Proceedings of Machine Learning Research vol XX:1–31, 2019

Computationally and Statistically Efficient Truncated Regression

Constantinos Daskalakis COSTIS@CSAIL.MIT.EDU
Massachusetts Institute of Technology

Themis Gouleakis TGOULE@MIT.EDU
University of Southern California

Christos Tzamos TZAMOS@WISC.EDU
University of Winscosin-Madison

Manolis Zampetakis MZAMPET@MIT.EDU
Massachusetts Institute of Technology

Abstract

We provide a computationally and statistically efficient estimator for the classical problem of truncated linear regression, where the dependent variable $y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + \varepsilon$ and its corresponding vector of covariates $\boldsymbol{x} \in \mathbb{R}^k$ are only revealed if the dependent variable falls in some subset $S \subseteq \mathbb{R}$; otherwise the existence of the pair $(\boldsymbol{x}, y)$ is hidden. This problem has remained a challenge since the early works of Tobin (1958); Amemiya (1973); Hausman and Wise (1977); Breen et al. (1996), its applications are abundant, and its history dates back even further to the work of Galton, Pearson, Lee, and Fisher Galton (1897); Pearson and Lee (1908); Lee (1914); Fisher (1931). While consistent estimators of the regression coefficients have been identified, the error rates are not well-understood, especially in high-dimensional settings.

Under a "thickness assumption" about the covariance matrix of the covariates in the revealed sample, we provide a computationally efficient estimator for the coefficient vector $\boldsymbol{w}$ from n revealed samples that attains ℓ_2 error $O(\sqrt{k/n})$, recovering the guarantees of least squares in the standard (untruncated) linear regression setting. Our estimator uses Projected Stochastic Gradient Descent (PSGD) on the negative log-likelihood of the truncated sample, and only needs oracle access to the set S, which may otherwise be arbitrary, and in particular may be non-convex. PSGD must be restricted to an appropriately defined convex cone to guarantee that the negative log-likelihood is strongly convex, which in turn is established using concentration of matrices on variables with sub-exponential tails. We perform experiments on simulated data to illustrate the accuracy of our estimator.

As a corollary of our work, we show that SGD provably learns the parameters of single-layer neural networks with noisy Relu activation functions Nair and Hinton (2010); Bengio et al. (2013); Gulcehre et al. (2016), given linearly many, in the number of network parameters, input-output pairs in the realizable setting.

Keywords: linear regression, truncated statistics, truncated regression, stochastic gradient descent

1. Introduction

A central challenge in statistics is estimation from truncated samples. Truncation occurs whenever samples that do not belong in some set S are not observed. For example, a clinical study of obesity will not contain samples with weight smaller than a threshold set by the study. The related notion of censoring is similar except that part of the sample may be observed even if it does not belong

to S. For example, the values that an insurance adjuster observes are right-censored as clients report their loss as equal to the policy limit when their actual loss exceeds the policy limit. In this case, samples below the policy limit are shown, and only the count of the samples that are above the limit is provided. Truncation and censoring have myriad manifestations in business, economics, manufacturing, engineering, quality control, medical and biological sciences, management sciences, social sciences, and all areas of the physical sciences. As such they have received extensive study.

In this paper, we revisit the classical problem of *truncated linear regression*, which has been a challenge since the early works of Tobin (1958); Amemiya (1973); Hausman and Wise (1977); Breen et al. (1996). Like standard linear regression, the dependent variable $y \in \mathbb{R}$ is assumed to satisfy a linear relationship $y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + \varepsilon$ with the vector of covariates $\boldsymbol{x} \in \mathbb{R}^k$, where $\varepsilon \sim \mathcal{N}(0,1)$, and $\boldsymbol{w} \in \mathbb{R}^k$ is some unknown vector of regression coefficients. Unlike standard linear regression, however, neither $\boldsymbol{x}$ nor y are observed, unless the latter belongs to some set $S \subseteq \mathbb{R}$. Given a collection $(\boldsymbol{x}^{(i)}, y^{(i)})_{i=1,\dots,n}$ of samples that survived truncation, the goal is to estimate $\boldsymbol{w}$. In the closely related and easier setting of *censored linear regression*, we are also given the set of covariates resulting in a truncated response.

Applications of truncated and censored linear regression are abundant, as in many cases observations are systematically filtered out during data collection, if the response variable lies below or above certain thresholds. An interesting example of truncated regression is discussed in Hausman and Wise (1977) where the effect of education and intelligence on the earnings of workers in "low level" jobs is studied, based on a data collected by surveying families whose incomes, during the year preceding the experiment, were smaller than one and one-half times the 1967 poverty line.

Truncated and censored linear regression have a long history, dating back to at least Tobin (1958), and the work of Amemiya (1973) and Hausman and Wise (1977). Amemiya (1973) studies censored regression, when the truncation set S is a half-line and shows consistency and asymptotic normality of the maximum likelihood estimator. He also proposes a two-step Newton method to compute a consistent and asymptotically normal estimator. Hausman and Wise study the harder problem of truncated regression also establishing consistency of the maximum likelihood estimator. An overview of existing work on the topic can be found in Breen et al. (1996). While known estimators achieve asymptotic rate of $O_k(\frac{1}{\sqrt{n}})$, at least for sets S that are half-lines, the dependence of $O_k(\cdot)$ on the dimension k is not well-understood. Moreover, while these weaker guarantees can be attained for censored regression, no efficient algorithm is known at all for truncated regression.

Our goal in this work is to obtain computationally and statistically efficient estimators for truncated linear regression. We make no assumptions about the set S that is used for truncation, except that we are given oracle access to this set, namely, given a point $\boldsymbol{x}$ the oracle outputs $\mathbf{1}_{\boldsymbol{x}\in S}$. We also make a couple of necessary assumptions about the covariates of the observable samples:

Assumption I: the first is that the probability, conditionally on $\boldsymbol{x}^{(i)}$, that the response variable $y^{(i)} = \boldsymbol{w}^{\top}\boldsymbol{x}^{(i)} + \varepsilon^{(i)}$ corresponding to a covariate $\boldsymbol{x}^{(i)}$ in our sample is not truncated is lower bounded by some absolute constant, say 1%, with respect to the choice of $\varepsilon^{(i)} \sim \mathcal{N}(0,1)$; this assumption is also necessary as was shown in Daskalakis et al. (2018) for the special case of our problem, pertaining to truncated Gaussian estimation,

Assumption II: the second is the same thickness assumption, also made in the some standard (untruncated) linear regression, that the average $\frac{1}{n}\sum_i \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}$ of the outer-products of the covariates in our sample has some absolute lower bound on its minimum singular value.

These assumptions are further discussed in Section 3, where weaker sufficient conditions, exploiting more structural information about the covariates, are also proposed. See Assumptions 1 and 2. As these are more cumbersome, the reader can just think of Assumptions I and II stated above.

Under Assumptions I and II (or Assumptions 1 and 2), we provide the first time and sample efficient estimation algorithm for truncated linear regression, whose estimation error of the coefficient vector $\boldsymbol{w}$ decays as $O(\sqrt{\frac{k}{n}})$, recovering the error rate of the least-squares estimator in the standard (untruncated) linear regression setting. For a formal statement see Theorem 2. Our algorithm is the first computationally efficient estimator for truncated linear regression. It is also the first, to the best of our knowledge, estimator that can accommodate arbitrary truncation sets S. This, in turn, enables statistical estimation in settings where set S is determined by a complex set of rules, as it happens in many important applications.

We present below a high-level overview of the techniques involved in proving our main result, and discuss further related work. Section 2 provides the necessary preliminaries, while Section 3 presents the truncated linear regresion model and contains a discussion of the assumptions made for our estimation. Section 4 states our main result and provides its proof. Finally, in Section 6, we perform experiments on simulated data to illustrate the accuracy of our method.

Learning Single-Layer Neural Networks with Noisy Activation Functions. Our main result implies as an immediate corollary the learnability, via SGD, of single-layer neural networks with noisy Relu activation functions Nair and Hinton (2010); Bengio et al. (2013); Gulcehre et al. (2016). The noisy Relu activations, considered in these papers for the purposes of improving the stability of gradient descent, are similar to the standard Relu activations, except that noise is added to their inputs before the application of the non-linearity. In particular, if z is the input to a noisy Relu activation, its output is $\max\{0, z+\varepsilon\}$, where $\varepsilon \sim \mathcal{N}(0,1)$. In turn, a single-layer neural network with noisy Relu activations is a random mapping, $f_{\boldsymbol{w}} : \boldsymbol{x} \mapsto \max\{0, \boldsymbol{w}^T\boldsymbol{x}+\varepsilon\}$, where $\varepsilon \sim \mathcal{N}(0,1)$.

We consider the learnability of single-layer neural networks of this type in the realizable setting. In particular, given a neural network $f_{\boldsymbol{w}}$ of the above form, and a sequence of inputs $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, suppose that $y^{(1)}, \ldots, y^{(n)}$ are the (random) outputs of the network on these inputs. Given the collection $(\boldsymbol{x}^{(i)}, y^{(i)})_{i=1}^n$ our goal is to recover $\boldsymbol{w}$. This problem can be trivially reduced to the main learning problem studied in this paper as a special case where: (i) the truncation set is very simple, namely the half open interval $[0, +\infty)$; and (ii) the identities of the inputs $\boldsymbol{x}^{(i)}$ resulting in truncation are also revealed to us, namely we are in a censoring setting rather than a truncation setting. As such, our more general results are directly applicable to this setting. For more information see Section 5.

1.1. Overview of the Techniques.

We present a high-level overview of our time- and statistically-efficient algorithm for truncated linear regression (Theorem 2). Our algorithm, shown in Figure 1, is Projected Stochastic Gradient Descent (PSGD) on the negative log-likelihood of the truncated samples. Notice that we cannot write a closed-form expression for the negative log-likelihood, as the set S is arbitrary and unknown to us. Indeed, we only have oracle access to this set and can thus not write down a formula for the measure of S under different estimates of the coefficient vector $\boldsymbol{w}$. While we cannot write a closed-form expression for the negative log-likelihood, we still show that it is convex with respect to $\boldsymbol{w}$ for arbitrary truncation sets $S \subseteq \mathbb{R}$.

To effectively run Gradient Descent on the negative log-likelihood, we need however to ensure that Gradient Descent remains within a region where it is *strongly convex*. To accomplish this we

define a convex set of vectors $(\boldsymbol{w})$ in Definition 4 and show in Theorem 9 that the negative log-likelihood is strongly convex on that set; see in particular (9) in the statement of the theorem, whose LHS is the Hessian of the log-likelihood. We also show that this set contains the true coefficient vector in Lemma 10, using the matrix Bernstein inequality. Finally, we show that we can efficiently project on this set; see Figure 1.

Thus we run our Projected Stochastic Gradient Descent procedure on this set. As we have already noted, we have no closed-form expression for the negative log-likelihood or its gradient. Nevertheless, we show that, given oracle access to set S, we can get an un-biased sample of the gradient. If $(\boldsymbol{x}_t, y_t)$ is a (randomly chosen) sample processed by PSGD at step t, and $\boldsymbol{w}_t$ the current iterate, we perform rejection sampling to obtain a sample from the Gaussian $\mathcal{N}(\boldsymbol{w}_t^T \boldsymbol{x}_t, 1)$ conditioned on the truncation set S, in order to compute an unbiased estimate of the gradient as per Eq. (6) in Section 4.1. Because we use rejection sampling, it is important to maintain that PSGD remains within a region where the rejection sampling will succeed with constant probability with respect to a random choice of $\boldsymbol{x}_t$, and this is guaranteed by a combination of Lemmas 6 and 10, Definition 4, and Assumption 1.

1.2. Further Related Work

We have already surveyed work on truncated and censored linear regression since the 1950s. Early precursors of this literature can be found in the simpler, non-regression version of our problem, where the $\boldsymbol{x}^{(i)}$'s are single-dimensional and equal, which corresponds to estimating a truncated Normal distribution. This problem goes back to at least Galton (1897), Pearson (1902), Pearson and Lee (1908), and Fisher (1931). Following these early works, there has been a large volume of research devoted to estimating truncated Gaussians or other truncated distributions in one or multiple dimensions; see e.g. Hotelling (1948); Tukey (1949), and Schneider (1986); Cohen (2016); Balakrishnan and Cramer (2014) for an overview of this work. There do exist consistent estimators for estimating the parameters of truncated distributions, but, as in the case of truncated and censored regression, the optimal estimation rates are mostly not well-understood. Only very recentwork of Daskalakis et al. (2018) provides computationally and statistically efficient estimators for the parameters of truncated high-dimensional Gaussians. Similar to the present work, Daskalakis et al. (2018) we use PSGD to optimize the negative log-likelihood of the truncated samples. Showing that the negative log-likelihood is convex in the truncated Gaussian setting follows immediately from the fact that a truncated Gaussian belongs to the exponential family. In our setting it is non-standard; see also discussion in Amemiya (1973). Moreover, identifying the set where the negative log-likelihood is strongly convex and establishing its strong convexity are also simpler tasks in the truncated Gaussian setting compared to the truncated regression setting, due to the shifting of the mean of the samples induced by the different covariate vectors $\boldsymbol{x}^{(i)}$. A further comparison between our techniques and those in Daskalakis et al. (2018) is provided in Section 4.

Last but not least, our work is related, albeit more loosely, to the literature on robust Statistics, which has recently been revived by a strand of fantastic works, discussed in Appendix H. For the most part, these works assume that an adversary perturbs a small fraction of the samples *arbitrarily*. Compared to truncation and censoring, these perturbations are harder to handle. As such only small amounts of perturbation can be accommodated, and the parameters cannot be estimated to arbitrary precision. In contrast, in our setting the truncation set S may very well have an ex ante probability

of obliterating most of the observations, say 99% of them, yet the parameters of the model can still be estimated to arbitrary precision.

2. Preliminaries

Notation. We use small bold letters $\boldsymbol{x}$ to refer to real vectors in finite dimension $\mathbb{R}^d$ and capital bold letters $\boldsymbol{A}$ to refer to matrices in $\mathbb{R}^{d\times\ell}$. Similarly, a function with image in $\mathbb{R}^d$ is represented by a small and bold letter $\boldsymbol{f}$. Let $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ be the inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. We use $\boldsymbol{I}_d$ to refer to the identity matrix in d dimensions. We may drop the subscript when the dimensions are clear. Let also $\mathcal{Q}_d$ be the set of all the symmetric $d \times d$ matrices. The covariance matrix between two vector random variables $\boldsymbol{x}, \boldsymbol{y}$ is $\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}]$.

Vector and Matrix Norms. We define the ℓ_p-norm of $\boldsymbol{x} \in \mathbb{R}^d$ to be $\|\boldsymbol{x}\|_p = (\sum_i x_i^p)^{1/p}$ and the ℓ_∞-norm of $\boldsymbol{x}$ to be $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$. We also define the *spectral norm* of a matrix $\boldsymbol{A}$ to be

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\in\mathbb{R}^d, \boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}.$$

It is well known that for $\boldsymbol{A} \in \mathcal{Q}_d$, $\|\boldsymbol{A}\|_2 = \max_i\{\lambda_i\}$, where λ_i's are the eigenvalues of $\boldsymbol{A}$. The *Frobenius norm* of a matrix $\boldsymbol{A} = (a_{ij}) \in \mathcal{Q}_d$ is defined as follows:

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

The *Mahalanobis distance* between two vectors $\boldsymbol{x}, \boldsymbol{y}$ given a covariance matrix Σ is defined as:

$$\|\boldsymbol{x} - \boldsymbol{y}\|_\Sigma = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{y})}$$

Truncated Gaussian Distribution. Let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, with the following probability density function

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{1}$$

Also, let $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)$ denote the *probability mass of a set* S under this Gaussian measure.

Let $S \subseteq \mathbb{R}^d$ be a subset of the d-dimensional Euclidean space, we define the *S-truncated normal distribution* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ conditioned on taking values in the subset S. The probability density function of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ is the following

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S; \boldsymbol{x}) = \begin{cases} \dfrac{1}{\int_S \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{y}) d\boldsymbol{y}} \cdot \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) & \boldsymbol{x} \in S \\ 0 & \boldsymbol{x} \notin S \end{cases}. \tag{2}$$

Membership Oracle of a Set. Let $S \subseteq \mathbb{R}^d$ be a subset of the d-dimensional Euclidean space. A *membership oracle* of S is an efficient procedure M_S that computes the characteristic function of S, i.e. $M_S(\boldsymbol{x}) = \mathbb{1}_{\boldsymbol{x}\in S}$.

In the Appendix F we present some concentration result that we need for our proof.

3. Truncated Linear Regression Model

Let $S \subseteq \mathbb{R}$ be a measurable subset of the real line. We assume that we have access to n truncated samples of the form $(\boldsymbol{x}^{(i)}, y^{(i)})$. Truncated samples are generated as follows:

1. one $\boldsymbol{x}^{(i)} \in \mathbb{R}^k$ is picked arbitrarily or randomly,
2. the value $y^{(i)}$ is computed according to
$$y^{(i)} = \boldsymbol{w}^{*T}\boldsymbol{x}^{(i)} + \varepsilon^{(i)}, \tag{3}$$
where $\varepsilon^{(i)}$ is sampled from a standard normal distribution $\mathcal{N}(0,1)$,
3. if $y^{(i)} \in S$ then return $(\boldsymbol{x}^{(i)}, y^{(i)})$, otherwise repeat from step 1 with the same index i.

Without any assumptions on the truncation set S, it is easy to see that no meaningful estimation is possible. When $k = 1$ the regression problem becomes the estimation of the mean of a Gaussian distribution that has been studied in Daskalakis et al. (2018). In this case the necessary and sufficient condition is that the Gaussian measure of the set S is at least a constant α. When $k > 1$ though, for every $\boldsymbol{x} \in \mathbb{R}^k$ we have a different $\alpha(\boldsymbol{x})$ defined as we can see in the following definition.

Definition 1 (SURVIVAL PROBABILITY) *Let S be a measurable subset of $\mathbb{R}$. Given $\boldsymbol{x} \in \mathbb{R}^k$ and $\boldsymbol{w} \in \mathbb{R}^k$ we define the survival probability $\alpha(\boldsymbol{w}, \boldsymbol{x}; S)$ of the sample with feature vector $\boldsymbol{x}$ and parameters $\boldsymbol{w}$ as*
$$\alpha(\boldsymbol{w}, \boldsymbol{x}; S) = \mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}, 1; S).$$
When S is clear from the context we may refer to $\alpha(\boldsymbol{w}, \boldsymbol{x}; S)$ simply as $\alpha(\boldsymbol{w}, \boldsymbol{x})$.

Since S has a different mass for every $\boldsymbol{x}$ the assumption that we need in this regime is more complicated than the assumption used by Daskalakis et al. (2018). A natural candidate assumption is that for every $\boldsymbol{x}^{(i)}$ the mass $\alpha(\boldsymbol{w}, \boldsymbol{x}^{(i)}; S)$ is large enough. We propose an even weaker condition which is sufficient for recovering the regression parameters.

Assumption 1 (CONSTANT SURVIVAL PROBABILITY ASSUMPTION) *Let $(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})$ be samples from the regression model (3). There exists a constant $a > 0$ such that*
$$\sum_{i=1}^{n} \log\left(\frac{1}{\alpha(\boldsymbol{x}^{(i)}, \boldsymbol{w}^*)}\right) \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} \preceq \log\left(\frac{1}{a}\right) \sum_{i=1}^{n} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}.$$

Our second assumption involves only the $\boldsymbol{x}^{(i)}$'s that we observe and is similar to the usual assumption in linear regression that covariance matrix of $\boldsymbol{x}^{(i)}$'s has high enough variance in every direction.

Assumption 2 (THICKNESS OF COVARIANCE MATRIX OF COVARIATES ASSUMPTION) *Let $\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}$, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^k$. Then for every $i \in [n]$, it holds that*
$$\boldsymbol{X} \succeq \frac{\log(k)}{n} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}.$$

There are two stronger conditions that can replace the aforementioned thickness assumption. The first one assumes that $\boldsymbol{X} \succeq \boldsymbol{I}$, as is usually assumed in classical linear regression applications, and also that $\left\|\boldsymbol{x}^{(i)}\right\|_2^2 \leq \frac{n}{\log(k)}$, for all i. This pair of conditions hold with high probability if the covariates are sampled from some well-behaved distribution, e.g. a Gaussian, when n is large enough.

In Appendix E we outline how Assumptions 1 and 2 can be proven using more mild assumption if the covariates $\boldsymbol{x}^{(i)}$ are sampled from some prior distribution.

4. Estimating the Parameters of a Truncated Linear Regression

We start with the formal statement of our main theorem about the parameter estimation of the truncated linear regression model (3).

Theorem 2 *Let $(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(n)}, y^{(n)})$ be n samples from the linear regression model (3) with parameters $\boldsymbol{w}^*$. If Assumptions 1 and 2 hold, then there exists a polynomial time algorithm with success probability at least $2/3$, that outputs an estimation $\hat{\boldsymbol{w}}$ such that*

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_{\boldsymbol{X}} \leq \mathrm{poly}(1/a)\sqrt{\frac{k}{n}\log(n)},$$

where $\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}$.

As we explained in the introduction, our estimation algorithm is projected stochastic gradient descent for maximizing the population log-likelihood function with the careful choice of the projection set. Because the proof of consistency and efficiency of our algorithm is technical we first present an outline of the proof and then we present the individual lemmas for each step.

The framework that we use for our analysis is based on the Chapter 14 of Shalev-Shwartz and Ben-David (2014). In this framework the goal is to minimize a convex function $f : \mathbb{R}^k \to \mathbb{R}$ by doing gradient steps but with noisy unbiased estimations of the gradient. The following algorithm describes projected stochastic gradient descent applied to a convex function f, with projection set $\mathcal{D}_r \subseteq \mathbb{R}^k$. We will define formally the particular set $\mathcal{D}_r$ that we consider in Definition 4. For now $\mathcal{D}_r$ should be thought of an arbitrary convex subset of $\mathbb{R}^k$.

Algorithm ($\star$). Projected SGD for Minimizing a λ-Strongly Convex Function.

1: $\boldsymbol{w}^{(0)} \leftarrow$ arbitrary point in $\mathcal{D}_r$ ▷ (a) *initial feasible point*
2: **for** $i = 1, \ldots, M$ **do**
3: Sample $\boldsymbol{v}^{(i)}$ such that $\mathbb{E}\left[\boldsymbol{v}^{(i)} \mid \boldsymbol{w}^{(i-1)}\right] \in \partial f(\boldsymbol{w}^{(i-1)})$ ▷ (b) *estimation of gradient*
4: $\boldsymbol{r}^{(i)} \leftarrow \boldsymbol{w}^{(i-1)} - \frac{1}{\lambda \cdot i}\boldsymbol{v}^{(i)}$
5: $\boldsymbol{w}^{(i)} \leftarrow \mathrm{argmin}_{\boldsymbol{w} \in \mathcal{D}_r} \left\|\boldsymbol{w} - \boldsymbol{r}^{(i)}\right\|$ ▷ (c) *projection step*
6: **return** $\bar{\boldsymbol{w}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{w}^{(i)}$

Our goal is to apply the Algorithm ($\star$) to the negative log-likelihood function of the truncated linear regression model. Although we formally define the negative log-likelihood function in Section 4.1, it is clear from the above description that in order to apply Algorithm ($\star$) we have to solve the following three algorithmic problems

(a) **initial feasible point:** efficiently compute an initial feasible point in $\mathcal{D}_r$,

(b) **unbiased gradient estimation:** efficiently sample an unbiased estimation of ∇f,

(c) **efficient projection:** design an efficient algorithm to project to the set $\mathcal{D}_r$.

Solving (a) - (c) is the first step in the proof of our main Theorem 2. Then our goal is to apply the following theorem of Shalev-Shwartz and Ben-David (2014).

Theorem 3 (Theorem 14.11 of Shalev-Shwartz and Ben-David (2014).) *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function, $\boldsymbol{v}^{(1)}, \dots, \boldsymbol{v}^{(M)}$ be a sequence of random vectors such that $\mathbb{E}\left[\boldsymbol{v}^{(i)} \mid \boldsymbol{w}^{(i-1)}\right] \in \partial f(\boldsymbol{w}^{(i-1)})$ and $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathcal{D}_r} f(\boldsymbol{w})$ be a minimizer of f. If we assume the following:*

(i) **bounded variance step:** $\mathbb{E}\left[\left\|\boldsymbol{v}^{(i)}\right\|_2^2\right] \leq \rho^2$,

(ii) **strong convexity:** *$f(\boldsymbol{w})$ is λ-strongly convex,*

then, $\mathbb{E}\left[f(\bar{\boldsymbol{w}})\right] - f(\boldsymbol{w}^*) \leq \frac{\rho^2}{2\lambda M}\left(1 + \log(M)\right)$, where $\bar{\boldsymbol{w}}$ is the output of the Algorithm (⋆).

Unfortunately, none of the properties (i), (ii) hold for all vectors $\boldsymbol{w} \in \mathbb{R}^d$ and hence we cannot use vanilla SGD. For this reason, we add the projection step. We identify a projection set $\mathcal{D}_{r^*}$ such that log-likelihood satisfies both (i) and (ii) for all vectors $\boldsymbol{w} \in \mathcal{D}_{r^*}$.

Definition 4 (PROJECTION SET) *We define*

$$D_r = \left\{ \boldsymbol{w} \in \mathbb{R}^k \mid \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)}\right)^2 \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)T} \preceq r \sum_{i=1}^{n} \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)T} \right\}.$$

*We set $r^* = 4\log(2/a) + 7$. Given $\boldsymbol{x} \in \mathbb{R}^k$ we say that $\boldsymbol{x}$ is **feasible** if and only if $\boldsymbol{x} \in \mathcal{D}_{r^*}$.*

Using the projection to $\mathcal{D}_{r^*}$, we can prove (i) and (ii) and hence we can apply Theorem 3. The last step is to transform the conclusions of Theorem 3 to guarantees in the parameter space. For this we use again the strong convexity of f which implies that closeness in the objective value translates to closeness in the parameter space. For this argument we also need the following property:

(iii) **feasibility of optimal solution:** $\boldsymbol{w}^* \in \mathcal{D}_{r^*}$.

OUTLINE OF THE PROOF OF THEOREM 2. The proof follows the following steps solving the aforementioned algorithmic problems (a) - (c) and the statistical problems (i) - (iii). For the problem (a) we observe that is reducible to (c) since once we have an efficient procedure to project we can start from an arbitrary point in $\mathbb{R}^k$, e.g. $\boldsymbol{w} = \boldsymbol{0}$ and project back to $\mathcal{D}_{r^*}$ and this is our initial point.

1. We start in Section 4.1 with the definition of the negative log-likelihood function and we compute its gradient and Hessian.
2. In Section 4.2 we provide some technical lemmas that are useful to understand the formal statements of the rest of the proof.
3. In Section 4.3 we present the details of the Algorithm 2 that is used to compute an unbiased estimation of the gradient, which gives a solution to the algorithmic problem (b).
4. In the Appendix C.1 we present a detailed analysis of our projection Algorithm 3, which gives a solution to algorithmic problems (a) and (c).
5. In Section 4.4 we present the statements that prove the (i) bounded variance and (ii) strong convexity of the log-likelihood function. This is the main technical contribution of the paper and uses all the results that we have proved in Section 4.2 together with Assumptions 1, 2.
6. In Section 4.5 we prove the feasibility of the optimal solution, i.e. $\boldsymbol{w}^* \in \mathcal{D}_{r^*}$ which resolves the problem (iii).
7. Finally in Section 4.6 we use all the mentioned results to prove our main Theorem 2.

In the Appendix G we provide a comparison of the techniques used to prove Theorem 2 with similar approaches in truncated statistics.

4.1. Log-Likelihood of Truncated Linear Regression

We first present the negative log-likelihood of a single sample $(\boldsymbol{x}, y)$ and then we present the population version of the negative log-likelihood function and its first two derivatives.

Given two vectors $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^k$, the log-likelihood that $(\boldsymbol{x}, y)$ is a sample of the form (3) is equal to

$$\begin{aligned}\ell(\boldsymbol{w}; \boldsymbol{x}, y) &= -\frac{1}{2}(y - \boldsymbol{w}^T\boldsymbol{x})^2 - \log\left(\int_S \exp\left(-\frac{1}{2}(z - \boldsymbol{w}^T\boldsymbol{x})^2\right) dz\right) \\ &= -\frac{1}{2}y^2 + y \cdot \boldsymbol{w}^T\boldsymbol{x} - \log\left(\int_S \exp\left(-\frac{1}{2}z^2 + z \cdot \boldsymbol{w}^T\boldsymbol{x}\right) dz\right) \end{aligned} \tag{4}$$

The population log-likelihood function with n samples is equal to

$$\bar{\ell}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)},1,S)}\left[\ell(\boldsymbol{w}; \boldsymbol{x}^{(i)}, y)\right]. \tag{5}$$

We now compute the gradient of $\bar{\ell}(\boldsymbol{w})$.

$$\nabla\bar{\ell}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{y\sim\mathcal{N}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)},1,S)}\left[y \cdot \boldsymbol{x}^{(i)}\right] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{w}^{T}\boldsymbol{x}^{(i)},1,S)}\left[z \cdot \boldsymbol{x}^{(i)}\right]. \tag{6}$$

Finally, we compute the Hessian $\mathbf{H}_{\bar{\ell}}$

$$\bar{\mathbf{H}}_{\ell} = -\frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}_{z\sim\mathcal{N}(\boldsymbol{w}^{T}\boldsymbol{x}^{(i)},1,S)}\left[z \cdot \boldsymbol{x}^{(i)}, z \cdot \boldsymbol{x}^{(i)}\right]. \tag{7}$$

Since the covariance matrix of a random variable is always positive semidefinite, we conclude that $\mathbf{H}_{\bar{\ell}}$ is negative semidefinite, which implies the following lemma.

Lemma 5 *The population log-likelihood function $\bar{\ell}(\boldsymbol{w})$ is a concave function with respect to $\boldsymbol{w}$ for all $\boldsymbol{x} \in \mathbb{R}^k$, $y \in \mathbb{R}$.*

4.2. Survival Probability of Feasible Points

One necessary technical lemma is how to correlate the survival probabilities $\alpha(\boldsymbol{w}, \boldsymbol{x})$, $\alpha(\boldsymbol{w}', \boldsymbol{x})$ for two different points $\boldsymbol{w}$, $\boldsymbol{w}'$. In Lemma 6 we show that this is possible based on their distance with respect to $\boldsymbol{x}$. Then in Lemma 7 we show how the expected second moment with respect to the truncated Guassian error is related with the value of the corresponding survival probability. We present the proofs of Lemma 6 and Lemma 7 in the Appendix D.3 and D.4 respectively.

Lemma 6 *Let $\boldsymbol{x}$, $\boldsymbol{w}$, $\boldsymbol{w}' \in \mathbb{R}^k$, then* $\log\left(\frac{1}{\alpha(\boldsymbol{w},\boldsymbol{x})}\right) \leq 2\log\left(\frac{1}{\alpha(\boldsymbol{w}',\boldsymbol{x})}\right) + \left|(\boldsymbol{w} - \boldsymbol{w}')^T\boldsymbol{x}\right|^2 + 2.$

Lemma 7 *Let $\boldsymbol{x} \in \mathbb{R}^k$, $\boldsymbol{w} \in \mathbb{R}^k$, then* $\mathbb{E}_{y\sim\mathcal{N}(\boldsymbol{w}^T\boldsymbol{x},1,S)}\left[(y - \boldsymbol{w}^T\boldsymbol{x})^2\right] \leq 2\log\left(\frac{1}{\alpha(\boldsymbol{w},\boldsymbol{x})}\right) + 4.$

One corollary of these lemmas is an interesting property of feasible points, i.e. points $\boldsymbol{w}$ inside $\mathcal{D}_{r^*}$, namely that they satisfy Assumption 1, under the assumption that $\boldsymbol{w}^* \in \mathcal{D}_{r^*}$, which we will prove later.

4.3. Unbiased Gradient Estimation

Using (6) we have that the gradient of the population log-likelihood function is equal to $\nabla\bar{\ell}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{y\sim\mathcal{N}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)},1,S)}\left[y\cdot\boldsymbol{x}^{(i)}\right] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{z\sim\mathcal{N}(\boldsymbol{w}^{T}\boldsymbol{x}^{(i)},1,S)}\left[z\cdot\boldsymbol{x}^{(i)}\right]$. Hence an unbiased estimation of $\nabla\bar{\ell}(\boldsymbol{w})$ can be computed given one sample from the distribution $\mathcal{N}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)},1,S)$ and one sample from $\mathcal{N}(\boldsymbol{w}^{T}\boldsymbol{x}^{(i)},1,S)$. For the first sample we can use $y^{(i)}$. For the second sample though we need a sampling procedure that given $\boldsymbol{w}$ and $\boldsymbol{x}$ produces a sample from $\mathcal{N}(\boldsymbol{w}^{T}\boldsymbol{x},1,S)$. For this we could simply use rejection sampling, but because we have not assumed that $\alpha(\boldsymbol{w},\boldsymbol{x})$ is always large, we use a more elaborate argument based on the following Lemma 8.

Lemma 8 *For any $\boldsymbol{w}\in\mathcal{D}_{r^*}$ there exists an $i\in[n]$ such that $\alpha\left(\boldsymbol{w},\boldsymbol{x}^{(i)}\right)\geq\text{poly}\left(\frac{1}{a}\right)$.*

We present the proof of Lemma 8 in the Appendix C.2. The exact rejection sampling algorithm is presented in the Appendix 2. From Lemma 8 we have that every iteration of the while loop of Algorithm 2 has probability of success at least $\frac{1}{n}\text{poly}\left(\frac{1}{a}\right)$ hence in $\text{poly}\left(n,\frac{1}{a}\right)$ iterations the rejection sampling algorithm succeeds with high probability.

4.4. Bounded Step Variance and Strong Convexity

We are now ready to prove the (i) bounded variance and (ii) strong convexity properties as descripted in the beginning of Section 4. The results are summarized in the following Theorem 9. The proof of Theorem 9 is presented in the Appendix D.5.

Theorem 9 *Let $(\boldsymbol{x}^{(1)},y^{(1)}),\ldots,(\boldsymbol{x}^{(n)},y^{(n)})$ be n samples from the linear regression model (3) with parameters $\boldsymbol{w}^*$. If Assumptions 1 and 2 hold, then*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(y^{(i)}-z^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right]\leq 10\cdot r^*k+20, \tag{8}$$

$$\mathbf{H}_{-\bar{\ell}}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(z^{(i)}-\mathbb{E}\left[z^{(i)}\right]\right)^2\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}\right]\succeq e^{-16}a^{10}\boldsymbol{I}, \tag{9}$$

where $y^{(i)}\sim\mathcal{N}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)},1,S)$, $z^{(i)}\sim\mathcal{N}(\boldsymbol{w}^{T}\boldsymbol{x}^{(i)},1,S)$ and $\boldsymbol{w}\in D_{r^*}$ with $r^*=4\log\left(2/a\right)+7$.

4.5. Feasibility of Optimal Solution

As described in the high level description of our proof in the beginning of the section, in order to be able to use strong convexity to prove the closeness in parameter space of our estimator, we have to prove that $\boldsymbol{w}^*\in\mathcal{D}_{r^*}$. This is also needed to prove that all the points $\boldsymbol{w}\in\mathcal{D}_r$ satisfy the Assumption 1, which we have used to prove the bounded variance and the strong convexity property in Section 4.4. The proof of the following Lemma can be found in the Appendix D.2.

Lemma 10 *Let $(\boldsymbol{x}^{(1)},y^{(1)}),\ldots,(\boldsymbol{x}^{(n)},y^{(n)})$ be n samples from the linear regression model (3) with parameters $\boldsymbol{w}^*$. If Assumptions 1 and 2 hold then, $\mathbb{P}\left(\boldsymbol{w}^*\in D_{r^*}\right)\geq 2/3$.*

4.6. Proof of Theorem 2

In this section we analyze Algorithm 1, which implements the projected stochastic gradient descent on the negative log-likelihood landscape.

First we observe that from Section 4.3 and Appendix C.1 we have a proof that we can: (a) efficiently find one initial feasible point, (b) we can compute efficiently an unbiased estimation of the gradient and (c) we can efficiently project to the set $\mathcal{D}_{r^*}$. These results prove that the Algorithm 1 that we analyze has running time that is polynomial in the number of steps M of stochastic gradient descent, and in the dimension k.

It remains to upper bound the number of steps that the PSGD algorithm needs to compute an estimate that is close in parameter space with $\boldsymbol{w}^*$. The analysis of this estimation algorithm will be based on Theorem 9 from Section 4.4 combined with the general theorem about the performance of the projected stochastic gradient descent (Theorem 3). This is summarized in Lemma 11 which follows directly from Theorem 3, and Theorem 9.

Lemma 11 *Let $\boldsymbol{w}^*$ be the underlying parameters of our model, $f = -\bar{\ell}$ and M be the number of PSGD steps of Algorithm 1 then, $\mathbb{E}\left[f(\bar{\boldsymbol{w}})\right] - f(\boldsymbol{w}^*) \leq \mathrm{poly}\left(\frac{1}{a}\right)\frac{k}{M}\left(1+\log(M)\right)$, where $\bar{\boldsymbol{w}}$ is the output of Algorithm 1.*

Proof of Theorem 2: Applying Markov's inequality to Lemma 11 we get that

$$\mathbb{P}\left(f(\bar{\boldsymbol{w}}) - f(\boldsymbol{w}^*) \geq \mathrm{poly}\left(\frac{1}{\delta a}\right)\frac{k}{M}\left(1+\log(M)\right)\right) \leq \delta. \tag{10}$$

Hence we can condition on the event $f(\hat{\boldsymbol{w}}) - f(\boldsymbol{w}^*) \leq \mathrm{poly}\left(\frac{1}{a}\right)\frac{k}{M}\left(1+\log(M)\right)$ and we only lose probability at most δ. Now we can the λ-strong convexity of f, which implies that $f(\boldsymbol{w}) - f(\boldsymbol{w}^*) \geq \frac{\lambda}{2}\left\|\boldsymbol{w}-\boldsymbol{w}^*\right\|_2^2$, to get that

$$\left\|\hat{\boldsymbol{w}}-\boldsymbol{w}^*\right\|_2^2 \leq \mathrm{poly}\left(\frac{1}{a}\right)\frac{k}{M}\left(1+\log(M)\right), \tag{11}$$

and our theorem follows. ■

5. Learning Single-Layer Neural Networks with Noisy Activation Function

In this section we will describe how we can use our truncated regression algorithm to provably learn the parameters of an one layer neural network with noisy activation functions. Noisy activation function have been explored by Nair and Hinton (2010), Bengio et al. (2013) and Gulcehre et al. (2016) as we have discussed in the introduction. The problem of estimating the parameters of such a neural network is a challenging problem and no theoretically rigorous methods are known. In this section we show that this problem can be formulated as a truncated linear regression problem which we can efficiently solve using our Algorithm 1.

Let $g : \mathbb{R} \to \mathbb{R}$ be a random map that corresponds to a noisy rectifier linear unit, i.e. $g(x) = \max\{0, x+\varepsilon\}$ where ε is a standard normal random variable. Then an one layer neural network with noisy activation functions is the multivalued function f parameterized by the vector $\boldsymbol{w} \in \mathbb{R}^k$ such that $f_{\boldsymbol{w}}(\boldsymbol{x}) = g(\boldsymbol{w}^T\boldsymbol{x})$. In the realizable, supervised setting we observe n labeled samples of the form $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ and we want to estimate the parameters $\boldsymbol{W}$ that better capture the samples we

have observed. We remind that the assumption that $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ is realizable means that there exists a $\boldsymbol{w}^*$ such that for all i it holds $y^{(i)} = f_{\boldsymbol{w}^*}\left(\boldsymbol{x}^{(i)}\right) = g(\boldsymbol{w}^{*T}\boldsymbol{x})$.

Our SGD algorithm then gives a rigorous method to estimate $\boldsymbol{W}^*$ if we assume that the inputs $\boldsymbol{x}^{(i)}$ together with the truncation of the activation function satisfy Assumption 1 and Assumption 2. Using Theorem 2 we can then bound the number of samples that we need for this learning task. These results are summarized in the following corollary which directly follows from Theorem 2.

Corollary 12 *Let $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ be n i.i.d. samples drawn according to the following distribution*

$$y^{(i)} = f_{\boldsymbol{w}^*}\left(\boldsymbol{x}^{(i)}\right) = \max\{0, \boldsymbol{w}^{*T}\boldsymbol{x}^{(i)} + \varepsilon^{(i)}\}$$

with $\varepsilon^{(i)} \sim \mathcal{N}(0,1)$. Assume also that $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)$ and $\boldsymbol{w}^*$ satisfy Assumption 1 and Assumption 2 with survival probability a. Then the SGD Algorithm 1 runs in polynomial time and outputs an estimate $\hat{\boldsymbol{w}}$ such that

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_{\boldsymbol{X}} \leq \text{poly}\left(\frac{1}{a}\right)\sqrt{\frac{k}{n}}$$

with probability at least $2/3$, where $\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}$.

We remark that the aforementioned problem is easier than the problem that we solve in Section 4. The reason is that in the neural network setting even the samples $y^{(i)}$ that are filtered by the activation function are available to us and hence we have the additional information that we can compute their percentage. In Corollary 12 we don't use at all this information.

6. Experiments

To validate the performance of our proposed algorithm we constructed a synthetic dataset with various datapoints $\boldsymbol{x}_i \in \mathbb{R}^{10}$ that were drawn uniformly at random from a Gaussian Distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. For each of these datapoints, we generated the corresponding y_i as $y_i = \boldsymbol{w}^T\boldsymbol{x}_i + \varepsilon_i$, where ε_i where drawn independently from $\mathcal{N}(0,1)$ and w was chosen to be the all-ones vector $\mathbf{1}$. We filtered the dataset to keep all samples with $y_i > 4$ and $\boldsymbol{w}^T\boldsymbol{x} < 2$. We note that to run the projection step of our algorithm we use the convex optimization library `cvxpy`.

Figure 2 shows the comparison with ordinary least squares. You can see that even though the OLS estimator quickly converges, its estimate is biased due to the data truncation. As a result, the estimates produced tend to be significantly larger in magnitude than the true $\boldsymbol{w} = \mathbf{1}$. In contrast, our proposed method is able to correct for this bias achieving an estimate that improves with the number of samples n at an optimal rate of $1/\sqrt{n}$, despite the adversarial nature of the filtering that kept only significantly high values of y.

References

Takeshi Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica: Journal of the Econometric Society*, pages 997–1016, 1973.

N Balakrishnan and Erhard Cramer. *The art of progressive censoring*. Springer, 2014.

Sivaraman Balakrishnan, Simon S. Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 169–212, 2017. URL http://proceedings.mlr.press/v65/balakrishnan17a.html.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Richard Breen et al. *Regression models: Censored, sample selected, or truncated data*, volume 111. Sage, 1996.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 47–60, 2017. doi: 10.1145/3055399.3055491. URL http://doi.acm.org/10.1145/3055399.3055491.

A Clifford Cohen. *Truncated and censored samples: theory and applications*. CRC press, 2016.

Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Efficient statistics, in high dimensions, from truncated samples. In *the 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2018.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 655–664, 2016. doi: 10.1109/FOCS.2016.85. URL https://doi.org/10.1109/FOCS.2016.85.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 999–1008, 2017. URL http://proceedings.mlr.press/v70/diakonikolas17a.html.

Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 2683–2702, 2018. doi: 10.1137/1.9781611975031.171. URL https://doi.org/10.1137/1.9781611975031.171.

Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *the 30th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2019.

RA Fisher. Properties and applications of Hh functions. *Mathematical tables*, 1:815–852, 1931.

Francis Galton. An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data. *Proceedings of the Royal Society of London*, 62(379-387):310–315, 1897.

Robert D Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

Caglar Gulcehre, Marcin Moczulski, Misha Denil, and Yoshua Bengio. Noisy activation functions. In *International Conference on Machine Learning*, pages 3059–3068, 2016.

Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Conference on Learning Theory*, pages 354–375, 2013.

Jerry A Hausman and David A Wise. Social experimentation, truncated distributions, and efficient estimation. *Econometrica: Journal of the Econometric Society*, pages 919–938, 1977.

Harold Hotelling. Fitting generalized truncated normal distributions. In *Annals of Mathematical Statistics*, volume 19, pages 596–596, 1948.

Kevin A. Lai, Anup B. Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 665–674, 2016. doi: 10.1109/FOCS.2016.76. URL https://doi.org/10.1109/FOCS.2016.76.

Alice Lee. Table of the Gaussian "Tail" Functions; When the "Tail" is Larger than the Body. *Biometrika*, 10(2/3):208–214, 1914.

Jerry Li. Robust sparse estimation tasks in high dimensions. *CoRR*, abs/1702.05860, 2017. URL http://arxiv.org/abs/1702.05860.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Karl Pearson. On the systematic fitting of frequency curves. *Biometrika*, 2:2–7, 1902.

Karl Pearson and Alice Lee. On the generalised probable error in multiple normal correlation. *Biometrika*, 6(1):59–68, 1908.

Helmut Schneider. *Truncated and censored samples from normal populations*. Marcel Dekker, Inc., 1986.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 45:1–45:21, 2018. doi: 10.4230/LIPIcs.ITCS.2018.45. URL https://doi.org/10.4230/LIPIcs.ITCS.2018.45.

James Tobin. Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36, 1958.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

John W Tukey. Sufficiency, truncation and selection. *The Annals of Mathematical Statistics*, pages 309–311, 1949.

Huan Xu, Constantine Caramanis, and Shie Mannor. Principal component analysis with contaminated data: The high dimensional case. *arXiv preprint arXiv:1002.4658*, 2010.

Appendix A. Full Description of the Algorithm

Algorithm 1 Projected Stochastic Gradient Descent. Given access to samples from $\mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}, 1, S)$.

procedure SGD(M, λ) ▷ *M: number of steps, λ: parameter*
  $\boldsymbol{w}^{(0)} \leftarrow$ PROJECTTODOMAIN($r^*, \mathbf{0}$)
  for $i = 1, \ldots, M$ **do**
    Sample $(\boldsymbol{x}^{(i)}, y^{(i)})$ from $\mathcal{O}$
    $\eta_i \leftarrow \frac{1}{\lambda \cdot i}$
    $\boldsymbol{v}^{(i)} \leftarrow$ GRADIENTESTIMATION($\boldsymbol{x}^{(i)}, \boldsymbol{w}^{(i-1}, y^{(i)})$)
    $\boldsymbol{r}^{(i)} \leftarrow \boldsymbol{w}^{(i-1)} - \eta_i \boldsymbol{v}^{(i)}$
    $\boldsymbol{w}^{(i)} \leftarrow$ PROJECTTODOMAIN($r^*, \boldsymbol{r}^{(i)}$)
  return $\bar{\boldsymbol{w}} \leftarrow \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{w}^{(i)}$ ▷ *output the average*

Algorithm 2 The function to estimate the gradient of log-likelihood as in (6).

function GRADIENTESTIMATION($\boldsymbol{r}, \boldsymbol{w}, y$)
  $b \leftarrow$ false
  while $b =$ false **do**
    pick $\boldsymbol{x}$ at random from the set $\left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right\}$
    Sample z from $\mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}, 1)$
    $b \leftarrow M_S(z)$ ▷ *M_S is the membership oracle of the set S*
  return $y\boldsymbol{r} - z\boldsymbol{x}$

Algorithm 3 The function that projects a current guess back to the domain $\mathcal{D}_r$ (see Appendix C).

function PROJECTTODOMAIN($r, \boldsymbol{w}$) ▷ *r is the parameter of the domain $\mathcal{D}_r$*
  $\hat{\tau} \leftarrow \arg\min_\tau \{$ELLIPSOID($\boldsymbol{w}, \tau, r$) $\neq$ "Empty"$\}$ ▷ *find τ via binary search*
  return ELLIPSOID($\boldsymbol{w}, \hat{\tau}, r$)

function ELLIPSOID($\boldsymbol{w}, \tau, r$) ▷ *return a point $\boldsymbol{z}$ in $\mathcal{D}_r$ with $\|\boldsymbol{z} - \boldsymbol{w}\|_2 \leq \tau$ or "Empty"*
  $\mathcal{E}_0 \leftarrow \{\boldsymbol{z} \mid \|\boldsymbol{z} - \boldsymbol{w}\|_2 \leq \tau\}$
  return the result of the ellipsoid method with initial ellipsoid $\mathcal{E}_0$ and FINDSEPARATION as a separation oracle

function FINDSEPARATION($\boldsymbol{u}, r$) ▷ *find a separating hyperplane between $\boldsymbol{u}$ and $\mathcal{D}_r$*
  $\boldsymbol{A} \leftarrow \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{u}^T\boldsymbol{x}^{(i)}\right)^2 \boldsymbol{x_i}^{(i)}\boldsymbol{x_i}^{(i)T}$
  if $\lambda_{\max}(\boldsymbol{A}) \leq r$ **then**
    return "is member"
  else
    $\boldsymbol{v}_m \leftarrow$ eigenvector of $\boldsymbol{A}$ that corresponds to the maximum eigenvalue $\lambda_{\max}(\boldsymbol{A})$
    $\mathcal{E} \leftarrow \left\{\boldsymbol{z} \in \mathbb{R}^k \mid \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{z}^T\boldsymbol{x}^{(i)}\right)^2 \left(\boldsymbol{v}_m^T\boldsymbol{x}^{(i)}\right)^2 \leq r\right\}$
    return a separating hyperplane between the vector $\boldsymbol{u}$ and the ellipsoid $\mathcal{E}$ (See (16))

Figure 1: Description of the Stochastic Gradient Descent (SGD) algorithm for estimating the parameters of a truncated linear regression.

Appendix B. Plots

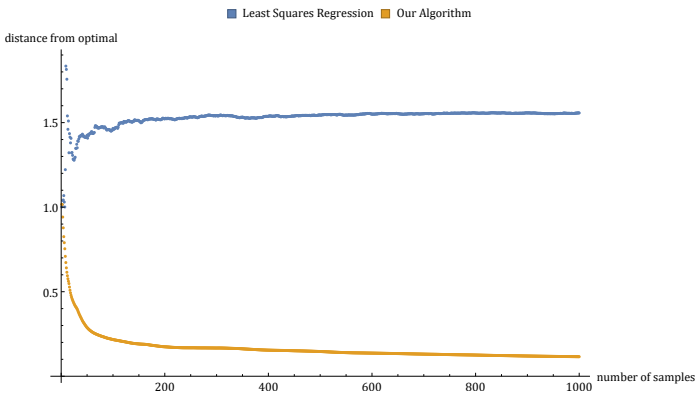

Figure 2: Comparison of the proposed method with ordinary least squares.

Appendix C. Technical Details on Algorithm 2 and Algorithm 3

C.1. Details on Algorithm 3

The convex problem we need to solve in this step is the following

$$\min_{\boldsymbol{z}\in\mathcal{D}_r} \|\boldsymbol{z}-\boldsymbol{w}\|_2 . \tag{12}$$

For simplicity of exposition we assume that we have normalized the space so that $\sum_{i=1}^{n} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} = \boldsymbol{I}$.

The main idea of the algorithm to solve 12 is to use the ellipsoid method with separating hyperplane oracle as descripted in Chapter 3 of Grötschel et al. (2012). This yields a polynomial time algorithm as it is proved in Grötschel et al. (2012). We now explain in more detail each step of the algorithm.

1. The binary search over τ is the usual procedure to reduce the minimization of the norm $\|\boldsymbol{z}-\boldsymbol{w}\|_2$ to satifiability queries of a set of convex constraints, in our case $\boldsymbol{z} \in \mathcal{D}_r$ and $\|\boldsymbol{z}-\boldsymbol{w}\|_2 \leq \tau$.

2. The fact that the constraint $\|\boldsymbol{z}-\boldsymbol{w}\|_2 \leq \tau$ is satisfied through the execution of the ellipsoid algorithm is guaranteed because of the selection of the initial ellipsoid to be

$$\mathcal{E}_0 = \{\boldsymbol{z} \mid \|\boldsymbol{z}-\boldsymbol{w}\|_2 \leq \tau\}.$$

3. The main technical difficulty is how to find a separating hyperplane between a vector $\boldsymbol{u}$ that is outside the set $\mathcal{D}_r$ and the convex set $\mathcal{D}_r$. First observe that $\boldsymbol{z} \in \mathcal{D}_r$ is equivalent with the following set of constraints

$$\sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{z}^T\boldsymbol{x}^{(i)}\right)^2 \left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \leq r \quad \forall \boldsymbol{v} \in \mathbb{R}^k,\ \|\boldsymbol{v}\|_2 = 1 \tag{13}$$

which after of simple calculations is equivalent with

$$\boldsymbol{z}^T \boldsymbol{P}_{\boldsymbol{v}} \boldsymbol{z} + \boldsymbol{q}_{\boldsymbol{v}}^T \boldsymbol{z} + s_{\boldsymbol{v}} \leq r \quad \forall \boldsymbol{v} \in \mathbb{R}^k, \ \|\boldsymbol{v}\|_2 = 1 \tag{14}$$

with

$$\begin{aligned} \boldsymbol{P}_{\boldsymbol{v}} &= \sum_{i=1}^{n} \left(\boldsymbol{v}^T \boldsymbol{x}^{(i)} \right)^2 \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)T}, \\ \boldsymbol{q}_{\boldsymbol{v}} &= -2 \sum_{i=1}^{n} y^{(i)} \left(\boldsymbol{v}^T \boldsymbol{x}^{(i)} \right)^2 \boldsymbol{x}^{(i)} \\ \text{and} \quad s_{\boldsymbol{v}} &= \sum_{i=1}^{n} \left(y^{(i)} \right)^2 \left(\boldsymbol{v}^T \boldsymbol{x}^{(i)} \right)^2. \end{aligned}$$

It is clear from the definition that $\boldsymbol{P}_{\boldsymbol{v}} \succeq \boldsymbol{0}$. Also observe that for any vector $\boldsymbol{z}$ such that $\boldsymbol{z}^T \boldsymbol{P}_{\boldsymbol{v}} \boldsymbol{z} = 0$ it also holds that $\boldsymbol{q}_{\boldsymbol{v}}^T \boldsymbol{z} = 0$. This holds because $\boldsymbol{z}^T \boldsymbol{P}_{\boldsymbol{v}} \boldsymbol{z} = 0$ is a sum of squares, hence it is zero if and only if all the terms are zero and $\boldsymbol{q}_{\boldsymbol{v}}^T \boldsymbol{z} = 0$ is a linear combination of these terms. Hence for any unit vector $\boldsymbol{v}$ the equivalent inequalities (13), (14) describe an ellipsoid with its interior.

Therefore if $\boldsymbol{u}$ is outside $\mathcal{D}_r$ then $\boldsymbol{u}$ violates some of the ellipsoid inequalities (14). To find such an ellipsoid it suffices to compute the eigenvector $\boldsymbol{v}_m$ that corresponds to the maximum eigenvalue of the following matrix

$$\boldsymbol{A} = \sum_{i=1}^{k} \left(y^{(i)} - \boldsymbol{u}^T \boldsymbol{x}^{(i)} \right)^2 \boldsymbol{x_i}^{(i)} \boldsymbol{x_i}^{(i)T}.$$

If $\lambda_{\max}(\boldsymbol{A}) \leq r$ then $\boldsymbol{u} \in \mathcal{D}_r$, otherwise it holds that

$$g \triangleq \boldsymbol{v}_m^T \left(\sum_{i=1}^{k} \left(y^{(i)} - \boldsymbol{u}^T \boldsymbol{x}^{(i)} \right)^2 \boldsymbol{x_i}^{(i)} \boldsymbol{x_i}^{(i)T} \right) \boldsymbol{v}_m > r. \tag{15}$$

This implies that $\boldsymbol{u}$ is outside the ellipsoid

$$\mathcal{E} = \left\{ \boldsymbol{z} \mid \boldsymbol{z}^T \boldsymbol{P}_{\boldsymbol{v}_m} \boldsymbol{z} + \boldsymbol{q}_{\boldsymbol{v}_m}^T \boldsymbol{z} + s_{\boldsymbol{v}_m} \leq r \right\}$$

and also from the definition of $\mathcal{D}_r$ we have $\mathcal{D}_r \subseteq \mathcal{E}$. Hence it suffices to find a hyperplane that separates $\boldsymbol{u}$ with $\mathcal{E}$. This is an easy task since we can define ellipsoid surface $\mathcal{S}$ that is parallel to $\mathcal{E}$ and passes through $\boldsymbol{u}$ as follows

$$\mathcal{S} = \left\{ \boldsymbol{z} \mid \boldsymbol{z}^T \boldsymbol{P}_{\boldsymbol{v}_m} \boldsymbol{z} + \boldsymbol{q}_{\boldsymbol{v}_m}^T + s_{\boldsymbol{v}_m} = g \right\}$$

where g is defined in (15) and the tangent hyperplane of $\mathcal{S}$ at $\boldsymbol{u}$ is a separating hyperplane between $\boldsymbol{u}$ and $\mathcal{E}$. To compute the tangent hyperplane we can compute the gradient $\boldsymbol{d} = \nabla_{\boldsymbol{z}} \left(\boldsymbol{z}^T \boldsymbol{P}_{\boldsymbol{v}_m} \boldsymbol{z} + \boldsymbol{q}_{\boldsymbol{v}_m}^T + s_{\boldsymbol{v}_m} \right) |_{\boldsymbol{z}=\boldsymbol{u}}$ and define the following hyperplane

$$\mathcal{H} = \{ \boldsymbol{z} \mid \boldsymbol{d}^T \boldsymbol{z} = \boldsymbol{d}^T \boldsymbol{u} \}. \tag{16}$$

C.2. Proof of Lemma 8

Without loss of generality we assume that we have normalized the space so that $\sum_{i=1}^{n} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} = \boldsymbol{I}$. Using Lemma 6 we have that

$$\begin{aligned}\log\left(\frac{1}{\alpha(\boldsymbol{w}, \boldsymbol{x}^{(i)})}\right) &\leq 2\log\left(\frac{1}{\alpha(\boldsymbol{w}^*, \boldsymbol{x}^{(i)})}\right) + \left|(\boldsymbol{w} - \boldsymbol{w}^*)^T \boldsymbol{x}^{(i)}\right|^2 + 2 \\ &\leq 2\log\left(\frac{1}{\alpha(\boldsymbol{w}^*, \boldsymbol{x}^{(i)})}\right) + \left|y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)}\right|^2 + \left|y^{(i)} - \boldsymbol{w}^{*T}\boldsymbol{x}^{(i)}\right|^2 + 2\end{aligned}$$

Let now $\boldsymbol{v} \in \mathbb{R}^k$ be any unit vector, then we can multiply the above inequality by $\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2$ and sum over all $i \in [n]$ using the fact that we have normalized the space so that $\sum_{i=1}^{n} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} = \boldsymbol{I}$ we get

$$\begin{aligned}\sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \log\left(\frac{1}{\alpha(\boldsymbol{w}, \boldsymbol{x}^{(i)})}\right) \leq 2\sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \log\left(\frac{1}{\alpha(\boldsymbol{w}^*, \boldsymbol{x}^{(i)})}\right) + \\ + \sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \left|y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)}\right|^2 + \\ + \sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \left|y^{(i)} - \boldsymbol{w}^{*T}\boldsymbol{x}^{(i)}\right|^2 + 2\end{aligned}$$

now we can use the fact that $\boldsymbol{w} \in \mathcal{D}_{r^*}$, $\boldsymbol{w}^* \in \mathcal{D}_{r^*}$, Assumption 1 and $\sum_{i=1}^{n} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} = \boldsymbol{I}$ to get

$$\sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \log\left(\frac{1}{\alpha(\boldsymbol{w}, \boldsymbol{x}^{(i)})}\right) \leq 2\log\left(\frac{1}{a}\right) + 2r^* + 2$$

since $\sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 = 1$ this implies that

$$\min_{i\in[n]}\left\{\log\left(\frac{1}{\alpha(\boldsymbol{w}, \boldsymbol{x}^{(i)})}\right)\right\} \leq 2\log\left(\frac{1}{a}\right) + 2r^* + 2$$

which directly implies

$$\max_{i\in[n]}\left\{\alpha(\boldsymbol{w}, \boldsymbol{x}^{(i)})\right\} \geq \text{poly}\left(\frac{1}{a}\right)$$

and the lemma follows.

Appendix D. Omitted Proofs

D.1. Auxiliary Lemmas

The following lemma will be useful in the proof of lemma 10.

Lemma 13 *Let z be a random variable that follows a truncated Gaussian distribution $\mathcal{N}(0, 1, S)$ with survival probability a then there exists a real value $q > 0$ such that*

1. $q \leq 2\log(2/a)$,
2. *the random variable* $z^2 - q$ *is stochastically dominated by a sub-gamma random variable* $u \in \Gamma_+(1,2)$ *with* $\mathbb{E}[u] = 1/2$.

Proof We will prove that the random variable z stochastically dominated by an exponential random variable. First observe that the distribution of z^2 for different sets S is stochastically dominated from the distribution pf z^2 when $S = S^* = S_q = \{z \mid z^2 \geq q\}$, where q is chosen such that $\mathcal{N}(0,1;S^*) = a$. To prove this let F_S be the cumulative distribution function of z^2 when the truncation set is S, we have that $\mathcal{N}(0,1;S) = \mathcal{N}(0,1;S^*) = a$ and hence

$$\mathbb{P}_{z\sim\mathcal{N}(0,1,S)}\left(z^2 \geq t\right) = \frac{1}{a}\mathcal{N}(0,1;S\cap S_t),$$

$$\mathbb{P}_{z\sim\mathcal{N}(0,1,S^*)}\left(z^2 \geq t\right) = \frac{1}{a}\mathcal{N}(0,1;S^*\cap S_t).$$

We now prove that $\mathcal{N}(0,1;S\cap S_t) \leq \mathcal{N}(0,1;S^*\cap S_t)$. If $t \geq q$ then $S^*\cap S_t = S_t$ and hence $S\cap S_t \subseteq S^*\cap S_t$ which implies $\mathcal{N}(0,1;S\cap S_t) \leq \mathcal{N}(0,1;S^*\cap S_t)$. If $t \leq q$ then $S^*\cap S_t = S^*$ and hence

$$\mathcal{N}(0,1;S^*\cap S_t) = \mathcal{N}(0,1;S^*) = \mathcal{N}(0,1;S) \geq \mathcal{N}(0,1;S\cap S_t).$$

Therefore $\mathcal{N}(0,1;S\cap S_t) \leq \mathcal{N}(0,1;S^*\cap S_t)$ and this implies $F_S(t) \geq F_{S^*}(t)$, which implies that the distribution of z^2 for different sets S is stochastically dominated from the distribution pf z^2 when $S = S^*$. Hence we can focus on the distribution of z^2 when $z \sim \mathcal{N}(0,1,S^*)$. First we have to get an upper bound on q. To do so we consider the Q-function of the standard normal distribution and we have that $a = \mathcal{N}(0,1;S^*) = 2Q(\sqrt{q})$. But by Chernoff bound we have that $Q(\sqrt{q}) \leq \exp(-q/2)$ which implies

$$q \leq 2\log\left(\frac{2}{a}\right).$$

Let F_z the cumulative density function of z and F_{z^2} the cumulative density function of z^2, we have that

$$F_{z^2}(t) = F_z(\sqrt{t}) - F_z(-\sqrt{t}),$$

but we know that

$$F_z(t) = \frac{1}{a}\begin{cases} \Phi(t) & t \leq -\sqrt{q} \\ \Phi(-\sqrt{q}) & -\sqrt{q} < t < \sqrt{q} \\ \Phi(t) - (\Phi(\sqrt{q}) - \Phi(-\sqrt{q})) & t \geq \sqrt{q} \end{cases}$$

Hence we have that

$$F_{z^2}(t) = \frac{1}{a}\begin{cases} 0 & t < q \\ \Phi(\sqrt{t}) - \Phi(-\sqrt{t}) - (\Phi(\sqrt{q}) - \Phi(-\sqrt{q})) & t \geq q \end{cases}$$

But we know that $\Phi(\sqrt{t}) - \Phi(-\sqrt{t})$ is the cumulative density function of the square of a Gaussian distribution, namely is a gamma distribution with both parameters equal to $1/2$. This means that $\Phi(\sqrt{t}) - \Phi(-\sqrt{t}) = \int_0^t \frac{\exp(-\tau/2)}{\sqrt{2\pi\tau}} d\tau$ and hence we get

$$F_{z^2}(t) = \frac{1}{a}\begin{cases} 0 & t < q \\ \displaystyle\int_q^t \frac{\exp(-\tau/2)}{\sqrt{2\pi\tau}} d\tau & t \geq q \end{cases}$$

If we define the random variable $v = z^2 - q$ then the cumulative density function F_v of v is equal to

$$F_v(t) = \frac{1}{a}\int_0^t \frac{\exp\left(-\tau+q/2\right)}{\sqrt{2\pi(\tau+q)}}d\tau$$

which implies that the probability density function f_v of v is equal to

$$f_v(t) = \frac{1}{a}\frac{\exp\left(-(t+q)/2\right)}{\sqrt{2\pi(t+q)}}d\tau.$$

It is easy to see that the density of f_v is stochastically dominated from the density of the exponential distribution $g(t) = \frac{1}{2}\exp\left(-\frac{x}{2}\right)$. The reason is that $f_v(t)$ and $g(t)$ are single crossing and $g(t)$ dominates when $t \to \infty$. Then it is easy to see that for the cumulative it holds that $G(t) \leq F_v(t)$. Hence $G(t)$ stochastically dominates $F_v(t)$. Finally we have that $G(t)$ is a sub-gamma $\Gamma_+(1,2)$ and hence v is also sub-gamma $\Gamma_+(1,2)$ and the claim follows. ■

The following lemma lower bounds the variance of $z \sim \mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}, 1, S)$, and will be useful for showing strong convexity of the log likelihood for all values of the parameters in the projection set.

Lemma 14 *Let $\boldsymbol{x} \in \mathbb{R}^k$, $\boldsymbol{w} \in \mathbb{R}^k$, and $z \sim \mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}, 1, S)$ then*

$$\mathbb{E}\left[\left(z - \mathbb{E}\left[z\right]\right)^2\right] \geq \frac{\alpha(\boldsymbol{w},\boldsymbol{x})^2}{12}.$$

Proof We want to bound the following expectation: $\lambda_1 = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I},S)}[(x_1-\mu_{S,1})^2] = \mathrm{Var}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I},S)}[g_1]$, where g_1 denotes the marginal distribution of g along the direction of $\boldsymbol{e_1}$. Since $\mathcal{N}(\boldsymbol{0},\boldsymbol{I};S) = \alpha$, the worst case set (i.e the one that minimizes $\mathrm{Var}[g_1]$) is the one that has α mass as close as possible to the hyperplane $x_1 = \mu_{S,1}$. However, the maximum mass that a gaussian places at the set $\{x_1 : |x_1 - \mu_{S,1}| < c\}$ is at most $2c$ as the density of the univariate gaussian is at most 1. Thus the $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I},S)}[(x_1 - \mu_{S,1})^2]$ is at least the variance of the uniform distribution $U[-\alpha/2, \alpha/2]$ which is $\alpha^2/12$. Thus $\lambda_i \geq \lambda_1 \geq \alpha^2/12$. ■

D.2. Proof of Lemma 10

As we explained before, without loss of generality we can assume $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} = \boldsymbol{I}$. Hence, we have to prove that

$$\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\varepsilon}^{(i)}\right)^2 \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} \preceq r\boldsymbol{I},$$

where $\boldsymbol{\varepsilon}^{(i)} = y^{(i)} - \boldsymbol{w}^{*T}\boldsymbol{x}^{(i)}$ and $r = \log\left(1/a\right)$. Observe that $\boldsymbol{\varepsilon}^{(i)}$ is a truncated standard normal random variable with the following truncation set

$$S^{(i)} = \left\{\boldsymbol{z} \mid \left(\boldsymbol{z} + \boldsymbol{w}^{*T}\boldsymbol{x}^{(i)}\right) \in S\right\}.$$

Therefore we have to prove that for any unit vector $\boldsymbol{v}$ it holds that

$$\frac{1}{n}\sum_{i=1}^n \left(\boldsymbol{\varepsilon}^{(i)}\right)^2 \left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \leq r.$$

For start we fix a unit vector $\boldsymbol{v}$ and we want to bound the following probability

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\varepsilon}^{(i)}\right)^2\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \geq t\right).$$

This means that we are interested in the independent random variables $\left(\boldsymbol{\varepsilon}^{(i)}\right)^2$. Let $q^{(i)}$ the real value that is guaranteed from Lemma 13 and corresponds to $\boldsymbol{\varepsilon}^{(i)}$ and $\boldsymbol{u}^{(i)}$ the corresponding random variable guaranteed from Lemma 13. We also define $\boldsymbol{\delta}^{(i)} = \left(\boldsymbol{\varepsilon}^{(i)}\right)^2 - q^{(i)}$, finally set $a^{(i)} = \alpha(\boldsymbol{w}^*, \boldsymbol{x}^{(i)})$. We have that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\varepsilon}^{(i)}\right)^2\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 &\leq 2\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}^{(i)}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 + 2\frac{1}{n}\sum_{i=1}^{n}q^{(i)}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \\ &\leq 2\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}^{(i)}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 + 4\frac{1}{n}\sum_{i=1}^{n}\log\left(2/a^{(i)}\right)\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2\end{aligned}$$

but from Assumption 1 this implies

$$\leq 2\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}^{(i)}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 + 4\log\left(2/a\right).$$

Now the random variables $\boldsymbol{\delta}^{(i)}$ are stochastically dominated by $\boldsymbol{v}^{(i)}$ and hence

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}^{(i)}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \geq t\right) \leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^{(i)}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \geq t\right)$$

But from the fact that $\mathbb{E}\left[\boldsymbol{v}^{(i)}\right] = 1/2$ we have that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}^{(i)}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 &\leq \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{u}^{(i)} - \mathbb{E}\left[\boldsymbol{u}^{(i)}\right]\right)\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{u}^{(i)} - \mathbb{E}\left[\boldsymbol{u}^{(i)}\right]\right)\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 + 1\end{aligned}$$

Therefore we have that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\varepsilon}^{(i)}\right)^2\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \geq t + 4\log\left(2/a\right) + 1\right) \leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{u}^{(i)} - \mathbb{E}\left[\boldsymbol{u}^{(i)}\right]\right)\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \geq t\right). \tag{17}$$

To bound the last probability we will use the matrix analog of Bernstein's inequality as we expressed in the preliminary section in Theorem 18. More precisely we will bound the following probability

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{u}^{(i)} - \mathbb{E}\left[\boldsymbol{u}^{(i)}\right]\right)\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}\right) \geq t\right).$$

We set $\boldsymbol{Z}_i = \frac{1}{n}\left(\boldsymbol{u}^{(i)} - \mathbb{E}\left[\boldsymbol{u}^{(i)}\right]\right)\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}$. Observe that

$$\left(\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}\right)^p = \left\|\boldsymbol{x}^{(i)}\right\|_2^{2(p-2)}\left(\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}\right)^2.$$

Also since $\boldsymbol{u}^{(i)}$ is an exponential random variable with rate $1/2$, it is well known that

$$\mathbb{E}\left[\left(\boldsymbol{u}^{(i)} - \mathbb{E}\left[\boldsymbol{u}^{(i)}\right]\right)^p\right] \leq \frac{p!}{2}2^p.$$

These two relations imply that

$$\mathbb{E}\left[\boldsymbol{Z}_i^p\right] \preceq \frac{p!}{2}\left(2\frac{\left\|\boldsymbol{x}^{(i)}\right\|_2^2}{n}\right)^{p-2}\left(\frac{2}{n}\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}\right)^2$$

Hence we can apply Theorem 18 with $R = 2\frac{\left\|\boldsymbol{x}^{(i)}\right\|_2^2}{n}$ and $\boldsymbol{A}_i = \frac{2}{n}\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}$. Now observe that by Assumption 2 and the fact that without loss of generality we can have $\boldsymbol{X} = \boldsymbol{I}$, we get that

$$\frac{\left\|\boldsymbol{x}^{(i)}\right\|_2^2}{n} \leq \frac{1}{\log k}.$$

We now compute the variance parameter

$$\sigma^2 = \left\|\sum_{i=1}^n \boldsymbol{A}_i\right\| = 4\frac{\left\|\boldsymbol{x}^{(i)}\right\|_2^2}{n}\left\|\frac{1}{n}\sum_{i=1^n}\boldsymbol{x}^{(i)}\right\|\boldsymbol{x}^{(i)T} \leq \frac{4}{\log k}.$$

Using the same reasoning we get $R \leq 2/\log k$, therefore applying Theorem 18 we get that

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}\left(\sum_{i=1}^n \boldsymbol{Z}_i \geq t\right)\right) &\leq k\exp\left(-\frac{t^2}{2\sigma^2 + Rt}\right)\\ &\leq k\exp\left(-\frac{t^2\log k}{8+2t}\right)\end{aligned}$$

From the last inequality we get that for $t \geq 5$ there is at most probability $1/3$ such that

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n\left(\boldsymbol{u}^{(i)} - \mathbb{E}\left[\boldsymbol{u}^{(i)}\right]\right)\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}\right) \geq t\right) \leq 1/3.$$

Then the lemma follows.

D.3. Proof of Lemma 6

We have that

$$\begin{aligned}
\alpha(\boldsymbol{w}, \boldsymbol{x}) &= \mathbb{E}_{y\sim\mathcal{N}(\boldsymbol{w}^T\boldsymbol{x},\boldsymbol{I})}[\mathbb{1}_{y\in S}] \\
&= \mathbb{E}_{y\sim\mathcal{N}(\boldsymbol{w}'^T\boldsymbol{x},\boldsymbol{I})}[\mathbb{1}_{y\in S}\exp(\left\|y-\boldsymbol{w}'^Tx\right\|^2/2 - \left\|y-\boldsymbol{w}^Tx\right\|^2/2)] \\
&= \alpha(\boldsymbol{w}', \boldsymbol{x})\cdot\mathbb{E}_{y\sim\mathcal{N}(\boldsymbol{w}'^T\boldsymbol{x},1,S)}[\exp(\frac{1}{2}\left\|y-\boldsymbol{w}'^Tx\right\|^2 - \frac{1}{2}\left\|y-\boldsymbol{w}^Tx\right\|^2)] \\
&\geq \alpha(\boldsymbol{w}', \boldsymbol{x})\cdot\exp(\mathbb{E}_{y\sim\mathcal{N}(\boldsymbol{w}'^T\boldsymbol{x},1,S)}[\frac{1}{2}\left\|y-\boldsymbol{w}'^Tx\right\|^2 - \frac{1}{2}\left\|y-\boldsymbol{w}^Tx\right\|^2]) \\
&\geq \alpha(\boldsymbol{w}', \boldsymbol{x})\cdot\exp(\mathbb{E}_{y\sim\mathcal{N}(\boldsymbol{w}'^T\boldsymbol{x},1,S)}[-\frac{1}{2}\left\|y-\boldsymbol{w}'^Tx\right\|^2 - \left|\left(\boldsymbol{w}-\boldsymbol{w}'\right)^T\boldsymbol{x}\right|^2]) \\
&\geq \exp\left(-\left|\left(\boldsymbol{w}-\boldsymbol{w}'\right)^T\boldsymbol{x}\right|^2 - 2\log\left(\frac{1}{\alpha(\boldsymbol{w}', \boldsymbol{x})}\right) - 2\right).
\end{aligned}$$

which implies the desired bound. The first inequality follows from Jensen's inequality. The second follows from the fact that $\|u-v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$. The final inequality follows from Lemma 7.

D.4. Proof of Lemma 7

Let $\alpha(\boldsymbol{w}, \boldsymbol{x}) = c$ for some fixed constant $c < 1$. Note that, $\mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}, 1, S)$ is a truncated version of the normal distribution $\mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}, 1)$. Assume, without loss of generality that $\boldsymbol{w}^T\boldsymbol{x} = 0$. Our goal is to upper bound the second moment of this distribution around 0. It is clear that for this moment to be maximized, we need to choose a set S of measure c that is located as far from μ as possible. Thus, the worst case set $S = (-\infty, -z]\cup[z, \infty)$ consists of both tails of $\mathcal{N}(0, \boldsymbol{I})$, each having mass $c/2$. We know that for the CDF of the normal distribution, the following holds:

$$\Phi(z) \geq 1 - e^{-\frac{z^2}{2}}$$

and we need to find the point z such that $\Phi(z) = 1 - \frac{c}{2}$. Thus, we have:

$$\frac{c}{2} \leq e^{-\frac{z^2}{2}} \Leftrightarrow z \leq \sqrt{2\log\left(\frac{2}{c}\right)}$$

It remains to upper bound $\mathbb{E}_{y\sim\mathcal{N}(0,1,S)}\left[(y-\boldsymbol{w}^T\boldsymbol{x})^2\right]$.

For the aforementioned worst case set S, this quantity is equal to the second non-central moment around $\boldsymbol{w}\boldsymbol{x}$ of the truncated Gaussian distribution in the interval $[z, \infty)$. Thus,

$$\mathbb{E}_{y\sim\mathcal{N}(0,1,S)}\left[(y-\boldsymbol{w}^T\boldsymbol{x})^2\right] = 1 + \frac{z\phi(z)}{1-\Phi(z)}$$

The term $M(z) = \frac{\phi(z)}{1-\Phi(z)}$ is the inverse Mills ratio which is bounded by $z \leq M(z) \leq z + 1/z$ for $z > 0$, see Gordon (1941).

Thus, $z^2 + 1 \leq \mathbb{E}_{y\sim\mathcal{N}(0,1,S)}\left[(y-\boldsymbol{w}^T\boldsymbol{x})^2\right] \leq z^2 + 2 \leq 2\log\left(\frac{2}{c}\right) + 2$.

D.5. Proof of Theorem 9

We now use our Lemmas 6, 7, 10 to prove some useful inequalities that we need to prove our strong convexity and bounded step variance.

Lemma 15 *Let $(\boldsymbol{x}^{(1)}, y^{(i)}), \dots, (\boldsymbol{x}^{(n)}, y^{(n)})$ be n samples from the linear regression model (3) with parameters $\boldsymbol{w}^*$. If Assumptions 1 and 2 hold, then*

$$\frac{1}{n}\sum_{i=1}^{n}\left\|((\boldsymbol{w}-\boldsymbol{w}^*)^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2 \leq 4r^*k, \tag{18}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(y^{(i)}-\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right] \leq 2\log\left(\frac{1}{a}\right)+4, \tag{19}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(z^{(i)}-\boldsymbol{w}^{T}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right] \leq 5r^*k+6, \tag{20}$$

where $y^{(i)} \sim \mathcal{N}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)}, 1, S)$, $z^{(i)} \sim \mathcal{N}(\boldsymbol{w}^{T}\boldsymbol{x}^{(i)}, 1, S)$ and $\boldsymbol{w} \in D_{r^*}$ with $r^* = 4\log(2/a)+7$.

Proof [Proof of Lemma 15.] Without loss of generality we assume that $\frac{1}{n}\sum_{i=1}\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T} = \boldsymbol{I}$ by multiplying our space with $\boldsymbol{X}^{-1}$.
Proof of (18). It is easy to see that

$$\frac{1}{n}\sum_{i=1}^{n}\left\|((\boldsymbol{w}-\boldsymbol{w}^*)^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2 \leq \frac{2}{n}\sum_{i=1}^{n}\left\|(y^{(i)}-\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2+\frac{2}{n}\sum_{i=1}^{n}\left\|(y^{(i)}-\boldsymbol{w}^{T}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2$$

now the first term of the right hand side is less than r^*k because of Lemma 10 and the second term is less than r^*k because of the assumption that $\boldsymbol{w} \in D_{r^*}$, hence (18) follows.
Proof of (19). Using Lemma 7 we get that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(y^{(i)}-\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right] \leq \frac{1}{n}\sum_{i=1}^{n}\left(2\log\left(\frac{1}{\alpha(\boldsymbol{w}^{*T},\boldsymbol{x}^{(i)})}\right)+4\right)\left\|\boldsymbol{x}^{(i)}\right\|_2^2$$

but now we can use Assumption 1 on the right hand side of the above and we get that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(y^{(i)}-\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right] \leq 2\log\left(\frac{1}{a}\right)+4$$

and (19) follows.
Proof of (20). Using Lemma 7 we get that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(z^{(i)}-\boldsymbol{w}^{T}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right] \leq \frac{1}{n}\sum_{i=1}^{n}\left(2\log\left(\frac{1}{\alpha(\boldsymbol{w},\boldsymbol{x}^{(i)})}\right)+4\right)\left\|\boldsymbol{x}^{(i)}\right\|_2^2.$$

Using Lemma 6 on the last inequality we get that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(z^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right] \leq \frac{1}{n}\sum_{i=1}^{n}\left(2\log\left(\frac{1}{\alpha(\boldsymbol{w}^*,\boldsymbol{x}^{(i)})}\right)+4\right)\left\|\boldsymbol{x}^{(i)}\right\|_2^2 + \\ + \frac{1}{n}\sum_{i=1}^{n}\left\|((\boldsymbol{w}-\boldsymbol{w}^*)^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2+2$$

Now using (18) and Assumption 1 we get that

$$\leq 2\log\left(\frac{1}{a}\right)+4+4r^*k+2 \\ \leq 5r^*k+6,$$

and (20) follows. ■

Now we are ready to prove Theorem 9.

Proof [Proof of Theorem 9] Without loss of generality we assume that $\frac{1}{n}\sum_{i=1}\boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)T}=\boldsymbol{I}$ by multiplying our space with $\boldsymbol{X}^{-1/2}$.

Proof of (8). Observe that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(y^{(i)}-z^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right] \leq \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(y^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right]+\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|(z^{(i)}-\boldsymbol{w}^T\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)}\right\|_2^2\right].$$

We can now use (18) and (19) to prove (8).

Proof of (9). Let $\boldsymbol{v}$ be an arbitrary unit vector in $\mathbb{R}^k$, using Lemma 14 from Appendix D.1 and then Lemma 6, we have that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(z^{(i)}-\mathbb{E}\left[z^{(i)}\right]\right)^2\right]\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \geq \frac{1}{n}\sum_{i=1}^{n}\alpha^2(\boldsymbol{w},\boldsymbol{x}^{(i)})\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \\ \geq \frac{e^{-2}}{n}\sum_{i=1}^{n}\exp\left(-\left\|(\boldsymbol{w}-\boldsymbol{w}^*)^T\boldsymbol{x}^{(i)}\right\|_2^2-2\log\left(\frac{1}{\alpha(\boldsymbol{w}^*,\boldsymbol{x}^{(i)})}\right)\right)\cdot \\ \cdot\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2$$

since it holds that $\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2=1$ we can now apply Jensen's inequality and we get

$$\geq \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\left(\left\|(\boldsymbol{w}-\boldsymbol{w}^*)^T\boldsymbol{x}^{(i)}\right\|_2^2+2\log\left(\frac{1}{\alpha(\boldsymbol{w}^*,\boldsymbol{x}^{(i)})}\right)\right)\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2\right)$$

now applying Assumption 1 we get

$$\geq a^2 \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\left\|(\boldsymbol{w}-\boldsymbol{w}^*)^T\boldsymbol{x}^{(i)}\right\|_2^2\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2\right)$$
$$\geq a^2 \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\left\|y^{(i)}-\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)}\right\|_2^2\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2\right)\cdot$$
$$\cdot \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\left\|y^{(i)}-\boldsymbol{w}^{T}\boldsymbol{x}^{(i)}\right\|_2^2\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2\right)$$

now we can use the fact the both $\boldsymbol{w}$ and $\boldsymbol{w}^*$ belong to the projection set D_{r^*} as we showed in Lemma 10 and we get

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(z^{(i)}-\mathbb{E}\left[z^{(i)}\right]\right)^2\right]\left(\boldsymbol{v}^T\boldsymbol{x}^{(i)}\right)^2 \geq a^2\exp\left(-2r^*\right)$$

and the relation (9) follows. ■

Appendix E. Assumptions in the Bayesian Setting

Let us assume that each $\boldsymbol{x}^{(i)}$ is sampled from the standard k-dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$. Let also $\mathcal{D}$ be the ex-ante probability distribution of $\boldsymbol{x}$'s after the truncation procedure described in (3). We will sketch the reduction of our main Assumption 1 and 2 to some simpler and more intuitive assumptions in this Bayesian setting.

It is not hard to see that Assumption 1 can be reduced to

Assumption 3 *Then there exists a positive constant a such that*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(\frac{1}{\alpha(\boldsymbol{x},\boldsymbol{w}^*)}\right)\right]\leq\log\left(\frac{1}{a}\right).$$

To see this observe that Assumption 1 in this setting becomes

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(\frac{1}{\alpha(\boldsymbol{x},\boldsymbol{w}^*)}\right)\boldsymbol{x}\boldsymbol{x}^T\right]\preceq\log\left(\frac{1}{a}\right)\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{x}\boldsymbol{x}^T\right].$$

which is not difficult to prove using the fact that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{x}\boldsymbol{x}^T\right]\succeq c\boldsymbol{I} \tag{21}$$

which holds under the Assumption 3.

The situation with Assumption 2 is even better. In this case the assumption can be proven if the following is true

$$\mathbb{P}\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\boldsymbol{x}\boldsymbol{x}^T\right]\succ\frac{\log k}{tn}\boldsymbol{x}\boldsymbol{x}^T\right)\leq\exp(-t).$$

Now using (21) this becomes

$$\mathbb{P}\left(c\boldsymbol{I} \succ \frac{\log k}{tn}\boldsymbol{x}\boldsymbol{x}^T\right) \leq \exp(-t)$$

which can be reduced to

$$\mathbb{P}\left(c > \frac{\log k}{tn}\left\|\boldsymbol{x}\right\|_2^2\right) \leq \exp(-t)$$

the last inequality it can be shown to be correct using Assumption 21 and the fact the $n = \omega(k \log k)$.

Appendix F. Useful Concentration Results

The following lemma is useful in cases where one wants to show concentration of a weighted sum of i.i.d *sub-gamma* random variables.

Definition 16 *(SUB-GAMMA RANDOM VARIABLES) A random variable x is called sub-gamma random variable if it satisfies*

$$\log\left(\mathbb{E}\left[\exp(\lambda x)\right]\right) \leq \frac{\lambda^2 v}{2(1-c\lambda)} \quad \forall \lambda \in [0, 1/c]$$

for some positive constants v, c. We call $\Gamma_+(v,c)$ the set of all sub-gamma random variables.

Theorem 17 (Section 2.4 of Boucheron et al. (2013).) *Let $X_1, \ldots, X_n$ independent identically distributed random variables such that $X_i \in \Gamma_+(v,c)$ and let $\boldsymbol{a} \in \mathbb{R}_+^d$. Then, the following inequalities hold for any $t \in \mathbb{R}_+$.*

$$\mathbb{P}\left(\sum_{i=1}^d a_i\left(X_i - \mathbb{E}\left[X_i\right]\right) \geq \left\|\boldsymbol{a}\right\|_2\sqrt{2vt} + \left\|\boldsymbol{a}\right\|_\infty ct\right) \leq \exp(-t),$$

$$\mathbb{P}\left(-\sum_{i=1}^d a_i\left(X_i - \mathbb{E}\left[X_i\right]\right) \geq \left\|\boldsymbol{a}\right\|_2\sqrt{2vt} + \left\|\boldsymbol{a}\right\|_\infty ct\right) \leq \exp(-t).$$

We also need a matrix concentration inequality analog to the Bernstein inequality for real valued random variables. For a proof of this inequality we refer to Section 6.2 of Tropp (2012).

Theorem 18 (Theorem 6.2 of Tropp (2012)) *Consider a finite sequence $\{\boldsymbol{Z}_i\}$ of independent, random, self-adjoint matrices with dimension k. Assume that*

$$\mathbb{E}\left[\boldsymbol{Z}_i\right] = \boldsymbol{0} \quad \textit{and} \quad \mathbb{E}\left[\boldsymbol{Z}_i^p\right] \preceq \frac{p!}{2}\cdot R^{p-2}\boldsymbol{A}_i^2 \quad \textit{for } p = 2, 3, 4, \ldots.$$

Compute the variance parameter

$$\sigma^2 := \left\|\sum_i \boldsymbol{A}_i^2\right\|.$$

Then the following chain of inequalities holds for all $t \geq 0$.

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_i \boldsymbol{Z}_i\right) \geq t\right) \leq k \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt}\right) \tag{i}$$

$$\leq \begin{cases} k \cdot \exp(-t^2/4\sigma^2) & \textit{for } t \leq \sigma^2/R; \\ k \cdot \exp(-t/4R) & \textit{for } t \geq \sigma^2/R. \end{cases} \tag{ii}$$

Appendix G. Comparison of Techniques with Previous Work

Estimating the parameters of a truncated linear regression model is a very challenging problem with very high practical relevance. Although similar approaches for solving estimation problems have been followed before, e.g. Daskalakis et al. (2018) used this method to estimate the parameters of a truncated multidimensional Gaussian distribution, the analysis of the truncated linear regression problem is very challenging in a lot of different ways.

1. It is very challenging to define the appropriate projection set $\mathcal{D}_t$ that is needed for the SGD to work. For this step we define a carefully selected positive semidefinite cone and we provide an efficient algorithmic procedure for projecting on this cone.

2. The initial estimation requires special non-trivial treatment, as we have to prove that the initial estimator is contained in the projection set. For this step, we use the least squares regression for our initial guess and we prove using some strong matrix concentration inequalities that it lies in the positive semidefinite cone that we defined in the previous step. The proof of this part involves proving some interesting properties of the least squares regression when applied to truncated samples.

3. Most importantly, the proof of strong convexity in this setting is much more complicated case. One of the main tools that we use to show this result is a strong matrix concentration inequality for matrices with subexponential entries.

Appendix H. Related work On Robust Statistics

We have already discussed work on censored and truncated Statistics. More broadly, our problem falls in the realm of robust Statistics, where there has been a strand of recent works studying robust estimation and learning in high dimensions. A celebrated result by Candès et al. (2011) computes the PCA of a matrix, allowing for a constant fraction of its entries to be adversarially corrupted, but they require the locations of the corruptions to be relatively evenly distributed. Related work of Xu et al. (2010) provides a robust PCA algorithm for arbitrary corruption locations, requiring at most 50% of the points to be corrupted.

Diakonikolas et al. (2016); Lai et al. (2016); Diakonikolas et al. (2017, 2018) do robust estimation of the parameters of multi-variate Gaussian distributions in the presence of arbitrary corruptions to a small ε fraction of the samples, allowing for both deletions of samples and additions of samples that can also be chosen adaptively (i.e. after seeing the sample generated by the Gaussian). The authors in Charikar et al. (2017) show that corruptions of an arbitrarily large fraction of samples can be tolerated as well, as long as we allow "list decoding" of the parameters of the Gaussian. In particular, they design learning algorithms that work when an $(1-\alpha)$-fraction of the samples can be adversarially corrupted, but output a set of $\mathrm{poly}(1/\alpha)$ answers, one of which is guaranteed to be accurate.

Closer to our work in this strand of literature are works studying robust linear regression Bhatia et al. (2015); Diakonikolas et al. (2019) where a small fraction of the response variables are arbitrarily corrupted. As we already discussed in Section 1.2, these results allow arbitrary corruptions yet a small number of them. We only allow filtering out observations, but an arbitrarily large fraction of them.

Other works in this literature include robust estimation under sparsity assumptions Li (2017); Balakrishnan et al. (2017). In Hardt and Moitra (2013), the authors study robust subspace recovery having both upper and lower bounds that give a trade-off between efficiency and robustness. Some general criteria for robust estimation are formulated in Steinhardt et al. (2018).