# THE PRICE OF PREDICTION ALGORITHMS

ANDY HAUPT

Abstract. An agent would like predict a quantity from characteristics. She can purchase a prediction algorithm from a monopolist. Her willingness to pay for predictions is heterogenous for different combinations of characteristics. The monopolist is constrained by a prediction technology using samples from an underlying distribution. We relate revenue maximising menus to optimal menus for the multi-goods monopolist problem. We give evidence that more flexible prediction technologies have ambiguous effects on the designer's revenue if the number of samples used in the technology is fixed. We use this to highlight the tradeoff between quality and specifiability, which we introduce.

## 1. Introduction

As a byproduct of more and more machine-readable data, prediction algorithms generate high societal value. In computer vision, prediction algorithms are used to recognize street signs for autonomous vehicles; and in natural language processing, topics are automatically extracted from texts.

An important feature of these algorithms is their transferability. For example, a natural language algorithm trained to recognize a class of cursewords can work well on the task to find subjects in a sentence. Techniques relating to such cross-usage are known by the term *transfer learning*.

Given the value generated through prediction algorithms and the cost of their creation, transactions involving them become important. In this paper, we study pricing of prediction algorithms in the presence of transfer learning.[1]

Transfer learning and incentives shape invention in fields like healthcare. The NEWDIGS initiative at MIT works for a consortium of pharmaceutical and health insurance companies to create a marketplace for prediction algorithms that recommend drugs for patients. Different buyers of such prediction algorithms differ considerably in their interest in different patient sub-populations. A pharmaceutical company working on a new drug, for

*Date*: May 13, 2020.

JEL *Classification.* D47, D83.

*Key words and phrases.* market design, information design, value of information.

[1]A researcher from the MIT Quest for Intelligence noted: "In AI, we do not have clear abstractions yet. Compared to the pretty independent markets for Operating Systems or word processors, there are tasks, such as image segmentation, NLP, which are very broad and cannot be seen as as neatly separated."

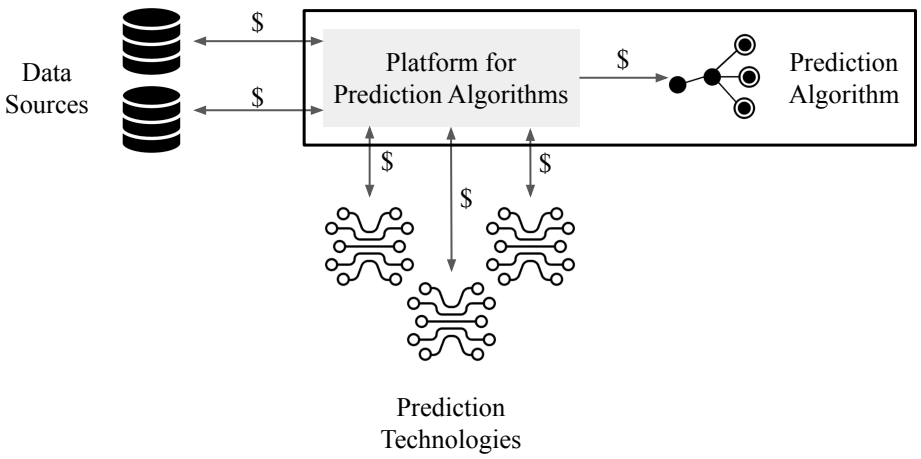


FIGURE 1. Participants in NEWDIGS' market design problem. Heterogenous data providers provide their data and receive reimbursement. Developers of prediction technologies are reimbursed for their services. Finally, prediction algorithms are sold to buyers. We consider a model of the sale of a prediction algorithm to a single buyer.

example, might be willing to pay for accurate predictions for young patients, whereas a health insurer might be more interested in predictions for a regional sub-population. The possibility of transfer learning shrinks the set of potential combinations of prediction quality on different sub-populations: If the quality is degraded on one sub-population, this might impact the quality on another sub-population. This introduces feasibility constraints into the seller's problem.

This proposal studies the pricing of predictions for sub-populations and is only a small part of the challenge of the marketplace upon which NEWDIGS aims to improve. Incentive design is, however, necessary also for other groups (see Figure 1): Data such as lists of claims or Electronic Health Records are dispersed among different data providers; the reimbursement of suppliers of prediction technologies is an open question. Also, in the market place, several buyers have overlapping interests in different patient subpopulations. We address a reduced version of the last of these three problems: A monopolist with data and a fixed prediction technology sells a prediction algorithm to maximize revenue.

Our contributions are threefold:

(1) We present a model of sale of prediction algorithms, or equivalently of information subject to a technology. We bring together notions from

statistical learning (hypothesis classes, empirical risk minimization) and combinatorial mechanism design.
(2) We relate the problem to the multiple-goods monopolist problem and give sufficient conditions for equivalence. Our reduction is constructive in that it allows to construct optimal menus using techniques (duality, algorithms) for the multi-goods monopolist problem.
(3) Finally, we show ambiguous effects of a more flexible prediction technology and introduce more broadly a *quality-specifiability* tradeoff.

**Related Literature.** Our study contributes to a literature on pricing of information. In its reduced setting, we are the closest to [BBS18], where an information buyer would like to acquire additional information to make a payoff-relevant decision. Our model differs from this paper through the addition of a learning technology. Also, methodologically, our model is more closely aligned with the multi-good monopolist problem than the model in [BBS18]. Where [BBS18] introduces heterogeneity via prior knowledge, our model does so by assuming different values for prediction quality.

Our paper is also related to [CLR$^+$15] from algorithmic game theory. In the design problem they study, a mechanism designer would like to efficiently aggregate noisy estimators of a random variable. The agents supply noisy observations with different variances subject to a cost. Cost profiles are private information. The mechanism uses Vickrey-Clarke-Groves payments to elicit costs from the agents. Our model differs in that we study revenue-maximizing mechanisms, whereas [CLR$^+$15] studies efficient ones. In our model, the structure of information is allowed to be more complex than a single quantity.

Also, our model is related to the design of marketplaces for data. [ADS19] shares with our model that agents have willingness to pay according to a loss function. Their focus, however, is on learning optimal pricing when agents arrive sequentially and submit their prediction tasks non-strategically. They propose an algorithm to learn the type distribution and set reserve prices to maximize revenue (see also [GJL19] for a similar model with sequential arrival of buyers). Our model differs from [ADS19] in that it considers strategic reports of prediction problems.

In the machine learning literature, our paper most closely aligns with the study of robustness against adversarial attacks. [STS$^+$18] show that prediction algorithms that are robust to misclassification due to small perturbations (adversarial attacks) require more training data than those that are not. We also propose a perturbation model, with the different goal to locally degrade the quality of prediction algorithms. Our concept differs in that we perturb labels ($Y$), whereas robustness against adversarial attacks adds perturbations to what we call populations ($X$).

Finally, our model introduces a variant of the multiple-goods monopolist problem [DDT17, KM19]. Our variant introduces a feasibility constraint and

hence introduces a mild complementarity of feasible goods combinations into the model.

**Outline.** The plan of the rest of the paper is as follows. In section 2 we introduce our model. We continue in section 3 by relating the problem of selling a prediction algorithm to the multiple-goods monopolist problem. In section 4 we show ambiguous effects of strictly richer technologies on the feasible set of the seller. We discuss policy implications and extensions and formulate conclusions in section 5. All proofs are in Appendix A.

## 2. Model

A monopolist (he) sells a *predictor* $f$ mapping *data* $x \in \mathcal{X}$ to *labels* $y \in \mathcal{Y}$ to a single buyer (she). We assume that $\mathcal{X}$ is a measure space that admits a uniform distribution $\lambda$. The monopolist is constrained to choose a (potentially random) predictor from a *prediction technology* $\mathcal{F} \subseteq \Delta(\{f \colon \mathcal{X} \to \mathcal{Y}\})$, which we specify later.

The *data-generating process* is a Markov kernel

$$p \in \mathcal{X} \times \Delta\mathcal{Y},$$

which is unknown both to the seller and the buyer. Denote the set of all such Markov kernels by $\Pi$. The setup fixes a *loss function* $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. The loss function is common knowledge.

The buyer has a type

$$\theta \in \mathbb{R}_+^{\mathcal{X}},$$

which is the *willingness to pay for subpopulation* $i$.

**Assumption 1.** *There is an exogenous partition* $\mathcal{X} = \bigcup_{i=1}^n \mathcal{X}_i$ *such that* $\theta$ *is constant on* $\mathcal{X}_i$*. We call the* $\mathcal{X}_i$ populations *and write, as an abuse of notation,* $\theta_i = \theta(x)$ *for some* $x \in \mathcal{X}_i$*.*

Using Assumption 1, we can identify types with elements of $\Theta \coloneqq [0,1]_+^n$. The mechanism designer has a prior $F \in \Delta(\Theta)$ which admits a density $f$ with bounded first derivative. For technical reasons, we assume that $f$, along with its first derivative, are zero at the boundary of $\Theta$. The seller offers several (potentially random) predictors $f$ for prices $t$. The agent has a quasi-linear utility

$$\begin{aligned} u((f,t);\theta) &= \int_{\mathcal{X}} \theta(x) \mathbb{E}_{Y \sim p(x,\bullet)}[\mathbb{E}_f[-\ell(f(X),Y)]] \mathrm{d}\lambda(x) - t \\ &= \sum_{i=1}^n \theta_i \int_{\mathcal{X}_i} \mathbb{E}_{Y \sim p(x,\bullet)}[\mathbb{E}_f[-\ell(f(X),Y)]] \mathrm{d}\lambda(x) - t. \end{aligned}$$

The agent has an outside option, whose value we normalize to zero. The agent hence derives utility from a prediction algorithm's negative loss, weighted according to $\theta$. Denote

$$q_i = \int_{\mathcal{X}_i} \mathbb{E}_{Y \sim p(x,\bullet)}[\mathbb{E}_f[-\ell(f(X),Y)]]$$

the *expected quality* of the classifier on population $i$. Denote by

$$\mathcal{Q}(\mathcal{F},p) \coloneqq \left\{ \left( \int_{\mathcal{X}_i} \mathbb{E}_{Y\sim p(x,\bullet)}[\mathbb{E}_f[-\ell(f(X),Y)]] \right)_{i=1,2,\dots,n} \right\} \cap \mathbb{R}^n_+$$

the *feasible positive quality profiles* for technology $\mathcal{F}$ and data-generating process $p$.

The timeline is as follows:

(1) The seller posts a menu $\{f,t\}$
(2) The buyer chooses a (random) prediction algorithm $f$ and pays price $t$, or leaves.
(3) The random function $f$ is realized.
(4) The expected quality of the predictor is realized.

**Program.** By virtue of the revelation principle, the mechanism can elicit the type $\theta$ from the agent if they satisfy incentive compatibility (IC) and individual rationality (IR) constraints. This means that we can rewrite the seller's problem as

$$\begin{aligned} &\max t(\theta) \text{ s.t} \\ \text{(IC)} \qquad &u(f(\theta),t(\theta);\theta) \geq u(f(\theta'),t(\theta');\theta) - t(\theta'), \quad \theta,\theta' \in \Theta \\ \text{(IR)} \qquad &u(f(\theta),t(\theta);\theta) \geq 0, \quad \theta \in \Theta \\ &f \in \mathcal{F}, \end{aligned}$$

**Reparametrization.** Observe that $u(f(\theta),t(\theta);\theta)$ depends on $f$ only through $q_i$. Indeed, using the notation $\theta \cdot q \coloneqq \sum_{i=1}^n \theta_i q_i$, we can rewrite the above program as

$$\begin{aligned} &\max t(\theta) \text{ s.t} \\ \text{(IC')} \qquad &q(\theta)\cdot\theta - t(\theta) \geq q(\theta')\cdot\theta - t(\theta'), \quad \theta,\theta' \in \Theta \\ \text{(IR')} \qquad &q(\theta)\cdot\theta - t(\theta) \geq 0,, \quad \theta \in \Theta \\ &q_x(\theta) \in Q(\mathcal{F},p) \end{aligned}$$

for any $\theta,\theta' \in \Theta$. We highlight that if $\mathcal{Q}(\mathcal{F},p) = [0,1]^n$, then this problem is mathematically equivalent to the multi-goods monopolist problem. The rest of this paper will study the relationship between $\mathcal{F}$ and $\mathcal{Q}$, and go first steps into solving this problem for $\mathcal{Q} \neq [0,1]^n$.

2.1. **Prediction Technologies.** In this subsection, we describe additional structure that we put on $\mathcal{F}$. As the seller's problem only depends on $\mathcal{F}$ through $\mathcal{Q}(\mathcal{F},p)$, we can formulate our theorems below purely in relation to properties of $\mathcal{Q}(\mathcal{F},p)$. We discuss generalizations of our discussion in Appendix B.

We assume that $\mathcal{Y} = \mathbb{R}$ and $\ell(y_1,y_2) = -2 + (y_1 - y_2)^2$. (The constant $-2$ can be seen as the value for agents of a perfect prediction. We do not allow for heterogeneity of these baseline loss values in the present model.) We call this the *regression benchmark*. Fix a *hypothesis class* $\mathcal{H} \subseteq \{f\colon \mathcal{X} \to \mathcal{Y}\}$ of

functions and distinct $x_1, x_2, \dots x_m \in \mathcal{X}$. (In econometrics/statistics this would correspond to the *fixed design* setting.) Assume that $Y_i \sim p(x, \bullet)$, $i = 1, 2, \dots, n$ are independent and identically distributed. Denote

$$(\mathbf{X}, \mathbf{Y}) = ((x_1, Y_1), (x_2, Y_2), \dots, (x_m, Y_m)).$$

We call

$$f_{\mathrm{ERM}}(\mathbf{Y}) \in \arg\max_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(Y_i, f(x_i))$$

the *empirical risk minimizer*, or an estimator from the class of $M$-estimators, which is the ordinary least squares analysis in the regression benchmark if the hypothesis class is all linear functions. (We highlight that this is a random function, as $Y_1, Y_2, \dots, Y_n$ are random.) If the empirical risk minimizer is not unique, we assume that the prediction seller can choose a minimizer. We denote by $Y_i^{\varepsilon}$ the random variable

$$Y_i^{\varepsilon} = Y_i + \nu_i,$$

where $\nu_i \sim N(0, \varepsilon_i)$, $i = 1, 2, \dots, m$ is independent Gaussian noise. We call this $\varepsilon$-perturbed data. Now consider the class of random functions

$$\mathcal{F} = \mathcal{F}(P, x_1, x_2, \dots, x_m) = \{f_{\mathrm{ERM}}(\mathbf{Y}^{\varepsilon}) | \varepsilon \in \mathbb{R}^m\}.$$

Our choice of prediction functions deserves some discussion. We choose $M$-estimators as a general class of algorithms widely used in econometrics, statistics and machine learning.

An alternative model of perturbation could be to add independent noise to prediction *labels* ($Y$) after an output of the algorithms. Our perturbation model perturbs data due to the following thought experiment. If an algorithm with such added noise were published, a buyer could relatively easily remove such noise and hence render the degradation of accuracy infeasible.

We characterize $\mathcal{Q}(\mathcal{F}, p)$ to make the concepts in this subsection more concrete.

**Example 1.** *Let $\mathcal{X} = [0, 1]$, $\mathcal{X}_1 = [0, 1]$, $\mathcal{X}_2 = [1, 2]$. Let $x_1 = 1$, $x_2 = 2$ and*

$$Y = \xi \overset{d}{=:} p(x, \bullet),$$

*where $\xi_i \sim N(0, 1)$. The hypothesis class is $\mathcal{H} = \{x \mapsto \beta x\}$. Observe that then*

$$Y_i^{\varepsilon} = \xi_i$$

*where $\xi_i \sim N(0, 1 + \varepsilon_i)$. Using known identities from statistics, we find that the posterior given $(\mathbf{Y}^{\varepsilon})$ is given by*

$$\beta \sim N\left(0, \frac{3}{5} + \frac{1}{5}\varepsilon_1 + \frac{2}{5}\varepsilon_2\right).$$

*and hence* $(\beta x - \xi)^2$ *is a multiple of a standard* $\chi^2$*-random variable with expectation*

$$x^2 \left( \frac{3}{5} + \frac{1}{5}\varepsilon_1 + \frac{2}{5}\varepsilon_2 \right)$$

*Integrating over* $\mathcal{X}_1$ *resp.* $\mathcal{X}_2$*, we get*

$$\mathcal{Q} = \left\{ \left( \frac{9}{5} - \lambda, \frac{3}{5} - \lambda \right) \middle| \lambda \in \left[ 0, \frac{3}{5} \right] \right\}$$

*This set is depicted in Figure 2. In this example, hence, the prediction technology allows only to degrade the quality on one subset if the other is degraded. In an effectively one-dimensional problem,* [Mye81]*'s characterization applies and a non-degraded predictor for a posted price maximizes revenue.*

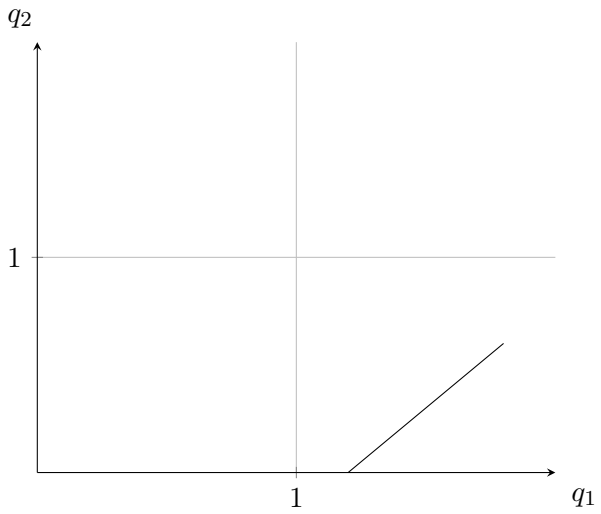


FIGURE 2. Realizable Qualities for a regression problem. The feasible qualities are perfect complementary in that an increase in quality of predictions on one population implies an increase in prediction accuracy in other populations.

Before we come to our results, we discuss limitations of our model. First, our model our model does not allow for heterogeneity concerning outside options of accuracy or prior knowledge. Any baseline accuracy values are in our model encoded in the loss function. This homogeneity allows us to consider incentive compatibility constraints independent from types. A second limitation concerns verifiability. The model assumes that the prediction accuracy can be perfectly evaluated. In the real world, this would typically happen via a test set. We do not model incentives relating to the provision of test sets. Therefore, our model applies in more standardized prediction problems in which test sets are readily available, such as object recognition.

## 3. Relation to the Multiple-Goods Monopolist Problem

In this section, we formalize the relationship of the seller's problem to the multi-goods monopolist problem.

In the multi-goods monopolist problem, a seller offers a bundle of $n$ goods, $q \in X = [0,1]^n$, to an buyer that has a type $\theta \in \Theta = [0,M]^n$ for some $M \in \mathbb{R}$. The agent has quasi-linear utility and the buyer has a prior $F \in \Delta\Theta$ which admits a density function $f$ that is differentiable with bounded derivative. The seller would like to design a menu $\mathcal{M}$ of allocations $q$ and payments $t$ that maximizes expected payments.

Before we formalize the relation of the two problems, we need one more definition. We call a function class $\mathcal{H}$ *unconstrained* if $\mathcal{H} = \{f \colon \mathcal{X} \to \mathcal{Y}\}$, i.e. if any function can be fit. We call the mechanism designer's problem unconstrained if the underlying hypothesis class is unconstrained.

**Proposition 1.** *Assume the mechanism designer's problem is unconstrained. Then, the problem of selling a prediction algorithm is equivalent to the multiple goods monopolist problem. More specifically, there is a linear function $A \colon [0,1]^n \to \mathcal{Q}$ such that if the optimal menu for the multi-goods monopolist problem with type density $f(A^{-1}(\theta))$ is given by $(q(\theta), t(\theta))$, then the optimal menu for the prediction seller's problem is $(q(A(\theta)), t(A(\theta)))$.*

This reduction has immediate corollaries such as the existence of infinite menus and randomization [DDT17] for some instances of the seller's problem. Relevant to our setting, a direct corollary of [DDT17, Theorem 4] gives an indication when it is revenue-maximizing to sell a non-degraded prediction algorithm for a posted price. Call the prediction algorithm $T_{\text{ERM}}(\mathbf{Y})$, i.e. the one that uses no perturbation, the *maximally informative prediction algorithm.*

**Corollary 1.** *Assume the seller's problem is unconstrained. For every $n$ there is a $c \in \mathbb{R}_+$ such that if values are distributed uniformly on $[c, c+1)$, it is optimal to offer the maximally informative prediction algorithm for a posted price.*

If willingness to pay for goods has hence a sufficiently high common component, then it is optimal to not degrade the predictor at all.

We highlight that unrestricted function classes are typically not employed, as they can fit arbitrary noise. The aforementioned results should hence be seen as limiting cases of independence of qualities in different populations.

To make progress on the case of constrained problems, we give a more general condition than the one in Proposition 1 to derive a similar result.

**Assumption 2.** *Let the set of attainable qualities be a polytope in $\mathbb{R}^n_+$ with at most $2n$ faces that contains a neighborhood of $\mathbf{0}$, the all-zero vector. Mathematically, this assumption says that*

$$\mathcal{Q}(\mathcal{F}, p) = \{q \in \mathbb{R}^n | q \geq 0, Aq \leq \mathbb{1}\},$$

*where $A \in \mathbb{R}^{n \times n}$ is invertible and $\mathbb{1}$ is the all-one vector.*

This assumption says that the feasible set is defined by linear constraints. This is satisfied by Example 1 (when projecting into one dimension), but neither of our examples.

**Proposition 2.** *Assume Assumption 2. Then, the problem of selling a prediction algorithm is equivalent to the multiple goods monopolist problem in the sense of Proposition 1.*

Proposition 2 lies the groundwork for the verification of optimal mechanisms, thanks to duality theory for the multi-goods monopolist problem [DDT17, Theorem 2]. Also, this allows for generalizations. We conjecture, e.g., a generalization of Corollary 1 for bounded complementarity. We leave this for further work.

## 4. Comparing Prediction Technologies

In this section, we continue to study the regression benchmark, and vary the hypothesis class from Example 1. Recall that in this setup, $\mathcal{Y} = \mathbb{R}$ and $\ell(y_1, y_2) = -2 + (y_1 - y_2)^2$, as well as $\mathcal{X} = [0, 1]$, $\mathcal{X}_1 = [0, 1]$, $\mathcal{X}_2 = [1, 2]$. Furthermore, $x_1 = 1$, $x_2 = 2$ and

$$Y = \xi_i \stackrel{d}{=:} p(x, \bullet),$$

where $\xi_i \sim N(0, 1)$. We now assume, however, that $\mathcal{H}' = \{x \to \beta_1 + \beta_2 x\}$. Observe that $\mathcal{H}'$ is a superset of the hypothesis class $\{x \mapsto \beta x | \beta \in \mathbb{R}\}$ considered in Example 1. We say that hypothesis class $\mathcal{H}'$ is more flexible than hypothesis class $\mathcal{H}$.

**Proposition 3.** *Fix some data $x_1, x_2, \ldots, x_m$. The mapping $(\mathcal{H}, \subseteq) \mapsto (\mathcal{Q}, \subseteq)$ is not monotonic.*

This result shows that the effects of a change in technology for the revenue of the seller might be ambiguous: The set of achievable qualities (which determines the optimal mechanism) need need not be related for more or less flexible classes. This tradeoff is illustrated in a concrete example in Figure 3. The achievable qualities in the case of the less flexible model allow for higher joint quality (point $a$), but does not allow arbitrarily low quality on one issue with still high quality on the other issue (point $b$). The qualities in the more flexible model give less accuracy, as given the same number of samples from a distribution, it has more parameters to fit. On the other hand, more parameters allow to more flexibly degrade accuracies on different sub-populations.

This tradeoff of *specifiability* versus *quality* can again be seen as a production constraint: In our example of a more flexible technology, the qualities are (imperfect) complements.

This tradeoff, together with our results on unconstrained problems shed light on the revenue maximizing choice of prediction technologies: To maximize revenue, these should be less flexible where accurate fit to data requires it *and* demand does not vary too much.

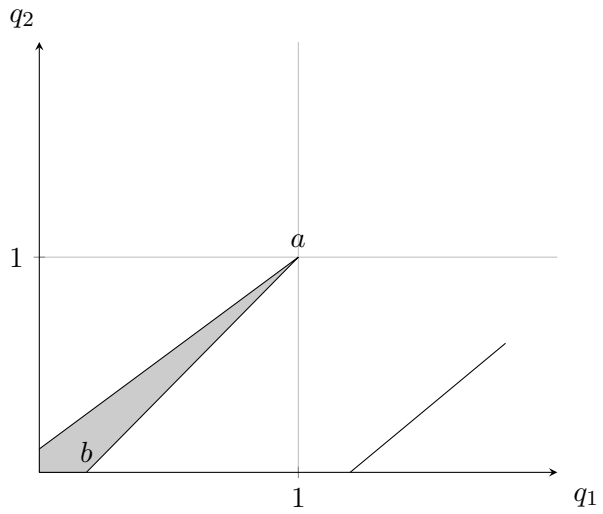


FIGURE 3. Comparison of sets of feasible qualities for two hypothesis classes. The line for hypothesis class $\{x \mapsto \beta x | \beta \in \mathbb{R}\}$, the shaded region for $\{x \mapsto \beta_1 + \beta_2 x | \beta \in \mathbb{R}\}$

## 5. Conclusion

We studied the sale of prediction algorithms subject to a prediction technology constraint. We derived that the structure of data can lead to complex menus even in the most restricted settings. Furthermore, we showed that our problem features a complementarity in feasible prediction qualities and that the prediction technology non-trivially influences the feasible qualities.

Our model has some implications for the design of prediction sale and regulation. Techniques that make prediction algorithms less accurate for some issues and hence partially excludable, can do so only to a certain degree due to constraints imposed by prediction technologies. Mandating more flexible prediction algorithms does not solve this problem, as they might perform worse with respect to accuracy overall. Therefore, designing prediction technologies should already take into account the demand structure if revenue maximization is the goal.

A first avenue for further research is on problems with multiple buyers. NEWDIGS' problem naturally features the allocation of one (non-excludable) prediction algorithm to several agents. Studying this constrained public goods model would introduce additional complexities due to free-riding and come closer to improving the marketplace.

A second avenue for further research is the characterization of sets of achievable qualities for broader classes of prediction technologies. This could shed light on the relation between the geometry of hypothesis classes (such as the Vapnik-Chervonenkis dimension [Wai19]) and the specifiability of technologies. In this vein, also connections of specifiability and quality to the

bias-variance decomposition could be derived [HTF14]. Finally, results for broader classes of prediction technologies could make the problem amenable to further research on algorithmic questions.

A third area for further research is the relationship of the proposed model to the sale of statistical experiments [BBS18]. We formulated our model without any prior knowledge. A uniform outside option regarding prediction accuracy is encoded in the loss function. It would be interesting to see whether our model can be framed as a variant of [BBS18] with an additional constraint due to a prediction technology. Such a formulation could also inspire the introduction of technology constraints into [BBS18]'s model.

## References


[ADS19] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar, *A marketplace for data: An algorithmic solution*, ACM EC 2019 - Proceedings of the 2019 ACM Conference on Economics and Computation (2019), 701–726.

[BBS18] Dirk Bergemann, Alessandro Bonatti, and Alex Smolin, *The design and price of information*, American Economic Review **108** (2018), no. 1, 1–48.

[CLR+15] Rachel Cummings, Katrina Ligett, Aaron Roth, Zhiwei Steven Wu, and Juba Ziani, *Accuracy for sale: Aggregating data with a variance constraint*, ITCS 2015 - Proceedings of the 6th Innovations in Theoretical Computer Science (2015), 317–324.

[DDT17] Constantinos Daskalakis, Alan Deckelbaum, and Christos Tzamos, *Strong Duality for a Multiple-Good Monopolist*, Econometrica **85** (2017), no. 3, 735–767.

[GJL19] Negin Golrezaei, Patrick Jaillet, and Jason Cheuk Nam Liang, *Incentive-aware Contextual Pricing with Non-parametric Market Noise* (2019), 1–68.

[HTF14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *Elements of Statistical Learning*, Springer, 2014.

[KM19] Andreas Kleiner and Alejandro Manelli, *Strong Duality in Monopoly Pricing*, Econometrica **87** (2019), no. 4, 1391–1396.

[Mye81] Roger B Myerson, *Optimal Auction Design*, Mathematics of Operations Research **6** (1981), no. February, 1–17.

[STS+18] Ludwig Schmidt, Kunal Talwar, Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry, *Adversarially robust generalization requires more data*, Advances in Neural Information Processing Systems (2018), 5014–5026.

[Wai19] Martin J. Wainwright, *High-Dimensional Statistics*, Cambridge University Press, 2019.

## Appendix A. Proofs

*Proof of Proposition 1.* We first show that the set of attainable qualities is equal to a hypercube, and then apply Proposition 2.

Observe that by unconstrainedness, the seller can choose the minimizers for $x_i$ such that $x_i \in \mathcal{X}_\rangle$ separately for $i = 1, 2, \ldots, n$,

$$f_{\mathrm{ERM}(\mathbf{Y})}|\mathcal{X}_i = f_i,$$

where $f_i$ depends only on $(Y_j)_{j:x_j \in \mathcal{X}_i}$. Observe that for $q \in \mathcal{Q}$, $q_i$ depends on $f_{\mathrm{ERM}(\mathbf{Y})}$ only through $f_i$. Hence, $(\varepsilon_i)_{j:x_j \in \mathcal{X}_i}$ only affect $q_i$, but nor $q_k$, $k \neq i$. This implies that $Q$ is a product set. Furthermore, it is not hard to see that $q_i$ (for an appropriate choice of $f_{\mathrm{ERM}(\mathbf{Y})}$) is continuous in $(\varepsilon_i)_{j:x_j \in \mathcal{X}_i}$ and that as for $\varepsilon \to \infty$, $q_i \to 0$ uniformly in $i = 1, 2, \ldots, n$, we find that it must be of the form

$$\mathcal{Q}(\mathcal{F}, p) = \{q \in \mathbb{R}^n | \mathbf{0} \leq q, Aq \leq \mathbb{1}\}$$

for some choice of $a \in \mathbb{R}^n$.

Then, the seller's problem can be formulated as

$$\begin{aligned} &\max t(\theta) \text{ s.t} \\ q(\theta) \cdot \theta - t(\theta) &\geq q(\theta') \cdot \theta - t(\theta'), \quad \theta, \theta' \in \Theta \\ q(\theta) \cdot \theta - t(\theta) &\geq 0, \quad \theta \in \Theta \\ q(\theta) &\in \{q \in \mathbb{R}^n | \mathbf{0} \leq q, Aq \leq \mathbb{1}\}, \end{aligned}$$

and we can apply Proposition 2. □

*Proof of Proposition 2.* Consider the program for the seller's problem:

$$\begin{aligned} &\max t(\theta) \text{ s.t} \\ q(\theta) \cdot \theta - t(\theta) &\geq q(\theta') \cdot \theta - t(\theta'), \quad \theta, \theta' \in \Theta \\ q(\theta) \cdot \theta - t(\theta) &\geq 0, \quad \theta \in \Theta \\ q(\theta) &\in \{q \in \mathbb{R}^n | \mathbf{0} \leq q, Aq \leq \mathbb{1}\}, \end{aligned}$$

Now define $\tilde{\theta} \coloneqq A(\theta) \in \mathbb{R}^n_+$, $\tilde{q} \coloneqq Aq \in [0,1]^n$, as well as $\tilde{t}(\theta) = t(A(\theta))$ and $\tilde{q}(\theta) = q(A^{-1}(\theta))$. Given these renamings, the problem reads

$$\begin{aligned} &\max \tilde{t}(\tilde{\theta}) \text{ s.t} \\ \tilde{q}(\tilde{\theta}) \cdot \tilde{\theta} - \tilde{t}(\tilde{\theta}) &\geq \tilde{q}(\tilde{\theta}') \cdot \tilde{\theta} - \tilde{t}(\tilde{\theta}'), \\ \tilde{q}(\tilde{\theta}) \cdot \tilde{\theta} - \tilde{t}(\theta) &\geq 0, \\ \mathbf{0} \leq \tilde{q}(\theta) &\leq \mathbb{1}. \end{aligned}$$

Let $\tilde{t}$ and $\tilde{q}$ be a solution to this problem. They are the solution to the multiple-goods monopolist problem with types distributed as $f(A^{-1}(\theta))$ for a density function $f$. Note that this distribution can be extended to $[0,1]^M$ while admitting bounded derivatives by our assumption on the derivative of the density at the boundary. Then, $q(\theta) = \tilde{q}(A(\theta))$, $t(\theta) = \tilde{t}(A(\theta))$ defines a solution to the seller's problem under type distribution $f(\theta)$. □

We highlight that this proof is not possible for affine, nonlinear transformations.

*Proof of Proposition 3.* For this, it suffices to compute the set of feasible qualities and show that it is not a super- or subset of the feasible qualities set in Example 1, and hence, to compute an example.

Recall that $\mathcal{X} = [0,2]$, $\mathcal{X}_1 = [0,1]$, $\mathcal{X}_2 = [1,2]$, $x_1 = 1$, $x_2 = 2$, $p(x, \bullet) = N(0,1)$ and $\ell(y_1, y_2) = 2 - (y_1 - y_2)^2$. We now consider the function class $\mathcal{H}_2 = \{\beta_1 + \beta_2 x | \beta_1, \beta_2 \in \mathbb{R}\}$. Note that $Y_i^\varepsilon \sim N(0, 1+\varepsilon_i)$. We get with formulas from econometrics that

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} \sim N(0, \Sigma)$$

where

$$\Sigma = \begin{pmatrix}4(1+\varepsilon_1)^2 & (1+\varepsilon_2)^2\\ (1+\varepsilon_1)^2 & (1+\varepsilon_2)^2\end{pmatrix}.$$

We then get that

$$(\beta_1 + \beta_2 x - \xi)^2$$

has expectation

$$1 + 4(1+\varepsilon_1)^2 + x((1+\varepsilon_1)^2 + (1+\varepsilon_2)^2) + (1+\varepsilon_2)^2 x^2.$$

Introduce the following parameters $\kappa_i \coloneqq (1+\varepsilon_i)^2$, $i = 1, 2$. This gives for the achievable qualities

$$\left(1 - \frac{9}{2}\kappa_1 - \frac{5}{6}\kappa_2, 1 - \frac{11}{2}\kappa_1 - \frac{23}{3}\kappa_2\right)$$

which can be written as the convex hull of $\{(0,0), (1,1), (\frac{2}{11}, 0), (0, \frac{5}{46})\}$, which we depict in Figure 3. □

## Appendix B. On Our Choice of Perturbation Model

The addition of noise is also possible in the case of categorical variables. In this case, we can define

$$Y_i^\varepsilon \begin{cases} = Y_i & \text{w.p. } 1-\varepsilon \\ \sim U(\mathcal{X}) & \text{w.p. } \varepsilon \end{cases},$$

and derive measures of quality in a similar way. We conjecture an axiomatic characterization of these perturbations, which we discuss in the following.

We call a function

$$T \colon (\Delta(\mathcal{X} \times \Delta(\mathcal{Y})), \mathcal{Y}^m) \to \mathcal{H}, (p, \mathbf{Y}) \mapsto f$$

a *learning algorithm*, which takes a prior and training samples to a function in the hypothesis class. We denote by $\mathcal{T}$ the set of learning algorithms. (We drop the dependence on the fixed values $x_1, x_2, \ldots, x_m$.) Note that empirical

risk minimization is a learning algorithm, which we denote by $T_{\mathrm{ERM}(p,\mathbf{Y})}$. Given a prior $p' \in \Delta(\mathcal{X} \times \Delta\mathcal{Y})$

$$A_{p',T} = \left\{ \left( \int_{\mathcal{X}_i} \mathbb{E}_{Y \sim p(x,\bullet)}[\mathbb{E}_f[-\ell(T(p,\mathbf{Y})(X),Y)]] \right)_{i=1,2,\ldots,n} \middle| f \in \mathcal{H} \right\}$$

the *set of achievable qualities of a learning algorithm.*

**Definition 1.** *Denote* $\Pi_p = \{p \in \Delta(\mathcal{X} \times \Delta(\mathcal{Y}) | p(x_i, \bullet) = p'(x_i, \bullet)\}$ *the set of priors that agree in law with the training data. Then define*

$$Q_p = \bigcap_{p' \in \Pi_p} \bigcup_{T \in \mathcal{T}} A_{p,T}$$

*the* set of prior-independent realizable qualities.

**Conjecture 1.** *In the regression benchmark, the set of prior-independent realizable qualities is equal to the set of empirical risk minimizers with $\varepsilon$-perturbed data. More technically,*

$$\mathcal{Q}_p = \mathcal{Q}(\mathcal{F}, p).$$

The last conjecture says that it is, for a prediction seller that has no information on the data-generating process beyond the information contained in samples, without loss to consider our model of perturbations. A proof would observe that any addition of a deterministic quantity that could be done with the knowledge of the prior reduces the quality on populations in the same way as appropriate noise would.

Massachusetts Institute of Technology, Institute for Data, Systems, and Society, 50 Ames Street, 02142 Cambridge MA, USA
*Email address*: haupt@mit.edu