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

Editors: Alina Beygelzimer and Daniel Hsu

[1]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 average covariate covariance matrix 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 $\tilde{O}(\sqrt{k/n})$, almost 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 set 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

[1] Extended abstract. Full version is available on arXiv with the same title.

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 $(\boldsymbol{x}^{(i)}, y^{(i)})_{i=1,\ldots,n}$ 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 assumptions about the covariates of the observable samples:

Assumption I: 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 $\varepsilon^{(i)} \sim \mathcal{N}(0,1)$; we can relax this assumption slightly, but an assumption of this type is necessary was shown in Daskalakis et al. (2018) for the special case of our problem, pertaining to truncated Gaussian estimation;

Assumption II: 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; this is the same thickness assumption, that is also commonly made in some standard (untruncated) linear regression settings.

Under Assumptions I and II, 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 $\tilde{O}(\sqrt{\frac{k}{n}})$, almost recovering the error rate of the least-squares estimator in the standard (untruncated) linear regression setting. For a formal statement see the full version of the paper. 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.

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.

1.1. Overview of the Techniques.

We present a high-level overview of our time- and statistically-efficient algorithm for truncated linear regression. Our algorithm 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 $\boldsymbol{w}$ and its $\boldsymbol{x}_i$. 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 $\mathcal{D}_r$ of vectors $(\boldsymbol{w})$ and show that the negative log-likelihood is strongly convex on that set. We also show that this set contains the true coefficient vector, using the matrix Bernstein inequality. Finally, we show that we can efficiently project on this set.

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. 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 our analysis.

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 $\boldsymbol{x}^{(i)}$.

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 Xu et al. (2010); Candès et al. (2011); Diakonikolas et al. (2016); Lai et al. (2016); Diakonikolas et al. (2017, 2018); Bhatia et al. (2015); Diakonikolas et al. (2019). 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.

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.

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.

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.

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.

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

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.

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.

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

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.