ABSTRACT

In the classical setting of self-selection, the goal is to learn $k$ models simultaneously, from observations $(x^{(i)}, y^{(i)})$ where $y^{(i)}$ is the output of one of $k$ underlying models on input $x^{(i)}$. In contrast to mixture models, where we observe the output of a randomly selected model (and therefore the selection of which model is observed is exogenous), in self-selection models the observed model depends on the realized outputs of the underlying models themselves, as determined by some known selection criterion (e.g., we might observe the highest output, the smallest output, or the median output of the $k$ models), and is thus endogenous. In known-index self-selection, the identity of the observed model output is observable; in unknown-index self-selection, it is not. Self-selection has a long history in Econometrics (going back to the works of Roy [54], Gronau [22], Lewis [40], Heckman [28] and others) and many applications in various theoretical and applied fields, including treatment effect estimation, imitation learning, learning from strategically reported data, and learning from markets at disequilibrium.

In this work, we present the first computationally and statistically efficient estimation algorithms for the most standard setting of this problem where the models are linear. In the known-index case, we require poly$(1/\epsilon, k, d)$ sample and time complexity to estimate all model parameters to accuracy $\epsilon$ in $d$ dimensions, and can accommodate quite general selection criteria. In the more challenging unknown-index case, even the identifiability of the linear models (from infinitely many samples) was not known. We show three results in this case for the commonly studied max self-selection criterion: (1) we show that the linear models are indeed identifiable, (2) for general $k$ we provide an algorithm with poly$(d) \cdot \exp(poly(k))$ sample and time complexity to estimate the regression parameters up to error $1/poly(k)$, and (3) for $k = 2$ we provide an algorithm for any error $\epsilon$ and poly$(d, 1/\epsilon)$ sample and time complexity.

CCS CONCEPTS

• Mathematics of computing → Multivariate statistics; Maximum likelihood estimation; • Computing methodologies → Machine learning.

KEYWORDS

selection bias, linear regression, endogeneity, self-selection

1 INTRODUCTION

To introduce our problem we present the following story adapted from the seminal work of Roy [54]. In a small village, two mutually exclusive occupations are available: hunting and fishing. An analyst visits the village with a simple question:

What makes a good fisherman and what makes a good hunter?

More precisely, the analyst wishes to construct a statistical model mapping villagers’ features (e.g., their height and weight) to their proficiency at hunting and fishing (as measured, e.g., by their income). To accomplish this, the analyst might collect a random sample of hunters and fishers from the village, record their relevant features as well as their income, and then use this data to estimate the parameters of two linear models—one for each occupation. For this purpose, it is natural for the analyst to use the OLS estimator on all the hunter data to estimate the hunter model, and the OLS estimator on all the fisher data to estimate the fisher model.

It turns out, however, that even with a perfectly representative sample of villagers, the resulting linear fits will likely be biased. Indeed, if the villagers are rational agents, they will choose their occupations based on which one generates more income for them: those who are better at hunting than fishing (in terms of earnings) will opt to hunt, and vice-versa. As a result, the analyst will never observe, e.g., the hunting earnings of an individual who is better at fishing than hunting, since that individual will choose to fish. This

Footnote 1: For the purposes of this discussion, we assume an abundance of game and fish, and that everything is exported at fixed prices so that the income from each occupation is not affected by how many villagers exercise each occupation.
induces bias in the observed hunting and fishing datasets which makes the outputs of naive estimators on these datasets biased as well: Figure 1 illustrates this effect in one dimension. In fact, this bias arises even in the simpler case where earnings from both occupations are normally distributed and fully independent of an individual’s features and of one another, as discussed in Roy [54].

Outcome self-selection. The above example (due to Roy [54]) is just one illustration of bias due to self-selection, wherein the outcome variable that we observe is selected, often due to strategic considerations from a set of potential outcomes. In this setting, we observe $n$ feature vectors $\{x^{(i)}\}_{i=1}^n$, each accompanied by a label $y^{(i)}$ that is the output of one out of $k$ underlying models:

$$y^{(i)} \in \left\{ f_{w_1}(x^{(i)}, \epsilon_1^{(i)}), \ldots, f_{w_k}(x^{(i)}, \epsilon_k^{(i)}) \right\},$$

where $\epsilon^{(i)} = (\epsilon_1^{(i)}, \ldots, \epsilon_k^{(i)})$ is a noise vector sampled independently across different observations, and $f_{w_1}, \ldots, f_{w_k}$ are $k$ unknown models from some class. The model whose output $y^{(i)}$ we observe is determined by some known function $S : \mathbb{R}^k \rightarrow \{1, \ldots, k\}$ called the self-selection criterion—in the village example, this criterion was the maximum, i.e., $y^{(i)} = \max_{j \in [k]} \{ f_{w_j}(x^{(i)}, \epsilon_j^{(i)}) \}$.

We will consider two instantiations of the self-selection problem that differ in the amount of information available to the statistician. In the easier version of the problem—the "known-index self-selection model"—the statistician observes the identity $j^{(i)}$ of the model that produced each output $y^{(i)}$, in addition to observing $y^{(i)}$ itself. This setting captures the hunting/fishing example of Roy [54], where we observe both the earnings and the occupation of each individual. In the harder version of the problem—the "unknown-index self-selection model"—the statistician does not observe the identity of the model that produced each output $y^{(i)}$.

Applications and prior work. It turns out that the above formulation captures a wide variety of settings wherein observed data is the output of a strategic or systematic selection process operating on some underlying data, the entirety of which is never observed. We discuss some examples:

1. **Imitation learning**: Consider the problem of learning an optimal policy in some contextual bandit setting wherein we observe the arms (e.g. treatments) pulled by an expert (e.g. doctor) in different contexts (e.g. patients). Modeling the reward (e.g. efficacy) from each arm $j$ as an unknown function $f_{w_j}(x, \epsilon_j)$ of the context $x$ and additional randomness $\epsilon_j$ that the expert might observe but we do not, we assume that the expert selects the arm $j$ with the highest reward $\max_j \{ f_{w_j}(x, \epsilon_j) \}$. Our goal is to learn the underlying models $w_1, \ldots, w_k$ of the arms by observing the expert make decisions in different contexts. This scenario is an instantiation of the known-index self-selection model with the maximum selection criterion.

2. Learning from strategically reported data: A widely studied setting featuring self-selected data is one wherein agents are incentivized to strategically choose which data to report. This is a standard challenge in Econometrics, which has recently received increased attention in machine learning literature due to the impact of learning-mediated decisions in various contexts; see e.g. [25, 34, 42] and their references. A common example is the reporting of standardized test scores in college admissions, where applicants have a variety of standardized tests available to them, and are only required to report a chosen subset of them. In a concrete setting of two tests, $A$ and $B$, let $S_A(x, \epsilon_A)$ and $S_B(x, \epsilon_B)$ denote the score of an applicant with features $x$ on each test. Upon completing both tests and receiving their scores $s_A$ and $s_B$, however, an applicant can compute their conditional probabilities of getting accepted based on each score, $P(\text{accept} | s_A)$ and $P(\text{accept} | s_B)$, and only report the score that is more likely to result in acceptance. Estimating parameters of the models determining the scores as a function of a student’s features now corresponds to a known-index self-selection problem with a complicated selection rule (albeit one that is handled by our results in Section 3).

3. Learning from market data: Markets are often in disequilibrium. Following Fair and Jaffee [16], consider a model of the housing market, wherein there is a supply $S(x, \epsilon_S) = w_S^* x + \epsilon_S$ and a demand $D(x, \epsilon_D) = w_D^* x + \epsilon_D$ of houses with features $x$, but the market is in disequilibrium and supply does not equal demand. So the quantity transacted is $Q(x) = \min \{ S(x, \epsilon_S), D(x, \epsilon_D) \}$, where $(\epsilon_S, \epsilon_D)$ are random shocks. This example can be captured by the general model discussed above by setting $k = 2$, considering linear models, and taking the selection criterion to be the minimum selection criterion. Moreover, this is an instance of the unknown-index model as we do not observe whether the disequilibrium is caused by lack of supply or lack of demand.

4. Learning from auction data: Athey and Haile [3] and a large body of literature in Econometrics consider the problem of learning bid (and valuation) distributions from auction data with partial observability, wherein only the winner of each auction and the price they paid are observed. Consider such observations in repeated first-price auctions. We can cast this problem as an instance of known-index self-selection where $k$ is the number of bidders, the models are parametric/non-parametric bid distributions mapping randomness $\epsilon_j$ to a bid, and the selection rule is the maximum function. A body of work in the literature has provided estimation and identification results in this setting [4], including recent work of Cherapanamjeri et al. [11] which demonstrates polynomial-time algorithms for estimating the bid distributions non-parametrically to within Kolmogorov distance $\varepsilon$.

As suggested by the diversity of examples above, models with self-selection bias have received extensive study due to their numerous applications. These include studies of participation in the labor force [13, 23, 24, 28, 29, 47], retirement decisions [20], returns to education [21, 32, 59], effects of unions on wages [1, 36], migration and income [6, 46], physician and lawyer behavior [49, 58], tenure choice and the demand for housing [33, 38, 51], identification of auction models under partial observability [3, 4], and more;
and the convex-inducing self-selection
where the missingness is not at random; see e.g. [19], but computationally and statistically efficient algorithms are lacking, even in very simple cases of the problem such as the housing market disequilibrium model described above, even in the known-index setting. More broadly, self-selection models fall under the literature of regression with missingness in the outcomes, where the missingness is not at random; see e.g. [52, 53, 55] and their references. However, the identification structures considered in this literature do not apply to our setting and/or the results are asymptotic. We discuss existing approaches for self-selection and related models more extensively in Section 1.3.

1.1 Our Results
In this work, we focus on the simple yet prevalent case where the potential outcomes are linear in the collected features (as in the housing market disequilibrium example), with a residual (random) error term that is uncorrelated across potential outcomes. More precisely, we consider a setting involving $n$ individuals (observations) and $k$ potential outcomes (models). Each individual $i \in [n]$ has a feature vector $x^{(i)} \in \mathbb{R}^d$ and each model $j \in [k]$ has a vector of regression parameters $w_{j}^* \in \mathbb{R}^d$. Then, each individual with feature vector $x$ receives a label for each outcome $j$ equaling $y_j = w_j^* x + \epsilon_j$, where $\epsilon$ is assumed to be an standard multivariate Gaussian random variable. Our setting will involve self-selection bias arising from the fact that every individual will only choose to reveal one of their labels, $j_* \in \{1, \ldots, k\}$, as determined by some function of that individual’s full set of labels $y_1, \ldots, y_k$. Our goal in this paper is to estimate the $k$ parameter vectors $w_{1}^*, \ldots, w_{k}^*$ under both the known-index and unknown-index observational models.

We begin with our results in the known-index setting, which we formally define in Definition 2:

**Known-index Setting:** We observe $n$ samples $(x^{(i)}, y^{(i)}_j, j^{(i)}_*)$, where $j^{(i)}_* = S(y^{(i)}_1, \ldots, y^{(i)}_k)$, and $y^{(i)}_j = y^{(i)}_{j^{(i)}_*}$ for some known self-selection rule $S : \mathbb{R}^k \rightarrow \{1, \ldots, k\}$.

Here, we can estimate the unknown parameter vectors $w_{1}^*, \ldots, w_{k}^*$ to arbitrary accuracy and we allow for quite general self-selection rules. In particular, we allow the self-selection rule $S$ to be any convex-inducing rule (Definition 3): letting $y \in \mathbb{R}^k$ be the vector of potential outcomes, if we fix the $j$-th coordinate $y_j$ for any $j \in [k]$, then deciding whether $j$ will be the winner (i.e., whether $j_* = j$) is the same as deciding whether $y_{-j}$ belongs to some convex set (this convex set can depend on $y_j$). We formally define the self-selection rule in Definition 1 and the convex-inducing self-selection rule in Definition 3. Below we present an informal version of our estimation theorem for the known-index case. The corresponding formal version of the theorem can be found in Theorem 1.

**Informal Theorem 1** (Known-Index Estimation – Theorem 1). Let $(x^{(i)}, y^{(i)}_j, j^{(i)}_*)_{i=1}^n$ be $n$ observations from the known-index self-selection model with $k$ linear models $w_{1}^*, \ldots, w_{k}^*$ as described in Section 1.1. If the self-selection rule is convex-inducing and that the probability of observing each model is lower bounded by $(a/k)$ for some $a > 0$, then there is an estimation algorithm that outputs $\hat{w}_{1}, \ldots, \hat{w}_{k}$.
with \( \| \hat{w}_i - w_i^* \| \leq \epsilon \), when \( n \geq \text{poly}(d, k, 1/\alpha, 1/\epsilon) \). Furthermore, the running time of the algorithm is \( \text{poly}(d, k, 1/\alpha, 1/\epsilon) \).

The case of unknown-index is significantly more challenging and as we already mentioned even the with infinite number of samples it is unclear if we have enough information to estimate \( w_j \)’s. We define the setting formally in Definition 4 and informally below:

**Unknown-index Setting:** We observe \( n \) samples of the form \((x^{(i)}, y^{(i)}) \), where \( x^{(i)} \sim N(0, L_i) \), and \( y^{(i)} = \max_{j \in [k]} y^{(i)}_j \).

Observe that in the unknown-index setting problem we assume a Gaussian prior distribution for the covariates \( x^{(i)} \). This is a classical assumption in other linear regression settings, e.g., mixtures of linear regressions, where even the identifiability of the parameters is unclear without prior distribution assumption on \( x^{(i)} \). We face a similar situation here, even our identifiability result for the unknown-index setting rely on the prior distribution of \( x^{(i)} \). This is not the case for the known-index setting where we can assume that \( x_i \) can be picked arbitrarily.

Our first result below shows this identifiability and its formal version is Theorem 2.

**Informal Theorem 2** *(Unknown-index Identifiability – Theorem 2)*. *If we have infinitely many samples from the unknown-index setting, then we can identify all \( k \) linear models \( w_1^*, \ldots, w_k^* \).*

Next we continue with finite-sample and finite-time algorithms for the unknown-index case. To achieve this problem we need some separability assumption between the \( w_j^* \).

**Informal Assumption 1** *(Separability Assumption – See Assumption 4)*. *The projection of any vector \( w_j^* \) to the direction of any other vector \( w_j^* \) cannot be larger that the norm of \( w_j^* \) (and in fact, must be at least \( \Delta \) smaller). Also, each \( w_j^* \) is bounded in norm.*

Our first result for arbitrary \( k \) guarantees the estimation of the parameter vectors \( w_j^* \) within accuracy \( 1/\text{poly}(k) \). For a formal statement of the theorem below we refer to Theorem 3.

**Informal Theorem 3** *(Unknown-index Estimation for General \( k \) – Theorem 3)*. *Let \((x^{(i)}, y^{(i)})_{i=1}^n \) be \( n \) observations from a self-selection setting with \( k \) linear models \( w_1^*, \ldots, w_k^* \) as described in the unknown-index setting of Section 1.1. If we also assume Informal Assumption 1, then there exists an estimation algorithm that outputs \( \hat{w}_1, \ldots, \hat{w}_k \) with \( \| \hat{w}_i - w_i^* \| \leq 1/\text{poly}(k) \), assuming that \( n \geq \exp(\text{poly}(k)) \cdot \text{poly}(d) \). Also, the running time of our algorithm is \( \exp(\text{poly}(k)) \cdot \text{poly}(d) \).*

We pause briefly to make a quick remark on the results of Theorem 3. Note that Theorem 3 (unlike Theorem 1) does not require a lower bound on the observation probability \( a/k \). This is due to the fact that Informal Assumption 1 implies a weaker exponential lower bound on \( a \) which in addition, does not hold uniformly over the regressors \( x \). This actually suffices for our unknown-index algorithm, but incurs a worse sample complexity than the one from Theorem 1. Finally, we study the case where \( k = 2 \) and show how to estimate the parameters to arbitrary accuracy:

**Informal Theorem 4** *(Unknown-index Estimation for \( k = 2 \) – Theorem 4)*. *Consider any \( \epsilon > 0 \) and let \((x^{(i)}, y^{(i)})_{i=1}^n \) be \( n \) observations from a self-selection setting with 2 linear models \( w_1^*, w_2^* \) as described in the unknown-index setting of Section 1.1. Under Informal Assumption 1, there exists an estimation algorithm that outputs \( \hat{w}_1, \hat{w}_2 \) with \( \| \hat{w}_i - w_i^* \| \leq \epsilon \), assuming that \( n \geq \text{poly}(d, 1/\epsilon) \). Furthermore, the running time of the algorithm is also \( \text{poly}(d, 1/\epsilon) \).*

### 1.2 Our Techniques

We initiate a line of work on attaining statistical and computational efficiency guarantees in the face of structured self-selection bias. Below we briefly explain the main ideas behind our main proofs.

**Known-index case.** In the known-index case, our algorithm requires \( \text{poly}(1/\epsilon, k, d) \) sample and time complexity to estimate all \( k \) model parameters to accuracy \( \epsilon \) in \( d \) dimensions, and can accommodate quite general selection criteria. To prove this known-index result, we construct a log-likelihood-inspired objective function that has the true set of parameters as optimum. The key difficulty associated with this formulation is that unlike “nice” settings (for example, the data generating model belongs to an exponential family), the log-likelihood involves an integration over possible outputs of the unobserved models. This scenario is reminiscent of latent-variable models where strong structural properties on the objective functions are uncommon. Nevertheless, we show that this objective function is strongly convex where we crucially rely on the variance reduction properties of log-concave densities conditioned on convex sets. Our next goal is to run projected stochastic gradient descent (PSGD) on this objective function. Unfortunately, in contrast to standard settings in stochastic optimization, we do not have simple access to unbiased stochastic estimates of the gradient due to the integrating out of the unobserved models in the objective function. Consequently, the gradient in this case involves sampling from the conditional distribution over outputs from the unobserved models given the observed sample at the candidate parameter set currently being considered. To sample from this conditional distribution, we show that a projected version of the Langevin Monte Carlo sampling algorithm due to Bubeck, Eldan and Lehec [8] mixes fast and produces an approximate stochastic gradient. Finally, we show that this approximate stochastic gradient suffices for the PSGD algorithm to converge. We provide the details of this algorithm and its analysis in Section 3.

**Identification with unknown index.** In the more challenging unknown index case, it is not even clear whether the parameters \( w_j \) are identifiable from the sample that we have. First, we show that this is in fact the case. Our proof uses an novel identification argument which we believe can be applied to other self-selection settings beyond the Max-selection criterion considered in this work. Formally speaking, we would like to exhibit the existence of a mapping \( f \) from \( \Phi \) to the set of parameters given access to the distribution function, \( \Phi \), of the pairs \((x, y)\) generated according to the self-selection model with unknown indices. Our construction of \( f \) is based on a conditional moment calculation where we analyze the moments of \( y \) conditioned on \( x \) lying in various one-dimensional subspaces. The main observation is that while closed form solutions are not known for the conditional moments, the higher order moments of \( y \) still determine the length of the projection of the parameter vector with the largest projection along \( x \). Concretely, we show that the higher order moments of \( y \) conditioned on \( x \) being parallel to a unit vector
are upper and lower bounded (up to constants) by the moments of normal distribution with 0 and variance $\max((a^T w_i)^2 + 1$. While a single direction does not uniquely determine any of the underlying parameter vectors, the direction maximizing this quantity over all one dimensional subspaces corresponds to the unit vector along the longest parameter vectors allowing recovery of one of $k$ vectors. In the next step, we show that we may effectively “peel off” the single identified model from the distribution function, $\Phi$, reducing the problem of recovering the remaining parameter vectors to a self-selection problem with $k-1$ parameter vectors. A recursive application of this argument allows identification of the remaining parameter vectors one by one.

**Estimation with unknown index.** We then move our attention to estimation with finite time and samples. We target the common max-selection criterion and, under some separability assumption among the $w_j$’s, we provide an algorithm with poly($d$) $\exp$(poly($k$)) sample and time complexity to estimate the regression parameters up to error $1/poly(k)$. Our technique to prove these finite-time and finite-sample results is to try to develop a finite accuracy version of our identifiability argument. Unfortunately, this requires an exponential (in the ambient dimension $d$) sized grid search over the unit sphere. The traditional approach to address such difficulties is to restrict our search to a suitably chosen $k$ dimensional subspace, $U$, identifiable from the data and crucially contains the parameter vectors, $w_j$. We identify this subspace through the spectrum of a suitably chosen matrix but we encounter an additional key difficulty, compared to the previous applications of this method. While our choice of matrix, $M \triangleq \mathbb{E}[y^T xx^T]$, is natural, $M$ does not decompose into a independently weighted sum of matrices each corresponding to a single parameter vector, in stark contrast to the scenario encountered in simpler problems such as mixtures of linear regressions. Hence, showing that the top singular subspace of $M$ contains the $w_j$ is significantly more involved and requires novel approximation ideas. Despite these difficulties, we derive a closed form lower bound for $M$ as a (positively) weighted sum of the identity matrix, $I$ and outer products $w_i w_i^T$ which is tight on the nullspace of the span of $w_i$. Notably, the coefficients of a single parameter vector in the bound depend on the other vectors – a scenario markedly different from other applications of this method such as mixtures of linear regressions. To obtain this closed form lower bound, we replace the max function in the self-selection criterion with a smooth maximum function resulting in a matrix $M'$ approximating $M$ and analyze its spectrum through several careful applications of Stein’s Lemma facilitated by the differentiability of the smooth maximization function. Applying a limiting argument to $M'$, we obtain our lower bound and consequently, show that the span of the parameter vectors $w_j$ is contained in the top-$k$ singular subspace of $M$ and the singular values associated with these directions are bounded away from those for the orthogonal complement establishing a strict spectral gap. Having identified the low-dimensional subspace containing the $w_j$, we may now restrict our search to this subspace. We conclude our estimation argument with a careful finite-sample adaptation of our identifiability argument highlighted above.

**Efficient unknown-index estimation for $k = 2$.** For the specific yet well-studied case of $k = 2$, e.g. [16], we develop a estimation algorithm based on the method of moments that achieves estimation error $\varepsilon$ with poly($d, 1/\varepsilon$) time and sample complexity. It is an interesting open problem whether a similar procedure may be derived for the general case.

### 1.3 Related Work

As previously discussed, bias due to outcome self-selection is a well-documented phenomenon across statistics, econometrics, and the social sciences (see Section 1 for a list of references). In this section, we discuss existing approaches to solve such problems as well as a few similar problems to the statistical and computational ones addressed in this work.

**Classical approaches to self-selection.** There are many parametric and semi-parametric methods from Econometrics for parameter estimation in the presence of self-selection (see [37] for an overview). To describe results in the literature let us consider a generalization of the problem we consider here, where the potential outcomes $y_i$ are generated as

$$y_i = x^T w_j + u_i \quad \text{and} \quad i^* = \max \{ (x^T w_j + \varepsilon_i : i \in [k]) \},$$

where the $\varepsilon_i$ and $u_i$ are all jointly normally distributed. Analytical algorithms for this setting (i.e., where the noise that determines selection is non-identical to the noise in the observed outcome [27, 39, 43, 45]) typically focus on the case when $k = 2$ for tractability reasons. The prevailing algorithm here is the two-stage estimator [29, 38], where one first estimates the model $x \rightarrow i^*$ (whose true parameter is $w_1 - w_2$) then uses a Heckman correction to estimate the outcome parameters $w_1$ and $w_2$. Alternatively, one can use likelihood-based approaches [21, 28, 47], which—as shown by Olsen [48]—are identified for a known correlation $\rho$ between $\varepsilon_i$ and $u_i$ above. For $k > 2$, the likelihood function involves several integrals, and so the prevailing approach is to use Markov-Chain Monte Carlo (MCMC) and simulation-based approaches, for which there are no established convergence rates and the underlying algorithms are not efficient (see [17] for an overview).

Thus, although the problems are similarly motivated, our approach to the known-index case differs from standard Econometrics approaches in a few ways. First, in our model the selection noise and observation noise are identical, which introduces bias in typical two-stage estimators. Second, our algorithm for the known-index case applies directly to $k > 2$ and has a (known) convergence rate that is nearly linear in $n$ and polynomial in $k$. Finally, we are unaware of any algorithms in Econometrics (whether analytical or simulation-based) that tackle the unknown-index case, where even identifiability (Section 4.1) is non-trivial.

**Truncated linear regression.** Both the known-index and unknown-index cases addressed in this work bear some similarity to the truncated and censored linear regression problem, wherein there is a single outcome that is only seen if it falls within a fixed observation window. Truncated regression problems date back to at least the works of Tobin [56], Amemiya [2], and Hausman and Wise [26], who all note the effect of omitting outcomes that fall below or above a certain threshold on resulting regression models. Recently, Daskalakis et al. [14] propose a gradient-based algorithm (and derive statistical and computational efficiency guarantees) for estimating parameters of truncated linear regression; Ilyas et al. [30] extend these results to truncated probit and logistic regression.
What makes self-selection bias more challenging than these settings is the inherent endogeneity in the process of selecting samples to be observed. That is, in the case of truncated regression, the learner is always aware of which samples would be truncated had they been generated by a given set of parameter estimates. In contrast, under the self-selection model, an incorrect set of parameter estimates can lead to an incorrect estimate of the truncation mechanism (since one depends on the other).

Mixtures of linear regressions. The problem that we consider—and specifically the unknown-index case—bears some similarity to the well-studied problem of learning (noisy) mixtures of linear regressions. In mixtures-of-regressions, the observations are also of the form \((x^{(i)}, y^{(i)})\) where \(y^{(i)}\) is the output of one of \(k\) linear models: unlike our setting, however, in mixtures of linear regressions the model from which \(y^{(i)}\) is observed is selected at random, and crucially, independently of the model outputs themselves.

To illustrate the significance of this distinction, observe that for mixtures of linear regressions, the corresponding “known-index” case (i.e., where we observe the regression from which each datapoint was generated) is trivial. In particular, it corresponds to estimating \(k\) independent ordinary least squares models. In contrast, the known-index case here still requires a more elaborate algorithmic approach and analysis, as the datasets remain correlated with one another even conditioned on the selection indices.

In terms of finite-sample estimation algorithms, for the mixtures of linear regressions problem both noiseless [41] and noisy cases [10, 15, 35] have been considered in the literature. The guarantee that we get for our unknown-index case is similar to the guarantee of [10] for mixtures of linear regression, i.e., to achieve error of order \(1/\text{poly}(k)\) when the magnitude of the noise is also \(1/\text{poly}(k)\), we require time and sample complexity \(\text{poly}(d) \cdot \exp(\text{poly}(k))\). For the mixtures of linear regression case this \(1/\text{poly}(k)\) accuracy is enough to get arbitrary small accuracy \(\varepsilon\) by applying the EM algorithm that admits local convergence in this case as shown by [35]. This local convergence result is missing in our unknown-index setting and it is a very interesting open problem that will lead to estimation with arbitrary small accuracy when combined with the results that we show in this paper.

Max-affine regression. Another related line of literature to the unknown-index case is the problem of max-affine regression [18]. The main difference between the settings is that in max-affine regression, the noise is added to the model after the maximum operator. Hence, the self-selection bias due to the noise does not appear in this setting. Although this seems like a small change it makes the problems completely different both technically and conceptually:

- Conceptually, the main application of max-affine regression is in settings where we want to regress with respect to the set of convex functions. The set of convex functions is non-parametric and in high-dimensions this will result in an exponential sample and computational complexity. On the other hand if we parametrize the convex function as the maximum of a set of affine functions then we get a finite number of parameters and we can achieve much better efficiency. The resulting statistical problem is max-affine regression.

But max-affine regression cannot capture the biases that we mention above due to strategic agents of imitation learning.
- Technically, the max-affine regression could be solved via analogs of OLS. The main problem there is that the optimization landscape is non necessarily convex but from a statistical point of view minimizing the average square loss is a meaningful thing to do. In our case, the main problem is that the solutions of naive optimization problem are biased and hence we need to find the correct optimization problem that effectively debiases the data. In the unknown-index setting this is particularly challenging and in fact even arguing about the identifiability is non-trivial.

Our unknown-index setting seems more difficult from this point of view, but the fact that there is noise added after the maximum makes the two problems not reducible from one to another.

In fact, one can view max-affine regression as a case of self-selection where the error terms \(\epsilon_j\) are perfectly correlated across potential outcomes \(j \in [k]\), as opposed to perfectly independent (as in our setting). This raises a question of whether we can design algorithms to handle more complex correlations between the model-specific error terms \(\epsilon_j\).

2 MODEL AND MAIN RESULTS

Notation. We use \(N(\mu, \Sigma)\) to denote the normal distribution with mean \(\mu\) and covariance matrix \(\Sigma\). For any measurable set \(\mathcal{K} \subseteq \mathbb{R}^d\) we denote with \(N(\mu, \Sigma; \mathcal{K})\) the normal distribution \(N(\mu, \Sigma)\) conditioned on that the output belongs to \(\mathcal{K}\). We will use \(f_{\mathcal{K}}\) and \(F_{\mathcal{K}}\) to denote the PDF and CDF, respectively, of the single-dimensional normal distribution \(N(0, \sigma^2)\). When \(\sigma\) is clear from the context we will just use \(f\) and \(F\). Let \(A\) be an \(n \times m\) matrix, we define \(A^\flat\) to be a vector in \(\mathbb{R}^n\) that is the flattening of \(A\), where for the flattening we use the lexicographic order of the coordinates of \(A\). Let \(\mathcal{K} \subseteq \mathbb{R}^d\) be a convex set let \(x \in \mathbb{R}^d\), we define \(\Pi_{\mathcal{K}}(x)\) to be the projection of \(x\) to \(\mathcal{K}\). We use \(B(x, r)\) to denote the Euclidean ball with center \(x\) and radius \(r\) and when \(x = 0\) we also use simply \(B(r)\). For a subspace \(V\), we use \(P_V\) and \(P_{\perp V}\) to denote the projection operators onto \(V\) and the orthogonal complement of \(V\) respectively. For a vector \(v\) and matrix \(V\), \(P_v\) and \(P_V\) denote the projection operators onto the one-dimensional subspace along \(v\) and the column space of \(V\) respectively and analogously for \(P_{v,\perp}\) and \(P_{V,\perp}\).

In this section we define the two models that we are solving: the known-index setting and the unknown-index setting, we describe the assumptions that we use for each of the settings and we formally state our main results.

2.1 Known-Index Setting

We start with the definition of a self-selection rule that is fundamental in the modeling of the linear regression problem with self-selection bias in the known-index setting.

Definition 1 (Self-Selection Rule). A self-selection rule is a function \(S : \mathbb{R}^k \rightarrow [k]\). We assume throughout this work that we have query access to \(S\), i.e. for every \(y \in \mathbb{R}^k\) there is an oracle that outputs \(S(y)\). We also define a slice \(C_j(a) \subseteq [k]\) of \(S\) as follows:

\[
C_j : \mathbb{R} \Rightarrow \mathbb{R}^{k-1} \quad C_j(a) = \{ y_{-j} \in \mathbb{R}^k \mid S(y) = j, y_j = a \}
\]

where \(\Rightarrow\) refers to a point-to-set map.
The first setting that we consider is the known-index setting, where the observed data for each covariate include the response variable, as well as the index of the corresponding regressor.

**Definition 2 (Self-Selection with Observed Index).** Self-selection with observed index is parameterized by an unknown set of weight vectors \( w^*_1, \ldots, w^*_k \in \mathbb{R}^d \), a known variance \( \sigma > 0 \), and a self-selection rule \( S : \mathbb{R}^k \rightarrow [k] \). For \( i \in [n] \), an observation \((x^{(i)}, y^{(i)}, j^{(i)})\) in this model is a triplet, comprising a feature vector \( x^{(i)} \), and a pair \((y^{(i)}, j^{(i)})\) sampled as follows conditioning on \( x^{(i)} \):

1. Sample latent variables \( y^{(i)} \sim \mathcal{N}(w^{*_i} x^{(i)}, \sigma^2) \) for each \( j \in [k] \).
2. **Reveal the observation index** \( j^{(i)} = S(y^{(i)}), \ldots, y^{(i)} \) and the response variable \( y^{(i)} = y^{(i)}_{j^{(i)}} \).

For a fixed \( x \) and \( W = (w^*_1, \ldots, w^*_k) \), we use \( \mathcal{D}(x; W^*) \) to denote the probability distribution of the pair \((i, y)\) sampled according to Steps (1) and (2) above. For example, if \( S(y^{(i)}), \ldots, y^{(i)} \) are \( \arg\max_{j \in [k]} y^{(i)}_j \), then under this model we observe only the largest \( y^{(i)} \) and its index. Our goal is to obtain an accurate estimate \( \hat{w}_j \) for each \( w^*_j \) given only samples from the above model.

In most of this paper we are mainly concerned with estimating the weights \( w^*_1, \ldots, w^*_k \) from observation of the covariates and the maximal response variable

\[
y = \max_{j \in [k]} \{ w^{*_j} x + \epsilon_j \}, \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2 : I_k).
\]

In the observed-index setting, however, it turns out that our efficient estimation can be applied to a much larger set of selection functions \( S(y) \), which we call convex-inducing.

**Definition 3 (Convex-Inducing Self-Selection Function).** We call a self-selection function \( S : \mathbb{R}^k \rightarrow [k] \) convex-inducing if, for each \( j \in [k] \) and \( a \in \mathbb{R} \), slice \( C_j(a) \) (Definition 1) is a convex set.

Notably, setting \( S(y) = \arg\max_{j \in [k]} y_j \) recovers the maximum-response observation model, and this choice of \( S(\cdot) \) also satisfies Definition 3 with \( C_j(a) = (-\infty, a]^{k-1} \). The definition also allows us to capture cases beyond the maximum; for example, convex-inducing functions also include \( S(y) = \arg\max_{j \in [k]} f_j(y) \) for any set of monotonic functions \( f_j \).

Our estimation procedure relies on the following assumptions on the feature and parameter vectors, which are classical and present even in many standard linear regression instances.

**Assumption 1 (Feature and Parameter Vectors).** For the set of feature vectors \( \{x^{(i)}\}_{i=1}^n \) that we have observed assume that:

\[
\left\| x^{(i)} \right\|_2 \leq C \quad \text{for all } i \in [n] \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n x^{(i)} x^{(i)\top} \succeq I. \quad (1)
\]

For the true parameter vectors \( w^*_1, \ldots, w^*_k \) we assume that \( \left\| w^*_j \right\|_2 \leq B. \)

Although our theorem allows for a wide range of self-selection rules, we need some additional assumptions that allow the recovery of every parameter vector \( w_j \). To see why this is needed imagine the setting where the self-selection rule \( S \) is the arg max and for some \( j \) the coordinates of \( w_j \) are extremely small. In this case it is impossible to hope to estimate \( w_j \) since there is a huge probability that we do not even observe one sample of the form \( y = x^\top w_j + \epsilon_j \).

For this reason, we need the following assumption.

**Assumption 2 (Survival Probability).** There exists a constant \( \alpha > 0 \) such that, for every sample \((x, y, j)\) that we have observed, the following properties hold:

1. For every \( j \in [k] \), the probability that \( j = j \) is at least \( \alpha k \).
2. The mass of the set \( C_j(\epsilon) \) with respect to \( \mathcal{N}((W^* x)_j^\top, \sigma^2 : I_{k-1}) \) is at least \( \alpha. \)

Finally, we assume oracle access to a self-selection rule \( S(\cdot) \) that is convex-inducing, in the same sense introduced in Definition 3.

**Assumption 3 (Self-Selection Rule).** We assume the self-selection rule \( S : \mathbb{R}^k \rightarrow [k] \) is convex-inducing, and that for every \( j \in [k] \) and \( y \in \mathbb{R} \) we have access to the following:

1. A membership oracle for \( S(y) \), i.e., for every \( y \in \mathbb{R}^k \) and \( j \in [k] \), we can know whether \( S(y) = j \).
2. A projection oracle to the (convex) slice \( C_j(a) \) of the self-selection rule, i.e., for every \( y \in \mathbb{R}^k \) and every \( a \in \mathbb{R} \), we can efficiently compute \( y = \arg\min_{y \in C_j(a)} \|y - s\| \).

Having established the above definitions and assumptions, we are now ready to state our main theorem for the known-index setting.

**Theorem 1 (Known-Index Estimation).** Let \( (x^{(i)}, y^{(i)})_{i=1}^n \) be \( n \) observations from a self-selection setting with \( k \) linear models \( w^*_1, \ldots, w^*_k \), as in Definition 2. Under Assumptions 1, 2(i), and 3(i), there exists an algorithm that outputs \( \hat{w}_1, \ldots, \hat{w}_k \) such that with probability at least 0.99, for every \( j \in [k] \)

\[
\left\| \hat{w}_j - w^*_j \right\|_2 \leq \text{poly}(\sigma, k, 1/\alpha, B, \epsilon) \cdot \frac{\log(n)}{n}.
\]

If we additionally assume the Assumptions 2(ii), and 3(ii), then the running time of the algorithm is \( \text{poly}(n, d, k, 1/\alpha, B, C, \sigma, 1/\epsilon) \).

In Section 3, we explain our algorithm for proving Theorem 1 and we describe the main ideas and techniques for the proof. Specifically, to construct an efficient algorithm we use an interesting combination of projected gradient descent and the Langevin algorithm, after we develop an appropriate objective function.

### 2.2 Unknown-Index Setting

We next consider the more challenging unknown-index setting, where we have sample access to the response variable, but do not observe the index of the corresponding weight vector. In this more challenging setting, we make a few additional assumptions on the structure of the problem, namely that the covariates \( x^{(i)} \) are drawn from an mean-zero identity-covariance Gaussian distribution (rather than being arbitrary); that the noise terms \( \epsilon^{(i)} \) are also identity-covariance (rather than \( \sigma \cdot I_k \) for \( \sigma > 0 \)); and that the self-selection rule is the maximum response rule \( S(y) = \arg\max_{j \in [k]} y_j \) (rather than an arbitrary convex-inducing rule).

**Definition 4 (Self-Selection with Unknown Index).** Just as for Definition 2, we have a set of weight vectors \( w^*_1, \ldots, w^*_k \in \mathbb{R}^d \). For each sample covariate \( x^{(i)} \), we:
allows us to design an algorithm that yields 
holds, then there exists an
data and 
exists an estimation algorithm that outputs a set of weights 
values
In order to transform the above information-theoretic result to a
model with unknown indices as described in Definition
probability of observing a response from a given weight vector
argument that might be of independent interest—we believe it may
one that shows (infinite-sample) identifiability for unknown-index
samples from the above model.

Our goal is to obtain an accurate estimate \( \hat{w}_j \) for each \( w_j \) given only
samples from the above model.

This model resembles mixtures of linear regressions, with the
key difference being that in the latter, the index of the weight
This model resembles mixtures of linear regressions, with the
parameters

Assumption 4

For some known real
values \( B, \Delta \) it holds that:

(A)

Assumption 4 allows us to design an algorithm that yields \( \epsilon \)-accurate estimates of the \( w_j \)’s.

Theorem 3. Let \((x^{(i)}, y^{(i)})_{i=1}^n\) be \( n \) observations from a self-selection
setting with \( k \) linear models \( w_1^*, \ldots, w_k^* \) as described in the unknown-index
setting in Definition 4. If we assume Assumption 4, then there
exists an estimation algorithm that outputs a set of weights \( \{w_j^*\}_{j=1}^k \)
and an ordering of these weights \( \hat{w}_1, \ldots, \hat{w}_k \) such that with probability
at least 0.99, for every \( j \in [k] \)

\[ \| \hat{w}_j - w_j^* \|_2 \leq \epsilon, \]

as long as \( n \geq \text{poly}(d, 1/\epsilon, B, 1/\Delta) \) and \( \epsilon \leq \Delta/4 \). Furthermore, the
running time of the algorithm is at most \( n \cdot \text{poly}(d, 1/\epsilon, B, 1/\Delta) \).

In Section 4 we describe the algorithms and proofs for Theorem 3 and Theorem 4.

Remark. (High-Probability Results) All of the above results are expressed in term of constant probability of error. We can boost
this probability to \( \delta \) by paying an additional \( \log(1/\delta) \) factor in the sample and time complexities. This boosting can be done because we are solving a parametric problem and it is a folklore idea that any probability of error less than \( 1/2 \) can be boosted to \( \delta \). Roughly, this boosting works by running the algorithm independently \( \log(1/\delta) \) times, and from the \( \log(1/\delta) \) different estimates keeping one that contains at least half of all the others within a ball of radius \( 2\epsilon \).

3 PARAMETER ESTIMATION FOR THE
KNOWN-INDEX SETTING

In this section we present and analyze our algorithm for estimating the parameters \( \{w_j\} \) from samples distributed according to Definition 2. At a high level, our approach is to run projected stochastic gradient descent (SGD) on the appropriate objective function whose optima coincide with the true set of parameters. We start with the definition of this function, then we show the design of the appropriate projection set and then we proceed with proving its main properties. We conclude with the proof of our estimation theorem.

3.1 Objective Function for Linear Regression

with Self-Selection Bias

The objective function that we use is inspired by the log-likelihood function. We show that our objective function is convex (even though linear regression with self-selection bias does not belong to any exponential family). Suppose we have a given parameter estimate for the \( \{w_j^*\}_{j=1}^k \) given by \( W = \{w_j\}_{j=1}^k \) then we define its objective value \( \tilde{f}(W) \) as follows.

\[ \tilde{f}(W) \leq \frac{1}{n} \sum_{i=1}^n E(y_{i,j}) \cdot (x^{(i)}, w^*) \left[ f(W; x^{(i)}, y_{i,j}) \right] \]

\[ \leq \frac{1}{n} \sum_{i=1}^n E(y_{i,j}) \left[ \log \left( \int_{C_{y}(y)} \prod_{j \in I_y} f_y(z_j - w_j^* x^{(i)}) \ dz_j \right) \right], \quad (2) \]
where we recall that $f_{\theta}$ is the density function of the standard normal distribution. The above expression is based on the population likelihood under the current estimate $W$ of the pair $(y, j_s)$ conditioned on the value of $x$: see Appendix A of the full version [12] for the exact derivation. The gradient of $\overline{\ell}$ can then be expressed in the following form:

$$
\nabla_{W} \overline{\ell}(W) = \frac{1}{n_{\text{adj}}} \sum_{i=1}^{n} \mathbb{E}_{(y_{j_s})} \left[ 1_{j = j_s} \cdot y \right] + \mathbb{E}_{z_{j \neq j_s} \sim N((W^T x^{(i)}), \sigma^2 I_{K-1})} \left[ z_j \cdot x_{j \neq j_s} \in C_j(y) \right] - \mathbf{w}^T x^{(i)} x^{(i)}. 
$$

(3)

The first thing to verify is that the set of true parameters $W^*$ is a stationary point of the objective function that we proposed above.

**Lemma 1.** It holds that $\nabla \overline{\ell}(W^*) = 0$, where $W^* = \{w^*_j\}_{j=1}^K$ is the set of true parameters of the known-index self-selection model described in Definition 2.

**Proof.** See Appendix B.1 in the full version [12].

Our goal is to apply projected stochastic gradient descent (PSGD) on $\overline{\ell}$. To this end, we need to prove that our objective function is actually strongly concave and hence the optimum of $\overline{\ell}$ is unique and equal to $W^*$. We show this strong convexity in Section 3.2. Next, we need to show that we actually apply PSGD and hence need to find a procedure to sample unbiased estimates of the gradient of $\overline{\ell}$. Unfortunately the form of the objective function does not allow us to find such an efficient procedure. For this reason we relax our requirement to finding approximately unbiased estimates of $\nabla \overline{\ell}$. To achieve this we use a projected version of Langevin dynamics as we show in Section 3.3. Additionally, we need to show that the second moment of our gradient estimates cannot be very large which we also show in Section 3.3. Finally, we need to adapt the proof of convergence of PSGD to show that the small bias that Langevin dynamics introduces can be controlled in a way that does not severely affect the quality of the output estimation which we show in Section 3.4. In Section 3.5 we combine everything together to prove our estimation result.

### 3.2 Strong Concavity

The Hessian of $\overline{\ell}$ is difficult to analyze directly. We thus start with the Hessian of the log-likelihood for a single sample $(x^{(i)}, y^{(i)}, j^{(i)})$. In particular, in Appendix A of the full version [12] we derive the Hessian of this function $\ell(W; x^{(i)}, y^{(i)}, j^{(i)})$, which comprises blocks $H_{W}$ such that

$$
(H_{W})_{ab} = \frac{\partial^2}{\partial w^a \partial w^b} \ell(W; x, y, j_s).
$$

It follows that for a single sample $x, y, j_s$, the matrix block $H_{j_s, j_s} = 0$ for all $j \neq j_s$. Thus, it remains to consider only the blocks $H_{j_s, j}$ and $H_{j, j}$ for $j, l \neq j_s$. Again, in Appendix A of the full version [12] we show that

$$
H_{j_s, j} = -\frac{1}{\sigma^2} xx^T
$$

and that

$$
H_{j, j} = \frac{1}{\sigma^2} \mathbf{w}^T x x^T \left[ \int_{\mathcal{D}} f_{\mathcal{X}}(x) dx \right] \mathbf{w}^T x x^T,
$$

where $\otimes$ represents the Kronecker product. Now, the key property of convex-inducing selection functions in our proof is that, for Gaussian random variables over $\mathbb{R}^{K-1}$, the variance is non-increasing when the variable is restricted to a convex set.

**Lemma 2** (Corollary 2.1 of [31]). Let $X \in \mathbb{R}^n$ be a random vector with Gaussian density $f_{\mathcal{X}}$. For a convex set $A \subseteq \mathbb{R}^n$ with positive mass under the distribution of $X$, define $X_A = X$ to be $X$ restricted to $A$, i.e., a random variable with density $f_{\mathcal{X}}(x) = f_{\mathcal{X}}(x) \cdot (\int_A f_{\mathcal{X}}(z) dz)^{-1}$. Then, for all $\mathbf{v} \in \mathbb{R}^n$,

$$
\text{Var}[\mathbf{v}^T X_A] \leq \text{Var}[\mathbf{v}^T X].
$$

In particular, together with our thickness assumption and properties of the Kronecker product, this implies that $H_{j_s, j_s} \preceq 0$. Thus, the complete Hessian of the function $\ell$ can be expressed as a block matrix of the form:

$$
H = \begin{bmatrix}
-\frac{1}{\sigma^2} xx^T & 0 \\
0 & \mathbf{H}_{W_{j_s, j_s}}
\end{bmatrix}
$$

We are now ready to upper bound the Hessian $H_{\text{pop}}$ of our objective function $\ell$. In particular, at this point we can use our minimum-probability assumption (Assumption 1) and our thickness of covariates assumption (Assumption 2) from which we get that for the Hessian $H_{\text{pop}}$ it holds that

$$
H_{\text{pop}} \preceq -\frac{\alpha}{\sigma^2} 1_{K} I
$$

From the above we conclude that the following lemma

**Lemma 3.** The objective function $\ell$ is $-\frac{\alpha}{\sigma^2}$-strongly-concave.

### 3.3 Approximate Stochastic Gradient Estimation

In this section we describe an algorithm for sampling approximate stochastic estimates of our objective function $\ell$. Our algorithm is based on projected Langevin dynamics. We start with the expression of the gradient of $\ell$ based on (3), where recall that $E_{(y, j_s)}$ denotes the expectation of the pair $(y^{(i)}, j_s^{(i)})$ conditioned on $x^{(i)}$ and $W^*$. To obtain stochastic gradient estimates, we will replace these expectations with their corresponding observed values as we will see below. The more difficult step of the gradient estimation process is sampling the last term of (3), for which it suffices to be able to sample the truncated normal distribution

$$
N((W^T x^{(i)})_{j_s}, \sigma^2 I_{K-1}; C_j(y))
$$

given some set of parameter $W$, a vector of covariates $x^{(i)}$ and a pair $(y, j_s)$ drawn from $D(x^{(i)}; W^*)$. The simplest way to get a sample from (4) is to first sample from $N((W^T x^{(i)})_{j_s}, \sigma^2 I_{K-1})$, and then apply rejection sampling until we get a sample inside $C_j(y)$. This is feasible information-theoretically but it might require a lot of computational steps if the survival probability of $C_j(y)$ is small. In particular, the rejection sampling might require time that is exponential in the norm of the $W_j$s. For this reason, if we require statistical efficiency we need to apply a more elaborate
technique. In particular, we use projected Langevin dynamics. Let $K = C_{ij}(y) \cap \mathcal{B}(R)$ for some sufficiently large constant $R$ and let $\mu_{-j} = (W^T x^{(i)})_{-j}$ for the rest of this section. The iteration of projected Langevin algorithm for sampling is the following [8]:

$$z^{(t+1)} = \Pi_K \left( z^{(t)} - \frac{y}{2} \cdot \sigma^2 (z^{(t)} - \mu_{-j}) + \sqrt{\gamma} \cdot \xi^{(t)} \right)$$

(5)

where $\xi^{(1)}, \xi^{(2)}, \ldots$ are i.i.d. samples from the standard normal distribution in $(k-1)$-dimensions. The next lemma describes the sampling guarantees of the Langevin algorithm (5).

**Lemma 4.** Let $L \subseteq \mathbb{R}^{k-1}$ be a convex set with $\mathbb{P} N(0, \sigma^2 I_{k-1}) \geq \alpha$ for $\alpha > 0$. Then, for any $\mu_{-j} \in \mathbb{R}^k$ and $\in \{0, 1/2\}$, the projected Langevin sampling algorithm (5) with $K = L \cap \mathcal{B}(R)$ for some appropriate value $R$, and initialized with $z^{(0)} = \Pi_K (0_{k-1})$, generates a random variable $\tilde{X} = z^{(m)}$ satisfying

$$TV \left( \tilde{X}, N(\mu_{-j}, \sigma^2 \cdot I_{k-1}, K) \right) \leq \epsilon$$

assuming the number of steps $m \geq \text{poly}(k, \|w\|, 1/\epsilon, 1/\alpha, \sigma^2, 1/\sigma^2)$.

**Proof.** The proof of this lemma can be found in Appendix B.2 of the full version of this work [12].

Now that we can sample from the distribution (4), we can move to approximately estimating a stochastic gradient of $\tilde{f}$. First, we sample uniformly $i \in [n]$ uniformly at random, and we fix the corresponding $x^{(i)}$. Then, we use the $i$-th sample from the true model to substitute in the pair $(y, j)$. Finally, we use the Langevin algorithm that we described above to sample (4). Before moving to bounding the bias of our estimator there is one more thing that we need to take care of, and this is that for every $x^{(i)}$ we only have one sample of the pair $(y, j)$. Hence, we need to make sure that during the execution of the algorithm while we pick the indices $i \in [n]$ uniformly at random we will never pick the same index $i$ twice. To ensure that we are going to require more samples than the ones we need.

Let $n$ be the total number of samples that we have and $T$ be total number of samples that we need for our PSGD algorithm. A straightforward birthday paradox calculation yields that the probability of sampling the same $i$ twice is at most $2T^2/n$. Thus, if we pick $n \geq 2T^2/\xi$, then the collision probability during the execution of the PSGD algorithm is at most $\xi$.

We are now ready to put everything together in algorithm that describes our combined estimation procedure. The following lemma whose proof can be found in Appendix B.3 of the full paper [12] describes the performance guarantees of this estimation algorithm.

**Lemma 5.** Let $g^{(1)}, \ldots, g^{(T)}$ be a sequence of outputs of Algorithm $i$ when used with input $W^{(1)}, \ldots, W^{(T)}$, where $\|W^{(p)}\|_2 \leq k \cdot B$ and $W^{(p)}$ can depend on $W^{(p-1)}, g^{(p-1)}$. If $n \geq 2T^2/\xi$, for the hyper-parameters $\eta, R$, it holds that $R, 1/\eta \leq \text{poly}(k, B, 1/\beta, 1/\alpha, \sigma^2, 1/\sigma^2)$, and $m \geq \text{poly}(k, B, 1/\beta, 1/\alpha, \sigma^2, 1/\sigma^2)$ then with probability at least $1 - \zeta$ it holds that for every $p \in [T]

$$\mathbb{E} \left[ \|g^p \cdot (W^{(p-1)}, g^{(p-1)}) - \nabla \tilde{f}(W^{(p)}) \|_2 \right] \leq \beta.$$

(6)

**Algorithm 1 Approximate Stochastic Gradient Estimation**

1. procedure **ESTIMATEGRADIENT**(W)
2. sample $i$ uniformly from $[n]$
3. $K \leftarrow C_{ij}^{(i)} (y^{(i)}) \cap \mathcal{B}(R)$
4. $\mu \leftarrow (W^T x^{(i)})_{-j}^{(i)}$
5. $z^{(0)} \leftarrow \Pi_K (0)$
6. for $t = 1, \ldots, m$
7. sample $\xi^{(t-1)}$ from $N(0, I)$
8. $z^{(t)} \leftarrow \Pi_K (z^{(t-1)} - \frac{y}{2} \cdot \sigma^2 (z^{(t-1)} - \mu) + \sqrt{\gamma} \cdot \xi^{(t-1)})$
9. end for
10. for $j = 1, \ldots, k$
11. $g_j \leftarrow \frac{1}{\sigma} (z^{(i)}_{j}^{(i)}) y^{(i)} + 1_{j_{i_{x_{j}}}} \cdot z^{(m)} - w_j^{(m)} x^{(i)}$
12. end for
13. return $g = (g_1, \ldots, g_k)$
14. end procedure

**Algorithm 2 Projected Stochastic Gradient Descent**

1. procedure **PSGD**
2. $W^{(0)} \leftarrow 0$
3. for $t = 1, \ldots, T$
4. $\eta_t \leftarrow 1/\lambda \cdot t$
5. $g \leftarrow $ **ESTIMATEGRADIENT**(W)
6. $w_j^{(t)} \leftarrow \Pi_K (w_j^{(t)} - \eta_t \cdot g_j^{(t)})$ for all $j \in [k]$
7. end for
8. return $\hat{W} \leftarrow \frac{1}{T} \sum_{t=1}^T w_j^{(t)}$
9. end procedure

### 3.4 Stochastic Gradient Descent with Biased Gradients

In the previous section we showed that we can compute approximate stochastic gradients of the strongly-concave function $f$. In this section we show that this is enough to approximately optimize $f$ using projected gradient descent. We start with a description of the PSGD algorithm.

**Lemma 6.** Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function, $\mathcal{K} \subset \mathbb{R}^k$ a convex set, and fix an initial estimate $w^{(0)} \in \mathcal{K}$. Now, let $x^{(1)}, \ldots, x^{(T)}$ be the iterates generated by running $T$ steps of projected SGD using gradient estimates $g^{(1)}, \ldots, g^{(T)}$ satisfying $\mathbb{E} [g^{(i)} | x^{(i-1)}] = \nabla f(x^{(i-1)}) + b^{(i)}$. Let $x_* = \text{arg min}_{x \in \mathcal{K}} f(x)$ be a minimizer of $f$. Then, if we assume:

- (i) Bounded step variance: $\mathbb{E} \|g^{(i)}\|_2^2 \leq \rho^2$,
- (ii) Strong convexity: $f$ is $\lambda$-strongly convex, and
- (iii) Bounded gradient bias: $\|b^{(i)}\|_2 \leq \frac{\rho^2}{\sqrt{T} \cdot \text{dim} \mathcal{K}}$,

then the average iterate $\hat{x} = \frac{1}{T} \sum_{t=1}^T x^{(t)}$ satisfies $\mathbb{E} |f(\hat{x}) - f(x_*)| \leq \frac{\rho^2}{\sqrt{T}} (1 + \log(T))$.

**Proof.** See Appendix B.4 of the full paper [12].
3.5 Proof of Theorem 1

We are now ready to combine the results of the previous sections into a recovery guarantee for $W^* = (w_j^*_i)_{j=1}^k$. In particular, we will apply Lemma 6 to show that Algorithm 2 converges to an average iterate $\tilde{W}$ that is close to $W^*$. First, observe that the norm of the gradient estimates outputted by Lemma 4 are bounded in norm by

$$\mathbb{E}\left[\|\tilde{\eta}(W)\|^2\right] \leq \mathbb{E}\left[\sum_{j=1}^k \|\nabla_w \tilde{f}(W; (x, i, y))\|^2\right] + \beta,$$

where $\beta$ is as in Lemma 5. Our bounds on the norm of the weights and covariates directly implies

$$\mathbb{E}\left[\|\tilde{\eta}(W)\|^2\right] \in O(k \cdot \text{poly}(B, C) ) + \beta.$$

Next, Lemma 3 guarantees that $f$ in Lemma 6 is strongly convex with $\lambda = \alpha/(\sigma k)$. Finally, Lemma 4 ensures access to gradients with appropriately bounded bias (i.e., satisfying assumptions (i) and (iii) in Lemma 6) in poly($k, B, C, T, 1/\alpha, \sigma^2, \sigma^2$)-time. We are thus free to apply Lemma 6 to our problem—after averaging $T$ steps of projected stochastic gradient descent, we will find $W = (w_j^*)_{j=1}^k$ such that

$$\mathbb{E}[\tilde{\eta}(W)] - \tilde{\eta}(W^*) \leq \frac{\sigma^2 \cdot k^2 \cdot \text{poly}(B, C)}{2 \cdot \alpha T} (1 + \log(T)).$$

An application of Markov’s inequality shows that, with probability at least $1 - \delta$,

$$\tilde{\eta}(W) - \tilde{\eta}(W^*) \leq \frac{\sigma^2 \cdot k^3 \cdot \text{poly}(B, C, 1/\delta)}{2 \cdot \alpha T} (1 + \log(T)).$$

Thus, we can condition on the event in (7) while only losing a factor of $1 - \delta$ in success probability. Finally, a parameter-space recovery bound follows from another application of convexity:

$$\|W - W^*\|_F \leq \frac{\sigma^4 \cdot k^3 \cdot \text{poly}(B, C, 1/\delta)}{2 \cdot \alpha T} (1 + \log(T)).$$

4 PARAMETER ESTIMATION FOR THE UNKNOWN-INDEX SETTING

In this section, we establish Theorems 3 and 4, our main results for parameter recovery from a self-selection model with unknown indices (Definition 4). Note that in this setting, even information theoretic identifiability of the parameters (i.e. parameter identification given access to infinite samples from the model) is not known. Hence, we start our discussion with a simple identifiability proof in the information theoretic setting in Chapter 3 of the book. In Subsection 4.1.1. We then extend these ideas in Subsections 4.2 and 4.3 to prove Theorem 3. A naive adaptation of our identifiability proof results in a runtime and sample complexity scaling exponentially in the dimension of the input points. Therefore, in Subsection 4.2, we show how one can efficiently identify a $k$-dimensional containing the span of the weight vectors, $(w_j^*)_{j=1}^k$. While similar approaches based on effective subspace identification have also been employed for other statistical learning tasks such as that of learning mixtures of well-separated gaussians [57], our analysis is significantly more intricate as the moments of the distributions under consideration do not have an obvious closed-form expression. Having identified a suitable low-dimensional subspace, we then carry out a finite sample analysis of our information theoretic identifiability proof from Subsection 2 in Subsection 4.3 to prove Theorem 3. Finally, in the special case where $k = 2$, we describe a procedure which enables parameter recovery from sample complexity and runtime scaling as poly(1/ε) improving on the exp(1/ε) sample complexity and runtime from Theorem 3, proving Theorem 4 and concluding the section.

4.1 Identifiability with Unknown Indices

Here, we establish the information theoretic identifiability of the self-selection model with unknown indices. Recall, that we receive samples generated according to $y^{(i)} = \max_{j \in \{k\}} w_j^* x^{(i)} + \eta_j^{(i)}$ where $x^{(i)} \sim N(0, I)$ and $\eta_j^{(i)} \iid N(0, 1)$. We now establish the following theorem:

**Theorem 2.** Let $W^* = (w_j^*)_{j=1}^k \in \mathbb{R}^{d \times k}$ and $\Phi_{W^*}$ be the distribution function of the pairs $(x^{(i)}, y^{(i)})$ associated with the self-selection model with unknown indices as described in Definition 4. Then, there exists a mapping $f$ satisfying:

$$\forall W^* = (w_j^*)_{j=1}^k \in \mathbb{R}^d \text{ it holds that } f(\Phi_{W^*}) = W^*.$$

**Proof.** Our proof will be based on an inductive argument on the number of components, $k$. We will use a peeling argument to reduce the parameter recovery problem with $k$ components to one with $k - 1$ components. The base case when $k = 1$, reduces to standard linear regression where, for example, $\mathbb{E}[x^{(i)} \cdot y^{(i)}] = w_1$. Sufficient for the inductive argument, suppose $k > 1$ and consider the following function:

$$\forall v \in \mathbb{R}^d, \|v\| = 1: F(v) = \lim_{\text{Even } p \rightarrow \infty} \lim_{\text{Odd } p \rightarrow \infty} \frac{\mathbb{E}[y^p \mid \|P_{\perp v} x\| \leq y]}{(p - 1)!!},$$

where $P_{\perp v}$ is the projection matrix orthogonal to the direction of $v$. We will now show that the above function is well defined for all $\|v\| = 1$. The conditional moments may be evaluated with access to the distribution function $\Phi_{W^*}$. Defining $f^* = \arg \max_{j \in \{k\}} |\langle v, w_j^* \rangle|$ and $s_j = |\langle v, w_j^* \rangle|$, we now lower bound the conditional moment:

$$\mathbb{E}[y^p \mid \|P_{\perp v} x\| \leq y] \geq \mathbb{E}[y^p \cdot \mathbb{1}\{w_j^* x + \eta_j^* \geq 0\} \mid \|P_{\perp v} x\| \leq y] \geq \frac{1}{2} \mathbb{E}[y^p \mid \|P_{\perp v} x\| \leq y] \geq \frac{1}{2} \mathbb{E}[y^p \mid \|P_{\perp v} x\| \leq y] \geq \frac{1}{2} \mathbb{E}\left[\sum_{j=0}^{p/2} \binom{p}{2j} (w_j^* P_{\perp v} x + \eta_j^*)^{p - 2j} (w_j^* P_{\perp v} x)^{2j} \mid \|P_{\perp v} x\| \leq y\right] \geq \frac{1}{2} \left(p - 1\right)!! \cdot (\sigma_j^2 + 1)^{p/2}.$$

Through a similar computation, we obtain an upper bound on the conditional moment:

$$\mathbb{E}[y^p \mid \|P_{\perp v} x\| \leq y] \leq \sum_{j=1}^k \mathbb{E}[(w_j^* x + \eta_j)^p \mid \|P_{\perp v} x\| \leq y] = \sum_{j=1}^k \mathbb{E}[(w_j^* P_{\perp v} x + \eta_j + w_j^* P_{\perp v} x)^p \mid \|P_{\perp v} x\| \leq y]$$

1709
Lemma 7 (Weighted covariance). Consider the matrix

\[ M = \mathbb{E} \left[ \max(0, y^2 \cdot \mathbf{x} \mathbf{x}^T) \right], \]

and let \( p_i = \mathbb{P} \left( i = \arg \max_{j \in [k]} \langle w_j, x + \eta_j \rangle \right) \) and \( \{w_j \cdot x + \eta_i > 0\} \). Then, if \( \sigma \) is a unit vector,

\[ \sigma \in \text{span}(w_1, \ldots, w_k) \Rightarrow \sigma^T M \sigma \geq \mathbb{E} \left[ \max(0, y^2) \right] + 2 \sum_{i=1}^k p_i (\sigma^T w_i)^2. \]

Proof. We present the proof of this Lemma in the full version of our work [12]. The main idea is to approximate \( \max(0, y^2) \) as the soft maximum function \( F_\beta(W) := \beta^{-1} \log(\sum_{k=1}^n \exp(\beta \cdot W_j)) \) applied to the \((k+1)\)-dimensional vector \((0, w_j^T x + \eta_1, \ldots)\). The remainder of the proof follows by repeated application of Stein's identity, stated below:

**Lemma 8 (Stein’s Identity).** Let \( W = (W_1, \ldots, W_p) \) be a centered Gaussian random vector in \( \mathbb{R}^p \). Let \( f : \mathbb{R}^p \rightarrow \mathbb{R} \) be a \( C_1 \)-function such that \( \mathbb{E}[|\partial_j f(W)|] < \infty \) for all \( j \in [p] \). Then, for every \( j \in [p] \),

\[ \mathbb{E}(W_j \cdot f(W)) = \sum_{a=1}^p \mathbb{E}[W_a \cdot W_j] \mathbb{E}[\partial_a f(W)]. \]

After applying Stein’s identity twice, we take the limit as \( \beta \rightarrow \infty \) and bound the resulting terms as needed to complete the proof. \( \square \)

We now establish lower bounds on the \( p_i \) from the previous lemma which will aid us in establishing a quantitative spectral gap. The proof of the following Lemma is found in the full paper [12]:

**Lemma 9.** Under Assumption 4, for some absolute constant \( C > 0 \):

\[ \forall i \in [k] : p_i \geq \exp(-Ck \log((2B)/\Lambda)). \]

We will now use Lemma 9 to establish a quantitative spectral gap bound on the principal subspaces approximately containing the \( \mathbf{w}_i \).

**Lemma 10.** Let \( \epsilon \in (0, 1/2) \) and \( \{u_i \}_{i \in [l]} \) be all the singular vectors of \( M \) whose corresponding singular values \( \{\sigma_i \}_{i \in [l]} \) satisfy:

\[ \forall i \in [l] : \sigma_i \geq \mathbb{E} \left[ \max(0, y^2) + \min_{i \in [k]} \frac{p_i e^2}{2} \right]. \]

Then, letting \( U = \text{span}(\{u_i \}_{i \in [l]}) \), we have:

\[ I \leq k, \forall i \in [k] : \frac{\|w_i - P_U w_i\|}{\|w_i\|} \leq \epsilon \text{ and } \|M\| \leq 3kd(8^2 + 1). \]

Proof. The first claim (i.e., that \( \ell \leq k \)) follows from Lemma 7. For the second, consider the contrary and suppose that for some \( i \in [k] \),

\[ \frac{\|w_i - P_U w_i\|}{\|w_i - P_U w_i\|} > \epsilon. \]

Now, define the vector \( \sigma \) as:

\[ \sigma = \frac{w_i - P_U w_i}{\|w_i - P_U w_i\|}. \]
We have $v \perp U$ and furthermore, Lemma 7 yields:
\[ \langle v^T M v, x \rangle \geq \mathbb{E} \left[ \max(0, y)^2 + p_j (v^T w_j)^2 \right] \geq \mathbb{E} \left[ \max(0, y)^2 \right] + p_j \epsilon^2 \]
yielding the contradiction, establishing the second claim. Finally, for the last claim, we have:
\[ \|M\| \leq \mathbb{E} \left[ \max(0, y)^2 \|xx^T\| \right] = \mathbb{E} \left[ \max(0, y)^2 \|x\|^2 \right] \]
\[ \leq \sum_{i=1}^{d} \sum_{j=1}^{d} \mathbb{E} \left[ (w_i^T x + \eta_j)^2 \cdot (x_j)^2 \right] \]
\[ \leq \sum_{i=1}^{d} \sum_{j=1}^{d} \mathbb{E} \left[ (w_i^T x + \eta_j)^4 \cdot (x_j)^4 \right] \leq 3kd(B^2 + 1). \]
\[ \square \]

We now combine the spectral gap shown in the previous Lemma with a matrix concentration argument to argue that k-SVD on $M = \mathbb{E}[y^2] \cdot I$ (where $M$ is as defined in Lemma 7) approximately recovers the span of the $\{w_i\}$. We use as a primitive k-SVD:

**Fact 1 (Anonymity).** Let $M \in \mathbb{R}^{k \times k}$, and let $x_1, x_2, x_3$ denote the non-zero singular values of $M$. For any $j \in [k - 1]$, define spectral gap $\gamma_j = \min(\sigma_j, \sigma_{j+1})$. Furthermore, suppose that we have an oracle that computes $M(x)$ for any $x \in \mathbb{R}^k$ in time $T$. Then, for any $\eta, \delta > 0$, there is an algorithm APPROX-SVD($M, \eta, \delta$) which runs in time $O(j R^{k \min(\log(\eta) - 1)}, \log(k/\sigma_1))$ and with probability at least $1 - \delta$ outputs $U \in \mathbb{R}^{k \times k}$ with orthonormal columns so that $\|U - U_j\|_2 < \eta$, where $U_j$ is the matrix whose columns are the top $j$ right singular vectors of $M$.

The combination of Lemma 7 and Fact 1 imply that it suffices to show concentration of the average of a sequence $M^\epsilon_i y_i$. The following Lemma helps us establish this concentration:

**Lemma 11.** Suppose we generate $n$ samples $(x^{(i)}, y^{(i)}) \sim X, Y$ from the self-selected linear regression model, i.e., $x^{(i)} \sim \mathcal{N}(0, I)$, then $y^{(i)} = w_i^T x^{(i)} + \mathcal{N}(0, 1)$. Define the empirical second-moment matrix
\[ \hat{M} = \frac{1}{n} \sum_{i=1}^{n} \max(0, y^{(i)})^2 \cdot x^{(i)} x^{(i)^T}. \]

Fix any $\delta \in (0, 1)$. Then, if $n \geq \Omega(\max(1/\delta, d))$, with probability at least $1 - \delta$,
\[ \|\hat{M} - \mathbb{E}[M]\|_2 \leq O \left( \frac{\log(kn)}{\sqrt{n}} \max \left\{ \sqrt{\log(2/\delta)}, \sqrt{d} \right\} \right). \]

**Proof.** See Appendix C.1 of the full version [12]. \[ \square \]

We can thus approximately identify the subspace in poly-time.

**4.3 Estimating Parameters using the Low-Dimensional Subspace**

Here we leverage the results of Subsections 4.2 and 4.1 to build an algorithm for estimating the weight vectors $\{w_i\}_{i=1}^k$ under the separability Assumption 4 on the $\{w_i\}$. Subsection 4.2 allows us to effectively reduce the dimensionality of the problem down to $k$ dimensions. We will then adapt the identifiability argument from Subsection 4.1 in the $k$-dimensional subspace to estimate the weight vectors and complete the proof of Theorem 3, restated below:

**Theorem 3.** Let $(x^{(i)}, y^{(i)}) \sim \mathcal{P}_{\mathcal{P}}$ be $n$ observations from a self-selection setting with $k$ linear models $w_1, \ldots, w_k$ as described in the unknown-index setting in Definition 4. If we assume Assumption 4, then there exists an estimation algorithm that outputs a set of weights $\hat{w}_1, \ldots, \hat{w}_k$ and an ordering of these weights $\hat{w}_1, \ldots, \hat{w}_k$ such that with probability at least $0.99$, for every $j \in [k]$,
\[ \|\hat{w}_j - w_j\|_2 \leq \epsilon, \]
as long as $n \geq \max\{\text{poly}(d) \cdot \text{exp} \left\{ \text{poly}(B/\epsilon) \cdot \tilde{O}(k) \right\} \}$ and $\epsilon \leq \Delta/4$. The runtime of the algorithm is at most $n \cdot \text{exp} \left\{ \text{poly}(B/\epsilon) \cdot \tilde{O}(k) \right\}$.

The proof of Theorem 3 (in the full version of our work [12]) is essentially a quantitative version of the identifiability argument from Theorem 2 restricted to the subspace identified in Subsection 4.2.

**4.4 Estimation in the $k = 2$ Case**

We now demonstrate how, when $k = 2$, we can use a moment-based algorithm to estimate $\{w_1, w_2\}$ in poly($1/\epsilon$) time. In particular, we provide the algorithm corresponding to Theorem 4, restated below:

**Theorem 4.** Let $(x^{(i)}, y^{(i)}) \sim \mathcal{P}_{\mathcal{P}}$ be $n$ observations from a self-selection setting with 2 linear models $w_1, w_2$ as described in the unknown-index setting of Section 2. If Assumption 4 holds, then there exists an estimation algorithm that outputs $\hat{w}_1, \hat{w}_2$ such that with probability at least $0.99$, for every $j \in [2]$,
\[ \|\hat{w}_j - w_j\|_2 \leq \epsilon, \]
as long as $n \geq \max\{\text{poly}(d, 1/\epsilon, B, 1/\delta)\}$ and $\epsilon \leq \Delta/4$. Furthermore, the running time of the algorithm is at most $n \cdot \text{poly}(d, 1/\epsilon, B, 1/\Delta)$.

The algorithm will operate as follows:

1. Using the procedure given in Section 4.2, we find an approximation $\hat{U}$ to the linear subspace $U^*$ containing span($w_1, w_2$).
2. Set up an $(\epsilon/6)$-covering over $\hat{U} \cap \mathcal{B}(0)$, where $\mathcal{B}(0)$ is the $\epsilon$ ball with radius $B$. Since $\|w_i\|_2 \leq B$ both vectors are contained in the covering, and the covering is of size $O(B^2/\epsilon^2)$.
3. For each element $\hat{w}$ of the covering, we collect samples $(x, y - x^T \hat{w})$ where $x, y$ are from the no-index self selection model.
4. Using moments of $y - x^T \hat{w}$, we estimate $\min_{\epsilon \in [0, 1]} \|w_1 - \hat{w}\|^2$.
5. We show that $O(\delta^{-1} \epsilon^{-2} \text{poly}(B/\epsilon))$ samples suffice to get an $\epsilon$-close approximation to this quantity with probability $1 - \delta$. Setting $\delta = O(\rho^2)$ for $\rho < 1$ ensures that we get accurate estimates of this quantity for each element $\hat{w}$ of our covering.
6. As long as $w_1$ and $w_2$ are sufficiently separated, we can estimate $\hat{w}_1$ to be the minimum of our estimate over the $\epsilon$-covering, i.e., $\hat{w}_1 = \arg \min_{\epsilon \in [0, 1]} \|w_1 - \hat{w}\|$. We can then estimate $\hat{w}_2$ to be the minimizer of the estimate over points that are far enough from $\hat{w}_1$.

Turning this outline into an efficient algorithm entails tackling a few distinct technical challenges, which we address in the full version of this work [12]. First, we show how to estimate $\min_{\epsilon \in [0, 1]} \|w_1 - \hat{w}\|$ using samples $(x, y)$ from our data-generating process. Then, we show that our sequential approach to estimating $w_1$ and $w_2$ indeed suffices to recover both with good enough accuracy. Finally, we show that the error incurred by the subspace-finding step does not adversely affect our estimation.
ACKNOWLEDGEMENTS

We thank Sitan Chen for pointing out an error in the original proof of Lemma 7. Work supported by: NSF CCF-1901292, DMS-2022448, DMS2134108, DOE PHILMs project (DE-AC05-76RL01830); a Simons Investigator Award; the Simons Collaboration on the Theory of Algorithmic Fairness; a DSTA grant; an Open Philanthropy AI Fellowship; and an MSR-BAIR Open Research Commons grant.

REFERENCES

[1] John M. Abowd and Henry S. Farber. 1982. Job queues and the union status of workers. JLR Review 35, 3 (1982), 354–367.
[2] Sathi Arunachalam. 1973. Regression analysis when the dependent variable is truncated normal. Econometrica: Journal of the Econometric Society (1973).
[3] Susan Athey and Philip A. Haile. 2002. Identification of standard auction models. Econometrica 70, 6 (2002), 2107–2140.
[4] Susan Athey and Philip A. Haile. 2007. Nonparametric approaches to auctions. Handbook of econometrics 6 (2007), 3847–3965.
[5] Ernst R. Berndt, Bronwyn H. Hall, Robert E. Hall, and Jerry A. Hausman. 1974. Estimation and inference in nonlinear structural models. In Annals of Economic and Social Measurement, Volume 3, number 4 NBER, 653–665.
[6] George J. Borjas. 1987. Self-selection and the earnings of immigrants. Technical Report. National Bureau of Economic Research.
[7] Chris Brooks. 2019. Introductory econometrics for finance. Cambridge University Press.
[8] S. Bubeck, Ronen Eldan, and Joseph Lehec. 2018. Sampling from a log-concave distribution with projected Langevin Monte Carlo. Discrete Comput. Geom. 59, 4 (2018), 571–583. doi:10.1007/s00454-018-9992-1
[9] A. Cameron and Pravin K. Trivedi. 2005. Microeconometrics: methods and applications. Cambridge university press.
[10] Sitan Chen, Jerry Li, and Zhao Song. 2020. Learning mixtures of linear regressions in subexponential time via Fourier moments. In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, 587–600.
[11] Yeshwanth Cherapanamjeri, Constantinos Daskalakis, Andrew Ilyas, and Manolis Zampetakis. 2022. Estimation of Standard Auction Models. (2022).
[12] Yeshwanth Cherapanamjeri, Constantinos Daskalakis, Andrew Ilyas, and Manolis Zampetakis. 2022. What Makes A Good Fisherman? Linear Regression under Self-Selection Bias. arXiv:2205.03246 [math.ST]
[13] John Cogan. 2014. Labor supply with costs of labor market entry. In Analysis of Career Choice, Self-selection, and Access in Algorithms, Mechanisms, and Optimization. J. Comput. Syst. Sci. 81, 2 (2015), 233–250.
[14] Giora Hanoch and James P. Smith. 2014. Strategic Classification. In Frank and others (Ed.). ACM, 111–122.
[15] Joel W Hay and Randall J. Olsen. 1984. Let them eat cake: a note on comparing income and housing as a joint decision. Journal of Econometrics 28, 2 (1984), 357–382.
[16] Robert J Willis and Sherwin Rosen. 1979. Education and self-selection. Econometrica: Journal of the Econometric Society (1979), 775–789.
[17] James Heckman. 1974. Shadow prices, market wages, and labor supply. Econometrica: journal of the econometric society (1974), 679–694.
[18] James J Heckman. 1979. Sample selection bias as a specification error. Econometrica: Journal of the econometric society (1979), 153–161.
[19] Andrew Ilyas, Emmanuelou Zampetakis, and Constantinos Daskalakis. 2020. A theoretical and practical framework for regression and classification from truncated samples. In International Conference on Artificial Intelligence and Statistics.
[20] Marek Kanter and Harold Proppe. 1977. Reduction of variance for Gaussian densities via extreme values to convex sets. Journal of Multivariate Analysis (1977).
[21] Lawrence W. Kenny, Lung-Fei Lee, Gisella Maddala, and Robert P. Trost. 1979. Returns to college education: An investigation of self-selection bias based on the project talent data. International Economic Review (1979), 775–789.
[22] Mervyn A King. 1980. An econometric model of tenure choice and demand for housing as a joint decision. Journal of Public Economics 14, 2 (1980), 137–159.
[23] Anilesh K. Krishnaswamy, Haoming Li, David Rein, Hannu Zhang, and Vincent Conitzer. 2020. Classification with Strategically Withheld Data. CoRR abs/2012.10203 (2020). arXiv:2012.10203 https://arxiv.org/abs/2012.10203
[24] Zhiping Kwon and Constantine Carvalho. 2020. EM converges for a mixture of many linear regressions. In International Conference on Artificial Intelligence and Statistics. PMLR, 1727–1736.
[25] Lung-Fei Lee. 1978. Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. International economic review (1978), 415–433.
[26] Lung-Fei Lee. 2001. Self-selection. Companion to Theoretical Econometrics (2001).
[27] Lung-Fei Lee and Robert P. Trost. 1978. Estimation of some limited dependent variables models with application to housing demand. Journal of Econometrics 8, 3 (1978), 357–382.
[28] Siu Fai Leung and Shihli Yu. 1996. On the choice between sample selection and two-part models. Journal of econometrics 72, 1-2 (1996), 197–229.
[29] H. Greg Lewis. 1974. Comments on selectivity biases in wage comparisons. Journal of Political Economy 82, 1 (1974), 1145–1155.
[30] Yuanchi Li and Yingyu Liang. 2018. Learning mixtures of linear regressions with nearly optimal complexity. In Conference On Learning Theory: PMLR, 1115–1144.
[31] Zaoji Li and Nikhil Garg. 2021. Test-optimal Policies: Overcoming Strategic Behavior and Informational Gaps. In EAAOM 2021: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. ACM, 11:1–11:13.
[32] Ganghadiharo S Maddala. 1985. A survey of the literature on selectivity bias as it pertains to health care markets. Advances in health economics and health services research 6 (1985), 3–26.
[33] Ganghadiharo S Maddala. 1986. Limited-dependent and qualitative variables in econometrics. National Bureau of Economic Research.
[34] Willard G Manning, Nahuala Duan, and William H Rogers. 1987. Monte Carlo evidence on the choice between sample selection and two-part models. Journal of econometrics 35, 1 (1987), 59–82.
[35] Robert A Nakosteen and Michael Zimmer. 1980. Migration and income: the question of self-selection. Southern Economic Journal 46, 3 (1980), 775–789.
[36] Forrest D. Nelson. 1977. Censored regression models with unobserved, stochastic censoring thresholds. Journal of econometrics 6, 3 (1977), 309–327.
[37] Randal J Olsen. 1982. Distributional tests for selectivity bias and a more robust likelihood estimator. International Economic Review (1982), 223–240.
[38] Dale J. Poirier. 1981. A Switching Simultaneous Equations Model of Behaviour in Ontario. (1981).
[39] Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. 2010. A randomized algorithm for principal component analysis. SIAM J. Matrix Anal. Appl. 31, 3 (2010).
[40] Harvey S. Rosen. 1979. Housing decisions and the US income tax: An econometric analysis. Journal of Public Economics 11, 1 (1979), 1–23.
[41] Andrew Rotnitzky and James M. Robins. 1995. Semiparametric regression estimation in the presence of dependent censoring. Biometrika 82, 4 (1995), 805–820.
[42] Andra Rotnitzky and James M Robins. 1998. Semiparametric regression for repeated outcomes with nonignorable nonresponse. Journal of the American statistical association 93, 444 (1998), 1321–1339.
[43] Andrew Donald Roy. 1951. Some thoughts on the distribution of earnings. Oxford economic papers 3, 2 (1951), 135–146.
[44] Eric J. Tchetgen Tchetgen, Linbo Wang, and BaoLuo Sun. 2018. Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. Statistical Science 28, 4 (2018), 2069.
[45] James Tobin. 1958. Estimation of relationships for limited dependent variables. Econometrica: Journal of the Econometric Society (1958), 245–258.
[46] Siu Fai Leung and Shihli Yu. 1996. On the choice between sample selection and two-part models. Journal of econometrics 72, 1-2 (1996), 197–229.
[47] Andrew J. Wiles and Sharon Rosen. 1979. Education and self-selection. Journal of political Economy 87, 5, Part 2 (1979), 57–536.

Received 2022-11-07; accepted 2023-02-06

Yeshwanth Cherapanamjeri, Constantinos Daskalakis, Andrew Ilyas, and Manolis Zampetakis