

A generalization bound for exit wave reconstruction via deep unfolding

MOUSSA ATWI* AND BENJAMIN BERKELS^{ID}

IGPM, RWTH AACHEN UNIVERSITY, SCHINKELSTR. 2, 52062 AACHEN, GERMANY

*Corresponding author: atwi@ssd.rwth-aachen.de

Transmission Electron Microscopy enables high-resolution imaging of materials, but the resulting images are difficult to interpret directly. One way to address this is exit wave reconstruction, i.e., the recovery of the complex-valued electron wave at the specimen's exit plane from intensity-only measurements. This is an inverse problem with a nonlinear forward model. We consider a simplified forward model, making the problem equivalent to phase retrieval, and propose a discretized regularized variational formulation. To solve the resulting non-convex problem, we employ the proximal gradient algorithm (PGA) and unfold its iterations into a neural network, where each layer corresponds to one PGA step with learnable parameters. This unrolling approach, inspired by LISTA, enables improved reconstruction quality, interpretability, and implicit dictionary learning from data. We analyze the effect of parameter perturbations and show that they can accumulate exponentially with the number of layers L . Building on proof techniques of Behboodi et al., originally developed for LISTA, i.e., for a linear forward model, we extend the analysis to our nonlinear setting and establish generalization error bounds of order $\mathcal{O}(\sqrt{L})$. Numerical experiments support the exponential growth of parameter perturbations.

Keywords: Phase retrieval; TEM imaging; sparse reconstruction; deep unfolding; proximal gradient algorithm; generalization error; Rademacher complexity.

1. Introduction

Transmission Electron Microscopy (TEM) is a microscopy technique in which two-dimensional images are created by recording a beam of electrons passing through a thin sample (specimen). After the electrons interact with the material and leave the exit plane of the specimen, they go through a system of electromagnetic lenses and finally their squared amplitude is recorded, forming a TEM image. However, the direct interpretation of TEM images is difficult because they are blurred by aberrations from the lenses and further affected by the loss of phase information. One possible way to deal with these limitations is to reconstruct the electron wave at the exit plane of the specimen, which is known as the *exit wave* [5]. Exit wave reconstruction provides access to phase information that is otherwise lost in direct TEM imaging. This results in a nonlinear inverse problem that can be seen as a generalized variant of phase retrieval. A range of strategies has been developed for this task, from early focus-variation approaches addressing both linear and nonlinear imaging effects [5] to exit wave reconstruction using the transport of intensity equation combined with self-consistent propagation [12]. More recent work introduced a variational framework for exit wave reconstruction [8] generalizing some of the older approaches, others proposed improved focal-series algorithms that incorporate noise suppression and allow tracking structural evolution in nanomaterials [19].

The variational approach from [8] forms the starting point for us. We consider a simplified forward model where the nonlinearity is just element-wise phase loss. Despite the simplification, this is still a challenging nonlinear inverse problem. To address it, we use the proximal gradient algorithm (PGA) to solve the regularized variational problem. Building on advances in algorithm unrolling, we unfold the iterative PGA into a neural network architecture inspired by the Iterative Shrinkage-Thresholding Algorithm (ISTA) and its learned counterpart LISTA. This unfolding maps each iteration to a network layer with learnable parameters, providing a structured and interpretable model that combines traditional optimization with deep learning [10, 15].

A critical aspect here is understanding the generalization ability of the unfolded network, i.e., how well a learned neural network performs on unseen data. Generalization bounds provide probabilistic estimates of the gap between the true error with respect to the unknown distribution and the empirical error on the training set, and thus form the theoretical basis of our analysis. Using tools from statistical learning theory [3, 17], particularly building on the strategy to derive generalization bounds for LISTA from [3], we derive such a bound for our exit wave reconstruction variant. The bound has the form

$$\mathcal{L}(h) - \hat{\mathcal{L}}(h) \lesssim \sqrt{\frac{N^2}{m}}(1 + \sqrt{K})\sqrt{L},$$

where L is the depth of the network, N^2 the number of trainable parameters, KN^2 the number of measurements and m the number of samples.

2. Forward Model and Variational Approach

As described in the introduction, exit wave reconstruction aims to recover the complex-valued exit wave from intensity-only measurements recorded at different defocus levels. The following presents both the full forward model and a simplified version suitable for deep unfolding, leading to a variational approach.

2.1. A Simplified Exit Wave Model - Phase Retrieval

Full forward model: Let $\Psi \in L^2(\mathbb{R}^d; \mathbb{C})$ denote the complex-valued exit wave, $T_j \in L^\infty(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{C})$ the transmission cross-coefficient (TCC). We assume a sequence of real-valued TEM images $(\tilde{g}_j)_{j=1}^K$ has been recorded at different defocus levels. Our objective is to reconstruct Ψ from these intensity-only measurements. This means to find Ψ such that

$$\Psi \star_{T_j} \Psi = \mathcal{F}(\tilde{g}_j) \text{ for } j = 1, \dots, K.$$

Here, $\Psi \star_{T_j} \Psi$ is the weighted auto-correlation, i.e.,

$$(\Psi \star_{T_j} \Psi)(x) = \int_{\mathbb{R}^d} \Psi(x+y) T_j(x+y, y) \Psi^*(y) dy.$$

This gives rise to the objective functional

$$\mathcal{E}[\Psi] = \frac{1}{K} \sum_{j=1}^K \left\| \Psi \star_{T_j} \Psi - \mathcal{F}(\tilde{g}_j) \right\|_{L^2}^2 + \mathcal{R}[\Psi],$$

where $\mathcal{R}[\Psi]$ is a regularization term.

Simplified forward model: To reduce computational complexity, we assume the TCCs are separable, i.e., that there exist weighting functions $w_j \in L^\infty(\mathbb{R}^d; \mathbb{C})$ such that $T_j(x, y) = \overline{w_j(x)} w_j(y)$. Under this assumption, the weighted auto-correlation simplifies to $\Psi \star_{T_j} \Psi = (\Psi w_j) \star (\Psi w_j)$, and using the convolution theorem $\mathcal{F}^{-1}(\Psi \star_{T_j} \Psi) = |\mathcal{F}^{-1}(\Psi w_j)|^2$, we get the simplified forward model

$$\tilde{g}_j = |\mathcal{F}^{-1}(\Psi w_j)|^2$$

and the simplified objective functional

$$\mathcal{E}[\Psi] = \frac{1}{K} \sum_{j=1}^K \left\| |\mathcal{F}^{-1}(\Psi w_j)|^2 - \tilde{g}_j \right\|_{L^2}^2 + \mathcal{R}[\Psi].$$

Despite being simplified compared to the original exit wave reconstruction problem, this is still an inverse problem with a nonlinear forward model.

2.2. Discretization

To solve the variational problem numerically, we discretize the domain using N spatial nodes. The continuous exit wave becomes a vector $\psi \in \mathbb{C}^N$, and each image \tilde{g}_j becomes a vector $\tilde{\mathbf{g}}_j \in \mathbb{R}^N$.

We define measurement operators $A_j \in \mathbb{C}^{N \times N}$ representing an inverse discrete Fourier transform followed by pointwise multiplication with the weighting function

$$A_j \psi = \mathcal{F}_d^{-1}(\psi \odot w_j),$$

where \odot denotes element-wise multiplication. The discretized objective function becomes:

$$\tilde{\mathcal{E}}(\psi) = \tilde{\mathcal{D}}(\psi) + \mathcal{R}(\psi), \text{ with data term } \tilde{\mathcal{D}}(\psi) := \frac{1}{2KN} \sum_{j=1}^K \|\tilde{\varphi}(A_j \psi) - \tilde{\mathbf{g}}_j\|_2^2.$$

Here, $\tilde{\varphi} : \mathbb{C} \rightarrow \mathbb{R}$, $\tilde{\varphi}(z) = |z|^2$, is applied element-wise, i.e.,

$$\tilde{\varphi}(\mathbf{z}) := (\tilde{\varphi}(z_1), \dots, \tilde{\varphi}(z_{KN})), \quad \mathbf{z} \in \mathbb{C}^{KN}.$$

By stacking the measurement operators and image vectors, i.e.,

$$A = \begin{pmatrix} A_1 \\ \vdots \\ A_K \end{pmatrix} \in \mathbb{C}^{KN \times N}, \quad \tilde{\mathbf{g}} = \begin{pmatrix} \tilde{\mathbf{g}}_1 \\ \vdots \\ \tilde{\mathbf{g}}_K \end{pmatrix} \in \mathbb{R}^{KN}, \quad (2.1)$$

the data term can be written compactly as

$$\tilde{\mathcal{D}}(\psi) := \frac{1}{2KN} \|\tilde{\varphi}(A \psi) - \tilde{\mathbf{g}}\|_2^2.$$

2.3. Variational Perspective

We introduce a general transformation strategy to improve robustness. Let $\gamma \in C^1([0, \infty), [0, \infty))$ be a strictly increasing function, and define

$$\varphi := \gamma \circ \tilde{\varphi}, \quad \mathbf{g} := \gamma(\tilde{\mathbf{g}}),$$

so that

$$\varphi(\mathbf{A}\psi) = \mathbf{g}.$$

A possible choice is $\gamma(x) = \sqrt{|x| + \delta^2} - \delta$, $\delta > 0$. In this case,

$$\varphi(x) = \sqrt{|x|^2 + \delta^2} - \delta, \quad \delta > 0.$$

i.e., φ is the scaled Pseudo Huber transformation. It has been observed that the smoothing of the absolute value done here not only improves the theoretical properties of the objective function, but can also improve the practical performance of reconstruction algorithms [18].

The resulting smoothed energy functional is: We write the energy as:

$$\mathcal{E}[\psi] = \mathcal{D}(\psi) + \mathcal{R}(\psi), \quad (2.2)$$

where the data term is

$$\mathcal{D}(\psi) := \frac{1}{2KN} \|\varphi(\mathbf{A}\psi) - \mathbf{g}\|_2^2,$$

and $\mathcal{R}(\psi)$ is a convex regularizer, e.g., promoting sparsity via ℓ_1 -norm.

Optimization via Proximal Gradient Method The minimization is performed using the proximal gradient method, adapted to complex variables via the Wirtinger derivative:

$$\psi^{k+1} = \text{prox}_{\tau_k \mathcal{R}} \left(\psi^k - \tau_k \nabla \mathcal{D}(\psi^k) \right),$$

with gradient

$$\nabla \mathcal{D}(\psi) = \frac{1}{KN} \overline{\mathbf{A}^\top} (\varphi(\mathbf{A}\psi) - \mathbf{g}) \odot \varphi'(\mathbf{A}\psi).$$

Since the \mathcal{D} is real-valued, its Wirtinger derivative coincides with the standard (real) derivative when interpreting \mathbb{C} as \mathbb{R}^2 .

2.4. Deep Unfolding

Deep unfolding, or algorithm unrolling, is a technique that transforms classical iterative algorithms into trainable neural networks and introduces parameters that can be learned from data, resulting in a hybrid model that leverages both domain knowledge and learning capabilities [15].

Traditional deep networks, while powerful, often operate as black boxes with complex structures that make their decision-making difficult to interpret. Their performance heavily relies on large, high-quality training datasets, which are often expensive to acquire, especially in fields such as medical imaging or electron microscopy. In contrast, classical iterative algorithms are transparent and theoretically grounded, offering a clear link between algorithm steps and the problem structure.

By unrolling such algorithms into neural networks, we retain interpretability and domain fidelity, while benefiting from the hybrid model's faster convergence and parameter efficiency. Here, the convergence speed improvement is relative to the baseline iterative method, and the reduced parameter count arises from the structured, hybrid design rather than from deep learning alone. Each layer of the network mimics one iteration, and the resulting model can be trained end-to-end on real data. This allows the network to generalize well even with limited data and provides a more explainable and data-efficient alternative to standard deep architectures.

The foundation of deep unfolding was notably laid by the unrolling of ISTA (Iterative Shrinkage-Thresholding Algorithm), originally formulated in [7], and later unrolled by Gregor and LeCun in their introduction of Learned ISTA (LISTA) [10]. In their experimental study, they demonstrated that LISTA significantly outperformed the classical ISTA algorithm in terms of reconstruction speed and accuracy. This substantial improvement in performance was primarily attributed to the unique network architecture enabled by the unrolling process itself, which reinterprets the optimization dynamics as a trainable feed-forward structure. Another notable example of algorithm unrolling is in blind image deblurring, where the iterative steps of a Half-Quadratic Splitting algorithm for gradient-domain deblurring are unrolled to construct a deep network, referred to as DUBLID [13]. In this approach, each iteration corresponds to a network layer, and key parameters such as filters and thresholds are learned from data, resulting in both interpretable and computationally efficient deblurring. Similarly, the Deep Griffin-Lim Iteration (DeGLI) framework for phase reconstruction [14] unrolls the classical Griffin-Lim algorithm, with each iteration augmented by a trainable residual DNN block. Another significant contribution to the field was made by Hershey et al. [11], who formalized a general deep unfolding framework. They demonstrated how iterative model-based inference algorithms, such as belief propagation and non-negative matrix factorization, can be unrolled into trainable deep network architectures.

2.5. Unrolling the Proximal Gradient Algorithm into a Neural Network

We now apply deep unfolding to the proximal gradient algorithm applied for the minimization of (2.2), i.e., for variational exit wave reconstruction as introduced above.

We model each exit wave $\psi \in \mathbb{C}^N$ as a sparse signal with respect to an (unknown) dictionary in the form of a unitary matrix $\Phi_0 \in \mathcal{U}(N)$, i.e., $\psi = \Phi_0 \mathbf{z}$ for some sparse vector $\mathbf{z} \in \mathbb{C}^N$. Here, $\mathcal{U}(N) := \left\{ \mathbf{U} \in \mathbb{C}^{N \times N} : \mathbf{U}^\top \mathbf{U} = \mathbf{I} \right\}$ is the set of unitary $N \times N$ matrices. We assume that this sparse generative model holds across the training data, i.e., for training samples $((\psi_j, \mathbf{g}_j))_{j=1}^m$, each fulfills $\psi_j = \Phi_0 \mathbf{z}_j$ and $\mathbf{g}_j = \varphi(\mathbf{A}\psi_j)$. The corresponding loss function measures how well the reconstruction of ψ_j from the estimated sparse code and dictionary matches the true exit wave. To obtain the sparse codes \mathbf{z}_j for one exit wave, only the measurements \mathbf{g}_j are used. In contrast, the dictionary Φ is learned from the full training set $((\psi_j, \mathbf{g}_j))_{j=1}^m$.

We design an unrolled neural network architecture that mimics L iterations of the proximal gradient algorithm, replacing the unknown fixed dictionary Φ_0 by a trainable parameter $\Phi \in \mathcal{U}(N)$. Each iteration is referred to as a *stage*, consisting of a gradient descent step and an application of the proximal operator. Each stage maps an input sparse code estimate to a refined estimate, guided by the observed intensity data $\mathbf{g} \in \mathbb{R}^{KN}$ and the learned dictionary Φ .

We assume that signals ψ are bounded in the ℓ_2 -norm by a fixed constant C_{in} , i.e., $\|\psi\|_2 \leq C_{\text{in}}$. Note that this constant will appear in the generalization bound we are going to derive.

The ℓ -th stage of the network is defined as:

$$f_{\ell+1}^{\mathbf{g}, \Phi}(\mathbf{z}) := \text{prox}_{\tau_\ell \mathcal{R}} \left(\mathbf{z} - \frac{\tau_\ell}{KN} (\overline{\mathbf{A}\Phi})^\top (\varphi(\mathbf{A}\Phi\mathbf{z}) - \mathbf{g}) \odot \varphi'(\mathbf{A}\Phi\mathbf{z}) \right), \quad (2.3)$$

Here, by replacing the original signal ψ with a different basis representation $\Phi\mathbf{z}$, we are effectively replacing A by $A\Phi$ in the data term \mathcal{D} . The matrix $\Phi \in \mathbb{C}^{N \times N}$ is searched over the space of unitary matrices to best fit the observed measurements. The step sizes τ_ℓ are fixed and not learned. φ' is the Wirtinger derivative of the real-valued φ . \mathcal{R} is the regularizer introduced earlier, e.g., the ℓ_1 -norm.

Stacking L such stages yields the full unrolled network:

$$f_{\Phi}^L(\mathbf{g}) = f_L^{\mathbf{g}, \Phi} \circ f_{L-1}^{\mathbf{g}, \Phi} \circ \cdots \circ f_1^{\mathbf{g}, \Phi}(\mathbf{z}_0), \quad (2.4)$$

with initialization \mathbf{z}_0 obtained via the spectral initialization from Wirtinger flow, cf. [4, Algorithm 1].

To ensure boundedness of the output of the network, the 1-Lipschitz function σ

$$\sigma(x) = \begin{cases} x & \text{if } \|x\|_2 \leq C_{\text{out}}, \\ C_{\text{out}} \frac{x}{\|x\|_2} & \text{otherwise,} \end{cases} \quad (2.5)$$

is applied after mapping the sparse code through the learned dictionary Φ . This way, the network output $\hat{\psi} = \sigma(\Phi f_{\Phi}^L(\mathbf{g}))$ lies within the ℓ_2 -ball of radius C_{out} . For the sake of simplicity, we choose $C_{\text{out}} = C_{\text{in}}$.

The associated hypothesis class is

$$\mathcal{H}_1^L = \{\sigma \circ (\Phi f_{\Phi}^L) : \Phi \in \mathcal{U}(N)\}, \quad \text{embedded in} \quad \mathcal{H}_2^L = \{\sigma \circ (\Psi f_{\Phi}^L) : \Phi, \Psi \in \mathcal{U}(N)\}. \quad (2.6)$$

3. Preliminaries

The following operator-theoretic definitions and lemmas provide the mathematical foundation to analyze the behavior of the operators underlying our network, in particular those associated with the nonlinearities resulting from the smoothed amplitude φ . Although these results are stated in real Hilbert spaces in [1], they remain valid in complex Hilbert spaces when replacing the inner product that occurs in the statements and definitions for real Hilbert spaces, by the real part of the inner product. These concepts will be used later to analyze the effect of perturbations on the unfolded network iterations.

In the following, let \mathcal{H} be a (real or complex) Hilbert space.

Definition 1 Let $D \subset \mathcal{H}$ nonempty and $T : D \rightarrow \mathcal{H}$.

- T is called firmly nonexpansive, if

$$(\forall x \in D)(\forall y \in D) \quad \|Tx - Ty\|^2 + \|(\text{id} - T)x - (\text{id} - T)y\|^2 \leq \|x - y\|^2.$$

- Let $\beta \in \mathbb{R}_{++} := (0, \infty)$. T is called β -cocoercive (or β -inverse strongly monotone), if

$$(\forall x \in D)(\forall y \in D) \quad \text{Re} \langle x - y | Tx - Ty \rangle \geq \beta \|Tx - Ty\|^2.$$

- Let $L \in \mathbb{R}_+ := [0, \infty)$. T is called L -Lipschitz continuous, if

$$(\forall x \in D)(\forall y \in D) \quad \|Tx - Ty\| \leq L \|x - y\|.$$

The properties defined above are central in analyzing proximal gradient iterations, since the gradient-like and proximal-like steps correspond to operators with Lipschitz and firmly nonexpansive behavior, respectively. By establishing these operator-theoretic properties, we can later show the effect of perturbations on the unrolled network iterations and ensure that the iterates remain bounded.

We summarize several important properties of operators and proximal mappings in \mathcal{H} below.

Proposition 1 (Properties of cocoercive and proximal operators) *Let $D \subseteq \mathcal{H}$ be nonempty, and let $T : D \rightarrow \mathcal{H}$ be an operator. Let $\alpha, \beta \in \mathbb{R}_{++}$ and $f \in \Gamma_0(\mathcal{H})$. Then,*

- T is firmly nonexpansive if and only if it is 1-cocoercive.
- If T is β -cocoercive, then αT is $\frac{\beta}{\alpha}$ -cocoercive.
- T is β -cocoercive if and only if βT is firmly nonexpansive.
- If T is β -cocoercive, then T is $\frac{1}{\beta}$ -Lipschitz continuous.
- $\text{prox}_f : \mathcal{H} \rightarrow \mathcal{H}$ is firmly nonexpansive, cf. [1, Proposition 12.28].
- By [1, Proposition 24.8(v)], one gets the identity

$$\alpha \text{prox}_f(x) = \text{prox}_{\alpha^2 f(\frac{1}{\alpha})}(\alpha x) \quad \text{for all } x \in \mathcal{H}. \quad (3.1)$$

- Let $0 \in \arg \min_{x \in \mathbb{R}^n} f(x)$. Then,

$$\text{prox}_f(0) = 0, \quad (3.2)$$

as an immediate consequence of [1, Proposition 12.29] and

$$0 \in \arg \min_{x \in \mathbb{R}^n} f^*(x). \quad (3.3)$$

Proposition 2 (Real variant: [1], Proposition 4.12) *Let I be a finite set. For every $i \in I$, let \mathcal{K}_i be real or complex Hilbert space, $L_i \in \mathcal{B}(\mathcal{H}, \mathcal{K}_i) \setminus \{0\}$, $\beta_i \in \mathbb{R}_{++}$, and $T_i : \mathcal{K}_i \rightarrow \mathcal{K}_i$ be β_i -cocoercive. Moreover, let*

$$T = \sum_{i \in I} L_i^* T_i L_i \quad \text{and} \quad \beta = \frac{1}{\sum_{i \in I} \frac{\|L_i\|^2}{\beta_i}}.$$

Then, T is β -cocoercive.

Corollary 1 (Real variant: [1], Corollary 4.13) *Let \mathcal{H}, \mathcal{K} be a real or complex Hilbert spaces, let $T : \mathcal{K} \rightarrow \mathcal{K}$ be firmly nonexpansive, and let $L \in \mathcal{B}(\mathcal{H}, \mathcal{K})$ be such that $\|L\| \leq 1$. Then, $L^* T L$ is firmly nonexpansive.*

Lemma 1 *Let*

$$\phi_1 : \mathbb{R} \rightarrow \mathbb{R} \cup \{\infty\}, \quad x \mapsto \begin{cases} -\frac{x^2}{2} - \sqrt{1-x^2} & \text{if } |x| \leq 1, \\ +\infty & \text{otherwise} \end{cases} \quad (3.4)$$

Then, $\phi_1 \in \Gamma_0(\mathbb{R})$, ϕ_1 is even, $\arg \min_{x \in \mathbb{R}} \phi_1(x) = 0$ and $\text{prox}_{\phi_1}(y) = \frac{y}{\sqrt{1+y^2}}$.

This lemma provides a key connection between the smoothed amplitude nonlinearity used in the loss function and a proximal operator, allowing us to apply the theoretical results and properties of proximal mappings in complex Hilbert spaces.

Lemma 2 (Norm Composition [2, Theorem 6.18]) *Let \mathcal{H} be a real Hilbert space, and let $f : \mathcal{H} \rightarrow \mathbb{R}$ be defined by*

$$f(x) = g(\|x\|),$$

where $g \in \Gamma_0(\mathbb{R})$ and $\text{dom}(g) \subseteq \mathbb{R}_+$. Then, the proximal operator of f is given by

$$\text{prox}_f(x) = \begin{cases} \text{prox}_g(\|x\|) \frac{x}{\|x\|}, & \text{if } x \neq 0, \\ \{y \in \mathcal{H} : \|y\| = \text{prox}_g(0)\}, & \text{if } x = 0. \end{cases}$$

In the following, we use these results to explicitly represent φ_δ and φ'_δ as proximal operators in the complex setting (Remark 2), which makes it possible to formulate the unfolded iterations in the form of proximal operators.

Remark 1 *Let $g := \phi_1 + I_{\mathbb{R}_+}$ with ϕ_1 from (3.4) and $I_{\mathbb{R}_+}(t) = \begin{cases} 0 & t \in \mathbb{R}_+, \\ \infty & t \notin \mathbb{R}_+ \end{cases}$. Then, $g \in \Gamma_0(\mathbb{R})$ and for $t \in \mathbb{R}_+$*

$$\begin{aligned} \text{prox}_g(t) &= \text{prox}_{\phi_1 + I_{\mathbb{R}_+}}(t) = \arg \min_{s \in \mathbb{R}} (\phi_1(s) + I_{\mathbb{R}_+}(s) + \frac{1}{2}(s-t)^2) \\ &= \arg \min_{s \in \mathbb{R}_+} (\phi_1(s) + \frac{1}{2}(s-t)^2) = P_{\mathbb{R}_+} \left(\arg \min_{s \in \mathbb{R}} (\phi_1(s) + \frac{1}{2}(s-t)^2) \right) \\ &= P_{\mathbb{R}_+} (\text{prox}_{\phi_1}(t)) \stackrel{\text{Lemma 1}}{=} P_{\mathbb{R}_+} \left(\frac{t}{\sqrt{1+t^2}} \right) \stackrel{t \in \mathbb{R}_+}{=} \frac{t}{\sqrt{1+t^2}}. \end{aligned}$$

Here, $P_{\mathbb{R}_+}$ denotes the Euclidean projection from \mathbb{R} to \mathbb{R}_+ . From this, we get

$$\{y \in \mathbb{R}^n : \|y\| = \text{prox}_g(0)\} = \{y \in \mathbb{R}^n : \|y\| = 0\} = \{0\}.$$

For $x \in \mathbb{R}^n \setminus \{0\}$, we get

$$\text{prox}_g(\|x\|) \frac{x}{\|x\|} = \frac{\|x\|}{\sqrt{1+\|x\|^2}} \frac{x}{\|x\|} = \frac{x}{\sqrt{1+\|x\|^2}}.$$

Combined with the above, we get from Lemma 2 for $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $x \mapsto g(\|x\|)$ that

$$\text{prox}_f(x) = \begin{cases} \text{prox}_g(\|x\|) \frac{x}{\|x\|}, & x \neq 0, \\ 0, & x = 0. \end{cases} = \frac{x}{\sqrt{1 + \|x\|^2}}.$$

Moreover, $f \in \Gamma_0(\mathbb{R}^n)$, since $\phi_1 \in \Gamma_0(\mathbb{R})$ and thus $g \in \Gamma_0(\mathbb{R})$.

Remark 2 Let $\varphi_\delta : \mathbb{C} \rightarrow \mathbb{R}$, $z \mapsto \sqrt{|z|^2 + \delta^2} - \delta$ for $\delta > 0$ and identify \mathbb{C} with \mathbb{R}^2 . Then, we have for the Wirtinger derivative of φ_δ that

$$\varphi'_\delta(z) = \frac{z}{\sqrt{|z|^2 + \delta^2}} = \frac{\frac{z}{\delta}}{\sqrt{\left|\frac{z}{\delta}\right|^2 + 1}} \stackrel{\text{Remark 1}}{=} \text{prox}_f\left(\frac{z}{\delta}\right) \stackrel{(3.1)}{=} \frac{1}{\delta} \text{prox}_{\delta^2 f(\frac{\cdot}{\delta})}(z).$$

Moreover, we have

$$\varphi_\delta(z)\varphi'_\delta(z) = \left(\sqrt{|z|^2 + \delta^2} - \delta\right) \frac{z}{\sqrt{|z|^2 + \delta^2}} = z - \frac{\delta z}{\sqrt{|z|^2 + \delta^2}} = z - \delta\varphi'_\delta(z) = z - \text{prox}_{\delta^2 f(\frac{\cdot}{\delta})}(z)$$

Let

$$f_\delta : \mathbb{C} \rightarrow \mathbb{R}, \quad z \mapsto \delta^2 f\left(\frac{z}{\delta}\right). \quad (3.5)$$

Since $f \in \Gamma_0(\mathbb{C})$, we also have $f_\delta \in \Gamma_0(\mathbb{C})$. Thus, by [1, Corollary 13.38], $f_\delta^* \in \Gamma_0(\mathbb{C})$, where f_δ^* denotes the convex conjugate (or Fenchel conjugate) of f_δ . By Moreau's decomposition ([6, Theorem 2.1]), we get

$$z - \text{prox}_{f_\delta}(z) = \text{prox}_{f_\delta^*}(z), \quad (3.6)$$

In total, we have shown

$$\varphi'_\delta(z) = \frac{1}{\delta} \text{prox}_{f_\delta}(z) \quad \text{and} \quad \varphi_\delta(z)\varphi'_\delta(z) = \text{prox}_{f_\delta^*}(z). \quad (3.7)$$

Again using (3.6), we get for $g \in \mathbb{R}$,

$$\varphi_\delta(z)\varphi'_\delta(z) - g\varphi'_\delta(z) = z - \text{prox}_{f_\delta}(z) - \frac{g}{\delta} \text{prox}_{f_\delta}(z) = z - \left(1 + \frac{g}{\delta}\right) \text{prox}_{f_\delta}(z).$$

Moreover, since $f_\delta \in \Gamma_0(\mathbb{C})$ and $0 \in \arg \min_{z \in \mathbb{C}} f_\delta(z)$, by (3.2), we get $\text{prox}_{f_\delta}(0) = 0$. Moreover, using (3.3), we can analogously conclude $\text{prox}_{f_\delta^*}(0) = 0$.

Remark 3 From Remark 2, one can obtain by taking the modulus of φ'_δ

$$|\varphi'_\delta(z)| = \frac{|z|}{\sqrt{|z|^2 + \delta^2}} = \frac{\sqrt{|z|^2}}{\sqrt{|z|^2 + \delta^2}} \leq \frac{\sqrt{|z|^2 + \delta^2}}{\sqrt{|z|^2 + \delta^2}} = 1. \quad (3.8)$$

4. Exponential Growth

In the following, we establish a bound on the network output as well as a perturbation bound for the network parameters. The latter shows that the network output is Lipschitz with respect to the network parameters.

Assumption 1. Let $\Phi \in \mathcal{U}(N)$, $\mathbf{A} \in \mathbb{C}^{KN \times N}$ and $\tau_\ell \in \mathbb{R}_{>0}$ with

$$\frac{\tau_\ell}{KN} \|\mathbf{A}\|_{2 \rightarrow 2}^2 \leq 1. \quad (4.1)$$

Definition 2 For $z \in \mathbb{C}^N$, define the operators

$$T_\Phi(z) := \frac{1}{KN} (\mathbf{A}\Phi)^* T(\mathbf{A}\Phi z) \quad \text{and} \quad S_\Phi^\mathbf{g}(z) := \frac{1}{KN} (\mathbf{A}\Phi)^* S^\mathbf{g}(\mathbf{A}\Phi z), \quad (4.2)$$

where

$$T(z) := \text{prox}_{f_\delta^*}(z) \quad \text{and} \quad S^\mathbf{g}(z) := \frac{\mathbf{g}}{\delta} \odot \text{prox}_{f_\delta}(z).$$

Remark 4 By Definition 2, we have for $z \in \mathbb{C}^N$ that

$$T_\Phi(z) = \frac{1}{KN} (\mathbf{A}\Phi)^* \text{prox}_{f_\delta^*}(\mathbf{A}\Phi z) \quad \text{and} \quad S_\Phi^\mathbf{g}(z) = \frac{1}{KN} (\mathbf{A}\Phi)^* \frac{\mathbf{g}}{\delta} \odot \text{prox}_{f_\delta}(\mathbf{A}\Phi z),$$

Moreover, since prox_{f_δ} and $\text{prox}_{f_\delta^*}$ vanish at zero (cf. Remark 2), we get

$$S_\Phi^\mathbf{g}(0) = \text{prox}_{f_\delta}(0) = 0 = \text{prox}_{f_\delta^*}(0) = T_\Phi(0).$$

Remark 5 The following inequalities will be used later.

- Let $\mathbf{A} \in \mathbb{C}^{m \times n}$ and $\mathbf{z} \in \mathbb{C}^n$. Then,

$$\|\mathbf{A}\mathbf{z}\|_2 \leq \|\mathbf{A}\|_{2 \rightarrow 2} \|\mathbf{z}\|_2, \quad (4.3)$$

which follows directly from the definition of the operator norm.

- Let $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{m \times n}$. Then,

$$\|\mathbf{x} \odot \mathbf{y}\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n |x_{ij} y_{ij}|^2 \leq \max_{i,j} |x_{ij}|^2 \sum_{i=1}^m \sum_{j=1}^n |y_{ij}|^2 = \|\mathbf{x}\|_\infty^2 \|\mathbf{y}\|_F^2, \quad (4.4)$$

where we used the entrywise inequality $|x_{ij} y_{ij}|^2 \leq (\max_{i,j} |x_{ij}|^2) |y_{ij}|^2$.

In particular, for vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$, we have

$$\|\mathbf{x} \odot \mathbf{y}\|_2^2 \leq \|\mathbf{x}\|_\infty^2 \|\mathbf{y}\|_2^2. \quad (4.5)$$

Lemma 3 *If Assumption 1 holds, $\tau_\ell T_\Phi$ is firmly nonexpansive and $\tau_\ell S_\Phi^g$ is $\frac{\|\mathbf{g}\|_\infty}{\delta}$ -Lipschitz continuous.*

Proof Recalling (4.2), we have

$$KNT_\Phi(z) = (\mathbf{A}\Phi)^* \text{prox}_{f_\delta^*}(\mathbf{A}\Phi z) = L^* \text{prox}_{f_\delta^*}(Lz),$$

where $L = \mathbf{A}\Phi$. By applying Proposition 2 and noting that $\text{prox}_{f_\delta^*}$ is 1-cocoercive and $\|\mathbf{A}\Phi\|_{2 \rightarrow 2} = \|\mathbf{A}\|_{2 \rightarrow 2}$, we obtain that KNT_Φ is $\frac{1}{\|\mathbf{A}\|_{2 \rightarrow 2}^2}$ -cocoercive. Thus, by using Proposition 1 with $\alpha = \frac{1}{KN}$, we obtain that T_Φ is $\frac{KN}{\|\mathbf{A}\|_{2 \rightarrow 2}^2}$ -cocoercive. Hence $\tau_\ell T_\Phi$ is $\frac{KN}{\tau_\ell \|\mathbf{A}\|_{2 \rightarrow 2}^2}$ -cocoercive. If Assumption 1 holds, we get $\frac{KN}{\tau_\ell \|\mathbf{A}\|_{2 \rightarrow 2}^2} \geq 1$. Thus, $\tau_\ell T_\Phi$ is 1-cocoercive, and thus firmly nonexpansive.

For the Lipschitz continuity of $\tau_\ell S_\Phi^g$, let $z_1, z_2 \in \mathbb{C}^N$. Then,

$$\|S_\Phi^g(z_1) - S_\Phi^g(z_2)\|_2 = \left\| \frac{1}{KN} (\mathbf{A}\Phi)^* \left(\frac{\mathbf{g}}{\delta} \odot (\text{prox}_{f_\delta}(\mathbf{A}\Phi z_1) - \text{prox}_{f_\delta}(\mathbf{A}\Phi z_2)) \right) \right\|_2.$$

Since prox_{f_δ} is 1-Lipschitz and using (4.3) and (4.5), we get

$$\begin{aligned} \|S_\Phi^g(z_1) - S_\Phi^g(z_2)\|_2 &\leq \frac{1}{KN} \|\mathbf{A}\Phi^*\|_{2 \rightarrow 2} \left\| \frac{\mathbf{g}}{\delta} \odot (\text{prox}_{f_\delta}(\mathbf{A}\Phi z_1) - \text{prox}_{f_\delta}(\mathbf{A}\Phi z_2)) \right\|_2 \\ &\leq \frac{1}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \left\| \frac{\mathbf{g}}{\delta} \right\|_\infty \|\text{prox}_{f_\delta}(\mathbf{A}\Phi z_1) - \text{prox}_{f_\delta}(\mathbf{A}\Phi z_2)\|_2 \\ &\leq \frac{1}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \left\| \frac{\mathbf{g}}{\delta} \right\|_\infty \|\mathbf{A}\Phi z_1 - \mathbf{A}\Phi z_2\|_2 \\ &\leq \frac{1}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \left\| \frac{\mathbf{g}}{\delta} \right\|_\infty \|\mathbf{A}\Phi\|_{2 \rightarrow 2} \|z_1 - z_2\|_2 \\ &= \frac{1}{KN} \|\mathbf{A}\|_{2 \rightarrow 2}^2 \left\| \frac{\mathbf{g}}{\delta} \right\|_\infty \|z_1 - z_2\|_2. \end{aligned}$$

Using Assumption 1, we get that $\tau_\ell S_\Phi^g$ is $\frac{\|\mathbf{g}\|_\infty}{\delta}$ -Lipschitz. \square

Before stating a corollary from the above, we extend the operators T_Φ and S_Φ^G from \mathbb{C}^N to $\mathbb{C}^{N \times m}$ by applying them column-wise. For $\mathbf{Z} = (\mathbf{z}_1 | \cdots | \mathbf{z}_m) \in \mathbb{C}^{N \times m}$, we set

$$\tau_\ell T_\Phi(\mathbf{Z}) := (\tau_\ell T_\Phi(\mathbf{z}_1) | \cdots | \tau_\ell T_\Phi(\mathbf{z}_m)),$$

and similarly, for $\mathbf{G} = (\mathbf{g}_1 | \cdots | \mathbf{g}_m) \in \mathbb{C}^{KN \times m}$,

$$\tau_\ell S_\Phi^G(\mathbf{Z}) := (\tau_\ell S_\Phi^{g_1}(\mathbf{z}_1) | \cdots | \tau_\ell S_\Phi^{g_m}(\mathbf{z}_m)).$$

These extensions are consistent with the Frobenius norm since

$$\|\mathbf{Z}\|_F^2 = \sum_{j=1}^m \|\mathbf{z}_j\|_2^2.$$

Moreover, since $T_\Phi(0) = 0$ and $S_\Phi^G(0) = 0$ by Remark 4, it follows that their column-wise extensions also vanish at $0 \in \mathbb{C}^{N \times m}$.

Corollary 2 Let $\mathbf{G} = (\mathbf{g}_1 | \cdots | \mathbf{g}_m) \in \mathbb{C}^{KN \times m}$ and $\mathbf{Z} = (\mathbf{z}_1 | \cdots | \mathbf{z}_m) \in \mathbb{C}^{N \times m}$. Under Assumption 1, the following hold:

- The operator $\tau_\ell T_{\Phi} : \mathbb{C}^{N \times m} \rightarrow \mathbb{C}^{N \times m}$ is firmly nonexpansive with respect to the Frobenius norm. Moreover, we have
$$\|(\text{id} - \tau_\ell T_{\Phi})(\mathbf{Z})\|_F \leq \|\mathbf{Z}\|_F.$$
- The operator $\tau_\ell S_{\Phi}^{\mathbf{G}} : \mathbb{C}^{N \times m} \rightarrow \mathbb{C}^{N \times m}$ is $\|\mathbf{G}\|_\infty / \delta$ -Lipschitz continuous with respect to the Frobenius norm. Moreover, we have
$$\|\tau_\ell S_{\Phi}^{\mathbf{G}}(\mathbf{Z})\|_F \leq \frac{\|\mathbf{G}\|_\infty}{\delta} \|\mathbf{Z}\|_F.$$

Proof From Lemma 3, the operator $\tau_\ell T_{\Phi} : \mathbb{C}^N \rightarrow \mathbb{C}^N$ is firmly nonexpansive. By construction, the extension to $\mathbb{C}^{N \times m}$ is column-wise and thus respects the Frobenius norm. Explicitly, for $\mathbf{Z} = (\mathbf{z}_1 | \cdots | \mathbf{z}_m)$, we obtain

$$\|(\text{id} - \tau_\ell T_{\Phi})(\mathbf{Z})\|_F^2 = \sum_{j=1}^m \|(\text{id} - \tau_\ell T_{\Phi})(\mathbf{z}_j)\|_2^2 \leq \sum_{j=1}^m \|\mathbf{z}_j\|_2^2 = \|\mathbf{Z}\|_F^2,$$

which shows the Frobenius nonexpansiveness directly.

For the second part, Lemma 3 states that for any $z_1, z_2 \in \mathbb{C}^N$ and corresponding $\mathbf{g} \in \mathbb{C}^{KN}$,

$$\|\tau_\ell S_{\Phi}^{\mathbf{g}}(z_1) - \tau_\ell S_{\Phi}^{\mathbf{g}}(z_2)\|_2 \leq \frac{\|\mathbf{g}\|_\infty}{\delta} \|z_1 - z_2\|_2.$$

Applying this column-wise to $\mathbf{Z}_1 = (\mathbf{z}_{1,1} | \cdots | \mathbf{z}_{1,m})$ and $\mathbf{Z}_2 = (\mathbf{z}_{2,1} | \cdots | \mathbf{z}_{2,m})$, we get

$$\begin{aligned} \|\tau_\ell S_{\Phi}^{\mathbf{G}}(\mathbf{Z}_1) - \tau_\ell S_{\Phi}^{\mathbf{G}}(\mathbf{Z}_2)\|_F^2 &= \sum_{j=1}^m \left\| \tau_\ell S_{\Phi}^{\mathbf{g}_j}(\mathbf{z}_{1,j}) - \tau_\ell S_{\Phi}^{\mathbf{g}_j}(\mathbf{z}_{2,j}) \right\|_2^2 \leq \sum_{j=1}^m \left(\frac{\|\mathbf{g}_j\|_\infty}{\delta} \right)^2 \|\mathbf{z}_{1,j} - \mathbf{z}_{2,j}\|_2^2 \\ &\leq \left(\frac{\|\mathbf{G}\|_\infty}{\delta} \right)^2 \sum_{j=1}^m \|\mathbf{z}_{1,j} - \mathbf{z}_{2,j}\|_2^2 = \left(\frac{\|\mathbf{G}\|_\infty}{\delta} \right)^2 \|\mathbf{Z}_1 - \mathbf{Z}_2\|_F^2. \end{aligned}$$

By Remark 4, we have $S_{\Phi}^{\mathbf{g}}(0) = 0$ for each \mathbf{g} , which implies

$$S_{\Phi}^{\mathbf{G}}(0) = (S_{\Phi}^{\mathbf{g}_1}(0) | \cdots | S_{\Phi}^{\mathbf{g}_m}(0)) = 0.$$

Taking $\mathbf{Z}_1 = \mathbf{Z}$ and $\mathbf{Z}_2 = 0$ in the Lipschitz inequality, we obtain:

$$\|\tau_\ell S_{\Phi}^{\mathbf{G}}(\mathbf{Z})\|_F = \|\tau_\ell S_{\Phi}^{\mathbf{G}}(\mathbf{Z}) - \tau_\ell S_{\Phi}^{\mathbf{G}}(0)\|_F \leq \frac{\|\mathbf{G}\|_\infty}{\delta} \|\mathbf{Z}\|_F. \quad \square$$

Corollary 3 Under Assumption 1, for any $z_1, z_2 \in \mathbb{C}^N$, we have

$$\|(z_1 - z_2) - \tau_\ell [(T_{\Phi} - S_{\Phi}^{\mathbf{g}})(z_1) - (T_{\Phi} - S_{\Phi}^{\mathbf{g}})(z_2)]\|_2 \leq \left(1 + \frac{\|\mathbf{g}\|_\infty}{\delta} \right) \|z_1 - z_2\|_2. \quad (4.6)$$

Proof This is an immediate consequence of Lemma 3 and that the firmly nonexpansiveness of an operator T implies the nonexpansiveness of $\text{id} - T$. \square

Remark 6 Let $\psi^0 \in \mathbb{C}^N$ with $\|\psi^0\|_2 = 1$. For $\Phi \in \mathcal{U}(N)$, $\ell \in \mathbb{N}_0$ and $\mathbf{g} \in \mathbb{C}^{KN}$, we define

$$f_{\Phi}^{\ell+1}(\mathbf{g}) := \text{prox}_{\tau_{\ell}\mathcal{R}} \left(f_{\Phi}^{\ell}(\mathbf{g}) - \frac{\tau_{\ell}}{KN} (\mathbf{A}\Phi)^T \left(\varphi_{\delta}(\mathbf{A}\Phi f_{\Phi}^{\ell}(\mathbf{g})) - \mathbf{g} \right) \odot \varphi'_{\delta}(\mathbf{A}\Phi f_{\Phi}^{\ell}(\mathbf{g})) \right).$$

with initialization

$$f_{\Phi}^0(\mathbf{g}) := \psi^0 \in \mathbb{C}^N, \quad \text{with } \|\psi^0\|_2 = 1.$$

For a matrix $\mathbf{G} = (\mathbf{g}_1 | \cdots | \mathbf{g}_m) \in \mathbb{C}^{KN \times m}$, we extend this definition column-wise, i.e., for $\ell \in \mathbb{N}_0$, we define

$$f_{\Phi}^{\ell}(\mathbf{G}) := (f_{\Phi}^{\ell}(\mathbf{g}_1) | \cdots | f_{\Phi}^{\ell+1}(\mathbf{g}_m)).$$

This implies,

$$\mathbf{f} := f_{\Phi}^0(\mathbf{G}) = (\psi^0 | \cdots | \psi^0) \in \mathbb{C}^{N \times m},$$

i.e., the same initialization ψ^0 is used for each column. Moreover,

$$\|\mathbf{f}\|_F^2 = \|(\psi^0 | \cdots | \psi^0)\|_F^2 = \sum_{j=1}^m \|\psi^0\|_2^2 = m.$$

Remark 7 Using (3.7) element-wise, we get the identity

$$(\varphi_{\delta}(z) - \mathbf{g}) \odot \varphi'_{\delta}(z) = \varphi_{\delta}(z) \odot \varphi'_{\delta}(z) - \mathbf{g} \odot \varphi'_{\delta}(z) = \text{prox}_{f_{\delta}^*}(z) - \frac{\mathbf{g}}{\delta} \odot \text{prox}_{f_{\delta}}(z).$$

Combined with $T_{\Phi}, S_{\Phi}^{\mathbf{g}}$ from Definition 2, the iteration $f_{\Phi}^{\ell+1}(\mathbf{g})$ can be expressed as:

$$\begin{aligned} f_{\Phi}^{\ell+1}(\mathbf{g}) &= \text{prox}_{\tau_{\ell}\mathcal{R}} \left(f_{\Phi}^{\ell}(\mathbf{g}) - \frac{\tau_{\ell}}{KN} (\mathbf{A}\Phi)^* \left(\varphi_{\delta}(\mathbf{A}\Phi f_{\Phi}^{\ell}(\mathbf{g})) - \mathbf{g} \right) \odot \varphi'_{\delta}(\mathbf{A}\Phi f_{\Phi}^{\ell}(\mathbf{g})) \right) \\ &= \text{prox}_{\tau_{\ell}\mathcal{R}} \left(f_{\Phi}^{\ell}(\mathbf{g}) - \frac{\tau_{\ell}}{KN} (\mathbf{A}\Phi)^* \left(\text{prox}_{f_{\delta}^*}(\mathbf{A}\Phi f_{\Phi}^{\ell}(\mathbf{g})) - \frac{\mathbf{g}}{\delta} \odot \text{prox}_{f_{\delta}}(\mathbf{A}\Phi f_{\Phi}^{\ell}(\mathbf{g})) \right) \right) \\ &= \text{prox}_{\tau_{\ell}\mathcal{R}} \left(f_{\Phi}^{\ell}(\mathbf{g}) - \frac{\tau_{\ell}}{KN} (\mathbf{A}\Phi)^* \left(T(\mathbf{A}\Phi f_{\Phi}^{\ell}(\mathbf{g})) - S_{\Phi}^{\mathbf{g}}(\mathbf{A}\Phi f_{\Phi}^{\ell}(\mathbf{g})) \right) \right) \end{aligned} \quad (4.7)$$

$$= \text{prox}_{\tau_{\ell}\mathcal{R}} \left(f_{\Phi}^{\ell}(\mathbf{g}) - \tau_{\ell} \left(T_{\Phi}(f_{\Phi}^{\ell}(\mathbf{g})) - S_{\Phi}^{\mathbf{g}}(f_{\Phi}^{\ell}(\mathbf{g})) \right) \right). \quad (4.8)$$

Lemma 4 Let $\mathbf{G} \in \mathbb{R}^{KN \times m}$ and $\Phi \in \mathcal{U}(N)$. Furthermore, let $\mathcal{R} \in \Gamma_0(\mathbb{C})$ with $0 \in \arg \min_{x \in \mathbb{R}^n} \mathcal{R}(x)$.

Then, if Assumption 1 holds, we have

$$\left\| f_{\Phi}^{\ell}(\mathbf{G}) \right\|_F \leq \sqrt{m} + \frac{1}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \|\mathbf{G}\|_{\infty} \sum_{k=0}^{\ell-1} \tau_k \quad \text{for } \ell \in \mathbb{N}_0. \quad (4.9)$$

Proof Let $k \in \mathbb{N}$. Using (4.8), the nonexpansiveness of proximal mappings, $\text{prox}_{\tau_k \mathcal{R}}(0) = 0$ (cf. (3.2), Proposition 1), and Corollary 2, we obtain that

$$\begin{aligned} \|f_{\Phi}^{k+1}(\mathbf{G})\|_F &\leq \left\| f_{\Phi}^k(\mathbf{G}) - \tau_k \left(T_{\Phi}(f_{\Phi}^k(\mathbf{G})) - S_{\Phi}^{\mathbf{G}}(f_{\Phi}^k(\mathbf{G})) \right) \right\|_F \\ &\leq \left\| (\text{id} - \tau_k T_{\Phi})(f_{\Phi}^k(\mathbf{G})) \right\|_F + \left\| \tau_k S_{\Phi}^{\mathbf{G}}(f_{\Phi}^k(\mathbf{G})) \right\|_F \\ &\leq \|f_{\Phi}^k(\mathbf{G})\|_F + \left\| \tau_k S_{\Phi}^{\mathbf{G}}(f_{\Phi}^k(\mathbf{G})) \right\|_F \\ &= \|f_{\Phi}^k(\mathbf{G})\|_F + \left\| \frac{\tau_k}{KN} (\mathbf{A}\Phi)^*(\mathbf{G} \odot \varphi'_{\delta}(\mathbf{A}\Phi f_{\Phi}^k(\mathbf{G}))) \right\|_F \end{aligned}$$

Since $\varphi'_{\delta}(\mathbf{A}\Phi f_{\Phi}^k(\mathbf{G})) \leq L_{\varphi_{\delta}} = \sup_{x \in \mathbb{R}^n} \|\varphi'_{\delta}(x)\|_2 \stackrel{(3.8)}{=} 1$, we obtain

$$\begin{aligned} \|f_{\Phi}^{k+1}(\mathbf{G})\|_F &\leq \|f_{\Phi}^k(\mathbf{G})\|_F + \frac{\tau_k}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot L_{\varphi_{\delta}} \|\mathbf{G}\|_{\infty} \\ &= \|f_{\Phi}^k(\mathbf{G})\|_F + \frac{\tau_k}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \|\mathbf{G}\|_{\infty} \end{aligned} \quad (4.10)$$

Note that above we use the boundedness of φ'_{δ} instead of its Lipschitz continuity to avoid getting $c \|f_{\Phi}^k(\mathbf{G})\|_F$ with $c > 1$ as bound for this step, which would lead to a bound for $\|f_{\Phi}^{\ell}(\mathbf{G})\|_F$ that grows exponentially with ℓ .

By starting with $\|f_{\Phi}^0(\mathbf{G})\|_F$ and applying the above inequality ℓ -times, we get

$$\|f_{\Phi}^{\ell}(\mathbf{G})\|_F \leq \|f_{\Phi}^0(\mathbf{G})\|_F + \frac{1}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \|\mathbf{G}\|_{\infty} \sum_{k=0}^{\ell-1} \tau_k.$$

Combined with $\|f_{\Phi}^0(\mathbf{G})\|_F = \|\mathbf{f}\|_F = \sqrt{m}$ (cf. Remark 6), this shows the statement. \square

Lemma 5 *Let $\mathbf{G} = (\mathbf{g}_1 \mid \dots \mid \mathbf{g}_m) \in \mathbb{R}^{KN \times m}$ and $\Phi_1, \Phi_2 \in \mathcal{U}(N)$. If Assumption 1 holds, then*

$$\begin{aligned} &\left\| f_{\Phi_1}^{\ell+1}(\mathbf{G}) - f_{\Phi_2}^{\ell+1}(\mathbf{G}) \right\|_F \\ &\leq \left(1 + \frac{\|\mathbf{G}\|_{\infty}}{\delta} \right) \left(\frac{2\tau_{\ell}}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \|\mathbf{A}\Phi_2 - \mathbf{A}\Phi_1\|_{2 \rightarrow 2} \left\| f_{\Phi_1}^{\ell}(\mathbf{G}) \right\|_F + \left\| f_{\Phi_1}^{\ell}(\mathbf{G}) - f_{\Phi_2}^{\ell}(\mathbf{G}) \right\|_F \right). \end{aligned}$$

Proof Starting from the iteration identity (4.7) applied columnwise, we bound the difference:

$$\begin{aligned} \left\| f_{\Phi_1}^{\ell+1}(\mathbf{G}) - f_{\Phi_2}^{\ell+1}(\mathbf{G}) \right\|_F &= \left\| \text{prox}_{\tau_{\ell} \mathcal{R}} \left(f_{\Phi_1}^{\ell}(\mathbf{G}) - \frac{\tau_{\ell}}{KN} (\mathbf{A}\Phi_1)^* (T - S^{\mathbf{G}}) (\mathbf{A}\Phi_1 f_{\Phi_1}^{\ell}(\mathbf{G})) \right) \right. \\ &\quad \left. - \text{prox}_{\tau_{\ell} \mathcal{R}} \left(f_{\Phi_2}^{\ell}(\mathbf{G}) - \frac{\tau_{\ell}}{KN} (\mathbf{A}\Phi_2)^* (T - S^{\mathbf{G}}) (\mathbf{A}\Phi_2 f_{\Phi_2}^{\ell}(\mathbf{G})) \right) \right\|_F. \end{aligned}$$

Adding cross terms and using the triangle inequality, we obtain

$$\begin{aligned}
\left\| f_{\Phi_1}^{\ell+1}(\mathbf{G}) - f_{\Phi_2}^{\ell+1}(\mathbf{G}) \right\|_F &\leq \left\| \text{prox}_{\tau_\ell \mathcal{R}} \left(f_{\Phi_1}^\ell(\mathbf{G}) - \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_1)^*(T - S^{\mathbf{G}})(\mathbf{A}\Phi_1 f_{\Phi_1}^\ell(\mathbf{G})) \right) \right. \\
&\quad \left. - \text{prox}_{\tau_\ell \mathcal{R}} \left(f_{\Phi_1}^\ell(\mathbf{G}) - \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_1)^*(T - S^{\mathbf{G}})(\mathbf{A}\Phi_2 f_{\Phi_1}^\ell(\mathbf{G})) \right) \right\|_F \\
&\quad + \left\| \text{prox}_{\tau_\ell \mathcal{R}} \left(f_{\Phi_1}^\ell(\mathbf{G}) - \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_1)^*(T - S^{\mathbf{G}})(\mathbf{A}\Phi_2 f_{\Phi_1}^\ell(\mathbf{G})) \right) \right. \\
&\quad \left. - \text{prox}_{\tau_\ell \mathcal{R}} \left(f_{\Phi_1}^\ell(\mathbf{G}) - \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_2)^*(T - S^{\mathbf{G}})(\mathbf{A}\Phi_2 f_{\Phi_1}^\ell(\mathbf{G})) \right) \right\|_F \\
&\quad + \left\| \text{prox}_{\tau_\ell \mathcal{R}} \left(f_{\Phi_1}^\ell(\mathbf{G}) - \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_2)^*(T - S^{\mathbf{G}})(\mathbf{A}\Phi_2 f_{\Phi_1}^\ell(\mathbf{G})) \right) \right. \\
&\quad \left. - \text{prox}_{\tau_\ell \mathcal{R}} \left(f_{\Phi_2}^\ell(\mathbf{G}) - \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_2)^*(T - S^{\mathbf{G}})(\mathbf{A}\Phi_2 f_{\Phi_2}^\ell(\mathbf{G})) \right) \right\|_F.
\end{aligned}$$

Using the 1-Lipschitz property of the proximal operator, we get

$$\begin{aligned}
&\left\| f_{\Phi_1}^{\ell+1}(\mathbf{G}) - f_{\Phi_2}^{\ell+1}(\mathbf{G}) \right\|_F \\
&\leq \underbrace{\left\| \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_1)^* \left[(T - S^{\mathbf{G}})(\mathbf{A}\Phi_2 f_{\Phi_1}^\ell(\mathbf{G})) - (T - S^{\mathbf{G}})(\mathbf{A}\Phi_1 f_{\Phi_1}^\ell(\mathbf{G})) \right] \right\|_F}_{=:(*^1)} \\
&\quad + \underbrace{\left\| \frac{\tau_\ell}{KN} ((\mathbf{A}\Phi_2)^* - (\mathbf{A}\Phi_1)^*)(T - S^{\mathbf{G}})(\mathbf{A}\Phi_2(f_{\Phi_1}^\ell(\mathbf{G}))) \right\|_F}_{=:(*^2)} \\
&\quad + \underbrace{\left\| \left(f_{\Phi_1}^\ell(\mathbf{G}) - f_{\Phi_2}^\ell(\mathbf{G}) \right) - \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_2)^* \left[(T - S^{\mathbf{G}})(\mathbf{A}\Phi_2 f_{\Phi_1}^\ell(\mathbf{G})) - (T - S^{\mathbf{G}})(\mathbf{A}\Phi_2 f_{\Phi_2}^\ell(\mathbf{G})) \right] \right\|_F}_{=:(*^3)}.
\end{aligned}$$

The third term can be bounded as

$$\begin{aligned}
(*^3) &= \left\| \left(f_{\Phi_1}^\ell(\mathbf{G}) - f_{\Phi_2}^\ell(\mathbf{G}) \right) - \tau_\ell \left[(T_{\Phi_2} - S_{\Phi_2}^{\mathbf{G}})(f_{\Phi_1}^\ell(\mathbf{G})) - (T_{\Phi_2} - S_{\Phi_2}^{\mathbf{G}})(f_{\Phi_2}^\ell(\mathbf{G})) \right] \right\|_F \\
&\stackrel{(4.6)}{\leq} \left(1 + \frac{\|\mathbf{G}\|_\infty}{\delta} \right) \left\| (f_{\Phi_1}^\ell(\mathbf{G})) - (f_{\Phi_2}^\ell(\mathbf{G})) \right\|_F.
\end{aligned}$$

Since proximal operator are firmly nonexpansive and thus also 1-Lipschitz, we get that $T - S^{\mathbf{G}}$ is $\left(1 + \frac{\|\mathbf{G}\|_\infty}{\delta}\right)$ -Lipschitz and we can bound the first term by:

$$\begin{aligned}
(*^1) &\leq \left(1 + \frac{\|\mathbf{G}\|_\infty}{\delta} \right) \cdot \left\| \frac{\tau_\ell}{KN} (\mathbf{A}\Phi_1)^* \right\|_{2 \rightarrow 2} \cdot \left\| \mathbf{A}\Phi_2 f_{\Phi_1}^\ell(\mathbf{G}) - \mathbf{A}\Phi_1 f_{\Phi_1}^\ell(\mathbf{G}) \right\|_F \\
&\leq \left(1 + \frac{\|\mathbf{G}\|_\infty}{\delta} \right) \frac{\tau_\ell}{KN} \|\mathbf{A}\Phi_1\|_{2 \rightarrow 2} \cdot \|\mathbf{A}\Phi_2 - \mathbf{A}\Phi_1\|_{2 \rightarrow 2} \left\| f_{\Phi_1}^\ell(\mathbf{G}) \right\|_F
\end{aligned}$$

Using $(T - S^{\mathbf{G}})$ is $\left(1 + \frac{\|\mathbf{G}\|_{\infty}}{\delta}\right)$ -Lipschitz together with $T(0) = 0 = S^{\mathbf{G}}(0)$ (Remark 4), we obtain

$$\begin{aligned} (*^2) &\leq \left\| \frac{\tau_{\ell}}{KN} ((\mathbf{A}\Phi_2)^* - (\mathbf{A}\Phi_1)^*) \right\|_{2 \rightarrow 2} \cdot \left\| (T - S^{\mathbf{G}})(\mathbf{A}\Phi_2(f_{\Phi_1}^{\ell}(\mathbf{G}))) \right\|_F \\ &\leq \left(1 + \frac{\|\mathbf{G}\|_{\infty}}{\delta}\right) \left\| \frac{\tau_{\ell}}{KN} ((\mathbf{A}\Phi_2)^* - (\mathbf{A}\Phi_1)^*) \right\|_{2 \rightarrow 2} \cdot \|\mathbf{A}\Phi_2\|_{2 \rightarrow 2} \cdot \left\| f_{\Phi_1}^{\ell}(\mathbf{G}) - 0 \right\|_F \end{aligned}$$

Noting $\|\mathbf{A}\Phi_1\|_{2 \rightarrow 2} = \|\mathbf{A}\|_{2 \rightarrow 2} = \|\mathbf{A}\Phi_2\|_{2 \rightarrow 2}$, the bounds for $(*^1)$ and $(*^2)$ are the same and we get

$$\begin{aligned} (*^1) + (*^2) + (*^3) &\leq \frac{2\tau_{\ell}}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \left(1 + \frac{\|\mathbf{G}\|_{\infty}}{\delta}\right) \|\mathbf{A}\Phi_2 - \mathbf{A}\Phi_1\|_{2 \rightarrow 2} \left\| (f_{\Phi_1}^{\ell}(\mathbf{G})) \right\|_F \\ &\quad + \left(1 + \frac{\|\mathbf{G}\|_{\infty}}{\delta}\right) \left\| (f_{\Phi_1}^{\ell}(\mathbf{G})) - (f_{\Phi_2}^{\ell}(\mathbf{G})) \right\|_F. \end{aligned}$$

Factoring out $\left(1 + \frac{\|\mathbf{G}\|_{\infty}}{\delta}\right)$ shows the desired inequality. \square

Theorem 1 *Let $\mathbf{A} \in \mathbb{C}^{KN \times N}$ and $L \in \mathbb{N}_0$. Furthermore, for $\ell \in \{0, \dots, L-1\}$, let $\tau_{\ell} \in \mathbb{R}_{>0}$ satisfy Assumption 1. Define the constants*

$$\gamma := 1 + \frac{\|\mathbf{G}\|_{\infty}}{\delta} \quad \text{and} \quad \mathcal{T}_L := \sum_{k=0}^{L-1} \tau_k.$$

Then, for all $\Phi_1, \Phi_2 \in \mathcal{U}(N)$, the following perturbation bound holds:

$$\left\| f_{\Phi_1}^L(\mathbf{G}) - f_{\Phi_2}^L(\mathbf{G}) \right\|_F \leq K_L \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2}, \quad (4.11)$$

where the constant $K_L \in (0, \infty)$ is given by

$$K_L = \sum_{\ell=0}^{L-1} \gamma^{L-\ell} B_{\ell}, \text{ where } B_{\ell} := \frac{2\tau_{\ell}}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \left(\sqrt{m} + \frac{\mathcal{T}_{\ell}}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \|\mathbf{G}\|_{\infty} \right). \quad (4.12)$$

Proof Using first Lemma 5 and then Lemma 4, we get for $\ell \in \mathbb{N}_0$

$$\begin{aligned} \left\| f_{\Phi_1}^{\ell+1}(\mathbf{G}) - f_{\Phi_2}^{\ell+1}(\mathbf{G}) \right\|_F &\leq \gamma \left(\left\| f_{\Phi_1}^{\ell}(\mathbf{G}) - f_{\Phi_2}^{\ell}(\mathbf{G}) \right\|_F + \frac{2\tau_{\ell}}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \cdot \left\| f_{\Phi_1}^{\ell}(\mathbf{G}) \right\|_F \right) \\ &\leq \gamma \left(\left\| f_{\Phi_1}^{\ell}(\mathbf{G}) - f_{\Phi_2}^{\ell}(\mathbf{G}) \right\|_F + \frac{2\tau_{\ell}}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \cdot \left(\sqrt{m} + \frac{\mathcal{T}_{\ell}}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \|\mathbf{G}\|_{\infty} \right) \right) \\ &= \gamma \left\| f_{\Phi_1}^{\ell}(\mathbf{G}) - f_{\Phi_2}^{\ell}(\mathbf{G}) \right\|_F + \gamma B_{\ell} \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \end{aligned} \quad (4.13)$$

We proceed by induction on $L \in \mathbb{N}_0$. For $L = 1$, using (4.13) with $\ell = L-1 = 0$, we have

$$\begin{aligned} \left\| f_{\Phi_1}^1(\mathbf{G}) - f_{\Phi_2}^1(\mathbf{G}) \right\|_F &\leq \gamma \left\| f_{\Phi_1}^0(\mathbf{G}) - f_{\Phi_2}^0(\mathbf{G}) \right\|_F + \gamma B_0 \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \\ &= \gamma B_0 \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2}. \end{aligned}$$

Here, we have used $f_{\Phi_1}^0(\mathbf{G}) = \mathbf{f} = f_{\Phi_2}^1(\mathbf{G})$.

Since $\mathcal{T}_1 = \tau_0$ and $K_1 = \gamma B_0$, this shows the statement for $L = 1$.

Assume now that the inequality holds for some $L \in \mathbb{N}_0$, i.e.,

$$\left\| f_{\Phi_1}^L(\mathbf{G}) - f_{\Phi_2}^L(\mathbf{G}) \right\|_F \leq \sum_{\ell=0}^{L-1} \gamma^{L-\ell} B_\ell \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2}.$$

Using (4.13) with $\ell = L = 0$ and then the inductive hypothesis, we have

$$\begin{aligned} \left\| f_{\Phi_1}^{L+1}(\mathbf{G}) - f_{\Phi_2}^{L+1}(\mathbf{G}) \right\|_F &\leq \gamma \left\| f_{\Phi_1}^L(\mathbf{G}) - f_{\Phi_2}^L(\mathbf{G}) \right\|_F + \gamma B_L \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \\ &\leq \gamma \sum_{\ell=0}^{L-1} \gamma^{L-\ell} B_\ell \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} + \gamma B_L \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \\ &= \left(\sum_{\ell=0}^{L-1} \gamma^{L+1-\ell} B_\ell + \gamma B_L \right) \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \\ &= \left(\sum_{\ell=0}^L \gamma^{L+1-\ell} B_\ell \right) \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \end{aligned} \quad \square$$

5. Generalization Error Bound via Covering Numbers

In this section, we derive a generalization error bound for the hypothesis class \mathcal{H}_2^L , closely following the technique from [3]. The approach relies on bounding the covering numbers of

$$\mathcal{M}_2 := \{\sigma(\Psi f_{\Phi}^L(\mathbf{G})) \in \mathbb{C}^{KN \times m} : \Psi, \Phi \in \mathcal{U}(N)\}. \quad (5.1)$$

which represents the family of network outputs indexed by unitary parameters. We begin by quantifying how changes in the parameters affect the outputs in \mathcal{M}_2 .

The following corollary provides a Lipschitz-type inequality showing that the Frobenius norm of output differences in \mathcal{M}_2 can be bounded in terms of the distances between parameter matrices:

Corollary 4 *Let $L \geq 1$, and assume that Assumption 1 holds for all τ_ℓ with $\ell = 0, \dots, L-1$. Moreover let $\mathcal{R} \in \Gamma_0(\mathbb{C})$ with $0 \in \arg \min_{x \in \mathbb{R}^n} \mathcal{R}(x)$. Then, for any $\Psi_1, \Psi_2 \in \mathcal{U}(N)$ and $\Phi_1, \Phi_2 \in \mathcal{U}(N)$, the following estimate holds:*

$$\left\| \sigma(\Psi_1 f_{\Phi_1}^L(\mathbf{G})) - \sigma(\Psi_2 f_{\Phi_2}^L(\mathbf{G})) \right\|_F \leq M_L \|\Psi_1 - \Psi_2\|_{2 \rightarrow 2} + M'_L \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2}.$$

with

$$M_L = \sqrt{m} + \frac{\mathcal{T}_L}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \|\mathbf{G}\|_\infty, \quad M'_L = K_L \quad (5.2)$$

with K_L and \mathcal{T}_L from Theorem 1.

Proof By applying the 1-Lipschitz continuity of the function σ (see (2.5)), together with the triangle inequality, and using that $\Psi_1, \Psi_2 \in \mathcal{U}(N)$, as well as (4.9) and (4.11), we obtain

$$\begin{aligned}
& \left\| \sigma(\Psi_1 f_{\Phi_1}^L(\mathbf{G})) - \sigma(\Psi_2 f_{\Phi_2}^L(\mathbf{G})) \right\|_F \leq \left\| \Psi_1 f_{\Phi_1}^L(\mathbf{G}) - \Psi_2 f_{\Phi_2}^L(\mathbf{G}) \right\|_F \\
& \leq \left\| \Psi_1 f_{\Phi_1}^L(\mathbf{G}) - \Psi_2 f_{\Phi_1}^L(\mathbf{G}) \right\|_F + \left\| \Psi_2 f_{\Phi_1}^L(\mathbf{G}) - \Psi_2 f_{\Phi_2}^L(\mathbf{G}) \right\|_F \\
& = \left\| (\Psi_1 - \Psi_2) f_{\Phi_1}^L(\mathbf{G}) \right\|_F + \left\| \Psi_2 \left(f_{\Phi_1}^L(\mathbf{G}) - f_{\Phi_2}^L(\mathbf{G}) \right) \right\|_F \\
& \leq \|\Psi_1 - \Psi_2\|_{2 \rightarrow 2} \left\| f_{\Phi_1}^L(\mathbf{G}) \right\|_F + \left\| f_{\Phi_1}^L(\mathbf{G}) - f_{\Phi_2}^L(\mathbf{G}) \right\|_F \\
& \leq \left(\sqrt{m} + \frac{\mathcal{T}_L}{KN} \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \|\mathbf{G}\|_\infty \right) \|\Psi_1 - \Psi_2\|_{2 \rightarrow 2} + K_L \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2}. \quad \square
\end{aligned}$$

Using the inequality from Corollary 4, we can define a metric on the parameter space $\mathcal{U}(N) \times \mathcal{U}(N)$ by appropriately scaling the operator norms. Let

$$d_1(\Psi_1, \Psi_2) := M_L \|\Psi_1 - \Psi_2\|_{2 \rightarrow 2}, \quad d_2(\Phi_1, \Phi_2) := M'_L \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2}.$$

Since $\|\cdot\|_{2 \rightarrow 2}$ is a norm and $M_L, M'_L \geq 0$, the function

$$d((\Phi_1, \Psi_1), (\Phi_2, \Psi_2)) := d_1(\Psi_1, \Psi_2) + d_2(\Phi_1, \Phi_2)$$

defines a metric on $\mathcal{U}(N) \times \mathcal{U}(N)$. Therefore, $(\mathcal{U}(N) \times \mathcal{U}(N), d)$ is a metric space. Moreover, the Frobenius norm between any two elements of \mathcal{M}_2 can be bounded above by this metric:

$$\begin{aligned}
\|m_1 - m_2\|_F &= \left\| \sigma(\Psi_1 f_{\Phi_1}^L(\mathbf{G})) - \sigma(\Psi_2 f_{\Phi_2}^L(\mathbf{G})) \right\|_F \\
&\leq M_L \|\Psi_1 - \Psi_2\|_{2 \rightarrow 2} + M'_L \|\mathbf{A}\Phi_1 - \mathbf{A}\Phi_2\|_{2 \rightarrow 2} \\
&= d_1(\Psi_1, \Psi_2) + d_2(\Phi_1, \Phi_2) = d((\Phi_1, \Psi_1), (\Phi_2, \Psi_2))
\end{aligned} \tag{5.3}$$

In this sense, the distance on \mathcal{M}_2 can be controlled by the metric d on $\mathcal{U}(N) \times \mathcal{U}(N)$.

From this, we immediately obtain that the covering numbers of \mathcal{M}_2 under the Frobenius norm can be upper bounded by those of its parameter space:

$$\mathcal{N}(\mathcal{M}_2, \|\cdot\|_F, \varepsilon) \leq \mathcal{N}(\mathcal{U}(N) \times \mathcal{U}(N), d, \varepsilon). \tag{5.4}$$

To see this, let $n := \mathcal{N}(\mathcal{U}(N) \times \mathcal{U}(N), d, \varepsilon)$. Then, there exists $(\Psi_i, \Phi_i)_{i=1}^n$ such that

$$\mathcal{U}(N) \times \mathcal{U}(N) \subset \bigcup_{i=1}^n \mathcal{B}_\varepsilon^d((\Psi_i, \Phi_i)).$$

Let $m \in \mathcal{M}_2$. Then, there exists $(\Psi, \Phi) \in \mathcal{U}(N) \times \mathcal{U}(N)$ such that $m = \sigma(\Psi f_\Phi^L(\mathbf{G}))$. Furthermore, there exists a $j \in \{1, \dots, n\}$ with

$$d((\Psi, \Phi_j), (\Phi, \Phi_j)) \leq \varepsilon.$$

Now, using (5.3), we obtain

$$\|m - m_j\|_F = \|\sigma(\Psi f_{\Phi}^L(\mathbf{G})) - \sigma(\Psi_j f_{\Phi_j}^L(\mathbf{G}))\| \leq d((\Psi, \Psi_j), (\Phi, \Phi_j)) \leq \epsilon.$$

Thus, $m \in \bigcup_{i=1}^n \mathcal{B}_{\epsilon}^{\|\cdot\|_F}(m_i)$ where $m_i = \sigma(\Psi_i f_{\Phi_i}^L(\mathbf{G}))$. Since $m \in M_2$, was arbitrary, we have $M \subset \bigcup_{i=1}^n \mathcal{B}_{\epsilon}^{\|\cdot\|_F}(m_i)$ and consequently

$$\mathcal{N}\left(\mathcal{M}_2, \|\cdot\|_F, \epsilon\right) \leq n = \mathcal{N}\left(\mathcal{U}(N) \times \mathcal{U}(N), d, \epsilon\right).$$

We now derive an explicit upper bound on the covering numbers of \mathcal{M}_2 based on the metric properties established above.

Lemma 6 (Covering Number Bound for \mathcal{M}_2) *Let the assumptions of Corollary 4 be fulfilled. Then,*

$$\log(\mathcal{N}(\mathcal{M}_2, \|\cdot\|_{2 \rightarrow 2}, \epsilon)) \leq 2N^2 \log\left(1 + \frac{4M_L}{\epsilon}\right) + 2KN^2 \log\left(1 + \frac{4M'_L \|\mathbf{A}\|_{2 \rightarrow 2}}{\epsilon}\right).$$

Proof We argue as follows. First estimate the covering number of the set

$$\mathcal{W} = \{\mathbf{A}\Phi : \Phi \in \mathcal{U}(N)\} = \left\{ \|\mathbf{A}\|_{2 \rightarrow 2} \cdot \frac{\mathbf{A}\Phi}{\|\mathbf{A}\|_{2 \rightarrow 2}} : \Phi \in \mathcal{U}(N) \right\} \subset \mathbb{C}^{KN \times N}.$$

To control $\mathcal{N}(\mathcal{W}, \|\cdot\|_{2 \rightarrow 2}, \xi)$ for $\xi > 0$, we extend [3, Lemma 6] to the complex case. Using the identity above for \mathcal{W} , we get

$$\mathcal{N}(\mathcal{W}, \|\cdot\|_{2 \rightarrow 2}, \xi) = \mathcal{N}\left(\left\{ \frac{\mathbf{A}\Phi}{\|\mathbf{A}\|_{2 \rightarrow 2}} : \Phi \in \mathcal{U}(N) \right\}, \|\cdot\|_{2 \rightarrow 2}, \frac{\xi}{\|\mathbf{A}\|_{2 \rightarrow 2}}\right) \leq \left(1 + \frac{2\|\mathbf{A}\|_{2 \rightarrow 2}}{\xi}\right)^{2KN^2}. \quad (5.5)$$

The last inequality follows from $\mathcal{U}(N) \subset B_{\|\cdot\|_{2 \rightarrow 2}}^{N \times N} := \{B \in \mathbb{C}^{N \times N} : \|B\|_{2 \rightarrow 2} \leq 1\}$ and [9, Proposition C.3]. Since the latter is formulated for the real case, we identify $\mathbb{C}^{KN \times N}$ with $\mathbb{R}^{2 \times KN \times N}$ to apply this proposition. Note that this doubles the dimension of the space, hence doubling the exponent in the covering bound.

Next, using (5.4) together with the product-space covering bound, we get

$$\begin{aligned} \mathcal{N}(\mathcal{M}_2, \|\cdot\|_F, \epsilon) &\leq \mathcal{N}(\mathcal{U}(N) \times \mathcal{U}(N), d, \epsilon) = \mathcal{N}(\mathcal{U}(N) \times \mathcal{U}(N), d_1 + d_2, \epsilon) \\ &\leq \mathcal{N}(\mathcal{U}(N), d_1, \frac{\epsilon}{2}) \times \mathcal{N}(\mathcal{U}(N), d_2, \frac{\epsilon}{2}) \\ &= \mathcal{N}(\mathcal{U}(N), M_L \|\cdot\|_{2 \rightarrow 2}, \frac{\epsilon}{2}) \times \mathcal{N}(\mathcal{W}, M'_L \|\cdot\|_{2 \rightarrow 2}, \frac{\epsilon}{2}) \\ &\leq \mathcal{N}(\mathcal{U}(N), \|\cdot\|_{2 \rightarrow 2}, \frac{\epsilon}{2M_L}) \times \mathcal{N}(\mathcal{W}, \|\cdot\|_{2 \rightarrow 2}, \frac{\epsilon}{2M'_L}). \end{aligned}$$

Since $\mathcal{U}(N) \subset \mathcal{B}_{\|\cdot\|_{2 \rightarrow 2}}^{N \times N}$, the covering number bound from [9, Proposition C.3] (doubling the dimension, since we apply it for complex spaces) gives for $\xi > 0$

$$\mathcal{N}(\mathcal{U}(N), \|\cdot\|_{2 \rightarrow 2}, \xi) \leq \left(1 + \frac{2}{\xi}\right)^{2N^2}$$

Substituting $\xi = \frac{\varepsilon}{2M_L}$ above and $\xi = \frac{\varepsilon}{2M'_L}$ in (5.5), we obtain

$$\mathcal{N}(\mathcal{M}_2, \|\cdot\|_F, \varepsilon) \leq \left(1 + \frac{4M_L}{\varepsilon}\right)^{2N^2} \times \left(1 + \frac{4M'_L \|\mathbf{A}\|_{2 \rightarrow 2}}{\varepsilon}\right)^{2KN^2}.$$

Taking logarithms gives the claimed result. \square

Using the covering number bound from Lemma 6, we now present the main generalization bound for the hypothesis class \mathcal{H}_2^L , based on Rademacher complexity estimates. Here, the *generalization error* is defined as the difference between the true risk of a hypothesis $h \in \mathcal{H}_2^L$ on unseen data and its empirical risk measured on the training set. This quantity measures how much worse a hypothesis $h \in \mathcal{H}_2^L$ may perform on unseen data compared to its training performance (see Subsection 2.1 of [3]).

Theorem 2 *Consider the hypothesis space \mathcal{H}_2^L in (2.6). Let $\alpha \in (0, 1)$. With a probability at least $1 - \alpha$, for all $h \in \mathcal{H}_2^L$, the generalization error is bounded by*

$$\mathcal{L}(h) - \hat{\mathcal{L}}(h) \leq 2\mathcal{R}_m(l \circ \mathcal{H}_2^L) + 4(C_{in} + C_{out}) \sqrt{\frac{2\log(4/\alpha)}{m}},$$

where C_{in} and C_{out} are the constants defined in Section 2.5. The Rademacher complexity term is further bounded by

$$\mathcal{R}_m(l \circ \mathcal{H}_2^L) \leq 4C_{out} \frac{N}{\sqrt{m}} \left(\sqrt{\log \left(e \left(1 + \frac{8M_L}{\sqrt{m}C_{out}} \right) \right)} + \sqrt{\log \left(e \left(1 + \frac{8\|\mathbf{A}M'_L\|_{2 \rightarrow 2}}{\sqrt{m}C_{out}} \right) \right)} \right),$$

where M_L and M'_L are as in (5.2).

Proof Noting that the loss in our case is bounded by $c = C_{in} + C_{out}$, the claimed upper bound for $\mathcal{L}(h) - \hat{\mathcal{L}}(h)$ follows from [16, Theorem 26.5]. The bound for the Rademacher complexity is based on Dudley's inequality [9, Theorem 8.23]. Following the derivation of the bound derived

in [3, Subsection 4.3, (33)], one gets

$$\mathcal{R}_m(l \circ \mathcal{H}_2^L) \leq \frac{4\sqrt{2}}{m} \int_0^{\sqrt{m}^{C_{\text{out}}/2}} \sqrt{\log(\mathcal{N}(\mathcal{M}_2, \|\cdot\|_{2 \rightarrow 2}, \varepsilon))} d\varepsilon.$$

Using this, together with Lemma 6 and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$, we get

$$\begin{aligned} \mathcal{R}_m(l \circ \mathcal{H}_2^L) &\leq \frac{4\sqrt{2}}{m} \int_0^{\sqrt{m}^{C_{\text{out}}/2}} \sqrt{\log(\mathcal{N}(\mathcal{M}_2, \|\cdot\|_{2 \rightarrow 2}, \varepsilon))} d\varepsilon \\ &\leq \frac{4\sqrt{2}}{m} \int_0^{\sqrt{m}^{C_{\text{out}}/2}} \sqrt{2N^2 \log\left(1 + \frac{4M_L}{\varepsilon}\right)} d\varepsilon \\ &\quad + \frac{4\sqrt{2}}{m} \int_0^{\sqrt{m}^{C_{\text{out}}/2}} \sqrt{2KN^2 \log\left(1 + \frac{4M'_L \|\mathbf{A}\|_{2 \rightarrow 2}}{\varepsilon}\right)} d\varepsilon \\ &\leq 4C_{\text{out}} \frac{N}{\sqrt{m}} \sqrt{\log\left(e\left(1 + \frac{4M_L}{\sqrt{m}^{C_{\text{out}}/2}}\right)\right)} \\ &\quad + 4C_{\text{out}} \sqrt{K} \frac{N}{\sqrt{m}} \sqrt{\log\left(e\left(1 + \frac{4M'_L \|\mathbf{A}\|_{2 \rightarrow 2}}{\sqrt{m}^{C_{\text{out}}/2}}\right)\right)}, \end{aligned}$$

where the last inequality holds due to [3, (47)] (which slightly generalizes [9, Lemma C.9]) with $\alpha = \sqrt{m}^{C_{\text{out}}/2}$ and $\beta = 4M_L$ or $\beta = 4M'_L \|\mathbf{A}\|_{2 \rightarrow 2}$. \square

We conclude this section with a discussion of how the generalization bound scales with the network depth L , the parameter dimension N^2 , the number of measurement operators K , and the number of training samples m .

Remark 8 *M_L grows (at most) linearly in L , while M'_L scales exponentially in L . This can be seen as follows. Noting*

$$\mathcal{T}_L = \sum_{k=0}^{L-1} \tau_k \leq L \max_{k \in [0, L-1]} \tau_k,$$

\mathcal{T}_L and thus M_L grows (at most) linearly in L . We recall

$$K_L = \sum_{\ell=0}^{L-1} \gamma^{L-\ell} B_\ell \quad \text{with} \quad \gamma = 1 + \frac{\|\mathbf{G}\|_\infty}{\delta} > 1,$$

and note that B_ℓ depends linearly on \mathcal{T}_ℓ . With the above, we get $B_\ell = \mathcal{O}(\ell)$. Factoring out γ^L in the expression for K_L above gives

$$K_L = \gamma^L \sum_{\ell=0}^{L-1} \gamma^{-\ell} B_\ell = \gamma^L \sum_{\ell=0}^{L-1} \mathcal{O}\left(\frac{\ell}{\gamma^\ell}\right).$$

Since ℓ/γ^ℓ decays, the inner sum is bounded. Therefore, $K_L = \mathcal{O}(\gamma^L)$, showing that $K_L = M'_L$ grows exponentially in L .

By isolating the dependence on K , N , m and L , and treating other terms as constants, the generalization error is bounded as

$$\mathcal{L}(h) - \hat{\mathcal{L}}(h) \lesssim \frac{N}{\sqrt{m}} \sqrt{\log(L)} + \frac{N\sqrt{K}}{\sqrt{m}} \sqrt{L} \lesssim \frac{N(1+\sqrt{K})}{\sqrt{m}} \sqrt{L}.$$

This reveals an overall square-root dependency on the depth L (since the logarithmic terms are dominated by \sqrt{L}). Moreover, the dependency on the number of trainable parameters can be made explicit: since the dimension of the parameter space is $\dim \mathcal{U}(N) = N^2 = \mathcal{O}(N^2)$, the bound can equivalently be expressed in terms of the ratio of trainable parameters to samples as

$$\mathcal{L}(h) - \hat{\mathcal{L}}(h) \lesssim \sqrt{\frac{P}{m}} (1 + \sqrt{K}) \sqrt{L}, \quad P \sim N^2.$$

Hence, the error decreases at the standard rate $1/\sqrt{m}$ with the number of samples, while growing with the square root of the ratio P/m , the depth L , and the measurement count K . This shows that any factor increase in N^2 must be compensated by the same factor increase in m to preserve generalization performance. Hence, the number of samples must scale proportionally to the number of trainable parameters to preserve generalization performance.

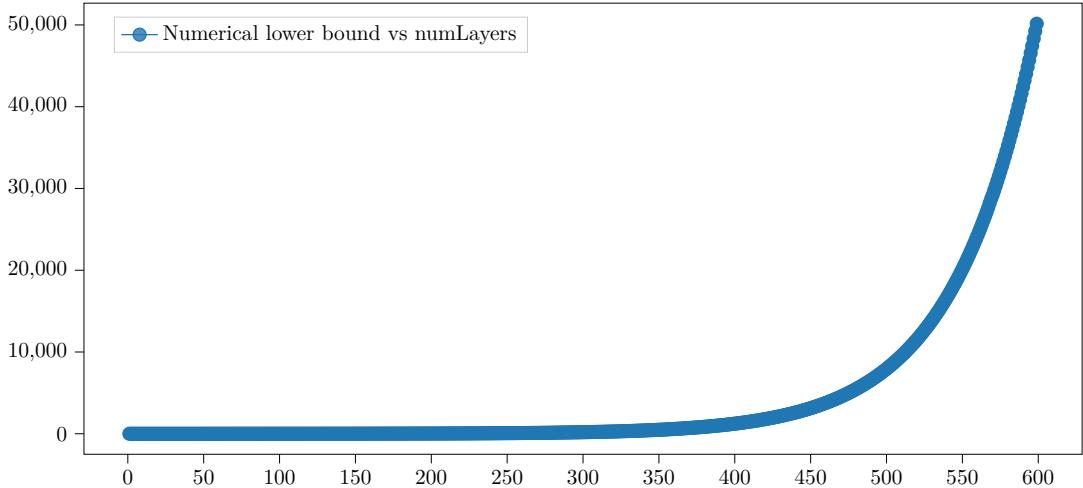


Figure 1. Lower bound for K_L computed by numerically maximizing the ratio $\frac{\|f_{\Phi_1}^L(\mathbf{G}) - f_{\Phi_2}^L(\mathbf{G})\|_F}{\|\Phi_1 - \Phi_2\|_{2 \rightarrow 2}}$ with respect to Φ_1 and Φ_2 for the case $N = 2$, $K = 1$. Here, the dictionary matrices were restricted to be rotations and the optimization was done over the rotation angles.

As we have seen, the resulting generalization bound is linear in the number of layers L (unlike the bound of [3], which is logarithmic in L). This deficiency comes from the perturbation bound, where K_L grows exponentially in L with our proof, not quadratically like in [3]. Thus, the question arises whether this is an artifact of our proof or a inevitable consequence of the nonlinear forward model. Figure 1 provides numerical evidence supporting the scaling with L .

found in Theorem 1. Without regularization ($\mathcal{R} = 0$), the numerically computed lower bound for the Lipschitz constant for network perturbations appears to grow exponentially with the network depth L , in agreement with the bound $K_L = \mathcal{O}(\gamma^L)$, where $\gamma = 1 + \|G\|_\infty/\delta$.

Using $\mathcal{R} = \lambda \|\cdot\|_1$ and strong regularization (i.e., λ large), it may be possible to get a logarithmic bound. The problem is that the forward step of the PGA is expanding in the vicinity of the origin. This is due to φ'_δ having a large local Lipschitz constant $\sim 1/\delta$ only near the origin. Since the backward step of the PGA with $\mathcal{R} = \lambda \|\cdot\|_1$ is soft thresholding, this could possibly be used to compensate for the behavior of the forward step near the origin.

Finally, we would like to point out that the generalization bound from [3] for LISTA also follows from our approach. By choosing $\varphi(z) = z$, $\mathcal{R} = \lambda \|\cdot\|_1$ and restricting to the real case, (2.2) is the objective from ISTA. In this case, one gets $\varphi'(z) = 1$ and $\varphi\varphi' = \text{Id}$, where Id denotes the identity mapping. Moreover, $\varphi\varphi' = \text{Id} = \text{prox}_0$ (the proximal operator of the constant zero mapping) and $\varphi' = \text{prox}_{I_{\{1\}}}$ (the proximal operator of indicator function $I_{\{1\}}(z) = 0$ for $z = 1$ and ∞ else.). Since the Lipschitz constant of $\varphi' \equiv 1$ is 0, one gets $\gamma = 1$ in Theorem 1 which restores the quadratic growth of the perturbation bound in L and thus the logarithmic scaling in L of the generalization bound.

Acknowledgments

Moussa Atwi and Benjamin Berkels were supported by the German research foundation (DFG) within the Collaborative Research Centre SFB 1481 “Sparsity and Singular Structures” (Project ID 442047500, Project A08).

REFERENCES

- Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2017.
- Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, October 2017.
- Arash Behboodi, Holger Rauhut, and Ekkehard Schnoor. *Compressive Sensing and Neural Networks from a Statistical Learning Perspective*, pages 247–277. Springer International Publishing, 2022.
- Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, April 2015.
- W.M.J. Coene, A. Thust, M. Op de Beeck, and D. Van Dyck. Maximum-likelihood method for focus-variation image reconstruction in high resolution transmission electron microscopy. *Ultramicroscopy*, 64:109–135, 1996.
- Patrick L. Combettes. Monotone operator theory in convex optimization. *Mathematical Programming*, 170(1):177–206, June 2018.
- Ingrid Daubechies, Michel Defrise, and Christine de Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, August 2004.
- Christian Doberstein and Benjamin Berkels. A least-squares functional for joint exit wave reconstruction and image registration. *Inverse Problems*, 35(5), 2019.
- Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.
- Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on*

- Machine Learning (ICML-10)*, pages 399–406, Haifa, Israel, June 2010. Omnipress.
- 11. John R. Hershey, Jonathan Le Roux, and Felix Weninger. Deep unfolding: Model-based inspiration of novel deep architectures, September 2014.
 - 12. Wen-Kuo Hsieh, Fu-Rong Chen, Ji-Jung Kai, and A.I Kirkland. Resolution extension and exit wave reconstruction in complex hrem. *Ultramicroscopy*, 98(2-4):99–114, January 2004.
 - 13. Yuelong Li, Mohammad Tofghi, Vishal Monga, and Yonina C. Eldar. An algorithm unrolling approach to deep image deblurring. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7675–7679. IEEE, May 2019.
 - 14. Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. Deep griffin-lim iteration: Trainable iterative phase reconstruction using neural network. *IEEE Journal of Selected Topics in Signal Processing*, 15(1):37–50, January 2021.
 - 15. Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, March 2021.
 - 16. Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, May 2014.
 - 17. Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, 2nd edition, 2021.
 - 18. Zhuolei Xiao, Ya Wang, and Guan Gui. Smoothed amplitude flow-based phase retrieval algorithm. *Journal of the Franklin Institute*, 358(14):7270–7285, September 2021.
 - 19. Xiaohan Zhang, Shaowen Chen, Shuya Wang, Ying Huang, Chuanhong Jin, and Fang Lin. Exit wave reconstruction of a focal series of images with structural changes in high-resolution transmission electron microscopy. *Journal of Microscopy*, 296(1):24–33, May 2024.