&\Rightarrow \frac{1}{\left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{N-2}{N}}} \log\frac{\theta^*_{\min}}{\epsilon} \le \frac{1}{\beta}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) \\
&\Rightarrow \frac{1}{\eta N(N-1)\left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}} \log\frac{\theta^*_{\min}}{\epsilon} \\
&\quad \le \frac{1}{\beta\eta N(N-1)\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{(N-2)/4}}\right) \\
&\Rightarrow \frac{25}{\eta N^2\left[\tau + \frac{4}{25}\eta\gamma N^2 (\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}} \log\frac{\theta^*_{\min}}{\epsilon} \\
&\quad \le \frac{100}{\beta\eta N(N-1)\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) \Rightarrow K_3 \le K_4.
\end{aligned}
$$

□

Remark 6. *Proposition 2 shows that, under the condition that $\|\boldsymbol{e}_k\|_\infty \le \alpha^{\frac{N}{4}}$, the approximation error can be further reduced from $\mathcal{O}(\theta^*_{\min})$ to $\mathcal{O}(\|\frac{1}{n}X^\top \boldsymbol{\xi}\|_\infty)$.*

Proposition 3 (Uniform Control of the Effective Perturbations under Incoherence)**.** *Assume that $\boldsymbol{\theta}^* \succ 0$ (all components are positive) and sparse parameters $\pi \ge 1$. Let the normalized design matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ satisfies σ-incoherence with $\sigma \le \frac{C_1}{\pi d}$, where C_1 is a sufficiently small constant, and $d := \frac{\theta^*_{\max}}{\theta^*_{\min}}$. Suppose that the iteration index lies in the range considered in Propositions 1 and 2. Then, for sufficiently small absolute constants C_1, C_2, the effective perturbations introduced in (10) and (11) satisfy*

$$\|\widetilde{\boldsymbol{p}}^{\boldsymbol{u}}_k\|_\infty \le \widetilde{C}_1\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty, \quad \|\widetilde{\boldsymbol{b}}^{\boldsymbol{u}}_k\|_\infty \le \widetilde{C}_2\|\boldsymbol{b}_k\|_\infty.$$

Consequently, the effective perturbation sequences satisfy the same conditions required in Propositions 1 and 2.

Proof. Since $\boldsymbol{p}_k = \boldsymbol{E}\boldsymbol{\rho}_k$ and $\boldsymbol{\rho}_k = \boldsymbol{s}_k - \boldsymbol{\theta}^*$ is supported on S, for every i, $(\boldsymbol{p}_k)_i = \sum_{j\in S} \boldsymbol{E}_{ij}\boldsymbol{\rho}_{k,j}$. Thus

$$|(\boldsymbol{p}_k)_i| \le \sum_{j\in S} |\boldsymbol{E}_{ij}||\boldsymbol{\rho}_{k,j}| \le \sum_{j\in S} \sigma\|\boldsymbol{\rho}_k\|_\infty = |S|\sigma\|\boldsymbol{\rho}_k\|_\infty = \pi\sigma\|\boldsymbol{\rho}_k\|_\infty.$$

Taking the maximum over i gives $\|\boldsymbol{p}_k\|_\infty \le \pi\sigma\|\boldsymbol{\rho}_k\|_\infty$. Since $\sigma \le \frac{C_1}{\pi d}$, we obtain $\pi\sigma \le \frac{C_1}{d}$. Therefore $\|\boldsymbol{p}_k\|_\infty \le \frac{C_1}{d}\|\boldsymbol{\rho}_k\|_\infty$.

We now control the additional signal-dependent correction

$$\boldsymbol{\delta}_k^{\boldsymbol{u},s} := \boldsymbol{R}_k^{\boldsymbol{u}} \boldsymbol{E}\left[(\boldsymbol{\rho}_k + \boldsymbol{p}_k) \odot \boldsymbol{u}_k^{2N-2}\right].$$

Recall that $\boldsymbol{R}_k^{\boldsymbol{u}} = \eta\gamma N^2 (\boldsymbol{H}_k^{\boldsymbol{u}})^{-1}$, where $\boldsymbol{H}_k^{\boldsymbol{u}} = \tau \boldsymbol{I} + \eta\gamma N^2 \operatorname{Diag}(\boldsymbol{u}_k^{2N-2})$. Since $\boldsymbol{H}_k^{\boldsymbol{u}} \succeq \tau \boldsymbol{I}$, we have $\left\|(\boldsymbol{H}_k^{\boldsymbol{u}})^{-1}\right\|_\infty \leq \frac{1}{\tau}$, and hence $\|\boldsymbol{R}_k^{\boldsymbol{u}}\|_\infty \leq \frac{\eta\gamma N^2}{\tau}$.

Therefore,

$$\begin{aligned}
\|\boldsymbol{\delta}_k^{\boldsymbol{u},s}\|_\infty &= \left\|\boldsymbol{R}_k^{\boldsymbol{u}} \boldsymbol{E}\left[(\boldsymbol{\rho}_k + \boldsymbol{p}_k) \odot \boldsymbol{u}_k^{2N-2}\right]\right\|_\infty \\
&\leq \|\boldsymbol{R}_k^{\boldsymbol{u}}\|_\infty \left\|\boldsymbol{E}\left[(\boldsymbol{\rho}_k + \boldsymbol{p}_k) \odot \boldsymbol{u}_k^{2N-2}\right]\right\|_\infty \\
&\leq \frac{\eta\gamma N^2}{\tau} \left\|\boldsymbol{E}\left[(\boldsymbol{\rho}_k + \boldsymbol{p}_k) \odot \boldsymbol{u}_k^{2N-2}\right]\right\|_\infty.
\end{aligned}$$

Since $\boldsymbol{\rho}_k$ is supported on S, i.e., $(\boldsymbol{\rho}_k)_{S^c} = 0$. Therefore $(\boldsymbol{\rho}_k + \boldsymbol{p}_k)_S = \boldsymbol{\rho}_{k,S} + \boldsymbol{p}_{k,S}$. Thus,

$$\begin{aligned}
\|(\boldsymbol{\rho}_k + \boldsymbol{p}_k)_S\|_\infty &\leq \|\boldsymbol{\rho}_k\|_\infty + \|\boldsymbol{p}_k\|_\infty \\
&\leq \|\boldsymbol{\rho}_k\|_\infty + \frac{C_1}{d}\|\boldsymbol{\rho}_k\|_\infty \\
&= \left(1 + \frac{C_1}{d}\right)\|\boldsymbol{\rho}_k\|_\infty.
\end{aligned}$$

On the complement S^c: $(\boldsymbol{\rho}_k + \boldsymbol{p}_k)_{S^c} = \boldsymbol{p}_{k,S^c}$. Therefore,$\|(\boldsymbol{\rho}_k + \boldsymbol{p}_k)_{S^c}\|_\infty \leq \frac{C_1}{d}\|\boldsymbol{\rho}_k\|_\infty$. We can obtain that

$$\left\|\mathbb{E}\left[(\boldsymbol{\rho}_k + \boldsymbol{p}_k) \odot \boldsymbol{u}_k^{2N-2}\right]\right\|_\infty \leq \sigma\left[\pi M_{S,k}^{\boldsymbol{u}}\left(1 + \frac{C_1}{d}\right) + p M_{S^c,k}^{\boldsymbol{u}} \frac{C_1}{d}\right]\|\boldsymbol{\rho}_k\|_\infty.$$

Therefore:

$$\|\boldsymbol{\delta}_k^{\boldsymbol{u},s}\|_\infty \leq \frac{\eta\gamma N^2 \sigma}{\tau}\left[\pi M_{S,k}^{\boldsymbol{u}}\left(1 + \frac{C_1}{d}\right) + \frac{pC_1}{d} M_{S^c,k}^{\boldsymbol{u}}\right]\|\boldsymbol{\rho}_k\|_\infty.$$

By definition: $\boldsymbol{\delta}_k^{\boldsymbol{u},\boldsymbol{b}} = \boldsymbol{R}_k^{\boldsymbol{u}} \boldsymbol{E}[\boldsymbol{b}_k \odot \boldsymbol{u}_k^{2N-2}]$. Therefore:

$$\|\boldsymbol{\delta}_k^{\boldsymbol{u},b}\|_\infty \leq \frac{\eta\gamma N^2}{\tau}\|\boldsymbol{E}[\boldsymbol{b}_k \odot \boldsymbol{u}_k^{2N-2}]\|_\infty \leq \frac{\eta\gamma N^2 \sigma}{\tau}[\pi M_{S,k}^{\boldsymbol{u}} + p M_{S^c,k}^{\boldsymbol{u}}]\|\boldsymbol{b}_k\|_\infty.$$

Therefore:

$$\|\boldsymbol{\delta}_k^{\boldsymbol{u},b}\|_\infty \leq \Gamma_k^{\boldsymbol{u}}\|\boldsymbol{b}_k\|_\infty,$$

where

$$\Gamma_k^{\boldsymbol{u}} := \frac{\eta\gamma N^2 \sigma}{\tau}[\pi M_{S,k}^{\boldsymbol{u}} + p M_{S^c,k}^{\boldsymbol{u}}].$$

Similarly, we define

$$\Lambda_k^{\boldsymbol{u}} := \frac{\eta\gamma N^2 \sigma}{\tau}\left[\pi M_{S,k}^{\boldsymbol{u}}\left(1 + \frac{C_1}{d}\right) + \frac{pC_1}{d} M_{S^c,k}^{\boldsymbol{u}}\right].$$

Then:

$$\begin{aligned}
\|\boldsymbol{\delta}_k^{\boldsymbol{u},s}\|_\infty &\leq \Lambda_k^{\boldsymbol{u}}\|\boldsymbol{\rho}_k\|_\infty, \\
\|\boldsymbol{\delta}_k^{\boldsymbol{u},\boldsymbol{b}}\|_\infty &\leq \Gamma_k^{\boldsymbol{u}}\|\boldsymbol{b}_k\|_\infty.
\end{aligned}$$

Thus, the new effective perturbations satisfy:

$$\|\widetilde{\boldsymbol{p}}_k^{\boldsymbol{u}}\|_\infty \leq \left(\frac{C_1}{d} + \Lambda_k^{\boldsymbol{u}}\right)\|\boldsymbol{\rho}_k\|_\infty,$$

and

$$\|\widetilde{\boldsymbol{b}}_k^{\boldsymbol{u}}\|_\infty \leq (1 + \Gamma_k^{\boldsymbol{u}})\|\boldsymbol{b}_k\|_\infty.$$

For every $j \in S$, $\boldsymbol{u}_{k,j}^N = s_{k,j} = \boldsymbol{\theta}_j^* + \boldsymbol{\rho}_{k,j}$. Since $|\boldsymbol{\rho}_{k,j}| \leq \|\boldsymbol{\rho}_k\|_\infty \leq \zeta \leq \theta_{\max}^*$, we obtain

$$\boldsymbol{u}_{k,j}^N \leq |\boldsymbol{\theta}_j^*| + |\boldsymbol{\rho}_{k,j}| \leq \theta_{\max}^* + \zeta \leq 2\theta_{\max}^*.$$

Therefore,

$$\boldsymbol{u}_{k,j}^{2N-2} = \left(\boldsymbol{u}_{k,j}^{N}\right)^{\frac{2N-2}{N}} \leq \left(2\theta_{\max}^{*}\right)^{\frac{2N-2}{N}}.$$

Hence

$$M_{S,k}^{\boldsymbol{u}} = \max_{j\in S} \boldsymbol{u}_{k,j}^{2N-2} \leq \left(2\theta_{\max}^{*}\right)^{\frac{2N-2}{N}}.$$

On the complement S^c, the corrected coordinate-wise dynamics,

$$\boldsymbol{u}_{k+1,j} = \boldsymbol{u}_{k,j}\left[1 - N\eta\left(\tau + \eta\gamma N^2 \boldsymbol{u}_{k,j}^{2N-2}\right)\left(\widetilde{\boldsymbol{p}}_{k,j}^{\boldsymbol{u}} + \widetilde{\boldsymbol{b}}_{k,j}^{\boldsymbol{u}}\right)\boldsymbol{u}_{k,j}^{N-2}\right].$$

Taking absolute values gives

$$\begin{aligned}|\boldsymbol{u}_{k+1,j}| &\leq |\boldsymbol{u}_{k,j}|\left[1 + N\eta\left(\tau + \eta\gamma N^2 |\boldsymbol{u}_{k,j}|^{2N-2}\right)\left|\widetilde{\boldsymbol{p}}_{k,j}^{\boldsymbol{u}} + \widetilde{\boldsymbol{b}}_{k,j}^{\boldsymbol{u}}\right||\boldsymbol{u}_{k,j}|^{N-2}\right] \\ &\leq |\boldsymbol{u}_{k,j}|\left[1 + N\eta\left(\tau + \eta\gamma N^2 |\boldsymbol{u}_{k,j}|^{2N-2}\right) B_{\text{off}}|\boldsymbol{u}_{k,j}|^{N-2}\right].\end{aligned}$$

Since $U_{k,j} = |\boldsymbol{u}_{k,j}|^N$, we have $|\boldsymbol{u}_{k,j}|^{N-2} = U_{k,j}^{\frac{N-2}{N}}$, and $|\boldsymbol{u}_{k,j}|^{2N-2} = U_{k,j}^{\frac{2N-2}{N}}$. Moreover, $|\boldsymbol{u}_{k+1,j}|^N = U_{k+1,j}$. Therefore,

$$U_{k+1,j} \leq U_{k,j}\left[1 + N\eta\left(\tau + \eta\gamma N^2 U_{k,j}^{\frac{2N-2}{N}}\right) B_{\text{off}} U_{k,j}^{\frac{N-2}{N}}\right]^N,$$

Since $U_{0,j} = \boldsymbol{u}_{0,j}^N = \alpha^N$, and according to Lemma 14, we have $U_{k,j} \leq \sqrt{U_{0,j}} = \sqrt{\alpha^N} = \alpha^{N/2}$. Finally,

$$\boldsymbol{u}_{k,j}^{2N-2} = \left(\boldsymbol{u}_{k,j}^{N}\right)^{\frac{2N-2}{N}} = U_{k,j}^{\frac{2N-2}{N}} \leq \left(\alpha^{N/2}\right)^{\frac{2N-2}{N}} = \alpha^{N-1}.$$

We have $M_{S^c,k}^{\boldsymbol{u}} = \max_{j\in S^c} \boldsymbol{u}_{k,j}^{2N-2} \leq \alpha^{N-1}$. Substituting $M_{S^c,k}^{\boldsymbol{u}} = \max_{j\in S^c} \boldsymbol{u}_{k,j}^{2N-2} \leq \alpha^{N-1}$ and $M_{S,k}^{\boldsymbol{u}} =\leq \left(2\theta_{\max}^{*}\right)^{\frac{2N-2}{N}}$ into $\Lambda_k^{\boldsymbol{u}}$ and $\Gamma_k^{\boldsymbol{u}}$ yields the final conclusion. □

G THE PROOF OF THE THEOREM 2

Proof. According to the initialization condition

$$\alpha \leq \left(\frac{\epsilon}{p+1}\right)^{\frac{4}{N}} \vee \text{Init}(\theta_{\max}^{*}, \theta_{\min}^{*}, \epsilon, N, \tau, \gamma)$$

Since for any vector $\boldsymbol{v}$ and $\frac{1}{\sqrt{n}}\boldsymbol{X}$ is an ℓ_2-normalized matrix satisfying the σ-incoherence condition, we have $\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}\right\|_\infty \leq p\|\boldsymbol{v}\|_\infty$. If $\|\boldsymbol{e}_k\|_\infty \leq \alpha^{\frac{4}{N}}$, then by Lemma 14 we have

$$\begin{aligned}\|\boldsymbol{b}_k\|_\infty + \alpha^{\frac{4}{N}} &\leq \left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty + \left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{e}_k\right\|_\infty + \alpha^{\frac{4}{N}} \leq 2\left(\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty \wedge \left(p\|\boldsymbol{e}_k\|_\infty\right) + \alpha^{\frac{4}{N}}\right) \\ &\leq 2\left(\left\|\frac{1}{n}\mathbf{X}^\top\boldsymbol{\xi}\right\|_\infty \wedge (p+1)\alpha^{\frac{4}{N}}\right) \leq C_2\frac{2}{C_2}\left(\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\epsilon}\right\|_\infty \wedge \epsilon\right) \leq C_2\zeta.\end{aligned}$$

where C_2 is a small positive constant. Furthermore, we have

$$\|\boldsymbol{p}_k\|_\infty \leq \frac{C_1}{\theta_{\max}^{*}/\theta_{\min}^{*}}\|\boldsymbol{s}_k - \boldsymbol{\theta}^*\|_\infty.$$

According to Proposition 3, the sequences $(\widetilde{\boldsymbol{b}}_k^{\boldsymbol{u}})_{k\geq 0}$ and $(\widetilde{\boldsymbol{p}}_k^{\boldsymbol{u}})_{k\geq 0}$ satisfy the conditions of Proposition 1 under the condition $\|\boldsymbol{e}_k\|_\infty < \alpha^{N/4}$. If $\zeta \geq \theta_{\max}^{*}$, then at $k = 0$ we have $\|\boldsymbol{s}_0 - \boldsymbol{\theta}^*\|_\infty \leq \zeta$. Otherwise, applying Proposition 1, after

$$K_1 = \frac{15}{4N(N-2)\eta\zeta\tau\alpha^{N-2}} + \frac{75}{4N^2\eta\left[\tau + \frac{25}{64}\eta\gamma N^2\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{2N-2}{N}}}\log\frac{|\theta_{\max}^{*} - \alpha^N|}{\epsilon},$$

iterations and before

$$K_2 = \frac{20}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}}\right),$$

iterations, we have

$$\|\boldsymbol{s}_{K_1} - \boldsymbol{\theta}^*\| \le \zeta, \quad \|\boldsymbol{e}_{K_1}\|_\infty \le \alpha^{\frac{N}{2}}.$$

If $\frac{1}{2}\theta^*_{\min} \le \frac{2}{C_2}\left|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right|_\infty \wedge \frac{2}{C_2}\epsilon$, then the proof is complete.

If $\frac{1}{2}\theta^*_{\min} \ge \frac{2}{C_2}\left|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right|_\infty \wedge \frac{2}{C_2}\epsilon$, then $\zeta = \frac{1}{5}\theta^*_{\min}$. Choosing $C_2 + C_1 \le \frac{1}{40}$, after K_1 iterations we have,

$$\left\|\widetilde{\boldsymbol{b}}^{\boldsymbol{u}}_k\right\|_\infty + \|\widetilde{\boldsymbol{p}}^{\boldsymbol{u}}_k\|_\infty \le C_2\frac{1}{5}\theta^*_{\min} + \frac{C_1}{\theta^*_{\max}/\theta^*_{\min}}\frac{1}{5}\theta^*_{\min} \le (C_1+C_2)\frac{1}{5}\theta^*_{\min} \le \frac{1}{200}\theta^*_{\min}.$$

From Proposition 2, after

$$K_3 = \frac{25}{N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}}\log\frac{\theta^*_{\min}}{\epsilon},$$

iterations and before

$$K_4 = \frac{100}{\beta N(N-1)\eta\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right),$$

iterations, we have

$$|s_{k,i} - \theta^*_i| \lesssim \pi\sigma\max_{j\in S} B_j \wedge B_i \wedge \epsilon, \qquad \|\boldsymbol{e}_k\|_\infty \le \alpha^{\frac{N}{4}}.$$

Therefore, the total required number of iterations is

$$\begin{aligned}
&K_3 + K_1\\
=&\frac{25}{N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}}\log\frac{\theta^*_{\min}}{\epsilon}\mathbb{I}\left\{\frac{1}{5}\theta^*_{\min} > \frac{2}{C_2}\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty \wedge \frac{2}{C_2}\epsilon\right\}\\
&+\frac{15}{4N(N-2)\eta\zeta\tau\alpha^{N-2}} + \frac{75}{4N^2\eta\left[\tau + \frac{25}{64}\eta\gamma N^2\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{2N-2}{N}}}\log\frac{|\theta^*_{\max} - \alpha^N|}{\epsilon},
\end{aligned} \tag{18}$$

and an upper bound on the total number of iterations is

$$\begin{aligned}
K_4 + K_2 =&\frac{100}{\beta N(N-1)\eta\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right)\mathbb{I}\left\{\frac{1}{5}\theta^*_{\min} > \frac{2}{C_2}\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty \wedge \frac{2}{C_2}\epsilon\right\}\\
&+\frac{20}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}}\right),
\end{aligned} \tag{19}$$

where $\mathbb{I}$ is the indicator function. □

H THE PROOF OF THE THEOREM 5

Theorem 5. *Assume the sparsity parameter $\pi \ge 1$ and the design matrix $\boldsymbol{X}/\sqrt{n}$ satisfies σ-incoherence with $\sigma \lesssim \frac{1}{\pi d}$. For any tolerance $\epsilon > 0$, suppose the initial value α satisfies*

$$0 < \alpha \le \left(\frac{\epsilon}{p+1}\right)^{4/N} \vee \mathrm{Init}(\theta^*_{\max}, \theta^*_{\min}, \epsilon, N, \tau, \gamma) \vee \tilde{\mathrm{Init}}(\theta^*_{\min}, N).$$

where $\tilde{\mathrm{Init}}(\theta^*_{\min}, N)$ is defined as

$$\tilde{\mathrm{Init}}(\theta^*_{\min}, N) := (2 - 2^{\frac{N-2}{N}})^{\frac{1}{N-2}}(\theta^*_{\min})^{\frac{1}{N}} \vee 2^{\frac{3}{N}}(2^{\frac{1}{N}} - 1)^{\frac{1}{N-2}}(\theta^*_{\min})^{\frac{1}{N}}.$$

For any iteration k satisfying $K_1+K_3 \le k \le K_2+K_4$, under the learning rate $\eta \le \frac{\alpha^N}{N^2\gamma(\theta^*_{\max})^{\frac{3N-2}{N}}}$, the k-th iterate $\boldsymbol{\theta}_k$ obtained by momentum gradient descent (4) and (5) satisfies the following properties:

$$|\theta_{k,i}-\theta_i^*| \lesssim \begin{cases} \left\|\frac{1}{n}\boldsymbol{X}^\top\xi\right\|_\infty \wedge \epsilon & \text{if } i\in S \text{ and } \theta^*_{\min} \lesssim \left\|\frac{1}{n}\boldsymbol{X}^\top\xi\right\|_\infty \wedge \epsilon, \\ \left|\frac{1}{n}(\boldsymbol{X}^\top\xi)_i\right| \wedge k\mu\|\frac{1}{n}\boldsymbol{X}^\top\xi\odot\mathbf{1}_S\|_\infty \wedge \epsilon & \text{if } i\in S \text{ and } \theta^*_{\min} \gtrsim \left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty \wedge \epsilon, \\ \alpha^{\frac{N}{4}} & \text{if } i\notin S. \end{cases}$$

In the special case $\theta^*=0$, if $\alpha \le \left(\frac{\epsilon}{p+1}\right)^{\frac{4}{N}}$, $\eta \le \frac{1}{N(N-1)\gamma\zeta\alpha^{(N-2)/2}}$ and $k \le K_2+K_4$, then for all i, $|\theta_{k,i}-\theta_i^*| \le \alpha^{\frac{N}{4}}$.

Proof. Theorem 2 can be extended to the general update rules (4) and (5). According to the initialization condition

$$\alpha \le \left(\frac{\epsilon}{p+1}\right)^{4/N} \vee \Phi(\theta^*_{\max},\theta^*_{\min},\epsilon,N,\tau,\gamma) \vee \Psi(\theta^*_{\min},N).$$

where $\mathrm{Init}(\theta^*_{\max},\theta^*_{\min},\epsilon,N,\tau,\gamma)$ is the initialization condition in Theorem 2.

The other settings are the same as in Theorem 2. From Lemma 13, Lemma 14, and Lemma 12, within the range where the number of iterations does not exceed the upper bound K_2+K_4 defined in Proposition 2, the following conclusions hold:

$$\theta_j^*>0 \Rightarrow 0\le\theta_k^-\le\alpha^{\frac{N}{4}},$$
$$\theta_j^*<0 \Rightarrow 0\le\theta_k^+\le\alpha^{\frac{N}{4}}.$$

When $\theta_j^*>0$, $(\theta_{j,k}^-)_{k\ge0}$ can be regarded as part of the bounded error b_k and thus ignored. Consequently, when $\theta_j^*>0$, the iteration of the sequence $(\theta_{j,k}^+)$ satisfies the following form:

$$\begin{aligned}&\theta_{k+1}^+\\ =&\theta_k^+\left\{1-N\eta\left[\tau+\eta\gamma\left(N^2(\theta_k^+)^{\frac{2N-2}{N}}\right)\right]\cdot\left(\left(\theta_{j,k}^+-\theta_j^*+\widetilde{p}_{j,k}+(\widetilde{b}_{j,k}-\theta_{j,k}^-)\right)\odot(\theta_k^+)^{\frac{N-2}{N}}\right)\right\}^N.\end{aligned}$$

When $\theta_j^*<0$, it suffices to replace θ_k^+ and θ_k^- accordingly. Therefore, following the proof idea of Theorem 2 and setting $C_2=\frac{1}{100}$, we obtain

$$\zeta=\frac{1}{5}\theta^*_{\min}\wedge 200\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty\wedge 200\epsilon.$$

For any number of iterations k satisfying $k\in[K_1+K_3,K_2+K_4]$, the conclusion of Theorem 2 holds. The definitions of K_1+K_2 and K_3+K_4 can be found in (18) and (19) of Theorem 2.

Finally, if all components of the target vector are zero, i.e., $\pi=0$, then $\theta^*=0$. Suppose $\alpha\le\left(\frac{\epsilon}{p+1}\right)^{\frac{4}{N}}$ and $\|\boldsymbol{e}_k\|_\infty\le\alpha^{\frac{4}{N}}$. Then we have

$$\begin{aligned}\|\boldsymbol{b}_k\|_\infty+\alpha^{N/4} &\le p\alpha^{N/4}+\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty+\alpha^{N/4}\\ &\le 2\left(\left\|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{\xi}\right\|_\infty\wedge\epsilon\right)\le C_2\zeta\le\frac{1}{40}\zeta.\end{aligned}$$

By Lemma 3, when the learning rate satisfies $\eta\le\frac{40}{N^2\gamma\zeta\alpha^{\frac{2N-2}{4}}}$ and the iteration number satisfies $k\le\frac{20}{cN(N-1)\eta\zeta}\left(\frac{1}{\alpha^{N-2}}-\frac{1}{\alpha^{(N-2)/2}}\right)$, we have $\|\boldsymbol{e}_k\|_\infty\le\alpha^{\frac{N}{4}}$. □

I THE PROOF OF THE THEOREM 3

Theorem 3. Let K_1, K_2, K_3 and K_4 be the numbers of iterations defined in Proposition 1 and Proposition 2. Assume that $\zeta\ge1$, $\theta^*_{\max}\ge1$ and the initialization $\alpha\le\exp(-\frac{5}{3})$. If α and η are fixed and do not vary with N and μ, then both K_2-K_1 and K_4-K_3 have a lower bound, and this lower bound increasing as the depth N increases ($N>2$).

Proof. Under the assumptions that $\zeta \geq 1$ and $\theta^*_{\max} \geq 1$,

$$\frac{75}{4N^2\eta\left[\tau + \frac{25}{64}\eta\gamma N^2\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{2N-2}{N}}} \log \frac{|\theta^*_{\max} - \theta_0|}{\epsilon}$$

and

$$K_3 = \frac{25}{N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}} \log \frac{\theta^*_{\min}}{\epsilon}$$

decrease as N increases. For the remaining part $K_2 - K_1$, we will show that it has a lower bound that increases with N.

$$\begin{aligned}
&\frac{20}{N(N-1)c\eta\zeta}\left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}}\right) - \frac{15}{4N(N-2)\eta\zeta\tau\alpha^{N-2}} \\
&\geq \frac{20}{4N(N-1)c\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{4}{\alpha^{\frac{N-2}{2}}}\right), \quad (N > 2).
\end{aligned}$$

Define the function

$$f(N) = \frac{1}{N(N-1)}\left(\frac{1}{\alpha^{N-2}} - \frac{4}{\alpha^{\frac{N-2}{2}}}\right).$$

Taking the derivative of this function with respect to N:

$$\begin{aligned}
f'(N) &= \frac{-(2N-1)}{N^2(N-1)^2}\left(\frac{1}{\alpha^{(N-2)}} - \frac{4}{\alpha^{\frac{N-2}{2}}}\right) + \frac{1}{N(N-1)}(-\log\alpha)\left(\frac{1}{\alpha^{(N-2)}} - \frac{4}{2\alpha^{\frac{N-2}{2}}}\right) \\
&= \frac{-(2N-1) - (N-1)N\log\alpha}{N^2(N-1)^2}\left(\frac{1}{\alpha^{(N-2)}} - \frac{2}{\alpha^{\frac{N-2}{2}}}\right) + \frac{2N-1}{N^2(N-1)^2}\frac{2}{\alpha^{\frac{N-2}{2}}}.
\end{aligned}$$

Since the second term is always positive, it suffices to ensure that the first term is positive:

$$-(2N-1) - (N-1)N\log\alpha \geq 0, \quad \text{and} \quad \frac{1}{\alpha^{(N-2)}} - \frac{2}{\alpha^{\frac{N-2}{2}}} \geq 0,$$

Therefore, the following conditions must be satisfied:

$$\begin{aligned}
\log\alpha \leq \min_{N\geq 3}\frac{(1-2N)}{N(N-1)} = \min_{N\geq 3}\left(\frac{1}{1-N} - \frac{1}{N}\right) = -\frac{5}{6}, \\
\alpha^{\frac{N-2}{2}} \leq 1/2.
\end{aligned}$$

When $\alpha \leq \exp(-5/6)\vee 1/4$, the lower bound of $K_2 - K_1$ increases with N. To prove that $K_4 - K_3$ increases with N, we first have

$$\frac{100}{\beta N(N-1)\eta\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) - \frac{10}{N^2\eta\left[\tau + \frac{1}{4}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}}\log\frac{\theta^*_{\min}}{\epsilon}.$$

Therefore, the second term decreases as N increases. Next, we only need to show that K_4 is increasing. Define

$$\tilde{f}(N) = \frac{1}{N(N-1)}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right).$$

Taking the derivative of the above function,

$$\begin{aligned}
\tilde{f}'(N) &= \frac{-(2N-1)}{N^2(N-1)^2}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) + \frac{1}{2N(N-1)}(-\log\alpha)\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{2\alpha^{\frac{N-2}{4}}}\right) \\
&= \frac{-2(2N-1) - N(N-1)\log\alpha}{2N^2(N-1)^2}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{2\alpha^{\frac{N-2}{4}}}\right) + \frac{2N-1}{N^2(N-1)^2}\frac{1}{2\alpha^{\frac{N-2}{2}}}.
\end{aligned}$$

Since the second term is always positive, it suffices to ensure that the first term is positive.

$$\begin{aligned}
-2(2N-1) - N(N-1)\log\alpha \geq 0, \\
\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{2\alpha^{\frac{N-2}{4}}} \geq 0,
\end{aligned}$$

This requires the following inequalities:

$$\log \alpha \leq \min_{N\geq 3} \frac{2(1-2N)}{N(N-1)} = \min_{N\geq 3} 2\left(\frac{1}{1-N} - \frac{1}{N}\right) = -\frac{5}{3},$$

$$\alpha^{\frac{N-2}{4}} \leq 2.$$

Therefore, we obtain: when $\alpha \leq \exp(-5/3)$, the lower bound of $K_2 - K_1$ increases with N. □

J THE PROOF OF THE THEOREM 4

Theorem 4. Let K_1, K_2, K_3 and K_4 be the numbers of iterations defined in Proposition 1 and Proposition 2. Assume that $\zeta \geq 1$, $\theta^*_{\max} \geq 1$. If α and η are fixed and do not vary with N and μ, then both $K_2 - K_1$ and $K_4 - K_3$ have a lower bound, and this lower bound decreases as the momentum parameter μ increases.

Proof. For any $N > 2$, we first show that $K_2 - K_1$ decreases as the momentum parameter μ increases. First, recall the definition of $\beta(\mu)$ in Lemma 14,

$$\beta(\mu) = \frac{\tau + \frac{1}{B}}{1 - \frac{\tau+\frac{1}{B}}{\gamma}} \quad \text{where} \quad \tau = \frac{1}{1-\mu}, \quad \gamma = \frac{1+\mu}{2(1-\mu)^3}.$$

we have

$$\begin{aligned} K_2 - K_1 &\geq \frac{10}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{2}{\alpha^{\frac{N-2}{2}}}\right) - \frac{75}{4N^2\eta\left[\tau + \frac{25}{64}\eta\gamma N^2 \zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{2N-2}{N}}} \log\frac{|\theta^*_{\max}|}{\epsilon} \\ &\geq \frac{10}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{2}{\alpha^{\frac{N-2}{2}}}\right) - \frac{75}{4N^2\eta\zeta^{\frac{2N-2}{N}}} \log\frac{|\theta^*_{\max}|}{\epsilon} \end{aligned}$$

Let $\frac{1}{1-\mu} = t$, then $\tau = t$ and $\gamma = (2 - \frac{1}{t})2t^3$, and we have

$$\beta(t) = \frac{t + \frac{1}{B}}{1 - \frac{2(t+1/B)}{t^2(2t-1)}}.$$

Since t increases as μ increases, there exists a constant $\mu_0 \in [0, 1)$ such that for $\mu \in [\mu_0, 1)$ we have $1 - \frac{2(t+1/B)}{t^2(2t-1)} > 0$. It is easy to see that $c(\mu)$ increases as μ increases. Consequently, the lower bound of $K_2 - K_1$ decreases as the momentum parameter increases.

For $K_4 - K_3$, we have

$$\begin{aligned} &\frac{100}{\beta N(N-1)\eta\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) - \frac{25}{N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}} \log\frac{\theta^*_{\min}}{\epsilon} \\ \geq &\frac{100}{\beta N(N-1)\eta\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) - \frac{25}{N^2\eta(\theta^*_{\min})^{\frac{2N-2}{N}}} \log\frac{\theta^*_{\min}}{\epsilon}. \end{aligned}$$

Similarly, the lower bound of $K_4 - K_3$ decreases as the momentum parameter increases. □

K THE PROOF OF THE THEOREM 6

Theorem 6. *Let K_1, K_2, K_3 and K_4 be the numbers of iterations defined in proposition 1 and 2. Suppose $\zeta \geq 1$, $\theta^*_{\max} \geq 1$ and the initialization $\alpha \leq \exp(-\frac{5}{3})$. If α and η are fixed and do not vary with N and μ, then both $K_2 - K_1$ and $K_4 - K_3$ have a tight lower bound that decreases as α increases.*

Proof. For any $N > 2$, we first show that $K_2 - K_1$ increases as the initial value α decreases.

$$\begin{aligned} K_2 - K_1 &= \frac{20}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{N-2}} - \frac{1}{\alpha^{\frac{N-2}{2}}}\right) - \frac{15}{4N(N-2)\eta\zeta\tau\alpha^{N-2}} \\ &\quad - \frac{75}{4N^2\eta\left[\tau + \frac{25}{64}\eta\gamma N^2\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{2N-2}{N}}}\log\frac{|\theta^*_{\max} - \alpha^N|}{\epsilon} \\ &\geq \frac{10}{\beta N(N-1)\eta\zeta}\left(\frac{1}{\alpha^{(N-2)}} - \frac{2}{\alpha^{\frac{N-2}{2}}}\right) - \frac{75}{4N^2\eta\left[\tau + \frac{25}{64}\eta\gamma N^2\zeta^{\frac{2N-2}{N}}\right]\zeta^{\frac{2N-2}{N}}}\log\frac{|\theta^*_{\max}|}{\epsilon}. \end{aligned}$$

Since the second term in the above expression is independent of α, it suffices to ensure that $f(\alpha) = \frac{1}{\alpha^{(N-2)}} - \frac{2}{\alpha^{(N-2)/2}}$ increases as α decreases. Under the restriction $\alpha < 1$, we obtain

$$\begin{aligned} f'(\alpha) &= -\frac{(N-2)}{\alpha^{(N-1)}} + \frac{2(N-2)/2}{\alpha^{N/2}} \\ &= (N-2)\left(\frac{1}{\alpha^{\frac{N}{2}}} - \frac{1}{\alpha^{(N-1)}}\right) = (N-2)\frac{\alpha^{\frac{N-2}{2}} - 1}{\alpha^{(N-1)}} < 0. \end{aligned}$$

For $K_4 - K_3$, we have

$$\frac{100}{\beta N(N-1)\eta\theta^*_{\min}}\left(\frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}\right) - \frac{25}{N^2\eta\left[\tau + \frac{4}{25}\eta\gamma N^2(\theta^*_{\min})^{\frac{2N-2}{N}}\right](\theta^*_{\min})^{\frac{2N-2}{N}}}\log\frac{\theta^*_{\min}}{\epsilon}.$$

It suffices to show that as α decreases, the function $f(\alpha) = \frac{1}{\alpha^{\frac{N-2}{2}}} - \frac{1}{\alpha^{\frac{N-2}{4}}}$ is increasing. Under the general condition $\alpha < 1$, we have:

$$f'(\alpha) = -\frac{(N-2)/2}{\alpha^{\frac{N}{2}}} + \frac{(N-2)/4}{\alpha^{\frac{N+2}{4}}} = \frac{N-2}{2}\left(\frac{1}{2\alpha^{\frac{N+2}{4}}} - \frac{1}{\alpha^{\frac{N}{2}}}\right) = \frac{N-2}{2}\frac{\alpha^{\frac{N-2}{4}} - 2}{2\alpha^{\frac{N}{2}}} < 0.$$

The proof is complete. □

L Additional Experiments

We supplement the experiments mentioned in the main paper.

L.1 Convergence

We illustrate the convergence of the algorithm under different momentum $\mu \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$ and depth N settings (see Figure 6). The results on MNIST are shown in Figure 7.

L.2 Early Stopping

We set different momentum values $\mu \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$ to comprehensively demonstrate the influence of momentum on the early stopping window. The experimental results show that increasing the momentum gradually reduces and shifts forward the entire early stopping window (see Figure 8 and 9).

L.3 Incremental learning

This section demonstrates the phenomenon of incremental learning under different momentum and network depths. Experiments show that increasing depth delays the convergence of each component, thereby widening and delaying the entire early-stopping time window. Conversely, increasing momentum accelerates the convergence of each component, thus narrowing and advancing the early stopping window (see Figure 10).

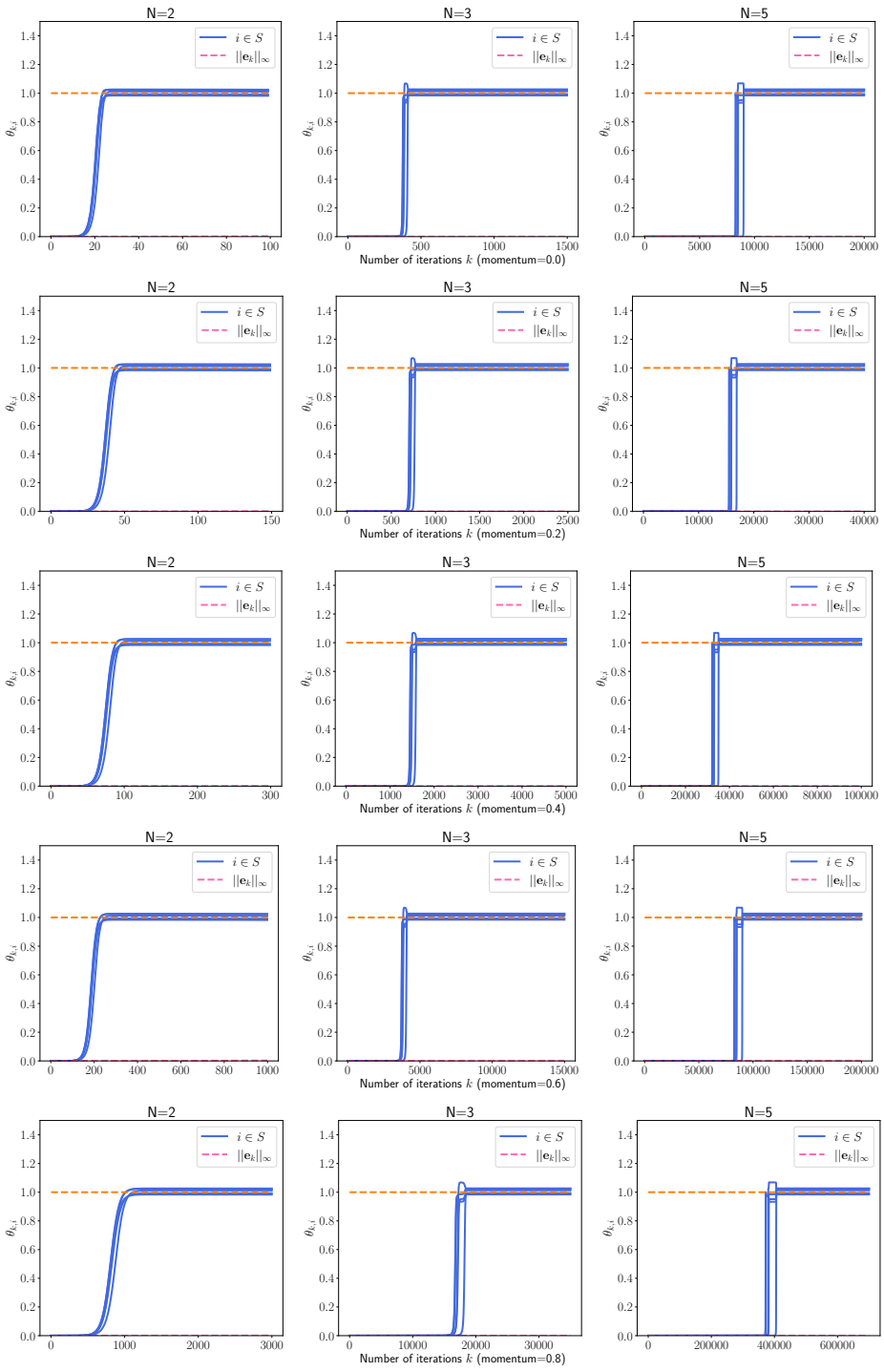
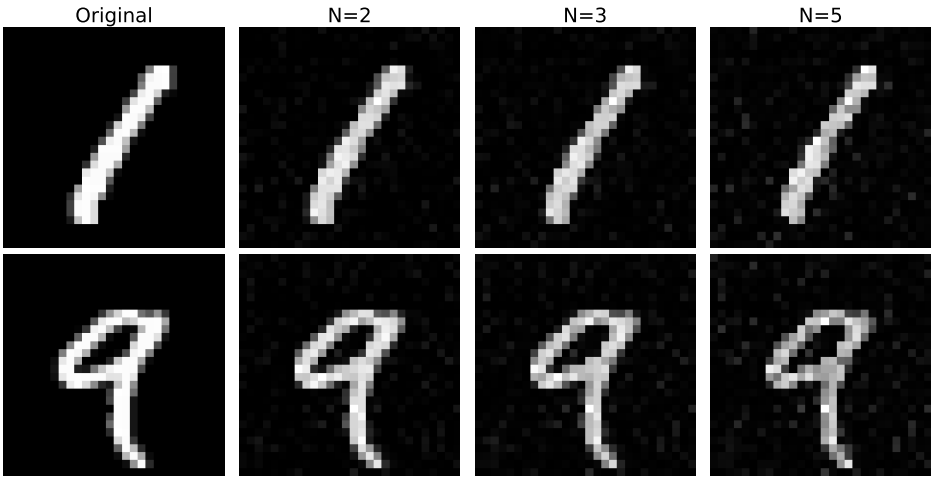


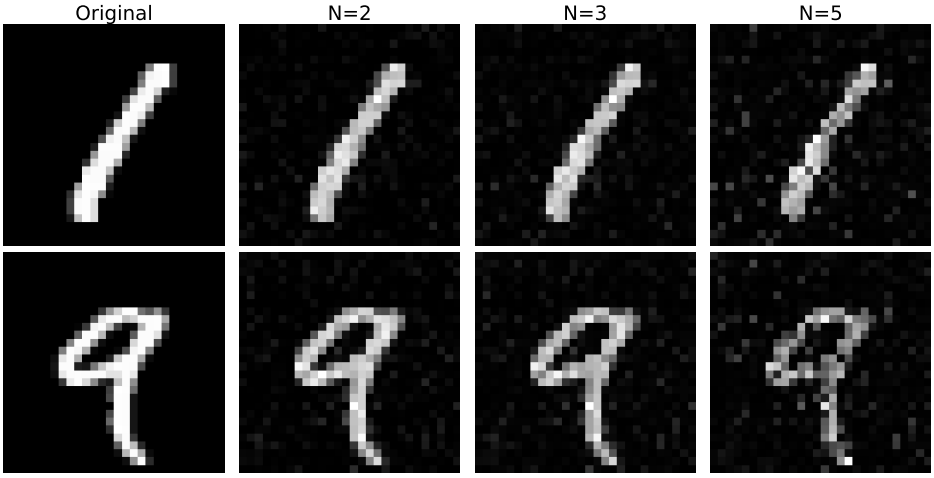
Figure 6: The evolution path of the components of $\boldsymbol{\theta}_k$ with $\alpha = 1e-6$ and $\eta = \frac{1}{N^2\gamma}$.

L.4 KERNEL REGIME

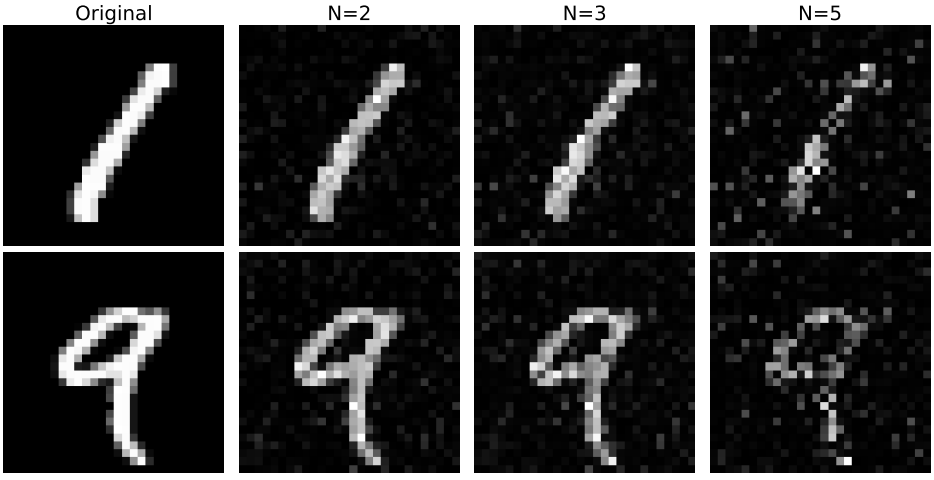
In this section, we empirically investigate the behavior of MGD in both the kernel and rich regimes under different depths N and momentum $\mu \in \{0.0, 0.2, 0.4, 0.6, 0.8, 0.9\}$. The results show that as the momentum increases, MGD requires earlier early stopping to establish a connection with ridge regression (see Figure 11).



(a) $\mu = 0.0$



(b) $\mu = 0.2$



(c) $\mu = 0.4$

Figure 7: Experimental results for different depth and momentum settings. Experimental setup: number of measurements $n = 392$, original image dimension $p = 784$, and the Rademacher sensing matrix is adopted. The learning rate $\eta = \frac{1}{N^2\gamma}$ and iteration $k = 15000$. The results show that larger depth and momentum lead to sparser image recovery.

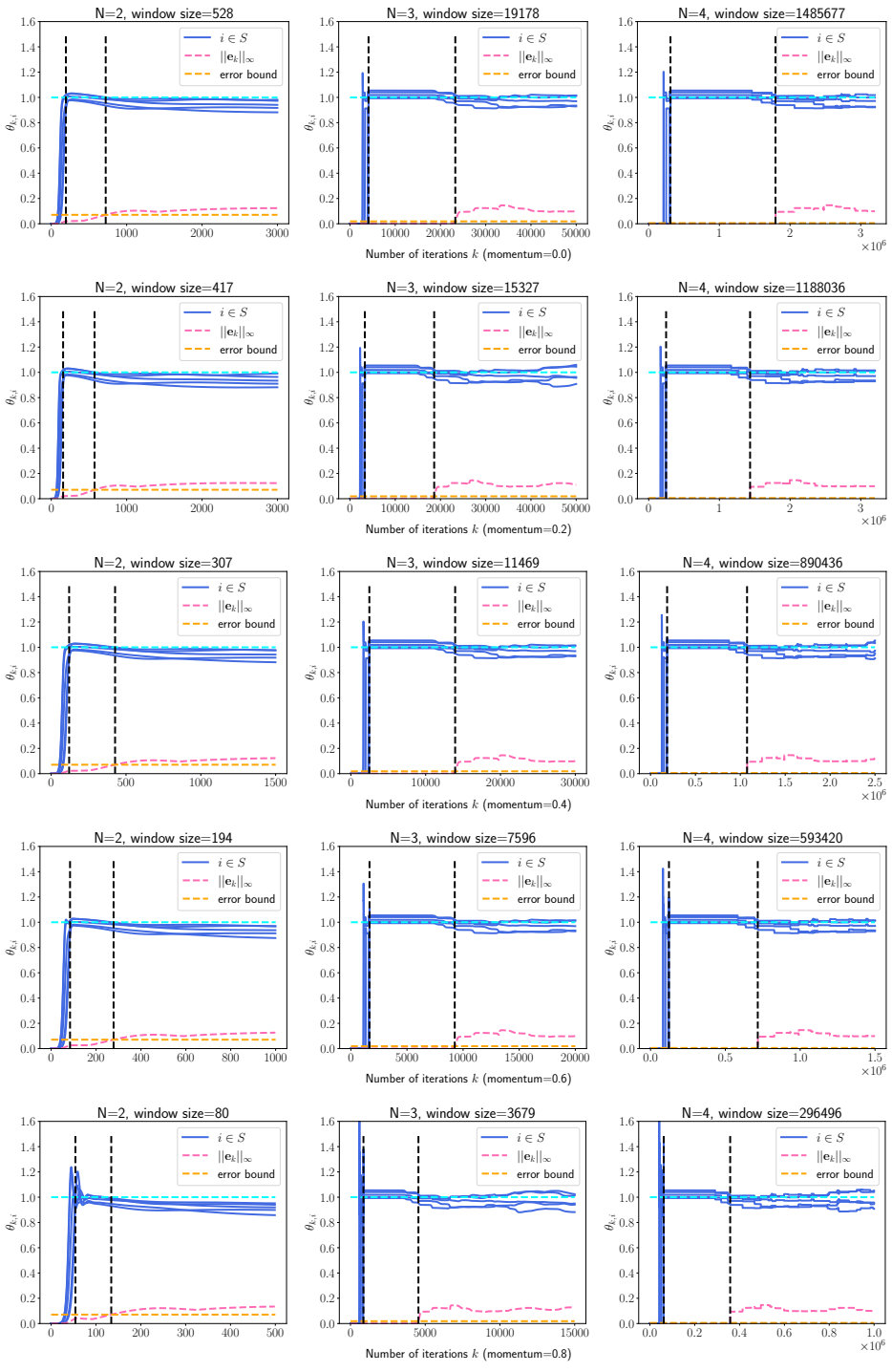


Figure 8: The effect of N and μ on the early stopping window.

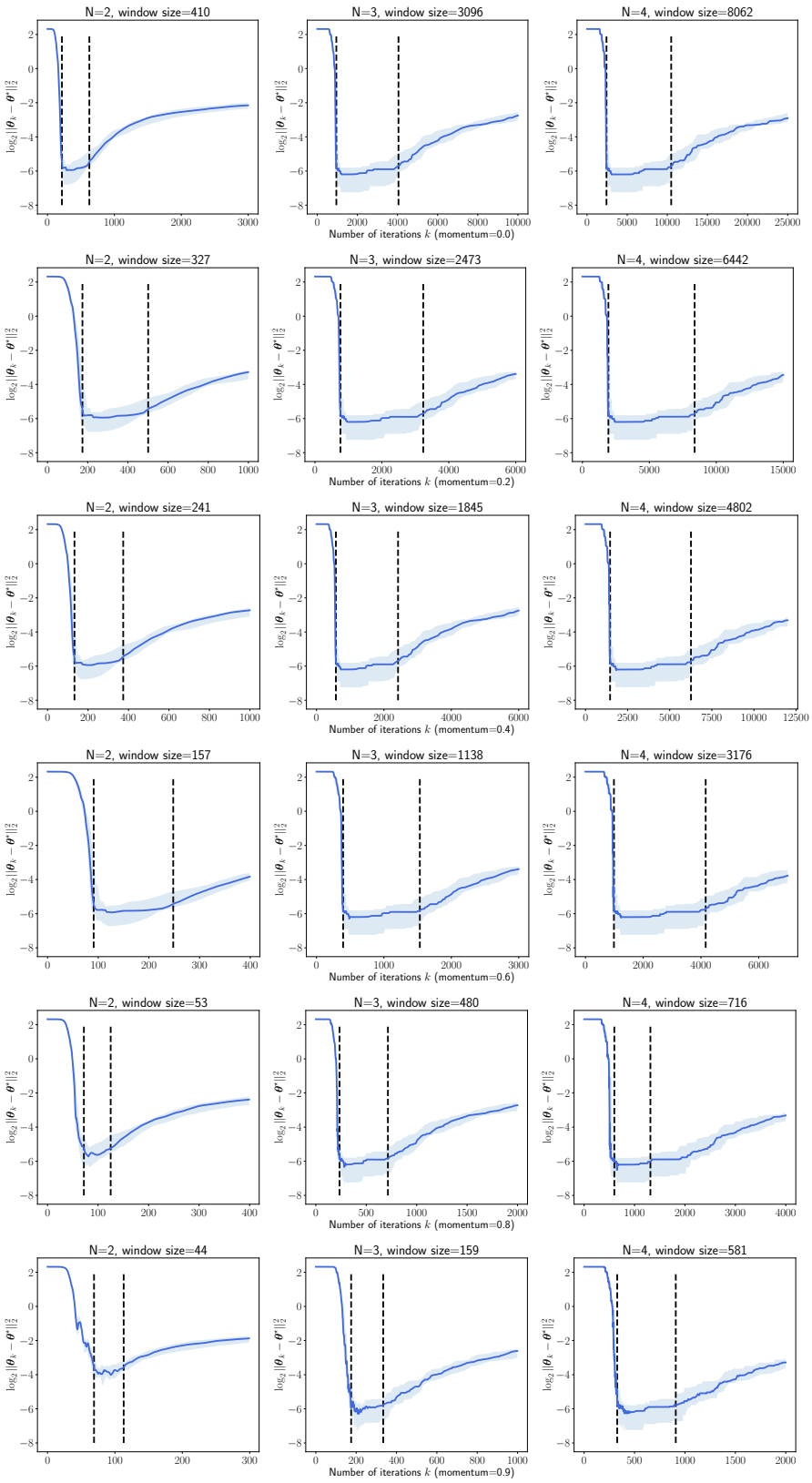


Figure 9: log-ℓ_2 error of $N \in \{2, 3, 4\}$ and $\mu \in \{0.0, 0.2, 0.4, 0.6, 0.8, 0.9\}$.

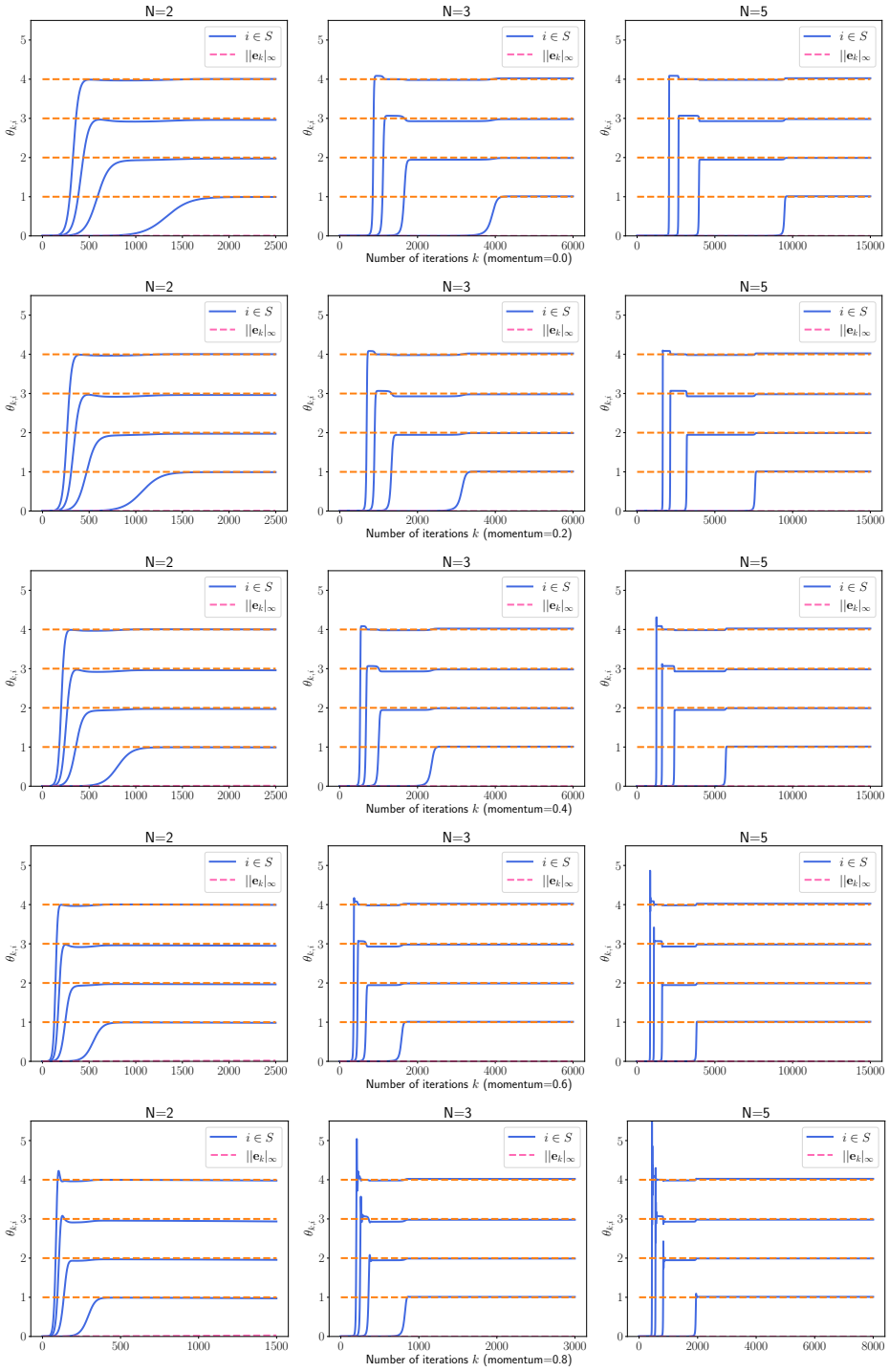


Figure 10: The evolution path of the non-zero component of $\boldsymbol{\theta}^*$ for $N \in \{2, 3, 5\}$ and $\mu \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$. The four nonzero components of $\boldsymbol{\theta}^*$ are set to $\{1, 2, 3, 4\}$, with initialization $\alpha = 1e-4$ and learning rate $\eta = 1e-3$.

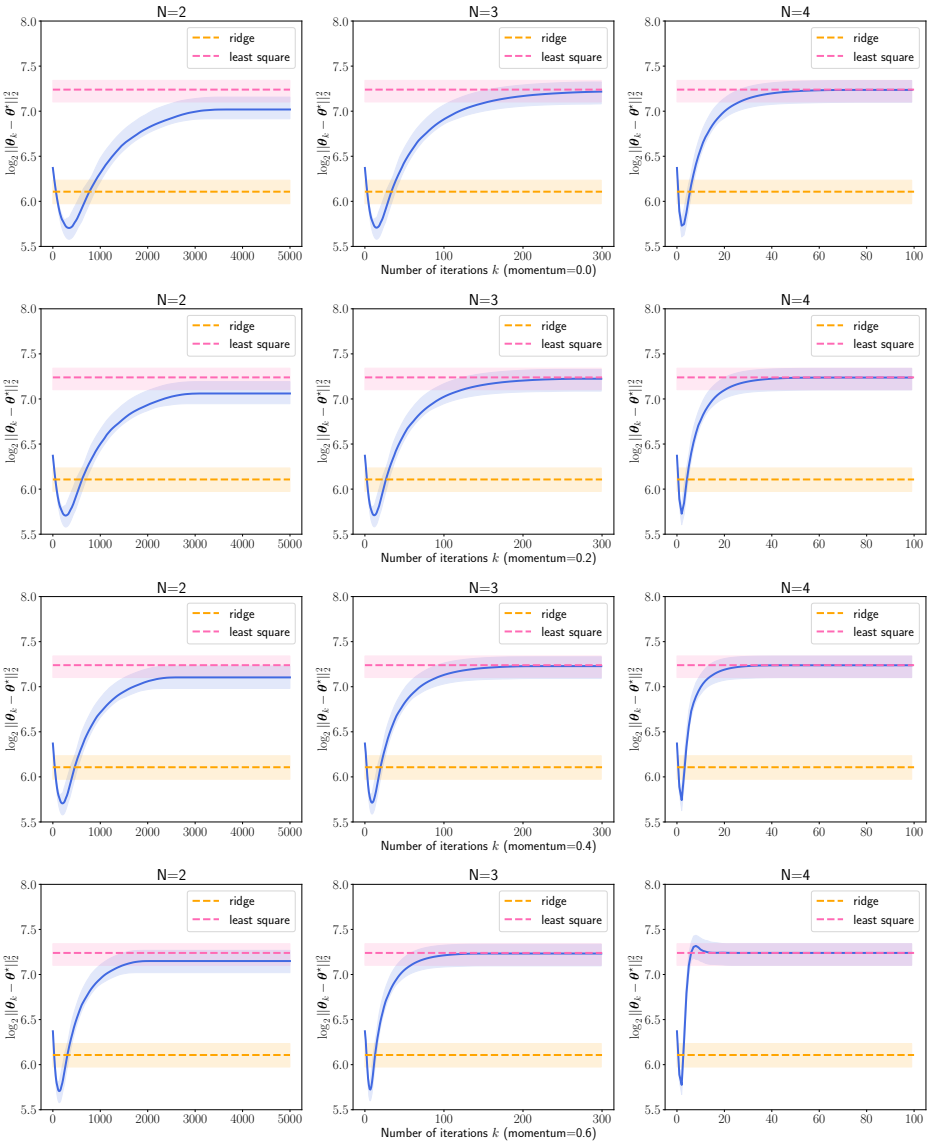


Figure 11: log-ℓ_2 error of $N \in \{2, 3, 4\}$ and $\mu \in \{0.0, 0.2, 0.4, 0.6\}$ for a ridge regression setting.

M LIMITATION

While the current work reveals the crucial role of momentum, its characterization remains mathematically intricate. The analysis of momentum is confined to diagonal networks and does not extend to more complex architectures, such as two-layer ReLU networks. Due to the specific properties of diagonal networks, experimental validation is limited to MNIST and synthetic data.