The Relative Value of Prediction in Algorithmic Decision Making

Juan Carlos Perdomo [1]

Abstract

Algorithmic predictions are increasingly used to inform the allocations of goods and interventions in the public sphere. In these domains, predictions serve as a means to an end. They provide stakeholders with insights into likelihood of future events as a means to improve decision making quality, and enhance social welfare. However, if maximizing welfare is the ultimate goal, prediction is only a small piece of the puzzle. There are various other policy levers a social planner might pursue in order to improve bottom-line outcomes, such as expanding access to available goods, or increasing the effect sizes of interventions. Given this broad range of design decisions, a basic question to ask is: What is the relative value of prediction in algorithmic decision making? How do the improvements in welfare arising from better predictions compare to those of other policy levers? The goal of our work is to initiate the formal study of these questions. Our main results are theoretical in nature. We identify simple, sharp conditions determining the relative value of prediction vis-à-vis expanding access, within several statistical models that are popular amongst quantitative social scientists. Furthermore, we illustrate how these theoretical insights can guide the design of algorithmic decision making systems in practice.

1. Introduction

Algorithmic predictions are increasingly used in the public sphere to improve the allocation of scarce resources. At their core, these prediction algorithms aim to provide decision makers with valuable information regarding the impacts of particular interventions, or the likelihood of future events, in order to determine who will receive a social good.

For instance, in education, over half of US public schools use early warning systems that predict which students are likely to drop out, as a means to target their limited counseling resources to those that need them the most (Balfanz & Byrnes, 2019; U.S Department of Education, 2016). Similarly, in Israel, officials build risk predictors to determine the likelihood that individuals will develop serious pulmonary disease in order to prioritize people for vaccines (Barda et al., 2020). Beyond these specific examples, this idea that better predictions implies better decisions, and hence higher welfare, is pervasive, and underlies the design of risk prediction tools used in numerous domains (see Section 2).

From a design perspective, however, we see that prediction is really a means to an end, rather than an end in of itself. Predictions help inform decisions with the ultimate goal of increasing some context-dependent notion of social welfare (e.g. the number of on-time high school graduates). And, if improving social welfare truly is the goal, there are of course many different policy levers one could experiment with. To name a few, apart from optimizing the quality of decision making via better predictions, one could keep the current predictor fixed and focus efforts on expanding access to the available resources (e.g., invest in more counselors so as to intervene on more students). Alternatively, one could also focus on improving the quality of the interventions themselves (e.g., boost the quality of tutoring).

Those tasked with solving these resource allocation problems do not always have the luxury to simultaneously pursue all these various avenues of improvement. Given the broad space of design decisions, we inevitably need to identify the most cost efficient ways of improving overall welfare, and understand: *To what extent is prediction truly the highest order bit in algorithmic decision making?*

In this work, we initiate the formal study of the relative value of prediction. We aim to develop a principled understanding of how improvements in welfare arising from better prediction compare to those induced through other policy levers such expanding access or enhancing quality. As a byproduct of our inquiry into the relative value of prediction, we also aim to shed light on a related question: What does it mean for a predictor to be "good enough" for decision making? At what point do we decide that we no longer need to collect more features or train a more accurate predictor?

Summarizing briefly, we argue that predictions are sufficient

---

[1]Harvard University, Center for Research on Computation and Society, Cambridge, MA. Correspondence to: Juan Carlos Perdomo <jcperdomo@g.harvard.edu>.

to enable reliable decisions if the relative value of prediction is *small*. That is, if it is relatively more cost efficient to improve welfare by keeping the current predictor fixed and instead focusing efforts on expanding access or pursuing other policy levers. Our theoretical results establish simple and precise criteria for when these conditions hold true.

1.1. Overview of Framework & Contributions

In this paper, we study problems where a social planner is tasked with making a single binary decision for every member of a population, subject to a resource constraint. The reader may think about the decision as whether or not an intervention, or social good, is allocated to a particular individual unit. Constraints reflect limits on the total amount of goods or interventions that can be allocated. While there is a single decision being made per unit, crucially, the effects of these decisions differ across units (alternatively, individuals). The goal of the planner is to efficiently select which units are allocated goods in order to maximize overall welfare.

More formally, we assume that there is a distribution $\mathcal{D}$ over units represented as pairs (x_i, w_i). Here, $x_i \in \mathcal{X}$ is the feature representation of unit i and $w_i \in \mathbb{R}$ is the *welfare improvement* achieved by making a positive decision (intervening) on unit i. The planner uses these features to decide on an assignment policy $\pi : \mathcal{X} \to \{0, 1\}$, where $\pi(x_i) = 1$ if unit i is intervened on and is 0 otherwise. We formalize the idea that resources are scarce by requiring that $\mathbb{E}[\pi(x)] \leq \alpha$ for some $\alpha \in (0, 1)$. Writing this out as an optimization problem, the planner solves:

$$\max_{\pi \in \{\mathcal{X} \to \{0,1\}\}} \mathbb{E}_{(x_i, w_i) \sim \mathcal{D}}[w_i \pi(x_i)] \text{ s.t } \mathbb{E}_{\mathcal{D}}[\pi(x)] \leq \alpha. \tag{1}$$

Learning plays a central role in these problems since the optimal solution is to first predict the expected welfare improvement $\mathbb{E}[w \mid x]$ and then allocate positive decisions to the top α fraction of units, $\pi_*(x) = \mathbf{1}\{\mathbb{E}[w \mid x] > t(\alpha)\}$, where $t(\alpha)$ is some threshold. In practice, $\mathbb{E}[w \mid x]$ is of course not known and we approximate the solution by finding a predictor f such that $f(x) \approx \mathbb{E}[w \mid x]$.

The Prediction-Access Ratio. We formalize the relative value of prediction through a technical notion we call the *prediction-access ratio*, or PAR. Put simply, this is the ratio between the marginal improvement in social welfare that comes from improving the predictor f, and the improvement achieved by keeping prediction fixed and expanding access:

$$PAR = \frac{\textit{Marginal Improvement from Expanding Access}}{\textit{Marginal Improvement from Improving Prediction}}$$

The reader can think of the marginal improvement from prediction as the increase in the objective function, $\mathbb{E}[w_i \pi(x_i)]$, that comes from finding a predictor with smaller error, $\mathbb{E}(w_i - f(x_i))^2$. On the other hand, the improvement from

expanding access corresponds to the change in the optimal value of $\mathbb{E}[w_i \pi(x_i)]$ that comes from increasing α in the constraint $\mathbb{E}[\pi(x_i)] \leq \alpha$ and being able to assign goods or interventions to more people.

Calculating this ratio is a fundamental step in deciding which system changes are (locally) the most cost-effective. In particular, to decide whether small improvements in access or prediction are more effective, we simply calculate the following cost benefit ratio:

$$\frac{\textit{Marginal Cost of Improving Prediction}}{\textit{Marginal Cost of Expanding Access}} \times PAR \tag{2}$$

The cost-efficiency of expanding access or improve prediction hinges on whether this expression is smaller or larger than 1. While costs are relatively simple to calculate across different settings (e.g., it costs y dollars to hire an extra teacher and expand access to tutoring), the prediction access ratio is more subtle and requires counterfactual analysis (e.g., what would the world look like if my decisions were slightly better versus if I could just treat more people?).

The main technical contribution of our work is to calculate this ratio in the context of workhorse statistical models that are commonly used amongst quantitative social scientists. We start by stating our main result for the case where the effects of interventions are real-valued (e.g., increases in life expectancy from receiving a hip replacement).

Theorem 1.1 (Informal). *Assume that welfare improvements follow a linear regression model where w is real-valued and normally distributed. If the predictor f explains a γ_s^2 fraction of the variance in outcomes w, and at most an α fraction of the population can be intervened on, then the prediction access ratio is equal to γ_s/α.*

Essentially, the theorem states that if resources are very scarce in the sense that α is very small, and the predictor explains a non-trivial fraction of the variance (say 1%), then the relative value of prediction is *small*: The gains in social welfare that arise from expanding access far outweigh the gains that arise from improving prediction.[1] In particular, as long as the costs of expanding access are not $1/\alpha$ times larger than those of collecting more data or more features, then the cost-efficient way of making progress is to focus efforts on treating more people.

Another way to interpret this result is to plug in the value of the prediction access ratio into the cost-benefit analysis from Equation (2). A direct corollary of this first theorem is

---

[1]Note that explaining 1% of the variance means $\gamma_s \geq .1$. Simple predictors of life outcomes that use basic features typically explain at least a single digit fraction of the variance (Salganik et al., 2020). Resources are typically, but not always, much more heavily constrained. See Section 2 for further discussion.

that as long as $\sqrt{\textit{Current Variance Explained}}$ is at least,

$$\frac{\textit{Marginal Cost of Expanding Access}}{\textit{Marginal Cost of Expanding Prediction}} \times \textit{Current Access},$$

then Equation (2) is greater than 1, implying that the relative value of prediction is small and that expanding access is locally optimal. As per the previous discussion, this also constitutes a sufficient condition for a predictor to be "good enough" for algorithmic decision making: From a welfare perspective, it is efficient to leave the predictor "as is" and focus on other aspects of the broader system.

To get a better sense of how these insights extend to other settings, we also calculate the prediction access ratio for a problem where welfare improvements are discrete (e.g., changes in unemployment status), rather than real-valued.

Theorem 1.2 (Informal). *Assume welfare improvements follow a probit regression model where $w = \mathbf{1}\{z > 0\}$, and z is a real-valued and normally distributed score. If the predictor f explains a γ_s^2 fraction of the variance in z, and at most an α fraction of the population can be intervened on, then the prediction access ratio is equal to, $\gamma_t/\alpha^{1/\gamma_t^2}$, where $\gamma_t^2 = 1 - \gamma_s^2$ is the fraction of <u>unexplained</u> variance in the underlying score z.*

Since $1/\gamma_t^2 > 1$ for all $\gamma_s \in (0,1)$, the prediction-access ratio is almost always *larger* in this discrete case than in the previous, real-valued setting. Hence, the relative value of prediction is smaller and one should be more willing to expand access vis-à-vis improving prediction. As before, one can similarly factor costs into the equation, and again invert the bound to identify simple conditions under which it is cost-efficient to expand access vs improve prediction, or vice versa (see Section 4).

1.2. Implications & Limitations

At a high level, our main finding is that if a predictor explains even a small fraction of the variance in target outcomes and resources are significantly constrained, the benefits of expanding access far outweigh the welfare benefits induced by improving prediction. We believe that these results regarding the relative value of prediction are counter-narrative, if not quite counter-intuitive. It makes sense that if interventions have mostly positive effects, we are better off intervening on two people, rather than painstakingly improving prediction to identify the single person who will benefit the most from an intervention. While prediction is a means to an end, expanding access is an end in and of itself.

Furthermore, our results show that you do not necessarily need to predict something very accurately, in order to act upon it effectively. Within the probit model, for instance, the prediction-access ratio can be larger than 1 for small values of α, *regardless* of the predictive value (r^2 value) of the model (see Section 4 and Figure 2). Structural improvements like expanding access can have significantly larger

"bang for buck" relative to the gains achieved from building a better prediction system, even if predictions only explain a vanishingly small fraction of outcomes.

Crucially, however, our results and implications are limited to domains where: (a) the interventions allocated can be truly though of as social *goods* (and not harms), (b) the social planner has an accurate sense of costs, and (c) prediction problems are well-specified.

Allocating Goods vs Harms. The models we study satisfy the property that there is a large fraction of the population for which welfare improvements are positive, $\Pr[w_i > 0] = b \gg 0$. Furthermore, all our bounds for the prediction-access ratio only hold in the regime where $\alpha < b$ ($\alpha = $ level of access). And, if $\alpha < b$, one is (mostly) expanding access to interventions for people who would benefit from them. We believe that several important resource allocation problems like vaccination campaigns in public health, cash transfers in development economics, and targeted tutoring campaigns in K-12 education fall into this paradigm where treatment effects are mostly positive. However, our results are by no means universal and they do not apply to problems where the effects of interventions are less clear. For instance, if predictions are used to prioritize patients for invasive medical exams, these may prevent future disease, but also cause harm through a misdiagnosis. The relevant theoretical analysis in these settings likely differs significantly from ours, and we hope it will be the subject of future work.[2]

Understanding Costs. Our analysis only applies to domains where the planner has an accurate sense of the costs of the relevant policy levers. Note that while expanding access is expensive, so is improving prediction. In particular, the salient way to improve a predictor in social settings is to collect more features for the people in your population (i.e, perform additional in-person surveys, medical tests) rather than changing the algorithm, or spending money on additional compute. These measurement endeavors come with significant privacy and financial costs. Here, we treat the costs as known, domain-specific functions (see discussion in Section 3.2) and focus our efforts on analyzing the relative *welfare* impacts of prediction and access. This decision is largely motivated by the fact that policy analysts in charge of designing targeting programs often have extensive experience understanding costs, but may lack the tools necessary to evaluate the counterfactual impacts of access versus prediction. Nevertheless, we believe it's an important direction for future work to develop new machinery to better estimate costs in domains where these may be less clear.

Well-Specified Models. Lastly, our insights regarding the

---

[2]One major difference here is that predictions are valuable not just because they identify people from whom interventions are welfare-enhancing, but also rule out people for whom the intervention is harmful.

limits to prediction are shown in a simple, well-specified setting –favorable to prediction– where a planner is able to directly estimate the effects of interventions. In the language of potential outcomes (Neyman, 1923; Rubin, 1974), the welfare improvements $w_i = y_i(1) - y_i(0)$ are treatment effects: the difference between outcomes under treatment, $y_i(1)$, and no treatment, $y_i(0)$. Directly predicting heterogenous treatment effects is in general a difficult problem that requires strong assumptions both on the data *generating* process $\mathcal{D}$ as well as the prediction model class $\mathcal{F}$ (Pearl, 2009; Hardt & Recht, 2022). In practice, one should be cautious when applying these insights to domains that may not strictly follow the causal models we study.

Given the lack of previous work in this area, we view these simple and foundational statistical models as a natural starting point to establish a new, formal language around prediction in social systems. We hope it may enable future research that develops a more general theory that analyzes quantities like the prediction-access ratio under broader assumptions on the underlying population, as well as under different choices of social objective functions.[3]

2. Motivating Applications and Related Work

Our work lies at the intersection of various threads of research from computer science, economics, and related fields. We briefly discuss some of these connections and list examples of resource allocation problems from the applied literature that motivate our framework.

Economics. This idea of using predictions to guide the allocation of public resources has been studied extensively within the *prediction policy problems* literature. Introduced by Kleinberg et al. (2015), the term refers to policy questions, which do not directly require causal inference, and can be solved via pure prediction techniques. This framing has led to several impactful analyses of problems in health policy (Obermeyer et al., 2019), legal decision making (Kleinberg et al., 2018), and hiring (Chalfin et al., 2016). A parallel line of work by Susan Athey and collaborators (Athey, 2017; 2018; Athey & Imbens, 2019; Athey & Wager, 2021) aims to establish further connections between machine learning, causality, and applied policy problems. On the methods side, Bhattacharya & Dupas (2012), as well as Manski (2004; 2013), and Kitagawa & Tetenov (2018), study related statistical questions regarding the design of optimal targeting and testing mechanisms under budget constraints. And, Viviano (2023) analyzes the design of targeting policies under network effects. Our work contributes to this agenda by introducing a formal framework studying how improvements in prediction compare to other structural

avenues for improving welfare.

Computer Science. The statistical targeting problem we consider is related to contextual bandits with knapsack (resource) constraints: a decision maker sees features describing a unit, makes a binary decision, and observes a reward (welfare improvement) for the limited units that were intervened on. Agrawal & Devanur (2016), Slivkins et al. (2023), and Verma et al. (2021) design sublinear regret algorithms for versions of this problem. These papers treat the design parameters (constraints) as fixed, whereas we view them as a design decision. Our work is also related to the literature on online calibration in that it explicitly aims to design predictions that have strong guarantees with respect to the goals of downstream decision makers (Foster & Vohra, 1998; Foster & Hart, 2021; Noarov et al., 2023). Recently, Liu et al. (2023) analyzed the extent to which accurate predictions of future outcomes can help a decision maker looking to choose from multiple different interventions. Relative to their work, we consider a different setting where there is just one treatment, the prediction task is well-specified, and there are a limits on who may be intervened on.

2.1. Applications & Examples of Targeting Systems

Education. Apart from the early warning systems example mentioned previously, there are a number of other resource allocation problems in education that follow our framework including: predicting teacher value added in order to use limited school funds most efficiently (Chalfin et al., 2016), and predicting youth gun violence to allocate limited spots in mentoring programs (Rockoff et al., 2011).

Healthcare. The use of risk predictors to infer treatment effects and prioritize patients for interventions is ubiquitous throughout healthcare. In Israel, one of the major public health agencies regularly predicts outcomes such as COVID-19 mortality risk (Barda et al., 2020), or the likelihood of hepatitis C (Leventer-Roberts et al., 2022), to prioritize patients for medical attention. Bhattacharya & Dupas (2012) study predicting the prevalence of malaria in order to distribute bug nets amongst households in Kenya. In some of these settings, like bug nets or vaccination campaigns, resources are significantly constrained relative to the total size of the population, especially during the initial rollout of the program (UNICEF, 2022).

Government. The 1993 Unemployment Compensation Act (Sen. Moynihan, 1993) *requires* that U.S. states use profiling tools to predict the likelihood someone will remain unemployed after exhausting regular benefits to decide who will receive additional job training resources. Similar ideas are applied within development initiatives. For instance, Aiken et al. (2022) study the possibility of targeting cash transfers to the poorest members of society by using satellite imagery to infer poverty. Black et al. (2022) analyze issues of fair-

---

[3]In particular, we focus on optimizing for average welfare, however, it is also natural to also study other objectives such as maximizing the minimum outcome.

ness and efficiency when using algorithmic predictions to target tax audits. Researchers have also studied the use of ML to efficiently target building safety inspections (Johnson et al., 2023; Hino et al., 2018; Glaeser et al., 2016).

Throughout these domains, there is often a single good or intervention being allocated, predictions are used to figure out which units would experience the largest welfare improvements if acted upon, and the decision maker allocates interventions to those with the highest predicted effects as per the targeting problem formalized in Equation (1).

3. Linear Regression Model

In this section, we present our first set of main results analyzing the prediction-access ratio for the case where welfare improvements follow a linear regression model. Linear regression is the workhorse method of analysis amongst economists, sociologists and other quantitative social scientists. It has been applied, with varying degrees of success (Freedman, 1999), to model causal effects (welfare improvements) in consequential domains such as education (Angrist & Lavy, 1999; Yule, 1899) and public health (Chandler et al., 2011). Furthermore, amongst theoreticians, linear models are the *drosophilia* of formal inquiry into more complex questions. According to Pearl, "many concepts and phenomena in causal analysis were first detected, quantified, and exemplified in linear structural equation models before they were understood in full generality and applied to nonparametric problems" (Pearl, 2013). As such, they are a natural starting point to study the prediction-access ratio.

3.1. Model Definition and Technical Preliminaries

Definition 3.1 (Linear Regression Model). We say that welfare improvements follow a linear regression model if [4]

$$w_i = \langle x_i,\ \beta \rangle + \mu \quad \text{where } x_i \sim \mathcal{N}(0, I), \qquad (3)$$

and $\beta \in \mathbb{R}^d$ is a vector of unknown coefficients. Here, $\mu = \mathbb{E}[w_i]$ is the average welfare improvement and $\|\beta\|_2^2 = \mathbb{E}[(w_i - \mu)^2]$ captures the heterogeneity in outcomes.[5]

While outcomes w_i are deterministic functions of x_i, the social planner does not observe the full vector of features, but rather only a subset. In particular, we assume that features x are partitioned into $x = (x_s, x_t) \in \mathcal{X}_s \times \mathcal{X}_t$: a set of observed features x_s and unobserved features x_t. For

---

[4]Note that Def. 3.1 is property of the data *generating* process, not just the model class used for prediction. As per the earlier discussion, welfare improvements w_i are treatment effects (causal quantities) and prediction of causal quantities requires assumptions on the underlying data generating process (Pearl, 2009).

[5]The condition that $x_i \sim \mathcal{N}(0, I)$ comes with no further loss of generality. In particular, our results also apply to the case where $x \sim \mathcal{N}(0, \Sigma)$, (the features are now correlated) by replacing β with $\Sigma^{1/2}\beta$. Since if $x \sim \mathcal{N}(0, \Sigma)$ then $\langle x, \beta \rangle \overset{d}{=} \langle x', \beta' \rangle$ where $x' \sim \mathcal{N}(0, I)$ and $\beta' = \Sigma^{1/2}\beta$.

example, the impacts of educational interventions on the likelihood of on-time graduation are the function of many variables. Some of these variables (e.g., student demographics, test scores, attendance rate) are observed, and some are typically unobserved (e.g., level of parental support).

Using the available data, the social planner aims to maximize average welfare, subject to a resource constraint that at most an α fraction of the population can be treated. We formally define the planner's targeting problem below:

Definition 3.2 (Planner's Targeting Problem). Given access to the observable features x_s, the planner solves the following optimization problem:

$$\max_{\pi \in \{\mathcal{X}_s \to \{0,1\}\}} \mathbb{E}_{(x_{i,s}, w_i) \sim \mathcal{D}}[w_i \pi(x_{i,s})] \text{ s.t } \mathbb{E}_{\mathcal{D}}[\pi(x_{i,s})] \le \alpha.$$

In this regression setting, we assume that given the partition of features into observables and unobservables,

$$w_i = \langle x_i,\ \beta \rangle + \mu = \underbrace{\langle x_{i,s},\ \beta_s \rangle + \mu}_{Observable} + \underbrace{\langle x_{i,t},\ \beta_t \rangle}_{Unobservable},$$

the planner has enough data that they can learn the coefficients β_s, corresponding to the observable components x_s exactly. In practice, one would estimate β_s in finite samples through procedures like ordinary least squares and find an estimate $\hat{\beta}_s$ such that $\|\beta_s - \hat{\beta}_s\|_2 \le \mathcal{O}(\sqrt{\dim(\beta_s)/n})$.

However, for the class of problems we are interested in, data typically comes from administrative databases collected by some centralized government authority, where the number of samples n is typically much larger than the number of features, $\dim(x_s)$. For example, in the Wisconsin early warning system studied in Perdomo et al. (2023), there are about 40 features and n is over 300k (all public school students in Wisconsin). Given these relevant scales, the salient axis for improving prediction in these settings is to collect more features (e.g conduct more tests, increase $\dim(x_s)$) rather than collecting more samples, since finite sample errors are essentially negligible. We reflect this reality in our model to streamline our presentation.

Technical Preliminaries. Let,

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \text{ and } \Phi(t) = \int_{-\infty}^{t} \phi(z)dz,$$

be the PDF and CDF for a standard normal random variable, $\mathcal{N}(0, 1)$, and define $\Phi^{-1}(\alpha) = \inf\{t \in \mathbb{R} : \alpha \le \Phi(t)\}$ to be the associated quantile function.

The value $\mathbb{E}[w_i \pi_s^*(x_{i,s})]$ achieved by the policy π_s^* is intimately tied to the variance explained by the observable features x_s. We say that the observable features x_s have an r^2 value of γ_s^2 if the associated predictor, $f(x_s) = \langle x_s,\ \beta_s \rangle$,

explains a γ_s^2 fraction of the variance in the outcomes w.[6]

Definition 3.3 (r^2 - Linear Regression)**.** For w_i satisfying the linear model (Def. 3.1), we say that observable features x_s have an r^2 value of γ_s^2, $r^2(x_s) = \gamma_s^2 \in [0, 1]$, if

$$r^2(x_s) = 1 - \frac{\mathbb{E}(w_i - \mu - \langle x_{i,s}, \ \beta_s \rangle)^2}{\mathbb{E}(w_i - \mu)^2} = \gamma_s^2. \quad (4)$$

A predictor that uses all the available features $x_s = x$, has an r^2 value of 1, while a model that uses no features has an r^2 value of 0. With this notation, we can define the value of a policy: the overall welfare improvement achieved as a function of the distribution of the w_i, the level of access α, and the variance explained by the observed features x_i.

Proposition 3.4 (Value Function, Linear Regression)**.** *For w_i satisfying the linear regression model (Definition 3.1), define $V_*^{\text{lin}}(\alpha, \gamma_s)$ to be,[7]*

$$V_*^{\text{lin}}(\alpha, \gamma_s) = \max_{\pi \in \{\mathcal{X}_s \to \{0,1\}\}} \mathbb{E}_{(x_{i,s}, w_i) \sim \mathcal{D}}[w_i \cdot \pi(x_{i,s})]$$

$$\text{s.t} \quad \mathbb{E}_{\mathcal{D}}[\pi(x_{i,s})] \le \alpha \text{ and } r^2(x_s) = \gamma_s^2, \quad (5)$$

the value achieved by the optimal policy π_s^* (see Lemma A.2) that observes features x_s and has an associated r^2 value equal to γ_s^2. Then, for $\alpha < 1/2$ and $\|\beta\|_2, \mu > 0$,

$$V_*^{\text{lin}}(\alpha, \gamma_s) = \alpha\mu + \gamma_s \|\beta\|_2 \phi(\Phi^{-1}(1-\alpha)). \quad (6)$$

In this linear setting, the value function for the optimal policy admits a simple closed-form solution (Equation (6)) and has an intuitive interpretation. In particular, $V_*^{\text{lin}}(\alpha, \gamma_s)$ consists of two terms. The first term, $\alpha\mu$, is exactly the value achieved by a random assignment policy. For $w_i \sim \mathcal{N}(\mu, \|\beta\|_2^2)$, if you select who gets treated on the basis of a random coin toss, $\pi_{\text{rand}}(x_{i,s}) \sim \text{Ber}(\alpha)$, the expected welfare is equal to $\mathbb{E}[w_i \pi_{\text{rand}}(x_{i,s})] = \mathbb{E}[w_i]\mathbb{E}[\pi_{\text{rand}}(x_{i,s})] = \alpha\mu$. [8] On the other hand, the second term in Equation (6) captures the value of predictions that are better than random (and hence have $\gamma_s > 0$). Recall that for $w_i = \langle \beta, \ x_i \rangle$ and $x_i \sim \mathcal{N}(0, I)$, the total variance in the outcomes is $\|\beta\|_2^2$ and the variance captured by the features is $\|\beta_s\|_2^2$ (since $\langle x_{i,s}, \ \beta_s \rangle \sim \mathcal{N}(0, \|\beta_s\|_2^2)$). Therefore, the marginal value of using prediction, versus random assignment, scales with the square root of the variance explained, $\|\beta_s\|_2 = \gamma_s\|\beta\|_2$, and a quantity that depends on the level of access: $\phi(\Phi^{-1}(1-\alpha))$.

---

[6]Due to the tight correspondence between features x_s and predictors $f(x_s) = \langle x_s, \beta_s \rangle$ in this linear model, we use the phrases, the features (or predictor) have an r^2 value of γ_s^2 interchangeably to refer to the condition in Eq. 4. For clarity, we abuse notation and write $r^2(x_s)$ instead of the usual $r^2(f)$ for a function f.

[7]We parametrize the value function in terms of γ_s instead of the set of observed features x_s since the value of the features in this model is exactly captured by γ_s.

[8]Here, $\text{Ber}(\alpha)$ is a Bernoulli random variable with parameter α. Under π_{rand}, each unit i receives the good, or intervention, independently with probability α.

When is Prediction Even Necessary. As a final note before discussing our main result for this section, we pause to point out that if there is very little variance in the outcome variable (i.e., $\|\beta\|_2^2$ is small) then, random assignment (i.e., deciding allocations according to $\pi_{\text{rand}}(x_{i,s}) \sim \text{Ber}(\alpha)$) is near optimal. Statistical targeting, above and beyond simple random allocations, only makes sense if there is significant variance in the outcomes w_i.

To see this, let $V_{\text{random}}(\alpha) = \mathbb{E}[w_i \pi_{\text{rand}}(x_{i,s})]$ for $\pi_{\text{rand}}(x_{i,s}) = \text{Ber}(\alpha)$. Then, for $\alpha, \mu, \|\beta\|_2 > 0$,

$$\frac{V_{\text{random}}(\alpha)}{V_*^{\text{lin}}(\alpha, 1)} = \frac{1}{1 + \frac{\|\beta\|_2}{\mu}\alpha^{-1}\phi(\Phi^{-1}(1-\alpha))},$$

where $V_*^{\text{lin}}(\alpha, 1)$ from Equation (6) is the value of the policy $\pi^*(x)$ with $r^2(x) = 1$ that observes <u>all</u> features. As $\|\beta\|_2/\mu \to 0$, then $V_{\text{random}}(\alpha)/V_*^{\text{lin}}(\alpha, 1) \to 1$ and π_{rand} is optimal. While somewhat trite, it is nevertheless an important, common-sense sanity check. Prediction-enabled targeting only makes sense if the variance in the target outcomes w is large relative to the mean. Since the goal of our work is to study the relative value of prediction – in settings where prediction makes sense to begin with – we will assume for the remainder that $\|\beta\|_2/\mu$ is not too small.

3.2. The Prediction-Access Ratio for Linear Regression

Defining Policy Levers. In the context of this linear model, various policy levers such as improving prediction and expanding access have simple, formal counterparts.

Improving Prediction: Optimizing prediction in this setting equates to enlarging the set of measured features. This yields a new partition of observed and unobserved covariates $(x_s, x_t) \to (x_{s'}, x_{t'})$ such that if $r^2(x_s) = \gamma_s^2$ then $r^2(x_{s'}) = (\gamma_s + \Delta_{r^2})^2$ for some $\Delta_{r^2} > 0$. As per our notation from Equation (5), the relevant improvement in welfare is equal to: $V_*^{\text{lin}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\text{lin}}(\alpha, \gamma_s)$.

Expanding Access. This corresponds to keeping the current policy π_s^* fixed and slackening the resource constraint α to $\alpha + \Delta_\alpha$. The increase in average welfare is equal to $V_*^{\text{lin}}(\alpha + \Delta_\alpha, \gamma_s) - V_*^{\text{lin}}(\alpha, \gamma_s)$.

Given current system parameters (α, γ_s), and proposed improvements $\Delta_{r^2}, \Delta_\alpha \in (0, 1)$, the prediction-access ratio, $\text{PAR}^{\text{lin}}(\alpha, \gamma_s, \Delta_\alpha, \Delta_{r^2})$, defined formally below, determines the relative impact of these changes on welfare:

$$\text{PAR}^{\text{lin}}(\alpha, \gamma_s, \Delta_\alpha, \Delta_{r^2}) = \frac{V_*^{\text{lin}}(\alpha + \Delta_\alpha, \gamma_s) - V_*^{\text{lin}}(\alpha, \gamma_s)}{V_*^{\text{lin}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\text{lin}}(\alpha, \gamma_s)}.$$
$$(7)$$

Values of the ratio larger than 1 indicate that a Δ_α increase in access yields a larger increase in welfare than a Δ_{r^2} improvement in prediction. The reverse is true if the ratio is
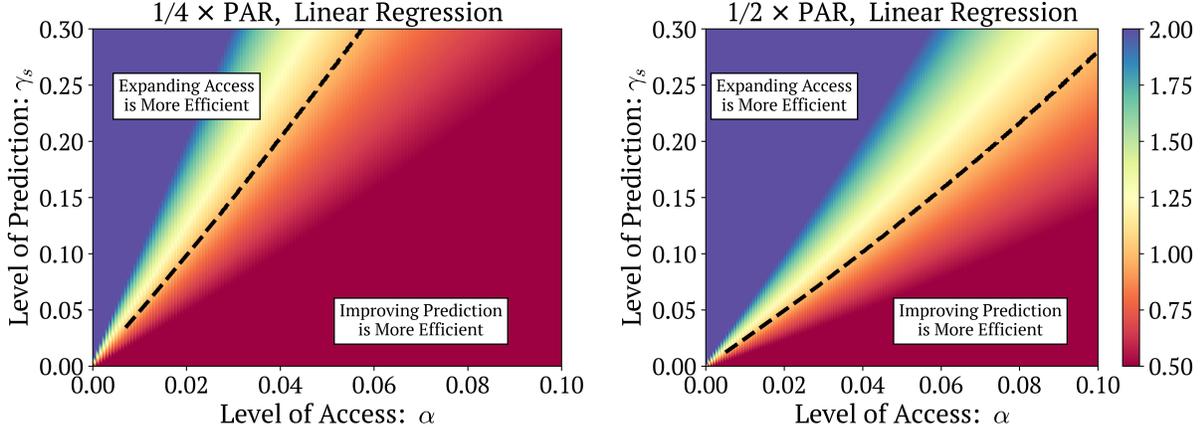
Figure 1. Visualization of the cost benefit ratio, Equation (8), for the linear regression model. We compute the ratio for each value of α (x-axis) and γ_s (y-axis), exactly via numerical simulation with $\Delta_\alpha = \Delta_{r^2}$. We display its value, clipped to $[1/2, 2]$, via the color bar. We set $C_{r^2}(\Delta_{r^2})/C_\alpha(\Delta_\alpha) = 1/4$ on the left and $1/2$ on the right. The black line indicates the set of points for which the ratio is equal to 1. As per Equation (9), the cutoff is approximate of the form $\gamma_s \propto \alpha$, where the slope is determined by the cost ratio. For values (α, γ_s) above the line, expanding access is relatively cost efficient, whereas improving prediction is efficient for points below the line.

smaller than 1. The following theorem is the main result of this section, effectively identifying this ratio (up to a small constant) for the linear regression model.

Theorem 3.5 (Prediction-Access Ratio, Linear Regression). *Assume that w_i satisfy the linear regression model (Definition 3.1) with $\|\beta\|_2, \mu > 0$. For any $\gamma_s, \Delta_{r^2} \in (0, 1)$ and Δ_α satisfying $\alpha + \Delta_\alpha < .05, \Delta_\alpha \leq 4\alpha$, define M_{lin} to be*

$$M_{\mathrm{lin}} = \frac{1}{\alpha}\left(\frac{\mu}{\|\beta\|_2 \Phi^{-1}(1-\alpha)} + \gamma_s\right).$$

Then, $.25 M_{\mathrm{lin}} \leq \mathrm{PAR}^{\mathrm{lin}}(\alpha, \gamma_s, \Delta_\alpha, \Delta_{r^2}) \leq M_{\mathrm{lin}}$.

Discussion. The main message of this theorem is that, if resources are scarce in the sense that α is small, then the local improvements induced by expanding access are far larger than the improvements in welfare that come from better targeting. More specifically, if we fix $\Delta_\alpha = \Delta_{r^2}$ and take γ_s to at least a constant (say 1%), the prediction-access ratio is order $1/\alpha$.[9] Hence, the amount we should be willing to pay to expand access *increases* the smaller α is. Said otherwise, the relative value of prediction in these targeting problems *decreases* with the level of scarcity.

Recall from the introduction that to decide which policy lever makes sense, we only need to factor costs into the equation. Here, we encourage the reader to think of costs in a broad sense. The costs of improving prediction are monetary (e.g., compute bills, labor) as well as *social* (e.g., the privacy costs of people releasing personal information to a central authority, or subjecting individuals to a means test). A similar comment applies to the costs of other policy levers. However, for the sake of modeling, we can imagine functions C_α and C_{r^2} from $[0, 1] \to \mathbb{R}_{\geq 0}$ that give costs

numeric values.[10] The decision to expand access vs improve prediction hinges on whether the cost-benefit ratio, formally defined for this linear regression model as,

$$\mathrm{PAR}^{\mathrm{lin}}(\alpha, \gamma_s, \Delta_\alpha, \Delta_{r^2}) \times \frac{C_{r^2}(\Delta_{r^2})}{C_\alpha(\Delta_\alpha)}, \qquad (8)$$

is larger than 1. Theorem 3.5 essentially states that as long as $C_\alpha(\Delta_\alpha) < \gamma_s/\alpha C_{r^2}(\Delta_{r^2})$, expanding access is the cost efficient decision. To see this in more detail, we can reinterpret the result as establishing a *threshold* for when a set of features is sufficiently "good" for targeting. As long as,

$$\gamma_s \gtrsim \frac{\Delta_{r^2} C_\alpha(\Delta_\alpha)}{\Delta_\alpha C_{r^2}(\Delta_{r^2})} \cdot \alpha - \frac{\mu}{\|\beta\|_2 \Phi^{-1}(1-\alpha)}, \qquad (9)$$

the cost-benefit ratio from Equation (8) is greater than 1 and expanding access makes economic sense.[11]

Visualization. We examine this threshold visually in Figure 1. In particular, in each plot we illustrate the cost benefit ratio (Equation (8)), for various choices of the ratio $C_{r^2}(\Delta_{r^2})/C_\alpha(\Delta_\alpha)$ where in each case $\Delta_\alpha = \Delta_{r^2} = .01$. As predicted by our theory (Theorem 3.5 and Equation (9)), we see that, for small values of α, the threshold at which one is indifferent between expanding access vs improving prediction is of the form $\gamma_s \propto \alpha$. That is, given any pairs of level of access α and prediction γ_s, the cost benefit ratio

---

[10]These functions return the *marginal* costs of increasing access from α to $\alpha + \Delta_\alpha$ and prediction (r^2) from γ_s^2 to $(\gamma_s + \Delta_{r^2})^2$. We could thus write them as $C_\alpha(\Delta_\alpha; \alpha)$ and $C_{r^2}(\Delta_{r^2}; \gamma_s)$ but omit the second argument as it is clear from context.

[11]In Theorem 3.5, we assume bounds on α to simplify the presentation of the main bound. From our analysis, we can prove a more general result that holds for general values α, yet the resulting expressions are unfortunately not as clean. As per our discussion, we largely neglect the second term in Eq. 9 since $\|\beta\|_2 \gg \mu$.

---

[9]$\Phi^{-1}(1-\alpha)$ is $\Theta(\sqrt{\log(1/\alpha)})$ for for small α, see Prop. A.8.

is above 1 if $\alpha \lesssim \gamma_s$ and below 1 if $\alpha \gtrsim \gamma_s$. The exact constant depends on the relative costs of prediction and expanding access. As per Equation (9), changing the relative costs only changes the slope of the line.

A different way to interpret this finding is that in the linear model, the optimal values of (α, γ_s) go "hand in hand". That is, if a predictor explains a γ_s^2 fraction of the the variance, then the *efficient* level of access is to set $\alpha \propto \gamma_s$: One should increase access until it matches the (square root of) the explained variance. This in effect establishes priorities for how a decision making system should be built out.

The Relative Value of Improving Quality. So far, we have focused on the relative impacts of prediction and access on welfare. However, there are other policy levers one might experiment with, for instance, improving the quality of the allocated goods. One way to formalize this, in this regression model it to consider increasing $\mu = \mathbb{E}[w]$, the average welfare improvement. Recall from Prop. 3.4 that the value function is equal to: $\mu\alpha + \gamma_s\|\beta\|_2\phi(\Phi^{-1}(1-\alpha))$. Since the value function is linear in the average quality μ, increasing μ to $\mu + \Delta_\mu$ increases the value function by an additive factor of $\Delta_\mu\alpha$. Hence, in this regression model, increasing the effect size of the average outcome yields an improvement in welfare that is on the same order as improving in prediction. Both are $\widetilde{\Theta}(\alpha)$ (see Equation (13)), and significantly smaller than improvements in access, which are $\widetilde{\Theta}(1)$.

Local vs Global Decision Making. The value of the prediction access ratio is that it helps guide which kind of *local* improvement is most effective. That is, it answers questions of the form: Given that my current system treats an α fraction of the population and explains a γ_s^2 fraction of the variance in treatment effects, is it better to expand access by 1% or improve prediction by 1%? This is motivated by the fact that that there is often significant inertia in the overall design of algorithmic decision making systems. Large government-run programs, like vaccine rollouts, cannot always be "restarted from scratch". Pragmatically speaking, only local changes are viable.

4. Probit Regression Model

Depending on how we measure outcomes, welfare improvements are often binary (e.g., on-time graduation) rather than real-valued as in the linear regression setting analyzed previously. Therefore, in this section we study the prediction-access ratio in the context of a discrete model for outcomes: probit regression. Like linear regression, probit regression is one of the most popular statistical models used in the quantitative social sciences. On a technical note, it belongs to the class of generalized linear models.[12]

---

[12]The other popular model in this class is logistic regression. Both probit and logistic regression transform the real-valued outputs of a regression model, $\langle \beta, x \rangle$, into a number in $[0, 1]$ (a prob-

4.1. Model Definition and Technical Preliminaries

Definition 4.1 (Probit Regression Model)**.** We say that welfare improvements follow a probit regression model if $w_i = \mathbf{1}\{\langle x_i,\ \beta\rangle + \mu > 0\}$ for $x_i \sim \mathcal{N}(0, I)$, where again $\beta \in \mathbb{R}^d$ is a vector of coefficients. Furthermore, we denote by $b = \Pr[w_i > 0]$ the base rate of positive welfare improvements in the population.

As before, we assume that features x are partitioned into observed and unobserved components, (x_s, x_t) and that the planner collects enough data to learn the coefficients β_s exactly. In the model outlined above, these parameters can be recovered in finite samples via convex programming at similar $\sqrt{\dim(x_s)/n}$ rates (Fahrmeir & Kaufmann, 1985).

Due to the nonlinearity (thresholding) being applied, parameters like μ and $\|\beta\|_2$ in this setting no longer directly correspond to natural quantities like mean and variance of w_i as in the linear case. Yet, the optimal policy for the probit case is conveniently the same as before (see Lemma B.1). Furthermore, given the binary nature of the outcome variable, there is no standard, agreed-upon measure for the coefficient of determination as in the continuous case. We find it convenient to deal with the following measure:

Definition 4.2 (r^2 - Probit Regression)**.** Assume that (w_i, x_i) satisfy the probit regression model (Definition 4.1). Features x_s have an r^2 value of $\gamma_s^2 \in [0, 1]$, $r^2(x_s) = \gamma_s^2$, if

$$r^2(x_s) = 1 - \frac{\mathbb{E}(z_i - \mu - \langle x_{i,s},\ \beta_s\rangle)^2}{\mathbb{E}(z_i - \mu)^2} = \gamma_s^2 \qquad (10)$$

for $z_i = \langle x_i, \beta\rangle + \mu$ and $w_i = \mathbf{1}\{z_i > 0\}$.

The value function for the probit case can again be parametrized in terms of γ_s and α.

Proposition 4.3. *For w_i satisfying the probit regression model (Definition 3.1), let $V_*^{\mathrm{pr}}(\alpha, \gamma_s)$ be*

$$V_*^{\mathrm{pr}}(\alpha, \gamma_s) = \max_{\pi \in \{\mathcal{X}_s \to \{0,1\}\}} \mathbb{E}_{(x_{i,s}, w_i) \sim \mathcal{D}}[w_i \cdot \pi(x_{i,s})]$$

$$\text{s.t} \quad \mathbb{E}_{\mathcal{D}}[\pi(x_{i,s})] \leq \alpha \text{ and } r^2(x_s) = \gamma_s^2 \qquad (11)$$

the value achieved by the optimal policy π_s^* (see Lemma B.1) that uses features x_s with $r^2(x_s) = \gamma_s^2$, then,

$$V_*^{\mathrm{pr}}(\alpha, \gamma_s) = \int_{\Phi^{-1}(1-\alpha)}^{\infty} \Phi\left(\frac{\gamma_s z_s + \mu\|\beta\|_2^{-1}}{\gamma_t}\right) \phi(z_s) dz_s$$

where $\gamma_t^2 = 1 - \gamma_s^2$ is the fraction of unexplained variance in the scores $z_i = \langle x_i,\ \beta\rangle + \mu$.

---

ability). The exact transformation is technically different, but morally they tend to provide similar "bottom-line" insights when applied in practice (Stock et al., 2003; Scott Long, 1997).
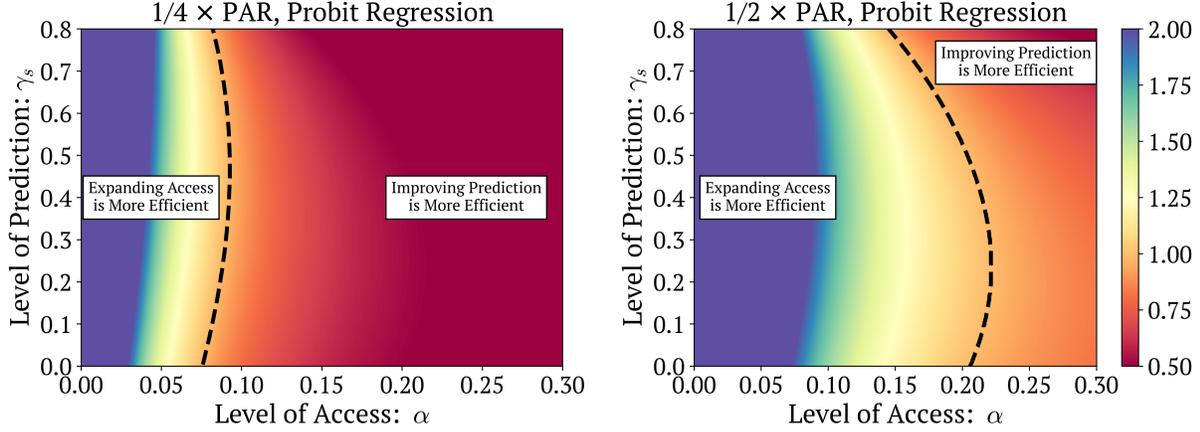
Figure 2. Visualization of the cost benefit ratio, Equation (12), for the probit model. As in Figure 1, we compute the ratio numerically with $\Delta_\alpha = \Delta_{r^2}$ and display its value, clipped to [1/2, 2], via the color bar. The black line indicates points for which the ratio is equal to 1. As per our analysis, the threshold between access and prediction is nonlinear, and the ratio is larger than 1 for small α, regardless of γ_s.

The nonlinear nature of the probit model complicates calculations so that the value functions no longer have simple expressions in terms of analytical functions. Interestingly enough, while V_*^{pr} has no closed-form solution, the prediction-access ratio does.

4.2. The Prediction Access Ratio for Probit Regression

We define policy levers for the probit model in a similar fashion as before. Improving prediction corresponds to expanding the set of measured features and increasing γ_s by some $\Delta_{r^2} > 0$. Expanding access corresponds to increasing the amount of resources from α to $\alpha + \Delta_\alpha$ for some $\Delta_\alpha > 0$. Formally, we define the prediction access ratio for the probit model to be:

$$\mathrm{PAR}^{\mathrm{pr}}(\alpha, \gamma_s, \Delta_\alpha, \Delta_{r^2}) = \frac{V_*^{\mathrm{pr}}(\alpha + \Delta_\alpha, \gamma_s) - V_*^{\mathrm{pr}}(\alpha, \gamma_s)}{V_*^{\mathrm{pr}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\mathrm{pr}}(\alpha, \gamma_s)}.$$

Theorem 4.4 (Prediction-Access Ratio, Probit Regression). *For any $\gamma_s \in (0, 1)$, $b = \Pr[w_i = 1] < .1$, and $\varepsilon \in (0, .1)$, let $\varepsilon' = \varepsilon/(1 - \varepsilon)$. There exists a threshold $t > 0$, such that for $\Delta_{r^2}, \alpha \leq t$ and $\Delta_\alpha \leq \alpha$, the following inequalities hold*

$$\frac{3}{10} \frac{\Delta_\alpha \gamma_t}{\Delta_{r^2}} \frac{1}{b \Phi^{-1}(1-b)} \left(\frac{1}{1.01\sqrt{2\pi}} \frac{1}{\alpha \Phi^{-1}(1-\alpha)} \right)^{1/\gamma_t^2 \cdot (1-\varepsilon)^2}$$
$$\leq \mathrm{PAR}^{\mathrm{pr}}(\alpha, \gamma_s, \Delta_\alpha, \Delta_{r^2}) \leq$$
$$3 \frac{\Delta_\alpha \gamma_t}{\Delta_{r^2}} \cdot \frac{1}{b \Phi^{-1}(1-b)} \left(\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\alpha \Phi^{-1}(1-\alpha)} \right)^{1/\gamma_t^2 \cdot (1+\varepsilon')^2}$$

Discussion. Ignoring constants and lower order terms, we see that the prediction-access ratio is essentially[13] $\widetilde{\Theta}(\gamma_t \alpha^{-1/\gamma_t^2})$. Hence, this ratio is significantly *larger* than in the linear case. Since $1/\gamma_t^2 > 1$, the ratio behaves like

---

[13]In particular, the ε and ε' are constants that can be chosen to be arbitrarily close to 0. Their exact value only affects the threshold t that upper bounds α and Δ_{r^2}.

α^{-c} for some $c > 1$ rather than α^{-1} as before. This implies that the relative value of prediction is *smaller* in this discrete case and one should be relatively *more* willing to expand access (even if the predictor explains only a very small fraction of the outcome variable). As motivated previously, the benefits of access outweigh those of prediction if the cost-benefit ratio, defined below for the probit model, is greater than 1.

$$\mathrm{PAR}^{\mathrm{pr}}(\alpha, \gamma_s, \Delta_\alpha, \Delta_{r^2}) \times \frac{C_{r^2}(\Delta_{r^2})}{C_\alpha(\Delta_\alpha)}. \qquad (12)$$

Plugging in our bounds from Theorem 4.4, and again ignoring lower order terms, the cost-benefit ratio (Equation (12)) is at least 1 if $\gamma_t \gtrsim \frac{C_\alpha(\Delta_\alpha)}{C_{r^2}(\Delta_{r^2})} \alpha^{1/\gamma_t^2}$. Therefore, if the level of access α is small (and in particular less than $\Pr[w_i > 0] = b$), as long as the costs of expanding access are not several orders of magnitude higher than those of improving prediction, expanding access is almost always the cost-efficient avenue of improvement.

Visualizations. As before, we complement these theoretical insights with numerical simulations (Figure 2). The plot displays values of the cost benefit ratio, Equation (12), for various choices of parameters (α, γ_s) and cost ratios $C_{r^2}(\Delta_{r^2})/C_\alpha(\Delta_\alpha)$. The visualizations further illustrate the technical point made previously: In this probit model the relative value of prediction is *smaller* than in the linear case. In particular, if α is well below the base rate of positive improvements, $\alpha \ll b = \Pr[w_i = 1]$, the cost benefit ratio is always significantly larger than 1, *regardless* of the level of prediction γ_s. Due to the nonlinear nature of the problem, the threshold at which one is indifferent between expanding prediction versus expanding access is no longer linear. In this model, prediction and access do not go "hand in hand" as before. Rather, as long as $\gamma_s > 0$, the focus should always be to expand access until $\alpha \approx b$, at which point the focus should switch to improving the predictor.

9

Impact Statement

This work analyzes the impact of machine learning systems within their broader social contexts and asks how predictions can improve welfare relative to other policy levers. That being said, it takes as given that the goals of prediction are broadly aligned with social welfare, which may or may not always be the case. Our contributions are largely theoretical, there are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

References

Agrawal, S. and Devanur, N. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 2016.

Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 2022.

Angrist, J. D. and Lavy, V. Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 1999.

Athey, S. Beyond prediction: Using big data for policy problems. *Science*, 2017.

Athey, S. The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2018.

Athey, S. and Imbens, G. W. Machine learning methods that economists should know about. *Annual Review of Economics*, 2019.

Athey, S. and Wager, S. Policy learning with observational data. *Econometrica*, 2021.

Balfanz, R. and Byrnes, V. Early warning indicators and intervention systems: State of the field. *Handbook of Student Engagement Interventions*, 2019.

Barda, N., Riesel, D., Akriv, A., Levy, J., Finkel, U., Yona, G., Greenfeld, D., Sheiba, S., Somer, J., Bachmat, E., et al. Developing a covid-19 mortality risk prediction model when individual-level data are not available. *Nature Communications*, 2020.

Bhattacharya, D. and Dupas, P. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 2012.

Black, E., Elzayn, H., Chouldechova, A., Goldin, J., and Ho, D. Algorithmic fairness and vertical equity: Income fairness with irs tax audit models. In *ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., and Mullainathan, S. Productivity and selection of human capital with machine learning. *American Economic Review*, 2016.

Chandler, D., Levitt, S. D., and List, J. A. Predicting and preventing shootings among at-risk youth. *American Economic Review*, 2011.

Fahrmeir, L. and Kaufmann, H. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 1985.

Foster, D. P. and Hart, S. Forecast hedging and calibration. *Journal of Political Economy*, 2021.

Foster, D. P. and Vohra, R. V. Asymptotic calibration. *Biometrika*, 1998.

Freedman, D. From association to causation: some remarks on the history of statistics. *Journal de la Société Française de Statistique*, 1999.

Glaeser, E. L., Hillis, A., Kominers, S. D., and Luca, M. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review*, 2016.

Hardt, M. and Recht, B. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

Hino, M., Benami, E., and Brooks, N. Machine learning for environmental monitoring. *Nature Sustainability*, 2018.

Johnson, M. S., Levine, D. I., and Toffel, M. W. Improving regulatory effectiveness through better targeting: Evidence from osha. *American Economic Journal: Applied Economics*, 2023.

Johnson, N. and Kotz, S. *Continuous Univariate Distributions*. J. Wiley, 1970.

Kitagawa, T. and Tetenov, A. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 2018.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. Prediction policy problems. *American Economic Review*, 2015.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 2018.

Leventer-Roberts, M., Dagan, N., Berent, J. M., Brufman, I., Hoshen, M., Braun, M., Balicer, R. D., and Feldman, B. S. Using population-level incidence of hepatitis c virus and immigration status for data-driven screening policies: a case study in israel. *Journal of Public Health*, 2022.

Liu, L. T., Barocas, S., Kleinberg, J., and Levy, K. On the actionability of outcome prediction. *arXiv preprint arXiv:2309.04470*, 2023.

Manski, C. F. Statistical treatment rules for heterogeneous populations. *Econometrica*, 2004.

Manski, C. F. Diagnostic testing and treatment under ambiguity: using decision analysis to inform clinical practice. *Proceedings of the National Academy of Sciences*, 2013.

Neyman, J. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 1923.

Noarov, G., Ramalingam, R., Roth, A., and Xie, S. High-dimensional prediction for sequential decision making. *arXiv*, 2023.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Pearl, J. Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, 2013.

Perdomo, J. C., Britton, T., Hardt, M., and Abebe, R. Difficult lessons on social prediction from wisconsin public schools. *arXiv preprint arXiv:2304.06205*, 2023.

Rockoff, J. E., Jacob, B. A., Kane, T. J., and Staiger, D. O. Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, 2011.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974.

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 2020.

Scott Long, J. Regression models for categorical and limited dependent variables. *Advanced Quantitative Techniques in the Social Sciences*, 1997.

Sen. Moynihan, D. P. Emergency unemployment compensation amendments of 1993, 1993. URL https://www.congress.gov/bill/103rd-congress/senate-bill/382.

Slivkins, A., Sankararaman, K. A., and Foster, D. J. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *Conference on Learning Theory*, 2023.

Stock, J. H., Watson, M. W., et al. *Introduction to econometrics*. 2003.

UNICEF. Fighting malaria with long-lasting insecticidal nets, 2022. URL https://shorturl.at/G8VZ3.

U.S Department of Education. Issue brief: Early warning systems, 2016. URL https://www2.ed.gov/rschstat/eval/high-school/early-warning-systems-brief.pdf.

Verma, A., Hanawal, M. K., Rajkumar, A., and Sankaran, R. Censored semi-bandits for resource allocation. *arXiv*, 2021.

Viviano, D. Policy targeting under network interference, 2023.

Waggoner, B. Tight bounds for gaussian tails and hazard rates. https://www.bowaggoner.com/blog/2018/03-17-gaussian-tails/, 2018.

Yule, G. U. An investigation into the causes of changes in pauperism in england, chiefly during the last two intercensal decades. *Journal of the Royal Statistical Society*, 1899.

A. Supporting Arguments for Linear Regression Model

The following result is a general statement regarding the optimal policy for targeting problems that aim to maximize the expected welfare in the population. The result are well-known and have appeared in numerous places (in particular, see Bhattacharya & Dupas (2012) for a stronger statement). We include a derivation here purely for the sake of having a self-contained presentation.

Proposition A.1. *For all distributions $\mathcal{D}$ over (x, w) supported on a discrete set of elements, the optimal policy maximizing the social planner's targeting problem,*

$$\max_{\pi \in \{\mathcal{X}_s \to \{0,1\}\}} \quad \mathbb{E}_{(x_{i,s}, w_i) \sim \mathcal{D}}[w_i \cdot \pi(x_{i,s})]$$

$$\text{subject to} \quad \mathbb{E}_{\mathcal{D}}[\pi(x_{i,s})] \leq \alpha,$$

is equal to,

$$\pi_s^*(x_{i,s}) = \mathbf{1}\{E[w_i \mid x_{i,s}] > F_s^{-1}(1 - \alpha)\} \cdot \mathbf{1}\{E[w_i \mid x_{i,s}] > 0\},$$

where F_s^{-1} is the quantile function for the random variable $\mathbb{E}[w_i \mid X_s]$.

Proof. This proof follows from viewing the optimization problem through the lens of linear programming. We start by expanding out the expectations as sums:

$$\mathbb{E}[w \cdot \pi(x_s)] = \sum_{x_s, w} w \cdot \pi(x_s) \cdot \Pr[x_s, w] = \sum_{x_s} \Pr[x_s]\pi(x_s) \sum_w w \cdot \Pr[w \mid x] = \sum_{x_s} \Pr[x_s]\pi(x_s)\mathbb{E}[w \mid x_s]$$

$$\mathbb{E}[\pi(x_s)] = \sum_x \pi(x_s)\Pr[x_s].$$

Therefore the optimization problem can be equivalently written as maximizing a linear cost function, while satisfying a linear constraint:

$$\max_{\pi \in \{\mathcal{X}_s \to \{0,1\}\}} \quad \sum_{x_s} \pi(x_s)v(x_s)$$

$$\text{subject to} \quad \sum \pi(x_s)c(x_s) \leq \alpha.$$

Here, $v(x_s) = \Pr[x_s]\mathbb{E}[w \mid x_s]$ and $c(x_s) = \Pr[x_s]$. Without loss of generality, we can assume that $\Pr[x_s] > 0$ for all x_s, otherwise the terms can be removed from both sums.

The optimal solution to this program is to set $\pi(x_s) = 1$, for the x_s that maximize the ratio $v(x_s)/c(x_s)$ until the constraint is reached. Intuitively, setting $\pi(x_s) = 1$ "costs" $c(x_s)$ units and "returns" $v(x_s)$ units. In this setting, this ratio is equal to the conditional expectation:

$$\frac{v(x_s)}{c(x_s)} = \mathbb{E}[w|x_s].$$

By definition of the quantile function, assigning $\pi(x_s) = 1$ to the x_s with the top α values of $\mathbb{E}[w \mid x_s]$ is equivalent to choosing

$$\pi(x_s) = \mathbf{1}\{\mathbb{E}[w|X_s] \geq F_s^{-1}(1 - \alpha)\}.$$

Since we can always choose to treat fewer than an α fraction of the units, the second indicator function ensure that each term adds a nonnegative amount to the objective. $\square$

We can specialize this result to the linear regression model:

Lemma A.2. *Assume w_i satisfy the linear regression model (Definition 3.1) and that $x = (x_s, x_t)$. The optimal assignment policy π_s^* solving the planner's targeting problem (Definition 3.2) is equal to:*

$$\pi_s^*(x_{i,s}) = \quad \mathbf{1}\{\langle x_{i,s}, \ \beta_s \rangle \geq \Phi^{-1}(1 - \alpha)\|\beta_s\|_2\} \cdot \mathbf{1}\{\langle x_{i,s}, \ \beta_s \rangle + \mu \geq 0\}.$$

Proof. As per Proposition A.1, the optimal solution to the planner's targeting problem is to intervene on the top α fraction of the population, conditional on those expected welfare effects being positive (see Proposition A.1):

$$\pi_s^*(x_i) = \mathbf{1}\{E[w_i \mid x_s] > F_s^{-1}(1-\alpha)\} \cdot \mathbf{1}\{E[w_i \mid x_s] > 0\}.$$

Here, F_s is the CDF of the random variable $\mathbb{E}[w_i \mid X_s]$, and F_s^{-1} is its quantile function. In the linear model (Definition 3.1), both of these functions have tractable expressions. Since,

$$w_i = \langle x_{i,s}, \ \beta_s \rangle + \langle x_{i,t}, \ \beta_t \rangle + \mu,$$

where $\langle x_{i,s}, \ \beta_s \rangle \sim \mathcal{N}(0, \|\beta_s\|_2^2)$ and $\langle x_{i,t}, \ \beta_t \rangle \sim \mathcal{N}(0, \|\beta_t\|_2^2)$ are independent, then:

$$E[w_i \mid x_{i,s}] = \langle x_{i,s}, \ \beta_s \rangle + \mu, \text{ and } F_s^{-1}(1-\alpha) = \Phi^{-1}(1-\alpha)\|\beta_s\|_2 + \mu.$$

The result follows from plugging these expressions into the first equation. $\qquad\square$

A.1. Value Function for Linear Regression Model: Proof of Proposition 3.4

Since w is Gaussian and we assume that $\alpha < .5$, $\mu > 0$, the condition that $\mathbb{E}[w_i \mid x_{i,s}] \geq F_s^{-1}(1-\alpha)$ implies that $\mathbb{E}[w_i \mid x_{i,s}] > 0$.

Plugging in our expression for π_s^* (Lemma A.2) into the objective function, $\mathbb{E}[w_i \pi_s^*(x_{i,s})]$, and using the functional form for w_i from the linear regression model, we have that,

$$\mathbb{E}[w_i \pi_s^*(x_{i,s})] = \mathbb{E}[(\langle x_{i,s}, \ \beta_s \rangle + \langle x_{i,t}, \ \beta_t \rangle + \mu) \cdot \mathbf{1}\{\langle x_{i,s}, \ \beta_s \rangle \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2\}]$$
$$= \alpha\mu + \mathbb{E}[\langle x_{i,s}, \ \beta_s \rangle \cdot \mathbf{1}\{\langle x_{i,s}, \ \beta_s \rangle \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2\}].$$

Here, we've used the fact that the unobserved $\langle x_{i,t}, \ \beta_t \rangle$ and observed components $\langle x_{i,s}, \ \beta_s \rangle$ are independent to conclude that,

$$\mathbb{E}[\langle x_{i,t}, \ \beta_t \rangle \cdot \mathbf{1}\{\langle x_{i,s}, \ \beta_s \rangle \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2\}] = \mathbb{E}[\langle x_{i,t}, \ \beta_t \rangle]\mathbb{E}[\mathbf{1}\{\langle x_{i,s}, \ \beta_s \rangle \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2\}] = 0,$$

since $\mathbb{E}[\langle x_{i,t}, \ \beta_t \rangle] = 0$. Now, using the general identity, that holds for any random variable Z and event A,

$$\mathbb{E}[Z \cdot \mathbf{1}_A] = \mathbb{E}[Z \mid A]\mathrm{Pr}[A],$$

we have that, for $w_{i,s} = \langle x_{i,s}, \ \beta_s \rangle$, the expectation $\mathbb{E}[w_{i,s} \cdot \mathbf{1}\{w_{i,s} \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2\}]$ is equal to

$$\mathbb{E}[w_{i,s} \mid w_{i,s} \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2] \cdot \mathrm{Pr}[w_{i,s} \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2].$$

By the Mills Ratio identity (Lemma A.4), since $w_{i,s} \sim \mathcal{N}(0, \|\beta_s\|_2^2)$, the first term is equal to,

$$\mathbb{E}[w_{i,s} \mid w_{i,s} \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2] = \|\beta_s\|_2 \frac{\phi\left(\frac{\Phi^{-1}(1-\alpha)\|\beta_s\|_2}{\|\beta_s\|_2}\right)}{1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha)\|\beta_s\|_2}{\|\beta_s\|_2}\right)} = \frac{\|\beta_s\|_2\phi(\Phi^{-1}(1-\alpha))}{\alpha},$$

since $1 - \Phi(\Phi^{-1}(1-\alpha)) = \alpha$. And, by definition of the quantile function,

$$\mathrm{Pr}[w_{i,s} \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2] = \alpha.$$

Therefore,

$$\mathbb{E}[w_{i,s} \mid w_{i,s} \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2] = \|\beta_s\|_2\phi(\Phi^{-1}(1-\alpha)),$$

and

$$\mathbb{E}[w_i \pi_s^*(x_{i,s})] = \alpha\mu + \|\beta_s\|_2\phi(\Phi^{-1}(1-\alpha)).$$

The exact expression follows by substituting $\|\beta_s\|_2 = \gamma_s\|\beta\|_2$ (Definition 3.3).

A.2. Prediction Access Ratio for Linear Regression: Proof of Theorem 3.5

Recall that the goal is to prove upper and lower bounds on the ratio,

$$\frac{V_*^{\text{lin}}(\alpha + \Delta_\alpha, \gamma_s) - V_*^{\text{lin}}(\alpha, \gamma_s)}{V_*^{\text{lin}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\text{lin}}(\alpha, \gamma_s)}.$$

To do so, we deal with the numerator and the denominator separately.

Quantifying Prediction Improvement.   We start with the denominator. This part is relatively simple because the value function is linear in γ_s. Recall from Proposition 3.4 that:

$$V_*^{\text{lin}}(\alpha, \gamma_s) = \alpha\mu + \gamma_s \|\beta\|_2 \phi(\Phi^{-1}(1 - \alpha)).$$

Therefore,

$$V_*^{\text{lin}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\text{lin}}(\alpha, \gamma_s) = \Delta_{r^2} \|\beta\|_2 \phi(\Phi^{-1}(1 - \alpha)).$$

Now, if we apply Lemma A.7 and our assumption on α, we get that,

$$\Delta_{r^2}\|\beta\|_2 \alpha \Phi^{-1}(1-\alpha) \le V_*^{\text{lin}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\text{lin}}(\alpha, \gamma_s) \le \Delta_{r^2}\|\beta\|_2 \alpha \Phi^{-1}(1-\alpha)(1 + f(\alpha)), \qquad (13)$$

where $f(\alpha) = \Phi^{-1}(1-\alpha)^2/(\Phi^{-1}(1-\alpha)^2 - 1) - 1$ is $o(1)$ and less than 1 for $\alpha < .05$.

Quantifying Access Improvement.   Using the closed form expression for the value function, we get that:

$$V_*^{\text{lin}}(\alpha + \Delta_\alpha, \gamma_s) - V_*^{\text{lin}}(\alpha, \gamma_s) = \Delta_\alpha \cdot \mu + \gamma_s \|\beta\|_2 \left(\phi(\Phi^{-1}(1 - \alpha - \Delta_\alpha)) - \phi(\Phi^{-1}(1 - \alpha)) \right).$$

Applying Lemma A.3, we get that:

$$\frac{1}{2}\Delta_\alpha \Phi^{-1}(1 - \alpha) \le \phi(\Phi^{-1}(1 - \alpha - \Delta_\alpha)) - \phi(\Phi^{-1}(1 - \alpha)) \le \Delta_\alpha \Phi^{-1}(1 - \alpha).$$

Therefore,

$$\Delta_\alpha(\mu + \frac{1}{2}\gamma_s\|\beta\|_2\Phi^{-1}(1-\alpha)) \le V_*^{\text{lin}}(\alpha + \Delta_\alpha, \gamma_s) - V_*^{\text{lin}}(\alpha, \gamma_s) \le \Delta_\alpha(\mu + \gamma_s\|\beta\|_2\Phi^{-1}(1-\alpha)). \qquad (14)$$

The statement follows from combining the bounds in Equation (14) and Equation (13).

Lemma A.3. *Define* $g(\alpha) = \phi(\Phi^{-1}(1 - \alpha))$, *then for* $\alpha + \Delta < .05$ *and* $\Delta \le 4\alpha$,

$$\Delta\Phi^{-1}(1 - \alpha) \ge g(\alpha + \Delta) - g(\alpha) \ge \frac{1}{2}\Phi^{-1}(1 - \alpha)\Delta.$$

Proof. From Lemma A.5, we have that

$$g'(\alpha) = \Phi^{-1}(1 - \alpha) \text{ and } g''(\alpha) = \frac{-1}{\phi(\Phi^{-1}(1 - \alpha))}.$$

By Taylor's theorem, there exists a value $c \in [0, 1]$ such that:

$$g(\alpha + \Delta) - g(\alpha) = g'(\alpha)\Delta + \frac{1}{2}\Delta^2 g''(\alpha + c\Delta)$$

$$= \Delta\Phi^{-1}(1 - \alpha) - \frac{1}{2}\Delta^2 \frac{1}{\phi(\Phi^{-1}(1 - \alpha - c\Delta))}$$

$$\le \Delta\Phi^{-1}(1 - \alpha).$$

To prove the lower bound on $g(\alpha + \Delta) - g(\alpha)$ it suffices to establish that

$$\frac{1}{2}\Delta^2 \frac{1}{\phi(\Phi^{-1}(1 - \alpha - c\Delta))} \le \frac{1}{2}\Delta\Phi^{-1}(1 - \alpha). \qquad (15)$$

Or equivalently, that $\Delta \leq \Phi^{-1}(1-\alpha)\phi(\Phi^{-1}(1-\alpha-c\Delta))$. To show this, we know that by Lemma A.7, for $\alpha + \Delta < .15$,

$$\phi(\Phi^{-1}(1-\alpha-c\Delta)) \geq (\alpha+c\Delta)\Phi^{-1}(1-\alpha-c\Delta) \geq \sqrt{2}\alpha.$$

The second inequality follows from the fact that $\Phi^{-1}(1-a) \geq \sqrt{2}$ for all $a < .05$ and that $c\Delta > 0$. Therefore, in order for,

$$\Delta \leq \Phi^{-1}(1-\alpha)\phi(\Phi^{-1}(1-\alpha-c\Delta)),$$

it suffices for $\Delta \leq 4\alpha$. This ensures that the inequality in Equation (15) is true and concludes the proof of the lower bound. $\qquad\square$

A.3. Supporting Technical Lemmas

Lemma A.4 (Inverse Mills Ratio). *Let $z \sim \mathcal{N}(\mu, \sigma^2)$, then*

$$\mathbb{E}[z \mid z > a] = \mu + \sigma\frac{\phi(a)}{1-\Phi(a)}.$$

Proof. This is a well known property of Gaussians, see e.g (Johnson & Kotz, 1970). $\qquad\square$

Lemma A.5.

$$\frac{\partial}{\partial\alpha}\Phi^{-1}(1-\alpha) = -\frac{1}{\phi(\Phi^{-1}(1-\alpha))}$$

$$\frac{\partial}{\partial\alpha}\phi(\Phi^{-1}(1-\alpha)) = \Phi^{-1}(1-\alpha)$$

Proof. These identities are also standard, we include a proof for the sake of completeness. By definition of inverse functions,

$$\frac{\partial}{\partial x}f^{-1}(f(x)) = \frac{\partial}{\partial x}x = 1,$$

and by the chain rule,

$$\frac{\partial}{\partial x}f^{-1}(f(x)) = f'(f^{-1}(x))(f^{-1})'(x).$$

Combining these two we get that:

$$\frac{\partial}{\partial x}f^{-1}(x) = \frac{1}{f'(f(x))} \quad \text{and} \quad \frac{\partial}{\partial x}f^{-1}(1-x) = -\frac{1}{f'(f(x))}.$$

Since $\Phi' = \phi$ (Φ is the CDF and ϕ is a PDF), then we can use the second identity above to conclude that:

$$\frac{\partial}{\partial\alpha}\Phi^{-1}(1-\alpha) = -\frac{1}{\phi(\Phi^{-1}(1-\alpha))}.$$

For the second calculation, we use the fact that $\phi'(x) = -x\phi(x)$ as well as the result from the first part. By the chain rule,

$$\frac{\partial}{\partial\alpha}\phi(\Phi^{-1}(1-\alpha)) = -\Phi^{-1}(1-\alpha)\phi(\Phi^{-1}(1-\alpha))\left(\frac{\partial}{\partial\alpha}\Phi^{-1}(1-\alpha)\right)$$

$$= \Phi^{-1}(1-\alpha).$$

$\qquad\square$

Lemma A.6. *Let $z \sim \mathcal{N}(0,1)$, then, for all $t > 0$:*

$$\frac{\phi(t)}{t}(1-t^{-2}) \leq \Pr(z \geq t) \leq \frac{\phi(t)}{t}.$$

Proof. These bounds appear in numerous sources, this particular result is drawn from a writeup from Bo Waggoner (Waggoner, 2018). □

Lemma A.7. *For all $\alpha < .15$,*

$$\alpha \Phi^{-1}(1 - \alpha) \leq \phi(\Phi^{-1}(1 - \alpha)) \leq \Phi^{-1}(1 - \alpha)\alpha(1 + f(\alpha)),$$

where $f(\alpha) = \Phi^{-1}(1-\alpha)^2/(\Phi^{-1}(1-\alpha)^2 - 1) - 1$ is o(1) and less than 1 for all $\alpha < .05$.

Proof. The main idea is to use Lemma A.6 and set $t = \Phi^{-1}(1 - \alpha)$. By definition of the quantile function, if z is a standard normal:

$$\Pr(z \geq \Phi^{-1}(1 - \alpha)) = \alpha.$$

From our assumption on α, $t = \Phi^{-1}(1 - \alpha) > 0$ and hence applying Lemma A.6 we get that,

$$\frac{\phi(\Phi^{-1}(1 - \alpha))}{\Phi^{-1}(1 - \alpha)} \left(1 - \frac{1}{\Phi^{-1}(1 - \alpha)^2}\right) \leq \alpha \leq \frac{\phi(\Phi^{-1}(1 - \alpha))}{\Phi^{-1}(1 - \alpha)}.$$

The final statement follows from rearranging these inequalities. □

Proposition A.8.

$$\Phi^{-1}(1 - \alpha) = \Theta(\sqrt{\log(1/\alpha)}) \text{ as } \alpha \to 0$$

Proof. From the definition of the quantile function, we have that

$$\Phi^{-1}(1 - \alpha) = \arg\max_t \Pr(z \geq t) \leq \alpha.$$

From the Lemma A.6, for $t \geq \sqrt{2}$,

$$\frac{\phi(t)}{2t} \leq \Pr[z \geq t] \leq \frac{\phi(t)}{t}.$$

Therefore, $\Phi^{-1}(1 - \alpha) \geq t_1$ where t_1 solves,

$$\frac{\phi(t_1)}{t_1} = \alpha,$$

and $\Phi^{-1}(1 - \alpha) \leq t_2$ where t_2 solves,

$$\frac{\phi(t_2)}{2t_2} = \alpha.$$

Using the definition of ϕ, it suffices to solve for the values of t that solve the equation,

$$\frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\frac{1}{c_0 t} = \alpha,$$

for $c_0 \in \{1, 2\}$. Moving everything into logs, this becomes

$$\frac{t^2}{2} - \log(\frac{1}{\sqrt{2\pi}}) - \log(\frac{1}{ct}) = \log(1/\alpha).$$

For small values of α, the t^2 term dominates and we get that

$$t \approx \sqrt{\log(1/\alpha)}.$$

□

B. Supporting Arguments for Probit Regression Model

B.1. Optimal Policy for the Probit Model

Lemma B.1. *Assume w_i satisfy the probit regression model (Definition 4.1) and that $x = (x_s, x_t)$. The optimal assignment policy π_s^* for the planner's targeting problem (Definition 3.2) is equal to:*

$$\pi_s^*(x_{i,s}) = \mathbf{1}\{\langle x_{i,s}, \beta_s \rangle \geq \Phi^{-1}(1-\alpha)\|\beta_s\|_2\}. \tag{16}$$

Proof. Since $\mathbb{E}[w_i \mid x_{i,s}]$ is always greater than 0, the optimal policy is simply to treat the units with the highest values of $\mathbb{E}[w_i \mid x_{i,s}]$. In this case,

$$\mathbb{E}[w_i \mid x_{i,s}] = \Pr[w_i = 1 \mid x_{i,s}] = \Phi(\mu + \langle x_{i,s}, \beta_s \rangle).$$

Because $\Phi(z)$ is strictly monotonic in its argument, $z = \mu + \langle x_{i,s}, \beta_s \rangle$, the top $1 - \alpha$ quantile for $\Phi(z)$ is the same as the top $1 - \alpha$ quantile for z itself. $\qquad\square$

B.2. Value Function for Probit Regression: Proof of Lemma B.1

Using the lemma regarding the functional form for the optimal targeting policy (Lemma B.1), and the definition of w_i from the probit model, we can expand out the value of the optimal policy as follows:

$$\mathbb{E}[w_i \pi_s^*(x_s)] = = \mathbb{E}[\mathbf{1}\{x_s^\top \beta_s + x_t^\top \beta_t + \mu > 0\} \cdot \mathbf{1}\{x_s^\top \beta_s + \mu \geq \mu + \|\beta_s\|_2 \Phi^{-1}(1-\alpha)\}$$
$$= \mathbb{E}[\mathbf{1}\{x_s^\top \beta_s + x_t^\top \beta_t + \mu > 0\} \cdot \mathbf{1}\{x_s^\top \beta_s \geq \|\beta_s\|_2 \Phi^{-1}(1-\alpha)\}. \tag{17}$$

Using the fact that the random variables $x_s^\top \beta_s \sim \mathcal{N}(0, \|\beta_s\|_2^2)$ and $x_t^\top \beta_t \sim \mathcal{N}(0, \|\beta_t\|_2^2)$ are independent and Gaussian, we can rewrite these as,

$$x_s^\top \beta_s = \|\beta_s\|_2 z_s \text{ and } x_t^\top \beta_t = \|\beta_t\|_2 z_t,$$

where $z_s, z_t \sim \mathcal{N}(0, 1)$ are i.i.d. Adopting this notation, we can rewrite the previous expression for the value function in Equation (17) as:

$$= \mathbb{E}[\mathbf{1}\{\|\beta_s\|_2 z_s + \|\beta_t\|_2 z_t + \mu > 0\} \cdot \mathbf{1}\{\|\beta_s\|_2 z_s \geq \|\beta_s\|_2 \Phi^{-1}(1-\alpha)\}]$$
$$= \mathbb{E}[\mathbf{1}\{\|\beta_s\|_2 z_s + \|\beta_t\|_2 z_t + \mu > 0\} \cdot \mathbf{1}\{z_s \geq \Phi^{-1}(1-\alpha)\}]$$
$$= \Pr[\mathbf{1}\{\|\beta_s\|_2 z_s + \|\beta_t\|_2 z_t + \mu > 0, \ z_s \geq \Phi^{-1}(1-\alpha)\}]$$
$$= \Pr[z_t \geq \frac{-\|\beta_s\|_2 z_s - \mu}{\|\beta_t\|_2}, \ z_s \geq \Phi^{-1}(1-\alpha)]$$
$$= \int_{\Phi^{-1}(1-\alpha)}^{\infty} \int_{\frac{-\|\beta_s\|_2 z_s - \mu}{\|\beta_t\|_2}}^{\infty} \phi(z_t)\phi(z_s) dz_t dz_s$$
$$= \int_{\Phi^{-1}(1-\alpha)}^{\infty} \left(1 - \Phi\left(\frac{-\|\beta_s\|_2 z_s - \mu}{\|\beta_t\|_2}\right)\right) \phi(z_s).$$

Here, we've again used ϕ denote the pdf for a standard Gaussian. Using the identity, $1 - \Phi(-a) = \Phi(a)$, this last expression is equal to the following integral:

$$\int_{\Phi^{-1}(1-\alpha)}^{\infty} \Phi\left(\frac{\|\beta_s\|_2 z_s + \mu}{\|\beta_t\|_2}\right) \phi(z_s) dz_s.$$

The exact statement follows from substituting $\|\beta_s\|_2 = \gamma_s \|\beta\|_2$ and $\|\beta_t\|_2 = \gamma_t \|\beta\|_2$.

B.3. Prediction Access Ratio for Probit Regression: Proof of Theorem 4.4

Recall from the definition of the probit model that,

$$\Pr[\langle x, \beta \rangle + \mu > 0] = b \in (0, 1),$$

and hence, $\mu \|\beta\|_2^{-1} = \Phi^{-1}(b) = -\Phi^{-1}(1 - b)$.

Quantifying Improvements by Expanding Access. By Taylor's theorem, there exists some $c \in (0,1)$ such that:

$$V_*^{\mathrm{pr}}(\alpha + \Delta_\alpha, \gamma_s) - V_*^{\mathrm{pr}}(\alpha, \gamma_s) = \Delta_\alpha \cdot \frac{\partial}{\partial \alpha} V_*^{\mathrm{pr}}(\alpha + c\Delta_\alpha, \gamma_s).$$

The last term is the derivative of the value function evaluated at the point $(\alpha + c\Delta_\alpha, \gamma_s)$. From, Lemma B.3, we know that

$$\frac{\partial}{\partial \alpha} V_*^{\mathrm{pr}}(\alpha + c\Delta_\alpha, \gamma_s) = \Phi\left(\frac{\gamma_s \Phi^{-1}(1 - \alpha - c\Delta_\alpha) + \mu \|\beta\|_2^{-1}}{\gamma_t} \right).$$

Using the identity, $\mu \|\beta\|_2^{-1} = -\Phi^{-1}(1-b)$, we can rewrite this as

$$\Phi\left(\frac{\gamma_s \Phi^{-1}(1 - \alpha - c\Delta_\alpha) + \mu \|\beta\|_2^{-1}}{\gamma_t} \right) = \Phi\left(\frac{1}{\gamma_t}\left(\gamma_s \Phi^{-1}(1 - \alpha - c\Delta_\alpha) - \Phi^{-1}(1-b) \right) \right). \tag{18}$$

Assuming that $\Delta_\alpha \leq \alpha$ and that α is small enough so that $\gamma_s \Phi^{-1}(1 - 2\alpha) \geq \Phi^{-1}(1-b)$, then,

$$\gamma_s \Phi^{-1}(1 - \alpha - c\Delta_\alpha) \geq \gamma_s \Phi^{-1}(1 - 2\alpha) \geq \Phi^{-1}(1-b),$$

and we have that the expression in Equation (18) is of the form $\Phi(x)$ for some $x \geq 0$. Therefore, by properties of the Gaussian CDF,

$$\frac{\partial}{\partial \alpha} V_*^{\mathrm{pr}}(\alpha + c\Delta_\alpha, \gamma_s) \in [1/2, 1].$$

Hence,

$$\frac{1}{2}\Delta_\alpha \leq V_*^{\mathrm{pr}}(\alpha + \Delta_\alpha, \gamma_s) - V_*^{\mathrm{pr}}(\alpha, \gamma_s) \leq \Delta_\alpha. \tag{19}$$

Quantifying Improvements via Prediction. We use the same strategy as in the previous part. By Taylor's theorem, there exists some $c \in (0,1)$ such that:

$$V_*^{\mathrm{pr}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\mathrm{pr}}(\alpha, \gamma_s) = \Delta_{r^2} \cdot \frac{\partial}{\partial \gamma_s} V_*^{\mathrm{pr}}(\alpha, \gamma_s + c\Delta_{r^2}).$$

Now, let $\gamma_{s,c} = \gamma_s + c\Delta_{r^2}$ and $\gamma_{t,c} = \sqrt{1 - \gamma_{s,c}^2}$. Then, by Lemma B.4,

$$\frac{\partial}{\partial \gamma_s} V_*^{\mathrm{pr}}(\alpha, \gamma_s + c\Delta_{r^2}) = \frac{1}{\gamma_{t,c}} \phi\left(\frac{\mu}{\|\beta\|_2} \right) \phi\left(\frac{\Phi^{-1}(1-\alpha) + \mu \|\beta\|_2^{-1} \gamma_{s,c}}{\gamma_{t,c}} \right)$$

$$= \underbrace{\frac{1}{\gamma_{t,c}} \phi\left(\frac{\mu}{\|\beta\|_2} \right)}_{T_1} \underbrace{\phi\left(\frac{\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-b)\gamma_{s,c}}{\gamma_{t,c}} \right)}_{T_2}.$$

We start by analyzing bounds on T_2. For any $\varepsilon_1 > 0$, there exists a value $t_1(\varepsilon_1, \gamma_t)$, depending on ε_1 and γ_t, such that for all $\Delta_{r^2} \leq t_1(\varepsilon_1, \gamma_t)$, we have that $.5\gamma_t \leq (1 - \varepsilon_1)\gamma_t \leq \gamma_{t,c} \leq \gamma_t$. Assuming this condition holds, since $\gamma_{s,c}\Phi^{-1}(1-b) > 0$, we get that:

$$\frac{\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-b)\gamma_{s,c}}{\gamma_{t,c}} \leq \Phi^{-1}(1-\alpha)\frac{1}{(1 - \varepsilon_1)\gamma_t}.$$

Furthermore, since $\gamma_{t,c} \leq \gamma_t$ and $\gamma_{s,c} \leq 1$, for any fixed ε_2, there exists a value $t_2(b, \gamma_t)$, depending on γ_t and b, such that for all $\alpha \leq t_2(b, \gamma_t, \varepsilon_2)$:

$$\frac{\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-b)\gamma_{s,c}}{\gamma_{t,c}} \geq \frac{\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-b)}{\gamma_{t,c}} \geq \frac{\Phi^{-1}(1-\alpha)}{\gamma_t}(1 - \varepsilon_2).$$

Using these last two lines, and the fact that $\phi(x') \le \phi(x)$ for $0 < x \le x'$, we get that

$$\phi\left(\frac{1}{(1-\varepsilon_3)}\frac{\Phi^{-1}(1-\alpha)}{\gamma_t}\right) \le \phi\left(\frac{\Phi^{-1}(1-\alpha) - \Phi^{-1}(1-b)\gamma_{s,c}}{\gamma_{t,c}}\right) \le \phi\left(\frac{\Phi^{-1}(1-\alpha)}{\gamma_t}(1-\varepsilon_3)\right).$$

where $\varepsilon_3 = \max\{\varepsilon_1, \varepsilon_2\}$. Now, by Lemma B.6, for α smaller than some value $t_3(\alpha, \varepsilon_3) > 0$,

$$T_2 \le \phi\left(\frac{\Phi^{-1}(1-\alpha)}{\gamma_t}(1-\varepsilon_3)\right) \le \frac{1}{\sqrt{2\pi}}\left(1.01\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{\frac{1}{\gamma_t^2}(1-\varepsilon_3)^2}.$$

And,

$$T_2 \ge \phi\left(\frac{1}{(1-\varepsilon_3)}\frac{\Phi^{-1}(1-\alpha)}{\gamma_t}\right) \ge \frac{1}{\sqrt{2\pi}}\left(\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{\frac{(1-\varepsilon_3)^{-2}}{\gamma_t^2}}$$

$$= \frac{1}{\sqrt{2\pi}}\left(\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{\frac{1}{\gamma_t^2}(1+\varepsilon_4)^2},$$

where we have rewritten the last expression in a more convenient form by letting $\varepsilon_4 = \varepsilon_3/(1-\varepsilon_3)$.

Analyzing T_1 is simple. Recall that $\mu\|\beta\|_2^{-1} = \Phi^{-1}(1-b)$ where $b = \Pr[w_i > 0]$. Hence,

$$\phi\left(\frac{\mu}{\|\beta\|_2}\right) = \phi(-\Phi^{-1}(1-b)) = \phi(\Phi^{-1}(1-b)).$$

Then for all $b < .15$, Lemma A.7 ensures that:

$$b\Phi^{-1}(1-b) \le \phi\left(\frac{\mu}{\|\beta\|_2}\right) \le 2b\Phi^{-1}(1-b).$$

Furthermore, by our initial calculation, we can set $\varepsilon_1 < .5$ so that $.5\gamma_t \le \gamma_{t,c} \le \gamma_t$. and hence:

$$\frac{1}{\gamma_t} \le \frac{1}{\gamma_{t,c}} \le \frac{2}{\gamma_t}.$$

This implies that T_1 satisfies the bounds:

$$b\Phi^{-1}(1-b) \le T_1 \le 4\frac{1}{\gamma_t}b\Phi^{-1}(1-b).$$

Combining our bounds for T_1 and T_2, we get that for appropriately small b, α, Δ_{r^2},

$$V_*^{\mathrm{pr}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\mathrm{pr}}(\alpha, \gamma_s) \le \Delta_{r^2}\frac{2}{\gamma_t}\sqrt{\frac{2}{\pi}} \cdot b\Phi^{-1}(1-b) \cdot \left(1.01\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{\frac{1}{\gamma_t^2}(1-\varepsilon_3)^2}$$

$$V_*^{\mathrm{pr}}(\alpha, \gamma_s + \Delta_{r^2}) - V_*^{\mathrm{pr}}(\alpha, \gamma_s) \ge \Delta_{r^2}\frac{1}{\gamma_t}\frac{1}{\sqrt{2\pi}} \cdot b\Phi^{-1}(1-b)\left(\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{\frac{1}{\gamma_t^2}(1+\varepsilon_4)^2}.$$

The final statement comes from combining these inequalities with those in Equation (19) and simplifying the constants.

B.4. Supporting Technical Lemmas

Lemma B.2 (Leibniz's Rule)**.**

$$\frac{d}{dx}\int_{a(x)}^{b(x)} f(x,t)\,dt = f(x,b(x))\cdot\frac{d}{dx}b(x) - f(x,a(x))\cdot\frac{d}{dx}a(x) + \int_{a(x)}^{b(x)}\frac{\partial}{\partial x}f(x,t)\,dt$$

Lemma B.3.

$$\frac{\partial}{\partial\alpha}V_*^{\mathrm{pr}}(\alpha, \gamma_s) = \Phi\left(\frac{\gamma_s\Phi^{-1}(1-\alpha) + \mu\|\beta\|_2^{-1}}{\gamma_t}\right)$$

19

Proof. By Leibniz's Rule (Lemma B.2), the derivative has only one nonzero term:

$$\frac{\partial}{\partial \alpha} \int_{\Phi^{-1}(1-\alpha)}^{\infty} \Phi\left(\frac{\gamma_s z_s + \mu\|\beta\|_2^{-1}}{\gamma_t}\right) \phi(z_s) dz_s = -\Phi\left(\frac{\gamma_s \Phi^{-1}(1-\alpha) + \mu\|\beta\|_2^{-1}}{\gamma_t}\right) \phi(\Phi^{-1}(1-\alpha)) \left(\frac{\partial}{\partial \alpha}\Phi^{-1}(1-\alpha)\right)$$

To finish, we can apply Lemma A.5, to evaluate,

$$\frac{\partial}{\partial \alpha}\Phi^{-1}(1-\alpha) = -(\phi(\Phi^{-1}(1-\alpha)))^{-1},$$

and then simplify. □

Lemma B.4.

$$\frac{\partial}{\partial \gamma_s} V_*^{\mathrm{pr}}(\alpha, \gamma_s) = \frac{1}{\gamma_t}\phi\left(\frac{\mu}{\|\beta\|_2}\right)\phi\left(\frac{\Phi^{-1}(1-\alpha) + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right)$$

Proof. Starting with the chain rule, we have that,

$$\frac{\partial}{\partial \gamma_s} \int_{\Phi^{-1}(1-\alpha)}^{\infty} \Phi\left(\frac{\gamma_s z_s + \mu\|\beta\|_2^{-1}}{\gamma_t}\right) \phi(z_s) dz_s,$$

is equal to:

$$\int_{\Phi^{-1}(1-\alpha)}^{\infty} \phi\left(\frac{\gamma_s z_s + \mu\|\beta^{-1}\|_2}{\gamma_t}\right) \phi(z_s) \left(z_s \cdot \frac{\partial}{\partial \gamma_s}\left(\frac{\gamma_s}{\gamma_t}\right) + \frac{\partial}{\partial \gamma_s}\left(\frac{\mu\|\beta\|_2^{-1}}{\gamma_t}\right)\right). \tag{20}$$

Now, we can calculate the remaining derivatives explicitly,

$$\frac{\partial}{\partial \gamma_s}\left(\frac{\gamma_s}{\gamma_t}\right) = \frac{\partial}{\partial \gamma_s}\left(\frac{\gamma_s}{\sqrt{1-\gamma_s^2}}\right) = \frac{1}{(1-\gamma_s^2)^{3/2}} = \frac{1}{\gamma_t^3}$$

$$\frac{\partial}{\partial \gamma_s}\left(\frac{\mu\|\beta\|_2^{-1}}{\gamma_t}\right) = \frac{\partial}{\partial \gamma_s}\left(\frac{\mu\|\beta\|_2^{-1}}{\sqrt{1-\gamma_s^2}}\right) = \frac{\mu\|\beta\|_2^{-1}\gamma_s}{(1-\gamma_s^2)^{3/2}} = \frac{\mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t^3},$$

and we can rewrite the product of pdfs as:

$$\phi(\frac{\gamma_s z_s + \mu\|\beta\|_2^{-1}}{\gamma_t})\phi(z_s) = \phi\left(\frac{\mu}{\|\beta\|_2}\right)\phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right).$$

Going back to the initial expression in Equation (20) and substituting in these last three identities, we have that the derivative can be written as:

$$\frac{1}{\gamma_t^3}\phi\left(\frac{\mu}{\|\beta\|_2}\right) \int_{\Phi^{-1}(1-\alpha)}^{\infty} \phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right)\left(z_s + \mu\|\beta\|_2^{-1}\gamma_s\right) dz_s. \tag{21}$$

These integrals we can indeed evaluate:

$$\int_{\Phi^{-1}(1-\alpha)}^{\infty} \phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right) dz_s = \gamma_s \cdot \Phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right)\Bigg|_{\Phi^{-1}(1-\alpha)}^{\infty}$$

$$= \gamma_t\left(1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha) + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right)\right).$$

And,

$$\int_{\Phi^{-1}(1-\alpha)}^{\infty} \phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right) z_s dz_s = -\gamma_t^2\left[\phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right) + \frac{\mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\Phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right)\right]\Bigg|_{\Phi^{-1}(1-\alpha)}^{\infty},$$

is equal to:

$$-\gamma_t^2 \left[-\phi\left(\frac{\Phi^{-1}(1-\alpha) + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t} \right) + \frac{\mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\left(1 - \Phi\left(\frac{\Phi^{-1}(1-\alpha) + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right)\right) \right].$$

Plugging in the solutions to the integrals back into (21), we get that the final expression is:

$$\frac{1}{\gamma_t}\phi\left(\frac{\mu}{\|\beta\|_2}\right)\phi\left(\frac{\Phi^{-1}(1-\alpha) + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right).$$

$\square$

Lemma B.5. *Assume that $\gamma_s^2 + \gamma_t^2 = 1$ and that γ_s, γ_t and both in $(0, 1)$. Then,*

$$\phi\left(\frac{\gamma_s z_s + \mu\|\beta\|_2^{-1}}{\gamma_t}\right)\phi(z_s) = \phi\left(\frac{\mu}{\|\beta\|_2}\right)\phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right).$$

Proof. The main idea is to expand out the exponents and complete the square:

$$
\begin{aligned}
\phi\left(\frac{\gamma_s z_s + \mu\|\beta\|_2^{-1}}{\gamma_t}\right)\phi(z_s) &= \frac{1}{2\pi}\exp\left(-\frac{(\gamma_s z_s + \mu\|\beta\|_2^{-1})^2}{2\gamma_t^2} - \frac{z_s^2}{2}\right)\\
&= \frac{1}{2\pi}\exp\left(-\frac{((\gamma_s^2 + \gamma_t^2)z_s^2 + \mu^2\|\beta\|_2^{-2} + 2\mu\|\beta\|_2^{-1}\gamma_s z_s)}{2\gamma_t^2}\right)\\
&= \frac{1}{2\pi}\exp\left(-\frac{\mu^2\|\beta\|_2^{-2}}{2\gamma_t^2}\right)\exp\left(-\frac{(z_s^2 + 2\mu\|\beta\|_2^{-1}\gamma_s z_s \pm \mu^2\|\beta\|_2^{-1}\gamma_s^2)}{2\sigma_t^2}\right)\\
&= \frac{1}{2\pi}\exp\left(-\frac{\mu^2\|\beta\|_2^{-2}}{2\gamma_t^2}\right)\exp\left(\frac{-(z_s + \mu\|\beta\|_2^{-1}\gamma_s)^2 + \mu^2\|\beta\|_2^{-2}\gamma_s^2)}{2\gamma_t^2}\right)\\
&= \frac{1}{2\pi}\exp\left(-\frac{\mu^2\|\beta\|_2^{-2}}{2\gamma_t^2}\right)\exp\left(-\frac{1}{2}\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right)^2\right)\exp\left(\frac{\mu^2\|\beta\|_2^{-2}\gamma_s^2}{2\gamma_t^2}\right).
\end{aligned}
$$

In the third line, we used the fact that $1 = \gamma_t^2 + \gamma_s^2$. Using this same identity, we can further simplify two of the remaining terms as:

$$\exp\left(-\frac{\mu^2\|\beta\|_2^{-2}}{2\gamma_t^2}\right)\exp\left(\frac{\mu^2\|\beta\|_2^{-2}\gamma_s^2}{2\gamma_t^2}\right) = \exp\left(\frac{-\mu\|\beta\|_2^{-2}(1 - \gamma_s^2)}{\gamma_t^2}\right) = \exp\left(-\frac{\mu^2}{\|\beta\|_2^2}\right).$$

Hence,

$$\phi\left(\frac{\gamma_s z_s + \mu\|\beta\|_2^{-1}}{\gamma_t}\right)\phi(z_s) = \phi\left(\frac{\mu}{\|\beta\|_2}\right)\phi\left(\frac{z_s + \mu\|\beta\|_2^{-1}\gamma_s}{\gamma_t}\right).$$

$\square$

Lemma B.6. *For any $k, \varepsilon > 0$ there exists a value $t(\varepsilon) > 0$, that depends only ε, such that for all $\alpha \leq t(\varepsilon)$, the followings inequalities hold:*

$$\frac{1}{\sqrt{2\pi}}(\sqrt{2\pi}\alpha)^{k^2}\left(\Phi^{-1}(1-\alpha)\right)^{k^2} \leq \phi(k \cdot \Phi^{-1}(1-\alpha)) \leq \frac{1}{\sqrt{2\pi}}\left((1+\varepsilon)\cdot\sqrt{2\pi}\alpha\Phi^{-1}(1-\alpha)\right)^{k^2}$$

Proof. The proof is just a direct calculation that follows from expanding out the definition of the Gaussian PDF $\phi(\cdot)$, and then applying Lemma A.7:

$$
\begin{aligned}
\phi(k \cdot \Phi^{-1}(1-\alpha)) &= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-k^2(\Phi^{-1}(1-\alpha))^2}{2}\right)\\
&= (2\pi)^{\frac{k^2-1}{2}}\left[\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(\Phi^{-1}(1-\alpha))^2}{2}\right)\right]^{k^2}\\
&= (2\pi)^{\frac{k^2-1}{2}}\left[\phi(\Phi^{-1}(1-\alpha))\right]^{k^2}.
\end{aligned}
$$

Moving onto the second part, for α small enough, we have that by Lemma A.7,

$$\alpha\Phi^{-1}(1-\alpha) \leq \phi(\Phi^{-1}(1-\alpha)) \leq \alpha\Phi^{-1}(1-\alpha)(1+\varepsilon).$$

Therefore,

$$(2\pi)^{\frac{k^2-1}{2}} \left[\alpha \cdot \Phi^{-1}(1-\alpha)\right]^{k^2} \leq \phi(k \cdot \Phi^{-1}(1-\alpha)) \leq (2\pi)^{\frac{k^2-1}{2}} \left[(1+\varepsilon)\alpha \cdot \Phi^{-1}(1-\alpha)\right]^{k^2}.$$

The precise inequality then follows from simplifying this expression above. $\square$

C. Simulation Details

For the visuaizations in Figure 1 and Figure 2, we compute the prediction access ratios numerically. Using our closed form expressions regarding the value functions V_*^{lin} and V_*^{pr} from Proposition 3.4 and Proposition 4.3 we compute the cost benefit ratios with $\Delta_\alpha = \Delta_{r^2} = .01$ and $\mu = 1$, $\beta = 10$ for the linear case. For the probit case, we set $b = .1$ and $\Delta_\alpha = \Delta_{r^2} = 1e - 3$. The simulation code is included in the submission.