FINDING GENERALIZATION MEASURES BY CONTRASTING SIGNAL AND NOISE

Anonymous authors
Paper under double-blind review

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 (RElative Fitting velocity 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) slow. Empirically, REF Complexity monotonically relates to test accuracy in real-world datasets without accessing additional validation sets, and achieves -0.988 correlation on CIFAR-10 and -0.960 correlation on CIFAR-100. We further theoretically verify the utility of REF Complexity under the regime of convex training with stochastic gradient descent.

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 where we only need to compare two models (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 widely 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 gradient 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 a theoretical backbone, (b) empirically works, and (c) is appliable in standard training scenarios, named REF Complexity (RElative Fitting velocity on signal and noise). The complexity is motivated by the intuition that a given model-algorithm pair may generalize better if it fits the signal faster while fitting the 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 a 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. 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 under the REF Complexity framework, which accords with reality.

From the theoretical perspective, we validate the utility of REF Complexity by deriving that generalization error can be bounded using REF Complexity under the regime of convex training[1] with SGD. 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 conclusion hold beyond SGD, 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).

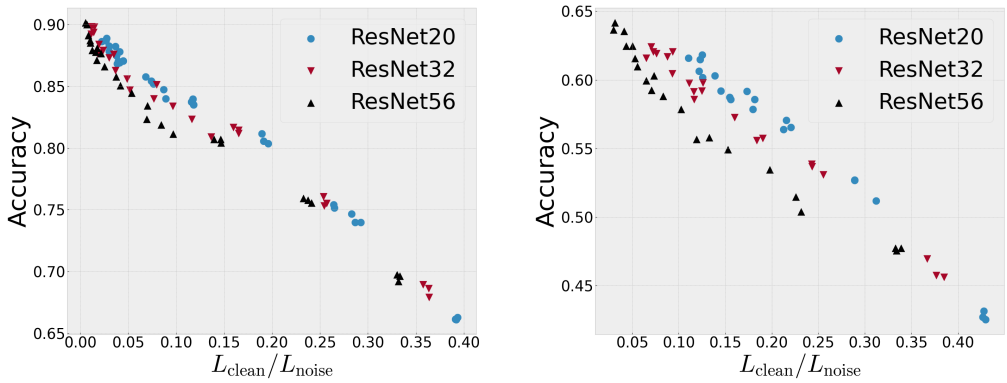
We list our contributions as follows:

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 the regime of convex training with SGD. Moreover, we show that similar arguments hold beyond SGD.

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 focuses on deriving high probability bound based on algorithmic stability (Feldman & Vondrák, 2019; Bousquet et al., 2020). Another line of works tries deriving 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).

[1] One may relax the convex assumption using Stochastic Gradient Langevin Dynamics (SGLD) algorithms.



(a) CIFAR-10 (b) CIFAR-100

Figure 1: Correlation between REF Complexity and test accuracy. We conduct over 144 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 6.

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 generalization 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 gener-

alization 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. 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 as $\mathcal{E}(\boldsymbol{\theta}; \mathcal{P}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}}\ell(\boldsymbol{\theta}; \boldsymbol{z}) - \ell(\boldsymbol{\theta}^*; \boldsymbol{z})$, 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,*

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

Then the following generalization bound holds

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

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*

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

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

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: $\mathcal{T}_n^{\beta}(\mathcal{D},\mathcal{A}_t) = \frac{\mathcal{L}_n(\mathcal{A}_t(\mathcal{D});\mathcal{D})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D});\mathcal{D})}{\frac{1}{m}\sum_{j\in[m]}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}^{(j)});\mathcal{D}_{\text{noi}}^{(j)})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}^{(j)});\mathcal{D}_{\text{noi}}^{(j)})}$.

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 Formal Definition of REF Complexity

This section introduces the formal definition of REF Complexity, which quantifies the ability of a model-algorithm pair to distinguish between signal and noise. Due to the practical restrictions, the notion of REF Complexity is slightly different in theory and in practice. We next introduce them separately, denoted as $\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)$ and $\mathcal{T}_n^{\beta}(\mathcal{D},\mathcal{A}_t)$.

4.1 Analyzing REF Complexity

In theoretical analysis, we can explicitly define the notion of signal and noise. Informally, if an output has the form $y = f(\boldsymbol{x};\theta^*) + \epsilon$, we can split it into $f(\boldsymbol{x};\theta^*)$ and ϵ, as defined in Section 3. Therefore, we directly define REF Complexity based on dataset $\mathcal{D}_{\text{noi}}$ and $\mathcal{D}_{\text{sig}}$ in the theoretical version. For a given training dataset $\mathcal{D}$ and training algorithm $\mathcal{A}_t$, its theoretical REF Complexity can be measured as

$$\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t) = \frac{1-\mathbb{E}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})}{1-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{sig}});\mathcal{D}_{\text{sig}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{sig}});\mathcal{D}_{\text{sig}})}, \tag{2}$$

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 Calculating REF Complexity

In practice, a real-world dataset usually mixes signal and noise. Unfortunately, it is impossible to split the dataset into signal and noise components perfectly. Therefore, we cannot obtain $\mathcal{D}_{\text{sig}}$ and calculate the REF Complexity like the theoretical version. Despite all this, a possible way is to quantify a data-algorithm pair's ability to distinguish the *real-world* dataset with 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. We formulate the practical version of REF Complexity in Equation equation 3, given the dataset $\mathcal{D}$ and algorithm $\mathcal{A}_t$,

$$\mathcal{T}_n^{\beta}(\mathcal{D},\mathcal{A}_t) = \frac{\mathcal{L}_n(\mathcal{A}_t(\mathcal{D});\mathcal{D})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D});\mathcal{D})}{\mathbb{E}\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})/\mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})}, \tag{3}$$

where n denotes the sample size, and the expectation is taken over the random noise in $\mathcal{D}_{\text{noi}}$. REF Complexity $\mathcal{T}_n^{\beta}(\mathcal{D},\mathcal{A}_t)$ becomes larger when fitting noise more (with smaller $\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})$), given the degree of fitting signal. The form in Equation equation 3 is different from that in Equation equation 2 due to computational stability. Since the real-world dataset $\mathcal{D}$ usually contains noise, it is safer to put the related term in the numerator instead of the denominator.

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.

Besides, REF Complexity $\mathcal{T}_n^\beta(\mathcal{D}, \mathcal{A}_t)$ is usually smaller than one, since machine learning models usually fit signal faster than noise (Arora et al., 2019; Zhang et al., 2021).

Comparison to 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 much noise, leading to a large generalization error.

5 Bounding Generalization via REF Complexity

In this section, we derive generalization bound using REF Complexity, providing theoretical guarantees for the metric. As the commonly-considered algorithm, we first study SGD under convex regimes in Section 5.1. We further validate that the bound form holds beyond SGD, by considering GD under overparameterized linear regression regimes in Section 5.2 During the analysis, we consider the metric of excess risk introduced before, which is generally considered in the related literature (Bartlett et al., 2020; Teng et al., 2022).

5.1 SGD under Convex 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

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

where $\boldsymbol{z}$ 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 5.1 (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 $t \geq T_1$,*

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

where $\psi_1(n) \to 0$ as $n \to \infty$.

Assumption 5.1 can hold in both linear and non-linear cases under some additional assumptions, as demonstrated in Teng et al. (2022). The next Assumption 5.2 sketches the properties of signal training and noise training.

Assumption 5.2 (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*

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

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

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

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 derive in Proposition 5.1 that overparameterized linear regression regime with MSE loss satisfies the above assumptions.

Proposition 5.1. *Overparameterized linear regression regimes satisfy both Assumption 5.1 and Assumption 5.2. Specifically, when the optimal parameter $\|\theta^*\| = O(1)$ and sample covariance $\|\Sigma_{\boldsymbol{x}}\| = O(1)$, we derive that*

(a.) For all step t, $\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{P}) \leq 2[\mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{sig}); \mathcal{P}_{sig}) + \mathbb{E}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{noi}); \mathcal{P}_{noi})]$;

(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 5.1 (Convex, smooth, with SGD)**.** *Assume that the loss $\ell(\theta; \boldsymbol{z})$ is convex and M-smooth with respect to θ for any sample $\boldsymbol{z}$. Consider the SGD training regime with constant stepsize $\eta \leq \frac{1}{\sqrt{t}}$. Under Assumption 5.1 and Assumption 5.2, the following inequality holds when $\sum_{j\in[t]} \sigma_w^2(j; \mathcal{D}_{noi}) = o(n^2)$ and $t \geq \max\{T_1, T_2, T_3\}$*

$$\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{P}) \leq \frac{8ec\max\{M,1\}}{\sqrt{t}} \max\left\{u, u^2\right\} \mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{T}_n^\alpha(\mathcal{D}, \mathcal{A}_t) + \psi(n),$$

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 in Assumption 5.1 and Assumption 5.2.

Derived from Theorem 5.1, 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 C for more discussions.

About the order in ψ. Besides the order of ψ_1, ψ_2 in Assumption 5.1 and Assumption 5.2, 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 5.1 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 5.1 and Assumption 5.2. 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 $t \in (\Omega(n), o(n^2))$[2].

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

We finally remark that we here provide the generalization bound with 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). 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 5.1 and Assumption 5.2. 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.

Remark 5.1 (Comparison to algorithmic stability)**.** *The measures proposed in Theorem 5.1 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 5.1 is not to provide a tight bound but to validate the utility of REF Complexity.*

We close the section by introducing how to generalize the results in Theorem 5.1 to general non-convex regimes, under the training algorithm of Stochastic Gradient Langevin Dynamics (SGLD). The main reason we need convex regimes 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 non-convex regimes. However, this could be avoided in SGLD training. A line of work (*e.g.*, Mou et al. (2018); Li et al. (2020)) derived that one can bound the generalization gap using the cumulative gradient. We leave the detailed discussion for future work.

5.2 GD UNDER OVERPARAMETERIZED LINEAR REGRESSION

To validate the generality of REF Complexity, this section proves a similar argument under overparameterized linear regression 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 $f(x;\theta^*) = x^\top\theta^*$ 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 $y - \boldsymbol{x}^\top\theta^*$ 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 5.2 (Overparameterized Linear Regression with Gradient Descent)**.** *Under overparameterized linear regression regimes, we assume that $r_0(\Sigma) = o(n)$ and $k^* = o(n)$. Besides, we assume that $\|\theta^*\|_2 = O(1)$, $\|\Sigma_{\boldsymbol{x}}\|_2 = O(1)$ in a constant scale. We consider 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 $1 - \delta$,*

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

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 5.1, Theorem 5.2 do not contains time dependency (t/n-type term). This is due to the different techniques used in the proof. Unfortunately, the techniques used in this section cannot be easily applied to general convex regimes. We defer the whole proof to Appendix B.

6 EXPERIMENTAL RESULTS

This section provides experimental results to validate the utility of REF Complexity. Specifically, we conduct over 144 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

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

Table 1: Experiments on different batch sizes and different learning rates. REF Complexity and norm-based bounds are both expected to be negatively related to accuracy. The W-norm fails to show correct correlation with accuracy (with even positive correlation), while REF Complexity works (with negative correlation).

BATCH SIZE	256	512	1024	2048	CORRELATION
ACCURACY	87.6±0.2	85.4 ±0.4	82.4±0.9	81.1±0.1	-
REF (↓)	0.021±0.003	0.051±0.003	0.111±0.016	0.155±0.002	**-0.987**
W-NORM (↓)	262.7±0.9	174±3.0	137±1.0	116±2.0	**0.960**
LR	0.005	0.01	0.05	0.1	CORRELATION
ACCURACY	81.9±0.2	82.4±0.9	86.3±0.3	88.1±0.4	-
REF (↓)	0.144±0.002	0.110±0.016	0.045±0.004	0.023±0.004	**-0.981**
W-NORM (↓)	115.9±0.5	137.1±1.3	304±4.0	536±12.0	**0.964**

negatively correlates to test accuracy with correlations of -0.988 and -0.960 on CIFAR-10 and CIFAR-100, respectively. Besides, we show how to use REF Complexity to explain some interesting phenomena in deep learning regimes, *e.g.*, the great success of stochastic algorithms.

Setup. Our experiments contain two parts: Firstly, we conduct experiments on CIFAR-10 and CIFAR-100 to show the correlation between test accuracy and REF Complexity. 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. Secondly, we conduct experiments on CIFAR-10 and evaluate the correlation between REF Complexity and test accuracy. Each configuration is trained three times, and we report the mean and standard deviation.

Discussion. Besides the monotonic relationship between REF Complexity and test accuracy, we find that the notion of REF Complexity helps explain the deep learning phenomenon from a different perspective. We here take stochastic algorithms as an example. The success of stochastic algorithms (*e.g.*, SGD and its variants) is widely observed in deep learning regimes. We empirically explain the phenomenon by REF Complexity (see Table 1), using different batch sizes as a surrogate. Fortunately, one can explain the phenomenon under REF Complexity. We defer the details below. Using a similar argument above, one can explain other phenomena empirically (*e.g.*, learning rate). We summarize the experimental results in Table 1.

Analyzing stochastic algorithms under REF Complexity. We end this section by discussing more on the stochastic algorithms under the framework of REF Complexity. For deterministic algorithms (*e.g.*, GD), each iteration sees all the samples, and therefore the training loss would decrease in each iterate for both signal and noise training. However, for stochastic algorithms (*e.g.*, SGD), each iterate 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 can even damage the training loss on the remaining samples. This leads to a better REF Complexity for stochastic algorithms. We illustrate this phenomenon in Appendix C (Figure 4).

7 CONCLUSION

In this paper, we propose a new generalization measure REF Complexity under the algorithmic stability framework, which contains a theoretical backbone and empirically works well in standard training scenarios. The complexity is motivated by the intuition that a model-algorithm pair would generalize better if it fits the signal fast while fitting the noise slowly. The success of REF Complexity may inspire some future directions. From the theoretical view, it would be interesting to relax the assumption on gradient noise used in Theorem 5.1. From the empirical view, one may find more generalization measures using signal-noise techniques. Another interesting direction is to predict exact generalization error using the REF Complexity framework. If this is done, it may become a new standard 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.

REFERENCES

Sumukh K. Aithal, Dhruva Kashyap, and Natarajan Subramanyam. Robustness to augmentations as a generalization metric. *CoRR*, abs/2101.06459, 2021. URL https://arxiv.org/abs/2101.06459.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (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.

Yamini Bansal, Gal Kaplun, and Boaz Barak. 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.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (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.

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

Raef Bassily, Vitaly Feldman, Cristóbal Guzmán, and Kunal Talwar. Stability of stochastic gradient descent on nonsmooth convex losses. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (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.

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

Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In Jacob D. Abernethy and Shivani Agarwal (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.

Sébastien Bubeck. 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.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (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.

Volkan Cevher and Bang Công Vu. 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.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In Gal Elidan, Kristian Kersting, and Alexander T. Ihler (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.

Gintare Karolina Dziugaite, Alexandre Drouin, Brady Neal, Nitarshan Rajkumar, Ethan Caballero, Linbo Wang, Ioannis Mitliagkas, and Daniel M. Roy. In search of robust measures of generalization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (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.

Frank Emmert-Streib and Matthias Dehmer. 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.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In Alina Beygelzimer and Daniel Hsu (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.

Benoît Frénay and Michel Verleysen. 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.

Saurabh Garg, Sivaraman Balakrishnan, J. Zico Kolter, and Zachary C. Lipton. RATT: leveraging unlabeled data to guarantee generalization. In Marina Meila and Tong Zhang (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.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (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.

Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Dan Roy. Towards a unified information-theoretic framework for generalization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (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.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In Maria-Florina Balcan and Kilian Q. Weinberger (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.

Yiding Jiang, Parth Natekar, Manik Sharma, Sumukh K. Aithal, Dhruva Kashyap, Natarajan Subramanyam, Carlos Lassance, Daniel M. Roy, Gintare Karolina Dziugaite, Suriya Gunasekar, Isabelle Guyon, Pierre Foret, Scott Yak, Hossein Mobahi, Behnam Neyshabur, and Samy Bengio. Methods and analysis of the first competition in predicting generalization of deep learning. In Hugo Jair Escalante and Katja Hofmann (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.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 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.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J. Zico Kolter. 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.

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

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 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.

Yunwen Lei and Yiming Ying. 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.

Jian Li, Xuanyuan Luo, and Mingda Qiao. 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.

Naresh Manwani and P. S. Sastry. 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.

David A. McAllester. Pac-bayesian model averaging. In Shai Ben-David and Philip M. Long (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.

Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of SGLD for non-convex learning: Two theoretical viewpoints. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (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.

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

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

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. 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.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In Peter Grünwald, Elad Hazan, and Satyen Kale (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.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (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.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. 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.

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

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert (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.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Shai Shalev-Shwartz and Shai Ben-David. *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.

Jiaye Teng, Jianhao Ma, and Yang Yuan. 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.

Colin Wei and Tengyu Ma. 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.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (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.

Jing Xu, Jiaye Teng, Yang Yuan, and Andrew Chi-Chih Yao. When do models generalize? a perspective from data-algorithm compatibility, 2022. URL https://arxiv.org/abs/2202.06054.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 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.

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

Appendix

A PROOF OF THEOREM 5.1

Theorem 5.1 (Convex, smooth, with SGD)**.** *Assume that the loss $\ell(\theta; \boldsymbol{z})$ is convex and M-smooth with respect to θ for any sample $\boldsymbol{z}$. Consider the SGD training regime with constant step-size $\eta \leq \frac{1}{\sqrt{t}}$. Under Assumption 5.1 and Assumption 5.2, the following inequality holds when $\sum_{j\in[t]} \sigma_w^2(j; \mathcal{D}_{noi}) = o(n^2)$ and $t \geq \max\{T_1, T_2, T_3\}$*

$$\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{P}) \leq \frac{8ec\max\{M,1\}}{\sqrt{t}} \max\left\{u, u^2\right\} \mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{T}_n^{\alpha}(\mathcal{D}, \mathcal{A}_t) + \psi(n),$$

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 in Assumption 5.1 and Assumption 5.2.

Proof. Firstly, due to Assumption 5.1 and Assumption 5.2, the difficulties of bounding the excess risk falls in the noise component, that is to say,

$$\mathbb{E}_{\mathcal{D},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D}); \mathcal{P}) \leq [c_1 + c_1c_2]\mathbb{E}_{\mathcal{D}_{\text{sig}},\mathcal{A}_t}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{sig}}); \mathcal{P}_{\text{sig}}) + \psi_1(n,t) + c_1\psi_2(n,t). \tag{4}$$

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,

$$\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) \leq \mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{P}_{\text{noi}}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})]. \tag{5}$$

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

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{P}_{\text{noi}}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}}) \\
\leq&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}[\frac{2e(M+c)\eta^2}{n}(1+\frac{t}{n}) + \frac{1}{2c}(\frac{1}{t}+\eta^2M)]\frac{1}{n}\sum_{i\in[n]}\sum_{j\in[t]}\mathbb{E}\|\nabla\ell(\mathcal{A}_j(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i)\|^2 + \frac{1}{2c}\sigma_w^2(t).
\end{aligned} \tag{6}$$

We finally apply Lemma A.3, which leads to

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

By using the fact that $\mathbb{E}[\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}})) - \mathcal{L}_n(\mathcal{A}_0(\mathcal{D}_{\text{noi}}))] \leq \mathcal{T}_n^{\alpha}(\mathcal{D}, \mathcal{A}_t)$ and taking $c = \sqrt{\frac{(1+1/t+\eta^2M)n}{(1+t/n)4e\eta^2}}$, it holds that

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

We consider the three parts separately:

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

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

For the second part, we derive similarly that

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

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

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

which goes to zero for bounded gradient norm.

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

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

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

For the third part, notice that

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

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

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t,\mathcal{D}_{\text{noi}}}\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{P}_{\text{noi}})-\mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}});\mathcal{D}_{\text{noi}})\\
\leq& 8e\max\{M,1\}\frac{1}{\sqrt{t}}\max\{\frac{1}{n}(1+\frac{t}{n}),\sqrt{\frac{1}{n}(1+\frac{t}{n})}\}\mathcal{T}_n^{\alpha}(\mathcal{D},\mathcal{A}_t)+\psi_3(n,t),
\end{aligned} \tag{7}$$

where $\psi_3(n,t) \to 0$ as $n, t \to 0$.

Combining Equation equation 4, Equation equation 5, Equation equation 7 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 5.1, we can bound the excess risk via generalization gap,*

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

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

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

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 5.2 , therefore,

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

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

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

Therefore, we conclude that

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

□

Lemma A.2. *Under the assumptions in Theorem 5.1, we derive that for any $c > 0$, we have that*

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

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

$$\begin{aligned}
&\mathbb{E}_{\mathcal{A}_t, \mathcal{D}_{\text{noi}}}[\mathcal{L}(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{P}_{\text{noi}}) - \mathcal{L}_n(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \mathcal{D}_{\text{noi}})] \\
=&\mathbb{E}_{\mathcal{A}_t, \mathcal{D}_{\text{noi}}, \mathcal{D}^{(i)}_{\text{noi}}} \frac{1}{n} \sum_{i \in [n]} \ell(\mathcal{A}_t(\mathcal{D}^{(i)}_{\text{noi}}); \boldsymbol{z}_i) - \ell(\mathcal{A}_t(\mathcal{D}_{\text{noi}}); \boldsymbol{z}_i),
\end{aligned} \tag{8}$$

where $\mathcal{D}^{(i)}_{\text{noi}}$ denotes the dataset with only the i-th sample different from $\mathcal{D}_{\text{noi}}$. The above equation holds because $\mathcal{D}^{(i)}_{\text{noi}}$ 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 $c > 0$,

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

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

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

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

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

Therefore, we have that

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

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

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

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,

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

where the expectation is taken over the randomness on 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

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

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

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

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

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

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

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

Therefore, we have

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

□

□

Lemma A.4 (One-expansion under convexity, from Hardt et al. (2016))**.** *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 $\boldsymbol{z}_i$,*

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

Lemma A.5 (Bound for stability parameter difference)**.** *Under the Assumptions in Theorem 5.1, We have that for any i*

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

Proof. We firstly notice that there exist ξ such that

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

Therefore, we have that

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

By taking expectation on the chosen index i, we have that

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

By iteration, we have that

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

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

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

Therefore, we have that

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

□

B Proof of Theorem 5.2

Theorem 5.2 (Overparameterized Linear Regression with Gradient Descent)**.** *Under overparameterized linear regression regimes, we assume that $r_0(\Sigma)=o(n)$ and $k^*=o(n)$. Besides, we assume that $\|\theta^*\|_2=O(1)$, $\|\Sigma_{\boldsymbol{x}}\|_2=O(1)$ in a constant scale. We consider 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 $t=\omega(1)$[4], with probability at least $1-\delta$,*

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

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

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

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

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

$$\lim_{n\to\infty}\mathcal{E}(\mathcal{A}_t(\mathcal{D}_{\text{sig}});\mathcal{D}_{\text{sig}})=0. \tag{9}$$

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

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

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

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

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

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

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

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

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

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

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

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

Note that

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

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

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

□

Lemma B.3 (Bound for noise component)**.** *Under the assumptions in Theorem 5.2, we have that with probability at least* $1-\delta$

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

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

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

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

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

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 $X\Sigma_{\boldsymbol{x}}X = \sum_i \lambda_i^2\boldsymbol{z}_i\boldsymbol{z}_i^\top$

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

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

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

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 $\lambda_i \leq \frac{1}{bn}\sum_{j>i}\lambda_j$

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

Therefore, for the remaining part, we have that

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

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

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

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

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

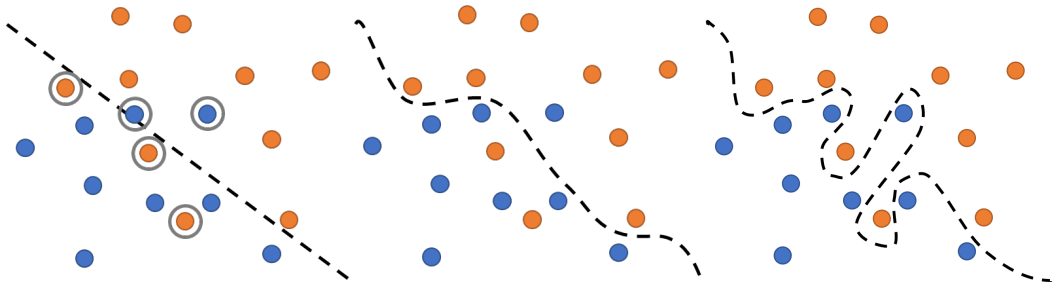
where we abuse the notation c as a constant different from the above text. Therefore, with high probability, we have that there exist constant c_3, such that

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

Therefore, we have that

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

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



(a) Noised Sample (b) Learn Signal Faster (c) Learn Noise Faster

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

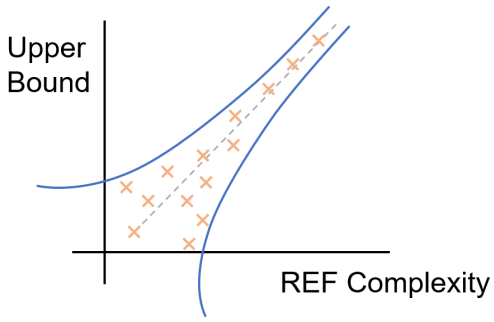
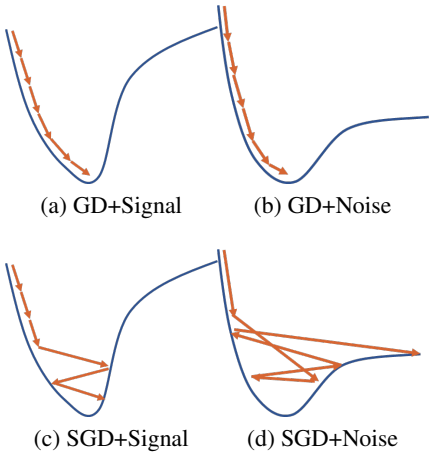


Figure 3: An illustration for the bound of REF Complexity



(a) GD+Signal (b) GD+Noise

(c) SGD+Signal (d) SGD+Noise

Figure 4: An illustration for Stochastic Algorithms

C 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 5.1 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 stochastic algorithms, signal training can reduce the training loss due to the same pattern, while noise training cannot.