Efficient Truncated Linear Regression with Unknown Noise Variance

Anonymous Author(s)
Affiliation
Address
email

Abstract

Truncated linear regression is a classical challenge in Statistics, wherein a label, $y = w^T x + \varepsilon$, and its corresponding feature vector, $x \in \mathbb{R}^k$, are only observed if the label falls in some subset $S \subseteq \mathbb{R}$; otherwise the existence of the pair (x, y) is hidden from observation. Linear regression with truncated observations has remained a challenge, in its general form, since the early works of Tobin (1958); Amemiya (1973). When the distribution of the error is normal with known variance, recent work of Daskalakis et al. (2019) provides computationally and statistically efficient estimators of the linear model, w. In this paper, we provide the first computationally and statistically efficient estimators for truncated linear regression when the noise variance is unknown, estimating both the linear model and the variance of the noise. Our estimator is based on an efficient implementation of Projected Stochastic Gradient Descent on the negative log-likelihood of the truncated sample. Importantly, we show that the error of our estimates is asymptotically normal, and we use this to provide explicit confidence regions for our estimates.

1 Introduction

A common challenge facing statistical estimation is the systematic omission of relevant data from the data used to train models. Data omission is a prominent form of dataset bias, which may occur for a variety of reasons, including incorrect experimental design or data collection campaigns, which prevent certain sub-populations from being observed, instrument errors or saturation, which make certain measurements unreliable, societal biases, which may suppress the realization of certain samples, as well as legal or privacy constraints, which might prevent the use of some data. In turn, it is well-understood that failure to account for the systematic omission of data in statistical inference may lead to incorrect models, and this has motivated a long line of research in Statistics, Econometrics, and a rage of other theoretical and applied fields, targeting statistical inference that is robust to missing data; see e.g. Galton (1897); Pearson (1902); Pearson & Lee (1908); Fisher (1931); Tukey (1949); Tobin (1958); Amemiya (1973); Hausman & Wise (1977); Heckman (1979); Breen et al. (1996); Hajivassiliou & McFadden (1998); Mohan et al. (2013); Balakrishnan & Cramer (2014).

In this paper, we revisit the classical challenge of *truncated linear regression*. As in the standard linear regression setting, we assume that the world produces pairs $(x^{(i)}, y^{(i)})_i$, where each $x^{(i)} \in \mathbb{R}^k$ is a feature vector drawn from some distribution $\mathcal{P}$ and where its corresponding label $y^{(i)}$ is linearly related to $x^{(i)}$ according to $y^{(i)} = w^{*T} x^{(i)} + \varepsilon^{(i)}$, for some unknown, but fixed across different i, coefficient vector $w^* \in \mathbb{R}^k$ and some random noise $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^{*2})$, where σ^{*2} is unknown, but fixed across different i, and the $\varepsilon^{(i)}$'s across different i's are independent. The difference to the standard setting is that we do not get to observe all pairs $(x^{(i)}, y^{(i)})_i$ that the world produces. Rather, a pair $(x^{(i)}, y^{(i)})$ is only observed if $y^{(i)} \in S$, where $S \subset \mathbb{R}$ is a fixed "observation set," that is known or given to us via oracle access, i.e. we can query an oracle about whether or not some point

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

belongs to S. If $y^{(i)} \notin S$, then $(x^{(i)}, y^{(i)})$ is removed from observation. Our goal is to estimate w^* and ${\sigma^*}^2$ from the pairs that survived removal, called "truncated samples."

Overview of our results. While both truncated and censored linear regression are fundamental problems, which commonly arise in practice, and which have been studied in the literature since at least the 1950s (Tobin, 1958; Amemiya, 1973), there are no known statistical inference methods that are computationally efficient and enjoy statistical rates that scale gracefully with the dimension k of the problem. Recent work of Daskalakis et al. (2019) has obtained such methods in the simpler setting where the noise variance ${\sigma^*}^2$ is known. Our goal here is to extend those results in two important ways. First, we study settings where ${\sigma^*}^2$ is unknown, providing computationally and statistically efficient estimators for the joint estimation of w^* and ${\sigma^*}^2$. Moreover, we establish the asymptotic normality of our estimators, and use this to derive confidence regions for our estimators.

We next provide informal statements of our results, postponing formal statements to Sections 3 and 5. It is important to note that, as shown in prior work (Daskalakis et al., 2018, 2019), for the model parameters to be identifiable from truncated samples we need to place some assumptions that limit how aggressively the truncation interacts with the data. The assumptions we make are listed below, and stated more precisely in Section 2. They combine normalization assumptions (Assumption 1.1), assumptions that are needed even without truncation (Assumption 1.3), and assumptions that are need for identifiability purposes in the presence of truncation (Assumption 1.2).

Assumption 1.1. We assume that $\|w^*\|_2^2 \leq \beta$ and $\left\|x^{(i)}\right\|_2^2 \leq 1$ for all i.

Assumption 1.2. For all $(x^{(i)}, y^{(i)})$ in the truncated dataset, conditioning on $x^{(i)}$, the probability that ${w^*}^T x^{(i)} + \varepsilon \in S$, with respect to the randomness in $\varepsilon \sim \mathcal{N}(0, {\sigma^*}^2)$, is larger than some a.

Assumption 1.3. If $(x^{(i)}, y^{(i)})_{i=1}^n$ is the truncated dataset, we assume that the feature covariance matrix, $X = \frac{1}{n}\sum_{i=1}^n x^{(i)}{x^{(i)}}^T$, satisfies: (i) $\mathbb{E}_{x\sim\mathcal{P}}[X] = \mathbb{E}_{x\sim\mathcal{P}}[xx^T] \succeq b \cdot I$ for some b; and (ii) there is no component $x^{(i)}{x^{(i)}}^T$ that contributes more than a $n/\log k$ fraction of X in any direction.

With these assumptions we prove our main results for the joint estimation of w^* and ${\sigma^*}^2$, listed as Informal Theorems 1.4–1.5 below, and respectively Theorems 3.2 and 5.1 in later sections.

Informal Theorem 1.4 (Estimation)**.** *Suppose we are given n samples $(x^{(i)}, y^{(i)})_{i=1}^n$ from the truncated linear regression model with parameters w^*, ${\sigma^*}^2$, and suppose Assumptions 1.1–1.3 hold. Then running Projected Stochastic Gradient Descent (Algorithm 1) on the negative log-likelihood of the sample, using the projection set described in Definition 3.1, we obtain with probability at least $2/3$ (which can be boosted to as large a constant as desirable) estimations $\hat{w}, \hat{\sigma}^2$ such that $\|\hat{w} - w^*\|_2 + \left|\hat{\sigma}^2 - {\sigma^*}^2\right| \leq \varepsilon$, as long as $N \geq \min(\frac{C_1}{\varepsilon^4}, \frac{C_2}{\varepsilon^2})$, where C_1 depends polynomially on β, $1/a$ and $1/b$, and C_2 depends linearly in $1/b$ and exponentially in β/a, where β, a and b are respectively the constants in Assumptions 1.1, 1.2 and 1.3.*

Informal Theorem 1.5 (Asymptotic Normality)**.** *Let $\hat{w}$, $\hat{\sigma}^2$ be the estimates of Informal Theorem 1.4. There exists a matrix A that depends on $\hat{w}$, $\hat{\sigma}^2$ (and is explicitly provided in Theorem 5.1) such that the random vector $(\hat{w}, \hat{\sigma}^2)$ converges asymptotically to a normal distribution with mean $(w^*, {\sigma^*}^2)$ and covariance matrix A. Hence the $1-\alpha$ confidence region of $(\hat{w}, \hat{\sigma}^2)$ can be computed using the matrix A and the quantiles of the chi-squared distribution.*

Techniques. Our estimates are computed by running Stochastic Gradient Descent (SGD) on the population negative log-likelihood of the truncated sample, which is a convex function of the parameters after an appropriate reparameterization. We provide two analyses of SGD resulting in two different bounds that are each more effective in a different range of parameters, as we discuss next.

Our first analysis hinges on establishing a lower bound on the convexity of the negative log-likelihood function. We can't however, expect that a strong bound holds over the whole parameter domain. For this reason, we add a projection step to our method, identifying a projection set wherein we show that the true parameters lie and wherein the negative log-likelihood is strongly convex. This approach is similar to that of Daskalakis et al. (2019), except that the unknown noise variance introduces lots of challenges in identifying an appropriate projection set. The main result bounding the strong convexity of the negative log likelihood function within the projection set is stated as Theorem 4.1. The main disadvantage of this analysis is that, when the variance is unknown, the bound on the

strong convexity of the negative log-likelihood becomes small and the resulting dependence of the sample size on the parameters β and $1/a$ of the problem is exponential. On the other hand, the dependence of the sample size on $1/b$ is linear and its dependence on the required estimation accuracy is $\Theta(1/\varepsilon^2)$, which is optimal. Our other analysis follows an approach that has not been pursued in prior work, and in particular does not rely on a global lower bound on the strong convexity of the negative log-likelihood. Instead, we only bound the strong convexity at the optimum and use an upper bound on the smoothness of the log-likelihood function to prove the convergence of SGD in terms of function value. Again, it is impossible to prove a bound on the smoothness over the whole domain, because large noise variance can lead to very bad smoothness bounds. Thus we use the same projection set as above and show that, within this set, we can get an upper bound on the smoothness that depends polynomially on all parameters $\beta, 1/a, 1/b$ of the problem. The main drawback of this method is that the dependence on the accuracy is $\Theta(1/\varepsilon^4)$. It is worth noting that obtaining an algorithm whose sample complexity has an optimal $\Theta(1/\varepsilon^2)$ dependence on the estimation accuracy and also scales polynomially in $1/a$ hits a fundamental bottleneck, which is already present in the much simpler special case of our problem of estimating the variance of a single-dimensional truncated Gaussian; see Lemma 7 of Daskalakis et al. (2018).

Finally, it helps build intuition about the intricacies arising in estimating truncated linear regression models with unknown noise variance to consider why natural approaches of reducing this problem to estimating truncated linear regression models with *known* noise variance – which can be done using the algorithm of Daskalakis et al. (2019) – fail. A simple such reduction might compute the empirical variance σ_0^2 of the truncated sample, and plug that into the algorithm of Daskalakis et al. (2019) to estimate the linear model pretending that σ_0^2 is the true noise variance. A more sophisticated approach might grid over candidate noise variances, plug each of these candidates into the algorithm of Daskalakis et al. (2019) to estimate a linear model, and then select the (linear model, noise variance) pair that is more consistent, e.g. the pair whose noise variance is closest to the empirical variance of the residuals between the linear model's predictions and the observed labels. Both of these approaches fail, as suggested by the example below. Suppose that the ground truth is $y = w^* x + w_0^* + \varepsilon$, where x is one-dimensional, $(w^*, w_0^*) = (0, 0)$, and the noise variance is $\sigma^{*2} = 1$. Suppose also that we create a truncated dataset by, repeatedly, sampling x_i from $\mathcal{N}(0, 1)$, generating y_i according to the ground truth, and truncating the resulting pair (x_i, y_i) unless $y_i \geq \tau$, where τ is some threshold. It is clear that, if we guess the true noise variance to be the empirical one, then as τ gets bigger, the empirical variance gets smaller, thus the linear model that would be estimated by treating the empirical variance as the true noise variance would have intercept that is close to τ rather than 0. So this naive approach would fail. Let us consider the more sophisticated one that grids over candidate noise variances. In particular, let us compare what happens when (1) we guess the noise variance to be $\sigma_1^2 = \sigma^{*2}$, and run the algorithm of Daskalakis et al. (2019) with this guess to get an estimate $(\hat{w}, \hat{w}_0)$; and (2) when we guess the noise variance to be $\sigma_2^2 \equiv \sigma_0^2$ (the empirical variance) and run the algorithm of Daskalakis et al. (2019) with this guess to get an estimate $(\hat{w}', \hat{w}_0')$. It is easy to see that, as τ gets large, $\left|\sigma^{*2} - \frac{1}{N}\sum_i (y_i - \hat{w}x_i - \hat{w}_0)^2\right| \gg \left|\sigma_0^2 - \frac{1}{N}\sum_i (y_i - \hat{w}'x_i - \hat{w}_0')^2\right|$ (even if the number of truncated samples goes to infinity). Thus the selection criterion proposed above will fail to prefer (1) over (2), even though the guess of the noise variance under (1) is the correct one. Our approach in this paper amounts to using a different criterion, namely selecting the pair with the highest log-likelihood. But analyzing when and how well this works, as well as obtaining an efficient algorithm to find the best pair is the main contribution of this work.

Roadmap. We present the truncated linear regression setting and our assumptions in Section 2. We state and prove our main estimation results Section 3 and 4, and prove the asymptotic normality of our estimators in Section 5. Finally, in Section 6 we assess the performance of our methods.

2 Models and assumptions

We study the truncated linear regression setting studied in Daskalakis et al. (2019), with the additional complexity that the variance of the noise model is unknown. In particular, let $S \subset \mathbb{R}$ be a measurable subset of the real line, to which oracle access is provided. Assume we have N truncated samples $(x^{(i)}, y^{(i)})_{i=1}^N$, each generated according to the following procedure, for some unknown $w^* \in \mathbb{R}^k$ and $\sigma^* \in \mathbb{R}$ that are fixed across different i's:

1. sample $x^{(i)}$ from some distribution $\mathcal{P}$ with support $\mathrm{supp}(\mathcal{P}) \subseteq \mathbb{R}^k$;

2. generate $y^{(i)}$ according to the following model, where $\varepsilon \sim \mathcal{N}(0, {\sigma^*}^2)$;

$$y^{(i)} = {w^*}^T x^{(i)} + \varepsilon \tag{2.1}$$

3. if $y^{(i)} \in S$ then return $(x^{(i)}, y^{(i)})$ as the ith sample, otherwise repeat from step 1.

As already established in prior work (Daskalakis et al., 2018, 2019; Kontonis et al., 2019), some assumptions must be placed on S and its interaction with the unknown parameters $(w^*, {\sigma^*}^2)$ and the $x^{(i)}$'s in order for the parameters to be identifiable. We state our assumptions after a useful definition.

Definition 2.1 (Survival Probability)**.** Let $S \subseteq \mathbb{R}$ be measurable, $x, w \in \mathbb{R}^k$, and $\sigma \in \mathbb{R}$. We define the survival probability $\alpha(w, \sigma, x, S)$ as $\alpha(w, \sigma, x, S) = \mathbb{P}\{Y \in S\}$; where $Y \sim \mathcal{N}(w^T x, \sigma^2)$. When S is clear from context we may omit S from the arguments of α and simply write $\alpha(w, \sigma, x)$.

We are now ready to state our assumptions. Our first assumption gives an upper bound on the norm of the true parameters and the feature vectors.

Assumption 2.2 (Normalization/Data Generation)**.** Assume that $\|w^*\|_2^2 < \beta$, and $\|x\|_2^2 < 1$ for all $x \in \mathrm{supp}(\mathcal{P})$, and that we have sample access to the, otherwise unknown, distribution $\mathcal{P}$. We also assume oracle access for set S, i.e. we have an oracle which determines, for any y, whether $y \in S$.

Our next assumption is the survival probability lower bound following Daskalakis et al. (2018, 2019).

Assumption 2.3 (Constant Survival Probability)**.** For all $x \in \mathrm{supp}(\mathcal{P})$, we have $\alpha(w^*, \sigma^*, x) > a$.

Our final assumption is about the spectrum of the covariance matrix of the features and is classical in many linear regression settings.

Assumption 2.4 (Thickness of Feature Covariance)**.** Let $X = \frac{1}{n}\sum_{i=1}^n x^{(i)} {x^{(i)}}^T$ be the $k \times k$ feature vector covariance matrix. We assume that, for some positive real number b, we have $\mathbb{E}_{x\sim\mathcal{P}}[X] = \mathbb{E}_{x\sim\mathcal{P}}[xx^T] \succeq b \cdot I$ and $X \succeq \frac{\log k}{n} x^{(i)} {x^{(i)}}^T, \forall i \in [n]$

3 Log likelihood functions and feasibility

When the variance parameter ${\sigma^*}^2$ is unknown the log-likelihood function of the plain linear regression becomes non-concave and hence it is not straightforward to optimize. For this reason we follow the common practice of re-parameterizing the problem with respect to the parameters $v = w/\sigma^2$ and $\lambda = 1/\sigma^2$. The algorithm that we use is Projected SGD on the population negative log-likelihood function $\bar{\ell}(v, \lambda)$ as presented in Algorithm 1. We formally define the loss function $\bar{\ell}$ in Section 4.1.

Algorithm 1 Projected SGD on $\bar{\ell}$

Input: $(x^{(i)}, y^{(i)})$ for $i \in [n]$, learning rates η_t, convex set D over pairs (v, λ) (from Def 3.1)

$(w_0, \sigma_0^2) \leftarrow$ OLS estimates; $(q, \lambda^{(0)}) \leftarrow (w_0/\sigma_0^2, 1/\sigma_0^2)$ ▷ *(a) Initial Feasible Point*
$v^{(0)} \leftarrow \mathrm{argmin}_{v:(v,\lambda^{(0)})\in D} \|v - q\|_2$
for $t = 1, \ldots, n$ **do** ▷ *(b) Unbiased Gradient Estimation*
  Sample gradient $u^{(t)}$ such that $\mathbb{E}\big[u^{(t)} | (v, \lambda)^{(t-1)}\big] = \nabla\bar{\ell}(v^{(t-1)}, \lambda^{(t-1)})$
  $(v^{(t)}, \lambda^{(t)}) \leftarrow (v^{(t-1)}, \lambda^{(t-1)}) - \eta_t \cdot u^{(t)}$ ▷ *(c) Efficient Projection*
  $(v^{(t)}, \lambda^{(t)}) \leftarrow \arg\min_{(v,\lambda)\in D} \big\|(v, \lambda) - (v^{(t)}, \lambda^{(t)})\big\|_2$
end for
$(\hat{v}, \hat{\lambda}) \leftarrow \big(v^{(n)}, \lambda^{(n)}\big)$; **return** $(\hat{w}, \hat{\sigma}^2) \leftarrow \Big(\hat{v}/\hat{\lambda}, 1/\hat{\lambda}\Big)$

Our goal is to apply the above algorithm to the negative log-likelihood function of the truncated linear regression model. To execute the above algorithm in polynomial time, we need to solve the following algorithmic problems.

(a) **initial feasible point:** compute an initial feasible point in some projection set D,

(b) **unbiased gradient estimation:** sample an unbiased estimation of $\nabla\bar{\ell}(v^{(t-1)}, \lambda^{(t-1)})$,

(c) **efficient projection:** design an algorithm to project to the set D.

Solving the computational problems (a)-(c) is the first step in the proof of our main result. First we define the appropriate projection set to use in our algorithm.

Definition 3.1 (Projection Set)**.** Let $(\bar{x}^{(1)}, \bar{y}^{(1)}), \ldots, (\bar{x}^{(n)}, \bar{y}^{(n)})$ be n samples from the truncated regression problem. We use the notation $(\bar{x}^{(i)}, \bar{y}^{(i)})$ to distinguish these samples from the samples that we use for the rest of the SGD Algorithm 1. We define the projection set be the (v, λ) satisfying

$$D_r = \left\{ (v, \lambda) = \left(\frac{w}{\sigma^2}, \frac{1}{\sigma^2} \right) \in \mathbb{R}^k \times \mathbb{R} \;\middle|\; \frac{1}{n} \sum_{i=1}^{n} \frac{(\bar{y}^{(i)} - w^T \bar{x}^{(i)})^2}{\sigma^2} \bar{x}^{(i)} \bar{x}^{(i)^T} \preceq r\bar{X}, \frac{1}{16r} \leq \frac{\sigma^2}{\sigma_0^2} \leq \frac{576}{a^2}, \|w\|_2^2 < \beta \right\}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}^{(i)} \bar{x}^{(i)^T}$ and σ_0 is the estimated variance from solving the classical linear regression problem on the data ignoring truncation. We typically take $r^* = 12 + 4\log(2/a)$, and we may use D to refer to D_{r^*}.

This projection set contains two parts: the first part that restricts the weights and the second part that restricts the variance. For the second part we use the variance predicted using classical linear regression technique of ordinary least squares.

Now we are ready to formally state our main theorem about the estimation of the parameters w and σ from truncated samples under the Assumptions 2.2- 2.4.

Theorem 3.2. *Let $(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})$ be n samples from the linear regression model (2.1) with parameters $w^*, {\sigma^*}^2$. If Assumptions 2.2- 2.4 hold then there exists an algorithm with success probability $\geq 2/3$, that outputs an estimation $\hat{w}, \hat{\sigma}^2$ such that both the following inequalities hold*

$$\|\hat{w} - w^*\|_2 + \left|\hat{\sigma}^2 - {\sigma^*}^2\right| \leq \frac{\mathrm{poly}(\frac{\sigma_0 \cdot \beta}{a \cdot b}, \frac{1}{\sigma_0}) \cdot \log(n)}{n^{1/4}} \tag{3.1}$$

$$\|\hat{w} - w^*\|_2 + \left|\hat{\sigma}^2 - {\sigma^*}^2\right| \leq \frac{\mathrm{poly}(\sigma_0) \cdot \exp\Big(\mathrm{poly}(\frac{\beta}{a \cdot \sigma_0})\Big)}{b\sqrt{n}} \tag{3.2}$$

where σ_0 is the estimated variance using ordinary least squares. Moreover, the running time of the algorithm is $\mathrm{poly}(n, \sigma^*) \cdot \exp\Big(\mathrm{poly}(\frac{\beta}{a \cdot \sigma^*})\Big)$ and if the survival set S is a union of k intervals then the running time of the algorithm becomes $\mathrm{poly}(k, n, \sigma^*, 1/\sigma^*, 1/a, \beta, 1/b)$.

Remark 3.3. We can prove Theorem 3.2 even if we replace Assumption 2.2 with a weaker one where we do not assume sample access to the distribution $\mathcal{P}$. The only difference in the guarantees is that the first guarantee becomes $\|\hat{w} - w^*\|_2 + \left|\hat{\sigma}^2 - {\sigma^*}^2\right| \leq \frac{\mathrm{poly}(\frac{\sigma_0 \cdot \beta}{a \cdot b}, \frac{1}{\sigma_0}) \cdot \log(n)}{n^{1/8}}$. We refer to the supplementary material for the full statement and proof of Theorem 3.2 with the weaker version of Assumption 2.2. We choose to present the stated version of Theorem 3.2 with the stronger assumption here, because the algorithm and its analysis are more direct and easier to understand and also because this is the same setting for which we prove our confidence intervals in the next section. Arguably access to the distribution $\mathcal{P}$ fits better with the censored regression setting described in Section 1.

Remark 3.4. It is easy to see that the probability of success in Theorem 3.2 can be boosted to $1 - \delta$ with an additional cost of order $\log(1/\delta)$ in the rates. We can achieve this using a folklore boosting technique for parameter estimation problems. In particular, we can run the algorithm of Theorem 3.2 $\log(1/\delta)$ times independently and then we can pick the estimate that is ε close to at least the half of the rest of the estimates, where ε is the target estimation error. It is easy to see that with this boosting technique we can get error at most 2ε with probability of success at least $1 - \delta$.

To prove Theorem 3.2 we provide two different analyses of the Algorithm 1. The first analysis uses only the fact that the stochastic gradients of the population negative log-likelihood function have bounded second moment and is based on the following theorem.

Theorem 3.5 (Theorem 2 of Shamir & Zhang (2013))**.** *Let $\bar{\ell}$ be a convex function with a minimizer (v^*, λ^*) and suppose that there exists a constant ρ such that*

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

(ii) **bounded domain:** $\max_{(v,\lambda) \in D} \|(v - v^*, \lambda - \lambda^*)\|_2^2 \leq \rho^2$,

(iii) **feasibility of optimal:** *the minimizer* $(v^*, \lambda^*) \in D$

then if we apply Algorithm 1 on $\bar{\ell}$ *with learning rates* $\eta_t = c/\sqrt{t}$ *where* c *is an absolute constant then for every* $t \in [n]$ *it holds that* $\mathbb{E}\big[\bar{\ell}(v^{(t)}, \lambda^{(t)}) - \bar{\ell}(v^*, \lambda^*)\big] \leq \frac{\mathrm{poly}(\rho)\cdot\log(t)}{\sqrt{t}}$.

For the second analysis of Algorithm 1 we also need to bound the strong convexity of $\bar{\ell}$. This is the reason that we get the exponential dependence on some of the parameters of the problem, but we can also get a better consistency rate. For this second analysis we use the following theorem.

Theorem 3.6 (Lemma 1 of Rakhlin et al. (2011))**.** *Let* $\bar{\ell}$ *be a convex function with a minimizer* (v^*, λ^*) *and suppose that there exist constants* ρ, ζ *such that*

(i) **bounded step variance:** $\mathbb{E}\Big[\big\|u^{(t)}\big\|_2^2\Big] \leq \rho^2$

(iii) **feasibility of optimal:** *the minimizer* $(v^*, \lambda^*) \in D$

(iv) **strong convexity:** $\bar{\ell}$ *is* ζ*-strongly convex*

then if we apply Algorithm 1 on $\bar{\ell}$ *with learning rates* $\eta_t = 1/(\zeta \cdot t)$ *then for every* $t \in [n]$ *it holds that*

$$\mathbb{E}\left[\left\|(v^{(t)}, \lambda^{(t)}) - (v^*, \lambda^*)\right\|_2^2\right] \leq \frac{4\rho^2}{\zeta^2 \cdot t}.$$

In sum, the main challenges for proving Theorem 3.2 are solving computational problems (a)–(c) and proving mathematical properties (i)–(iv).

4 Overview of the proof of Theorem 3.2.

The steps that we follow in the proof of Theorem 3.2, are the following.

1. In Section 4.1 we derive the negative log-likelihood, its gradient, and its Hessian matrix.
2. In Section 4.2 we discuss the computational problems (a) - (c).
3. In Section 4.3 we formally state our results that prove the properties (i), (ii), and (iv).
4. In Section 4.4 we state our results for property (iii), the feasibility of the optimal solution.
5. Finally in Section 4.5 we use all the above results to complete the proof of 3.2.

4.1 Log-Likelihood of Truncated Linear Regression

Given a sample (x, y), the negative log-likelihood that this is a sample from the truncated linear regression with parameters w, σ^2 is equal to

$$\ell(w, \sigma^2; x, y) = \frac{1}{2\sigma^2}(y^2 - 2yw^Tx) + \log\left(\int_S \exp(-\frac{1}{2\sigma^2}(z^2 - 2zw^Tx))\mathrm{d}z\right)$$

If we reparameterize the negative log-likelihood with $\lambda = 1/\sigma^2$ and $v = w/\sigma^2$, we have

$$\ell(v, \lambda; x, y) = \frac{1}{2}(\lambda y^2 - 2yv^Tx) + \log\left(\int_S \exp(-\frac{1}{2}(\lambda z^2 - 2zv^Tx))\mathrm{d}z\right) \tag{4.1}$$

We define the distributions $F_x = \mathcal{N}({w^*}^Tx, {\sigma^*}^2, S)$ and $Q_x = \mathcal{N}(w^Tx, \sigma^2, S)$, and the negative population log-likelihood function $\bar{\ell}$ as follows

$$\bar{\ell}(v, \lambda) = \mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{y\sim F_x}[\ell(v, \lambda; x, y)]\right]. \tag{4.2}$$

We can then easily compute

$$\frac{\partial\bar{\ell}}{\partial v} = \mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{z\sim Q_x}[z\cdot x] - \mathop{\mathbb{E}}_{y\sim F_x}[y\cdot x]\right] \quad \frac{\partial\bar{\ell}}{\partial\lambda} = \frac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left[\mathop{\mathbb{E}}_{y\sim F_x}[y^2] - \mathop{\mathbb{E}}_{z\sim Q_x}[z^2]\right],$$

and also From these we also compute the Hessian matrix which is

$$\mathbf{H}(v, \lambda) = \mathop{\mathbb{E}}_{x\sim\mathcal{P}}[\mathbf{H}_x(v, \lambda)] = \begin{bmatrix} \mathrm{Var}_{z\sim D_x}[z]xx^T & -\mathrm{Cov}_{z\sim D_x}\big[\frac{1}{2}z^2, z\big]x \\ -\mathrm{Cov}_{z\sim D_x}\big[\frac{1}{2}z^2, z\big]x^T & \mathrm{Var}_{z\sim D_x}\big[\frac{1}{2}z^2\big] \end{bmatrix} \tag{4.3}$$

Here, $\mathbf{H}$ is the expectation of covariance matrices, which is positive definite. Thus, $\bar{\ell}$ is convex.

4.2 Computational Problems

The computational problems (a) - (c) can be tackled in the following way

(a) We start with computing (w_0, σ_0) by applying OLS to the truncated data.

(b) As we see from the expression of the gradient of $\bar{\ell}$, we can compute an unbiased estimate of the gradient using one of our samples $(x^{(i)}, y^{(i)})$, our sample access to the distribution $\mathcal{P}$ of x's, and by doing rejection sampling until we find a point with (x, z) with $z \in S$. To bound the time needed for this step we need to understand what is the mass of the survival set S for some $x \sim \mathcal{P}$ and some vector of parameters $(v, \lambda) \in D$. For this we refer to Section A.1.

(c) We show in the supplementary material that because D is a convex set we can define an efficient algorithm to project to D.

For more details on the way we solve the computational problems (a) - (c) we refer to the supplementary material.

4.3 Bounded Step Variance, Bounded Domain, and Strong Convexity

In this section we present the statements that establish the properties (i), (ii), (iv) that we need in order to apply Theorem 3.5 and Theorem 3.6.

Theorem 4.1. *For every $(v, \lambda) \in D$, we have*

$$\mathbb{E}\Big[\|(y-z)x\|_2^2\Big] \leq \mathrm{poly}(1/a)(1+\beta^2+\sigma_0^4), \;\; \mathbf{H}(v,\lambda) \succeq b\cdot\exp\left(-\mathrm{poly}(1/a)(1+\frac{\beta}{\sigma_0^2})\right)\begin{pmatrix}\sigma_0^2 I & 0\\ 0 & \sigma_0^4\end{pmatrix}$$

where, $x \sim \mathcal{P}$, $y \sim F_x$ and $z \sim Q_x$ and $\mathbf{H}$ from (4.3).

Also, it is easy to see that the diameter of D is bounded:

$$\|v\|_2 < \|w\|_2/\sigma^2 \leq 16(12+4\log(2/a))\sqrt{\beta}/\sigma_0^2 \quad \text{and} \quad |\lambda| < 1/\sigma^2 \leq 16(12+4\log(2/a))/\sigma_0^2$$

4.4 Feasibility of Optimal Solution

We need to calculate the probability for the initial point to be in the projection set, i.e., $w^*, \sigma^* \in D$.

Lemma 4.2. *Under Assumptions 2.3 and 2.2 we have that, $\mathbb{P}((w^*, \sigma^*) \in D) \geq 27/40$.*

4.5 Proof of Theorem 3.2

The algorithm that proves Theorem 3.2 starts with splitting the samples n in three parts of equal size. The first part of $n/3$ samples is used to define the projection set D, and is used at every step where a projection to D is needed. The second part of $n/3$ samples is used to compute the OLS estimates w_0, σ_0, where w_0 is used in the initialization step and σ_0 is used both in the initialization and in the definition of D. Finally, the third part of the $n/3$ samples is used in our main SGD algorithm. For the rest of the proof we abuse the notation and we use $n \leftarrow n/3$ for simplicity. The next thing that we need to decide is the learning rates that we are going to use for the SGD. To do this we compare the two different guarantees of our algorithm $A_1 = \frac{\mathrm{poly}(\frac{\sigma_0\cdot\beta}{a\cdot b}, \frac{1}{\sigma_0})\cdot\log(n)}{n^{1/4}}$, $A_2 = \frac{\mathrm{poly}(\sigma_0)\cdot\exp\left(\mathrm{poly}(\frac{\beta}{a\cdot\sigma_0})\right)}{b\sqrt{n}}$ for the specific values of the parameters of the current instantiation. If $A_1 \leq A_2$ then we use rates $\eta_t = c/\sqrt{t}$, otherwise we use $\eta_t = 1/(\zeta \cdot t)$.

Case $\eta_t = c/\sqrt{t}$. First, combining Theorem 3.5 and Theorem 4.1, we get the following corollary.

Corollary 4.3. *Suppose the true parameters $(v^*, \lambda^*) = (w^*/{\sigma^*}^2, 1/{\sigma^*}^2) \in D$ and $\bar{\ell}$ is defined as in (4.2). Then, for the output $\hat{v}, \hat{\lambda}$ of Algorithm 1: $\mathbb{E}\left[\bar{\ell}(\hat{v}, \hat{\lambda}) - \bar{\ell}(v^*, \lambda^*)\right] \leq \frac{\mathrm{poly}(\frac{\sigma_0\cdot\beta}{a\cdot b}, \frac{1}{\sigma_0})\cdot\log(n)}{n^{1/2}}$.*

Our next step is to transform the optimality in function values to closeness to the true parameters. To do so we use the strong convexity of $\bar{\ell}$ at the optimum.

Lemma 4.4. *The Hessian at the true (v^*, λ^*) satisfies $\mathbf{H}(v^*, \lambda^*) \succeq \frac{b\times\mathrm{poly}(a)}{\mathrm{poly}(1/a, \beta/\sigma_0^2)}\begin{pmatrix}{\sigma_0}^2 I & 0\\ 0 & {\sigma_0}^4\end{pmatrix}$.*

We also utilize the following lemma to calculate an upper bound on $\bar{\ell}(v^*, \lambda^*)$:

Lemma 4.5. (Lemma 3 in Daskalakis et al. (2019)) *Let $x, w \in \mathbb{R}^k$, $\sigma > 0$, then $\mathbb{E}_{y\sim\mathcal{N}(w^T x, \sigma^2, S)}(y - w^T x)^2 \leq (4 - 2\log\alpha(w, \sigma, x))\sigma^2$.*

By the above lemma, we have:

$$\bar{\ell}(v^*, \lambda^*) = \mathop{\mathbb{E}}_{x\in\mathcal{P}} \mathop{\mathbb{E}}_{y} \left[\frac{(y - w^{*T} x)^2}{2\sigma^{*2}} + \log\alpha(w^*, \sigma^*, x) \right] \leq 4 - 2\log \min_{x\in\mathcal{P}} \alpha(w^*, \sigma^*, x) \leq 4 - 2\log a$$

Applying Markov's inequality we have

$$\mathbb{P}\left(\bar{\ell}(\hat{v}, \hat{\lambda}) - \bar{\ell}(v^*, \lambda^*) > 120\mathbb{E}\Big[\bar{\ell}(\hat{v}, \hat{\lambda}) - \bar{\ell}(v^*, \lambda^*)\Big] \right) < \frac{1}{120}. \tag{4.4}$$

The inequality implies that, with at least $119/120$ probability, the actual difference is at most 120 times its expectation. Also, the probability that $(v^*, \lambda^*) \in D$ is at least $27/40$ by Lemma 4.2. Combining these two, we have at least $27/40 - 1/120 = 2/3$ probability to hold:

$$\bar{\ell}(\hat{v}, \hat{\lambda}) - \bar{\ell}(v^*, \lambda^*) < \frac{\mathrm{poly}(\frac{\sigma_0 \cdot \beta}{ab}, \frac{1}{\sigma_0}) \cdot \log(n)}{n^{1/2}} \tag{4.5}$$

Now applying Lemma 4.4, since $\hat{v}, \hat{\lambda}$ converge to v^*, λ^*, when they have small distance, we have

$$\|\hat{v} - v^*\|_2^2 + (\hat{\lambda} - \lambda^*)^2 \leq \frac{2}{\delta}\left(\bar{\ell}(\hat{v}, \hat{\lambda}) - \bar{\ell}(v^*, \lambda^*) \right) \leq \frac{\mathrm{poly}(\frac{\sigma_0 \cdot \beta}{ab}, \frac{1}{\sigma_0}) \cdot \log(n)}{\delta \cdot n^{1/2}} \tag{4.6}$$

where $\mathbf{H}(v^*, \lambda^*) \succeq \delta I$, and $\delta^{-1} = \mathrm{poly}\left(\frac{\sigma_0 \cdot \beta}{a \cdot b}, \frac{1}{\sigma_0}\right)$.

Finally, doing some simple calculations for the reparameterization (see Appendix A.6) we have that

$$\|\hat{w} - w^*\|_2 + \left|\hat{\sigma}^2 - \sigma^{*2}\right| \leq \sqrt{2\left(\|\hat{w} - w^*\|_2^2 + \left|\hat{\sigma}^2 - \sigma^{*2}\right|^2\right)} \leq \mathrm{poly}(1/a, \sigma^*) \left\|(\hat{v}, \hat{\lambda}) - (v^*, \lambda^*)\right\|_2. \tag{4.7}$$

Combining inequalities (4.4), (4.5), (4.6), (4.7), inequality (3.1) of Theorem 3.2 follows.

Case $\eta_t = 1/(\zeta t)$. We combine Theorem 3.6 and Theorem 4.1 to get the following corollary.

Corollary 4.6. *Suppose the true parameters $(v^*, \lambda^*) = (w^*/\sigma^{*2}, 1/\sigma^{*2}) \in D$ and $\bar{\ell}$ is defined as equation (4.2). Then, for $\hat{v}, \hat{\lambda}$ being the output of Algorithm 1, it holds that*

$$\mathbb{E}\left[\left\|(\hat{v}, \hat{\lambda}) - (v^*, \lambda^*)\right\|_2^2 \right] \leq \mathrm{poly}\left(\beta, \sigma_0 + \frac{1}{\sigma_0}, \frac{1}{a}\right) \times \exp\left(\mathrm{poly}(1/a)\left(1 + \frac{\beta}{\sigma_0^2}\right)\right) \frac{1}{b^2 n}.$$

Using similar techniques (details are in Appendix A.6) to those used to prove inequality (3.1) upstairs, combined with the above corollary, we get that with probability $\geq 2/3$ it holds that $\left\|(\hat{v}, \hat{\lambda}) - (v^*, \lambda^*)\right\|_2^2 \leq \frac{\mathrm{poly}(\sigma_0) \cdot \exp\left(\mathrm{poly}(\frac{\beta}{a \cdot (\sigma_0)})\right)}{b^2 \cdot n}$. From the latter we can prove the inequality (3.2) of Theorem 3.2 via simple calculations and using the definition of the projection set D.

5 Inference and Confidence Regions

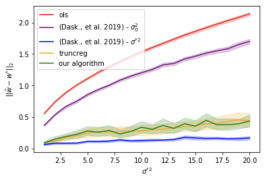
In addition to providing estimates $\hat{w}$, $\hat{\sigma}$ for the truncated linear regression model, we prove the asymptotic normality of our estimates, which we use in turn to obtain confidence regions for our estimates. Our result is summarized in the following theorem whose proof can be found in the supplementary material.

Theorem 5.1. *Let $(\hat{w}, \hat{\sigma}^2)$ be the estimates obtained from Algorithm 1 with learning rates $\eta_t = 1/(\zeta \cdot t)$ for some constant ζ. Then $\sqrt{n}((\hat{w}, \hat{\sigma}^2) - (w^*, \sigma^{*2}))$ is asymptotically normal, with mean zero and covariance matrix $J^{-1^T}\zeta^{-1}\Sigma J^{-1}$, where $J^{-1} = \begin{pmatrix} 1/\hat{\sigma}^2 I & -\hat{w}/\hat{\sigma}^4 \\ 0 & -1/\hat{\sigma}^4 \end{pmatrix}$ is the inverse of*

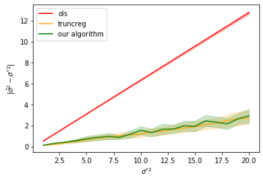
the Jacobian matrix of the mapping from (v, λ) *to* (w, σ^2)*, and* Σ *is the matrix that satisfies the following system of linear equations* $\left(\hat{\mathbf{H}} - \frac{\zeta}{2}I\right)\Sigma + \Sigma\left(\hat{\mathbf{H}} - \frac{\zeta}{2}I\right) = \Gamma(\hat{v}, \hat{\lambda})$ *where* $\hat{\mathbf{H}}$ *is the empirical estimation of the Hessian matrix defined by* (4.3) *at the point* $(\hat{v}, \hat{\lambda})$*,* ζ *is the minimum eigenvalue of* $\hat{\mathbf{H}}$*, and* Γ *is* $\Gamma(v, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial l(v,\lambda;x^{(i)},y^{(i)})}{\partial(v,\lambda)} \frac{\partial l(v,\lambda;x^{(i)},y^{(i)})}{\partial(v,\lambda)}^T$.

Corollary 5.2. *Let* q_α *be the* $1-\alpha$ *quantile of the chi distribution* χ_{k+1}*. Then the* $1-\alpha$ *asymptotic confidence region of* (w, σ^2) *is* $\left(\begin{pmatrix} w \\ \sigma^2 \end{pmatrix} - \begin{pmatrix} \hat{w} \\ \hat{\sigma}^2 \end{pmatrix}\right)^T J^{-T}\Sigma J^{-1} \left(\begin{pmatrix} w \\ \sigma^2 \end{pmatrix} - \begin{pmatrix} \hat{w} \\ \hat{\sigma}^2 \end{pmatrix}\right) \leq \frac{q_\alpha}{\zeta n}$.

6 Experiments



(a)



(b)

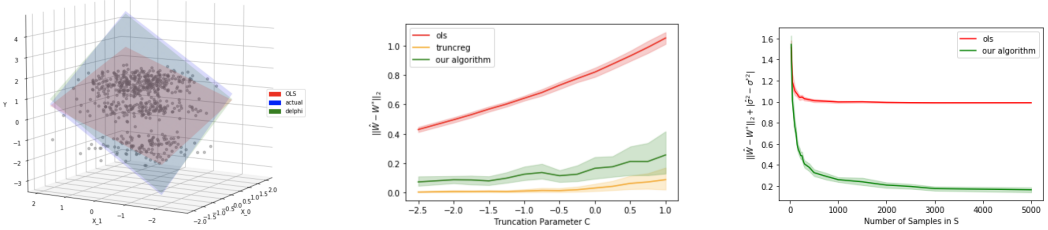
Figure 1: Figure (a) shows $\hat{w}$ estimation errors and Figure (b) shows the $\hat{\sigma}^{*2}$ estimation errors for each σ^{*2}.

In Figure 1, we show how our algorithm performs against models with large noise variances. For this experiment, we sample a 2 dimensional ground-truth weight vector $w^* \overset{i.i.d.}{\sim} \mathcal{U}(-1, 1)$ and generate 1000 $(x^{(i)}, y^{(i)})$ pairs according to $x^{(i)} \overset{i.i.d.}{\sim} \mathcal{U}(-5, 5)$ and $y^{(i)} = w^{*T}x^{(i)} + \varepsilon^{(i)}$; where $\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^{*2})$. We fix w^* and $x^{(0)}, ..., x^{(n)}$ across all trials, but vary σ^{*2}. After adding Gaussian noise to our ground-truth predictions, we left truncate at zero, removing all of the pairs who's $y^{(i)}$ is negative; retaining approximately 50% of the original samples. We vary σ^{*2} over the interval $[1, 20]$ and evaluate how well our procedure recovers w^* and σ^{*2} in comparison to OLS, Daskalakis et al. (2019), assuming σ_0^2 is σ^{*2}, Daskalakis et al. (2019), given σ^{*2}, and *truncreg*, an R package for truncated regression. We acknowledge that when the noise variance is known, Daskalakis et al. (2019) performs the best, but both our algorithm and *truncreg* perform comparably. We emphasize that running Daskalakis et al. (2019) with an incorrect σ^{*2} produces erroneous results.

In Figure 2 (a), we show that our method removes significant amounts of bias even when given a complex k-interval filtering mechanism. For this experiment, w^* is a two dimensional vector of ones, and 1500 $(x^{(i)}, y^{(i)})$ pairs are sampled according to $x^{(i)} \overset{i.i.d}{\sim} N(0, I/2)$ and $\varepsilon^{(i)} \overset{i.i.d.}{\sim} \mathcal{N}(0, 5)$. We then truncate according to $S = ([-2, -1] \cup [0, .5] \cup [1, 2])$, and retain 500 samples post truncation.

In Figure 2 (b), we show that our procedure also works even when the linear model's noise distribution is fat-tailed. In this experiment, we sample a 5 dimensional $w^* \overset{i.i.d}{\sim} \mathcal{U}(-1, 1)$, and generate our data according to $x^{(i)} \overset{i.i.d.}{\sim} \mathcal{U}(-5, 5)$ and $\varepsilon^{(i)} \overset{i.i.d.}{\sim}$ Laplace$(0, 5)$ $(\sigma^{*2} = 5)$. We then left truncate at varying points over the interval [-2.5, 1.0].

In Figure 2 (c), we show that our theoretical error bounds hold in practice. For this experiment, we set w^* to a ten dimensional weight vector of ones, sample n $(x^{(i)}, y^{(i)})$ pairs, where $x^{(i)} \overset{i.i.d.}{\sim} N(0, I)$ and $\varepsilon^{(i)} \overset{i.i.d}{\sim} N(0, 1)$, and right truncate at zero. We generate our data once. However, for each experiment, we randomly select a k samples that exceed zero, and run our experiment. We vary k over the interval $[10, 5000]$. Our results show that the error bound asymptotically approaches zero at a rate of $O(n^{\frac{1}{4}})$, as we increase the number of surviving samples; reaffirming our theoretical analysis.



(a) (b) (c)

Figure 2: Stress testing our method and others against various truncated regression models.

References

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

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

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

Carbery, A. and Wright, J. Distributional and L^q norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical research letters*, 8(3):233–248, 2001.

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. Efficient statistics, in high dimensions, from truncated samples. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 639–649. IEEE, 2018.

Daskalakis, C., Gouleakis, T., Tzamos, C., and Zampetakis, M. Computationally and statistically efficient truncated regression. In *Conference on Learning Theory (COLT)*, pp. 955–960, 2019.

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

Galton, F. 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.

Hajivassiliou, V. A. and McFadden, D. L. The method of simulated scores for the estimation of ldv models. *Econometrica*, pp. 863–896, 1998.

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

Heckman, J. J. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pp. 153–161, 1979.

Kontonis, V., Tzamos, C., and Zampetakis, M. Efficient truncated statistics with unknown truncation. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 1578–1595. IEEE, 2019.

Leluc, R. and Portier, F. Towards asymptotic optimality with conditioned stochastic gradient descent. *arXiv preprint arXiv:2006.02745*, 2020.

Mohan, K., Pearl, J., and Tian, J. Graphical models for inference with missing data. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2013.

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

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

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*, 2011.

Shamir, O. and Zhang, T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pp. 71–79. PMLR, 2013.

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

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

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We have made several assumptions to the dataset
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Yes, and the assumptions are listed in section 1 and section 2
   (b) Did you include complete proofs of all theoretical results? [Yes] Some missing proofs not presented in the paper and proof without details are fully presented in the supplementary materials
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the main paper, we describe the experiments that we conducted for this project. In the supplementary materials, we also inlclude a URL to anonymous repository with the code and more specific details to reproduce the same results seen in our work.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes, we report all of the hyperparameters used for our experiments in the supplementary materials
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes, we report error bars for our experimental results. check section 6 to confirm.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Yes, we describe the hardware used for our experiments in the supplementary materials.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] In our work, we build off of previous work instantiated in the literature. We cite the author's accordingly in our paper.
   (b) Did you mention the license of the assets? [Yes] The only third party asset that we used in our project was Scikit-learn, which a free software package under a BSD license. We cite it accordingly in the supplementary materials.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We will submit additional experiments in the supplementary materials. Nonetheless, all of the data and software that we use for our experiments is free and publicly available. There should be no issues with the additional assets that we choose.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]