Electronic Journal of Statistics
Vol. 16 (2022) 4547–4603
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2055

Dimension independent excess risk by stochastic gradient descent*

Xi Chen†[1]

[1]*Stern School of Business, New York University*
e-mail: xichen@nyu.edu

Qiang Liu†[2]

[2]*School of Statistics and Management, Shanghai University of Finance and Economics*
e-mail: liuqiang@mail.sufe.edu.cn

Xin T. Tong†[3]

[3]*Department of Mathematics, National University of Singapore*
e-mail: mattxin@nus.edu.sg

Abstract: One classical canon of statistics is that large models are prone to overfitting, and model selection procedures are necessary for high dimensional data. However, many overparameterized models, such as neural networks, perform very well in practice, although they are often trained with simple online methods and regularization. The empirical success of overparameterized models, which is often known as benign overfitting, motivates us to have a new look at the statistical generalization theory for online optimization. In particular, we present a general theory on the excess risk of stochastic gradient descent (SGD) solutions for both convex and locally non-convex loss functions. We further discuss data and model conditions that lead to a "low effective dimension". Under these conditions, we show that the excess risk either does not depend on the ambient dimension p or depends on p via a poly-logarithmic factor. We also demonstrate that in several widely used statistical models, the "low effective dimension" arises naturally in overparameterized settings. The studied statistical applications include both convex models such as linear regression and logistic regression and non-convex models such as M-estimator and two-layer neural networks.

MSC2020 subject classifications: Primary 49N60; secondary 49M25, 49M37.
Keywords and phrases: Dimension independent, excess risk, weak non-convexity, stochastic gradient descent, regularization.

Received March 2021.

*Xi Chen would like to thank the support from NSF via the grant IIS-1845444. Qiang Liu was a postdoc Research Fellow at the Department of Mathematics, National University of Singapore when the manuscript was submitted, with the financial support from Singapore MOE via the grant R-146-000-258-114. Now, he is an Assistant Professor at the School of Statistics and Management, Shanghai University of Finance and Economics, and his work is supported by Fundamental Research Funds for the Central Universities, Shanghai University of Finance and Economics (No. 2021110482, No. 2022110007). Xin T. Tong would like to thank the support from Singapore MOE via the grant R-146-000-292-114.

†The authors are listed alphabetically.

Contents

1. Introduction

The study of overfitting phenomenon has been an important topic in statistics and machine learning. From classical statistical learning theory, we understand that when the number of model parameters is large compared to the amount of

data, the test error can be excessively large even if the training error is small. This phenomenon is usually known as overfitting. For this reason, dimension reduction or feature selection mechanisms such as principal component analysis (PCA) and shrinkage methods are often required in the training phase to reduce model dimension and avoid overfitting.

In recent years, deep neural networks have achieved great successes in practical applications. Researchers have found out that overparameterized neural networks usually achieve superior performance [27, 39, 49, 2, 3]. Moreover, these models are often trained with simple regularization and do not need dimension reduction procedures. This phenomenon is sometimes referred to as *benign overfitting* [7]. To understand it, we need a new statistical framework to study generalization ability.

Although there is much practical evidence on the benefit of overparameterization, the existing theoretical study mainly focuses on linear models (see, e.g., [7, 47, 1]) or neural networks with certain special data structures (see, e.g., [39]). The main purpose of our paper is to systematically investigate the excess risk for a risk minimization problem when the number of parameters p is much larger than the sample size N. In particular, we establish an excess risk bound for stochastic gradient descent (SGD) solutions for both convex (e.g., linear regression and logistic regression) and non-convex problems (some M-estimators and neural networks). We focus our study on the SGD algorithm because it has been widely used in large-scale data learning due to its computational and memory efficiency.

Let us briefly introduce our setup of the overfitting problem and the SGD algorithm. We consider the following population risk minimization problem under a loss function F, which can be either convex or non-convex:

$$w^* = \operatorname*{argmin}_w F(w), \quad F(w) := \mathbb{E}_\zeta f(w, \zeta). \tag{1.1}$$

In (1.1), $w \in \mathbb{R}^p$ is a p-dimensional parameter vector, ζ denotes a random sample from a certain probability distribution, and $f(\cdot, \zeta)$ is the loss function on each individual data ζ. The global minimizer w^* is often the true model parameter in statistical estimation problems. In practice, the distribution of ζ is usually unknown, and one only has the access to N *i.i.d.* samples $\zeta_1, \ldots, \zeta_N$ from the population. Instead of minimizing the population risk $F(w)$ in (1.1), it is more practical to minimize the empirical loss function

$$\hat{F}(w) = \frac{1}{N} \sum_{i=1}^{N} f(w, \zeta_i). \tag{1.2}$$

Often, instead of directly minimizing the empirical loss, an extra regularization term is sometimes added to the empirical loss to avoid overfitting. On one hand, such a regularization is helpful to obtain a tighter convergence rate for convex optimization problem. On the other hand, we will see later that such a regularization is necessary for deriving dimension independent excess risk bound when the population risk F is non-convex. In this paper, we consider the most

commonly used ridge or Tikhonov regularization. The corresponding regularized empirical loss function takes the following form,

$$\widehat{F}_\lambda(w) := \widehat{F}(w) + \frac{\lambda}{2}\|w\|^2 = \frac{1}{N}\sum_{i=1}^{N} f_\lambda(w,\zeta_i), \quad f_\lambda(w,\zeta) := f(w,\zeta) + \frac{\lambda}{2}\|w\|^2. \tag{1.3}$$

The weight of regularization is controlled by $\lambda > 0$, which is a tuning parameter. When $\lambda = 0$, this corresponds to the ridgeless regression or "implicit regularization", we will also discuss this setup in this paper. One popular way to optimize $\widehat{F}_\lambda$ in machine learning is via SGD. In particular, for a generic initialization parameter w_0, SGD is an iterative algorithm, where the $(n+1)$-th iterate w_{n+1} is updated according to the following equation,

$$w_{n+1} := w_n - \eta \nabla f_\lambda(w_n, \zeta_n) = w_n - \eta\left(\nabla f(w_n,\zeta_n) + \lambda w_n\right). \tag{1.4}$$

By running through N samples, SGD outputs the N-th iterate w_N as the final estimator of w^*. Notably, SGD iterates are affected by the stochasticity of the data samples ζ. To reduce such noise and improve accuracy, the averaged SGD (ASGD) method uses the average iterate

$$\bar{w}_N = \frac{1}{N}\sum_{i=1}^{N} w_i$$

as the final estimator of w^*. In SGD iterations (1.4), the hyper-parameter η is known as the stepsize. In our paper, we consider using a constant stepsize, which is a popular choice in practice [5]. The value of η will be discussed later in our theoretical results. Moreover, in (1.4), the gradient is taken with respect to the parameter vector w. For notational simplicity, we will use "∇" as a short notation for "∇_w" throughout the paper.

When the sample size N is much larger than the dimensionality p, it is expected that w_N would be close to w^* under certain conditions. However, in an overparameterized setting where N is less than p, the solution w_N can be far away from w^*. In this case, estimating the underlying parameter accurately usually requires strong assumptions. However, for many machine learning tasks, it is of more interest in achieving small excess risk, which is defined as follows,

$$G(w_N) = F(w_N) - F(w^*). \tag{1.5}$$

The main purpose of the paper is to provide an upper bound of the excess risk in (1.5) in overparameterized settings. We will characterize the scenarios where such an excess risk bound is independent of p or only involves in poly-logarithmic factors of p.

1.1. Main results and paper organization

The main message of this paper is as follows. For a large class of statistical learning problems where the *effective dimension* is low (see the rigorous definition in Section 3), the stochastic gradient descent (SGD) algorithm with proper

ridge regularization will not overfit even if the ambient model dimension is much larger than the sample size. In particular, we will show that the excess risk has at most poly-logarithmic dependence on the ambient model dimension p.

In Section 2, we present a framework for excess risk analysis. We will separately discuss two scenarios. In the first scenario, the true parameter is sparse or weakly dense with a dimension independent l_2 norm. In this case, we show ridgeless SGD has been sufficient to obtain dimension independent excess risk (Theorem 2.3). In the second scenario, the true parameter is dimension independent only under some problem-specific norms. In this more challenging case, we show SGD can achieve dimension independent excess risk by proper amount of ridge penalty (Theorem 2.4). The upper bound of the excess risk is provided. Using linear regression as an illustrative example, we show that each term in the excess risk has a strong statistical interpretation (see Section 2.5). The upper bounds also lead to practical guidelines on the rates of problem-related parameters, which are given by Corollary 2.5 and Corollary 2.6.

While Theorem 2.4 provides an upper bound on the excess risk, for this bound to be almost dimension-independent, we require the *effective dimension* to be small. Section 3 first formally defines the general notion of *low effective dimension*, which can essentially be described by 1) the loss function has a fast decaying Hessian spectrum, and 2) the true parameter is either weak with bounded l_2 norm or uniformly bounded along Hessian's eigen-directions. Figure 1 shows the relationships between the main theoretical results derived in this paper.

In Section 4, we carefully investigate the excess risk in various linear models. We consider the cases of finite projections of infinite-dimensional models and linear regression with redundant features. In these scenarios, we quantify when the overparameterization does not hurt the generalization performance.

Our generalization result can also be applied to a wide range of nonlinear models. In Section 5, we study both convex nonlinear models such as logistic regression and non-convex models such as M-estimator with the Tukey's biweight loss function [61] and two-layer neural networks. We show that the low effective dimension naturally occurs in these applications.

1.2. Related works

In recent years, understanding the phenomenons including *benign overfitting* via the excess risk bounds for different models in overparameterized settings have been carefully investigated in the literature, especially for liner models as in [7, 47, 31, 60, 18, 10, 42, 46, 43] and references therein. Our result is different from these existing results in the following perspectives:

1) Nonlinearity: Our results can be applied to general nonlinear models while most of these works focus on linear models. For example, [7] established non-asymptotic excess risk bound of the *minimum-norm interpolator* for overparameterized linear regression and [60] further generalized the results

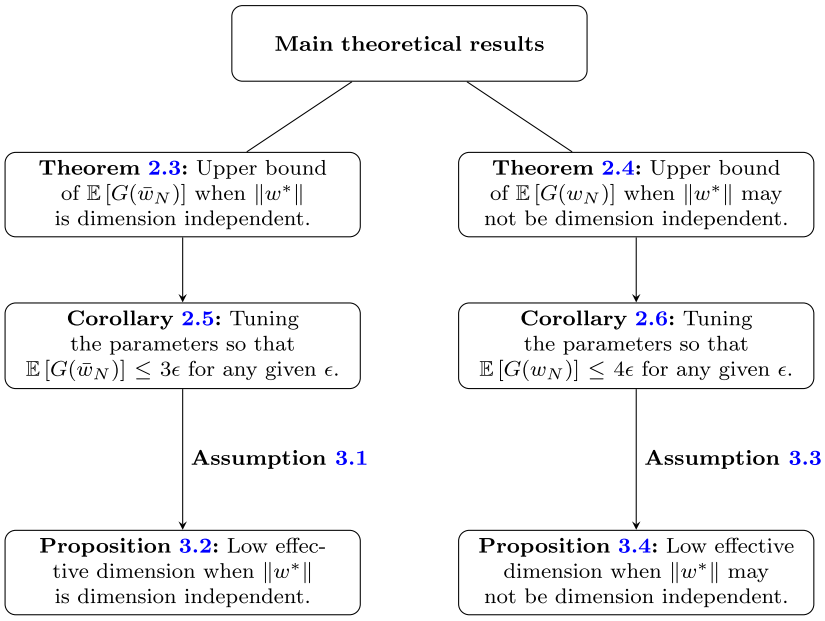


FIG 1. *The roadmap of the main theoretical results.*

to the case with ridge regularization. They defined some notions of *effective ranks* based on the feature covariance matrix and established their sufficient and necessary conditions for guaranteeing the excess risk to be small. The main insight is that, the feature family needs to satisfy a delicate balance between having a few important directions that favor the true signal (unknown function) and a large number of unimportant directions that absorb the noise in a harmless manner. Such a trade-off was also found for linear regression with different feature families under different settings ([10, 42, 46, 43]). Our result does not require such a trade-off, and a small excess risk can be achieved by thresholding the eigenvalues of the covariance matrix by the regularization parameter. Moreover, our defined *effective dimension* is simpler in formulation and can be easily checked, a more detailed comparison between our low effective dimension conditions and the low effective ranks in [7, 60] will be provided in Section 3.3. In addition, [31] developed the excess risk bound for the composition of an activation function and a linear model.

2) Anisotropic spectrum and regularization: [47] and [31] focused their studies on cases with isotropic or well-conditioned regressor covariance matrices. Our study focuses on cases with anisotropic regressor covariance, of which the minimum eigenvalue decays to zero. Moreover, since online learning has the implicit regularization effect and the ridge penalty is applied in certain scenarios, we do not have the "double descent" phenomenon as

in other literature [9, 8]. A recent paper by [48] also showed that for certain linear regression models with isotropic data distribution, the ridge penalty regularized regression (in the offline setting) can avoid the "double descent" phenomenon.

3) Online optimization: The aforementioned works mainly focus on offline optimization. For example, [7], [47], and [31] showed that the excess risk of the offline linear regression solution is closely related to the spectrum of the design matrix. In particular, when the minimum eigenvalue is close to zero, the offline learning results become unstable because of singular matrix-inversions. Since the design matrix relies on data realization, such instability can only be studied through random matrix theories (RMT). While these studies of offline linear regression are interesting and technically deep, their dependence on RMT makes the extensions to nonlinear models difficult. In comparison, online learning methods process one data point at a time and do not involve the inversion of design matrices, which facilitates the study of general nonlinear models. Moreover, in terms of practical applications, online optimization methods such as SGD are appealing due to their low per-iteration complexity compared to offline optimization. Therefore, this paper focuses on the excess risk for online learning in overparameterized settings.

For online optimization, the stochastic gradient descent (SGD), which dates back to [57], is perhaps the most widely used method in practice. The convergence rates of the SGD for different models have been well studied in the literature, especially on finite dimension (see, e.g., [19, 34, 32, 51, 50, 6]). For example, for the averaged SGD with constant stepsize, [6] provided an excess risk bound with the optimal convergence rate of $O(1/N)$. However, their bound has an explicit linear dependence on p, which is not applicable to the overparameterized setting. To this end, under similar setting, [20, 11] extended the optimal convergence rate to infinite-dimensional case under the framework of reproducing kernel Hilbert space (RKHS), [67] did their analysis from the operator view of averaged SGD and provided a sharp excess risk bound incorporating the full eigenspectrum of the data covariance matrix. It was shown in [22] that the convergence rate of SGD can even be accelerated by employing regularization. [40] obtained the optimal convergence rate for SGD when multiple passes over the data and mini-batches are allowed, and a further averaging procedure is considered by [45] for faster convergence rate. Although tight excess risk bounds are established in these mentioned works, they only considered the loss function of the least square loss or slightly more general. To go beyond this setting, [41] studied learning methods based on the regularized convex empirical risks for generalized linear model. As they noted, it would be interesting to extend the research to algorithms used to minimize the empirical risk such as SGD, which is exactly the scope of this paper. [36] extended the learning rate analysis of averaged SGD with multiple passes to general convex loss functions. But they considered the case of decaying stepsize, which is less challenging and practical compared with the constant one. [21] did an analy-

sis of the convergence for constant-step-size SGD in the strongly convex case and the properties of the limit distribution. The core technical contribution of this paper is providing dimension independent excess risk bounds of SGD with constant stepsize for general loss function (even non-convex) with regularization. Based on them, one can better understand how the generalization ability of SGD depends on the convexity, regularization, stepsize and sample size. We do not consider the mini-batch, multiple passes and tail averaging in this paper and leave them for our future consideration. Furthermore, most of the results in the aforementioned works are established on the so called *capacity condition* (CC) that quantifies the rate at which the covariance operator's eigenvalues decay and *source condition* (SC) that quantifies the rate at which the coefficients of the optimal predictor decay in an eigenbasis of eigenvectors of the covariance operator. The definitions of these conditions are highly correlated with the *effective dimension* given in this paper, as compared in Section 3.3.

Overparameterized neural network (NN) is a very active research direction. There are several existing works explaining why overfitting does not happen in large NN [27, 39, 49, 2, 3, 24, 25]. Interestingly, the conditions they impose are largely similar to the ones we will use. Namely, they require the high dimensional input data and the Frobenius norm of true weight matrices to be bounded by constants. For this to be true, only a small portion of the data or model components can be significantly active, which satisfies the concept of low effective dimension.

On the other hand, our study of two-layer NN in Section 5.3 is different from existing results in the following perspectives. One popular way to analyze generalization performance is by the Rademacher complexity in an offline optimization setting [27, 49, 24] or the sequential Rademacher complexity for online learning [54, 12, 53, 55], which can be used to establish an upper bound of generalization error, namely the difference between the empirical risk and the population risk (see, e.g., Theorem 3.3 of [44]). However, it is well known that the computation of the empirical Rademacher complexity is NP-hard for some hypothesis sets. Our paper focuses on the specific online algorithm SGD and derives the bound by exploiting the data covariance structure, which can be easily computed. [39] and [2] both studied NN generalization property with SGD iterations. But they mainly focused on classification scenarios where the loss function is bounded. Moreover, [39] required the loss function to be of a logistic form, and [2] studied the running average generalization error. Our results can be applied to regression NN with unbounded loss functions. [3] and [25] studied NN population risk bound with gradient descent (GD) and the neural tangent kernel (NTK) regime. Moreover, their studies assume a certain data angle or Gram matrix to have a strictly positive minimal eigenvalue. Please see Section 5.3 for more detailed comparisons with these works.

2. Excess risk bound

In this section, we present a general result on the excess risk bound for the SGD solution from (1.4).

2.1. Preliminaries: high dimensional norms and non-convex energy landscape

The main issue this paper tries to understand is the high dimensional excess risk when using SGD as the training method. In some simple scenarios, the true model parameters are "sparse" or dense but many of them are small, so that $\|w^*\|$ does not grow with the dimension. This allows us to measure the distance between SGD iterate w_n and w^* directly. In other scenarios, w^* can be very dense, and $\|w^*\|$ may grow linearly or even faster with the dimension. To resolve this issue, we define the following norms:

Definition 2.1. *Given a matrix $A \in \mathbb{R}^{p\times p}$ and λ, we decompose $\mathbb{R}^p = S_\lambda \oplus S_\perp$, where S_λ consists of eigenvectors of A with eigenvalues above $\lambda \geq 0$ and $S_\perp$ is the orthogonal complement of S_λ. Given any vector v, denote its decomposition as $v = v_\lambda + v_\perp$, where $v_\lambda \in S_\lambda$ and $v_\perp \in S_\perp$. Then define*

$$\|v\|_A^2 = v^T A v, \quad \|v\|_{A,\lambda}^2 := \lambda \|v_\lambda\|^2 + v_\perp^T A v_\perp. \tag{2.1}$$

We introduce the norm $\|v\|_A^2$, whose value can be independent of the ambient dimension p. The second norm $\|v\|_{A,\lambda}^2$ is a truncated version of the first norm, it essentially truncates all eigenvalues of A above λ to λ. It is easy to see that $\|v\|_{A,\lambda}^2 \leq \|v\|_A^2$. We introduce the second norm because $\|v\|_{A,\lambda}^2$ converges to zero when the regularization parameter λ does, while $\|v\|_A^2$ is independent of λ.

It is well known that SGD works well for convex problems. This is also true for our theoretical analysis, which works best in convex settings. On the other hand, many practical problems are non-convex, which makes the statistical learning problem technically more challenging. First, the function F can have multiple local minima, and each local minimum has an attraction basin, which is a "valley" in the graph of F. Within each valley, we assume that F is locally convex. Machine learning and theoretical deep learning literature often study the scenarios that local minima lie in large and shallow valleys, which lead to more stable generalization performance. Suppose $\mathcal{D}$ is the attraction basin of the optimal solution w^* in (1.1), initializing SGD in $\mathcal{D}$ will generate iterates converging to w^* with high probability. So a natural question is the gap between the estimator learned from SGD and w^*. On the other hand, if the SGD iterates take place outside $\mathcal{D}$, the output can be irrelevant to the properties of w^*. Therefore, we need to introduce a stopping time τ, which describes the first time SGD exits $\mathcal{D}$:

$$\tau = \min\{n : w_n \notin \mathcal{D}\},$$

where w_n is the n-th SGD iterate defined in (1.4). Our generalization analysis will assume the initialization $w_0 \in \mathcal{D}$ and the SGD iterates always stay within $\mathcal{D}$. We will also provide bounds of probability that SGD leaves $\mathcal{D}$.

While it is reasonable to assume that w^* as a local minimum is in the valley $\mathcal{D}$, verifying this assumption can be difficult. For example, because we do not have access to the population loss function F or the Hessian directly, to check the convexity of F, we need to investigate $\widehat{F}$ instead. There will be a certain inaccuracy due to the randomness in $\widehat{F}$. As another example, while we know the Hessian of F is positive semidefinite at w^* because w^* is a local minimum, F does not have to be convex in the neighborhood of w^*. For both these examples, it is more accurate to say F is "approximately convex" in $\mathcal{D}$. Our analysis can also extend to such very challenging cases by introducing proper regularizations. On the other hand, if F is very non-convex in $\mathcal{D}$, then one should not expect that SGD will produce good learning results. Therefore, to achieve a reasonable excess risk, we need F to be very close to convex. In many applications, this can be done by choosing either a very large sample size N or a very small neighborhood around w^*.

Based on our discussion, we have the following assumption on the population risk function.

Assumption 2.2. *The optimal solution w^* of the population risk F has a neighborhood $\mathcal{D}$, such that for some positive semidefinite (PSD) matrix A and $\delta \in [0, 1/2)$,*

$$-\delta A \preceq \nabla^2 F(w) \preceq A, \quad \forall\, w \in \mathcal{D}. \tag{2.2}$$

In Assumption 2.2 and in the sequel, for two symmetric matrices C, D, $C \preceq D$ indicates that $D - C$ is positive semidefinite (PSD). The upper bound on the Hessian matrix $\nabla^2 F(w) \preceq A$ is widely assumed in the statistical literature. The parameter δ above describes the level of non-convexity. In particular, $\delta = 0$ indicates that F is convex within $\mathcal{D}$. However, our condition in (2.2) is more general since δ can be strictly positive, which allows F to be non-convex. On the other hand, although our excess risk bound holds for $\delta \in [0, 1/2)$, for this upper bound to be smaller than a certain threshold, δ needs to be small. Please see Corollary 2.5 and Corollary 2.6 for the exact dependence of δ in the upper bound.

2.2. Excess risk bound with weak true parameter

We will discuss two possible scenarios with high dimensional machine learning. In the first case, the true parameter w^* is sparse or weak so that its l_2 norm does not depend on the dimension. We will refer such setting as a "weak" true parameter setting, which can be found in the literature under more involved setups [28, 62]. In this case, we have the following results:

Theorem 2.3. *Under Assumption 2.2, suppose $w_0 \in \mathcal{D}$ and there are constants r and c_r such that*

$$\mathbb{E}\|\nabla f(w, \zeta) - \nabla F(w)\|^2 \le r^2 + c_r |(w - w^*)^T \nabla F(w)|, \quad \forall w \in \mathcal{D}. \tag{2.3}$$

Then if the SGD stepsize η and the regularization parameter λ satisfy

$$\eta \leq \min\left\{\frac{1}{4(1+c_r)(\|A\|+\lambda)}, 1\right\}, \quad 2\delta\|A\| \leq \lambda \leq 1,$$

we have the excess risk for the averaged SGD,

$$\mathbb{E}\left[1_{\tau \geq N-1} G(\bar{w}_N)\right] \leq \frac{2\mathbb{E}\|w_0 - w^*\|^2}{N\eta} + 2\eta r^2 + 8\lambda\|w^*\|^2, \tag{2.4}$$

where the excess risk G is defined in (1.5) and $\bar{w}_N = \frac{1}{N}\sum_{i=0}^{N-1} w_i$ is the averaged SGD iterate.

The excess risk bound in (2.4) contains three terms. The first term decays with the sample size N. The second term is controlled by the stepsize η and the variance r^2 of the gradient on each individual data ζ. The last term $8\lambda\|w^*\|^2$ is a bias caused by the regularization. Since F_λ is different from F, the minimizer of F_λ is also different from w^*. As introduced in [14], the excess risk can be decomposed into three errors: The *approximation error* measures how closely the attraction basin $\mathcal{D}$ can approximate the optimal solution w^* (our paper does not have this error since we assume $w^* \in \mathcal{D}$); the *estimation error* measures the effect of minimizing the empirical risk instead of the population risk; the *optimization error* measures the distance between the minima of the empirical risk and the output generated by some optimization algorithms such as SGD. Instead of separately considering these errors as in [14] and analyzing their effects on the generalization performance, our result provides a unified upper bound for the estimation error and optimization error, and the regularization also brings in extra error, as we can see from the last term. For convex case, [30] obtained an excess risk bound by decomposing the risk estimates into an optimization error term and a stability term, and according to their Theorem 5.2,

$$\mathbb{E}\left[G(\bar{w}_N)\right] \leq \frac{\mathbb{E}\|w_0 - w^*\|^2}{2N\eta} + \frac{1}{2}\eta L^2,$$

with $\|\nabla f(w,\zeta)\| \leq L$. Similar bounding result with different constant coefficients is also established for general stochastic mirror descent methods in [35] (Theorem 4.1). Our result (2.4) does not require the boundedness of the stochastic gradient, and such a condition is usually not satisfied in many standard contexts, such as the simple least squares regression when the model parameter belongs to an unbounded domain. Although [37] removed the bounded gradient condition, Theorem 4 therein implies that their excess risk bound increases linearly with the value of $F(w^*)$, which may not be dimension independent. The excess risk bounds in [14, 6] are even looser than the one in [30] since they involve a term p/N increasing linearly with the dimension p thus can not be used in the overparameterized setting. Moreover, as discussed in [13], classical results are established on the smoothness of the objective function while we do not have this requirement.

One special case is when F is locally convex in $\mathcal{D}$ (i.e., $\delta = 0$ in (2.2)), then it is actually better to do rigdeless regression (i.e. $\lambda = 0$) since then the last

term in (2.4) can be removed. Note that in this case, the excess risk bound in (2.4) is dimension independent when $\|w^*\|^2$ is dimension independent with the initialization $w_0 = \mathbf{0}$. We see that the overfitting can be avoided even without explicit regularization, this is because SGD has a similar regularizing effect so that the minimizer of the empirical risk can achieve good generalization with dimension independent error bound. Such effect is often referred to as *implicit regularization* (see the discussions in [19, 22] and Section 10 of [4]). Moreover, without the regularization, our bound in (2.4) leads to a convergence rate of $O(1/N)$. Such a rate is known to be optimal for least-squares regression, as discussed in [6]. When Assumption 2.2 holds with $\delta > 0$, namely the population loss function F is non-convex within $\mathcal{D}$, then the regularization is necessary since we require $0 < 2\delta\|A\| \leq \lambda \leq 1$, and it will increase the excess risk accordingly.

Another problem is the guarantee regarding the initialization $w_0 \in \mathcal{D}$ since we have no prior information of w^* and its attraction basin $\mathcal{D}$. One possible way would be separating the data into two parts. On the first part of data, we implement Langevin dynamics (LD) or some other sampling-based algorithms to approximate the global minimal point w^* and find its neighbourhood $\mathcal{D}$. Sampling-based algorithms escape the local minima and explore the state space by adding stochasticity on the searching direction (see, e.g., [17, 63, 52, 23] and references therein). Specifically, the algorithms work by simulating an ergodic stochastic process for which the invariant measure is proportional to $\exp(-\frac{1}{\gamma}F(x))$, where γ is referred to as the "temperature" controlling the strength of stochasticity. It is easy to see that the sampling points concentrate around the global minimal w^*, especially for a smaller value of γ. Then for the second part of data, we take the output of the first stage as an initial point and run SGD on it. Such procedure can be further leveraged to outsource the computational cost during the first stage, while keeping the second part of data private from the outsourced agents. See [65] for more details. Based on the estimated attraction basin $\mathcal{D}$, we can also guarantee that the SGD iterates always stay within this region by abandoning those ones escaping $\mathcal{D}$. We do not discuss the detailed realization of this idea since this is not the scope of this paper. Interested readers can also refer to [23] for a collaboration scheme between LD and SGD to find the global minima.

2.3. Excess risk bound with strong true parameter

In some scenarios, $\|w^*\|$ may grow with the dimension. While the results in Theorem 2.3 still hold, the estimate (2.4) is no longer dimension independent. Since the derivation of the excess risk bound for the averaged SGD $\bar{w}_N$ in (2.4) is based on the bound of $\|w_N - w^*\|$, which is not realizable when $\|w^*\|$ is not dimension independent, this makes us to consider the excess risk bound for the final iterate w_N instead. In this case, the learning cannot rely only on implicit regularization. We will need to introduce regularization and high dimensional norms as in Definition 2.1 for these problems.

Theorem 2.4. *Under Assumption 2.2, suppose $w_0 \in \mathcal{D}$ and there are constants r and c_r such that*

$$\mathbb{E}\|\nabla f(w,\zeta) - \nabla F(w)\|^2 \le r^2 + c_r r^2 \min\{G(w), \|w\|^2\}, \quad \forall w \in \mathcal{D}. \tag{2.5}$$

Then if the SGD hyper-parameters, the stepsize η and the regularization parameter λ, satisfy

$$\eta \le \min\left\{1, \frac{\lambda}{12\|A\|^2 + 6\lambda^2 + 6c_r r^2}, \frac{1}{12\|A\|}, \frac{\lambda}{6c_r\|A\|r^2}\right\}, \quad 4\delta\|A\| \le \lambda \le 1,$$

we have

$$\begin{aligned}\mathbb{E}[G(w_N)1_{\tau\ge N}] \le 4\|w^*\|_{A,\lambda}^2 + \frac{C_1}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^n)(\eta+\delta) \\ + \exp(-\tfrac{1}{4}\lambda N\eta)\mathbb{E}[G(w_0) + 4N\|A\|\|w_0\|^2],\end{aligned} \tag{2.6}$$

with $C_1 = 60\|A\|\left(r^2 + \|w^*\|_A^2\right) + 10\|w^*\|_A^2$.

In the upper bound (2.6), each term carries a strong statistical interpretation, which will be illustrated via linear models in Section 2.5. In particular, the term $\|w\|_{A,\lambda}^2$ in (2.6) can be interpreted as the bias caused by minimizing F_λ instead of F, it decays with the regularization parameter λ shrinking to zero. The term $\frac{C_1\eta}{\lambda}$ is the variance induced by the SGD algorithm, which increases as λ decreases. This reveals that under our current problem setting, λ controls a bias-variance tradeoff. Ideally, we can choose small λ and stepsize η to make both the bias and variance small. However, this comes with a price. As the convergence rate scales with $\lambda\eta$, so using small λ and η need to be compensated with a large sample size N (i.e., the number of iterations in SGD).

While it is not completely new that SGD on convex ridge regression has dimension independent generalization error [58, 66, 16], in general, these results need the norm of gradient ∇F to be dimension independent (See Section 14.5.3 in [58]). Theorem 2.4 does not have this restriction. In fact, the population gradient ∇F can be unbounded in many applications (e.g., linear regression). In contrast, our assumption for the stochastic gradient is imposed on its variance, see (2.5). This is a much relaxed assumption because the variance can be reduced by various techniques (e.g., [33]) or simply by increasing the mini-batch size for stochastic gradient computation.

2.4. SGD configuration with given excess risk target

In Section 3, we will explicitly define the low effective dimension so that C_1 and the upper bound in (2.6) are independent of dimension p, or depend on p only via a polynomial logarithmic factor. This differentiates our result from the estimates in existing literature on SGD, e.g., [5].

We then quantify the tradeoffs in (2.4) and (2.6) by considering a practical scenario where the excess risk is pre-fixed to be ϵ, then our results provide

guidelines on how to tune the parameters of the regularization λ, the stepsize λ, the non-convexity δ and the sample size N. We will see that, with proper parameterization, if the SGD manages to fit the training data in a reasonable number of iterations, the overfitting can be avoided without any restriction on the dimensionality p.

Corollary 2.5 (Corollary of Theorem 2.3)**.** *Suppose there is an universal constant C_0 such that*

$$\|w^*\|^2, \|w_0 - w^*\|^2 \le C_0.$$

Given any $\epsilon > 0$, if the regularization parameter $\lambda(\epsilon)$, the stepsize $\eta(\epsilon)$, the non-convexity parameter $\delta(\epsilon)$, and the sample size $N(\epsilon)$ satisfy

$$\begin{aligned} &\lambda(\epsilon) \le \min\left\{\frac{\epsilon}{8C_0}, 1\right\}, \quad \delta(\epsilon) \le \min\left\{\frac{\epsilon}{16C_0\|A\|}, \frac{1}{2\|A\|}\right\}, \\ &\eta(\epsilon) < \min\left\{\frac{1}{2(1+c_r)(\|A\|+\lambda(\epsilon))}, \frac{\epsilon}{2r^2}, 1\right\}, \quad N(\epsilon) > \frac{2C_0}{\epsilon\eta(\epsilon)}, \end{aligned} \tag{2.7}$$

and the conditions of Theorem 2.3 hold, then $\mathbb{E}[G(\bar{w}_N)1_{\tau \ge N}] \le 3\epsilon$.

Corollary 2.6 (Corollary of Theorem 2.4)**.** *Given any $\epsilon > 0$, if the regularization parameter $\lambda(\epsilon)$, the stepsize $\eta(\epsilon)$, the non-convexity parameter $\delta(\epsilon)$, and the sample size $N(\epsilon)$ satisfy*

$$\begin{aligned} &4\|w^*\|^2_{A,\lambda(\epsilon)} < \epsilon, \quad \delta(\epsilon) \le \frac{\lambda(\epsilon)\epsilon}{C_1}, \quad \eta(\epsilon) < \frac{\lambda(\epsilon)\epsilon}{C_1}, \\ &N(\epsilon) > \max\left\{\frac{-4\log\{\epsilon/2\mathbb{E}[G(w_0)]\}}{\lambda(\epsilon)\eta(\epsilon)}, \frac{-8\log\{\epsilon\lambda(\epsilon)\eta(\epsilon)/(64\|A\|\mathbb{E}[\|w_0\|^2])\}}{\lambda(\epsilon)\eta(\epsilon)}\right\}, \end{aligned} \tag{2.8}$$

and the conditions of Theorem 2.4 hold, then $\mathbb{E}[G(w_N)1_{\tau \ge N}] \le 4\epsilon$.

It is noteworthy that (2.6) only discusses the scenario where SGD iterates stay in the domain $\mathcal{D}$. This is necessary since all our conditions are imposed only within $\mathcal{D}$. Once an SGD iterate leaves $\mathcal{D}$, there is no particular reason it can get back to $\mathcal{D}$. Another possible improvement is to find an upper bound for the conditional excess risk $\mathbb{E}[G(w_N)|\tau \ge N]$. But this is not feasible when $\mathcal{D}$ is a general region. For example, if $\mathcal{D}$ is the intersection between any set and $\{w : G(w) \ge G(w_0) - 1\}$, the conditional expectation will be larger than $G(w_0) - 1$, while $\{\tau \ge N\}$ may have a nonzero occurrence probability.

In practice, SGD is often implemented with mini-batch data to reduce the noise within stochastic gradient. With regularization, the mini-batch SGD with batch-size J can be formally written as

$$w_{n+1} := w_n - \frac{1}{J}\eta \sum_{k=Jn+1}^{J(n+1)} (\nabla f(w_n, \zeta_k) + \lambda w_n),$$

where *i.i.d.* data $\{\zeta_{Jn+1}, \ldots, \zeta_{J(n+1)}\}$ forms the n-th batch of data. Our results can be extended to mini-batch SGD as well. To see this, we simply let $z_n =$

$\{\zeta_{Jn+1}, \ldots, \zeta_{J(n+1)}\}$ and

$$\tilde{f}(w, z_n) = \frac{1}{J} \sum_{k=Jn+1}^{J(n+1)} f(w_n, \zeta_k).$$

Note that $\mathbb{E}\tilde{f}(w, z_n) = \mathbb{E}f(w, \zeta_i) = F(w)$. It is also straightforward to see that applying our SGD formulation (1.4) on $\tilde{f}$ with z_n leads to the mini-batch SGD. Applying our results, e.g. Corollaries 2.5 and 2.6, to mini-batch SGD requires simple modifications for only two parameters. First, the variance of stochastic gradient $\nabla\tilde{f}(w, z)$ is only $\frac{1}{J}$ of the variance of $\nabla f(w, \zeta)$, so the parameter r^2 in mini-batch SGD should be $\frac{1}{J}$ of r^2 in the standard SGD. Second, because each iteration of the mini-batch SGD requires J data samples, so the overall sample size should be NJ. The simple analysis demonstrates that the consideration of mini-batch does not sacrifice the convergence rate, the same conclusion is also made in [40] with multiple pass. Rigorously analyzing how the excess error bound depends on the stepsize, the mini-batch size, the number of passes and the regularization parameter may be our future work.

2.5. Statistical interpretation in linear models and bias-variance tradeoff

To facilitate better understanding our results, we will use linear models to illustrate the statistical interpretation of each term in the excess risk upper bound in (2.6). In linear regression, each *i.i.d.* observation contains a pair of dependent and response variables, $\zeta_i = (x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, where y_i is generated by the following model

$$y_i = x_i^T w^* + \xi_i. \tag{2.9}$$

In (2.9), w^* is the true regression coefficient, and the noise term ξ_i is independent of x_i with zero mean and a finite variance σ^2. For the ease of illustration, we assume $x_i \sim \mathcal{N}(0, \Sigma)$. The excess risk of this linear model takes the following form,

$$G(w) = \mathbb{E}\Big[\frac{1}{2}(y_i - w^T x_i)^2 - \frac{1}{2}(y_i - (w^*)^T x_i)^2\Big] = \frac{1}{2}(w - w^*)^T \Sigma (w - w^*). \tag{2.10}$$

From (2.10), we can see that $G(w)$ has a strong dependence on the structure of Σ. Let us denote the eigenvalues of Σ by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, and the eigenvector corresponding to λ_i by v_i. Then the parameter error, $v_i^T(w - w^*)$, contributes to $G(w)$ via the factor λ_i.

It is well known that SGD can be interpreted as a stochastic approximation of the gradient descent [57], namely we can rewrite (1.4) as

$$w_{n+1} = w_n - \eta \nabla F_\lambda(w_n) + \eta \xi_n, \tag{2.11}$$

where $\xi_n = -\nabla f(w_n, \zeta_n) + \nabla F(w_n)$ is the noise in stochastic gradient. For the quadratic loss with the ridge penalty F_λ, the SGD iterates in (2.11) take the following form,

$$w_{n+1} = w_n - \eta\Sigma(w_n - w^*) - \eta\lambda w_n + \eta\xi_n = (I - (\Sigma + \lambda I)\eta)(w_n - w_\lambda^*) + w_\lambda^* + \eta\xi_n, \tag{2.12}$$

where $w_\lambda^* := (\Sigma + \lambda I)^{-1}\Sigma w^*$ is the minimizer of $F_\lambda(w) = \mathbb{E}\frac{1}{2}(y - w^T x)^2 + \frac{\lambda}{2}\|w\|^2$.

It is easy to see that w_n then follows a vector autoregressive (VAR) model (see, e.g., Chapter 8 of [59]). For the ease of discussion, we simply treat ξ_n as $\mathcal{N}(0, \frac{r^2}{p}I)$, so $\mathbb{E}\|\xi_n\|^2 = r^2$ as we assumed in Theorem 2.4. Then the stationary distribution of w_n in (2.12) is a Gaussian $\mathcal{N}(\mu, V)$. After taking expectation and covariance on both sides of (2.12), we obtain μ and V with the following form (see Chapter 8.2.2 of [59]),

$$\mu = (I - (\Sigma + \lambda I)\eta)(\mu - w_\lambda^*) + w_\lambda^* \Rightarrow \mu = w_\lambda^*,$$
$$V = (I - (\Sigma + \lambda I)\eta)V(I - (\Sigma + \lambda I)\eta) + \frac{\eta^2 r^2}{p}I.$$

When the stepsize $\eta \le \|\Sigma + \lambda I\|^{-1}$, $(\Sigma + \lambda I)^2\eta \preceq \Sigma + \lambda I$, we get

$$V = \frac{r^2\eta}{p}(2(\Sigma + \lambda I) - (\Sigma + \lambda I)^2\eta)^{-1} \preceq \frac{r^2\eta}{p}(\Sigma + \lambda I)^{-1}.$$

These results give us the limiting average excess risk

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}G(w_n) &= \frac{1}{2}(w_\lambda^* - w^*)^T\Sigma(w_\lambda^* - w^*) + \frac{1}{2}\operatorname{tr}(V\Sigma) \\ &\le G(w_\lambda^*) + \frac{r^2\eta}{2p}\operatorname{tr}((\Sigma + \lambda I)^{-1}\Sigma).\end{aligned} \tag{2.13}$$

The first term $G(w_\lambda^*)$ is the bias caused by using regularization. Indeed, the optimizer of F_λ is w_λ^* rather than w^*. Recall that (λ_i, v_i) are the eigenvalues and eigenvectors of Σ. We define $a_i = \langle v_i, w^*\rangle$ and further express $G(w_\lambda^*)$ in (2.13) as follows,

$$\begin{aligned}G(w_\lambda^*) &= \frac{1}{2}((\Sigma + \lambda I)^{-1}\Sigma w^* - w^*)^T\Sigma((\Sigma + \lambda I)^{-1}\Sigma w^* - w^*) \\ &= \frac{1}{2}\lambda^2(w^*)^T(\Sigma + \lambda I)^{-1}\Sigma(\Sigma + \lambda I)^{-1}w^* = \frac{1}{2}\sum_{i=1}^p \frac{\lambda^2\lambda_i a_i^2}{(\lambda + \lambda_i)^2}.\end{aligned}$$

Note that when $\lambda_i \ge 0$, $\frac{\lambda^2\lambda_i}{(\lambda_i+\lambda)^2} \le \frac{\lambda^2\lambda_i}{\lambda^2} = \lambda_i$, and by Young's inequality $\frac{\lambda^2\lambda_i}{(\lambda_i+\lambda)^2} \le \frac{\lambda^2\lambda_i}{4\lambda_i\lambda} \le \lambda$. Therefore, we have the following upper bound of $G(w_\lambda^*)$

$$G(w_\lambda^*) \le \frac{1}{2}\sum_{i=1}^p (\lambda \wedge \lambda_i)a_i^2 = \frac{1}{2}\|w^*\|_{\Sigma,\lambda}. \tag{2.14}$$

The upper bound in (2.14) is essentially the first term in (2.6) by noticing that $\Sigma = A = \nabla^2 F(w)$ in linear regression, which gives an upper bound for the bias.

For the second variance term in (2.13),

$$\mathrm{var}(\lambda) := \frac{r^2\eta}{2p}\,\mathrm{tr}((\Sigma+\lambda I)^{-1}\Sigma) = \frac{r^2\eta}{2p}\sum_{i=1}^{p}\frac{\lambda_i}{\lambda_i+\lambda} \le \frac{r^2\eta}{2p}\sum_{i=1}^{p}\frac{\lambda_i}{\lambda} = \frac{\eta r^2\lambda_1}{2\lambda}. \tag{2.15}$$

This upper bound is essentially the second term in the excess risk upper bound in (2.6), as it depends linearly on $\eta, \lambda^{-1}, \lambda_1 r^2$.

The first two terms in (2.6) are based on the limiting average excess risk. With finite SGD iterations, the iterate w_n may not reach the limiting distribution. On the other hand, for VAR models, it is well known that the speed of convergence for w_n is exponential, and the convergence rate is closely related to the minimum eigenvalue $\lambda_{\min}((\Sigma+\lambda I)\eta) = \lambda\eta$ (see, e.g., [59] Chapter 8.2.2). The finite iterate error leads to the third term of $\exp(-\frac{1}{4}\lambda\eta N)$ in excess risk bound in (2.6).

In the special case that $\lambda = 0$, (2.13) reduces to

$$\lim_{n\to\infty}\mathbb{E}G(w_n) \le G(w^*) + \frac{r^2\eta}{2p}\,\mathrm{tr}(I) = r^2\eta.$$

Finally, we consider the scenario where Σ is indefinite with $\delta = -\lambda_{\min}(\Sigma) > 0$. While the population loss F is non-convex, by adopting $\lambda > 2\delta$, we have that $\Sigma + \lambda I$ is positive definite and F_λ is convex. Then the excess risk upper bounds need to be updated by replacing λ with $\lambda - \delta$, which leads to a perturbation on the order of δ. In particular, note that by Young's inequality, the derivative of the bias term with respect to λ is bounded by

$$|\partial_\lambda G(w_\lambda^*)| = \sum_{i=1}^{p}\frac{\lambda\lambda_i^2 a_i^2}{(\lambda+\lambda_i)^3} \le \sum_{i=1}^{p}\frac{\lambda_i a_i^2}{4(\lambda+\lambda_i)} \le \frac{1}{4\lambda}\sum_{i=1}^{p}\lambda_i a_i^2 = \frac{\|w^*\|_\Sigma^2}{4\lambda},$$

the derivative of the variance term with respect to λ in (2.15) is bounded by

$$|\partial_\lambda \mathrm{var}(\lambda)| = \frac{r^2\eta}{2p}\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+\lambda)^2} \le \frac{r^2\eta}{2p}\sum_{i=1}^{p}\frac{\lambda_1}{\lambda^2} \le \frac{r^2\eta\lambda_1}{2\lambda^2}.$$

Therefore, replacing λ with $\lambda - \delta$ to handle non-convexity, we need to add the following term in the excess risk bound,

$$\delta|\partial_\lambda F(w_\lambda^*)| + \delta|\partial_\lambda \mathrm{var}(\lambda)| \le \delta\left(\frac{\|w^*\|_\Sigma^2}{4\lambda} + \frac{r^2\lambda_1\eta}{2\lambda^2}\right).$$

This term can be further upper bounded by the second term of (2.6).

3. Low effective dimension

Given the excess risk bounds in Theorem 2.3 and Theorem 2.4, we introduce the concept of "low effective dimension" and show that the excess risk bounds

in (2.4) and (2.6) can be independent (or dependent poly-logarithmically) of the ambient dimension p in an overparameterized regime. We will use the O and Ω notations to hide constants independent of p and use the $\widetilde{O}$ and $\widetilde{\Omega}$ notations to hide constants depend poly-logarithmically on p. In particular, we introduce the following standard asymptotic notations: $A_\epsilon = O(f(\epsilon)), B_\epsilon = \widetilde{O}(f(\epsilon)), C_\epsilon = \Omega(f(\epsilon)), D_\epsilon = \widetilde{\Omega}(f(\epsilon))$. These notations mean that there exist some universal constants c and $C > 0$ such that,

$$A_\epsilon \le Cf(\epsilon), \quad B_\epsilon \le C(\log p)^c f(\epsilon), \quad C_\epsilon \ge Cf(\epsilon), \quad D_\epsilon \ge C(\log p)^c f(\epsilon).$$

3.1. Initialization and stochastic gradient variance

We investigate the terms that appear in the excess risk bound (2.6): whether they can be independent of p; and how they affect the necessary sample size $N(\epsilon)$ in (2.8).

First, we notice that the terms related to initialization w_0, i.e., $\mathbb{E}\|w_0\|^2$ and $\mathbb{E}G(w_0)$, appear in the sample size $N(\epsilon)$ in (2.8). If the region $\mathcal{D} = \mathbb{R}^p$, we can often choose appropriate w_0 so that $\mathbb{E}\|w_0\|^2$ and $\mathbb{E}G(w_0)$ are independent of p. For example, for linear regression loss function in (2.10), we can pick $w_0 = \mathbf{0}$, then $\mathbb{E}G(w_0) = \frac{1}{2}\|w^*\|_A^2$ with $A = \Sigma$, which will be bounded by an $O(1)$ constant as shown below. For a restrictive region $\mathcal{D}$, although $\mathbb{E}\|w_0\|^2$ and $\mathbb{E}G(w_0)$ may scale as a polynomial function of p, $N(\epsilon)$ only depends logarithmically on these two terms. Therefore, the dimension dependence of $N(\epsilon)$ is only logarithmic.

Second, we consider the stochastic gradient variance r^2, which contributes to the term C_1 in (2.6). In a typical setting, it scales roughly as the squared population gradient, i.e.,

$$\begin{aligned}\mathbb{E}\|\nabla f(w,\zeta) - \nabla F(w)\|^2 &\approx O(\mathbb{E}\|\nabla F(w)\|^2)\\ &= O(\mathbb{E}\|\nabla F(w) - \nabla(F(w^*))\|^2)\\ &= O(\mathbb{E}\|\nabla^2 F(w)(w - w^*)\|^2)\\ &\quad \text{Assume that } w \sim \mathcal{N}(0, I_p) \text{ and } \nabla^2 F \preceq A\\ &= O\big(\|A\|(\|w^*\|_A^2 + \operatorname{tr}(A))\big).\end{aligned}$$

We will see such an approximation holds for many applications of interest. Moreover, we have $\|A\| \le \operatorname{tr}(A)$, which can often be p-independent as discussed below. The scale of $\|w^*\|_A$ will also be discussed next.

From the discussion above, we only need to focus on two terms in (2.6), $\|w^*\|_A$ and $\|w^*\|_{A,\lambda}$. For the excess risk to be small and independent of p, we need to show $\|w^*\|_A$ is dimension independent and $\|w^*\|_{A,\lambda}$ decreases as λ decreases.

3.2. Low effective dimension settings

In this section, we formally define two settings of low effective dimension as Assumptions 3.1 and 3.3. In Sections 4 and 5, we will show that these assumptions easily hold for a wide range of convex and non-convex statistical models.

3.2.1. Weak true parameter

The first setting is characterized in the following assumption.

Assumption 3.1. *The followings are true*

1) $\|A\|$ *with* A *defined in Assumption 2.2 is bounded by an* $O(1)$ *constant.*
2) $\|w^*\|$ *is bounded by an* $O(1)$ *constant.*
3) r^2, c_r *defined in Theorem 2.3 are bounded by* $O(1)$ *constants.*
4) The initial values $\mathbb{E}\|w_0\|^2$ *and* $\mathbb{E}G(w_0)$ *grow polynomially with* p.

Assumption 3.1 can be interpreted as a weak sparsity condition for w^*, since there can be only a few significant components in w^*. Sparsity assumption is a very common condition in the statistical literature. However, our assumption only assumes that the ℓ_2-norm of w^*, instead of the ℓ_0-norm, is bounded. As compared to the ℓ_0-norm, the ℓ_2-norm is rotation-free. In addition, we do not need to apply any projection or shrinkage procedures on the SGD iterates.

Under Assumption 3.1, Corollary 2.5 can be simplified as the following excess risk bound, which shows that the necessary sample size depends on p in a poly-logarithmic factor.

Proposition 3.2. *Under the conditions in Corollary 2.5 and Assumption 3.1, given any* $\epsilon > 0$, *when*

$$\lambda(\epsilon) = O(\epsilon),\ \delta(\epsilon) = O(\epsilon),\ \eta(\epsilon) = O(\epsilon),\ N(\epsilon) = \Omega\Big(\frac{1}{\epsilon^2}\Big), \tag{3.1}$$

we have $\mathbb{E}[G(\bar{w}_N)1_{\tau \geq N-1}] \leq 3\epsilon$. *Moreover, for any* $p_0 \geq 0$, *there are* $a = O(1/p_0)$, *if* $\mathcal{D} = \{w : \|w - w^*\|^2 \leq a\}$, *we have*

$$\mathbf{P}(\tau \leq N) \leq p_0.$$

Alternatively, if $\delta = 0$, *for any* $\alpha > 0$, *we take*

$$\lambda(\epsilon) = 0,\ \eta(\epsilon) = O(\epsilon^{1+\alpha}),\ N(\epsilon) = \Omega\Big(\frac{1}{\epsilon^{2+\alpha}}\Big),$$

we have $\mathbb{E}[G(\bar{w}_N)1_{\tau \geq N-1}] \leq 3\epsilon$. *Meanwhile, for any* $a > \mathbb{E}\|w_0 - w^*\|^2$, *if* $\mathcal{D} = \{w : \|w - w^*\|^2 \leq a\}$, *we have*

$$\mathbf{P}(\tau \leq N) \leq \frac{\mathbb{E}\|w_0 - w^*\|^2 + O(\epsilon^\alpha)}{a}.$$

Proposition 3.2 consists of two parts. The first part shows that the excess risk is of order $O(1/\sqrt{N})$ if the non-convexity is of the same order. The second part demonstrates how to bound the probability of SGD escaping the convex region $\mathcal{D}$ if it is a ball centered at w^*. For all $p_0 \geq 0$, a needs to be of order $1/p_0$ so that the chance of escaping is less than p_0. If the problem is convex in $\mathcal{D}$, $\mathcal{D}$ just need to include w_0 to ensure the chance of no-escaping is nonzero. In both cases, the escape is harder when the radius $\sqrt{a}$ is larger. This also explains why machine learning literature is in favor of local-minima in large valleys.

3.2.2. Strong true parameter

The second setting is technically more interesting, which assumes the data has a low effective dimension in the following sense. When we say a component or a linear combination of components of w is effective, it means that the loss function F has a significant dependence on it. This can be analyzed through the eigen-decomposition of $\nabla^2 F(w)$ or its upper bound A in (2.2). Let (λ_i, v_i) be the eigenvalue-eigenvectors of A, where λ_i are arranged in decreasing order. Then a small λ_i indicates that F has a weak dependence along the direction of v_i. For the model to have a low effective dimension, there will be only constantly many λ_i being significant, while the remaining eigenvalues in sum have a negligible contribution to the overall loss function. We formally formulate this setting into the following assumption.

Assumption 3.3. *The followings are true*

1) $\mathrm{tr}(A)$ *with* A *defined in Assumption 2.2 is bounded by an* $\widetilde{O}(1)$ *constant.*
2) In each of A*'s eigen-direction, the true parameter* w^* *is bounded by an* $\widetilde{O}(1)$ *constant, in the sense that for the following spectrum-based quantity* $\|w^*\|_{A,S}$*,*

$$\|w^*\|_{A,S} := \max_i \{|\langle v_i, w^*\rangle|, i = 1, \ldots, p\} = \widetilde{O}(1). \tag{3.2}$$

3) r^2, c_r *defined in Theorem 2.4 are bounded by* $\widetilde{O}(1)$ *constants.*
4) The initial values $\mathbb{E}\|w_0\|^2$ *and* $\mathbb{E}G(w_0)$ *grow polynomially with* p*.*

By Cauchy Schwartz inequality, we have $\|w^*\|_{A,S} \le \|w^*\|$. So Assumption 3.3 condition 2) is weaker than Assumption 3.1 condition 2). In particular, it can include important cases where we only have upper and lower bounds on each of w^*'s components, and A is known to be a diagonal matrix. These cases are not covered by Assumption 3.1. On the other hand, the spectrum profile of A will be required to choose the regularization parameter as shown in the following proposition.

Proposition 3.4. *By the following inequalities,*

$$\|w^*\|_A^2 \le \mathrm{tr}(A)\|w^*\|_{A,S}^2, \quad \|w^*\|_{A,\lambda}^2 \le \|w^*\|_{A,S}^2 \sum_{i=1}^p \lambda \wedge \lambda_i.$$

Assumption 3.3 implies that $\|w^*\|_A = \widetilde{O}(1)$ *and* $\|w^*\|_{A,\lambda}^2 = \widetilde{O}\left(\sum_{i=1}^p \lambda \wedge \lambda_i\right)$*. Moreover, under the conditions in Corollary 2.6 and Assumption 3.3, given any* $\epsilon > 0$*, if the eigenvalues of* A *follows,*

1) Exponential decay: $\lambda_i = e^{-ci}$ *for some constant* $c > 0$*, and setting*

$$\lambda = \widetilde{O}\left(\frac{\epsilon}{|\log \epsilon|}\right), \ \delta(\epsilon) = \widetilde{O}\left(\frac{\epsilon^3}{|\log \epsilon|^2}\right),$$
$$\eta(\epsilon) = \widetilde{O}\left(\frac{\epsilon^2}{|\log \epsilon|}\right), \ N(\epsilon) = \widetilde{\Omega}\left(\frac{|\log \epsilon|^3}{\epsilon^3}\right),$$

we have $\mathbb{E}[G(w_N)1_{\tau\geq N}] \leq 4\epsilon$.

2) Polynomial decay: $\lambda_i = i^{-c}$ *for some constant* $c > 0$, *and setting*

$$\lambda(\epsilon) = \widetilde{O}\Big(\epsilon^{\frac{c+1}{c}}\Big),\ \delta(\epsilon) = \widetilde{O}\Big(\epsilon^{\frac{3c+2}{c}}\Big),$$
$$\eta(\epsilon) = \widetilde{O}\Big(\epsilon^{\frac{2c+1}{c}}\Big),\ N(\epsilon) = \widetilde{\Omega}\Big(\frac{|\log(\epsilon)|}{\epsilon^{\frac{3c+2}{c}}}\Big),$$

we have $\mathbb{E}[G(w_N)1_{\tau\geq N}] \leq 4\epsilon$.

In both cases, we have

$$\mathbb{E}[G(w_{N\wedge\tau})] \leq \mathbb{E}[G(w_0)] + \widetilde{O}(\epsilon|\log\epsilon|).$$

So if $\mathcal{D} = \{w : G(w) \leq (1+a)\mathbb{E}[G(w_0)]\}$, *then*

$$\mathbf{P}(\tau < N) \leq \frac{1}{1+a} + \widetilde{O}(\epsilon|\log\epsilon|).$$

We remark that the parameter of the spectrum decay (e.g., the constant c in polynomial decay spectrum) is often assumed to be known for many functional data analysis problems [29, 15]. From Proposition 3.4, for both exponential decay and polynomial decay of the Hessian spectrum, the sample size N only depends on p in a poly-logarithmic factor.

Similar to Proposition 3.2, the second part of this result demonstrates how to bound the probability of SGD escaping the convex region $\mathcal{D}$ if it is the sub-level set with w^* inside. The parameter a controls the size of $\mathcal{D}$. A larger a produces a larger $\mathcal{D}$ and hence a smaller escape probability.

3.3. Comparisons with related works

The excess risk bounds of SGD have been well studied under different scenarios for least square regression and other convex setting, the main purpose of this paper is extending the analysis to general setting (convex or non-convex) for overparameterized model. We show that, with the defined low effective dimension, our excess risk can be independent of the dimension. It is also interesting to compare our low effective dimension settings with the conditions used in other literature.

Some existing works established the excess risk bounds via a quantity called *effective rank* k^*, which is the minimum integer k such that the cumulative summation of the first k eigenvalues of the data covariance exceeds some value. For the main result, Theorem 1 in [7], to yield dimension-independent excess risk bound, three conditions (formulated in our notation) need to hold: 1) $\|w^*\|^2$ is bounded by a constant; 2) $\mathrm{tr}(A)$ is bounded by a constant; 3) the spectrum of A decays not so fast so that the effective rank $k^* := \min\{k \geq 0 : \sum_{i\geq k}\lambda_i \geq bN\lambda_k\}$ for some constant $b > 0$ is not so large, meanwhile, the spectrum of A should decay fast enough so that $\lambda^* := \sum_{i>k^*}\lambda_i^2/(\sum_{i>k^*}\lambda_i)^2$ is small. In comparison, our Assumption 3.1 only requires conditions 1) and 2), but not the technical

condition 3). Moreover, our result can also work under Assumption 3.3 where only $\|w^*\|_{A,S}^2$, instead of $\|w^*\|^2$, needs to be bounded. Similar condition as 3) is also needed in [60] when regularization is considered, with $k^* := \min\{k \geq 0 : (\sum_{i\geq k} \lambda_i + \lambda) \geq bN\lambda_k\}$ and $\lambda^* := \sum_{i>k^*} \lambda_i^2/(\lambda + \sum_{i>k^*} \lambda_i)^2$. And Theorem 2.1 in [67] requires the spectrum decays fast enough so that both k^*/N and $\lambda^* := \sum_{i>k^*} \lambda_i^2$, with $k^* := \max\{k \geq 0 : \lambda_k \geq 1/(\eta N)\}$, are small. On one hand, it is not clear how fast should the spectrum decays. On the other hand, finding the unknown quantities k^* and λ^* to guarantee the excess risk to be small is not easy. Our conclusion (2.6) works for any decaying spectrum with finite $\|A\|$ and a small excess risk can be achieved by choosing a small enough λ so that $\|w^*\|_{A,\lambda}$ is small and tuning other parameters, under our defined effective dimension.

Tighter excess risk bounds can be achieved under the so-called *capacity condition* (CC) and *source condition* (SC) for problems with linear structure (see e.g. [20]). CC quantifies the rate at which the covariance operator's eigenvalues decay, and SC quantifies the rate at which the coefficients of the optimal predictor decay in eigen-directions of the covariance operator. Moreover, these capacity and source conditions can be further exploited to gain improved variance and bias terms, thus obtaining faster convergence rate. Mathematically, CC is defined as $\lambda_k \propto 1/k^\alpha$ for some $\alpha \geq 1$ or $\text{tr}(\Sigma^\beta) < \infty$ for some $\beta \leq 1$, and SC is defined as $\|\Sigma^{\gamma/2} w^*\| < \infty$ for some $\gamma \leq 1$. Here Σ is data covariance matrix, and it plays a similar role as A in our framework. Different excess risk bounds or convergence rates are established for different values of α, β, γ in [20, 11, 22, 67, 40, 45, 41] and references therein. These conditions are close to our Assumption 3.1 of bounded $\|\Sigma\|$ and $\|w^*\|$. But our Assumption 3.3 is more general since $\lambda_k \propto 1/k^\alpha$ for some $\alpha \geq 1$ implies $\text{tr}(\Sigma) < \infty$, but not vice versa. Meanwhile, our condition (3.2) imposes bounds on w^*, which is similar to SC.

4. Overparameterization in linear regression

In general, overparameterization may lead to overfitting, but this sometimes can be avoided. Our main result, Theorem 2.4, provides a general tool to understand why overfitting sometimes happens and sometimes does not. In this section, we will demonstrate how to apply our results on linear regression models in various high dimensional settings. This section is technically straightforward and is mainly used for pedagogical purpose. The discussions on more technically challenging cases for nonlinear and non-convex models are provided in the next section.

4.1. Linear regression

First of all, we will find out the problem related parameters in Theorem 2.4 when applying to linear regression models. As in Section 2.5, we consider *i.i.d.* data points form $\zeta_i = (x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, where the response is generated by

$$y_i = x_i^T w^* + \xi_i. \tag{4.1}$$

In (4.1), $w^* \in \mathbb{R}^p$ is the true model-parameter to be estimated. $\xi_i \in \mathbb{R}$ are observation noise terms in the observation process, and we assume they are *i.i.d.* with zero mean and variance σ^2. For simplicity, we assume that the data x_i are *i.i.d.* Gaussian distributed, i.e., $x_i \sim \mathcal{N}(0, \Sigma)$. As a remark, our proof also allows the non-Gaussian distribution with finite fourth moments.

The regression loss of parameter w on data ζ_i is

$$f(w, \zeta_i) = \frac{1}{2}(x_i^T w - y_i)^2. \tag{4.2}$$

Plugging (4.1) into (4.2) and taking expectation, and we find the population loss function

$$F(w) = \frac{1}{2}(w - w^*)^T \Sigma (w - w^*) + \frac{1}{2}\sigma^2. \tag{4.3}$$

Now we show the problem related parameters in Theorem 2.4 can be set as below:

Proposition 4.1. *For linear regression, Assumption 2.2 holds with $A = \Sigma$, $\delta = 0$, $\mathcal{D} = \mathbb{R}^p$. When $w_0 = 0, \mathbb{E}G(w_0) = \frac{1}{2}\|w^*\|_\Sigma^2$, the stochastic gradient variance bounds in (2.3) and (2.5) hold with*

$$r^2 = 2\sigma^2 \operatorname{tr}(\Sigma) + 12 \operatorname{tr}(\Sigma)\|w^*\|_\Sigma^2, \quad c_r = \frac{6}{\sigma^2}\max\{\|\Sigma\|, 1\}.$$

As a consequence, Assumption 3.1 holds if $\|w^*\|^2$ and $\operatorname{tr}(\Sigma)$ are $O(1)$ constants, and by Proposition 3.2, the necessary sample size $N(\epsilon)$ is independent of p. Similarly, Assumption 3.3 holds if $\|w^*\|_{A,S}^2$ and $\operatorname{tr}(\Sigma)$ are $\widetilde{O}(1)$ constants, and by Proposition 3.4, the sample size $N(\epsilon)$ depends on p only via a polynomial logarithmic factor.

4.2. High dimensional data with principal components

The low effective dimension settings in Section 3 naturally rise in many high dimensional problems. For example, in image processing or functional data analysis (see e.g. [56, 29, 15, 26, 64]), the data are in general assumed to take place in a Hilbert space $(H, \langle \cdot, \cdot \rangle)$ with potentially infinitely many orthonormal basis functions $\{e^j, j = 1, 2, \ldots\}$. Each data can be written as

$$x_i = \sum_{j=1}^{\infty} a_i^j e^j. \tag{4.4}$$

Suppose a_i^j are independent Gaussian random variables with mean zero and variance σ_j^2. Note that $\mathbb{E}\langle x_i, x_i \rangle = \sum_{j=1}^{\infty} \sigma_j^2$. Therefore, for each data $x_i \in H$, we assume that $\sum_{j=1}^{\infty} \sigma_j^2 < \infty$ so that the norm of the data is bounded, which implicitly requires σ_j decaying to zero ([29]). Given the form of x_i in (4.4), the linear regression model takes the following form,

$$y_i = \langle x_i, w^* \rangle + \xi_i, \tag{4.5}$$

where $w^* = \sum_{j=1}^{\infty} w^{*,j} e^j$. If we assume $w^* \in H$, then $\langle w^*, w^* \rangle = \sum_{j=1}^{\infty} (w^{*,j})^2 < \infty$.

When training this "infinite dimensional" linear regression model in (4.5), we would need a finite projection $\mathcal{P}_p : H \mapsto \mathbb{R}^p$. When the basis functions are available, one natural choice of the projection is

$$\mathcal{P}_p x_i = \mathcal{P}_p \left(\sum_{j=1}^{\infty} a_i^j e^j \right) := [a_i^1, \dots a_i^p]^T.$$

Then the p-dimensional linear regression model is formulated as

$$y_i = (\mathcal{P}_p x_i)^T w_p^* + \xi_i^p. \tag{4.6}$$

It is worthwhile noticing that the true infinite dimensional model in (4.5) is compatible with the finite dimensional model in (4.6), in the sense that

$$w_p^* = \mathcal{P}_p w^* = [w^{*,1}, \dots, w^{*,p}]^T, \quad \xi_i^p = \xi_i + \sum_{j=p+1}^{\infty} w^{*,j} a_i^j.$$

Since a_i^j are independent Gaussian random variables, we have $\xi_i^p \sim \mathcal{N}(0, \sigma_{\xi,p}^2)$ with

$$\sigma_{\xi,p}^2 := \sigma^2 + \sum_{j=p+1}^{\infty} \sigma_j^2 (w^{*,j})^2 \le \sigma^2 + \|w^*\|_{\Sigma}^2, \quad \|w^*\|_{\Sigma}^2 := \sum_{j=1}^{\infty} \sigma_j^2 (w^{*,j})^2.$$

In the finite dimensional model (4.6), the data $\mathcal{P}_p x_i$ has the population covariance matrix $\Sigma_p = \operatorname{diag}(\sigma_1^2, \dots, \sigma_p^2)$, whose trace is bounded by

$$\operatorname{tr}(\Sigma_p) = \sum_{j=1}^{p} \sigma_j^2 \le \sum_{j=1}^{\infty} \sigma_j^2.$$

Therefore, by Proposition 4.1, the problem related parameters in Theorem 2.4 are

$$\begin{aligned} &A_p = \Sigma_p, \quad \text{with} \quad \operatorname{tr}(A_p) \le \sum_{j=1}^{\infty} \sigma_j^2, \quad \|A_p\| = \sigma_1^2, \\ &r_p^2 = 2\sigma_{\xi,p}^2 \operatorname{tr}(\Sigma_p) + 12 \operatorname{tr}(\Sigma_p) \|w_p^*\|_{\Sigma_p}^2 \le 2(\sigma^2 + 7\|w^*\|_{\Sigma}^2) \sum_{j=1}^{\infty} \sigma_j^2, \\ &c_{r,p}^2 = \frac{6}{\sigma_{\xi,p}^2} \max\{1, \sigma_1^2\} \le \frac{6}{\sigma^2} \max\{1, \sigma_1^2\}. \end{aligned} \tag{4.7}$$

Moreover, if we use $w_0 = 0$, then $\mathbb{E}\|w_0\|^2 = 0$ and

$$\mathbb{E}G(w_0) = \frac{1}{2}\|w_p^*\|_{\Sigma_p}^2 \le \|w^*\|_{\Sigma}^2.$$

Note that

$$\|w^*\|_\Sigma^2 = \sum_{j=1}^{\infty} \sigma_j^2 (w^{*,j})^2 \le \|w^*\|_\infty^2 \sum_{j=1}^{\infty} \sigma_j^2, \quad \|w^*\|_\infty := \max_{1\le j} |w^{*,j}|.$$

So as long as $\|w^*\|_\infty$ is finite, the upper bounds above are independent of dimension p.

When the true loading parameter w^* is an element of H, $\|w_p^*\|^2 \le \langle w^*, w^*\rangle < \infty$. Then we can check that all items of Assumption 3.1 hold. So by Proposition 3.2, we know that the excess risk is dimension-independent. Moreover, this does not require any information of the spectrum decay profile.

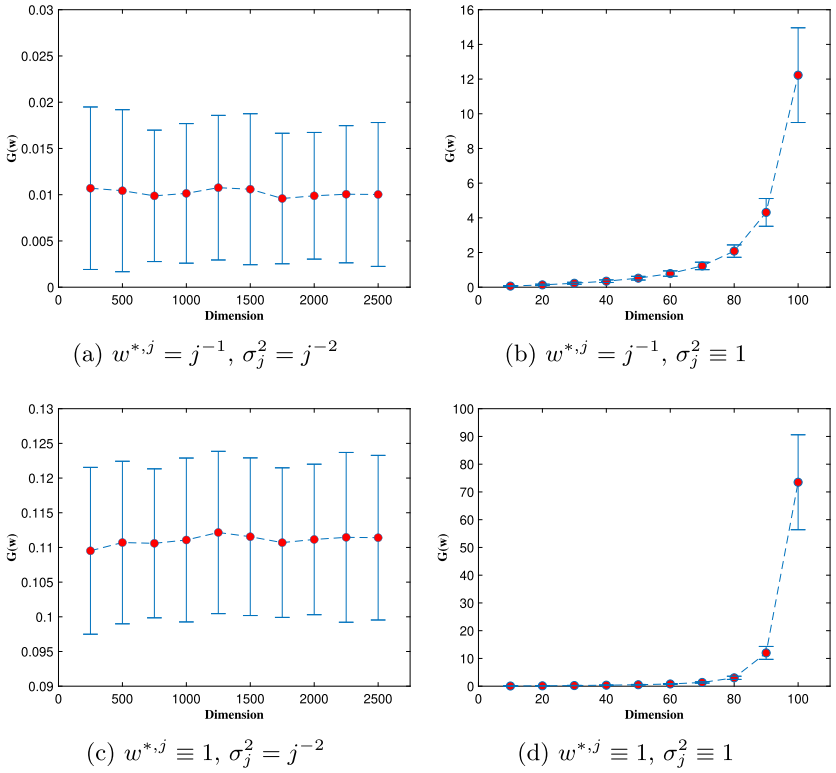
More generally, we only need that each component of the true loading parameter w^* is bounded, and w^* does not need to be an element of H itself. In particular, we note that

$$\|w_p^*\|_{\Sigma_p,S} = \max_{1\le j\le p} |w^{*,j}| \le \|w^*\|_\infty.$$

Therefore, if $\|w^*\|_\infty$ is finite, Assumption 3.3 holds (but in general Assumption 3.1 does not). Then by Proposition 3.4, the excess risk can be dimension independent when we know the spectrum decay profile.

As a simple demonstration, we run some simulations of SGD on linear regression model (4.6) and present them in Figure 2. We run SGD on (4.6) with the sample size $N = 500$, the initial value $w_0 = \mathbf{0}$, the stepsize $\eta = 0.02$, the variance of the noise $\sigma^2 = 1$ and the regularization parameter $\lambda = 0.01$. The covariance spectrum of predictors is set to be $\sigma_j^2 = j^{-2}$ so that $\mathrm{tr}(A_p)$ in (4.7) is a constant, and the true parameter is set to be $w^{*,j} = j^{-1}$ for $1 \le j \le p$ so that $\|w^*\|_\Sigma$ is bounded. The problem dimension ranges from $p = 250$ to $p = 2500$, which can be larger than the sample size. We use the final SGD output w_{500} as the estimator and compute the excess risk as in (2.10). We repeat this experiment 1000 times and compute the mean and standard deviation. We plot the error bar plot in the upper left panel of Figure 2. As one can see, the excess risk does not increase as the dimension increases, even when $p \gg N$. As a comparison experiment, we run simulations with the same settings except for $\sigma_j^2 \equiv 1$. We plot the excess risk in the upper right panel of Figure 2, which clearly shows the overfitting phenomenon, even when the dimension is in a lower range.

The similar story repeats when the true parameter $w^{*,j} \equiv 1$ (i.e., the case when $\|w^*\|_\infty$ is bounded), where the plots are given by the lower panels in Figure 2. As one can see, the excess risk with the decaying spectrum still remains stable against the increase of dimension, and it does not change much from the previous setting where the components of w^* are decaying. Meanwhile, overfitting with constant spectrum (i.e., $\sigma_j^2 \equiv 1$) becomes stronger. This simple illustrative example justifies that the low effective dimension helps to address the overfitting issue, even when the dimension p is much larger than the sample size N.



(a) $w^{*,j} = j^{-1}$, $\sigma_j^2 = j^{-2}$ (b) $w^{*,j} = j^{-1}$, $\sigma_j^2 \equiv 1$

(c) $w^{*,j} \equiv 1$, $\sigma_j^2 = j^{-2}$ (d) $w^{*,j} \equiv 1$, $\sigma_j^2 \equiv 1$

FIG 2. *Excess risk bar plot with high dimensional linear regression for different settings of* w_p^* *and* Σ_p. *The x-axis is the dimension and y-axis is the excess risk.*

4.3. Overfitting with redundant features

Another interesting setting of overparameterization is to consider adding redundant predictors to an existing model. In this scenario, the true model is low dimensional with the true parameter $w^* \in \mathbb{R}^d$. Suppose we do not know the true model and collect additional features $z \in \mathbb{R}^{p-d}$, so that the overparameterized linear model is written as

$$y_i = x_i^T w^* + z_i^T u^* + \xi_i, \tag{4.8}$$

where $w^* \in \mathbb{R}^d$, $u^* = 0$, and $[x_i; z_i]$ is jointly Gaussian with mean zero and covariance

$$\Sigma_p = \begin{bmatrix} \Sigma_x & B \\ B^T & \Sigma_z \end{bmatrix}.$$

We assume that $\|\Sigma_z\| \leq \|\Sigma_x\|$ for the ease of discussion. Then, Σ_p being PSD implies that $\|B\| \leq \|\Sigma_x\|$. Since we do not impose any restriction on B other than Σ_p being PSD, our setting allows the possibility that some components of

z_i to be highly correlated or even identical with the ones of x_i. This, in general, leads to highly singular design matrices and unstable offline learning results.

We apply Proposition 4.1 and find $A_p = \Sigma_p$. By triangular inequality, for any vectors x and z,

$$\|\Sigma_p[x;z]\| = \|[\Sigma_x x + Bz; B^T x + \Sigma_z z]\| \le 2\|\Sigma_x\|\|[x;z]\| \Rightarrow \|\Sigma_p\| \le 2\|\Sigma_x\|.$$

For simplicity, we initialize with $[w_0; u_0] = 0$, so

$$\mathbb{E}G(w_0, u_0) = \frac{1}{2}\|w_0 - w^*\|^2_{\Sigma_x} + \frac{1}{2}\|u_0\|^2_{\Sigma_z} = \frac{1}{2}\|w^*\|^2_{\Sigma_x}.$$

Moreover, we have

$$c_{r,p} = \frac{6}{\sigma^2}\max\{\|\Sigma_p\|, 1\} \le \frac{6}{\sigma^2}\max\{2\|\Sigma_x\|, 1\},$$

and

$$\|[w^*, u^*]\|_{\Sigma_p, S} = \|w^*\|_\infty, \quad \|[w^*, u^*]\|^2 = \|w^*\|^2.$$

These upper bounds are all independent of p, or the choice of Σ_z and B.

Meanwhile,

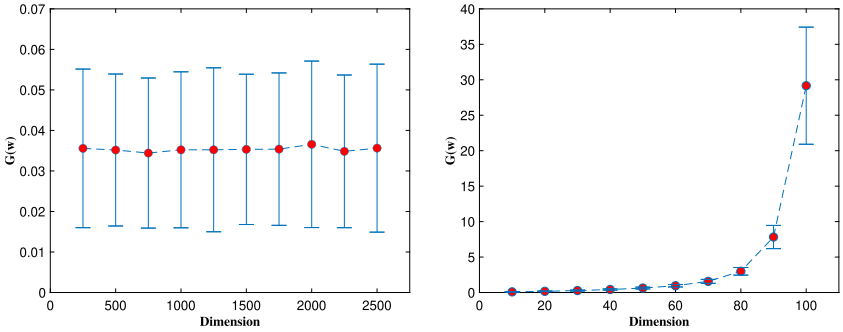
$$r_p^2 = 2(\sigma^2 + 6\|w^*\|^2_{\Sigma_x})(\mathrm{tr}(\Sigma_x) + \mathrm{tr}(\Sigma_z)), \quad \mathrm{tr}(A_p) = \mathrm{tr}(\Sigma_x) + \mathrm{tr}(\Sigma_z). \tag{4.9}$$

Given these simple calculation, we find that the only problem related parameters that depend on z are r_p^2 and $\mathrm{tr}(A_p)$ in (4.9) through $\mathrm{tr}(\Sigma_z)$. Therefore, our theory indicates that there is a simple dichotomy on whether the model (4.8) will overfit.

If $\mathrm{tr}(\Sigma_z)$ is bounded by a constant independent of p, Proposition 3.2 applies, which indicates that the excess risk is also independent of the ambient dimension p. This can happen if we select data features in z as PCA components. For example, suppose that the redundant data is in the form of $\sum_{j=1}^\infty a_i^j e^j$ as in the setting of (4.4), and we collect the $p-d$ dimensional principal components as $z_i = [a_i^1, \ldots, a_i^{p-d}]^T$. Then $\mathrm{tr}(\Sigma_z) = \sum_{j=1}^{p-d}\sigma_j^2 < \sum_{j=1}^\infty \sigma_j^2$, which is independent of p.

If $\mathrm{tr}(\Sigma_z)$ grows with p, model (4.8) may overfit. For simplicity, we consider a special case where $\Sigma_z = I_{p-d}, B = 0$. In other words, the redundant features are independent with each other and the features of x. Then our derivation shows that $r_p^2 = O(p)$. This indicates that the learning results may overfit.

To demonstrate this dichotomy, we simulate the SGD learning results and present their excess risk in Figure 3. In particular, we set $d = 5$ with true parameter $w^* = [1, 1, 1, 1, 1], \Sigma_x = I_5, \sigma^2 = 1$. We let $B = 0$ and choose first that Σ_z to be diagonal with decaying entries $\frac{1}{j^2}$. We run SGD with 500 iterations and compute the excess risk of the final iterate. We repeat this 1000 times and plot the error bar plot in the left panel of Figure 3. As we can see, the excess risk is stable against the increase of the dimension p. In comparison, if we use $\Sigma_z = I_{p-d}$, the learning results overfit, as we can see from the right panel of Figure 3.



(a) Σ_z is a diagonal matrix with decaying entry $(\Sigma_z)_{jj} = \frac{1}{j^2}$.

(b) $\Sigma_z = I_{p-d}$.

FIG 3. *Excess risk bar plot with high dimensional redundant features for two different cases of* Σ_z. *The x-axis is the dimension p and y-axis is the excess risk.*

5. Overparameterization for nonlinear and non-convex models

In this section, we apply our main theorems and corollaries in Section 3 to several important nonlinear and non-convex statistical problems, such as logistic regression, M-estimator with Tukey's biweight loss function, and two-layer neural networks.

5.1. Logistic regression

We consider the logistic regression for binary classification with N *i.i.d.* data $\zeta_i = (x_i, y_i)$. The binary response y_i takes values within $\{-1, 1\}$ with probability

$$\mathbf{P}(y_i = y|x_i) = \frac{1}{1 + \exp(-yx_i^T w^*)}, \quad y = \pm 1,$$

where $w^* \in \mathbb{R}^p$ is the true parameter to be estimated. We assume the predictors x_i are *i.i.d.* with $\mathbb{E}x_i x_i^T = \Sigma$. For each data, we adopt the negative log-likelihood as the loss function

$$f(w, \zeta_i) := \log(1 + \exp(-y_i x_i^T w)),$$

and the corresponding population loss is given by

$$F(w) = \mathbb{E}f(w, \zeta) = \mathbb{E}\log(1 + \exp(-yx^T w)).$$

The problem related parameters in Theorem 2.4 can be set by the following proposition.

Proposition 5.1. *For logistic regression, Assumption 2.2 holds with* $A = \Sigma, \delta = 0, \mathcal{D} = \mathbb{R}^p$. *When* $w_0 = 0, \mathbb{E}G(w_0) = \log 2 = O(1)$, *the stochastic gradient variance bounds in* (2.3) *and* (2.5) *hold with*

$$r^2 = \mathrm{tr}(\Sigma), \quad c_r = 0.$$

As a consequence, Assumption 3.1 holds if $\|w^*\|^2$ and $\mathrm{tr}(\Sigma)$ are $O(1)$ constants, and by Proposition 3.2, the sample size $N(\epsilon)$ is independent of p. Similarly, Assumption 3.3 holds if $\|w^*\|^2_{\Sigma,S}$ and $\mathrm{tr}(\Sigma)$ are $\widetilde{O}(1)$ constants, and by Proposition 3.4, the sample size $N(\epsilon)$ depends on p only via a poly-logarithmic factor.

5.2. M-estimator with Tukey's biweight loss function

In this non-convex example, we assume that the data $\zeta_i = (x_i, y_i)$ are generated from a linear model

$$y_i = x_i^T w^* + \xi_i. \tag{5.1}$$

We assume $x_i \sim \mathcal{N}(0, \Sigma)$, and ξ_i are *i.i.d.* mean-zero noises with finite fourth moment. We adopt the *non-convex* Tukey's biweight loss function as follows for the purpose of robust estimation

$$\rho(u) = \begin{cases} \frac{c^2}{6}[1-(1-(u/c)^2)^3] & \text{if } |u| \le c; \\ \frac{c^2}{6} & \text{if } |u| > c. \end{cases}$$

Then the individual data loss function and the population loss are given by,

$$f(w,\zeta) = \rho(x^T w - y) = \rho(x^T(w-w^*) - \xi), \quad F(w) = \mathbb{E}\rho(x^T(w-w^*)-\xi).$$

Proposition 5.2. *For the M-estimator with Tukey's biweight loss in (5.1), the model true parameter w^* is a local minimum if and only if*

$$c_0 = \mathbb{E}[(1-(\xi/c)^2)(1-5(\xi/c)^2)1_{|\xi|\le c}] > 0.$$

In that case, Assumption 2.2 holds with any $\delta \ge 0$, $A = \Sigma$ and

$$\mathcal{D} = \{w : \|w - w^*\|_\Sigma \le \tfrac{c_0+\delta}{16}\}.$$

Moreover, the stochastic gradient variance bounds in (2.3) and (2.5) hold with

$$r^2 = \mathrm{tr}(\Sigma), \quad c_r = 0.$$

Since $G(w_0) \le \max_u \rho(u) = \frac{c^2}{6}$, Assumption 3.1 holds if $\|w^*\|^2, \|w_0\|^2$ and $\mathrm{tr}(\Sigma)$ are $O(1)$ constants, and by Proposition 3.2, the sample size $N(\epsilon)$ is independent of p. We also see that for satisfying $\mathbb{E}[G(\bar{w}_n)] \le 3\epsilon$, the order of the radius of $\mathcal{D}$ only need to be $\frac{c_0+\epsilon}{16}$. Similarly, Assumption 3.3 holds if $\|w_0\|_\infty, \|w^*\|^2_{\Sigma,S}$ and $\mathrm{tr}(\Sigma)$ are $\widetilde{O}(1)$ constants, and by Proposition 3.4, the sample size $N(\epsilon)$ depends on p only via a polynomial logarithmic factor.

5.3. Two-layer neural network

In this example, we consider applying our result to two-layer neural networks (NN). We assume that every data point $\zeta = (x, y)$ consists of a p-dimensional

predictor $x \sim \mathcal{N}(0, \Sigma)$ and a univariate response $y \in \mathbb{R}$. We assume that the response is generated by

$$y = g(w, x) + \xi, \quad \mathbb{E}\xi = 0, \quad \mathbb{E}\xi^2 = \sigma_0^2.$$

The function g takes the form of a two-layer NN:

$$g(w, x) = c^T \psi(bx + a) = \sum_{i=1}^{k} c_i \psi(b_i^T x + a_i). \tag{5.2}$$

In (5.2), a and c are k-dimensional vectors with a_i and c_i being their components. The notation b is a p by k matrix, and b_i denotes the i-th column of b with $i = 1, \ldots, k$. We impose no restriction on k and it can depend on p in general. We denote all the parameters by $w = [a; b_1, \ldots, b_k; c] \in \mathbb{R}^{(p+2)k}$. In (5.2), ψ denotes the activation function. Popular choices of ψ include the rectified linear unit (ReLu), sigmoid function, and the hyperbolic tangent. Here, we do not require ψ to take a specific form but only satisfy certain regularity assumptions for some constant $C > 0$,

$$\psi(0) = 0, \quad |\dot{\psi}(x)| \le C, \quad |\ddot{\psi}(x)| \le C. \tag{5.3}$$

It is easy to verify that hyperbolic tangent satisfies these requirements, and the sigmoid also satisfies these if we shift its center to zero. The condition $\psi(0) = 0$ is mainly for the ease of technical derivations. Although ReLu does not have continuous derivatives, one can find a smooth approximation to meet these requirements.

Since we consider the regression problem, the squared loss function is given by

$$f(w, \zeta) = (y - g(w, x))^2 = (g(w^*, x) + \xi - g(w, x))^2.$$

We also introduce the following $(p + 2)k$ by $(p + 2)k$ block-diagonal matrix

$$\Sigma^\star = \operatorname{diag}\{I_k, \Sigma, \Sigma \ldots, \Sigma, I_k\}.$$

This matrix introduces a high dimensional norm

$$\|w\|_{\Sigma^\star}^2 = w^T \Sigma^\star w = \|a\|^2 + \sum_{i=1}^{k} \|b_i\|_\Sigma^2 + \|c\|^2.$$

Recall that b_i is of dimension p, its contribution to $\|w\|_{\Sigma^\star}^2$ is $\|b_i\|_\Sigma^2$. By Proposition 3.4, $\|b_i\|_\Sigma^2 \le \operatorname{tr}(\Sigma)\|b_i\|_{\Sigma,S}^2$, which can be independent of p under suitable conditions.

We are ready to show that the two-layer NN will not overfit in some overparameterized settings.

Proposition 5.3. *Assume the activation function satisfies the condition in* (5.3). *With the two-layer NN defined in* (5.2), *Assumption 2.2 holds for any* $\delta \in (0, 1/4]$ *with*

$$A = C_0(w^*)\Sigma^\star, \quad \mathcal{D} = \{w : \|w - w^*\|_{\Sigma^\star} \le \delta C_1(w^*)\|w^*\|_{\Sigma^\star}\}.$$

For any $w_0 \in \mathcal{D}, G(w_0) \le C_2(w^*)\|w^*\|_{\Sigma^\star}^4$, *the stochastic gradient variance bounds in* (2.3) *and* (2.5) *hold with*

$$r^2 = C_3(w^*), c_r = 0.$$

The exact values of the problem related parameters are given by

$$C_0(w^*) = 7C^2\|w^*\|_{\Sigma^\star}^2, \quad C_1(w^*) = \frac{2}{9\sqrt{2}(2\|w^*\|_{\Sigma^\star}+1)}, \quad C_2(w^*) = C^2\|w^*\|_{\Sigma^\star}^4,$$
$$C_3(w^*) = 8\sqrt{3}(1+\operatorname{tr}(\Sigma))C^2\|w^*\|_{\Sigma^\star}^2\big(C^2\|w^*\|_{\Sigma^\star}^4+\sigma_0^2\big).$$

Applying Corollary 2.5 and 2.6 yields that for satisfying $\mathbb{E}[G(\bar{w}_N)1_{\tau\ge N}] \le 3\epsilon$ and $\mathbb{E}[G(w_N)1_{\tau\ge N}] \le 4\epsilon$, the radius of the attraction basin $\mathcal{D}$ should be $O(\epsilon)$. Although this is too small, theoretically this is reasonable for NN with general structure. Figure 1 in [38] shows that for some network architectures, the loss landscapes can be very rough, thus it is unlikely that we can prove rigorously that NN all have minimizers reside in the valley with large radius. On the other hand, the empirical studies in [38] and references therein have shown that with some NNs will yield "flat valleys" in their landscape, which often are related to good generalization. Under our setup, such NNs have $\mathcal{D}$ with large radius, so our generalization theory can apply with much better excess risk.

As a consequence of Proposition 5.3, Assumption 3.1 holds if

$$\max\left\{\|a^*\|^2+\sum_{i=1}^k\|b_i^*\|^2+\|c^*\|^2, \operatorname{tr}(\Sigma), \|w_0\|_\infty\right\} = O(1), \tag{5.4}$$

and by Proposition 3.2, the sample size $N(\epsilon)$ is independent of p. Similarly, Assumption 3.3 holds if

$$\max\big\{k, \operatorname{tr}(\Sigma), \|w_0\|_\infty, |a_i|, |c_i|, |v_j^T b_i|, i=1,\dots,k, j=1,\dots,p\big\} = \widetilde{O}(1), \tag{5.5}$$

where v_j are the eigenvectors of Σ. Then by Proposition 3.4, the sample size $N(\epsilon)$ depends on p only via a polynomial logarithmic factor.

It is worthwhile mentioning that a similar version of Condition (5.4) can also be found in [49]. In particular, [49] assumed $\|c^*\|^2, \sum_{i=1}^k\|b_i^*\|^2$ to be $O(1)$ while the parameter a is set to be 0. There is no variance assumption of x_i in [49], but it is assumed that $\mathbb{E}\|x_i\|^2$ is $O(1)$, which is equivalent to requesting $\operatorname{tr}(\Sigma) = O(1)$ in our setting.

When the width of the hidden layer k is a fixed constant, the second condition (5.5) is in general less restrictive than the first one (5.4), since it allows $\|b_i^*\|$ to grow with p. When k grows with p, only the first condition is applicable, and it requires that $\|a^*\|^2+\sum_{i=1}^k\|b_i^*\|^2+\|c^*\|^2$ is bounded by $O(1)$. In other words, we need either $k = O(1)$ or the true parameters to be bounded by $O(1)$ to prevent overfitting. This can also be understood intuitively. Note that the output of the two-layer NN in (5.2) is a sum of k objects. Therefore, if k grows with p, the output of (5.2) will diverge, which contradicts the common assumption that g is bounded (see, e.g., [39, 49, 2]). Our result is consistent with the results

in [2], in the sense that [2] also showed that the sample size needs to grow with k. It is also possible to rescale g by multiplying (5.2) with a factor $\frac{1}{k}$ or $\frac{1}{\sqrt{k}}$, as done by [3], so that the excess risk is independent of the parameter k.

6. Conclusions and future works

One classical canon of statistics is that high dimensional models are prone to overfitting when the data sample size is not sufficiently large. However, many existing models, such as neural networks (NN), exhibit stable generalization performance despite being overparameterized. This paper developed an analysis framework of the excess risk bound for high dimensional regularized online learning. The error bound can be interpreted as a bias-variance tradeoff through a simplified stochastic approximation. This result indicates that overparameterization does not lead to overfitting if the model has a low effective dimension. We demonstrated how to apply this framework on various models such as linear regression, logistic regression, M-estimator with Tukey's biweight loss, and two-layer NN.

There are a few future directions. First, our excess risk bounds only apply when the starting point and the SGD iterates stay in a local region $\mathcal{D}$ near the true parameter w^*, it would be an interesting future work on how to realize our idea introduced in Section 2.2. Second, our framework indicates that the ambient model dimension itself may not be a good indicator of model complexity, especially in overparameterized settings. The quantity that characterizes the data variability may lead to new information criterion for model selection in the overparameterized setting. Such results may extend the classical criteria such as the AIC and BIC. Last but not least, regularization and overparameterization are good tools to handle misspecified models. [31] has discussed this issue for linear regression problems. How to extend our results to nonlinear misspecified models will be very interesting.

Appendix A: Proof of the main results in Section 2

A.1. Preliminaries

Lemma A.1. *For any vector $v \in \mathbb{R}^p$ and PSD matrix $A \in \mathbb{R}^{p\times p}$, the following results hold*

1) *For any $-\delta A \preceq B \preceq A$, let $B = V\Lambda V^T$ be the eigenvalue decomposition of B, and denote $|\Lambda|$ as taking absolute value on each element of the diagonal matrix Λ. Denote $|B| = V|\Lambda|V^T$. Then for any vectors v and w, $a > 0$*

$$2\langle v, Bw\rangle \le a\langle v, |B|v\rangle + \frac{1}{a}\langle w, |B|w\rangle \le a(1+\delta)\|v\|_A^2 + \frac{1+\delta}{a}\|w\|_A^2.$$

2) *For any $-\delta A \preceq B \preceq A$, and any vectors u and v, $a > 0$*

$$2|\langle u, Bv\rangle| \le a\langle u, Bu\rangle + 2a\delta\|u\|_A^2 + \frac{1+2\delta}{a}\|v\|_A^2.$$

Proof. Claim 1). Let (l_i, u_i) be the eigenvalue-eigenvectors of B. Assume also that

$$v = \sum_{i=1}^{p} a_i u_i, \quad w = \sum_{i=1}^{p} b_i u_i.$$

Then by Young's inequality

$$2\langle v, Bw\rangle = 2\sum_{i=1}^{p} l_i a_i b_i \le a\sum_{i=1}^{p} |l_i||a_i|^2 + \frac{1}{a}\sum_{i=1}^{p} |l_i||b_i|^2 = a\langle v, |B|v\rangle + \frac{\langle w, |B|w\rangle}{a}.$$

Next, we denote the positive part of Λ as Λ_+ and the negative part as Λ_-, so that

$$\Lambda = \Lambda_+ + \Lambda_-, \quad |\Lambda| = \Lambda_+ - \Lambda_-, \quad \Lambda_- \preceq 0 \preceq \Lambda_+.$$

Then by checking eigen-space with nonnegative eigenvalues, $B \preceq A$ indicates that $V\Lambda_+ V^T \preceq A$. Likewise, we have $-\delta A \preceq V\Lambda_- V^T$. In combination, we have

$$|B| = V\Lambda_+ V^T - V\Lambda_- V^T \preceq (1+\delta)A.$$

Therefore

$$a\langle v, |B|v\rangle + \frac{1}{a}\langle w, |B|w\rangle \le (1+\delta)a\|v\|_A^2 + \frac{1+\delta}{a}\|w\|_A^2.$$

For claim 2), denote $B_\delta = B + \delta A \succeq 0$. Then

$$\begin{aligned}
2|\langle u, Bv\rangle| &\le 2|\langle u, B_\delta v\rangle| + 2\delta|\langle u, Av\rangle| \\
&\le a\|u\|_{B_\delta}^2 + \frac{1}{a}\|v\|_{B_\delta}^2 + a\delta\|u\|_A^2 + \frac{\delta}{a}\|v\|_A^2 \\
&= a\langle u, Bu\rangle + a\delta\|u\|_A^2 + \frac{1}{a}\langle v, Bv\rangle + \frac{\delta}{a}\|v\|_A^2 + a\delta\|u\|_A^2 + \frac{\delta}{a}\|v\|_A^2 \\
&\le a\langle u, Bu\rangle + 2a\delta\|u\|_A^2 + \frac{1+2\delta}{a}\|v\|_A^2.
\end{aligned}$$

□

A.2. Proof of the main results

Proof of Theorem 2.3. We rewrite SGD update as

$$w_{n+1} = w_n - \eta\nabla f_\lambda(w_n, \zeta_n) = w_n - \eta\nabla F_\lambda(w_n) + \eta\xi_n, \tag{A.1}$$

where

$$\xi_n = \nabla F_\lambda(w_n) - \nabla f_\lambda(w_n, \zeta_n) = \nabla F(w_n) - \nabla f(w_n, \zeta_n).$$

Let $\mathcal{F}_n$ be the σ-algebra generated by $\{w_{i+1}, \zeta_i, i = 1, \ldots, n-1\}$. We use $\mathbb{E}_n(\cdot)$ to denote the conditional expectation $\mathbb{E}(\cdot|\mathcal{F}_n)$. Then ξ_n is a martingale sequence since $\mathbb{E}_n\xi_n \equiv 0$.

From (A.1), we find

$$\|w_{n+1} - w^*\|^2 = \|w_n - w^*\|^2 - 2\eta\langle w_n - w^*, \nabla F_\lambda(w_n) - \xi_n\rangle + \eta^2\|\nabla F_\lambda(w_n) - \xi_n\|^2. \tag{A.2}$$

We first try to find a bound of $\langle -(w_n - w^*), \nabla F_\lambda(w_n)\rangle$. We define

$$B_n := \int_0^1 \nabla^2 F_\lambda(sw_n + (1-s)w^*)ds = \lambda I + \int_0^1 \nabla^2 F(sw_n + (1-s)w^*)ds,$$

and apply fundamental theorem of calculus on ∇F_λ. Note that $\nabla F(w^*) = 0$, we obtain

$$\begin{aligned}\nabla F_\lambda(w_n) &= \nabla F_\lambda(w^*) + \int_0^1 \nabla^2 F_\lambda(sw_n + (1-s)w^*)(w_n - w^*)ds \\ &= \lambda w^* + B_n(w_n - w^*).\end{aligned} \tag{A.3}$$

Note that $\frac{1}{2}\lambda I \preceq -\delta A + \lambda I \preceq B_n \preceq A + \lambda I$ and B_n is symmetric, we have

$$\begin{aligned}\langle -(w_n - w^*), \nabla F_\lambda(w_n)\rangle &= -\|w_n - w^*\|_{B_n}^2 - \lambda\langle w^*, w_n - w^*\rangle \\ &\le -\frac{1}{2}\|w_n - w^*\|_{B_n}^2 - \frac{\lambda}{4}\|w_n - w^*\|^2 - \lambda\langle w^*, w_n - w^*\rangle \\ &\le -\frac{1}{2}\|w_n - w^*\|_{B_n}^2 + \lambda\|w^*\|^2.\end{aligned} \tag{A.4}$$

Furthermore, note that by $B_n \preceq A + \lambda I$, we have

$$\|\nabla F_\lambda(w_n)\|^2 = \|\lambda w^* + B_n(w_n - w^*)\|^2 \le 2\lambda^2\|w^*\|^2 + 2(\|A\| + \lambda)\|w_n - w^*\|_{B_n}^2. \tag{A.5}$$

Similarly, we find

$$\nabla F(w_n) = \int_0^1 \nabla^2 F(sw_n + (1-s)w^*)ds(w_n - w^*) = (B_n - \lambda I)(w_n - w^*),$$

thus

$$\|\nabla F(w_n)\|^2 \le \|A\|\|w_n - w^*\|_{B_n}^2. \tag{A.6}$$

Recall that w_n is $\mathcal{F}_n$-measurable and $\mathbb{E}_n\xi_n = 0$. Also

$$\mathbb{E}\|\xi_n\|^2 \le r^2 + c_r|(w_n - w^*)^T\nabla F(w_n)| \le r^2 + c_r\|w_n - w^*\|_{B_n}^2.$$

So plugging (A.4) and (A.5) into (A.2), using (2.3) with (A.6), and by Cauchy Schwarz inequality, we then have

$$\begin{aligned}&\mathbb{E}_n\|w_{n+1} - w^*\|^2 \\ &= \|w_n - w^*\|^2 - 2\eta\mathbb{E}_n\langle w_n - w^*, \nabla F_\lambda(w_n) - \xi_n\rangle + \eta^2\mathbb{E}_n\|\nabla F_\lambda(w_n) - \xi_n\|^2 \\ &= \mathbb{E}_n[\|w_n - w^*\|^2 - 2\eta\langle w_n - w^*, \nabla F_\lambda(w_n)\rangle + \eta^2(\|\nabla F_\lambda(w_n)\|^2 + \|\xi_n\|^2)] \\ &\le \|w_n - w^*\|^2 - \eta\|w_n - w^*\|_{B_n}^2 + 2\lambda\eta(1+\lambda\eta)\|w^*\|^2 \\ &\quad + \eta^2 2(1 + c_r)(\|A\| + \lambda)\|w_n - w^*\|_{B_n}^2 + \eta^2 r^2.\end{aligned}$$

Under the condition

$$\eta \le \min\Big\{\frac{1}{4(1+c_r)(\|A\|+\lambda)}, 1\Big\}, \quad \lambda \le 1,$$

and since $0 \le 1_{\tau \ge n+1} \le 1_{\tau \ge n}$, we have

$$\begin{aligned}
&\mathbb{E}[1_{\tau \ge n+1}\|w_{n+1} - w^*\|^2] \\
&\le \mathbb{E}[1_{\tau \ge n}\|w_{n+1} - w^*\|^2] = \mathbb{E}[1_{\tau \ge n}\mathbb{E}_n\|w_{n+1} - w^*\|^2] \\
&\le \mathbb{E}[1_{\tau \ge n}(\|w_n - w^*\|^2 - \tfrac{1}{2}\eta\|w_n - w^*\|^2_{B_n})] + 4\lambda\eta\|w^*\|^2 + \eta^2 r^2.
\end{aligned}$$

Summing this inequality over all $n = 0, \ldots, (N \wedge \tau) - 1$, we find that

$$\mathbb{E}[\|w_{\tau\wedge N} - w^*\|^2] \le \mathbb{E}\|w_0 - w^*\|^2 + N(4\lambda\eta\|w^*\|^2 + \eta^2 r^2). \tag{A.7}$$

Summing the same inequality over all $n = 0, \ldots, N-1$, we find that

$$\mathbb{E}\left[1_{\tau \ge N-1}\left(\frac{1}{2}\eta\sum_{n=0}^{N-1}\|w_n - w^*\|^2_{B_n}\right)\right] \le \mathbb{E}\|w_0 - w^*\|^2 + N(4\lambda\eta\|w^*\|^2 + \eta^2 r^2). \tag{A.8}$$

To continue, recall $G(w_n) = F(w_n) - F(w^*)$. Apply fundamental theorem of calculus to $F(w)$, we obtain

$$\begin{aligned}
&G(w_n) = F(w_n) - F(w^*) \\
&= \Big[\int_0^1 \nabla F(sw_n + (1-s)w^*)ds\Big]^T (w_n - w^*) \\
&= \Big[\int_0^1 \Big(\nabla F(w^*) + \int_0^s \nabla^2 F(tw_n + (1-t)w^*)(w_n - w^*)dt\Big)ds\Big]^T (w_n - w^*) \\
&= (w_n - w^*)^T\Big[\int_0^1 (1-s)\nabla^2 F(sw_n + (1-s)w^*)ds\Big](w_n - w^*) \\
&= (w_n - w^*)^T A_n (w_n - w^*),
\end{aligned} \tag{A.9}$$

with

$$A_n = \int_0^1 (1-s)\nabla^2 F(sw_n + (1-s)w^*)ds.$$

Under Assumption 3.1, we observe that

$$\begin{aligned}
A_n + \frac{1}{2}\lambda I &= \int_0^1 (1-s)\nabla^2 F_\lambda(sw_n + (1-s)w^*)ds \\
&\preceq \int_0^1 \nabla^2 F_\lambda(sw_n + (1-s)w^*)ds = B_n.
\end{aligned}$$

Namely, we have $H(w_n) = G(w_n) + \frac{\lambda}{2}\|w_n - w^*\|^2 \leq \|w_n - w^*\|^2_{B_n}$. Together with (A.8), we obtain

$$\mathbb{E}\left[1_{\tau \geq N-1}\left(\frac{1}{2}\eta \sum_{n=0}^{N-1} H(w_n)\right)\right] \leq \mathbb{E}\|w_0 - w^*\|^2 + N(4\lambda\eta\|w^*\|^2 + \eta^2 r^2).$$

Then because H is convex within $\mathcal{D}$, we have $H(\bar{w}_N) \leq \frac{1}{N}\sum_{n=0}^{N-1} H(w_n)$ and

$$\mathbb{E}\left[1_{\tau \geq N-1}\left(\eta N H(\bar{w}_N)\right)\right] \leq 2\mathbb{E}\|w_0 - w^*\|^2 + 2N(4\lambda\eta\|w^*\|^2 + \eta^2 r^2).$$

This leads to our claim

$$\mathbb{E}\left[1_{\tau \geq N-1} G(\bar{w}_N)\right] \leq \mathbb{E}\left[1_{\tau \geq N-1} H(\bar{w}_N)\right] \leq \frac{2\mathbb{E}\|w_0 - w^*\|^2}{N\eta} + 8\lambda\|w^*\|^2 + 2\eta r^2.$$

□

Proof of Theorem 2.4. **Step 1:** we build a bound for $\|w_n\|^2$. We rewrite SGD update as

$$w_{n+1} = w_n - \eta \nabla f_\lambda(w_n, \zeta_n) = w_n - \eta \nabla F_\lambda(w_n) + \eta \xi_n, \tag{A.10}$$

where

$$\xi_n = \nabla F_\lambda(w_n) - \nabla f_\lambda(w_n, \zeta_n) = \nabla F(w_n) - \nabla f(w_n, \zeta_n).$$

Let $\mathcal{F}_n$ be the σ-algebra generated by $\{w_{i+1}, \zeta_i, i = 1, \ldots, n-1\}$. We use $\mathbb{E}_n(\cdot)$ to denote the conditional expectation $\mathbb{E}(\cdot|\mathcal{F}_n)$. Then ξ_n is a martingale sequence since $\mathbb{E}_n \xi_n \equiv 0$.

From (A.10), we find

$$\|w_{n+1}\|^2 = \|w_n\|^2 - 2\eta\langle w_n, \nabla F_\lambda(w_n) - \xi_n\rangle + \eta^2\|\nabla F_\lambda(w_n) - \xi_n\|^2. \tag{A.11}$$

To continue, we try to find a bound of $-2\eta\langle w_n, \nabla F_\lambda(w_n)\rangle$. We define

$$B_n := \int_0^1 \nabla^2 F(s w_n + (1-s)w^*) ds,$$

and apply fundamental theorem of calculus on ∇F. Note that $\nabla F(w^*) = 0$, we obtain

$$\nabla F(w_n) = \nabla F(w^*) + \int_0^1 \nabla^2 F(s w_n + (1-s)w^*)(w_n - w^*) ds = B_n(w_n - w^*). \tag{A.12}$$

Note that $-\delta A \preceq B_n \preceq A$. We have

$$\begin{aligned}
&\langle -w_n, \nabla F(w_n)\rangle = -\langle w_n, B_n(w_n - w^*)\rangle \\
&= -\langle w_n, (B_n + \delta A)w_n\rangle + \langle w_n, (B_n + \delta A)w^*\rangle + \delta(\|w_n\|^2_A - \langle w_n, A w^*\rangle)
\end{aligned}$$

$$\leq \frac{1}{4}\|w^*\|^2_{B_n+\delta A} + \delta(2\|w_n\|^2_A + \frac{1}{4}\|w^*\|^2_A)$$
$$\leq 2\delta\|A\|\|w_n\|^2 + \frac{1}{2}\|w^*\|^2_A \text{ since } \delta \leq \frac{1}{2} \text{ and } \|w_n\|^2_A \leq \|A\|\|w_n\|^2.$$

Recall that $\delta\|A\| \leq \frac{\lambda}{4}$, we find

$$\begin{aligned} -2\eta\langle w_n, \nabla F_\lambda(w_n)\rangle &= -2\eta\langle w_n, \nabla F(w_n) + \lambda w_n\rangle \\ &= -2\lambda\eta\|w_n\|^2 + 2\eta\langle -w_n, \nabla F(w_n)\rangle \\ &\leq -\lambda\eta\|w_n\|^2 + \eta\|w^*\|^2_A. \end{aligned} \tag{A.13}$$

If $B_n = Q\Lambda Q^T$ is the eigendecomposition of B_n, let $|B_n| = Q|\Lambda|Q^T$, where $|\Lambda|$ takes absolute value on each element of the diagonal matrix Λ. From the proof of Lemma A.1 claim 1), we know $|B_n| \preceq (1+\delta)A \preceq (1+\delta)\|A\|I$. Thus by $B_n \preceq A$, we have

$$\begin{aligned} \|\nabla F(w_n)\|^2 = \|B_n(w^* - w_n)\|^2 &\leq 2\|B_n w^*\|^2 + 2\|B_n w_n\|^2 \\ &\leq 2(w^*)^T B_n^{1/2}|B_n|B_n^{1/2}w^* + 2\|A\|^2\|w_n\|^2 \\ &\leq 2(1+\delta)\|A\|\|B_n^{1/2}w^*\|^2 + 2\|A\|^2\|w_n\|^2 \\ &\leq 4\|A\|\|w^*\|^2_A + 2\|A\|^2\|w_n\|^2. \end{aligned} \tag{A.14}$$

Recall that w_n is $\mathcal{F}_n$-measurable and $\mathbb{E}_n\xi_n = 0$, we have $\mathbb{E}_n\langle w_n, \xi_n\rangle = 0$. So plugging (A.13) and (A.14) into (A.11), and by Cauchy Schwarz inequality, we then have

$$\begin{aligned} &\mathbb{E}_n\|w_{n+1}\|^2 \\ &= \mathbb{E}_n[\|w_n\|^2 - 2\eta\langle w_n, \nabla F_\lambda(w_n)\rangle + \eta^2\|\nabla F(w_n) + \lambda w_n - \xi_n\|^2] \\ &\leq \mathbb{E}_n[\|w_n\|^2 - 2\eta\langle w_n, \nabla F_\lambda(w_n)\rangle + 3\eta^2(\|\nabla F(w_n)\|^2 + \|\xi_n\|^2 + \lambda^2\|w_n\|^2)] \\ &\leq \|w_n\|^2 - \lambda\eta\|w_n\|^2 + 6\eta^2\|A\|^2\|w_n\|^2 + 3\eta^2\lambda^2\|w_n\|^2 \\ &\quad + \left(\eta + 12\|A\|\eta^2\right)\|w^*\|^2_A + 3\eta^2 r^2(1 + c_r\|w_n\|^2). \end{aligned}$$

Under the condition

$$\eta \leq \frac{\lambda}{12\|A\|^2 + 6\lambda^2 + 6c_r r^2},$$

we have that if $\tau \geq n$

$$\mathbb{E}_n\|w_{n+1}\|^2 \leq (1 - \tfrac{1}{2}\lambda\eta)\|w_n\|^2 + \eta M_w,$$

which can also leads to

$$\begin{aligned} &\mathbb{E}_n\|w_{n+1}\|^2 1_{\tau\geq n+1} \leq \mathbb{E}_n\|w_{n+1}\|^2 1_{\tau\geq n} \leq (1 - \tfrac{1}{2}\lambda\eta)\|w_n\|^2 1_{\tau\geq n} + \eta M_w, \\ &\mathbb{E}_n\|w_{n+1}\|^2 1_{\tau\geq n+1} - \|w_n\|^2 1_{\tau\geq n} \leq \eta M_w, \end{aligned}$$

where

$$M_w := (1 + 12\|A\|\eta)\|w^*\|^2_A + 3\eta r^2 = \|w^*\|^2_A + \eta\left(12\|A\|\|w^*\|^2_A + 3r^2\right).$$

Then iterating the inequalities above gives us

$$\mathbb{E}[\|w_n\|^2 1_{\tau\geq n}] \leq \left(1-\frac{\lambda\eta}{2}\right)^n \mathbb{E}\|w_0\|^2 + \frac{2}{\lambda}(1-(1-\tfrac{1}{2}\lambda\eta)^n)M_w, \tag{A.15}$$

$$\mathbb{E}[\|w_{n\wedge\tau}\|^2] \leq \mathbb{E}\|w_0\|^2 + \eta n M_w. \tag{A.16}$$

Step 2: we derive how does the excess risk evolve. According to Taylor's expansion and (A.10), we know that there exists a v_n, such that

$$\begin{aligned}
F(w_{n+1}) &= F(w_n) - \eta\|\nabla F(w_n)\|^2 - \eta\lambda\nabla F(w_n)^T w_n + \eta\xi_n^T\nabla F(w_n) \\
&\quad + \frac{\eta^2}{2}(\nabla F(w_n) + \lambda w_n - \xi_n)^T\nabla^2 F(v_n)(\nabla F(w_n) + \lambda w_n - \xi_n) \\
&\leq F(w_n) - \eta\|\nabla F(w_n)\|^2 - \eta\lambda\nabla F(w_n)^T w_n + \eta\xi_n^T\nabla F(w_n) \\
&\quad + \frac{\eta^2}{2}(\nabla F(w_n) + \lambda w_n - \xi_n)^T A(\nabla F(w_n) + \lambda w_n - \xi_n) \\
&\leq F(w_n) - \eta\|\nabla F(w_n)\|^2 - \eta\lambda\nabla F(w_n)^T(w_n - w^*) + \eta\lambda|\nabla F(w_n)^T w^*| \\
&\quad + \eta\xi_n^T\nabla F(w_n) + \frac{3\eta^2}{2}\|A\|(\|\nabla F(w_n)\|^2 + \lambda^2\|w_n\|^2 + \|\xi_n\|^2).
\end{aligned} \tag{A.17}$$

Step 3: we bound each term in (A.17) through interpolation. We observe that

$$|\lambda\nabla F(w_n)^T w^*| \leq |\lambda\nabla F(w_n)^T w_\perp^*| + |\lambda\nabla F(w_n)^T w_\lambda^*|, \tag{A.18}$$

where $w^* = w_\perp^* + w_\lambda^*$ is the decomposition introduced in Definition 2.1. Next

$$|\lambda\nabla F(w_n)^T w_\lambda^*| \leq \frac{1}{2}\|\nabla F(w_n)\|^2 + \frac{1}{2}\lambda^2\|w_\lambda^*\|^2. \tag{A.19}$$

Recall $\nabla F(w_n) = B_n(w_n - w^*)$ in (A.12), and by Lemma A.1 claim 2), we further have

$$\begin{aligned}
|\lambda\nabla F(w_n)^T w_\perp^*| &\leq \frac{1}{2}\lambda(w_n - w_*)^T B_n(w_n - w_*) \\
&\quad + \delta\lambda\|w_n - w_*\|_A^2 + \frac{1+2\delta}{2}\lambda w_\perp^{*T} A w_\perp^*.
\end{aligned} \tag{A.20}$$

Plugging (A.19), (A.20) into (A.18), applying the result to (A.17) gives us

$$\begin{aligned}
&F(w_{n+1}) \\
&\leq F(w_n) - \frac{1}{2}\eta\|\nabla F(w_n)\|^2 - \frac{1}{2}\lambda\eta\nabla F(w_n)^T(w_n - w_*) + \frac{1+2\delta}{2}\lambda\eta\|w^*\|_{A,\lambda}^2 \\
&\quad + \delta\lambda\eta\|w_n - w^*\|_A^2 + \eta\xi_n^T\nabla F(w_n) + \frac{3\eta^2\|A\|}{2}(\|\nabla F(w_n)\|^2 + \lambda^2\|w_n\|^2 + \|\xi_n\|^2) \\
&\text{(Recall that } \eta \leq 1,\ \delta < \frac{1}{2} \text{ and } \eta/2 - 3\eta^2\|A\|/2 \geq 0) \\
&\leq F(w_n) - \frac{1}{2}\lambda\eta\nabla F(w_n)^T(w_n - w_*) + \lambda\eta\|w^*\|_{A,\lambda}^2
\end{aligned} \tag{A.21}$$

$$+\delta\lambda\eta\|w_n - w^*\|_A^2 + \eta\xi_n^T\nabla F(w_n) + \frac{3\eta^2}{2}\|A\|(\lambda^2\|w_n\|^2 + \|\xi_n\|^2).$$

To continue, recall $G(w_n) = F(w_n) - F(w^*)$. Apply fundamental theorem of calculus to $F(w)$, we obtain

$$\begin{aligned}
G(w_n) &= F(w_n) - F(w^*) \\
&= \Big[\int_0^1 \nabla F(sw_n + (1-s)w^*)ds\Big]^T (w_n - w^*) \\
&= \Big[\int_0^1 \Big(\nabla F(w^*) + \int_0^s \nabla^2 F(tw_n + (1-t)w^*)(w_n - w^*)dt\Big)ds\Big]^T (w_n - w^*) \\
&= (w_n - w^*)^T\Big[\int_0^1 (1-s)\nabla^2 F(sw_n + (1-s)w^*)ds\Big](w_n - w^*) \\
&= (w_n - w^*)^T A_n(w_n - w^*), \qquad \text{(A.22)}
\end{aligned}$$

with

$$A_n = \int_0^1 (1-s)\nabla^2 F(sw_n + (1-s)w^*)ds.$$

Under Assumption 3.1, namely $0 \preceq \nabla^2 F(w_n) + \delta A \preceq A + \delta A$, we observe that

$$\begin{aligned}
\frac{1}{2}\delta A + A_n &= \frac{1}{2}\delta A + \int_0^1 (1-s)\nabla^2 F(sw_n + (1-s)w^*)ds \\
&= \int_0^1 (1-s)(\nabla^2 F(sw_n + (1-s)w^*) + \delta A)ds \\
&\preceq \int_0^1 (\nabla^2 F(sw_n + (1-s)w^*) + \delta A)ds \\
&= B_n + \delta A \preceq (1+\delta)A.
\end{aligned}$$

Namely, we have

$$A_n \preceq B_n + \frac{1}{2}\delta A \preceq (1 + \tfrac{\delta}{2})A.$$

Thus

$$G(w_n) - \frac{1}{2}\delta\|w_n - w^*\|_A^2 \le (w_n - w^*)^T B_n(w_n - w^*) = \nabla F(w_n)^T(w_n - w_*).$$

Plug this into (A.21), together with $\|w_n - w^*\|_A^2 \le 2\|w_n\|_A^2 + 2\|w^*\|_A^2 \le 2\|A\|\|w_n\|^2 + 2\|w^*\|_A^2$, we have

$$\begin{aligned}
G(w_{n+1}) &\le G(w_n) - \frac{1}{2}\eta\lambda G(w_n) + \eta\xi_n^T\nabla F(w_n) \\
&\quad + \frac{5}{4}\delta\lambda\eta\|w_n - w^*\|_A^2 + \lambda\eta\|w^*\|_{A,\lambda}^2 + \frac{3\eta^2}{2}\|A\|(\lambda^2\|w_n\|^2 + \|\xi_n\|^2). \\
&\le G(w_n) - \frac{1}{2}\eta\lambda G(w_n) + \eta\xi_n^T\nabla F(w_n) + \frac{3\eta^2}{2}\|A\|\|\xi_n\|^2
\end{aligned}$$

$$+\lambda\eta(\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2)+\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)\|A\|\|w_n\|^2.$$

Step 4: summarizing arguments. We will first establish a rough estimate, which is useful to the escape probability. Since $\mathbb{E}_n\xi_n^T\nabla F(w_n)\equiv 0$, we have

$$\begin{aligned}\mathbb{E}_{n\wedge\tau}[G(w_{n\wedge\tau+1})] &\le (1-\frac{1}{2}\eta\lambda)G(w_{n\wedge\tau})+\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)\|A\|\|w_{n\wedge\tau}\|^2\\ &\quad+\lambda\eta\left(\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2\right)+\frac{3\eta^2\|A\|}{2}r^2(1+c_rG(w_{n\wedge\tau})])\\ &\quad\text{Because }\eta\le\frac{\lambda}{6c_r\|A\|r^2}\\ &\le G(w_{n\wedge\tau})+\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)\|A\|\|w_{n\wedge\tau}\|^2\\ &\quad+\lambda\eta\left(\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2\right)+\frac{3\eta^2\|A\|}{2}r^2.\end{aligned}$$

Recall that $\mathbb{E}[\|w_{n\wedge\tau}\|^2]\le\mathbb{E}\|w_0\|^2+\eta nM_w$ with

$$M_w=\|w^*\|_A^2+\eta\left(12\|A\|\|w^*\|_A^2+3r^2\right).$$

Iterating above result gives us

$$\begin{aligned}\mathbb{E}[G(w_{n\wedge\tau})]\le&\mathbb{E}[G(w_0)]+\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)\|A\|(n\mathbb{E}\|w_0\|^2+n^2\eta M_w)\\ &+\lambda n\eta\left(\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2\right)+\frac{3n\eta^2\|A\|}{2}r^2.\end{aligned}\tag{A.23}$$

We can further improve this bound by using $0\le 1_{\tau\ge n+1}\le 1_{\tau\ge n}\le 1$ and taking conditional expectation for both sides. Since $\mathbb{E}_n\xi_n^T\nabla F(w_n)\equiv 0$, we have

$$\begin{aligned}&\mathbb{E}[G(w_{n+1})1_{\tau\ge n+1}]\\ &\le\mathbb{E}[G(w_{n+1})1_{\tau\ge n}]=\mathbb{E}[1_{\tau\ge n}\mathbb{E}_n[G(w_{n+1})]]\\ &\le(1-\frac{1}{2}\eta\lambda)\mathbb{E}[G(w_n)1_{\tau\ge n}]+\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)\|A\|\mathbb{E}[\|w_n\|^2 1_{\tau\ge n}]\\ &\quad+\lambda\eta\left(\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2\right)+\frac{3\eta^2\|A\|}{2}r^2(1+c_r\mathbb{E}[G(w_n)1_{\tau\ge n}])\\ &\quad\text{Because }\eta\le\frac{\lambda}{6c_r\|A\|r^2}\\ &\le(1-\frac{1}{4}\eta\lambda)\mathbb{E}[G(w_n)1_{\tau\ge n}]+\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)\|A\|\mathbb{E}[\|w_n\|^2 1_{\tau\ge n}]\\ &\quad+\lambda\eta\left(\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2\right)+\frac{3\eta^2\|A\|}{2}r^2.\end{aligned}\tag{A.24}$$

Since $\eta\lambda \leq 1$, we have $0 \leq 1 - \frac{1}{4}\lambda\eta \leq \exp(-\frac{1}{4}\lambda\eta)$, then iterating above result gives us

$$\begin{aligned}
&\mathbb{E}[G(w_n)1_{\tau\geq n}] \\
&\leq \exp(-\frac{1}{4}\lambda n\eta)\mathbb{E}[G(w_0)] + 4\|w^*\|_{A,\lambda}^2 + 10\delta\|w^*\|_A^2 + \frac{6\eta\|A\|}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^n)r^2 \\
&\quad + \left(\frac{3\eta^2\lambda^2}{2} + \frac{5}{2}\lambda\eta\delta\right)\|A\|\sum_{i=0}^{n}(1-\frac{1}{4}\eta\lambda)^{n-i}\mathbb{E}[\|w_i\|^2 1_{\tau\geq i}].
\end{aligned}$$

Applying (A.15), together with $\lambda \leq 1, \eta \leq 1$, $\delta \leq \frac{1}{2}$, $12\eta\|A\| \leq 1$ and $1 - \frac{1}{4}\lambda\eta \leq \exp(-\frac{1}{4}\lambda\eta)$, we obtain

$$\begin{aligned}
&\mathbb{E}[G(w_n)1_{\tau\geq n}] \\
&\leq \exp(-\frac{1}{4}\lambda n\eta)\mathbb{E}[G(w_0)] + 4\|w^*\|_{A,\lambda}^2 + 10\delta\|w^*\|_A^2 + \frac{6\eta\|A\|}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^n)r^2 \\
&\quad + \left(\frac{3\eta^2\lambda^2}{2} + \frac{5}{2}\lambda\eta\delta\right)\|A\|\sum_{i=0}^{n}\Big((1-\frac{1}{4}\lambda\eta)^n\mathbb{E}[\|w_0\|^2] \\
&\qquad\qquad + (1-\frac{1}{4}\lambda\eta)^{n-i}\frac{2}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^i)M_w\Big) \\
&\leq \exp(-\frac{1}{4}\lambda n\eta)\mathbb{E}[G(w_0) + 4n\|A\|\|w_0\|^2] \\
&\quad + \frac{6\eta\|A\|}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^n)r^2 + 4\|w^*\|_{A,\lambda}^2 + 10\delta\|w^*\|_A^2 \\
&\quad + \frac{(12\lambda\eta+20\delta)\|A\|}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^n)\left(\|w^*\|_A^2 + \eta\left(12\|A\|\|w^*\|_A^2 + 3r^2\right)\right) \\
&\leq \exp(-\frac{1}{4}\lambda n\eta)\mathbb{E}[G(w_0) + 4n\|A\|\|w_0\|^2] \\
&\quad + \frac{6\eta\|A\|}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^n)r^2 + 4\|w^*\|_{A,\lambda}^2 + 10\delta\|w^*\|_A^2 \\
&\quad + \frac{(12\lambda\eta+20\delta)\|A\|}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^n)\left(2\|w^*\|_A^2 + 3r^2\eta\right) \\
&\leq 4\|w^*\|_{A,\lambda}^2 + \frac{C_1}{\lambda}(1-(1-\frac{1}{4}\eta\lambda)^n)(\eta+\delta) \\
&\quad + \exp(-\frac{1}{4}\lambda n\eta)\mathbb{E}[G(w_0) + 4n\|A\|\|w_0\|^2],
\end{aligned}$$

with $C_1 = 60\|A\|\left(r^2 + \|w^*\|_A^2\right) + 10\|w^*\|_A^2$. □

Proof of Corollary 2.5. By Theorem 2.3, $\mathbb{E}[G(\bar{w}_N)1_{\tau\geq N}] \leq 3\epsilon$ holds if we choose λ, η, N, δ such that the following results hold

$$\frac{2\mathbb{E}\|w_0 - w^*\|^2}{N\eta} \leq \epsilon,\ 8\lambda\|w^*\|^2 \leq \epsilon,\ 2\eta r^2 \leq \epsilon,$$

and the following conditions are satisfied

$$\eta \leq \Big\{\frac{1}{2(1+c_r)(\|A\|+\lambda)}, 1\Big\}, \quad 2\delta\|A\| \leq \lambda \leq 1.$$

Solving $8\lambda\|w^*\|^2 \le 8\lambda C_0 \le \epsilon$ gives us $\lambda(\epsilon) \le \frac{\epsilon}{8C_0}$. The condition on $\delta(\epsilon)$ is obtained from $\lambda \ge 2\delta\|A\|$. The condition of $\eta(\epsilon)$ ensures that $2\eta r^2 \le \epsilon$ and $\eta \le \frac{1}{2(1+c_r)(\|A\|+\lambda)}$. With chosen $\eta(\epsilon)$, the scale of $N(\epsilon)$ is obtained by solving $\frac{2\mathbb{E}\|w_0-w^*\|^2}{N\eta} \le \frac{2C_0}{N\eta} \le \epsilon$. □

Proof of Corollary 2.6. By Theorem 2.4, $\mathbb{E}[G(w_N)1_{\tau\ge N}] \le 4\epsilon$ holds if we choose λ, η, N, δ such that the following results hold

$$4\|w^*\|^2_{A,\lambda} \le \epsilon,\ \frac{C_1\eta}{\lambda} \le \epsilon,\ \frac{C_1\delta}{\lambda} \le \epsilon,\ \exp(-\frac{1}{4}\lambda N\eta)\mathbb{E}[G(w_0) + 4N\|A\|\|w_0\|^2] \le \epsilon.$$

We first choose $\lambda(\epsilon)$ such that $4\|w^*\|^2_{A,\lambda(\epsilon)} < \epsilon$. The conditions on $\eta(\epsilon), \delta(\epsilon)$ ensure that $\frac{C_1\eta}{\lambda} \le \epsilon$ and $\frac{C_1\delta}{\lambda} \le \epsilon$. With chosen $\lambda(\epsilon), \eta(\epsilon)$, the scale of $N(\epsilon)$ is obtained by solving $\exp(-\frac{1}{4}\lambda N\eta)\mathbb{E}[G(w_0)] \le \frac{\epsilon}{2}$ and $\exp(-\frac{1}{4}\lambda N\eta)4N\|A\|\mathbb{E}[\|w_0\|^2] \le \frac{\epsilon}{2}$ by using

$$\exp(-\frac{1}{4}\lambda N\eta)N = \frac{4}{\lambda\eta}\exp(-\frac{1}{4}\lambda N\eta)\frac{1}{4}N\lambda\eta \le \frac{8}{\lambda\eta}\exp(-\frac{1}{8}\lambda N\eta),$$

which is derived from $\exp(-x)x \le 2\exp(-\frac{1}{2}x)$, since by Taylor expansion $x \le 2\exp(\frac{x}{2})$. □

Appendix B: Proof for results in low effective dimension in Section 3.2

Proof of Proposition 3.2. Under Assumption 3.1, applying Corollary 2.5 results in

$$\lambda(\epsilon) = O(\epsilon),\ \delta(\epsilon) = O(\epsilon),\ \eta(\epsilon) = O(\epsilon),\ N(\epsilon) = \Omega\Big(\frac{1}{\epsilon^2}\Big),$$

for guaranteeing $\mathbb{E}[G(\bar{w}_N)1_{\tau\ge N}] \le 3\epsilon$. In this case, by (A.7), according to Chebyshev's inequality and recall that

$$\mathbb{E}\|w_{N\wedge\tau} - w^*\|^2 \le \mathbb{E}\|w_0 - w^*\|^2 + N(4\lambda\eta\|w^*\|^2 + \eta^2 r^2) = O(1),$$

we have

$$\begin{aligned}\mathbf{P}(\tau \le N) &= \mathbf{P}(\{w_{N\wedge\tau} \notin \mathcal{D}\}) = \mathbf{P}(\{\|w_{N\wedge\tau} - w^*\|^2 > a\}) \\ &\le \frac{\mathbb{E}\|w_{N\wedge\tau} - w^*\|^2}{a} \le \delta.\end{aligned}$$

If $\delta = 0$, according to Theorem 2.3,

$$\mathbb{E}\left[1_{\tau\ge N-1}G(\bar{w}_N)\right] \le \frac{2\mathbb{E}\|w_0 - w^*\|^2}{N\eta} + 8\lambda\|w^*\|^2 + 2\eta r^2,$$

taking

$$\lambda(\epsilon) = 0,\ \eta(\epsilon) = O(\epsilon^{1+\alpha}),\ N(\epsilon) = \Omega\Big(\frac{1}{\epsilon^{2+\alpha}}\Big),$$

we obtain $\mathbb{E}[G(\bar{w}_N)1_{\tau\geq N}] \leq 3\epsilon$. In this case, according to Chebyshev's inequality and recall that

$$\mathbb{E}\|w_{N\wedge\tau} - w^*\|^2 \leq \mathbb{E}\|w_0 - w^*\|^2 + N(4\lambda\eta\|w^*\|^2 + \eta^2 r^2) = \mathbb{E}\|w_0 - w^*\|^2 + O(\epsilon^\alpha),$$

we have

$$\begin{aligned}\mathbf{P}(\tau \leq N) &= \mathbf{P}(\{w_{N\wedge\tau} \notin \mathcal{D}\}) = \mathbf{P}(\{\|w_{N\wedge\tau} - w^*\|^2 > a\}) \\ &\leq \frac{\mathbb{E}\|w_0 - w^*\|^2 + O(\epsilon^\alpha)}{a} < 1,\end{aligned}$$

if $a > \mathbb{E}\|w_0 - w^*\|^2$. □

Proof of Proposition 3.4. Recall that (λ_i, v_i), for $i = 1, \ldots, p$, are the eigenvalue-eigenvectors of A with λ_i decreasingly sorted. Therefore, we have

$$\|w^*\|_A^2 = w^{*T}Aw^* = \sum_{i=1}^p \lambda_i \langle w^*, v_i\rangle^2 \leq \|w^*\|_{A,S}^2 \operatorname{tr}(A),$$

$$\|w^*\|_{A,\lambda}^2 = \sum_{i=1}^p \lambda_i \wedge \lambda \langle w^*, v_i\rangle^2 \leq \|w^*\|_{A,S}^2 \sum_{i=1}^p \lambda_i \wedge \lambda.$$

For an exponential spectrum, given any k and p,

$$\sum_{i=k+1}^p \lambda_i = \sum_{i=k+1}^p e^{-ci} = \frac{e^{-(k+1)c}(1 - e^{(k-p)c})}{1 - e^{-c}} \leq \frac{1}{e^{kc}(e^c - 1)}.$$

Thus, to make $\sum_{i=k+1}^p \lambda_i \leq \frac{\epsilon}{8}\|w^*\|_{A,S}^2$, it is sufficient for us to take $k \geq \frac{1}{c}\log\left\{\frac{8\|w^*\|_{A,S}^2}{\epsilon(e^c-1)}\right\}$. And to make $k\lambda = \frac{\epsilon}{8\|w^*\|_{A,S}^2}$, we take $\lambda = \frac{\epsilon}{8k\|w^*\|_{A,S}^2} = \widetilde{O}(\frac{\epsilon}{|\log\epsilon|})$. By these choices, we have

$$\|w^*\|_{A,\lambda}^2 \leq \sum_{i=1}^p \lambda \wedge \lambda_i \|w^*\|_{A,S}^2 \leq \frac{\epsilon}{4}.$$

Next, we find that $\|A\| \leq \operatorname{tr}(A) = \widetilde{O}(1)$, so $C_1 = \widetilde{O}(1), C_2 = \widetilde{O}(1)$. We implement Corollary 2.6 and find

$$\delta(\epsilon) = \widetilde{O}\left(\frac{\epsilon^3}{|\log\epsilon|^2}\right), \quad \eta(\epsilon) = \widetilde{O}\left(\frac{\epsilon^2}{|\log\epsilon|}\right), \quad N(\epsilon) = \widetilde{\Omega}\left(\frac{|\log\epsilon|^3}{\epsilon^3}\right).$$

Recall that

$$\begin{aligned}\mathbb{E}[G(w_{n\wedge\tau})] \leq& \mathbb{E}[G(w_0)] + \left(\frac{3\eta^2\lambda^2}{2} + \frac{5}{2}\lambda\eta\delta\right) n^2\eta C_1 \\ &+ \lambda n\eta\left(\|w^*\|_{A,\lambda}^2 + \frac{5}{2}\delta\|w^*\|_A^2\right) + \frac{3n\eta^2\|A\|}{2} r^2,\end{aligned}$$

together with

$$\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)N^2\eta=\widetilde{O}(\epsilon^2|\log\epsilon|),\ N\eta^2\frac{3\|A\|r^2}{2}=\widetilde{O}(\epsilon|\log\epsilon|),$$
$$\lambda N\eta\left(\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2\right)=\widetilde{O}(\epsilon|\log\epsilon|),$$

we have

$$\mathbb{E}[G(w_{N\wedge\tau})]\le\mathbb{E}[G(w_0)]+\widetilde{O}(\epsilon|\log\epsilon|).$$

For a polynomial spectrum, the derivation is similar. Given any k and p

$$\sum_{i=k+1}^{p}\lambda_i=\sum_{i=k+1}^{p}i^{-(1+c)}$$
$$\le\sum_{i=k+1}^{p}\int_{i-1}^{i}\frac{1}{x^{1+c}}dx=\sum_{i=k+1}^{p}\frac{-1}{c}x^{-c}\Big|_{i-1}^{i}=\frac{1}{c}(k^{-c}-p^{-c})\le\frac{1}{ck^c}.$$

Thus to make $\|w^*\|_{A,S}^2\sum_{i=k+1}^{p}\lambda_i\le\frac{1}{8}\epsilon$, we take $k\ge(\frac{8\|w^*\|_{A,S}^2}{c\epsilon})^{1/c}$. Next, we take $\lambda(\epsilon)=\frac{\epsilon}{8\|w^*\|_{A,S}^2k}=\widetilde{O}\Big(\epsilon^{\frac{c+1}{c}}\Big)$. This leads to $\|w^*\|_{A,\lambda}^2\le\epsilon/4$. Again we find that $C_1=\widetilde{O}(1),C_2=\widetilde{O}(1)$. The order of $\delta(\epsilon),\eta(\epsilon)$ and $N(\epsilon)$ can be derived by Corollary 2.6, that is,

$$\delta(\epsilon)=\widetilde{O}\Big(\epsilon^{\frac{3c+2}{c}}\Big),\ \eta(\epsilon)=\widetilde{O}\Big(\epsilon^{\frac{2c+1}{c}}\Big),\ N(\epsilon)=\widetilde{\Omega}\Big(\frac{|\log(\epsilon)|}{\epsilon^{\frac{3c+2}{c}}}\Big).$$

Recall that

$$\mathbb{E}[G(w_{n\wedge\tau})]\le\mathbb{E}[G(w_0)]+\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)n^2\eta C_1$$
$$+\lambda n\eta\left(\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2\right)+\frac{3n\eta^2\|A\|}{2}r^2,$$

together with

$$\left(\frac{3\eta^2\lambda^2}{2}+\frac{5}{2}\lambda\eta\delta\right)N^2\eta=\widetilde{O}(\epsilon^{\frac{2c+1}{c}}(\log\epsilon)^2),\ \lambda N\eta=\widetilde{O}(|\log\epsilon|),$$
$$\|w^*\|_{A,\lambda}^2+\frac{5}{2}\delta\|w^*\|_A^2=O(\epsilon),\ N\eta^2\frac{3\|A\|r^2}{2}=\widetilde{O}(\epsilon|\log\epsilon|),$$

we have

$$\mathbb{E}[G(w_{N\wedge\tau})]\le\mathbb{E}[G(w_0)]+\widetilde{O}(\epsilon|\log\epsilon|).$$

Given $\mathbb{E}[G(w_{N\wedge\tau})]\le\mathbb{E}[G(w_0)]+\widetilde{O}(\epsilon|\log\epsilon|)$ for both cases, according to Chebyshev's inequality, we have

$$\mathbf{P}(\tau\le N)=\mathbf{P}(\{w_{N\wedge\tau}\notin\mathcal{D}\})=\mathbf{P}(\{G(w_{N\wedge\tau})>(1+a)\mathbb{E}[G(w_0)]\})$$
$$\le\frac{\mathbb{E}[G(w_{N\wedge\tau})]}{(1+a)\mathbb{E}[G(w_0)]}\le\frac{\mathbb{E}[G(w_0)]+\widetilde{O}(\epsilon|\log\epsilon|)}{(1+a)\mathbb{E}[G(w_0)]}\le\frac{1}{1+a}+\widetilde{O}(\epsilon|\log\epsilon|).\qquad\square$$

Appendix C: Proofs of results for overparameterization in statistical models

C.1. Linear regression

Proof of Proposition 4.1. It is straightforward to find the gradient and Hessian of F as:

$$\nabla F(w) = \Sigma(w - w^*), \quad \nabla^2 F(w) = \Sigma. \tag{C.1}$$

This leads to $A = \Sigma, \delta = 0, \mathcal{D} = \mathbb{R}^p$.

Next, note that

$$\nabla f(w, \zeta) = (x^T w - y)x = (x^T(w - w^*) - \xi)x.$$

By Cauchy Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}\|\nabla f(w,\zeta) - \nabla F(w)\|^2 &\le \mathbb{E}[\|\nabla f(w,\zeta)\|^2] \\ &\le 2\mathbb{E}[\|xx^T(w - w^*)\|^2] + 2\mathbb{E}[\|x\xi\|^2] \\ &= 2(w - w^*)^T\mathbb{E}[xx^Txx^T](w - w^*) + 2\sigma^2 \operatorname{tr}(\Sigma). \end{aligned} \tag{C.2}$$

Next we compute $\mathbb{E}[xx^Txx^T]$. Since $x \sim \mathcal{N}(0, \Sigma)$, it can be decomposed as $x = \Sigma^{1/2}z$ with $z \sim \mathcal{N}(0, I_p)$. Let the eigen-decomposition of Σ be $V^T\Lambda V$ and denote $\Sigma^{1/2} = V^T\Lambda^{1/2}V$. We notice that $z' = Vz \sim \mathcal{N}(0, I_p)$, then the (i, j)-th element of $Vzz^TV^T\Lambda Vzz^TV^T$ is $\sum_{k=1}^p \lambda_k z'_i z'_j (z'_k)^2$, and taking expectation results in

$$\mathbb{E}[Vzz^TV^T\Lambda Vzz^TV^T] = \operatorname{diag}\left[2\lambda_1 + \sum_{j=1}^p \lambda_j, \ldots, 2\lambda_j + \sum_{j=1}^p \lambda_j, \ldots, 2\lambda_p + \sum_{j=1}^p \lambda_j\right].$$

Thus we have

$$\begin{aligned}&\mathbb{E}[xx^Txx^T] = V^T\Lambda^{1/2}\mathbb{E}[Vzz^TV^T\Lambda Vzz^TV^T]\Lambda^{1/2}V \\ &= V^T\Lambda^{1/2}\operatorname{diag}\left[2\lambda_1 + \sum_{j=1}^p \lambda_j, \ldots, 2\lambda_j + \sum_{j=1}^p \lambda_j, \ldots, 2\lambda_p + \sum_{j=1}^p \lambda_j\right]\Lambda^{1/2}V \\ &\preceq 3\operatorname{tr}(\Sigma)\Sigma.\end{aligned}$$

Plugging this upper bound in (C.2) gives us

$$\mathbb{E}\|\nabla f(w,\zeta) - \nabla F(w)\|^2 \le 6\operatorname{tr}(\Sigma)\|w - w^*\|_\Sigma^2 + 2\sigma^2\operatorname{tr}(\Sigma).$$

Finally, we note that $\|w - w^*\|_\Sigma^2 = 2G(w)$, and by Young's inequality

$$\begin{aligned}\|w - w^*\|_\Sigma^2 &\le 2\|w\|_\Sigma^2 + 2\|w^*\|_\Sigma^2 \le 2\|\Sigma\|\|w\|^2 + 2\|w^*\|_\Sigma^2, \\ \|w - w^*\|_\Sigma^2 &= (w - w^*)^T\Sigma(w - w^*) = \frac{1}{2}(w - w^*)^T\nabla F(w).\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}&\mathbb{E}\|\nabla f(w,\zeta) - \nabla F(w)\|^2 \\ &\le 2\sigma^2\operatorname{tr}(\Sigma) + 12\operatorname{tr}(\Sigma)\|w^*\|_\Sigma^2 + 12\operatorname{tr}(\Sigma)\min\{G(w), \|\Sigma\|\|w\|^2\}.\end{aligned}$$

□

Remark C.1. *In the proof above, we used the Gaussian distribution assumption only to obtain the first, second and fourth moments of x. This proof can be extended to scenarios where x has a non-Gaussian distribution, as long as an upper bound of $\mathbb{E}[xx^Txx^T]$ is available. Similar extensions can be made for other proofs below as well.*

C.2. Logistic regression

Proof for Proposition 5.1. By Fubini's theorem,

$$\nabla F(w) = \mathbb{E}\nabla f(w,\zeta) = \mathbb{E}\frac{-yx}{1+\exp(yx^Tw)},$$

and

$$\nabla^2 F(w) = \mathbb{E}\nabla\frac{-yx}{1+\exp(yx^Tw)} = \mathbb{E}\frac{y^2\exp(yx^Tw)xx^T}{(1+\exp(yx^Tw))^2}.$$

Because $0 < \frac{y^2\exp(yx^Tw)}{(1+\exp(yx^Tw))^2} < 1$ and $0 \preceq xx^T$, we find $0 \preceq \nabla^2 F(w) \preceq \Sigma$.

Next, we observe

$$\nabla f(w,\zeta) = \frac{-yx}{1+\exp(yx^Tw)}.$$

Then, because $y = \pm 1$, we obtain

$$\mathbb{E}[\|\nabla f(w,\zeta)\|^2] = \mathbb{E}\left[\left(\frac{-y}{1+\exp(yx^Tw)}\right)^2\|x\|^2\right] \le \mathbb{E}[\|x\|^2] = \mathrm{tr}(\Sigma). \qquad \square$$

C.3. M-estimator with Tukey's biweight loss

Proof for Proposition 5.2. First of all, let $v = w - w^*, u = x^Tv - \xi$. We find that

$$\nabla f(w,\zeta) = (1-(u/c)^2)^2 ux1_{|u|\le c}.$$

Then, by Fubini theorem, we have

$$\nabla F(w) = \nabla_v F(w) = \mathbb{E}\nabla[\rho(x^Tv-\xi)] = \mathbb{E}[(1-(u/c)^2)^2 ux1_{\{|u|\le c\}}],$$
$$\nabla^2 F(w) = \mathbb{E}[xx^T(1-(u/c)^2)(1-5(u/c)^2)1_{|u|\le c}].$$

For the first two claims, note that

$$\mathbb{E}\|\nabla f(w,\zeta)\|^2 = \mathbb{E}[(1-(u/c)^2)^4(u/c)^2 1_{|u|\le c}\|x\|^2] \le \mathbb{E}\|x\|^2 = \mathrm{tr}(\Sigma),$$
$$\nabla^2 F(w) = \mathbb{E}[xx^T(1-(u/c)^2)(1-5(u/c)^2)1_{|u|\le c}] \preceq \mathbb{E}xx^T = \Sigma.$$

At $w = w^*$,

$$\nabla^2 F(w^*) = \mathbb{E}[xx^T(1-(\xi/c)^2)(1-5(\xi/c)^2)1_{|\xi|\le c}] = c_0\Sigma.$$

We consider the directional derivative along the v direction

$$\begin{aligned}
&\langle v, \nabla^3 F(w)\rangle := \lim_{\epsilon\to 0}\frac{1}{\epsilon}(\nabla^2 F(w+\epsilon v) - \nabla^2 F(w)) \\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}(\mathbb{E}[xx^T(1-(u/c+\epsilon x^T v)^2)(1-5(u/c+\epsilon x^T v)^2)1_{|u|\le c}] \\
&\qquad - \mathbb{E}[xx^T(1-(u/c)^2)(1-5(u/c)^2)1_{|u|\le c}]) \\
&= \mathbb{E}[4xx^T x^T v(3u/c - 5(u/c)^3)1_{|u|\le c}].
\end{aligned}$$

We find

$$\begin{aligned}
\pm\langle v, \nabla^3 F(w)\rangle &= \pm\mathbb{E}[4xx^T x^T v(3u/c - 5(u/c)^3)1_{|u|\le c}] \\
&\preceq \mathbb{E}[4xx^T|x^T v||5(u/c)^3 - 3u/c|1_{|u|\le c}] \preceq \mathbb{E}[8xx^T|x^T v|].
\end{aligned}$$

For any test vector ψ,

$$\begin{aligned}
|\psi^T\langle v, \nabla^3 F(w)\rangle\psi| &\le 8\mathbb{E}[(x^T\psi)^2|x^T v|] \le 8\sqrt{\mathbb{E}[(x^T\psi)^4]\mathbb{E}[|x^T v|^2]} \\
&= 8\sqrt{3(\psi^T\Sigma\psi)^2(v^T\Sigma v)} \le 16\|v\|_\Sigma \psi^T\Sigma\psi.
\end{aligned}$$

Therefore,

$$-16\|v\|_\Sigma\Sigma \preceq \langle v, \nabla^3 F(w)\rangle \preceq 16\|v\|_\Sigma\Sigma.$$

Furthermore, since $w = w^* + v$, from

$$\nabla^2 F(v+w^*) = \nabla^2 F(w^*) + \int_0^1 \langle v, \nabla^3 F(w^* + sv)\rangle ds,$$

we find

$$\nabla^2 F(w) \succeq -\delta\Sigma,$$

if $16\|v\|_\Sigma \le c_0 + \delta$. □

C.4. Two-layer neural network

First of all, we provide a simple upper bound when computing the fourth order moments of Gaussian random variables.

Lemma C.2. *If $x \in \mathbb{R}^p$ is Gaussian with mean being zero, for any PSD $A \in \mathbb{R}^{p\times p}$ and $a > 0$*

$$\mathbb{E}(x^T Ax + a)^2 \le 3(\mathbb{E}(x^T Ax + a))^2.$$

Proof. Let Σ be the covariance matrix of x. Since replacing x with $\Sigma^{-1/2}x$, the statement of the Lemma remains the same, therefore we can assume $x \sim \mathcal{N}(0, I_p)$. Let $A = V^T\Lambda V$ be the eigenvalue decomposition of A, and the eigenvalues of A be $\lambda_1, \ldots, \lambda_p$. Let $z = Vx \sim \mathcal{N}(0, I_p)$. Note that

$$\mathbb{E}(x^T Ax + a)^2 = \mathbb{E}(\|z\|_\Lambda^4 + 2a\|z\|_\Lambda^2 + a^2),$$

and further,

$$\mathbb{E}\|z\|_\Lambda^4 = \sum_{i,j} \lambda_i\lambda_j \mathbb{E}(z_i^2 z_j^2) \le 3\sum_{i,j} \lambda_i\lambda_j \mathbb{E}z_i^2 \mathbb{E}z_j^2 = 3(\mathbb{E}\|z\|_\Lambda^2)^2.$$

As a consequence, we obtain

$$\mathbb{E}(x^T Ax + a)^2 \le 3(\mathbb{E}\|z\|_\Lambda^2 + a)^2 = 3(\mathbb{E}(x^T Ax + a))^2. \qquad \square$$

Lemma C.3. *Assume that* $\psi(0) = 0$ *and* $|\dot\psi|, |\ddot\psi| \le C$. *Denote*

$$\Sigma^\star = \mathrm{diag}\{I_k, \Sigma, \cdots, \Sigma, I_k\} \in \mathbb{R}^{(p+2)k\times(p+2)k},$$

and $\Delta w = w - w^*$. *Then the followings hold*

1) $\mathbb{E}\|\nabla f(w)\|^2 \le 8\sqrt{3}(1+\mathrm{tr}(\Sigma))(6C^2\|\Delta w\|_{\Sigma^\star}^2(\|w^*\|_{\Sigma^\star}^2+\|w\|_{\Sigma^\star}^2)+\sigma_0^2)C^2\|w\|_{\Sigma^\star}^2$.
2) $\mathbb{E}\nabla g(w,x)\nabla g(w,x)^T \preceq 6C^2\|w\|_{\Sigma^\star}^2\Sigma^\star$.
3) $-M_w \preceq \mathbb{E}(g(w,x) - g(w^*,x) - \xi)\nabla^2 g \preceq M_w$, *where*

$$M_w := 6\sqrt{2}C^2(\|c\|_\infty + 1)\|\Delta w\|_{\Sigma^\star}(\|w^*\|_{\Sigma^\star} + \|w\|_{\Sigma^\star})\Sigma^\star,$$

with $\|c\|_\infty := \max_i\{|c_i|\}$.
4) $G(w) \le 6C^2\|\Delta w\|_{\Sigma^\star}^2(\|w^*\|_{\Sigma^\star}^2 + \|w\|_{\Sigma^\star}^2)$.

Proof. For simplicity of discussion, we denote $z_i = b_i^T x + a_i$ and $z = bx + a$.

Proof for Claim 1): We note that $\nabla f(w) = 2(g(w,x) - g(w^*,x) - \xi)\nabla g(w,x)$, thus

$$\begin{aligned}\mathbb{E}\|\nabla f(w)\|^2 &= 4\mathbb{E}[(g(w,x) - g(w^*,x))^2\|\nabla g(w,x)\|^2] + 4\sigma_0^2\mathbb{E}\|\nabla g(w,x)\|^2 \\ &\le 4\sqrt{\mathbb{E}(g(w,x) - g(w^*,x))^4}\sqrt{\mathbb{E}\|\nabla g(w,x)\|^4} + 4\sigma_0^2\sqrt{\mathbb{E}\|\nabla g(w,x)\|^4}. \qquad \text{(C.3)}\end{aligned}$$

Note that

$$\nabla g = \left[c\circ\dot\psi(z); c_1\dot\psi(z_1)x; \cdots; c_k\dot\psi(z_k)x; \psi(z)\right]^T \in \mathbb{R}^{2k+kp},$$

as a consequence, we have

$$\begin{aligned}\mathbb{E}\|\nabla g(w,x)\|^4 &= \mathbb{E}\left(\|c\circ\dot\psi(z)\|^2 + \sum_{i=1}^k \|c_i\dot\psi(z_i)x\|^2 + \sum_{i=1}^k \|\psi(z_i)\|^2\right)^2 \\ &\le \mathbb{E}\left(C^2\|c\|^2 + \sum_{i=1}^k (c_i)^2C^2\|x\|^2 + 2C^2\sum_{i=1}^k (b_i^T x)^2 + 2C^2\|a\|^2\right)^2\end{aligned}$$

Since x is mean zero Gaussian, by Lemma C.2

$$\le 3\left(C^2\|c\|^2 + \sum_{i=1}^k (c_i)^2C^2\mathbb{E}\|x\|^2 + 2C^2\mathbb{E}\sum_{i=1}^k (b_i^T x)^2 + 2C^2\|a\|^2\right)^2$$

$$\leq 3C^4\left(\|c\|^2+\|c\|^2\operatorname{tr}(\Sigma)+2\sum_{i=1}^k\|b_i\|_\Sigma^2+2\|a\|^2\right)^2$$
$$\leq 12C^4(1+\operatorname{tr}(\Sigma))^2\|w\|_{\Sigma^\star}^4. \tag{C.4}$$

Next, we let $w^s=sw+(1-s)w^*$ and $C_w^2=\|w\|_{\Sigma^\star}^2+\|w^*\|_{\Sigma^\star}^2$. By the convexity of $\|\cdot\|_{\Sigma^\star}^2$, we get

$$\|w^s\|_{\Sigma^\star}^4\leq\max\{\|w\|_{\Sigma^\star}^4,\|w^*\|_{\Sigma^\star}^4\}\leq\|w\|_{\Sigma^\star}^4+\|w^*\|_{\Sigma^\star}^4\leq C_w^4.$$

Then, we have

$$|g(w,x)-g(w^*,x)|^2=\left(\int_0^1\Delta w^T\nabla g(w^s,x)ds\right)^2$$
$$\leq\int_0^1\left(\Delta a^Tc^s\circ\dot\psi(z^s)+\sum_{i=1}^k c_i^s\dot\psi(z_i^s)\Delta b_i^Tx+\Delta c^T\psi(z^s)\right)^2ds$$
$$\leq\int_0^1\left(C\|\Delta a\|\|c^s\|+C\sum_{i=1}^k|c_i^s||\Delta b_i^Tx|+\|\Delta c\|\|\psi(z^s)\|\right)^2ds$$
$$\leq\int_0^1\left(C_w^2\|\Delta a\|^2+C_w^2\sum_{i=1}^k|\Delta b_i^Tx|^2+\|\Delta c\|^2\|\psi(z^s)\|^2/C^2\right)ds$$
$$\cdot\int_0^1\left(\frac{C^2\|c^s\|^2}{C_w^2}+\frac{C^2}{C_w^2}\sum_{i=1}^k|c_i^s|^2+C^2\right)ds$$
$$\leq\int_0^1\left(C_w^2\|\Delta a\|^2+C_w^2\sum_{i=1}^k|\Delta b_i^Tx|^2+\|\Delta c\|^2\|\psi(z^s)\|^2/C^2\right)ds$$
$$\cdot\int_0^1\left(2C^2\|c^s\|^2/C_w^2+C^2\right)ds$$
$$\leq 3C^2\int_0^1\left(C_w^2\|\Delta a\|^2+C_w^2\sum_{i=1}^k|\Delta b_i^Tx|^2+\|\Delta c\|^2\|\psi(z^s)\|^2/C^2\right)ds. \tag{C.5}$$

By Lemma C.2 and $\mathbb{E}|\Delta b_i^Tx|^2=\Delta b_i^T\Sigma\Delta b_i$, we have

$$\mathbb{E}\left(C_w^2\|\Delta a\|^2+C_w^2\sum_{i=1}^k|\Delta b_i^Tx|^2+\|\Delta c\|^2\|\psi(z^s)\|^2/C^2\right)^2$$
$$\leq\mathbb{E}\left(C_w^2\|\Delta a\|^2+C_w^2\sum_{i=1}^k|\Delta b_i^Tx|^2+2\|\Delta c\|^2(\|a^s\|^2+\sum_{i=1}^k|(b_i^s)^Tx|^2)\right)^2$$

Note that $\|a^s\|^2+\sum_{i=1}^k|(b_i^s)^Tx|^2\leq\max\{\|w\|_{\Sigma^\star}^2,\|w^*\|_{\Sigma^\star}^2\}\leq C_w^2$

$$\leq \left(C_w^2 \|\Delta a\|^2 + C_w^2 \sum_{i=1}^{k} \|\Delta b_i\|_\Sigma^2 + 2C_w^2 \|\Delta c\|^2 \right)^2$$
$$\leq 4(\|w\|_{\Sigma^\star}^2 + \|w^*\|_{\Sigma^\star}^2)^2 \|\Delta w\|_{\Sigma^\star}^4.$$

Replace these bounds into the square of (C.5), we find

$$\mathbb{E}|g(w,x) - g(w^*,x)|^4 \leq 36C^4 \|\Delta w\|_{\Sigma^\star}^4 (\|w^*\|_{\Sigma^\star}^2 + \|w\|_{\Sigma^\star}^2)^2. \tag{C.6}$$

Furthermore, we combine this with (C.4) into (C.3), we find that

$$\mathbb{E}\|\nabla f(w)\|^2 \leq 8\sqrt{3}(1+\mathrm{tr}(\Sigma))(6C^2\|\Delta w\|_{\Sigma^\star}^2(\|w^*\|_{\Sigma^\star}^2 + \|w\|_{\Sigma^\star}^2) + \sigma_0^2)C^2\|w\|_{\Sigma^\star}^2.$$

Proof for Claim 2): Recall that

$$\nabla g = \left[c \circ \dot{\psi}(z); c_1 \dot{\psi}(z_1)x; \cdots; c_k \dot{\psi}(z_k)x; \psi(z)\right] \in \mathbb{R}^{2k+kp}.$$

With $u \in \mathbb{R}^k,\ v_1 \in \mathbb{R}^p, \ldots, v_k \in \mathbb{R}^p,\ w \in \mathbb{R}^k$, we define

$$W = \left[u; v_1; \cdots; v_k; w\right] \in \mathbb{R}^{2k+kp},$$

and show that $W^T \mathbb{E}\nabla g \nabla g^T W \preceq 6C^2 \|w\|_{\Sigma^\star}^2 W^T \Sigma^\star W$. Note that

$$\begin{aligned} & W^T \mathbb{E}\nabla g \nabla g^T W \\ &= \mathbb{E}u^T[c \circ \dot{\psi}(z)(c \circ \dot{\psi}(z))^T]u + v_i^T \mathbb{E}[c_i^2 \dot{\psi}(z_i)\dot{\psi}(z_i)xx^T]v_i + w^T \mathbb{E}[\psi(z)(\psi(z))^T]w \\ &\quad + 2u^T \mathbb{E}c \circ \dot{\psi}(z)(\psi(z))^T w + 2\sum_{i=1}^{k} u^T \mathbb{E}c \circ \dot{\psi}(z) c_i \dot{\psi}(z_i) x^T v_i \\ &\quad + 2\sum_{i<j} v_i^T \mathbb{E}c_i \dot{\psi}(z_i) c_j \dot{\psi}(z_j) xx^T v_j + 2\sum_{i=1}^{k} v_i^T \mathbb{E}c_i \dot{\psi}(z_i) x(\psi(z))^T w. \end{aligned} \tag{C.7}$$

For the diagonal terms, note that

$$\mathbb{E}[c \circ \dot{\psi}(z)(c \circ \dot{\psi}(z))^T] \preceq \mathbb{E}[\|c \circ \dot{\psi}(z)\|^2 I_k] \preceq C^2\|c\|^2 I_k,$$
$$\mathbb{E}[c_i^2 \dot{\psi}(z_i)\dot{\psi}(z_i)xx^T] \preceq C^2 c_i^2 \mathbb{E}[xx^T] = C^2 c_i^2 \Sigma,$$
$$\mathbb{E}[\psi(z)(\psi(z))^T] \preceq \mathbb{E}[\|\psi(z)\|^2 I_k] \preceq 2C^2\Big(\|a\|^2 + \sum_{i=1}^{k} \|b_i\|_\Sigma^2\Big) I_k.$$

For the cross terms, note that by Cauchy Schwarz inequality

$$\begin{aligned} & u^T c \circ \dot{\psi}(z) c_i \dot{\psi}(z_i) x^T v_i \\ &\leq |u^T c \circ \dot{\psi}(z)||c_i \dot{\psi}(z_i) x^T v_i| \\ &= (u^T c \circ \dot{\psi}(z)(c \circ \dot{\psi}(z))^T u)^{1/2} (v_i^T (c_i \dot{\psi}(z_i))^2 xx^T v_i)^{1/2} \\ &\leq \frac{c_i^2}{2\|c\|^2}(u^T c \circ \dot{\psi}(z)(c \circ \dot{\psi}(z))^T u) + \frac{\|c\|^2}{2}(v_i^T(\dot{\psi}(z_i))^2 xx^T v_i), \end{aligned}$$

and similarily,

$$u^T c\circ\dot\psi(z)(\psi(z))^T w \le \frac{1}{2}u^T c\circ\dot\psi(z)(c\circ\dot\psi(z))^T u + \frac{1}{2}w^T\psi(z)(\psi(z))^T w,$$
$$v_i^T c_i\dot\psi(z_i)c_j\dot\psi(z_j)xx^T v_j \le \frac{c_j^2}{2}v_i^T(\dot\psi(z_i))^2xx^T v_i + \frac{c_i^2}{2}v_j^T(\dot\psi(z_j))^2xx^T v_j,$$
$$v_i^T c_i\dot\psi(z_i)x(\psi(z))^T w \le \frac{\|c\|^2}{2}v_i^T(\dot\psi(z_i))^2xx^T v_i + \frac{c_i^2}{2\|c\|^2}w^T(\psi(z))(\psi(z))^T w.$$

Plugging the results above into (C.7) gives us

$$\begin{aligned}
&\mathbb{E}\nabla g(w,x)\nabla g(w,x)^T\\
&\preceq C^2\begin{bmatrix} 2\|c\|^2 I_k & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times k}\\ \mathbf{0}_{p\times k} & 3\|c\|^2\Sigma & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times k}\\ \mathbf{0}_{p\times k} & \mathbf{0}_{p\times p} & \ddots & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times k}\\ \mathbf{0}_{p\times k} & \mathbf{0}_{p\times p} & \mathbf{0}_{p\times p} & 3\|c\|^2\Sigma & \mathbf{0}_{p\times k}\\ \mathbf{0}_{k\times k} & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & \mathbf{0}_{k\times p} & 6\left(\|a\|^2+\sum_{i=1}^k\|b_i\|_\Sigma^2\right)I_k \end{bmatrix}\\
&\preceq 6C^2\|w\|_{\Sigma^\star}^2\Sigma^\star.
\end{aligned}$$

Proof for Claim 3): First of all, we find that

$$\nabla^2 g = \begin{bmatrix} D_{c\circ\ddot\psi(z)} & c_1\ddot\psi(z_1)e_1x^T & c_2\ddot\psi(z_2)e_2x^T & \cdots & c_k\dot\psi(z_k)xe_k^T & D_{\dot\psi(z)}\\ c_1\ddot\psi(z_1)xe_1^T & c_1\ddot\psi(z_1)xx^T & \mathbf{0}_{p\times p} & \cdots & \mathbf{0}_{p\times p} & \dot\psi(z_1)xe_1^T\\ c_2\ddot\psi(z_2)xe_2^T & \mathbf{0}_{p\times p} & c_2\ddot\psi(z_2)xx^T & \cdots & \mathbf{0}_{p\times p} & \dot\psi(z_2)xe_2^T\\ \vdots & \vdots & & & \vdots & \vdots\\ c_k\ddot\psi(z_k)xe_k^T & \mathbf{0}_{p\times p} & \cdots & \mathbf{0}_{p\times p} & c_k\ddot\psi(z_k)xx^T & \dot\psi(z_k)xe_k^T\\ D_{\dot\psi(z)} & \dot\psi(z_1)e_1x^T & \dot\psi(z_2)e_2x^T & \cdots & \dot\psi(z_k)e_kx^T & \mathbf{0}_{k\times k} \end{bmatrix}.$$

In above, we use D_v to denote the diagonal matrix with diagonal entries being components of v. We will first show that $\nabla^2 g \preceq Q_x \preceq (2\|c\|_\infty+2)C\Sigma_x^\star$, where

$$\begin{aligned}
Q_x &:= C\,\mathrm{diag}\{(2\|c\|_\infty+1)I_k, (2\|c\|_\infty+1)xx^T,\ldots,(2\|c\|_\infty+1)xx^T, 2I_k\},\\
\Sigma_x^\star &:= \mathrm{diag}\{I_k, xx^T,\ldots,xx^T, I_k\}.
\end{aligned}$$

Recall $W=[u;v_1;\cdots;v_k;w]\in\mathbb{R}^{2k+kp}$. Note that

$$\begin{aligned}
W^T\nabla^2 gW =& u^T D_{c\circ\ddot\psi(z)}u + \sum_{i=1}^k c_i\ddot\psi(z_i)(v_i^Tx)^2 + 2w^T D_{\dot\psi(z)}u\\
&+ 2\sum_{i=1}^k c_i\ddot\psi(z_i)(v_i^Tx)(u^Te_i) + 2\sum_{i=1}^k\dot\psi(z_i)(v_i^Tx)(w^Te_i). \qquad \text{(C.8)}
\end{aligned}$$

And further

$$\begin{aligned} u^T D_{c\circ\ddot{\psi}(z)} u &\le \|D_{c\circ\ddot{\psi}(z)}\|\|u\|^2 \le C\|c\|_\infty \|u\|^2, \\ c_i\ddot{\psi}(z_i)(v_i^T x)^2 &\le C\|c\|_\infty (v_i^T x)^2, \\ 2w^T D_{\dot{\psi}(z)} u &\le C\|w\|^2 + C\|u\|^2, \\ 2c_i\ddot{\psi}(z_i)(v_i^T x)(u^T e_i) &\le \|c\|_\infty C((v_i^T x)^2 + (u^T e_i)^2), \\ 2\dot{\psi}(z_i)(v_i^T x)(w^T e_i) &\le C((v_i^T x)^2 + (w^T e_i)^2). \end{aligned}$$

Replace these upper bounds to terms in (C.8), we find

$$W^T \nabla^2 g W \le W^T Q_x W,$$

because $\sum_i (u^T e_i)^2 = \|u\|^2$. Since this holds for all W, we have $\nabla^2 g \preceq Q_x$. Finally, we note that

$$\begin{aligned} &|W^T \mathbb{E}(g(w,x) - g(w^*,x) - \xi)\nabla^2 g W| \\ &= |W^T \mathbb{E}[(g(w,x) - g(w^*,x))\nabla^2 g] W| \\ &\le \sqrt{\mathbb{E}(g(w,x) - g(w^*,x))^2}\sqrt{\mathbb{E}(W^T\nabla^2 g W)^2}. \end{aligned}$$

Recall (C.6), we have

$$\begin{aligned} \sqrt{\mathbb{E}(g(w,x) - g(w^*,x))^2} &\le (\mathbb{E}(g(w,x) - g(w^*,x))^4)^{1/4} \\ &\le \sqrt{6}C\|\Delta w\|_{\Sigma^\star}(\|w^*\|_{\Sigma^\star} + \|w\|_{\Sigma^\star}). \end{aligned}$$

Then, by Lemma C.2, we have

$$\mathbb{E}(W^T\nabla^2 g W)^2 \le 4(\|c\|_\infty + 1)^2 C^2 \mathbb{E}(W^T \Sigma_x^\star W)^2 \le 12C^2(\|c\|_\infty + 1)^2 (W^T\Sigma^\star W)^2.$$

In combination, we find

$$\begin{aligned} &|W^T \mathbb{E}(g(w,x) - g(w^*,x) - \xi)\nabla^2 g W| \\ &\le 6\sqrt{2}C^2(\|c\|_\infty + 1)\|\Delta w\|_{\Sigma^\star}(\|w^*\|_{\Sigma^\star} + \|w\|_{\Sigma^\star}) W^T\Sigma^\star W. \end{aligned}$$

This verifies our claim 3).

Proof for Claim 4): Simply note that by (C.6), we get

$$G(w) = \mathbb{E}|g(w,x) - g(w^*,x)|^2 \le 6C^2\|\Delta w\|_{\Sigma^\star}^2(\|w^*\|_{\Sigma^\star}^2 + \|w\|_{\Sigma^\star}^2). \qquad \square$$

Proof for Proposition 5.3. First, we find that, when $w \in \mathcal{D}$,

$$\|c\|_\infty^2 \le \|w\|_{\Sigma^\star}^2 \le (1 + \tfrac{1}{4})^2 \|w^*\|_{\Sigma^\star}^2 \le 2\|w^*\|_{\Sigma^\star}^2. \tag{C.9}$$

Note that

$$\nabla^2 F = \mathbb{E}\nabla g(w,x)\nabla g(w,x)^T + \mathbb{E}(g(w,x) - g(w^*,x) - \xi)\nabla^2 g(w,x)$$

$$= \mathbb{E}\nabla g(w,x)\nabla g(w,x)^T + \mathbb{E}(g(w,x) - g(w^*,x))\nabla^2 g(w,x).$$

By Lemma C.3 claim 2) and claim 3) and (C.9), we have

$$\mathbb{E}\nabla g(w,x)\nabla g(w,x)^T \preceq 6C^2\|w^*\|_{\Sigma^\star}^2\Sigma^\star,$$

and

$$\begin{aligned}
&\mathbb{E}(g(w,x) - g(w^*,x))\nabla^2 g(w,x)\\
&\preceq 6\sqrt{2}C^2(\|c\|_\infty + 1)\delta C_1(w^*)\|w^*\|_{\Sigma^\star}(\|w^*\|_{\Sigma^\star} + \|w\|_{\Sigma^\star})\Sigma^\star\\
&\preceq 18\sqrt{2}C^2(2\|w^*\|_{\Sigma^\star} + 1)\delta C_1(w^*)\|w^*\|_{\Sigma^\star}^2\Sigma^\star\\
&\preceq 4\delta C^2\|w^*\|_{\Sigma^\star}^2\Sigma^\star.
\end{aligned}$$

So $\nabla^2 F \preceq C_0(w^*)\Sigma^\star$. Also note that $\mathbb{E}\nabla g(w,x)\nabla g(w,x)^T \succeq 0$, we have

$$\nabla^2 F \succeq \mathbb{E}(g(w,x) - g(w^*,x))\nabla^2 g(w,x) \succeq -4\delta C^2\|w^*\|_{\Sigma^\star}^2\Sigma^\star \succeq -\delta A.$$

Then, by Lemma C.3 claim 1) and (C.9), we find that

$$\begin{aligned}
&\mathbb{E}\|\nabla f(w) - \nabla F(w)\|^2\\
&\leq \mathbb{E}\|\nabla f(w)\|^2\\
&\leq 8\sqrt{3}(1 + \mathrm{tr}(\Sigma))(6C^2\|\Delta w\|_{\Sigma^\star}^2(\|w^*\|_{\Sigma^\star}^2 + \|w\|_{\Sigma^\star}^2) + \sigma_0^2)C^2\|w\|_{\Sigma^\star}^2\\
&\leq 8\sqrt{3}(1 + \mathrm{tr}(\Sigma))(18C^2\|\Delta w\|_{\Sigma^\star}^2\|w^*\|_{\Sigma^\star}^2 + \sigma_0^2)C^2\|w^*\|_{\Sigma^\star}^2\\
&\leq 8\sqrt{3}(1 + \mathrm{tr}(\Sigma))(18\delta^2(C_1(w^*))^2C^2\|w^*\|_{\Sigma^\star}^2\|w^*\|_{\Sigma^\star}^2 + \sigma_0^2)C^2\|w^*\|_{\Sigma^\star}^2\\
&\leq 8\sqrt{3}(1 + \mathrm{tr}(\Sigma))(C^2\|w^*\|_{\Sigma^\star}^4 + \sigma_0^2)C^2\|w^*\|_{\Sigma^\star}^2.
\end{aligned}$$

Finally, by claim 4) of Lemma C.3, when $w_0 \in \mathcal{D}$, we have

$$G(w_0) \leq 18C^2(C_1(w^*))^2\delta^2\|w^*\|_{\Sigma^\star}^4 \leq C^2\|w^*\|_{\Sigma^\star}^4.$$ □

Acknowledgments

The authors thank the suggestions made by the editor and anonymous reviewers.

References

[1] Ali, A., Kolter, Z., and Tibshirani, R. (2019). A continuous-time view of early stopping for least squares regression. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

[2] Allen-Zhu, Z., Li, Y., and Liang, Y. (2019). Learning and generalization in overparameterized neural networks, going beyond two layers. In *Proccedings of the Advance of Neural Information Processing Systems*.

[3] Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *Proceedings of the International Conference on Machine Learning*.
[4] Bach, F. (2022). *Learning Theory from First Principles*. The MIT Press.
[5] Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Proceedings of the Advances in Neural Information Processing Systems*.
[6] Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Proceedings of the Advances in Neural Information Processing Systems*.
[7] Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences 117*, 30063–30070. MR4263288
[8] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 32, 15849–15854. MR3997901
[9] Belkin, M., Hsu, D., and Mitra, P. (2018). Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Proccedings of the Advance of Neural Information Processing Systems*.
[10] Belkin, M., Hsu, D., and Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science* **2**, 1167–1180. MR4186534
[11] Berthier, R., Bach, F., and Gaillard, P. (2020). Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. arXiv:2006.08212.
[12] Block, A., Dagan, Y., and Rakhlin, A. (2021). Majorizing measures, sequential complexities, and online learning. arXiv:2102.01729.
[13] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review* **60**, 2, 223–311. MR3797719
[14] Bousquet, O. and Bottou, L. (2007). The tradeoffs of large scale learning. In *Proceedings of the Advances in Neural Information Processing Systems*.
[15] Cai, T. and Hall, P. (2008). Prediction in function linear regression. *Ann. Statist.* **34**, 2159–2179. MR2291496
[16] Carratino, L., Rudi, A., and Rosasco, L. (2018). Learning with sgd and random features. In *Advances in Neural Information Processing Systems*. 10192–10203.
[17] Chen, X., Du, S. S., and Tong, X. T. (2020). On stationary-point hitting time and ergodicity of stochastic gradient langevin dynamics. *Journal of Machine Learning Research* **21**, 1–41. MR4095347
[18] Dar, Y., Muthukumar, V., and Baraniuk, R. G. (2021). A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. arXiv:2109.02355.

[19] Defossez, A. and Bach, F. (2015). Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Vol. **38**, 205–213.

[20] Dieuleveut, A. and Bach, F. (2016). Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.* **44**, 1363–1399. MR3519927

[21] Dieuleveut, A., Durmus, A., and Bach, F. (2020). Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Ann. Statist.* **48**, 1348–1382. MR4124326

[22] Dieuleveut, A., Flammarion, N., and Bach, F. (2017). Harder, better, faster, stronger convergence rates for least-squares regression. *Journal of Machine Learning Research* **18**, 3520–3570. MR3725440

[23] Dong, J. and Tong, X. T. (2021). Replica exchange for non-convex optimization. *Journal of Machine Learning Research* **22**, 1–59. MR4318529

[24] E, W., Ma, C., and Wang, Q. (2019). A priori estimates of the population risk for residual networks. arXiv:1903.02154. MR4044196

[25] E, W., Ma, C., and Wu, L. (2019). A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics. arXiv:1904.04326. MR4119555

[26] Fan, Y., James, M., and Radchenko, P. (2015). Functional additive regression. *Ann. Statist.* **43**, 2296–2325. MR3396986

[27] Golowich, N., Rakhlin, A., and Shamir, O. (2018). Size-independent sample complexity of neural networks. In *Proceedings of the Annual Conference on Learning Theory*. MR4108976

[28] Guo, Z., Wang, W., Cai, T. T., and Li, H. (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association* **114**, 525, 358–369. MR3941260

[29] Hall, P. and Horowitz, J. (2007). Methodology and convergence rates for funcational linear regeression. *Ann. Statist.* **35**, 70–91. MR2332269

[30] Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on Machine Learning*.

[31] Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. (2022). Suprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.* **50**, 949–986. MR4404925

[32] Jain, P., Kakade, S., Kidambi, R., Netrapalli, P., Pillutla, V., and Sidford, A. (2017). A Markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). arXiv:1710.09430. MR3774046

[33] Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems* **26**, 315–323.

[34] Lakshminarayanan, C. and Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the Twenty-First International Conference on Artificial*

Intelligence and Statistics, Vol. **84**, 1347–1355.

[35] Lan, G. (2020). *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature Switzerland AG. MR4219819

[36] Lei, Y., Hu, T., and Tang, K. (2021). Generalization performance of multi-pass stochastic gradient descent with convex loss functions. *Journal of Machine Learning Research* **22**, 1–41. MR4253718

[37] Lei, Y. and Ying, Y. (2020). Fine-grained analysis of stability and generalization for stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning*. MR4420771

[38] Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. In *32nd Conference on Neural Information Pricessing Systems (NeurIPS 2018)*.

[39] Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Proccedings of the Advance of Neural Information Processing Systems*. MR3796895

[40] Lin, J. and Rosasco, L. (2017). Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research* **18**, 1–47. MR3714260

[41] Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. (2019). Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Conference on Learning Theory*.

[42] Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. arXiv:1908.05355. MR4400901

[43] Mitra, P. (2019). Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. arXiv:1906.03667.

[44] Mohri, M., Rostamizadeh, A., and Talwakar, A. (2018). *Foundations of machine learning*. The MIT Press. MR3931734

[45] Mücke, N., Neu, G., and Rosasco, L. (2019). Beating SGD saturation with tail-averaging and minibatching. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*.

[46] Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. (2019). Harmless interpolation of noisy data in regression. arXiv:1903.09139.

[47] Nakkiran, P. (2019). More data can hurt for linear regression: samplewise double descent. arXiv:1912.07242.

[48] Nakkiran, P., Venkat, P., Kakade, S., and Ma, T. (2020). Optimal regularization can mitigate double descent. arXiv:2003.01897.

[49] Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro., N. (2019). Towards understanding the role of over-parametrization in generalization of neural networks. In *Proceedings of the International Conference on Learning Representations*.

[50] Prateek Jain, Sham M. Kakade, R. K. P. N. and Sidford, A. Accelerating stochastic gradient descent for least squares regression. In *Proceedings of Machine Learning Research, the 31st Conference On Learning*

Theory, Vol. **75**. MR3827111

[51] Prateek Jain, Praneeth Netrapalli, S. M. K. R. K. and Sidford, A. (2017). Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research* **18**, 8258–8299. MR3827111

[52] Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proceedings of the Conference on Learning Theory*.

[53] Rakhlin, A., Sridharan, K., and Tewari, A. (2010). Online learning: Random averages, combinatorial parameters, and learnability. In *Advances in Neural Information Processing Systems 23*.

[54] Rakhlin, A., Sridharan, K., and Tewari, A. (2015a). Online learning via sequential complexities. *Journal of Machine Learning Research* **16**, 155–186. MR3333006

[55] Rakhlin, A., Sridharan, K., and Tewari, A. (2015b). Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields* **161**, 111–153. MR3304748

[56] Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer. MR2168993

[57] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22**, 3, 400–407. MR0042668

[58] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

[59] Tsay, R. (2010). *Analysis of financial time series* (3rd ed.). Wiley. MR2778591

[60] Tsigler, A. and Bartlett, P. L. (2020). Benign overfitting in ridge regression. arXiv:2009.14286. MR4388407

[61] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics* **2**, 448–485. MR0120720

[62] Wang, W., Wu, J., and Yao, Z. (2021). Phase transitions for high-dimensional quadratic discriminant analysis with rare and weak signals. arXiv:2108.10802. MR3931395

[63] Xu, P., Chen, J., Zou, D., and Gu, Q. (2018). Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In *Advances in Neural Information Processing Systems*.

[64] Xue, K. and Yao, F. (2018). Hypothesis testing in large-scale functional linear regression. Accepted by *Statistica Sinica*, SS-2018-0456. MR4286208

[65] Yi Chen, Dong, J. and Tong, X. T. (2022). Can we do better than random start? The power of data outsourcing. arXiv:2205.08098.

[66] Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation* **17**, 9, 2077–2098. MR2175849

[67] Zou, D., Wu, J., Braverman, V., Gu, Q., and Kakade, S. M. (2021). Benign overfitting of constant-stepsize SGD for linear regression. arXiv:2103.12692.