A Theoretical and Practical Framework for Regression and Classification from Truncated Samples

Andrew Ilyas
MIT

Manolis Zampetakis
MIT

Constantinos Daskalakis
MIT

Abstract

Machine learning and statistics are invaluable for extracting insights from data. A key assumption of most methods, however, is that they have access to independent samples from the distribution of relevant data. As such, these methods often perform poorly in the face of *biased data* which breaks this assumption. In this work, we consider the classical challenge of bias due to truncation, wherein samples falling outside of an "observation window" cannot be observed [Coh16; Bre96]. We present a general framework for regression and classification from samples that are truncated according to the value of the dependent variable. The framework argues that stochastic gradient descent (SGD) can be efficiently executed on the population log-likelihood of the truncated sample. While our framework is broadly applicable, we also provide end-to-end guarantees for the well-studied problems of truncated logistic and probit regression, where we use it to argue that the true model parameters can be identified computationally and statistically efficiently from truncated data, extending recent work on truncated linear regression [Das+19]. We also provide experiments to illustrate the practicality of our framework on synthetic and real data.

1 Introduction

An emergent threat to the practical use of machine learning is the presence of *bias* in the data used to train ML models. Biased training data can result in models which make incorrect or disproportionately correct decisions, or that reinforce the biases reflected in their training data. For example, some recent works [Bol+16; CBN17] show that semantics derived automatically from language corpora contain human-like biases; others [Kla+12; NG15; BG18] show that the accuracies of face and gender recognition systems are systematically lower for people of color and for women; and, lots of works outline the dangers arising from the use of AI systems trained on biased data ; see e.g. [CP14; GBF16; Bol+16; HPS+16; Kil+17; Kle+17; AMT17; ONe17; BG18; Ren+18] and their references.

While the root causes of AI bias are mutlifaceted, a common source of bias is the violation of the pervasive "i.i.d. assumption"—the assumption that training and test data are independent samples from the same distribution, i.e. that the training examples are a representative sample of the conditions that the trained model will encounter in the future. Of course, this assumption is very commonly violated. Measurement limitations, experimental design, data collection practices, and legal or privacy constraints on data use often make it so that that the training data are a biased sample of the distribution of interest. Unfortunately, training a model naively on a biased sample from a distribution is well-known to potentially lead to very poor performance on an unbiased sample.

In this work, we focus on a specific type of bias that is commonly found in data, in particular bias due to *output truncation*, i.e., when samples are filtered out according to the value of the response variable. Truncation occurs very commonly in practice due to saturation of measurement devices, poor data collection, or legal and privacy constraints preventing the use of some of the data, and has been known to bias statistical inference, since at least the work of Berkson [Ber46] in their eponymous paradox. A classical example of biased inference due to output truncation can be found in a line of work in Econometrics studying the importance of IQ in predicting the earnings of low-skill workers. Studies based on surveys of families whose income was at most 1.5 times the poverty line concluded that there is a negligible contribution of IQ to earnings [Hau72],

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

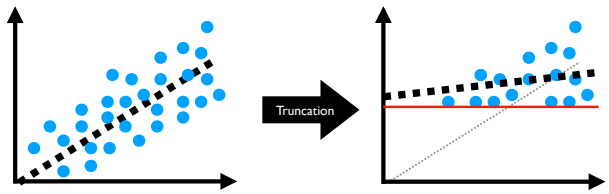


Figure 1: An illustrated example of bias that can be induced by truncation in training inputs.

only to be disproved once the underlying truncation was taken into account [HW77]—the truncation arose from the fact that low-skill workers (individuals earning at most certain amount per hour) who were making incomes larger than 1.5 times the poverty line, were truncated in the surveys.

To illustrate mathematically how output truncation leads to biased estimation, we consider the following simple regression. We have covariates $x_i \sim \mathcal{N}(0,1)$, and for each covariate a response is computed as $y_i = m \cdot x_i + \varepsilon_i$, where $m > 0$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is sampled independently for each observation i. Estimating m given examples $(x_i, y_i)_i$ generated from this model is very well understood, and can be done with ordinary least squares regression. Now, consider the following twist to the plot: suppose that some filtering mechanism filters out observations i whose response variable y_i is negative. Performing ordinary least squares regression on the data surviving truncation will no longer lead to an unbiased estimate of m—OLS will return an estimated coefficient that is smaller than the truth, regardless how many truncated observations we collect! (Figure 1 illustrates this phenomenon.)[1]

How do we recover from this? The topic of statistical estimation from truncated samples is the focus of the field of truncated statistics [Sch86; Coh16; BC14], which finds its roots in Bernoulli's 1766 analysis of smallpox morbidity and mortality to demonstrate the efficacy of vaccination [Ber60], and which was further developed since the beginning of the twentieth century starting with the work of Galton [Gal97], Pearson and Lee [Pea02; PL08], and Fisher [Fis31]. Truncated statistics is widely applicable in Econometrics and many other theoretical and applied fields. Despite intense work, however, there are still numerous outstanding challenges, targeting density estimation, regression and classification tasks. For example, attaining efficient statistical rates for truncated linear regression was only achieved recently [Das+19] despite lots of work in the past several decades [Tob58; Ame73; HW77; Mad86; Kea93; Bre96; HM98], while attaining efficient statistical rates in other basic truncated problems such as truncated logistic and probit regression still remained open prior to our work.

[1]Note that in the truncation setting, the learner is not made aware of dropped samples—this makes it substantially more challenging than the *censored* setting in which the learner is notified when samples are dropped.

Our Contributions. The goal of our paper is to develop a general framework of truncated statistics, pushing both the frontier of what is currently known in theory, and developing generally applicable methods that are practical.

We first propose a general model that encapsulates truncation in a broad range of scenarios (Definition 1). We demonstrate that the model is very general and effectively captures problems spanning truncated classification and regression (Section 2.3).

We next present a general, simple-to-implement SGD-based framework for regression and classification under our presented model. Under two concrete instantiations of our model, namely truncated logistic and truncated probit regression, we provide end-to-end theoretical guarantees, showing that in both cases our framework computationally and statistically efficiently identifies the model parameters (Theorems 2 and 4). We then demonstrate the generality of our framework experimentally; we show that the framework is both easy to instantiate and highly effective even in settings with no theoretical guarantees.

2 A General Framework for Learning from Truncated Data

In this section, we develop the preliminaries and general approach to estimation in the presence of truncated samples. We begin by defining a general form of the truncated learning problem. We show that the framework is sufficiently general, and can be instantiated in the form of logistic, linear, and probit regression. We then outline a general approach to truncated estimation problems based on log-likelihood maximization.

2.1 Preliminaries

We begin with a general definition of the truncated estimation model, which is similar to the classical regression/classification framework.

Definition 1 (Truncated Regression/Classification Model)**.** *The truncated regression/classification model consists of co-variate vectors* $\mathcal{X} = \{x_1, \ldots, x_n | x_i \subseteq \mathbb{R}^d\}$ *and corresponding response vectors* $z_i \in \mathbb{R}^q$*, where the* z_i *are sampled according to*

$$z_i = h_{\theta^\star}(x_i) + \varepsilon_i,$$

where h *is a response function parameterized by an* unknown *parameter* $\theta^\star \in \Theta$*, and* $\varepsilon_i \sim \mathcal{D}_N$ *is mean-zero*

independently sampled noise. The covariate-response pair (x_i, z_i) is silently removed with probability $1-\phi(z_i)$, for a fixed known filtering mechanism $\phi : \mathbb{R}^q \to [0, 1]$. Otherwise (w.p. $\phi(z_i)$), we observe a pair (x_i, y_i), where y_i is determined by a projection $y_i = \pi(z_i)$ and $\pi : \mathbb{R}^q \to \mathbb{R}^q$ is a Markov kernel.

For regression we assume that π is given by a deterministic measurable map and denote by $\pi^{-1}(y)$ the pre-image of y under π. For classification we assume that there is a finite set A such that $\operatorname{supp} \pi(z) \subseteq A$ for all z.

We include π in the model primarily for classification tasks, to allow truncation on the intermediate value z before observing the discrete value y. For regression tasks π will generally be the identity function, although our results for regression easily extend to the case when π is any deterministic function. Using an arbitrary noise distribution $\mathcal{D}_N$ allows us to capture linear and probit regression (which use Gaussian noise), and also logistic regression (which uses logistic noise). For simplicity of exposition, we will consider the case where $\mathcal{D}_N$ has fixed variance—the framework can also be extended to the setting where $\mathcal{D}_N$ has variable/learned variance. The goal of truncated regression is to recover the true parameters θ given a truncated dataset generated according to Definition 1.

Definition 2 (Truncated Regression Problem)**.** *Given a filtering mechanism ϕ, a projection mechanism π, a space of parameters Θ, and a training set $(x_i, y_i)_{i=1}^N$ generated according to the procedure of Definition 1 using some $\theta_* \in \Theta$, our goal is to identify θ_*.*

The truncated regression problem of Definition 2 is quite general and captures many standard problems in the literature. For example:

- **Untruncated linear regression**: To cast the standard (untruncated) linear regression model in the framework of Definition 2, we take the output dimension $q = 1$; the response function parameter space $\Theta = \mathbb{R}^d$ with the class of response functions being $h_\theta(x) = \theta^\top x$ for all x, θ. The filtering mechanism is simply $\phi(z) = 1$ (nothing is truncated) and $\pi(z) = z$ for all z (there is no projection). Lastly, we take $\mathcal{D}_N$ to be Gaussian $\mathcal{N}(0, \boldsymbol{I})$.
- **Untruncated logistic regression**: To cast the standard (untruncated) logistic regression model in the framework of Definition 2, we make use of the latent variable formulation of logistic regression. In particular, we take $q = 1$, $\Theta = \mathbb{R}^d$, and $h_\theta(x) = \theta^\top x$. The added noise $\mathcal{D}_N$ is the Logistic distribution Logistic$(0, 1)$ with scale 1. Lastly, we define $\pi(z_i) = \mathbf{1}_{z_i \geq 0}$. (One can show the equivalence of this latent variable formulation to the standard formulation by noting that the CDF of Logistic$(0, 1)$ is the logistic sigmoid function). Finally, since we are in the untruncated setting, $\phi(z) = 1$ for all z.
- **Untruncated probit regression**: The probit regression model can be captured in an almost identical manner to logistic regression model. The only difference is that instead of having $\mathcal{D}_N =$ Logistic$(0, 1)$, the noise $\mathcal{D}_N = \mathcal{N}(0, 1)$ is Gaussian.
- **Adapting to truncation**: The truncated versions of the afore-described problems arise by taking the filtering mechanism $\phi(\cdot)$ to not be identically 1.

In the next sections, we outline our general framework for solving such truncated estimation problems (Section 2.2), and give concrete instantiations of this framework for linear, probit, and logit regression (Section 2.3). Then in Section 3 we present instances of the framework for which we can give theoretical guarantees.

2.2 Approach

We will approach such truncated regression and classification problems via optimization of the population log-likelihood with respect to the parameters θ:

$$\max_\theta \bar{\ell}(\theta; \theta_*) \qquad \text{where}$$
$$\bar{\ell}(\theta; \theta_*) = \sum_{x_i \in \mathcal{X}} \mathbb{E}_{z_i \overset{\phi}{\sim} h_{\theta_*}(x_i) + \mathcal{D}_N} \left[\ell(\theta; x_i, \mathbf{1}_{z_i \geq 0}, \phi) \right].$$

This log-likelihood for an observation (x, y) is determined by the choice of noise distribution $\mathcal{D}_N$. In a slight abuse of notation, we use $\mathcal{D}_N(\cdot)$ to denote the probability density function of the chosen noise distribution. Recall that the likelihood for an estimation problem of this kind in the non-truncated case for a single sample (x, y) and choice of parameter $\theta \in \Theta$ is given by

$$p(\theta; x, y, \phi) = \int_{z \in \pi^{-1}(y)} \mathcal{D}_N(z - h_\theta(x))\ dz. \tag{1}$$

In the case of truncation, the likelihood becomes:

$$p(\theta; x, y, \phi) = \frac{\int_{z \in \pi^{-1}(y)} \mathcal{D}_N(z - h_\theta(x)) \cdot \phi(z)\ dz}{\int_z \mathcal{D}_N(z - h_\theta(x)) \cdot \phi(z)\ dz}, \tag{2}$$

and so the log-likelihood is

$$\begin{aligned} \ell(\theta; x, y, \phi) &= \log \left(\frac{\int_{z \in \pi^{-1}(y)} \mathcal{D}_N(z - h_\theta(x)) \cdot \phi(z)\ dz}{\int_z \mathcal{D}_N(z - h_\theta(x)) \cdot \phi(z)\ dz} \right) \\ &= \log \left(\int_{z \in \pi^{-1}(y)} \mathcal{D}_N(z - h_\theta(x)) \cdot \phi(z)\ dz \right) \\ &\quad - \log \left(\int_z \mathcal{D}_N(z - h_\theta(x)) \cdot \phi(z)\ dz \right). \end{aligned} \tag{3}$$

The gradient of the first term in (3) w.r.t. θ is

$$\nabla_\theta \log\left(\int_{z\in\pi^{-1}(y)} \mathcal{D}_N(z-h_\theta(x))\cdot\phi(z)\,dz\right)$$
$$= \frac{\int_z \nabla_\theta \mathcal{D}_N(z-h_\theta(x))\cdot\phi(z)\cdot\mathbf{1}_{\pi^{-1}(y)}(z)\,dz}{\int_z \mathcal{D}_N(z-h_\theta(x))\cdot\phi(z)\cdot\mathbf{1}_{\pi^{-1}(y)}(z)\,dz}$$
$$= -\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[\frac{\nabla_\varepsilon \mathcal{D}_N(\varepsilon)^\top \boldsymbol{J}_\theta[h_\theta(x)]}{\mathcal{D}_N(\varepsilon)}\middle|\ h_\theta(x)+\varepsilon\in\pi^{-1}(y)\right] \tag{4}$$

where $\boldsymbol{J}_\theta[h_\theta(x)] \in \mathbb{R}^{q\times d}$ is the Jacobian of h_θ with respect to θ. The gradient of the second term of (3) is nearly identical, except that there is no conditioning on the event $z\in\pi^{-1}(y)$ (since the integral is over the entire domain, rather than just the pre-image of y). Thus, the full gradient of (3) with respect to θ is

$$\nabla_\theta\ \ell(\theta;x,y,\phi) = \left(\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[\frac{\nabla_\varepsilon\mathcal{D}_N(\varepsilon)}{\mathcal{D}_N(\varepsilon)}\right]\right.$$
$$\left.-\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[\frac{\nabla_\varepsilon\mathcal{D}_N(\varepsilon)}{\mathcal{D}_N(\varepsilon)}\middle|\ h_\theta(x)+\varepsilon\in\pi^{-1}(y)\right]\right)^\top \boldsymbol{J}_\theta[h_\theta(x)], \tag{5}$$

where in each expectation ϵ is distributed just as in (4). Now, note that under our setup (c.f. Definition 2), we have sample access to the distribution $\varepsilon\overset{\phi}{\sim}\mathcal{D}_N$, and thus can compute estimates of the gradient above and use first-order methods.

2.3 Instantiations: Linear, Logistic, and Probit Regression

In the last section, we introduced the general SGD framework that we use to approach the truncated learning setting. Here, we consider concrete problems expressible under in this setting (namely linear, logit, and probit regression) and provide their corresponding SGD update steps.

Linear regression. We begin by demonstrating that our truncated estimation framework captures the truncated linear regression setting studied in [Das+19]. Recall from Section 2.1 that linear regression is expressible under the problem setting of Definition 1. In particular, we have $h_\theta(x)=\theta^\top x$ and $\boldsymbol{J}_\theta[h_\theta(x)]=x$, $\pi(y)=y$ and $\mathcal{D}_N=\mathcal{N}(0,1)$, where $\mathcal{N}(x;\mu,\Sigma)$ corresponds to the PDF of the normal distribution with mean μ and covariance Σ, evaluated at x. Recall that $\nabla_x\mathcal{N}(x;0,1)=-x\cdot\mathcal{N}(x;0,1)$. Thus, the gradient for linear regression truncated by $\phi:\mathbb{R}^q\to[0,1]$ can be computed to be:

$$\nabla_\theta\ell(\theta;x,y,\phi) = \quad = \left(y-\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[h_\theta(x)+\varepsilon\right]\right)^\top x,$$

which matches the gradient calculation of Daskalakis et al. [Das+19]. As shown in that work, by computing the Hessian of the truncated likelihood, one can find that under mild conditions, the truncated linear regression problem is in fact strongly convex.

Logistic regression. Next, we demonstrate how to instantiate the truncated learning setup in the setting of logistic regression. As with linear regression, we have $h_\theta(x)=\theta^T x$ and $\boldsymbol{J}_\theta[h_\theta(x)]=x$. We denote $f_\ell(\cdot)$ to be the PDF of the logistic distribution with mean zero and scale one—then, direct calculation shows that $\nabla_x f_\ell(x)=f_\ell(x)\cdot(1-2\cdot\sigma(x))$, where σ is the sigmoid function $\sigma(x)=[1+e^{-x}]^{-1}$. Thus, the gradient of the log-likelihood with respect to θ becomes:

$$\nabla_\theta\ell(\theta;x,y,\phi) = \left(\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[2\cdot\sigma(\varepsilon)\middle|\ h_\theta(x)+\varepsilon\in\pi^{-1}(y)\right]\right.$$
$$\left.-\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[2\cdot\sigma(\varepsilon)\right]\right)^\top x$$
$$= 2\cdot\left(\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[\sigma(\varepsilon)\middle|\ \mathbf{1}_{h_\theta(x)+\varepsilon\geq 0}=y\right]-\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[\sigma(\varepsilon)\right]\right)^\top x \tag{6}$$

Unlike for linear regression where $\pi^{-1}(y)=\{y\}$, here the inverse of π is a non-singleton set, and thus we will need to assume that it has non-zero measure under f_ℓ in order to be able to sample from the first expectation.

Probit regression. All parameters in probit regression are the same as for linear regression, except for the noise distribution $\mathcal{D}_N$, which is now $\mathcal{N}(\cdot;0,1)$ instead of $f_\ell(\cdot)$. Thus, the gradient can be computed as

$$\nabla_\theta\ell(\theta;x,y,\phi) =$$
$$\left(\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[\varepsilon\mid\mathbf{1}_{h_\theta(x)+\varepsilon\geq 0}=y\right]-\mathbb{E}_{\epsilon\overset{\phi}{\sim}\mathcal{D}_N}\left[\varepsilon\right]\right)^\top x. \tag{7}$$

In the next section, we present instances of logistic and probit regression where we can derive theoretical guarantees for the convergence of stochastic first-order methods on the truncated log-likelihood.

3 Theoretical Analysis for Probit and Logistic Regression

In this section we present theoretical results for identifying the parameters of logistic and probit regression, given access to samples from the corresponding models. Results for truncated linear regression have been shown in the recent work of Daskalakis et al. [Das+19].

Our goal in both probit and logistic regression is to maximize the population version of the log-likelihood function. In the population version, where we have infinite many samples, we can compute the expected value of the log-likelihood function for all the possible values of the parameters and we can prove that the optimum of the likelihood is consistent. The next step is

to observe that for both probit and logistic regression, we have enough statistical information available to compute an unbiased estimate of the gradient of the population log-likelihood function. The main technical difficulty is to prove the upper bound on the samples that we need in order for SGD to find an estimate for which the expected gradient of the population log-likelihood is small.

Recall from the definitions of logistic and probit regression in Section 2 that the response variables are one-dimensional, i.e. $q = 1$. In this section, we consider the case where the filtering mechanism is of the form $\phi(x) = \mathbf{1}_{x\in[a,b]}$. For brevity, we write $\ell(\theta; x, y)$ in place of $\ell(\theta; x, y, \phi)$. In both logistic and probit regression, the projection is $\pi(x) = \mathbf{1}_{x\geq 0}$. If $a \geq 0$ (or $b \leq 0$) the label will be $y = 1$ (or $y = 0$) regardless of x, and so no inference is possible. Hence, we assume throughout that $a \leq 0$. In fact, we show that the relevant quantity to lower bound is the probability of any particular data point being be classified as positive or negative; that is, no data point should be classified positively/negatively with 100% probability:

$$\alpha^\star = \min_{x\in\mathcal{X}} \min_{y\in\{0,1\}} \int_{z\in\pi^{-1}(y)} \mathcal{D}_N(z - h_{\theta_*}(x)) \cdot \phi(z)\, dz, \tag{8}$$

where we recall from Section 2.1 that $\mathcal{D}_N(z)$ is the distribution of the response variables. In particular, we will use a lower bound α on α_* to construct our projection set as:

$$\mathcal{R}_{\bar\alpha} = \left\{\theta \middle| \min_{x\in\mathcal{X}, y\in\{0,1\}} \int_{z\in\pi^{-1}(y)} \mathcal{D}_N(z; \theta^\top x) \cdot \phi(z)\, dz > \bar\alpha\right\}, \tag{9}$$

which, by definition contains θ_*, and by the continuity of the noise distribution, can be enforced as constraint on $\theta^\top x$ for all x. We discuss how this set can be efficiently projected to in Appendix A. We also require the following (standard) assumptions on the covariates:

Assumption 1. (*Norm bound*) $\max_{x\in\mathcal{X}} \|x\|_2 \leq B$,

Assumption 2. (*Thickness of covariance matrix of covariates*) $\sum_{x_i} x_i x_i^\top \succeq cI$.

3.1 Probit Regression

We begin by showing identifiability of parameters in the case of probit regression. Our approach is to use the following theorem, which describes the performance of (projected) SGD on a strongly convex function:

Theorem 1 (Theorem 14.11 of [SB14].)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, let $v^{(1)}, \ldots, v^{(M)}$ be a sequence of random vectors such that $\mathbb{E}\left[v^{(i)} \mid \theta^{(i-1)}\right] \in \partial f(\theta^{(i-1)})$ and let $\theta_* = \arg\min_{\theta\in\mathcal{D}} f(\theta)$ be a minimizer of f. If we assume the following:*

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

(ii) **strong convexity:** *$f(\theta)$ is λ-strongly convex,*

then $\mathbb{E}\left[f(\bar\theta)\right] - f(\theta_*) \leq \frac{\rho^2}{2\lambda M}(1 + \log(M))$, where $\bar\theta$ is the average of the steps of the projected stochastic gradient descent algorithm with projection set $\mathcal{D}$ and learning rate at step t, $\eta_t = 1/(\lambda \cdot t)$.

To apply Theorem 1, we need to prove that (i) gradient steps have bounded variance and (ii) the function being optimized is strongly convex. We begin with (ii); for a current value of θ and a sample (x, y), the log-likelihood in our setting can be written as:

$$\begin{aligned}\ell(\theta; x, y) = y \cdot \log\left(\int_0^b \mathcal{N}\left(z; \theta^\top x\right) dz\right) \\ +(1-y)\cdot\log\left(\int_a^0 \mathcal{N}\left(z; \theta^\top x\right) dz\right) \\ -\log\left(\int_a^b \mathcal{N}\left(z; \theta^\top x\right) dz\right).\end{aligned} \tag{10}$$

In order to prove strong concavity of the population log-likelihood, we demonstrate that the likelihood for a single sample is strongly concave for both $y = 1$ and $y = 0$. Recall that we found the gradient of the log-likelihood in Section 2.3; through a calculation shown in Appendix B.1, the Hessian is given by:

$$\begin{aligned}\nabla^2_\theta \ell(\theta; x, y) = \Bigg[(1-y)\cdot \operatorname{Var}_{z \overbrace{\sim}^{z\in[a,0]} \mathcal{N}(\cdot;\mu)}[z] \\ + y \cdot \operatorname{Var}_{z \overbrace{\sim}^{z\in[0,b]} \mathcal{N}(\cdot;\mu)}[z] - \operatorname{Var}_{z \overbrace{\sim}^{z\in[a,b]} \mathcal{N}(\cdot;\mu)}[z]\Bigg] \cdot xx^\top.\end{aligned} \tag{11}$$

We proceed by proving that the inner term of (11) is strictly negative, which can then be combined with the thickness assumption to yield strong concavity. To accomplish this, we make use of the following Lemma, which we prove in Appendix B.2:

Lemma 1. *Let θ_0 be the density of the standard normal distribution truncated to a half-open interval (r, ∞). Now, we define the function $V(s)$ to be the variance of the left-truncated distribution when* also *right-truncated at $s > r$, that is, $V(s) = Var_{x\sim\mathcal{N}(0,1)}[x | r \leq x \leq s]$. Then, for any $b_1, b_2, \gamma \in \mathbb{R}$ such that $s > b_2 > b_1 > r$ and $P_{x\sim\mathcal{N}(0,1)}([b_1, b_2]) > \gamma > 0$, we have $V(b_2) - V(b_1) > poly(\gamma, b_2 - b_1)$.*

In order to be able to apply Lemma 1 to our setting, we need to ensure that there exists a lower bound γ such that $\min(P_{z\sim\mathcal{N}(\theta^\top x,1)}([0, b]), P_{z\sim\mathcal{N}(\theta^\top x,1)}([a, 0])) > \gamma$. Note that α as defined in (8) bounds precisely this value for $\theta = \theta_*$. Also, from the analysis of [Das+18] it follows that because $\alpha^* \geq \alpha$ we have $|a|, |b| \geq \text{poly}(\alpha)$.

If we additionally assume that $\mu = \theta^\top x$ and $\theta \in \mathcal{R}_\alpha$ then together with Lemma 1 we get that

$$\mathrm{Var}_{z \overset{z\in[a,b]}{\sim} \mathcal{N}(\cdot;\mu)}[z] - \mathrm{Var}_{z \overset{z\in[0,b]}{\sim} \mathcal{N}(\cdot;\mu)}[z] > \mathrm{poly}(\alpha), \tag{12}$$

$$\mathrm{Var}_{z \overset{z\in[a,b]}{\sim} \mathcal{N}(\cdot;\mu)}[z] - \mathrm{Var}_{z \overset{z\in[a,0]}{\sim} \mathcal{N}(\cdot;\mu)}[z] > \mathrm{poly}(\alpha). \tag{13}$$

Thus, $\ell(\theta^\top x; y)$ is strongly convex in θ within the projection set. In order to apply the theoretical results about the convergence of stochastic gradient descent we also need the following lemma which follows directly from the log-likelihood gradient, the definition of the set $\mathcal{R}_\alpha$, the Assumption 1 and $\alpha^* \geq \alpha$:

Lemma 2. *For any $\theta \in \mathcal{R}_\alpha$ we have that $\mathbb{E}[|\ell(\theta; x, y)|^2] \leq \mathrm{poly}(1/\alpha, B)$.*

Combining Lemmata 1 and 2 with equations (12) and (13) and Theorem 1 yields the following theorem.

Theorem 2 (Main result for probit regression)**.** *Let $\mathcal{X} = \{x_1 \dots x_n\}$ be covariates satisfying Assumptions 1 and 2 and such that $\alpha^* \geq \alpha$. Let $\{y_1, \dots, y_n\}$ be independent samples from the probit regression model with vector of parameters $\theta_* \in \mathbb{R}^d$. If $n \geq$ poly$(1/\varepsilon, 1/\alpha, B, d, c)$ then projected SGD on the log-likelihood with projection set $\mathcal{R}_\alpha$, outputs an estimate θ such that $\|\theta - \theta_*\|_2 \leq \varepsilon$ with probability at least 99%.*

3.2 Logistic Regression

We now move to the setting of logistic regression, for which we briefly derived the gradient in Section 2.3. Just as for probit regression, we take the truncation function ϕ to be a thresholding function to the interval $[a, b]$. We demonstrate that in contrast to the probit model, the log-likelihood for truncated logistic regression can in fact be *non-concave*, but fortunately is everywhere *quasi-concave*. Our main tool in showing convergence will thus be the following result of Hazan, Levy, and Shalev-Shwartz [HLS15], which guarantees the convergence of stochastic normalized gradient descent (SNGD) for sequences of strictly locally quasi-concave (SLQC) functions (the definition of SLQC functions is given in Definition 3 of Appendix C.1):

Theorem 3 (Normalized PGD on locally quasi-concave functions; [HLS15])**.** *Fix δ, ε, G, M, $\kappa > 0$. Suppose we run projected stochastic normalized gradient descent with $T \geq \kappa^2 \|x_1 - x^*\|^2/\varepsilon^2$ iterations and batch size $b \geq \max\left\{\frac{M^2 \log(4T/\delta}{2\varepsilon^2}, b_0(\varepsilon, \delta, T)\right\}$, where $b_0 : \mathbb{R}^3 \to \mathbb{N}$ is a function defining the minimum batch size necessary such that w.p. $1 - \delta$ and for all $t \in [T]$, the minibatch average at time t is M-bounded and (ϵ, κ, x^*)-SLQC in x_t. Then, with probability of at least $1 - 2\delta$, we have that $f(x^T) - f(x^*) \leq 3\varepsilon$.*

To apply Theorem 3, if suffices to show that with high probability, the log-likelihood for a finite batch of samples is strictly locally quasi-concave (SLQC). To do so, we will first prove that the *population* log-likelihood is SLQC, then exploit concentration of measure to show that for large enough batch sizes, replacing expectations with finite averages (i.e. calculating the empirical log-likelihood) does not break the (local) quasi-concavity.

Recall from Section 2.3 the gradient for truncated logistic regression which we can use to write the corresponding population version (for brevity here we omit an all-enclosing $\sum_{x_i \in \mathcal{X}}[\cdot]$):

$$\nabla_\theta \ell(\theta; x, y) = 2 \cdot \left(\mathbb{E}_{\epsilon \overset{\phi}{\sim} \mathcal{D}_N} \left[\sigma(\varepsilon) \middle| \mathbf{1}_{h_\theta(x) + \varepsilon \geq 0} = y \right] - \mathbb{E}_{\epsilon \overset{\phi}{\sim} \mathcal{D}_N} \left[\sigma(\varepsilon) \right] \right) \boldsymbol{x}$$

$$\nabla_\theta \bar{\ell}(\theta; \theta_*) = \left(\left[\sum_{y \in \{0,1\}} p^*_{(x,y)} \cdot \mathbb{E}_{z \overset{\phi}{\sim} f_\ell(\cdot; \theta^\top x)} \left[\sigma(z - \theta^\top x) \middle| \mathbf{1}_{z \geq 0} = y \right] \right] - \mathbb{E}_{z \overset{\phi}{\sim} f_\ell(\cdot; \theta^\top x)} \left[\sigma(z - \theta^\top x) \right] \right) \boldsymbol{x}$$

where $\quad p^*_{(x,y)} = \mathbb{P}_{z \overset{\phi}{\sim} f_\ell(\theta_*^\top x)} (\mathbf{1}_{z \geq 0} = y)$

Through some calculation (c.f. Appendix C.2), we can derive a closed-form for the above gradient as follows:

$$\nabla_\theta \bar{\ell}(\theta; \theta_*) = \left(\sigma\left(a - \theta^\top x\right) - \sigma\left(b - \theta^\top x\right) \right) \cdot \left(\frac{\sigma\left(b - \theta^\top x\right) - \sigma\left(-\theta^\top x\right)}{\sigma\left(b - \theta^\top x\right) - \sigma\left(a - \theta^\top x\right)} \right. \tag{14}$$

$$\left. - \frac{\sigma(b - \theta_*^\top x) - \sigma(-\theta_*^\top x)}{\sigma(b - \theta_*^\top x) - \sigma(a - \theta_*^\top x)} \right) \boldsymbol{x}. \tag{15}$$

We can use (15) to show that the population likelihood $\bar{\ell}(\theta; \theta_*)$ is strictly quasi-convex in θ.

Lemma 3. *The population log-likelihood from logistic regression truncated to the interval $[a, b]$ is strictly quasi-concave—in particular, we have that for any $\theta \in \mathcal{R}_\alpha$:*

$$\langle \bar{\ell}(\theta; \theta_*), \theta - \theta_* \rangle \leq -\frac{\alpha^2 \cdot \varepsilon^2}{4 \cdot B^2} \cdot \frac{(1 - e^a)(e^b - 1)}{e^b - e^a},$$

where B is an upper bound on $\|x\|_2$.

The proof of Lemma 3 can be found in Appendix C. Finally, we can exploit concentration of single-dimensional bounded random variables to show that, for sufficiently large batch sizes, the *minibatch* log-likelihoods are also strictly locally quasi-concave, allowing us to apply Theorem 3:

Lemma 4. *For a minibatch size* $b \geq \bar{\Theta}\left(poly(B, \frac{1}{\alpha}) \cdot \frac{1}{\varepsilon^2}\right)$, *the minibatch log-likelihood,* $\ell_b(\theta; \theta_*) = \sum_{i=1}^{b} \ell(\theta; x_i, y_i)$, *is* $(\varepsilon, \kappa, \theta_*)$*-SLQC where* $\kappa \in \Theta\left(\frac{1}{\epsilon} \cdot poly\left(B, \frac{1}{\alpha}\right)\right)$.

The proof of Lemma 4 can be found in Appendix C. Together, these results suffice to show that normalized gradient descent can recover the true regression parameters θ_* in the presence of truncation.

Theorem 4 (Main result for logistic regression)**.** *Let* $\mathcal{X} = \{x_1 \dots x_n\}$ *be covariates satisfying Assumptions 1 and 2 and such that* $\alpha^* \geq \alpha$. *Let* $\{y_1, \dots, y_n\}$ *be independent samples from the truncated logistic regression model with vector of parameters* $\theta_* \in \mathbb{R}^d$. *If* $n \geq poly(1/\varepsilon, 1/\alpha, B, d, c)$ *then projected normalized SGD with batch size as in Lemma 4 on the log-likelihood with projection set* $\mathcal{R}_\alpha$, *outputs an estimate* θ *such that* $\|\theta - \theta_*\|_2 \leq \varepsilon$ *with probability at least* 99%.

4 Experimental Results

In the previous sections we introduced a general SGD framework for learning from truncated samples. Here, we conduct experiments demonstrating the practicality of the resulting algorithms. We present the experiments and their results here, with precise details and hyperparameters in Appendix D.

Logistic/Probit Regression on Synthetic Data. In Section 3 we provided theoretical guarantees for truncated logistic and probit regression. Here, we test the practicality of these instantiations of the truncated learning framework on synthetic data, generated using the underlying latent variable model of each algorithm. In particular, we generate a dataset $(\boldsymbol{X}, \boldsymbol{y})$ as:

1. We compute $X_0 = \{x_{[1,\dots,n]} : x_i \sim \mathcal{U}([0, 100])^d\}$ is an $n \times d$ random data matrix. A "ground-truth" θ_* is selected at random ($\theta_* \sim \mathcal{U}([-1, 1])$).
2. For each x_i, a latent z_i is sampled as $z_i = \theta_*^\top x_i + \varepsilon$, where ε is a centered Gaussian (Logistic) random variable for probit (logistic) regression. We remove all (x, z) where z is less than some threshold C.
3. We set $\boldsymbol{X}$ to be a matrix comprised of the vectors not removed in the last step. Each x_i is labeled as $y_i = \mathbf{1}_{z_i \geq 0}$, and we set $\boldsymbol{y} = \{y_1, \dots, y_n\}$.

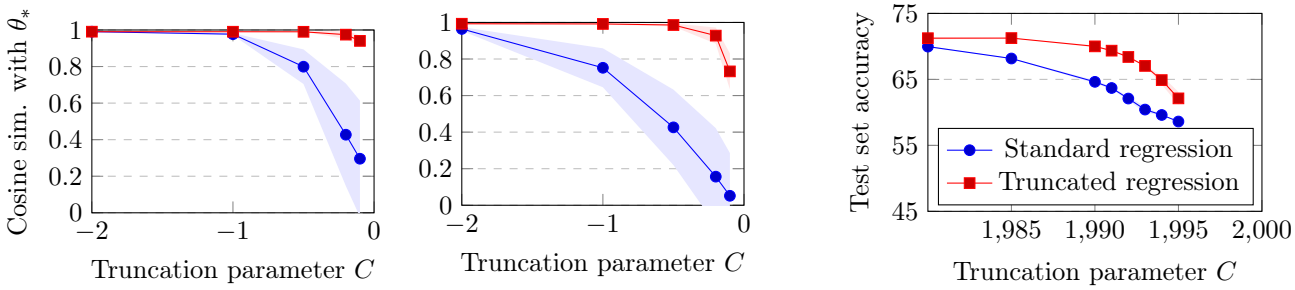
This dataset represents one that has been biased by output truncation, for varying amounts of output truncation C (which corresponds to truncation based on the probability that a point gets classified positively). We compare the performance of our truncation-aware algorithm, to standard logistic/probit regression (which is done obliviously to the bias in the data). We vary C within $[-10, -0.1]$, and study the ability of the two algorithms to recover the parameter θ_*.

In Figures 2a and 2b we evaluate the performance of each algorithm based on how precisely they can recover θ_* (the ground-truth parameter) from truncated data. The results indicate that the truncated learning algorithms presented in Section 2.2 are able to correctly recover the parameters even when the truncation nearly eliminates an entire class of outputs.

Logistic Regression on UCI data. To further demonstrate the practicality of our framework, we test the truncated classification algorithm on data from the UCI machine learning repository [DG17]. Specifically, we use the MillionSongDataset [Ber+11] where the task is to predict, given a set of attributes, whether the song was recorded before or after the millenium. The truncation mechanism is as follows: for varying values of C, we remove all songs from the training set that were recorded on or before the year C. We then train a logistic regression classifier on the resulting training set, comparing standard likelihood minimization to our first-order algorithm in terms of accuracy on the untruncated test set. Figure 2c shows that bias in the training set is much slower to affect test results when the truncated regression algorithm is used.

Neural Networks. Finally, we demonstrate that our framework provides compelling practical performance, even in the absence of theoretical guarantees. Specifically, we study classification and regression settings in which *deep neural networks* are trained on noisy, biased (truncated) data. Our results indicate that the general framework we present Section 2 indeed extends beyond simple settings with provable convergence. To implement our algorithm in this setting, we implement custom versions of the cross-entropy (for classification) and squared-error (for regression) loss functions which return the gradient of the truncated likelihood (c.f. Section 2.3) on the backwards pass We provide the code for this in Appendix E.

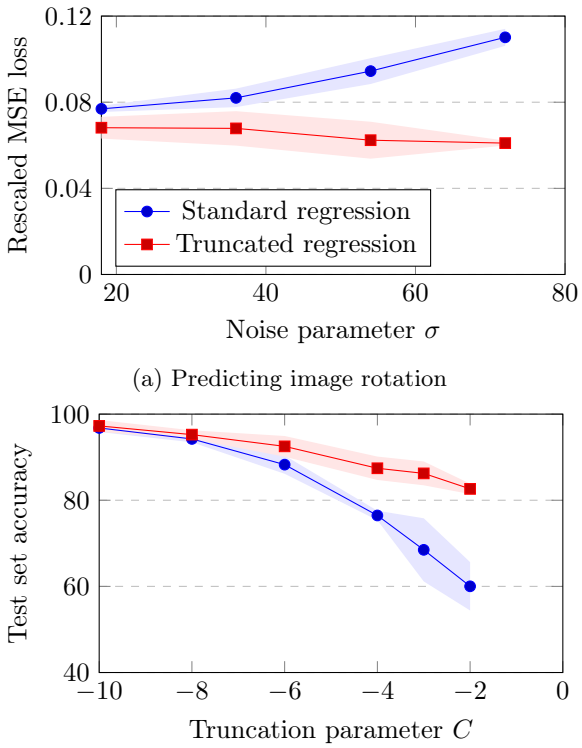
First, we consider a regression task where a CNN (ResNet-50) is tasked with predicting the angle of rotation of randomly rotated images from the CIFAR-10 dataset. A varying amount of label noise σ is added—note that conceptually, as more label noise is added, the bias incurred from truncation tends to increase (c.f. Figure 1). The training data is then truncated based on the labeld angle of rotation, so that only images labeled $0 \leq y \leq 180$ degrees are present in the training set. The results (c.f. Figure 3a) demonstrate that standard log-likelihood maximization can lead to poor performance in the presence of truncation, while the algorithm derived from the truncated likelihood maximization framework maintains good performance.



(a) Probit regression (b) Logistic regression (c) Logistic regression (UCI dataset)

Figure 2: **(a) and (b)**: Comparing estimation methods (standard vs truncation-aware) in terms of their ability to recover the true parameter θ_* on synthetic data generated according to the probit latent variable model and biased via truncation. 95% confidence intervals are shown. **(c)**: comparing standard vs truncation-aware logistic regression on the *YearPredictionMSD* dataset, in terms of test set accuracy.

We then consider a two-class classification task (CIFAR dogs vs. cats) with truncated samples. Concretely, a *base model* is first trained on the untruncated data—samples are removed from the training set if the log-probability assigned to them by the base model is less than or equal to a C (which we vary). We then use the truncated training set to train a new network, and measure its performance on the untruncated test set. The results, presented in Figure 3b, demonstrate that without knowing the relative frequency of each class in the test set, maximizing the truncated log-likelihood still manages to alleviate a significant amount of the bias affecting classifiers obtained via standard training (i.e. minimizing cross-entropy loss).



(a) Predicting image rotation

(b) Classification with truncated data

Figure 3: Comparing neural networks obtained via standard loss minimization vs our framework when trained on truncated data. We train networks to (a) estimate the angle that an image has been rotated; (b) classify between CIFAR dogs and cats. In both cases, the training set undergoes label truncation—networks trained using our framework perform significantly better in the presence of truncation.

5 Conclusion

We introduce a general framework for training machine learning models on *truncated samples*, i.e., on training data that has been biased by omitting samples whose labels fall outside of an observation window.

We instantiate this framework for two specific model classes on which we can derive provable guarantees. We then demonstrate the practicality of our method at learning correctly from biased data in both provable and complex settings. Our findings thus put forth a promising avenue for accounting for known biases in the data generation or collection process.

6 Acknowledgements

We thank Sujit Rao for valuable discussion and input. AI supported by Open Philanthropy Project AI Fellowship; MZ by Google Ph.D. Fellowship, CD by NSF IIS-1741137, CCF-1617730 and CCF-1901292, Simons Investigator Award, DOE PhILMs DE-AC05-76RL01830, DARPA HR00111990021, MIT Frank Quick Fellowship.

References

[Ame73] Takeshi Amemiya. "Regression analysis when the dependent variable is truncated normal". In: *Econometrica: Journal of the Econometric Society* (1973), pp. 997–1016.

[AMT17] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. "Physiognomy's New Clothes". In: *Medium,* `https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a` (2017).

[BC14] N Balakrishnan and Erhard Cramer. *The art of progressive censoring*. Springer, 2014.

[Ber+11] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. "The Million Song Dataset". In: *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*. 2011.

[Ber46] Joseph Berkson. "Limitations of the Application of Fourfold Table Analysis to Hospital Data". In: *Biometrics Bulletin* 2.3 (1946), pp. 47–53. ISSN: 00994987. URL: `http://www.jstor.org/stable/3002000`.

[Ber60] Daniel Bernoulli. "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir". In: *Histoire de l'Acad., Roy. Sci.(Paris) avec Mem* (1760), pp. 1–45.

[BG18] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on Fairness, Accountability and Transparency*. 2018, pp. 77–91.

[Bol+16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4349–4357.

[Bre96] Richard Breen. *Regression models: Censored, sample selected, or truncated data*. Vol. 111. Sage University Paper Series; Quantitative Applications in the Social Sciences, 1996.

[Bur96] Kenneth Burdett. "Truncated means and variances". In: *Economics Letters* 52.3 (Sept. 1996), pp. 263–267.

[CBN17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[Coh16] A Clifford Cohen. *Truncated and censored samples: theory and applications*. CRC press, 2016.

[CP14] Danielle Keats Citron and Frank Pasquale. "The scored society: Due process for automated predictions". In: *Wash. L. Rev.* 89 (2014), p. 1.

[Das+18] Constantinos Daskalakis, Themis Gouleakis, Chistos Tzamos, and Manolis Zampetakis. "Efficient statistics, in high dimensions, from truncated samples". In: *the 59th Annual Symposium on Foundations of Computer Science (FOCS)*. 2018.

[Das+19] Constantinos Daskalakis, Themis Gouleakis, Chistos Tzamos, and Manolis Zampetakis. "Computationally and Statistically Efficient Truncated Regression". In: *the 32nd Annual Conference on Learning Theory (COLT)*. 2019.

[DG17] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[Fis31] RA Fisher. "Properties and applications of Hh functions". In: *Mathematical tables* 1 (1931), pp. 815–852.

[Gal97] Francis Galton. "An examination into the registered speeds of american trotting horses, with remarks on their value as hereditary data". In: *Proceedings of the Royal Society of London* 62.379-387 (1897), pp. 310–315.

[GBF16] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. "The Perpetual Line-Up. Unregulated Police Face Recognition in America". In: *Georgetown Law Center on Privacy & Technology, October 18* (2016).

[GLS80] Martin Grotsche, Laszlo Lovasz, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer, 1980.

[Gup15] Anupam Gupta. 2015. URL: `http://www.cs.cmu.edu/~anupamg/advalgos15/lectures/lecture17.pdf`.

[Hau72] John C Hause. "Earnings profile: Ability and schooling". In: *Journal of Political Economy* 80.3, Part 2 (1972), S108–S138.

[HLS15] Elad Hazan, Kfir Y. Levy, and Shai Shalev-Shwartz. "Beyond Convexity: Stochastic Quasi-Convex Optimization".

In: *Neural Information Processing Systems (NeurIPS)*. 2015.

[HM98] Vassilis A Hajivassiliou and Daniel L McFadden. "The method of simulated scores for the estimation of LDV models". In: *Econometrica* (1998), pp. 863–896.

[HPS+16] Moritz Hardt, Eric Price, Nati Srebro, et al. "Equality of opportunity in supervised learning". In: *Advances in Neural Information Processing Systems*. 2016, pp. 3315–3323.

[HW77] Jerry A Hausman and David A Wise. "Social experimentation, truncated distributions, and efficient estimation". In: *Econometrica: Journal of the Econometric Society* (1977), pp. 919–938.

[Kea93] Michael P Keane. "20 simulation estimation for panel data models with limited dependent variables". In: (1993).

[Kil+17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. "Avoiding discrimination through causal reasoning". In: *Advances in Neural Information Processing Systems*. 2017, pp. 656–666.

[Kla+12] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. "Face recognition performance: Role of demographic information". In: *IEEE Transactions on Information Forensics and Security* 7.6 (2012), pp. 1789–1801.

[Kle+17] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. "Human decisions and machine predictions". In: *The quarterly journal of economics* 133.1 (2017), pp. 237–293.

[Mad86] Gangadharrao S Maddala. *Limited-dependent and qualitative variables in econometrics*. 3. Cambridge university press, 1986.

[NG15] Mei Ngan and Patrick Grother. *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015.

[ONe17] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

[Pea02] Karl Pearson. "On the systematic fitting of frequency curves". In: *Biometrika* 2 (1902), pp. 2–7.

[PL08] Karl Pearson and Alice Lee. "On the generalised probable error in multiple normal correlation". In: *Biometrika* 6.1 (1908), pp. 59–68.

[Pré73] András Prékopa. "On logarithmic concave measures and functions". In: *Acta Scientiarum Mathematicarum* 34 (1973), pp. 335–343.

[Ren+18] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. "Learning to Reweight Examples for Robust Deep Learning". In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. 2018.

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[Sch86] Helmut Schneider. *Truncated and censored samples from normal populations*. Marcel Dekker, Inc., 1986.

[Tob58] James Tobin. "Estimation of relationships for limited dependent variables". In: *Econometrica: journal of the Econometric Society* (1958), pp. 24–36.