Learning Provably Improves the Convergence of Gradient Descent

Qingyu Song
Xiamen University
simmonssong96@gmail.com

Wei Lin, Hong Xu
The Chinese University of Hong Kong
wlin23@cse.cuhk.edu.hk, hongxu@cuhk.edu.hk

Abstract

Learn to Optimize (L2O) trains deep neural network based solvers for optimization, achieving success in accelerating convex problems and improving non-convex solutions. However, L2O lacks rigorous theoretical backing for its own training convergence, as existing analyses often use unrealistic assumptions–a gap this work highlights empirically. We bridge this gap by proving the training convergence of L2O models that learn Gradient Descent (GD) hyperparameters for quadratic programming, leveraging the Neural Tangent Kernel (NTK) theory. We propose a deterministic initialization strategy to support our theoretical results and promote stable training over extended optimization horizons by mitigating gradient explosion. Our L2O framework demonstrates over 50% better optimality against GD and superior robustness over state-of-the-art L2O methods on synthetic datasets.

1 Introduction

Learn to optimize (L2O) represents an increasingly influential paradigm for tackling optimization problems [6]. Numerous studies have demonstrated the efficacy of employing learning-based models to achieve superior performance across a spectrum of optimization tasks. These encompass convex problems, exemplified by LASSO [7, 8, 21] and logistic regression [22, 32], and non-convex scenarios such as MIMO sum-rate maximization [33] and network resource allocation [31].

Distinct from black-box approaches [5, 34, 38], which directly derive solutions to optimization problems from a neural network (NN), the so-called "white-box" methodologies are garnering increased attention. This heightened interest stems from their inherent advantages, such as enhanced trustworthiness [13] and theoretical guarantees [32]. A key characteristic of these white-box strategies is the integration of mechanisms to ensure the "controllability" of the generated solutions. For instance, Lv et al. [24] employ a NN to predict the step size for the gradient descent (GD) algorithm, where the inherent structure of GD stabilizes the optimization trajectory. Similarly, Heaton et al. [13] integrate a conventional solver within an L2O framework to act as a safeguard, thereby preventing the learning-based model from producing solutions with extreme violations. This principle of guided or constrained learning has also been extended to the training phase of L2O models [37].

Further, "unrolling" has emerged as a prominent technique within L2O [6], characterized by the strategic replacement of components of conventional optimization algorithms with neural network (NN) blocks [11, 14, 19]. For instance, Liu et al. [22] introduce Math-L2O that imposes architectural constraints on unrolled L2O models by deriving necessary conditions for their convergence. Their analysis revealed that for a L2O model to achieve optimality, its embedded NN must effectively perform a linear combination of input feature vectors, weighted by learnable parameter matrices. Empirical validation demonstrates that the proposed methods exhibit strong generalization capabilities when trained using a coordinate-wise input-to-output strategy. Subsequent research by Song et al. [32] further enhance this generalization performance by reducing the magnitude of input features.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Despite these advancements, to the best of our knowledge, a formal demonstration of the convergence for unrolling-based L2O methods in solving general optimization problems remains elusive. While LISTA-CPSS [7] establishes convergence for the well-known LISTA framework [11], its analysis is based on the assumption that neural network (NN) outputs are confined to a specific subspace, a condition that is often not met in practical implementations. Similarly, while Math-L2O [22] derives necessary conditions for convergence, the mechanisms by which the training process itself can guarantee such convergence are not elucidated. Subsequent analysis by Song et al. [32] investigates the inference-time convergence of Math-L2O. However, this work relies on a stringent training assumption, effectively constraining the L2O model to emulate the behavior of a conventional Gradient Descent (GD) algorithm.

This apparent deficiency in comprehensively demonstrating L2O convergence stems from two fundamental, unresolved technical challenges. First, unrolling-based L2O models [8, 11, 21] represent a specialized class of NN architectures. Despite much progress in understanding the training convergence of general neural networks (NNs), notably through the Neural Tangent Kernel (NTK) theory since 2019 [2, 3, 10, 23, 27, 28], a formal proof of training convergence remains conspicuously absent. Such a proof is an essential precursor to establishing the convergence of the L2O model in its primary task of solving optimization problems. Second, the precise relationship between the training convergence achieved during the L2O model's training phase (i.e., optimizing the NN parameters) and the convergence of the L2O model when applied to the target optimization problem (i.e., finding the optimal solution) is not well understood. For instance, Math-L2O [22] is designed to learn the step size for an underlying GD algorithm. While the problem-solving efficacy of Math-L2O is naturally evaluated based on the progression of GD iterations, its training convergence is measured in terms of training steps (e.g., epochs). These two notions of convergence: one on model parameter optimization and the other on problem-solving iterations, are largely decoupled and operate on fundamentally different scales.

In this work, we present the first rigorous demonstration that an unrolling framework can achieve theoretical convergence in solving optimization problems. Our analysis focuses on the state-of-the-art (SOTA) Math-L2O framework, wherein a NN functions as a recurrent block, iteratively generating hyperparameters for an underlying optimization algorithm. The solution obtained at each iteration, which utilizes these generated hyperparameters, is then incorporated as an input feature for the subsequent iteration [22]. This inherent recurrence imparts RNN-like characteristics to Math-L2O, significantly complicating the analysis of its training convergence. Specifically, the recurrent structure causes the NN to manifest as a high-order polynomial function with respect to its input features [3]. This characteristic poses challenges for establishing tight analytical bounds, potentially leading to looser convergence rates compared to non-recurrent architectures, as highlighted in related NTK analyses for RNNs [3]. Moreover, the Math-L2O architecture introduces an additional layer of complexity: the emergence of high-order polynomial dependencies not only on the input features but also on the learnable parameters themselves. This distinct feature renders the convergence proof for Math-L2O arguably more intricate than those for conventional RNNs, where such parameter-dependent high-order terms are typically less pronounced.

We address the pivotal connection between the NN's training convergence and the ultimate problem-solving convergence of the L2O model. Within the Math-L2O framework, we establish this critical linkage by explicitly demonstrating an alignment between the convergence dynamics exhibited during the NN's training phase and the convergence characteristics of its underlying backbone optimization algorithm. This alignment provides a novel theoretical bridge, ensuring that a successfully trained L2O model translates to effective convergence when applied to optimization tasks. Our contributions are summarized as follows:

1. We provide a formal proof that the Math-L2O training framework substantially enhances the convergence performance of its underlying backbone algorithms. This is achieved by rigorously establishing an explicit alignment between the convergence rates of the training process and the iterative steps of the backbone algorithm.

2. We establish the first linear convergence rate for Math-L2O training. Inspired by [27], we employ a NN architecture with a single wide layer and utilize NTK to prove the boundedness of NN outputs, gradients, and the training loss function within the Math-L2O framework.

3. We introduce a novel deterministic parameter initialization scheme, coupled with a specific learning rate configuration strategy. This combined approach is proven to guarantee the training convergence of the Math-L2O model across all iterations.
4. We empirically validate our theoretical findings through comprehensive experiments. The results showcase significant performance advantages, including up to a 50% improvement in solution optimality over the standard GD algorithm post-training, and superior robustness compared to SOTA L2O models and the Adam optimizer [9]. Furthermore, ablation studies empirically confirm the practical efficacy and individual contributions of our proposed theorems.

2 Preliminary

This section first defines the optimization problem objective and the L2O framework. The L2O training loss is then formulated based on these definitions. Then, the NN's computational graph is employed to detail the forward pass and the derivation of parameter gradients.

2.1 Definitions

Let $d > b$, suppose $x \in \mathbb{R}^{d\times 1}$, $y \in \mathbb{R}^{b\times 1}$, and $\mathbf{M} \in \mathbb{R}^{b\times d}$, we define the optimization objective as:

$$\min_{x\in\mathbb{R}^d} f(x) = \tfrac{1}{2}\|\mathbf{M}x - y\|_2^2. \tag{1}$$

This objective function is commonly selected for convergence analysis [4]. The least-squares problem, a frequent subject in NN convergence studies [2, 3, 10, 20, 27], is a specific instance of the minimization in Equation (1) where $d = b$ and $\mathbf{M}_i = \mathbf{I}$.

We assume f to be β-smooth, such that $\|\mathbf{M}^\top\mathbf{M}\|_2 \leq \beta$, and $\mathbf{M}$ to possess full row rank, with $\lambda_{\min}(\mathbf{M}\mathbf{M}^\top) = \beta_0 > 0$. This setting often favors numerical algorithms (e.g., GD) over analytical solutions due to computational complexity. GD's $\mathbf{O}(bd)$ complexity is typically lower than the $\mathbf{O}(b^3)$ of analytical methods involving costly matrix inversions. The loss function is then defined as the sum of N objectives specified in Equation (1):

$$F(X) = \tfrac{1}{2}\|\mathbf{M}X - Y\|_2^2, \tag{2}$$

where F, $\mathbf{M} \in \mathbb{R}^{Nb\times Nd}$, $X \in \mathbb{R}^{Nd\times 1}$, and $Y \in \mathbb{R}^{Nb\times 1}$ represent the concatenated objectives, parameters, variables, and labels, respectively, from N optimization problems (see Appendix A.1 for details). F is also β-smooth, given that $\|\mathbf{M}^T\mathbf{M}\|_2 \leq \max_{i=1,\ldots,N}\{\|\mathbf{M}_i^T\mathbf{M}_i\|_2\} = \beta$.

Learn to Optimize (L2O). Let g_W denote an L-layer NN with parameters $W = \{W_1, \ldots, W_L\}$. For each step $t \in [T]$ in solving problem Equation (1), the Math-L2O model, following [22], is defined as $g_W(X_{t-1}, \nabla F(X_{t-1}))$. The inputs to the NN are the current variable X_{t-1} and its gradient $\nabla F(X_{t-1})$. Denoting the Hadamard product by $\odot$, the T-step iterative update from an initial X_0 is:

$$X_t = X_{t-1} - \tfrac{1}{\beta} P_t \odot \nabla F(X_{t-1}), \tag{3}$$

$P_t = g_W(X_{t-1}, \nabla F(X_{t-1}))$ is a vector whose entries represent the learned step sizes.

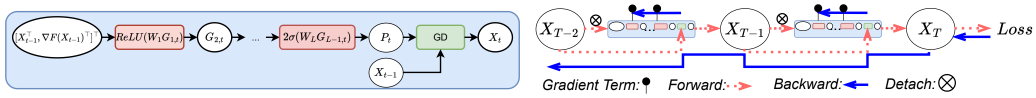
The neural network g_W is structured layer-wise. It employs a coordinate-wise architecture, processing each input dimension independently, recognized for its robustness in L2O applications [22, 32]. For layer $\ell \in [L]$ with parameters $W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ (where $n_L = 1$ for the output layer), the output $G_{\ell,t}$ at step t, utilizing ReLU (ReLU) [1] and Sigmoid (σ) [26] activations, is defined as:

$$G_{\ell,t} = \begin{cases} [X_{t-1}, \nabla F(X_{t-1})]^\top & \ell = 0, \\ \mathrm{ReLU}(W_\ell G_{\ell-1,t}) & \ell \in [L-1], \\ P_t = 2\sigma(W_L G_{L-1,t})^\top & \ell = L. \end{cases} \tag{4}$$

2.2 Layer-Wise Derivative of NN's Parameters

Let k represent a training iteration for loss Equation (2) minimization, distinct from an optimization step t for solving objective Equation (1). The computational graph in Figure 1 illustrates the

Math-L2O forward and backward operations, which parallel those of Recurrent Neural Networks (RNNs) [12]. Figure 1a details the NN block (see Equation (4)). Figure 1b depicts the overall process: the block takes an input solution, performs T internal optimization steps to produce an updated solution (red dashed arrows), and each training iteration k triggers a full backward pass (blue bold lines). As per [22], the gradient flow from the input features to the NN block is detached.



(a) NN Block

(b) Forward and Backward Processes

Figure 1: Computational Graph of Math-L2O

The derivative of an objective F with respect to (w.r.t.) the parameters W_ℓ of layer ℓ is determined via the computational graph, paralleling Back-Propagation-Through-Time (BPTT) for RNNs [25]:

$$\frac{\partial F}{\partial W_\ell} = \frac{\partial F(X_T)}{\partial X_T}\Big(\sum_{t=1}^{T}\big(\prod_{j=T}^{t+1}\frac{\partial X_j}{\partial X_{j-1}}\big)\frac{\partial X_t}{\partial P_t}\frac{\partial P_t}{\partial W_\ell}\Big). \tag{5}$$

The summation aggregates gradients across T optimization steps. $\prod_{j=T}^{t+1}(\partial X_j/\partial X_{j-1})$ represents the chain rule application from the final output X_T to an intermediate state X_t.

Moreover, we derive two key gradients, instrumental for establishing the theoretical results in the ensuing section. Following Definition 2.2 in [2], the gradient of the ReLU is represented by a diagonal matrix $\mathbf{D}_\ell^t$, where its i-th diagonal element is $[\mathbf{D}_\ell^t]_{i,i} := \mathbf{1}_{(W_\ell G_{\ell-1,t})_i \geq 0}$ for $i \in [n_\ell]$. Let $\Gamma_t := \mathbf{M}^\top(\mathbf{M}X_t - Y)$ and $\Xi_\ell := (\mathbf{I}_d \otimes W_L)(\prod_{j=L-1}^{\ell+1}\mathbf{D}_{j,t}(\mathbf{I}_d \otimes W_j))\mathbf{I}_{n_\ell}$. Defining $\mathcal{D}(\cdot)$ as the operator that constructs a diagonal matrix from a vector, the gradients for an inner layer W_ℓ ($\ell < L$) and the final layer W_L are:

$$\frac{\partial F}{\partial W_\ell} = -\frac{1}{\beta}\Gamma_T^\top\sum_{t=1}^{T}\big(\prod_{j=T}^{t+1}(\mathbf{I}_d - \tfrac{1}{\beta}\mathbf{M}^\top\mathbf{M}\mathcal{D}(P_j))\big)\mathcal{D}(\Gamma_t)\mathcal{D}\big(P_t \odot (1-P_t/2)\big)\Xi_\ell \otimes G_{\ell-1,t}^\top, \tag{6}$$

$$\frac{\partial F}{\partial W_L} = -\frac{1}{\beta}\Gamma_T^\top\sum_{t=1}^{T}\big(\prod_{j=T}^{t+1}(\mathbf{I}_d - \tfrac{1}{\beta}\mathcal{D}(P_j)\mathbf{M}^\top\mathbf{M})\big)\mathcal{D}(\Gamma_T)\mathcal{D}\big(P_t \odot (1-P_t/2)\big)G_{L-1,t}^\top, \tag{7}$$

where $\otimes$ denotes the Kronecker product. Equation (7) (for W_L) differs from Equation (6) (for W_ℓ) in its final terms: $G_{L-1,t}^\top$ replaces $\Xi_L \otimes G_{\ell-1,t}^\top$. This simplification arises as W_L is the terminal layer, and $G_{L-1,t}$ is its direct input from layer $L-1$. Thus, its gradient calculation does not involve a subsequent layer propagation factor analogous to Ξ_L.

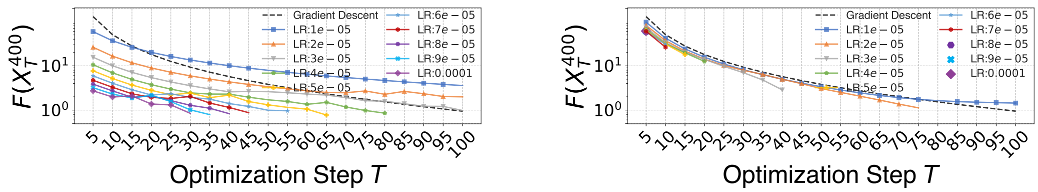
3 Convergence of L2O: Improved Convergence from Training

This section rigorously substantiates the convergence of the L2O framework, Math-L2O. We first expose theoretical and numerical instabilities prevalent in current SOTA L2O methods. Then, we demonstrate Math-L2O's accelerated training convergence compared to GD and then present a formal methodology to establish its convergence.

3.1 Limitations Analysis of Existing SOTA L2O Frameworks

We analyze limitations in the convergence guarantees of two SOTA L2O frameworks: LISTA-CPSS [7] and Math-L2O [22]. LISTA-CPSS [7] constructively proves that its predecessor, LISTA [11], can attain a linear convergence rate. However, this theoretical guarantee is contingent upon several stringent conditions. Math-L2O [22] proposes an L2O framework derived from the GD algorithm, incorporating necessary conditions for convergence. Both frameworks employ sequential solution updates and utilize BPTT for parameter optimization.

Initially, we assess training loss across varying optimization steps. This is pertinent due to the well-documented issue of gradient explosion of BPTT arising from long-term gradient accumulation [18]. Both models are trained on 10 randomly sampled optimization problems for 400 epochs. Figure 2 depicts training losses (y-axis) against optimization steps (x-axis) for several learning rates (distinguished by line color). Data points exhibiting numerical overflow (indicative of gradient explosion at first training iteration) are excluded, resulting in plot lines terminating before 100 steps for affected configurations. The results demonstrate that both frameworks suffer from poor convergence at low learning rates (LRs) and training instability at high LRs.



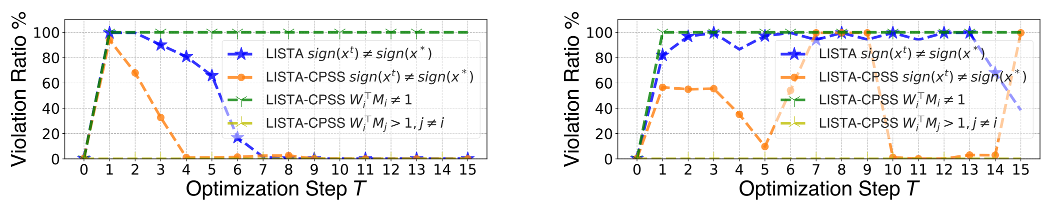
(a) LISTA-CPSS [7]

(b) Math-L2O [22]

Figure 2: Training Loss of SOTA L2O Frameworks

Further, we examine the convergence conditions outlined for LISTA-CPSS [7], illustrating their propensity for violation during typical training procedures. The first condition mandates asymptotic sign consistency between iterates X_t and the solution X^*, requiring $\text{sign}(X_t) = \text{sign}(X^*)$ for all t. The second condition imposes constraints on the columns of the learned parameter matrix $\mathbf{W}$ relative to the columns of the objective coefficient matrix $\mathbf{M}$. Specifically, denoting column indices by i and j, it necessitates that $\mathbf{W}_i^\top \mathbf{M}_i = 1$ and $\mathbf{W}_i^\top \mathbf{M}_j > 1$ for all $j \neq i$.

Following the experimental design in [22], we quantify the violation percentage of the aforementioned conditions during inference. Results are in Figure 3. We consider two configurations: (i) shared parameters W across iterations (Figure 3a), and (ii) unique parameters W_t per step t (Figure 3b). Both scenarios reveal that the specified conditions are frequently violated post-training. For instance, in the shared W case (Figure 3a), while the conditions hold in later steps, substantial violations occur in early steps. This divergence contradicts the convergence rate analysis presented in [7].



(a) Shared W

(b) Unique W_t per Step

Figure 3: Violation Ratio of LISTA-CPSS Conditions During Inference

The preceding observations highlight that training is indispensable for L2O convergence analysis. Three fundamental questions arise in L2O: (i) *What is the impact of training on convergence?* (ii) *How can training be incorporated into the convergence analysis framework?* (iii) *What mechanisms ensure a stable training process?* We propose a concise approach to address these questions, establishing a direct alignment between the training's convergence rate and an existing algorithm's rate.

3.2 Align Convergence of L2O Training to Backbone Algorithm

First, we introduce a general convergence analysis framework. Let X^* be the optimal solution, r_t an iteration-dependent rate term, and $C(X_0)$ a constant dependent on the initial point X_0 (and X^*), the convergence rate of an algorithm (whether learned or classical) for minimizing an objective $F(X)$ (e.g., the objective in Equation (1) or the loss in Equation (2)) is often formulated as: $F(X_t) \leq r_t C(X_0)$. For example, standard GD has a rate of $F(X_t) \leq \frac{\beta}{t}\|X_0 - X^*\|_2^2$ [4].

The performance of L2O models, stabilized via training, is typically assessed after T iterations [22, 32]. We formulate the L2O training convergence rate w.r.t. training iteration k as:

$$F(X_T^k) \leq r_k C(X_T^0), \quad \text{where } X_T^0 = \text{L2O}_W(X_0), \tag{8}$$

with X_T^0 being the initial solution from the L2O model and C a constant. Based on the proof in [36], For the non-learning GD algorithm, its convergence rate, corresponding to the initial L2O state, is:

$$F(X_T^0) \leq \frac{\beta}{T}\|X_T^0 - X^*\|_2^2. \tag{9}$$

Given the independence of training iteration k and optimization step T, we align the LHS of Equation (8) with the RHS of Equation (9) by setting $C(X_T^0) = F(X_T^0)$. This yields the combined training convergence rate:

$$F(X_T^k) \le r_k \frac{\beta}{T} \|X_T^0 - X^*\|_2^2. \tag{10}$$

Here, the LHS represents the objective value after k training iterations, while the RHS is a constant term dependent on the initial point X_0. W.r.t. T, Equation (10) demonstrates a sub-linear convergence rate of at least $\mathcal{O}(1/T^2)$. The rate indicates that integrating L2O with an existing algorithm via training can enhance its convergence. Such integration is achieved by the Math-L2O framework [22], which utilizes a NN to learn hyperparameters for non-learning algorithms (e.g., step size for GD, step size and momentum for Nesterov Accelerated Gradient [4]).

Further, we construct the Math-L2O training rate r_k (see Equation (8)). Section 4 establishes its linear convergence. Subsequently, Section 5 proposes a deterministic initialization strategy to ensure the alignment ($C(X_T^0) = F(X_T^0)$) and uphold the theoretical conditions for this linear rate.

4 Math-L2O Training Linearly Converges

In this section, we establish the linear convergence rate for training a Math-L2O model employing an over-parameterized NN, w.r.t. the loss defined in Equation (2). By training the NN (Equation (4)) using GD, we establish its linear convergence rate via NTK theory. Classical NTK theory [15] requires infinite NN width to maintain a non-singular kernel matrix, which facilitates a gradient lower bound akin to the Polyak-Lojasiewicz condition [27, 30]. Applying the relaxation from [27] and the rigorous NN formalizations (Section 2), we demonstrate that an NN width of $\mathcal{O}(Nd)$ is sufficient.

To derive the rate, we first introduce an lemma bounding Math-L2O's gradients. Then, we prove that appropriate initialization leads to deterministic loss minimization in initial training iteration. After that, we develop a strategy to maintain this property throughout training, thereby ensuring convergence. This approach culminates in a linear convergence rate for an $\mathcal{O}(Nd)$-width NN. The main results are summarized herein, with detailed proofs deferred to Appendix A.4 and Appendix A.5.

4.1 Bound Outputs of Math-L2O

Let $\alpha_0 := \sigma_{\min}(G_{L-1,T}^0)$, $\bar{\lambda}_\ell \in \mathbb{R}^+$ be some constants. Let $C_\ell > 0$ for $\ell \in [L]$ be a sequence of positive numbers. For $t, j \in [T]$, we define the following quantities:

$$\begin{aligned}
&\bar{\lambda}_\ell = \|W_\ell^0\|_2 + C_\ell, \Theta_L = \textstyle\prod_{\ell=1}^L \bar{\lambda}_\ell, \Phi_j = \|X_0\|_2 + \frac{2j-1}{\beta}\|\mathbf{M}^\top Y\|_2, \\
&\Lambda_j = (1+\beta)\|X_0\|_2^2 + \tfrac{(4j-3)(1+\beta)+\beta}{\beta}\|X_0\|_2\|\mathbf{M}^\top Y\|_2 + \tfrac{(2j-1)(\beta(2j-1)+(2j-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2, \\
&S_{\Lambda,T} = \textstyle\sum_{t=1}^T \Lambda_t, \qquad \delta_1^t = \sum_{s=1}^t \left(\prod_{j=s+1}^t (1 + \frac{1+\beta}{2}\Theta_L\Phi_j)\right)\Lambda_s, \\
&S_{\bar{\lambda},L} = \textstyle\sum_{\ell=1}^L \bar{\lambda}_\ell^{-2}, \qquad \delta_2 = \sum_{s=1}^{T-1}\left(\prod_{j=s+1}^{T-1}(1 + \frac{1+\beta}{2}\Theta_L\Phi_j)\right)\Lambda_s, \\
&\zeta_1 = \sqrt{\beta}\|X_0\|_2 + (2T+1)\|Y\|_2, \quad \delta_3 = (1+\beta)\|X_0\|_2 + \left(2T - 1 + \tfrac{2T-2}{\beta}\right)\|\mathbf{M}^\top Y\|_2, \\
&\zeta_2 = \|X_0\|_2 + \tfrac{2T-2}{\beta}\|\mathbf{M}^\top Y\|_2, \qquad \delta_4 = \sigma(\delta_3\Theta_L)(1 - \sigma(\delta_3\Theta_L)),
\end{aligned} \tag{11}$$

where X_0 denotes the initial point, and $\mathbf{M}$ (parameter matrix) and Y (labels) are input features from Equation (2). The defined quantities are positive under the conditions $j \ge 1$ and $\bar{\lambda}_\ell > 0$.

First, we derive a bound for the training gradients by considering them as perturbations from initialization. This bound relates the gradient magnitude to the objective function in Equation (2), as detailed in the following lemma. Despite the derivative for inner layers (Equation (6)) containing an additional term compared to that of the last layer (Equation (7)), a uniform bound as stated applies. The proof is provided in Appendix A.4.4.

Lemma 4.1. *Assuming* $\max(\|W_\ell^{k+1}\|_2, \|W_\ell^k\|_2) \le \bar{\lambda}_\ell$ *for* $\ell \in [L]$*, for any training iteration k, the gradient of the* ℓ*-th layer parameters* W_ℓ^k *is bounded by:* $\left\|\frac{\partial F}{\partial W_\ell^k}\right\|_2 \le \frac{\sqrt{\beta}\Theta_L S_{\Lambda,T}}{2\bar{\lambda}_\ell}\|\mathbf{M}X_T^k - Y\|_2$.

Building upon Lemmas 4.1 and A.6 and auxiliary results (see Appendix A.4), we now analyze the dynamics of the final solution X_T w.r.t. parameter updates during training. The subsequent lemma establishes a rigorous formulation for the fluctuation of X_T in response to changes in parameters between adjacent training iterations. This result demonstrates that Math-L2O, viewed as a function of its learnable parameters, exhibits semi-smoothness, aligning with findings for ReLU-Nets in [27]. The proof is provided in Appendix A.4.3.

The semi-smoothness of the Math-L2O NN is preserved despite its recurrent operations. The coefficient associated with $\|W_\ell^{k+1} - W_\ell^k\|_2$ exhibits $\mathcal{O}(e^{LT})$ scaling, where e is an initialization parameter detailed in Section 5. This represents a looser bound compared to that for ReLU-Nets [27], which is a consequence of Math-L2O's greater architectural complexity, specifically the T-fold execution of an L-layer NN block (see Equation (7)). However, this scaling behavior is consistent with observations for other deep architectures [2].

Lemma 4.2. *For any training iteration k, assume there exist constants $\bar{\lambda}_\ell \in \mathbb{R}^+$ for $\ell \in [L]$ such that $\max_{k' \in \{k,k+1\}} \|W_\ell^{k'}\|_2 \leq \bar{\lambda}_\ell$. Let X_t^{k+1} and X_t^k be outputs of the Math-L2O (defined in Equations* (3) *and* (4)*) corresponding to parameters $W^{k+1} = \{W_\ell^{k+1}\}_{\ell=1}^L$ and $W^k = \{W_\ell^k\}_{\ell=1}^L$, respectively. Then, Math-L2O exhibits the following semi-smoothness property:*

$$\|X_t^{k+1} - X_t^k\|_2 \leq \tfrac{1}{2} \textstyle\sum_{s=1}^{t-1} \big(\prod_{j=s+1}^{t} (1 + (1+\beta)/2\Theta_L \Phi_j) \big) \Lambda_s \Theta_L \big(\sum_{\ell=1}^{L} \bar{\lambda}_\ell^{-1} \|W_\ell^{k+1} - W_\ell^k\|_2 \big).$$

Lemma 4.2 demonstrates that Math-L2O solutions exhibit a bounded response to perturbations in its NN parameters. This finding, in conjunction with Lemma 4.1, facilitates a more nuanced analysis of the loss dynamics. Further, judicious selection of learning rates enables control over the evolution of NN parameters. Such control is instrumental in bounding the constant quantities from these lemmas, thereby establishing the desired convergence rate presented in the subsequent theorem.

4.2 Linear Training Convergence Rate of Math-L2O

Leveraging the bounds on Math-L2O's output (Lemma A.6) and its gradient (Lemma 4.1), the following theorem establishes the linear convergence rate for training the Math-L2O model. The proof is provided in Appendix A.5.

Theorem 4.3. *Consider the NN defined in Equation* (4)*, using quantities from Equation* (11)*, suppose the following conditions hold at initialization:*

$$\alpha_0 \geq 8(1+\beta)\zeta_2, \quad (12a) \quad \alpha_0^2 \geq \tfrac{\beta^3}{4\beta_0^2} \delta_4^{-2} \big(-\tfrac{1}{2} \Theta_{L-1}^2 \Lambda_T S_{\Lambda,T-1} + \Theta_L^2 (\Lambda_T + \delta_2) S_{\bar{\lambda},L} S_{\Lambda,T} \big). \quad (12b)$$

$$\alpha_0^2 \geq \max_{\ell \in [L]} \tfrac{\Theta_L}{C_\ell \bar{\lambda}_\ell} \tfrac{\beta^2 \sqrt{\beta}}{8\beta_0^2} \delta_4^{-2} \zeta_1 S_{\Lambda,T}, \quad (12c) \quad \alpha_0^3 \geq \tfrac{(1+\beta)\beta^2\sqrt{\beta}}{2\beta_0^2} \delta_4^{-2} \Theta_L \Theta_{L-1} \zeta_1 \zeta_2 S_{\bar{\lambda},L} S_{\Lambda,T}, \quad (12d)$$

Let the learning rate η satisfy:

$$\eta < \tfrac{8}{\beta} (\delta_2 + \Lambda_T) \big(\delta_2 + \Theta_L S_{\Lambda,T} S_{\bar{\lambda},L} \big)^{-1} S_{\Lambda,T}^{-2}, \quad (13a) \qquad \eta < \tfrac{1}{4} \tfrac{\beta^2}{\beta_0^2} \delta_4^{-2} \alpha_0^{-2}. \quad (13b)$$

Then, for weights $W^k = \{W_\ell^k\}_{\ell=1}^L$ at training iteration k, the loss function $F(W^k)$ converges linearly to a global minimum:

$$F(W^k) \leq \big(1 - 4\eta \tfrac{\beta_0^2}{\beta^2} \delta_4 \alpha_0^2 \big)^k F(W^0).$$

The conditions specified in Equation (12) impose additional lower bounds on α_0, the minimal singular value of the $(L-1)$-th layer's inner output. The bounds stipulated in Equations (12b) to (12d) are influenced by both the network depth L and the number of gradient descent (GD) iterations T. In contrast, the constraint in Equation (12a) primarily depends on T. An initialization strategy ensuring these conditions are met is proposed in Section 5.

5 Deterministic Initialization

This section introduces an initialization strategy ensuring the alignment between Math-L2O and GD (see Section 3) while also satisfying the conditions presented in Section 4. The proposed initialization strategy first establishes Math-L2O to operate as a standard GD algorithm, and then guarantees the uniform convergence of Math-L2O throughout subsequent training iterations.

5.1 Initialization for Alignment

Our initialization follows the ReLU-Net scheme [27], with $C_\ell = 1$ for $\ell \in [L]$ and parameters $\theta_0 = \{W_1^0, \ldots, W_{L-1}^0, W_L^0 = \mathbf{0}\}$. The specific initialization $W_L^0 = \mathbf{0}$, combined with the 2σ activation detailed in Equation (4), results in $P_T = \mathbf{I}$. Consequently, the learning proceeds with a uniform step size of $1/\beta$, emulating standard GD and its typical sub-linear convergence rate [36]. Moreover, this zero-initialization of W_L^0 ensures that initial gradients for the inner layers are null (as shown in Equation (6)), which serves to mitigate gradient explosion.

To satisfy the condition $\alpha_0 > 0$ (cf. Theorem 4.3) for the initial weight matrices $\{W_k^0\}_{k=1}^{L-1}$, these are drawn from a standard Gaussian distribution. This approach generally ensures full row rank for fat matrices (more columns than rows) [35]. Each matrix W_k^0 then undergoes QR decomposition. Non-negativity is subsequently enforced upon the elements of the resulting upper triangular factor (e.g., via its element-wise absolute value, achieved in PyTorch using its `sign` function).

5.2 Enhancing Singular Values for Linear Convergence of Training

Motivated by properties of minimal singular values in ReLU-Nets identified in [27], we analyze the order-gap for α_0 between the left-hand side (LHS) and right-hand side (RHS) of the inequalities in Equation (12). To satisfy these inequalities, we propose increasing α_0. This is achieved by applying a constant *expansion coefficient* $e \geq 1$ to the initial NN parameters $\{W_1^0, \dots, W_{L-1}^0\}$, transforming them to $\{eW_1^0, \dots, eW_{L-1}^0\}$. This parameter expansion scales the minimal singular value α_0 to $e^{L-1}\alpha_0$, reflecting the cumulative impact across $L-1$ layers. However, other terms on the RHS of Equation (12) also depend on e. We then establish four lemmas to demonstrate that the conditions for linear convergence, as specified in Theorem 4.3, are met for an appropriately chosen value of e.

First, we set the initial point to the origin, $X_0 = \mathbf{0}$, a choice mostly adopted in L2O literature [22, 32]. Then, with $C_\ell = 1$ for $\ell \in [L]$, we present four lemmas demonstrating that the conditions for linear convergence (see Theorem 4.3) are satisfied for an appropriately chosen constant e. The lemmas indicate that a larger e is required as the number of optimization steps (T) increases. Specifically, Lemma 5.2 establishes that e scales exponentially with T. Conversely, increasing the network depth (L) alleviates the need for a large e. The proofs are provided in Appendix B.

Lemma 5.1. *Assuming $X_0 = \mathbf{0}$, if $e = \mathbf{\Omega}(T^{\frac{1}{L-1}})$, then the inequality Equation* (12a) *holds.*

Lemma 5.2. *If $e = \mathbf{\Omega}(T^{\frac{3T+6}{TL-T-4L+6}})$, then the inequality Equation* (12b) *holds.*

Lemma 5.3. *Assuming $X_0 = \mathbf{0}$, if $e = \mathbf{\Omega}(T^{\frac{4}{L-1}})$, then the inequality Equation* (12c) *holds.*

Lemma 5.4. *Assuming $X_0 = \mathbf{0}$, if $e = \mathbf{\Omega}(T^{\frac{5}{L-1}} L^{\frac{1}{L-1}})$, then the inequality Equation* (12d) *holds.*

6 Empirical Evaluation

This section presents an empirical evaluation of the framework proposed in Section 3 and the theoretical results from Section 4. Experiments are conducted using Python 3.9 and PyTorch 1.12.0 on an Ubuntu 20.04 system equipped with 128GB of RAM and two NVIDIA RTX 3090 GPUs.

Data Generation. Due to GPU memory constraints, vectors $X \in \mathbb{R}^{5120\times 1}$ and $Y \in \mathbb{R}^{4000\times 1}$ for Equation (2) are generated by sampling from a standard Gaussian distribution. These represent ten problem instances with respective dimensional components of 512 (for X) and 400 (for Y). Following Liu et al. [22]'s coordinate-wise approach, we formed an input feature matrix of 5120×2. This setup is equivalent to a training batch of 5120 two-feature samples.

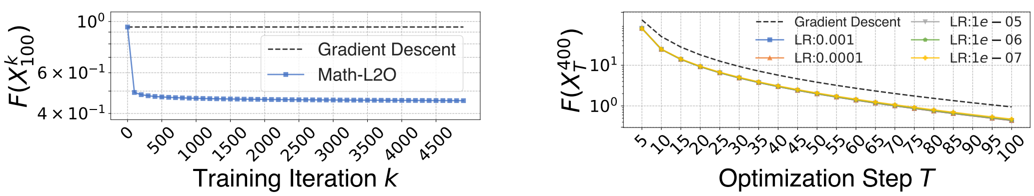
Math-L2O Model Architecture. The Math-L2O model is configured with $T = 100$ optimization steps (Equation (2)). Its architecture comprises a $L = 3$-layer DNN, as formulated in Equation (4). The first layer has an output dimension of 2. To ensure over-parameterization, the $(L-1)$-th (i.e., second) layer's output dimension is set to $512 \times 10 = 5120$. The final layer produces a scalar output (dimension 1). Three specific model configurations are designed for ablation studies, foundational experiments, and robustness evaluations. These are detailed in Appendix C.1.

Training and Initialization Configurations. L2O models are trained using the Stochastic Gradient Descent (SGD) optimizer. For the $L = 3$ layer network configuration, parameters for the initial two layers ($l = 1, 2$) are initialized according to the methodology presented in Section 5.1, while parameters for the final layer ($l = 3$) are zero-initialized.

6.1 Training Performance

We evaluated the mean training loss in Equation (2) across all samples. Figure 4a illustrates this loss at $T = 100$, benchmarked against the standard GD objective (black dashed line). The results demonstrate that Math-L2O consistently achieves fast training convergence, corroborating the theoretical linear convergence established in Theorem 4.3.

Further, we investigated the robustness of our proposed L2O method to variations in optimization steps and learning rates (LRs). Models corresponding to different step/LR configurations are trained for 400 epochs. Figure 4b presents the training objectives for these configurations, benchmarked against standard GD (black dashed line). In contrast to the instability observed for Math-L2O [22] and LISTA-CPSS [7] under certain settings (Figure 2), the consistent convergence across all tested configurations in Figure 4b demonstrates the robustness of our proposed L2O approach.



(a) Loss with Training Iteration (b) Loss with Optimization Step

Figure 4: Training Loss of Our L2O

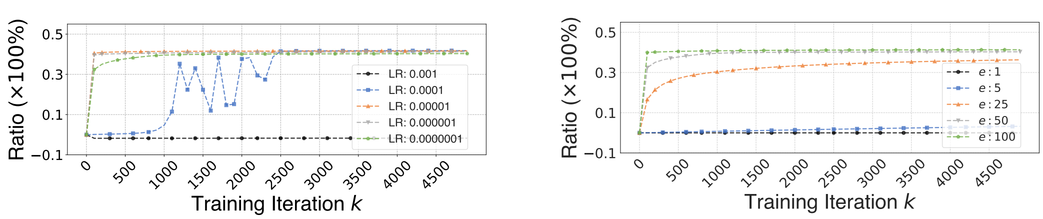
Moreover, we evaluate the inference performance of our framework against baseline methods. Experimental results (in Appendix C.3) demonstrate the framework's robustness to hyperparameters.

6.2 Ablation Study for Learning Rate η and Expansion Coefficient e

We conduct ablation studies to assess the impact of the LR η, theoretically bounded in Equations (13a) and (13b) (Theorem 4.3), and the initialization coefficient e, defined in Section 5. The experimental configuration employs $T = 20$, input $X \in \mathbb{R}^{32\times 32}$, output $Y \in \mathbb{R}^{32\times 20}$, and a neural network width of 1024. Performance is measured by the relative improvement of the proposed L2O method over standard GD at iteration $T = 20$, calculated as $\frac{\text{obj}_{\text{GD}} - \text{obj}_{\text{L2O}}}{\text{obj}_{\text{GD}}}$. These studies further validate Corollary C.1, which establishes an inverse relationship between the viable LR η and the coefficient e, implying that a larger e necessitates a smaller η to ensure convergence.

With the initialization coefficient fixed at $e = 50$, we evaluate the impact of varying the LR η on the relative objective improvement. The results in Figure 5a demonstrate that while LRs such as 10^{-4} and smaller achieve convergence, $\eta = 10^{-3}$ leads to unstable behavior or divergence. This finding empirically supports the existence of an operational upper bound on the LR, consistent with the theoretical constraints outlined in Equations (13a) and (13b). Moreover, reducing the LR below this stability threshold results in slower convergence rates. This observation aligns with the implication of Theorem 4.3 that, under the specified conditions, larger permissible LRs yield faster convergence.

Fixing the LR at $\eta = 10^{-7}$, we examine the influence of the initialization coefficient e on performance. The results, presented in Figure 5b, demonstrate that the relative objective improvement consistently increases with larger values of e. Additional results exploring different e and LR combinations are deferred to Appendix C owing to space constraints. These findings validate the proposed strategies for selecting the initialization coefficient and learning rate.



(a) Ratio with Learning Rate η, $e = 50$ (b) Ratio with Expansion Coefficient e, $\eta = 10^{-7}$

Figure 5: Ablation Studies of Improve Ratio to Learning Rate and Expansion Coefficient

7 Conclusion

This work analyzes a Learning-to-Optimize (L2O) framework that accelerates Gradient Descent (GD) through adaptive step-size learning. We theoretically prove that the L2O training enhances GD's convergence rate by linking network training bounds to GD's performance. Leveraging Neural Tangent Kernel (NTK) theory and over-parameterization via wide layers, we establish convergence guarantees for the complete L2O system. A principled initialization strategy is introduced to satisfy the theoretical requirements for these guarantees. Empirical results across various optimization problems validate our theory and demonstrate substantial practical efficacy.

References

[1] AF Agarap. Deep Learning Using Rectified Linear Units (ReLU). *arXiv preprint arXiv:1803.08375*, 2018.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

[3] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the Convergence Rate of Training Recurrent Neural Networks. *Advances in neural information processing systems*, 32, 2019.

[4] Amir Beck and Marc Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[5] Yanmei Cao, Guomei Zhang, Guobing Li, and Jia Zhang. A Deep Q-Network Based-Resource Allocation Scheme for Massive MIMO-NOMA. *IEEE Communications Letters*, 25(5):1544–1548, 2021.

[6] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Zhangyang Wang, Howard Heaton, Jialin Liu, and Wotao Yin. Learning to Optimize: A Primer and a Benchmark. *The Journal of Machine Learning Research*, 23(1):8562–8620, 2022.

[7] Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Theoretical Linear Convergence of Unfolded ISTA and Its Practical Weights and Thresholds. *Advances in Neural Information Processing Systems*, 31, 2018.

[8] Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Hyperparameter Tuning is All You Need for LISTA. *Advances in Neural Information Processing Systems*, 34:11678–11689, 2021.

[9] Kingma Diederik. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-Parameterized Neural Networks. *arXiv preprint arXiv:1810.02054*, 2018.

[11] Karol Gregor and Yann LeCun. Learning Fast Approximations of Sparse Coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pages 399–406, 2010.

[12] Stephen Grossberg. Recurrent neural networks. *Scholarpedia*, 8(2):1888, 2013.

[13] Howard Heaton, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Safeguarded Learned Convex Optimization. In *AAAI*, pages 7848–7855, 2023.

[14] Qiyu Hu, Yunlong Cai, Qingjiang Shi, Kaidi Xu, Guanding Yu, and Zhi Ding. Iterative Algorithm Induced Deep-Unfolding Neural Networks: Precoding Design for Multiuser MIMO Systems. *IEEE TWC*, 20(2):1394–1410, 2020.

[15] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[16] Sungyoon Kim and Mert Pilanci. Convex Relaxations of ReLU Neural Networks Approximate Global Optima in Polynomial Time. *arXiv preprint arXiv:2402.03625*, 2024.

[17] Sungyoon Kim, Aaron Mishkin, and Mert Pilanci. Exploring the Loss Landscape of Regularized Neural Networks via Convex Duality. *arXiv preprint arXiv:2411.07729*, 2024.

[18] Timothy P Lillicrap and Adam Santoro. Backpropagation Through Time and the Brain. *Current Opinion in Neurobiology*, 55:82–89, 2019.

[19] Wei Lin, Qingyu Song, and Hong Xu. Adaptive Coordinate-Wise Step Sizes for Quasi-Newton Methods: A Learning-to-Optimize Approach. *arXiv preprint arXiv:2412.00059*, 2024.

[20] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss Landscapes and Optimization in Over-Parameterized Non-Linear Systems and Neural Networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

[21] Jialin Liu and Xiaohan Chen. ALISTA: Analytic Weights Are As Good As Learned Weights in LISTA. In *International Conference on Learning Representations (ICLR)*, 2019.

[22] Jialin Liu, Xiaohan Chen, Zhangyang Wang, Wotao Yin, and HanQin Cai. Towards Constituting Mathematical Structures for Learning to Optimize. In *International Conference on Machine Learning*, pages 21426–21449. PMLR, 2023.

[23] Xin Liu, Zhisong Pan, and Wei Tao. Provable Convergence of Nesterovs Accelerated Gradient Method for Over-Parameterized Neural Networks. *Knowledge-Based Systems*, 251:109277, 2022.

[24] Kaifeng Lv, Shunhua Jiang, and Jian Li. Learning Gradient Descent: Better Generalization and Longer Horizons. In *ICML*, pages 2247–2255. PMLR, 2017.

[25] Michael C Mozer. A Focused Backpropagation Algorithm for Temporal Pattern Recognition. In *Backpropagation*, pages 137–169. Psychology Press, 2013.

[26] Sridhar Narayan. The Generalized Sigmoid Activation Function: Competitive Supervised Learning. *Information sciences*, 99(1-2):69–82, 1997.

[27] Quynh Nguyen. On the Proof of Global Convergence of Gradient Descent for Deep ReLU Networks with Linear Widths. In *International Conference on Machine Learning*, pages 8056–8062. PMLR, 2021.

[28] Quynh N Nguyen and Marco Mondelli. Global Convergence of Deep Networks with One Wide Layer Followed by Pyramidal Topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020.

[29] Mert Pilanci. From Complexity to Clarity: Analytical Expressions of Deep Neural Network Weights via Clifford Algebra and Convexity. *Transactions on Machine Learning Research*, 2024.

[30] Boris Teodorovich Polyak. Minimization of Unsmooth Functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

[31] Yifei Shen, Yuanming Shi, Jun Zhang, and Khaled B. Letaief. Graph Neural Networks for Scalable Radio Resource Management: Architecture Design and Theoretical Analysis. *IEEE JSAC*, 39(1):101–115, 2021.

[32] Qingyu Song, Wei Lin, Juncheng Wang, and Hong Xu. Towards Robust Learning to Optimize with Theoretical Guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27498–27506, 2024.

[33] Qingyu Song, Juncheng Wang, Jingzong Li, Guochen Liu, and Hong Xu. A Learning-only Method for Multi-Cell Multi-User MIMO Sum Rate Maximization. In *International Conference on Computer Communications*. IEEE, 2024.

[34] Haoran Sun, Xiangyi Chen, Qingjiang Shi, Mingyi Hong, Xiao Fu, and Nicholas D Sidiropoulos. Learning to Optimize: Training Deep Neural Networks for Interference Management. *IEEE TSP*, 66(20):5438–5453, 2018.

[35] Terence Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

[36] Ryan Tibshirani. 10-725: Convex Optimization — lecture 6: Gradient descent, convergence analysis, and subgradients. Course lecture notes, Carnegie Mellon University, September 2013. URL `https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/lec6.pdf`. Scribed by Micol Marchetti-Bowick. Lecture date: 12 Sep 2013. Accessed 11 May 2025.

[37] Junjie Yang, Tianlong Chen, Mingkang Zhu, Fengxiang He, Dacheng Tao, Yingbin Liang, and Zhangyang Wang. Learning to Generalize Provably in Learning to Optimize. In *International Conference on Artificial Intelligence and Statistics*, pages 9807–9825, 2023.

[38] Yu Zhao, Ignas G. Niemegeers, and Sonia M.Heemstra De Groot. Dynamic Power Allocation for Cell-Free Massive MIMO: Deep Reinforcement Learning Methods. *IEEE Access*, 9: 102953–102965, 2021.

A Appendix

A.1 Details for Definitions

General L2O. Given X_0, we have the following L2O update with NN g to generate X_T:

$$X_t = X_{t-1} + g(W_1, W_2, \ldots, W_L, X_{t-1}, \nabla F(X_{t-1})), t \in [T]. \tag{14}$$

Concatenation of N Problems. For $t \in [T]$, we make the following denotations to represent the concatenation of N samples (each is a unique optimization problem):

$$\mathbf{M} := \begin{bmatrix} \mathbf{M}_1 & & \\ & \ldots & \\ & & \mathbf{M}_N \end{bmatrix}, X_t := [x_{1,t}^\top | x_{2,t}^\top | \ldots | x_{N,t}^\top]^\top, Y := [y_1^\top | y_2^\top | \ldots | y_N^\top]^\top.$$

X_t and Y are still column vectors since we take the coordinate-wise setting from [22].

A.2 Derivative of General L2O

In this section, we derive a general framework for any L2O models by the chain rule, which gives us a complete workflow of each component in the derivatives within the chain. Then, we apply it to the Math-L2O framework [22] to get the formulation for the L2O model defined in equation 4.

Due to chain rule, we have following general formulation of the derivative in L2O model:

$$\frac{\partial F(X_T)}{\partial W_\ell} = \frac{\partial F(X_T)}{\partial X_T} \left(\frac{\partial X_T}{\partial X_{T-1}} \frac{\partial X_{T-1}}{\partial W_\ell} + \frac{\partial X_T}{\partial G_{L,t}} \frac{\partial G_{L,t}}{\partial W_\ell} \right).$$

We calculate each terms in the right-hand side (RHS) in the above formulation. First, we calculate $\frac{\partial X_{T-1}}{\partial W_\ell}$ as:

$$\frac{\partial X_{T-1}}{\partial W_\ell} = \frac{\partial X_{T-1}}{\partial X_{T-2}} \frac{\partial X_{T-2}}{\partial W_\ell} + \frac{\partial X_{T-1}}{\partial G_{L,T-1}} \frac{\partial G_{L,T-1}}{\partial W_\ell}.$$

Thus, we can iteratively derive the gradient until X_1. After arrangement, we have the following complete formulation of $\frac{\partial F}{\partial W_\ell}$:

$$\frac{\partial F(X_T)}{\partial W_\ell} = \frac{\partial F(X_T)}{\partial X_T} \left(\sum_{t=1}^T \left(\prod_{j=T}^{t+1} \frac{\partial X_j}{\partial X_{j-1}} \right) \frac{\partial X_t}{\partial G_{L,t}} \frac{\partial G_{L,t}}{\partial W_\ell} \right). \tag{15}$$

We note that $\frac{\partial X_j}{\partial X_{j-1}}$ relies on different implementations. For example, for general L2O model that the update in each step is directly the output of neural networks (NNs), we have $\frac{\partial X_j}{\partial X_{j-1}} := \mathbf{I} + \frac{\partial G_{L,j}}{\partial X_{j-1}}$. Then, Equation (15) is derived by:

$$\frac{\partial F}{\partial W_\ell} = \frac{\partial F(X_T)}{\partial X_T} \left(\sum_{t=1}^T \left(\prod_{j=T}^{t+1} (\mathbf{I} + \frac{\partial G_{L,j}}{\partial X_{j-1}}) \right) \frac{\partial X_T}{\partial G_{L,t}} \frac{\partial G_{L,t}}{\partial W_\ell} \right). \tag{16}$$

$\frac{\partial G_{L,j}}{\partial X_{j-1}}$ depends on specific implementation of NNs. Liu et al. [22] simplify $\frac{\partial G_{L,j}}{\partial X_{j-1}}$ by detaching input tensor out of back-propagation process, which truncate the branches in the chain from $F(X_T)$ to W_ℓ. The detaching operation yields more simple $\frac{\partial X_j}{\partial X_{j-1}}$. As will be introduced in the following sections, $\frac{\partial X_j}{\partial X_{j-1}}$ depends only on NN's output.

Further, definition of $\frac{\partial X_T}{\partial G_{L,t}}$ relies on different L2O frameworks as well. For example, in the general L2O model, $\frac{\partial X_T}{\partial G_{L,t}} := \mathbf{I}$. In Math-L2O [22], $\frac{\partial X_T}{\partial G_{L,t}}$ is defined by the FISTA algorithm [4]. Next, we conduct a layer-by-layer derivative for every $\frac{\partial G_{L,j}}{\partial X_{t-1}}$ and $\frac{\partial G_{L,t}}{\partial W_\ell}$.

First, we derive $\frac{\partial G_{L,t}}{\partial G_{L-1,t}}$ by:

$$\frac{\partial G_{L,t}}{\partial G_{L-1,t}} = \begin{cases} \nabla \operatorname{ReLU}(G_{L-1,t}) W_\ell & \ell \in [L-1], \\ \nabla 2\sigma(G_{\ell,t}) W_\ell & \ell = L. \end{cases}$$

For simplification, we use $\nabla \operatorname{ReLU}$ and $\nabla 2\sigma$ to represent derivatives $\nabla \operatorname{ReLU}(G_{L-1,t})$ and $\nabla 2\sigma(G_{\ell,t})$, respectively, which are corresponding diagonal matrices of coordinate-wise activation function' derivatives. Next, $\frac{\partial G_{L,t}}{\partial X_{t-1}}$ is given by:

$$\frac{\partial G_{L,j}}{\partial X_{T-1}} = (\prod_{\ell=L}^{2} \frac{\partial G_{l,j}}{\partial G_{l,j-1}}) \frac{\partial G_{1,j-1}}{\partial X_{T-1}} = \nabla 2\sigma w_L (\prod_{\ell=L-1}^{2} \nabla \operatorname{ReLU} W_\ell)[\mathbf{I}, \mathbf{H}^\top], \tag{17}$$

where $\mathbf{H} := \mathbf{M}^\top \mathbf{M}$ denotes the Hessian matrix of loss in Equation (2).

Second, $\frac{\partial G_{L,t}}{\partial W_\ell}$ is given by:

$$\begin{aligned}\frac{\partial G_{l,t}}{\partial W_\ell} &= (\prod_{j=L}^{\ell+1} \frac{\partial G_{j,t}}{\partial G_{j-1,t}}) \frac{\partial G_{l,t}}{\partial W_\ell} \\ &= \begin{cases} \nabla 2\sigma w_L (\prod_{j=L-1}^{\ell+1} \nabla \operatorname{ReLU} W_j) \nabla \operatorname{ReLU}(\mathbf{I}_{n_\ell} \otimes {G_{\ell-1,t}}^\top) & \ell \in [L-1], \\ \nabla 2\sigma (\mathbf{I}_{n_\ell} \otimes {G_{L-1,t}}^\top) & \ell = L, \end{cases}\end{aligned} \tag{18}$$

where $\mathbf{I}_{n_\ell} \in \mathbb{R}^{n_\ell \times n_\ell}$, $\otimes$ denotes Kronecker Product, and $\mathbf{I}_{n_\ell} \otimes {G_{\ell-1,t}}^\top \in \mathbb{R}^{n_\ell \times n_\ell n_{\ell-1}}$.

Substituting Equation (17) and Equation (18) into Equation (16) yields following final derivative formulation of general L2O model:

$$\begin{aligned}&\frac{\partial F}{\partial W_\ell} \\ =&\frac{\partial F(X_T)}{\partial X_T} \Big(\sum_{t=1}^{T} \big(\prod_{j=T}^{t+1} (\mathbf{I} + \frac{\partial G_{L,j}}{\partial X_{j-1}}) \big) \frac{\partial X_T}{\partial G_{L,t}} \frac{\partial G_{L,t}}{\partial W_\ell} \Big), \\ =&\begin{cases} \mathbf{K}_{n_\ell, n_{\ell-1}} \Bigg(({X_T^k}^\top \mathbf{M}^\top - Y^\top) \mathbf{M} \\ \quad \Big(\sum_{t=1}^{T} \big(\mathbf{I} + \nabla 2\sigma w_L (\prod_{\ell=L-1}^{2} \nabla \operatorname{ReLU} W_\ell)[\mathbf{I}, \mathbf{H}^\top] \big)^{T-t} & \ell \in [L-1], \\ \quad \nabla 2\sigma w_L^\top \big(\prod_{j=L-1}^{\ell+1} \nabla \operatorname{ReLU} W_j \big) \nabla \operatorname{ReLU}(\mathbf{I}_{n_\ell} \otimes {G_{\ell-1,t}}^\top) \Big) \Bigg)^\top \\ \mathbf{K}_{n_\ell, n_{\ell-1}} \Bigg(({X_T^k}^\top \mathbf{M}^\top - Y^\top) \mathbf{M} \\ \quad \Big(\sum_{t=1}^{T} \big(\mathbf{I} + \nabla 2\sigma w_L (\prod_{\ell=L-1}^{2} \nabla \operatorname{ReLU} W_\ell)[\mathbf{I}, \mathbf{H}^\top] \big)^{T-t} & l = L, \\ \quad \nabla 2\sigma (\mathbf{I}_{n_\ell} \otimes {G_{L-1,t}}^\top) \Big) \Bigg)^\top \end{cases}\end{aligned} \tag{19}$$

where $\mathbf{K}_{n_\ell, n_{\ell-1}}$ denotes a commutation matrix, which is a $n_\ell * n_{\ell-1} \times n_\ell * n_{\ell-1}$ permutation matrix that swaps rows and columns in the vectorization process.

A.3 Derivative of Coordinate-Wise Math-L2O

Based on the results in Appendix A.2, in this section, we construct the gradient formulations for Math-L2O model. We present the results in Equation (6) and Equation (7).

As defined in equation 3 and equation 4, Math-L2O [22] learns to choose hyperparameters of existing non-learning algorithms [22, 32]. Suppose $P_i \in \mathbb{R}^{N*d}, i \in [0, \dots, T]$ is the hyperparameter vector generated by NNs. Suppose $X_{-1} := X_0$, based on Equation 3, the solution update process from initial step is defined by:

$$\begin{aligned} X_1 &= X_0 - \tfrac{1}{\beta} P_1 \odot \nabla F(X_0), \\ X_2 &= X_1 - \tfrac{1}{\beta} P_2 \odot \nabla F(X_1), \\ &\dots, \\ X_T &= X_{T-1} - \tfrac{1}{\beta} P_T \odot \nabla F(X_{T-1}), \end{aligned} \tag{20}$$

We use this formulation of L2O in the Appendix.

We re-use the definition in Section 2 that defines $\mathcal{D}(\cdot)$ as the operator that constructs a diagonal matrix from a vector, we calculate the following one-line and linear-like formulation of X_T with

X_0:

$$X_T = \prod_{t=T}^{1}(\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_t)\mathbf{M}^\top\mathbf{M})X_0 + \tfrac{1}{\beta}\sum_{t=1}^{T}\prod_{s=T}^{t+1}(\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_s)\mathbf{M}^\top\mathbf{M})\mathcal{D}(P_t)\mathbf{M}^\top Y. \quad (21)$$

It is worth-noting that P_t is generated by non-linear NN with the input about X_{t-1}. Thus, it cannot be formulated into the above linear dynamic system. Moreover, we note that the uncertain subgradient can be replaced with gradient map for non-smooth problems to get similar formulations [32].

Due to the above computational graph in Figure 1, the gradient of X_t comes from X_{t-1} and P_t, which yields the following framework of each layer's derivative (Equation (5)):

$$\frac{\partial F}{\partial W_\ell} = \frac{\partial F(X_T)}{\partial X_T}\Big(\sum_{t=1}^{T}\big(\prod_{j=T}^{t+1}\frac{\partial X_j}{\partial X_{j-1}}\big)\frac{\partial X_t}{\partial P_t}\frac{\partial P_t}{\partial W_\ell}\Big). \quad (22)$$

We obtain the above equation by counting the number of formulations from F to W_ℓ. From the Figure 1, we conclude that each timestamp t leads to the gradient of $\frac{\partial X_T}{\partial X_{T-1}}$. Thus, there are $\prod_{j=T}^{t+1}\frac{\partial X_j}{\partial X_{j-1}}$ blocks of formulation in total.

We start with deriving the formulation of gradient w.r.t. the GD algorithm, which yields the gradient of $\frac{\partial X_T}{\partial P_T}$. Due to the GD formulation in Equation (20), we derive $\frac{\partial X_t}{\partial X_{t-1}}$ as:

$$\begin{aligned}
\frac{\partial X_t}{\partial X_{t-1}} =& \mathbf{I}_d - \frac{1}{\beta}\frac{\partial\Big(P_t\odot\nabla F(X_{t-1})\Big)}{\partial X_{t-1}} \\
=& \mathbf{I}_d - \frac{1}{\beta}\frac{\partial P_t\odot\Big(\mathbf{M}^\top(\mathbf{M}X_{t-1}-Y)\Big)}{\partial X_{t-1}}, \\
=& \mathbf{I}_d - \frac{1}{\beta}\mathcal{D}(P_t)\mathbf{M}^\top\mathbf{M} - \frac{1}{\beta}\frac{\partial P_t\odot\Big(\mathbf{M}^\top(\mathbf{M}X_{t-1}-Y)\Big)}{\partial P_t}\frac{\partial P_t}{\partial X_{t-1}}, \\
=& \mathbf{I}_d - \frac{1}{\beta}\mathcal{D}(P_t)\mathbf{M}^\top\mathbf{M} - \frac{1}{\beta}\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}-Y)\big)\frac{\partial P_t}{\partial X_{t-1}}.
\end{aligned} \quad (23)$$

Next, we calculate $\frac{\partial P_t}{\partial X_{t-1}}$. Similarly, we derive $\frac{\partial \mathrm{vec}(G_{L,t})}{\partial W_\ell}$ and each $\frac{\partial \mathrm{vec}(G_{L,j})}{\partial X_{j-1}}$ of Math-L2O layer-by-layer. $\frac{\partial \mathrm{vec}(G_{L,t})}{\partial \mathrm{vec}(G_{L-1,t})}$ in Math-L2O is similar to Equation (18). We calculate:

$$\begin{cases}
\frac{\partial P_t}{\partial W_\ell} = \mathcal{D}\big(P_t\odot(1-P_t/2)\big)(\mathbf{I}_d\otimes W_L)\prod_{j=L-1}^{\ell+1}\mathbf{D}_{j,t}\mathbf{I}_d\otimes W_j\mathbf{I}_{n_\ell}\otimes {G_{\ell-1,t}}^\top & \ell\in[L-1], \\
\frac{\partial P_t}{\partial W_L} = \mathcal{D}\big(P_t\odot(1-P_t/2)\big){G_{L-1,t}}^\top & \ell = L.
\end{cases} \quad (24)$$

Similarly, we calculate the following derivative of output of Math-L2O to it input at step t:

$$\frac{\partial P_t}{\partial X_{t-1}} = \mathcal{D}\big(P_t\odot(1-P_t/2)\big)W_L\big(\prod_{\ell=L-1}^{2}\mathbf{D}_{\ell,t}W_\ell\big)[\mathbf{I},\mathbf{H}^\top]^\top. \quad (25)$$

Substituting Equation (25) into Equation (23) yields $\frac{\partial X_t}{\partial X_{t-1}}$:

$$\begin{aligned}
\frac{\partial X_t}{\partial X_{t-1}} =& \mathbf{I}_d - \frac{1}{\beta}\mathcal{D}(P_t)\mathbf{M}^\top\mathbf{M} \\
&- \frac{1}{\beta}\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}-Y)\big)\mathcal{D}\big(P_t\odot(1-P_t/2)\big)W_L\big(\prod_{\ell=L-1}^{2}\mathbf{D}_{\ell,t}W_\ell\big)[\mathbf{I},\mathbf{H}^\top]^\top.
\end{aligned} \quad (26)$$

We note that in [22], the gradient formulations are simplified in the implementation by detaching the input feature from computational graph. Thus, we can eliminate the complicated last term in the above formulation, which leads to the following compact version:

$$\frac{\partial X_t}{\partial X_{t-1}} = \mathbf{I}_d - \frac{1}{\beta}\mathcal{D}(P_t)\mathbf{M}^\top\mathbf{M}. \quad (27)$$

In this paper, we take the gradient formulation in Equation (27).

Next, we calculate the $\frac{\partial X_t}{\partial P_t}$ component in Equation (22). We calculate derivative of GD's output to its input hyperparameter P (generated by NNs) as:

$$\frac{\partial X_t}{\partial P_t} = -\frac{1}{\beta}\mathcal{D}(\nabla F(X_{t-1})) = -\frac{1}{\beta}\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}-Y)\big), \quad (28)$$

where $\nabla F(X_{t-1}) := \mathbf{M}^\top(\mathbf{M}X_{t-1} - Y)$ the first-order derivative of the objective in Equation 1.

Substituting Equation (24), Equation (27), and Equation (28) into Equation (22) yields the final derivative of all layers' parameters.

First, for $\ell = L$, since there is no cumulative gradients of later layers, Equation 7 is directly calculated by:

$$\begin{aligned}\frac{\partial F}{\partial W_L} =& -\frac{1}{\beta}\sum_{t=1}^{T}\big(\mathbf{M}^\top(\mathbf{M}X_T - Y)\big)^\top\big(\prod_{j=T}^{t+1}\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_j)\mathbf{M}^\top\mathbf{M}\big)\\ &\mathcal{D}\big((\mathbf{M}^\top(\mathbf{M}X_{t-1} - Y))\big)\mathcal{D}\big(P_t \odot (1 - P_t/2)\big)G_{L-1,t}{}^\top.\end{aligned}$$

And its transpose is given by:

$$\begin{aligned}\frac{\partial F}{\partial W_L}^\top =& -\frac{1}{\beta}\sum_{t=1}^{T}G_{L-1,t}\mathcal{D}\big(P_t \odot (1 - P_t/2)\big)\mathcal{D}\big((\mathbf{M}^\top(\mathbf{M}X_{t-1} - Y))\big)\\ &\big(\prod_{j=t+1}^{T}\mathbf{I} - \frac{1}{\beta}\mathbf{M}^\top\mathbf{M}\mathcal{D}(P_j)\big)\mathbf{M}^\top(\mathbf{M}X_T - Y).\end{aligned} \tag{29}$$

When $\ell \in [L-1]$, the derivative is calculated by:

$$\begin{aligned}\frac{\partial F}{\partial W_\ell} =& \frac{\partial F(X_T)}{\partial X_T}\left(\sum_{t=1}^{T}\left(\prod_{j=T}^{t+1}\frac{\partial X_j}{\partial X_{j-1}}\right)\frac{\partial X_t}{\partial P_t}\frac{\partial P_t}{\partial W_\ell}\right),\\ =& -\frac{1}{\beta}\sum_{t=1}^{T}(\mathbf{M}^\top(\mathbf{M}X_T - Y))^\top\big(\prod_{j=T}^{t+1}\mathbf{I}_d - \frac{1}{\beta}\mathbf{M}^\top\mathbf{M}\mathcal{D}(P_j)\big)\\ &\mathcal{D}\big((\mathbf{M}^\top(\mathbf{M}X_{t-1} - Y))\big)\mathcal{D}\big(P_t \odot (1 - P_t/2)\big)\\ &(\mathbf{I}_d \otimes W_L)\prod_{j=L-1}^{\ell+1}\mathbf{D}_{j,t}\mathbf{I}_d \otimes W_j\mathbf{I}_{n_\ell} \otimes G_{\ell-1,t}{}^\top.\end{aligned}$$

Remark 1. The only difference between Equation (7) and Equation (6) lies in the last term, where Equation (6) is more complicated due to the cumulated gradients from later layers.

The above two formulations are used in the next section to derive the gradient bound for each layer.

A.4 Tools

In this section, before the bound constructions, we start to derive several tools for constructing the convergence rate, which gives several important properties of the L2O models. In the following sections, we use superscript k to denote the parameters and variables at training iteration k and use subscript $_t$ to denote the optimization step.

A.4.1 NN's Outputs are Bounded

First, we demonstrate that the outputs and inner outputs of NN layers within the L2O model are bounded.

Bound $\left\|\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}\right\|_2, \forall k, t$.

Lemma A.1. *Suppose $\|\mathbf{M}^\top\mathbf{M}\|_2 \le \beta$ and $0 < P_t^k < 2$, we have the following bound:*

$$\left\|\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}\right\|_2 < 1. \tag{30}$$

Proof. Suppose eigenvectors of $\mathbf{M}^\top\mathbf{M}$ are σ_i and v_i, $i \in [1, \dots, N*d]$ respectively, we calculate:

$$\frac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}v_i = \frac{\sigma_i}{\beta}\mathcal{D}(P_t^k)v_i.$$

Due to $0 < P_t^k < 2$, we have following spectral norm definition:

$$\left\|\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}\right\|_2 = \max_{x\in\mathbb{R}^d}\frac{x^\top(\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M})x}{x^\top x}$$

Then, by taking $x = v_i$, we calculate:

$$v_i^\top(\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M})v_i = 1 - \frac{1}{\beta}v_i^\top\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}v_i = 1 - \frac{\sigma_i}{\beta}v_i^\top\mathcal{D}(P_t^k)v_i \overset{①}{\le} 1,$$

where ① is due to $0 < P_t^k < 2$. □

Bound $\|\mathcal{D}(P_t^k)\|_2, \forall k, t$. Similar to the bound of $\left\|\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}\right\|_2, \forall k, t$, due to the Sigmoid function, we directly have:

Lemma A.2. *Suppose $0 < P_t^k < 2$, we have the following bound:*

$$\|\mathcal{D}(P_t^k)\|_2 < 2. \tag{31}$$

Proof. Since $\mathcal{D}$ is the diagonalization operation and $0 < P_t^k < 2$, we directly have $\|\mathcal{D}(P_t^k)\|_2 < 2$. □

Besides, we can derive another bound from the Lipschitz property for the Sigmoid activation function:

$$\begin{aligned}
\|\mathcal{D}(P_t^k)\|_2 =& \|2\sigma(\mathrm{ReLU}(\mathrm{ReLU}([X_{t-1}^k, \mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)]{W_1^k}^\top)\cdots {W_{L-1}^k}^\top){W_L^k}^\top)\|_\infty, \\
\overset{①}{\leq}& \tfrac{1}{2}\|[X_{t-1}^k, \mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)]\|_2 \textstyle\prod_{s=1}^{L-1}\|W_s^k\|_2 + 1, \\
\overset{②}{\leq}& \tfrac{1}{2}(\|X_t^k\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_t^k - Y)\|_2)\textstyle\prod_{s=1}^{L-1}\|W_s^k\|_2 + 1.
\end{aligned} \tag{32}$$

① is from equation (17), Lemma 4.2 of [28]. ② is from triangle inequality.

Remark 2. Different from the Lipschitz continuous property of ReLU, the above bound implies that the Sigmoid function prohibits us from obtaining numerical results only. To bound NN's output to achieve the convergence rate of GD, we need a more tight bound. A possible selection is the convex cone defined by W_L^k for the last hidden layer. However, such cone invokes an unbounded space for all learnable parameters.

Bound Semi-Smoothness of NN's Output, i.e., $\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2$, $\forall k, t$. Since our L2O model is a coordinate-wise model [22], suppose $P_i = \alpha_i(P_t^{k+1})_i + (1-\alpha_i)(P_t^k)_i$, $\alpha_p \in [0,1]$, based on Mean Value Theorem, we have $(\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k))_i = \frac{\partial F}{P_i}((P_t^{k+1})_i - (P_t^k)_i)$. Thus, we bound $\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2$ by the following lemma:

Lemma A.3. *Denote $j \in [L]$, for some $\bar{\lambda}_j \in \mathbb{R}$, we assume $\|W_j^{k+1}\|_2 \leq \bar{\lambda}_j$. Using quantities from Equation* (11)*, we have:*

$$\begin{aligned}
&\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2 \\
\leq& \tfrac{1}{2}(1+\beta)\|X_{t-1}^{k+1} - X_{t-1}^k\|_2\Theta_L \\
&+ \tfrac{1}{2}(\|X_{t-1}^k\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\|_2)\Theta_L\textstyle\sum_{\ell=1}^L \bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^k\|_2.
\end{aligned} \tag{33}$$

Remark 3. The above lemma shows the output of NN is a "mixed" Lipschitz continuous on input feature and learnable parameters. The first term illustrates the Lipschitz property on input feature. The second term can be regarded as a Lipschitz property on learnable parameters with a stable input feature.

Proof. Due to Mean Value Theorem, we have:

$$
\begin{aligned}
&\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2 \\
=&\|\mathcal{D}(2\sigma(\mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k+1}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)]{W_1^{k+1}}^\top) \cdots {W_{L-1}^{k+1}}^\top) W_L^{k+1})) \\
&- \mathcal{D}(2\sigma(\mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]{W_1^{k}}^\top) \cdots {W_{L-1}^{k}}^\top) W_L^{k}))\|_2, \\
\leq&(2\sigma(P_i)(1-\sigma(P_i)))_{\max} \\
&\| \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k+1}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)]{W_1^{k+1}}^\top) \cdots {W_{L-1}^{k+1}}^\top) W_L^{k+1} \\
&- \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]{W_1^{k}}^\top) \cdots {W_{L-1}^{k}}^\top) W_L^{k}\|_\infty, \\
\leq&\tfrac{1}{2}\| \mathrm{ReLU}(\mathrm{ReLU}([X_{t-1}^{k+1}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)]{W_1^{k+1}}^\top) \cdots {W_{L-1}^{k+1}}^\top) W_L^{k+1} \\
&\quad - \mathrm{ReLU}(\mathrm{ReLU}([X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]{W_1^{k}}^\top) \cdots {W_{L-1}^{k}}^\top) W_L^{k}\|_\infty, \\
\overset{①}{\leq}&\tfrac{1}{2}\| \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k+1}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)]{W_1^{k+1}}^\top) \cdots {W_{L-1}^{k+1}}^\top) \\
&\quad - \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]{W_1^{k}}^\top) \cdots {W_{L-1}^{k}}^\top)\|_\infty \|W_L^{k+1}\|_2 \\
&+ \tfrac{1}{2}\| \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]) {W_{L-1}^{k}}^\top)\|_2 \|W_L^{k+1} - W_L^k\|_2, \\
\overset{②}{\leq}&\tfrac{1}{2}\| \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k+1}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)]{W_1^{k+1}}^\top) \cdots {W_{L-2}^{k+1}}^\top) {W_{L-1}^{k+1}}^\top \\
&\quad - \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]{W_1^{k}}^\top) \cdots {W_{L-2}^{k}}^\top) {W_{L-1}^{k}}^\top\|_\infty \bar{\lambda}_L \\
&+ \tfrac{1}{2}\|[X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]\|_2 \textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \|W_L^{k+1} - W_L^k\|_2, \\
\overset{③}{\leq}&\tfrac{1}{2}\| \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k+1}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)]{W_1^{k+1}}^\top) \cdots {W_{L-2}^{k+1}}^\top) \\
&\quad - \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]{W_1^{k}}^\top) \cdots {W_{L-2}^{k}}^\top)\|_\infty \bar{\lambda}_{L-1}\bar{\lambda}_L \\
&+ \tfrac{1}{2}\|[X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]\|_2 \textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \|W_L^{k+1} - W_L^k\|_2, \\
&+ \tfrac{1}{2}\|[X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]\|_2 \textstyle\prod_{j=1}^{L-2}\bar{\lambda}_j \bar{\lambda}_L \|W_{L-1}^{k+1} - W_{L-1}^k\|_2, \\
\overset{④}{=}&\tfrac{1}{2}\| \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k+1}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)]{W_1^{k+1}}^\top) \cdots {W_{L-2}^{k+1}}^\top) \\
&\quad - \mathrm{ReLU}(\cdots \mathrm{ReLU}([X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]{W_1^{k}}^\top) \cdots {W_{L-2}^{k}}^\top)\|_\infty \bar{\lambda}_{L-1}\bar{\lambda}_L \\
&+ \tfrac{1}{2}\|[X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]\|_2 \Theta_L(\bar{\lambda}_L^{-1}\|W_L^{k+1} - W_L^k\|_2 + \bar{\lambda}_{L-1}^{-1}\|W_{L-1}^{k+1} - W_{L-1}^k\|_2), \\
&\cdots, \\
\overset{⑤}{\leq}&\tfrac{1}{2}\|[X_{t-1}^{k+1}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)] - [X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]\|_2 \Theta_L \\
&+ \tfrac{1}{2}\|[X_{t-1}^{k}, \mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)]\|_2 \Theta_L \Big(\textstyle\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^k\|_2\Big), \\
\overset{⑥}{\leq}&\tfrac{1}{2}(1+\beta)\|X_{t-1}^{k+1} - X_{t-1}^k\|_2 \Theta_L \\
&+ \tfrac{1}{2}(\|X_{t-1}^k\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_{t-1}^{k} - Y)\|_2) \Theta_L \Big(\textstyle\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^k\|_2\Big).
\end{aligned}
$$

① is due to triangle and Cauchy Schwarz inequalities, where we make a upper bound relaxation from ∞-norm to 2-norm. ② is due to 1-Lipschitz property of ReLU and $\max(\|W_L^{k+1}\|_2, \|W_L^k\|_2) \leq \bar{\lambda}_L$ in the definition. ③ is due to triangle and Cauchy Schwarz inequalities as well. We make a arrangement in ④ and eliminate inductions in $\cdots$. In ⑤. we make another upper bound relaxation from ∞-norm to 2-norm. ⑥ is due to triangle inequality, the definition of Frobenius norm, and $\|\mathbf{M}^\top\mathbf{M}\|_2 \leq L$ of objective's L-smooth property. □

Semi-Smoothness of Inner Output of NN, i.e., Bound $\|G_{\ell,t}^a - G_{\ell,t}^b\|_2, \ell \in [L-1], \forall a, b, t$.

Lemma A.4. *Denote $\ell \in [L-1]$, for some $\bar{\lambda}_\ell \in \mathbb{R}$, we assume* $\max(\|W_\ell^a\|_2, \|W_\ell^b\|_2) \leq \bar{\lambda}_\ell$. *Using quantities from Equation* (11)*, we have:*

$$\begin{aligned}\|G_{\ell,t}^a - G_{\ell,t}^b\|_2 \leq &(1+\beta)\|X_{t-1}^a - X_{t-1}^b\|_2 \textstyle\prod_{j=1}^{\ell} \bar{\lambda}_j \\ &+ (\|X_{t-1}^b\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_{t-1}^b - Y)\|_2) \textstyle\prod_{j=1}^{\ell} \bar{\lambda}_j \sum_{s=1}^{\ell} \bar{\lambda}_s^{-1} \|W_s^a - W_s^b\|_2.\end{aligned}$$

Proof. Since the bounding target in Lemma A.4 is a degenerated version of that in Lemma A.3. Similar to the proof of Lemma A.3, we calculate:

$$\begin{aligned}&\|G_{\ell,t}^a - G_{\ell,t}^b\|_2 \\ =&\| \mathrm{ReLU}(\mathrm{ReLU}([X_{t-1}^a, \mathbf{M}^\top(\mathbf{M}X_{t-1}^a - Y)]{W_1^a}^\top) \cdots {W_\ell^a}^\top) \\ &- \mathrm{ReLU}(\mathrm{ReLU}([X_{t-1}^b, \mathbf{M}^\top(\mathbf{M}X_{t-1}^b - Y)]{W_1^b}^\top) \cdots {W_\ell^b}^\top)\|_2, \\ \leq&\|[X_{t-1}^a, \mathbf{M}^\top(\mathbf{M}X_{t-1}^a - Y)] - [X_{t-1}^b, \mathbf{M}^\top(\mathbf{M}X_{t-1}^b - Y)]\|_2 \textstyle\prod_{j=1}^{\ell} \bar{\lambda}_j \\ &+ \|[X_{t-1}^b, \mathbf{M}^\top(\mathbf{M}X_{t-1}^b - Y)]\|_2 \textstyle\prod_{j=1}^{\ell} \bar{\lambda}_j \sum_{s=1}^{\ell} \bar{\lambda}_s^{-1} \|W_s^a - W_s^b\|_2, \\ \leq&(1+\beta)\|X_{t-1}^a - X_{t-1}^b\|_2 \textstyle\prod_{j=1}^{\ell} \bar{\lambda}_j \\ &+ (\|X_{t-1}^b\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_{t-1}^b - Y)\|_2) \textstyle\prod_{j=1}^{\ell} \bar{\lambda}_j \sum_{s=1}^{\ell} \bar{\lambda}_s^{-1} \|W_s^a - W_s^b\|_2.\end{aligned}$$

□

Bound NN's Inner Output $G_{l,t}^k, l = [L-1], \forall k, t$.

Lemma A.5. *Denote $\ell \in [L-1]$, for some $\bar{\lambda}_\ell \in \mathbb{R}$, we assume $\|W_\ell^k\|_2 \leq \bar{\lambda}_\ell$. Using quantities from Equation* (11)*, we have:*

$$\|G_{\ell,t}^k\|_2 \leq \big((1+\beta)\|X_0\|_2 + \big(2t - 1 + \tfrac{2t-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big) \textstyle\prod_{s=1}^{\ell} \bar{\lambda}_s.$$

Proof.

$$\begin{aligned}\|G_{\ell,t}^k\|_2 =&\| \mathrm{ReLU}(\mathrm{ReLU}([X_{t-1}^k, \mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)]{W_1^k}^\top) \cdots {W_\ell^k}^\top)\|_2, \\ \overset{①}{\leq}&\|[X_{t-1}^k, \mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)]\|_2 \textstyle\prod_{s=1}^{\ell} \|W_s^k\|_2, \\ \overset{②}{\leq}&(\|X_{t-1}^k\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\|_2) \textstyle\prod_{s=1}^{\ell} \|W_s^k\|_2, \\ \overset{③}{\leq}&\big((1+\beta)\|X_0\|_2 + \Big(\tfrac{(1+\beta)2(t-1)}{\beta} + 1\Big)\|\mathbf{M}^\top Y\|_2\big) \textstyle\prod_{s=1}^{\ell} \|W_s^k\|_2, \\ \leq&\big((1+\beta)\|X_0\|_2 + \big(2t - 1 + \tfrac{2t-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big) \textstyle\prod_{s=1}^{\ell} \bar{\lambda}_s.\end{aligned}$$

① is from equation (17), Lemma 4.2 of [28]. ② is from triangle inequality. ③ is due to definition of β-smoothness of objective and upper bound of $\|X_t\|_2$ in Lemma A.6. □

A.4.2 Outputs of L2O are Bounded

Next, we establish bounds for the Math-L2O's outputs. Leveraging the momentum-free setting, we formulate the dynamics from X_0 to X_t as a *semi-linear* system, where parameters are non-linearly generated by the NN block (see Figure 1a). Application of the Cauchy-Schwarz and triangle inequalities to this system yields the following explicit bound.

Lemma A.6 (Bound on Math-L2O Output)**.** *For any training iteration k, the t-th output X_t^k of Math-L2O (as per Equation* (3)*) is bounded by:* $\|X_t^k\|_2 \leq \|X_0\|_2 + \frac{2t}{\beta}\|\mathbf{M}^\top Y\|_2$.

Proof. We calculate the upper bound based on the one-line formulation from X_0 in Equation (21).

$$
\begin{aligned}
&\|X_t^k\|_2 \\
=&\Big\|\prod_{s=t}^{1}(\mathbf{I}-\tfrac{1}{\beta}\mathcal{D}(P_s^k)\mathbf{M}^\top\mathbf{M})X_0+\tfrac{1}{\beta}\sum_{s=1}^{t}\prod_{j=t}^{s+1}(\mathbf{I}-\tfrac{1}{\beta}\mathcal{D}(P_s^k)\mathbf{M}^\top\mathbf{M})\mathcal{D}(P_s^k)\mathbf{M}^\top Y\Big\|_2 \\
\overset{①}{\le}&\Big\|\prod_{s=1}^{t}(\mathbf{I}-\tfrac{1}{\beta}\mathcal{D}(P_s^k)\mathbf{M}^\top\mathbf{M})X_0\Big\|_2+\Big\|\tfrac{1}{\beta}\sum_{s=1}^{t}\prod_{j=t}^{s+1}(\mathbf{I}-\tfrac{1}{\beta}\mathcal{D}(P_s^k)\mathbf{M}^\top\mathbf{M})\mathcal{D}(P_s^k)\mathbf{M}^\top Y\Big\|_2 \\
\overset{②}{\le}&\prod_{s=1}^{t}\Big\|\mathbf{I}-\tfrac{1}{\beta}\mathcal{D}(P_s^k)\mathbf{M}^\top\mathbf{M}\Big\|_2\|X_0\|_2 \\
&+\tfrac{1}{\beta}\sum_{s=1}^{t}\prod_{j=t}^{s+1}\Big\|\mathbf{I}-\tfrac{1}{\beta}\mathcal{D}(P_s^k)\mathbf{M}^\top\mathbf{M}\Big\|_2\|\mathcal{D}(P_s^k)\|_2\|\mathbf{M}^\top Y\|_2, \\
\overset{③}{\le}&\|X_0\|_2+\tfrac{2}{\beta}\sum_{s=1}^{t}\|\mathbf{M}^\top Y\|_2=\|X_0\|_2+\tfrac{2t}{\beta}\|\mathbf{M}^\top Y\|_2,
\end{aligned}
$$

where ① is from the triangle inequality, ② is due to Cauchy Schwarz inequalities, and ③ is due to Lemma A.1 and Lemma A.2. □

This lemma demonstrates that Math-L2O outputs remain bounded independently of the training iteration k and the specific learnable parameters.

A.4.3 L2O is Semi-Smooth to Its Parameters

In this section, we regard L2O model defined in Equation (20) and corresponding neural network are functions with input as learnable parameters. We prove that the functions are semi-smooth w.r.t. the different parameters. This is the foundation of proving the convergence of gradient descent algorithm since the algorithm leads to two adjacent parameters.

First, we give the following explicit formulation of P:

$$
\begin{aligned}
P_t^k &= 2\sigma(W_L^k\,\mathrm{ReLU}(W_{L-1}^k(\cdots\mathrm{ReLU}(W_1^k[X_{t-1}^k,\mathbf{M}^\top(\mathbf{M}X_{t-1}^k-Y)]^\top)\cdots)))^\top, \\
&= 2\sigma(\mathrm{ReLU}(\cdots\mathrm{ReLU}([X_{t-1}^k,\mathbf{M}^\top(\mathbf{M}X_{t-1}^k-Y)]W_1^\top)\cdots W_{L-1}^{k}{}^\top)W_L^k).
\end{aligned}
$$

Moreover, we present ReLU activation function with signal matrices defined in Section 2. We denote $\cdot_K$ as the entry-wise product to the matrices, which is also equivalent to reshape a matrix to a vector then product a diagonal signal matrix and reshape back afterward.

$$
\begin{aligned}
P_t^k &= 2\sigma(W_L^k\mathbf{D}_{L-1}\cdot_K W_{L-1}^k(\cdots\mathbf{D}_1\cdot_K(W_1^k[X_{t-1}^k,\mathbf{M}^\top(\mathbf{M}X_{t-1}^k-Y)]^\top)\cdots))^\top, \\
&= 2\sigma((\cdots\cdots([X_{t-1}^k,\mathbf{M}^\top(\mathbf{M}X_{t-1}^k-Y)]W_1^\top)\cdot_K\mathbf{D}_1\cdots)W_{L-1}^{k}{}^\top\cdot_K\mathbf{D}_{L-1}W_L^k).
\end{aligned}
$$

Proof for Lemma 4.2. We demonstrate the semi-smoothness of Math-L2O's output, i.e., bound $\|X_t^{k+1}-X_t^k\|_2, \forall k,t$

Proof. Different from [28], X_T^{k+1} and X_T^k are outputs of a NN with different inputs. We cannot direct write down a subtraction between two linear-like NNs, using quantities from Equation (11), we choose to make up such subtractions by upper bound relaxation of norm and calculate that:

$$
\begin{aligned}
&\|X_t^{k+1} - X_t^k\|_2 \\
=&\Big\|X_{t-1}^{k+1} - \tfrac{1}{\beta}\mathcal{D}(P_t^{k+1})\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^{k+1} - Y)\big) - \big(X_{t-1}^k - \tfrac{1}{\beta}\mathcal{D}(P_t^k)(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y))\big)\Big\|_2, \\
=&\Big\|\Big(\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_t^{k+1})\mathbf{M}^\top\mathbf{M}\Big)X_{t-1}^{k+1} - \Big(\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}\Big)X_{t-1}^k \\
&\quad + \tfrac{1}{\beta}(\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k))\mathbf{M}^\top Y\Big\|_2 \\
\overset{①}{\leq}&\Big\|\Big(\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_t^{k+1})\mathbf{M}^\top\mathbf{M}\Big) - \Big(\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}\Big)\Big\|_2\|X_{t-1}^{k+1}\|_2 \\
&+ \Big\|\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_t^k)\mathbf{M}^\top\mathbf{M}\Big\|_2\|X_{t-1}^{k+1} - X_{t-1}^k\|_2 + \tfrac{1}{\beta}\|\mathbf{M}^\top Y\|_2\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2, \\
\overset{②}{\leq}&\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2\|X_{t-1}^{k+1}\|_2 + \|X_{t-1}^{k+1} - X_{t-1}^k\|_2 + \tfrac{1}{\beta}\|\mathbf{M}^\top Y\|_2\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2, \\
\overset{③}{\leq}&\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2(\|X_0\|_2 + \tfrac{2t-2}{\beta}\|\mathbf{M}^\top Y\|_2) + \|X_{t-1}^{k+1} - X_{t-1}^k\|_2 \\
&+ \tfrac{1}{\beta}\|\mathbf{M}^\top Y\|_2\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2, \\
=&(\|X_0\|_2 + \tfrac{2t-1}{\beta}\|\mathbf{M}^\top Y\|_2)\|\mathcal{D}(P_t^{k+1}) - \mathcal{D}(P_t^k)\|_2 + \|X_{t-1}^{k+1} - X_{t-1}^k\|_2, \\
\overset{④}{\leq}&(\|X_0\|_2 + \tfrac{2t-1}{\beta}\|\mathbf{M}^\top Y\|_2) \\
&\Big(\tfrac{1}{2}(1+\beta)\|X_{t-1}^{k+1} - X_{t-1}^k\|_2\Theta_L \\
&\quad + \tfrac{1}{2}(\|X_{t-1}^k\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\|_2)\Theta_L\textstyle\sum_{\ell=1}^L\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^k\|_2\Big) \\
&+ \|X_{t-1}^{k+1} - X_{t-1}^k\|_2, \\
=&\Big(1 + (\|X_0\|_2 + \tfrac{2t-1}{\beta}\|\mathbf{M}^\top Y\|_2)\tfrac{1+\beta}{2}\Theta_L\Big)\|X_{t-1}^{k+1} - X_{t-1}^k\|_2, \\
&+ \tfrac{1}{2}(\|X_0\|_2 + \tfrac{2t-1}{\beta}\|\mathbf{M}^\top Y\|_2) \\
&\quad(\|X_{t-1}^k\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\|_2)\Theta_L\textstyle\sum_{\ell=1}^L\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^k\|_2, \\
\overset{⑤}{\leq}&\tfrac{1}{2}\textstyle\sum_{s=1}^t\Big(\prod_{j=s+1}^t\big(1 + (\|X_0\|_2 + \tfrac{2j-1}{\beta}\|\mathbf{M}^\top Y\|_2)\tfrac{1+\beta}{2}\Theta_L\big)\Big) \\
&\quad\underbrace{(\|X_0\|_2 + \tfrac{2s-1}{\beta}\|\mathbf{M}^\top Y\|_2)\big((1+\beta)\|X_0\|_2 + (2s-1+\tfrac{2s-2}{\beta})\|\mathbf{M}^\top Y\|_2\big)}_{\Lambda_s} \\
&\quad\Theta_L\textstyle\sum_{\ell=1}^L\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^k\|_2,
\end{aligned}
$$

where ① is from triangle inequality. ② is from Lemma A.6. ③ is due to inductive summation to $t = 1$. ④ is due to the semi-smoothness of NN's output in Lemma A.3. ⑤ is from induction.

Remark 4. We note that the above upper bound relaxation is non-loose. Current existing approaches derive semi-smoothness in terms of NN functions, where parameters matrices are linearly applied and activation functions are Lipschitz continuous. However, in our setting under [22], the sigmoid activation is not Lipschitz continuous. Moreover, the input that is utilized to generate X_t^{k+1} is from X_{t-1}^{k+1}, which is not identical to the X_{t-1}^k for generating X_{t-1}^k.

□

A.4.4 Gradients are Bounded

In this section, we derive bound for the gradient of each layer's parameter at the given iteration k.

Proof for Lemma 4.1 We demonstrate that the gradients of Math-L2O's each layer are bounded.

Proof. For $\ell = L$, we calculate the gradient on W_L^k (Equation (7)):

$$
\begin{aligned}
&\left\|\frac{\partial F}{\partial W_L^k}\right\|_2 \\
=&\frac{1}{\beta}\Big\|\sum_{t=1}^{T}\big(\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\big)^\top \\
&\qquad \big(\prod_{j=T}^{t+1}\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_j^k)\mathbf{M}^\top\mathbf{M}\big)\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big){G_{L-1,t}^k}^\top\Big\|_2, \\
\overset{①}{\leq}&\frac{1}{\beta}\sum_{t=1}^{T}\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2\prod_{j=T}^{t+1}\Big\|(\mathbf{I}_d - \frac{1}{\beta}\mathcal{D}(P_j^k)\mathbf{M}^\top\mathbf{M})\Big\|_2 \\
&\qquad\qquad \|\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\|_2\|\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big)\|_2\|G_{L-1,t}^k\|_2, \\
\overset{②}{\leq}&\frac{1}{2\sqrt{\beta}}\|\mathbf{M}X_T^k - Y\|_2\sum_{t=1}^{T}(\|\mathbf{M}^\top\mathbf{M}X_{t-1}^k\|_2 + \|\mathbf{M}^\top Y\|_2)\|G_{L-1,t}^k\|_2, \\
\overset{③}{\leq}&\frac{\sqrt{\beta}}{2}\|\mathbf{M}X_T^k - Y\|_2\prod_{\ell=1}^{L-1}\bar{\lambda}_\ell\sum_{t=1}^{T}\big((1+\beta)\|X_0\|_2 + \big(2t - 1 + \frac{2t-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big) \\
&\quad (\|X_0\|_2 + \frac{2t-1}{\beta}\|\mathbf{M}^\top Y\|_2), \\
=&\frac{\sqrt{\beta}}{2}\|\mathbf{M}X_T^k - Y\|_2\prod_{\ell=1}^{L-1}\bar{\lambda}_\ell\sum_{t=1}^{T} \\
&\underbrace{(1+\beta)\|X_0\|_2^2 + \big((4t-3)(1+\frac{1}{\beta}) + 1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2 + \frac{(2T-1)(\beta(2T-1)+(2T-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2}_{\Lambda_t}, \\
=&\frac{\sqrt{\beta}\Theta_L S_{\Lambda,T}}{2\bar{\lambda}_L}\|\mathbf{M}X_T^k - Y\|_2,
\end{aligned}
$$

where ① is from triangle and Cauchy-Schwarz inequalities. ② is from the bound of "p" in Lemma A.1. ③ is from the bound of L2O model's output in Lemma A.6 and inner outputs in Lemma A.5.

For $\ell \in [L-1]$, we calculate gradient on W_ℓ^k (Equation (6)) at iteration k by:

$$
\begin{aligned}
&\left\|\frac{\partial F}{\partial W_\ell^k}\right\|_2 \\
=&\Big\| - \frac{1}{\beta}\sum_{t=1}^{T}(\mathbf{M}^\top(\mathbf{M}X_T^k - Y))^\top\big(\prod_{j=T}^{t+1}\mathbf{I}_d - \frac{1}{\beta}\mathbf{M}^\top\mathbf{M}\mathcal{D}(P_j^k)\big) \\
&\qquad\qquad \mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big)(\mathbf{I}_d \otimes W_L^k) \\
&\qquad\qquad \prod_{j=L-1}^{\ell+1}\mathbf{D}_{j,t}^k\mathbf{I}_d \otimes W_j^k\mathbf{I}_{n_\ell} \otimes {G_{\ell-1,t}^k}^\top\Big\|_2, \\
\overset{①}{\leq}&\frac{1}{\beta}\sum_{t=1}^{T}\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2\prod_{j=T}^{t+1}\|\mathbf{I}_d - \frac{1}{\beta}\mathbf{M}^\top\mathbf{M}\mathcal{D}(P_j^k)\|_2\|\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\|_2 \\
&\qquad \|\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big)(\mathbf{I}_d \otimes W_L^k)\|_2\Big\|\prod_{j=L-1}^{\ell+1}\mathbf{D}_{j,t}^k\mathbf{I}_d \otimes W_j^k\mathbf{I}_{n_\ell} \otimes {G_{\ell-1,t}^k}^\top\Big\|_2, \\
\overset{②}{\leq}&\frac{\sqrt{\beta}}{2}\|\mathbf{M}X_T^k - Y\|_2\prod_{j=\ell+1}^{L}\|W_j^k\|_2\sum_{t=1}^{T}(\|\mathbf{M}^\top\mathbf{M}X_{t-1}^k\|_2 + \|\mathbf{M}^\top Y\|_2)\|G_{\ell-1,t}^k\|_2, \\
\overset{②}{=}&\frac{\sqrt{\beta}}{2}\|\mathbf{M}X_T^k - Y\|_2\prod_{j=1,j\neq\ell}^{L}\bar{\lambda}_j\sum_{t=1}^{T} \\
&\underbrace{(1+\beta)\|X_0\|_2^2 + \big((4t-3)(1+\frac{1}{\beta}) + 1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2 + \frac{(2T-1)(\beta(2T-1)+(2T-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2}_{\Lambda_t}, \\
=&\frac{\sqrt{\beta}\Theta_L}{2\lambda_\ell}S_{\Lambda,T}\|\mathbf{M}X_T^k - Y\|_2,
\end{aligned}
$$

① is from triangle and Cauchy-Schwarz inequalities. Inequality ② is from bounds of "p" in Lemma A.1 and we make a rearrangement in it. In inequality ②, we use norm's triangle inequality of dot product and Kronecker product, bounds of NN's inner output in Lemma A.5, and we calculate $\prod_{j=1,j\neq\ell}^{L}\|W_j^k\|_2 = \prod_{j=\ell+1}^{L}\|W_j^k\|_2 * \prod_{s=1}^{\ell-1}\|W_j^k\|_2$. We reuse the result in the proof for the last layer's gradient upper bound for case $\ell = L$ in equality ③ to get the final result. □

A.5 Bound Linear Convergence Rate

Now we are able to substitute the above formulation into three bounding targets in Equation (42) and bound them one-by-one.

Proof. We start to prove the Theorem 4.3 by proving the following lemma.

Lemma A.7.

$$\begin{cases} \|W_\ell^r\|_2 \leq \bar{\lambda}_\ell, \quad \ell \in [L], \quad r \in [0,k], \\ \sigma_{\min}(G_{L-1,T}^r) \geq \frac{1}{2}\alpha_0, \quad r \in [0,k], \\ F([W]^r) \leq (1-\eta 4\eta \frac{\beta_0^2}{\beta^2}\delta_4)^r F([W]^0), \quad r \in [0,k]. \end{cases} \tag{34}$$

Remark 5. The first inequality means that there exists a scalar $\bar{\lambda}_\ell$ that upper bounds each layer's learnable parameter. The second inequality means that the last inner output is lower bounded. The last inequality is the linear rate of training.

A.5.1 Induction Part 1: NN's Parameter and the Last Inner Output are Bounded

For $k = 0$, Equation (34) degenerates and holds by nature. Assume Equation (34) holds up to iteration k, we aim to prove it still holds for iteration $k+1$. We calculate:

$$\begin{aligned} \|W_\ell^{k+1} - W_\ell^0\|_2 &\overset{①}{\leq} \sum_{s=0}^k \|W_\ell^{s+1} - W_\ell^s\|_2 \\ &\overset{②}{=} \eta \sum_{s=0}^k \left\| \frac{\partial F}{W_\ell^s} \right\|_2 \\ &\overset{③}{\leq} \eta \sum_{s=0}^k \frac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell} S_{\Lambda,T} \|\mathbf{M}X_T^s - Y\|_2, \\ &\overset{④}{\leq} \eta \frac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell} S_{\Lambda,T} \sum_{s=0}^k (1-\eta 4\eta \frac{\beta_0^2}{\beta^2}\delta_4)^{s/2} \|\mathbf{M}X_T^0 - Y\|_2, \end{aligned}$$

where ① is due to triangle inequality. ② is due to definition of gradient descent. ③ is due the gradient is upperly bounded in Lemma 4.1 and our assumption that $\|W_\ell^r\|_2 \leq \bar{\lambda}_\ell, \quad \ell \in [L], \forall r \in [0,k]$. ④ is due to the linear rate in our induction assumption.

Define $u := \sqrt{1-\eta 4\eta \frac{\beta_0^2}{\beta^2}\delta_4}$, we calculate the sum of geometric sequence by:

$$\begin{aligned} \eta \frac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell} S_{\Lambda,T} \sum_{s=0}^k u^s \|\mathbf{M}X_T^0 - Y\|_2 &= \eta \frac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell} S_{\Lambda,T} \frac{1-u^{k+1}}{1-u} \|\mathbf{M}X_T^0 - Y\|_2, \\ &\overset{①}{=} \frac{1}{4\eta \frac{\beta_0^2}{\beta^2}\delta_4} \frac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell} S_{\Lambda,T} (1-u^2) \frac{1-u^{k+1}}{1-u} \|\mathbf{M}X_T^0 - Y\|_2, \\ &\overset{②}{\leq} \frac{1}{4\eta \frac{\beta_0^2}{\beta^2}\delta_4} \frac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell} S_{\Lambda,T} \|\mathbf{M}X_T^0 - Y\|_2, \\ &\overset{③}{\leq} \frac{1}{4\eta \frac{\beta_0^2}{\beta^2}\delta_4} \frac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell} S_{\Lambda,T} \big(\sqrt{\beta}\|X_0\|_2 + (2T+1)\|Y\|_2\big), \\ &\overset{④}{\leq} C_\ell, \end{aligned}$$

where ① is due to $1 - u^2 = \eta 4\eta \frac{\beta_0^2}{\beta^2}\delta_4$. ② is due to $0 \leq u \leq 1$. ③ is due to NN's output's bound in Lemma A.6. ④ is due to the lower bound of singular value of last inner output layer in Equation (12c).

Due to Weyl's inequality [27], we have:

$$\|W_\ell^{k+1}\|_2 \leq \|W_\ell^0\|_2 + C_\ell = \bar{\lambda}_\ell.$$

Next, we bound $G^{k+1}_{L-1,T}$ by calculating:

$$\begin{aligned}
&\|G^{k+1}_{L-1,T} - G^0_{L-1,T}\|_2 \\
\overset{①}{\leq}&(1+\beta)\|X^{k+1}_{T-1} - X^0_{T-1}\|_2 \textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \\
&+ (\|X^0_{T-1}\|_2 + \|\mathbf{M}^\top(\mathbf{M}X^0_{T-1} - Y)\|_2)\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W^{k+1}_\ell - W^0_\ell\|_2, \\
\overset{②}{\leq}&(1+\beta)2(\|X_0\|_2 + \tfrac{2T-2}{\beta}\|\mathbf{M}^\top Y\|_2)\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \\
&+ (\|X^0_{T-1}\|_2 + \|\mathbf{M}^\top(\mathbf{M}X^0_{T-1} - Y)\|_2)\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W^{k+1}_\ell - W^0_\ell\|_2, \\
\overset{③}{\leq}&(1+\beta)\textstyle\sum_{i=0}^{k}\frac{1}{2}\Theta_L \underbrace{\textstyle\sum_{s=1}^{T-1}\Big(\prod_{j=s+1}^{T-1}\big(1+\frac{1+\beta}{2}\Theta_L\Phi_j\big)\Big)\Lambda_s}_{\delta_1^{T-1}} \textstyle\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\|W^{i+1}_\ell - W^i_\ell\|_2 \prod_{j=1}^{L-1}\bar{\lambda}_j \\
&+ (\|X^0_{T-1}\|_2 + \|\mathbf{M}^\top(\mathbf{M}X^0_{T-1} - Y)\|_2)\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\|W^{k+1}_\ell - W^0_\ell\|_2,
\end{aligned} \tag{35}$$

where ① is due to the semi-smoothness of NN's inner output in Lemma A.4. ② is due to the triangle inequality. ③ is due to semi-smoothness of L2O in Lemma 4.2.

Further, based on the inner results in the former demonstration for $\|W^{k+1}_\ell - W^0_\ell\|_2$, we have:

$$\textstyle\sum_{i=0}^{k}\|W^{i+1}_\ell - W^i_\ell\|_2 \leq \frac{1}{4\eta\frac{\beta_0^2}{\beta^2}\delta_4}\frac{\sqrt{\beta}\Theta_L}{2\lambda_\ell}S_{\Lambda,T}\|\mathbf{M}X^0_T - Y\|_2.$$

Substituting above result back into Equation (35) yields:

$$\begin{aligned}
&\|G^{k+1}_{L-1,T} - G^0_{L-1,T}\|_2 \\
\leq&(1+\beta)2(\|X_0\|_2 + \tfrac{2T-2}{\beta}\|\mathbf{M}^\top Y\|_2)\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \\
&+ (\|X^0_{T-1}\|_2 + \|\mathbf{M}^\top(\mathbf{M}X^0_{T-1} - Y)\|_2)\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\frac{1}{4\eta\frac{\beta_0^2}{\beta^2}\delta_4}\frac{\sqrt{\beta}\Theta_L}{2\lambda_\ell}S_{\Lambda,T}\|\mathbf{M}X^0_T - Y\|_2, \\
\overset{①}{\leq}&\tfrac{1}{4\eta\frac{\beta_0^2}{\beta^2}\delta_4}(1+\beta)\zeta_2\big(\sqrt{\beta}\|X_0\|_2 + (2T+1)\|Y\|_2\big)S_{\Lambda,T}\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\frac{\sqrt{\beta}\Theta_L}{2\lambda_\ell} \\
&+ 2(1+\beta)(\|X_0\|_2 + \tfrac{2T-2}{\beta}\|\mathbf{M}^\top Y\|_2)\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j, \\
\overset{②}{\leq}&\tfrac{1}{4\eta\frac{\beta_0^2}{\beta^2}\delta_4}(1+\beta)\zeta_2\big(\sqrt{\beta}\|X_0\|_2 + (2T+1)\|Y\|_2\big)S_{\Lambda,T}\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\frac{\sqrt{\beta}\Theta_L}{2\lambda_\ell} \\
&+ \tfrac{1}{4}\alpha_0, \\
\overset{③}{\leq}&\tfrac{1}{2}\alpha_0,
\end{aligned} \tag{36}$$

where ① is due to NN's output's bound in Lemma A.6 and ② and ③ are due to the other lower bound for minimal singular value of NN's inner output in Equation (12a) and Equation (12d). The inequality in Equation (36) implies $\sigma_{\min}(G^{k+1}_{L-1}) \geq \frac{1}{2}\alpha_0$ since $\sigma_{\min}(G^0_{L-1}) = \alpha_0$.

Based on the above two inequalities, we prove the last linear rate step-by-step in the following sub-section.

A.5.2 Induction Part 2: Linear Convergence

In this section, we aim to prove that $F([W]^{k+1}) \leq (1 - \eta 4\eta\frac{\beta_0^2}{\beta^2}\delta_4)^{k+1}F([W]^0)$.

Step 1: Split Perfect Square Leveraging term $\mathbf{M}X^k_T$, we can split the perfect square in objective $F([W]^{k+1})$ as:

$$F([W]^{k+1}) = F([W]^k) + \tfrac{1}{2}\|\mathbf{M}X^{k+1}_T - \mathbf{M}X^k_T\|_2^2 + (\mathbf{M}X^{k+1}_T - \mathbf{M}X^k_T)^\top(\mathbf{M}X^k_T - Y). \tag{37}$$

Based on [27], we aim to demonstrate that $F([W]^{k+1})$ can be upperly bounded by $c_k F([W]^k)$, where $c_k < 1$ is a coefficient related to training iteration k.

Step 2: Bound Term-by-Term We aim to upperly bound all terms in Equation (37) by $F([W]^k)$.

Bound the first term $\frac{1}{2}\|\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k\|_2^2$. First, based on the β-smoothness of objective F, we calculate

$$\begin{aligned}\frac{1}{2}\|\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k\|_2^2 =& \frac{1}{2}(X_T^{k+1} - X_T^k)^\top \mathbf{M}^\top \mathbf{M}(X_T^{k+1} - X_T^k),\\ \leq& \frac{1}{2}\|X_T^{k+1} - X_T^k\|_2^2 \|\mathbf{M}^\top \mathbf{M}\|_2,\\ \leq& \frac{\beta}{2}\|X_T^{k+1} - X_T^k\|_2^2.\end{aligned}$$

The above inequality shows that we need to bound the distance between outputs of two iterations. Moreover, since our target is to construct linear convergence rate, we need to find the upper bound of above inequality w.r.t. the objective $F([W]^k)$, i.e., $\frac{1}{2}\|\mathbf{M}X_T^k - Y\|_2^2$. We apply Lemma 4.2 to derive the following lemma.

Lemma A.8. *Denote $\ell \in [L]$, for some $\bar{\lambda}_\ell \in \mathbb{R}$, we assume $\max(\|W_\ell^{k+1}\|_2, \|W_\ell^k\|_2) \leq \bar{\lambda}_\ell, \forall k$. Using quantities from Equation (11), we further define the following quantities with $i, j \in [T]$:*

$$\begin{aligned}\Lambda_i =& (1+\beta)\|X_0\|_2^2 + \big((4i-3)(1+\tfrac{1}{\beta}) + 1\big)\|X_0\|_2 \|\mathbf{M}^\top Y\|_2\\ &+ \tfrac{(2i-1)(\beta(2i-1)+(2i-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2,\\ \Phi_j =& \|X_0\|_2 + \tfrac{2j-1}{\beta}\|\mathbf{M}^\top Y\|_2,\\ {\delta_1}^T =& \Big(\textstyle\sum_{s=1}^T \big(\prod_{j=s+1}^T (1 + \frac{1+\beta}{2}\Theta_L \Phi_j)\big)\Big(\sum_{j=1}^s \Lambda_j\Big)\Big).\end{aligned}$$

We have the following upperly bounding property:

$$\frac{1}{2}\|\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k\|_2^2 \leq \frac{\beta^2\eta^2}{16}({\delta_1}^T)^2 \Big(S_{\Lambda,T}\Big)^2 \Big(\Theta_L^2 \textstyle\sum_{\ell=1}^L \bar{\lambda}_\ell^{-2}\Big)^2 \frac{1}{2}\|\mathbf{M}X_T^k - Y\|_2. \tag{38}$$

Proof. We calculate:

$$\begin{aligned}&\tfrac{1}{2}\|\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k\|_2^2 \leq \tfrac{\beta}{2}\|X_T^{k+1} - X_T^k\|_2^2,\\ \overset{①}{\leq}& \tfrac{\beta}{2}\Big(\textstyle\sum_{s=1}^T \Big(\prod_{j=s+1}^T (1 + \frac{1+\beta}{2}\Theta_L \Phi_j)\Big)\frac{1}{2}\Lambda_s \Theta_L \sum_{\ell=1}^L \bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^k\|_2\Big)^2,\\ \overset{②}{=}& \tfrac{\beta\eta^2}{2}\Big(\textstyle\sum_{s=1}^T \Big(\prod_{j=s+1}^T (1 + \frac{1+\beta}{2}\Theta_L \Phi_j)\Big)\frac{1}{2}\Lambda_s \Theta_L \sum_{\ell=1}^L \bar{\lambda}_\ell^{-1}\Big\|\frac{\partial F}{\partial W_\ell^k}\Big\|_2\Big)^2,\\ \overset{③}{\leq}& \tfrac{\beta\eta^2}{2}\Big(\textstyle\sum_{s=1}^T \Big(\prod_{j=s+1}^T (1 + \frac{1+\beta}{2}\Theta_L \Phi_j)\Big)\frac{1}{2}\Lambda_s \Theta_L \sum_{\ell=1}^L \bar{\lambda}_\ell^{-1}\frac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell}\Big(S_{\Lambda,T}\Big)\|\mathbf{M}X_T^k - Y\|_2\Big)^2,\\ =& \tfrac{\beta^2\eta^2}{32}\Bigg(\underbrace{\Big(\textstyle\sum_{s=1}^T (\prod_{j=s+1}^T (1 + \frac{1+\beta}{2}\Theta_L \Phi_j))\Lambda_s\Big)\Big(S_{\Lambda,T}\Big)\Theta_L^2 \sum_{\ell=1}^L \bar{\lambda}_\ell^{-2}}_{{\delta_1}^T}\|\mathbf{M}X_T^k - Y\|_2\Bigg)^2,\\ =& \tfrac{\beta^2\eta^2}{16}({\delta_1}^T)^2\Big(S_{\Lambda,T}\Big)^2\Big(\Theta_L^2 \textstyle\sum_{\ell=1}^L \bar{\lambda}_\ell^{-2}\Big)^2 \frac{1}{2}\|\mathbf{M}X_T^k - Y\|_2,\end{aligned} \tag{39}$$

① is from semi-smoothness of L2O's output in Lemma 4.2, Appendix A.4.3. ② is due to gradient descent with learning rate η. ③ is from gradient bounds in Lemma 4.1. □

Bound the second term $(\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k)^\top(\mathbf{M}X_T^k - Y)$. We calculate:

$$\begin{aligned}&(\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k)^\top(\mathbf{M}X_T^k - Y)\\ =&(X_T^{k+1} - X_T^k)^\top \mathbf{M}^\top(\mathbf{M}X_T^k - Y),\\ =&(X_T^{k+1} - X_T^k)^\top \mathbf{M}^\top(\mathbf{M}X_T^k - Y).\end{aligned} \tag{40}$$

Inspired by the method in [27], we stabilize all other learnable parameters and concentrate on gradient on last layers W_L to construct a non-singular NTK, which invokes the PL-condition for a linear convergence rate.

Given last NN layer's learnable parameter W_L^{k+1} at iteration $k+1$, due to the GD formulation in Equation (20), we define the following quantity:

$$Z = X_{T-1}^k - \tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^k)^\top)\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y), \tag{41}$$

where $G_{L-1,T}^k$ represents inner output of layer $L-1$ at training iteration k.

With Z, we reformulate Equation (40) as:

$$\begin{aligned}
&(X_T^{k+1} - X_T^k)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y),\\
=&(X_T^{k+1} - Z + Z - X_T^k)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y),\\
=&(X_T^{k+1} - Z)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y) + (Z - X_T^k)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y),
\end{aligned} \tag{42}$$

where X_T^{k+1} at training iteration $k+1$ with W_L^{k+1} and solution X_T^k at training iteration k with W_L^k are defined as:

$$X_T^{k+1} = X_{T-1}^{k+1} - \tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)\mathbf{M}^\top(\mathbf{M}X_{T-1}^{k+1} - Y).$$

$$X_T^k = X_{T-1}^k - \tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^kG_{L-1,T}^k)^\top)\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y).$$

Then, we have following lemmas to bound the two terms, respectively:

Lemma A.9. *Denote $\ell \in [L]$, for some $\bar{\lambda}_\ell, \in \mathbb{R}$ with $j \in [T]$, we assume $\max(\|W_\ell^{k+1}\|_2, \|W_\ell^k\|_2) \leq \bar{\lambda}_\ell$. Define the following quantities with $t \in [T]$:*

$$\begin{aligned}
\Lambda_t =&(1+\beta)\|X_0\|_2^2 + \big((4t-3)(1+\tfrac{1}{\beta}) + 1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2\\
&+ \tfrac{(2T-1)(\beta(2T-1)+(2T-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2,\\
\Phi_j =&\|X_0\|_2 + \tfrac{2j-1}{\beta}\mathbf{M}^\top Y\|_2,\\
\Theta_L =&\Theta_L,\\
\delta_2 =&{\textstyle\sum_{s=1}^{T-1}}\Big(\textstyle\prod_{j=s+1}^{T}\big(1 + \tfrac{1+\beta}{2}\Theta_L\Phi_j\big)\Big)\Lambda_s.
\end{aligned}$$

We have the following upperly bounding property:

$$(X_T^{k+1} - Z)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y) \leq \tfrac{\beta\eta}{2}(\Lambda_T + \delta_2)\Theta_L^2 S_{\bar{\lambda},L}S_{\Lambda,T}\tfrac{1}{2}\|\mathbf{M}X_T^k - Y\|_2^2.$$

Proof. We straightforwardly apply upper bound relaxation in this part, where we reuse the results of the first term $\frac{1}{2}\|\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k\|_2^2$'s upper bound in Lemma A.8.

To reuse the results, we would like to construct the $X_{T-1}^{k+1} - X_{T-1}^k$ term. We substitute Equation (44) into above equation and use the Cauchy-Schwarz inequality for vectors to split our bounding targets into two parts and relax the L_2-norm of vector summations into each element by triangle inequalities:

$$
\begin{aligned}
&(X_T^{k+1}-Z)^\top \mathbf{M}^\top(\mathbf{M}X_T^k-Y) \\
=&\Big(X_{T-1}^{k+1}-\tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)\mathbf{M}^\top(\mathbf{M}X_{T-1}^{k+1}-Y) \\
&\quad -\big(X_{T-1}^{k}-\tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k})^\top)\mathbf{M}^\top(\mathbf{M}X_{T-1}^{k}-Y)\big)\Big)^\top \mathbf{M}^\top(\mathbf{M}X_T^k-Y), \\
\overset{①}{\leq}&\Big(\Big\|\big(\mathbf{I}_d-\tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)\mathbf{M}^\top\mathbf{M}\big)X_{T-1}^{k+1} \\
&\qquad -\big(\mathbf{I}_d-\tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k})^\top)\mathbf{M}^\top\mathbf{M}\big)X_{T-1}^{k}\Big\|_2 \\
&\quad +\tfrac{1}{\beta}\Big\|\Big(\underbrace{\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)-\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k})^\top)}_{C_{k+1}}\Big)\mathbf{M}^\top Y\Big\|_2\Big) \\
&\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2, \\
\overset{②}{\leq}&\Big(\Big\|\big(\mathbf{I}_d-\tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)\mathbf{M}^\top\mathbf{M}\big)(X_{T-1}^{k+1}-X_{T-1}^{k})\Big\|_2 \\
&\quad +\Big\|\Big(\big(\mathbf{I}_d-\tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)\mathbf{M}^\top\mathbf{M}\big) \\
&\qquad -\big(\mathbf{I}_d-\tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k})^\top)\mathbf{M}^\top\mathbf{M}\big)\Big)X_{T-1}^{k}\Big\|_2 \\
&\quad +\tfrac{1}{\beta}\|C_{k+1}\mathbf{M}^\top Y\|_2\Big)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2, \\
\overset{③}{\leq}&\Big(\Big\|\big(\mathbf{I}_d-\tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)\mathbf{M}^\top\mathbf{M}\big)\Big\|_2\|X_{T-1}^{k+1}-X_{T-1}^{k}\|_2 \\
&\quad +\|\tfrac{1}{\beta}C_{k+1}\mathbf{M}^\top\mathbf{M}\|_2\|X_{T-1}^{k}\|_2+\tfrac{1}{\beta}\|C_{k+1}\|_2\|\mathbf{M}^\top Y\|_2\Big)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2, \\
\overset{④}{\leq}&\Big(\|X_{T-1}^{k+1}-X_{T-1}^{k}\|_2+\|X_{T-1}^{k}\|_2\|C_{k+1}\|_2+\tfrac{1}{\beta}\|\mathbf{M}^\top Y\|_2\|C_{k+1}\|_2\Big)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2, \\
\overset{⑤}{\leq}&\Big(\|X_{T-1}^{k+1}-X_{T-1}^{k}\|_2+\big(\|X_0\|_2+\tfrac{2T-1}{\beta}\|\mathbf{M}^\top Y\|_2\big)\|C_{k+1}\|_2\Big)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,
\end{aligned}
\tag{43}
$$

where ① is due to triangle and Cauchy-Schwarz inequalities. ② is due to triangle inequality. ③ is due to Cauchy-Schwarz inequality. ④ is due to β-smooth definition that $\mathbf{M}^\top\mathbf{M}\leq\beta$ and $\|\mathbf{I}_d-\frac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)\mathbf{M}^\top\mathbf{M}\|_2\leq 1$ in Lemma A.1. ④ is due to the upper bound of X_{T-1} in Lemma A.6.

Further, we bound $C_{k+1}:=\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)-\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k})^\top)$. We apply the Mean Value Theorem and assume a point v_1^k. For v_1^k's each entry $(v_1^k)_i$, for some $\alpha_{1i}^k\in[0,1]$, we calculate $(v_1^k)_i$ as:

$$
(v_1^k)_i=\alpha_{1i}^k((W_L^{k+1}G_{L-1,T}^{k+1})^\top)_i+(1-\alpha_{1i}^k)((W_L^{k+1}G_{L-1,T}^{k})^\top)_i.
$$

Then, we can represent quantity $\|C_{k+1}\|_2$ by:

$$
\begin{aligned}
&\|\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k+1})^\top)-\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^{k})^\top)\|_2 \\
\overset{①}{\leq}&\Big\|\tfrac{\partial 2\sigma}{\partial v_1^k}\odot(W_L^{k+1}G_{L-1,T}^{k+1}-W_L^{k+1}G_{L-1,T}^{k})^\top\Big\|_\infty, \\
\overset{②}{\leq}&\tfrac{1}{2}\Big\|(W_L^{k+1}G_{L-1,T}^{k+1}-W_L^{k+1}G_{L-1,T}^{k})^\top\Big\|_\infty, \\
\overset{③}{\leq}&\tfrac{1}{2}\|W_L^{k+1}\|_2\|G_{L-1,T}^{k+1}-G_{L-1,T}^{k}\|_2\leq\tfrac{1}{2}\bar{\lambda}_L\|G_{L-1,T}^{k+1}-G_{L-1,T}^{k}\|_2,
\end{aligned}
$$

where ① is from the Mean Value Theorem. ② is from the gradient upper bound of Sigmoid function. ③ is from triangle inequality and definition of learnable parameter W_L.

We further substitute the upper bound of $\|G_{L-1,T}^{k+1} - G_{L-1,T}^{k}\|_2$ in Lemma A.4 and calculate:

$$
\begin{aligned}
&\tfrac{1}{2}\bar{\lambda}_L\|G_{L-1,T}^{k+1} - G_{L-1,T}^{k}\|_2 \\
\leq&\tfrac{1}{2}\bar{\lambda}_L\Big((1+\beta)\|X_{T-1}^{k+1} - X_{T-1}^{k}\|_2\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j \\
&\quad + (\|X_{T-1}^{k}\|_2 + \|\mathbf{M}^\top(\mathbf{M}X_{T-1}^{k} - Y)\|_2)\textstyle\prod_{j=1}^{L-1}\bar{\lambda}_j\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^{k}\|_2\Big) \\
\overset{①}{\leq}&\tfrac{1}{2}(1+\beta)\Theta_L\|X_{T-1}^{k+1} - X_{T-1}^{k}\|_2 \\
&+ \tfrac{1}{2}\Big((1+\beta)\|X_0\|_2 + (2T-1+\tfrac{2T-2}{\beta})\|\mathbf{M}^\top Y\|_2\Big)\Theta_L\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^{k}\|_2.
\end{aligned}
$$

where ① is due to upper bound of X_{T-1} in Lemma A.6.
Substituting the above inequality back into Equation (43) yields:

$$
\begin{aligned}
&(X_T^{k+1} - Z)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y) \\
\leq&\Big(\|X_{T-1}^{k+1} - X_{T-1}^{k}\|_2 + \big(\|X_0\|_2 + \tfrac{2T-1}{\beta}\|\mathbf{M}^\top Y\|_2\big)\|C_{k+1}\|_2\Big)\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2, \\
\leq&\Bigg(\|X_{T-1}^{k+1} - X_{T-1}^{k}\|_2 \\
&\quad + \big(\|X_0\|_2 + \tfrac{2T-1}{\beta}\|\mathbf{M}^\top Y\|_2\big) \\
&\Big(\tfrac{1}{2}(1+\beta)\Theta_L\|X_{T-1}^{k+1} - X_{T-1}^{k}\|_2 \\
&+ \tfrac{1}{2}\big((1+\beta)\|X_0\|_2 + (2T-1+\tfrac{2T-2}{\beta})\|\mathbf{M}^\top Y\|_2\big)\Theta_L\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^{k}\|_2\Big)\Bigg) \\
&\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2, \\
=&\Bigg(\big(1 + \tfrac{1+\beta}{2}\Theta_L(\|X_0\|_2 + \tfrac{2T-1}{\beta}\|\mathbf{M}^\top Y\|_2)\big)\|X_{T-1}^{k+1} - X_{T-1}^{k}\|_2 \\
&\quad + \Big(\tfrac{1}{2}\big((1+\beta)\|X_0\|_2 + (2T-1+\tfrac{2T-2}{\beta})\|\mathbf{M}^\top Y\|_2\big) \\
&\qquad (\|X_0\|_2 + \tfrac{2T-1}{\beta}\|\mathbf{M}^\top Y\|_2)\Theta_L\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^{k}\|_2\Big)\Bigg) \\
&\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2, \\
=&\Bigg(\big(1 + \tfrac{1+\beta}{2}\Theta_L(\underbrace{\|X_0\|_2 + \tfrac{2T-1}{\beta}\|\mathbf{M}^\top Y\|_2}_{\Phi_T})\big)\|X_{T-1}^{k+1} - X_{T-1}^{k}\|_2 + \\
&\Big(\tfrac{1}{2}\underbrace{(1+\beta)\|X_0\|_2^2 + \big((4T-3)(1+\tfrac{1}{\beta})+1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2 + \tfrac{(2T-1)(\beta(2T-1)+(2T-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2}_{\Lambda_T}\Big) \\
&\qquad \Theta_L\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^{k}\|_2\Bigg)\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2, \\
=&\Big(\big(1 + \tfrac{1+\beta}{2}\Theta_L\Phi_T\big)\|X_{T-1}^{k+1} - X_{T-1}^{k}\|_2 + \tfrac{1}{2}\Lambda_T\Theta_L\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1} - W_\ell^{k}\|_2\Big) \\
&\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2,
\end{aligned}
$$

Further, we apply semi-smoothness of L2O model in Lemma 4.2 and upper bound of gradient in Lemma 4.1 to derive the upper bound. We calculate:

$$
\begin{aligned}
&(X_T^{k+1}-Z)^\top \mathbf{M}^\top(\mathbf{M}X_T^k-Y)\\
\leq&\Big(\big(1+\tfrac{1+\beta}{2}\Theta_L\Phi_T\big)\|X_{T-1}^{k+1}-X_{T-1}^k\|_2+\tfrac{1}{2}\Lambda_T\Theta_L\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1}-W_\ell^k\|_2\Big)\\
&\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,\\
\overset{①}{\leq}&\Big(\big(1+\tfrac{1+\beta}{2}\Theta_L\Phi_T\big)\tfrac{1}{2}\Theta_L\textstyle\sum_{s=1}^{T-1}\Big(\prod_{j=s+1}^{T-1}\big(1+\tfrac{1+\beta}{2}\Theta_L\Phi_j\big)\Big)\Lambda_s\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1}-W_\ell^k\|_2\\
&+\tfrac{1}{2}\Lambda_T\Theta_L\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1}-W_\ell^k\|_2\Big)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,\\
\leq&\Bigg(\tfrac{1}{2}\Theta_L\underbrace{\textstyle\sum_{s=1}^{T-1}\Big(\prod_{j=s+1}^{T}\big(1+\tfrac{1+\beta}{2}\Theta_L\Phi_j\big)\Big)\Lambda_s}_{\delta_2}\textstyle\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1}-W_\ell^k\|_2\\
&\quad+\tfrac{1}{2}\Lambda_T\Theta_L\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1}-W_\ell^k\|_2\Bigg)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,\\
=&\tfrac{1}{2}\Theta_L\Big(\delta_2\bar{\lambda}_L^{-1}\|W_L^{k+1}-W_L^k\|_2+(\Lambda_T+\delta_2)\textstyle\sum_{\ell=1}^{L-1}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1}-W_\ell^k\|_2\Big)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,\\
\overset{②}{\leq}&\tfrac{1}{2}\Theta_L(\Lambda_T+\delta_2)\textstyle\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1}-W_\ell^k\|_2\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,
\end{aligned}
$$

where ① is due to Lemma 4.2. ② is due to $\Lambda_T\geq 0$.

Further, based on the gradient descent, i.e., $W_\ell^{k+1}=W_\ell^k-\eta\frac{\partial F}{\partial W_\ell^k}$, we substitute the bound of gradient in Lemma 4.1 and calculate:

$$
\begin{aligned}
&(X_T^{k+1}-Z)^\top \mathbf{M}^\top(\mathbf{M}X_T^k-Y)\\
\leq&\tfrac{1}{2}\Theta_L(\Lambda_T+\delta_2)\textstyle\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\|W_\ell^{k+1}-W_\ell^k\|_2\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,\\
\leq&\tfrac{\eta}{2}\Theta_L(\Lambda_T+\delta_2)\textstyle\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\Big\|\tfrac{\partial F}{\partial W_\ell^k}\Big\|_2\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,\\
\overset{①}{\leq}&\tfrac{\eta}{2}\Theta_L(\Lambda_T+\delta_2)\textstyle\sum_{\ell=1}^{L}\bar{\lambda}_\ell^{-1}\tfrac{\sqrt{\beta}\Theta_L}{2\bar{\lambda}_\ell}S_{\Lambda,T}\|\mathbf{M}X_T^k-Y\|_2\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2,\\
\overset{②}{\leq}&\tfrac{\beta\eta}{2}(\Lambda_T+\delta_2)\Theta_L^2S_{\bar{\lambda},L}S_{\Lambda,T}\tfrac{1}{2}\|\mathbf{M}X_T^k-Y\|_2^2,
\end{aligned}
$$

where ① is due to Lemma 4.1 and ② is due to $\|M\|_2\leq\sqrt{\beta}$. □

Lemma A.10. *Define the following quantities with $t\in[T]$:*

$$
\begin{aligned}
\Lambda_t=&(1+\beta)\|X_0\|_2^2+\big((4t-3)(1+\tfrac{1}{\beta})+1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2\\
&+\tfrac{(2T-1)(\beta(2T-1)+(2T-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2,\\
\Phi_j=&\|X_0\|_2+\tfrac{2j-1}{\beta}\mathbf{M}^\top Y\|_2,\\
\Theta_L=&\Theta_L,\\
\delta_3=&\big((1+\beta)\|X_0\|_2+\big(2T-1+\tfrac{2T-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big).
\end{aligned}
$$

We have the following upperly bounding property:

$$
\begin{aligned}
&(Z-X_T^k)^\top\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\\
\leq&\Big(-\eta 8\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\tfrac{\beta_0^2}{\beta^2}\alpha_0^2+\tfrac{\eta\beta}{2}\Theta_{L-1}^2\Lambda_T\textstyle\sum_{t=1}^{T-1}\Lambda_t\Big)\tfrac{1}{2}\|\mathbf{M}X_T^k-Y\|_2^2.
\end{aligned}
$$

Proof. In our above demonstrations, we have construct a non-negative coefficient of the upper bound w.r.t. the objective $\frac{1}{2}\|\mathbf{M}X_T^k-Y\|_2^2$. To achieve the requirement of the linear convergence rate, we would like a negative one from our remaining bounding target. We calculate:

$$
\begin{aligned}
&(Z - X_T^k)^\top \mathbf{M}^\top (\mathbf{M}X_T^k - Y) \\
=&\Big(X_{T-1}^k - \tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^{k+1}G_{L-1,T}^k)^\top)\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big) \\
&\quad - \Big(X_{T-1}^k - \tfrac{1}{\beta}\mathcal{D}(2\sigma(W_L^k G_{L-1,T}^k)^\top)\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\Big)\Big)^\top \mathbf{M}^\top(\mathbf{M}X_T^k - Y), \\
=& -\tfrac{1}{\beta}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)^\top \mathcal{D}\big(2\sigma(W_L^{k+1}G_{L-1,T}^k)^\top - 2\sigma(W_L^k G_{L-1,T}^k)^\top\big) \\
&\quad \big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big).
\end{aligned} \tag{44}
$$

Similarly, due to Mean Value Theorem, suppose $v_{2,i}^k = \alpha_i(W_L^{k+1}G_{L-1,T}^k)_i + (1-\alpha_i)(W_L^k G_{L-1,T}^k)_i$, $v_{2,i}^k \in [0,1]$, based on Mean Value Theorem, we calculate:

$$
2\sigma(W_L^{k+1}G_{L-1,T}^k)_i^\top - 2\sigma(W_L^k G_{L-1,T}^k)_i^\top = \tfrac{\partial(2\sigma(v_{2,i}^k))}{\partial(v_{2,i}^k)_i}(W_L^{k+1}G_{L-1,T}^k)_i - (W_L^k G_{L-1,T}^k)_i.
$$

Denote $v_{2,i}^k := [\frac{\partial(2\sigma(v_{2,i}^k))}{\partial(v_{2,i}^k)_i}]$, we calculate:

$$
\begin{aligned}
&\mathcal{D}\big(2\sigma(W_L^{k+1}G_{L-1,T}^k)^\top - 2\sigma(W_L^k G_{L-1,T}^k)^\top\big) \\
=&\mathcal{D}\Big(\Big[\tfrac{\partial 2\sigma(v_{2,i}^k)}{\partial v_{2,i}^k}((W_L^{k+1}G_{L-1,T}^k)_i - (W_L^k G_{L-1,T}^k)_i)\Big]^\top\Big), \\
=&\mathcal{D}\Big(\Big[2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))((W_L^{k+1} - W_L^k)G_{L-1,T}^k)_i\Big]^\top\Big), \\
=&\mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\mathcal{D}\Big(\big((W_L^{k+1} - W_L^k)G_{L-1,T}^k\big)^\top\Big), \\
\overset{①}{=}& -\eta\mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\mathcal{D}\Big(\tfrac{\partial F}{\partial W_L^k}{G_{L-1,T}^k}^\top\Big),
\end{aligned}
$$

where $v_{2\,i}^k := \alpha_i(W_L^{k+1}G_{L-1,T}^k)_i + (1-\alpha)_i(W_L^k G_{L-1,T}^k)_i$ is an interior point between the corresponding entries of $W_L^{k+1}G_{L-1,T}^k$ and $W_L^k G_{L-1,T}^k$. ① is from gradient descent formulation of W_L^k in Equation (7).

Substituting above into Equation (44) yields:

$$
\begin{aligned}
&(Z - X_T^k)^\top \mathbf{M}^\top(\mathbf{M}X_T^k - Y) \\
=&\tfrac{\eta}{\beta}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)^\top \mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\mathcal{D}(\tfrac{\partial F}{\partial W_L^k}G_{L-1,T}^k)^\top\big(\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\big), \\
=&\tfrac{\eta}{\beta}\tfrac{\partial F}{\partial W_L^k}G_{L-1,T}^k \mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\big(\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\big),
\end{aligned}
$$

Further, we substitute the gradient formulation in Equation (7) and calculate:

$$
\begin{aligned}
&(Z - X_T^k)^\top \mathbf{M}^\top(\mathbf{M}X_T^k - Y) \\
=& -\tfrac{\eta}{\beta^2}\textstyle\sum_{t=1}^T \big(\mathbf{M}^\top(\mathbf{M}X_T - Y)\big)^\top \Big(\prod_{j=T}^{t+1}\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_j)\mathbf{M}^\top\mathbf{M}\Big) \\
&\quad \mathcal{D}\big((\mathbf{M}^\top(\mathbf{M}X_{t-1} - Y))\big)\mathcal{D}\big(P_t \odot (1 - P_t/2)\big){G_{L-1,t}}^\top G_{L-1,T}^k \\
&\quad \mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\big(\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\big), \\
=& -\tfrac{\eta}{\beta^2}(\mathbf{M}X_T^k - Y)^\top \mathbf{M}\mathbf{B}_T^k \mathbf{M}^\top(\mathbf{M}X_T^k - Y),
\end{aligned} \tag{45}
$$

where $\mathbf{B}_T^k$ is defined by:

$$
\begin{aligned}
&\mathbf{B}_T^k \\
=&\textstyle\sum_{t=1}^T \big(\prod_{j=T}^{t+1}\mathbf{I} - \tfrac{1}{\beta}\mathcal{D}(P_j^k)\mathbf{M}^\top\mathbf{M}\big)\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big){G_{L-1,t}^k}^\top \\
&\quad G_{L-1,T}^k \mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big).
\end{aligned}
$$

We discuss the definite property of $\mathbf{B}_T^k$ case-by-case.

Case 1: $t = T$. $\prod_{j=T}^{T+1}\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_j)\mathbf{M}^\top\mathbf{M}$ degenerates to be 1. The Equation (45) becomes:

$$\begin{aligned}
&[(Z - X_T^k)^\top \mathbf{M}^\top(\mathbf{M}X_T^k - Y)]_{\text{Part 1}} \\
=& -\tfrac{\eta}{\beta^2}(\mathbf{M}X_T^k - Y)^\top\mathbf{M} \\
&\quad \mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big) \\
&\quad \mathcal{D}\big(P_T^k \odot (1 - P_T^k/2)\big) \\
&\quad {G_{L-1,T}^k}^\top G_{L-1,T}^k \\
&\quad \mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big) \\
&\quad \mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\mathbf{M}^\top(\mathbf{M}X_T^k - Y),
\end{aligned} \tag{46}$$

We first present the following corollary to show that there exists a negative upper bound of $[(Z - X_T^k)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y)]_{\text{Part 1}}$:

Corollary A.11. *RHS of Equation* (46) < 0 *if* $\lambda_{\min}({G_{L-1,T}^k}^\top G_{L-1,T}^k) > 0$.

Proof. Due to definition of eigenvalue and Cauchy-Schwarz inequality, we calculate:

$$\begin{aligned}
&(\mathbf{M}X_T^k - Y)^\top\mathbf{M} \\
&\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big) \\
&\mathcal{D}\big(P_T^k \odot (1 - P_T^k/2)\big){G_{L-1,T}^k}^\top G_{L-1,T}^k\mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big) \\
&\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\mathbf{M}^\top(\mathbf{M}X_T^k - Y), \\
\geq& \big(P_T^k \odot (1 - P_T^k/2)\big)_{\min}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)_{\min} \\
&\quad \lambda_{\min}({G_{L-1,T}^k}^\top G_{L-1,T}^k)\lambda_{\min}(\mathbf{M}\mathbf{M}^\top)\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2^2, \\
\overset{①}{>}& 0,
\end{aligned}$$

where ① is due to Sigmoid function is non-negative, $\lambda_{\min}({G_{L-1,T}^k}^\top G_{L-1,T}^k) > 0$, and $\lambda_{\min}(\mathbf{M}\mathbf{M}^\top) > 0$ by definition. Thus, $(Z - X_T^k)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y) < 0$ by nature. $()_{\min}$ means the minimal value among all entries. ☐

To get a upper bound, we expect ${G_{L-1,T}^k}^\top G_{L-1,T}^k$ to be positive definition, in which we require $n_{L-1} \geq N$. Thus, we can easily get the upper bound from its minimal eigenvalue.

Based on corollary Corollary A.11, we calculate the negative lower bound of Equation (46) by:

$$\begin{aligned}
&(Z - X_T^k)^\top\mathbf{M}^\top(\mathbf{M}X_T^k - Y) \\
\leq& -\tfrac{\eta}{\beta^2}\big(P_T^k \odot (1 - P_T^k/2)\big)_{\min}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)_{\min} \\
&\quad \lambda_{\min}({G_{L-1,T}^k}^\top G_{L-1,T}^k)\lambda_{\min}(\mathbf{M}\mathbf{M}^\top)\|\mathbf{M}^\top(\mathbf{M}X_T^k - Y)\|_2^2,
\end{aligned} \tag{47}$$

The remaining task is to calculate $\big(P_T^k \odot (1 - P_T^k/2)\big)_{\min}$ and $\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)_{\min}$. We achieve that by calculating the values on the boundary of close sets.

First, denote $v_3^k := W_L^k G_{L-1,T}^k$, we represent $P_T^k \odot (1 - P_T^k/2)$ by:

$$P_T^k \odot (1 - P_T^k/2) = 2\sigma(v_3^k)^\top \odot (1 - \sigma(v_3^k))^\top.$$

Since the Sigmoid function is a coordinate-wise non-decreasing function, we can straightforwardly find $\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)_{\min}$ and $(2\sigma(v_3^k)^\top \odot (1 - \sigma(v_3^k))^\top)_{\min}$ by on the closed sets of v_2^k and v_3^k, respectively, which is achieved by the following lemma.

Lemma A.12. *For some $b, B \in \mathbb{R}^{k}$[1], $\forall v^k, b \le v^k \le B$, we calculate $(2\sigma(v^k)^\top \odot (1-\sigma(v^k))^\top)_{\min}$ by:*

$$(2\sigma(v^k)^\top \odot (1-\sigma(v^k))^\top)_{\min} = \begin{cases} \min\left(2\sigma(b)(1-\sigma(b))^\top, 2\sigma(B)(1-\sigma(B))^\top\right) & -b \neq B, \\ 2\sigma(B)(1-\sigma(B)) & -b = B. \end{cases}$$

Proof. Since $\sigma(x) \in (0,1)\forall x$, $\mathcal{D}(2\sigma(x) \odot (1-\sigma(x)))$ is a quadratic function w.r.t. x. Since $\sigma(x) \in (0,1)\forall x$, $\mathcal{D}(2\sigma(x) \odot (1-\sigma(x))) > 0$. Since the coefficient before the x^2 term is negative, its lower bound is either the value on the boundary or 0.

Since $\sigma(b), \sigma(B) \in (0,1)$, if $-b \neq B$, the lower bound is the smaller one, i.e., $\min(2\sigma(b) \odot (1-\sigma(b)), 2\sigma(B) \odot (1-\sigma(B)))$. Otherwise, since both $\sigma(x)$ and $\mathcal{D}(2\sigma(x) \odot (1-\sigma(x)))$ are symmetric around $\frac{1}{2}$, we have $2\sigma(B) \odot (1-\sigma(B)) = 2\sigma(b) \odot (1-\sigma(b))$. ☐

Further, we calculate the bounds of v_2^k and v_3^k and invoke Lemma A.12 to get $\left([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\right)_{\min}$ and $(2\sigma(v_3^k)^\top \odot (1-\sigma(v_3^k))^\top)_{\min}$.

We present the following two lemmas to show the close sets that v_2^k and v_3^k belong to.

Lemma A.13. *Denote $\ell \in [L]$, for some $\bar{\lambda}_\ell \in \mathbb{R}$, we assume $\|W_\ell^k\|_2 \le \bar{\lambda}_\ell$. We define the following quantity:*

$$\delta_3 = \left((1+\beta)\|X_0\|_2 + \left(2T - 1 + \tfrac{2T-2}{\beta}\right)\|\mathbf{M}^\top Y\|_2\right),$$
$$\Theta_L = \textstyle\prod_{\ell=1}^{L} \bar{\lambda}_\ell.$$

For $v_{2\,i}^k := \alpha_i (W_L^{k+1} G_{L-1,T}^k)_i + (1-\alpha_i)(W_L^k G_{L-1,T}^k)_i, \alpha_i \in [0,1]$, v_2^k belongs to the following close set:

$$v_2^k \in [-\delta_3\Theta_L, \delta_3\Theta_L].$$

Proof. We calculate v_2^k's upper bound by:

$$\begin{aligned}
\|v_2^k\|_\infty =& \|\alpha \odot (W_L^{k+1} G_{L-1,T}^k) + (1-\alpha) \odot (W_L^k G_{L-1,T}^k)\|_\infty, \\
=& \max_i \|\alpha_i (W_L^{k+1} G_{L-1,T}^k)_i + (1-\alpha_i)(W_L^k G_{L-1,T}^k)_i\|_\infty, \\
\overset{①}{\le}& \max_i \alpha_i \|(W_L^{k+1} G_{L-1,T}^k)_i\|_\infty + (1-\alpha_i)\|(W_L^k G_{L-1,T}^k)_i\|_\infty, \\
\overset{②}{\le}& \max_i \max(\|(W_L^{k+1} G_{L-1,T}^k)_i\|_\infty, \|(W_L^k G_{L-1,T}^k)_i\|_\infty), \\
=& \max(\max_i \|(W_L^{k+1} G_{L-1,T}^k)_i\|_\infty, \max_i \|(W_L^k G_{L-1,T}^k)_i\|_\infty), \\
\le& \max(\|W_L^{k+1} G_{L-1,T}^k\|_\infty, \|W_L^k G_{L-1,T}^k\|_\infty),
\end{aligned} \tag{48}$$

where ① is due to triangle inequality and ② is due to $\alpha_i \in [0,1]$ and upper bound of NN's inner output in Lemma A.5.

We calculate the bound of $\|W_L^{k+1} G_{L-1,T}^k\|_2$ by:

$$\begin{aligned}
\|W_L^{k+1} G_{L-1,T}^k\|_\infty \overset{①}{\le}& \|W_L^{k+1}\|_2 \|G_{L-1,T}^k\|_2, \\
\overset{②}{\le}& \bar{\lambda}_L \left((1+\beta)\|X_0\|_2 + \left(2T-1+\tfrac{2T-2}{\beta}\right)\|\mathbf{M}^\top Y\|_2\right) \textstyle\prod_{\ell=1}^{L-1} \bar{\lambda}_\ell, \\
=& \underbrace{\left((1+\beta)\|X_0\|_2 + \left(2T-1+\tfrac{2T-2}{\beta}\right)\|\mathbf{M}^\top Y\|_2\right)}_{\delta_3} \underbrace{\textstyle\prod_{\ell=1}^{L} \bar{\lambda}_\ell}_{\Theta_L},
\end{aligned}$$

where ① is due to Cauchy-Schwarz inequality and ② is due to definition and upper bound of NN's inner output in Lemma A.5. Similarly, we can get $\|W_L^{k+1} G_{L-1,T}^k\|_2 \le \delta_3 \Theta_L$.

[1] $\mathbb{R}^k$ means the space at training iteration k.

Substituting back to Equation (48) yields:

$$\|v_2^k\|_\infty \le \delta_3\Theta_L.$$

Thus, we have the following bound for vector v_2^k by nature:

$$-\delta_3\Theta_L \le v_2^k \le \delta_3\Theta_L.$$

It is note-worthing that the above lower bound is non-trivial since we cannot have $v_2^k \ge 0$, which can be easily violated by a little perturbation from gradient descent.

□

Lemma A.14. *Denote $\ell \in [L]$, for some $\bar{\lambda}_\ell \in \mathbb{R}$, we assume $\|W_\ell^k\|_2 \le \bar{\lambda}_\ell$. We define the following quantity:*

$$\begin{aligned}\delta_3 &= \big((1+\beta)\|X_0\|_2 + \big(2T-1+\tfrac{2T-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big),\\ \Theta_L &= \textstyle\prod_{\ell=1}^{L}\bar{\lambda}_\ell.\end{aligned}$$

For $v_3^k := W_L^k G_{L-1,T}^k, \forall k$, v_3^k belongs to the following close set:

$$v_3^k \in [-\delta_3\Theta_L, \delta_3\Theta_L].$$

Proof. We prove the lemma by a similar method. We calculate the bound of $\|W_L^k G_{L-1,T}^k\|_2$ by:

$$\begin{aligned}\|v_3^k\|_\infty =&\|W_L^k G_{L-1,T}^k\|_\infty\\ \overset{①}{\le}&\|W_L^k\|_2\|G_{L-1,T}^k\|_2,\\ \overset{②}{\le}&\bar{\lambda}_L\big((1+\beta)\|X_0\|_2 + \big(2T-1+\tfrac{2T-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big)\textstyle\prod_{\ell=1}^{L-1}\bar{\lambda}_\ell,\\ =&\underbrace{\big((1+\beta)\|X_0\|_2 + \big(2T-1+\tfrac{2T-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big)}_{\delta_3}\underbrace{\textstyle\prod_{\ell=1}^{L}\bar{\lambda}_\ell}_{\Theta_L},\end{aligned}$$

where ① is due to Cauchy-Schwarz inequality and ② is due to definition and upper bound of NN's inner output in Lemma A.5.

We have the following bound for v_3^k by nature:

$$-\delta_3\Theta_L \le v_3^k \le \delta_3\Theta_L.$$

□

We calculate $\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)_{\min}$ by substituting Lemma A.13 into Lemma A.12:

$$\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)_{\min} = 2\sigma(\delta_3\Theta_L)(1-\sigma(\delta_3\Theta_L)).$$

Similarly, we get $\big(P_T^k \odot (1-P_T^k/2)\big)$ by substituting Lemma A.14 into Lemma A.12:

$$\big(P_T^k \odot (1-P_T^k/2)\big)_{\min} = 2\sigma(\delta_3\Theta_L)(1-\sigma(\delta_3\Theta_L)).$$

Substituting the above results into Equation (47) and Equation (46) yields:

$$\begin{aligned}&[(Z-X_T^k)^\top\mathbf{M}^\top(\mathbf{M}X_T^k-Y)]_{\text{Part 1}}\\ \le& -\tfrac{\eta}{\beta^2}\big(P_T^k\odot(1-P_T^k/2)\big)_{\min}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)_{\min}\\ &\quad\lambda_{\min}({G_{L-1,T}^k}^\top G_{L-1,T}^k)\lambda_{\min}(\mathbf{M}\mathbf{M}^\top)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2^2,\\ \le& -\tfrac{\eta}{\beta^2}4\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\lambda_{\min}({G_{L-1,T}^k}^\top G_{L-1,T}^k)\lambda_{\min}(\mathbf{M}\mathbf{M}^\top)\|\mathbf{M}^\top(\mathbf{M}X_T^k-Y)\|_2^2,\\ \overset{①}{\le}& -\eta 8\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\tfrac{\beta_0^2}{\beta^2}\alpha_0^2\tfrac{1}{2}\|\mathbf{M}X_T^k-Y\|_2^2,\end{aligned}\tag{49}$$

where ① is from definition.

Case 2: $t < T$**.** We derive the upper bound of above term by Cauchy-Schwarz inequality:

$$
\begin{aligned}
&[(Z - X_T^k)^\top \mathbf{M}^\top (\mathbf{M}X_T^k - Y)]_{\text{Part 2}} \\
=& -\frac{\eta}{\beta^2}(\mathbf{M}X_T^k - Y)^\top \mathbf{M}\Big(\textstyle\sum_{t=1}^{T-1}\big(\prod_{j=T}^{t+1}\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_j^k)\mathbf{M}^\top\mathbf{M}\big) \\
&\qquad \mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big){G_{L-1,t}^k}^\top G_{L-1,T}^k \\
&\qquad \mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\Big)\mathbf{M}^\top(\mathbf{M}X_T^k - Y), \\
\overset{①}{\leq}& \frac{\eta}{\beta^2}\Big\|\textstyle\sum_{t=1}^{T-1}\big(\prod_{j=T}^{t+1}\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_j^k)\mathbf{M}^\top\mathbf{M}\big) \\
&\qquad \mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big){G_{L-1,t}^k}^\top G_{L-1,T}^k \\
&\qquad \mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\Big\|_2 \|\mathbf{M}\mathbf{M}^\top\|_2\|\mathbf{M}X_T^k - Y\|_2^2, \\
\leq& \frac{\eta}{\beta^2}\textstyle\sum_{t=1}^{T-1}\Big\|\big(\prod_{j=T}^{t+1}\mathbf{I} - \frac{1}{\beta}\mathcal{D}(P_j^k)\mathbf{M}^\top\mathbf{M}\big)\Big\|_2 \\
&\qquad \|\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big)\|_2\|G_{L-1,t}^k\|_2\|G_{L-1,T}^k\|_2\|\mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\|_2 \\
&\qquad \|\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\|_2\|\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\|_2\|\mathbf{M}\mathbf{M}^\top\|_2\|\mathbf{M}X_T^k - Y\|_2^2, \\
\overset{②}{\leq}& \frac{\eta}{\beta}\textstyle\sum_{t=1}^{T-1}\|\mathcal{D}\big(P_t^k \odot (1 - P_t^k/2)\big)\|_2\|G_{L-1,t}^k\|_2\|G_{L-1,T}^k\|_2\|\mathcal{D}\big([2\sigma(v_{2,i}^k)(1-\sigma(v_{2,i}^k))]^\top\big)\|_2 \\
&\qquad \|\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\|_2\|\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\|_2\|\mathbf{M}X_T^k - Y\|_2^2, \\
\overset{③}{\leq}& \frac{\eta}{4\beta}\textstyle\sum_{t=1}^{T-1}\|G_{L-1,t}^k\|_2\|G_{L-1,T}^k\|_2\|\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{t-1}^k - Y)\big)\|_2\|\mathcal{D}\big(\mathbf{M}^\top(\mathbf{M}X_{T-1}^k - Y)\big)\|_2 \\
&\qquad \|\mathbf{M}X_T^k - Y\|_2^2, \\
\leq& \frac{\eta}{4\beta}\big(\beta(\|X_0\|_2 + \frac{2T}{\beta}\|\mathbf{M}^\top Y\|_2) + \|\mathbf{M}^\top Y\|_2\big)\|G_{L-1,T}^k\|_2 \\
&\qquad \textstyle\sum_{t=1}^{T-1}\|G_{L-1,t}^k\|_2\big(\beta(\|X_0\|_2 + \frac{2t}{\beta}\|\mathbf{M}^\top Y\|_2) + \|\mathbf{M}^\top Y\|_2\big)\|\mathbf{M}X_T^k - Y\|_2^2, \\
\leq& \frac{\eta}{4\beta}\big(\beta(\|X_0\|_2 + \frac{2T-2}{\beta}\|\mathbf{M}^\top Y\|_2) + \|\mathbf{M}^\top Y\|_2\big)\big((1+\beta)\|X_0\|_2 + \big(2T - 1 + \frac{2T-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big) \\
&\qquad \textstyle\prod_{s=1}^{L-1}\bar{\lambda}_s\sum_{t=1}^{T-1}\big((1+\beta)\|X_0\|_2 + \big(2t - 1 + \frac{2t-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\big) \\
&\qquad \textstyle\prod_{s=1}^{L-1}\bar{\lambda}_s\big(\beta(\|X_0\|_2 + \frac{2t-2}{\beta}\|\mathbf{M}^\top Y\|_2) + \|\mathbf{M}^\top Y\|_2\big)\|\mathbf{M}X_T^k - Y\|_2^2,
\end{aligned}
$$

where ① is due to Cauchy-Schwarz inequality. It note-worthing that ① is non-trivial since $\mathbf{B}_{T-1}^k$ is non-necessarily to be positive definite. ② is due to upper bound of NN's output in Lemma A.1. ③ is due to the Sigmoid function is bounded.

Further, due to the definition of the quantities, we calculate:

$$
\begin{aligned}
&[(Z - X_T^k)^\top \mathbf{M}^\top (\mathbf{M}X_T^k - Y)]_{\text{Part 2}} \\
\leq& \frac{\eta\beta}{4} \\
&\underbrace{\Big((1+\beta)\|X_0\|_2^2 + \big((4T-3)(1+\tfrac{1}{\beta}) + 1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2 + \tfrac{(2T-1)(\beta(2T-1)+(2T-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big)}_{\Lambda_T} \\
&\textstyle\sum_{t=1}^{T-1} \\
&\underbrace{\Big((1+\beta)\|X_0\|_2^2 + \big((4t-3)(1+\tfrac{1}{\beta}) + 1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2 + \tfrac{(2T-1)(\beta(2T-1)+(2T-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big)}_{\Lambda_t} \\
&\Theta_{L-1}^2\|\mathbf{M}X_T^k - Y\|_2^2, \\
=& \frac{\eta\beta}{2}\Theta_{L-1}^2\Lambda_T\textstyle\sum_{t=1}^{T-1}\Lambda_t\frac{1}{2}\|\mathbf{M}X_T^k - Y\|_2^2.
\end{aligned}
\tag{50}
$$

Combining the two parts in Equation (49) and Equation (50) yields:

$$
\begin{aligned}
&(Z - X_T^k)^\top \mathbf{M}^\top (\mathbf{M}X_T^k - Y) \\
\leq& \Big(\frac{\eta\beta}{2}\Theta_{L-1}^2\Lambda_T\textstyle\sum_{t=1}^{T-1}\Lambda_t - \eta 8\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\frac{\beta_0^2}{\beta^2}\alpha_0^2\Big)\frac{1}{2}\|\mathbf{M}X_T^k - Y\|_2^2.
\end{aligned}
$$

□

Using quantities from Equation (11), substituting the upper bounds in Lemma A.8, Lemma A.9, and Lemma A.10 into Equation (37), we calculate:

$$
\begin{aligned}
&F([W]^{k+1})\\
=&F([W]^k) + \tfrac{1}{2}\|\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k\|_2^2 + (\mathbf{M}X_T^{k+1} - \mathbf{M}X_T^k)^\top(\mathbf{M}X_T^k - Y),\\
\le&F([W]^k) + \tfrac{\beta^2\eta^2}{16}({\delta_1}^T)^2\Big(S_{\Lambda,T}\Big)^2\Big(\Theta_L^2\textstyle\sum_{\ell=1}^L\bar{\lambda}_\ell^{-2}\Big)^2\tfrac{1}{2}\|\mathbf{M}X_T^k - Y\|_2\\
&+\tfrac{\beta\eta}{2}(\Lambda_T+\delta_2)\Theta_L^2 S_{\bar{\lambda},L}S_{\Lambda,T}\tfrac{1}{2}\|\mathbf{M}X_T^k - Y\|_2^2\\
&+\Big(-\eta 8\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\tfrac{\beta_0^2}{\beta^2}\alpha_0^2 + \tfrac{\eta\beta}{2}\Theta_{L-1}^2\Lambda_T\textstyle\sum_{t=1}^{T-1}\Lambda_t\Big)\tfrac{1}{2}\|\mathbf{M}X_T^k - Y\|_2^2,\\
\overset{①}{=}&F([W]^k) + \tfrac{\beta^2\eta^2}{16}({\delta_1}^T)^2\Big(S_{\Lambda,T}\Big)^2\Big(\Theta_L^2\textstyle\sum_{\ell=1}^L\bar{\lambda}_\ell^{-2}\Big)^2 F([W]^k)\\
&+\tfrac{\beta\eta}{2}(\Lambda_T+\delta_2)\Theta_L^2 S_{\bar{\lambda},L}S_{\Lambda,T}F([W]^k)\\
&+\Big(-\eta 8\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\tfrac{\beta_0^2}{\beta^2}\alpha_0^2 + \tfrac{\eta\beta}{2}\Theta_{L-1}^2\Lambda_T\textstyle\sum_{t=1}^{T-1}\Lambda_t\Big)F([W]^k),\\
=&\Bigg(1 + \tfrac{\eta^2\beta^2}{16}({\delta_1}^T)^2\Big(S_{\Lambda,T}\Big)^2\Big(\Theta_L^2\textstyle\sum_{\ell=1}^L\bar{\lambda}_\ell^{-2}\Big)^2 + \tfrac{\eta\beta}{2}(\Lambda_T+\delta_2)S_{\Lambda,T}\Theta_L^2 S_{\bar{\lambda},L}\\
&\quad+\tfrac{\eta\beta}{2}\Theta_{L-1}^2\Lambda_T\textstyle\sum_{t=1}^{T-1}\Lambda_t - \eta 8\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\tfrac{\beta_0^2}{\beta^2}\alpha_0^2\Bigg)F([W]^k),\\
\overset{②}{\le}&\Big(1 + \eta\beta(\Lambda_T+\delta_2)S_{\Lambda,T}\Theta_L^2 S_{\bar{\lambda},L} + \tfrac{\eta\beta}{2}\Theta_{L-1}^2\Lambda_T\textstyle\sum_{t=1}^{T-1}\Lambda_t - \eta 8\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\tfrac{\beta_0^2}{\beta^2}\alpha_0^2\Big)\\
&F([W]^k),\\
=&\Big(1 - \eta\big(8\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\tfrac{\beta_0^2}{\beta^2}\alpha_0^2 - \beta(\Lambda_T+\delta_2)S_{\Lambda,T}\Theta_L^2 S_{\bar{\lambda},L} - \tfrac{\beta}{2}\Theta_{L-1}^2\Lambda_T\textstyle\sum_{t=1}^{T-1}\Lambda_t\big)\Big)\\
&F([W]^k),\\
\overset{③}{\le}&(1-\eta\underbrace{4\sigma(\delta_3\Theta_L)^2(1-\sigma(\delta_3\Theta_L))^2\tfrac{\beta_0^2}{\beta^2}\alpha_0^2}_{4\eta\frac{\beta_0^2}{\beta^2}\delta_4})F([W]^k),
\end{aligned}
$$

where ① is due to the definition of objective. ② is due to upper bound of learning rate in Equation (13a) and ${\delta_1}^T = \delta_2 + \sum_{j=1}^T \Lambda_j$ in definition. ③ is due to the lower bound of the least eigenvalue α_0 in Equation (12b).

Due to learning rate's upper bound in Equation (13b), we know $0 < 1 - \eta 4\eta\frac{\beta_0^2}{\beta^2}\delta_4 < 1$, which yields the following linear rate by nature:

$$F([W]^k) \le (1 - \eta 4\eta\tfrac{\beta_0^2}{\beta^2}\delta_4)^k F([W]^0).$$

□

B Details for Initialization

B.1 Preliminary

To begin with, we define the following quantities:

$$
\begin{aligned}
&\delta_5 = \sigma\Big(\big(2T-1+\tfrac{2T-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\Theta_L\Big)^{-2}\Big(1-\sigma\big(\big(2T-1+\tfrac{2T-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\Theta_L\big)\Big)^{-2},\\
&\delta_6 = \sigma_{\min}\big([\textstyle\sum_{t=1}^{T-1}(\mathbf{I}-\tfrac{1}{\beta}\mathbf{M}^\top\mathbf{M})^{T-t}\mathbf{M}^\top Y|\mathbf{M}^\top(\mathbf{M}(\sum_{t=1}^{T-1}(\mathbf{I}-\tfrac{1}{\beta}\mathbf{M}^\top\mathbf{M})^{T-t}\mathbf{M}^\top Y)-Y)]\big),\\
&\delta_7 = \sigma_{\min}(\textstyle\sum_{t=1}^{T-1}(\mathbf{I}-\tfrac{1}{\beta}\mathbf{M}^\top\mathbf{M})^{T-t}).
\end{aligned}
$$

Analysis for the numerical stalibilty of δ_5. δ_5 is a function with Λ_t, which is also enlarge w.r.t. e^L. In general, it is possible to push $\sigma(1-\sigma(\big(2T-1+\frac{2T-2}{\beta}\big)\|\mathbf{M}^\top Y\|_2\Theta_L))$ to zero and let RHS of above inequality to be ∞ when $e^L\to\infty$. As presented in the lemma, we claim that the required e is not necessarily to be ∞. Thus, δ_5 can be regarded as a $\mathcal{O}(e^{L-1})\ll\infty$ constant. In the following proofs, we demonstrate that it holds since e is finite.

We calculate the following exact formulations of the quantities defined in Theorem 4.3:

$$
\begin{aligned}
\Lambda_T =&(1+\beta)\|X_0\|_2^2+\big((4T-3)(1+\tfrac{1}{\beta})+1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2\\
&+\tfrac{(2T-1)(\beta(2T-1)+(2T-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2,\\
=&\tfrac{4(\beta+1)}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T^2+\Big(\tfrac{4(1+\beta)}{\beta}\|X_0\|_2\|\mathbf{M}^\top Y\|_2-\tfrac{4\beta+6}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big)T\\
&+(1+\beta)\|X_0\|_2^2-(2+\tfrac{3}{\beta})\|X_0\|_2\|\mathbf{M}^\top Y\|_2+\tfrac{\beta+2}{\beta^2}\|\mathbf{M}^\top Y\|_2^2,\\
\overset{①}{=}&\tfrac{4(\beta+1)}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T^2-\tfrac{4\beta+6}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T+\tfrac{\beta+2}{\beta^2}\|\mathbf{M}^\top Y\|_2^2,
\end{aligned}
\tag{51}
$$

where ① is due to $X_0=0$ and

$$
\begin{aligned}
\textstyle\sum_{i=1}^T\Lambda_i =&\textstyle\sum_{i=1}^T(1+\beta)\|X_0\|_2^2+\big((4i-3)(1+\tfrac{1}{\beta})+1\big)\|X_0\|_2\|\mathbf{M}^\top Y\|_2\\
&+\tfrac{(2i-1)(\beta(2i-1)+(2i-2))}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\\
=&\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2T^3+\Big(\tfrac{2(1+\beta)}{\beta}\|X_0\|_2\|\mathbf{M}^\top Y\|_2-\tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big)T^2\\
&+\Big((1+\beta)\|X_0\|_2^2-\tfrac{1}{\beta}\|X_0\|_2\|\mathbf{M}^\top Y\|_2-\tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big)T,\\
\overset{①}{=}&\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2T^3-\tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T^2-\tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2T,
\end{aligned}
\tag{52}
$$

where ① is due to $X_0=0$.

Then, we analyze the expansion of $\sigma_{\min}(G^0_{L-1,T})$ w.r.t. $[W]_L=e[W]_L$. Due to the one line form equation of L2O model in Equation (21), we have $\sigma_{\min}(G^0_{L-1,T})$ is calculated by:

$$
\sigma_{\min}(G^0_{L-1,T})=\sigma_{\min}\big(\,\mathrm{ReLU}(\mathrm{ReLU}([X^0_{T-1},\mathbf{M}^\top(\mathbf{M}X^0_{T-1}-Y)]{W^0_1}^\top)\cdots {W^0_{L-1}}^\top)\big),
$$

where due to Equation (21), X^0_{T-1} is given by:

$$
\begin{aligned}
X^0_{T-1}=&\textstyle\prod_{t=T-1}^1(\mathbf{I}-\tfrac{1}{\beta}\mathcal{D}(P^0_t)\mathbf{M}^\top\mathbf{M})X_0+\tfrac{1}{\beta}\sum_{t=1}^{T-1}\prod_{s=T-1}^{t+1}(\mathbf{I}-\tfrac{1}{\beta}\mathcal{D}(P^0_s)\mathbf{M}^\top\mathbf{M})\mathcal{D}(P^0_t)\mathbf{M}^\top Y,\\
\overset{①}{=}&(\mathbf{I}-\tfrac{1}{\beta}\mathbf{M}^\top\mathbf{M})^{T-1}X_0+\tfrac{1}{\beta}\textstyle\sum_{t=1}^{T-1}(\mathbf{I}-\tfrac{1}{\beta}\mathbf{M}^\top\mathbf{M})^{T-t}\mathbf{M}^\top Y,\\
\overset{②}{=}&\tfrac{1}{\beta}\textstyle\sum_{t=1}^{T-1}(\mathbf{I}-\tfrac{1}{\beta}\mathbf{M}^\top\mathbf{M})^{T-t}\mathbf{M}^\top Y,
\end{aligned}
\tag{53}
$$

where ① is due to $P_t=\sigma(\mathbf{0})=\mathbf{I}$ since $W_L=0$. The result shows that X^0_{T-1} is unrelated to $[W]_L$ with $W_L=0$. ② is due to $X_0=0$.

Further, for $t\in[T]$, denote the angel between X^0_{t-1} and $\mathbf{M}^\top(\mathbf{M}X^0_{t-1}-Y)$ as θ_{t-1}, we have $\sin(\theta_{t-1})\in(0,1)$, setting $[W]_L=e[W]_L$, we calculate $\sigma_{\min}(\tilde{G}^0_{L-1,T})$ by:

$$
\begin{aligned}
\sigma_{\min}(\tilde{G}^0_{L-1,T})=&\sigma_{\min}\big(\,\mathrm{ReLU}(\mathrm{ReLU}([X^0_{T-1},\mathbf{M}^\top(\mathbf{M}X^0_{T-1}-Y)]e{W^0_1}^\top)\cdots e{W^0_{L-1}}^\top)\big),\\
\geq&\sigma_{\min}\big([X^0_{T-1}|\mathbf{M}^\top(\mathbf{M}X^0_{T-1}-Y)]\big)\textstyle\prod_{\ell=1}^{L-1}\sigma_{\min}(eW^0_\ell),\\
\geq&\tfrac{\|X^0_{T-1}\|_2\|\mathbf{M}^\top(\mathbf{M}X^0_{T-1}-Y)\|_2\sin(\theta_{T-1})}{\|X^0_{T-1}\|_2+|\mathbf{M}^\top(\mathbf{M}X^0_{T-1}-Y)\|_2}\textstyle\prod_{\ell=1}^{L-1}\sigma_{\min}(eW^0_\ell),\\
=&\frac{\sin(\theta_{T-1})}{\frac{1}{\|X^0_{T-1}\|_2}+\frac{1}{\|\mathbf{M}^\top(\mathbf{M}X^0_{T-1}-Y)\|_2}}\textstyle\prod_{\ell=1}^{L-1}\sigma_{\min}(eW^0_\ell),\\
\geq&\sin(\theta_{T-1})\textstyle\prod_{\ell=1}^{L-1}\sigma_{\min}(W^0_\ell)\Theta_L\|X^0_{T-1}\|_2.
\end{aligned}
\tag{54}
$$

Based on the definition of X_{T-1}^0 in Equation (53), we calculate following bound:

$$\begin{aligned}\sigma_{\min}(\tilde{G}_{L-1,T}^0) \geq& \frac{\sin(\theta_{T-1})}{\beta}\|\textstyle\sum_{t=1}^{T-1}(\mathbf{I}-\frac{1}{\beta}\mathbf{M}^\top\mathbf{M})^{T-t}\mathbf{M}^\top Y\|_2\prod_{\ell=1}^{L-1}\sigma_{\min}(eW_\ell^0),\\ \geq& \frac{\sin(\theta_{T-1})}{\beta}\underbrace{\sigma_{\min}(\textstyle\sum_{t=1}^{T-1}(\mathbf{I}-\frac{1}{\beta}\mathbf{M}^\top\mathbf{M})^{T-t})}_{\delta_7}\|\mathbf{M}^\top Y\|_2 e^{L-1}\prod_{\ell=1}^{L-1}\sigma_{\min}(W_\ell^0),\end{aligned} \tag{55}$$

where X_{T-1}^0 is a constant related to problem definition.

Substituting Equation (53), we calculate a more tight lower bound of $\|X_{T-1}^0\|_2$ by:

$$\begin{aligned}\|X_{T-1}^0\|_2 =& \left\|\frac{1}{\beta}\textstyle\sum_{t=1}^{T-1}\prod_{s=T-1}^{t+1}(\mathbf{I}-\frac{1}{\beta}\mathcal{D}(P_s^0)\mathbf{M}^\top\mathbf{M})\mathcal{D}(P_t^0)\mathbf{M}^\top Y\right\|_2,\\ \geq& \frac{1}{\beta}\|\mathbf{M}^\top Y\|_2\sigma_{\min}\Big(\textstyle\sum_{t=1}^{T-1}\prod_{s=T-1}^{t+1}(\mathbf{I}-\frac{1}{\beta}\mathcal{D}(P_s^0)\mathbf{M}^\top\mathbf{M})\mathcal{D}(P_t^0)\Big),\\ \overset{①}{\geq}& \frac{1}{\beta}\|\mathbf{M}^\top Y\|_2\textstyle\sum_{t=1}^{T-1}\sigma_{\min}\Big(\prod_{s=T-1}^{t+1}(\mathbf{I}-\frac{1}{\beta}\mathcal{D}(P_s^0)\mathbf{M}^\top\mathbf{M})\Big)\sigma_{\min}(\mathcal{D}(P_t^0)),\\ \geq& \frac{1}{\beta}\|\mathbf{M}^\top Y\|_2\textstyle\sum_{t=1}^{T-1}\Big(\prod_{s=T-1}^{t+1}\sigma_{\min}\big(\mathbf{I}-\frac{1}{\beta}\mathcal{D}(P_s^0)\mathbf{M}^\top\mathbf{M}\big)\Big)\sigma_{\min}(\mathcal{D}(P_t^0)),\end{aligned} \tag{56}$$

where ① is due to all matrices in the summation are positive semi-definite by definition.

We calculate lower bound for $\sigma_{\min}\big(\mathbf{I}-\frac{1}{\beta}\mathcal{D}(P_s^0)\mathbf{M}^\top\mathbf{M}\big)$ by:

$$\begin{aligned}\sigma_{\min}\big(\mathbf{I}-\frac{1}{\beta}\mathcal{D}(P_s^0)\mathbf{M}^\top\mathbf{M}\big) \geq& 1-\frac{1}{\beta}\sigma_{\max}\big(\mathcal{D}(2\sigma(eW_L^0\tilde{G}_{L-1,s}^0))\mathbf{M}^\top\mathbf{M}\big)\\ \geq& 1-2\underbrace{\sigma(\delta_3\Theta_L)(1-\sigma(\delta_3\Theta_L))}_{\delta_4}\sigma_{\max}(eW_L^0\tilde{G}_{L-1,s}^0),\end{aligned} \tag{57}$$

It is easy to verify that the above equation equal to 1 when $e\to+\infty$ and it decreases with e. Also, a large e ensures the RHS of above inequality to be positive.

Similarly, we calculate lower bound for $\sigma_{\min}(P_t^0)$ by:

$$\begin{aligned}\sigma_{\min}(\mathcal{D}(P_t^0)) \overset{①}{=}& \min\big(2\sigma(eW_L^0\tilde{G}_{L-1,t}^0)\big),\\ \overset{②}{=}& \min\big(\tfrac{\partial 2\sigma}{\partial v_4}(eW_L^0\tilde{G}_{L-1,t}^0)\big),\\ \overset{③}{\geq}& 2\delta_4\sigma_{\min}\big(eW_L^0\tilde{G}_{L-1,t}^0\big),\\ \overset{④}{\geq}& 2\delta_4 e\|W_L^0\|_2\sigma_{\min}(\tilde{G}_{L-1,t}^0),\\ \geq& 2\Theta_L\delta_4\textstyle\prod_{\ell=1}^{L}\|W_\ell^0\|_2\sigma_{\min}\big([X_{t-1}^0|\mathbf{M}^\top(\mathbf{M}X_{t-1}^0-Y)]\big),\\ \overset{⑤}{\geq}& 2\Theta_L\delta_4\textstyle\prod_{\ell=1}^{L}\|W_\ell^0\|_2\sin(\theta_{T-1})\|X_{t-1}^0\|_2\end{aligned} \tag{58}$$

where ① means we apply the expansion here. ② is due to Mean Value Theorem and v_4 denotes a inner point between 0 and $eW_L^0\tilde{G}_{L-1,T}^0$. ③ is due to Lemma A.12 and Lemma A.14. ④ is due to W_L^0 is a vector in definition. ⑤ is similar to the workflow in Equation (54).

Substituting Equation (57) and Equation (58) back into Equation (56) yields:

$$\begin{aligned}\|X_{t-1}^0\|_2 \geq& \frac{1}{\beta}\|\mathbf{M}^\top Y\|_2\textstyle\sum_{s=1}^{t-1}2\Theta_L\delta_4\prod_{\ell=1}^{L}\|W_\ell^0\|_2\sigma_{\min}\big([X_{s-1}^0|\mathbf{M}^\top(\mathbf{M}X_{s-1}^0-Y)]\big),\\ \geq& \frac{2}{\beta}\|\mathbf{M}^\top Y\|_2\Theta_L\textstyle\sum_{s=1}^{t-1}\delta_4\prod_{\ell=1}^{L}\|W_\ell^0\|_2\sin(\theta_{s-1})\|X_{s-1}^0\|_2,\end{aligned}$$

Similarly, we can get the following lower bound of $\|X_{t-1}^0\|_2$:

$$\|X_{t-1}^0\|_2 \geq \frac{2}{\beta}\|\mathbf{M}^\top Y\|_2\Theta_L\textstyle\sum_{s=1}^{t-1}\delta_4\prod_{\ell=1}^{L}\|W_\ell^0\|_2\sin(\theta_{t-1})\|X_{s-1}^0\|_2,$$

Based on the above results, we calculate the $\mathbf{\Omega}$ of $\|X_{T-1}^0\|_2$ as in terms of T and Θ_L as:

$$\|X_{T-1}^0\|_2 \geq \mathbf{\Omega}(\underbrace{\Theta_L \textstyle\sum_{t=1}^{T-1}\Theta_L \sum_{s=1}^{t-1}\Theta_L \sum_{j=1}^{s-1}\dots\sum_{j=1}^{2}}_{\text{T-2 terms}}) = \mathbf{\Omega}(\Theta_L^{T-2}).$$

Substituting back into Equation (54) yields:

$$\sigma_{\min}(\tilde{G}_{L-1,T}^0) = \mathbf{\Omega}(e^{L-1}e^{(T-2)(L-1)}) = \mathbf{\Omega}(e^{(T-1)(L-1)}). \tag{59}$$

B.2 Proof of Lemma 5.1

Proof. Making up the lower bounding relationship with Equation (55) and Equation (60) yields:

$$\begin{aligned} e^{L-1}\|\mathbf{M}^\top Y\|_2\delta_7\textstyle\prod_{\ell=1}^{L-1}\sigma_{\min}(W_\ell^0) \geq & 8(1+\beta)(\|X_0\|_2 + \tfrac{2T-2}{\beta}\|\mathbf{M}^\top Y\|_2), \\ = & \tfrac{8(1+\beta)}{\beta}(2T-2)\|\mathbf{M}^\top Y\|_2, \end{aligned}$$

which yields:

$$e \geq \sqrt[L-1]{\tfrac{8(1+\beta)}{\beta}\delta_7^{-1}\sigma_{\min}(W_\ell^0)^{-1}(2T-2)}.$$

□

B.3 Proof of Lemma 5.4

We apply a similar workflow to prove Lemma 5.4.

Proof. With $X_0 = 0$, we find the upper bound of the RHS of Equation (12d) by substituting the quantity δ_5:

$$\begin{aligned} & \tfrac{(1+\beta)\beta^2\sqrt{\beta}}{2\beta_0^2}\delta_5\big(\sqrt{\beta}\|X_0\|_2 + (2T+1)\|Y\|_2\big)\zeta_2 S_{\Lambda,T}\Theta_{L-1}\Big(\textstyle\sum_{\ell=1}^{L}\frac{\Theta_L}{\bar\lambda_\ell^2}\Big) \\ \overset{①}{=} & \tfrac{(1+\beta)\beta^2\sqrt{\beta}}{2\beta_0^2}\delta_5\big(\sqrt{\beta}\|X_0\|_2 + (2T+1)\|Y\|_2\big)\zeta_2 \\ & \Big(\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^3 - \tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^2 - \tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2 T\Big)\Theta_{L-1}\Big(\textstyle\sum_{\ell=1}^{L}\frac{\Theta_L}{\bar\lambda_\ell^2}\Big), \\ \overset{②}{=} & \tfrac{(1+\beta)\beta\sqrt{\beta}}{2\beta_0^2}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2(2T-2)(2T+1) \\ & \Big(\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^3 - \tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^2 - \tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2 T\Big)\Theta_{L-1}\Big(\textstyle\sum_{\ell=1}^{L}\frac{\Theta_L}{\bar\lambda_\ell^2}\Big), \\ \overset{③}{\leq} & \tfrac{(1+\beta)\sqrt{\beta}}{6\beta_0^2\beta}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2^3 \\ & \Big(16(\beta+1)T^5 - (8\beta+20)T^4 - 6(2\beta+1)T^3 + 2(\beta+4)T^2 + 2(\beta+1)T\Big)L\Theta_{L-1}^2, \end{aligned} \tag{60}$$

where ① is due to Equation (52) and definition of quantity δ_1^{T-1} in Theorem 4.3. ② is due to $X_0 = 0$. ③ is due to $\bar\lambda_L = 1$ and $\bar\lambda_\ell > 1, \ell \in [L-1]$.

Making up the lower bounding relationship with Equation (55) and Equation (60) yields:

$$\begin{aligned} & \Big(e^{L-1}\|\mathbf{M}^\top Y\|_2\delta_7\textstyle\prod_{\ell=1}^{L-1}\sigma_{\min}(W_\ell^0)\Big)^3 \\ \geq & e^{2L-2}\tfrac{(1+\beta)\sqrt{\beta}}{6\beta_0^2\beta}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2^3 L\textstyle\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2+1)^2 \\ & \Big(16(\beta+1)T^5 - (8\beta+20)T^4 - 6(2\beta+1)T^3 + 2(\beta+4)T^2 + 2(\beta+1)T\Big), \end{aligned} \tag{61}$$

which yields:

$$e \geq \sqrt[L-1]{C_{2,\delta_5}\Big(16(\beta+1)T^5 - (8\beta+20)T^4 - 6(2\beta+1)T^3 + 2(\beta+4)T^2 + 2(\beta+1)T\Big)}.$$

where C_{2,δ_5} denotes the $\frac{(1+\beta)\sqrt{\beta}}{6\beta_0^2\beta\delta_7^3}\delta_5\|Y\|_2 L\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2+1)^2\prod_{\ell=1}^{L-1}\sigma_{\min}(W_\ell^0)^{-3}$ term.

Similarly, the finite RHS of above inequality ensures $\delta_5 \ll \infty$. □

B.4 Proof of Lemma 5.2

Proof. Using quantities from Equation (11), with $X_0 = 0$, we find the upper bound of the RHS of Equation (12b) by substituting the quantity δ_5:

$$
\begin{aligned}
&\tfrac{\beta^3}{4\beta_0^2}\delta_5\Big(- \tfrac{1}{2}\Theta_{L-1}^2\Lambda_T\big(\textstyle\sum_{t=1}^{T-1}\Lambda_t\big) + \Theta_L^2 S_{\bar{\lambda},L}(\Lambda_T+\delta_2)S_{\Lambda,T}\Big) \\
\overset{①}{=}&\tfrac{\beta^3}{4\beta_0^2}\delta_5\Big(- \tfrac{1}{2}\Theta_{L-1}^2\Big(\tfrac{4(\beta+1)}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T^2 - \tfrac{4\beta+6}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T + \tfrac{\beta+2}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big) \\
&\qquad\Big(\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2(T-1)^3 - \tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2(T-1)^2 - \tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2(T-1)\Big) \\
&\qquad + \Theta_L^2 S_{\bar{\lambda},L}\Big(\Big(\tfrac{4(\beta+1)}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T^2 - \tfrac{4\beta+6}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T + \tfrac{\beta+2}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big) \\
&\qquad\qquad + \textstyle\sum_{s=1}^{T-1}\Big(\prod_{j=s+1}^{T}\big(1+\tfrac{1+\beta}{2\beta}(2j-1)\Theta_L\|\mathbf{M}^\top Y\|_2\big)\Big) \\
&\qquad\qquad\Big(\tfrac{4(\beta+1)}{\beta^2}\|\mathbf{M}^\top Y\|_2^2s^2 - \tfrac{4\beta+6}{\beta^2}\|\mathbf{M}^\top Y\|_2^2s + \tfrac{\beta+2}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big)\Big) \\
&\qquad\qquad\Big(\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2T^3 - \tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T^2 - \tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2T\Big)\Big), \\
\le&\mathcal{O}(e^{2L-2}T^5 + e^{2L-4}T^5 + e^{2L-4}T^6\textstyle\sum_{s=1}^{T-1}s^2\prod_{j=s+1}^{T}je^{L-1}), \\
=&\mathcal{O}(e^{TL-T+2L-4}T^{3T+6}).
\end{aligned}
\tag{62}
$$

where ① is due to Equation (52) and definition of quantity δ_1^{T-1} in Theorem 4.3. ② is due to $X_0 = 0$. ③ is due to $\bar{\lambda}_L = 1$ and $\bar{\lambda}_\ell > 1, \ell \in [L-1]$.

Making up the lower bounding relationship with Equation (59) and Equation (60) yields:

$$(\mathbf{\Omega}(e^{(T-1)(L-1)}))^2 \ge \mathcal{O}(e^{TL-T+2L-4}T^{3T+6}),$$

which yields:

$$e = \mathbf{\Omega}(T^{\frac{3T+6}{TL-T-4L+6}}).$$

□

B.5 Proof of Lemma 5.3

Proof. Using quantities from Equation (11), with $X_0 = 0$, we find the upper bound of the RHS of Equation (12c) by substituting the quantity δ_5:

$$
\begin{aligned}
&\max_{\ell\in[L]}\tfrac{\Theta_L}{C_\ell\bar{\lambda}_\ell}\tfrac{\beta^2\sqrt{\beta}}{8\beta_0^2} \\
&\qquad\underbrace{\sigma\big((2T-1+\tfrac{2T-2}{\beta})\|\mathbf{M}^\top Y\|_2\Theta_L\big)^{-2}\big(1-\sigma((2T-1+\tfrac{2T-2}{\beta})\|\mathbf{M}^\top Y\|_2\Theta_L)\big)^{-2}}_{\delta_5} \\
&\qquad S_{\Lambda,T}(2T+1)\|Y\|_2, \\
\overset{①}{\le}&\tfrac{\beta^2\sqrt{\beta}}{8\beta_0^2}\delta_5S_{\Lambda,T}(2T+1)\|Y\|_2\textstyle\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2+1), \\
\overset{②}{=}&\tfrac{\beta^2\sqrt{\beta}}{8\beta_0^2}\delta_5\big(\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2T^3 - \tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2T^2 - \tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2T\big) \\
&(2T+1)\|Y\|_2\textstyle\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2+1), \\
=&\tfrac{\beta^2\sqrt{\beta}}{8\beta_0^2}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2^2\Big(\tfrac{8(\beta+1)}{3\beta^2}T^4 + \big(\tfrac{4(\beta+1)}{3\beta^2} - \tfrac{2}{\beta^2}\big)T^3 - \big(\tfrac{1}{\beta^2}+2\tfrac{\beta+1}{3\beta^2}\big)T^2 - \tfrac{\beta+1}{3\beta^2}T\Big) \\
&\textstyle\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2+1),
\end{aligned}
\tag{63}
$$

where ① is due to $\bar{\lambda}_\ell > 1, \ell \in [L-1]$ and $\bar{\lambda}_L = 1$. ② is due to Equation (52).

We analyze the two sides of the above inequality when $[W]_L = e[W]_L$ to demonstrate a sufficient lower bound of e to ensure Equation (63) holds.

If $[W]_L = e[W]_L$, since $e \geq 1$, Equation (63) is upperly bounded by:

$$\begin{aligned}&\tfrac{\beta^2\sqrt{\beta}}{8\beta_0^2}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2^2\Big(\tfrac{8(\beta+1)}{3\beta^2}T^4 + \big(\tfrac{4(\beta+1)}{3\beta^2} - \tfrac{2}{\beta^2}\big)T^3 - \big(\tfrac{1}{\beta^2} + 2\tfrac{\beta+1}{3\beta^2}\big)T^2 - \tfrac{\beta+1}{3\beta^2}T\Big)\\ &\textstyle\prod_{\ell=1}^{L-1}(e\|W_\ell^0\|_2 + e)\\ =&e^{L-1}\tfrac{\beta^2\sqrt{\beta}}{8\beta_0^2}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2^2\\ &\left(\tfrac{8(\beta+1)}{3\beta^2}T^4 + \left(\tfrac{4(\beta+1)}{3\beta^2} - \tfrac{2}{\beta^2}\right)T^3 - \left(\tfrac{1}{\beta^2} + 2\tfrac{\beta+1}{3\beta^2}\right)T^2 - \tfrac{\beta+1}{3\beta^2}T\right)\textstyle\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2 + 1).\end{aligned} \tag{64}$$

If RHS (lower bound) of Equation (55) greater than the RBS (upper bound) of above result, lower bound condition for minimal singular value in Equation (63) sufficiently holds, which yields:

$$\begin{aligned}&\left(e^{L-1}\|\mathbf{M}^\top Y\|_2\delta_7\textstyle\prod_{\ell=1}^{L-1}\sigma_{\min}(W_\ell^0)\right)^2\\ \geq& e^{L-1}\tfrac{\beta^2\sqrt{\beta}}{8\beta_0^2\delta_6}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2^2\left(\tfrac{8(\beta+1)}{3\beta^2}T^4 + \left(\tfrac{4(\beta+1)}{3\beta^2} - \tfrac{2}{\beta^2}\right)T^3 - \left(\tfrac{1}{\beta^2} + 2\tfrac{\beta+1}{3\beta^2}\right)T^2 - \tfrac{\beta+1}{3\beta^2}T\right)\\ &\textstyle\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2 + 1),\end{aligned}$$

which yields:

$$e \geq \sqrt[L-1]{C_{1,\delta_5}\left(\tfrac{8(\beta+1)}{3}T^4 + \left(\tfrac{4(\beta+1)}{3} - 2\right)T^3 - \left(1 + 2\tfrac{\beta+1}{3}\right)T^2 - \tfrac{\beta+1}{3}T\right)},$$

where C_{1,δ_5} denotes the $\frac{\sqrt{\beta}}{8\beta_0^2\delta_6\delta_7^2}\delta_5\|Y\|_2\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2+1)\prod_{\ell=1}^{L-1}\sigma_{\min}(W_\ell^0)^{-2}$ term, which is a "constant" w.r.t. δ_5.

In the end, it is trivial to evaluate that the RHS of above δ_5 is finite with such e. □

C Additional Experimental Results

In this section, we present detailed experimental settings and corresponding results. We define problems at three distinct scales, as described in Appendix C.1. The smaller scale is utilized for ablation studies (Section 6.2), whereas the larger scales is adopted for training experiments (Section 6.1 and Appendix C.2) and inference experiments (Appendix C.3).

C.1 Configurations for Different Experiments

Details of the three experimental configurations are presented in Table 1. **Scale 1** involves a DNN trained with input $X \in \mathbb{R}^{32\times32}$ and output $Y \in \mathbb{R}^{32\times25}$, featuring an $(L-1)$-th layer dimension of 1024. **Scale 2** utilizes input $X \in \mathbb{R}^{10\times512}$ and output $Y \in \mathbb{R}^{10\times400}$, with the $(L-1)$-th layer dimension established at 5120. **Scale 3** employs input $X \in \mathbb{R}^{2048\times512}$ and output $Y \in \mathbb{R}^{2048\times400}$. This configuration is designed as an under-parameterized system, with an $(L-1)$-th layer dimension of 5120, specifically to evaluate the robustness of our proposed L2O framework.

Table 1: Configurations with Different Scales

Index	d	b	**Dimension of $L-1$ Layer's Output**	**Training Samples**
1	32	25	1024	32
2	512	400	5120	10
3	512	400	20	2048

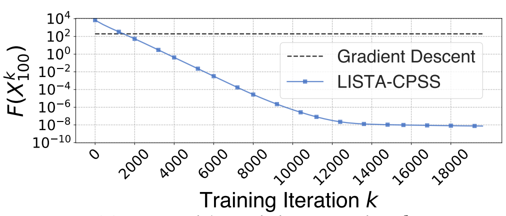
C.2 Additional Training Experiments

For these experiments, the **Scale 3** configuration is utilized. Both baseline state-of-the-art (SOTA) methods and our proposed L2O framework are trained for 2000 epochs using a learning rate of 0.001. However, the inherent model construction and training scheme of a prominent SOTA method, LISTA-CPSS [7], diverge considerably from the requirements of our problem. Direct application of its original settings to our scenario results in over-fitting and poor training convergence, indicating a lack of robustness for this specific application. The following discussion elaborates on these incompatibilities and the modifications undertaken.

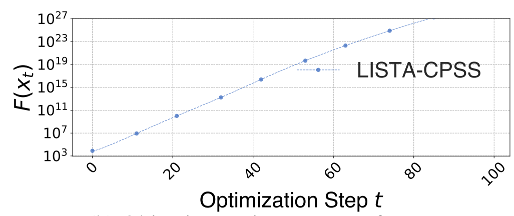
The original LISTA-CPSS framework possesses two key characteristics pertinent to this discussion. First, regarding its model construction, LISTA-CPSS addresses inverse problems by formulating a learnable Least Absolute Shrinkage and Selection Operator (LASSO) problem, wherein it learns a scalar coefficient for the L_1 regularization term [7]. However, our objective in Equation (1) is quadratic. Second, its training protocol is supervised, utilizing an L_2 loss against pre-generated optimal solutions, and employs a layer-wise training scheme. In this scheme, one layer is progressively added to the set of trainable parameters per training iteration, and these parameters are updated using four back-propagation (BP) steps [7]. To adapt LISTA-CPSS for our purposes, we modify both its model architecture and original training scheme to enable unsupervised optimization of our loss function (defined in Equation (2)) and to better align with our established training configuration.

First, to demonstrate the challenges of applying LISTA-CPSS's original training paradigm to unsupervised quadratic objectives, we evaluate a minimally adapted version. This version is trained unsupervisedly by defining the loss as the objective function value from the final optimization step. Given our quadratic loss in Equation (2), any model components in LISTA-CPSS specifically designed for non-quadratic terms are not directly applicable. Moreover, a critical aspect of the publicly available LISTA-CPSS implementation is its initialization of the neural network (NN) with a fixed matrix $\mathbf{M}$. This initialization inherently restricts the trained model's utility to problems featuring this identical, predetermined $\mathbf{M}$.

We train this minimally adapted LISTA-CPSS variant for 50 epochs (corresponding to 20000 BPs due to its layer-wise updates) using the Adam optimizer[2] on a dataset of 2048 randomly generated samples. The loss function defined in Equation (2) is evaluated at an optimization step of $T = 100$. The experimental results, depicted in Figure 6, reveal that this configuration leads to severe overfitting on the training samples. Specifically, Figure 6a illustrates the convergence of the objective function (at $T = 100$) as a function of the training iteration k. Concurrently, Figure 6b displays the mean objective value across 100 optimization steps during inference. These results indicate that while LISTA-CPSS achieves rapid convergence on the training data (which used a fixed $\mathbf{M}$), its performance degrades catastrophically (i.e., fails to generalize) when evaluated with a different matrix, $\mathbf{M}'$, during inference.



(a) Loss with Training Iteration k



(b) Objective Trajectory on Inference

Figure 6: Training Loss and Inference Trajectory of LISTA-CPSS [7] with Fixed $\mathbf{M}$

Informed by the above observation, a more robust approach is achieved through the random initialization of LISTA-CPSS. Specifically, weights are sampled from a standard Gaussian distribution and subsequently scaled by a factor of $\frac{1}{d \cdot b}$ to mitigate potential numerical overflow in cumulative products. The LISTA-CPSS model is then trained using this initialization strategy.

For our proposed L2O framework, the expansion coefficient e is set to 100. As detailed in **Scale 3** in Table 1, we implement an under-parameterized system wherein the dimension of the $(L-1)$-th layer is configured to 20. This implementation intentionally deviates from the theoretical requirements stipulated by our proposed theorems, which necessitate that the dimension of the $(L-1)$-th layer must be larger than the input dimension. This particular experiment is conducted to demonstrate the

[2]Our preliminary experiments indicates that SGD fails to converge with LISTA-CPSS's original layer-wise training scheme.

robustness of the proposed L2O framework, especially under such conditions that depart from our established theoretical framework.

The training losses of LISTA-CPSS and our proposed L2O framework are depicted in Figure 7, with the performance of non-learnable gradient descent (indicated by a horizontal line in the figure) serving as a baseline. Under scenarios with varied $\mathbf{M}$ configurations, LISTA-CPSS exhibits markedly slower convergence compared to both our proposed L2O framework and the gradient descent baseline. Moreover, the fast convergence observed for our L2O framework underscores the robustness and efficacy of its proposed initialization strategy, particularly when applied to under-parameterized models.

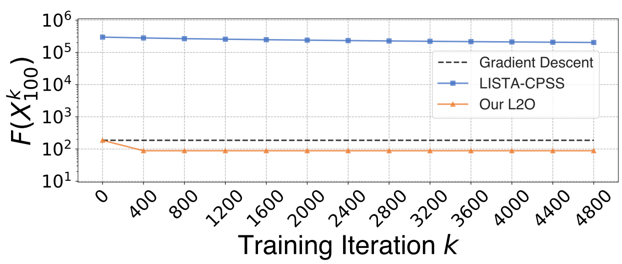


Figure 7: Training Losses with Varied $\mathbf{M}$

C.3 Inference Experiment

Beyond analyzing training outcomes, we extend our evaluation to the robustness of the proposed L2O framework by assessing its performance in inference-stage optimization. This involves comparing the convergence characteristics of L2O against the Adam optimizer [9] and standard gradient descent (GD). It should be noted that while our theorems provide convergence guarantees for the training phase, such guarantees do not explicitly extend to this inference optimization context. For this empirical investigation, both our L2O framework and the Adam optimizer are executed across a range of hyperparameter settings for 3000 iterations (longer than 100 iterations in training), and their respective objective function trajectories are plotted as a function of the iteration count.

Adam utilizes momentum to accelerate gradient descent. In addition to the learning rate η, Adam employs two crucial hyperparameters, β_1 and β_2, which control the exponential moving averages of past gradients and their squared magnitudes, respectively. For the Adam optimizer in our experiments, we set the learning rate $\eta = \frac{1}{\beta}$ (β-smoothness of objective) and explored hyper-parameters $\beta_1 \in \{0.1, 0.3, \dots, 0.9\}$ and $\beta_2 \in \{0.95, 0.955, \dots, 1.0\}$.

Regarding our proposed L2O framework and consistent with the initialization strategy detailed in Section 5, we selected a large expansion coefficient $e = 100$ to enhance training stability. The L2O model is then trained with learning rates η chosen from the set $\{10^{-3}, 10^{-4}, \dots, 10^{-7}\}$.

As illustrated in Figure 8, we present the objective trajectory over 3000 optimization steps, where each point is a mean value of 30 randomly generated problems' objectives. While the objective function initially exhibits rapid decay, the Adam optimizer fails to maintain this convergence, ultimately settling at sub-optimal values and not converging on average. In contrast, our proposed framework demonstrates superior performance compared to the Gradient Descent (GD) algorithm and exhibits robustness across various learning rates.

C.4 Corollary in Ablation Study

Corollary C.1 (LR's upper bound w.r.t. e)**.**

$$\eta = \mathcal{O}(e^{3-L}T^{-6}) \cap \mathcal{O}(e^{1-L}T^{-4}) \cap \mathcal{O}(e^{\frac{4}{3}(1-L)}T^{-\frac{10}{3}}) \cap \mathcal{O}(e^{-TL-2L+T+4}T^{-3T-6}) \cap \mathcal{O}(T^{-2}).$$

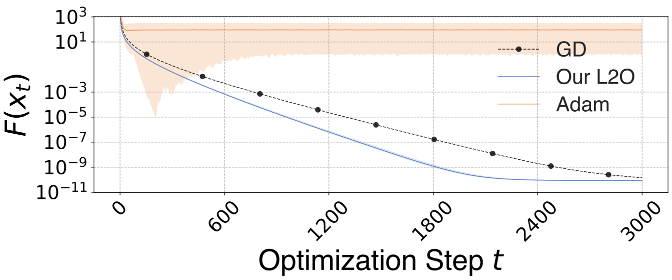


Figure 8: Inference Trajectory of Our Proposed L2O

Proof. From Equation (13a), we calculate:

$$
\begin{aligned}
&\eta \\
<&\tfrac{8}{\beta}(\delta_2+\Lambda_T)\Big(\delta_2+S_{\Lambda,T}\Big)^{-1}S_{\Lambda,T}^{-2}\Theta_L^{-1}S_{\bar{\lambda},L}^{-1}, \\
<&\Bigg(\textstyle\sum_{s=1}^{T-1}\Big(\prod_{j=s+1}^{T}\big(1+\frac{1+\beta}{2\beta}(2j-1)\Theta_L\|\mathbf{M}^\top Y\|_2\big)\Big) \\
&\Big(\tfrac{4(\beta+1)}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 s^2-\tfrac{4\beta+6}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 s+\tfrac{\beta+2}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big) \\
&\quad+\Big(\tfrac{4(\beta+1)}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^2-\tfrac{4\beta+6}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 T+\tfrac{\beta+2}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big)\Bigg) \\
&\Bigg(\textstyle\sum_{s=1}^{T-1}\Big(\prod_{j=s+1}^{T}\big(1+\frac{1+\beta}{2\beta}(2j-1)\Theta_L\|\mathbf{M}^\top Y\|_2\big)\Big) \\
&\Big(\tfrac{4(\beta+1)}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 s^2-\tfrac{4\beta+6}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 s+\tfrac{\beta+2}{\beta^2}\|\mathbf{M}^\top Y\|_2^2\Big) \\
&\quad+\Big(\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^3-\tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^2-\tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2 T\Big)\Bigg)^{-1} \\
&\Big(\tfrac{4(\beta+1)}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^3-\tfrac{1}{\beta^2}\|\mathbf{M}^\top Y\|_2^2 T^2-\tfrac{\beta+1}{3\beta^2}\|\mathbf{M}^\top Y\|_2^2 T\Big)^{-2}\Big(e^{L-1}\textstyle\prod_{\ell=1}^{L-1}\bar{\lambda}_\ell\Big)^{-1}S_{\bar{\lambda},L}^{-1}, \\
=&\mathcal{O}(e^{3-L}T^{-6}).
\end{aligned}
$$

From Equation (13b), due to the four lower bounds in Equation (12), we calculate following four upper bounds:

$$
\begin{aligned}
&\eta \\
<&\tfrac{1}{4}\tfrac{\beta^2}{\beta_0^2}\delta_4^{-2}\alpha_0^{-2}, \\
\overset{64}{<}&\tfrac{1}{4}\tfrac{\beta^2}{\beta_0^2}\delta_5\Bigg(e^{L-1}\tfrac{\beta^2\sqrt{\beta}}{8\beta_0^2}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2^2 \\
&\qquad\Big(\tfrac{8(\beta+1)}{3\beta^2}T^4+\Big(\tfrac{4(\beta+1)}{3\beta^2}-\tfrac{2}{\beta^2}\Big)T^3-\Big(\tfrac{1}{\beta^2}+2\tfrac{\beta+1}{3\beta^2}\Big)T^2-\tfrac{\beta+1}{3\beta^2}T\Big)\textstyle\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2+1)\Bigg)^{-1}, \\
=&\mathcal{O}(e^{1-L}T^{-4}).
\end{aligned}
$$

$$
\begin{aligned}
&\eta \\
&< \frac{1}{4}\frac{\beta^2}{\beta_0^2}\delta_4^{-2}\alpha_0^{-2}, \\
&\overset{\text{Equation (61)}}{<} \frac{1}{4}\frac{\beta^2}{\beta_0^2}\delta_5\Bigg(e^{2L-2}\frac{(1+\beta)\sqrt{\beta}}{6\beta_0^2\beta}\delta_5\|Y\|_2\|\mathbf{M}^\top Y\|_2^3 \\
&\qquad \Big(16(\beta+1)T^5 - (8\beta+20)T^4 - 6(2\beta+1)T^3 + 2(\beta+4)T^2 + 2(\beta+1)T\Big) \\
&\qquad L\prod_{\ell=1}^{L-1}(\|W_\ell^0\|_2+1)^2\Bigg)^{-\frac{2}{3}} \\
&= \mathcal{O}(e^{\frac{4}{3}(1-L)}T^{-\frac{10}{3}}).
\end{aligned}
$$

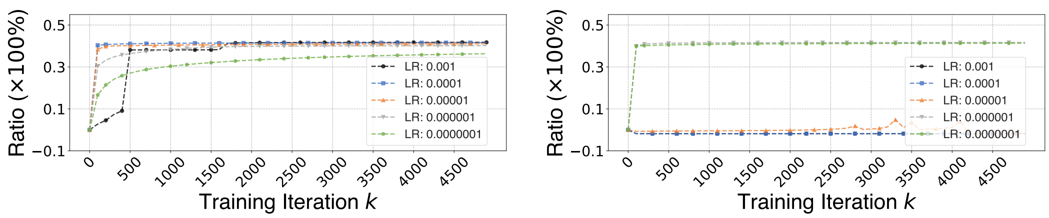
$$
\eta < \frac{1}{4}\frac{\beta^2}{\beta_0^2}\delta_4^{-2}\alpha_0^{-2} \overset{\text{Equation (62)}}{<} \frac{1}{4}\frac{\beta^2}{\beta_0^2}\delta_5\mathcal{O}((e^{TL-T+2L-4}T^{3T+6})^{-1}) = \mathcal{O}(e^{-TL-2L+T+4}T^{-3T-6}).
$$

$$
\eta < \frac{1}{4}\frac{\beta^2}{\beta_0^2}\delta_4^{-2}\alpha_0^{-2} \overset{\text{Equation (12a)}}{<} \frac{1}{4}\frac{\beta^2}{\beta_0^2}\delta_5\Bigg(8(1+\beta)(\|X_0\|_2 + \frac{2T-2}{\beta}\|\mathbf{M}^\top Y\|_2)\Bigg)^{-2} = \mathcal{O}(T^{-2}).
$$

□

C.5 Additional Ablation Study for Learning Rates

We present two additional ablation study with e of 25 and 100. Both use the configuration 1 in Table 1. The results are in Figure 9, which shows a deterministic relationship between LR and expansion coefficient. For $e = 25$ in Figure 9a, the 10^{-7} LR is too small and leads to worse optimality. The large LRs, i.e., 10^{-3}, 10^{-4}, causes unstable convergence. Similarly, for $e = 100$ in Figure 9b, a proper LR is 10^{-4}.



(a) $e = 25$ (b) $e = 100$

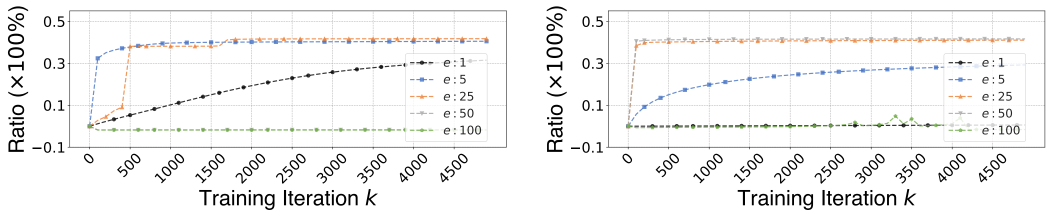
Figure 9: Additional Ablation Studies of Learning Rate with Different e.

C.6 Additional Ablation Study for Expansion Coefficient e in Initialization

We present two additional ablation study for e with learning rates of 0.001 and 0.00001. Both use the configuration 1 in Table 1. The results are in Figure 10. For a large LR, a large e may cause poor convergence due to Theorem 4.3. From Figure 10a, $e = 25$ is a proper setting for best convergence with $\eta = 0.001$. Similarly, for $\eta = 0.00001$, $e = 5$ is enough.

D Discussion

Scope of Theoretical Guarantees. Our theoretical analysis establishes convergence guarantees and demonstrates superior convergence rates specifically for *over-parameterized* Math-L2O systems compared to baseline optimization algorithms. While we acknowledge the empirical effectiveness of



(a) $\eta = 0.001$

(b) $\eta = 0.00001$

Figure 10: Additional Ablation Studies of e with Different Learning Rates.

certain *under-parameterized* Math-L2O systems [22, 32], providing theoretical convergence proofs for them remains challenging due to the inherent non-convexity of the underlying neural network training. Alternative theoretical approaches, such as convex dualization [16, 17, 29], have been explored. However, these methods typically necessitate the inclusion of regularization terms within the loss function, which may deviate from the original optimization objective we aim to solve.

Choice of Base Algorithm. Our framework utilizes Gradient Descent (GD) as the core algorithm primarily because it admits a direct analytical formulation relating the initial point X_0 to the iterate X_T. This tractability is crucial for our analysis. In contrast, accelerated variants like Nesterov Accelerated Gradient Descent (NAG) [4] generally lack such closed-form expressions for X_T. This absence significantly complicates the derivation of the output bounds required to analyze the L2O system's dynamics and prove convergence guarantees. Consequently, rigorously extending our current theoretical framework to momentum-based methods, despite attempts using inductive approaches, remains an open challenge.

E Impact Statement

This paper presents work whose goal is to advance the field of Learning Theory and its combination with optimization. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 6 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: If accepted, we will open source the codes.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 6 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 6 and Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: See Appendix E.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: See Appendix E.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.