Finite-Sample Analysis of Learning High-Dimensional Single ReLU Neuron

Jingfeng Wu [*1] **Difan Zou** [*2] **Zixiang Chen** [*3] **Vladimir Braverman** [4] **Quanquan Gu** [3] **Sham M. Kakade** [5]

Abstract

This paper considers the problem of learning a single ReLU neuron with squared loss (a.k.a., ReLU regression) in the overparameterized regime, where the input dimension can exceed the number of samples. We analyze a Perceptron-type algorithm called GLM-tron (Kakade et al., 2011) and provide its dimension-free risk upper bounds for high-dimensional ReLU regression in both well-specified and misspecified settings. Our risk bounds recover several existing results as special cases. Moreover, in the well-specified setting, we provide an instance-wise matching risk lower bound for GLM-tron. Our upper and lower risk bounds provide a sharp characterization of the high-dimensional ReLU regression problems that can be learned via GLM-tron. On the other hand, we provide some negative results for stochastic gradient descent (SGD) for ReLU regression with symmetric Bernoulli data: if the model is well-specified, the excess risk of SGD is provably no better than that of GLM-tron ignoring constant factors, for each problem instance; and in the noiseless case, GLM-tron can achieve a small risk while SGD unavoidably suffers from a constant risk in expectation. These results together suggest that GLM-tron might be preferable to SGD for high-dimensional ReLU regression.

1. Introduction

In modern machine learning such as deep learning, the number of model parameters often exceeds the amount of training data, which is often referred to as overparameterization. Yet, overparameterized models (when properly optimized) can still achieve strong generalization performance in practice. Understanding the statistical learning mechanism in the overparameterized regime has drawn great attention in the learning theory community.

Recently, overparameterized linear regression problems have been extensively investigated. Dimensional-free, finite-sample, and instance-wise excess risk bounds have been established for various algorithms, including the minimal ℓ_2-norm interpolator (Bartlett et al., 2020), ridge regression (Tsigler & Bartlett, 2020; Cheng & Montanari, 2022), low-norm interpolator (Zhou et al., 2020; 2021; Koehler et al., 2021) and the online stochastic gradient descent (SGD) methods (Zou et al., 2021b; Wu et al., 2022a). These results together deliver a relatively comprehensive picture of when and how high-dimensional linear regression problems can be learned with finite samples.

However, when the model is not linear, the overparameterized regime is much less well understood, even for the arguably simplest *ReLU regression* problems (see (1)). This work aims to fill this gap by providing sharp risk bounds for learning high-dimensional ReLU regression problems with finite samples.

High-Dimensional ReLU Regression. The problem of ReLU Regression aims to minimize the following risk:

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}\big(\mathtt{ReLU}(\mathbf{x}^\top \mathbf{w}) - y\big)^2, \quad \mathbf{w} \in \mathbb{H}, \tag{1}$$

where $\mathbb{H}$ is a Hilbert space that can be either d-dimensional for a finite d or countably infinite dimensional; $\mathtt{ReLU}(\cdot) := \max\{\cdot, 0\}$ is the *Rectified Linear Unit* (ReLU); $(\mathbf{x}, y) \in \mathbb{H} \otimes \mathbb{R}$ denotes a pair of an input feature vector and the corresponding scalar response; the expectation is taken over some unknown distribution of $(\mathbf{x}, y)$; and $\mathbf{w} \in \mathbb{H}$ denotes the model parameter. It is worth noting that in general $\mathcal{R}(\cdot)$ is non-convex due to the non-linearity of $\mathtt{ReLU}$. Therefore, ReLU regression is significantly harder than linear regression.

Given N i.i.d. samples, $(\mathbf{x}_t, y_t)_{t=1}^N$, two iterative algorithms will be considered for optimizing (1). The first algorithm is *stochastic gradient descent* (SGD), which is initialized from

[*]Equal contribution [1]Department of Computer Science, Johns Hopkins University [2]Department of Computer Science, The University of Hong Kong [3]Department of Computer Science, University of California, Los Angeles [4]Department of Computer Science, Rice University [5]Department of Computer Science and Department of Statistics, Harvard University. Correspondence to: Vladimir Braverman <vb21@rice.edu>, Quanquan Gu <qgu@cs.ucla.edu>, Sham M. Kakade <sham@seas.harvard.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

$\mathbf{w}_0$ and then makes the following update: for $t = 1, \ldots, N$,

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \gamma_t \cdot \mathbf{g}_t, \text{ and} \quad \text{(SGD)}$$
$$\mathbf{g}_t := \big(\mathtt{ReLU}(\mathbf{x}_t^\top \mathbf{w}_{t-1}) - y_t\big)\mathbf{x}_t \cdot \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]$$

where $(\gamma_t)_{t=0}^N$ refers to a stepsize scheduler, e.g., a geometrically decaying stepsize scheduler (Ge et al., 2019; Wu et al., 2022a),

$$\text{for } t \geq 1,\ \gamma_t = \begin{cases} \gamma_{t-1}/2, & t \,\%\, \big(N/\log(N)\big) = 0; \\ \gamma_{t-1}, & \text{otherwise}; \end{cases} \quad (2)$$

and the output is the last iterate, i.e., $\mathbf{w}_N$. The second algorithm is known as *Generalized Linear Model Perceptron* (GLM-tron) (Kalai & Sastry, 2009; Kakade et al., 2011), which is also initialized from $\mathbf{w}_0$ and makes the following update: for $t = 1, \ldots, N$,

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \gamma_t \cdot \big(\mathtt{ReLU}(\mathbf{x}_t^\top \mathbf{w}_{t-1}) - y_t\big)\mathbf{x}_t, \text{ (GLM-tron)}$$

where $(\gamma_t)_{t=0}^N$ is a stepsize scheduler, e.g., (2); and the output is the last iterate, i.e., $\mathbf{w}_N$. Comparing these two algorithms, the only difference is that (GLM-tron) ignores the derivative of $\mathtt{ReLU}(\cdot)$ in its updates.

Contribution 1 (Well-Specified Setting). We first consider the well-specified setting (also known as the "noisy teacher" setting (Frei et al., 2020)), where the expectation of the label conditioned on the input is a linear function followed by ReLU. In this setting, we provide a risk upper bound, $\min \mathcal{R}(\cdot) + \mathcal{O}(D_{\mathtt{eff}}/N)$, for (GLM-tron), where $D_{\mathtt{eff}}$ is an *effective* dimension jointly determined by the sample size, stepsize, and the data covariance matrix, and is independent of the ambient dimension. In particular, $D_{\mathtt{eff}}$ is small when the spectrum of the data covariance matrix decays fast. Moreover, we provide an instance-wise nearly-matching risk lower bound, demonstrating the tightness of our analysis. These bounds are in a similar flavor as the benign-overfitting-type bounds established for high-dimensional linear models (see, e.g., Bartlett et al. (2020); Tsigler & Bartlett (2020); Zou et al. (2021b)), but are the first of their kind for high-dimensional non-linear models.

Contribution 2 (Misspecified Setting). We then consider the misspecified setting (also known as the agnostic setting, see, e.g., Diakonikolas et al. (2020)), where no distributional assumption is made on the label generation. In this case, we provide an $\mathcal{O}(\min \mathcal{R}(\cdot) + D_{\mathtt{eff}}/N)$ risk upper bound for (GLM-tron), where the $D_{\mathtt{eff}}$ is the same effective dimension defined in the well-specified setting. Therefore, we can characterize when (GLM-tron) achieves a constant-factor approximation for misspecified ReLU regression in the overparameterized regime. In particular, when specialized to the finite-dimensional case, our upper bound improves an existing analysis for GLM-tron by Diakonikolas et al. (2020).

Contribution 3 (Comparison with SGD). We also show some negative results on (SGD) for ReLU regression with symmetric Bernoulli data: in the well-specified case, we show that the excess risk achieved by (SGD) is always no better than that achieved by (GLM-tron) ignoring constant factors, for every problem instance; in the noiseless case, (SGD) unavoidably suffers from a constant risk (in expectation) while (GLM-tron) is able to attain an arbitrarily small risk. These together suggest a potentially more preferable algorithmic bias of (GLM-tron) (compared with (SGD)) in ReLU regression.

Contribution 4 (Techniques). From a technical perspective, we introduce new analysis techniques which extend the operator method initially developed for linear models (see, e.g., Jain et al. (2017); Zou et al. (2021b); Wu et al. (2022a) and references therein) to handle the non-linearity of ReLU. The key idea is, instead of controlling the entire covariance matrix of the iterates as in the linear case, one should work with the *diagonal* matrix to better deal with the non-linearity of ReLU. Our novel development of the operator method can be of independent interest.

Paper Organization. The remaining paper is organized as follows. We first review related literature in Section 2. Then we set up the preliminaries in Section 3. We present our main results for well-specified, misspecified ReLU regression, and the comparison between GLM-tron and SGD in Sections 4, 5 and 6, respectively. We sketch our proof techniques in Section 7. Finally, the paper is concluded in Section 8. All proofs are deferred to the appendix.

2. Related Work

ReLU Regression. We first review a set of literature on the hardness results and achievable bounds for ReLU regression. On the negative side, Goel et al. (2020) showed that learning ReLU regression is NP-hard without distributional assumption. Moreover, Goel et al. (2019) showed that even for Gaussian features, learning ReLU regression with small *excess* risk is as hard as the learning sparse parities with noise problem, which is believed to be computationally intractable. On the positive side, Frei et al. (2020) showed that under certain conditions (e.g., bounded and well-spread features), GD or SGD can learn ReLU regression problems with $\min \mathcal{R}(\cdot) + o(1)$ risk in the well-specified cases and $\mathcal{O}(\min \mathcal{R}(\cdot) + o(1))$ risk in the misspecified cases. Compared to Frei et al. (2020), our risk bounds for (GLM-tron) are more general in both settings and can recover their bounds. For finite-dimensional misspecified ReLU regression, Diakonikolas et al. (2022) showed that a constant-factor approximation is possible with only poly-logarithmic samples. However, their result becomes vacuous in the overparameterized regime. Finally, in a significantly easier,

noiseless setting where $y = \texttt{ReLU}(\mathbf{w}_*^\top \mathbf{x})$ for some $\mathbf{w}_* \in \mathbb{H}$, there are far more results (see, e.g., Soltanolkotabi (2017); Du et al. (2017); Yehudai & Shamir (2020); Frei et al. (2020) and the references therein). Although our results can be directly applied, the noiseless setting is not the main focus of our paper.

Tangibly related to ReLU regression, the problem of learning leaky ReLU regression has been studied by Mei et al. (2018); Foster et al. (2018); Frei et al. (2020); Yehudai & Shamir (2020). Since ReLU is not a strictly increasing function (unlikely leaky ReLU), these results for leaky ReLU regression cannot be applied to ReLU regression.

Recent work by Zhou et al. (2022) provided dimension-free bounds on the generalization gap between the *Moreau envelope* of the empirical and population loss for general GLMs including ReLU regression. But their analysis is limited to Gaussian data while our analysis imposes much fewer constraints on the data distribution.

GLM-Tron. The GLM-tron algorithm dates back to at least Kalai & Sastry (2009); Kakade et al. (2011) for learning the well-specified generalized linear model (GLM), where the expectation of the label conditioning on the feature is generated through a GLM. As a special case, their results apply to well-specified ReLU regression as well. However, our results are significantly different from theirs. First of all, in the well-specified regime, we show nearly matching upper and lower excess risk bounds for (GLM-tron), which can recover the excess risk upper bounds from Kalai & Sastry (2009); Kakade et al. (2011). Moreover, from a technical standpoint, their analysis is motivated by the classical analysis for the perceptron algorithm (see, e.g., Section 4.1.7 in Bishop & Nasrabadi (2006)), while we take a completely different approach by analyzing (GLM-tron) in ReLU regression with the operator methods developed for analyzing SGD in linear regression (see, e.g., Zou et al. (2021b); Wu et al. (2022a)). We refer the reader to Section 7 for a detailed overview of our techniques. On the other hand, we remark that our analysis is specialized to ReLU regression and may not directly apply to general GLMs covered by Kalai & Sastry (2009); Kakade et al. (2011).

More recently, Diakonikolas et al. (2020) revisited GLM-tron for learning misspecified ReLU regression and showed a risk upper bound of $\mathcal{O}(\min \mathcal{R}(\cdot) + \sqrt{d/N})$, where d is the ambient dimension and N is the sample size. Their bound becomes vacuous in the overparameterized regime. In comparison, our bound in the misspecified setting can be applied in the overparameterized setting. Moreover, when specialized to the finite-dimensional cases, our bound improves the bound in Diakonikolas et al. (2020).

3. Preliminaries

In this part, we set up some additional preliminaries before presenting our results. The following defines the data covariance matrix.

Definition 3.1 (Data covariance matrix)**.** Assume that each entry and the trace of the $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$ are finite. Define $\mathbf{H} := \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$. Denote the eigenvalues of $\mathbf{H}$ by $(\lambda_i)_{i\geq 1}$, sorted in non-increasing order.

In what follows, we will make the following assumption about the symmetricity of the feature vector.

Assumption 3.2 (Symmetricity conditions)**.** Assume that for every $\mathbf{u} \in \mathbb{H}$ and $\mathbf{v} \in \mathbb{H}$, it holds that

$$\begin{aligned}
&\mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0, \mathbf{x}^\top \mathbf{v} > 0]\big] \\
&\qquad = \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} < 0, \mathbf{x}^\top \mathbf{v} < 0]\big]; \\
&\mathbb{E}\big[(\mathbf{x}^\top \mathbf{v})^2 \mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0, \mathbf{x}^\top \mathbf{v} > 0]\big] \\
&\qquad = \mathbb{E}\big[(\mathbf{x}^\top \mathbf{v})^2 \mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} < 0, \mathbf{x}^\top \mathbf{v} < 0]\big].
\end{aligned}$$

Assumption 3.2 requires that both the second and fourth moments of $\mathbf{x}$, when projected into a sector, are invariant under sign flipping. Clearly, Assumption 3.2 holds when $\mathbf{x}$ follows a symmetric distribution, i.e., $\mathbf{x}$ and $-\mathbf{x}$ satisfy the same distribution, which covers Gaussian or symmetric Bernoulli distributions. We also remark that Assumption 3.2 can be slightly relaxed; see more discussions in Appendix A.

Most existing results for ReLU regression impose some distributional conditions on the feature vectors. For example, Frei et al. (2020); Yehudai & Shamir (2020) assumed that the p.d.f. of $\mathbf{x}$ is "well-spreaded" along every two-dimensional projection. Diakonikolas et al. (2020; 2022) assumed concentration and anti-concentration (and anti-anti-concentration) conditions on $\mathbf{x}$. Our Assumption 3.2 only involves up to the fourth moments of $\mathbf{x}$ and is not directly comparable to theirs that involve the entire p.d.f. of $\mathbf{x}$.

Notation. We reserve upper-case calligraphic letters for linear operators on symmetric matrices. For two positive-value functions $f(x)$ and $g(x)$ we write $f(x) \lesssim g(x)$ or $f(x) \gtrsim g(x)$ if $f(x) \leq cg(x)$ or $f(x) \geq cg(x)$ for some absolute constant $c > 0$ respectively; we write $f(x) \eqsim g(x)$ if $f(x) \lesssim g(x) \lesssim f(x)$. For two vectors $\mathbf{u}$ and $\mathbf{v}$ in a Hilbert space, their inner product is denoted by $\langle \mathbf{u}, \mathbf{v} \rangle$ or equivalently, $\mathbf{u}^\top \mathbf{v}$. For a matrix $\mathbf{A}$, its spectral norm is denoted by $\|\mathbf{A}\|_2$. For two matrices $\mathbf{A}$ and $\mathbf{B}$ of appropriate dimension, their inner product is defined as $\langle \mathbf{A}, \mathbf{B} \rangle := \texttt{tr}(\mathbf{A}^\top \mathbf{B})$. For a positive semi-definite (PSD) matrix $\mathbf{A}$ and a vector $\mathbf{v}$ of appropriate dimension, we write $\|\mathbf{v}\|_{\mathbf{A}}^2 := \mathbf{v}^\top \mathbf{A} \mathbf{v}$. The Kronecker/tensor product is denoted by $\otimes$. Moreover, $\log(\cdot)$ refers to logarithm base 2.

Denote the eigen decomposition of the data covariance by $\mathbf{H} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, where $(\lambda_i)_{i\geq 1}$ are eigenvalues in a non-increasing order and $(\mathbf{v}_i)_{i\geq 1}$ are the corresponding eigenvectors. We denote $\mathbf{H}_{k^*:k^\dagger} := \sum_{k^*<i\leq k^\dagger} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, where $0 \leq k^* \leq k^\dagger$ are two integers, and we allow $k^\dagger = \infty$. For example,

$$\mathbf{H}_{0:k} = \sum_{1\leq i\leq k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \mathbf{H}_{k:\infty} = \sum_{i>k} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top.$$

Similarly, we denote $\mathbf{I}_{k^*:k^\dagger} := \sum_{k^*<i\leq k^\dagger} \mathbf{v}_i \mathbf{v}_i^\top$.

4. Well-Specified ReLU Regression

In this part, we present our results for well-specified ReLU regression. In the literature, the well-specified setting is also extensively referred to as the "noisy teacher" setting (Frei et al., 2020). We formally define a well-specified noise as follows.

Assumption 4.1 (Well-specified noise)**.** Assume that there exists a parameter $\mathbf{w}_* \in \mathbb{H}$ such that

$$\mathbb{E}[y|\mathbf{x}] = \mathtt{ReLU}(\mathbf{x}^\top \mathbf{w}_*).$$

Moreover, denote the variance of the additive noise by

$$\sigma^2 := \mathcal{R}(\mathbf{w}_*) = \mathbb{E}[(y - \mathtt{ReLU}(\mathbf{x}^\top \mathbf{w}_*))^2].$$

Clearly, in the well-specified case, we have

$$\mathcal{R}(\mathbf{w}) = \mathcal{R}(\mathbf{w}_*) + \mathbb{E}[(\mathtt{ReLU}(\mathbf{x}^\top \mathbf{w}) - \mathtt{ReLU}(\mathbf{x}^\top \mathbf{w}_*))^2],$$

which implies that $\mathbf{w}^* \in \arg\min \mathcal{R}(\cdot)$. In this case, we will work with the *excess risk*, defined by

$$\Delta(\mathbf{w}) := \mathcal{R}(\mathbf{w}) - \mathcal{R}(\mathbf{w}_*). \tag{3}$$

Excess Risk Landscape. Our first observation is that the landscape of the excess risk (3) in ReLU regression is closely related to that in linear regression, i.e., a quadratic landscape. The following lemma rigorously characterizes this connection.

Lemma 4.2 (Excess risk landscape)**.** *Under Assumptions 3.2 and 4.1, the following holds for* (3)*:*

$$0.25 \cdot \|\mathbf{w} - \mathbf{w}_*\|_{\mathbf{H}}^2 \leq \Delta(\mathbf{w}) \leq \|\mathbf{w} - \mathbf{w}_*\|_{\mathbf{H}}^2.$$

Even though the excess risk (3) could be non-convex locally, Lemma 4.2 suggests that the landscape of the excess risk in a large scale is "approximately" quadratic in the sense of ignoring some multiplicative factors. This landscape enables us to build sharp upper and lower bounds on the excess risk by bounding a simpler quadratic, $\|\mathbf{w} - \mathbf{w}_*\|_{\mathbf{H}}^2$.

Operators. We follow Zou et al. (2021b); Wu et al. (2022a) and introduce some matrix operators for applying the operator methods to analyze (GLM-tron). Firstly, we denote the covariance of the (GLM-tron) iterates by

$$\mathbf{A}_t := \mathbb{E}(\mathbf{w}_t - \mathbf{w}_*)(\mathbf{w}_t - \mathbf{w}_*)^\top, \quad t \geq 0. \tag{4}$$

We next define a set of linear operators on the matrix space:

$$\begin{aligned} \mathcal{I} := \mathbf{I} \otimes \mathbf{I}, \quad \mathcal{M} := \mathbb{E}[\mathbf{x}^{\otimes 4}], \quad \widetilde{\mathcal{M}} := \mathbf{H} \otimes \mathbf{H}, \\ \mathcal{T}(\gamma) := \mathbf{I} \otimes \mathbf{H} + \mathbf{H} \otimes \mathbf{I} - \gamma \cdot \mathcal{M}, \\ \widetilde{\mathcal{T}}(\gamma) := \mathbf{I} \otimes \mathbf{H} + \mathbf{H} \otimes \mathbf{I} - \gamma \cdot \widetilde{\mathcal{M}}. \end{aligned} \tag{5}$$

A Key Lemma. The next lemma is the key to our analysis, which relates the covariance of a sequence of (GLM-tron) iterates for a ReLU regression problem with the covariance of a sequence of "imaginary" SGD iterates for an "imaginary" linear regression problem.

Lemma 4.3 (Generic bounds on the GLM-tron iterates)**.** *Under Assumptions 3.2 and 4.1, the following holds for* (4)*:*

(A) $\mathbf{A}_{t+1} \preceq \left(\mathcal{I} - \frac{\gamma_t}{2} \cdot \mathcal{T}(2\gamma_t)\right) \circ \mathbf{A}_t + \gamma_t^2 \sigma^2 \cdot \mathbf{H}$;
(B) $\mathbf{A}_{t+1} \succeq \left(\mathcal{I} - \frac{\gamma_t}{2} \cdot \mathcal{T}\left(\frac{\gamma_t}{2}\right)\right) \circ \mathbf{A}_t + \frac{\gamma_t^2 \sigma^2}{4} \cdot \mathbf{H}$.

In the remaining part of this section, we will derive sharp risk bounds for (GLM-tron) in high-dimensional ReLU regression based on Lemma 4.3 and the results for SGD in high-dimensional linear regression developed by Zou et al. (2021b;a); Wu et al. (2022a;b).

4.1. Symmetric Bernoulli Distributions

In order to gain intuitions on the behaviors of (GLM-tron), we start with a simple, symmetric Bernoulli data model defined as follows. Note that this is just the symmetrization of the one-hot data model considered in Zou et al. (2021a).

Assumption 4.4 (Symmetric Bernoulli distribution)**.** Let $(\mathbf{e}_i)_{i\geq 1}$ be a set of orthogonal basis for $\mathbb{H}$. Assume that $\mathbb{P}\{\mathbf{x} = \mathbf{e}_i\} = \mathbb{P}\{\mathbf{x} = -\mathbf{e}_i\} = \lambda_i/2$ for $i \geq 1$, where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$.

Clearly Assumption 4.4 implies Assumption 3.2. We now present our instance-wise sharp excess risk bounds for (GLM-tron) under Assumption 4.4.

Theorem 4.5 (Risk Bounds for GLM-tron)**.** *Suppose that Assumptions 4.1 and 4.4 hold. Let* $\mathbf{w}_N$ *be the output of* (GLM-tron) *with stepsize scheduler* (2)*. Assume that* $N > 100$*. Let* $N_{\mathtt{eff}} := N/\log(N)$*. Suppose that* $\gamma_0 < 1/2$*.*

(A) For every $k^* \geq 0$ *it holds that*

$$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \left\|\mathbf{w}_0 - \mathbf{w}_*\right\|^2_{\prod_{t=1}^N (\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H})\mathbf{H}} + \sigma^2 \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},$$

where $D_{\mathtt{eff}}$ *is defined by*

$$D_{\mathtt{eff}} := k^* + N_{\mathtt{eff}}^2 \gamma_0^2 \cdot \sum_{i>k^*} \lambda_i^2. \tag{6}$$

(B) For $D_{\texttt{eff}}$ defined by (6) *with*

$$k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 N_{\texttt{eff}})\}, \tag{7}$$

it holds that

$$\mathbb{E}\Delta(\mathbf{w}_N) \gtrsim \big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\prod_{t=1}^N (\mathbf{I}-\gamma_t\mathbf{H})\mathbf{H}} + \sigma^2 \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}.$$

Proof Sketch. We first use Lemmas 4.2 and 4.3 to relate (GLM-tron) for ReLU regression problems to SGD for linear regression problems. Then we invoke the one-hot analysis in Zou et al. (2021a) to get the results. □

4.2. Hypercontractive Distributions

We are ready to present our results for the more interesting distributions that satisfy the *hypercontractivity* conditions.

Assumption 4.6 (Hypercontractivity conditions)**.** Assume that the fourth moment of $\mathbf{x}$ is finite and:

(A) There is a constant $\alpha > 0$, such that for every PSD matrix $\mathbf{A}$, we have

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] \preceq \alpha \cdot \texttt{tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H}.$$

Clearly, it must hold that $\alpha \geq 1$.

(B) There is a constant $\beta > 0$, such that for every PSD matrix $\mathbf{A}$, we have

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top\mathbf{A}\mathbf{x}\mathbf{x}^\top] - \mathbf{H}\mathbf{A}\mathbf{H} \succeq \beta \cdot \texttt{tr}(\mathbf{H}\mathbf{A}) \cdot \mathbf{H}.$$

One can verify that Assumption 4.6 holds with $\alpha = 3$ and $\beta = 1$ when $\mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$. Moreover, Assumption 4.6(A) holds when $\mathbf{H}^{-1/2}\mathbf{x}$ is sub-Gaussian or sub-Exponential and Assumption 4.6(B) holds when $\mathbf{H}^{-1/2}\mathbf{x}$ follows a multi-dimensional spherically symmetric distribution (Zou et al., 2021b; Wu et al., 2022a). For more examples of Assumption 4.6 we refer the readers to Zou et al. (2021b); Wu et al. (2022a).

Theorem 4.7 (Risk Bounds for GLM-tron)**.** *Suppose that Assumptions 3.2 and 4.1 hold. Let $\mathbf{w}_N$ be the output of* (GLM-tron) *with stepsize scheduler* (2)*. Assume that $N >$ 100. Let $N_{\texttt{eff}} := N/\log(N)$.*

(A) If in addition Assumption 4.6(A) holds, then for $\gamma_0 < 1/(4\alpha(\texttt{tr}(\mathbf{H})))$ it holds that

$$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \bigg\|\prod_{t=1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*)\bigg\|^2_{\mathbf{H}} + \Big(\alpha\big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\frac{\mathbf{I}_{0:k^*}}{N_{\texttt{eff}}\gamma_0} + \mathbf{H}_{k^*:\infty}} + \sigma^2\Big) \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}},$$

where $D_{\texttt{eff}}$ is defined by (6) *and $k^* \geq 0$ is arbitrary.*

(B) If in addition Assumption 4.6(B) holds, then for $\gamma_0 < 1/\lambda_1$, it holds that

$$\mathbb{E}\Delta(\mathbf{w}_N) \gtrsim \bigg\|\prod_{t=1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*)\bigg\|^2_{\mathbf{H}} + \big(\beta\|\mathbf{w}_0 - \mathbf{w}_*\|^2_{\mathbf{H}_{k^*:\infty}} + \sigma^2\big) \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}},$$

where $D_{\texttt{eff}}$ is defined by (6) *and k^* is defined by* (7)*.*

Proof Sketch. We first use Lemmas 4.2 and 4.3 to relate (GLM-tron) for ReLU regression problems to SGD for linear regression problems. Then we invoke Corollary 3.4 in Wu et al. (2022b) to get the results. □

These bounds in Theorem 4.7 match those of SGD for high-dimensional linear regression shown in Wu et al. (2022b) (and also Zou et al. (2021b); Wu et al. (2022a)) and can be interpreted in a similar manner. Specifically, in the upper bound, the first error term shows that $\mathbf{w}_N$ recovers the true model parameter geometrically at each dimension and is at most

$$\|\mathbf{w}_0 - \mathbf{w}_*\|_2^2/(\gamma_0 N_{\texttt{eff}}),$$

and the second error term is at most

$$\Big(\alpha\big\|\mathbf{w}_0 - \mathbf{w}_*\big\|_2^2 + \sigma^2\Big) \cdot \frac{D_{\texttt{eff}}}{N_{\texttt{eff}}}.$$

Provided a bounded signal-to-noise ratio and a constant initial stepsize (which might not be optimal), the expected risk decreases at a rate of $\mathcal{O}(D_{\texttt{eff}}/N_{\texttt{eff}})$. Moreover, the lower bound justifies the sharpness of the upper bound.

We remark that $D_{\texttt{eff}}$ is independent of the ambient dimension, and is small so long as the spectrum of $\mathbf{H}$ decays fast. This enables (GLM-tron) to achieve a small excess risk even in the overparameterized regime.

The following corollary provides three concrete examples.

Corollary 1. *Under the same conditions as Theorem 4.7, suppose that $\gamma_0 = 1/(4\alpha\,\texttt{tr}(\mathbf{H}))$, and $\|\mathbf{w}_0 - \mathbf{w}^*\|_2$ is finite. Recall the eigenspectrum of $\mathbf{H}$ is $(\lambda_k)_{k\geq 1}$.*

1. *If $\lambda_k = k^{-(1+r)}$ for some constant $r > 0$, then the excess risk is $\mathcal{O}\big(N^{\frac{-r}{1+r}} \cdot \log^{\frac{r}{1+r}}(N)\big)$.*
2. *If $\lambda_k = k^{-1}\log^{-r}(k+1)$ for some constant $r > 1$, then the excess risk is $\mathcal{O}\big(\log^{-r}(N)\big)$.*
3. *If $\lambda_k = 2^{-k}$, then the excess risk is $\mathcal{O}\big(N^{-1}\log^2(N)\big)$.*

Iterate Averaging. Theorem 4.7 focuses on the last iterate of (GLM-tron) with decaying stepsize (2). We remark that this theorem can also be extended to constant stepsize GLM-tron with iterate averaging. See Theorem B.5 in Appendix B.6, where we show matching upto constant factor upper and lower risk bounds for constant stepsize GLM-tron with iterate averaging. It is proved similarly by invoking Lemmas 4.2 and 4.3 and related results from Zou et al. (2021b).

Applications in the Classical Regime. In the next corollary, we apply our instance-dependent risk bounds to the classical regime, i.e., finite dimension and/or bounded ℓ_2-norm.

Corollary 4.8 (Classical regime)**.** *Under the setting of Theorem 4.7, in addition assume that* $\sigma^2 \lesssim 1,\ \|\mathbf{w}_0 - \mathbf{w}_*\|_2 \lesssim 1,\ \lambda_1 \lesssim 1$. *We then have the following:*

(A) If $\mathtt{tr}(\mathbf{H}) \lesssim 1$, *then by choosing* $\gamma_0 \asymp 1/\sqrt{N_{\mathtt{eff}}}$ *and* $k^* := \max\{k : \lambda_k \geq 1/\sqrt{N_{\mathtt{eff}}}\}$, *we have*

$$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \frac{1}{\sqrt{N_{\mathtt{eff}}}} = \sqrt{\frac{\log(N)}{N}}.$$

(B) If d *is finite, then by choosing* $\gamma_0 \asymp 1/\mathtt{tr}(\mathbf{H})$ *and* $k^* = d$, *we have*

$$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \frac{d}{N_{\mathtt{eff}}} = \frac{d\log(N)}{N}.$$

It is worth remarking that the $\log(N)$ *factors in the above rates can be removed when considering constant-stepsize GLM-tron with iterate-averaging (see Theorem B.5).*

In Corollary 4.8, the condition $\|\mathbf{w}_0 - \mathbf{w}_*\|_2 \lesssim 1$ corresponds to the bounded ℓ_2-norm condition of $\mathbf{w}_*$ made in Kakade et al. (2011); Frei et al. (2020) (by taking initialization $\mathbf{w}_0 = 0$). The condition $\mathtt{tr}(\mathbf{H}) \lesssim 1$ corresponds to the bounded ℓ_2-norm condition of features made in Kakade et al. (2011); Frei et al. (2020) (because $\mathbb{E}[\|\mathbf{x}\|_2^2] = \mathtt{tr}(\mathbf{H})$). Then Corollary 4.8(A) matches the $\tilde{\mathcal{O}}(1/\sqrt{N})$ rate for GLM-tron in Kakade et al. (2011), and nearly matches the $\mathcal{O}(1/\sqrt{N})$ rate for GD in Frei et al. (2020). Corollary 4.8(B) shows a faster $\tilde{\mathcal{O}}(d/N)$ rate in the finite-dimensional regime.

5. Misspecified ReLU Regression

In this part, we present our results for misspecified ReLU regression. This setting is also known as the agnostic setting in literature (Goel et al., 2019; Diakonikolas et al., 2020). We first define a misspecified noise as follows.

Assumption 5.1 (Misspecified noise)**.** Denote the minimum population risk by

$$\mathtt{OPT} := \min_{\mathbf{w}' \in \mathbb{H}} \mathcal{R}(\mathbf{w}').$$

Moreover, assume that there exists an optimal model parameter $\mathbf{w}^* \in \arg\min_{\mathbf{w}' \in \mathbb{H}} \mathcal{R}(\mathbf{w}')$ such that

$$\mathbb{E}\big[(y - \mathtt{ReLU}(\mathbf{x}^\top \mathbf{w}_*))^2 \mathbf{x}\mathbf{x}^\top\big] \preceq \sigma^2 \cdot \mathbf{H} \tag{8}$$

holds for some constant $\sigma^2 > 0$.

Different from the well-specified case, Assumption 5.1 does not directly impose any probability condition on the label-generating process. In particular, it captures the situation when $1 - \mathtt{OPT}$ fraction of the label is generated without noise while the rest $\mathtt{OPT}$ fraction of the label is *adversarially* given (Diakonikolas et al., 2020).

Moreover, we empathize that the condition (8) in Assumption 5.1 is very weak and conservative. In particular condition (8) holds trivially when y is bounded, $\|\mathbf{w}_*\|_{\mathbf{H}}$ is finite and $\mathbf{x}$ satisfies the hypercontractivity condition in Assumption 4.6(A), because:

$$\begin{aligned}\text{l.h.s. of (8)} &\preceq 2\mathbb{E}\big[y^2 \mathbf{x}\mathbf{x}^\top\big] + 2\mathbb{E}\big[(\mathbf{x}^\top \mathbf{w}_*)^2 \mathbf{x}\mathbf{x}^\top\big] \\ &\preceq \big(2(\sup\{y\})^2 + 2\alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2\big) \cdot \mathbf{H}.\end{aligned}$$

The above requirements on y, $\mathbf{w}_*$ and $\mathbf{x}$ are already weaker than that required in the literature for learning miss-specified ReLU regression (Frei et al., 2020; Diakonikolas et al., 2020; Goel et al., 2019).

In the misspecified setting, the label can correlate with data in an arbitrary manner. This breaks our nice Lemma 4.3 proved in the well-specified setting. In order to analyze (GLM-tron) in the misspecified setting, we extend the operator methods from considering PSD matrices to considering only the *diagonals* of PSD matrices (see Section 7 for more discussions). With the new techniques, we obtain the following instance-dependent risk bound.

Theorem 5.2 (Risk Bounds for GLM-tron)**.** *Suppose that Assumptions 3.2, 4.6(A) and 5.1 hold. Let* $\mathbf{w}_N$ *be the output of* (GLM-tron) *with stepsize scheduler* (2). *Assume that* $N > 100$. *Let* $N_{\mathtt{eff}} := N/\log(N)$. *Then for* $\gamma_0 < 1/(8\alpha(\mathtt{tr}(\mathbf{H})))$, *it holds that*

$$\begin{aligned}\mathbb{E}\mathcal{R}(\mathbf{w}_N) \lesssim \mathtt{OPT} &+ \Big\| \prod_{t=1}^{N} \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*) \Big\|_{\mathbf{H}}^2 \\ &+ (1 + \mathtt{SNR}) \cdot \sigma^2 \cdot \frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}},\end{aligned}$$

where $D_{\mathtt{eff}}$ *is defined by* (6), $k^* \geq 0$ *is arbitrary, and*

$$\begin{aligned}&\mathtt{SNR} \\ &:= \alpha\Big(\mathtt{OPT} + \|\mathbf{w}_*\|_{\mathbf{H}}^2 + \big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\frac{\mathbf{I}_{0:k^*}}{N_{\mathtt{eff}}\gamma_0} + \mathbf{H}_{k^*:\infty}}\Big)/\sigma^2 \\ &\leq \alpha(\mathtt{OPT} + \|\mathbf{w}_*\|_{\mathbf{H}}^2 + \|\mathbf{w}_0 - \mathbf{w}_*\|_{\mathbf{H}}^2)/\sigma^2.\end{aligned}$$

Similar to the well-specified setting, Theorem 5.2 allows (GLM-tron) to achieve a constant-factor approximation even in the overparameterized regime, as long as the spectrum of $\mathbf{H}$ decays fast such that $D_{\mathtt{eff}}$ is small compared to $N_{\mathtt{eff}}$.

Applications in the Finite-Dimensional Regime. The next corollary shows that, when applied to the finite-dimensional regime, our bound improves an existing bound, $\mathcal{O}(\mathtt{OPT} + \sqrt{d/N})$, of GLM-tron for misspecified ReLU regression proved by Diakonikolas et al. (2020).

Corollary 5.3 (Finite-dimensional regime)**.** *Under the setting of Theorem 5.2, in addition assume that d is finite and*

$$\sigma^2 \lesssim 1,\ \|\mathbf{w}_0 - \mathbf{w}_*\|_2 \lesssim 1,\ \|\mathbf{w}_*\|_2 \lesssim 1,\ \lambda_1 \lesssim 1.$$

Then by choosing $\gamma_0 \asymp 1/\,\mathtt{tr}(\mathbf{H})$ and $k^* = d$, we have

$$\mathbb{E}\mathcal{R}(\mathbf{w}_N) \lesssim \mathtt{OPT} + \frac{d}{N_{\mathtt{eff}}} = \mathtt{OPT} + \frac{d\log(N)}{N}.$$

6. Comparing GLM-tron with SGD

In this part, we show some negative results for (SGD) in ReLU regression with symmetric Bernoulli data.

Well-Specified Case. We first consider well-specified ReLU regression with symmetric Bernoulli data. We provide the following risk lower bound for (SGD).

Theorem 6.1 (Risk lower bound for SGD)**.** *Suppose that Assumptions 4.1 and 4.4 hold. Let $\mathbf{w}_N$ be the output of* (SGD) *with stepsize scheduler* (2)*. Assume that $N > 100$. Let $N_{\mathtt{eff}} := N/\log(N)$. Then for $\gamma_0 < 1$, it holds that*

$$\mathbb{E}\Delta(\mathbf{w}_N) \gtrsim \big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\prod_{t=1}^N(\mathbf{I}-\gamma_t\mathbf{H})\mathbf{H}} + \sigma^2\cdot\frac{D_{\mathtt{eff}}}{N_{\mathtt{eff}}} + \Psi,$$

where $D_{\mathtt{eff}}$ is defined by (6) *with k^* defined by* (7)*, and*

$$\Psi := \Big\langle \sum_{t=0}^{N-1}\gamma_t(1-\gamma_t)\prod_{k=t+1}^{N-1}(1-\gamma_k\mathbf{H})\mathbf{H},\ \mathbf{F}_t\Big\rangle$$

and $\mathbf{F}_t \succeq 0$ is a PSD matrix.

The excess risk lower bound for (SGD) in Theorem 6.1 is in sharp contrast to the excess risk upper bound for (GLM-tron) in Theorem 4.5: the bias and variance error lower bounds for (SGD) is comparable to the bias and variance error upper bounds for (GLM-tron); in addition, there is an extra non-negative error term Ψ for (SGD). This seems to suggest that (SGD) is no better than (GLM-tron). Our next theorem formalizes this observation.

Theorem 6.2 (GLM-tron vs. SGD)**.** *Fix an initialization $\mathbf{w}_0$. Consider a set of well-specified ReLU regression problems with symmetric Bernoulli data (denoted by $\mathscr{E}$) such that: Assumption 4.1 and 4.4 hold and $\|\mathbf{w}_0 - \mathbf{w}_*\|_2^2 \lesssim \sigma^2$. Let $\mathbf{w}_N^{\mathrm{sgd}}(\gamma_0^{\mathrm{sgd}}, \mathcal{P})$ and $\mathbf{w}_N^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}}, \mathcal{P})$ be the outputs of* (SGD) *and* (GLM-tron) *with the same stepsize scheduler* (2)*, initialization $\mathbf{w}_0$, sample size $N > 100$, and on the same problem instance $\mathcal{P} \in \mathscr{E}$, respectively, where $\gamma_0^{\mathrm{sgd}} < 1$ and $\gamma_0^{\mathrm{tron}} < 1/2$ denote their initial stepsizes, respectively. Then for every problem $\mathcal{P} \in \mathscr{E}$, it holds that*

$$\min_{\gamma_0^{\mathrm{tron}}<1/2}\mathbb{E}\Delta\big(\mathbf{w}_N^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}},\mathcal{P})\big) \lesssim \min_{\gamma_0^{\mathrm{sgd}}<1}\mathbb{E}\Delta\big(\mathbf{w}_N^{\mathrm{sgd}}(\gamma_0^{\mathrm{sgd}},\mathcal{P})\big).$$

This theorem shows that for every problem instance in $\mathcal{E}$, the excess risk achieved by (SGD) is no better than that achieved by (GLM-tron) ignoring constant factors.

Noiseless Case. Our final result shows that for the noiseless ReLU regression with symmetric Bernoulli data, (SGD) unavoidably suffers from a constant risk in expectation, while (GLM-tron) can still obtain a small risk.

Theorem 6.3 (Failure of SGD)**.** *Consider a noiseless ReLU regression problem with symmetric Bernoulli data, i.e., Assumptions 4.1 and 4.4 hold with $\sigma^2 = 0$. Let $\mathbb{E}_{\mathbf{w}_*}$ denote the expectation over the randomness of flipping the sign in each component of $\mathbf{w}_*$ uniformly and let $\mathbb{E}_{\mathrm{alg}}$ denote the expectation over the randomness of an algorithm. Let $N > 100$ be the sample size. Then:*

(A) For $\mathbf{w}_N^{\mathrm{tron}}$, the (GLM-tron) *output with stepsize scheduler* (2) *and initial stepsize $\gamma_0 < 1/2$, it holds that*

$$\mathbb{E}_{\mathbf{w}_*}\mathbb{E}_{\mathrm{alg}}\mathcal{R}(\mathbf{w}_N^{\mathrm{tron}}) \lesssim \big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\prod_{t=1}^N\left(\mathbf{I}-\frac{\gamma_t}{2}\mathbf{H}\right)\mathbf{H}}.$$

(B) For $\mathbf{w}_N^{\mathrm{sgd}}$, the (SGD) *output with stepsize scheduler* (2) *and any initial stepsize $\gamma_0 < 1$, it holds that*

$$\mathbb{E}_{\mathbf{w}_*}\mathbb{E}_{\mathrm{alg}}\mathcal{R}(\mathbf{w}_N^{\mathrm{sgd}}) \geq \frac{1}{2}\cdot\|\mathbf{w}_*\|^2_{\mathbf{H}} \geq \frac{1}{2}\cdot\mathcal{R}(0).$$

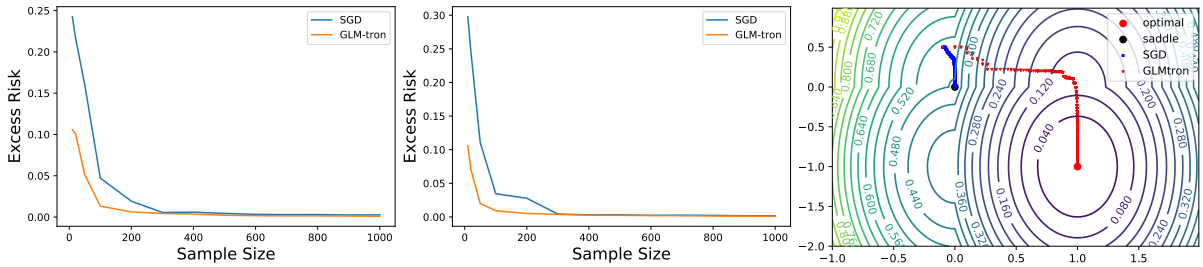
Simulations. Furthermore, we empirically compare the performance of (GLM-tron) and (SGD) for ReLU regression with symmetric Bernoulli data. Simulation results are presented in Figure 1. In the well-specified setting, Figures 1(a) and 1(b) show that the excess risk of (GLM-tron) is no worse than that of (SGD), even when both algorithms are tuned with their hyperparameters (initial stepsizes) respectively. This verifies our Theorem 6.2. In the noiseless setting, Figure 1(c) clearly illustrates that (SGD) can converge to a critical point with constant risk, while (GLM-tron) successfully recovers the true parameters $\mathbf{w}_*$. This verifies our Theorem 6.3.

7. Proof Sketch

We now overview our techniques for analyzing (GLM-tron) iterates in both well-specified and misspecified cases.

For simplicity let us denote the label noise by $\epsilon_t := y_t - \mathtt{ReLU}(\mathbf{w}_*^\top\mathbf{x}_t)$. We first reformulate (GLM-tron) as

$$\begin{aligned}&\mathbf{w}_t - \mathbf{w}_*\\ &= \underbrace{\Big(\mathbf{I} - \gamma_t\mathbb{1}[\mathbf{x}_t^\top\mathbf{w}_{t-1} > 0]\mathbf{x}_t\mathbf{x}_t^\top\Big)(\mathbf{w}_{t-1} - \mathbf{w}_*)}_{\mathbf{c}}\\ &\quad + \underbrace{\gamma_t\big(\mathbb{1}[\mathbf{x}_t^\top\mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top\mathbf{w}_{t-1} > 0]\big)\mathbf{x}_t\mathbf{x}_t^\top\mathbf{w}_*}_{\mathbf{f}}\\ &\quad + \underbrace{\gamma_t\epsilon_t\mathbf{x}_t}_{\mathbf{n}},\end{aligned}$$



(a) $\lambda_i \propto i^{-2}$, well-specified setting (b) $\lambda_i \propto i^{-3}$, well-specified setting (c) 2D illustration, noiseless setting

Figure 1. (a) and (b): Excess risk comparison between (SGD) and (GLM-tron) in well-specified ReLU regression with symmetric Bernoulli data. Here $d = 1,024$, $\sigma^2 = 0.01$ and $\mathbf{w}_* = (i^{-1})_{i=1}^d$. The eigen spectrum is $\lambda_i \propto i^{-2}$ and $\lambda_i \propto i^{-3}$ for (a) and (b), respectively. For each algorithm and each sample size, we do a grid search on the initial stepsize $\gamma_0 \in \{0.5, 0.25, 0.1, 0.075, 0.05, 0.025, 0.01\}$ and report the best excess risk. The plots are averaged over 20 independent runs. (c): Training trajectories of (SGD) and (GLM-tron) on a 2D noiseless ReLU regression with symmetric Bernoulli data. Here $(\lambda_1, \lambda_2) = (0.8, 0.2)$ and $\mathbf{w}_* = (1, -1)$.

where the three parts can be understood as a *contraction* term ($\mathbf{c}$), a *fluctuation* term ($\mathbf{f}$) and a *noise* term ($\mathbf{n}$), respectively. So we have

$$\begin{aligned}\mathbf{A}_t &:= \mathbb{E}(\mathbf{w}_t - \mathbf{w}_*)^{\otimes 2} = \mathbb{E}[(\mathbf{c} + \mathbf{f} + \mathbf{n})^{\otimes 2}] \\ &= \mathbb{E}[\mathbf{c}^{\otimes 2} + \mathbf{f}^{\otimes 2} + \mathbf{n}^{\otimes 2} + \text{cross terms}].\end{aligned}$$

We begin with computing the three quadratic terms. For the contraction term, by Assumption 3.2 we have

$$\begin{aligned}&\mathbb{E}[\mathbf{c}^{\otimes 2}] \\ &= \mathbf{A}_{t-1} - \frac{\gamma_t}{2}(\mathbf{H}\mathbf{A}_{t-1}^\top + \mathbf{A}_{t-1}\mathbf{H}^\top) + \frac{\gamma_t^2}{2}\mathcal{M} \circ \mathbf{A}_{t-1}.\end{aligned}$$

For the fluctuation term, we have

$$\begin{aligned}&\mathbb{E}[\mathbf{f}^{\otimes 2}] \\ &= \gamma_t^2 \cdot \mathbb{E}\big[(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0])^2 \cdot \\ &\qquad (\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t^{\otimes 2}\big] \\ &= 2\gamma_t^2 \cdot \mathbb{E}\big[\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* < 0, \mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \\ &\qquad (\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t^{\otimes 2}\big],\end{aligned}$$

where in the last inequality we use Assumption 3.2. As for the noise term, we simply apply Assumption 4.1 in the well-specified setting or Assumption 5.1 in the misspecified setting to obtain

$$\mathbb{E}[\mathbf{n}^{\otimes 2}] \preceq \gamma_t^2 \sigma^2 \mathbf{H}.$$

In what follows, we utilize the symmetricity condition (Assumption 3.2) to compute the cross terms.

Well-Specified Setting. In the well-specified setting we have that ϵ_t is mean zero conditional on $\mathbf{x}_t$, so all the cross terms involving $\mathbf{n}$ is mean zero, then we have

$$\mathbb{E}[\text{cross terms}] = \mathbb{E}[\mathbf{c}\mathbf{f}^\top + \mathbf{f}\mathbf{c}^\top].$$

Moreover, under Assumption 3.2 it holds that $\mathbb{E}[\mathbf{f}] = 0$, so the part in $\mathbf{c}$ that does not involve $\mathbf{x}_t$ will disappear in the expected crossing terms, i.e.,

$$\begin{aligned}&\mathbb{E}[\text{cross terms}] = \mathbb{E}[\mathbf{c}\mathbf{f}^\top + \mathbf{f}\mathbf{c}^\top] \\ &= -\gamma_t \mathbb{E}\big[\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*)(\mathbf{x}_t \mathbf{f}^\top + \mathbf{f}\mathbf{x}_t^\top)\big] \\ &= 2\gamma_t^2 \mathbb{E}\big[\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \\ &\qquad \mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*) \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t \mathbf{x}_t^\top\big].\end{aligned}$$

Combining the cross term and $\mathbb{E}[\mathbf{f}^{\otimes 2}]$ we obtain

$$\begin{aligned}&\mathbb{E}[\mathbf{f}^{\otimes 2} + \text{cross terms}] = \mathbb{E}[\mathbf{f}^{\otimes 2} + \mathbf{c}\mathbf{f}^\top + \mathbf{f}\mathbf{c}^\top] \\ &= 2\gamma_t^2 \mathbb{E}\big[\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \\ &\qquad \mathbf{x}_t^\top \mathbf{w}_{t-1} \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t \mathbf{x}_t^\top\big] \\ &\preceq 0,\end{aligned}$$

where the last inequality is because the random variable inside the expectation is always non-positive.

Putting everything together, we have shown that

$$\begin{aligned}\mathbf{A}_t &= \mathbb{E}[\mathbf{c}^{\otimes 2} + \mathbf{f}^{\otimes} + \mathbf{n}^{\otimes 2} + \text{cross terms}] \\ &= \mathbb{E}[\mathbf{c}^{\otimes 2} + \mathbf{f}^{\otimes} + \mathbf{n}^{\otimes 2} + \mathbf{c}\mathbf{f}^\top + \mathbf{f}\mathbf{c}^\top] \\ &\preceq \mathbf{A}_{t-1} - \frac{\gamma_t}{2} \cdot (\mathbf{H}\mathbf{A}_{t-1}^\top + \mathbf{A}_{t-1}\mathbf{H}^\top) \\ &\quad + \frac{\gamma_t^2}{2} \cdot \mathcal{M} \circ \mathbf{A}_{t-1} + \gamma_t^2 \sigma^2 \cdot \mathbf{H}.\end{aligned} \tag{9}$$

This matrix recursion has been well-understood thanks to the works by Zou et al. (2021b); Wu et al. (2022a;b).

Misspecified Setting. Now we consider the misspecified setting. Compared to the well-specified setting, the difference is that the part of the cross terms that involve ϵ_t is no longer zero mean, as ϵ_t could correlate with $\mathbf{x}_t$ in an arbitrary manner. The extra work is to understand this part

of the cross terms:

$$\begin{aligned}
&\mathbb{E}[\mathbf{c}\mathbf{n}^\top + \mathbf{n}\mathbf{c}^\top + \mathbf{f}\mathbf{n}^\top + \mathbf{n}\mathbf{f}^\top] \\
&= \underbrace{\gamma_t \mathbb{E}\big[\epsilon_t\big((\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{x}_t^\top + \mathbf{x}_t(\mathbf{w}_{t-1} - \mathbf{w}_*)\big)\big]}_{\text{leading order}} \\
&\quad + \underbrace{2\gamma_t^2 \mathbb{E}[\texttt{IndFunc1} \cdot \epsilon_t \cdot \mathbf{x}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}_*) \cdot \mathbf{x}_t\mathbf{x}_t^\top]}_{\text{higher order 1}} \\
&\quad + \underbrace{2\gamma_t^2 \mathbb{E}[\texttt{IndFunc2} \cdot \epsilon_t \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t\mathbf{x}_t^\top]}_{\text{higher order 2}},
\end{aligned}$$

where IndFunc1 and IndFunc2 are two functions of indicators, both bounded between -1 and 1. For the first higher order term, notice the following by Cauchy inequality:

$$\begin{aligned}
&\texttt{IndFunc1} \cdot \epsilon_t \cdot \mathbf{x}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}_*) \\
&\qquad \le \frac{1}{2}\big(\epsilon_t^2 + (\mathbf{x}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}_*))^2\big),
\end{aligned}$$

so we have

$$\begin{aligned}
&\text{higher order 1} \\
&\preceq \gamma_t^2 \cdot \mathbb{E}[\epsilon_t^2 \mathbf{x}_t\mathbf{x}_t^\top + (\mathbf{x}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}_*))^2 \cdot \mathbf{x}_t\mathbf{x}_t^\top] \\
&\preceq \gamma_t^2\sigma^2\mathbf{H} + \gamma_t^2 \mathcal{M} \circ \mathbf{A}_{t-1},
\end{aligned}$$

where in the last inequality we use Assumption 5.1. We bound the second higher order term in the same manner:

$$\begin{aligned}
\text{higher order 2} &\preceq \gamma_t^2 \cdot \mathbb{E}[\epsilon_t^2 \cdot \mathbf{x}_t\mathbf{x}_t^\top + (\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t\mathbf{x}_t^\top] \\
&\preceq \gamma_t^2\sigma^2 \cdot \mathbf{H} + \alpha\gamma_t^2 \|\mathbf{w}_*\|_{\mathbf{H}}^2 \cdot \mathbf{H},
\end{aligned}$$

where the last inequality is by Assumptions 5.1 and 4.6(A).

The leading order term needs some special treatments. In fact, it is hard to sharply control the leading order term by a PSD matrix. Alternatively, it is possible to sharply bound the *diagonal* of the leading order term by a diagonal matrix (here we assume that $\mathbf{H}$ is diagonal, without loss of generality). The following bound is proved in Lemma C.4 in Appendix C:

$$\texttt{diag}(\text{leading order}) \preceq \frac{\gamma_t}{2} \cdot \mathbf{H}\,\texttt{diag}(\mathbf{A}_{t-1}) + 2\gamma_t \cdot \boldsymbol{\Xi},$$

where $\boldsymbol{\Xi}$ is a fixed diagonal PSD matrix and $\texttt{tr}(\boldsymbol{\Xi}) \le \texttt{OPT}$.

Putting things together with (9), we have

$$\begin{aligned}
&\texttt{diag}(\mathbf{A}_t) \\
&= \texttt{diag}(\mathbb{E}[\mathbf{c}^{\otimes 2} + \mathbf{f}^{\otimes} + \mathbf{n}^{\otimes 2} + \mathbf{c}\mathbf{f}^\top + \mathbf{f}\mathbf{c}^\top]) \\
&\quad + \texttt{diag}(\mathbb{E}[\mathbf{c}\mathbf{n}^\top + \mathbf{n}\mathbf{c}^\top + \mathbf{f}\mathbf{n}^\top + \mathbf{n}\mathbf{f}^\top]) \\
&\preceq \texttt{diag}(\mathbf{A}_{t-1}) - \gamma_t \mathbf{H}\,\texttt{diag}(\mathbf{A}_{t-1}) \\
&+ \frac{\gamma_t^2}{2}\texttt{diag}(\mathcal{M} \circ \mathbf{A}_{t-1}) + \gamma_t^2\sigma^2\mathbf{H} + \gamma_t^2\sigma^2\mathbf{H} \\
&+ \gamma_t^2\texttt{diag}(\mathcal{M} \circ \mathbf{A}_{t-1}) + \gamma_t^2\sigma^2\mathbf{H} + \alpha\gamma_t^2\|\mathbf{w}_*\|_{\mathbf{H}}^2\mathbf{H} \\
&+ \frac{\gamma_t}{2}\mathbf{H}\,\texttt{diag}(\mathbf{A}_{t-1}) + 2\gamma_t\boldsymbol{\Xi} \\
&\preceq \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big)\texttt{diag}(\mathbf{A}_{t-1}) + 2\gamma_t^2\,\texttt{diag}(\mathcal{M} \circ \mathbf{A}_{t-1}) \\
&+ 3\gamma_t^2(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\mathbf{H} + 2\gamma_t\boldsymbol{\Xi}.
\end{aligned}$$

The remaining efforts are to bound the above recursion using techniques developed from Zou et al. (2021b); Wu et al. (2022a;b). It is crucial to remark that $\texttt{tr}(\boldsymbol{\Xi}) \le \texttt{OPT}$, which ensures that the cumulation of the extra "noise term", $2\gamma_t\boldsymbol{\Xi}$, would cause an additive error of at most $\mathcal{O}(\texttt{OPT})$ in the final risk bound.

8. Conclusion

We consider the problem of learning high-dimensional ReLU regression with well-specified or misspecified noise. In the well-specified setting, we provide instance-wise sharp excess risk upper and lower bounds for GLM-tron, that can be applied in the overparameterized regime. In the misspecified setting, we also provide sharp instance-dependent risk upper bound for GLM-tron. In addition, negative results are shown for SGD in well-specified or noiseless ReLU regression with symmetric Bernoulli data, suggesting that GLM-tron might be more effective in ReLU regression.

Acknowledgements

We would like to thank the anonymous reviewers and area chairs for their helpful comments. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. JW and VB are partially supported by the National Science Foundation awards #2244870, #2107239, and #2244899. ZC and QG are partially supported by the National Science Foundation awards IIS-1906169 and IIS-2008981. SK acknowledges funding from the Office of Naval Research under award N00014-22-1-2377 and the National Science Foundation Grant under award #CCF-2212841. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Cheng, C. and Montanari, A. Dimension free ridge regression. *arXiv preprint arXiv:2210.08571*, 2022.

Diakonikolas, I., Goel, S., Karmalkar, S., Klivans, A. R., and Soltanolkotabi, M. Approximation schemes for relu regression. In *Conference on Learning Theory*, pp. 1452–1485. PMLR, 2020.

Diakonikolas, I., Kontonis, V., Tzamos, C., and Zarifis, N. Learning a single neuron with adversarial label noise via gradient descent. In *Conference on Learning Theory*, pp. 4313–4361. PMLR, 2022.

Du, S. S., Lee, J. D., and Tian, Y. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017.

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Frei, S., Cao, Y., and Gu, Q. Agnostic learning of a single neuron with gradient descent. *Advances in Neural Information Processing Systems*, 33:5417–5428, 2020.

Ge, R., Kakade, S. M., Kidambi, R., and Netrapalli, P. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *arXiv preprint arXiv:1904.12838*, 2019.

Goel, S., Karmalkar, S., and Klivans, A. Time/accuracy tradeoffs for learning a relu with respect to gaussian marginals. *Advances in Neural Information Processing Systems*, 32, 2019.

Goel, S., Klivans, A. R., Manurangsi, P., and Reichman, D. Tight hardness results for training depth-2 relu networks. In *Information Technology Convergence and Services*, 2020.

Jain, P., Netrapalli, P., Kakade, S. M., Kidambi, R., and Sidford, A. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(1):8258–8299, 2017.

Kakade, S. M., Kanade, V., Shamir, O., and Kalai, A. Efficient learning of generalized linear and single index models with isotonic regression. *Advances in Neural Information Processing Systems*, 24, 2011.

Kalai, A. T. and Sastry, R. The isotron algorithm: High-dimensional isotonic regression. In *COLT*, 2009.

Koehler, F., Zhou, L., Sutherland, D. J., and Srebro, N. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Soltanolkotabi, M. Learning relus via gradient descent. *Advances in neural information processing systems*, 30, 2017.

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

Wu, J., Zou, D., Braverman, V., Gu, Q., and Kakade, S. M. Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression. *The 39th International Conference on Machine Learning*, 2022a.

Wu, J., Zou, D., Braverman, V., Gu, Q., and Kakade, S. M. The power and limitation of pretraining-finetuning for linear regression under covariate shift. *The 36th Conference on Neural Information Processing Systems*, 2022b.

Yehudai, G. and Shamir, O. Learning a single neuron with gradient methods. In *Conference on Learning Theory*, pp. 3756–3786. PMLR, 2020.

Zhou, L., Sutherland, D. J., and Srebro, N. On uniform convergence and low-norm interpolation learning. *Advances in Neural Information Processing Systems*, 33: 6867–6877, 2020.

Zhou, L., Koehler, F., Sutherland, D. J., and Srebro, N. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *arXiv preprint arXiv:2112.04470*, 2021.

Zhou, L., Koehler, F., Sur, P., Sutherland, D. J., and Srebro, N. A non-asymptotic moreau envelope theory for high-dimensional generalized linear models. *arXiv preprint arXiv:2210.12082*, 2022.

Zou, D., Wu, J., Braverman, V., Gu, Q., Foster, D. P., and Kakade, S. The benefits of implicit regularization from sgd in least squares problems. *Advances in Neural Information Processing Systems*, 34:5456–5468, 2021a.

Zou, D., Wu, J., Braverman, V., Gu, Q., and Kakade, S. Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory*, pp. 4633–4635. PMLR, 2021b.

A. Weaker Symmetricity Assumptions

In fact, Assumption 3.2 can be relaxed into some moment symmetricity conditions:

Assumption A.1 (Moment symmetricity conditions)**.** Assume that

(A) For every $\mathbf{u} \in \mathbb{H}$, it holds that

$$\mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0]\big] = \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} < 0]\big].$$

(B) For every $\mathbf{u} \in \mathbb{H}$ and $\mathbf{v} \in \mathbb{H}$, it holds that

$$\mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0, \mathbf{x}^\top \mathbf{v} > 0]\big] = \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} < 0, \mathbf{x}^\top \mathbf{v} < 0]\big].$$

(C) For every $\mathbf{u} \in \mathbb{H}$, it holds that

$$\mathbb{E}\big[\mathbf{x}^{\otimes 4} \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0]\big] = \mathbb{E}\big[\mathbf{x}^{\otimes 4} \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} < 0]\big].$$

(D) For every $\mathbf{u} \in \mathbb{H}$ and $\mathbf{v} \in \mathbb{H}$, it holds that

$$\mathbb{E}\big[(\mathbf{x}^\top \mathbf{v})^2 \mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0, \mathbf{x}^\top \mathbf{v} > 0]\big] = \mathbb{E}\big[(\mathbf{x}^\top \mathbf{v})^2 \mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} < 0, \mathbf{x}^\top \mathbf{v} < 0]\big].$$

Clearly all the conditions in Assumption A.1 holds when Assumption 3.2 is true. Assumption A.1(A) is crucial to our analysis. Assumption A.1(B) is only useful for deriving lower bounds. Note that Assumption A.1(B) implies Assumption A.1(A). Assumption A.1(C) is only useful for deriving lower bounds, too. Assumption A.1(D) is only made for technical simplicity; without using Assumption A.1(D) one can still derive an upper bound for GLM-tron, the only difference will be replacing σ^2 in the current upper bound with $\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2$.

Some Moments Results. The following moments results are direct consequences of Assumption A.1.

Lemma A.2. *The following holds:*

(A) Under Assumption A.1 (A), it holds that: for every vector $\mathbf{u} \in \mathbb{H}$*,*

$$\mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0]\big] = \frac{1}{2} \cdot \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big] =: \frac{1}{2} \cdot \mathbf{H}.$$

(B) Under Assumption A.1 (C), it holds that: for every vector $\mathbf{u} \in \mathbb{H}$*,*

$$\mathbb{E}\big[\mathbf{x}^{\otimes 4} \cdot \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0]\big] = \frac{1}{2} \cdot \mathbb{E}\big[\mathbf{x}^{\otimes 4}\big] =: \frac{1}{2} \cdot \mathcal{M}.$$

Proof of Lemma A.2. By Assumption A.1(A), we have

$$\mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0]\big] = \mathbb{E}\big[(-\mathbf{x})(-\mathbf{x})^\top \mathbb{1}[(-\mathbf{x})^\top \mathbf{u} > 0]\big] = \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \mathbb{1}[\mathbf{x}^\top \mathbf{u} < 0]\big].$$

Moreover, notice that

$$\mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0]\big] + \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \mathbb{1}[\mathbf{x}^\top \mathbf{u} < 0]\big] = \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big].$$

The above two equations together imply that

$$\mathbb{E}\big[\mathbf{x}\mathbf{x}^\top \mathbb{1}[\mathbf{x}^\top \mathbf{u} > 0]\big] = \frac{1}{2}\mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big].$$

Similarly, we can prove the second equality in the lemma. □

B. Well-Specified Setting

In this section, we focus on the well-specified setting and always assume Assumption 4.1 holds.

B.1. Proof of Lemma 4.2

We will prove a slightly stronger lemma.

Lemma B.1 (Loss landscape, restated Lemma 4.2)**.** *Suppose that Assumption 4.1 holds. Consider* (3)*, we have:*

(A) $\Delta(\mathbf{w}) \le \|\mathbf{w}-\mathbf{w}_*\|_{\mathbf{H}}^2$;

(B) if in addition Assumption A.1(B) holds, then $\Delta(\mathbf{w}) \ge \frac{1}{4}\cdot\|\mathbf{w}-\mathbf{w}_*\|_{\mathbf{H}}^2$.

Proof. Under Assumption 4.1, it holds that

$$\Delta(\mathbf{w}) = \mathbb{E}\big(\mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}_*)\big)^2.$$

The upper bound follows from the fact that $\mathtt{ReLU}(\cdot)$ is 1-Lipschitz, i.e., $|\mathtt{ReLU}(a)-\mathtt{ReLU}(b)| \le |a-b|$.

For the lower bound, we first expand the excess risk to obtain that

$$\begin{aligned}
&\mathbb{E}\big(\mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}_*)\big)^2\\
&= \mathbb{E}\big(\mathbf{x}^\top\mathbf{w}\cdot\mathbb{1}[\mathbf{x}^\top\mathbf{w}>0] - \mathbf{x}^\top\mathbf{w}_*\cdot\mathbb{1}[\mathbf{x}^\top\mathbf{w}_*>0]\big)^2\\
&= \mathbb{E}\big[\mathbf{w}^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}\cdot\mathbb{1}[\mathbf{x}^\top\mathbf{w}>0]\big] + \mathbb{E}\big[\mathbf{w}_*^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}_*\cdot\mathbb{1}[\mathbf{x}^\top\mathbf{w}_*>0]\big]\\
&\quad - 2\mathbb{E}\big[\mathbf{w}^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}_*\cdot\mathbb{1}[\mathbf{x}^\top\mathbf{w}>0, \mathbf{x}^\top\mathbf{w}_*>0]\big].
\end{aligned}$$

In the above equation, we use Assumption A.1(B) to obtain that

$$\begin{aligned}
&\mathbb{E}\big(\mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}_*)\big)^2\\
&= \mathbb{E}\big[\mathbf{w}^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}\cdot\mathbb{1}[\mathbf{x}^\top\mathbf{w}<0]\big] + \mathbb{E}\big[\mathbf{w}_*^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}_*\cdot\mathbb{1}[\mathbf{x}^\top\mathbf{w}_*<0]\big]\\
&\quad - 2\mathbb{E}\big[\mathbf{w}^\top\mathbf{x}\mathbf{x}^\top\mathbf{w}_*\cdot\mathbb{1}[\mathbf{x}^\top\mathbf{w}<0, \mathbf{x}^\top\mathbf{w}_*<0]\big]\\
&= \mathbb{E}\big(\mathtt{ReLU}(-\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(-\mathbf{x}^\top\mathbf{w}_*)\big)^2.
\end{aligned}$$

Moreover, notice the following by Cauchy inequality:

$$\begin{aligned}
&\big(\mathbf{x}^\top\mathbf{w} - \mathbf{x}^\top\mathbf{w}_*\big)^2\\
&= \big(\mathbf{x}^\top\mathbf{w}\mathbb{1}[\mathbf{x}^\top\mathbf{w}>0] - \mathbf{x}^\top\mathbf{w}_*\mathbb{1}[\mathbf{x}^\top\mathbf{w}_*>0] + \mathbf{x}^\top\mathbf{w}\mathbb{1}[\mathbf{x}^\top\mathbf{w}<0] - \mathbf{x}^\top\mathbf{w}_*\mathbb{1}[\mathbf{x}^\top\mathbf{w}_*<0]\big)^2\\
&\le 2\big(\mathbf{x}^\top\mathbf{w}\mathbb{1}[\mathbf{x}^\top\mathbf{w}>0] - \mathbf{x}^\top\mathbf{w}_*\mathbb{1}[\mathbf{x}^\top\mathbf{w}_*>0]\big)^2 + 2\big(\mathbf{x}^\top\mathbf{w}\mathbb{1}[\mathbf{x}^\top\mathbf{w}<0] - \mathbf{x}^\top\mathbf{w}_*\mathbb{1}[\mathbf{x}^\top\mathbf{w}_*<0]\big)^2\\
&= 2\big(\mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}_*)\big)^2 + 2\big(\mathtt{ReLU}(-\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(-\mathbf{x}^\top\mathbf{w}_*)\big)^2.
\end{aligned}$$

Then taking an expectation on both sides we obtain that

$$\begin{aligned}
\mathbb{E}\big(\mathbf{x}^\top\mathbf{w} - \mathbf{x}^\top\mathbf{w}_*\big)^2 &\le 2\mathbb{E}\big(\mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}_*)\big)^2 + 2\mathbb{E}\big(\mathtt{ReLU}(-\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(-\mathbf{x}^\top\mathbf{w}_*)\big)^2\\
&= 4\mathbb{E}\big(\mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}) - \mathtt{ReLU}(\mathbf{x}^\top\mathbf{w}_*)\big)^2,
\end{aligned}$$

which concludes the proof. □

B.2. Proof of Lemma 4.3

We will prove a stronger result.

Lemma B.2 (Generic bounds on the GLM-tron iterates, restated Lemma 4.3)**.** *Suppose that Assumption 4.1 holds. Consider* (GLM-tron)*. Then:*

(A) If in addition Assumptions A.1(A) and A.1(D) hold, then $\mathbf{A}_{t+1} \preceq \left(\mathcal{I} - \frac{\gamma_t}{2}\cdot\mathcal{T}(2\gamma_t)\right)\circ\mathbf{A}_{t-1} + \gamma_t^2\sigma^2\mathbf{H}$;

(B) If in addition Assumptions A.1(A), A.1(C) and A.1(D) hold, then $\mathbf{A}_{t+1} \succeq \Big(\mathcal{I} - \frac{\gamma_t}{2} \cdot \mathcal{T}\big(\frac{\gamma_t}{2}\big)\Big) \circ \mathbf{A}_t + \frac{\gamma_t^2 \sigma^2}{4} \cdot \mathbf{H}$.

Proof. From (GLM-tron) we have

$$\begin{aligned}
\mathbf{w}_t &= \mathbf{w}_{t-1} - \gamma_t \cdot \big(\texttt{ReLU}(\mathbf{x}_t^\top \mathbf{w}_{t-1}) - y_t\big)\mathbf{x}_t \\
&= \mathbf{w}_{t-1} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_{t-1} + \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_* + \gamma_t \epsilon_t \mathbf{x}_t \\
&= \mathbf{w}_{t-1} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*) \\
&\quad + \gamma_t \big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_* + \gamma_t \epsilon_t \mathbf{x}_t,
\end{aligned}$$

which implies that

$$\begin{aligned}
\mathbf{w}_t - \mathbf{w}_* = \Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \mathbf{x}_t \mathbf{x}_t^\top\Big)(\mathbf{w}_{t-1} - \mathbf{w}_*) \\
+ \gamma_t \big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_* + \gamma_t \epsilon_t \mathbf{x}_t.
\end{aligned} \tag{10}$$

Let us consider the expected outer product:

$$\begin{aligned}
&\mathbb{E}\big(\mathbf{w}_t - \mathbf{w}_*\big)^{\otimes 2} \\
&= \mathbb{E} \underbrace{\Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big)^{\otimes 2} \circ (\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2}}_{\texttt{(quadratic term 1)}} \\
&\quad + \gamma_t^2 \cdot \mathbb{E} \underbrace{\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_* \mathbf{w}_*^\top \mathbf{x}_t \mathbf{x}_t^\top}_{\texttt{(quadratic term 2)}} \\
&\quad + \gamma_t \cdot \mathbb{E} \underbrace{\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_* (\mathbf{w}_{t-1} - \mathbf{w}_*)^\top \Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big)}_{\texttt{(crossing term 1)}} \\
&\quad + \gamma_t \cdot \mathbb{E} \underbrace{\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big)(\mathbf{w}_{t-1} - \mathbf{w}_*) \mathbf{w}_*^\top \mathbf{x}_t \mathbf{x}_t^\top}_{\texttt{(crossing term 2)}} \\
&\quad + \gamma_t^2 \cdot \mathbb{E}\big(\epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^\top\big),
\end{aligned} \tag{11}$$

where the crossing terms involving ϵ_t has zero expectation because $\mathbb{E}[\epsilon_t | \mathbf{x}_t] = 0$.

For the second quadratic term in (11), notice that

$$\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big)^2 = \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] + \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} < 0, \mathbf{x}_t^\top \mathbf{w}_* > 0],$$

then we have

$$\begin{aligned}
&\mathbb{E}(\texttt{quadratic term 2}) \\
&= \mathbb{E}\Big(\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big)^2 \cdot \big(\mathbf{x}_t^\top \mathbf{w}_*\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&= \mathbb{E}\Big(\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] + \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} < 0, \mathbf{x}_t^\top \mathbf{w}_* > 0]\Big) \cdot \big(\mathbf{x}_t^\top \mathbf{w}_*\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \qquad (12) \\
&= 2 \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \big(\mathbf{x}_t^\top \mathbf{w}_*\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big), \qquad (13)
\end{aligned}$$

where the last equation is by Assumption A.1(D). For the crossing terms in (11) we have that

$$\begin{aligned}
&(\texttt{crossing term 1}) + (\texttt{crossing term 2}) \\
&= \big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \Big(\mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_* (\mathbf{w}_{t-1} - \mathbf{w}_*)^\top + (\mathbf{w}_{t-1} - \mathbf{w}_*) \mathbf{w}_*^\top \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&\quad - 2\gamma_t \big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*) \cdot \mathbf{x}_t \mathbf{x}_t^\top
\end{aligned}$$

$$= \big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \Big(\mathbf{x}_t\mathbf{x}_t^\top \mathbf{w}_*(\mathbf{w}_{t-1} - \mathbf{w}_*)^\top + (\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{w}_*^\top \mathbf{x}_t\mathbf{x}_t^\top\Big)$$
$$+ 2\gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}_*) \cdot \mathbf{x}_t\mathbf{x}_t^\top, \tag{14}$$

where in the last equality we use

$$\begin{aligned}
&- \big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \\
&= \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] \cdot \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \\
&= \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0].
\end{aligned}$$

Now we take expectation on both sides of (14). By Assumption A.1(A) (or Lemma A.2(A)) the first term in (14) has zero expectation, therefore we obtain

$$\begin{aligned}
&\mathbb{E}\big((\texttt{crossing term 1}) + (\texttt{crossing term 2})\big) \\
&= 2\gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}_*) \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big) \\
&= 2\gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top \mathbf{w}_{t-1} \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big) \\
&\quad - 2\gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \big(\mathbf{x}_t^\top \mathbf{w}_*\big)^2 \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big),
\end{aligned} \tag{15}$$

Now considering (11) and applying (13) and (15), we obtain

$$\begin{aligned}
\mathbb{E}\big(\mathbf{w}_t - \mathbf{w}_*\big)^{\otimes 2} &= \mathbb{E}\Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big)^{\otimes 2} \circ (\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2} + \gamma_t^2\sigma^2\mathbf{H} \\
&\quad + 2\gamma_t^2 \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top \mathbf{w}_{t-1} \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big).
\end{aligned} \tag{16}$$

An Upper Bound. In (16), we can use the indicator function to show that

$$\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top \mathbf{w}_{t-1} \le 0,$$

so we have

$$\begin{aligned}
\mathbb{E}\big(\mathbf{w}_t - \mathbf{w}_*\big)^{\otimes 2} &\preceq \mathbb{E}\Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big)^{\otimes 2} \circ (\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2} + \gamma_t^2\sigma^2\mathbf{H} \\
&= \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \\
&\quad - \gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big) \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \\
&\quad - \gamma_t \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big) \\
&\quad + \gamma_t^2 \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t\mathbf{x}_t^\top \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big) \\
&\quad + \gamma_t^2\sigma^2\mathbf{H}.
\end{aligned} \tag{17}$$

By Assumption A.1(A) (or Lemma A.2(A)) we have

$$\mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big) = \frac{1}{2}\mathbf{H},$$

moreover

$$\mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t\mathbf{x}_t^\top \otimes \mathbf{x}_t\mathbf{x}_t^\top\Big) \preceq \mathbb{E}\Big(\mathbf{x}_t\mathbf{x}_t^\top \otimes \mathbf{x}_t\mathbf{x}_t^\top\Big) = \mathcal{M}.$$

Then under notations of $\mathbf{A}_t$, $\mathcal{T}$ and $\mathcal{M}$, (17) can be written as

$$\begin{aligned}
\mathbf{A}_t &\preceq \mathbf{A}_{t-1} - \frac{\gamma_t}{2}\big(\mathbf{H}\mathbf{A}_{t-1} + \mathbf{A}_{t-1}\mathbf{H}\big) + \gamma_t^2\mathcal{M} \circ \mathbf{A}_{t-1} + \gamma_t^2\sigma^2\mathbf{H} \\
&= \Big(\mathcal{I} - \frac{\gamma_t}{2} \cdot \mathcal{T}(2\gamma_t)\Big) \circ \mathbf{A}_{t-1} + \gamma_t^2\sigma^2\mathbf{H}.
\end{aligned}$$

A Lower Bound. We now derive a lower bound for (16). We first notice the following fact: for every two vectors $\mathbf{v}$ and $\mathbf{u}$, it holds that

$$\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top = \frac{1}{2}\big((\mathbf{u}+\mathbf{v})^{\otimes 2} - (\mathbf{u}-\mathbf{v})^{\otimes 2}\big) \succeq -\frac{1}{2}(\mathbf{u}-\mathbf{v})^{\otimes 2}. \tag{18}$$

Applying (18), we obtain that

$$\begin{aligned}
&2\gamma_t^2 \cdot \mathbb{E}\bigg(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top \mathbf{w}_{t-1} \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg) \\
&= \gamma_t^2 \cdot \mathbb{E}\bigg(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \cdot \big(\mathbf{w}_* \mathbf{w}_{t-1}^\top + \mathbf{w}_{t-1}\mathbf{w}_*^\top\big) \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg) \\
&\succeq -\frac{\gamma_t^2}{2} \cdot \mathbb{E}\bigg(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \cdot (\mathbf{w}_{t-1} - \mathbf{w}_*)(\mathbf{w}_{t-1} - \mathbf{w}_*)^\top \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg) \\
&\succeq -\frac{\gamma_t^2}{2} \cdot \mathbb{E}\bigg(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \cdot \mathbb{E}(\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2} \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg).
\end{aligned}$$

We now bring this into (16), then we get

$$\begin{aligned}
\mathbb{E}\big(\mathbf{w}_t - \mathbf{w}_*\big)^{\otimes 2} &\succeq \mathbb{E}\Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big)^{\otimes 2} \circ (\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2} + \gamma_t^2 \sigma^2 \mathbf{H} \\
&\quad - \frac{\gamma_t^2}{2} \cdot \mathbb{E}\bigg(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \cdot \mathbb{E}(\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2} \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg) \\
&= \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \\
&\quad - \gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \\
&\quad - \gamma_t \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&\quad + \gamma_t^2 \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&\quad + \gamma_t^2 \sigma^2 \mathbf{H} \\
&\quad - \frac{\gamma_t^2}{2} \cdot \mathbb{E}\bigg(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \cdot \mathbb{E}(\mathbf{w}_* - \mathbf{w}_{t-1})^{\otimes 2} \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg) \\
&= \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \\
&\quad - \gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \\
&\quad - \gamma_t \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_{t-1}^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&\quad + \frac{\gamma_t^2}{2} \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \cdot \mathbb{E}\big(\mathbf{w}_{t-1} - \mathbf{w}_*\big)^{\otimes 2} \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&\quad + \gamma_t^2 \sigma^2 \mathbf{H}.
\end{aligned} \tag{19}$$

By Assumptions A.1(A) and A.1(C) (or Lemma A.2(B)) we have

$$\mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) = \frac{1}{2}\mathbf{H}, \quad \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top \otimes \mathbf{x}_t \mathbf{x}_t^\top\Big) = \frac{1}{2}\mathcal{M}.$$

Then under notations of $\mathbf{A}_t$, $\mathcal{T}$ and $\mathcal{M}$, (19) can be written as

$$\begin{aligned}
\mathbf{A}_t &\succeq \mathbf{A}_{t-1} - \frac{\gamma_t}{2}\big(\mathbf{H}\mathbf{A}_{t-1} + \mathbf{A}_{t-1}\mathbf{H}\big) + \frac{\gamma_t^2}{4}\mathcal{M} \circ \mathbf{A}_{t-1} + \gamma_t^2 \sigma^2 \mathbf{H} \\
&= \bigg(\mathcal{I} - \frac{\gamma_t}{2} \cdot \mathcal{T}\bigg(\frac{\gamma_t}{2}\bigg)\bigg) \circ \mathbf{A}_{t-1} + \gamma_t^2 \sigma^2 \mathbf{H}.
\end{aligned}$$

We have completed the proof. □

B.3. Proof of Theorem 4.5

Notations. In this section, we always assume that $\mathbf{H}$ is diagonal. For a PSD matrix $\mathbf{A}$, we use $\mathring{\mathbf{A}}$ to refer to the diagonal of $\mathbf{A}$.

Proof of Theorem 4.5. The proof is by combing Lemma B.2, Lemma B.1 and the analysis for one-hot data in Zou et al. (2021a).

Note that for symmetric Bernoulli distribution, or under Assumption 4.4, it holds that (see also the proof of Lemma A.1 in Zou et al. (2021a)): for any PSD matrix $\mathbf{A}$,

$$\mathcal{M} \circ \mathbf{A} = \mathbb{E}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) \cdot \mathbf{x}\mathbf{x}^\top = \mathtt{diag}(\mathbf{H}\mathbf{A}) = \mathbf{H}\mathring{\mathbf{A}}. \tag{20}$$

Upper Bound. We first show the upper bound. By Lemma B.2 and (20) we have

$$\begin{aligned}\mathbf{A}_t &\preceq \mathbf{A}_{t-1} - \frac{\gamma_t}{2}\big(\mathbf{H}\mathbf{A}_{t-1} + \mathbf{A}_{t-1}\mathbf{H}\big) + \gamma_t^2 \mathcal{M} \circ \mathbf{A}_{t-1} + \gamma_t^2\sigma^2\mathbf{H} \\ &= \mathbf{A}_{t-1} - \frac{\gamma_t}{2}\big(\mathbf{H}\mathbf{A}_{t-1} + \mathbf{A}_{t-1}\mathbf{H}\big) + \gamma_t^2 \mathbf{H}\mathring{\mathbf{A}}_{t-1} + \gamma_t^2\sigma^2\mathbf{H}.\end{aligned}$$

Taking diagonal on both sides we get

$$\begin{aligned}\mathring{\mathbf{A}}_t &\preceq \mathring{\mathbf{A}}_{t-1} - \gamma_t \mathbf{H}\mathring{\mathbf{A}}_{t-1} + \gamma_t^2 \mathbf{H}\mathring{\mathbf{A}}_{t-1} + \gamma_t^2\sigma^2\mathbf{H} \\ &\preceq \Big(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\Big)\cdot\mathring{\mathbf{A}}_{t-1} + \gamma_t^2\sigma^2\mathbf{H},\end{aligned}$$

where we use the assumption that $\gamma < 1/2$. Solving the above recursion and apply Lemma C.7, we obtain

$$\begin{aligned}\mathring{\mathbf{A}}_N &\preceq \prod_{t=1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\Big)\cdot\mathring{\mathbf{A}}_0 + \sigma^2\sum_{t=1}^N \gamma_t^2 \prod_{k=t+1}^N \Big(\mathbf{I} - \frac{\gamma_k}{2}\cdot\mathbf{H}\Big)\mathbf{H} \\ &\preceq \prod_{t=1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\Big)\cdot\mathring{\mathbf{A}}_0 + \frac{\sigma^2}{8}\cdot\bigg(\frac{1}{N_{\mathtt{eff}}}\mathbf{H}_{0:k}^{-1} + N_{\mathtt{eff}}\gamma_0^2\mathbf{H}_{k:\infty}\bigg).\end{aligned}$$

Taking inner product with $\mathbf{H}$ gives the upper bound on the excess risk.

Lower Bound. We next show the lower bound. By Lemma B.2 and (20) we have

$$\begin{aligned}\mathbf{A}_t &\succeq \mathbf{A}_{t-1} - \frac{\gamma_t}{2}\big(\mathbf{H}\mathbf{A}_{t-1} + \mathbf{A}_{t-1}\mathbf{H}\big) + \frac{\gamma_t^2}{4} \mathcal{M} \circ \mathbf{A}_{t-1} + \gamma_t^2\sigma^2\mathbf{H} \\ &\succeq \mathbf{A}_{t-1} - \frac{\gamma_t}{2}\big(\mathbf{H}\mathbf{A}_{t-1} + \mathbf{A}_{t-1}\mathbf{H}\big) + \gamma_t^2\sigma^2\mathbf{H}.\end{aligned}$$

Taking diagonal on both sides we get

$$\mathring{\mathbf{A}}_t \succeq \big(\mathbf{I} - \gamma_t\mathbf{H}\big)\cdot\mathring{\mathbf{A}}_{t-1} + \gamma_t^2\sigma^2\mathbf{H}.$$

Solving the above recursion and apply Lemma C.7, we obtain

$$\begin{aligned}\mathring{\mathbf{A}}_N &\succeq \prod_{t=1}^N \Big(\mathbf{I} - \gamma_t\mathbf{H}\Big)\cdot\mathring{\mathbf{A}}_0 + \sigma^2\sum_{t=1}^N \gamma_t^2 \prod_{k=t+1}^N \Big(\mathbf{I} - \gamma_k\mathbf{H}\Big)\mathbf{H} \\ &\succeq \prod_{t=1}^N \Big(\mathbf{I} - \gamma_t\mathbf{H}\Big)\cdot\mathring{\mathbf{A}}_0 + \frac{\sigma^2}{400}\cdot\bigg(\frac{1}{N_{\mathtt{eff}}}\mathbf{H}_{0:k^*}^{-1} + N_{\mathtt{eff}}\gamma_0^2\mathbf{H}_{k^*:\infty}\bigg),\end{aligned}$$

where $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 N_{\mathtt{eff}})\}$. Taking inner product with $\mathbf{H}$ gives the lower bound on the excess risk. □

B.4. Proof of Theorem 4.7

We first restate Corollary 3.4 in Wu et al. (2022b) under our notations.

Corollary (Corollary 3.4 in Wu et al. (2022b), restated)**.** *Consider a sequence of PSD matrices $(\mathbf{A}_t)_{t=0}^N$ that describes the covariance of the SGD iterates for linear regression, i.e.,*

$$\mathbf{A}_0 := (\mathbf{w}_0 - \mathbf{w}_*)^{\otimes 2}, \quad \mathbf{A}_t := \mathbb{E}(\mathbf{I} - \gamma_t \mathbf{x}\mathbf{x}^\top)\mathbf{A}_{t-1}(\mathbf{I} - \gamma_t \mathbf{x}\mathbf{x}^\top) + \gamma_t^2 \cdot \sigma^2 \cdot \mathbf{H},\ t = 1, \ldots, N,$$

where $(\gamma_t)_{t=0}^N$ is a stepsize scheduler as defined in (2)*. Assume that $N > 100$. Let $N_{\mathtt{eff}} := N/\log(N)$.*

(A) If Assumption 4.6(A) holds, then for $\gamma_0 < 1/(4\alpha(\mathtt{tr}(\mathbf{H})))$ it holds that

$$\langle \mathbf{H}, \mathbf{A}_N \rangle \lesssim \Big\| \prod_{t=1}^N \Big(\mathbf{I} - \gamma_t \mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*) \Big\|_{\mathbf{H}}^2 + \Big(\alpha \big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\frac{\mathbf{I}_{0:k^*}}{N_{\mathtt{eff}}\gamma_0} + \mathbf{H}_{k^*:\infty}} + \sigma^2\Big) \cdot \frac{k^* + N_{\mathtt{eff}}^2 \gamma_0^2 \cdot \sum_{i>k^*} \lambda_i^2}{N_{\mathtt{eff}}},$$

where $k^* \geq 0$ is an arbitrary index.

(B) If Assumption 4.6(B) holds, then for $\gamma_0 < 1/(4\alpha(\mathtt{tr}(\mathbf{H})))$ it holds that

$$\langle \mathbf{H}, \mathbf{A}_N \rangle \gtrsim \Big\| \prod_{t=1}^N \Big(\mathbf{I} - \gamma_t \mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*) \Big\|_{\mathbf{H}}^2 + \big(\beta \|\mathbf{w}_0 - \mathbf{w}_*\|^2_{\mathbf{H}_{k^*:\infty}} + \sigma^2\big) \cdot \frac{k^* + N_{\mathtt{eff}}^2 \gamma_0^2 \cdot \sum_{i>k^*} \lambda_i^2}{N_{\mathtt{eff}}},$$

where $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 N_{\mathtt{eff}})\}$.

Proof. See Corollary 3.4 in Wu et al. (2022b). □

We restate Theorem 4.7 in a slightly stronger version.

Theorem B.3 (Risk Bounds for GLM-tron, restated Theorem 4.7)**.** *Suppose that Assumption 4.1 holds. Let $\mathbf{w}_N$ be the output of* (GLM-tron) *with stepsize scheduler* (2)*. Assume that $N > 100$. Let $N_{\mathtt{eff}} := N/\log(N)$.*

(A) If in addition Assumption 4.6(A) and Assumption A.1(A)(D) hold, then for $\gamma_0 < 1/(4\alpha(\mathtt{tr}(\mathbf{H})))$ it holds that

$$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \Big\| \prod_{t=1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2} \mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*) \Big\|_{\mathbf{H}}^2 + \Big(\alpha \big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\frac{\mathbf{I}_{0:k^*}}{N_{\mathtt{eff}}\gamma_0} + \mathbf{H}_{k^*:\infty}} + \sigma^2\Big) \cdot \frac{k^* + N_{\mathtt{eff}}^2 \gamma_0^2 \cdot \sum_{i>k^*} \lambda_i^2}{N_{\mathtt{eff}}},$$

where $k^* \geq 0$ is an arbitrary index.

(B) If in addition Assumption 4.6(B) and Assumption A.1 hold, then for $\gamma_0 < 1/\lambda_1$, it holds that

$$\mathbb{E}\Delta(\mathbf{w}_N) \gtrsim \Big\| \prod_{t=1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2} \mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*) \Big\|_{\mathbf{H}}^2 + \big(\beta \|\mathbf{w}_0 - \mathbf{w}_*\|^2_{\mathbf{H}_{k^*:\infty}} + \sigma^2\big) \cdot \frac{k^* + N_{\mathtt{eff}}^2 \gamma_0^2 \cdot \sum_{i>k^*} \lambda_i^2}{N_{\mathtt{eff}}},$$

where $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 N_{\mathtt{eff}})\}$.

Proof. We first use Lemma B.1 and Lemma B.2 to relate GLM-tron for ReLU regression problems to SGD for linear regression problems. Then we invoke Corollary 3.4 in Wu et al. (2022b) (see above) to get the results. □

B.5. Proof of Corollary 1

Proof of Corollary 1. For all these examples one can verify that $\mathtt{tr}(\mathbf{H}) \eqsim 1$. Therefore $\gamma_0 \eqsim 1$.

We can verify that

$$\Big\| \prod_{t=1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2} \mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*) \Big\|_{\mathbf{H}}^2 \leq \Big\| \Big(\mathbf{I} - \frac{\gamma_0}{2} \mathbf{H}\Big)^{N_{\mathtt{eff}}} (\mathbf{w}_0 - \mathbf{w}_*) \Big\|_{\mathbf{H}}^2$$

$$
\begin{aligned}
&= \sum_i \lambda_i \cdot \left(1 - \frac{\gamma_0}{2} \cdot \lambda_i\right)^{2N_{\mathtt{eff}}} \cdot (\mathbf{w}_0[i] - \mathbf{w}_*[i])^2 \\
&\lesssim \sum_i \lambda_i \cdot \frac{1}{\gamma_0 \lambda_i N_{\mathtt{eff}}} \cdot (\mathbf{w}_0[i] - \mathbf{w}_*[i])^2 \\
&\eqsim \frac{\left\|(\mathbf{w}_0 - \mathbf{w}^*)\right\|_2^2}{\gamma_0 N_{\mathtt{eff}}} \\
&\eqsim \frac{1}{N_{\mathtt{eff}}} \eqsim \frac{\log(N)}{N},
\end{aligned}
$$

and that

$$
\left\|\mathbf{w}_0 - \mathbf{w}_*\right\|^2_{\frac{\mathbf{I}_{0:k^*}}{N_{\mathtt{eff}}\gamma_0} + \mathbf{H}_{k^*:\infty}} \lesssim \left\|\mathbf{w}_0 - \mathbf{w}_*\right\|_2^2 \lesssim 1.
$$

Therefore in Theorem 4.7 we have

$$
\begin{aligned}
\mathbb{E}\Delta(\mathbf{w}_N) &\lesssim \left\| \prod_{t=1}^{N} \left(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\right)(\mathbf{w}_0 - \mathbf{w}_*) \right\|_{\mathbf{H}}^2 + \left(\alpha \left\|\mathbf{w}_0 - \mathbf{w}_*\right\|^2_{\frac{\mathbf{I}_{0:k^*}}{N_{\mathtt{eff}}\gamma_0} + \mathbf{H}_{k^*:\infty}} + \sigma^2\right) \cdot \frac{k^* + N_{\mathtt{eff}}^2 \gamma_0^2 \cdot \sum_{i>k^*} \lambda_i^2}{N_{\mathtt{eff}}} \\
&\lesssim \frac{1}{N_{\mathtt{eff}}} + \frac{k^* + N_{\mathtt{eff}}^2 \gamma_0^2 \cdot \sum_{i>k^*} \lambda_i^2}{N_{\mathtt{eff}}} \\
&\lesssim \frac{k^* + N_{\mathtt{eff}}^2 \cdot \sum_{i>k^*} \lambda_i^2}{N_{\mathtt{eff}}}.
\end{aligned}
$$

We next examine each case. Recall that $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 N_{\mathtt{eff}})\}$.

1. By definitions we have

   $$k^* \eqsim (N_{\mathtt{eff}})^{\frac{1}{1+r}},$$

   therefore we have

   $$
   \begin{aligned}
   k^* + N_{\mathtt{eff}}^2 \cdot \sum_{i>k^*} \lambda_i^2 &\eqsim k^* + (N_{\mathtt{eff}})^2 \cdot (k^*)^{-1-2r} \\
   &\eqsim (N_{\mathtt{eff}})^{\frac{1}{1+r}}.
   \end{aligned}
   $$

   This implies that

   $$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim (N_{\mathtt{eff}})^{\frac{-r}{1+r}} \eqsim (N/\log(N))^{\frac{-r}{1+r}}.$$

2. By definitions we have

   $$k^* \eqsim N_{\mathtt{eff}} \cdot \log^{-r}(N_{\mathtt{eff}}),$$

   therefore we have

   $$
   \begin{aligned}
   k^* + N_{\mathtt{eff}}^2 \cdot \sum_{i>k^*} \lambda_i^2 &\eqsim k^* + (N_{\mathtt{eff}})^2 \cdot (k^*)^{-1} \log^{-2r}(k^*) \\
   &\eqsim N_{\mathtt{eff}} \cdot \log^{-r}(N_{\mathtt{eff}}).
   \end{aligned}
   $$

   This implies that

   $$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \log^{-r}(N_{\mathtt{eff}}) \eqsim \log^{-r}(N/\log(N)) \eqsim \log^{-r}(N).$$

3. By definitions we have

   $$k^* \eqsim \log(N_{\mathtt{eff}}),$$

   therefore we have

   $$
   \begin{aligned}
   k^* + N_{\mathtt{eff}}^2 \cdot \sum_{i>k^*} \lambda_i^2 &\eqsim k^* + (N_{\mathtt{eff}})^2 \cdot 2^{-k^*} \\
   &\eqsim \log(N_{\mathtt{eff}}).
   \end{aligned}
   $$

   This implies that

   $$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \log(N_{\mathtt{eff}})/N_{\mathtt{eff}} \eqsim \log^2(N)/N.$$

We have completed the proof. □

B.6. Iterate Average

We may also consider constant-stepsize GLM-tron with iterate averaging, i.e., (GLM-tron) is run with constant stepsize γ and outputs the average of the iterates:

$$\bar{\mathbf{w}}_N := \frac{1}{N}\sum_{t=0}^{N-1}\mathbf{w}_t. \tag{21}$$

Lemma B.4 (Iterate averaging)**.** *Suppose that Assumption 4.1 and Assumption A.1(A) hold. For $\bar{\mathbf{w}}_N$ defined in* (21)*, we have that*

$$\mathbb{E}\langle \mathbf{H}, (\bar{\mathbf{w}}_N - \mathbf{w}_*)^{\otimes 2}\rangle \le \frac{1}{\gamma N^2}\left\langle \mathbf{I} - \left(\mathbf{I} - \frac{\gamma}{2}\mathbf{H}\right)^N, \sum_{t=0}^{N}\mathbf{A}_t\right\rangle;$$

$$\mathbb{E}\langle \mathbf{H}, (\bar{\mathbf{w}}_N - \mathbf{w}_*)^{\otimes 2}\rangle \ge \frac{1}{2\gamma N^2}\left\langle \mathbf{I} - \left(\mathbf{I} - \frac{\gamma}{2}\mathbf{H}\right)^{N/2}, \sum_{t=0}^{N/2}\mathbf{A}_t\right\rangle.$$

Proof. In (10), we take conditional expectation to obtain

$$\begin{aligned}
\mathbb{E}[\mathbf{w}_t - \mathbf{w}_*|\mathbf{w}_{t-1}] &= \mathbb{E}\left[\left(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\mathbf{x}_t\mathbf{x}_t^\top\right)(\mathbf{w}_{t-1} - \mathbf{w}_*)|\mathbf{w}_{t-1}\right] \\
&\quad + \gamma_t \cdot \mathbb{E}\left[\left(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\right)\mathbf{x}_t\mathbf{x}_t^\top \mathbf{w}_*|\mathbf{w}_{t-1}\right] + \gamma_t \mathbb{E}[\epsilon_t \mathbf{x}_t|\mathbf{w}_{t-1}] \\
&= \mathbb{E}\left[\left(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\mathbf{x}_t\mathbf{x}_t^\top\right)(\mathbf{w}_{t-1} - \mathbf{w}_*)|\mathbf{w}_{t-1}\right] \\
&= \left(\mathbf{I} - \frac{\gamma}{2}\mathbf{H}\right)(\mathbf{w}_{t-1} - \mathbf{w}_*),
\end{aligned}$$

where the second equation is due to Assumption A.1(A) (or Lemma A.2(A)) and Assumption 4.1, and the third equation is due to Assumption A.1(A) (or Lemma A.2(A)). Applying the above recursively we obtain that: for $t > s$,

$$\mathbb{E}[\mathbf{w}_t - \mathbf{w}_*|\mathbf{w}_s] = \left(\mathbf{I} - \frac{\gamma}{2}\mathbf{H}\right)^{t-s}(\mathbf{w}_s - \mathbf{w}_*),$$

which also implies that

$$\mathbb{E}[(\mathbf{w}_t - \mathbf{w}_*) \otimes (\mathbf{w}_s - \mathbf{w}_*)] = \left(\mathbf{I} - \frac{\gamma}{2}\mathbf{H}\right)^{t-s} \cdot \mathbb{E}(\mathbf{w}_s - \mathbf{w}_*)^{\otimes 2} = \left(\mathbf{I} - \frac{\gamma}{2}\mathbf{H}\right)^{t-s} \cdot \mathbf{A}_s. \tag{22}$$

Now let us consider $\mathbb{E}(\bar{\mathbf{w}}_N - \mathbf{w}_*)^{\otimes 2}$:

$$\begin{aligned}
&\mathbb{E}(\bar{\mathbf{w}}_N - \mathbf{w}_*)^{\otimes 2} \\
&= \frac{1}{N^2} \cdot \left(\mathbb{E}\sum_{t=0}^{N-1}(\mathbf{w}_t - \mathbf{w}_*)^{\otimes 2} + \mathbb{E}\sum_{s=0}^{N-1}\sum_{t=s+1}^{N-1}\Big((\mathbf{w}_t - \mathbf{w}_*) \otimes (\mathbf{w}_s - \mathbf{w}_*) + (\mathbf{w}_s - \mathbf{w}_*) \otimes (\mathbf{w}_t - \mathbf{w}_*)\Big)\right) \\
&= \frac{1}{N^2} \cdot \left(\sum_{s=0}^{N-1}\mathbf{A}_s + \sum_{s=0}^{N-1}\sum_{t=s+1}^{N-1}\left(\left(\mathbf{I} - \frac{\gamma}{2}\mathbf{H}\right)^{t-s} \cdot \mathbf{A}_s + \mathbf{A}_s \cdot \left(\mathbf{I} - \frac{\gamma}{2}\mathbf{H}\right)^{t-s}\right)\right).
\end{aligned}$$

The remaining proof simply follows from Zou et al. (2021b).

□

We next present the risk bounds for constant-stepsize GLM-tron with iterate averaging as follows.

Theorem B.5 (Risk Bounds for constant-stepsize GLM-tron)**.** *Suppose that Assumption 4.1 holds. Consider* $\bar{\mathbf{w}}_N$ *defined in* (21)*, i.e., the iterate average of constant stepsize* (GLM-tron)*. Suppose* $N > 100$*.*

(A) If in addition Assumption 4.6(A) and Assumption A.1(A)(D) hold, then for $\gamma < 1/(4\alpha(\mathtt{tr}(\mathbf{H})))$ *it holds that*

$$\mathbb{E}\Delta(\bar{\mathbf{w}}_N) \lesssim \frac{1}{N^2\gamma^2}\cdot\left\|\mathbf{w}_0-\mathbf{w}_*\right\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \left\|\mathbf{w}_0-\mathbf{w}_*\right\|^2_{\mathbf{H}_{k^*:\infty}} + \left(\alpha\cdot\frac{\left\|\mathbf{w}_0-\mathbf{w}_*\right\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\cdot\left\|\mathbf{w}_0-\mathbf{w}_*\right\|^2_{\mathbf{H}_{k^*:\infty}}}{N\gamma} + \sigma^2\right)\cdot\frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\cdot\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}},$$

where $k^* \geq 0$ *is an arbitrary index.*

(B) If in addition Assumption 4.6(B) and Assumption A.1 hold, then for $\gamma_0 < 1/\lambda_1$*, it holds that*

$$\mathbb{E}\Delta(\mathbf{w}_N) \gtrsim \frac{1}{N^2\gamma^2}\cdot\left\|\mathbf{w}_0-\mathbf{w}_*\right\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \left\|\mathbf{w}_0-\mathbf{w}_*\right\|^2_{\mathbf{H}_{k^*:\infty}} + \left(\beta\cdot\frac{\left\|\mathbf{w}_0-\mathbf{w}_*\right\|^2_{\mathbf{I}_{0:k^*}} + N\gamma\cdot\left\|\mathbf{w}_0-\mathbf{w}_*\right\|^2_{\mathbf{H}_{k^*:\infty}}}{N\gamma} + \sigma^2\right)\cdot\frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\cdot\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}},$$

where $k^* := \max\{k : \lambda_k \geq 1/(\gamma_0 N_{\mathtt{eff}})\}$.

Proof. We first use Lemma B.1 and Lemma B.2 to relate GLM-tron for ReLU regression problems to SGD for linear regression problems. Then we invoke Lemma B.4 and the proof of Theorems 2.1 and 2.2 in Zou et al. (2021b) to get the results. □

B.7. Proof of Corollary 4.8

Proof of Corollary 4.8. According to the stepsize scheduler (2) and the assumptions, we have that

$$\begin{aligned}\mathbb{E}\Delta(\mathbf{w}_N) &\lesssim \|e^{-0.5N_{\mathtt{eff}}\gamma_0\mathbf{H}}\mathbf{H}(\mathbf{w}_0-\mathbf{w}_*)\|_2^2 + \frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}}\\ &\lesssim \frac{1}{N_{\mathtt{eff}}\gamma_0} + \frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}},\end{aligned}$$

where k^* can be arbitrary.

For the first part, we choose $\gamma_0 = 1/\sqrt{N_{\mathtt{eff}}}$ and $k^* = \max\{k : \lambda_k > 1/\sqrt{N_{\mathtt{eff}}}\}$, then from $\mathtt{tr}(\mathbf{H}) \lesssim 1$ we know that

$$k^* \lesssim \sqrt{N_{\mathtt{eff}}}, \quad \sum_{i>k^*}\lambda_i^2 \lesssim \frac{1}{\sqrt{N_{\mathtt{eff}}}}.$$

Then we have

$$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \frac{1}{N_{\mathtt{eff}}\gamma_0} + \frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}} \lesssim \frac{1}{\sqrt{N_{\mathtt{eff}}}} + \frac{\sqrt{N_{\mathtt{eff}}} + N_{\mathtt{eff}}^2\cdot\frac{1}{N_{\mathtt{eff}}}\cdot\frac{1}{\sqrt{N_{\mathtt{eff}}}}}{N_{\mathtt{eff}}} \eqsim \frac{1}{\sqrt{N_{\mathtt{eff}}}}.$$

As for the second part, we choose $\gamma \eqsim 1/\,\mathtt{tr}(\mathbf{H})$, and $k^* := d$, then

$$\mathbb{E}\Delta(\mathbf{w}_N) \lesssim \frac{1}{N_{\mathtt{eff}}\gamma_0} + \frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}} \lesssim \frac{\mathtt{tr}(\mathbf{H})}{N_{\mathtt{eff}}} + \frac{d}{N_{\mathtt{eff}}} \eqsim \frac{d}{N_{\mathtt{eff}}}.$$

□

C. Misspecified Setting

In this part, we consider the misspecified setting and assume Assumption 5.1.

Notations. In this section, we assume that $\mathbf{H}$ is diagonal. For $\mathbf{A}_t := \mathbb{E}(\mathbf{w}_t - \mathbf{w}_*)^2$, we use $\mathring{\mathbf{A}}_t$ to refer to the diagonal of $\mathbf{A}_t$. For simplicity, we will use

$$\epsilon_t := y_t - \texttt{ReLU}(\mathbf{w}_*^\top \mathbf{x}_t)$$

to refer to the misspecified noise in this section.

One technique we used for dealing with misspecified cases is to study the diagonal, instead of the matrix itself, of the expected outer product of the error iterates. The following lemma is useful for translating inequalities about PSD matrices to inequalities about their diagonals.

Lemma C.1. *For every pair of symmetric matrices* $\mathbf{A}$ *and* $\mathbf{B}$*,* $\mathbf{A} \preceq \mathbf{B}$ *implies* $\mathring{\mathbf{A}} \preceq \mathring{\mathbf{B}}$*.*

Proof. We only need to show that $\texttt{diag}(\mathbf{B} - \mathbf{A})$ is PSD. This holds because every diagonal entry of a PSD matrix must be non-negative. □

C.1. Risk Landscape

We first show the following lemma about an upper bound on the risk.

Lemma C.2 (Risk landscape, misspecified case)**.** *Under Assumption 5.1, it holds that*

$$\mathcal{R}(\mathbf{w}) \le 2 \cdot \|\mathbf{w} - \mathbf{w}_*\|_{\mathbf{H}}^2 + 2 \cdot \texttt{OPT}.$$

Proof. We prove the conclusion as follows:

$$\begin{aligned}
\mathcal{R}(\mathbf{w}) &:= \mathbb{E}\big(\texttt{ReLU}(\mathbf{w}^\top \mathbf{x}) - y\big)^2 \\
&= \mathbb{E}\big(\texttt{ReLU}(\mathbf{w}^\top \mathbf{x}) - \texttt{ReLU}(\mathbf{w}_*^\top \mathbf{x}) + \texttt{ReLU}(\mathbf{w}_*^\top \mathbf{x}) - y\big)^2 \\
&\le 2 \cdot \mathbb{E}\big(\texttt{ReLU}(\mathbf{w}^\top \mathbf{x}) - \texttt{ReLU}(\mathbf{w}_*^\top \mathbf{x})\big)^2 + 2 \cdot \mathbb{E}\big(\texttt{ReLU}(\mathbf{w}_*^\top \mathbf{x}) - y\big)^2 \\
&\le 2 \cdot \mathbb{E}\big(\mathbf{w}^\top \mathbf{x} - \mathbf{w}_*^\top \mathbf{x}\big)^2 + 2 \cdot \mathbb{E}\big(\texttt{ReLU}(\mathbf{w}_*^\top \mathbf{x}) - y\big)^2 \\
&= 2 \cdot \|\mathbf{w} - \mathbf{w}_*\|_{\mathbf{H}}^2 + 2 \cdot \texttt{OPT},
\end{aligned}$$

where in the last inequality we use the fact that $\texttt{ReLU}(\cdot)$ is 1-Lipschitz. □

C.2. Iterate Bounds

Lemma C.3 (Iterate upper bound)**.** *Suppose that Assumption 5.1, Assumption A.1(A) and Assumption 4.6(A) hold, then the following holds for* (GLM-tron)*:*

$$\mathring{\mathbf{A}}_t \preceq \Big(\mathbf{I} - \frac{\gamma_t}{2} \cdot \mathbf{H}\Big) \cdot \mathring{\mathbf{A}}_{t-1} + 2\alpha\gamma_t^2 \cdot \langle \mathbf{H}, \mathring{\mathbf{A}}_{t-1} \rangle \cdot \mathbf{H} + 3\gamma_t^2(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2) \cdot \mathbf{H} + 2\gamma_t \cdot \mathbf{\Xi},$$

where $\mathbf{\Xi}$ *is a diagonal deterministic matrix and* $\texttt{tr}(\mathbf{\Xi}) \le \texttt{OPT}$*.*

Proof. We first consider the expected outer product of (10) in the misspecified setting:

$$\begin{aligned}
\mathbf{A}_t &:= \mathbb{E}\big(\mathbf{w}_t - \mathbf{w}_*\big)^{\otimes 2} \\
&= \left.\begin{aligned}
&\mathbb{E}\Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big)^{\otimes 2} \circ (\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2} \\
&+\gamma_t^2 \cdot \mathbb{E}\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_* \mathbf{w}_*^\top \mathbf{x}_t \mathbf{x}_t^\top \\
&+\gamma_t \cdot \mathbb{E}\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \mathbf{x}_t \mathbf{x}_t^\top \mathbf{w}_* (\mathbf{w}_{t-1} - \mathbf{w}_*)^\top \Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&+\gamma_t \cdot \mathbb{E}\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big)(\mathbf{w}_{t-1} - \mathbf{w}_*) \mathbf{w}_*^\top \mathbf{x}_t \mathbf{x}_t^\top
\end{aligned}\right\} =: \mathbf{S} \\
&\left.\begin{aligned}
&+\gamma_t^2 \cdot \mathbb{E}\big[\epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^\top\big] \\
&+\gamma_t \cdot \mathbb{E}\Big[\epsilon_t \Big((\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{x}^\top + \mathbf{x}(\mathbf{w}_{t-1} - \mathbf{w}_*)^\top\Big)\Big] \\
&-2\gamma_t^2 \cdot \mathbb{E}\Big[\epsilon_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*) \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big] \\
&+2\gamma_t^2 \cdot \mathbb{E}\Big[\epsilon_t \big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big],
\end{aligned}\right\} =: \mathbf{N}
\end{aligned}$$

where we decompose $\mathbf{A}_t$ into a signal part and a noise part, i.e., $\mathbf{A}_t := \mathbf{S} + \mathbf{N}$. We next upper bound these two parts separately.

Signal Part. The analysis of this part is similar to the derivation of (17) in the proof of Theorem 4.3. However this time we only use Assumption A.1(A) and do not use Assumption A.1(D). In specific, under Assumption A.1(A), (12) and (15) still hold, and applying which to the signal part $\mathbf{S}$ we obtain

$$\begin{aligned}
\mathbf{S} &= \mathbb{E}\Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big)^{\otimes 2} \circ (\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2} \\
&\quad + 2\gamma_t^2 \cdot \mathbb{E}\bigg(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top \mathbf{w}_{t-1} \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg) \\
&\quad - 2\gamma_t^2 \cdot \mathbb{E}\bigg(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot \big(\mathbf{x}_t^\top \mathbf{w}_*\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg) \\
&\quad + \gamma_t^2 \cdot \mathbb{E}\bigg(\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] + \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} < 0, \mathbf{x}_t^\top \mathbf{w}_* > 0]\Big) \cdot \big(\mathbf{x}_t^\top \mathbf{w}_*\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg).
\end{aligned}$$

In the above, the second term is always non-positive due to the property of the indicator function; and the third and fourth terms together is equal to

$$\begin{aligned}
&\gamma_t^2 \cdot \mathbb{E}\bigg(\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} < 0, \mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\Big) \cdot \big(\mathbf{x}_t^\top \mathbf{w}_*\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\bigg) \\
&\le \gamma_t^2 \cdot \mathbb{E}\big((\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\big) = \gamma_t^2 \cdot \mathcal{M} \circ (\mathbf{w}_* \mathbf{w}_*^\top),
\end{aligned}$$

so the signal part can be bounded by

$$\mathbf{S} \preceq \mathbb{E}\Big(\mathbf{I} - \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big)^{\otimes 2} \circ (\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2} + \gamma_t^2 \cdot \mathcal{M} \circ (\mathbf{w}_* \mathbf{w}_*^\top).$$

Now use Assumption A.1(A) (or Lemma A.2(A)) and Assumption 4.6(A), we obtain

$$\begin{aligned}
\mathbf{S} &\preceq \mathbf{A}_{t-1} - \frac{\gamma_t}{2}\big(\mathbf{H}\mathbf{A}_{t-1} + \mathbf{A}_{t-1}\mathbf{H}\big) + \gamma_t^2 \mathcal{M} \circ \mathbf{A}_{t-1} + \gamma_t^2 \cdot \mathcal{M} \circ (\mathbf{w}_* \mathbf{w}_*^\top) \\
&\preceq \mathbf{A}_{t-1} - \frac{\gamma_t}{2}\big(\mathbf{H}\mathbf{A}_{t-1} + \mathbf{A}_{t-1}\mathbf{H}\big) + \gamma_t^2 \mathcal{M} \circ \mathbf{A}_{t-1} + \alpha\gamma_t^2 \|\mathbf{w}_*\|_{\mathbf{H}}^2 \cdot \mathbf{H}. \qquad (23)
\end{aligned}$$

Noise Part. For the noise part, we apply Cauchy inequality to obtain

$$\begin{aligned}
\mathbf{N} &:= \gamma_t^2 \cdot \mathbb{E}\big[\epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^\top\big] + \gamma_t \cdot \mathbb{E}\Big[\epsilon_t \Big((\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{x}^\top + \mathbf{x}(\mathbf{w}_{t-1} - \mathbf{w}_*)^\top\Big)\Big] \\
&\quad - 2\gamma_t^2 \cdot \mathbb{E}\Big[\epsilon_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*) \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big] \\
&\quad + 2\gamma_t^2 \cdot \mathbb{E}\Big[\epsilon_t \big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big) \cdot \mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big] \\
&\preceq \gamma_t^2 \cdot \mathbb{E}\big[\epsilon_t^2 \mathbf{x}_t \mathbf{x}_t^\top\big] + \gamma_t \cdot \mathbb{E}\Big[\epsilon_t \Big((\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{x}^\top + \mathbf{x}(\mathbf{w}_{t-1} - \mathbf{w}_*)^\top\Big)\Big] \\
&\quad + \gamma_t^2 \cdot \mathbb{E}\Big[\Big(\epsilon_t^2 + \big(\mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*)\big)^2\Big) \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big] + \gamma_t^2 \cdot \mathbb{E}\Big[\big(\epsilon_t^2 + (\mathbf{x}_t^\top \mathbf{w}_*)^2\big) \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big].
\end{aligned}$$

Next we apply Assumption 4.6(A) and Assumption 5.1 to obtain

$$\begin{aligned}
\mathbf{N} &\preceq \gamma_t^2 \sigma^2 \cdot \mathbf{H} + \gamma_t \cdot \mathbb{E}\Big[\epsilon_t \cdot \Big((\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{x}^\top + \mathbf{x}(\mathbf{w}_{t-1} - \mathbf{w}_*)^\top\Big)\Big] \\
&\quad + \gamma_t^2 \cdot \Big(\sigma^2 \cdot \mathbf{H} + \mathcal{M} \circ \mathbf{A}_{t-1}\Big) + \gamma_t^2 \cdot \Big(\sigma^2 \cdot \mathbf{H} + \alpha \,\mathtt{tr}(\mathbf{H}\mathbf{w}_*\mathbf{w}_*^\top) \cdot \mathbf{H}\Big) \\
&= 3\gamma_t^2 \sigma^2 \cdot \mathbf{H} + \alpha\gamma_t^2 \|\mathbf{w}_*\|_{\mathbf{H}}^2 \cdot \mathbf{H} + \gamma_t^2 \cdot \mathcal{M} \circ \mathbf{A}_{t-1} + \gamma_t \cdot \mathbb{E}\Big[\epsilon_t \cdot \Big((\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{x}^\top + \mathbf{x}(\mathbf{w}_{t-1} - \mathbf{w}_*)^\top\Big)\Big].
\end{aligned}$$

Next, we take diagonal over the above inequality and apply Lemma C.1 and Lemma C.4, then we obtain

$$\mathring{\mathbf{N}} \preceq 3\gamma_t^2 \sigma^2 \cdot \mathbf{H} + \alpha\gamma_t^2 \|\mathbf{w}_*\|_{\mathbf{H}}^2 \cdot \mathbf{H} + \gamma_t^2 \cdot \mathtt{diag}(\mathcal{M} \circ \mathbf{A}_{t-1})$$

$$
\begin{aligned}
&+ \gamma_t \cdot \mathbb{E}\Big[\epsilon_t \cdot \mathtt{diag}\left((\mathbf{w}_{t-1}-\mathbf{w}_*)\mathbf{x}^\top + \mathbf{x}(\mathbf{w}_{t-1}-\mathbf{w}_*)^\top\right)\Big] \\
&\preceq 3\gamma_t^2\sigma^2 \cdot \mathbf{H} + \alpha\gamma_t^2\|\mathbf{w}_*\|_{\mathbf{H}}^2 \cdot \mathbf{H} + \gamma_t^2 \cdot \mathtt{diag}(\mathcal{M}\circ\mathbf{A}_{t-1}) + \frac{\gamma_t}{2}\cdot\mathbf{H}\mathring{\mathbf{A}}_{t-1} + 2\gamma_t\cdot\mathbf{\Xi}, \qquad (24)
\end{aligned}
$$

where $\mathbf{\Xi}$ is a deterministic diagonal PSD matrix and that $\mathtt{tr}(\mathbf{\Xi}) \le \mathtt{OPT}$.

Combining Two Parts. Combining the diagonal of (23) with (24), we have

$$
\begin{aligned}
\mathring{\mathbf{A}}_t &= \mathring{\mathbf{S}} + \mathring{\mathbf{N}} \\
&\preceq \mathring{\mathbf{A}}_{t-1} - \gamma_t\cdot\mathbf{H}\mathring{\mathbf{A}}_{t-1} + \gamma_t^2\cdot\mathtt{diag}(\mathcal{M}\circ\mathbf{A}_{t-1}) + \alpha\gamma_t^2\|\mathbf{w}_*\|_{\mathbf{H}}^2\cdot\mathbf{H} \\
&\quad + 3\gamma_t^2\sigma^2\cdot\mathbf{H} + \alpha\gamma_t^2\|\mathbf{w}_*\|_{\mathbf{H}}^2\cdot\mathbf{H} + \gamma_t^2\cdot\mathtt{diag}(\mathcal{M}\circ\mathbf{A}_{t-1}) + \frac{\gamma_t}{2}\cdot\mathbf{H}\mathring{\mathbf{A}}_{t-1} + 2\gamma_t\cdot\mathbf{\Xi} \\
&\preceq \left(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\right)\cdot\mathring{\mathbf{A}}_{t-1} + 2\gamma_t^2\cdot\mathtt{diag}(\mathcal{M}\circ\mathbf{A}_{t-1}) + 3\gamma_t^2(\sigma^2+\alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\cdot\mathbf{H} + 2\gamma_t\cdot\mathbf{\Xi} \\
&\preceq \left(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\right)\cdot\mathring{\mathbf{A}}_{t-1} + 2\alpha\gamma_t^2\cdot\langle\mathbf{H},\mathring{\mathbf{A}}_{t-1}\rangle\cdot\mathbf{H} + 3\gamma_t^2(\sigma^2+\alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\cdot\mathbf{H} + 2\gamma_t\cdot\mathbf{\Xi},
\end{aligned}
$$

where in the last inequality we applied Assumption 4.6(A). We have completed the proof. □

Lemma C.4. *In the setting of Lemma C.3, it holds that*

$$
\gamma_t\cdot\mathbb{E}\Big[\epsilon_t\cdot\mathtt{diag}\left((\mathbf{w}_{t-1}-\mathbf{w}_*)\mathbf{x}_t^\top + \mathbf{x}_t(\mathbf{w}_{t-1}-\mathbf{w}_*)^\top\right)\Big] \preceq \frac{\gamma_t}{2}\cdot\mathbf{H}\mathring{\mathbf{A}}_{t-1} + 2\gamma_t\cdot\mathbf{\Xi},
$$

where $\mathbf{\Xi}$ is a fixed diagonal matrix and that $\mathtt{tr}(\mathbf{\Xi}) \le \mathtt{OPT}$.

Proof. Define a fixed vector

$$
\mathbf{a} := \mathbb{E}[\epsilon_t\mathbf{H}^{-\frac{1}{2}}\mathbf{x}_t].
$$

Recall that $\mathbf{H}$ is a diagonal matrix, so $\mathbf{H}$ commutes with any diagonal matrix. Then we have

$$
\begin{aligned}
\mathbb{E}\Big[\epsilon_t\cdot\mathtt{diag}\left((\mathbf{w}_{t-1}-\mathbf{w}_*)\mathbf{x}_t^\top + \mathbf{x}_t(\mathbf{w}_{t-1}-\mathbf{w}_*)^\top\right)\Big] &= \mathbb{E}\Big[2\epsilon_t\cdot\mathtt{diag}\left((\mathbf{w}_{t-1}-\mathbf{w}_*)\mathbf{x}_t^\top\right)\Big] \\
&= \mathbb{E}\Big[2\cdot\mathtt{diag}\left(\mathbf{H}^{\frac{1}{2}}(\mathbf{w}_{t-1}-\mathbf{w}_*)\cdot\epsilon_t\mathbf{x}_t^\top\mathbf{H}^{-\frac{1}{2}}\right)\Big] \\
&= \mathbb{E}\Big[2\cdot\mathtt{diag}\left(\mathbf{H}^{\frac{1}{2}}(\mathbf{w}_{t-1}-\mathbf{w}_*)\cdot\mathbf{a}^\top\right)\Big],
\end{aligned}
$$

where in the last equation we take (conditional) expectation over the fresh randomness introduced by ϵ_t and $\mathbf{x}_t$. Now use the fact that: for every two vectors $\mathbf{u},\mathbf{v}$ it holds that

$$
\mathbf{u}\mathbf{v}^\top + \mathbf{v}\mathbf{u}^\top \preceq \mathbf{u}\mathbf{u}^\top + \mathbf{v}\mathbf{v}^\top,
$$

we then obtain

$$
\begin{aligned}
&\mathbb{E}\Big[\epsilon_t\cdot\mathtt{diag}\left((\mathbf{w}_{t-1}-\mathbf{w}_*)\mathbf{x}_t^\top + \mathbf{x}_t(\mathbf{w}_{t-1}-\mathbf{w}_*)^\top\right)\Big] \\
&= \mathbb{E}\Big[2\cdot\mathtt{diag}\left(\frac{1}{\sqrt{2}}\mathbf{H}^{\frac{1}{2}}(\mathbf{w}_{t-1}-\mathbf{w}_*)\cdot\sqrt{2}\mathbf{a}^\top\right)\Big] \\
&\preceq \mathbb{E}\Big[\mathtt{diag}\left(\frac{1}{2}\mathbf{H}^{\frac{1}{2}}(\mathbf{w}_{t-1}-\mathbf{w}_*)(\mathbf{w}_{t-1}-\mathbf{w}_*)^\top\mathbf{H}^{\frac{1}{2}} + 2\mathbf{a}\mathbf{a}^\top\right)\Big] \\
&= \frac{1}{2}\cdot\mathtt{diag}(\mathbf{H}\mathbf{A}_{t-1}) + 2\cdot\mathtt{diag}\left(\mathbf{a}\mathbf{a}^\top\right).
\end{aligned}
$$

Moreover, notice that

$$
\mathbf{a}^\top\mathbf{a} = \mathbb{E}[\epsilon_t\mathbf{x}_t^\top\mathbf{H}^{-\frac{1}{2}}\mathbf{a}] \le \frac{1}{2}\mathbb{E}[\epsilon_t^2 + \mathbf{a}^\top\mathbf{H}^{-\frac{1}{2}}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{H}^{-\frac{1}{2}}\mathbf{a}] = \frac{1}{2}\mathtt{OPT} + \frac{1}{2}\mathbf{a}^\top\mathbf{a},
$$

which implies that $\mathbf{a}^\top\mathbf{a} \le \mathtt{OPT}$, so it holds that

$$
\mathtt{tr}(\mathtt{diag}\left(\mathbf{a}\mathbf{a}^\top\right)) = \mathtt{tr}(\mathbf{a}\mathbf{a}^\top) = \mathbf{a}^\top\mathbf{a} \le \mathtt{OPT}.
$$

We have completed the proof by setting $\mathbf{\Xi} := \mathtt{diag}(\mathbf{a}\mathbf{a}^\top)$ and noting that $\mathtt{diag}(\mathbf{H}\mathbf{A}_{t-1}) = \mathbf{H}\mathring{\mathbf{A}}_{t-1}$. □

C.3. Proof of Theorem 5.2

We will prove the following slightly stronger version.

Theorem C.5 (Risk Bounds for GLM-tron, restated Theorem 5.2)**.** *Suppose that Assumption 5.1, Assumption A.1(A) and Assumption 4.6(A) hold. Let* $\mathbf{w}_N$ *be the output of* (GLM-tron) *with stepsize scheduler* (2)*. Assume that* $N > 100$*. Let* $N_{\mathtt{eff}} := N/\log(N)$*. Then for* $\gamma_0 < 1/(8\alpha(\mathtt{tr}(\mathbf{H})))$*, it holds that*

$$\begin{aligned}\mathbb{E}[\mathcal{R}(\mathbf{w}_N)] \lesssim \mathtt{OPT} + \Big\| \prod_{t=0}^{N-1} \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*)\Big\|_{\mathbf{H}}^2 \\ + \left(\alpha\Big(\mathtt{OPT} + \|\mathbf{w}_*\|_{\mathbf{H}}^2 + \big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\frac{\mathbf{I}_{0:k^*}}{N_{\mathtt{eff}}\gamma} + \mathbf{H}_{k^*:\infty}}\Big) + \sigma^2\right) \cdot \frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}},\end{aligned}$$

where $k^* \geq 0$ *can be any index.*

Proof. First of all, by Lemma C.2, it holds that

$$\begin{aligned}\mathbb{E}[\mathcal{R}(\mathbf{w}_N)] &\leq 2 \cdot \mathbb{E}\|\mathbf{w}_N - \mathbf{w}_*\|_{\mathbf{H}}^2 + 2 \cdot \mathtt{OPT} \\ &= 2 \cdot \langle \mathbf{H}, \mathring{\mathbf{A}}\rangle + 2 \cdot \mathtt{OPT}.\end{aligned}$$

Now consider the recursion of $\mathring{\mathbf{A}}_t$ given in Lemma C.3. Note that $\mathring{\mathbf{A}}_t$ is related to $\mathring{\mathbf{A}}_{t-1}$ through a linear operator, therefore $\mathring{\mathbf{A}}_t$ can be understood as the sum of two iterates, i.e., $\mathring{\mathbf{A}}_t := \mathring{\mathbf{B}}_t + \mathring{\mathbf{C}}_t$, where

$$\begin{cases}\mathring{\mathbf{B}}_t \preceq \big(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\big)\cdot\mathring{\mathbf{B}}_{t-1} + 2\alpha\gamma_t^2\cdot\langle\mathbf{H}, \mathring{\mathbf{B}}_{t-1}\rangle\cdot\mathbf{H}; \\ \mathring{\mathbf{B}}_0 := \mathtt{diag}((\mathbf{w}_0 - \mathbf{w}_*)^{\otimes 2}),\end{cases}$$

and

$$\begin{cases}\mathring{\mathbf{C}}_t \preceq \big(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\big)\cdot\mathring{\mathbf{C}}_{t-1} + 2\alpha\gamma_t^2\cdot\langle\mathbf{H}, \mathring{\mathbf{C}}_{t-1}\rangle\cdot\mathbf{H} + 3\gamma_t^2(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\cdot\mathbf{H} + 2\gamma_t\cdot\boldsymbol{\Xi}; \\ \mathring{\mathbf{C}}_0 := 0.\end{cases}$$

Then we have

$$\mathbb{E}[\mathcal{R}(\mathbf{w}_N)] \leq 2\cdot\langle\mathbf{H}, \mathring{\mathbf{B}}\rangle + 2\cdot\langle\mathbf{H}, \mathring{\mathbf{C}}\rangle + 2\cdot\mathtt{OPT}.$$

Bounding the Bias Error $\langle\mathbf{H}, \mathring{\mathbf{B}}\rangle$. Note that $\mathring{\mathbf{B}}_t$ is exactly the diagonal of the bias iterate in Wu et al. (2022a;b), ignoring a difference in constant factors in the stepsizes. So by the proof of the bias part of Corollary 3.3 in Wu et al. (2022b), we have

$$\langle\mathbf{H}, \mathring{\mathbf{B}}\rangle \lesssim \Big\| \prod_{t=1}^{N}\Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big)(\mathbf{w}_0 - \mathbf{w}_*)\Big\|_{\mathbf{H}}^2 + \alpha\cdot\big\|\mathbf{w}_0 - \mathbf{w}_*\big\|^2_{\frac{\mathbf{I}_{0:k^*}}{N_{\mathtt{eff}}\gamma} + \mathbf{H}_{k^*:\infty}}\cdot\frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}}.$$

Bounding the Variance Error $\langle\mathbf{H}, \mathring{\mathbf{C}}\rangle$. However $\mathring{\mathbf{C}}_t$ is slightly different from the variance iterate in Wu et al. (2022a;b), as the noise structure is different due to the appearance of $\boldsymbol{\Xi}$. But a similar analysis idea applies here.

We first derive a crude upper bound on $\mathring{\mathbf{C}}_t$ in Lemma C.6:

$$\mathring{\mathbf{C}}_t \preceq \rho\gamma\cdot\mathbf{I} + 4\cdot\mathbf{H}^{-1}\boldsymbol{\Xi}, \quad \text{where } \rho := \frac{16\alpha\mathtt{OPT} + 6(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)}{1 - 4\gamma\alpha\,\mathtt{tr}(\mathbf{H})}, \quad t \geq 0.$$

Then we establish a sharper bound based on Lemma C.6 as follows:

$$\begin{aligned}\mathring{\mathbf{C}}_t &\preceq \Big(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\Big)\cdot\mathring{\mathbf{C}}_{t-1} + 2\alpha\gamma_t^2\cdot\langle\mathbf{H}, \mathring{\mathbf{C}}_{t-1}\rangle\cdot\mathbf{H} + 3\gamma_t^2(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\cdot\mathbf{H} + 2\gamma_t\cdot\boldsymbol{\Xi} \\ &\preceq \Big(\mathbf{I} - \frac{\gamma_t}{2}\cdot\mathbf{H}\Big)\cdot\mathring{\mathbf{C}}_{t-1} + 2\alpha\gamma_t^2\big(\rho\gamma\,\mathtt{tr}(\mathbf{H}) + 4\mathtt{OPT}\big)\cdot\mathbf{H} + 3\gamma_t^2(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\cdot\mathbf{H} + 2\gamma_t\cdot\boldsymbol{\Xi}\end{aligned}$$

$$= \left(\mathbf{I} - \frac{\gamma_t}{2} \cdot \mathbf{H}\right) \cdot \mathring{\mathbf{C}}_{t-1} + \left(2\alpha\rho\gamma\,\mathtt{tr}(\mathbf{H}) + 8\alpha\mathtt{OPT} + 3(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\right) \cdot \gamma_t^2 \cdot \mathbf{H} + 2\gamma_t \cdot \mathbf{\Xi}$$
$$\preceq \left(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\right)\mathring{\mathbf{C}}_{t-1} + \left(16\alpha\mathtt{OPT} + 6(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\right) \cdot \gamma_t^2 \cdot \mathbf{H} + 2\gamma_t \cdot \mathbf{\Xi},$$

where the second inequality is by Lemma C.6; and in the last inequality we use the assumption that

$$\gamma < \frac{1}{8\alpha\,\mathtt{tr}(\mathbf{H})},$$

so that

$$\rho := \frac{16\alpha\mathtt{OPT} + 6(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)}{1 - 4\gamma\alpha\,\mathtt{tr}(\mathbf{H})} \le 32\alpha\mathtt{OPT} + 12(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2),$$

which together imply

$$2\alpha\rho\gamma\,\mathtt{tr}(\mathbf{H}) \le \frac{\rho}{4} \le 8\alpha\mathtt{OPT} + 3(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2).$$

We then solve the recursion and obtain

$$\mathring{\mathbf{C}}_N \preceq \left(16\alpha\mathtt{OPT} + 6(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\right) \cdot \sum_{t=1}^{N} \gamma_t^2 \prod_{i=t+1}^{N} \left(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\right) \cdot \mathbf{H} + 2\sum_{t=1}^{N} \gamma_t \prod_{i=t+1}^{N} \left(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\right) \cdot \mathbf{\Xi}.$$

Finally we use Lemma C.7 and obtain

$$\mathring{\mathbf{C}}_N \preceq 8\left(16\alpha\mathtt{OPT} + 6(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\right) \cdot \left(\frac{1}{N_{\mathtt{eff}}}\mathbf{H}_{0:k}^{-1} + N_{\mathtt{eff}}\gamma^2\mathbf{H}_{k:\infty}\right) + 32\mathbf{H}^{-1}\mathbf{\Xi}.$$

So it holds that

$$\left\langle \mathbf{H}, \mathring{\mathbf{C}}_N \right\rangle \le 8\left(16\alpha\mathtt{OPT} + 6(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\right) \cdot \frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}} + 32\,\mathtt{tr}(\mathbf{\Xi})$$
$$\le 8\left(16\alpha\mathtt{OPT} + 6(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\right) \cdot \frac{k^* + N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}} + 32\mathtt{OPT}.$$

Putting everything together completes the proof. □

C.4. Some Auxiliary Lemmas

Lemma C.6 (A crude variance upper bound)**.** *Consider a sequence of variance iterates defined as follows:*

$$\begin{cases} \mathring{\mathbf{C}}_t \preceq \mathring{\mathbf{C}}_{t-1} - \frac{\gamma_t}{2} \cdot \mathbf{H}\mathring{\mathbf{C}}_{t-1} + 2\alpha\gamma_t^2 \cdot \langle \mathbf{H}, \mathring{\mathbf{C}}_{t-1}\rangle \cdot \mathbf{H} + 3\gamma_t^2(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2) \cdot \mathbf{H} + 2\gamma_t \cdot \mathbf{\Xi}; \\ \mathring{\mathbf{C}}_0 := 0, \end{cases}$$

where $\mathbf{\Xi}$ is deterministic and $\mathtt{tr}(\mathbf{\Xi}) \le \mathtt{OPT}$. Then for $\gamma < 1/(4\alpha\,\mathtt{tr}(\mathbf{H}))$, it holds that

$$\mathring{\mathbf{C}}_t \preceq \rho\gamma \cdot \mathbf{I} + 4 \cdot \mathbf{H}^{-1}\mathbf{\Xi}, \quad \textit{where } \rho := \frac{16\alpha\mathtt{OPT} + 6(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)}{1 - 4\gamma\alpha\,\mathtt{tr}(\mathbf{H})}, \quad t \ge 0.$$

Proof. We show it by induction. For $t = 0$ the conclusion holds because $\mathring{\mathbf{C}}_0 = 0$. Now suppose that

$$\mathring{\mathbf{C}}_{t-1} \preceq \rho\gamma\mathbf{I} + 4\mathbf{H}^{-1}\mathbf{\Xi},$$

then

$$\langle \mathbf{H},\ \mathring{\mathbf{C}}_{t-1}\rangle \le \rho\gamma\,\mathtt{tr}(\mathbf{H}) + 4\,\mathtt{tr}(\mathbf{\Xi}) \le \rho\gamma\,\mathtt{tr}(\mathbf{H}) + 4\mathtt{OPT}.$$

Then

$$\mathring{\mathbf{C}}_t \preceq \left(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\right)\mathring{\mathbf{C}}_{t-1} + 2\alpha\gamma_t^2 \cdot \langle \mathbf{H}, \mathring{\mathbf{C}}_{t-1}\rangle \cdot \mathbf{H} + 3\gamma_t^2(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2) \cdot \mathbf{H} + 2\gamma_t \cdot \mathbf{\Xi}$$

$$
\begin{aligned}
&\preceq \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big)\big(\rho\gamma\mathbf{I} + 4\mathbf{H}^{-1}\mathbf{\Xi}\big) + 2\alpha\gamma_t^2\big(\rho\gamma\,\mathtt{tr}(\mathbf{H}) + 4\mathtt{OPT}\big)\cdot\mathbf{H} + 3\gamma_t^2(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\cdot\mathbf{H} + 2\gamma_t\cdot\mathbf{\Xi} \\
&= \big(\rho\gamma\mathbf{I} + 4\mathbf{H}^{-1}\mathbf{\Xi}\big) + \gamma_t\mathbf{H}\cdot\Big(-\frac{\rho\gamma}{2} + 2\alpha\gamma_t\big(\rho\,\mathtt{tr}(\mathbf{H}) + 4\mathtt{OPT}\big) + 3\gamma_t(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\Big) \\
&\le \big(\rho\gamma\mathbf{I} + 4\mathbf{H}^{-1}\mathbf{\Xi}\big) + \gamma_t\mathbf{H}\cdot\Big(-\frac{\rho\gamma}{2} + 2\alpha\gamma\big(\rho\,\mathtt{tr}(\mathbf{H}) + 4\mathtt{OPT}\big) + 3\gamma(\sigma^2 + \alpha\|\mathbf{w}_*\|_{\mathbf{H}}^2)\Big) \\
&= \rho\gamma\mathbf{I} + 4\mathbf{H}^{-1}\mathbf{\Xi}.
\end{aligned}
$$

We have completed the proof. □

Lemma C.7 (Some technical bounds)**.** *It holds that*

(A) $\sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big)\cdot\mathbf{H} \preceq 8\cdot\Big(\frac{1}{N_{\mathtt{eff}}}\mathbf{H}_{0:k}^{-1} + N_{\mathtt{eff}}\gamma_0^2\mathbf{H}_{k:\infty}\Big)$.

(B) $\sum_{t=1}^N \gamma_t \prod_{i=t+1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big) \preceq 16\cdot\mathbf{H}^{-1}$.

(C) For $k^* := \max\{k : \lambda_k \ge 1/(\gamma_0 N_{\mathtt{eff}})\}$*, it holds that*

$$
\sum_{t=1}^N \gamma_t^2 \prod_{i=t+1}^N (\mathbf{I} - \gamma_t\mathbf{H})\cdot\mathbf{H} \succeq \frac{1}{400}\cdot\Big(\frac{1}{N_{\mathtt{eff}}}\mathbf{H}_{0:k^*}^{-1} + N_{\mathtt{eff}}\gamma_0^2\mathbf{H}_{k^*:\infty}\Big).
$$

Proof. The first result is from the proof of Theorem 5 in Wu et al. (2022a). The third result is from the proof of Theorem 7 in Wu et al. (2022a). The second result can be proved in a similar manner. By definition, we have

$$
\begin{aligned}
\sum_{t=1}^N \gamma_t \prod_{i=t+1}^N \Big(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\Big) &= \sum_{\ell=0}^{L-1}\frac{\gamma}{2^\ell}\cdot\sum_{i=1}^{N_{\mathtt{eff}}}\Big(\mathbf{I} - \frac{\gamma}{2^{\ell+1}}\mathbf{H}\Big)^{N_{\mathtt{eff}}-i}\cdot\prod_{j=\ell+1}^{L-1}\Big(\mathbf{I} - \frac{\gamma}{2^{j+1}}\mathbf{H}\Big)^{N_{\mathtt{eff}}} \\
&= 2\mathbf{H}^{-1}\cdot\sum_{\ell=0}^{L-1}\Big(\mathbf{I} - \Big(\mathbf{I} - \frac{\gamma}{2^{\ell+1}}\mathbf{H}\Big)^{N_{\mathtt{eff}}}\Big)\cdot\prod_{j=\ell+1}^{L-1}\Big(\mathbf{I} - \frac{\gamma}{2^{j+1}}\mathbf{H}\Big)^{N_{\mathtt{eff}}} \\
&\preceq 2\mathbf{H}^{-1}\cdot\sum_{\ell=0}^{L-1}\Big(N_{\mathtt{eff}}\cdot\frac{\gamma}{2^{\ell+1}}\mathbf{H}\Big)\cdot\prod_{j=\ell+1}^{L-1}\Big(\mathbf{I} - \frac{\gamma}{2^{j+1}}\mathbf{H}\Big)^{N_{\mathtt{eff}}} \\
&=: 2N_{\mathtt{eff}}\mathbf{H}^{-1}\cdot f(\gamma\mathbf{H}),
\end{aligned}
$$

where

$$
f(x) := \sum_{\ell=0}^{L-1}\frac{x}{2^{\ell+1}}\cdot\prod_{j=\ell+1}^{L-1}\Big(1 - \frac{x}{2^{j+1}}\Big)^{N_{\mathtt{eff}}}, \quad 0 < x < 1.
$$

We then upper bound $f(x)$ as follows:

- For $x \in (0, 4/N_{\mathtt{eff}})$ it holds that
$$
f(x) \le \sum_{\ell=0}^{L-1}\frac{x}{2^{\ell+1}} \le x \le \frac{4}{N_{\mathtt{eff}}}.
$$
- As for $x \in [4/N_{\mathtt{eff}}, 1]$, there is an
$$
\ell^* := \lfloor\log(N_{\mathtt{eff}}x)\rfloor - 2 \in [0,\ L-1),
$$
such that
$$
2^{\ell^*+2}/N_{\mathtt{eff}} \le x < 2^{\ell^*+3}/N_{\mathtt{eff}}.
$$

by which and the definition of $f(x)$ we obtain:

$$\begin{aligned}
f(x) &= \sum_{\ell=0}^{\ell^*} \frac{x}{2^{\ell+1}} \cdot \prod_{j=\ell+1}^{L-1} \left(1 - \frac{x}{2^{j+1}}\right)^{N_{\texttt{eff}}} + \sum_{\ell=\ell^*+1}^{L-1} \frac{x}{2^{\ell+1}} \cdot \prod_{j=\ell+1}^{L-1} \left(1 - \frac{x}{2^{j+1}}\right)^{N_{\texttt{eff}}} \\
&\le \sum_{\ell=0}^{\ell^*} \frac{x}{2^{\ell+1}} \cdot \left(1 - \frac{x}{2^{\ell+2}}\right)^{N_{\texttt{eff}}} + \sum_{\ell=\ell^*+1}^{L-1} \frac{x}{2^{\ell+1}} \cdot 1 \\
&\le \sum_{\ell=0}^{\ell^*} \frac{2^{\ell^*-\ell+2}}{N_{\texttt{eff}}} \cdot \left(1 - \frac{2^{\ell^*-\ell}}{N_{\texttt{eff}}}\right)^{N_{\texttt{eff}}} + \sum_{\ell=\ell^*+1}^{L-1} \frac{2^{\ell^*-\ell+2}}{N_{\texttt{eff}}} \\
&\le \frac{4}{N_{\texttt{eff}}} \cdot \sum_{\ell=0}^{\ell^*} 2^{\ell^*-\ell} \cdot e^{-2^{\ell^*-\ell}} + \frac{4}{N_{\texttt{eff}}} \\
&\le \frac{4}{N_{\texttt{eff}}} \cdot 1 + \frac{4}{N_{\texttt{eff}}} = \frac{8}{N_{\texttt{eff}}}.
\end{aligned}$$

In sum we have shown $f(x) \le 8/N_{\texttt{eff}}$ for $x \in (0, 1)$. Therefore

$$\sum_{t=1}^{N} \gamma_t \prod_{i=t+1}^{N} \left(\mathbf{I} - \frac{\gamma_t}{2}\mathbf{H}\right) = 2N_{\texttt{eff}}\mathbf{H}^{-1} \cdot f(\gamma\mathbf{H}) \preceq 2N_{\texttt{eff}}\mathbf{H}^{-1} \cdot \frac{8}{N_{\texttt{eff}}} = 16\mathbf{H}^{-1}.$$

We have completed the proof. □

C.5. Proof of Corollary 5.3

Proof of Corollary 5.3. According to the stepsize scheduler (2) and the assumptions, we have that

$$\begin{aligned}
\mathbb{E}\mathcal{R}(\mathbf{w}_N) &\lesssim \texttt{OPT} + \|e^{-0.5N_{\texttt{eff}}\gamma_0\mathbf{H}}\mathbf{H}(\mathbf{w}_0 - \mathbf{w}_*)\|_2^2 + \frac{k^* + N_{\texttt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\texttt{eff}}} \\
&\lesssim \texttt{OPT} + \frac{1}{N_{\texttt{eff}}\gamma_0} + \frac{k^* + N_{\texttt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\texttt{eff}}},
\end{aligned}$$

where k^* can be arbitrary. We then simply choose $k^* = d$, and $\gamma_0 \eqsim 1/\texttt{tr}(\mathbf{H})$, then

$$\mathbb{E}\mathcal{R}(\mathbf{w}_N) \lesssim \texttt{OPT} + \frac{\texttt{tr}(\mathbf{H})}{N_{\texttt{eff}}} + \frac{d}{N_{\texttt{eff}}} \lesssim \texttt{OPT} + \frac{d}{N_{\texttt{eff}}},$$

where we use that $\lambda_1 \lesssim 1$. □

D. GLM-tron versus SGD

In this section, we compare GLM-tron and SGD in learning well-specified ReLU regression with symmetric Bernoulli data. We assume that Assumption 4.1 and Assumption 4.4 hold in this part.

Notations. In this section, we assume that $\mathbf{H}$ is diagonal. For $\mathbf{A}_t := \mathbb{E}(\mathbf{w}_t - \mathbf{w}_*)^2$, we use $\mathring{\mathbf{A}}_t$ to refer to the diagonal of $\mathbf{A}_t$. For simplicity, we will use

$$\epsilon_t := y_t - \texttt{ReLU}(\mathbf{w}_*^\top \mathbf{x}_t)$$

to refer to the additive noise in this section.

D.1. Proof of Theorem 6.1

Proof of Theorem 6.1. Consider (SGD).

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \gamma_t\big(\texttt{ReLU}(\mathbf{x}_t^\top \mathbf{w}_{t-1}) - \texttt{ReLU}(\mathbf{x}_t^\top \mathbf{w}_*) - \epsilon_t\big) \cdot \mathbf{x}_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]$$

$$
\begin{aligned}
&= \mathbf{w}_{t-1} - \gamma_t\big(\mathbf{x}_t^\top \mathbf{w}_{t-1}\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] - \mathbf{x}_t^\top \mathbf{w}_*\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_* > 0] - \epsilon_t\big)\cdot \mathbf{x}_t\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \\
&= \mathbf{w}_{t-1} - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\mathbf{w}_{t-1} + \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* > 0]\mathbf{w}_* \\
&\quad + \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\epsilon_t \mathbf{x}_t \\
&= \mathbf{w}_{t-1} - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0](\mathbf{w}_{t-1} - \mathbf{w}_*) - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\mathbf{w}_* \\
&\quad + \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\epsilon_t \mathbf{x}_t,
\end{aligned}
$$

which implies that

$$
\begin{aligned}
\mathbf{w}_t - \mathbf{w}_* = \big(\mathbf{I} - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big)(\mathbf{w}_{t-1} - \mathbf{w}_*) \\
- \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\mathbf{w}_* + \gamma_t \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\epsilon_t \mathbf{x}_t.
\end{aligned}
$$

Let us compute the expected outer product:

$$
\begin{aligned}
&\mathbb{E}\big(\mathbf{w}_t - \mathbf{w}_*\big)^{\otimes 2} \\
&= \mathbb{E}\underbrace{\Big(\big(\mathbf{I} - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big)(\mathbf{w}_{t-1} - \mathbf{w}_*)\Big)^{\otimes 2}}_{\texttt{quadratic term 1}} \\
&\quad + \gamma_t^2 \cdot \mathbb{E}\underbrace{\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\cdot (\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t^{\otimes 2}\big)}_{\texttt{quadratic term 2}} \\
&\quad - \gamma_t \cdot \mathbb{E}\underbrace{\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\mathbf{x}_t^\top \mathbf{w}_* \cdot \big(\mathbf{I} - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big)(\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{x}_t^\top\Big)}_{\texttt{crossing term 1}} \\
&\quad - \gamma_t \cdot \mathbb{E}\underbrace{\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t(\mathbf{w}_{t-1} - \mathbf{w}_*)^\top \big(\mathbf{I} - \gamma_t \mathbf{x}_t\mathbf{x}_t^\top \mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\big)\Big)}_{\texttt{crossing term 2}} \\
&\quad + \gamma_t^2 \cdot \mathbb{E}\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0]\epsilon_t^2 \cdot \mathbf{x}_t^{\otimes 2}\big),
\end{aligned}
\tag{25}
$$

where the crossing terms involving ϵ has zero expectation because $\mathbb{E}[\epsilon_t|\mathbf{x}_t] = 0$.

Now we use Assumption 4.4 and compute each part in (25). Notice that under Assumption 4.4, $\mathbf{x}_t \in \{\pm \mathbf{e}_i\}_{i\geq 1}$, then one can verify that

$$
\text{for every } \mathbf{u} \in \mathbb{H},\ \mathtt{diag}(\mathbf{u}\mathbf{x}_t^\top) = \mathtt{diag}(\mathbf{x}_t\mathbf{u}^\top) = \mathbf{x}_t^\top \mathbf{u}\cdot \mathbf{x}_t\mathbf{x}_t^\top. \tag{26}
$$

By (26) we see that

$$
\begin{aligned}
&\mathtt{diag}\left(\gamma_t^2 \cdot \mathbb{E}(\texttt{quadratic term 2}) - \gamma_t \cdot \mathbb{E}(\texttt{crossing term 1}) - \gamma_t \cdot \mathbb{E}(\texttt{crossing term 2})\right) \\
&= \gamma_t^2 \cdot \mathbb{E}\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\cdot (\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t\mathbf{x}_t^\top\big) \\
&\quad - 2\gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\mathbf{x}_t^\top \mathbf{w}_* \cdot \mathtt{diag}\left((\mathbf{w}_{t-1} - \mathbf{w}_*)\mathbf{x}_t^\top\right)\Big) \\
&\quad + 2\gamma_t^2 \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top(\mathbf{w}_{t-1} - \mathbf{w}_*)\cdot \mathbf{x}_t\mathbf{x}_t^\top\Big) \\
&= \gamma_t^2 \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\cdot (\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t\mathbf{x}_t^\top\Big) \\
&\quad + (2\gamma_t - 2\gamma_t^2) \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]\mathbf{x}_t^\top \mathbf{w}_* \cdot \mathbf{x}_t^\top(\mathbf{w}_* - \mathbf{w}_{t-1})\cdot \mathbf{x}_t\mathbf{x}_t^\top\Big),
\end{aligned}
$$

where in the last equality we use (26). Define

$$
F(\mathbf{w}) := \mathbb{E}_{\mathbf{x}}\Big(\mathbb{1}[\mathbf{x}^\top \mathbf{w} > 0, \mathbf{x}^\top \mathbf{w}_* < 0]\mathbf{x}^\top \mathbf{w}_* \cdot \mathbf{x}^\top(\mathbf{w}_* - \mathbf{w})\cdot \mathbf{x}\mathbf{x}^\top\Big),
$$

where the expectation is only taken with respect to the randomness of $\mathbf{x}$. Then by the property of the indicator function, $\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0]$, we observe that

$$0 \preceq \mathbb{E}_{\mathbf{x}_t}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot (\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \preceq F(\mathbf{w}_{t-1}).$$

So when $0 < \gamma_t < 1$ it holds that

$$\begin{aligned}
&\mathtt{diag}\left(\gamma_t^2 \cdot \mathbb{E}(\mathtt{quadratic\ term}) - \gamma_t \cdot \mathbb{E}(\mathtt{crossing\ term\ 1}) - \gamma_t \cdot \mathbb{E}(\mathtt{crossing\ term\ 2})\right) \\
&= \gamma_t^2 \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0, \mathbf{x}_t^\top \mathbf{w}_* < 0] \cdot (\mathbf{x}_t^\top \mathbf{w}_*)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) + (2\gamma_t - 2\gamma_t^2) \cdot \mathbb{E}[F(\mathbf{w}_{t-1})] \\
&\begin{cases} \leq (2\gamma_t - \gamma_t^2) \cdot \mathbb{E}[F(\mathbf{w}_{t-1})] \leq 2\gamma_t \cdot \mathbb{E}[F(\mathbf{w}_{t-1})]; \\ \geq (2\gamma_t - 2\gamma_t^2) \cdot \mathbb{E}[F(\mathbf{w}_{t-1})] = 2\gamma_t(1-\gamma_t) \cdot \mathbb{E}[F(\mathbf{w}_{t-1})]. \end{cases}
\end{aligned} \tag{27}$$

Similarly, we calculate the diagonal of the expectation of $(\mathtt{quadratic\ term\ 1})$ in (25):

$$\begin{aligned}
&\mathtt{diag}\big(\mathbb{E}(\mathtt{quadratic\ term\ 1})\big) \\
&= \mathtt{diag}\big(\mathbb{E}(\mathbf{w}_{t-1} - \mathbf{w}_*)^{\otimes 2}\big) - 2\gamma_t \cdot \mathtt{diag}\left(\mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*) \cdot \mathbf{x}_t (\mathbf{w}_{t-1} - \mathbf{w}_*)^\top\Big)\right) \\
&\quad + \gamma_t^2 \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \big(\mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*)\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&= \mathring{\mathbf{A}}_{t-1} + (\gamma_t^2 - 2\gamma_t) \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \big(\mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*)\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) \\
&\begin{cases} \preceq \mathring{\mathbf{A}}_{t-1} - \gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \big(\mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*)\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big); \\ \succeq \mathring{\mathbf{A}}_{t-1} - 2\gamma_t \cdot \mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \big(\mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*)\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big), \end{cases}
\end{aligned}$$

where in the second equality we use (26) and in the inequality we use $0 < \gamma_t < 1$. We now use Assumption 4.4 to obtain that

$$\mathbb{E}\Big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \cdot \big(\mathbf{x}_t^\top (\mathbf{w}_{t-1} - \mathbf{w}_*)\big)^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\Big) = \sum_i \frac{\lambda_i}{2} \cdot \mathbb{E}\big(\mathbf{w}_{t-1}[i] - \mathbf{w}_*[i]\big)^2 \cdot \mathbf{e}_i = \frac{1}{2} \cdot \mathbf{H}\mathring{\mathbf{A}}_{t-1}.$$

So we have

$$\mathtt{diag}\big(\mathbb{E}(\mathtt{quadratic\ term\ 1})\big) \begin{cases} \preceq (\mathbf{I} - \frac{\gamma_t}{2} \cdot \mathbf{H}) \cdot \mathring{\mathbf{A}}_{t-1}; \\ \succeq (\mathbf{I} - \gamma_t \cdot \mathbf{H}) \cdot \mathring{\mathbf{A}}_{t-1}. \end{cases} \tag{28}$$

Bring (27), (28) and that

$$\mathbb{E}\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \epsilon_t^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\big) = \frac{\sigma^2}{2} \cdot \mathbf{H}$$

into (25) we obtain

$$\begin{aligned}
\mathring{\mathbf{A}}_t &= \mathtt{diag}\big(\mathbb{E}(\mathbf{w}_t - \mathbf{w}_*)^{\odot 2}\big) \\
&= \mathtt{diag}\big(\mathbb{E}(\mathtt{quadratic\ term\ 1})\big) + \mathbb{E}\big(\mathbb{1}[\mathbf{x}_t^\top \mathbf{w}_{t-1} > 0] \epsilon_t^2 \cdot \mathbf{x}_t \mathbf{x}_t^\top\big) \\
&\quad + \mathtt{diag}\left(\gamma_t^2 \cdot \mathbb{E}(\mathtt{quadratic\ term\ 2}) - \gamma_t \cdot \mathbb{E}(\mathtt{crossing\ term\ 1}) - \gamma_t \cdot \mathbb{E}(\mathtt{crossing\ term\ 2})\right) \\
&\begin{cases} \preceq (\mathbf{I} - \frac{\gamma_t}{2} \cdot \mathbf{H}) \cdot \mathring{\mathbf{A}}_{t-1} + \frac{\gamma_t^2 \sigma^2}{2} \mathbf{H} + 2\gamma_t \cdot \mathbb{E}[F(\mathbf{w}_{t-1})], \\ \succeq (\mathbf{I} - \gamma_t \cdot \mathbf{H}) \cdot \mathring{\mathbf{A}}_{t-1} + \frac{\gamma_t^2 \sigma^2}{2} \mathbf{H} + 2\gamma_t(1-\gamma_t) \cdot \mathbb{E}[F(\mathbf{w}_{t-1})]. \end{cases}
\end{aligned}$$

By solving the above recursion we obtain

$$\mathring{\mathbf{A}}_N \preceq \prod_{t=1}^N (\mathbf{I} - \frac{\gamma_t}{2} \cdot \mathbf{H}) \cdot \mathring{\mathbf{A}}_0 + \frac{\sigma^2}{2} \sum_{t=1}^N \gamma_t^2 \prod_{k=t+1}^N (1 - \frac{\gamma_k}{2}\mathbf{H})\mathbf{H} + 2\sum_{t=1}^N \gamma_t \prod_{k=t+1}^N (1 - \frac{\gamma_k}{2}\mathbf{H})\mathbb{E}[F(\mathbf{w}_t)],$$

$$\mathring{\mathbf{A}}_N \succeq \prod_{t=1}^{N}(\mathbf{I}-\gamma_t\cdot\mathbf{H})\cdot\mathring{\mathbf{A}}_0+\frac{\sigma^2}{2}\sum_{t=1}^{N}\gamma_t^2\prod_{k=t+1}^{N}(1-\gamma_k\mathbf{H})\mathbf{H}+2\sum_{t=1}^{N}\gamma_t(1-\gamma_t)\prod_{k=t+1}^{N}(1-\gamma_k\mathbf{H})\mathbb{E}[F(\mathbf{w}_t)].$$

The remain efforts are taking inner product with $\mathbf{H}$ and using Lemma C.7 to show that

$$\begin{aligned}\langle\mathbf{H},\mathring{\mathbf{A}}_N\rangle &\lesssim \Big\langle\mathbf{H},\prod_{t=1}^{N}\Big(\mathbf{I}-\frac{\gamma_t}{2}\cdot\mathbf{H}\Big)\cdot\mathring{\mathbf{A}}_0\Big\rangle+\sigma^2\Big\langle\mathbf{H},\sum_{t=1}^{N}\gamma_t^2\prod_{k=t+1}^{N}\Big(1-\frac{\gamma_k}{2}\mathbf{H}\Big)\mathbf{H}\Big\rangle\\&\quad+\Big\langle\mathbf{H},\sum_{t=1}^{N}\gamma_t\prod_{k=t+1}^{N}\Big(1-\frac{\gamma_k}{2}\mathbf{H}\Big)\mathbb{E}[F(\mathbf{w}_t)]\Big\rangle\\&\lesssim \big\|\mathbf{w}_0-\mathbf{w}_*\big\|^2_{\prod_{t=1}^{N}(\mathbf{I}-\frac{\gamma_t}{2}\mathbf{H})\mathbf{H}}+\sigma^2\cdot\frac{k^*+N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}}\\&\quad+\sum_{t=1}^{N}\Big\langle\gamma_t\prod_{k=t+1}^{N}\Big(1-\frac{\gamma_k}{2}\mathbf{H}\Big)\mathbf{H},\mathbb{E}[F(\mathbf{w}_t)]\Big\rangle,\end{aligned}$$

and

$$\begin{aligned}\langle\mathbf{H},\mathring{\mathbf{A}}_N\rangle &\gtrsim \Big\langle\mathbf{H},\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\cdot\mathring{\mathbf{A}}_0\Big\rangle+\sigma^2\sum_{t=1}^{N}\gamma_t^2\prod_{k=t+1}^{N}(1-\gamma_k\mathbf{H})\mathbf{H}\\&\quad+\sum_{t=1}^{N}\gamma_t(1-\gamma_t)\prod_{k=t+1}^{N}(1-\gamma_k\mathbf{H})\mathbb{E}[F(\mathbf{w}_t)],\\&\gtrsim \big\|\mathbf{w}_0-\mathbf{w}_*\big\|^2_{\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\mathbf{H}}+\sigma^2\cdot\frac{k^*+N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k^*}\lambda_i^2}{N_{\mathtt{eff}}}\\&\quad+\sum_{t=1}^{N}\Big\langle\gamma_t(1-\gamma_t)\prod_{k=t+1}^{N}\big(1-\gamma_k\mathbf{H}\big)\mathbf{H},\mathbb{E}[F(\mathbf{w}_t)]\Big\rangle.\end{aligned}$$

We have completed the proof. □

D.2. Proof of Theorem 6.2

Proof of Theorem 6.2. We now compare the risk upper bound for (GLM-tron) shown in Theorem 4.5 and the risk lower bound for (SGD) shown in Theorem 6.1. Denote $\gamma_0^{\mathrm{sgd}}<1$ as the initial stepsize for (SGD), and

$$\begin{aligned}\mathtt{BIAS}^{\mathrm{sgd}}(\gamma_0)&:=\big\|(\mathbf{w}_0-\mathbf{w}_*)\big\|^2_{\prod_{t=1}^{N}(\mathbf{I}-\gamma_t\mathbf{H})\mathbf{H}},\\\mathtt{VAR}^{\mathrm{sgd}}(\gamma_0)&:=\sigma^2\cdot\frac{\#\{i:\lambda_i\geq\frac{1}{N_{\mathtt{eff}}\gamma_0}\}+N_{\mathtt{eff}}^2\gamma_0^2\sum_{\lambda_i<\frac{1}{N_{\mathtt{eff}}\gamma_0}}\lambda_i^2}{N_{\mathtt{eff}}},\end{aligned}$$

then Theorem 6.1 implies that for every $\gamma_0^{\mathrm{sgd}}<1$,

$$\mathbb{E}\Delta(\mathbf{w}_N^{\mathrm{sgd}})\gtrsim\mathtt{BIAS}(\gamma_0^{\mathrm{sgd}})+\mathtt{VAR}(\gamma_0^{\mathrm{sgd}})+\Psi,$$

where $\Psi\geq 0$. Similarly, denote $\gamma_0^{\mathrm{tron}}<1/2$ as the initial stepsize for (GLM-tron), and

$$\mathtt{BIAS}^{\mathrm{tron}}(\gamma_0):=\big\|(\mathbf{w}_0-\mathbf{w}_*)\big\|^2_{\prod_{t=1}^{N}(\mathbf{I}-\frac{\gamma_t}{2}\mathbf{H})\mathbf{H}},\qquad\mathtt{VAR}^{\mathrm{tron}}(\gamma_0,k):=\sigma^2\cdot\frac{k+N_{\mathtt{eff}}^2\gamma_0^2\sum_{i>k}\lambda_i^2}{N_{\mathtt{eff}}},$$

then Theorem 4.5 implies that for every $\gamma_0^{\mathrm{tron}}<1/2$,

$$\mathbb{E}\Delta(\mathbf{w}_N^{\mathrm{tron}})\lesssim\mathtt{BIAS}(\gamma_0^{\mathrm{tron}})+\mathtt{VAR}(\gamma_0^{\mathrm{tron}},k),$$

where $k\geq 0$ can be an arbitrary index.

We prove the theorem by discussing two cases on whether or not the (SGD) initial stepsize is large or not.

SGD with Small Initial Stepsize. If the initial stepsize for (SGD) is $\gamma_0^{\mathrm{sgd}} < 1/8$, then one can choose an initial stepsize $\gamma_0^{\mathrm{tron}} = 2\gamma_0^{\mathrm{sgd}} < 1/4$ for (GLM-tron). Then we have $\texttt{BIAS}^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}}) = \texttt{BIAS}^{\mathrm{sgd}}(\gamma_0^{\mathrm{sgd}})$. Moreover by choosing $k := \#\{i : \lambda_i \geq 1/(N_{\texttt{eff}}\gamma_0^{\mathrm{sgd}})\}$, we have $\texttt{VAR}^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}}, k) = \texttt{VAR}^{\mathrm{sgd}}(\gamma_0^{\mathrm{sgd}})$. These together imply that

$$\mathbb{E}\Delta(\mathbf{w}_N^{\mathrm{tron}}) \lesssim \mathbb{E}\Delta(\mathbf{w}_N^{\mathrm{tron}})$$

holds if $\gamma_0^{\mathrm{sgd}} < 1/8$.

SGD with Large Initial Stepsize. Now we discuss the case when $\gamma_0^{\mathrm{sgd}} > 1/8$. In this case we choose $\gamma_0^{\mathrm{tron}} = 1/8 < 1/4$.

- If $N_{\texttt{eff}} < \frac{1}{8\lambda_1}$, i.e., $\gamma_0^{\mathrm{sgd}}\lambda_i \leq \gamma_0^{\mathrm{sgd}}\lambda_1 \leq \frac{1}{8N_{\texttt{eff}}}$, which implies that $(\mathbf{I} - \gamma_0^{\mathrm{sgd}}\mathbf{H})^{2N_{\texttt{eff}}} \succeq (1 - \frac{1}{8N_{\texttt{eff}}}\mathbf{I})^{2N_{\texttt{eff}}} \succeq 0.01 \cdot \mathbf{I}$, then

$$\begin{aligned}\texttt{BIAS}^{\mathrm{sgd}}(\gamma_0^{\mathrm{sgd}}) &= \big\|(\mathbf{w}_0 - \mathbf{w}_*)\big\|^2_{\prod_{t=1}^N(\mathbf{I}-\gamma_t\mathbf{H})\mathbf{H}} \geq \big\|(\mathbf{w}_0 - \mathbf{w}_*)\big\|^2_{(\mathbf{I}-\gamma_0^{\mathrm{sgd}}\mathbf{H})^{2N_{\texttt{eff}}}\mathbf{H}} \\ &\geq 0.01 \cdot \big\|(\mathbf{w}_0 - \mathbf{w}_*)\big\|^2_{\mathbf{H}}.\end{aligned}$$

  So we have $\texttt{BIAS}^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}}) \lesssim \texttt{BIAS}^{\mathrm{sgd}}(\gamma_0^{\mathrm{sgd}})$. Similarly we choose $k := \#\{i : \lambda_i \geq 1/(N_{\texttt{eff}}\gamma_0^{\mathrm{sgd}})\}$, we have $\texttt{VAR}^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}}, k) \lesssim \texttt{VAR}^{\mathrm{sgd}}(\gamma_0^{\mathrm{sgd}})$, because $\gamma_0^{\mathrm{tron}} = 1/8 \leq \gamma_0^{\mathrm{sgd}}$.

- If $N_{\texttt{eff}} > \frac{1}{8\lambda_1}$, then it holds that $\lambda_1 > 1/(8N_{\texttt{eff}}) = 1/(N_{\texttt{eff}}\gamma_0^{\mathrm{tron}})$ then we must have

$$\texttt{VAR}^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}}, k) = \sigma^2 \cdot \frac{k + N_{\texttt{eff}}^2(\gamma_0^{\mathrm{tron}})^2\sum_{i>k}\lambda_i^2}{N_{\texttt{eff}}} \geq \frac{\sigma^2}{N_{\texttt{eff}}},$$

  for every $k \geq 0$. But we also have $\big(\mathbf{I} - \frac{\gamma_0^{\mathrm{tron}}}{2}\mathbf{H}\big)^{N_{\texttt{eff}}} \leq \frac{2}{\gamma_0^{\mathrm{tron}}N_{\texttt{eff}}}\mathbf{H}^{-1} = \frac{16}{N_{\texttt{eff}}}\mathbf{H}^{-1}$, which implies that

$$\begin{aligned}\texttt{BIAS}^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}}) &= \big\|(\mathbf{w}_0 - \mathbf{w}_*)\big\|^2_{\prod_{t=1}^N(\mathbf{I}-\frac{\gamma_t}{2}\mathbf{H})\mathbf{H}} \leq \big\|(\mathbf{w}_0 - \mathbf{w}_*)\big\|^2_{\big(\mathbf{I}-\frac{\gamma_0^{\mathrm{tron}}}{2}\mathbf{H}\big)^{N_{\texttt{eff}}}\mathbf{H}} \\ &\leq \frac{16}{N_{\texttt{eff}}} \cdot \|\mathbf{w}_0 - \mathbf{w}_*\|_2^2 \lesssim \texttt{VAR}^{\mathrm{tron}}(\gamma_0^{\mathrm{tron}}, k),\end{aligned}$$

  for every $k \geq 0$. Therefore we have

$$\mathbb{E}\Delta(\mathbf{w}_N^{\mathrm{tron}}) \lesssim \texttt{BIAS}(\gamma_0^{\mathrm{tron}}) + \texttt{VAR}(\gamma_0^{\mathrm{tron}}, k) \lesssim \texttt{VAR}(\gamma_0^{\mathrm{tron}}, k) \lesssim \texttt{VAR}(\gamma_0^{\mathrm{sgd}}) \lesssim \mathbb{E}\Delta(\mathbf{w}_N^{\mathrm{sgd}}),$$

  where the third inequality is by choosing $k := \#\{i : \lambda_i \geq 1/(N_{\texttt{eff}}\gamma_0^{\mathrm{sgd}})\}$ and the fact that $\gamma_0^{\mathrm{tron}} = 1/8 \leq \gamma_0^{\mathrm{sgd}}$.

Putting everything together, we have completed the proof. □

D.3. Proof of Theorem 6.3

Proof of Theorem 6.3. We only need to show that for SGD (SGD) it holds that

$$\mathbb{E}_{\mathbf{w}_*}\mathbb{E}_{\mathrm{alg}}\mathcal{R}(\mathbf{w}_N) \geq \frac{1}{2} \cdot \mathcal{R}(\mathbf{w}_0).$$

According to the SGD iterate (SGD) and the noiseless assumption ($\sigma^2 = 0$), we can write the gradient as

$$\begin{aligned}\mathbf{g}_t &:= \big(\texttt{ReLU}(\mathbf{x}_t^\top\mathbf{w}_{t-1}) - y_t\big) \cdot \mathbb{1}[\mathbf{x}_t^\top\mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t \\ &= \big(\mathbf{x}_t^\top\mathbf{w}_{t-1}\mathbb{1}[\mathbf{x}_t^\top\mathbf{w}_{t-1} > 0] - \mathbf{x}_t^\top\mathbf{w}_*\mathbb{1}[\mathbf{x}_t^\top\mathbf{w}_* > 0]\big) \cdot \mathbb{1}[\mathbf{x}_t^\top\mathbf{w}_{t-1} > 0] \cdot \mathbf{x}_t.\end{aligned}$$

Let us focus on the i-th component from now on. Without loss of generality, we assume for now that $\mathbf{w}_*[i] \geq 0$. We use Assumption 4.4 to obtain

$$\mathbf{g}_t[i] = \begin{cases} 0, & \mathbf{x}_t \notin \{\pm\mathbf{e}_i\} \text{ or } \mathbf{x}_t^\top\mathbf{w}_{t-1} \leq 0; \\ \mathbf{w}_{t-1}[i] - \mathbf{w}_*[i], & \mathbf{w}_{t-1}[i] \geq 0 \text{ and } \mathbf{x}_t = \mathbf{e}_i; \\ \mathbf{w}[i], & \mathbf{w}_{t-1}[i] < 0 \text{ and } \mathbf{x}_t = -\mathbf{e}_i. \end{cases}$$

So the i-th component of the SGD iterate is updated by

$$\mathbf{w}_t[i] = \begin{cases} \mathbf{w}_{t-1}[i], & \mathbf{x}_t \notin \{\pm \mathbf{e}_i\} \text{ or } \mathbf{x}_t^\top \mathbf{w}_{t-1} \le 0; \\ \mathbf{w}_{t-1}[i] - \gamma(\mathbf{w}_t[i] - \mathbf{w}_*[i]), & \mathbf{w}_{t-1}[i] \ge 0 \text{ and } \mathbf{x}_t = \mathbf{e}_i; \\ \mathbf{w}_{t-1}[i] - \gamma \mathbf{w}_{t-1}[i], & \mathbf{w}_{t-1}[i] < 0 \text{ and } \mathbf{x}_t = -\mathbf{e}_i. \end{cases}$$

Recall that $\gamma < 1/\,\mathtt{tr}(\mathbf{H}) = 1$. We next show that: if $\mathbf{w}_0[i] \le 0$, then $\mathbf{w}_t[i] \le 0$ for all t. This is done by induction: when $\mathbf{w}_{t-1}[i] \le 0$, two possible updates happen: $\mathbf{w}_t[i] = \mathbf{w}_{t-1}[i] \le 0$ or $\mathbf{w}_t[i] = (1-\gamma)\mathbf{w}_{t-1}[i] \le 0$. In both cases, it holds that $\mathbf{w}_t[i] \le 0$. We have completed the induction. Moreover, recall that we assume $\mathbf{w}_*[i] \ge 0$, so if $\mathbf{w}_t[i] \le 0$, it holds that

$$(\mathbf{w}_t[i] - \mathbf{w}_*[i])^2 \ge (\mathbf{w}_*[i])^2, \quad t \ge 0.$$

Similarly we can prove that when $\mathbf{w}_*[i] \le 0$, if $\mathbf{w}_t[i] \ge 0$, it holds that

$$(\mathbf{w}_t[i] - \mathbf{w}_*[i])^2 \ge (\mathbf{w}_*[i])^2, \quad t \ge 0.$$

Now recall that $\mathbf{w}_*[i]$ is initialized with a uniformly random sign, so with half probability $\mathbf{w}_*[i]$ and $\mathbf{w}_[i]$ will have different signs. Therefore we have

$$\mathbb{E}_{\mathbf{w}_*}\mathbb{E}_{\mathrm{alg}}\big[(\mathbf{w}_t[i] - \mathbf{w}_*[i])^2\big] \ge \mathbb{E}_{\mathbf{w}_*}\mathbb{E}_{\mathrm{alg}}\big[(\mathbf{w}_t[i] - \mathbf{w}_*[i])^2 \cdot \mathbb{1}[\mathbf{w}_0[i] \cdot \mathbf{w}_*[i] \le 0]\big] \ge \frac{1}{2} \cdot (\mathbf{w}_*[i])^2, \quad t \ge 0.$$

Therefore we have shown that for every $t \ge 0$,

$$\mathbb{E}_{\mathbf{w}_*}\mathbb{E}_{\mathrm{alg}}\mathcal{R}(\mathbf{w}_t) = \mathbb{E}_{\mathbf{w}_*} \sum_i \lambda_i (\mathbf{w}_t[i] - \mathbf{w}_*[i])^2 \ge \frac{1}{2} \sum_i \lambda_i (\mathbf{w}_*[i])^2 = \frac{1}{2} \cdot \|\mathbf{w}_*\|_{\mathbf{H}}^2 \ge \frac{1}{2} \cdot \mathcal{R}(0),$$

where the last inequality is due to Lemma 4.2.

□

E. Additional Experiments

Figures 2, 3 and 4 show the additional experimental results, where we compare the excess risk achieved by (GLM-tron) and (SGD) on Bernoulli and Gaussian data. Figure 2 provides the experimental results on Bernoulli data in the noiseless setting. We can clearly see that SGD finally reaches a point with constant risk, while GLM-tron achieves nearly zero excess risk. This backs up our Theorem 6.3. Figures 3 and 4 visualize the learning performance of GLM-tron and SGD on Gaussian data. We can also see that GLM-tron achieves smaller excess risk than SGD, which also supports our claim that GLM-tron is preferable to SGD for high-dimensional ReLU regression.

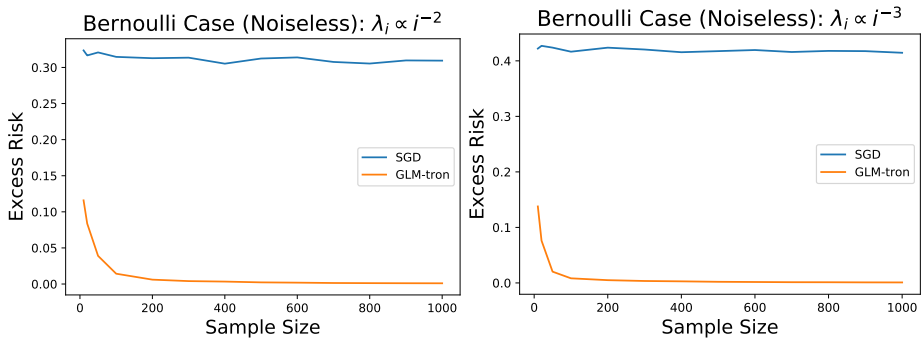


Figure 2. Excess risk comparison between SGD and GLM-tron on Bernoulli Distribution. The problem dimension $d = 1024$, and the regression model is well-specified without noise. We consider two different symmetric Bernoulli distributions and set true model parameters $\mathbf{w}[i]^* = i^{-1}$. For each algorithm and each sample size, we do a grid search and report the best excess risk achieved by $\gamma_0 \in \{0.5, 0.25, 0.1, 0.075, 0.05, 0.025, 0.01\}$. The plots are averaged over 20 independent runs.

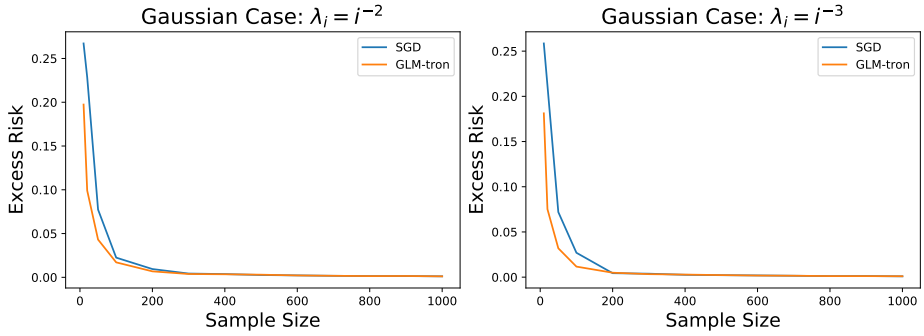


Figure 3. Excess risk comparison between SGD and GLM-tron on Gaussian Distribution. The regression model is well-specified with noise variance $\sigma = 0.1$. Other problem parameters and algorithm designs are the same as those in Figure 2.

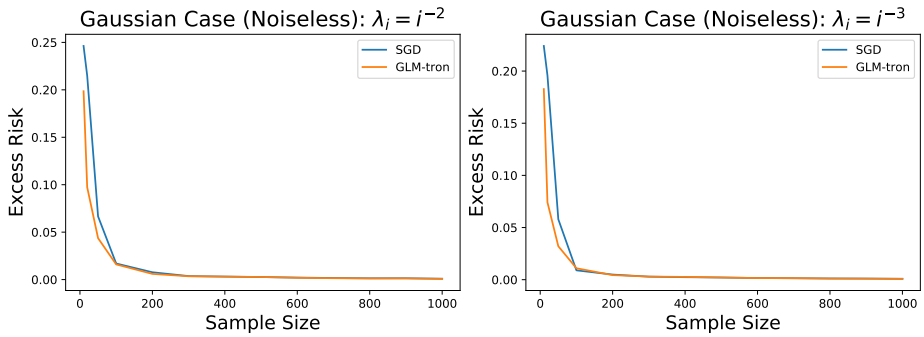


Figure 4. Excess risk comparison between SGD and GLM-tron on Gaussian Distribution. The regression model is well-specified without noise. Other problem and algorithm parameters are the same as those in Figure 3.