Finding Generalization Measures by Contrasting Signal and Noise

Jiaye Teng [*1] **Bohang Zhang** [*2] **Ruichen Li** [*2] **Haowei He** [*1] **Yequan Wang** [3] **Yan Tian** [3] **Yang Yuan** [1 4 5]

Abstract

Generalization is one of the most fundamental challenges in deep learning, aiming to predict model performances on unseen data. Empirically, such predictions usually rely on a validation set, while recent works showed that an unlabeled validation set also works. Without validation sets, it is extremely difficult to obtain non-vacuous generalization bounds, which leads to a weaker task of finding generalization measures that monotonically relate to generalization error. In this paper, we propose a new generalization measure REF Complexity (REIative Fitting degree between signal and noise), motivated by the intuition that a given model-algorithm pair may generalize well if it fits signal (*e.g.*, true labels) fast while fitting noise (*e.g.*, random labels) slowly. Empirically, REF Complexity monotonically relates to test accuracy in real-world datasets without accessing additional validation sets, achieving -0.988 correlation on CIFAR-10 and -0.960 correlation on CIFAR-100. We further theoretically verify the utility of REF Complexity under three different cases, including convex and smooth regimes with stochastic gradient descent, smooth regimes (not necessarily convex) with stochastic gradient Langevin dynamics, and linear regimes with gradient descent. The code is available at `https://github.com/962086838/REF-complexity`.

1. Introduction

Generalization is one of the most fundamental mysteries in deep learning, measuring how the trained model performs on unseen data. By convention, people empirically estimate generalization error via validation data that are independently drawn from the population distribution. However, such validation data are obtained by splitting a portion of training data, causing a shrink in the training set. Recently, a line of work argues that labeled validation sets are unnecessary in predicting generalization, and proposes to predict generalization via an unlabeled validation set, *e.g.*, RATT approach (Garg et al., 2021), disagreement-based approaches (Jiang et al., 2022). However, the additional dataset, even unlabeled, might be expensive. This naturally leads to a question: can we estimate generalization error without any additional dataset?

Directly answering the question can be extremely challenging (Jiang et al., 2020a). As a surrogate, people consider a weaker task of finding *generalization measures* that monotonically relate to generalization error (Jiang et al., 2020b; Dziugaite et al., 2020). Unlike the predicting task that calculates the *exact* value of generalization error, generalization measures are only required to sketch its *trend*. Such relaxation is meaningful in many scenarios, *e.g.*, model selection tasks (Zucchini, 2000; Johnson & Omland, 2004; Emmert-Streib & Dehmer, 2019).

There are various types of generalization measures in the existing literature, which can be roughly split into four branches (Jiang et al., 2020b): (a) empirical measures, (b) norm-based measures, (c) PAC-Bayesian and information-based measures, (d) stability-based measures. However, (a) may imply a spurious causal relationship between the measure and generalization (Dziugaite & Roy, 2017), (b) even negatively correlate with generalization error (Jiang et al., 2020b), (c) only applies in stochastic models instead of standard training scenarios. Therefore, (d) stands out due to its algorithm-dependent property and is considered a potential approach to generalization measure analysis (Nagarajan & Kolter, 2019; Jiang et al., 2020b). Existing works have proposed meaningful generalization measures based on algorithmic stability. For example, Hardt et al. (2016) theoretically study algorithmic stability and argue that "train faster, generalize better", and Jiang et al. (2020b) observe that the initial phase of optimization benefits the final generalization. Although these arguments perform well empirically, there still exist phenomena that the existing stability-based measures cannot explain. For example, stochastic gradi-

[*]Equal contribution [1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China [2]Peking University [3]Beijing Academy of Artificial Intelligence, Beijing, China [4]Shanghai Artificial Intelligence Laboratory [5]Shanghai Qi Zhi Institute. Correspondence to: Yang Yuan <yuanyang@tsinghua.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

ent descent (SGD) usually generalizes better while trained slower (with more iterations) than gradient descent (GD).

In this paper, we propose a new measure following stability-based approaches, which (a) has theoretical backbones, (b) empirically works, and (c) is applicable in standard training scenarios, named REF Complexity (RElative Fitting degree on signal and noise). The complexity is motivated by the intuition that *a given model-algorithm pair may generalize better if it fits signal faster while fitting noise slower during the training process*. Empirically, one can treat the real-world dataset as the signal and the same dataset with random labels as the noise. Given the training set $\mathcal{D}$ and training algorithm $\mathcal{A}$, REF Complexity is informally derived as

$$\mathcal{T}_n(\mathcal{D}, \mathcal{A}) = \frac{\text{The degree of fitting noise}}{\text{The degree of fitting signal}}, \tag{1}$$

where n denotes the sample size. Intuitively, REF Complexity measures the degree to which a model-algorithm pair can distinguish between signal and noise during training, and $\mathcal{T}_n(\mathcal{D}, \mathcal{A})$ is anticipated to monotonically increase with respect to generalization error since fitting noise usually hurts generalization. Besides the property (a, b, c) above, REF Complexity (d) does not require an additional dataset, and (e) increases with the noise scale. Property (e) meets the requirement that the generalization bound (and its corresponding measure) should increase with the degree of noisy labels, proposed in Nagarajan & Kolter (2019).

From the experimental perspective, REF Complexity monotonically correlates with the generalization error (See Figure 1), demonstrated by experiments on CIFAR-10 and CIFAR-100. We further show that REF Complexity explains several phenomena in deep learning. We take the comparison between stochastic algorithms (*e.g.*, SGD) and deterministic algorithms (*e.g.*, GD) as an example. SGD usually fits signal and noise both slower than GD. However, we observe that SGD is trained significantly slower when fitting noise compared to signal, leading to a smaller REF Complexity. Therefore, SGD generalizes better than GD under REF Complexity frameworks, which accords with reality. As a comparison, existing measures including stability-based measures cannot explain the phenomenon.

From the theoretical perspective, we validate the utility of REF Complexity by deriving that generalization error can be bounded using REF Complexity under several different cases, including convex and smooth regimes with SGD, and smooth regimes (not necessarily convex) with Stochastic Gradient Langevin Dynamics (SGLD). The derivation is inspired by the stability-based techniques in generalization analysis. Informally, the degree of fitting noise ensures that the training gradient cannot be extremely large, leading to a guarantee for algorithmic stability. Similar conclusions hold beyond SGD and SGLD, and we also derive a similar bound under the regime of GD with overparameterized linear regression, following the benign overfitting techniques proposed in Bartlett et al. (2020).

We list our contributions as follows:

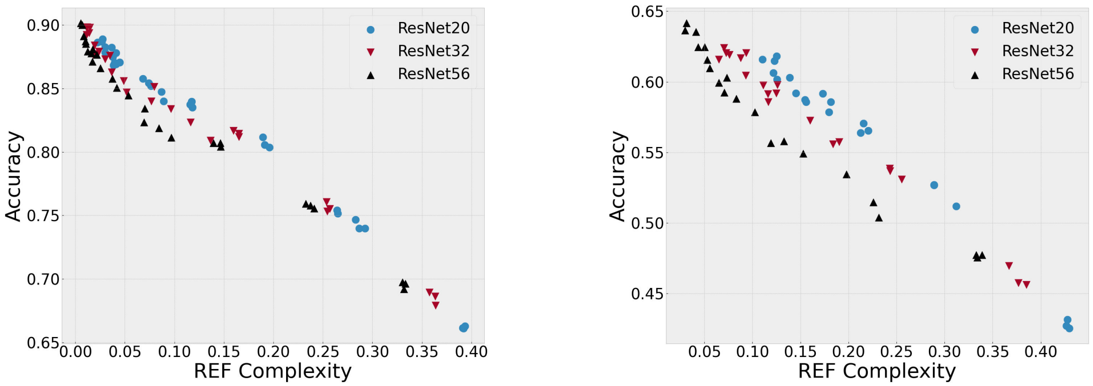
1. We propose a new generalization measure named REF Complexity, which quantifies how well a given model-algorithm pair distinguishes between signal and noise during training. REF Complexity extends the scope of stability-based measures.
2. Experimental results on CIFAR-10 and CIFAR-100 demonstrate the effectiveness of REF Complexity, where REF Complexity monotonically decreases with respect to test accuracy with correlations of -0.988 and -0.960 on CIFAR-10 and CIFAR-100, respectively.
3. We further theoretically validate the utility of REF Complexity under several different regimes, including convex and smooth loss with SGD, smooth loss (not necessarily convex) with SGLD, and linear regimes with GD.

2. Related Work

Algorithmic Stability is one of the most popular techniques in generalization analysis (Bousquet & Elisseeff, 2002; Hardt et al., 2016). A line of works derives high probability bounds based on algorithmic stability (Feldman & Vondrák, 2019; Bousquet et al., 2020). Another line of works derives algorithmic stability under various regimes, e.g., unbounded gradient (Lei & Ying, 2020), non-smooth loss (Bassily et al., 2020), stochastic gradient Langevin dynamics (Mou et al., 2018; Li et al., 2020). One of the properties of algorithmic stability is that the corresponding bound usually increases with time, motivating the optimization-based measures which quantify the number of iterations to reach a given loss threshold (Jiang et al., 2020b).

Theoretical generalization measures. Besides stability-based measures, there are many other theory-motivated measures. A line of work focuses on the norm-based measures (Neyshabur et al., 2015; Bartlett et al., 2017; Neyshabur et al., 2018; Wei & Ma, 2020), but it may dramatically fail to show monotonically correlation with test errors (Nagarajan & Kolter, 2019; Jiang et al., 2020b). Another line of work focus on PAC-Bayesian (McAllester, 1999; Dziugaite & Roy, 2017; Neyshabur et al., 2017) and information-based analysis (Russo & Zou, 2016; Xu & Raginsky, 2017; Haghifam et al., 2020; 2021). This line of work performs well numerically but requires changing the training scheme with stochastic models (Jiang et al., 2020b).

Predicting generalization errors. Compared to generalization measure approaches, predicting the exact generaliza-



(a) CIFAR-10

(b) CIFAR-100

Figure 1. Correlation between REF Complexity and test accuracy. We conduct over one hundred experiments with ResNet20, ResNet32, and RseNet56 on CIFAR-10 and CIFAR-100, showing that REF Complexity negatively relates to test accuracy with correlations of -0.988 and -0.960 on CIFAR-10 and CIFAR-100, respectively. We defer the experiment details to Section 5.

tion error is a more difficult task. Traditional approaches split a holdout partition (namely, validation set) from the available labeled data, where performances on the validation set directly imply generalization error. However, this approach restricts the number of labeled data in the training process. Recently, Garg et al. (2021) leveraged an unlabeled dataset (with random labels) to augment the labeled dataset and predict generalization via the different performances on the two datasets. Besides, a line of work (Jiang et al., 2022) focuses on the relationship between disagreement and generalization, where the disagreement comes from the different model performances (*e.g.*, trained with different training schemes) on unlabeled data. Despite not requiring additional labeled datasets (validation set), these approaches still need additional unlabeled datasets.

Empirical generalization measures. Besides those measures motivated by theoretical analysis, there are also empirical approaches to finding generalization measures or predicting generalization errors, including sharpness-based techniques (Keskar et al., 2017), robustness on representations (Natekar & Sharma, 2020) and robustness on augmentation (Aithal et al., 2021).

Distinguishing signal and noise. The structure of the response is one of the basic data properties in generalization analysis. For example, Nagarajan & Kolter (2019) argues that the generalization bound should increase with the noise levels (*e.g.*, the portion of random labels). However, some generalization measures do not even distinguish signal and noise (*e.g.*, Rademacher complexity (Shalev-Shwartz & Ben-David, 2014)), and therefore only return vacuous generalization bound when the model can fit arbitrary random noise (Zhang et al., 2021). A line of work implicitly considers different performances of signal and noise, *e.g.*, algorithmic stability can extract the output structure since neural networks usually fit signal faster than fitting noise (Zhang et al., 2021), and NTK-based data-dependent measure grows with the potion of noise (Arora et al., 2019). Besides, another line of work focuses on bounding the noise tolerance (Rudin, 2005; Manwani & Sastry, 2013; Frénay & Verleysen, 2014; Bansal et al., 2021), which analyzes the training accuracy decrease when adding a portion of label noise. This differs from our approach, where we aim at bounding generalization using noise tolerance. Of particular relevance here is Teng et al. (2022), which explicitly split the effects of signal and noise during the generalization analysis. Both bounds are inspired by the intuition that neural networks may learn simple patterns (*e.g.*, signal) faster than complex patterns (*e.g.*, noise) (Arpit et al., 2017; Rolnick et al., 2017; Rahaman et al., 2019). However, the bound in Teng et al. (2022) cannot directly lead to a simple generalization measure.

3. Preliminaries

This section introduces basic notations and necessary assumptions. Some of the notations differ from the existing literature because besides the original data distribution, we also consider two parallel types of distributions: signal distribution and noise distribution. We subscript them by `sig` and `noi`, respectively.

3.1. Basic Notations

Data Distribution. Let $(\boldsymbol{x}, y) \sim \mathcal{P} \subset \mathbb{R}^d \times \mathbb{R}$ denote the input and the corresponding response. We consider the ground truth function $y = f(\boldsymbol{x}; \boldsymbol{\theta}^*) + \epsilon$ where $\epsilon \in \mathbb{R}$ denotes the random noise, $\boldsymbol{\theta}^* \in \mathbb{R}^p$ denotes the best parameter, and $f(\cdot; \boldsymbol{\theta}^*)$ denotes a function f indexed by parameter $\boldsymbol{\theta}^*$. In such regimes, we assume that $\mathbb{E}[\epsilon|\boldsymbol{x}] = 0$. Without loss of generality, assume that $f(\boldsymbol{x}; \mathbf{0}) \equiv 0$. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$ denote the dataset with n data points sampled from distribution $\mathcal{P}$, where we omit the dependency of n for simplicity. The corresponding signal dataset and noise dataset are denoted by $\mathcal{D}_{\text{sig}} = \{(\boldsymbol{x}_i, f(\boldsymbol{x}_i; \boldsymbol{\theta}^*))\}_{i\in[n]}$ and $\mathcal{D}_{\text{noi}} = \{(\boldsymbol{x}_i, \epsilon_i)\}_{i\in[n]}$ with distribution $\mathcal{P}_{\text{sig}}$ and $\mathcal{P}_{\text{noi}}$.

Loss. Let $\ell(\boldsymbol{\theta}; \boldsymbol{z})$ denote the loss function with parameter $\boldsymbol{\theta}$ on sample $\boldsymbol{z} = (\boldsymbol{x}, y)$, given the prediction $f(\boldsymbol{x}; \boldsymbol{\theta})$. The training loss is then denoted by $\mathcal{L}_n(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{n}\sum_{\boldsymbol{z}_i\in\mathcal{D}} \ell(\boldsymbol{\theta}; \boldsymbol{z}_i)$. The corresponding excess risk is then denoted by $\mathcal{E}(\boldsymbol{\theta}; \mathcal{P}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}}\ell(\boldsymbol{\theta}; \boldsymbol{z}) - \ell(\boldsymbol{\theta}^*; \boldsymbol{z})$, measuring the distance between $\boldsymbol{\theta}$ and the best parameter $\boldsymbol{\theta}^*$. We assume that the excess risk is well-behaved, namely, $\mathcal{E}(\boldsymbol{\theta}^*; \mathcal{P}) \le \mathcal{E}(\boldsymbol{\theta}; \mathcal{P})$, $\mathcal{E}(\boldsymbol{\theta}^*; \mathcal{P}_{\text{sig}}) \le \mathcal{E}(\boldsymbol{\theta}; \mathcal{P}_{\text{sig}})$, and $\mathcal{E}(\mathbf{0}; \mathcal{P}_{\text{noi}}) \le \mathcal{E}(\boldsymbol{\theta}; \mathcal{P}_{\text{noi}})$ for all $\boldsymbol{\theta}$.

Algorithm. Let $\mathcal{A}_t$ denote the algorithm which takes a dataset $\mathcal{D}$ as an input and returns a parameter $\boldsymbol{\theta}^{(t)} = \mathcal{A}_t(\mathcal{D}) \in \mathbb{R}^p$ at step t. In the following text, we prefer the notation $\mathcal{A}_t(\mathcal{D})$ to emphasize the dependency on dataset $\mathcal{D}$. The algorithm can be either deterministic (e.g., gradient descent) or randomized (e.g., stochastic gradient descent). When the context is clear, let $\mathcal{A} = \{\mathcal{A}_j\}_{j\in[t]}$ denote algorithms in all steps. To simplify the discussion, we assume that the algorithm starts from zero, namely, $\mathcal{A}_0(\mathcal{D}) = \mathbf{0}$. During the discussion, we are interested in the excess risk of $\mathcal{A}_t(\mathcal{D})$, namely, $\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{P})$. Without loss of generality, assume that $\mathcal{A}_0(\mathcal{D}_{\text{sig}}) = \mathcal{A}_0(\mathcal{D})$ and $\mathcal{A}_0(\mathcal{D}_{\text{noi}}) = \mathbf{0}$.

3.2. Algorithmic Stability

Algorithmic stability is one of the most popular approaches to generalization (Bousquet & Elisseeff, 2002; Hardt et al., 2016). Informally, algorithmic stability measures how the model performance alters when changing a training sample, which leads to generalization bound via Proposition 3.1.

Proposition 3.1 (Algorithmic stability, from Hardt et al. (2016))**.** *Assume that the algorithm $\mathcal{A}_t$ is γ-uniformly-stable, namely, for any two datasets $\mathcal{D}$ and $\mathcal{D}'$ with only one different data point,*

$$\sup_{\tilde{\boldsymbol{z}}} \mathbb{E}_{\mathcal{A}}[\ell(\mathcal{A}_t(\mathcal{D}); \tilde{\boldsymbol{z}}) - \ell(\mathcal{A}_t(\mathcal{D}'); \tilde{\boldsymbol{z}})] \le \gamma. \tag{2}$$

Then the following generalization bound holds

$$\mathbb{E}_{\mathcal{A},\mathcal{D}}[\mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}}\ell(\mathcal{A}_t(\mathcal{D}); \boldsymbol{z}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}); \mathcal{D})] \le \gamma. \tag{3}$$

One can generalize the results in Proposition 3.1 using other types of algorithmic stability, e.g., on-average algorithmic stability (Lei & Ying, 2020). A line of research derives generalization measures under specific regimes based on Proposition 3.1. Among them, the most popular one is the bound derived in general convex and smooth regimes, proposed in Proposition 3.2.

Proposition 3.2 (Convex and smooth regimes, from Hardt et al. (2016))**.** *Assume that the loss function $\ell(\cdot; \boldsymbol{z})$ is convex, M-smooth and L-Lipschitz for any sample $\boldsymbol{z}$, it holds that*

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}}[\mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}}\ell(\mathcal{A}_t(\mathcal{D}); \boldsymbol{z}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}); \mathcal{D})] \le \frac{2\eta t}{n} L^2, \tag{4}$$

where η denotes the constant stepsize satisfying $\eta \le 2/M$.

Based on Proposition 3.2, a t/n-type generalization measure directly follows, leading to the argument *train faster, generalize better* (Hardt et al., 2016). In the next section, we show a different generalization measure under the stability-based framework, contrasting the signal and noise during the training process.

4. Bounding Generalization via REF Complexity

In this section, we derive generalization bound using REF Complexity, providing theoretical guarantees for the metric. We first introduce the formal notion of REF Complexity in Section 4.1 and derive generalization bound under convex and smooth regimes with SGD in Section 4.2. We then relax the convex assumptions by considering smooth regimes with SGLD in Section 4.3. We further validate that such bounds hold beyond SGD and SGLD by considering GD under overparameterized linear regression regimes in Section 4.4. During the analysis, we consider the metric of excess risk introduced before, which is widely considered in the related literature (Bartlett et al., 2020; Teng et al., 2022).

4.1. REF Complexity

We first introduce the formal definition of REF Complexity, which quantifies the ability of a model-algorithm pair to distinguish between signal dataset $\mathcal{D}_{\text{sig}}$ and noise dataset $\mathcal{D}_{\text{noi}}$. Motivated from Equation 1, for a given training dataset $\mathcal{D}$ and training algorithm $\mathcal{A}_t$, its theoretical REF Complexity can be measured as

$$\begin{aligned}&\mathcal{T}_n^{\alpha}(\mathcal{D}, \mathcal{A}_t)\\ &= \frac{1 - \mathbb{E}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})}{1 - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{sig}}); \mathcal{D}_{\text{sig}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{sig}}); \mathcal{D}_{\text{sig}})},\end{aligned} \tag{5}$$

where the expectation is taken over the random noise in $\mathcal{D}_{\text{noi}}$. The metric $\mathcal{T}_n^{\alpha}(\mathcal{D}, \mathcal{A}_t)$ becomes larger when fitting noise more (with smaller $\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{sig}}); \mathcal{D}_{\text{sig}})$), given the degree of fitting signal.

4.2. SGD under Convex and Smooth Regimes

This section introduces a generalization bound via REF Complexity in convex cases with SGD, starting from the basic notations. The core technique in the proof is algorithmic stability. The key intuition is that, one can bound the algorithm stability using a cumulative gradient, which is further bounded by the degree of fitting noise.

Settings. We follow the notations in Section 3 when the context is clear. Additionally, we consider a specific algorithm $\mathcal{A}_t$: constant-stepsize SGD with replacement, where the iteration performs as

$$\mathcal{A}_{t+1}(\mathcal{D}) = \mathcal{A}_t(\mathcal{D}) - \eta\nabla\ell(\mathcal{A}_t(\mathcal{D}); \boldsymbol{z}_t), \tag{6}$$

where $\boldsymbol{z}_t$ is sampled uniformly from dataset $\mathcal{D}$. We sketch the gradient noise in step t as $\sigma_w^2(t;\mathcal{D}) = \mathbb{E}_{\mathcal{A},\mathcal{D}}\frac{1}{n}\sum_{i\in[n]}\|\ell(\mathcal{A}_t(\mathcal{D});\boldsymbol{z}_i)\|^2 - \|\nabla\mathcal{L}_n(\mathcal{A}_t(\mathcal{D});\mathcal{D})\|^2$. Similar notations are also used in optimization-relevant papers (Shalev-Shwartz & Ben-David, 2014). We assume a bounded gradient noise regime in the noise training, where $\sigma_w^2(t;\mathcal{D}_{\text{noi}}) \leq \sigma_w^2 = \mathcal{O}(1)$ for any step t. Besides, we assume that the gradient noise is non-increasing during the noisy training process, namely, $\mathbb{E}_{\mathcal{A},\mathcal{D}_{\text{noi}}}\sigma_w^2(t;\mathcal{D}_{\text{noi}}) \leq \mathbb{E}_{\mathcal{A},\mathcal{D}_{\text{noi}}}\sigma_w^2(j;\mathcal{D}_{\text{noi}})$ for any $j \leq t$. This assumption is valid under convex regimes where the gradient is approximately non-increasing (Li et al., 2020).

Additionally, we assume the following Decomposition condition for the excess risk, aiming to decompose the influence of signal and noise in the generalization analysis.

Assumption 4.1 (Excess Risk Decomposition)**.** We assume that the excess risk can be decomposed into its signal component and noise component, namely, there exists a constant c_1 such that for any given time $t \geq T_1$,

$$\begin{aligned}&\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D});\mathcal{P})\\ \leq& c_1\left[\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}}) + \mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{sig}});\mathcal{P}_{\text{sig}})\right] + \psi_1(n),\end{aligned}$$

where $\psi_1(n) \to 0$ as $n \to \infty$, and the expectation is taken over both algorithm $\mathcal{A}_t$ and the dataset $\mathcal{D}$.

Assumption 4.1 can hold in both linear and non-linear cases under some additional assumptions, as demonstrated in Teng et al. (2022). The next Assumption 4.2 sketches the properties of signal training and noise training.

Assumption 4.2 (Signal and Noise Training)**.** We assume that the signal training component satisfies for any $t \geq T_2$, there exists a constant c_2 such that

$$\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{sig}});\mathcal{P}_{\text{sig}}) \leq c_2\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}}) + \psi_2(n), \tag{7}$$

where $\psi_2(n) \to 0$ as $n \to \infty$. Besides, we assume that the noise training component satisfies that for any $t \geq T_3$,

$$\mathbb{E}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}}) \leq \mathbb{E}\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}}). \tag{8}$$

We additionally assume that the initial loss in noise training is bounded, namely, there exists a constant c_3 such that

$$\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}}) \leq c_3. \tag{9}$$

The first part on signal implies that signal training is a relatively simpler task than noise training, which is demonstrated empirically (*e.g.*, Arora et al. (2019); Zhang et al. (2021)) and theoretically (*e.g.*, Gaussian Mixture Models (Cao et al., 2021), overparameterized linear regression and Hypercube Classifier (Negrea et al., 2020)). The second part on noise requires that the training loss decreases during noise training in expectation, without which REF Complexity might become negative. This holds with a sufficiently small learning rate, guaranteed by optimization theory (Shalev-Shwartz & Ben-David, 2014).

We next show in Proposition 4.3 that overparameterized linear regression regime with MSE loss satisfies the above assumptions (Teng et al., 2022).

Proposition 4.3. *Overparameterized linear regression regimes satisfy both Assumption 4.1 and Assumption 4.2. Specifically, when the optimal parameter $\|\boldsymbol{\theta}^*\| = O(1)$ and sample covariance $\|\Sigma_{\boldsymbol{x}}\| = O(1)$, we derive that*

- *(a.) For all step t, it holds that $\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D});\mathcal{P}) \leq 2[\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{sig});\mathcal{P}_{sig}) + \mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{noi});\mathcal{P}_{noi})]$;*
- *(b.) For $t \geq n$, $\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{sig});\mathcal{P}_{sig}) = O(\frac{1}{\sqrt{n}})$;*
- *(c.) With sufficiently small η, $\mathbb{E}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{noi});\mathcal{D}_{noi}) \leq \mathbb{E}\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{noi});\mathcal{D}_{noi})$.*

We are now ready to introduce the main theorem, which bounds the excess risk using REF Complexity $\mathcal{T}_n^\alpha(\mathcal{D},\mathcal{A}_t)$ in general convex regimes.

Theorem 4.4 (Convex, smooth, with SGD)**.** *Assume that the loss $\ell(\boldsymbol{\theta};\boldsymbol{z})$ is convex and M-smooth with respect to $\boldsymbol{\theta}$ for any sample $\boldsymbol{z}$. Consider the SGD training regime with constant stepsize η. Under Assumption 4.1 and Assumption 4.2, the following inequality holds when $\sum_{j\in[t]}\sigma_w^2(j;\mathcal{D}_{noi}) = o(n^2)$, $t \geq \max\{T_1, T_2, T_3\}$, and $\eta \leq \frac{1}{\sqrt{t}}$,*

$$\begin{aligned}&\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D});\mathcal{P})\\ \leq&\frac{c\max\left\{u,u^2\right\}}{\sqrt{t}}\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{T}_n^\alpha(\mathcal{D},\mathcal{A}_t) + \psi(n),\end{aligned} \tag{10}$$

where we define $u \triangleq \sqrt{\frac{1}{n}(1+\frac{t}{n})}$ for simplicity, and the term $\psi(n) \to 0$ as $n \to \infty$. The constant $c > 0$ denotes a constant related to the constant c_1, c_2, c_3, M in Assumption 4.1 and Assumption 4.2.

Derived from Theorem 4.4, REF Complexity $\mathcal{T}_n^\alpha(\mathcal{D},\mathcal{A}_t)$ is valid from two aspects: (a) if $\mathcal{T}_n^\alpha(\mathcal{D},\mathcal{A}_t)$ is relatively small, the excess risk is consistent and, therefore, would be relatively small. Here we use consistency to represent a bound that converges to zero as the sample size goes to infinity. (b) if $\mathcal{T}_n^\alpha(\mathcal{D}, \mathcal{A}_t)$ is relatively large, the bound is dominated by the first term. Therefore, $\mathcal{T}_n^\alpha(\mathcal{D}, \mathcal{A}_t)$ is a proper index for generalization since the proposed upper bound shows an approximate correlation. We refer to Figure 3 in Appendix F for more illustration.

About the order in ψ. Besides the order of ψ_1, ψ_2 in Assumption 4.1 and Assumption 4.2, the order of ψ is also closely related to the term $\frac{1}{n^2}\sum_{j\in[t]}\sigma_w^2(j;\mathcal{D}_{\text{noi}})$. To ensure the consistency, we assume that $\sum_{j\in[t]}\sigma_w^2(j;\mathcal{D}_{\text{noi}}) = o(n^2)$. If $t = o(n^2)$ the assumption directly holds since $\sigma_w^2(j;\mathcal{D}_{\text{noi}}) = O(1)$. However, the estimation on $\sum_{j\in[t]}\sigma_w^2(j;\mathcal{D}_{\text{noi}})$ can be much better, since the gradient norm usually decreases in expectation along the trajectory under convex regimes (*e.g.*, strong growth assumption in Schmidt & Roux (2013); Cevher & Vu (2019). This would lead to a weaker requirement on t.

About other assumptions. The convex and smooth assumption used in Theorem 4.4 are also used in algorithmic stability relevant papers (*e.g.*, Lei & Ying (2020)). Besides, the stepsize assumption is valid in SGD-relevant analysis (*e.g.*, section 6.2 in Bubeck (2015)). We also remark that the assumption $\sum_{j\in[t]}\sigma_w^2(j;\mathcal{D}_{\text{noi}}) = o(n^2)$ usually do not contradict to the time requirement T_1, T_2, T_3 used in Assumption 4.1 and Assumption 4.2. For example, in overparameterized linear regression cases, the first assumption is weaker than $t = o(n^2)$, and the second assumption requires that $t \geq \max\{T_1, T_2, T_3\} = n$. Therefore the bound is at least valid in the region $t \in (\Omega(n), o(n^2))$[1].

Here are three key steps during the proof. The first is to decompose the excess risk into signal component and noise component based on Assumption 4.1 and Assumption 4.2. The second is to bound the algorithmic stability of the noise part using the cumulative gradient, based on the convex and smooth assumption. And the third is to bound cumulative gradient using REF Complexity, which is derived by smoothness assumption. We defer the whole proof to Appendix A. We finally remark that we here provide the generalization bound with the expectation version instead of the high probability version, due to the inherent properties of stability-based techniques. One can generalize the results to high probability versions following Feldman & Vondrák (2019); Bousquet et al. (2020).

Remark 4.5 (Comparison to algorithmic stability). The measures proposed in Theorem 4.4 are fundamentally different from the stability-based approaches, although our bound is derived via stability-based techniques. The measure proposed in this paper explicitly quantifies the ability to distinguish signal and noise, which differs from the existing measures. We finally remark that the goal of Theorem 4.4 is not to provide a tight bound but to validate the utility of REF Complexity.

4.3. SGLD under Smooth Regime

In this section, we relax the convexity assumptions required in Section 4.2. The reason we need convexity in Section 4.2 is the one-expansion property under convexity with SGD, required by algorithmic stability analysis (See Lemma A.4 in Appendix). This property is easily violated under nonconvex regimes. Fortunately, the convexity assumption can be avoided in Stochastic Gradient Langevin Dynamics (SGLD) training (Mou et al., 2018; Li et al., 2020).

Settings. We follow the notations and basic assumptions in Section 3 and Section 4.2 when the context is clear. Unlike the SGD settings, the iteration of SGLD performs as

$$\mathcal{A}_{t+1}(\mathcal{D}) = \mathcal{A}_t(\mathcal{D}) - \eta\nabla\ell(\mathcal{A}_t(\mathcal{D}); \boldsymbol{z}_t) + \frac{\sigma}{\sqrt{2}}\boldsymbol{n}_t,$$

where $\boldsymbol{n}_t \in \mathbb{R}^p$ follows a standard Gaussian distribution, and $\boldsymbol{z}_t$ is sampled uniformly from dataset $\mathcal{D}$. We next introduce the theorem based on SGLD, which does not require convexity assumptions.

Theorem 4.6 (Smooth, with SGLD)**.** *Assume that the loss $\ell(\boldsymbol{\theta}; \boldsymbol{z})$ is $O(1)$-bounded, L-Lipschitz, and M-smooth with respect to $\boldsymbol{\theta}$ for any sample $\boldsymbol{z}$. Consider SGLD with noise scale σ and stepsize $\eta < \min\{\frac{\sigma}{20L}, \frac{1}{M}\}$. Under Assumption 4.1 and Assumption 4.2, if $p = o(\frac{n^2}{\eta t})$, the following inequality holds when $\sum_{j\in[t]}\sigma_w^2(j;\mathcal{D}_{noi}) = o(\frac{n^2\sigma^2}{\eta^2})$ and $t \geq \max\{T_1, T_2, T_3\}$,*

$$\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{P}) \leq c'\frac{\sqrt{\eta}}{n\sigma}\sqrt{\mathcal{T}_n^\alpha(\mathcal{D}, \mathcal{A}_t)} + \psi'(n), \quad (11)$$

where the term $\psi'(n) \to 0$ as $n \to \infty$. The constant $c' > 0$ denotes a constant related to the constant c_1, c_2, c_3, M in Assumption 4.1 and Assumption 4.2.

Compared to Theorem 4.4, Theorem 4.6 has a milder assumption on the gradient noise $\sum_{j\in[t]}\sigma_w^2(j;\mathcal{D}_{\text{noi}})$ for a small learning rate. The benefit comes from the tighter bound of SGLD compared to SGD. Unfortunately, one may notice that there is an additional dimension-dependent term in Theorem 4.6, namely, $p = o(\frac{n^2}{\eta t})$. This comes from the noise term $\boldsymbol{n}_t \in \mathbb{R}^p$ where a large dimension would bring more noise. We defer the whole proofs to Appendix B.

4.4. GD under Overparameterized Linear Regression

To validate the generality of REF Complexity, we prove a similar argument under overparameterized linear regression

[1] We here use the notation $(\Omega(n), o(n^2))$ to represent an interval with lower bound in order $\Omega(n)$ and upper bound in order $o(n^2)$.

Algorithm 1 Estimate REF Complexity in practice

```
Input: Training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i\in[n]}$, optimization algorithm $\mathcal{A}_t$, training loss function $\mathcal{L}_n(\cdot,\cdot)$.
1: Calculate the training loss on step $0$ and step $t$ for the real-world dataset, namely, $\mathcal{L}_n(\mathcal{A}_0(\mathcal{D});\mathcal{D})$ and $\mathcal{L}_n(\mathcal{A}_t(\mathcal{D});\mathcal{D})$;
2: Generate $m$ randomly labeled datasets $\mathcal{D}_{\text{noi}}^{(j)} = \{(\boldsymbol{x}_i, \tilde{y}_i^{(j)})\}_{i\in[n]}, j \in [m]$, where $\tilde{y}_i^{(j)}$ denotes a random noise;
3: Calculate the training loss on step $0$ and step $t$ for the randomly labeled dataset, namely, $\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}^{(j)});\mathcal{D}_{\text{noi}}^{(j)})$ and $\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(j)});\mathcal{D}_{\text{noi}}^{(j)})$;
Output: $\mathcal{T}_n^{\beta}(\mathcal{D}, \mathcal{A}_t) = \frac{\mathcal{L}_n(\mathcal{A}_t(\mathcal{D});\mathcal{D})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D});\mathcal{D})}{\frac{1}{m}\sum_{j\in[m]}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(j)});\mathcal{D}_{\text{noi}}^{(j)})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}^{(j)});\mathcal{D}_{\text{noi}}^{(j)})}$.
```

regimes. One may generalize the results to kernel regression regimes (*e.g.*, neural tangent kernel), which is left for future work. Our techniques in this section are inspired by Bartlett et al. (2020); Xu et al. (2022).

Settings. We follow the notations in Section 3 when the context is clear. Additionally, set $f(x;\boldsymbol{\theta}^*) = x^\top \boldsymbol{\theta}^*$ as the ground truth function. Let $\Sigma_{\boldsymbol{x}} \triangleq \mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top$ denote the covariance matrix with non-increasing eigenvalues $\lambda_i, i \in [d]$. Let $r_k(\Sigma) = \frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}$ denote the corresponding effective rank, and $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ for some constant $b > 0$. Assume that the noise $y - \boldsymbol{x}^\top\boldsymbol{\theta}^*$ is σ_y^2-subGaussian, and $\boldsymbol{x} = \Sigma_{\boldsymbol{x}}^{1/2}\boldsymbol{z}$ can be represented as linear transformation of $\boldsymbol{z}$ where $\boldsymbol{z}$ denotes a random vector with independent and $\sigma_{\boldsymbol{x}}^2$-subGaussian coordinate.

Theorem 4.7 (Overparameterized Linear Regression with GD)**.** *Under overparameterized linear regression regimes, assume that* $r_0(\Sigma) = o(n)$ *and* $k^* = o(n)$. *Besides, assume that* $\|\boldsymbol{\theta}^*\|_2 = O(1)$, $\|\Sigma_{\boldsymbol{x}}\|_2 = O(1)$ *in a constant scale. We consider the GD training process with zero initialization and constant stepsize* η. *For any given* $\delta > 0$ *which does not vary with sample size* n *and satisfies* $\log(1/\delta) = o(n)$, *for* $t = \omega(1)$[2], *with probability at least* $1 - \delta$,

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D});\mathcal{D}) \leq c\log(1/\delta)\sigma_y^2\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t) + \tilde{\psi}(n), \quad (12)$$

where $\tilde{\psi}(n) \to 0$ *as* $n \to \infty$ *and* $c > 0$ *denotes a constant.*

We remark that the bound proposed here can be consistent if $\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t) \to 0$ as $n \to \infty$ for some given fixed t. This usually holds when $t = o(n)$ with constant stepsize. Besides, different from the results proposed in Theorem 4.4, Theorem 4.7 does not contain time dependency (t/n-type term). This is due to the different techniques used in the proof. Unfortunately, to our best known, the techniques used in this section cannot be easily applied to general convex regimes. We defer the whole proofs to Appendix C.

5. Experiment

This section provides experimental results to validate the utility of REF Complexity, where we defer the experiment details in Appendix D. Before showing the experiment results, we first revisit the formal definition of REF Complexity in Section 4.1. Notice that Equation 5 requires a clean dataset $\mathcal{D}_{\text{sig}}$. This simplifies the theoretical discussion but is nearly impossible in practice, since real-world datasets usually mix signal and noise. Despite all this, a possible way is to quantify a data-algorithm pair's ability to distinguish the *real-world* dataset and the *randomly labeled* dataset. Such a metric implies the ability to distinguish between signal and noise, since the real-world dataset usually contains enough signal information. Besides, since we use the real-world dataset to surrogate the signal dataset, it is safer to put the related term in the numerator instead of the denominator. Based on the discussion above, we formulate the REF Complexity used in practice as $\mathcal{T}_n^{\beta}(\mathcal{D},\mathcal{A}_t)$, given the dataset $\mathcal{D}$ and algorithm $\mathcal{A}_t$,

$$\mathcal{T}_n^{\beta}(\mathcal{D},\mathcal{A}_t) = \frac{\mathcal{L}_n(\mathcal{A}_t(\mathcal{D});\mathcal{D})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D});\mathcal{D})}{\mathbb{E}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})},$$

where n denotes the sample size, and the expectation is taken over the randomness in $\mathcal{D}_{\text{noi}}$. REF Complexity $\mathcal{T}_n^{\beta}(\mathcal{D},\mathcal{A}_t)$ becomes larger when fitting noise more.

We summarize the algorithm in Algorithm 1, which returns the REF Complexity value $\mathcal{T}_n^{\beta}(\mathcal{D},\mathcal{A}_t)$. The construction of random noise (Step 2) varies from task to task. For example, we can use Gaussian random noise in regression problems and uniform random labels in classification problems.

5.1. REF Complexity in CIFAR-10 and CIFAR-100

The first experiment aims to show the correlation between REF Complexity and generalization metrics. Specifically, we conduct over one hundred experiments on CIFAR-10 and CIFAR-100, and plot each regime's test accuracy and REF Complexity in Figure 1. Experimental results in Figure 1 illustrate that REF Complexity negatively correlates to test accuracy with correlations of -0.988 and -0.960 on CIFAR-10 and CIFAR-100, respectively.

As a comparison, the measures without noise terms (*REF w/o noi*), $\frac{\mathcal{L}_n(\mathcal{A}_t(\mathcal{D});\mathcal{D})}{\mathcal{L}_n(\mathcal{A}_0(\mathcal{D});\mathcal{D})}$ only return correlations -0.585 and -0.481, showing that the noise dataset is crucial in REF Complexity. Besides, the measures without initial terms

[2] The statement $t = \omega(1)$ means that $t \to \infty$ as $n \to \infty$.

Table 1. Correlation between REF Complexity and test accuracy. A generalization measure performs well if the correlations are all positive/negative, and their absolute values are close to one. The baseline measures follow Jiang et al. (2020b) (see Appendix D).

Type	Norm-based Measures					
Measure	L2-Norm	L2-Dist	F-norm	Inv-Margin	Spectral	Spectral/Margin
batch size	-0.935	0.899	-0.938	0.944	-0.957	-0.942
learning rate	-0.910	0.980	-0.911	-0.959	-0.653	-0.982
dropout	0.452	-0.072	0.449	0.676	-0.764	0.473

Type	Norm-based	Sharpness-based				
Measure	Path-Norm	PB-I	PB-O	PB-Flatness	PB-M-I	PB-M-O
batch size	-0.996	0.830	0.824	0.835	0.700	-0.909
learning rate	-0.927	0.976	0.978	0.977	0.992	0.981
dropout	0.452	-0.899	-0.895	-0.898	-0.647	-0.651

Type	Sharpness-based	Stability-based		Ours		
Measure	PB-M-Flatness	Steps(1)	Steps(1.5)	REF (w/o noi)	REF (w/o init)	**REF Complexity**
batch size	0.638	0.911	0.910	-0.782	-0.997	**-0.998**
learning rate	0.977	-0.985	-0.734	-0.501	-0.996	**-0.997**
dropout	-0.898	-0.908	-0.898	-0.766	-0.964	**-0.965**

(*REF w/o init*) $\frac{\mathcal{L}_n(\mathcal{A}_t(\mathcal{D});\mathcal{D})}{\mathbb{E}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})}$ also performs well (with correlation -0.988 and -0.959), meaning that the initial loss does not affect much.

5.2. Varying Only One Parameter

Besides the experiments in Section 5.1, we also evaluate REF Complexity with only one varying parameter in Table 1. Experiment results show the success of REF Complexity from two aspects: (a) consistency, where the correlation of REF Complexity is *always* negative, and (b) effectiveness, where the correlation values of REF Complexity are stably large (exceed 0.95).

Besides varying the batch size, learning rate, and dropout as mentioned in Table 1, we also conduct additional experiments to validate the effectiveness of REF Complexity by varying sample size, label noise, width, and structure. The results of these experiments are summarized in Table 2. These additional experiments aim to demonstrate the consistency and effectiveness of REF Complexity in different scenarios. Since the baseline results were already inconsistent, we do not present them in Table 2.

5.3. Phenomenon Explanation Under REF Complexity

We next show that REF Complexity helps explain the deep learning phenomenon from a different perspective, taking stochastic algorithms as an example. The success of stochastic algorithms (*e.g.*, SGD and its variants) is widely observed in deep learning regimes. REF Complexity explains such success, as Table 1 shows (where different batch sizes to surrogate GD and SGD). For deterministic algorithms (*e.g.*, GD), each iteration sees all samples, and therefore the training loss decreases in each iterate for both signal and noise training. For stochastic algorithms (*e.g.*, SGD), each iteration only sees part of the samples. For signal training, the model still learns useful information since each sample shares the same pattern. However, things can be much more different in noise training. The model may even oscillate since the pattern in the first batch may damage the training loss on the remaining samples. This leads to a better REF Complexity for stochastic algorithms. We illustrate this phenomenon in Appendix F (Figure 4). As a comparison, original stability-based approaches (steps required to reach a given loss) may even show a wrong correlation.

5.4. Effect on Training Epoch T

Previous experiment results in Section 5.1 and Section 5.2 demonstrate the effectiveness of REF Complexity under a fixed training epoch. These findings align with the theoretical results presented in Section 4, where the upper bound is associated with the training epoch.

In this section, we investigate the impact of the training epoch T on the efficacy of complexity. Specifically, we first establish that the correlation between REF Complexity and test accuracy remains valid for epochs that are not excessively large. However, the correlation may substantially deteriorate for extremely large epochs. Fortunately, even for epochs where the correlation holds, the training loss has already reached a considerably low level. For example, with an epoch of $T = 150$, the training loss is around 0.01 and

Table 2. Correlation between REF Complexity and test accuracy with varying parameters including sample size (n), additional label noise (s), width (w) and structure. The correlation performs consistently for REF Complexity.

VARYING PARAMETER	CORRELATION	NOTE
SAMPLE SIZE n	-0.9929	$n = 4000, 10000, 20000, 40000$
ADDITIONAL LABEL NOISE s	-0.9824	$s = 0.1, 0.2, 0.3, 0.4$
WIDTH w	-0.9204	$w = 1, 2, 4, 8$
STRUCTURE	-0.9189	RESNET20, RESNET32, RESNET56, VGG16, VGG19

Table 3. Correlation between REF Complexity and test accuracy varies with different epochs. As the time T increases, the correlation tends to decrease. Nevertheless, it is worth noting that even for relatively small T values, the training loss is already significantly minimized. Consequently, despite the diminishing correlation, the concept of complexity remains practical and valuable in real-world scenarios.

EPOCH T	REF	REF (W/O INIT)	REF (W/O NOISE)	REF (W/O SIGNAL)	TRAINING LOSS
50	-0.9743	-0.9824	-0.2819	-0.0544	0.29067
100	-0.9487	-0.9551	-0.6394	0.2555	0.1465
150	-0.9044	-0.9141	-0.5129	0.1066	0.0871
200	-0.8866	-0.8974	-0.4487	0.0189	0.0633
250	-0.7733	-0.7816	-0.5873	0.353	0.0495
300	-0.6525	-0.664	-0.4353	0.2856	0.0351
400	-0.4958	-0.4983	-0.4682	0.3279	0.0284
500	-0.2202	-0.2142	-0.3107	0.1945	0.0250

the training accuracy exceeds 0.95. Thus, we argue that it is still meaningful to utilize complexity in practical applications. We summarize the experiment results in Table 3.

5.5. Comparison to Rademacher Complexity

We close the section with the comparison between REF Complexity and Rademacher complexity. Both metrics focus on the ability to fit noise. However, Rademacher complexity measures the noise-fitting ability for a given *function class*, while REF Complexity measures it for a given model-algorithm pair. Besides, REF Complexity distinguishes the signal influence and the noise influence, which is not covered in Rademacher Complexity. As an algorithm-independent and output-independent measure, Rademacher complexity is inconsistent and vacuous since neural networks can fit arbitrary random noise (Zhang et al., 2021). In comparison, REF Complexity is noise recognizable since $\mathcal{T}_n^\beta(\mathcal{D}, \mathcal{A}_t)$ becomes larger when the real-world dataset $\mathcal{D}$ contains more noise, leading to a large generalization error.

6. Conclusion

This paper proposes a new generalization measure REF Complexity under algorithmic stability frameworks, which contains theoretical backbones and empirically works well in standard training scenarios. The complexity is motivated by the intuition that a model-algorithm pair may generalize better if it quickly captures the signal while adapting to the noise slowly. The success of REF Complexity inspires several future directions. From the empirical view, one may find more generalization measures using signal-noise techniques. From the theoretical view, one may relax the assumption on gradient noise used in Theorem 4.4, and the dimension dependency in Theorem 4.6. Another interesting direction is predicting exact generalization errors instead of only the trend using REF Complexity frameworks. This may potentially inspire new standards in practice parallel to cross-validation. One can track the algorithmic performance on a randomly labeled dataset during training, and compare different models based on it. Additionally, REF Complexity might have some inherent limitations. For example, it may fail when the model completely fits both the signal and the noise. It would be interesting to modify and improve REF Complexity in such scenarios.

Acknowledgements

This work is supported by the National Key R&D Program of China (2020AAA0105200). Jiaye Teng would express his sincere thanks to Ruiqi Gao, Jianhao Ma and Jingzhao Zhang for their insightful discussion.

References

Aithal, S. K., Kashyap, D., and Subramanyam, N. Robustness to augmentations as a generalization metric. *CoRR*, abs/2101.06459, 2021. URL `https://arxiv.org/abs/2101.06459`.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 322–332. PMLR, 2019. URL http://proceedings.mlr.press/v97/arora19a.html.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017. URL http://proceedings.mlr.press/v70/arpit17a.html.

Bansal, Y., Kaplun, G., and Barak, B. For self-supervised learning, rationality implies generalization, provably. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=Srmggo3b3X6.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. Spectrally-normalized margin bounds for neural networks. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6240–6249, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/b22b257ad0519d4500539da3c8bcf4dd-Abstract.html.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bassily, R., Feldman, V., Guzmán, C., and Talwar, K. Stability of stochastic gradient descent on nonsmooth convex losses. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/2e2c4bf7ceaa4712a72dd5ee136dc9a8-Abstract.html.

Bousquet, O. and Elisseeff, A. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002. URL http://jmlr.org/papers/v2/bousquet02a.html.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. In Abernethy, J. D. and Agarwal, S. (eds.), *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pp. 610–626. PMLR, 2020. URL http://proceedings.mlr.press/v125/bousquet20b.html.

Bubeck, S. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, 2015. doi: 10.1561/2200000050. URL https://doi.org/10.1561/2200000050.

Cao, Y., Gu, Q., and Belkin, M. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8407–8418, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/46e0eae7d5217c79c3ef6b4c212b8c6f-Abstract.html.

Cevher, V. and Vu, B. C. On the linear convergence of the stochastic gradient method with constant step-size. *Optim. Lett.*, 13(5):1177–1187, 2019. doi: 10.1007/s11590-018-1331-1. URL https://doi.org/10.1007/s11590-018-1331-1.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In Elidan, G., Kersting, K., and Ihler, A. T. (eds.), *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017. URL http://auai.org/uai2017/proceedings/papers/173.pdf.

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., and Roy, D. M. In search of robust measures of generalization. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/86d7c8a08b4aaa1bc7c599473f5dddda-Abstract.html.

Emmert-Streib, F. and Dehmer, M. Evaluation of regression models: Model assessment, model selection and generalization error. *Mach. Learn. Knowl. Extr.*, 1(1): 521–551, 2019. doi: 10.3390/make1010032. URL https://doi.org/10.3390/make1010032.

Feldman, V. and Vondrák, J. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Beygelzimer, A. and Hsu, D. (eds.), *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1270–1279. PMLR, 2019. URL http://proceedings.mlr.press/v99/feldman19a.html.

Frénay, B. and Verleysen, M. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894. URL https://doi.org/10.1109/TNNLS.2013.2292894.

Garg, S., Balakrishnan, S., Kolter, J. Z., and Lipton, Z. C. RATT: leveraging unlabeled data to guarantee generalization. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3598–3609. PMLR, 2021. URL http://proceedings.mlr.press/v139/garg21a.html.

Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., and Dziugaite, G. K. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/712a3c9878efeae8ff06d57432016ceb-Abstract.html.

Haghifam, M., Dziugaite, G. K., Moran, S., and Roy, D. Towards a unified information-theoretic framework for generalization. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 26370–26381, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ddbc86dc4b2fbfd8a62e12096227e068-Abstract.html.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In Balcan, M. and Weinberger, K. Q. (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1225–1234. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/hardt16.html.

Jiang, Y., Natekar, P., Sharma, M., Aithal, S. K., Kashyap, D., Subramanyam, N., Lassance, C., Roy, D. M., Dziugaite, G. K., Gunasekar, S., Guyon, I., Foret, P., Yak, S., Mobahi, H., Neyshabur, B., and Bengio, S. Methods and analysis of the first competition in predicting generalization of deep learning. In Escalante, H. J. and Hofmann, K. (eds.), *NeurIPS 2020 Competition and Demonstration Track, 6-12 December 2020, Virtual Event / Vancouver, BC, Canada*, volume 133 of *Proceedings of Machine Learning Research*, pp. 170–190. PMLR, 2020a. URL http://proceedings.mlr.press/v133/jiang21a.html.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020b. URL https://openreview.net/forum?id=SJgIPJBFvH.

Jiang, Y., Nagarajan, V., Baek, C., and Kolter, J. Z. Assessing generalization of SGD via disagreement. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=WvOGCEAQhxl.

Johnson, J. B. and Omland, K. S. Model selection in ecology and evolution. *Trends in ecology & evolution*, 19(2):101–108, 2004.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Lei, Y. and Ying, Y. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5809–5819. PMLR, 2020. URL http://proceedings.mlr.press/v119/lei20c.html.

Li, J., Luo, X., and Qiao, M. On generalization error bounds of noisy gradient methods for non-convex learning. In *8th*

International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkxxtgHKPS.

Manwani, N. and Sastry, P. S. Noise tolerance under risk minimization. *IEEE Trans. Cybern.*, 43(3): 1146–1151, 2013. doi: 10.1109/TSMCB.2012.2223460. URL https://doi.org/10.1109/TSMCB.2012.2223460.

McAllester, D. A. Pac-bayesian model averaging. In Ben-David, S. and Long, P. M. (eds.), *Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT 1999, Santa Cruz, CA, USA, July 7-9, 1999*, pp. 164–170. ACM, 1999. doi: 10.1145/307400.307435. URL https://doi.org/10.1145/307400.307435.

Mou, W., Wang, L., Zhai, X., and Zheng, K. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pp. 605–638. PMLR, 2018. URL http://proceedings.mlr.press/v75/mou18a.html.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 11611–11622, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/05e97c207235d63ceb1db43c60db7bbb-Abstract.html.

Natekar, P. and Sharma, M. Representation based complexity measures for predicting generalization in deep learning. *CoRR*, abs/2012.02775, 2020. URL https://arxiv.org/abs/2012.02775.

Negrea, J., Dziugaite, G. K., and Roy, D. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7263–7272. PMLR, 2020. URL http://proceedings.mlr.press/v119/negrea20a.html.

Neyshabur, B., Tomioka, R., and Srebro, N. Norm-based capacity control in neural networks. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pp. 1376–1401. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Neyshabur15.html.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5947–5956, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Pitas, K., Davies, M. E., and Vandergheynst, P. Pac-bayesian margin bounds for convolutional neural networks - technical report. *CoRR*, abs/1801.00171, 2018. URL http://arxiv.org/abs/1801.00171.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. C. On the spectral bias of neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019. URL http://proceedings.mlr.press/v97/rahaman19a.html.

Rolnick, D., Veit, A., Belongie, S. J., and Shavit, N. Deep learning is robust to massive label noise. *CoRR*, abs/1705.10694, 2017. URL http://arxiv.org/abs/1705.10694.

Rudin, C. Stability analysis for regularized least squares regression. *CoRR*, abs/cs/0502016, 2005. URL http://arxiv.org/abs/cs/0502016.

Russo, D. and Zou, J. Controlling bias in adaptive data analysis using information theory. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pp. 1232–1240.

JMLR.org, 2016. URL http://proceedings.mlr.press/v51/russo16.html.

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

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-10-705713-5. URL http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms.

Teng, J., Ma, J., and Yuan, Y. Towards understanding generalization via decomposing excess risk dynamics. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=rS9-7AuPKWK.

Wei, C. and Ma, T. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJe_yR4Fwr.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2524–2533, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/ad71c82b22f4f65b9398f76d8be4c615-Abstract.html.

Xu, J., Teng, J., Yuan, Y., and Yao, A. C.-C. When do models generalize? a perspective from data-algorithm compatibility, 2022. URL https://arxiv.org/abs/2202.06054.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.

Zucchini, W. An introduction to model selection. *Journal of mathematical psychology*, 44(1):41–61, 2000.

Appendix

We show the deferred proof of Theorem 4.4 in the Appendix A, the deferred proof of Theorem 4.6 in the Appendix B, and the deferred proof of Theorem 4.7 in the Appendix C. When the notations are clear, $\sigma_w^2(j)$ denotes the gradient noise in noisy training, namely, $\sigma_w^2(j;\mathcal{D}_{\text{noi}})$. We then introduce experiment details in Appendix D, and show several special cases to illustrate REF Complexity in Appendix E. We finally show the omitted illustration graph in Appendix F.

A. Proof of Theorem 4.4

Theorem 4.4 (Convex, smooth, with SGD)**.** *Assume that the loss $\ell(\boldsymbol{\theta};\boldsymbol{z})$ is convex and M-smooth with respect to $\boldsymbol{\theta}$ for any sample $\boldsymbol{z}$. Consider the SGD training regime with constant stepsize η. Under Assumption 4.1 and Assumption 4.2, the following inequality holds when $\sum_{j\in[t]}\sigma_w^2(j;\mathcal{D}_{noi}) = o(n^2)$, $t \ge \max\{T_1, T_2, T_3\}$, and $\eta \le \frac{1}{\sqrt{t}}$,*

$$\begin{aligned}&\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D});\mathcal{P})\\ \le&\frac{c\max\left\{u,u^2\right\}}{\sqrt{t}}\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)+\psi(n),\end{aligned}\tag{10}$$

where we define $u \triangleq \sqrt{\frac{1}{n}(1+\frac{t}{n})}$ for simplicity, and the term $\psi(n)\to 0$ as $n\to\infty$. The constant $c>0$ denotes a constant related to the constant c_1, c_2, c_3, M in Assumption 4.1 and Assumption 4.2.

Proof. Firstly, due to Assumption 4.1 and Assumption 4.2, the difficulties of bounding the excess risk falls in the noise component, that is to say,

$$\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D});\mathcal{P}) \le [c_1+c_1c_2]\mathbb{E}_{\mathcal{D}_{\text{noi}},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})+\psi_1(n)+c_1\psi_2(n).\tag{13}$$

We next focus on the excess risk of the noise component. The first step is to bound the excess risk via the generalization gap via Lemma A.1,

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}}) \le \mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})].\tag{14}$$

The next step is to bound the generalization gap via Lemma A.2, where we use the notion of on-average model stability proposed in Lei & Ying (2020). We derive that

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\\ \le&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\frac{2e(M+c)\eta^2}{n}(1+\frac{t}{n})+\frac{1}{2c}(\frac{1}{t}+\eta^2M^2)]\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2+\frac{1}{2c}\sigma_w^2(t).\end{aligned}\tag{15}$$

We finally apply Lemma A.3, which leads to

$$\begin{aligned}&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\\ \le&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\frac{2e(M+c)\eta^2}{n}(1+\frac{t}{n})+\frac{1}{2c}(\frac{1}{t}+\eta^2M^2)]\frac{2}{\eta}\mathbb{E}[\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}))-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}))]\\ &+[\frac{2e(M+c)\eta^2}{n}(1+\frac{t}{n})+\frac{1}{2c}(\frac{1}{t}+\eta^2M^2)][2\sum_{j\in[t]}\sigma_w^2(j)]+\frac{1}{2c}\sigma_w^2(t).\end{aligned}\tag{16}$$

Due to Assumption 4.2 where the initial loss has a bound c_3,

$$\begin{aligned}&\mathbb{E}[\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}))-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}))]\\ =&\mathbb{E}[\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}))[1-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}))/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}))]]\\ \le&c_3\mathbb{E}[1-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}))/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}))]\\ \le&c_3\frac{1-\mathbb{E}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})}{1-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{sig}});\mathcal{D}_{\text{sig}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{sig}});\mathcal{D}_{\text{sig}})}\\ =&c_3\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t).\end{aligned}\tag{17}$$

The last inequality here would not cost much when the signal training performs well (that is, a small $\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{sig}});\mathcal{D}_{\text{sig}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{sig}});\mathcal{D}_{\text{sig}})$).

Therefore, taking $c=\sqrt{\frac{(1+1/t+\eta^2M^2)n}{(1+t/n)4e\eta^2}}$, it holds that

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\\
\leq&\mathbb{E}[\frac{2eM\eta}{n}(1+\frac{t}{n})+\frac{2\sqrt{e}}{\sqrt{n}}\sqrt{(1+\frac{t}{n})(\frac{1}{t}+\eta^2M^2)}]2c_3\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)\\
&+[\frac{2eM^2\eta^2}{n}(1+\frac{t}{n})+\frac{2\sqrt{e}\eta}{\sqrt{n}}\sqrt{(1+\frac{t}{n})(\frac{1}{t}+\eta^2M^2)}]2\sum_{j\in[t]}\sigma_w^2(j)+\frac{\sqrt{e}\eta}{\sqrt{n}}\sqrt{\frac{1+t/n}{1/t+\eta^2M^2}}\sigma_w^2(t).
\end{aligned}
\tag{18}
$$

We consider the three parts separately:

For the first part, by setting $\eta\leq(1/\sqrt{t})$, we derive that

$$
\begin{aligned}
&\mathbb{E}[\frac{2eM\eta}{n}(1+\frac{t}{n})+\frac{2\sqrt{e}}{\sqrt{n}}\sqrt{(1+\frac{t}{n})(\frac{1}{t}+\eta^2M^2)}]2c_3\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)\\
\leq&4e\max\{M,1\}[\frac{1}{n\sqrt{t}}(1+t/n)+\frac{1}{\sqrt{n}}\sqrt{(1+\frac{t}{n})(1/t+1/t)}]2c_3\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)\\
\leq&4e\max\{M,1\}\frac{1}{\sqrt{nt}}[\frac{1}{\sqrt{n}}(1+\frac{t}{n})+\sqrt{(1+\frac{t}{n}}]c_3\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)\\
=&4e\max\{M,1\}\frac{1}{\sqrt{t}}[\frac{1}{n}(1+\frac{t}{n})+\sqrt{\frac{1}{n}(1+\frac{t}{n}})]c_3\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)\\
\leq&8e\max\{M,1\}\frac{1}{\sqrt{t}}\max\{\frac{1}{n}(1+\frac{t}{n}),\sqrt{\frac{1}{n}(1+\frac{t}{n})}\}c_3\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t).
\end{aligned}
\tag{19}
$$

For the second part, we derive similarly that

$$
\begin{aligned}
&[\frac{2eM\eta^2}{n}(1+\frac{t}{n})+\frac{2\sqrt{e}\eta}{\sqrt{n}}\sqrt{(1+\frac{t}{n})(\frac{1}{t}+\eta^2M^2)}]2\sum_{j\in[t]}\sigma_w^2(j)\\
\leq&8e\max\{M,1\}\frac{1}{\sqrt{n}}[\frac{1}{\sqrt{n}}(1+\frac{t}{n})+\sqrt{(1+\frac{t}{n}}]\frac{1}{t}\sum_{j\in[t]}\sigma_w^2(j)\\
=&8e\max\{M,1\}[\frac{1}{n}(1+\frac{t}{n})+\sqrt{\frac{1}{n}(1+\frac{t}{n}})]\frac{1}{t}\sum_{j\in[t]}\sigma_w^2(j).
\end{aligned}
\tag{20}
$$

If $t\leq n^2$, it holds that

$$
[\frac{1}{n}(1+\frac{t}{n})+\sqrt{\frac{1}{n}(1+\frac{t}{n}})]\frac{1}{t}\sum_{j\in[t]}\sigma_w^2(j)\leq 4\frac{1}{t}\sum_{j\in[t]}\sigma_w^2(j),
\tag{21}
$$

which goes to zero for bounded gradient norm.

If $t\geq n^2$, it holds that

$$
[\frac{1}{n}(1+\frac{t}{n})+\sqrt{\frac{1}{n}(1+\frac{t}{n}})]\frac{1}{t}\sum_{j\in[t]}\sigma_w^2(j)\leq 4\frac{1}{n^2}\sum_{j\in[t]}\sigma_w^2(j),
\tag{22}
$$

which goes to zero as long as $\sum_{j\in[t]}\sigma_w^2(j)=o(n^2)$.

For the third part, notice that

$$\begin{aligned}
&\frac{\sqrt{e}\eta}{\sqrt{n}}\sqrt{\frac{1+t/n}{1/t+\eta^2M^2}\sigma_w^2(t)} \\
=&\frac{\sqrt{e}}{\sqrt{n}}\sqrt{\frac{1+t/n}{1/(t\eta^2)+M^2}\sigma_w^2(t)} \\
\leq&\frac{\sqrt{e}}{\sqrt{n}}\sqrt{\frac{1+t/n}{1/(t\eta^2)+M^2}\frac{1}{t}\sum_{j\in[t]}\sigma_w^2(j)} \\
\leq&\frac{1}{M}\sqrt{e}\sqrt{\frac{1}{n}(1+\frac{t}{n})\frac{1}{t}\sum_{j\in[t]}\sigma_w^2(j)},
\end{aligned} \tag{23}$$

which also goes to zero as n goes to infinity, given that $\sum_{j\in[t]}\sigma_w^2(j)=o(n^2)$. Therefore, summarizing the above equations, we have that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}}) \\
\leq&8ec_3\max\{M,1\}\frac{1}{\sqrt{t}}\max\{\frac{1}{n}(1+\frac{t}{n}),\sqrt{\frac{1}{n}(1+\frac{t}{n})}\}\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)+\psi_3(n),
\end{aligned} \tag{24}$$

where $\psi_3(n)\to 0$ as $n\to\infty$.

Combining Equation equation 13, Equation equation 14, Equation equation 24 leads to the conclusion.

□

Lemma A.1 (Bounding excess risk via Generalization Gap)**.** *Let $\mathcal{L}(\boldsymbol{\theta};\mathcal{P})$ denote the population risk of $\boldsymbol{\theta}$ on distribution $\mathcal{P}$ and $\mathcal{L}_n(\boldsymbol{\theta};\mathcal{D})$ denote the empirical risk of $\boldsymbol{\theta}$ on dataset $\mathcal{D}$. Under the Assumptions in Theorem 4.4, we can bound the excess risk via generalization gap,*

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{noi}}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{noi});\mathcal{P}_{noi})\leq\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{noi}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{noi});\mathcal{P}_{noi})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{noi});\mathcal{D}_{noi})]. \tag{25}$$

Proof. Notice that the noise excess risk can be decomposed as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}}) \\
=&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}_{\text{noi}}}\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z})-\ell(\boldsymbol{\theta}^*_{\text{noi}};\boldsymbol{z}) \\
\triangleq&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}(\boldsymbol{\theta}^*_{\text{noi}};\mathcal{P}_{\text{noi}}) \\
=&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})] \\
&+\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})-\mathcal{L}_n(\boldsymbol{\theta}^*_{\text{noi}};\mathcal{D}_{\text{noi}})]+\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}_n(\boldsymbol{\theta}^*_{\text{noi}};\mathcal{D}_{\text{noi}})-\mathcal{L}(\boldsymbol{\theta}^*_{\text{noi}};\mathcal{P}_{\text{noi}})],
\end{aligned} \tag{26}$$

where $\boldsymbol{\theta}^*_{\text{noi}}$ denotes the parameter to minimize the excess risk on noise part. For the second term, note that $\mathcal{A}_0(\mathcal{D}_{\text{noi}})=\mathbf{0}$ and $\boldsymbol{\theta}^*_{\text{noi}}=\mathbf{0}$, and $\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})\leq 0$ by Assumption 4.2 , therefore,

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\boldsymbol{\theta}^*_{\text{noi}};\mathcal{P}_{\text{noi}})\leq 0. \tag{27}$$

Besides, notice that since $\boldsymbol{\theta}^*_{\text{noi}}$ is unrelated to the training set $\mathcal{D}_{\text{noi}}$, we have

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}_n(\boldsymbol{\theta}^*_{\text{noi}};\mathcal{D}_{\text{noi}})-\mathcal{L}(\boldsymbol{\theta}^*_{\text{noi}};\mathcal{P}_{\text{noi}})=0. \tag{28}$$

Therefore, we conclude that

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\leq\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})]. \tag{29}$$

□

Lemma A.2. *Under the assumptions in Theorem 4.4, we derive that for any $c > 0$, we have that*

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{noi}}\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{noi});\mathcal{P}_{noi}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{noi});\mathcal{D}_{noi})\\
\leq&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{noi}}[\frac{2e(M+c)\eta^2}{n}(1+\frac{t}{n}) + \frac{1}{2c}(\frac{1}{t}+\eta^2 M^2)]\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{noi});\boldsymbol{z}_i)\|^2 + \frac{1}{2c}\sigma_w^2(t).
\end{aligned} \tag{30}$$

Proof. Here we use the notion of on-average model stability proposed in Lei & Ying (2020), where we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})]\\
=&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}},\mathcal{D}_{\text{noi}}^{(i)}}\frac{1}{n}\sum_{i\in[n]}\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)});\boldsymbol{z}_i) - \ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i),
\end{aligned} \tag{31}$$

where $\mathcal{D}_{\text{noi}}^{(i)}$ denotes the dataset with only the i-th sample different from $\mathcal{D}_{\text{noi}}$. The above equation holds because $\mathcal{D}_{\text{noi}}^{(i)}$ does not contain any information of $\boldsymbol{z}_i$, and therefore is equal to the test loss in expectation. Due to smoothness assumption, we have that for any constant $c > 0$,

$$\begin{aligned}
&\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)});\boldsymbol{z}_i) - \ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\\
\leq&\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|\|\nabla\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\| + \frac{M}{2}\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2\\
\leq&\frac{c}{2}\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2 + \frac{1}{2c}\|\nabla\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2 + \frac{M}{2}\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2\\
=&[\frac{M+c}{2}]\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2 + \frac{1}{2c}\|\nabla\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2.
\end{aligned} \tag{32}$$

where the first inequality is due to smoothness, the second inequality is due to $2ab \leq ca^2 + c^{-1}b^2$,

We note that due to Lemma A.5, we have that for constant stepsize η

$$\mathbb{E}\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2 \leq 4e(\frac{1}{n}+\frac{t}{n^2})\eta^2\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2. \tag{33}$$

Besides, due to Lemma A.6, we have that

$$\frac{1}{n}\sum_{i\in[n]}\|\nabla\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2 \leq (\frac{1}{t}+\eta^2 M^2)\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2 + \sigma_w^2(t). \tag{34}$$

Therefore, we have that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}},\mathcal{D}_{\text{noi}}^{(i)}}\frac{1}{n}\sum_{i\in[n]}\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)});\boldsymbol{z}_i) - \ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\\
\leq&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}},\mathcal{D}_{\text{noi}}^{(i)}}\frac{1}{n}\sum_{i\in[n]}[\frac{M+c}{2}]\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2 + \frac{1}{2c}\|\nabla\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2\\
\leq&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}},\mathcal{D}_{\text{noi}}^{(i)}}\frac{1}{n}\sum_{i\in[n]}[\frac{M+c}{2}]4e(\frac{1}{n}+\frac{t}{n^2})\eta^2\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2\\
&+\frac{1}{2c}(\frac{1}{t}+\eta^2 M)\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2 + \frac{1}{2c}\sigma_w^2(t)\\
=&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}},\mathcal{D}_{\text{noi}}^{(i)}}[\frac{2e(M+c)\eta^2}{n}(1+\frac{t}{n}) + \frac{1}{2c}(\frac{1}{t}+\eta^2 M^2)]\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2 + \frac{1}{2c}\sigma_w^2(t)
\end{aligned} \tag{35}$$

□

Lemma A.3 (Bounding Cumulative Gradient)**.** *Assuming that $\mathcal{L}_n(\cdot;\boldsymbol{z})$ is M-smooth, if the training stepsize $\eta < 1/M$ (constant stepsize), it holds that*

$$\mathbb{E}\frac{1}{n}\sum_{i\in[n]}\sum_{[j\in[t]]}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{noi}),z_i)\|^2 \le \frac{2}{\eta}\mathbb{E}[\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{noi}))-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{noi}))]+2\sum_{j\in[t]}\sigma_w^2(j). \tag{36}$$

where $\sigma_w^2(j)$ denotes the variance in gradient at step j.

Proof of Lemma A.3. Due to the smoothness assumption on the empirical loss (it could be done by the smoothness assumption on each sample), we have that for all i,

$$\begin{aligned}\mathbb{E}\mathcal{L}_n(\mathcal{A}_{i+1}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}}) \le& \mathcal{L}_n(\mathcal{A}_i(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})+\mathbb{E}(\mathcal{A}_{i+1}(\mathcal{D}_{\text{noi}})-\mathcal{A}_i(\mathcal{D}_{\text{noi}}))^\top\nabla\mathcal{L}(\mathcal{A}_i(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\\ &+\mathbb{E}(\frac{M}{2}\|\mathcal{A}_{i+1}(\mathcal{D}_{\text{noi}})-\mathcal{A}_i(\mathcal{D}_{\text{noi}})\|^2),\end{aligned} \tag{37}$$

where the expectation is taken over the randomness on the gradient. Plugging in the iteration $\mathcal{A}_{i+1}(\mathcal{D}_{\text{noi}}) = \mathcal{A}_i(\mathcal{D}_{\text{noi}}) + \eta_i\nabla\ell(\mathcal{A}_i(\mathcal{D}_{\text{noi}}),\boldsymbol{z}_{[i]})$, where $\boldsymbol{z}_{[i]}$ denotes the chosen sample, we have

$$\mathbb{E}[\mathcal{L}_n(\mathcal{A}_{i+1}(\mathcal{D}_{\text{noi}}))] \le \mathcal{L}_n(\mathcal{A}_i(\mathcal{D}_{\text{noi}}))-\eta\|\nabla\mathcal{L}_n(\mathcal{A}_i(\mathcal{D}_{\text{noi}}))\|^2+\mathbb{E}(\frac{M}{2}\eta^2\|\nabla\ell(\mathcal{A}_i(\mathcal{D}_{\text{noi}}),z_{[i]})\|^2). \tag{38}$$

Due to the definition of variance that $\sigma_w^2 = \mathbb{E}\|\nabla\ell(\mathcal{A}_i(\mathcal{D}_{\text{noi}}),z_{[i]})\|^2 - \|\nabla\mathcal{L}_n(\mathcal{A}_i(\mathcal{D}_{\text{noi}}))\|^2$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}_n(\mathcal{A}_{i+1}(\mathcal{D}_{\text{noi}}))] \le& \mathcal{L}_n(\mathcal{A}_i(\mathcal{D}_{\text{noi}}))-\eta\|\nabla\ell(\mathcal{A}_i(\mathcal{D}_{\text{noi}}),z_{[i]})\|^2+\eta\sigma_w^2+\mathbb{E}(\frac{M}{2}\eta^2\|\nabla\ell(\mathcal{A}_i(\mathcal{D}_{\text{noi}}),z_{[i]})\|^2).\\ \le& \mathcal{L}_n(\mathcal{A}_i(\mathcal{D}_{\text{noi}}))+\eta\sigma_w^2+(-\eta+\frac{M}{2}\eta^2)\mathbb{E}\|\nabla\ell(\mathcal{A}_i(\mathcal{D}_{\text{noi}}),z_{[i]})\|^2\\ \le& \mathcal{L}_n(\mathcal{A}_i(\mathcal{D}_{\text{noi}}))+\eta\sigma_w^2-\frac{\eta}{2}\mathbb{E}\|\nabla\ell(\mathcal{A}_i(\mathcal{D}_{\text{noi}}),z_{[i]})\|^2,\end{aligned} \tag{39}$$

where the last equation is due to $\eta < 1/M$. By telescoping and taking expectation, we rewrite it as

$$\mathbb{E}\eta\sum_{[j\in[t]]}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}),z_{[j]})\|^2 \le 2\mathbb{E}[\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}))-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}))]+2\sum_{j\in[t]}\eta\sigma_w^2(j). \tag{40}$$

Since each sample is sampled uniformly with probability $1/n$, taking expectation leads to

$$\mathbb{E}\sum_t\|\nabla\ell(\mathcal{A}_i(\mathcal{D}_{\text{noi}}),z_{[i]})\|^2 = \frac{1}{n}\sum_{i\in[n]}\sum_{[j\in[t]]}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}),z_i)\|^2. \tag{41}$$

Therefore, we have

$$\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}),z_i)\|^2 \le \frac{2}{\eta}\mathbb{E}[\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}))-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}))]+2\sum_{j\in[t]}\sigma_w^2(j). \tag{42}$$

□

Lemma A.4 (One-expansion under convexity, from Hardt et al. (2016) (arxiv version), Lemma 3.7 (argument 2))**.** *Assume that for all $\boldsymbol{z}$, the function $\ell(\boldsymbol{z};w)$ is convex with respect to w and M-smooth, then for step size $\eta < 2/M$ we have that when not choosing the sample $\boldsymbol{z}_i$,*

$$\|\mathcal{A}_{j+1}(\mathcal{D}_{noi}^{(i)})-\mathcal{A}_{j+1}(\mathcal{D}_{noi})\| \le \|\mathcal{A}_j(\mathcal{D}_{noi}^{(i)})-\mathcal{A}_j(\mathcal{D}_{noi})\|. \tag{43}$$

Lemma A.5 (Bound for stability parameter difference)**.** *Under the Assumptions in Theorem 4.4, We have that for any i*

$$\mathbb{E}\|\mathcal{A}_t(\mathcal{D}_{noi}^{(i)})-\mathcal{A}_t(\mathcal{D}_{noi})\|^2 \le 4e(\frac{1}{n}+\frac{t}{n^2})\sum_{j\in[t]}\eta_j^2\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{noi});\boldsymbol{z}_i)\|^2. \tag{44}$$

where i denotes the sample index and j denotes the time index.

Proof of Lemma A.5. The proof is partly inspired by the proof of Lemma C.2 in Lei & Ying (2020).

Note that for any step j, if the chosen index is i, we have that

$$\begin{aligned}
&\|\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}})\| \\
=&\|\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}}) - \eta_t \nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}); \tilde{\boldsymbol{z}}_i) + \eta_t \nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|
\end{aligned} \tag{45}$$

Therefore, due to the inequality $(a+b)^2 \leq (1+p)a^2 + (1+1/p)b^2$ for any $p > 0$, we have that

$$\begin{aligned}
&\|\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}})\|^2 \\
\leq&(1+p)\|\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}})\|^2 + \eta_t^2(1+1/p)\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}); \tilde{\boldsymbol{z}}_i) - \nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2 \\
\leq&(1+p)\|\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}})\|^2 + 2\eta_t^2(1+1/p)[\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}); \tilde{\boldsymbol{z}}_i)\|^2 + \|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2],
\end{aligned} \tag{46}$$

for any $p > 0$.

If the chosen index is not i, due to the convexity of the loss , according to Lemma A.4, we have that

$$\|\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}})\|^2 \leq \|\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}})\|^2. \tag{47}$$

Therefore, since each index is chosen uniformly, we have that

$$\begin{aligned}
&\mathbb{E}\|\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}})\|^2 \\
\leq&\frac{1}{n}[(1+p)\mathbb{E}\|\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}})\|^2 + 2\eta_t^2(1+1/p)\mathbb{E}[\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}); \tilde{\boldsymbol{z}}_i)\|^2 + \|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2]] \\
&+ \frac{n-1}{n}\mathbb{E}\|\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}})\|^2 \\
=&(1+\frac{p}{n})\mathbb{E}\|\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}})\|^2 + 2\frac{\eta_t^2}{n}(1+1/p)\mathbb{E}[\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}); \tilde{\boldsymbol{z}}_i)\|^2 + \|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2] \\
=&(1+\frac{p}{n})\mathbb{E}\|\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}})\|^2 + 4\frac{\eta_t^2}{n}(1+1/p)\mathbb{E}\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2.
\end{aligned} \tag{48}$$

where the expectation is taken over the algorithm for the last step, and the dataset $\mathcal{D}_{\text{noi}}, \mathcal{D}_{\text{noi}}^{(i)}$. We use the fact that $\mathbb{E}\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2 = \mathbb{E}\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}^{(i)}); \tilde{\boldsymbol{z}}_i)\|^2$. By iteration, we have that

$$\begin{aligned}
&\mathbb{E}\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2 \\
\leq&\frac{4(1+p^{-1})}{n}\sum_{j\in[t]} \eta_j^2 (1+p/n)^{t-j}\mathbb{E}\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2.
\end{aligned} \tag{49}$$

By choosing $p = n/t$, we have that

$$(1+p/n)^{t-j} \leq (1+p/n)^t = (1+1/t)^t \leq e.$$

Therefore, we have that

$$\mathbb{E}\|\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(i)}) - \mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2 \leq \frac{4e(1+t/n)}{n}\sum_{j\in[t]} \eta_j^2 \mathbb{E}\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2. \tag{50}$$

□

Lemma A.6 (Bound for the last iterate gradient)**.**

$$\frac{1}{n}\sum_{i\in[n]} \|\nabla \ell(\mathcal{A}_t(\mathcal{D}_{noi}); \boldsymbol{z}_i)\|^2 \leq (\frac{1}{t} + \eta^2 M)\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]} \mathbb{E}\|\nabla \ell(\mathcal{A}_j(\mathcal{D}_{noi}); \boldsymbol{z}_i)\|^2 + \sigma_w^2(t). \tag{51}$$

Proof. We first notice that there exists ξ such that

$$\begin{aligned}
&\nabla\mathcal{L}_n(\mathcal{A}_{t+1}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\\
=&\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})-\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})[\mathcal{A}_{t+1}(\mathcal{D}_{\text{noi}})-\mathcal{A}_{t}(\mathcal{D}_{\text{noi}})]\\
=&\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})-\eta\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})[\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_t)]
\end{aligned} \tag{52}$$

Therefore, we have that

$$\begin{aligned}
&\|\nabla\mathcal{L}_n(\mathcal{A}_{t+1}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2\\
=&\|\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})-\eta\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})[\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_t)]\|^2\\
=&\|\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2-2\eta\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_t)\\
&+\eta^2\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_t)\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_t).
\end{aligned} \tag{53}$$

By taking expectation on the chosen sample $\boldsymbol{z}_t$, we have that

$$\begin{aligned}
&\mathbb{E}\|\nabla\mathcal{L}_n(\mathcal{A}_{t+1}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2\\
=&\|\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2-2\eta\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\\
&+\eta^2\mathbb{E}\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_t)\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})\nabla^2\mathcal{L}_n(\xi;\mathcal{D}_{\text{noi}})\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_t)\\
\le&\|\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2+\eta^2M^2\mathbb{E}\|\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_t)\|^2\\
=&\|\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2+\eta^2M^2\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\|\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});z_i)\|^2
\end{aligned} \tag{54}$$

By iteration, we have that

$$\mathbb{E}\|\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2\le\eta^2M^2\frac{1}{n}\sum_{i\in[n]}\sum_{j=k}^{t}\mathbb{E}\|\nabla\ell(\mathcal{A}_{j}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2+\|\nabla\mathcal{L}_n(\mathcal{A}_{k}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2. \tag{55}$$

The above equation indeed holds for any iteration k, and therefore by taking an average over all iterations, we have that

$$\begin{aligned}
&\mathbb{E}\|\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2\\
\le&\eta^2M^2\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_{j}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2+\frac{1}{t}\sum_{j\in[t]}\|\nabla\mathcal{L}_n(\mathcal{A}_{j}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2\\
\le&\eta^2M^2\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_{j}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2+\frac{1}{t}\sum_{j\in[t]}\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\|\nabla\ell(\mathcal{A}_{j}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2\\
=&(\frac{1}{t}+\eta^2M^2)\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_{j}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2
\end{aligned} \tag{56}$$

Therefore, we have that

$$\begin{aligned}
&\mathbb{E}\frac{1}{n}\sum_{i\in[n]}\|\nabla\ell(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2\\
=&\mathbb{E}\|\nabla\mathcal{L}_n(\mathcal{A}_{t}(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\|^2+\sigma_w^2(t)\\
\le&(\frac{1}{t}+\eta^2M^2)\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_{j}(\mathcal{D}_{\text{noi}});\boldsymbol{z}_i)\|^2+\sigma_w^2(t).
\end{aligned} \tag{57}$$

□

B. Proof of Theorem 4.6

Theorem 4.6 (Smooth, with SGLD)**.** *Assume that the loss $\ell(\boldsymbol{\theta}; \boldsymbol{z})$ is $O(1)$-bounded, L-Lipschitz, and M-smooth with respect to $\boldsymbol{\theta}$ for any sample $\boldsymbol{z}$. Consider SGLD with noise scale σ and stepsize $\eta < \min\{\frac{\sigma}{20L}, \frac{1}{M}\}$. Under Assumption 4.1 and Assumption 4.2, if $p = o(\frac{n^2}{\eta t})$, the following inequality holds when $\sum_{j\in[t]} \sigma_w^2(j; \mathcal{D}_{noi}) = o(\frac{n^2\sigma^2}{\eta^2})$ and $t \geq \max\{T_1, T_2, T_3\}$,*

$$\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{P}) \leq c' \frac{\sqrt{\eta}}{n\sigma}\sqrt{\mathcal{T}_n^{\alpha}(\mathcal{D}, \mathcal{A}_t)} + \psi'(n), \tag{11}$$

where the term $\psi'(n) \to 0$ as $n \to \infty$. The constant $c' > 0$ denotes a constant related to the constant c_1, c_2, c_3, M in Assumption 4.1 and Assumption 4.2.

Proof of Theorem 4.6. Similar to the proofs of Theorem 4.4, under Assumption 4.1 and Assumption 4.2, the core of the proof is the bound for the generalization gap $\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{P}_{\text{noi}}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})]$. We next only focus on the proof of the generalization gap.

Firstly, as shown in Li et al. (2020) (Theorem 11), if the loss $\ell(\theta; \boldsymbol{z})$ is C-bounded, L-Lipschitz, and the learning rate $\eta < \sigma/20L$, the generalization bound can be bounded using the cumulative gradient norm:

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{P}_{\text{noi}}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})] \leq \frac{8.12C}{n}\frac{\eta}{\sigma}\sqrt{\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\sum_{j\in[t]}\frac{1}{n}\sum_{i\in[n]}\|\nabla\ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}}), \boldsymbol{z}_i)\|^2]}. \tag{58}$$

Besides, we derive by Lemma B.1 that:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{P}_{\text{noi}}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})]\\
\leq&\frac{8.12C}{n}\frac{\eta}{\sigma}\sqrt{\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\sum_{j\in[t]}\frac{1}{n}\sum_{i\in[n]}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), \boldsymbol{z}_i)\|^2]}\\
\leq&\frac{8.12C}{n}\frac{\eta}{\sigma}\sqrt{\frac{2}{\eta}\mathbb{E}\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}})) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}})) + 2\sum_{j\in[t]}\sigma_w^2(j) + \frac{pt\sigma^2}{\eta}}.
\end{aligned} \tag{59}$$

Notice that if $\sum_{j\in[t]} \sigma_{\boldsymbol{w}}^2(j) = o(\frac{n^2\sigma^2}{\eta^2})$ and $p = o(\frac{n^2}{t\eta})$, the final several terms becomes o(1), and therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{P}_{\text{noi}}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})]\\
\leq&\frac{8.12\sqrt{c_3}\sqrt{2}C}{n}\frac{\sqrt{\eta}}{\sigma}\sqrt{\mathbb{E}\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}})) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}))} + \psi'(n)\\
\leq&\frac{8.12\sqrt{2}C}{n}\frac{\sqrt{\eta}}{\sigma}\sqrt{\mathcal{T}_n^{\alpha}(\mathcal{D}, \mathcal{A}_t)} + \psi'(n).
\end{aligned} \tag{60}$$

The results of Theorem 4.6 directly holds by adjusting the constant term to c_3'.

□

The following Lemma B.1 bounds the cumulative gradient, inspired by the proofs in Lemma A.3. Compared to the results in Lemma A.3, the additional term $\frac{pt\sigma^2}{\eta}$ comes from the Gaussian noise in SGLD.

Lemma B.1. *Under the Assumptions in Theorem 4.6, it holds that in SGLD, if $\eta < 1/M$,*

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{noi}}\sum_{j\in[t]}\frac{1}{n}\sum_{i\in[n]}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{noi}), \boldsymbol{z}_i)\|^2 \leq \frac{2}{\eta}\mathbb{E}\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{noi})) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{noi})) + 2\sum_{j\in[t]}\sigma_w^2(j) + \frac{pt\sigma^2}{\eta}. \tag{61}$$

Proof of Lemma B.1. The proof is similar to the proof of Lemma A.3. Firstly, due to the M-smoothness conditions, it holds for all i that

$$\begin{aligned}
\mathbb{E}\mathcal{L}_n(\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) \leq&\mathcal{L}_n(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) + \mathbb{E}(\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}}))^\top\nabla\mathcal{L}(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})\\
&+\mathbb{E}(\frac{M}{2}\|\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}) - \mathcal{A}_j(\mathcal{D}_{\text{noi}})\|^2),
\end{aligned} \tag{62}$$

where the expectation is taken over the randomness on the gradient. Plugging in the iteration $\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}) = \mathcal{A}_j(\mathcal{D}_{\text{noi}}) + \eta\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), \boldsymbol{z}_{[j]}) + \frac{\sigma}{\sqrt{2}}\boldsymbol{n}_t$, where $\boldsymbol{z}_{[j]}$ denotes the chosen sample at iteration j, we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}_n(\mathcal{A}_{j+1}(\mathcal{D}_{\text{noi}}))] &\leq \mathcal{L}_n(\mathcal{A}_j(\mathcal{D}_{\text{noi}})) - \eta\|\nabla\mathcal{L}_n(\mathcal{A}_j(\mathcal{D}_{\text{noi}}))\|^2 + \mathbb{E}(\frac{M}{2}\eta^2\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), z_{[j]})\|^2) + \frac{\sigma^2}{2}p \\ &= \mathcal{L}_n(\mathcal{A}_j(\mathcal{D}_{\text{noi}})) - \eta(\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), z_{[j]})\|^2 - \sigma_w^2(j) + \mathbb{E}(\frac{M}{2}\eta^2\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), z_{[j]})\|^2) + \frac{\sigma^2}{2}p,\end{aligned} \tag{63}$$

that is to say, by telescoping and taking expectation, it holds that

$$\eta(1-\frac{M}{2}\eta)\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), \boldsymbol{z}_{[j]})\|^2 \leq \mathbb{E}\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}})) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}})) + \eta\sum_{j\in[t]}\sigma_w^2(j) + \frac{pt\sigma^2}{2}. \tag{64}$$

Besides, since $\eta < 1/M$, we have that

$$\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), \boldsymbol{z}_{[j]})\|^2 \leq \frac{2}{\eta}\mathbb{E}\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}})) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}})) + 2\sum_{j\in[t]}\sigma_w^2(j) + \frac{pt\sigma^2}{\eta}. \tag{65}$$

Note that the expectation here includes the randomness of the algorithm and the dataset here, by plugging into the randomness of the chosen sample $\boldsymbol{z}_{[j]}$,

$$\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), \boldsymbol{z}_{[j]})\|^2 = \mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\sum_{j\in[t]}\frac{1}{n}\sum_{i\in[n]}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}), \boldsymbol{z}_i)\|^2. \tag{66}$$

□

C. Proof of Theorem 4.7

Theorem 4.7 (Overparameterized Linear Regression with GD)**.** *Under overparameterized linear regression regimes, assume that $r_0(\Sigma) = o(n)$ and $k^* = o(n)$. Besides, assume that $\|\boldsymbol{\theta}^*\|_2 = O(1)$, $\|\Sigma_{\boldsymbol{x}}\|_2 = O(1)$ in a constant scale. We consider the GD training process with zero initialization and constant stepsize η. For any given $\delta > 0$ which does not vary with sample size n and satisfies $\log(1/\delta) = o(n)$, for $t = \omega(1)$[3], with probability at least $1-\delta$,*

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{D}) \leq c\log(1/\delta)\sigma_y^2\mathcal{T}_n^\alpha(\mathcal{D}, \mathcal{A}_t) + \tilde{\psi}(n), \tag{12}$$

where $\tilde{\psi}(n) \to 0$ as $n \to \infty$ and $c > 0$ denotes a constant.

Proof. Due to Lemma C.2, we derive that

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{D}) \leq 2\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{sig}}); \mathcal{D}_{\text{sig}}) + 2\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}). \tag{67}$$

According to Lemma C.1, since $t = \omega(1)$, $r_0(\Sigma_{\boldsymbol{x}}) = o(n)$ and $\log(1/\delta) = o(n)$, we have that

$$\lim_{n\to\infty}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{sig}}); \mathcal{D}_{\text{sig}}) = 0. \tag{68}$$

Besides, due to Lemma C.3, we have that

$$\lim_{n\to\infty}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) \leq c_1\log(1/\delta)\sigma_y^2\mathcal{T}_n^\alpha(\mathcal{D}, \mathcal{A}_t), \tag{69}$$

where we use the assumption that $k^* = o(n)$, and δ is unrelated to n. Therefore, we summarize the results as

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{D}) \leq c\log(1/\delta)\sigma_y^2\mathcal{T}_n^\alpha(\mathcal{D}, \mathcal{A}_t) + \tilde{\psi}(n), \tag{70}$$

where $\tilde{\psi}(n) \to 0$ as $n \to \infty$. □

[3] The statement $t = \omega(1)$ means that $t \to \infty$ as $n \to \infty$.

Lemma C.1 (Bound for signal component, Lemma (A.7) in Xu et al. (2022))**.** *Under the overparameterized linear regression regimes,*

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{sig}); \mathcal{D}_{sig}) \leq c\|\theta^*\|^2 \left(\frac{1}{\lambda t} + \|\Sigma_{\boldsymbol{x}}\| \max\{ \sqrt{\frac{r_0(\Sigma_{\boldsymbol{x}})}{n}}, \frac{r_0(\Sigma_{\boldsymbol{x}})}{n}, \sqrt{\frac{\log(1/\delta)}{n}}, \frac{\log(1/\delta)}{n} \} \right). \tag{71}$$

Lemma C.2 (Decomposition lemma, Lemma 18 in Bartlett et al. (2020))**.** *In overparameterized linear regression regimes, we have that*

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{D}) \leq 2\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{sig}); \mathcal{D}_{sig}) + 2\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{noi}); \mathcal{D}_{noi}). \tag{72}$$

Proof. Due to the iteration of GD which is linear in y, we have that

$$\mathcal{A}_t(\mathcal{D}) = \mathcal{A}_t(\mathcal{D}_{\text{sig}}) + \mathcal{A}_t(\mathcal{D}_{\text{noi}}). \tag{73}$$

Note that

$$\begin{aligned} \mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{D}) &= \|\mathcal{A}_t(\mathcal{D}) - \theta^*\|^2_{\Sigma_{\boldsymbol{x}}}, \\ \mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{sig}}); \mathcal{D}_{\text{sig}}) &= \|\mathcal{A}_t(\mathcal{D}_{\text{sig}}) - \theta^*\|^2_{\Sigma_{\boldsymbol{x}}}, \\ \mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) &= \|\mathcal{A}_t(\mathcal{D}_{\text{noi}})\|^2_{\Sigma_{\boldsymbol{x}}}. \end{aligned} \tag{74}$$

Therefore, due to the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we have that

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{D}) \leq 2\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{sig}}); \mathcal{D}_{\text{sig}}) + 2\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}). \tag{75}$$

□

Lemma C.3 (Bound for noise component)**.** *Under the assumptions in Theorem 4.7, we have that with probability at least* $1-\delta$

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{noi}); \mathcal{D}_{noi}) \leq c\log(1/\delta)\sigma_y^2 \mathcal{T}_n^{\alpha}(\mathcal{D}, \mathcal{A}_t) + c\log(1/\delta)\sigma_y^2 \frac{k^*}{n}, \tag{76}$$

for a given constant $c > 0$ *which is related to* $\log(1/\delta)$.

Proof. For the noise component, we first notice that from Lemma C.1 in Teng et al. (2022), we have that

$$\mathcal{A}_t(\mathcal{D}_{\text{sig}}) = X^\top [XX^\top]^{-1} [I - [I - \frac{\lambda}{n} XX^\top]^t][Y - X\beta^*]. \tag{77}$$

Therefore, due to the subGaussian assumption on $Y - X\beta^*$, we have that (we refer to Lemma 7 in Bartlett et al. (2020))

$$\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) \leq c\sigma_y^2 \log(1/\delta) \text{Tr}[C], \tag{78}$$

where $C = [[I - [I - \frac{\lambda}{n} XX^\top]^t]^2 [XX^\top]^{-1} X\Sigma_{\boldsymbol{x}} X^\top [XX^\top]^{-1}]$.

By denoting $\boldsymbol{z}_i = Xv_i/\sqrt{\lambda_i}$, where λ_i, v_i denotes the i-th eigenvalue and the corresponding eigenvector of matrix $\Sigma_{\boldsymbol{x}}$, we have that $X\Sigma_{\boldsymbol{x}}X = \sum_i \lambda_i^2 \boldsymbol{z}_i \boldsymbol{z}_i^\top$

$$\text{Tr}C = \text{Tr} \sum_i \lambda_i^2 [I - [I - \frac{\lambda}{n} XX^\top]^t]^2 [XX^\top]^{-1} \boldsymbol{z}_i \boldsymbol{z}_i^\top [XX^\top]^{-1}. \tag{79}$$

We split the summation operator into two parts by $k^* = \min\{k \geq 0, r_k(\Sigma_{\boldsymbol{x}}) \geq bn\}$. For the first part

$$\begin{aligned}
&\mathrm{Tr}\sum_{i\leq k^*}\lambda_i^2[I-[I-\frac{\lambda}{n}XX^\top]^t]^2[XX^\top]^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top[XX^\top]^{-1}\\
=&\sum_{i\leq k^*}\lambda_i^2\mathrm{Tr}[I-[I-\frac{\lambda}{n}XX^\top]^t]^2[XX^\top]^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top[XX^\top]^{-1}\\
\leq&\sum_{i\leq k^*}\lambda_i^2\mathrm{Tr}[XX^\top]^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top[XX^\top]^{-1}\\
=&\sum_{i\leq k^*}\mathrm{Tr}\lambda_i^2[XX^\top]^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top[XX^\top]^{-1}\\
=&\sum_{i\leq k^*}\lambda_i^2 z_i^\top[XX^\top]^{-2}z_i\\
\leq&\frac{k^*}{n},
\end{aligned}\tag{80}$$

where the first inequality comes from the fact that $\mathrm{Tr}AB \geq \mathrm{Tr}AC$ if A and $B-C$ are both positive semi-definite. The second inequality comes from Lemma 11 in Bartlett et al. (2020), given that $\log(1/\delta) = o(n)$.

Before considering the remaining part, we first notice that when $i > k^*$, we have that $\lambda_i \leq \frac{1}{bn}\sum_{j>i}\lambda_j$

$$\begin{aligned}
&\sum_{i>k^*}\lambda_i^2\boldsymbol{z}_i\boldsymbol{z}_i^\top\\
\leq&\sum_{i>k^*}[\frac{1}{bn}\sum_{j>i}\lambda_j]\lambda_i\boldsymbol{z}_i\boldsymbol{z}_i^\top\\
\leq&[\frac{1}{bn}\sum_{j>k^*}\lambda_j]\sum_{i>k^*}\lambda_i\boldsymbol{z}_i\boldsymbol{z}_i^\top\\
=&[\frac{1}{bn}\sum_{j>k^*}\lambda_j]XX^\top.
\end{aligned}\tag{81}$$

Therefore, for the remaining part, we have that

$$\begin{aligned}
&\mathrm{Tr}\sum_{i>k^*}\lambda_i^2[I-[I-\frac{\lambda}{n}XX^\top]^t]^2[XX^\top]^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^\top[XX^\top]^{-1}\\
=&\mathrm{Tr}[XX^\top]^{-1}[I-[I-\frac{\lambda}{n}XX^\top]^t]^2[XX^\top]^{-1}\sum_{i>k^*}\lambda_i^2\boldsymbol{z}_i\boldsymbol{z}_i^\top\\
\leq&\mathrm{Tr}[XX^\top]^{-1}[I-[I-\frac{\lambda}{n}XX^\top]^t]^2[XX^\top]^{-1}[\frac{1}{bn}\sum_{j>k^*}\lambda_j]XX^\top\\
=&\mathrm{Tr}[\frac{1}{bn}\sum_{j>k^*}\lambda_j][XX^\top]^{-1}[I-[I-\frac{\lambda}{n}XX^\top]^t]^2\\
\leq&\frac{c_1}{bn}\mathrm{Tr}[I-[I-\frac{\lambda}{n}XX^\top]^t]^2.
\end{aligned}\tag{82}$$

The last inequality uses the fact that $XX^\top \geq \frac{1}{c_1}\sum_{j>k^*}\lambda_j$ for a given constant c_1 (see Lemma 10 in Bartlett et al. (2020)). Besides, notice that since $I - [I - \frac{\lambda}{n}XX^\top]^t$ is positive semi-definite, we have that

$$\frac{1}{n}\mathrm{Tr}[I-[I-\frac{\lambda}{n}XX^\top]^t]^2 \leq \frac{1}{n}\mathrm{Tr}[I-[I-\frac{\lambda}{n}XX^\top]^{2t}].\tag{83}$$

Besides, we notice that with high probability (concentration on $y - x^\top \theta^*$), we have that there exists constant c_2 such that

$$\begin{aligned}&\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) \leq (1+c_2)\sigma_y^2 \text{Tr}[I - [I - \frac{\lambda}{n} XX^\top]^{2t}, \\ &\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) \geq (1-c_2)\sigma_y^2 > 0.\end{aligned} \tag{84}$$

where we abuse the notation c as a constant different from the above text. Therefore, with high probability, we have that there exists constant c_3, such that

$$c_3 \text{Tr}[I - [I - \frac{\lambda}{n} XX^\top]^{2t}] \leq 1 - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) / \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}). \tag{85}$$

Therefore, we have that

$$\begin{aligned}&\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) \\ \leq& c\sigma_y^2 \log(1/\delta)[\frac{k^*}{n} + 1 - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) / \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})] \\ \leq& c\sigma_y^2 \log(1/\delta)\frac{k^*}{n} + c\log(1/\delta)\sigma_y^2 \mathcal{T}_n^\alpha(\mathcal{D}, \mathcal{A}_t).\end{aligned} \tag{86}$$

for a constant probability δ, where we abuse the notation c as a constant independent of the data distribution and time t. □

D. Detailed Experiment Setting

Our experiments contain two parts: Firstly, we conduct experiments on CIFAR-10 and CIFAR-100 with different hyperparameter settings to show the correlation between test accuracy and REF Complexity. More specifically, we use ResNets as the basic architecture, and evaluate the test accuracy with different learning rates, batch sizes, weight decay, and depths. We train each model for 150 epochs. To evaluate REF Complexity correctly, each noise training process is trained five times, and we calculate the averaged REF Complexity as the metric. The results are shown in Figure 1. Here we list all the hyperparameters in Table 4.

Table 4. Hyperparameters

hyperparameter	value
learning rate	0.1, 0.01, 0.001
batch size	256, 512, 1024
weight decay	1e-5, 1e-6, 1e-7
architecture	ResNet20, ResNet32, ResNet44, ResNet56

Secondly, we conduct experiments on CIFAR-10 to compare our REF Complexity and other relevant generalization measures. We train ResNet-32 for 150 epochs using SGD with a weight decaying of 1e-5. The baseline algorithms can be categorized into four classes as shown in Table 1. More specifically, the baselines include L2-norm of the final model (L2), L2-distance from initialization (L2-DIST), Frobenius norm of the model (F-NORM) (Jiang et al., 2020b), inverse of the margin of the logits between the labels (INV-MARGIN) (Dziugaite et al., 2020), spectral norm of the model (SPECTRAL) (Pitas et al., 2018; Bartlett et al., 2017), sum of spectral norm over margin (SPECTRAL/MARGIN) (Jiang et al., 2020b), path-norm of the model (PATH-NORM) (Neyshabur et al., 2015), PAC-Bayesian bounds using the origin and initialization as reference tensors (PB-I and PB-O), PAC-Bayesian flatness (PB-FLATNESS) PAC-Bayesian Magnitude-aware Perturbation Bounds (PB-M-I, PB-M-O) and Magnitude-aware PAC-Bayesian flatness (PB-M-FLATNESS) (Keskar et al., 2017; Neyshabur et al., 2017; Jiang et al., 2020b), number of iterations required to reach cross-entropy equals 1.0/1.5 (STEP(1), STEP(1.5)) (Jiang et al., 2020b). Additionally, we also report the results of F-distance from initialization (F-DIST), and path norm over margin (PATH-NORM/MARGIN), which are omitted in the main text due to space limitations. Again, we evaluate the correlation between generalization measures and test accuracy. Same as above, we searched hyperparameters on batch size, learning rate and drouout rate. We repeated the experiment three times for every configuration, and the noise training process of our algorithm runs three times. Full mean and standard deviation results are reported in Table 5. The full correlation between generalization measures and test accuracy is summarized in Table 6. We run all the experiments on a RTX2080Ti graphic card.

Table 5. Detailed experiment results of REF Complexity and other baseline metrics.

batch size	learning rate	dropout	test accuracy	L2-NORM	L2-DIST	F-NORM	F-DIST	INV-MARGIN	SPECTRAL	SPECTRAL/MARGIN	PATH-NORM	PATH-NORM/MARGIN	PB-I	PB-O	PB-FLATNESS	PB-M-I	PB-M-O	PB-M-FLATNESS	STEPS(1)	STEPS(1.5)	REF (W/O NOI)	REF (W/O INIT)	REF COMPLEXITY
1024	0.01	0	0.783±0.005	0.029±0	0.02±0	0.194±0.001	0.09±0.001	0.004±0	11.463±0.195	11.379±0.194	6.377±0.43	5.868±0.495	0.553±0.025	1.194±0.055	0.055±0.003	2.984±0.042	3.533±0.033	0.014±0.001	0.014±0.001	0.004±0	0.221±0.009	0.223±0.009	0.012±0.004
2048	0.01	0	0.716±0.016	0.03±0	0.019±0	0.199±0.001	0.083±0	0.003±0	12.673±0.215	12.319±0.171	11.374±1.106	7.955±0.379	0.444±0.047	1.062±0.117	0.048±0.005	2.974±0.032	3.591±0.028	0.013±0.001	0.012±0	0.003±0	0.319±0.017	0.321±0.021	0.016±0.01
256	0.005	0	0.822±0.014	0.029±0	0.021±0	0.189±0.001	0.1±0.001	0.007±0.001	10.214±0.07	10.597±0.207	3.532±0.276	5.232±0.89	0.663±0.122	1.247±0.228	0.059±0.011	3.047±0.086	3.496±0.075	0.016±0.002	0.027±0	0.008±0	0.158±0.017	0.159±0.018	0.009±0.007
256	0.01	0	0.835±0.015	0.029±0	0.023±0	0.182±0.001	0.115±0.001	0.007±0.002	8.936±0.158	9.386±0.331	1.508±0.102	2.423±0.578	1.002±0.43	1.588±0.68	0.078±0.034	3.194±0.307	3.515±0.28	0.027±0.022	0.019±0.002	0.004±0	0.127±0.029	0.126±0.028	0.005±0.004
256	0.01	0.1	0.839±0.007	0.028±0	0.022±0	0.175±0.001	0.112±0.001	0.005±0.001	8.562±0.333	8.658±0.55	1.264±0.076	1.42±0.285	2.453±2.381	3.844±3.756	0.196±0.192	3.38±0.27	3.674±0.25	0.035±0.02	0.021±0.002	0.004±0	0.123±0.012	0.124±0.012	0.007±0.003
256	0.01	0.2	0.835±0.012	0.028±0	0.022±0	0.173±0	0.112±0.001	0.004±0	8.645±0.197	8.471±0.21	1.128±0.055	0.948±0.04	184.434±367.217	285.501±568.463	14.707±29.282	3.447±0.376	3.73±0.35	14.675±29.298	0.023±0	0.007±0.002	0.14±0.022	0.138±0.021	0.008±0.004
256	0.01	0.5	0.825±0.016	0.028±0	0.022±0	0.172±0.001	0.113±0	0.002±0	8.979±0.201	8.276±0.123	1.107±0.07	0.55±0.058	369.751±451.485	562.758±687.147	29.362±35.852	3.572±0.456	3.84±0.424	29.322±35.884	0.033±0.002	0.01±0.002	0.169±0.021	0.167±0.023	0.009±0.003
256	0.05	0	0.871±0.008	0.028±0	0.027±0	0.171±0.001	0.166±0.001	0.002±0	8.172±0.166	7.173±0.246	0.396±0.024	0.15±0.033	1361.635±4.512	1398.945±8.207	73.271±0	4.31±0.002	4.325±0.003	73.271±0	0.008±0	0.004±0	0.067±0.013	0.066±0.012	0.004±0.001
256	0.1	0	0.874±0.01	0.028±0	0.03±0	0.177±0.001	0.195±0.001	0.001±0	9.321±0.232	7.611±0.17	0.387±0.024	0.07±0.007	1596.66±4.923	1448.384±7.388	73.271±0	4.395±0.002	4.343±0.003	73.271±0	0.008±0	0.004±0	0.065±0.013	0.065±0.013	0.007±0.006
512	0.01	0	0.825±0.009	0.029±0	0.021±0	0.188±0.001	0.1±0.001	0.007±0.001	10.036±0.06	10.45±0.18	3.066±0.098	4.738±0.972	0.673±0.065	1.262±0.121	0.06±0.006	3.024±0.061	3.473±0.054	0.015±0.002	0.016±0	0.004±0	0.152±0.014	0.154±0.015	0.013±0.003

D.1. Additional Experiment on noise-fitting epochs

To reduce the computation of REF Complexity, we test REF Complexity under different noise-fitting epochs in Table 7 and found that our techniques are robust to the number of noise-fitting epochs. This suggests that one can use a decreasing number of noise-fitting rounds to reduce computation costs without significant degradation in performance. The reason might be that the initial fitting speed of noise might have contained enough information.

E. Discussion

We next consider several special cases, which help better understand REF Complexity in practice.

Table 6. Correlation between REF Complexity and test accuracy. A generalization measure performs well if the correlations are all positive/negative, and their absolute values are close to one. This table can be derived by Table 5.

Type	Norm-based Measures					
Measure	L2-Norm	L2-Dist	F-norm	F-Dist	Inv-Margin	Spectral
batch size	-0.935	0.899	-0.938	0.889	0.944	-0.957
learning rate	-0.910	0.980	-0.911	0.973	-0.959	-0.653
dropout	0.452	-0.072	0.449	-0.071	0.676	-0.764
Type	Norm-based			Sharpness-based		
Measure	Spectral/Margin	Path-Norm	Path-Norm/Margin	PB-I	PB-O	PB-Flatness
batch size	-0.942	-0.996	-0.934	0.830	0.824	0.835
learning rate	-0.982	-0.927	-0.960	0.976	0.978	0.977
dropout	0.473	0.452	0.568	-0.899	-0.895	-0.898
Type	Sharpness-based			Stability-based		
Measure	PB-M-I	PB-M-O	PB-M-Flatness	Steps(1)	Steps(1.5)	
batch size	0.700	-0.909	0.638	0.911	0.910	
learning rate	0.992	0.981	0.977	-0.985	-0.734	
dropout	-0.647	-0.651	-0.898	-0.908	-0.898	
Type	Ours					
Measure	REF (w/o noi)	REF (w/o init)	**REF Complexity**			
batch size	-0.782	-0.997	**-0.998**			
learning rate	-0.501	-0.996	**-0.997**			
dropout	-0.766	-0.964	**-0.965**			

Table 7. Correlation between REF Complexity and test accuracy with different noise-fitting epochs.

Epoch	Correlation
25	-0.9883
50	-0.9881
75	-0.9877
100	-0.9871
150	-0.9824

Case 1: algorithm with constant output. If an algorithm performs like $\mathcal{A}_t(\mathcal{D}) = \boldsymbol{\theta}$, it will return constant REF Complexity in the training process. This matches the fact that the generalization metric would not change during training.

Case 2: algorithm which returns constant output when detecting noise. In this case, an algorithm returns constant output when detecting noise, and otherwise performs as usual. In this case, REF Complexity would ignore the model's performance on the noise dataset, and degenerate into stability-based measures. This will, of course, hurt the effectiveness of REF Complexity. Fortunately, this algorithm is artificially designed and usually does not actually appear in practice (since detecting whether the response is pure noise is not an easy task in practice).

Case 3: memorization case. If an algorithm memorizes both the real-world dataset and the noise dataset, REF Complexity ($\mathcal{T}_n^\beta(\mathcal{D}, \mathcal{A}_t)$) would become a $0/0$ type. This is invalid in the experiment. Fortunately, this usually does not appear in practice. The reason is that: machine learning models usually fit noise slowly and require many epochs to fit noise datasets. However, practical models usually cannot be trained with such many epochs and, therefore, cannot indeed memorize all the noisy labels. Therefore, it is usually safe to apply REF Complexity in practical cases.

F. Illustration

This section introduces some intuitions omitted in the main text. We first show in Figure 2 the intuition of REF Complexity. Specifically, for a noisy dataset, if a model-algorithm pair learn signal faster (small REF Complexity), it generalizes better (Figure (b)), and vice versa. We also show in Figure 3 the relationship between the bound proposed in Theorem 4.4 and REF Complexity. Additionally, we show in Figure 3 the comparison between stochastic algorithms (*e.g.*, SGD) and deterministic algorithms (*e.g.*, GD). Specifically, for deterministic algorithms, each iteration reduces the training loss. However, for

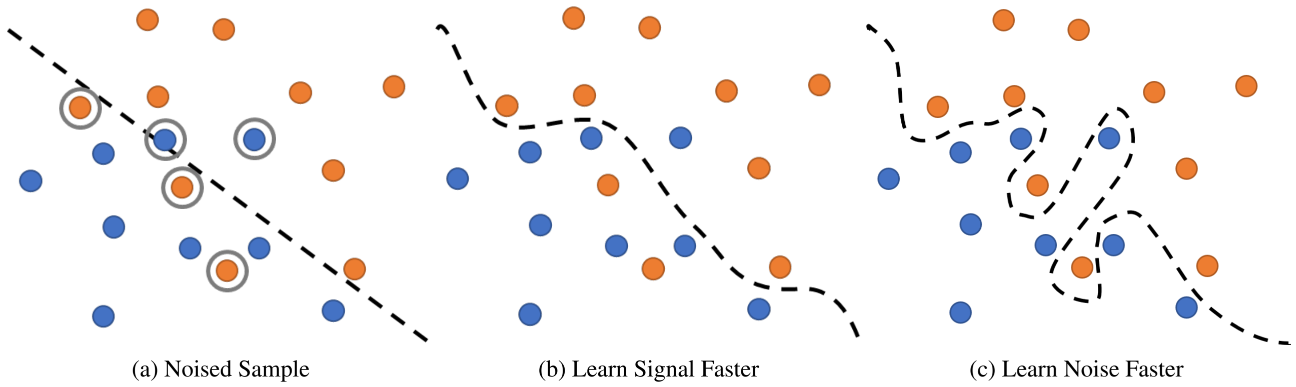


Figure 2. An illustration for REF Complexity. When the signal learning is faster, the learned decision boundary becomes close to the ground truth. In opposite, if the noise learning is faster, the decision boundary becomes close to the noise thus hard to generalize

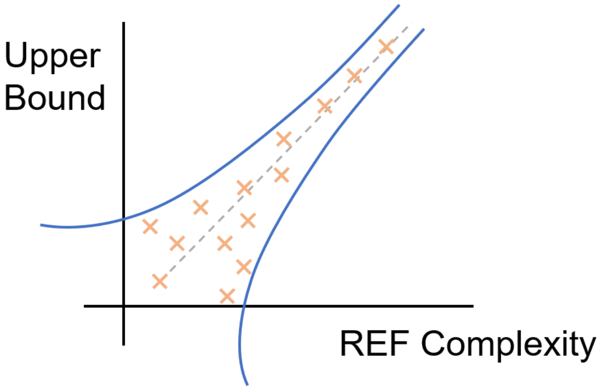


Figure 3. An illustration for the bound of REF Complexity

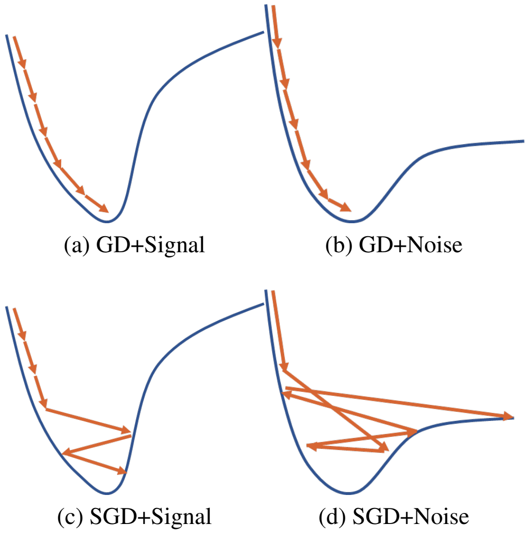


Figure 4. An illustration for Stochastic Algorithms

stochastic algorithms, signal training can reduce the training loss due to the same pattern, while noise training cannot.