WHEN DO MODELS GENERALIZE? A PERSPECTIVE FROM DATA-ALGORITHM COMPATIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

One of the major open problems in machine learning is to characterize generalization in the overparameterized regime, where most traditional generalization bounds become inconsistent even for overparameterized linear regression (Nagarajan & Kolter, 2019). In many scenarios, their failure can be attributed to obscuring the crucial interplay between the training algorithm and the underlying data distribution. To address this issue, we propose a concept named compatibility, which quantitatively characterizes generalization in a both data-relevant and algorithm-relevant manner. By considering the entire training trajectory and focusing on early-stopping iterates, compatibility exploits the data and the algorithm information and is therefore a suitable notion for generalization of overparameterized models. We validate this by theoretically studying compatibility under the setting of solving overparameterized linear regression with gradient descent. Specifically, we perform a data-dependent trajectory analysis and derive a sufficient condition for compatibility in such a setting. Our theoretical results demonstrate that in the sense of compatibility, generalization holds with significantly weaker restrictions on the problem instance than the previous at-convergence analysis.

1 INTRODUCTION

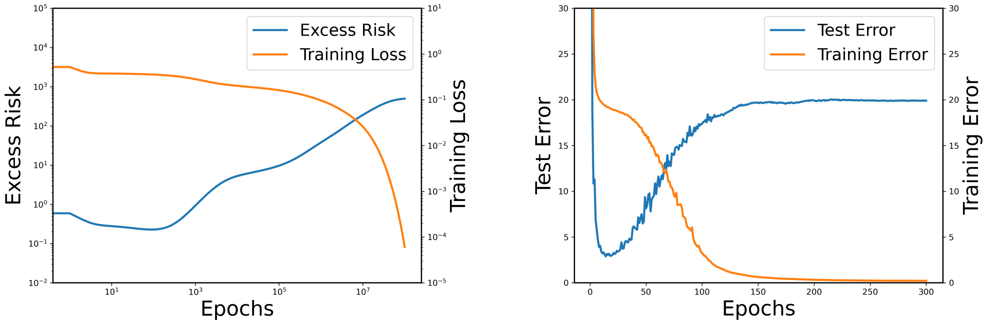
Although deep neural networks achieve great success in practice (Silver et al., 2017; Devlin et al., 2019; Brown et al., 2020), their remarkable generalization ability is still among the essential mysteries in the deep learning community. One of the most intriguing features of deep neural networks is overparameterization, which confers a level of tractability to the training problem, but leaves traditional generalization theories failing to work. In generalization analysis, both the training algorithm and the data distribution play essential roles (Jiang et al., 2020). For instance, a line of work (Zhang et al., 2021; Nagarajan & Kolter, 2019) highlights the role of the algorithm by showing that the algorithm-irrelevant uniform convergence bounds can become inconsistent in deep learning regimes. Another line of work (Bartlett et al., 2019; Tsigler & Bartlett, 2020) on benign overfitting emphasizes the role of data distribution via profound analysis of specific overparameterized models.

Despite the significant role of data and algorithm in generalization analysis, existing theories usually focus on either the data factor (*e.g.*, uniform convergence (Nagarajan & Kolter, 2019) and last iterate analysis (Bartlett et al., 2019; Tsigler & Bartlett, 2020)) or the algorithm factor (*e.g.*, stability-based bounds (Hardt et al., 2016)). [1] Combining both data and algorithm factor into generalization analysis can help derive tighter generalization bounds and explain the generalization ability of overparameterized models observed in practice. In this sense, a natural question arises:

How to incorporate both data factor and algorithm factor into generalization analysis?

To gain insight into the interplay between data and algorithms, we provide motivating examples of a synthetic overparameterized linear regression task and a classification task on the corrupted MNIST dataset in figure 1. In both scenarios, the final iterate with less algorithmic information,

[1] The taxonomy of data-dependent techniques does not mean that they totally ignore all the information of algorithm, but mean that it loses some important algorithm information which makes it vacuous in generalization analysis. Similar arguments hold for algorithm-dependent techniques.



(a) Linear Regression (b) Corrupted MNIST

Figure 1: **(a) The training plot for linear regression with spectrum $\lambda_i = 1/i^2$ using GD.** Note that the axes are in the log scale. **(b) The training plot of CNN on corrupted MNIST with 20% label noise using SGD.** Both models successfully learn the useful features in the initial phase of training, but it takes a long time for them to fit the noise in the dataset. The observations demonstrate the power of data-dependent trajectory analysis, since the early stopping solution on the trajectory generalizes well but the final iterate fails to generalize. See Appendix C for details.

which may include the algorithm type (*e.g.*, GD or SGD), hyperparameters (*e.g.*, learning rate, number of epochs), generalizes much worse than the early stopping solutions (see the Blue Line). In the linear regression case, the generalization error of the final iterate can be more than $\times 100$ larger than that of the early stopping solution. In the MNIST case, the final iterate on the SGD trajectory has 19.9% test error, much higher than the 2.88% test error of the best iterate on the GD trajectory. Therefore, the almost ubiquitous strategy of early stopping is a key ingredient in generalization analysis for overparameterized models, whose benefits have been demonstrated both theoretically and empirically (Yao et al., 2007; Ali et al., 2019; Li et al., 2020b; Ji et al., 2021). By focusing on the entire optimization trajectory and performing data-dependent trajectory analysis, both data information and the dynamics of the training algorithm can be exploited to yield consistent generalization bounds.

To analyze the data-dependent trajectory, we introduce a new concept named **data-algorithm-compatibility**, which jointly characterizes the role of the data and the algorithm in generalization analysis. Informally speaking, an algorithm is compatible with a data distribution if as the sample size goes to infinity, the minimum excess risk of the iterates on the training trajectory converges to zero.

The significance of compatibility comes in three folds. Firstly, compatibility incorporates both data and algorithm factors into generalization analysis, and brings new messages into generalization in the overparameterization regime (see Definition 3.1). Secondly, compatibility serves as a minimal condition for generalization, without which one cannot expect to find a consistent solution via standard learning procedures. Consequently, compatibility holds with only mild assumptions and applies to a wide range of problem instances (see Theorem 4.1). Thirdly, compatibility captures the algorithmic significance of early stopping in generalization. By exploiting the algorithm information along the entire trajectory, we arrive at better generalization bounds than the at-convergence analysis (see Table 1 and 2 for examples).

To theoretically validate compatibility, we study it under overparameterized linear regression setting. Analysis of the overparameterized linear regression is a reasonable starting point to study compatibility for more complex models like deep neural networks, since many phenomena of the high dimensional non-linear model are also observed in the linear regime (*e.g.*, Figure 1). Furthermore, the recent neural tangent kernel (NTK) framework demonstrates that very wide neural networks trained using gradient descent with appropriate random initialization can be approximated by kernel regression in a reproducing kernel Hilbert space, which rigorously establishes a close relationship between overparameterized linear regression and deep neural network training (Jacot et al., 2018; Arora et al., 2019).

Specifically, we investigate solving overparameterized linear regression using gradient descent with constant step size, and prove that under some mild regularity conditions, gradient descent is compatible with overparameterized linear regression if the effective dimensions of the feature covariance matrix are asymptotically bounded by the sample size. In this setting, the assumptions needed for generalization in the sense of compatibility are significantly weaker than that in the at-convergence analysis (Bartlett et al., 2019), which demonstrates the benefits of data-relevant and algorithm-relevant generalization analysis.

We summarize our contributions as follows:

- We formalize the notion of compatibility, which highlights the interaction between data and algorithm and serves as a minimal condition for generalization.
- We derive a sufficient condition for compatibility in solving overparameterized linear regression with gradient descent. Our theory substantiates the meaningfulness of compatibility by showing that generalization in the sense of compatibility typically requires much weaker restrictions in the problem instance.
- Technically, we derive time-variant generalization bounds for overparameterized linear regression via data-dependent trajectory analysis. Empirically, various experiment results verify the motivation of compatibility and demonstrate the benefits of early stopping.

2 Related Works

Data-Dependent Techniques mainly focus on the data distribution condition for generalization. One of the most popular bounds among data-dependent bounds is uniform convergence (Koltchinskii & Panchenko, 2000; Bartlett et al., 2017; Zhou et al., 2020; Zhang et al., 2021). However, recent works(Nagarajan & Kolter, 2019; Negrea et al., 2020) point out that uniform convergence may not be powerful enough to explain generalization, because it may only yield inconsistent bound in even linear regression cases. Another line of works investigates benign overfitting, which mainly involves generalization at convergence (Bartlett et al., 2019; Zou et al., 2021; Tsigler & Bartlett, 2020; Li et al., 2020c; Wang & Thrampoulidis, 2021; Frei et al., 2022).

Algorithm-Dependent Techniques measure the role of the algorithmic information in generalization. A line of works derives generalization bounds via algorithm stability (Hardt et al., 2016; Feldman & Vondrák, 2018; Mou et al., 2018; Feldman & Vondrák, 2019; Bousquet et al., 2020; Li et al., 2020a; Lei & Ying, 2020; Bassily et al., 2020; Teng et al., 2021). A parallel line of works analyzes the implicit bias of algorithmic information (Bousquet & Elisseeff, 2002; Soudry et al., 2018; Shah et al., 2020; Hu et al., 2020; Lyu & Li, 2020; Lyu et al., 2021), which are mainly based on analyzing a specific data distribution (*e.g.*, linear separable).

Other Generalization Techniques. Besides the techniques discussed above, there are many other approaches. For example, PAC-Bayes theory performs well empirically and theoretically (Shawe-Taylor & Williamson, 1997; Seeger, 2002; McAllester, 2003; Parrado-Hernández et al., 2012; McAllester, 2013; Dziugaite & Roy, 2017; Neyshabur et al., 2018) and can even yield non-vacuous bounds in deep learning regimes (Rivasplata et al., 2020; Pérez-Ortiz et al., 2021). Furthermore, there are other promising techniques including information theory (Russo & Zou, 2016; Xu & Raginsky, 2017; Banerjee & Montúfar, 2021), and compression-based bounds (Arora et al., 2018).

Early Stopping has the potential to improve generalization for various machine learning problems (Raskutti et al., 2014; Vaskevicius et al., 2020; Zhang et al., 2021; Li et al., 2021; Kuzborskij & Szepesvári, 2021; Bai et al., 2021; Shen et al., 2022). A line of works studies the rate of early stopping in linear regression and kernel regression with different algorithms, *e.g.*, gradient descent (Yao et al., 2007), stochastic gradient descent (Tarres & Yao, 2014; Rosasco & Villa, 2015; Dieuleveut & Bach, 2016; Lin & Rosasco, 2017; Pillaud-Vivien et al., 2018), gradient flow (Ali et al., 2019), conjugate gradient (Blanchard & Krämer, 2016) and spectral algorithms (Gerfo et al., 2008; Lin & Cevher, 2018). Of the most relevance here is Yao et al. (2007), which proves an optimal excess bound for a certain class of kernel regression problems solved using early stopping gradient descent, Beyond linear models, early-stopping is also effective for training deep neural networks (Li et al., 2020b; Ji et al., 2021). Another line of research focuses on the signal for early stopping (Prechelt, 2012; Forouzesh & Thiran, 2021).

3 Compatibility

In this section, we formally define compatibility between the data distribution and the training algorithm, starting from the basic notations.

3.1 Notations

Data Distribution. Let $\mathcal{D}$ denote the population distribution and $\boldsymbol{z} \sim \mathcal{D}$ denote a data point sampled from distribution $\mathcal{D}$. Usually, $\boldsymbol{z}$ contains a feature and its corresponding response. Besides, we denote the dataset with n samples as $\boldsymbol{Z} \triangleq \{\boldsymbol{z}_i\}_{i\in[n]}$, where $\boldsymbol{z}_i \sim \mathcal{D}$ are i.i.d. sampled from distribution $\mathcal{D}$.

Loss and Excess Risk. Let $\ell(\boldsymbol{\theta}; \boldsymbol{z})$ denote the loss on sample $\boldsymbol{z}$ with parameter $\boldsymbol{\theta} \in \mathbb{R}^p$. The corresponding population loss is defined as $L(\boldsymbol{\theta}; \mathcal{D}) \triangleq \mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}\ell(\boldsymbol{\theta}; \boldsymbol{z})$. When the context is clear, we omit the dependency on $\mathcal{D}$ and denote the population loss by $L(\boldsymbol{\theta})$. Our goal is to find the optimal parameter $\boldsymbol{\theta}^*$ which minimizes the population loss, i.e., $L(\boldsymbol{\theta}^*) = \min_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$. Measuring how a parameter $\boldsymbol{\theta}$ approaches $\boldsymbol{\theta}^*$ relies on a term *excess risk* $R(\boldsymbol{\theta})$, defined as $R(\boldsymbol{\theta}) \triangleq L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$.

Algorithm. Let $\mathcal{A}(\cdot)$ denote a iterative algorithm that takes training data $\boldsymbol{Z}$ as input and outputs a sequence of parameters $\{\boldsymbol{\theta}_n^{(t)}\}_{t\geq 0}$, where t is the iteration number. The algorithm can be either deterministic or stochastic, *e.g.*, variants of (S)GD.

3.2 Definitions of Compatibility

Based on the above notations, we introduce the notion of compatibility between data distribution and algorithm in Definition 3.1. Informally, compatibility measures whether a consistent excess risk can be reached along the training trajectory. Note that we omit the role of the loss function in the definition, although the algorithm depends on the loss function.

Definition 3.1 (**Compatibility**). *Given a loss function $\ell(\cdot)$ with corresponding excess risk $R(\cdot)$, a data distribution $\mathcal{D}$ is **compatible** with an algorithm $\mathcal{A}$ if there exists nonempty subsets T_n of $\mathbb{N}$, such that $\sup_{t\in T_n} R(\boldsymbol{\theta}_n^{(t)})$ converges to zero in probability as sample size n tends to infinity, where $\{\boldsymbol{\theta}_n^{(t)}\}_{t\geq 0}$ denotes the output of algorithm $\mathcal{A}$, and the randomness comes from the sampling of training data $\boldsymbol{Z}$ from distribution $\mathcal{D}$ and the execution of algorithm $\mathcal{A}$. That is to say, $(\mathcal{D}, \mathcal{A})$ is compatible if there exists nonempty sets T_n, such that*

$$\sup_{t\in T_n} R(\boldsymbol{\theta}_n^{(t)}) \xrightarrow{P} 0 \quad \textit{as} \quad n \to \infty. \tag{1}$$

We call $\{T_n\}_{n>0}$ the compatibility region of $(\mathcal{D}, \mathcal{A})$. The distribution $\mathcal{D}$ is allowed to change with n. In this case, $\mathcal{D}$ should be understood as a sequence of distributions $\{\mathcal{D}_n\}_{n\geq 1}$. We also allow the dimension of model parameter $\boldsymbol{\theta}$ to be infinity or to grow with n. We omit this dependency on n when the context is clear.

Compatibility serves as a minimal condition for generalization, since if a data distribution is incompatible with the algorithm, one cannot expect to reach a small excess risk even if we allow for *arbitrary* early stopping. However, we remark that considering only the minimal excess risk is insufficient for a practical purpose, as one cannot exactly find the t that minimizes $R(\boldsymbol{\theta}_n^{(t)})$ due to the noise in the validation set. Therefore, it is meaningful to consider a region of time t on which the excess risk is consistent as in Definition 3.1. The larger the compatibility region is, the more robust the algorithm will be to the noise in its execution.

Comparisons with Other Notions. Compared to classic definitions of learnability, *e.g.*, PAC learning, the definition of compatibility is data-specific and algorithm-specific, and is thus a more fine-grained notion. Compared to the concept of *benign* proposed in the recent paper (Bartlett et al., 2019), which studies whether the excess risk at $t = \infty$ converges to zero in probability as the sample size goes to infinity, compatibility only requires that there exists a time to derive a consistent excess risk. We will show later in Section 4.2 that in the overpamameterized linear regression setting, there exist cases such that the problem instance is compatible but not benign.

4 Compatibility Analysis of Overparameterized Linear Regression with Gradient Descent

This paper mainly analyzes compatibility in the overparameterized linear regression regime. We first introduce the data distribution, loss, and training algorithm, and then present the main theorem, which provides a sufficient condition for compatibility in this setting.

4.1 Preliminaries for Overparameterized Linear Regression

Notations. Let O, o, Ω, ω denote asymptotic notations, with their usual meaning. For example, the argument $a_n = O(b_n)$ means that there exists a large enough constant C, such that $a_n \leq C b_n$. We use $\lesssim$ with the same meaning as the asymptotic notation O. Besides, let $\|\boldsymbol{x}\|$ denote the ℓ_2 norm for vector $\boldsymbol{x}$, and $\|\boldsymbol{A}\|$ denote the operator norm for matrix $\boldsymbol{A}$. We allow the vector to belong to a countably infinite-dimensional Hilbert space $\mathcal{H}$, and with a slight abuse of notation, we use $\mathbb{R}^\infty$ interchangeably with $\mathcal{H}$. In this case, $x^\top z$ denotes inner product and $xz^\top$ denotes tensor product for $x, z \in \mathcal{H}$.

Data Distribution. Let $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$ denote the feature vector and the response, following a joint distribution $\mathcal{D}$. Let $\boldsymbol{\Sigma} \triangleq \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$ denote the feature covariance matrix, whose eigenvalue decomposition is $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top = \sum_{i>0} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$ with decreasing eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$. We make the following assumptions on the distribution of the feature vector.

Assumption 1 (Assumptions on feature distribution)**.** *We assume that*

1. $\mathbb{E}[\boldsymbol{x}] = 0$.
2. $\lambda_1 > 0, \sum_{i>0} \lambda_i < C$ *for some absolute constant* C.
3. *Let* $\tilde{\boldsymbol{x}} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{V}^\top \boldsymbol{x}$. *The random vector* $\tilde{\boldsymbol{x}}$ *has independent* σ_x*-subgaussian entries.*[2]

Loss and Excess Risk. We choose square loss as the loss function ℓ, i.e. $\ell(\boldsymbol{\theta}, (\boldsymbol{x}, y)) = 1/2(y - \boldsymbol{x}^\top\boldsymbol{\theta})^2$. The corresponding population loss is denoted by $L(\boldsymbol{\theta}) = \mathbb{E}\ell(\boldsymbol{\theta}, (\boldsymbol{x}, y))$ and the optimal parameter is denoted by $\boldsymbol{\theta}^* \triangleq \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^p} L(\boldsymbol{\theta})$. We assume that $\|\boldsymbol{\theta}^*\| < C$ for some absolute constant C. If there are multiple such minimizers, we choose an arbitrary one and fix it thereafter. We focus on the excess risk of parameter $\boldsymbol{\theta}$, defined as

$$R(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*) = \frac{1}{2}\mathbb{E}(y - \boldsymbol{x}^\top\boldsymbol{\theta})^2 - \frac{1}{2}\mathbb{E}(y - \boldsymbol{x}^\top\boldsymbol{\theta}^*)^2 = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \boldsymbol{\Sigma}(\boldsymbol{\theta} - \boldsymbol{\theta}^*). \quad (2)$$

Let $\varepsilon = y - \boldsymbol{x}^\top\boldsymbol{\theta}^*$ denote the noise in data point $(\boldsymbol{x}, y)$. The following assumptions involve the conditional distribution of the noise.

Assumption 2 (Assumptions on noise distribution)**.** *We assume that*

1. *The conditional noise* $\varepsilon|\boldsymbol{x}$ *has zero mean.*
2. *The conditional noise* $\varepsilon|\boldsymbol{x}$ *is* σ_y*-subgaussian.*

Note that both Assumption 1 and Assumption 2 are commonly considered in the related literatures (Bartlett et al., 2019; Tsigler & Bartlett, 2020; Zou et al., 2021).

Training Set. Given a training set $\{(\boldsymbol{x}_i, y_i)\}_{1 \leq i \leq n}$ with n pairs independently sampled from the population distribution $\mathcal{D}$, we define $\boldsymbol{X} \triangleq (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}$ as the feature matrix, $\boldsymbol{Y} \triangleq (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$ as the corresponding noise vector, and $\boldsymbol{\varepsilon} \triangleq \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}^*$ as the residual vector. Let the singular value decomposition (SVD) of $\boldsymbol{X}$ be $\boldsymbol{X} = \boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{W}^\top$, with $\tilde{\boldsymbol{\Lambda}} = \operatorname{diag}\{\mu_1 \cdots, \mu_n\} \in \mathbb{R}^{n \times n}$, $\mu_1 \geq \cdots \geq \mu_n$.

We consider the overparameterized regime where the feature dimension is larger than the sample size, namely, $p > n$. In this regime, we assume that $\operatorname{rank}(\boldsymbol{X}) = n$ almost surely as in Bartlett et al. (2019). This assumption is equivalent to the invertibility of $XX^\top$.

[2] A random variable X is σ-subgaussian if $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for any λ.

Assumption 3 (Linear independent training set)**.** *For any $n < p$, we assume that the features in the training set $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ is linearly independent almost surely.*

Algorithm. Given the dataset $(\boldsymbol{X}, \boldsymbol{Y})$, define the empirical loss function as $\hat{L}(\boldsymbol{\theta}) \triangleq \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{Y}\|^2$. We choose full-batch gradient descent on the empirical risk with a constant learning rate λ as the algorithm $\mathcal{A}$ in the previous template. In this case, the update rule for the optimization trajectory $\{\boldsymbol{\theta}_t\}_{t \geq 0}$ is formulated as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\lambda}{n}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\theta}_t - \boldsymbol{Y}). \tag{3}$$

Without loss of generality, we consider zero initialization $\boldsymbol{\theta}_0 = \mathbf{0}$ in this paper. In this case, for a sufficiently small learning rate λ, $\boldsymbol{\theta}_t$ converges to the *min-norm interpolator* $\hat{\boldsymbol{\theta}} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{Y}$ as t goes to infinity, which was well studied previously (Bartlett et al., 2019). This paper takes one step further and discuss the excess risk along the entire training trajectory $\{R(\boldsymbol{\theta}_t)\}_{t \geq 0}$.

Effective Rank and Effective Dimensions. We define the effective rank of the feature matrix $\boldsymbol{\Sigma}$ as $r(\boldsymbol{\Sigma}) \triangleq \frac{\sum_{i>0}\lambda_i}{\lambda_1}$. Our results on compatibility depend on two notions of effective dimension of the feature covariance $\boldsymbol{\Sigma}$, defined as

$$k_0 \triangleq \min\left\{l \geq 0 : \lambda_{l+1} \leq \frac{c_0 \sum_{i>l}\lambda_i}{n}\right\}, \tag{4}$$

$$k_1 \triangleq \min\left\{l \geq 0 : \lambda_{l+1} \leq \frac{c_1 \sum_{i>0}\lambda_i}{n}\right\}, \tag{5}$$

where c_0, c_1 are constants independent of the dimension p, sample size n, and time t [3]. We omit the dependency of k_0, k_1 on $c_0, c_1, n, \boldsymbol{\Sigma}$ when the context is clear.

4.2 Compatibility for Overparameterized Linear Regression with Gradient Descent

Next, we present the main result of this section, which provides a clean condition for compatibility between gradient descent and overparameterized linear regression.

Theorem 4.1 (**Compatibility for Overparameterized Linear Regression with Gradient Descent**)**.** *Consider the overparameterized linear regression setting defined in section 4.1. Let Assumption 1,2 and 3 hold. Assume the learning rate satisfies $\lambda = O\left(\frac{1}{\mathrm{Tr}(\boldsymbol{\Sigma})}\right)$. Then under the condition that $k_0 = O(n), k_1 = o(n), r(\boldsymbol{\Sigma}) = o(n)$, gradient descent is compatible with overparameterized linear regression in the region $T_n = \left(\omega\left(\frac{1}{\lambda}\right), o\left(\frac{n}{\lambda}\right)\right)$, namely,*

$$\sup_{t \in T_n} R(\boldsymbol{\theta}_t) \xrightarrow{P} 0 \quad as \quad n \to \infty. \tag{6}$$

Furthermore, if the feature dimension $p = \infty$ and the data distribution does not change with n, then the condition $k_0 = O(n)$ alone suffices for compatibility.

The proof of Theorem 4.1 is given in Appendix A and sketched in Section 5. The theorem shows that gradient descent is compatible with overparameterized linear regression under some mild regularity conditions on the learning rate, effective rank and effective dimensions. The condition on the learning rate is natural for optimizing a smooth objective. We conjecture that the condition $k_0 = O(n)$ can not be removed in general cases, since the effective dimension k_0 characterizes the concentration of the singular values of the data matrix $\boldsymbol{X}$ and plays a crucial role in the excess risk of the gradient descent dynamics. We discuss extensions to the kernel regression setting in Appendix B.6.

Comparison with Benign Overfitting. The recent paper Bartlett et al. (2019) studies overparameterized linear regression and gives the condition for min-norm interpolator to generalize. They prove that the feature covariance $\boldsymbol{\Sigma}$ is benign if and only if

$$k_0 = o(n), \quad R_{k_0}(\boldsymbol{\Sigma}) \triangleq \frac{(\sum_{i>k_0}\lambda_i)^2}{\sum_{i>k_0}\lambda_i^2} = \omega(n), \quad r(\boldsymbol{\Sigma}) = o(n) \tag{7}$$

[3] Constants may depend on σ_x, and we omit this dependency thereafter for clarity.

As discussed in Section 3.2, benign problem instance also satisfies compatibility, since benign overfitting requires a stronger condition on k_0 and an additional assumption on $R_{k_0}(\mathbf{\Sigma})$. The following example shows that this inclusion relationship is strict.

Example 4.1. *Under the same assumption as in Theorem 4.1, if the spectrum of $\mathbf{\Sigma}$ satisfies*

$$\lambda_k = \frac{1}{k^\alpha}, \tag{8}$$

for some $\alpha > 1$, we derive that $k_0 = \Theta(n)$. Therefore, this problem instance satisfies compatibility, but does not satisfy benign overfitting.

Example 4.1 shows the existence of a case where the early stopping solution can generalize but interpolating solution fails to generalize. Therefore, compatibility can characterize generalization for a much wider range of problem instances. [4]

Comparisons with Existing Methods. Theorem 4.1 cannot be directly implied by off-the-shelf stability-based generalization bounds (Hardt et al., 2016; Feldman & Vondrák, 2019) or uniform convergence bounds (Koltchinskii & Panchenko, 2000; Bartlett et al., 2017). The main reason is that both methods rely on a *high probability* bound of the parameter norm, which requires a non-trivial data-dependent and algorithm-dependent analysis. Even if this can be done, both methods will provably give looser bounds with smaller compatibility regions than that in Theorem 4.1. See Appendix B.3 for a detailed comparison with stability-based bounds and Appendix B.4 for discussions on uniform convergence bounds. We also provide a comparison with previous analysis of early stopping (Yao et al., 2007; Lin & Rosasco, 2017; Pillaud-Vivien et al., 2018) in Appendix B.5.

5 Proof Sketch and Techniques

5.1 A Time Variant Bound

We further introduce an additional type of effective dimension besides k_0, k_1, which is time variant and is utilized to track the optimization dynamics.

Definition 5.1 (Effective Dimensions). *Given a feature covariance matrix $\mathbf{\Sigma}$, define the effective dimension k_2 as*

$$k_2 \triangleq \min\left\{ l \geq 0 : \sum_{i>l} \lambda_i + n\lambda_{l+1} \leq c_2 c(t,n) \sum_{i>0} \lambda_i \right\}, \tag{9}$$

where c_2 is a constant independent of the dimension p, sample size n, and time t. The term $c(t,n)$ is a function to be discussed later. When the context is clear, we omit its dependencies on $c_2, c(t,n), n, \mathbf{\Sigma}$ and only denote it by k_2.

Based on the effective rank and effective dimensions defined above, we provide a time-variant bound in Theorem 5.1 for overparameterized linear regression, which further leads to the compatibility argument in Theorem 4.1. Compared to the existing bound (Bartlett et al., 2019), Theorem 5.1 focuses on investigating the role of training epoch t in the excess risk, and is of independent interest.

Theorem 5.1 (Time Variant Bound). *Suppose Assumption 1, 2 and 3 hold. Fix a function $c(t,n)$. Given $\delta \in (0,1)$, assume that $k_0 \leq \frac{n}{c}$, $\log\frac{1}{\delta} \leq \frac{n}{c}$, $0 < \lambda \leq \frac{1}{c\sum_{i>0}\lambda_i}$ for a large enough constant c. Then with probability at least $1-\delta$, we have for any $t \in \mathbb{N}$,*

$$R(\boldsymbol{\theta}_t) \lesssim B(\boldsymbol{\theta}_t) + V(\boldsymbol{\theta}_t), \tag{10}$$

where

$$B(\boldsymbol{\theta}_t) = \|\boldsymbol{\theta}^*\|^2 \left(\frac{1}{\lambda t} + \|\mathbf{\Sigma}\| \max\left\{ \sqrt{\frac{r(\mathbf{\Sigma})}{n}}, \frac{r(\mathbf{\Sigma})}{n}, \sqrt{\frac{\log(\frac{1}{\delta})}{n}} \right\} \right), \tag{11}$$

$$V(\boldsymbol{\theta}_t) = \sigma_y^2 \log\left(\frac{1}{\delta}\right) \left(\frac{k_1}{n} + \frac{k_2}{c(t,n)n} + c(t,n)\left(\frac{\lambda t}{n} \sum_{i>0} \lambda_i \right)^2 \right). \tag{12}$$

[4] It is worth mentioning that Bartlett et al. (2019) and this paper study the generalization behavior of different models. The comparison here and in the following sections aims to validate the meaningfulness of compatibility analysis, instead of beating their results.

Table 1: **Comparisons of excess risk bound with Bartlett et al. (2019) and Zou et al. (2021).** We provide four types of feature covariance with eigenvalues λ_k, including *Inverse Polynomial* ($\lambda_k = \frac{1}{k^\alpha}$, $\alpha > 1$), *Inverse Log Polynomial* ($\lambda_k = \frac{1}{k \log^\beta (k+1)}$, $\beta > 1$), *Constant* ($\lambda_k = \frac{1}{n^{1+\varepsilon}}$, $1 \le k \le n^{1+\varepsilon}$, $\varepsilon > 0$), and *Piecewise Constant* ($\lambda_k = \frac{1}{s}$ if $1 \le k \le s$ and $\lambda_k = \frac{1}{d-s}$ if $s+1 \le k \le d$, where $s = n^r, d = n^q, 0 < r \le 1, q > 1$). In light of these bounds, ours outperforms Bartlett et al. (2019) in all the cases, and outperforms Zou et al. (2021) in Constant / Piecewise Constant cases if $\varepsilon < \frac{1}{2}$ and $q < \min\{2-r, \frac{3}{2}\}$. We refer to Appendix B for more details.

Distributions	Ours	Bartlett et al. (2019)	Zou et al. (2021)
Inverse polynomial	$O\left(n^{-\min\left\{\frac{\alpha-1}{\alpha}, \frac{1}{2}\right\}}\right)$	$O(1)$	$O\left(n^{-\frac{\alpha-1}{\alpha}}\right)$
Inverse log polynomial	$O\left(\frac{1}{\log^\beta n}\right)$	$o(1)$	$O\left(\frac{1}{\log^\beta n}\right)$
Constant	$O\left(n^{-\frac{1}{2}}\right)$	$O\left(n^{-\min\left\{\varepsilon, \frac{1}{2}\right\}}\right)$	$O\left(n^{-\min\{\varepsilon, 1\}}\right)$
Piecewise constant	$O\left(n^{-\min\left\{1-r, \frac{1}{2}\right\}}\right)$	$O\left(n^{-\min\left\{1-r, q-1, \frac{1}{2}\right\}}\right)$	$O\left(n^{-\min\{1-r, q-1\}}\right)$

We provide a high-level intuition behind Theorem 5.1. We decompose $R(\boldsymbol{\theta}_t)$ into the bias term and the variance term. The variance term is then split into the leading part and tailing part based on the sprctrum of the feature covariance $\boldsymbol{\Sigma}$. The eigenvalues in the tailing part will cause the variance term in the excess risk of the min-norm interpolating solution to be $\Omega(1)$ for fast decaying spectrum, as is the case in (Bartlett et al., 2019). However, since the convergence in the tailing eigenspace is slower compared with the leading eigenspace, a proper early stopping strategy will prevent the overfitting in the tailing eigenspace and meanwhile avoid underfitting in the leading eigenspace.

The $c(t,n)$ Principle. It is worth emphasizing that our bound holds for arbitrary positive function $c(t,n)$. Therefore, one can fine-tune the generalization bound by choosing a proper $c(t,n)$. In the subsequent sections, we show how to derive consistent risk bounds for different time t, based on different choices of $c(t,n)$. We present the case of choosing a constant $c(t,n)$ in the next section. We leave the case of choosing a varying $c(t,n)$ to Appendix B.7.

5.2 Varying t, Constant $c(t,n)$

Theorem 5.1 provides an excess risk upper bound uniformly for $t \in \mathbb{N}$. However, it is still non-trivial to derive Theorem 4.1, where the remaining question is to decide the term $c(t,n)$. The following corollary shows the generalization bound when setting $c(t,n)$ to a constant.

Corollary 5.1 (Constant $c(t,n)$)**.** *Let Assumption 1, 2 and 3 hold. Fix a constant $c(t,n)$. Suppose $k_0 = O(n)$, $k_1 = o(n)$, $r(\boldsymbol{\Sigma}) = o(n)$, $\lambda = O\left(\frac{1}{\sum_{i>0} \lambda_i}\right)$. Then there exists a sequence of positive constants $\{\delta_n\}_{n \ge 0}$ which converge to 0, such that with probability at least $1 - \delta_n$, the excess risk is consistent for $t \in \left(\omega\left(\frac{1}{\lambda}\right), o\left(\frac{n}{\lambda}\right)\right)$, i.e.*

$$R(\boldsymbol{\theta}_t) = o(1). \tag{13}$$

Furthermore, for any positive constant δ, with probability at least $1 - \delta$, the minimal excess risk on the training trajectory can be bounded as

$$\min_t R(\boldsymbol{\theta}_t) \lesssim \frac{\max\{\sqrt{r(\boldsymbol{\Sigma})}, 1\}}{\sqrt{n}} + \frac{\max\{k_1, 1\}}{n}. \tag{14}$$

Lemma 5.1 below shows that $k_1 = o(n)$ always holds for fixed distribution. Therefore, combining Corollary 5.1 and Lemma 5.1 completes the proof of Theorem 4.1.

Lemma 5.1. *For any fixed (i.e. independent of sample size n) feature covariance $\boldsymbol{\Sigma}$ satisfying assumption 1, we have $k_1(n) = o(n)$.*

Example Distributions We apply the bound in Corollary 5.1 to several data distributions. These distributions are widely discussed in (Bartlett et al., 2019; Zou et al., 2021). We also derive the

Table 2: **The effective dimension k_1, the optimal early stopping excess risk, and the min-norm excess risk for different feature distributions, with sample size $n = 100$, $p = 1000$.** The table shows that early stopping solutions generalize significantly better than min-norm interpolators, and reveals a positive correlation between the effective dimension k_1 and excess risk of early stopping solution. We calculate the 95% confidence interval for each excess risk.

DISTRIBUTIONS	k_1	OPTIMAL EXCESS RISK	MIN-NORM EXCESS RISK
$\lambda_i = \frac{1}{i}$	$\Theta\left(n\right)$	2.399 ± 0.0061	24.071 ± 0.2447
$\lambda_i = \frac{1}{i^2}$	$\Theta\left(n^{\frac{1}{2}}\right)$	0.214 ± 0.0050	43.472 ± 0.6463
$\lambda_i = \frac{1}{i^3}$	$\Theta\left(n^{\frac{1}{3}}\right)$	0.077 ± 0.0005	10.401 ± 0.2973
$\lambda_i = \frac{1}{i\log(i+1)}$	$\Theta\left(\frac{n}{\log n}\right)$	0.697 ± 0.0053	89.922 ± 0.9591
$\lambda_i = \frac{1}{i\log^2(i+1)}$	$\Theta\left(\frac{n}{\log^2 n}\right)$	0.298 ± 0.0054	82.413 ± 0.9270
$\lambda_i = \frac{1}{i\log^3(i+1)}$	$\Theta\left(\frac{n}{\log^3 n}\right)$	0.187 ± 0.0047	38.145 ± 0.5862

existing excess risk bounds, which focus on the min-norm interpolator (Bartlett et al., 2019) and one-pass SGD iterate (Zou et al., 2021), of these distributions and compare them with our theorem. The results are summarized in Table 1, which shows that the bound in Corollary 5.1 outperforms previous results for a general class of distributions. We defer the details to Appendix B.

6 EXPERIMENTS

In this section, we provide numerical studies of overparameterized linear regression problems. We consider overparameterized linear regression instances with input dimension $p = 1000$, sample size $n = 100$. The features are sampled from Gaussian distribution with different covariances. The empirical results **(a.) demonstrate the benefits of trajectory analysis underlying the definition of compatibility,** since the optimal excess risk along the algorithm trajectory is significantly lower than that of the min-norm interpolator **(b.) validate the statements in Corollary 5.1**, since the optimal excess risk is lower when the eigenvalues of feature covariance decay faster. We refer to Appendix C for detailed setups, additional results and discussions.

Observation One: Early stopping solution along the training trajectory generalizes significantly better than the min-norm interpolator. We calculate the excess risk of optimal early stopping solutions and min-norm interpolators from 1000 independent trials and list the results in Table 2. The results illustrate that the early stopping solution on the algorithm trajectory enjoys much better generalization properties. This observation corroborates the meaningfulness of compatibility and the importance of data-dependent training trajectory in generalization analysis.

Observation Two: The faster covariance spectrum decays, the lower optimal excess risk is. Table 2 also illustrates a positive correlation between the decaying rate of λ_i and the generalization performance of the early stopping solution. This accords with Theorem 5.1, showing that the excess risk is better for a smaller effective dimension k_1, where small k_1 indicates a faster-decaying eigenvalue λ_i. We additionally note that such a phenomenon also illustrates the difference between min-norm and early stopping solutions in linear regression, since Bartlett et al. (2019) demonstrate that the min-norm solution is not consistent when the eigenvalues decay too fast. By comparison, early stopping solutions do not suffer from this restriction.

7 CONCLUSION

In this paper, we propose the concept of data-algorithm compatibility, and study compatibility for overparameterized linear regression with gradient descent. Our theoretical and empirical results demonstrate that compatibility eases the assumptions and broadens the scope of generalization by fully exploiting the data information and the algorithm information. Despite linear cases in this paper, compatibility can be a much more general concept. Therefore, we believe this paper will motivate more work on data-dependent trajectory analysis.

REFERENCES

Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A continuous-time view of early stopping for least squares regression. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 2019. URL `http://proceedings.mlr.press/v89/ali19a.html`.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 254–263. PMLR, 2018. URL `http://proceedings.mlr.press/v80/arora18b.html`.

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`.

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *CoRR*, abs/2106.15853, 2021. URL `https://arxiv.org/abs/2106.15853`.

Pradeep Kr. Banerjee and Guido Montúfar. Information complexity and generalization bounds. *CoRR*, abs/2105.01747, 2021. URL `https://arxiv.org/abs/2105.01747`.

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. *CoRR*, abs/1906.11300, 2019. URL `http://arxiv.org/abs/1906.11300`.

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`.

Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

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`.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 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/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large stepsizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

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.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1): 1–50, 2010.

Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9770–9780, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/05a624166c8eb8273b8464e8d9cb5bd9-Abstract.html.

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.

Mahsa Forouzesh and Patrick Thiran. Disparity between batches as a signal for early stopping. In Nuria Oliver, Fernando Pérez-Cruz, Stefan Kramer, Jesse Read, and José Antonio Lozano (eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part II*, volume 12976 of *Lecture Notes in Computer Science*, pp. 217–232. Springer, 2021. doi: 10.1007/978-3-030-86520-7_14. URL https://doi.org/10.1007/978-3-030-86520-7_14.

Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *CoRR*, abs/2202.05928, 2022. URL https://arxiv.org/abs/2202.05928.

L. Lo Gerfo, Lorenzo Rosasco, Francesca Odone, Ernesto De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Comput.*, 20(7):1873–1897, 2008. doi: 10.1162/neco.2008.05-07-517. URL https://doi.org/10.1162/neco.2008.05-07-517.

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.

Wei Hu, Lechao Xiao, Ben Adlam, and Jeffrey Pennington. The surprising simplicity of the early-time learning dynamics of neural networks. 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/c6dfc6b7c601ac2978357b7a81e2d7ae-Abstract.html.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

Ziwei Ji, Justin D. Li, and Matus Telgarsky. Early-stopped neural networks are consistent. 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. 1805–1817, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0e1ebad68af7f0ae4830b7ac92bc3c6f-Abstract.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, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pp. 443–457. Springer, 2000.

Ilja Kuzborskij and Csaba Szepesvári. Nonparametric regression with shallow overparameterized neural networks trained by GD with early stopping. In Mikhail Belkin and Samory Kpotufe (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2853–2890. PMLR, 2021. URL http://proceedings.mlr.press/v134/kuzborskij21a.html.

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, 2020a. URL https://openreview.net/forum?id=SkxxtgHKPS.

Jiangyuan Li, Thanh V. Nguyen, Chinmay Hegde, and Ka Wai Wong. Implicit sparse regularization: The impact of depth and early stopping. 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. 28298–28309, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ee188463935a061dee6df8bf449cb882-Abstract.html.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 4313–4324. PMLR, 2020b. URL http://proceedings.mlr.press/v108/li20j.html.

Zhu Li, Weijie Su, and Dino Sejdinovic. Benign overfitting and noisy features. *CoRR*, abs/2008.02901, 2020c. URL https://arxiv.org/abs/2008.02901.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel ”ridgeless” regression can generalize. *CoRR*, abs/1808.00387, 2018. URL http://arxiv.org/abs/1808.00387.

Junhong Lin and Volkan Cevher. Optimal rates for spectral-regularized algorithms with least-squares regression over hilbert spaces. *CoRR*, abs/1801.06720, 2018. URL http://arxiv.org/abs/1801.06720.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *J. Mach. Learn. Res.*, 18:97:1–97:47, 2017. URL http://jmlr.org/papers/v18/17-176.html.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. 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=SJeLIgBKPS.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. 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. 12978–12991, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/6c351da15b5e8a743a21ee96a86e25df-Abstract.html.

David A. McAllester. Simplified pac-bayesian margin bounds. In Bernhard Schölkopf and Manfred K. Warmuth (eds.), *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, volume 2777 of *Lecture Notes in Computer Science*, pp. 203–215. Springer, 2003. doi: 10.1007/978-3-540-45167-9_16. URL https://doi.org/10.1007/978-3-540-45167-9_16.

David A. McAllester. A pac-bayesian tutorial with A dropout bound. *CoRR*, abs/1307.2118, 2013. URL http://arxiv.org/abs/1307.2118.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012. ISBN 978-0-262-01825-8. URL http://mitpress.mit.edu/books/foundations-machine-learning-0.

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.

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, 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.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. Pac-bayes bounds with data dependent priors. *J. Mach. Learn. Res.*, 13:3507–3531, 2012. URL http://dl.acm.org/citation.cfm?id=2503353.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *J. Mach. Learn. Res.*, 22:227:1–227:40, 2021. URL http://jmlr.org/papers/v22/20-879.html.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis R. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8125–8135, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/10ff0b5e85e5b85cc3095d431d8c08b4-Abstract.html.

Lutz Prechelt. Early stopping - but when? In Grégoire Montavon, Genevieve B. Orr, and Klaus-Robert Müller (eds.), *Neural Networks: Tricks of the Trade - Second Edition*, volume 7700 of *Lecture Notes in Computer Science*, pp. 53–67. Springer, 2012. doi: 10.1007/978-3-642-35289-8_5. URL https://doi.org/10.1007/978-3-642-35289-8_5.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15(1):335–366, 2014. doi: 10.5555/2627435.2627446. URL https://dl.acm.org/doi/10.5555/2627435.2627446.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. Pac-bayes analysis beyond the usual bounds. 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/c3992e9a68c5ae12bd18488bc579b30d-Abstract.html.

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 1630–1638, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/1587965fb4d4b5afe8428a4a024feb0d-Abstract.html.

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.

Matthias W. Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *J. Mach. Learn. Res.*, 3:233–269, 2002. URL http://jmlr.org/papers/v3/seeger02a.html.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. 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/6cfe0e6127fa25df2a0ef2ae1067d915-Abstract.html.

Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. Evaluating machine accuracy on imagenet. 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. 8634–8644. PMLR, 2020. URL `http://proceedings.mlr.press/v119/shankar20c.html`.

John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a bayesian estimator. In Yoav Freund and Robert E. Schapire (eds.), *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT 1997, Nashville, Tennessee, USA, July 6-9, 1997*, pp. 2–9. ACM, 1997. doi: 10.1145/267460.267466. URL `https://doi.org/10.1145/267460.267466`.

Ruoqi Shen, Liyao Gao, and Yi-An Ma. On optimal early stopping: Over-informative versus under-informative parametrization. *CoRR*, abs/2202.09885, 2022. URL `https://arxiv.org/abs/2202.09885`.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nat.*, 550(7676):354–359, 2017. doi: 10.1038/nature24270. URL `https://doi.org/10.1038/nature24270`.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The implicit bias of gradient descent on separable data. 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=r1q7n9gAb`.

Pierre Stock and Moustapha Cissé. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pp. 504–519. Springer, 2018. doi: 10.1007/978-3-030-01231-1_31. URL `https://doi.org/10.1007/978-3-030-01231-1_31`.

Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Trans. Inf. Theory*, 60(9):5716–5735, 2014. doi: 10.1109/TIT.2014.2332531. URL `https://doi.org/10.1109/TIT.2014.2332531`.

Jiaye Teng, Jianhao Ma, and Yang Yuan. Towards understanding generalization via decomposing excess risk dynamics. *CoRR*, abs/2106.06153, 2021. URL `https://arxiv.org/abs/2106.06153`.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. The statistical complexity of early-stopped mirror descent. 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/024d2d699e6c1a82c9ba986386f4d824-Abstract.html`.

Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of gaussian mixtures. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pp. 4030–4034. IEEE, 2021. doi: 10.1109/ICASSP39728.2021.9413946. URL `https://doi.org/10.1109/ICASSP39728.2021.9413946`.

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.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: From single to multi-labels, from global to localized labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 2340–2350. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yun_Re-Labeling_ImageNet_From_Single_to_Multi-Labels_From_Global_to_Localized_CVPR_2021_paper.html.

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.

Lijia Zhou, Danica J. Sutherland, and Nati Srebro. On uniform convergence and low-norm interpolation learning. 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/4cc5400e63624c44fadeda99f57588a6-Abstract.html.

Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Benign overfitting of constant-stepsize SGD for linear regression. In Mikhail Belkin and Samory Kpotufe (eds.), *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4633–4635. PMLR, 2021. URL http://proceedings.mlr.press/v134/zou21a.html.

Supplementary Materials

A Proofs for the Main Results

We first sketch the proof in section A.1 and give some preliminary lemmas A.2. The following sections A.3, A.4 and A.5 are devoted to the proof of Theorem 5.1. The proof of Theorem 4.1 is given in A.6.

A.1 Proof Sketch

We start with a standard bias-variance decomposition following Bartlett et al. (2019), which derives that the time-variant excess risk $R(\boldsymbol{\theta}_t)$ can be bounded by a bias term and a variance term. We refer to Appendix A.3 for more details.

For the bias part, we first decompose it into an optimization error and an approximation error. For the optimization error, we use the spectrum analysis to bound it with $O\left(1/t\right)$ where t denotes the time. For the approximation error, we bound it with $O\left(1/\sqrt{n}\right))$ where n denotes the sample size, inspired by Bartlett et al. (2019). We refer to Appendix A.4 for more details.

For the variance part, a key step is to bound the term $(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top)^t$, where $\boldsymbol{X}$ is the feature matrix. The difficulty arises from the different scales of the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$, where the largest eigenvalue has order $\Theta(n)$ while the smallest eigenvalue has order $O(1)$, according to Lemma 10 in Bartlett et al. (2019). To overcome this issue, we divide the matrix $\boldsymbol{X}\boldsymbol{X}^\top$ based on whether its eigenvalues is larger than $c(t,n)$, which is a flexible term dependent on time t and sample size n. Therefore, we split the variance term based on eigenvalues of covariance matrix $\boldsymbol{\Sigma}$ (leading to the k_1-related term) and based on the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$ (leading to the k_2-related term). We refer to Appendix A.5 for more details.

A.2 Preliminaries

The following result comes from Bartlett et al. (2019), which bounds the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$.

Lemma A.1. *(Lemma 10 in Bartlett et al. (2019)) For any σ_x, there exists a constant c, such that for any $0 \le k < n$, with probability at least $1 - e^{-\frac{n}{c}}$,*

$$\mu_{k+1} \le c\left(\sum_{i>k} \lambda_i + \lambda_{k+1} n\right). \tag{15}$$

This implies that as long as the step size λ is small than a threshold independent of sample size n, gradient descent is stable.

Corollary A.1. *There exists a constant c, such that with probability at least $1 - e^{-\frac{n}{c}}$, for any $0 \le \lambda \le \frac{1}{c\sum_{i>0}\lambda_i}$ we have*

$$\boldsymbol{O} \preceq \boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X} \preceq \boldsymbol{I}. \tag{16}$$

Proof. The right hand side of the inequality is obvious since $\lambda > 0$. For the left hand side, we have to show that the eigenvalues of $\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}$ is non-negative. since $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^\top$ have the same non-zero eigenvalues, we know that with probability at least $1 - e^{-\frac{n}{c}}$, the smallest eigenvalue of $\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}$ can be lower bounded by

$$1 - \frac{\lambda}{n}\mu_1 \ge 1 - c\lambda\left(\frac{\sum_{i>0}\lambda_i}{n} + \lambda_{k+1}\right) \ge 1 - 2c\lambda\sum_{i>0}\lambda_i \ge 0. \tag{17}$$

where the second inequality uses lemma A.1, and the last inequality holds if $\lambda \le \frac{1}{2c\sum_{i>0}\lambda_i}$. □

A.3 Proof for the Bias-Variance Decomposition

Let $\boldsymbol{X}^\dagger$ denote the Moore–Penrose pseudoinverse of matrix $\boldsymbol{X}$. The following lemma gives a closed form expression for $\boldsymbol{\theta}_t$.

Lemma A.2. *The dynamics of $\{\boldsymbol{\theta}_t\}_{t\geq 0}$ satisfies*

$$\boldsymbol{\theta}_t = \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t (\boldsymbol{\theta}_0 - \boldsymbol{X}^\dagger\boldsymbol{Y}) + \boldsymbol{X}^\dagger\boldsymbol{Y}. \tag{18}$$

Proof. We prove the lemma using induction. The equality holds at $t = 0$ as both sides are $\boldsymbol{\theta}_0$. Recall that $\boldsymbol{\theta}_t$ is updated as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \frac{\lambda}{n}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}_t). \tag{19}$$

Suppose that the dynamic holds up to the t-th step. Plug the expression for $\boldsymbol{\theta}_t$ into the above recursion and note that $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\dagger = \boldsymbol{X}^\top$, we get

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)\boldsymbol{\theta}_t + \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{Y} \\ &= \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{t+1}(\boldsymbol{\theta}_0 - \boldsymbol{X}^\dagger\boldsymbol{Y}) + \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)\boldsymbol{X}^\dagger\boldsymbol{Y} + \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{Y} \\ &= \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{t+1}(\boldsymbol{\theta}_0 - \boldsymbol{X}^\dagger\boldsymbol{Y}) + \boldsymbol{X}^\dagger\boldsymbol{Y}.\end{aligned} \tag{20}$$

which finishes the proof. □

Next we prove two identities which will be used in further proof.

Lemma A.3. *The following two identities hold for any matrix X and non-negative integer t:*

$$\boldsymbol{I} - \boldsymbol{X}^\dagger\boldsymbol{X} + \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\boldsymbol{X}^\dagger\boldsymbol{X} = \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t, \tag{21}$$

$$\left[\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\right]\boldsymbol{X}^\dagger\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{X}^\top\left[\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top\right)^t\right]. \tag{22}$$

Proof. Note that $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\dagger = \boldsymbol{X}^\top$, we can expand the left hand side of the first identity above using binomial theorem and eliminate the pseudo-inverse $\boldsymbol{X}^\dagger$:

$$\begin{aligned}&\boldsymbol{I} - \boldsymbol{X}^\dagger\boldsymbol{X} + \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\boldsymbol{X}^\dagger\boldsymbol{X} \\ &= \boldsymbol{I} - \boldsymbol{X}^\dagger\boldsymbol{X} + \sum_{k=0}^{t}\binom{t}{k}\left(-\frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^k\boldsymbol{X}^\dagger\boldsymbol{X} \\ &= \boldsymbol{I} - \boldsymbol{X}^\dagger\boldsymbol{X} + \boldsymbol{X}^\dagger\boldsymbol{X} + \sum_{k=1}^{t}\binom{t}{k}\left(-\frac{\lambda}{n}\right)^k(\boldsymbol{X}^\top\boldsymbol{X})^{k-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\dagger\boldsymbol{X} \\ &= \boldsymbol{I} + \sum_{k=1}^{t}\binom{t}{k}\left(-\frac{\lambda}{n}\right)^k(\boldsymbol{X}^\top\boldsymbol{X})^k \\ &= \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t.\end{aligned} \tag{23}$$

The second identity can be proved in a similar way:

$$\begin{aligned}
&\left[\boldsymbol{I}-\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\right]\boldsymbol{X}^\dagger\boldsymbol{X}\boldsymbol{X}^\top \\
&=-\sum_{k=1}^{t}\binom{t}{k}\left(-\frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^k\boldsymbol{X}^\dagger\boldsymbol{X}\boldsymbol{X}^\top \\
&=-\sum_{k=1}^{t}\binom{t}{k}\left(-\frac{\lambda}{n}\right)^k(\boldsymbol{X}^\top\boldsymbol{X})^{k-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\dagger\boldsymbol{X}\boldsymbol{X}^\top \\
&=-\sum_{k=1}^{t}\binom{t}{k}\left(-\frac{\lambda}{n}\right)^k(\boldsymbol{X}^\top\boldsymbol{X})^{k-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top \\
&=-\sum_{k=1}^{t}\binom{t}{k}\left(-\frac{\lambda}{n}\right)^k\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^k \\
&=\boldsymbol{X}^\top\left[\boldsymbol{I}-\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top\right)^t\right].
\end{aligned} \tag{24}$$

□

We are now ready to prove the main result of this section.

Lemma A.4. *The excess risk at the t-th epoch can be upper bounded as*

$$R(\boldsymbol{\theta}_t)\le\boldsymbol{\theta}^{*\top}\boldsymbol{B}\boldsymbol{\theta}^*+\boldsymbol{\varepsilon}^\top\boldsymbol{C}\boldsymbol{\varepsilon}, \tag{25}$$

where

$$\boldsymbol{B}=\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\boldsymbol{\Sigma}\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t, \tag{26}$$

$$\boldsymbol{C}=\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\left[\boldsymbol{I}-\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top\right)^t\right]\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\left[\boldsymbol{I}-\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top\right)^t\right]\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}, \tag{27}$$

which characterizes bias term and variance term in the excess risk. Furthermore, there exists constant c such that with probability at least $1-\delta$ over the randomness of $\boldsymbol{\varepsilon}$, we have

$$\boldsymbol{\varepsilon}^\top\boldsymbol{C}\boldsymbol{\varepsilon}\le c\sigma_y^2\log\frac{1}{\delta}\operatorname{Tr}[\boldsymbol{C}]. \tag{28}$$

Proof. First note that $\boldsymbol{X}\boldsymbol{X}^\top$ is invertible by Assumption 3. Express the excess risk as follows

$$\begin{aligned}
R(\boldsymbol{\theta}_t)&=\frac{1}{2}\mathbb{E}[(y-\boldsymbol{x}^\top\boldsymbol{\theta}_t)^2-(y-\boldsymbol{x}^\top\boldsymbol{\theta}^*)^2] \\
&=\frac{1}{2}\mathbb{E}[(y-\boldsymbol{x}^\top\boldsymbol{\theta}^*+\boldsymbol{x}^\top\boldsymbol{\theta}^*-\boldsymbol{x}^\top\boldsymbol{\theta}_t)^2-(y-\boldsymbol{x}^\top\boldsymbol{\theta}^*)^2] \\
&=\frac{1}{2}\mathbb{E}[(\boldsymbol{x}^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*))^2+2(y-\boldsymbol{x}^\top\boldsymbol{\theta}^*)(\boldsymbol{x}^\top\boldsymbol{\theta}^*-\boldsymbol{x}^\top\boldsymbol{\theta}_t)] \\
&=\frac{1}{2}\mathbb{E}[\boldsymbol{x}^\top(\boldsymbol{\theta}_t-\boldsymbol{\theta}^*)]^2.
\end{aligned} \tag{29}$$

Recall that $\boldsymbol{\theta}_0=0$ and $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\theta}^*+\boldsymbol{\varepsilon}$ and we can further simplify the formula for $\boldsymbol{\theta}_t$ in lemma A.2:

$$\begin{aligned}
\boldsymbol{\theta}_t&=\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t(\boldsymbol{\theta}_0-\boldsymbol{X}^\dagger\boldsymbol{Y})+\boldsymbol{X}^\dagger\boldsymbol{Y} \\
&=\left[\boldsymbol{I}-\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\right]\boldsymbol{X}^\dagger(\boldsymbol{X}\boldsymbol{\theta}^*+\boldsymbol{\varepsilon}).
\end{aligned} \tag{30}$$

Plug it into the above expression for $R(\boldsymbol{\theta}_t)$, we have

$$\begin{aligned} R(\boldsymbol{\theta}_t) &= \frac{1}{2}\mathbb{E}\left[\boldsymbol{x}^\top \left[\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\right]\boldsymbol{X}^\dagger(\boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}) - \boldsymbol{x}^\top\boldsymbol{\theta}^*\right]^2 \\ &= \frac{1}{2}\mathbb{E}\left[\boldsymbol{x}^\top\left(\boldsymbol{X}^\dagger\boldsymbol{X} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\boldsymbol{X}^\dagger\boldsymbol{X} - \boldsymbol{I}\right)\boldsymbol{\theta}^*\right. \\ &\left. + \boldsymbol{x}^\top\left[\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\right]\boldsymbol{X}^\dagger\boldsymbol{\varepsilon}\right]^2. \end{aligned} \tag{31}$$

Applying lemma A.3, we obtain

$$\begin{aligned} R(\boldsymbol{\theta}_t) &= \frac{1}{2}\mathbb{E}\left[-\boldsymbol{x}^\top\left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\boldsymbol{\theta}^* + \boldsymbol{x}^\top\boldsymbol{X}^\top\left[\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top\right)^t\right](\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\varepsilon}\right]^2 \\ &\leq \mathbb{E}\left[\boldsymbol{x}^\top\left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^t\boldsymbol{\theta}^*\right]^2 + \mathbb{E}\left[\boldsymbol{x}^\top\boldsymbol{X}^\top\left[\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top\right)^t\right](\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{\varepsilon}\right]^2 \\ &:= \boldsymbol{\theta}^{*\top}\boldsymbol{B}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}^\top\boldsymbol{C}\boldsymbol{\varepsilon}. \end{aligned} \tag{32}$$

which proves the first claim in the lemma. The second part of the theorem directly follows from lemma 18 in Bartlett et al. (2019). □

A.4 Proof for the Bias Upper Bound

The next lemma guarantees that the sample covariace matrix $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}$ concentrates well around $\boldsymbol{\Sigma}$.

Lemma A.5. *(Lemma 35 in Bartlett et al. (2019)) There exists constant c such that for any* $0 < \delta < 1$ *with probability as least* $1 - \delta$,

$$\left\|\boldsymbol{\Sigma} - \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right\| \leq c\|\boldsymbol{\Sigma}\|\max\left\{\sqrt{\frac{r(\boldsymbol{\Sigma})}{n}}, \frac{r(\boldsymbol{\Sigma})}{n}, \sqrt{\frac{\log(\frac{1}{\delta})}{n}}, \frac{\log(\frac{1}{\delta})}{n}\right\}. \tag{33}$$

The following inequality will be useful in our proof to characterize the decaying rate of the bias term with t.

Lemma A.6. *For any positive semidefinite matrix* $\boldsymbol{P}$ *which satisfies* $\|\boldsymbol{P}\| \leq 1$, *we have*

$$\|\boldsymbol{P}(1 - \boldsymbol{P})^t\| \leq \frac{1}{t}. \tag{34}$$

Proof. Assume without loss of generality that $\boldsymbol{P}$ is diagonal. Then it suffices to consider seperately each eigenvalue σ of $\boldsymbol{P}$, and show that $\sigma(1-\sigma)^t \leq \frac{1}{t}$.

In fact, by AM-GM inequality we have

$$\sigma(1-\sigma)^t \leq \frac{1}{t}\left[\frac{t\sigma + (1-\sigma)t}{t+1}\right]^{t+1} \leq \frac{1}{t}, \tag{35}$$

which completes the proof. □

Next we prove the main result of this section.

Lemma A.7. *There exists constant c such that if* $0 \leq \lambda \leq \frac{1}{c\sum_{i>0}\lambda_i}$, *then for any* $0 < \delta < 1$, *with probability at least* $1 - \delta$ *the following bound on the bias term holds for any t*

$$\boldsymbol{\theta}^{*\top}\boldsymbol{B}\boldsymbol{\theta}^* \leq c\,\|\boldsymbol{\theta}^*\|^2\left(\frac{1}{\lambda t} + \|\boldsymbol{\Sigma}\|\max\left\{\sqrt{\frac{r(\boldsymbol{\Sigma})}{n}}, \frac{r(\boldsymbol{\Sigma})}{n}, \sqrt{\frac{\log(\frac{1}{\delta})}{n}}, \frac{\log(\frac{1}{\delta})}{n}\right\}\right). \tag{36}$$

Proof. The bias can be decomposed into the following two terms

$$\begin{aligned}\boldsymbol{\theta}^{*\top}\boldsymbol{B}\boldsymbol{\theta}^{*} = \boldsymbol{\theta}^{*\top}\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{t}\left(\boldsymbol{\Sigma}-\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{t}\boldsymbol{\theta}^{*} \\ +\,\boldsymbol{\theta}^{*\top}\left(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{2t}\boldsymbol{\theta}^{*}.\end{aligned} \tag{37}$$

For sufficiently small learning rate λ as given by corollary A.1, we know that with high probability

$$\left\|\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right\| \leq 1, \tag{38}$$

which together with lemma A.5 gives a high probability bound on the first term:

$$\begin{aligned}&\boldsymbol{\theta}^{*\top}\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{t}\left(\boldsymbol{\Sigma}-\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{t}\boldsymbol{\theta}^{*} \\ &\leq c\|\boldsymbol{\Sigma}\|\,\|\boldsymbol{\theta}^{*}\|^{2}\max\left\{\sqrt{\frac{r(\boldsymbol{\Sigma})}{n}},\frac{r(\boldsymbol{\Sigma})}{n},\sqrt{\frac{\log(\frac{1}{\delta})}{n}},\frac{\log(\frac{1}{\delta})}{n}\right\}.\end{aligned} \tag{39}$$

For the second term, invoke lemma A.6 with $\boldsymbol{P}=\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}$ and we get

$$\begin{aligned}\boldsymbol{\theta}^{*\top}\left(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{2t}\boldsymbol{\theta}^{*} &\leq \frac{1}{\lambda}\left\|\boldsymbol{\theta}^{*}\right\|^{2}\left\|\left(\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)\left(\boldsymbol{I}-\frac{\lambda}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{2t}\right\| \\ &\leq \frac{1}{2\lambda t}\left\|\boldsymbol{\theta}^{*}\right\|^{2}.\end{aligned} \tag{40}$$

Putting these two bounds together gives the proof for the main theorem.

□

A.5 Proof for the Variance Upper Bound

Recall that $\boldsymbol{X}=\boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{W}^{\top}$ is the singular value decomposition of data matrix $\boldsymbol{X}$, where $\boldsymbol{U}=(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n)$, $\boldsymbol{W}=(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_n)$, $\tilde{\boldsymbol{\Lambda}}=\text{diag}\{\mu_1,\cdots,\mu_n\}$ with $\mu_1\geq\mu_2\geq\cdots\mu_n$.

Recall that

$$\begin{aligned}k_0 &= \min\{l\geq 0:\lambda_{l+1}\leq\frac{c_0\sum_{i>l}\lambda_i}{n}\}, \\ k_1 &= \min\{l\geq 0:\lambda_{l+1}\leq\frac{c_1\sum_{i>0}\lambda_i}{n}\}, \\ k_2 &= \min\{l\geq 0:\sum_{i>l}\lambda_i+n\lambda_{l+1}\leq c_2c(t,n)\sum_{i>0}\lambda_i\}\},\end{aligned} \tag{41}$$

for some constant c_0,c_1,c_2 and function $c(t,n)$.

We further define

$$k_3=\min\{l\geq 0:\mu_{l+1}\leq c_3c(t,n)\sum_{i>0}\lambda_i\}, \tag{42}$$

for some constant c_3.

The next lemma shows that we can appropriately choose constants to ensure that $k_3\leq k_2$ holds with high probability, and in some specific cases we have $k_2\leq k_1$.

Lemma A.8. *For any function $c(t,n)$ and constant c_2, there exists constants c,c_3, such that $k_3\leq k_2$ with probability at least $1-e^{-\frac{n}{c}}$. Furthermore, if $c(t,n)$ is a positive constant function, for any c_1, there exists c_2 such that $k_2\leq k_1$.*

Proof. According to lemma A.1, there exists a constant c, with probability at least $1-e^{-\frac{n}{c}}$ we have

$$\mu_{k_2+1} \leq c(\sum_{i>k_2} \lambda_i + n\lambda_{k_2+1}) \leq cc_2 c(t,n) \sum_{i>0} \lambda_i. \tag{43}$$

Therefore, we know that $k_3 \leq k_2$ for $c_3 = cc_2$.

By the definition of k_1, we have

$$\sum_{i>k_1} \lambda_i + n\lambda_{k_1+1} \leq (c_1+1) \sum_{i>0} \lambda_i, \tag{44}$$

which implies that $k_2 \leq k_1$ for $c_2 = \frac{c_1+1}{c(t,n)}$, if $c(t,n)$ is a positive constant. □

Next we bound $\mathrm{Tr}[\boldsymbol{C}]$, which implies an upper bound on the variance term.

Theorem A.1. *There exist constants c, c_0, c_1, c_2 such that if $k_0 \leq \frac{n}{c}$, then with probability at least $1-e^{-\frac{n}{c}}$, the trace of the variance matrix C has the following upper bound for any t:*

$$\mathrm{Tr}[\boldsymbol{C}] \leq c\left(\frac{k_1}{n} + \frac{k_2}{c(t,n)n} + c(t,n)\left(\frac{\lambda t}{n}\sum_{i>0}\lambda_i\right)^2\right). \tag{45}$$

Proof. We divide the eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$ into two groups based on whether they are greater than $c_3 c(t,n)\sum_{i>0}\lambda_i$. The first group consists of $\mu_1 \cdots \mu_{k_3}$, and the second group consists of $\mu_{k_3+1}\cdots\mu_n$. For $1 \leq j \leq k_3$, we have

$$1 - \left(1 - \frac{\lambda}{n}\mu_j\right)^t \leq 1. \tag{46}$$

Therefore we have the following upper bound on $\left[\boldsymbol{I} - (\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top)^t\right]^2$:

$$\begin{aligned}
&\left[\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top\right)^t\right]^2 \\
=&\,\boldsymbol{U}\mathrm{diag}\left\{\left[1-\left(1-\frac{\lambda}{n}\mu_1\right)^t\right]^2 \cdots \left[1-\left(1-\frac{\lambda}{n}\mu_n\right)^t\right]^2\right\}\boldsymbol{U}^\top \\
\preceq&\,\boldsymbol{U}\mathrm{diag}\left\{\overbrace{1,\cdots 1}^{k_3 \text{ times}}, \overbrace{\left[1-\left(1-\frac{\lambda}{n}\mu_{k_3+1}\right)^t\right]^2, \cdots \left[1-\left(1-\frac{\lambda}{n}\mu_n\right)^t\right]^2}^{n-k_3 \text{ times}}\right\}\boldsymbol{U}^\top \\
=&\,\boldsymbol{U}\mathrm{diag}\left\{\overbrace{1,\cdots 1}^{k_3 \text{ times}}, \overbrace{0,\cdots 0}^{n-k_3 \text{ times}}\right\}\boldsymbol{U}^\top \\
+&\,\boldsymbol{U}\mathrm{diag}\left\{\overbrace{0,\cdots 0}^{k_3 \text{ times}}, \overbrace{\left[1-\left(1-\frac{\lambda}{n}\mu_{k_3+1}\right)^t\right]^2, \cdots \left[1-\left(1-\frac{\lambda}{n}\mu_n\right)^t\right]^2}^{n-k_3 \text{ times}}\right\}\boldsymbol{U}^\top.
\end{aligned} \tag{47}$$

For positive semidefinite matrices $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{R}$ which satisfies $\boldsymbol{Q} \preceq \boldsymbol{R}$, it holds that $\mathrm{Tr}[\boldsymbol{P}\boldsymbol{Q}] \leq \mathrm{Tr}[\boldsymbol{P}\boldsymbol{R}]$. It implies the following upperbound of $\mathrm{Tr}[\boldsymbol{C}]$:

$$
\begin{aligned}
&\operatorname{Tr}[\boldsymbol{C}] \\
&= \operatorname{Tr}\left[\left[\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n}\boldsymbol{X}\boldsymbol{X}^\top\right)^t\right]^2 \left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\right] \\
&\le \underbrace{\operatorname{Tr}\left[\boldsymbol{U}\operatorname{diag}\left\{\overbrace{1,\cdots 1}^{k_3\text{ times}},\ \overbrace{0,\cdots 0}^{n-k_3\text{ times}}\right\}\boldsymbol{U}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\right]}_{\textcircled{1}} \\
&\quad \underbrace{+ \operatorname{Tr}\left[\boldsymbol{U}\operatorname{diag}\left\{\overbrace{0,\cdots 0}^{k_3\text{ times}},\ \overbrace{\left[1-\left(1-\frac{\lambda}{n}\mu_{k_3+1}\right)^t\right]^2,\cdots\left[1-\left(1-\frac{\lambda}{n}\mu_n\right)^t\right]^2}^{n-k_3\text{ times}}\right\}\boldsymbol{U}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\right]}_{\textcircled{2}}.
\end{aligned}
\tag{48}
$$

Bounding $\textcircled{1}$

Noticing $\boldsymbol{X} = \boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{W}^\top$ and $\boldsymbol{\Sigma} = \sum_{i\ge 1}\lambda_i\boldsymbol{v}_i\boldsymbol{v}_i^\top$, we express the first term as sums of eigenvector products,

$$
\begin{aligned}
\textcircled{1} &= \operatorname{Tr}\left[\boldsymbol{U}\operatorname{diag}\left\{\overbrace{1,\cdots 1}^{k_3\text{ times}},\ \overbrace{0,\cdots 0}^{n-k_3\text{ times}}\right\}\boldsymbol{U}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{-2}\boldsymbol{X}\boldsymbol{\Sigma}\boldsymbol{X}^\top\right] \\
&= \operatorname{Tr}\left[\boldsymbol{U}\operatorname{diag}\left\{\overbrace{1,\cdots 1}^{k_3\text{ times}},\ \overbrace{0,\cdots 0}^{n-k_3\text{ times}}\right\}\boldsymbol{U}^\top\boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{-2}\boldsymbol{U}^\top\boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{W}^\top\boldsymbol{\Sigma}\boldsymbol{W}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{U}^\top\right] \\
&= \operatorname{Tr}\left[\operatorname{diag}\left\{\overbrace{1,\cdots 1}^{k_3\text{ times}},\ \overbrace{0,\cdots 0}^{n-k_3\text{ times}}\right\}\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{W}^\top\boldsymbol{\Sigma}\boldsymbol{W}\right] \\
&= \sum_{i\ge 1}\lambda_i\operatorname{Tr}\left[\operatorname{diag}\left\{\overbrace{1,\cdots 1}^{k_3\text{ times}},\ \overbrace{0,\cdots 0}^{n-k_3\text{ times}}\right\}\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{W}^\top\boldsymbol{v}_i\boldsymbol{v}_i^\top\boldsymbol{W}\right] \\
&= \sum_{i\ge 1}\sum_{1\le j\le k_3}\frac{\lambda_i}{\mu_j}\left(\boldsymbol{v}_i^\top\boldsymbol{w}_j\right)^2.
\end{aligned}
\tag{49}
$$

Next we divide the above summation into $1 \le i \le k_1$ and $i > k_1$. For the first part, notice that

$$
\begin{aligned}
\sum_{1\le j\le k_3}\frac{\lambda_i}{\mu_j}\left(\boldsymbol{v}_i^\top\boldsymbol{w}_j\right)^2 &\le \sum_{1\le j\le n}\frac{\lambda_i}{\mu_j}\left(\boldsymbol{v}_i^\top\boldsymbol{w}_j\right)^2 \\
&= \lambda_i\boldsymbol{v}_i^\top\left(\sum_{1\le j\le n}\frac{1}{\mu_j}\boldsymbol{w}_j\boldsymbol{w}_j^\top\right)\boldsymbol{v}_i \\
&= \lambda_i\boldsymbol{v}_i^\top\boldsymbol{W}\tilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{W}^\top\boldsymbol{v}_i \\
&= \lambda_i\boldsymbol{v}_i^\top\boldsymbol{W}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{U}^\top\boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{-2}\boldsymbol{U}^\top\boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{W}^\top\boldsymbol{v}_i \\
&= \lambda_i^2\tilde{\boldsymbol{x}}_i^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-2}\tilde{\boldsymbol{x}}_i,
\end{aligned}
\tag{50}
$$

where $\tilde{\boldsymbol{x}}_i$ is defined as $\tilde{\boldsymbol{x}}_i = \frac{\boldsymbol{X}\boldsymbol{v}_i}{\sqrt{\lambda_i}} = \frac{\boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{W}^\top\boldsymbol{v}_i}{\sqrt{\lambda_i}}$.

From the proof of lemma 11 in Bartlett et al. (2019), we know that for any σ_x, there exists a constant c_0 and c such that if $k_0 \le \frac{n}{c}$, with probability at least $1 - e^{-\frac{n}{c}}$ the first part can be bounded as

$$\sum_{1\le i\le k_1}\sum_{1\le j\le k_3}\frac{\lambda_i}{\mu_j}\left(\boldsymbol{v}_i^\top\boldsymbol{w}_j\right)^2 \le \sum_{1\le i\le k_1}\lambda_i^2\tilde{\boldsymbol{x}}_i(\boldsymbol{X}\boldsymbol{X}^\top)^{-2}\tilde{\boldsymbol{x}}_i \le c\frac{k_1}{n}, \tag{51}$$

which gives a bound for the first part.

For the second part we interchange the order of summation and have

$$\begin{aligned}
\sum_{i\ge k_1}\sum_{1\le j\le k_3}\frac{\lambda_i}{\mu_j}\left(\boldsymbol{v}_i^\top\boldsymbol{w}_j\right)^2 &= \sum_{1\le j\le k_3}\sum_{i\ge k_1}\frac{\lambda_i}{\mu_j}\left(\boldsymbol{v}_i^\top\boldsymbol{w}_j\right)^2\\
&\le \frac{1}{c_3c(t,n)\sum_{i>0}\lambda_i}\sum_{1\le j\le k_3}\sum_{i\ge k_1}\lambda_i\left(\boldsymbol{v}_i^\top\boldsymbol{w}_j\right)^2\\
&= \frac{\lambda_{k_1+1}}{c_3c(t,n)\sum_{i>0}\lambda_i}\sum_{1\le j\le k_3}\sum_{i\ge k_1}\left(\boldsymbol{v}_i^\top\boldsymbol{w}_j\right)^2\\
&\le \frac{\lambda_{k_1+1}}{c_3c(t,n)\sum_{i>0}\lambda_i}\sum_{1\le j\le k_3}1\\
&= \frac{\lambda_{k_1+1}k_3}{c_3c(t,n)\sum_{i>0}\lambda_i}\\
&\le c\frac{k_3}{c(t,n)n}.
\end{aligned} \tag{52}$$

for c large enough.

Putting 51 and 52 together, and noting that $k_3 \le k_2$ with high probability as given in lemma A.8, we know there exists a constant c that with probability at least $1-e^{-\frac{n}{c}}$,

$$\textcircled{1} \le c\frac{k_1}{n} + c\frac{k_2}{c(t,n)n}. \tag{53}$$

Bounding ②

Similar to the first step in bounding ①, we note that

$$\begin{aligned}
\textcircled{2} = \operatorname{Tr}\Bigg[&\boldsymbol{U}\operatorname{diag}\left\{\overbrace{0,\cdots 0}^{k_3\text{ times}},\overbrace{\left[1-\left(1-\frac{\lambda}{n}\mu_{k_3+1}\right)^t\right]^2,\cdots,\left[1-\left(1-\frac{\lambda}{n}\mu_n\right)^t\right]^2}^{n-k_3\text{ times}}\right\}\\
&\boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{-2}\boldsymbol{U}^\top\boldsymbol{U}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{W}^\top\boldsymbol{\Sigma}\boldsymbol{W}\tilde{\boldsymbol{\Lambda}}^{\frac{1}{2}}\boldsymbol{U}^\top\Bigg]\\
= \operatorname{Tr}\Bigg[&\operatorname{diag}\left\{\overbrace{0,\cdots 0}^{k_3\text{ times}},\overbrace{\frac{1}{\mu_{k_3+1}}\left[1-\left(1-\frac{\lambda}{n}\mu_{k_3+1}\right)^t\right]^2,\cdots,\frac{1}{\mu_n}\left[1-\left(1-\frac{\lambda}{n}\mu_n\right)^t\right]^2}^{n-k_3\text{ times}}\right\}\\
&\boldsymbol{W}^\top\boldsymbol{\Sigma}\boldsymbol{W}\Bigg].
\end{aligned} \tag{54}$$

From Bernoulli's inequality and the definition of k_3, for any $k_3+1\le j\le n$, we have

$$\frac{1}{\mu_k}\left[1-\left(1-\frac{\lambda}{n}\mu_k\right)^t\right]^2 \le \frac{1}{\mu_k}\left(\frac{\lambda}{n}\mu_k t\right)^2 = \left(\frac{\lambda t}{n}\right)^2\mu_k \le c_3\left(\frac{\lambda t}{n}\right)^2 c(t,n)\sum_{i>0}\lambda_i, \tag{55}$$

Hence,

$$
\begin{aligned}
② &\leq c_3 c(t,n) \left(\frac{\lambda t}{n}\right)^2 \sum_{i>0} \lambda_i \operatorname{Tr}[\boldsymbol{W}^\top \boldsymbol{\Sigma} \boldsymbol{W}] \\
&= c_3 c(t,n) \left(\frac{\lambda t}{n} \sum_{i>0} \lambda_i\right)^2 .
\end{aligned} \tag{56}
$$

Putting things together

From the bounds for ① and ② given above, we know that there exists a constant c such that with probability at least $1 - e^{-\frac{n}{c}}$, the trace of the variance matrix C has the following upper bound

$$
\operatorname{Tr}[C] \leq c \left(\frac{k_1}{n} + \frac{k_2}{c(t,n)n} + c(t,n) \left(\frac{\lambda t}{n} \sum_{i>0} \lambda_i \right)^2 \right). \tag{57}
$$

□

Proof of theorem 5.1. Lemma A.4, A.7 and Theorem A.1 gives the complete proof. Note that the high probability events in the proof are independent of the epoch number t, and this implies that the theorem holds uniformly for all $t \in \mathbb{N}$. □

A.6 Proof of Compatibility Results

Corollary A.2 (Corollary 5.1 restated). *Let Assumption 1, 2 and 3 hold. Fix a constant $c(t,n)$. Suppose $k_0 = O(n)$, $k_1 = o(n)$, $r(\boldsymbol{\Sigma}) = o(n)$, $\lambda = O\left(\frac{1}{\sum_{i>0}\lambda_i}\right)$. Then there exists a sequence of positive constants $\{\delta_n\}_{n\geq 0}$ which converge to 0, such that with probability at least $1-\delta_n$, the excess risk is consistent for $t \in \left(\omega\left(\frac{1}{\lambda}\right), o\left(\frac{n}{\lambda}\right)\right)$, i.e.*

$$
R(\boldsymbol{\theta}_t) = o(1).
$$

Furthermore, for any positive constant δ, with probability at least $1-\delta$, the minimal excess risk on the training trajectory can be bounded as

$$
\min_t R(\boldsymbol{\theta}_t) \lesssim \frac{\max\{\sqrt{r(\boldsymbol{\Sigma})}, 1\}}{\sqrt{n}} + \frac{\max\{k_1, 1\}}{n}.
$$

Proof. According to Lemma A.7, with probability at least $1 - \frac{\delta_n}{2}$, the following inequality holds for all t:

$$
B(\boldsymbol{\theta}_t) \lesssim \left(\frac{1}{\lambda t} + \max\left\{ \sqrt{\frac{r(\boldsymbol{\Sigma})}{n}}, \frac{r(\boldsymbol{\Sigma})}{n}, \sqrt{\frac{\log(\frac{1}{\delta_n})}{n}}, \frac{\log(\frac{1}{\delta_n})}{n} \right\} \right). \tag{58}
$$

If δ_n is chosen such that $\log\frac{1}{\delta_n} = o(n)$, we have that with probability at least $1 - \frac{\delta_n}{2}$, we have for all $t = \omega\left(\frac{1}{\lambda}\right)$:

$$
B(\boldsymbol{\theta}_t) = o(1), \tag{59}
$$

in the sample size n.

When $c(t,n)$ is a constant, we have $k_2 \leq k_1$ with high probability as given in lemma A.8. Therefore, according to Lemma A.4 and Theorem A.1, we know that if $\log\frac{1}{\delta_n} = O(n)$, with probability at least $1 - \frac{\delta_n}{2}$, the following bound holds for all t:

$$
V(\boldsymbol{\theta}_t) \lesssim \log\left(\frac{1}{\delta_n}\right) \left(\frac{k_1}{n} + \frac{\lambda^2 t^2}{n^2} \right). \tag{60}
$$

Since $k_1 = o(n)$, $t = o\left(\frac{n}{\lambda}\right)$, we have $\frac{k_1}{n} + \frac{\lambda^2 t^2}{n^2} = o(1)$. Therefore, there exists a mildly decaying sequence of δ_n with $\log\left(\frac{1}{\delta_n}\right)\left(\frac{k_1}{n} + \frac{\lambda^2 t^2}{n^2}\right) = o(1)$, i.e.,

$$
V(\boldsymbol{\theta}_t) = o(1). \tag{61}
$$

To conclude, δ_n can be chosen such that

$$\log\left(\frac{1}{\delta_n}\right) = \omega(1), \log\left(\frac{1}{\delta_n}\right) = O(n), \log\left(\frac{1}{\delta_n}\right) = O\left(\frac{1}{\frac{k_1}{n} + \frac{\lambda^2 t^2}{n^2}}\right), \tag{62}$$

and then with probability at least $1 - \delta_n$, the excess risk is consistent for all $t \in \left(\omega\left(\frac{1}{\lambda}\right), o\left(\frac{n}{\lambda}\right)\right)$:

$$R(\boldsymbol{\theta}_t) = B(\boldsymbol{\theta}_t) + V(\boldsymbol{\theta}_t) = o(1). \tag{63}$$

This completes the proof for the first claim. The second claim follows from Equation 58 and 60 by setting $t = \Theta\left(\frac{\sqrt{n}}{\lambda}\right)$. □

Lemma A.9 (Lemma 5.1 restated). *For any fixed (i.e. independent of sample size n) feature covariance $\boldsymbol{\Sigma}$ satisfying assumption 1, we have $k_1(n) = o(n)$.*

Proof. Suppose there exists constant c, such that $k_1(n) \geq cn$. By definition of k_1, we know that $\lambda_l \geq \frac{c_1 \sum_{i>0} \lambda_i}{n}$ holds for $1 \leq l \leq k_1(n)$. Hence we have

$$\sum_{l=\lfloor cn2^i \rfloor + 1}^{\lfloor cn2^{i+1} \rfloor} \lambda_l \gtrsim \frac{c_1 \sum_{i>0} \lambda_i}{n 2^{i+1}} cn2^i \gtrsim \sum_{i>0} \lambda_i. \tag{64}$$

summing up all l leads to a contradiction since $\sum_{i>0} \lambda_i < \infty$, which finishes the proof. □

Theorem A.2 (Theorem 4.1 restated). *Consider the overparameterized linear regression setting defined in section 4.1. Let Assumption 1,2 and 3 hold. Assume the learning rate satisfies $\lambda = O(\frac{1}{\mathrm{Tr}(\boldsymbol{\Sigma})})$. Then under the condition that $k_0 = O(n), k_1 = o(n), r(\boldsymbol{\Sigma}) = o(n)$, gradient descent is compatible with overparameterized linear regression in the region $T_n = \left(\omega\left(\frac{1}{\lambda}\right), o\left(\frac{n}{\lambda}\right)\right)$, namely,*

$$\sup_{t \in T_n} R(\boldsymbol{\theta}_t) \xrightarrow{P} 0 \quad as \quad n \to \infty.$$

Furthermore, if the feature dimension $p = \infty$ and the data distribution does not change with n, then the condition $k_0 = O(n)$ alone suffices for compatibility.

Proof. According to Corollary 5.1 and Lemma 5.1, for any $\varepsilon > 0$, there exists $\{\delta_n\}_{n>0}$ and N such that for any sample size $n > N$, we have

$$\Pr\left[\left|\sup_{t \in T_n} R(\boldsymbol{\theta}_t)\right| > \varepsilon\right] \leq \delta_n. \tag{65}$$

Let $n \to \infty$ shows that $\sup_{t \in T_n} R(\boldsymbol{\theta}_t)$ converges to 0 in probability, which completes the proof. □

B Examples and Discussions

B.1 Example Distributions

We apply our bound in Corollary 5.1 to several examples. In each example, we show the data distribution, the time interval, and the corresponding generalization bound. These distributions are widely discussed in (Bartlett et al., 2019; Zou et al., 2021).

Example B.1. *Under the same conditions as Theorem 5.1, let $\boldsymbol{\Sigma}$ denote the feature covariance matrix. We show the following examples:*

1. **(Inverse Polynominal).** *If the spectrum of $\boldsymbol{\Sigma}$ satisfies*

$$\lambda_k = \frac{1}{k^\alpha},$$

for some $\alpha > 1$, we derive that $k_0 = \Theta(n)$, $k_1 = \Theta\left(n^{\frac{1}{\alpha}}\right)$. Therefore, $\min_t V(\boldsymbol{\theta}_t) = O\left(n^{\frac{1-\alpha}{\alpha}}\right)$ and

$$\min_t R(\boldsymbol{\theta}_t) = O\left(n^{-\min\left\{\frac{\alpha-1}{\alpha}, \frac{1}{2}\right\}}\right).$$

2. **(Inverse Log-Polynominal).** *If the spectrum of* $\mathbf{\Sigma}$ *satisfies*

$$\lambda_k = \frac{1}{k \log^{\beta}(k+1)},$$

for some $\beta > 1$*, we derive that* $k_0 = \Theta\left(\frac{n}{\log n}\right)$*,* $k_1 = \Theta\left(\frac{n}{\log^{\beta} n}\right)$*. Therefore,* $\min_t V(\boldsymbol{\theta}_t) = O\left(\frac{1}{\log^{\beta} n}\right)$ *and*

$$\min_t R(\boldsymbol{\theta}_t) = O\left(\frac{1}{\log^{\beta} n}\right).$$

3. **(Constant).** *If the spectrum of* $\mathbf{\Sigma}$ *satisfies*

$$\lambda_k = \frac{1}{n^{1+\varepsilon}}, 1 \le k \le n^{1+\varepsilon},$$

for some $\varepsilon > 0$*, we derive that* $k_0 = 0$*,* $k_1 = 0$*. Therefore,* $\min_t V(\boldsymbol{\theta}_t) = O\left(\frac{1}{n}\right)$ *and*

$$\min_t R(\boldsymbol{\theta}_t) = O\left(\frac{1}{\sqrt{n}}\right).$$

4. **(Piecewise Constant).** *If the spectrum of* $\mathbf{\Sigma}$ *satisfies*

$$\lambda_k = \begin{cases} \frac{1}{s} & 1 \le k \le s, \\ \frac{1}{d-s} & s+1 \le k \le d, \end{cases}$$

where $s = n^r, d = n^q, 0 < r \le 1, q \ge 1$*. We derive that* $k_0 = n^r$*,* $k_1 = n^r$*. Therefore,* $\min_t V(\boldsymbol{\theta}_t) = O(n^{r-1})$ *and*

$$\min_t R(\boldsymbol{\theta}_t) = O\left(n^{-\min\left\{1-r, \frac{1}{2}\right\}}\right).$$

B.2 Comparisons with Benign Overfitting Results

We summarize the results in Bartlett et al. (2019); Zou et al. (2021) and our results in Table 1, and provide a detailed comparison with them below.

Comparison to Bartlett et al. (2019). In this seminal work, the authors study the excess risk of the min-norm interpolator. As discussed before, the min-norm interpolator is the convergence point of gradient descent in the overparameterized linear regression setting. One of the main results in Bartlett et al. (2019) is providing a tight bound for the variance part[5] in excess risk as

$$V(\hat{\boldsymbol{\theta}}) = O\left(\frac{k_0}{n} + \frac{n}{R_{k_0}(\mathbf{\Sigma})}\right), \tag{66}$$

where $\hat{\boldsymbol{\theta}} = \boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{Y}$ denotes the min-norm interpolator, and $R_k(\mathbf{\Sigma}) = (\sum_{i>k}\lambda_i)^2/(\sum_{i>k}\lambda_i^2)$ denote another type of effective rank.

By introducing the time factor, Theorem 5.1 improves over Equation (66) in at least two aspects. Firstly, Theorem 5.1 guarantees the consistency of the gradient descent dynamics for a broad range of step number t, while Bartlett et al. (2019) only study the limiting behavior of the dynamics of $t \to \infty$. Secondly, Theorem 5.1 implies that the excess risk of early stopping gradient descent solution can be much better than the min-norm interpolator. Compared to the bound in Equation (66), the bound in Corollary 5.1 (a.) replaces k_0 with a much smaller quantity k_1; and (b.) drops the second term involving $R_{k_0}(\mathbf{\Sigma})$. Therefore, we can derive a consistent bound for overparameterized linear regression for an early stopping solution, even though the excess risk of limiting point (min-norm interpolator) can be $\Omega(1)$ (See the first example in B.1). We can further derive a data-dependent time interval in which the bound is consistent, which cannot be directly obtained in Bartlett et al. (2019).

[5] We do not compare the bias component in the excess risk bound here since our bound for the bias component follows Bartlett et al. (2019).

Comparison to Zou et al. (2021). Zou et al. (2021) study a different setting, which focuses on the one-pass stochastic gradient descent solution of linear regression. The authors prove a bound for the excess risk as

$$R(\bar{\boldsymbol{\theta}}_t) = O\left(\frac{k_1}{n} + \frac{n\sum_{i>k_1}\lambda_i^2}{(\sum_{i>0}\lambda_i)^2}\right), \tag{67}$$

where $\bar{\boldsymbol{\theta}}_t$ denotes the parameter obtained using stochastic gradient descent (SGD) with constant step size at epoch t. Similar to our bound, Equation 67 also uses the effective dimension k_1 to characterize the variance term. However, we emphasize that Zou et al. (2021) derive the bound in a pretty different scenario from ours, which is one-pass SGD scenario. During the one-pass SGD training, we use a fresh data point to perform stochastic gradient descent in each epoch, and therefore they set $t = \Theta(n)$ by default. As a comparison, we apply the standard full-batch gradient descent, and thus the time can be more flexible. Besides, our results in Corollary 5.1 improve the bound in Equation (67) by dropping the second term. We refer to the third and fourth example in Example B.1 for more details, where[6] our bound outperforms Zou et al. (2021) when $\varepsilon < 1/2$ or $q < \min\{2-r, 3/2\}$.

B.3 Comparisons with Stability-Based Bounds

In this section, we show that Theorem 5.1 gives provably better upper bounds than the stability-based method. We cite a result from Teng et al. (2021), which uses stability arguments to tackle overparameteried linear regression under similar assumptions.

Theorem B.1 (modified from Theorem 1 in Teng et al. (2021)). *Under the overparameterized linear regression settings, assume that* $\|\boldsymbol{x}\| \le 1$, $|\varepsilon| \le V$, $w = \frac{\boldsymbol{\theta}^{*,\top}\boldsymbol{x}}{\sqrt{\boldsymbol{\theta}^{*,\top}\boldsymbol{\Sigma}\boldsymbol{\theta}^*}}$ *is* σ_w^2*-subgaussian. Let* $B_t = \sup_{\tau\in[t]}\|\boldsymbol{\theta}_t\|$. *the following inequality holds with probability at least* $1-\delta$*:*

$$R(\boldsymbol{\theta}_t) = \tilde{O}\left(\max\{1, \boldsymbol{\theta}^{*,\top}\boldsymbol{\Sigma}\boldsymbol{\theta}^*\sigma_w^2, (V+B_t)^2\}\sqrt{\frac{\log(4/\delta)}{n}} + \frac{\|\boldsymbol{\theta}^*\|^2}{\lambda t} + \frac{\lambda t(V+B_t)^2}{n}\right). \tag{68}$$

Theorem B.1 applies the general stability-based results (Hardt et al., 2016; Feldman & Vondrák, 2019) in the overparameterized linear regression setting, by replacing the bounded Lipschitz condition with the bounded domain condition. A fine-grained analysis (Lei & Ying, 2020) may remove the bounded Lipschitz condition, but it additionally requires zero noise or decaying learning rate, which is different from our setting. We omit the excess risk decomposition technique adopted in Teng et al. (2021) for presentation clarity.

Theorem B.1 can not directly yield the stability argument in Theorem 4.1, since obtaining a high probability bound of B_t requires a delicate trajectory analysis and is a non-trivial task. Therefore, data-irrelevant methods such as stability-based bounds can not be directly applied to our setting. Even if one can replace B_t in Equation 68 with its expectation that is easier to handle (this modification will require adding concentration-related terms, and make the bound in Equation 68 looser), we can still demonstrate that Theorem 5.1 is tighter than the corresponding stability-based analysis by providing a lower bound on $\mathbb{E}[B_t^2]$, which will imply a lower bound on the righthand side of Equation 68.

Theorem B.2. *Let Assumption 1, 2, 3 holds. Suppose* $\lambda = O\left(\frac{1}{\sum_{i>0}\lambda_i}\right)$. *Suppose the conditional variance of the noise* $\varepsilon|\boldsymbol{x}$ *is lower bounded by* σ_ε^2. *There exists constant c, such that with probability at least* $1 - ne^{-\frac{n}{c}}$, *we have for* $t = o(n)$,

$$\mathbb{E}\|\boldsymbol{\theta}_t\|^2 = \Omega\left(\frac{\lambda^2 t^2}{n}\left(\sum_{i>k_0}\lambda_i\right)\right) \tag{69}$$

First we prove the following lemma, bounding the number of large μ_i.

[6] Due to the bias term in Theorem 5.1, the overall excess risk bound cannot surpass the order $O(1/\sqrt{n})$, which leads to the cases that Zou et al. (2021) outperforms our bound. However, we note that such differences come from the intrinsic property of GD and SGD, which may be unable to avoid in the GD regimes.

Lemma B.1. *Suppose $t = o(n)$. Let l denote the number of μ_i, such that $\mu_i = \Omega\left(\frac{n}{t}\right)$. Then with probability at least $1 - ne^{-\frac{n}{c}}$, we have $l = O(t)$.*

Proof. According to Lemma A.1, we know that with probability at least $1 - ne^{-\frac{n}{c}}$, Equation 15 holds for all $0 \leq k \leq n-1$. Conditioned on this, we have

$$\frac{n}{t} l \lesssim \sum_{k=1}^{l} \mu_i \lesssim \sum_{k=1}^{l} (\sum_{i \geq k} \lambda_i + \lambda_k n) \lesssim (l+n) \sum_{i>0} \lambda_i \lesssim l + n. \tag{70}$$

Since $t = o(n)$, we have $l = O(t)$ as claimed. □

We also need the result from Bartlett et al. (2019), which gives a lowerbound of μ_n.

Lemma B.2. *(Lemma 10 in Bartlett et al. (2019)) For any σ_x, there exists a constant c, such that with probability at least $1 - e^{-\frac{n}{c}}$ we have,*

$$\mu_n \geq c \left(\sum_{i > k_0} \lambda_i \right). \tag{71}$$

We are now ready to prove Theorem B.2.

Proof. We begin with the calculation of $\|\boldsymbol{\theta}_t\|^2$. By Lemma A.2, the conditional unbiasedness of noise in Assumption 2 and the noise variance lower bound, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{\theta}_t\|^2 &= \left\| \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t (\boldsymbol{\theta}_0 - \boldsymbol{X}^\dagger \boldsymbol{Y}) + \boldsymbol{X}^\dagger Y \right\|^2 \\
&= \mathbb{E} \left\| \left(\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t \right) \boldsymbol{X}^\dagger \left(\boldsymbol{X} \boldsymbol{\theta}^* + \boldsymbol{\varepsilon} \right) \right\|^2 \\
&= \mathbb{E} \left\| \left(\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t \right) \boldsymbol{X}^\dagger \boldsymbol{X} \boldsymbol{\theta}^* \right\|^2 + \mathbb{E} \left\| \left(\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t \right) \boldsymbol{X}^\dagger \boldsymbol{\varepsilon} \right\|^2 \\
&\geq \mathbb{E} \left\| \left(\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t \right) \boldsymbol{X}^\dagger \boldsymbol{\varepsilon} \right\|^2 \\
&= \mathbb{E} \operatorname{Tr} \left[\left(\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t \right) \boldsymbol{X}^\dagger \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \boldsymbol{X}^{\dagger,\top} \left(\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t \right) \right] \\
&\geq \sigma_\varepsilon^2 \mathbb{E} \operatorname{Tr} \left[\left(\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t \right) \boldsymbol{X}^\dagger \boldsymbol{X}^{\dagger,\top} \left(\boldsymbol{I} - \left(\boldsymbol{I} - \frac{\lambda}{n} \boldsymbol{X}^\top \boldsymbol{X} \right)^t \right) \right] \\
&= \sigma_\varepsilon^2 \sum_{i=1}^{n} \frac{[1 - (1 - \frac{\lambda}{n} \mu_i)^t]^2}{\mu_i}.
\end{aligned} \tag{72}$$

When $\mu_i = o\left(\frac{n}{t}\right)$, we have

$$1 - (1 - \frac{\lambda}{n} \mu_i)^t = 1 - 1 + \frac{\lambda}{n} \mu_i t + O\left(\left(\frac{\lambda}{n} \mu_i t \right)^2 \right) = \Theta\left(\frac{\lambda}{n} \mu_i t \right). \tag{73}$$

Plugging it into Equation 72 and then use Lemma B.1, B.2, we know that under the high probability event in Lemma B.1 and B.2,

$$\mathbb{E}\|\boldsymbol{\theta}_t\|^2 = \Omega\left((n - l) \frac{\lambda^2}{n^2} \mu_n t^2 \right) = \Omega\left(\frac{\lambda^2}{n} \mu_n t^2 \right) = \Omega\left(\frac{\lambda^2 t^2}{n} \left(\sum_{i > k_0} \lambda_i \right) \right) \tag{74}$$

□

Therefore, the stability-based bound, i.e., the right hand side of Equation 68, can be lower bounded in expectation as $\Omega\left(\frac{\lambda^3 t^3}{n^2}\sum_{i>k_0}\lambda_i\right)$. This implies that the stability-based bound is vacuous when $t=\Omega\left(\frac{n^{\frac{2}{3}}\left(\sum_{i>k_0}\lambda_i\right)^{-\frac{1}{3}}}{\lambda}\right)$. Thus, stability-based methods will provably yield smaller compatibility region than $\left(\omega\left(\frac{1}{\lambda}\right), o\left(\frac{n}{\lambda}\right)\right)$ in Theorem 4.1 when $\sum_{i>k_0}\lambda_i$ is not very small, as demonstrated in the examples below.

Example B.2. *Let Assumption 1, 2, 3 holds. Assume without loss of generality that $\lambda=\Theta(1)$. We have the following examples:*

1. **(Inverse Polynominal).** *If the spectrum of $\mathbf{\Sigma}$ satisfies*
$$\lambda_k = \frac{1}{k^\alpha},$$
for some $\alpha > 1$, we derive that $k_0 = \Theta(n)$, $\sum_{i>k_0}\lambda_i = \Theta(\frac{1}{n^{\alpha-1}})$. Therefore, the stability bound in Theorem B.1 is vacuous when
$$t = \Omega\left(n^{\frac{\alpha+1}{3}}\right),$$
which is outperformed by the compatibility region in Theorem 5.1 when $\alpha < 2$.

2. **(Inverse Log-Polynominal).** *If the spectrum of $\mathbf{\Sigma}$ satisfies*
$$\lambda_k = \frac{1}{k\log^\beta(k+1)},$$
for some $\beta > 1$, we derive that $k_0 = \Theta\left(\frac{n}{\log n}\right)$, $\sum_{i>k_0}\lambda_i = \tilde{\Theta}(1)$. Therefore, the stability bound in Theorem B.1 is vacuous when
$$t = \tilde{\Omega}\left(n^{\frac{2}{3}}\right),$$
which is outperformed by the compatibility region in Theorem 5.1.

3. **(Constant).** *If the spectrum of $\mathbf{\Sigma}$ satisfies*
$$\lambda_k = \frac{1}{n^{1+\varepsilon}}, 1 \le k \le n^{1+\varepsilon},$$
for some $\varepsilon > 0$, we derive that $k_0 = 0$, $\sum_{i>k_0}\lambda_i = 1$. Therefore, the stability bound in Theorem B.1 is vacuous when
$$t = \Omega\left(n^{\frac{2}{3}}\right),$$
which is outperformed by the compatibility region in Theorem 5.1.

4. **(Piecewise Constant).** *If the spectrum of $\mathbf{\Sigma}$ satisfies*
$$\lambda_k = \begin{cases} \frac{1}{s} & 1 \le k \le s, \\ \frac{1}{d-s} & s+1 \le k \le d, \end{cases}$$
where $s = n^r, d = n^q, 0 < r \le 1, q \ge 1$, we derive that $k_0 = n^r$, $\sum_{i>k_0}\lambda_i = 1$. Therefore, the stability bound in Theorem B.1 is vacuous when
$$t = \Omega\left(n^{\frac{2}{3}}\right),$$
which is outperformed by the compatibility region in Theorem 5.1.

B.4 Comparisons with Uniform Convergence Bounds

We give a standard bound on the Rademacher complexity of linear models.

Theorem B.3 (Theorem in Mohri et al. (2012))**.** *Let $S \subseteq \{\boldsymbol{x} : \|x\|_2 \leq r\}$ be a sample of size n and let $\mathcal{H} = \{x \mapsto \langle w, x \rangle : \|w\|_2 \leq \Lambda\}$. Then, the empirical Rademacher complexity of $\mathcal{H}$ can be bounded as follows:*

$$\hat{\mathcal{R}}_S(\mathcal{H}) \leq \sqrt{\frac{r^2\Lambda^2}{n}}. \tag{75}$$

Furthermore, Talagrand's Lemma (See Lemma 5.7 in Mohri et al. (2012)) indicates that

$$\hat{\mathcal{R}}_S(l \circ \mathcal{H}) \leq L\hat{\mathcal{R}}_S(\mathcal{H}) = \frac{\Theta(\Lambda^2)}{\sqrt{n}}, \tag{76}$$

where $L = \Theta(\Lambda)$ is the Lipschitz coefficient of the square loss function l in our setting. Therefore, the Rademacher generalization bound is vacuous when $\Lambda = \Omega(n^{\frac{1}{4}})$. By Theorem B.2, we know that $\mathbb{E}\|\boldsymbol{\theta}_t\|^2 = \Omega(n^{\frac{1}{2}})$ when $t = \Omega\left(\frac{n^{\frac{3}{4}}}{\lambda(\sum_{i>k_0}\lambda_i)^{\frac{1}{2}}}\right)$. A similar comparison as in Example B.2 can demonstrate that uniform stability arguments will provably yield smaller compatibility region than that in Theorem 5.1 for example distributions.

B.5 Comparison with Previous Works on Early Stopping

A line of works focuses on deriving the excess risk guarantee of linear regression or kernel regression with early stopping (stochastic) gradient descent. We refer to Section 2 for a detailed discussion. Here we compare our results with some most relevant works, including (Yao et al., 2007; Lin & Rosasco, 2017; Pillaud-Vivien et al., 2018).

Comparison with Yao et al. (2007). Yao et al. (2007) study kernel regression with early stopping gradient descent and share some similarities with our paper. However, their approaches cannot cover ours, and are fundamentally different from ours in the following aspects.

Firstly, the assumptions used in the two approaches are different, due to different goals and techniques. Yao et al. (2007) assume that the input feature and data noise have bounded norm (see Section 2.1 Definitions and Notations in Yao et al. (2007)), while we require that the input feature is subgaussian with independent entries. The assumption used in our paper is widely-used in benign overfitting analysis, following Bartlett et al. (2019).

Furthermore, although Yao et al. (2007) obtain a minimax bound in terms of the convergence rate, it is suboptimal in terms of compatibility region. Specifically, The results in our paper show a compatibility region like $(0, n)$ while the techniques Yao et al. (2007) can only lead to a compatibility region like $(0, \sqrt{n})$. See Proof of the Main Theorem in section 2 in Yao et al. (2007) for details. Such differences come from different goals of the two approaches, where Yao et al. (2007) focus on providing the optimal early-stopping time while we focus on providing a larger compatibility region.

Comparison with Pillaud-Vivien et al. (2018). Pillaud-Vivien et al. (2018) study kernel regression with multi-pass stochastic gradient descent and derive optimal excess risk guarantee. Different from our approaches with full-batch gradient descent, they study *averaged* stochastic gradient descent with batchsize equal to 1.

Comparison with Lin & Rosasco (2017). Lin & Rosasco (2017) study stochastic gradient descent with arbitrary batchsize, which is reduced to full batch gradient descent when setting the batchsize to sample size n. However, their results are still fundamentally different from ours, since they require the boundness assumption, and focus more on the optimal early stopping time rather than the largest compatibility region, in the same spirit of Yao et al. (2007). Specifically, Lin & Rosasco (2017) derive a compatibility region like $(0, n^{\frac{\zeta+1}{2\zeta+\gamma}})$, where ζ and γ are problem dependent constants (See Theorem 1 in Lin & Rosasco (2017) for details). The following examples demonstrate that this paper's results yield larger compatibility regions for a wide range of distribution classes.

Example B.3. (Inverse Polynominal). *If the spectrum of $\boldsymbol{\Sigma}$ satisfies*

$$\lambda_k = \frac{1}{k^\alpha},$$

for some $\alpha > 1$. For this distribution, we have $\zeta = \frac{1}{2}$, $\gamma = \frac{1}{\alpha}$, and the compatibility region is $(0, n^{\frac{3\alpha}{2\alpha+1}})$, which is smaller than the compatibility region $(0, n^{\frac{3\alpha+1}{2\alpha+1}})$ given in Example B.5.

Example B.4. (Inverse Log-Polynominal). *If the spectrum of $\boldsymbol{\Sigma}$ satisfies*

$$\lambda_k = \frac{1}{k \log^{\beta}(k+1)},$$

for some $\beta > 1$. For this distribution, we have $\zeta = \frac{1}{2}$, $\gamma = 1$, and the compatibility region is $(0, n^{\frac{3}{4}})$, which is smaller than the compatibility region $(0, n)$ given Corollary 5.1.

B.6 Discussions on Extensions to Kernel Regression

In this section, we discuss extending the analysis in overparameterized linear regression to kernel regression setting.

Let $\mathcal{H}$ denote a infinite dimensional Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and $\phi : \mathbb{R}^d \to \mathcal{H}$ denote a feature map. Consider the following class of functions:

$$\mathcal{F} = \{ f : \mathbb{R}^d \to \mathbb{R} | f(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \phi(\boldsymbol{x}) \rangle_{\mathcal{H}} \}. \tag{77}$$

Let $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ denote the data vector and the response, following a joint distribution $\mathcal{D}$. The goal of kernel regression is to find a function f parameterized by $\boldsymbol{\theta}$, that minimizes the following population loss

$$L(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}\left(y - f(\boldsymbol{x})\right)^2 = \frac{1}{2}\mathbb{E}\left(y - \langle \boldsymbol{\theta}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle_{\mathcal{H}}\right)^2 \tag{78}$$

Therefore, kernel regression is equivalent to solving a linear regression problem on transformed data $(\phi(\boldsymbol{x}), y)$. By replacing $\boldsymbol{x}$ with $\phi(\boldsymbol{x})$, the notations and results in Section 4 can be naturally extended to the kernel regression case, which is detailed as follows.

Given a dataset $\{(\boldsymbol{x}_i, y_i)\}_{1 \le i \le n}$ sampled i.i.d from $\mathcal{D}$, consider the following dynamics of gradient descent analogous to Equation (3),

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\lambda}{n}\phi(\boldsymbol{X})^{\top}(\phi(\boldsymbol{X})\boldsymbol{\theta}_t - \boldsymbol{Y}), \tag{79}$$

where $\phi(\boldsymbol{X}) = (\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_n))^{\top}$ and $\boldsymbol{Y} = (y_1, \cdots, y_n)^{\top}$. Note that this dynamic takes place in the feature space $\mathcal{H}$. The following corollary characterizes the compatibility between kernel regression and gradient descent.

Corollary B.1. *Assume the distribution of $(\phi(\boldsymbol{x}), y)$ satisfies Assumptions 1, 2, 3, and does not change with sample size n. Let $\boldsymbol{\Sigma}_{\phi} = \mathbb{E}\left[\phi(\boldsymbol{x})\phi(\boldsymbol{x})^{\top}\right]$ denote the feature covariance. Then under the condition that the effective dimension $k_0(\boldsymbol{\Sigma}_{\phi}) = O(n)$ and the learning rate $\lambda = O(\frac{1}{\mathrm{Tr}(\boldsymbol{\Sigma}_{\phi})})$, gradient descent using 79 is compatible with kernel regression.*

Proof. The corollary follows from Theorem 4.1, by noting that its proof holds for feature vectors in an infinite dimensional Hilbert space. □

We give several remarks regarding the difference between the generalization analysis of kernel regression and overparameterized linear regression.

Firstly, Corollary B.1 assumes $\phi(\boldsymbol{x})$ has subgaussian i.i.d entries after normalization (see Assumption 1.3), which can be hard to satisfy due to the non-linearity of feature mapping $\phi(\cdot)$.

Secondly, Corollary B.1 focuses on gradient descent in the feature space $\mathcal{H}$. Another perspective is to apply the Representer theorem (Mohri et al., 2012) to express the weight $\boldsymbol{\theta}$ into a linear combination of transformed inputs as $\boldsymbol{\theta} = \sum_{1 \le j \le n} \alpha_j \phi(\boldsymbol{x}_j)$, and analyze the gradient descent dynamics of the n-dimensional vector $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)$. Although they both converge to the min-norm solution, these two gradient descent dynamics are different, and Theorem 4.1 is not directly applicable to gradient descent on $\boldsymbol{\alpha}$.

Thirdly, previous works (El Karoui, 2010; Liang & Rakhlin, 2018; Mei & Montanari, 2022) prove that the spectrum of feature covariance $\boldsymbol{\Sigma}_{\phi}$ has an approximately linear relationship with the spectrum of data covariance $\boldsymbol{\Sigma} = \mathbb{E}\left[\boldsymbol{x}\boldsymbol{x}^{\top}\right]$ when feature mapping $\phi(\cdot)$ corresponds to an inner product type kernel. This can be used to derive a more intrinsic compatibility condition using only data distribution rather than feature distribution. We leave a more refined analysis for data-algorithm compatibility in the kernel regression regime for future works.

B.7 Varying t, Varying $c(t, n)$

Although setting $c(t, n)$ to a constant as in Corollary 5.1 suffices to prove Theorem 4.1, in this section we show that the choice of $c(t, n)$ can be much more flexible. Specifically, we provide a concrete example and demonstrate that by setting $c(t, n)$ to a non-constant, Theorem 5.1 can indeed produce larger compatibility regions.

Example B.5. *Under the same conditions as Theorem 5.1, let $\mathbf{\Sigma}$ denote the feature covariance matrix. If the spectrum of $\mathbf{\Sigma}$ satisfies $\lambda_k = \frac{1}{k^\alpha}$ for some $\alpha > 1$, we set $c(t, n) = \Theta\left(n^{\frac{\alpha+1-2\alpha\tau}{2\alpha+1}}\right)$ for a given $\frac{\alpha+1}{2\alpha} \leq \tau \leq \frac{3\alpha+1}{2\alpha+2}$. Then for $t = \Theta(n^\tau)$, we derive that $V(\boldsymbol{\theta}_t) = O\left(n^{\frac{2\alpha\tau-3\alpha+2\tau-1}{2\alpha+1}}\right)$.*

Example B.5 shows that by choosing $c(t, n)$ as a non-constant, we exploit the full power of Theorem 5.1, and extend the compatibility region to $t = o\left(n^{\frac{3\alpha+1}{2\alpha+2}}\right) = \omega(n)$. In this example, Theorem 5.1 outperforms all $O\left(\frac{t}{n}\right)$-type bounds, which become vacuous when $t = \omega(n)$.

B.8 Calculations in Section B.1

We calculate the quantities $r(\Sigma), k_0, k_1, k_2$ for the example distributions in B.1. The results validate that k_1 is typically a much smaller quantity than k_0, and k_1 serves as a proxy for k_2 in the constant $c(t, n)$ setting.

1. **Calculations for $\lambda_k = \frac{1}{k^\alpha}, \alpha > 1$.**
   Define $r_k(\mathbf{\Sigma}) = \frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}$ as in Bartlett et al. (2019). Since $\sum_{i>k}\frac{1}{i^\alpha} = \Theta(\frac{1}{k^{\alpha-1}})$, we have $r_k(\mathbf{\Sigma}) = \Theta\left(\frac{\frac{1}{k^{\alpha-1}}}{\frac{1}{k^\alpha}}\right) = \Theta(k)$. Hence, $k_0 = \Theta(n)$ [7], and the conditions of theorem 5.1 is satisfied.
   As $\sum_{i>0}\lambda_i < \infty$, By its definition we know that k_1 is the smallest l such that $\lambda_{l+1} = O(\frac{1}{n})$. Therefore, $k_1 = \Theta(n^{\frac{1}{\alpha}})$. Similarly, $k_2 = \Theta(n^{\frac{1}{\alpha}})$.
2. **Calculations for $\lambda_k = \frac{1}{k\log^\beta(k+1)}, \beta > 1$.**
   $\sum_{i>k}\frac{1}{i\log^\beta(i+1)} = \Theta(\int_k^\infty \frac{1}{x\log^\beta x}dx) = \Theta(\frac{1}{\log^{\beta-1}k})$, which implies $r_k(\mathbf{\Sigma}) = k\log k$. Solving $k_0\log k_0 \geq \Theta(n)$, we have $k_0 = \Theta(\frac{n}{\log n})$.

   By the definition of k_1, we know that k_1 is the smallest l such that $l\log^\beta(l+1) \geq \Theta(n)$. Therefore, $k_1 = \Theta(\frac{n}{\log^\beta n})$. $k_2 = \Theta(\frac{n}{\log^\beta n})$ by similar calculations.
3. **Calculations for $\lambda_i = \frac{1}{n^{1+\varepsilon}}, 1 \leq i \leq n^{1+\varepsilon}, \varepsilon > 0$.**
   Since $r_0(\mathbf{\Sigma}) = n^{1+\varepsilon}$, we have $k_0 = 0$. By the definition of k_1 and k_2, we also have $k_1 = k_2 = 0$.
4. **Calculations for $\lambda = \begin{cases}\frac{1}{s} & 1 \leq k \leq s \\ \frac{1}{d-s} & s+1 \leq k \leq d\end{cases}, s = n^r, d = n^q, 0 < r \leq 1, q \geq 1$.**
   For $0 \leq k < s$, $r_k(\mathbf{\Sigma}) = \Theta(\frac{1}{\frac{1}{s}}) = \Theta(n^r) = o(n)$, while $r_s(\mathbf{\Sigma}) = \frac{1}{\frac{1}{d-s}} = \Theta(n^q) = \omega(n)$. Therefore, $k_0 = s = n^r$.
   Similarly, noting that $\lambda_k = \frac{1}{n^r} = \omega(n)$ for $0 \leq k < s$ and $\lambda_s = \Theta(\frac{1}{n^d}) = o(n)$, we know that $k_1 = s = n^r$. Similarly, $k_2 = s = n^r$.

B.9 Calculations for $\lambda_k = \frac{1}{k^\alpha}, \alpha > 1$ in Section B.7

Set $c(t, n) = \frac{1}{n^\beta}$, where $\beta > 0$ will be chosen later. First we calculate k_2 under this choice of $c(t, n)$. Note that $\sum_{i>k}\frac{1}{i^\alpha} = \Theta\left(\frac{1}{k^{\alpha-1}}\right)$. Therefore, k_2 is the smallest k such that $\frac{1}{k^{\alpha-1}} + \frac{n}{k^\alpha} = O(\frac{1}{n^\beta})$. For the bound on $V(\theta_t)$ to be consistent, we need $k_2 = o(n)$. Hence, $\frac{1}{k^{\alpha-1}} = O(\frac{n}{k^\alpha})$, which implies $k_2 = n^{\frac{\beta+1}{\alpha}}$.

[7] The calculations for k_0, k_1 and k_2 in this section only apply when n is sufficiently large.

Plugging the value of $c(t,n)$ and k_2 into our bound, we have

$$V(\theta_t) = O\left(n^{(\frac{1}{\alpha}-1)+(\frac{1}{\alpha}+1)\beta} + n^{2\tau-\beta-2}\right)$$

which attains its minimum $\Theta(n^{\frac{2\alpha\tau-3\alpha+2\tau-1}{2\alpha+1}})$ at $\beta = \Theta\left(\frac{2\alpha\tau-\alpha-1}{2\alpha+1}\right)$.

For $V(\theta) = O(1)$, we need $\tau \leq \frac{3\alpha+1}{2\alpha+2}$. For $\beta \geq 0$, we need $\tau \geq \frac{\alpha+1}{2\alpha}$. Putting them together gives the range of t in which the above calculation applies.

B.10 Discussion on $\mathcal{D}_n$

In this paper, the distribution $\mathcal{D}$ is regarded as a sequence of distribution $\{\mathcal{D}_n\}$ which may dependent on sample size n. The phenomenon comes from overparameterization and asymptotic requirements. In the definition of compatibility, we require $n \to \infty$. In this case, overparameterization requires that the dimension p (if finite) cannot be independent of n since $n \to \infty$ would break overparameterization. Therefore, the covariance Σ also has n-dependency, since Σ is closely related to p.

Several points are worth mentioning: (1) Similar arguments generally appear in related works. For example, Bartlett et al. (2019) use similar arguments when discussing the definition of benign covariance (Page 7 in the arxiv version). (2) One can avoid such issues by letting $p \to \infty$. This is why we discuss the special case $p = \infty$ in Theorem 4.1. (3) If p is a fixed finite constant that does not alter with n, the problem becomes underparameterized and thus trivial to get a consistent generalization bound.

B.11 Comparisons with PAC-Bayes Bounds

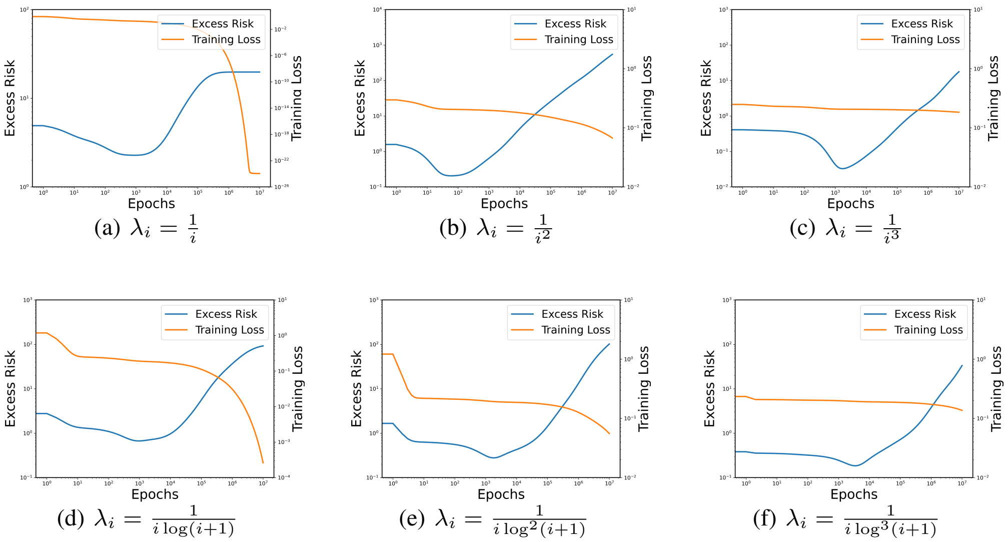
As one of the most exciting techniques in generalization analysis, PAC-Bayes theory works both theoretically and empirically. Usually, the form of PAC-Bayes bounds relies on the distance between the *prior distribution* which is unrelated to the training procedure, and the *posterior distribution* which is the distribution after training (*e.g.*, isotropic Gaussian distribution centered at a trained parameter).

In this sense, PAC-Bayes considers different regimes from this paper, since the returned classifier in this paper is not a distribution. Moreover, PAC-Bayes usually do not explicitly focus on early-stopping iterates, while the notion of compatibility heavily relies on the early-stopping arguments. By considering early-stopping that encodes algorithm information, we can derive tighter generalization bound relying on weaker assumptions.

More interestingly, PAC-Bayes framework is not mutually exclusive with the trajectory analysis. One can indeed introduce trajectory analysis into PAC-Bayes techniques and derive a compatibility region based on trajectory-based PAC-Bayes theory. By doing so, one can explicitly incorporate more algorithm information into PAC-Bayes framework. We leave more detailed discussions for future work.

We next show the differences in more detail:

1. **Different settings.** In PAC-Bayes analysis, the returned classifier is a distribution instead of a fixed parameter. Forcing PAC-Bayes analysis in the fixed parameter regime would cost much because it is hard to define distribution distance (*e.g.*, KL divergence) on two *single-point distributions*.
2. **Different classifiers.** Although both methods consider "a bag of" classifiers, they are fundamentally different. In PAC-Bayes framework, the trained random classifiers are regarded to be *independently* drawn from the posterior distribution. However, in trajectory analysis, all classifiers are *dependent* during the training process. Therefore, trajectory analysis is more challenging in this sense.
3. **Different characterizations.** PAC-Bayes framework characterizes the *expectation* of generalization loss, over the randomness of the posterior distribution over parameters, which do not explicitly sketch the time factor. In comparison, the trajectory analysis in this paper focus on providing a *compatibility region*, where the generalization error is uniformly consistent and considers the time factor.



(a) $\lambda_i = \frac{1}{i}$ (b) $\lambda_i = \frac{1}{i^2}$ (c) $\lambda_i = \frac{1}{i^3}$

(d) $\lambda_i = \frac{1}{i\log(i+1)}$ (e) $\lambda_i = \frac{1}{i\log^2(i+1)}$ (f) $\lambda_i = \frac{1}{i\log^3(i+1)}$

Figure 2: **The training plot for overparameterized linear regression with different covariances using GD.**

C Additional Experiment Results

C.1 Details and Discussions for Linear Regression Experiments

In this section, we provide the experiment details for linear regression experiments and present additional empirical results.

In Section 6, We consider six overparameterized linear regression instances with input dimension $p = 1000$ and sample size $n = 100$. The feature vectors are independently sampled from zero-mean Gaussian distributions, whose covariances are diagonal with entries $\lambda_i = \frac{1}{i}, \frac{1}{i^2}, \frac{1}{i^3}, \frac{1}{i\log(i+1)}, \frac{1}{i\log^2(i+1)}, \frac{1}{i\log^3(i+1)}$, respectively. For each feature $\boldsymbol{x}$, we construct the response as $y = \boldsymbol{x}^\top \boldsymbol{\theta}^* + \varepsilon$, where $\boldsymbol{\theta}^*$ is sampled from a p-dimensional standard Gaussian distribution for each instance, and ε is sampled from a standard Gaussian distribution. We conduct gradient descent with learning rate $\lambda = 0.001$ on the above instances. We also compute min-norm excess risk via its closed form, with 1×10^{-4} weight decay on parameter $\boldsymbol{\theta}$ to avoid numerical instability. Note that the regularization will only reduce the final excess and will not jeopardize our final conclusion's correctness.

The linear regression experiment in Figure 1 follows the setting described in section 6. Although the final iterate does not interpolate the training data, the results suffice to demonstrate the gap between the early-stopping and final-iterate excess risk. The training plot for different covariances are given in Figure 2.

We provide the experiment results of sample sizes $n = 50$, $n = 200$ and feature dimensions $p = 500$, $p = 2000$ analogous to those in Section 6. The settings are the same as described in the main text, except for the sample size. The optimal excess risk and min-norm excess risk are provided in Table 3, 4, 6 and 5. The tables indicate that the three observations given in section 6 hold for different sample size n.

C.2 Details for MNIST Experiments

In this section, we provide the experiment details and additional results in MNIST dataset.

The MNIST experiment details are described below. We create a noisy version of MNIST with label noise rate 20%, i.e. randomly perturbing the label with probability 20% for each training data,

Table 3: **The effective dimension k_1, the optimal early stopping excess risk and min-norm excess risk for different feature distributions, with sample size $n = 50$, $p = 1000$.** We calculate the 95% confidence interval for the excess risk.

DISTRIBUTIONS	k_1	OPTIMAL EXCESS RISK	MIN-NORM EXCESS RISK
$\lambda_i = \frac{1}{i}$	$\Theta(n)$	2.515 ± 0.0104	12.632 ± 0.1602
$\lambda_i = \frac{1}{i^2}$	$\Theta(n^{\frac{1}{2}})$	0.269 ± 0.0053	50.494 ± 0.9378
$\lambda_i = \frac{1}{i^3}$	$\Theta(n^{\frac{1}{3}})$	0.083 ± 0.0011	13.208 ± 0.4556
$\lambda_i = \frac{1}{i \log(i+1)}$	$\Theta(\frac{n}{\log n})$	0.808 ± 0.0090	46.706 ± 0.6983
$\lambda_i = \frac{1}{i \log^2(i+1)}$	$\Theta(\frac{n}{\log^2 n})$	0.381 ± 0.0076	74.423 ± 1.1472
$\lambda_i = \frac{1}{i \log^3(i+1)}$	$\Theta(\frac{n}{\log^3 n})$	0.233 ± 0.0052	43.045 ± 0.8347

Table 4: **The effective dimension k_1, the optimal early stopping excess risk and min-norm excess risk for different feature distributions, with sample size $n = 200$, $p = 1000$.** We calculate the 95% confidence interval for the excess risk.

DISTRIBUTIONS	k_1	OPTIMAL EXCESS RISK	MIN-NORM EXCESS RISK
$\lambda_i = \frac{1}{i}$	$\Theta(n)$	2.173 ± 0.0065	52.364 ± 0.4009
$\lambda_i = \frac{1}{i^2}$	$\Theta(n^{\frac{1}{2}})$	0.161 ± 0.0039	36.855 ± 0.4833
$\lambda_i = \frac{1}{i^3}$	$\Theta(n^{\frac{1}{3}})$	0.068 ± 0.0012	8.1990 ± 0.2313
$\lambda_i = \frac{1}{i \log(i+1)}$	$\Theta(\frac{n}{\log n})$	0.628 ± 0.0034	152.70 ± 1.1073
$\lambda_i = \frac{1}{i \log^2(i+1)}$	$\Theta(\frac{n}{\log^2 n})$	0.241 ± 0.0036	83.550 ± 0.7596
$\lambda_i = \frac{1}{i \log^3(i+1)}$	$\Theta(\frac{n}{\log^3 n})$	0.146 ± 0.0108	33.469 ± 0.4540

to simulate the label noise which is common in real datasets, e.g ImageNet (Stock & Cissé, 2018; Shankar et al., 2020; Yun et al., 2021). We do not inject noise into the test data.

We choose a standard four layer convolutional neural network as the classifier. We use a vanilla SGD optimizer without momentum or weight decay. The initial learning rate is set to 0.5. Learning rate is decayed by 0.98 every epoch. Each model is trained for 300 epochs. The training batch size is set to 1024, and the test batch size is set to 1000. We choose the standard cross entropy loss as the loss function.

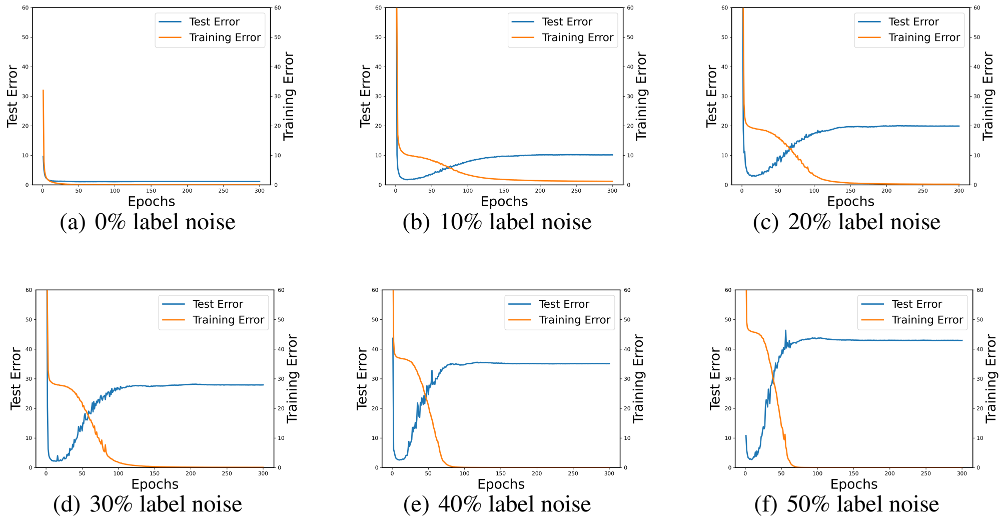
We provide the plot for different levels of label noise in Figure 3. We present the corresponding test error of the best early stopping iterate and the final iterate in Table 7. Since the theoretical part of this paper focuses on GD, we also provide a corresponding plot of GD training in Figure 4 for completeness.

Table 5: **The effective dimension k_1, the optimal early stopping excess risk and min-norm excess risk for different feature distributions, with sample size $n = 100$, $p = 500$.** We calculate the 95% confidence interval for the excess risk.

DISTRIBUTIONS	k_1	OPTIMAL EXCESS RISK	MIN-NORM EXCESS RISK
$\lambda_i = \frac{1}{i}$	$\Theta(n)$	1.997 ± 0.0876	27.360 ± 0.3019
$\lambda_i = \frac{1}{i^2}$	$\Theta(n^{\frac{1}{2}})$	0.211 ± 0.0050	43.531 ± 0.7025
$\lambda_i = \frac{1}{i^3}$	$\Theta(n^{\frac{1}{3}})$	0.076 ± 0.0011	10.062 ± 0.3022
$\lambda_i = \frac{1}{i\log(i+1)}$	$\Theta(\frac{n}{\log n})$	0.645 ± 0.0056	96.465 ± 1.0594
$\lambda_i = \frac{1}{i\log^2(i+1)}$	$\Theta(\frac{n}{\log^2 n})$	0.289 ± 0.0055	83.694 ± 0.9827
$\lambda_i = \frac{1}{i\log^3(i+1)}$	$\Theta(\frac{n}{\log^3 n})$	0.181 ± 0.0045	38.090 ± 0.6378

Table 6: **The effective dimension k_1, the optimal early stopping excess risk and min-norm excess risk for different feature distributions, with sample size $n = 100$, $p = 2000$.** We calculate the 95% confidence interval for the excess risk.

DISTRIBUTIONS	k_1	OPTIMAL EXCESS RISK	MIN-NORM EXCESS RISK
$\lambda_i = \frac{1}{i}$	$\Theta(n)$	2.662 ± 0.0066	23.111 ± 0.2278
$\lambda_i = \frac{1}{i^2}$	$\Theta(n^{\frac{1}{2}})$	0.219 ± 0.0050	43.130 ± 0.6421
$\lambda_i = \frac{1}{i^3}$	$\Theta(n^{\frac{1}{3}})$	0.077 ± 0.0010	10.031 ± 0.2942
$\lambda_i = \frac{1}{i\log(i+1)}$	$\Theta(\frac{n}{\log n})$	0.749 ± 0.0055	88.744 ± 0.9414
$\lambda_i = \frac{1}{i\log^2(i+1)}$	$\Theta(\frac{n}{\log^2 n})$	0.312 ± 0.0057	82.859 ± 0.9394
$\lambda_i = \frac{1}{i\log^3(i+1)}$	$\Theta(\frac{n}{\log^3 n})$	0.190 ± 0.0047	37.782 ± 0.5945



(a) 0% label noise

(b) 10% label noise

(c) 20% label noise

(d) 30% label noise

(e) 40% label noise

(f) 50% label noise

Figure 3: **The training plot for corrupted MNIST with different levels of label noise using SGD.** Figure (c) is copied from Figure 1.

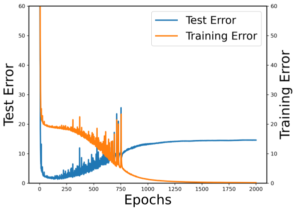


Figure 4: **The training plot for corrupted MNIST with 20% label noise using GD.**

Table 7: **The test error of optimal stopping iterate and final iterate on MNIST dataset with different levels of label noise.** The results demonstrate that stopping iterate can have significantly better generalization performance than interpolating solutions for real datasets.

NOISE LEVEL	OPTIMAL TEST ERROR	FINAL TEST ERROR
0%	1.07%	1.13%
10%	1.75%	10.16%
20%	2.88%	19.90%
30%	2.18%	27.94%
40%	2.57%	35.15%
50%	2.71%	42.95%