Learning to Personalize Treatments When Agents Are Strategic *

Evan Munro †

February 23, 2021

Abstract

There is increasing interest in allocating treatments based on observed individual data: examples include heterogeneous pricing, individualized credit offers, and targeted social programs. Policy targeting introduces incentives for individuals to modify their behavior to obtain a better treatment. We show standard risk minimization-based estimators are sub-optimal when observed covariates are endogenous to the treatment allocation rule. We propose a dynamic experiment that converges to the optimal treatment allocation function without parametric assumptions on individual strategic behavior, and prove that it has regret that decays at a linear rate. We validate the method in simulations and in a small MTurk experiment.

Keywords: Design of Experiments, Robustness, Optimization
JEL Codes: D82, C90, C61

*I thank Mohammad Akbarpour, Susan Athey, Anirudha Balasubramanian, Martino Banchio, Alex Frankel, Guido Imbens, Stefan Wager, Bob Wilson, and Kuang Xu for helpful comments and discussions, and the staff at the GSB Behavioral Lab for assistance with running the Amazon Mechanical Turk experiment. Replication code and data for the analysis in the paper is available at http://github.com/evanmunro/personalized-policy.

†Stanford Graduate School of Business. munro@stanford.edu

1 Introduction

Personalizing treatments on the basis of individual-level data is increasingly common. Examples include using neural networks to construct heterogeneous pricing functions from internet browsing data, allocating credit based on unconventional characteristics like phone usage (Björkegren and Grissen, 2019), and predicting quality of content on social media platforms using click-through rates. When treatment effects are heterogeneous (Nie and Wager, 2017; Wager and Athey, 2018; Künzel et al., 2019), and observed characteristics are informative about individual level treatment effects, appropriately designed personalized policy can improve outcomes compared to a uniform policy.

If an individual's treatment assignment impacts their utility, they have an incentive to change their behavior to receive a better treatment. This means that the distribution of observed characteristics is dependent on how treatments are targeted. For example, in response to heterogeneous pricing of consumer products, individuals may adjust device usage to receive cheaper prices. In response to heterogeneous treatment of media content on social media platforms based on engagement metrics, some publishers may purchase fake clicks and comments, while others may invest more to improve the quality of their content. Compared to when observed covariates are exogenous, manipulation can make characteristics a less reliable basis for allocating treatments, see Ball (2020), while other behavioral changes may have a positive impact on the planner objective. For example, Jin and Vasserman (2019) show that drivers who opt-in to a program that makes insurance discounts dependent on driving behavior exhibit safer driving after the monitoring is installed.

The first contribution of the paper is providing a general model of how policy targeting induces endogeneity in observed individual-level data and why, as a result, we require new methods for estimating treatment allocation rules. We define the optimal policy as the parametric function of observed covariates that maximizes a specified planner objective. We propose estimating the optimal policy using a novel dynamic experiment that randomizes how treatments depend on observed covariates and takes gradient steps towards the optimum. It is difficult to specify a parametric model of individual strategic behavior in response to policy targeting that is valid in a variety of different settings. A key feature of the estimation method is that it does not require a parametric model of individual strategic behavior, so it is robust to a variety of different assumptions on how targeting affects the distribution of observed covariates and outcomes.

Standard statistical approaches used to estimate treatment allocation rules fall short. Empirical risk minimization approaches, which are standard in industry, assume that observed characteristics are exogeneous, and fail to model that the distribution of observed data will shift in response to changes in the treatment allocation function. This paper formalizes a type of Lucas Critique (Lucas et al., 1976) for causal inference. An online seller may be interested in exploiting correlation between browsing history and willingness to pay for a consumer good. However, switching from a uniform pricing policy to one that varies pricing for different individuals based on their browsing history introduces incentives for individuals to mimic the browsing history of an individual with low willingness to pay, and receive a lower price. As a result, a policy that switches from uniform to heterogeneous pricing may not raise as much revenue as expected, unless the impact of strategic responses on price discrimination is taken into account.

After a brief review of the related literature, we describe formally the problem of a planner who would like to allocate a continuous treatment W_i optimally using some parametric function of observed individual characteristics $\boldsymbol{X}_i$. Individual outcomes Y_i are observed post-treatment and may depend on the treatment received. Individuals have some unobserved multidimensional type $\boldsymbol{\theta}_i$ which influences their outcome Y_i and how the individual reports $\boldsymbol{X}_i$ in order to receive a better treatment. The goal is to find the treatment allocation function that maximizes the expected value of some observable function of the treatment and outcome for each individual. This objective may be a revenue or engagement metric for a company, or a welfare objective for a government.

We then present in detail two special cases of the general model presented. The first is the setting of strategic classification, see Hardt et al. (2016), Perdomo et al. (2020), Björkegren et al. (2020), and Frankel and Kartik (2020). In the strategic classification setting, the treatment is the prediction of an outcome Y_i using reported characteristics $\boldsymbol{X}_i$. In the absence of incentives to manipulate, the reported characteristics $\boldsymbol{X}_i$ are correlated with the outcome Y_i. The planners' goal is to minimize the classification error when agents have the incentive to manipulate $\boldsymbol{X}_i$ to receive a higher prediction. This example shows that the methodology introduced solves the version of the problem that has been studied most frequently in the literature, without relying on parametric assumptions that are standard in existing work.

The second example is that of a profit maximizing seller of a consumer good who would like to price discriminate based on observed characteristics $\boldsymbol{X}_i$. The treatment is the price shown to each individual and the outcome is how much that individual purchases. This example shows how the method in this paper applies to a variety of more general optimal policy problems with distinct treatments, planner objectives, and agent behavior compared to the strategic classification setting. For example, in the price discrimination setting, the planner objective is revenue maximization, and both the outcome and the observed characteristics are endogenous to the treatment received.

We next describe formally the algorithm that solves for the optimal personalized policy over time. We assume at each step in time, a different batch of n agents arrive, so agents optimize a static decision problem. At each step, a noisy estimate of the gradient of the objective function is estimated by using small perturbations to randomize how the treatment depends on observed characteristics for each individual, and observing how the objective function is affected by the perturbations. In the price discrimination example, some customers would receive a price that is more dependent on their browsing history, and others would receive a price that is slightly less dependent on their browsing history. A key insight for experiment design is that for this setting, randomizing prices only, rather than the dependence of prices on observed data, is not sufficient to learn the optimal policy. The relationship between these perturbations and the revenue received from each customer leads to a consistent estimate of the gradient of the objective. Over time, the planner uses these gradient estimates to take steps towards the optimal personalized policy. When the objective is strongly concave, the algorithm is guaranteed to converge to the global maximum. We prove that the regret of the proposed algorithm decays at a linear rate. Furthermore, under monotonicity assumptions on the treatment allocation function and the agents' strategic reporting function, we prove that any fixed points of the repeated risk minimization approach, analyzed in Frankel and Kartik (2020) and Perdomo et al. (2020) in the context of the strategic classification problem, are sub-optimal.

After evaluating the theoretical properties of our proposed approach, we return to the two main examples of price discrimination and strategic classification. We show that in simulations, the iterative experimental approach converges to the optimal prediction function and the optimal price discrimination function, and has negligible average regret. In contrast, the risk-minimization based approaches have significantly lower revenue and higher MSE. In the final section of the paper we report the results of a small MTurk experiment. In the experiment, we predict a respondent's self-reported income from their self-reported age, education, and car ownership. We introduce incentives, through a variable bonus payment disclosed on the second page of the survey, for individuals to manipulate their self-reported education and car ownership in order to receive a higher predicted income. We find that a risk-minimization approach leads to higher out of sample mean-squared error compared to a prediction function derived from our proposed iterative algorithm, which downweights characteristics susceptible to manipulation and upweights characteristics that are not as susceptible to manipulation.

Literature Review The proposed algorithm applies to the problem of strategic classification analyzed theoretically by both computer scientists and economic theorists (Perdomo et al., 2020; Frankel and Kartik, 2020; Ball, 2020), as well as to price discrimination (Varian, 1989) and heterogeneous taxation based on manipulable covariates (Roberts, 1984). A contribution of this paper is introducing a general methodology for policy targeting, so the solution applies to a variety of treatment allocation problems studied in the theory literature, with different planner objectives and models of strategic behavior.

There is existing work in the computer science literature on estimation methods for the strategic classification problem: Hardt et al. (2016) provides a near-optimal classifier based on an assumed cost of gaming and Dong et al. (2018) provides an estimation method based on zero-th order optimization in an online setting when a single agent arrives at each step. The most closely related paper is Björkegren et al. (2020), who introduce and test empirically an estimation method that solves for the optimal prediction rule with strategic agents. Björkegren et al. (2020) use a structural model for the estimation, which allows for both estimating the optimal policy and quantifying strategic behavior under various counterfactual policies, but requires making parametric assumptions on the agents' costs and benefits of manipulating their behavior. In contrast, the method in this paper is designed for settings where estimating the optimal personalized policy robustly is more important than counterfactual analysis. As a result, our method converges to the optimal personalized policy without parametric assumptions on the agents' strategic behavior, but is not designed for counterfactual analysis away from the optimum.

The paper is also related to previous work on causal inference without strategic agents. The literature on empirical welfare maximization (EWM) studies the problem of estimating the optimal treatment assignment function based on covariates, in both experimental and observational settings, see Manski (2004), Kitagawa and Tetenov (2018), Kallus and Zhou (2020), Athey and Wager (2020), among many others. In this literature, the distribution of pre-treatment covariates are usually assumed exogenous to how treatments are allocated. This paper provides an estimation method for the treatment allocation problem with low regret when the distribution of pre-treatment covariates is endogenous.

Among econometricians, there has been recent interest in applying adaptive experiments to select treatments, usually in settings with discrete-valued treatments, see Kasy and Sautmann (2019), Hadad et al. (2019), and the large theoretical literature on multi-armed bandits. This paper introduces a type of adaptive experiment for allocating continuous treatments based on observed covariates.

The proposed algorithm is related to the literature on derivative-free optimization, where a convex function is optimized by observing sequential function evaluations, see Spall (2005) for a basic overview, and Duchi et al. (2015) for analysis of a method based on multiple function evaluations. We use a similar perturbation technique as Wager and Xu (2019), who use an experimental approach to optimize a fixed price when there are equilibrium effects in a one-sided market.

More broadly, this work is related to the sufficient statistics approach of Chetty (2009), who shows that certain welfare analyses can be performed with estimates of a few key derivatives, and without full specification of a structural model. It is also related to the literature on robust mechanism design, where optimal mechanisms are derived when there is some uncertainty over the form of agent behavior, see Bergemann and Morris (2013).

2 Model and Experiment Design

The framework considered is as follows. Individuals have three sets of characteristics:

- Pre-treatment characteristics $\boldsymbol{X}_i \in \mathbb{R}^{d_x}$, which are observed before the treatment is allocated.
- Outcomes $Y_i \in \mathbb{R}^{d_y}$, which are observed after the treatment is allocated.
- Unobserved i.i.d. type $\boldsymbol{\theta}_i \in \Theta \sim G$, where G is unknown to the planner.

The treatment $W_i = w(\boldsymbol{X}_i; \boldsymbol{\beta}_i)$ is the outcome of a known continuous function w, parameterized by $\boldsymbol{\beta}_i \in \mathbb{R}^k$, that depends on observed pre-treatment characteristics.

$$w(\boldsymbol{x}; \boldsymbol{\beta}_i) : \mathbb{R}^{d_x} \times \mathbb{R}^k \to \mathbb{R}$$

$\boldsymbol{\beta}_i$ may be stochastic, which allows for the planner to run experiments that vary how the treatment is allocated to different individuals. Each individual's outcome depends on their treatment w_i and their type $\boldsymbol{\theta}_i$:

$$Y_i = y(W_i, \boldsymbol{\theta}_i)$$

In some settings the continuous function y is known, while in other settings it must be estimated. This function defines the potential outcomes for each individual, see Imbens and Rubin (2015), as $Y_i(W_i) = y(W_i, \boldsymbol{\theta}_i)$. The key difference from a standard causal inference setting, however, is that individuals strategically report $\boldsymbol{X}_i$. $\boldsymbol{X}_i$ is the result of the maximization of a utility function U, which is unknown to the planner:

$$\boldsymbol{X}_i = \boldsymbol{x}(\boldsymbol{\beta}_i, \boldsymbol{\theta}_i) = \arg\max_{\boldsymbol{x}} U(w_i(\boldsymbol{x}; \boldsymbol{\beta}_i), \boldsymbol{x}, y_i, \boldsymbol{\theta}_i)$$

Since U is unknown to the planner, then the function $\boldsymbol{x}$ is unknown as well. Although we allow the planner can randomize $\boldsymbol{\beta}$ across individuals in the learning process, the goal of the planner is to choose the fixed $\boldsymbol{\beta}$ that optimizes an objective. The objective is constrained to take the form of the expected value of a known function $\pi(W_i, Y_i)$ of each individual's treatment and outcome. Most typical objectives in this setting are of this form. The objective may be as simple as maximizing the expected outcome $E[Y_i(W_i)]$. In other cases, such as if the planner is a platform designer, the objective may be slightly more complex, such as revenue maximization in the price discrimination setting, or prediction error minimization in the strategic classification setting. The objective function $\Pi(\boldsymbol{\beta})$ is defined as follows, where $W_i = w(\boldsymbol{x}(\boldsymbol{\beta}, \boldsymbol{\theta}_i); \boldsymbol{\beta})$:

$$\Pi(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\theta}_i \sim G}\Big[\pi(W_i, y(W_i, \boldsymbol{\theta}_i))\Big] \tag{1}$$

The effect of $\boldsymbol{\beta}$ on the objective function is complex. First, there is the direct effect on the treatment allocations W_i and the outcomes Y_i through the treatment allocation. There is also an indirect effect, however, since the distribution of observed characteristics X_i depends on $\boldsymbol{\beta}$. With a single sample of data $(W_i, Y_i, X_i)_{i=1,\ldots,n}$, the planner can evaluate the empirical version of the true objective function. We define this for a sample of n individuals with treatments $W_i = w(X_i; \boldsymbol{\beta})$ as:

$$\Pi_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \pi(W_i, Y_i)$$

with $\Pi_n(\boldsymbol{\beta}) \to_p \Pi(\boldsymbol{\beta})$ by the Law of Large Numbers. The planner would like to set $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, which are the parameters for the treatment assignment function w that optimize the planner's objective:

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}} \Pi(\boldsymbol{\beta})$$

In the next subsection, we will specify assumptions such that $\boldsymbol{\beta}^*$ is unique and introduce our algorithm for estimating $\boldsymbol{\beta}^*$. First, we show how this framework applies to two concrete examples: the problem of strategic classification and the problem of optimal price discrimination based on observed characteristics.

Example 1. Strategic Classification

The first example is a social media operator who would like to predict content quality (Y_i) from click-through rates, which are the observed X_i. A higher predicted content quality rewards sellers through better placement in a users feed, for example. As a result, sellers may purchase fake clicks. The goal is to come up with an optimal prediction function when sellers manipulate their engagement metrics to receive a better quality prediction. We assume that the planner uses a linear prediction function of Y_i, so $w(X_i; \boldsymbol{\beta}) = \beta_0 + \beta_1 X_i$. A content producer's unobserved type is $\boldsymbol{\theta}_i = [Z_i, \gamma_i]$, where γ_i is their manipulation ability and Z_i is the engagement metric of the post in the absence of manipulation. In this case, as is common in the strategic manipulation literature, we assume that the treatment W_i does not affect the true quality Y_i, which we can assume is observed sometime after the prediction about the content's quality is made.

The planner's objective is to minimize the squared error from predicting Y_i using X_i.

$$\boldsymbol{\beta}^* = \arg\max_{\beta} \mathbb{E}_{\boldsymbol{\theta}_i \sim G} \left[(Y_i - \beta_0 - x(\boldsymbol{\beta}, \boldsymbol{\theta}_i)\beta_1)^2 \right]$$

If X_i is assumed to be exogenous, then a sub-optimal $\boldsymbol{\beta}$ will be chosen, as shown in Frankel and Kartik (2020) under a parametric assumption on $X_i(\boldsymbol{\beta})$. The method described in this paper can be applied to find the optimal prediction function $\boldsymbol{\beta}^*$ in this setting without making any parametric assumptions on $X_i(\boldsymbol{\beta})$.

Example 2. Price Discrimination

In this example, the planner is a company that is selling insurance at a fixed rate per dollar of coverage. The company observes customer search history metric $X_i \in \mathbb{R}$, which is correlated with the customer's value for the insurance policy in the absence of manipulation. The unobserved type $\theta_i = [V_i, \gamma_i, Z_i]$ determines the customer's valuation for insurance (V_i), their search behavior in the absence of manipulation (Z_i), and their manipulation ability (γ_i). The treatment W_i

is the price offered to customer i. Given a price W_i, an individual demands a certain amount of coverage $y(W_i, \boldsymbol{\theta}_i)$. The planner chooses a linear pricing policy $W_i = p_0 + p_1 X_i$, so that $\boldsymbol{\beta} = [p_0, p_1]$. Every customer is charged a fixed price p_0 plus a variable price $p_1 X_i$ that depends on their search history. The planner's goal is to set $\boldsymbol{\beta}$ to maximize their expected revenue across all customer types, without observing $\boldsymbol{\theta}_i$ or knowing the functional form of x or y.

$$\boldsymbol{\beta}^* = [p_0^*, p_1^*] = \arg\max \mathbb{E}_{\boldsymbol{\theta}_i \sim G}[(p_0 + p_1 x(\boldsymbol{\beta}, \theta_i)) \cdot y(W_i, \boldsymbol{\theta}_i)]$$

2.1 Optimizing Personalized Policy

We have now introduced a model that describes the environment in which the planner would like to determine the optimal personalized policy. Before exploring methods to optimize in this environment, we first make some basic assumptions on the data generating process, which ensures that the optimal parameterization of the treatment allocation function, defined as β^*, is unique.

Assumption 1. The functions $w(\boldsymbol{X}_i; \beta)$, $x(\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$, $y(W_i, \boldsymbol{X}_i)$ and $\pi(W_i, Y_i)$ are continuously differentiable. $\boldsymbol{\theta}_i$ is a continuous random variable.

Lemma 1. Given Assumption 1, $\Pi(\boldsymbol{\beta})$ is continuously differentiable.

Assumption 1 implies that $\Pi(\boldsymbol{\beta})$ is a continuously differentiable function so that the gradient exists for all $\boldsymbol{\beta} \in \mathbb{R}^k$. The gradient is defined as follows:

$$\nabla\Pi(\boldsymbol{\beta}) = \mathbb{E}\left[\left(\frac{\partial \pi}{\partial w} + \frac{\partial \pi}{\partial y}\frac{\partial y}{\partial w}\right)\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\beta}}\frac{\partial w}{\partial \boldsymbol{x}} + \frac{\partial w}{\partial \boldsymbol{\beta}}\right)\right]$$

To ensure that $\Pi(\boldsymbol{\beta})$ has a unique maximizer, we make a stronger assumption:

Assumption 2. $\Pi(\boldsymbol{\beta})$ is $\sigma-$strongly concave.

Under Assumption 2, then $\boldsymbol{\beta}^*$ is unique, so it is the unique solution to $\nabla\Pi(\boldsymbol{\beta}^*) = 0$. We briefly review two standard approaches for finding the solution to this equation and indicate why they are insufficient in this case.

Mechanism Design Approach Both π and w are specified by the planner, so their partial derivatives are known. If appropriate assumptions are made on the functional forms of the functions y and $\boldsymbol{x}$, as well as the distribution of $\boldsymbol{\theta}_i$, then it is possible to calculate the solution to $\nabla\Pi(\boldsymbol{\beta}) = 0$ directly through standard numerical optimization of the non-linear function $\Pi(\boldsymbol{\beta})$. In the economics literature, it is common to make assumptions on strategic behavior so that $\boldsymbol{\beta}^*$ has an analytical solution, see Frankel and Kartik (2020). While this is useful for gaining economic intuition about how incentives to manipulate data affect the optimal treatment allocation, if any of the assumptions about strategic behavior or the distribution of unobserved types are incorrect, then in practice the estimated $\boldsymbol{\beta}$ will be sub-optimal. Since it is challenging to make assumptions on strategic behavior and unobserved characteristics that hold across multiple environments, we focus our attention on approaches that do not require such restrictions.

Risk Minimization Approach An alternative which requires less assumptions and is more common in practice is an empirical risk minimization approach. The empirical gradient for a batch of n agents is:

$$\nabla\Pi_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\partial \pi_i}{\partial w} + \frac{\partial \pi_i}{\partial y}\frac{\partial Y_i}{\partial w}\right)\left(\frac{\partial \boldsymbol{X}_i}{\partial \boldsymbol{\beta}}\frac{\partial W_i}{\partial \boldsymbol{x}} + \frac{\partial W_i}{\partial \boldsymbol{\beta}}\right)\right]$$

In a standard risk minimization approach, the observed characteristics $\boldsymbol{X}_i$ are treated as exogenous. Estimating the derivative of the outcome y with respect to the treatment w is standard in the literature on causal inference and is not the focus of this paper. As a result, we assume the planner has a good estimate of the individual treatment effects $\frac{\partial Y_i}{\partial w}$, for example through an experiment that randomizes W_i. Then, the empirical risk minimization approach would assume $\frac{\partial \boldsymbol{X}_i}{\partial \boldsymbol{\beta}} = 0$, and find the solution, which we will call $\tilde{\boldsymbol{\beta}}$, to the following equation:

$$0 = g(\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\partial \pi_i}{\partial w} + \frac{\partial \pi_i}{\partial y}\frac{\partial Y_i}{\partial w}\right)\left(\frac{\partial W_i}{\partial \boldsymbol{\beta}}\right)\right]$$

As long as the treatment effect of W_i on the outcomes Y_i does not depend on $\boldsymbol{\beta}$, then it is possible to find $\tilde{\boldsymbol{\beta}}$ using a single sample of data. The derivatives of π and w are known and depend on observed data, and we assume we have a good estimate of the derivatives of y. However, unless $\mathbb{E}\left[\frac{\partial x(\boldsymbol{\beta}^*, \theta_i)}{\partial \boldsymbol{\beta}}\frac{\partial W_i}{\partial \boldsymbol{x}}\right] = 0$, which is not in general true, then $\boldsymbol{\beta}^* \neq \tilde{\boldsymbol{\beta}}$, so strategic reporting of characteristics affects the optimal policy function, and a new approach is needed to estimate $\boldsymbol{\beta}$.

Iterative Experiment Given that the typical methods that might solve for $\boldsymbol{\beta}^*$ using a single sample of data do not meet the requirements of our setting, we instead turn to an alternative approach that optimizes $\boldsymbol{\beta}$ over time. We assume that at each time $t = 1, \ldots, T$, a different batch of n agents arrive and are treated by the planner. We do not observe agents repeatedly over time, so their decision problem is static. Then, we can approximate $\boldsymbol{\beta}^*$ by starting from an initial naive estimate of $\hat{\boldsymbol{\beta}}$, perturbing $\boldsymbol{\beta}$ in a zero-mean way and observing how the individual objective values $\pi_i = \pi(W_i, Y_i)$ correlate with the perturbations of $\boldsymbol{\beta}$, and take gradient steps towards the optimal policy. This is inspired by the approach in Wager and Xu (2019) that finds the optimal fixed price in a one-sided market with equilibrium effects. The experiment is described formally in Algorithm 1.

Algorithm 1: Online Experiment for Optimizing Personalized Policy

Input: Initial estimate $\hat{\boldsymbol{\beta}}^0$, sample size n, step size η, perturbation h, and steps T

Output: Updated estimate $\hat{\boldsymbol{\beta}}^T$

$t = 1$; $K = \dim(\hat{\boldsymbol{\beta}}^0)$;

while $t \leq T$ **do**

  New batch of n agents arrive, agent i has unobserved type $\boldsymbol{\theta}_i^t \in \Theta$;

  for $i \in \{1, \ldots, n\}$ **do**

    Sample $\boldsymbol{\epsilon}_i$ randomly from $\{-1, 1\}^K$;

    Announce $\boldsymbol{\beta}_i = \hat{\boldsymbol{\beta}}^{t-1} + h\boldsymbol{\epsilon}_i$;

    Agent reports $\boldsymbol{X}_i^t = x(\boldsymbol{\beta}_i, \boldsymbol{\theta}_i^t)$;

    Treat agent with $W_i^t = w(X_i^t; \boldsymbol{\beta}_i)$;

    Agent reports outcome $Y_i^t = y(W_i^t, \boldsymbol{\theta}_i^t)$;

    Calculate objective value $\pi_i^t = \pi(W_i^t, Y_i^t)$;

  end

  $\boldsymbol{Q}_t = h\epsilon^t$ is the $n \times K$ matrix of perturbations;

  $\boldsymbol{\pi}^t$ is the n-length vector of objective values;

  Run OLS of $\boldsymbol{\pi}^t$ on $\boldsymbol{Q}_t$: $\hat{\boldsymbol{\Gamma}}^t = (\boldsymbol{Q}_t'\boldsymbol{Q}_t)^{-1}(\boldsymbol{Q}_t'\boldsymbol{\pi}^t)$;

  $\hat{\boldsymbol{\beta}}^t = \hat{\boldsymbol{\beta}}^{t-1} + 2\eta\frac{\hat{\boldsymbol{\Gamma}}^t}{t+1}$;

  $t \leftarrow t + 1$;

end

return $\hat{\boldsymbol{\beta}}^T$

This algorithm estimates $\hat{\boldsymbol{\beta}}^T$ in T steps without the planner specifying any functional form for how individuals report observed characteristics $x(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$, or how the outcomes depend on the treatment $y(W_i, \boldsymbol{\theta}_i)$, which are the two unknown processes in our environment. The proposed experiment has two notable characteristics. First, it is dynamic; without strict assumptions on how individuals manipulate characteristics, it is not possible to optimize Equation 1 using a single sample of data without an infeasibly large experiment size that varies $\boldsymbol{\beta}$ globally rather than locally. Second, in contrast to traditional experiment approaches, the design perturbs $\boldsymbol{\beta}_i$, rather than W_i. Randomizing the treatment alone does not allow the planner to observe how the dependence of the treatment allocation on observed characteristics affects the objective, which is required to optimize the objective in this setting.

3 Theoretical Results

In this section, we analyze the convergence of Algorithm 1 to the optimum and compare its performance to alternative approaches. The first result is that the estimate $\hat{\boldsymbol{\Gamma}}^t$ from each step of the perturbation experiment converges in probability to the true gradient $\nabla\Pi(\boldsymbol{\beta})$ as the sample size in each step grows large. This result relies on the perturbation size going to zero as $n \to \infty$ at a sufficiently slow rate.

Theorem 1. Fix some $\hat{\boldsymbol{\beta}}^t$. If the perturbation size $h = cn^{-\alpha}$ for $0 < \alpha < 0.5$, then $\hat{\boldsymbol{\Gamma}}^t$ from Algorithm 1 converges to the k-dimensional gradient of the objective:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\hat{\boldsymbol{\Gamma}}^t - \nabla\Pi(\hat{\boldsymbol{\beta}}^t)\right| > \epsilon\right) = 0$$

for any $\epsilon > 0$.

The proof, in Appendix A, uses the LLN and Slutsky's theorem to show that $\hat{\Gamma}^t$ converges to the centered difference approximation of the total derivative of the objective with respect to the policy parameter.

The average regret of a policy in place for T time periods is the average difference in the objective function between the realized policy path and the policy that maximizes the average objective value over the T time periods. In Example 2, the average regret of Algorithm 1 corresponds to the average revenue loss of a policy that learns through an iterative experiment compared to the revenue of a planner with full information who implements the optimal policy immediately from the first step.

Theorem 2. Under Assumption 2, also assume the norm of the gradient of Π is bounded by M, so $||\nabla\Pi(\boldsymbol{\beta})||_2 \leq M$, and the step size $\eta > \sigma^{-1}$. If a planner runs Algorithm 1 for T time periods, then we have that the regret decays at rate $O(1/t)$, so for any $\boldsymbol{\beta} \in \mathbb{R}^k$:

$$\lim_{n\to\infty} P\left[\frac{1}{T}\sum_{t=1}^{T} t(\Pi(\boldsymbol{\beta}) - \Pi(\hat{\boldsymbol{\beta}}^t)) \leq \frac{\eta M^2}{2}\right] = 1$$

A corollary of this is that the the procedure converges to $\boldsymbol{\beta}^*$ in probability at a linear rate as the sample size at each time step grows large.

Corollary 1. Under the conditions of Theorem 2,

$$\lim_{n \to \infty} P\left[||\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^T||_2^2 \leq \frac{2\eta M^2}{\sigma T} \right] = 1$$

The proof of Theorem 2 and Corollary 1 is in Appendix A. Theorem 1 shows that gradient estimate is consistent. The proof of Theorem 2 then applies results for convergence of gradient descent when a consistent, but not necessarily unbiased, gradient oracle is available.

One caveat is that the strong-concavity of the objective function in $\boldsymbol{\beta}$ is not something that can be verified by the planner in advance. If the objective function is not strongly concave, then there may be multiple local maxima in the objective. Algorithm 1 may then converge to a critical point of the objective, which may be a local maximizer or a saddle point rather than the global maximizer of the objective. Although the theoretical results are weaker, in practice gradient descent approaches have been used successfully in the optimization of non-convex objective functions, for example in the training of neural networks.

The combination of these two results indicates that the suggested dynamic experiment successfully recovers $\boldsymbol{\beta}^*$. In contrast, risk minimization based approaches do not converge to $\boldsymbol{\beta}^*$.

3.1 Comparison with Repeated Risk Minimization

We argued informally in Section 2.1 that an empirical risk minimization approach based on a single sample of data results in a suboptimal treatment allocation policy. Rather than performing risk minimization once, it is common to conduct repeated risk minimization, which is described formally in Algorithm 2. To isolate the issues with ignoring strategic reporting of $\boldsymbol{X}_i$, we assume when conducting repeated risk minimization that the planner has some way, for example through randomized assignment of W_i or an ex-ante correct model of $y(W_i, \boldsymbol{\theta}_i)$, of estimating the Individual Treatment Effect (ITE) $\frac{\partial Y_i}{\partial w}$. Repeated risk minimization can be modeled as an adversarial game, where the agents choose a distribution of

Algorithm 2: Repeated Risk Minimization

Input: Initial estimate $\hat{\boldsymbol{\beta}}^0$, sample size n, steps T, procedure to estimate $\frac{\partial Y_i}{\partial w}$

Output: Updated estimate $\tilde{\boldsymbol{\beta}}^T$

while $t \leq T$ **do**

New batch of n agents arrive, agent i has unobserved type $\boldsymbol{\theta}_i^t \in \Theta$;

Announce $\tilde{\boldsymbol{\beta}}^{t-1}$;

for $i \in \{1, \ldots, n\}$ **do**

Agent reports $\boldsymbol{X}_i^t = x(\tilde{\boldsymbol{\beta}}^{t-1}, \boldsymbol{\theta}_i^t)$;

Treat agent with $W_i^t = w(X_i^t; \tilde{\boldsymbol{\beta}}^{t-1})$;

Agent reports outcome $Y_i^t = y(W_i^t, \boldsymbol{\theta}_i^t)$;

Calculate objective value $\pi_i^t = \pi(W_i^t, Y_i^t)$;

end

Define $W_i(\tilde{\boldsymbol{\beta}}^t) = w(X_i^t; \tilde{\boldsymbol{\beta}}^t)$, $Y_i(\tilde{\boldsymbol{\beta}}^t) = y(W_i(\tilde{\boldsymbol{\beta}}^t), \boldsymbol{\theta}_i)$ and $\pi_i(\tilde{\boldsymbol{\beta}}^t) = \pi(W_i(\tilde{\boldsymbol{\beta}}^t), Y_i(\tilde{\boldsymbol{\beta}}^t))$;

Given estimate of $\frac{\partial Y_i(\hat{\boldsymbol{\beta}}^t)}{\partial w}$;

$\tilde{\boldsymbol{\beta}}^t$ solves:

$$0 = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\partial \pi_i(\tilde{\boldsymbol{\beta}}^t)}{\partial w} + \frac{\partial \pi_i(\tilde{\boldsymbol{\beta}}^t)}{\partial y} \frac{\partial Y_i(\tilde{\boldsymbol{\beta}}^t)}{\partial w} \right] \frac{\partial W_i(\tilde{\boldsymbol{\beta}}^t)}{\partial \boldsymbol{\beta}}$$

end

return $\tilde{\boldsymbol{\beta}}^T$

characteristics based on the planner's policy, and then the planner chooses their policy based on the latest set of observed characteristics.

There are a variety of existing results on repeated risk minimization from the strategic classification literature. Perdomo et al. (2020) prove that as long as the distribution of observed characteristics does not respond too much to changes in the policy, then risk minimization converges to a fixed point. Furthermore, the difference in the optimality of this fixed point compared to the global optimum can be bounded by this sensitivity of the distribution of the observed characteristics to the policy. In a more specific setting with functional form restrictions on agents' response functions, Frankel and Kartik (2020) prove that any fixed point is strictly suboptimal to the globally optimal $\boldsymbol{\beta}$. It is worth examining how these results extend to our more general setting of personalized policy. Without further restrictions on the data generating process, there may not be a unique fixed point or a fixed point at all. If there is a fixed point of Algorithm 2, it can be defined as follows. Let $W_i(\boldsymbol{\beta}^{FP}) = w(x(\boldsymbol{\beta}^{FP}, \boldsymbol{\theta}_i); \boldsymbol{\beta}^{FP})$, $Y_i(\boldsymbol{\beta}^{FP}) = y(W_i(\boldsymbol{\beta}^{FP}), \boldsymbol{\theta}_i)$ and

$\pi_i(\boldsymbol{\beta}^{FP}) = \pi(W_i(\boldsymbol{\beta}^{FP}), Y_i(\boldsymbol{\beta}^{FP}))$. Then $\boldsymbol{\beta}^{FP}$ is any solution to the following equation:

$$\lim_{n\to\infty} 0 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial \pi_i(\boldsymbol{\beta}^{FP})}{\partial w} + \frac{\partial \pi_i(\boldsymbol{\beta}^{FP})}{\partial y}\frac{\partial y_i(\boldsymbol{\beta}^{FP})}{\partial w}\right]\frac{\partial w_i(\boldsymbol{\beta}^{FP})}{\partial \beta} = 0 \qquad (2)$$

With some additional assumptions on the treatment allocation function w and the strategic reporting function $\boldsymbol{x}$, we can show that any fixed point of a repeated risk minimization procedure is suboptimal in a personalized policy setting.

Theorem 3. Under the conditions of Theorem 2, also assume that $\boldsymbol{\beta}^* \neq 0$, that $w(\boldsymbol{X}_i; \beta)$ is strictly monotonic in $\boldsymbol{X}_i$ for $\boldsymbol{\beta} \neq 0$, and for all $\boldsymbol{\theta}_i$ we have that $\boldsymbol{x}(\boldsymbol{\beta}, \boldsymbol{\theta}_i)$ is strictly monotonic in $\boldsymbol{\beta}$ for some region of the distribution of $\boldsymbol{\theta}_i$ that occurs with non-zero probability. Then, we have that

$$\Pi(\boldsymbol{\beta}^{FP}) < \Pi(\boldsymbol{\beta}^*)$$

The proof, in Appendix A, is a simple proof by contradiction showing that $\boldsymbol{\beta}^*$, which sets the limit of $\nabla_n \Pi(\boldsymbol{\beta})$ to zero and the $\boldsymbol{\beta}^{FP}$, which sets the limit of Equation 2 to zero, cannot be equal. By the strong-convexity of Π, this implies that $\Pi(\boldsymbol{\beta}^{FP}) < \Pi(\boldsymbol{\beta}^*)$.

The monotonicity assumptions are natural in many settings with personalized policy. For example, in the price discrimination setting, a linear pricing rule is monotonic in each of the observed characteristics. In addition, most economic models of manipulation, for example see Frankel and Kartik (2020) and Roberts (1984), imply that an increase in incentives to manipulate a characteristic will result in a monotonic change in the report of that characteristic for those with non-zero manipulation ability.

The interpretation of Theorem 3 is straightforward. In an adversarial game between the planner and strategic agents, the planner does not learn about strategic behavior. Instead, the planner simply reacts to the strategic behavior that occurs, and in some cases this can converge to a fixed point. In contrast, in the iterative experiment of Algorithm 1, the planner perturbs the dependence of the treatment on the observed characteristics so that the relevant aspects of strategic behavior that impact the objective function are learned over time. We next illustrate the theoretical results of this section in simulations based on Example 1 and Example 2.

4 Simulations

The first simulation is of the strategic classification setting, which has received a significant amount of attention in the literature. In the strategic classification literature, the target Y_i is exogenous to the treatment, which is the prediction W_i. The approach we propose in this paper applies to more general settings where the post-treatment outcome depends on the treatment. This includes a wider variety of both private and public sector policies. As a result, the second simulation is of the price discrimination setting, where the amount purchased Y_i depends on the treatment, which is the individual-specific price W_i.

In both simulations, we examine the convergence and the regret of the estimated policy over T periods with n agents at each step, for four different approaches:

- Full information benchmark: Calculate $\boldsymbol{\beta}^*$ based on full knowledge of the data generating process.
- Iterative learning following Algorithm 1
- Naive risk minimization: Assume X_i is exogenous and estimate $\hat{\boldsymbol{\beta}}$ based on sample of data collected where individuals do not have an incentive to manipulate their reports of X_i.
- Repeated risk minimization following Algorithm 2

4.1 Strategic Classification Simulation

We impose the following assumptions on the framework in Example 1, where the planner's goal is to find the MSE-minimizing prediction function. The assumptions imposed lead to a similar data generating process as the one analyzed in Frankel and Kartik (2020). First we assume that the unobserved type $\boldsymbol{\theta}_i = [Z_i, \gamma_i, R_i]$ has the following distribution:

$$Z_i \sim \text{Normal}(0,1), \qquad \gamma_i \sim \text{Uniform}(0,1.5), \qquad R_i \sim \text{Normal}(0,1)$$

The reported characteristics are the result of the maximization of a linear - quadratic utility function that is increasing in the quality prediction $W_i = \beta_0 + X_i\beta_1$ and decreasing in the amount of manipulation of the click-through-rate:

$$X_i = \arg\max_{\tilde{x}_i} \beta_0 + \tilde{x}_i\beta_1 - \frac{(\tilde{x}_i - Z_i)^2}{2\gamma_i}$$

Solving this leads to a reporting function of $x(\beta, \theta_i) = Z_i + \gamma_i \beta_1$. The outcome, which is the realized content quality, is $y(\boldsymbol{\theta}_i) = Z_i + R_i$.

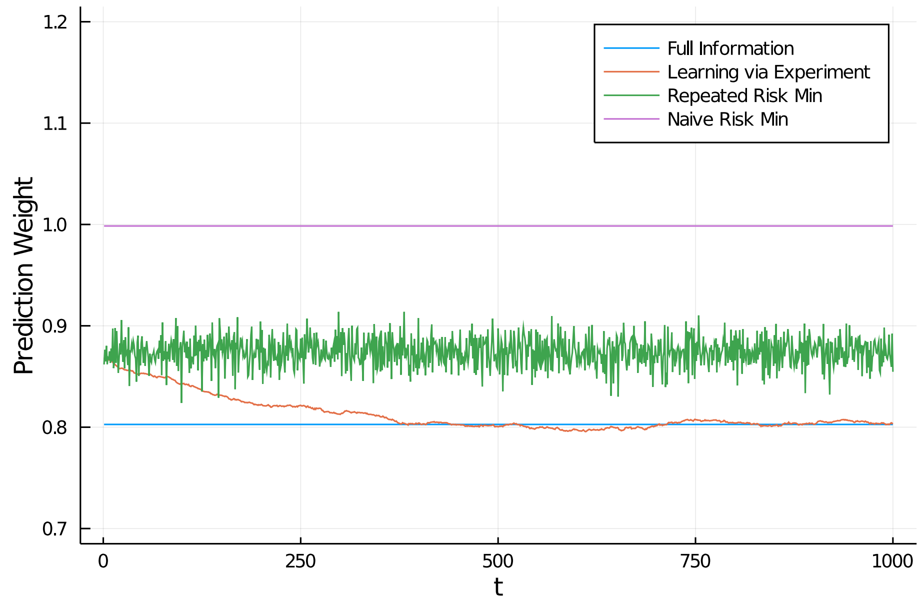


Figure 1: Convergence of Iterative Learning to Optimal Prediction Function

Figure 1 plots the estimated β_1 for 1000 periods in which a batch of 1000 agents arrive in each period. β_1 indicates how much the prediction of the outcome Y_i depends on the reported X_i. The purple line is the estimated β_1 from OLS on a sample of data where individuals are not rewarded based on $\beta_1 X_i$. When agents are strategic, this coefficient is too high. It places too much weight on X_i, which is susceptible to manipulation. The green line is the estimated β_1 when the planner updates β_1 at each step t using OLS, and agents best-respond in the next period. The prediction functions reached by this adversarial game are still sub-optimal, with a higher realized mean-squared error than the full information solution. The blue line is the full information solution, where the planner has access to the manipulation ability of each agent and can solve for β_1^* in one step using non-linear optimization. The orange line is the experimental estimate of β_1^*, which uses mean-zero perturbation of β and observations of X_i and Y_i to take gradient steps towards the optimum. The experimental approach converges to the MSE-optimal full knowledge solution without relying on any prior assumptions on the structure of manipulation. The risk minimization approaches induce a distribution of characteristics that is tainted by manipulation in a way that makes prediction more difficult. The iterative learning approach induces a distribution

Method	**Average MSE**
Full Information	1.1176
Iterative Learning	1.1180
Repeated Risk Minimization	1.1261
Naive Risk Minimization	1.7448

Table 1: Average MSE of Each Approach

that is more amenable to prediction, leading to a lower average MSE over the 1000 periods compared to the risk minimization approaches, despite taking some time to converge to the optimal value, see Table 1.

4.2 Price Discrimination Simulation

In Example 2, the planner would like to set the optimal linear pricing rule $W_i = w(X_i; \boldsymbol{\beta})$ to maximize revenue when agents may misreport X_i to receive a better price. To generate data in this setting, we impose assumptions on the distribution of the agents' type $\boldsymbol{\theta}_i = [V_i, Z_i, \gamma_i]$ and the form of the agents' utility function, which determines their demand $Y_i = y(W_i, \boldsymbol{\theta}_i)$ and their strategic responses $X_i = x(\boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$.

$$Z_i \sim \text{Uniform}(10, 20), \qquad V_i \sim \text{Normal}(5 + Z_i, 2), \qquad \gamma_i \sim \text{Uniform}(0, 3)$$

We assume the following utility function, which is a function of an individual's reported characteristic X_i, the amount purchased Y_i, the price W_i, and the type $\boldsymbol{\theta}_i$:

$$u(X_i, Y_i, W_i, \boldsymbol{\theta}_i) = Y_i(V_i - W_i) - \frac{1}{2}Y_i^2 - \frac{(X_i - Z_i)^2}{2\gamma_i}$$

This leads to a demand that is linear in price:

$$y(W_i, \boldsymbol{\theta}_i) = V_i - W_i$$

Given the demand function, the optimal report of X_i for the individual is:

$$x(\boldsymbol{\beta}, \boldsymbol{\theta}_i) = \frac{Z_i - \gamma_i p_1 [V_i - p_0]}{1 - p_1^2 \gamma_i}$$

An individual will shade their report of X_i by an amount that depends on the degree of price discrimination p_1 and their manipulation ability γ_i. Despite the more complex setting, in Figure 2, the simulation shows a similar pattern

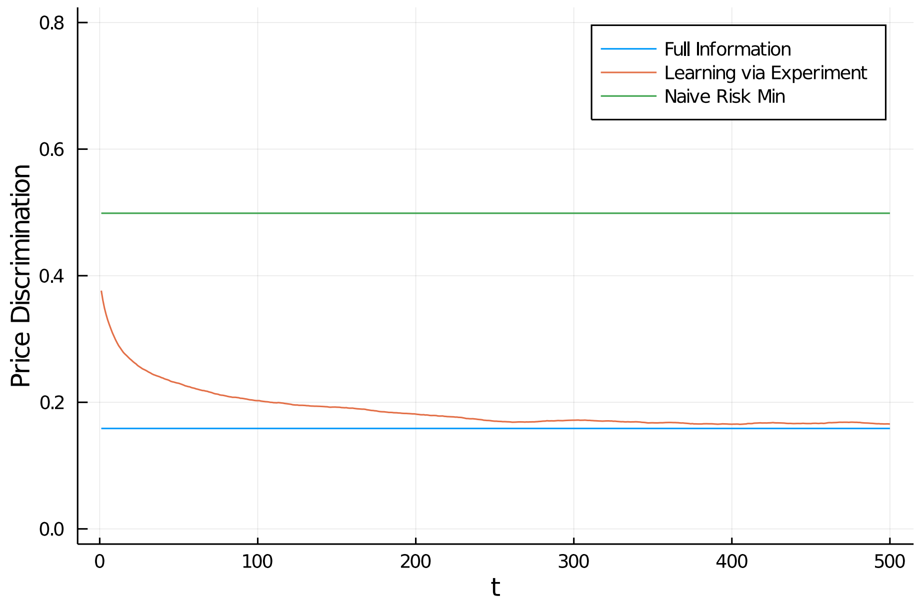


Figure 2: Convergence of Iterative Learning to Optimal Price Discrimination

Method	**Regret**
Full Information	0.00
Iterative Learning	-0.25
Repeated Risk Minimization	-24.45
Naive Risk Minimization	-48.21

Table 2: Average Revenue of Each Approach Compared to Full Information Optimum

to the strategic classification simulation. The chart plots p_1, which indicates how much the price shown to a customer depends on X_i, over 500 steps. In this case the repeated risk minimization approach does not converge to a fixed point, and instead flips back and forth between the naive risk minimization solution and a solution with close to zero price discrimination. As a result, it is omitted from the chart. The naive risk minimization approach price discriminates too much, not anticipating that the manipulation that occurs in response to this price discrimination leads to sub-optimal revenue. In contrast, Algorithm 1 converges to the revenue-optimal price discrimination function.

The difference in average revenues over the course of the 500 periods of the simulation is significant. In Table 2, we show that the risk minimization approaches have a cost that is 100 times larger than the negligible difference in average revenues between the iterative learning and full information approach.

5 Empirical Results

We also evaluate the impact of the methodology in a small experiment inspired by the setting in Example 1. We set up a Qualtrics survey on Amazon Mechanical Turk, where on Page 1, the respondent is asked to self-report their income and their age range. From Page 2, it is not possible to return to Page 1. On Page 2, the respondent self-reports their education and whether or not they own a car.

For the baseline survey, there is no incentive to misreport, and respondents are paid a small fee to complete the survey. For the experimental steps, respondents are paid the same fixed fee plus a variable bonus equal to the respondents' rescaled predicted income. The bonus is not mentioned on page 1, so that there are no incentives to misreport income or age. Respondents are made aware of the bonus' value and that it depends on their predicted income on Page 2, which may impact reports of education and car ownership. The bonus value dynamically updates as individuals choose their responses on Page 2. See Appendix B for an image of the survey Page 1, which has the appearance of a normal demographic survey, and of the incentives to manipulate introduced on survey Page 2. The prediction rule is linear, and the objective is to find the β that minimizes expected mean squared error of the prediction of the respondents' incomes, when the distributions of Educ and Car may depend on β:

$$\text{Income}_i = \beta_1 + \beta_2 \text{Age}_i + \beta_3 \text{Educ}(\beta)_i + \beta_4 \text{Car}(\beta)_i + \epsilon_i$$

We estimate two sets of coefficients. The first follows a risk minimization approach. $\hat{\boldsymbol{\beta}}^0$ is the OLS estimate on a sample of 100 individuals whose data is collected from the baseline survey without an incentive to manipulate. $\hat{\boldsymbol{\beta}}^1 = \hat{\boldsymbol{\beta}}^0 - \eta \hat{\boldsymbol{\Gamma}}$ is the estimated coefficients from a single step of the iterative learning algorithm. $\hat{\boldsymbol{\Gamma}}$ is estimated from zero-mean perturbation of $\hat{\boldsymbol{\beta}}^0$ in a sample of 400 individuals who receive a bonus $\hat{\boldsymbol{\beta}}^0 \boldsymbol{X}_i$. This means some individuals receive a prediction rule that is slightly more dependent on observed characteristics than the OLS rule, and some individuals receive a prediction rule that is slightly less dependent on observed characteristics than the OLS rule.

In Table 3, the first column reports the OLS coefficients and standard errors, estimated from the baseline sample of 100 coefficients. As expected, all three characteristics are useful in predicting income in the absence of incentives for individuals to misreport characteristics. The second column reports the estimated gradients derived from a single step of Algorithm 1. In the third column, we report

	$\hat{\boldsymbol{\beta}}^0$	$\hat{\nabla}\Pi(\boldsymbol{\beta})$	$\hat{\boldsymbol{\beta}}^1$
Intercept	-31.43 (26.2)		-30.96
Age	0.248 (0.262)	997	0.261
Education	3.694 (1.55)	43	3.979
Car	17.927 (8.05)	-27	9.002

Table 3: Experimental Update of β

the updated prediction function, where we used a characteristic-specific step size. The theoretical results in Frankel and Kartik (2020) and Ball (2020) indicate that under certain assumptions on agent behavior, the prediction-optimal coefficient on characteristics with a high variance in manipulation ability across individuals will have a lower weight compared to the OLS coefficient. Since we provide incentives to manipulate Car and Education but not Age, we might expect that the prediction rule would react to strategic behavior by increasing the coefficient on Age, and decreasing it on Education and Car. In the single gradient step conducted in the experiment, we do see the coefficient on Car decrease and Age increase, although we also see the coefficient on Education, which is a manipulable characteristic, increase.

We then calculate out-of-sample MSE in three different ways.

1. Risk Minimization: Predict Y_i using $W_i = \hat{\boldsymbol{\beta}}^0 \boldsymbol{X}_i$ for a sample of 400 individuals who report data without incentives to manipulate
2. Naive Risk Minimization: Predict Y_i using $W_i = \hat{\boldsymbol{\beta}}^0 \boldsymbol{X}_i$ for a sample of 400 individuals who receive a bonus $\hat{\boldsymbol{\beta}}^0 \boldsymbol{X}_i$
3. Iterative Learning: Predict Y_i using $W_i = \hat{\boldsymbol{\beta}}^1 \boldsymbol{X}_i$ for a sample of 400 individuals who receive a bonus $\hat{\boldsymbol{\beta}}^1 \boldsymbol{X}_i$

In Figure 3, we show that, as expected, the out of sample MSE is lowest under the risk minimization approach when individuals are not incentivized to misreport their characteristics. The MSE is highest under the naive risk minimization approach, when OLS, which incorrectly assumes reported characteristics are exogenous, is used to estimate the prediction function. The out of sample MSE takes an intermediate value under the iterative learning approach, when some of the impact of strategic responses is taken into account when formulating the prediction rule.

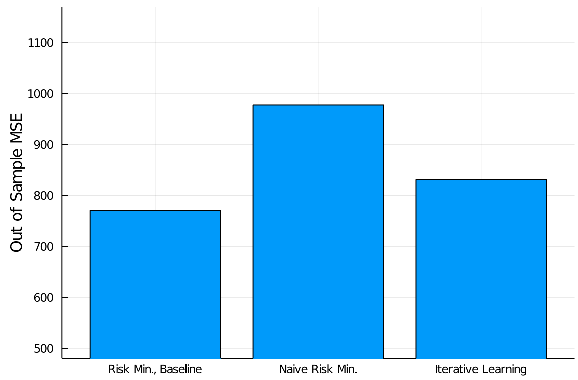


Figure 3: Out of Sample MSE ($n = 400$ for each)

6 Conclusion

When a planner treats individuals in a heterogeneous way based on some observed characteristics about that individual, incentives are introduced for individuals to manipulate their behavior to receive a better treatment. We have shown theoretically, in simulations, and in practice that this impacts how treatments should be optimally allocated based on observed individual level data. We propose an iterative method that converges to the optimal treatment assignment function, without making parametric assumptions on the structure of individuals' strategic behavior. The key to the success of this method is the dynamic approach, and randomizing how the treatment depends on observed characteristics rather than randomizing the treatment itself. There is a variety of potential future work involving the combination of economic models with advances in stochastic optimization to design policy that adjusts optimally without strict assumptions on the environment.

References

Athey, Susan and Stefan Wager, "Policy Learning with Observational Data," 2020.

Ball, Ian, "Scoring Strategic Agents," *Job Market Paper*, 2020.

Bergemann, Dirk and Stephen Morris, "Robust predictions in games with incomplete information," *Econometrica*, 2013, *81* (4), 1251–1308.

Björkegren, Daniel and Darrell Grissen, "Behavior revealed in mobile phone usage predicts credit repayment," 2019.

_, Joshua E Blumenstock, and Samsun Knight, "Manipulation-Proof Machine Learning," *arXiv preprint arXiv:2004.03865*, 2020.

Chetty, Raj, "Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods," *Annu. Rev. Econ.*, 2009, *1* (1), 451–488.

Dong, Jinshuo, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu, "Strategic classification from revealed preferences," in "Proceedings of the 2018 ACM Conference on Economics and Computation" 2018, pp. 55–70.

Duchi, John C, Michael I Jordan, Martin J Wainwright, and Andre Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, 2015, *61* (5), 2788–2806.

Frankel, Alex and Navin Kartik, "Improving Information from Manipulable Data," *arXiv preprint arXiv:1908.10330*, 2020.

Hadad, Vitor, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey, "Confidence intervals for policy evaluation in adaptive experiments," *arXiv preprint arXiv:1911.02768*, 2019.

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters, "Strategic classification," in "Proceedings of the 2016 ACM conference on innovations in theoretical computer science" 2016, pp. 111–122.

Imbens, Guido W and Donald B Rubin, *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press, 2015.

Jin, Yizhou and Shoshana Vasserman, "Buying Data from Consumers," 2019.

Kallus, Nathan and Angela Zhou, "Minimax-Optimal Policy Learning Under Unobserved Confounding," *Management Science*, 2020.

Kasy, Maximilian and Anja Sautmann, "Adaptive treatment assignment in experiments for policy choice," 2019.

Kitagawa, Toru and Aleksey Tetenov, "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 2018, *86* (2), 591–616.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the national academy of sciences*, 2019, *116* (10), 4156–4165.

Lucas, Robert E et al., "Econometric policy evaluation: A critique," in "Carnegie-Rochester conference series on public policy," Vol. 1 1976, pp. 19–46.

Manski, Charles F, "Statistical treatment rules for heterogeneous populations," *Econometrica*, 2004, *72* (4), 1221–1246.

Nie, Xinkun and Stefan Wager, "Quasi-oracle estimation of heterogeneous treatment effects," *arXiv preprint arXiv:1712.04912*, 2017.

Orabona, Francesco, Koby Crammer, and Nicolò Cesa-Bianchi, "A Generalized Online Mirror Descent with Applications to Classification and Regression," 2014.

Perdomo, Juan C, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt, "Performative prediction," *arXiv preprint arXiv:2002.06673*, 2020.

Roberts, Kevin, "The theoretical limits to redistribution," *The review of economic studies*, 1984, *51* (2), 177–195.

Spall, James C, *Introduction to stochastic search and optimization: estimation, simulation, and control*, Vol. 65, John Wiley & Sons, 2005.

Varian, Hal R, "Price discrimination," *Handbook of industrial organization*, 1989, *1*, 597–654.

Wager, Stefan and Kuang Xu, "Experimenting in equilibrium," *arXiv preprint arXiv:1903.02124*, 2019.

_ **and Susan Athey**, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 2018, *113* (523), 1228–1242.

A Proofs

A.1 Proof of Theorem 1

Proof. We can drop the t subscripts since we are fixing $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^t$. Let $\boldsymbol{Q} \in \{-h, h\}^{n \times k}$ be the $n \times k$ matrix of experimental perturbations, where $Q_{ik} = \epsilon_{ik}^t h$. Let $\boldsymbol{\pi}$ be the $n \times 1$ vector of objective values as a function of each individual's treatment and outcome, where $\pi_i = \pi(w_i, y_i)$.

Then we have that

$$\hat{\boldsymbol{\Gamma}} = \frac{\boldsymbol{Q}'\boldsymbol{\pi}}{\boldsymbol{Q}'\boldsymbol{Q}}$$

Since Q_{ik} is drawn i.i.d. for each i and each k, then $\mathbb{E}[Q_{ik}Q_{ij}] = 0$ unless $j = k$, in which case $\mathbb{E}[Q_{ik}^2] = h^2$. Since Q_{ik} is drawn randomly for each individual, by the Law of Large Numbers,

$$\frac{\boldsymbol{Q}'\boldsymbol{Q}}{h^2 n} \to_p \mathbb{E}\left[\boldsymbol{Q}'\boldsymbol{Q}/h^2\right] = I_k$$

,

The denominator $h^n = cn^{1-2\alpha}$. The condition $\alpha < 0.5$ ensures that we are not dividing by a term that goes to zero as $n \to \infty$. $\boldsymbol{Q}'\boldsymbol{\pi}$ is a $k \times 1$ vector, where

$$\frac{[\boldsymbol{Q}'\boldsymbol{\pi}]_j}{h^2} = \sum_{i=1}^{n} \frac{\pi_i Q_{ij}}{h^2} = \frac{\sum\limits_{i:Q_{ij}=h} \pi_i - \sum\limits_{i:Q_{ij}=-h} \pi_i}{h}$$

Let $\boldsymbol{e_j}$ be the length k basis vector with 1 at position j and 0 everywhere else.

Fixing h, we have that

$$\sum_{i:Q_{ij}=h} \pi_i / \sum_{i=1}^{n} \mathbb{1}(Q_{ij} = h) \to_p \Pi(\boldsymbol{\beta} + h\boldsymbol{e}_j)$$

by the Law of Large Numbers. Similarly, we have that:

$$\sum_{i:Q_{ij}=-h} \pi_i / \sum_{i=1}^{n} \mathbb{1}(Q_{ij} = -h) \to_p \Pi(\boldsymbol{\beta} - h\boldsymbol{e}_j)$$

We can then take the limit as $h \to 0$. We can apply the result that the centered difference approximation converges to the true partial derivative as $h \to 0$, which uses a Taylor expansion of $\Pi(\boldsymbol{\beta})$.

$$\lim_{h\to 0} \frac{\Pi(\boldsymbol{\beta} + h\boldsymbol{e}_j) - \Pi(\boldsymbol{\beta} - h\boldsymbol{e}_j)}{2h} = \frac{\partial \Pi(\boldsymbol{\beta})}{\partial \beta_j}$$

Since we have that $h = cn^{-\alpha}$ with $\alpha > 0$, then as $n \to \infty$, $h \to 0$. As a result, we can now apply the LLN and Slutsky's theorem to show that:

$$\frac{[\boldsymbol{Q}'\boldsymbol{\pi}]_j}{h^2 n} \to_p \lim_{h\to 0} \frac{\Pi(\boldsymbol{\beta} + h\boldsymbol{e}_j) - \Pi(\boldsymbol{\beta} - h\boldsymbol{e}_j)}{2h} = \frac{\partial \Pi(\boldsymbol{\beta})}{\partial \beta_j}$$

Now taking the expression for $\hat{\boldsymbol{\Gamma}}$ and dividing both the numerator and denominator by $h^2 n$, and applying Slutsky's theorem, we have that:

$$\hat{\boldsymbol{\Gamma}} = \left(\frac{\boldsymbol{Q}'\boldsymbol{Q}}{h^2 n}\right)^{-1} \left(\frac{\boldsymbol{Q}'\boldsymbol{\pi}}{h^2 n}\right) \to_p (\boldsymbol{I}_K)^{-1} \nabla \Pi(\hat{\boldsymbol{\beta}}^t) = \nabla \Pi(\hat{\boldsymbol{\beta}}^t)$$

□

A.2 Proof of Theorem 2

Proof. Follows the approach of the proof of Theorem 7 in Wager and Xu (2019). The first step is to use Lemma 1 from Orabona et al. (2014) to show that:

$$\sum_{t=1}^{T} t(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t)'\hat{\boldsymbol{\Gamma}}^t \leq \frac{1}{2\eta} \sum_{t=1}^{T} t||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t||_2^2 + \frac{\eta}{2} \sum_{t=1}^{T} ||\hat{\boldsymbol{\Gamma}}^t||_2^2$$

In order to use Lemma 1, we can define:

$$\boldsymbol{f}_t(\boldsymbol{\beta}) = \frac{1}{2\eta} \sum_{s=1}^{t} s||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^s||_2^2, \qquad \boldsymbol{\theta}_t = \sum_{s=1}^{t} s\hat{\Gamma}^s$$

We can also define the gradient of the Fenchel conjugate of $\boldsymbol{f}_t$:

$$\nabla \boldsymbol{f}_t^*(\boldsymbol{\theta}_t) = \arg\min_{\beta} f_t(\boldsymbol{\beta}) - \boldsymbol{\beta}\boldsymbol{\theta}_t \tag{3}$$

The next step is to show that $\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^{\boldsymbol{t}} + 2\eta\hat{\Gamma}^t/(t+1) = \nabla \boldsymbol{f}_t^*(\boldsymbol{\theta}_t)$. Setting the FOC of the term that is minimized in Equation 3 to zero:

$$\frac{1}{2\eta} \sum_{s=1}^{t} s(\hat{\boldsymbol{\beta}}^{t+1} - \hat{\boldsymbol{\beta}}^s) = \boldsymbol{\theta}_t$$

We can easily verify that the LHS of the equation is equal to the RHS of the equation when we have $\hat{\boldsymbol{\beta}}^{t+1} = \hat{\boldsymbol{\beta}}^{\boldsymbol{t}} + 2\eta\hat{\Gamma}^t/(t+1)$ so that $\hat{\boldsymbol{\beta}}^{t+1} - \hat{\boldsymbol{\beta}}^s =$

$\sum_{q=s}^{t} 2\eta\hat{\mathbf{\Gamma}}^q/(q+1)$:

$$\frac{1}{2\eta}\sum_{s=1}^{t} s(\hat{\boldsymbol{\beta}}^{t+1} - \hat{\boldsymbol{\beta}}^s) = 2\sum_{q=1}^{t}\sum_{b=1}^{q} b\hat{\mathbf{\Gamma}}^q/(q+1) = \sum_{s=1}^{t} s\hat{\mathbf{\Gamma}}^s = \boldsymbol{\theta}_t$$

The previous derivation has now shown that the gradient step in Algorithm 1 of this paper is in the form of Algorithm 1 (Online Mirror Descent) of Orabona et al. (2014), where we can map notation with $\boldsymbol{z}_t$ replaced by $t\hat{\mathbf{\Gamma}}^t$, $\boldsymbol{u}$ replaced by $\boldsymbol{\beta}$, and w_t replaced by $\hat{\boldsymbol{\beta}}^t$. We can now directly apply Lemma 1, where the third summand of Lemma 1 can be dropped since it is negative by Equation (4) of Orabona et al. (2014), to show that:

$$\sum_{t=1}^{T} t(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t)'\hat{\mathbf{\Gamma}}^t \leq \frac{1}{2\eta}\sum_{t=1}^{T} t||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t||_2^2 + \frac{\eta}{2}\sum_{t=1}^{T} ||\hat{\mathbf{\Gamma}}^t||_2^2$$

Then, we can replace the gradient estimate $\hat{\mathbf{\Gamma}}^t$ with its limit value $\nabla\Pi(\hat{\boldsymbol{\beta}}^t)$ and add an appropriate error term.

$$\sum_{t=1}^{T} t(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t)'\nabla\Pi(\hat{\boldsymbol{\beta}}^t) \leq \frac{1}{2\eta}\sum_{t=1}^{T} t||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t||_2^2 + \frac{\eta}{2}\sum_{t=1}^{T} ||\nabla\Pi(\hat{\boldsymbol{\beta}}^t)||_2^2$$
$$+ \frac{\eta}{2}\sum_{t=1}^{T}(||\hat{\mathbf{\Gamma}}^t||_2^2 - ||\nabla\Pi(\hat{\boldsymbol{\beta}}^t)||_2^2) + \sum_{t=1}^{T} t(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t)'(\nabla\Pi(\hat{\boldsymbol{\beta}}^t) - \hat{\mathbf{\Gamma}}^t)$$

From Lemma 1, we know that with probability approaching 1 as $n \to \infty$, we have that for any $\epsilon > 0$ that :

$$\sum_{t=1}^{T} t(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t)'\nabla\Pi(\hat{\boldsymbol{\beta}}^t) \leq \frac{1}{2\eta}\sum_{t=1}^{T} t||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t||_2^2 + \frac{\eta}{2}\sum_{t=1}^{T} ||\nabla\Pi(\hat{\boldsymbol{\beta}}^t)||_2^2 + \epsilon$$

Then, given that the gradient is bounded by M, we have that:

$$\sum_{t=1}^{T} t(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t)'\nabla\Pi(\hat{\boldsymbol{\beta}}^t) \leq \frac{1}{2\eta}\sum_{t=1}^{T} t||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t||_2^2 + \frac{\eta T M^2}{2}$$

Next, we use the σ-strong concavity of Π, which implies that:

$$\Pi(\boldsymbol{\beta}) \leq \Pi(\hat{\boldsymbol{\beta}}^t) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t)'\nabla\Pi(\hat{\boldsymbol{\beta}}^t) + \frac{\sigma}{2}||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t||_2^2$$

As well as the choice of $\sigma > \eta^{-1}$, to replace the LHS of the expression since we have that:

$$\sum_{t=1}^{T} t(\Pi(\boldsymbol{\beta}) - \Pi(\hat{\boldsymbol{\beta}}^t)) + \frac{1}{2\eta}||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t||_2^2 \leq \sum_{t=1}^{T} t(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^t)'\nabla\Pi(\hat{\boldsymbol{\beta}}^t)$$

This then gives the result:

$$\frac{1}{T}\sum_{t=1}^{T} t(\Pi(\boldsymbol{\beta}) - \Pi(\hat{\boldsymbol{\beta}}^t)) \leq \frac{\eta M^2}{2}$$

with probability approaching 1 as $n \to \infty$. □

A.3 Proof of Corollary 1

From Theorem 2, we have that

$$\frac{1}{T}\sum_{t=1}^{T} t(\Pi(\boldsymbol{\beta}) - \Pi(\hat{\boldsymbol{\beta}}^t)) \leq \frac{\eta M^2}{2}$$

with probability 1 as $n \to \infty$

Using σ- strong-concavity of Π, and the fact that $\nabla\Pi(\boldsymbol{\beta}^*) = 0$, we can rewrite this as:

$$\frac{\sigma}{2}\frac{1}{T}\sum_{t=1}^{T} t||\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^t||_2^2 \leq \frac{\eta M^2}{2}$$

Then, note that:

$$\frac{T^2}{2}||\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^T|| \leq \sum_{t=1}^{T} t||\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^T||_2^2 \leq \sum_{t=1}^{T} t||\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^t||$$

We can substitute this, which implies the result:

$$\frac{\sigma}{4}T||\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^T||_2^2 \leq \frac{\eta M^2}{2}$$

A.4 Proof of Theorem 3

Proof. The result follows from very simple proof by contradiction. Assume that $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{FP}$.

Then the solution to :

$$\lim_{n\to\infty} 0 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{\partial \pi_i(\boldsymbol{\beta}^*)}{\partial w} + \frac{\partial \pi_i}{\partial y}\frac{\partial Y_i(\boldsymbol{\beta}^*)}{\partial w}\right)\left(\frac{\partial \boldsymbol{X}_i(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}\frac{\partial W_i(\boldsymbol{\beta}^*)}{\partial \boldsymbol{x}} + \frac{\partial W_i(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}\right)\right]$$

and the solution to:

$$\lim_{n\to\infty} \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\partial \pi_i(\boldsymbol{\beta}^{FP})}{\partial w} + \frac{\partial \pi_i(\boldsymbol{\beta}^{FP})}{\partial y}\frac{\partial Y_i(\boldsymbol{\beta}^{FP})}{\partial w}\right]\frac{\partial W_i(\boldsymbol{\beta}^{FP})}{\partial \boldsymbol{\beta}} = 0$$

is the same. This implies that

$$\mathbb{E}\left[\frac{\partial \boldsymbol{X}_i(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}\frac{\partial W_i(\boldsymbol{\beta}^*)}{\partial \boldsymbol{x}}\right] = \mathbf{0}$$

But by the monotonicity assumptions in the Theorem, we know that is not the case, so it is not possible that $\boldsymbol{\beta}^{FP} = \boldsymbol{\beta}^*$

□

B Survey Questionnaire

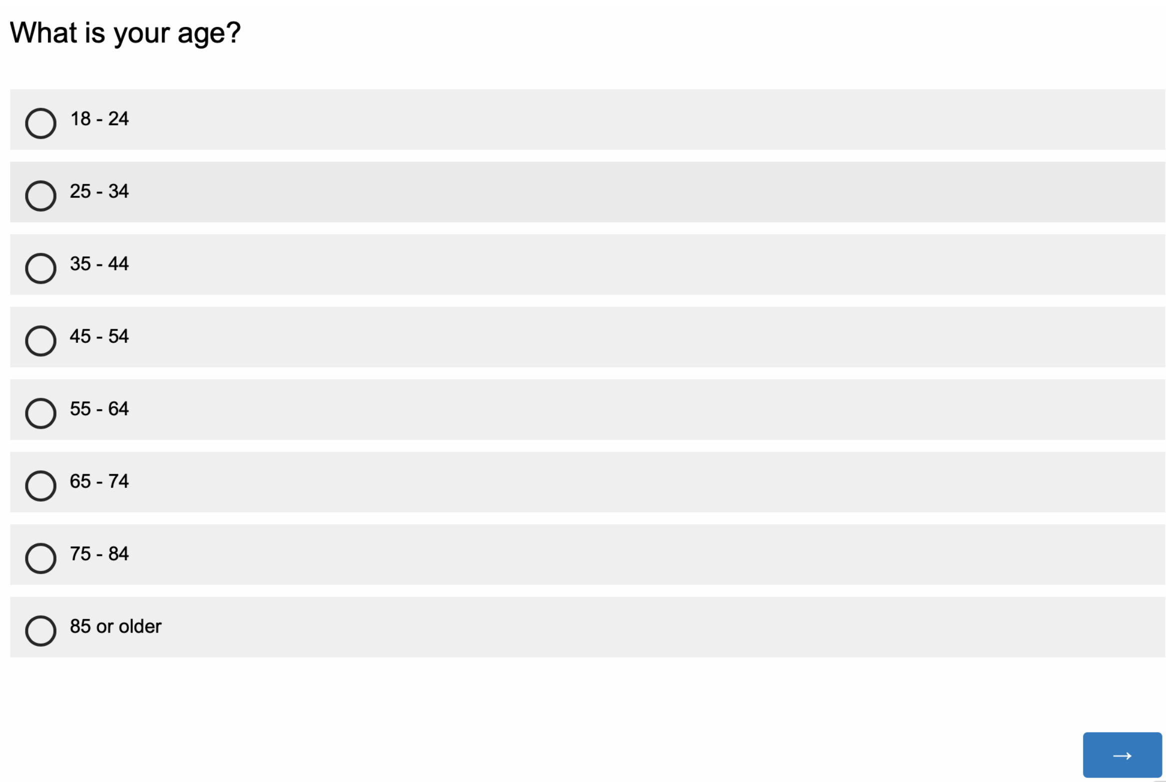
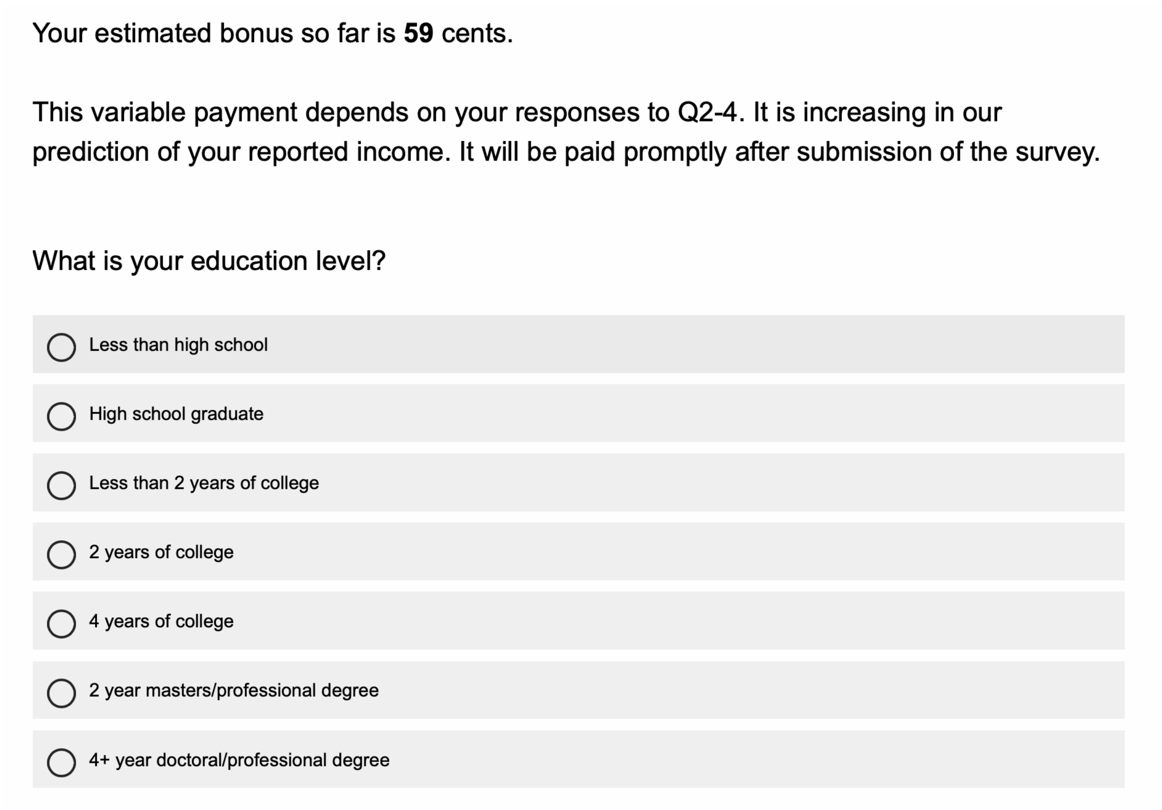


Figure 4: Page 1 of the survey is structured as a normal demographic survey.



Your estimated bonus so far is **59** cents.

This variable payment depends on your responses to Q2-4. It is increasing in our prediction of your reported income. It will be paid promptly after submission of the survey.

What is your education level?

- ◯ Less than high school
- ◯ High school graduate
- ◯ Less than 2 years of college
- ◯ 2 years of college
- ◯ 4 years of college
- ◯ 2 year masters/professional degree
- ◯ 4+ year doctoral/professional degree

Figure 5: In certain survey waves, incentives are introduced on Page 2 for individuals to misreport their education and car ownership.