Obscurable Fishermen

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

In the process of applying for a job across several similar firms, applicants often have the option to exclude certain features from a CV, e.g., photo, GPA, standardized test scores, etc. If applicants desire the best income offer possible and can submit multiple applications to similar positions, they may exclude or include various of these optional features on different applications to see which yields the best results, eventually accepting the highest offer. But if an analyst then would like to estimate what makes a good worker using the applications (features) and incomes (outcomes) of the finally accepted offers, she will have an endogeneity problem! The excluded features, which we term "obscured" will be missing not at random, meaning simple imputation methods such as the conditional expectation will result in biased estimates. We formalize this problem and present a preliminary result in which we reduce our obscured setting to a high-dimensional instantiation of the setting from Cherapanamjeri et al. [1]. Unfortunately, this reduction increases the number of variables by an amount combinatorial in the dimension of the problem, meaning the algorithmic tool for this setting will not be efficient in the original parameters. We present possible next steps such as approximate SGD on the MLE and kernelization to get around the increase in variables.

1 Introduction

Borrowing the fisherman occupation from Roy [5], we introduce what we call our "obscurable fisherman" problem much like Cherapanamjeri et al. [1]:

Suppose agents in a small village have only *one* industry available to them: fishing. They may apply to be fishermen at various very similar firms and receive an offer of income according to some common policy based on the features presented in their application. However, all applications may *optionally* include FAT (Fishing Aptitude Test) scores among other features. Every agent sends applications to various firms including/excluding various optional features. Because all agents test the waters by including/excluding features, to calculate job offers, firms just take their best guess (a conditional expectation) as to what the values of the missing features are on each application. Eventually, each agent accepts the fisherman job offer the gives the highest income.

A statistician gets access to accepted offers and applications (with obscured values). She asks:

What makes a good fisherman?

This anecdote sets up an endogeneity problem for the statistician. If applicants desire the best income offer possible and can submit multiple applications to similar firms, the application of the offer they eventually select will have *strategically* obscured features. These features are missing not at random (MNAR)[6] and so should not be thrown out or imputed with conditional averages by the statistician. Are there algorithms efficient in time and sample complexity that the statistician can use? In Section 2, we formally present a model of this strategic feature obscuration, which we call the obscurable fisherman setting. The statistician must estimate $\mathbf{w} \in \mathbb{R}^d$, the coefficients of a linear policy assigning income offers based on features. Any subset of the first k features can be obscured. Agents may test every possible obscuration pattern and then accept the one that yields the highest

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

(noisy) outcome. In Section 3, we present preliminary results as to the estimation of $\mathbf{w}$ using the linear estimation under model self-selection tools from Cherapanamjeri et al. [1]. Our obscurable fisherman setting, while a strategic missing data problem, can also be viewed as a model selection problem. As such, we create a reduction of a generic obscurable fisherman dataset, D, to a "good fisherman" (à la Cherapanamjeri et al. [1]) dataset, $\tilde{D}$, that is the best response to a set of models in the form of their setting. Unfortunately, the reduction requires an increase in the number of features that is combinatorial in the original dimension. Thus, directly using the algorithm they present is not efficient in the parameters of the obscurable fisherman setting. In Section 4 we discuss future work we hope will yield better results.

1.1 Related works

Cherapanamjeri et al. [1] is the most direct inspiration for our model; they consider agents who select (using a function such as max) between k linear models and a statistician that estimates the $\mathbf{w}_j^\star$ coefficient for each model. In our version, there is only one underlying linear coefficient vector, $\mathbf{w}$, and instead agents select from obscuration patterns. The strategic selection of obscuration patterns means that we consider estimation under missing not at random data (MNAR) which was first formally defined by Rubin [6] and cannot generally be fixed with imputation of conditional averages. See Little [3] for a taxonomy and survey of estimation methods under various missing data patterns. Additionally, while we focus on a linear coefficient statistical estimation problem, there are similar questions that involve creating an optimal *classifier* given strategically obscured data. Krishnaswamy et al. [2] design classification algorithms that perform well under strategically obscured data and Liu and Garg [4] evaluate whether it is possible to build a classifier that does not implicitly penalize agents who choose to obscure test score data in university admissions.

2 Model

2.1 Agents

Each agent (she), $i \in [n]$ has feature vector: $\mathbf{x}^{(i)} \in \mathbb{R}^d$ drawn from a joint distribution $\mathcal{D}(\mathbf{x})$. The first $k < d$ of d features are optional. That is, features at any subset $\mathcal{O}_j \subseteq [k]$ of indices may be obscured. We will call $\mathcal{O}_j$, a set obscured indices, an *obscuration pattern*. Let $\mathcal{O}_j \in \mathcal{P}$ where $\mathcal{P}$ is the set of all obscuration patterns. Clearly, $|\mathcal{P}| = \sum_{l=0}^{k} \binom{k}{l}$.

Definition 2.1 (Obscured feature vector)**.** *For a true feature vector,* $\mathbf{x}^{(i)}$*, and obscuration pattern,* $\mathcal{O}_j$*, an* obscured feature vector $\mathbf{x}_j^{(i)} \in \mathbb{R}^d$ *is the same as* $\mathbf{x}^{(i)}$ *except all elements at indices in the obscuration pattern are obscured. Formally:* $x_{j,u}^{(i)} = x_u^{(i)}\ \forall u \in [d] \setminus \mathcal{O}_j$ *and* $x_{j,h}^{(i)} = o\ \forall h \in \mathcal{O}_j$*. Where* o *(for* obscured*) indicates that this a missing value and holds no inherent numerical meaning.*

When os are replaced with conditional expectations:

Definition 2.2 (Expected feature vector)**.** *For an obscured feature vector,* $\mathbf{x}_j^{(i)}$*, and obscuration pattern,* $\mathcal{O}_j$*, an* expected feature vector*,* $\hat{\mathbf{x}}_j^{(i)} \in \mathbb{R}^d$*, is the same as* $\mathbf{x}^{(i)}$ *except all elements at indices in the obscuration pattern are expectations conditioned on all unobscured variables. Formally:* $\hat{x}_{j,u}^{(i)} = x_u^{(i)}\ \forall u \in [d] \setminus \mathcal{O}_j$ *and* $\hat{x}_{j,h}^{(i)} = \mathbb{E}[x_h | U(\mathcal{O}_j)] \quad \forall h \in \mathcal{O}_j$ *where* $U(\mathcal{O}_j)$ *are the* elements *at* un*obscured indices, i.e.,*$U(\mathcal{O}_j) := \{x_u^{(i)} | u \in [d] \setminus \mathcal{O}_j\}$

The agent privately tests a given linear model on each *expected feature vector* and selects the best outcome and obscuration pattern. That is she selects:

$$y^{(i)} := \max_{j \in |\mathcal{P}|} f_j(\mathbf{x}^{(i)}); j^{\star(i)} := \arg\max_{j \in |\mathcal{P}|} f_j(\mathbf{x}^{(i)}) \quad \text{where} \quad f_j(\mathbf{x}^{(i)}) := \mathbf{w}^\top \hat{\mathbf{x}}_j^{(i)} + \varepsilon_j$$

Noise $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$ is iid and drawn separately for each model. Notice that obscuration pattern and model are functionally the same. That is, if an agent chooses obscuration pattern j, she has chosen model j. We will use these terms interchangeably.

2.2 Learner

The learner (he) receives a dataset of the selected *obscured feature vectors* and best outcomes:

$$D := \{\mathbf{x}_{j^\star}^{(i)}, y^{(i)}, j^{\star(i)}\}_{i \in [n]}$$

First, note D will have data that is missing not at random (MNAR). Second, note that the obscuration pattern can be directly gleaned from $\mathbf{x}_{j^\star}^{(i)}$, thus receiving an obscured feature vector also allows the learner to know which model was selected, $j^\star$.

The learner would like to know what makes a good outcome, i.e., estimate $\mathbf{w}$, despite the non-randomness of the missing data. It is clear to see that the obscured setting creates endogeneity due to correlated errors and thus standard OLS estimates (with either conditional expectation imputations or dropping of missing data) would be biased.

Example 2.1 (Learner does biased OLS)**.** *Suppose* $\mathbf{w} := (1,1)$*,* $\sigma^2 = \frac{1}{5}$*, and both* $x_1, x_2 \sim$ $\text{UNIF}(-1,2)$*. Thus,* x_1, x_2 *are independent of one another and* $\mathbb{E}[x_2, | x_1 = x_1^{(i)}] = .5$ *for all* $x_1^{(i)}$*. We simulate* $n = 200$ *of this example and imagine the learner does OLS on the full data set (i.e. allowing* $o = .5$*) and also on just the points that have no missing data. This is presented in Figure 1. Clearly both OLS estimators are biased.*

What time and sample efficient algorithms may the learner run such that he achieves an ε-unbiased estimator of $\mathbf{w}$ despite strategically obscured data?

3 A reduction to Cherapanamjeri et al. [1] self-selection

In these results, we will detail a (relatively inefficient) approach to estimating $\mathbf{w}$ when conditional expectations are known using existing model selection tools from Cherapanamjeri et al. [1]. Improved methods and future work are discussed in Section 4.

Assumption 3.1 (Known Conditional Expectations)**.** $\mathbb{E}[x_h | U(\mathcal{O}_j)]$ *is known* $\forall h \in \mathcal{O}_j, \forall \mathcal{O}_j \in \mathcal{P}$

Assumption 3.1 is a strong assumption stating that the expectation for all obscurable features conditioned on any possible set of unobscured features is known.

3.1 Constructing a good fisherman setting

In the known-index model selection setting of Cherapanamjeri et al. [1], agents select a linear model, $f_j(\mathbf{x}) = \mathbf{w}_j^{\star\top} \mathbf{x}^{(i)} + \varepsilon_j$, that provides the best sampled outcome. Importantly, the resulting dataset provides $\{\mathbf{x}^{(i)}, y^{(i)}, j^{\star(i)}\}_{i \in [n]}$. Thus, while the provided *outcome* depends on the selected model, the *feature set* does not. We will transform our learner's dataset, D, which contains the problematic $\mathbf{x}_{j^\star}^{(i)}$ obscured features, into $\tilde{D}$, a dataset that could have come from a good fisherman setting. In constructing $\tilde{D}$ we will shift each obscurable feature such that $w_h x_h \geq 0 \forall h \in [k]$. Thus we need:

Assumption 3.2 (Obscurable features are sufficiently bounded)**.** *The following must hold for all obscurable feature indices,* $h \in [k]$*: If* $w_h > 0$ *then* $l_h \leq x_h \quad \forall x_h$*. If* $w_h < 0$ *then* $u_h \geq x_h \quad \forall x_h$

Definition 3.1 ($\tilde{D}$, good fisherman transformed dataset)**.** $\tilde{D} := \{1, \tilde{\mathbf{x}}^{(i)}, y^{(i)}, j^{\star(i)}\}_{i \in [n]}$ *where: each* $\tilde{\mathbf{x}}^{(i)} \in \mathbb{R}^{g(k,d)}$*,* $g(k,d) := k \sum_{l=0}^{k-1} \binom{k-1}{l} + d$ *and is constructed according to Algorithm 1*

Notice that $\tilde{\mathbf{x}}^{(i)}$ no longer depends on the model selection! The constructed feature set is the original with two key changes: (1) a shift on obscured variables (2) $k \sum_{l=0}^{k-1} \binom{k-1}{l}$ additional variables to "one-hot encode" for every relevant conditional expectation. For a given obscurable variable, x_h, Algorithm 1 adds a variable for every obscuration pattern it could be a part of. We will now show that $\tilde{D}$ could have come from a valid good fisherman setting.

Theorem 3.3 (Reduction to good fisherman self-selection)**.** *Using the same* ε_j *as those from the obscured models, dataset* $\tilde{D}$ *would be the best response to a maximizing self-selection over* $|\mathcal{P}|$ *linear models where:* $\tilde{f}_j(\tilde{\mathbf{x}}^{(i)}) := w_0 + \tilde{\mathbf{w}}_j^\top \tilde{\mathbf{x}}^{(i)} + \varepsilon_j$*, each* $\tilde{\mathbf{w}}_j$ *is constructed according to Algorithm 2, and*

$$w_0 := \sum_{h \in [k]} w_h \left(-\mathbb{1}_{w_h \geq 0} |\min\{0, l_h\}| + \mathbb{1}_{w_h < 0} |\max\{0, u_h\}| \right)$$

To prove this, we need to show that for every agent of $\tilde{D}$, a best response in this good fisherman setting would indeed still be the $j^\star$th model and the $j^\star$th model would produce that outcome. The intuition of this result can be seen directly from the following lemma statements. First, the transformed features, when multiplied by the $\tilde{\mathbf{w}}_{j^\star}$ and added to $\varepsilon_j + w_0$, produce the same outcome as $f_{j^\star}(\mathbf{x}^{(i)})$!

Lemma 3.1 (Output of $j^\star$ model is stable)**.** *For a point,* $\mathbf{x}_{j^\star}^{(i)}$ *we have:* $\tilde{f}_{j^\star}(\tilde{\mathbf{x}}^{(i)}) = f_{j^\star}(\mathbf{x}_{j^\star}^{(i)})$

Second, due to the construction of $\tilde{\mathbf{x}}^{(i)}$ and $\tilde{\mathbf{w}}_{j'}$ for all $j' \neq j^\star$, the inner product corresponding to each good fisherman model $+\varepsilon_{j'} + w_0$, will yield either the same output or less than the private tests the agent did for obscuration pattern j'.

Lemma 3.2 (Output of j' models is lowered)**.** *For a point,* $\mathbf{x}_{j^\star}^{(i)}$*:* $\tilde{f}_{j'}(\tilde{\mathbf{x}}^{(i)}) \leq f_{j'}(\mathbf{x}_{j'}^{(i)}) \quad \forall j' \neq j^\star$

With these lemmas, the proof of Theorem 3.3 is very direct, clearly

$$\tilde{f}_{j^\star}(\tilde{\mathbf{x}}^{(i)}) = f_{j^\star}(\mathbf{x}_{j^\star}^{(i)}) \geq f_{j'}(\mathbf{x}_{j'}^{(i)}) \geq \tilde{f}_{j'}(\tilde{\mathbf{x}}^{(i)}) \quad \forall j' \neq j^\star$$

Thus $j^\star$ is the best response and we still have the same $y^{(i)}$!

3.2 Estimating $\mathbf{w}$

After converting the dataset to one that could be the result of a maximum selection problem over linear models, with a few additional assumptions, the learner can run the algorithm presented by Cherapanamjeri et al. [1] to estimate $\tilde{\mathbf{w}}_j \quad \forall j \in |\mathcal{P}|$ and thus have estimates for $\mathbf{w}$! Recall that from Algorithm 2, we know which elements of $\tilde{\mathbf{w}}_j$ are equivalent to which elements of $\mathbf{w}$, so we can directly construct good estimates of $\mathbf{w}$ from good estimates of $\tilde{\mathbf{w}}_j$.

Corollary 3.1 (Corollary of Thm 3.3 and Thm 1 [1])**.** *Let* $\{\mathbf{x}_{j^\star}, y^{(i)}, j^{\star(i)}\}_{i\in[n]}$ *be* n *observations from an obscurable fisherman model as described in Section 2. Let* $\hat{\mathbf{w}}$ *be the estimator of the* $\mathbf{w}$*. Given assumptions 3.1 and 3.2, as well as the additional assumptions 1, 2, and 3 from Cherapanamjeri et al. [1], there exists an algorithm such that with probability at least .99,*

$$\|\mathbf{w} - \hat{\mathbf{w}}\|_2^2 \leq poly(\sigma, |\mathcal{P}|, 1/\alpha, B, C)\frac{\log n}{n}$$

under $poly(n, g(k,d), |\mathcal{P}|, 1/\alpha, B, C, \sigma, 1/\sigma)$ *running time.*

Where α, B, C *are constants defined by assumptions 1, 2, and 3 from Cherapanamjeri et al. [1]*

Unfortunately, in the parameters of the obscured problem, this is not a very efficient result. Recall that $|\mathcal{P}| = \sum_{l=0}^{k}\binom{k}{l}$ and $g(k,d) := k\sum_{l=0}^{k-1}\binom{k-1}{l} + d$. The number of obscuration patterns, i.e., models and the number of variables is combinatorial in the number of obscurable variables, which could be as large as $d-1$!

4 Conclusion and Future Work

We present a model of agents being able to self-select their set of obscurable features. We provide preliminary results of the estimation of linear model coefficients despite the selection bias that arises from strategic obscuration. Estimation in this setting can be viewed with both a missing not at random (MNAR) problem lens and model self-selection lens. Importantly, under the model-selection perspective, we can reduce the problem to a high-dimensional version of good fisherman setting[1]. Unfortunately, the reduction increases the number of data dimensions such that known algorithms will not be efficient in the original dimensions of the problem. Further, the reduction requires knowledge of conditional expectations, which is a strong assumption.

In the extended work, we hope to prove an alternate $\mathbf{w}$ estimation method through a more direct MLE estimation similar to that which done by Cherapanamjeri et al. [1]. Because the presented result has shown that the obscurable fisherman setting could be reduced to a version of the good fisherman one, it may be that there exists an analogous population likelihood function that is strongly concave with a stationary point at $\mathbf{w}$, which could be approximately optimized via SGD. Alternatively, as there is a combinatorial (in d) variable problem in the reduction, there may be applications of kernelization that remove this issue.

References

[1] Yeshwanth Cherapanamjeri, Constantinos Daskalakis, Andrew Ilyas, and Manolis Zampetakis. What makes a good fisherman? linear regression under self-selection bias. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, page 1699–1712, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399135. doi: 10.1145/3564246.3585177. URL `https://doi.org/10.1145/3564246.3585177`.

[2] Anilesh K. Krishnaswamy, Haoming Li, David Rein, Hanrui Zhang, and Vincent Conitzer. Classification with strategically withheld data. *Proceedings of the AAAI Conference on Artificial*

Intelligence, 35(6):5514–5522, May 2021. doi: 10.1609/aaai.v35i6.16694. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16694`.

[3] Roderick J. A. Little. Regression with missing x's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237, 1992. ISSN 01621459, 1537274X. URL `http://www.jstor.org/stable/2290664`.

[4] Zhi Liu and Nikhil Garg. Test-optional policies: Overcoming strategic behavior and informational gaps. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385534. doi: 10.1145/3465416.3483293. URL `https://doi.org/10.1145/3465416.3483293`.

[5] A. D. Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146, 1951. ISSN 00307653, 14643812. URL `http://www.jstor.org/stable/2662082`.

[6] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444, 14643510. URL `http://www.jstor.org/stable/2335739`.

A Supplementary material

A.1 Supplementary material for Section 2

A.1.1 Example

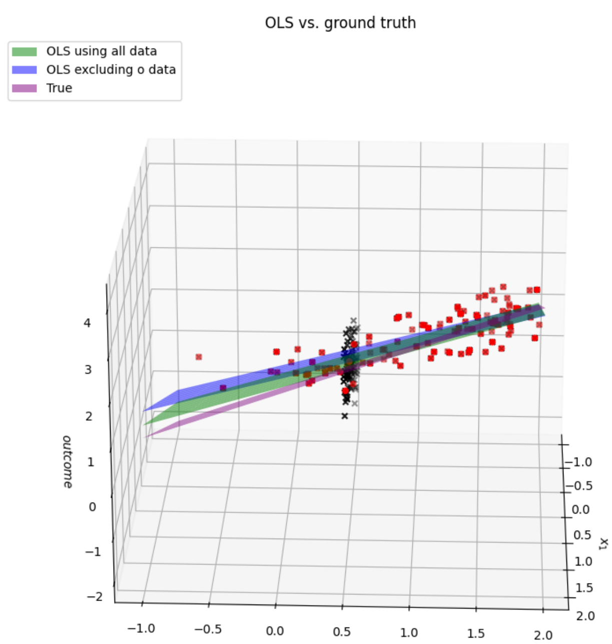

Figure 1: Learner runs OLS on $n = 200$ datapoints detailed in Example 2.1. Black Xs represent the points with obscured x_2 elements (missing x_2 is imputed as .5 for the green OLS). Red points represent those which are not obscured at all.

A.2 Supplementary material for Section 3

A.2.1 Algorithms to compute the reduction

Algorithm 1 Construct $\tilde{\mathbf{x}}^{(i)}$

Require: $\mathbf{x}_{j^\star}^{(i)}; \mathbb{E}[x_h|U(\mathcal{O}_j)] \quad \forall h \in \mathcal{O}_j, \forall j \in |\mathcal{P}|; l_h, u_h \quad \forall h \in [k]$

  $\tilde{\mathbf{x}}^{(i)} = []$
  for $h \leftarrow 1$ to k **do** ▷ loop adds k elements
    if $x_{j^\star,h}^{(i)} \neq o$ **then**
      if $w_h \geq 0$ **then**
        Append $x_{j^\star,h}^{(i)} + |\min\{0, l_h\}|$ **to** $\tilde{\mathbf{x}}^{(i)}$ ▷ if not obscured, add shifted known value
      else
        Append $x_{j^\star,h}^{(i)} - |\max\{0, u_h\}|$ **to** $\tilde{\mathbf{x}}^{(i)}$ ▷ if not obscured, add shifted known value
    else
      Append 0 **to** $\tilde{\mathbf{x}}^{(i)}$ ▷ if obscured, add 0
  for $h \leftarrow 1$ to k **do** ▷ loop adds $k \sum_{j=0}^{k-1} \binom{k-1}{j}$ elements
    for $l \leftarrow 0$ to $k-1$ **do**
      for all $S \in \binom{[k]\setminus\{h\}}{l}$ **do** ▷ loop through all obscuration patterns that include h
        $\mathcal{O} \leftarrow S \cup [h]$ ▷ construct obscuration pattern
        $U(\mathcal{O}) \leftarrow \{x_{j^\star,u}^{(i)} | u \in [d] \setminus \mathcal{O}\}$ ▷ construct set of unobscured elements
        if $U(\mathcal{O})$ contains elements s.t. $x_{j^\star,u}^{(i)} = o$ **then** ▷ check if these unobscured are o
          Append 0 **to** $\tilde{\mathbf{x}}^{(i)}$ ▷ if yes, then conditional exp incomputable
        else
          if $w_h \geq 0$ **then**
            Append $\mathbb{E}[x_h|U(\mathcal{O})] + |\min\{0, l_h\}|$ **to** $\tilde{\mathbf{x}}^{(i)}$ ▷ if no, add shifted cond exp
          else
            Append $\mathbb{E}[x_h|U(\mathcal{O})] - |\max\{0, u_h\}|$ **to** $\tilde{\mathbf{x}}^{(i)}$ ▷ if no, add shifted cond exp
  for $u \leftarrow k+1$ to d **do** ▷ loop adds $d-k$ elements
    Append $x_u^{(i)}$ **to** $\tilde{\mathbf{x}}^{(i)}$ ▷ add unobscured value
  return $\tilde{\mathbf{x}}^{(i)}$ ▷ constructed feature vector $\in \mathbb{R}^{g(k,d)}$

Algorithm 2 Construct $\tilde{\mathbf{w}}_j$ to match obscuration pattern, $\mathcal{O}_j$

Require: $\mathcal{O}_j$, the obscuration pattern of model j

  $\tilde{\mathbf{w}}_j = []$
  for $h \leftarrow 1$ to k **do** ▷ loop adds k elements
    if $h \in \mathcal{O}_j$ **then**
      Append 0 **to** $\tilde{\mathbf{w}}_j$ ▷ if h is obscured in this model don't turn on w
    else
      Append w_h **to** $\tilde{\mathbf{w}}_j$ ▷ if h is in this model turn on w
  for $h \leftarrow 1$ to k **do** ▷ loop adds $k \sum_{j=0}^{k-1} \binom{k-1}{j}$ elements
    for $l \leftarrow 0$ to $k-1$ **do**
      for all $S \in \binom{[k]\setminus\{h\}}{l}$ **do** ▷ loop through all obscuration patterns that include h
        $\mathcal{O} \leftarrow S \cup [h]$
        if $\mathcal{O} = \mathcal{O}_j$ **then**
          Append w_h **to** $\tilde{\mathbf{w}}_j$ ▷ if this conditional exp is in this model, turn on w
        else
          Append 0 **to** $\tilde{\mathbf{w}}_j$ ▷ if this conditional exp is not in this model, don't turn on w
  for $u \leftarrow k+1$ to d **do** ▷ loop adds $d-k$ elements
    Append w_u **to** $\tilde{\mathbf{w}}_j$ ▷ unobscurable vars are always in the model, so always have their coefficients on.
  return $\tilde{\mathbf{w}}_j$

A.2.2 Missing proofs

Proof of Lemma 3.1. First, note that Algorithm 1 shifts all obscurable variables, x_h by $\mathbb{1}_{w_h \geq 0}|\min\{0, l_h\}| - \mathbb{1}_{w_h<0}|\max\{0, u_h\}|$ and then $\tilde{f}_j$ adds a constant term

$$w_0 := \sum_{h \in [k]} w_h \left(-\mathbb{1}_{w_h \geq 0}|\min\{0, l_h\}| + \mathbb{1}_{w_h<0}|\max\{0, u_h\}|\right)$$

We can also do this without changing the outcome of any model (or model selection) to the obscurable setting because this is equivalent to adding and subtracting terms. For the remainder of the proof, we will refer to this affine version of the model (with w_0) and treat the obscurable variables from D as if they are shifted.

First we shall consider the function of Algorithm 1 and 2. Notice that, for every i, Algorithm 1 constructs a vector such that the first k elements correspond to [shifted] actual values of x_h where possible. Then $k\sum_{j=0}^{k-1}\binom{k-1}{j}$ elements are added to correspond to every obscurable variable's possible [shifted] conditional expectation. Finally $d-k$ elements at the end are simply the unobscurable values that must be present. Algorithm 2 on the other hand follows the same construction pattern, but instead, for a given $\mathcal{O}_j$, or equivalently, for an given model, places a w_h in the element spot that represents which conditional expectation (or unobscured value) appears in the model. This is conceptually very similar to a one-hot encoding!

Thus, for $\mathcal{O}_{j^\star}$, Algorithm 2 constructs a $\tilde{\mathbf{w}}$ that

1. For unobscurable variables, indexed by u, assigns $\tilde{w}_u = w_u$ to $\tilde{\mathbf{x}}$ element slots corresponding to each said unobscurable variable
2. For each obscurable variable, indexed by h, only assigns $\tilde{w}_h = w_h$ to the $\tilde{\mathbf{x}}^{(i)}$ element slot corresponding to obscurable variable OR conditional expectation appearing in the given $\mathbf{x}_{j^\star}^{(i)}$.

As a result,

$$\varepsilon_{j^\star} + w_o + \tilde{\mathbf{w}}_{j^\star}^\top \tilde{\mathbf{x}}^{(i)} = \varepsilon_{j^\star} + w_o + \mathbf{w}^\top \hat{\mathbf{x}}_{j^\star}^{(i)}$$

Where $\hat{\mathbf{x}}_{j^\star}^{(i)}$ is the *shifted* version of the expected feature vector corresponding to obscured feature vector. This is equivalent to the statement in the lemma. □

Proof of Lemma 3.2. As in the proof of Lemma 3.1, note that Algorithm 1 shifts all obscurable variables, x_h by $\mathbb{1}_{w_h \geq 0}|\min\{0, l_h\}| - \mathbb{1}_{w_h<0}|\max\{0, u_h\}|$ and then $\tilde{f}_j$ adds a constant term

$$w_0 := \sum_{h \in [k]} w_h \left(-\mathbb{1}_{w_h \geq 0}|\min\{0, l_h\}| + \mathbb{1}_{w_h<0}|\max\{0, u_h\}|\right)$$

We can also do this without changing the outcome of any model (or model selection) to the obscurable setting without changing the outcome of any model (or model selection) because this is equivalent to adding and subtracting terms. For the remainder of the proof, we will refer to this affine version of the model (with w_0) and treat the obscurable variables from D as if they are shifted.

First we shall consider the function of Algorithm 1 and 2. Notice that, for every i, Algorithm 1 constructs a vector such that the first k elements correspond to [shifted] actual values of x_h where possible. Then $k\sum_{j=0}^{k-1}\binom{k-1}{j}$ elements are added to correspond to every obscurable variable's possible [shifted] conditional expectation. Finally $d-k$ elements at the end are simply the unobscurable values that must be present. Algorithm 2 on the other hand follows the same construction pattern, but instead, for a given $\mathcal{O}_j$, or equivalently, for an given model, places a w_h in the element spot that represents which conditional expectation (or unobscured value) appears in the model. This is conceptually very similar to a one-hot encoding!

An important nuance happens when the obscuration pattern of $\tilde{\mathbf{w}}_{j'}$ does not match the obscuration pattern implicit to $\mathbf{x}_{j^\star}^{(i)}$. Algorithm 1 sets as 0 any elements of $\tilde{\mathbf{x}}^{(i)}$ that represent conditional expectations (or obscurable values) that cannot be computed from $\mathbf{x}_{j^\star}^{(i)}$, which may have missing values. For example, if $\mathbf{x}_{j^\star}^{(i)} = (o, 1, 4)$ and the first two variables obscurable, one of the elements in the corresponding $\tilde{\mathbf{x}}^{(i)}$ will be for $\mathbb{E}[x_2|x_1 =?, x_3 = 4]$, but this will be incomputable since obviously x_1 is obscured.

Consider an arbitrary element $\tilde{x}_q^{(i)}$ associated with the obscurable element at index h. That is, element at index q of $\tilde{\mathbf{x}}$ is some conditional expectation or value of obscurable element at index h of $\mathbf{x}^{(i)}$. As a result of Algorithm 1, if this conditional expectation or value is incomputable as a result of the the obscuration pattern of $\mathbf{x}_{j^\star}^{(i)}$ because relevant values are missing, $\tilde{x}_q^{(i)} = 0$. This means, for any obscuration patterns, $\mathcal{O}_{j'}$, that $\tilde{x}_q^{(i)}$ is represented in, while Algorithm 2 will construct a $\tilde{\mathbf{w}}$ that sets $\tilde{w}_q = w_h$, $\tilde{w}_q\tilde{x}_q^{(i)} = 0$! Meanwhile, in the earlier private test for that obscuration done by the agent, she tested $\varepsilon_{j'} + w_o + \mathbf{w}^\top\hat{\mathbf{x}}_{j'}^{(i)}$, and she would have:

$$w_h\hat{x}_{j',h} = w_h\left(\mathbb{E}[x_h|U(\mathcal{O}_{j'})] + \mathbb{1}_{w_h\geq 0}|\min\{0, l_h\}| - \mathbb{1}_{w_h<0}|\max\{0, u_h\}|\right) \geq 0$$

(again, for this proof we redefine $\hat{\mathbf{x}}_{j'}$ as the *shifted* expected feature vector) because she had access to missing variables and by construction of the shift its greater than or equal to zero. As a result we see that:

$$\varepsilon_{j'} + w_o + \mathbf{w}^\top\tilde{\mathbf{x}}_{j'}^{(i)} \leq \varepsilon_{j'} + w_o + \mathbf{w}^\top\hat{\mathbf{x}}_{j'}^{(i)} \quad \forall j' \neq j^\star$$

Where $\hat{\mathbf{x}}_{j^\star}^{(i)}$ is the *shifted* version of the expected feature vector corresponding to obscured feature vector. This is equivalent to the statement in the lemma. □

Proof of Theorem 3.3. We need to show that, for every i, were $\tilde{\mathbf{x}}^{(i)}$ the underlying true features generated, then $\max_{j\in|\mathcal{P}|}\tilde{f}_j(\tilde{\mathbf{x}}^{(i)})$ would generate the $y^{(i)}$ and the $j^{\star(i)}$ given. Equivalently, that a best response would indeed be the $j^\star$th model and the $j^\star$th model would produce that outcome.

The result directly follows from Lemma 3.1 and 3.2. First, Lemma 3.1 confirms that for all agents i, model $j^\star$ does yield the same output under both $f_{j^\star}$ and $\tilde{f}_{j^\star}$ settings. All that remains to show is that $\tilde{f}_{j^\star}$ is in fact the best outcome of all $\tilde{f}_j$. Notice that for an point $\mathbf{x}_{j^\star}^{(i)}$, we know that $f_{j^\star}(\mathbf{x}_{j^\star}^{(i)}) > f_{j'}(\mathbf{x}_{j'}^{(i)}) \quad \forall j' \neq j^\star$ because the agent selected $j^\star$. From Lemmas 3.1 and 3.2:

$$\tilde{f}_{j^\star}(\tilde{\mathbf{x}}^{(i)}) = f_{j^\star}(\mathbf{x}_{j^\star}^{(i)}) > f_{j'}(\mathbf{x}_{j'}^{(i)}) \geq \tilde{f}_{j'}(\tilde{\mathbf{x}}^{(i)}) \quad \forall j' \neq j^\star$$

Thus $j^\star$ is the best response in the transformed good fisherman model set as well! □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: we have the results and proofs described

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: we've included concerns about the efficiency and assumptions of the results

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: assumptions are delineated. proofs are in the appendix

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes] .

Justification: no experiments, we do have a very simple example, the explanation would be enough to recreate it

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: we have a very simple example of biased OLS, our main contribution is the theory and this example could be easily reconstructed. We are happy to provide code by request though.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA] .

Justification: no experiments

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA] .

   Justification: no experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA] .

   Justification: no experiments

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: all requirements are conformed to

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA] .

    Justification: This is a preliminary theory paper with a result that is inefficient, we don't expect this will cause any societal impact, haha

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: preliminary theory paper, no risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring

that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA] .

Justification: preliminary theory paper, no existing assets

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: preliminary theory paper, no new assets

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: preliminary theory paper, no human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: preliminary theory paper, no IRB needed

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: no LLM usage

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.