Algorithmic Decision-Making Safeguarded by Human Knowledge

Ningyuan Chen, Ming Hu, Wenhao Li
Rotman School of Management, University of Toronto, Toronto, Ontario, Canada M5S 3E6
ningyuan.chen@utoronto.ca, ming.hu@rotman.utoronto.ca, wwenhao.li@mail.utoronto.ca

Commercial AI solutions provide analysts and managers with data-driven business intelligence for a wide range of decisions, such as demand forecasting and pricing. However, human analysts may have their own insights and experiences about the decision-making that is at odds with the algorithmic recommendation. In view of such a conflict, we provide a general analytical framework to study the augmentation of algorithmic decisions with human knowledge: the analyst uses the knowledge to set a guardrail by which the algorithmic decision is clipped if the algorithmic output is out of bound and seems unreasonable. We study the conditions under which the augmentation is beneficial relative to the raw algorithmic decision. We show that when the algorithmic decision is asymptotically optimal with large data, the non-data-driven human guardrail usually provides no benefit. However, we point out three common pitfalls of the algorithmic decision: (1) lack of domain knowledge, such as the market competition, (2) model misspecification, and (3) data contamination. In these cases, even with sufficient data, the augmentation from human knowledge can still improve the performance of the algorithmic decision.

1. Introduction

The Russia-Ukraine war has sent a seismic wave to the energy market and brought soaring gas prices under the spotlight. Ideally, the optimal retail fuel price at the pump needs to take into account many factors, including the crude oil price, the transportation cost, the brand value, and the local competition. Because of the complexity of the pricing problem, in practice, it is not surprising that the station managers would rely on some heuristics or simple rules (such as a constant markup over the cost) to set prices instead of using a sophisticated pricing algorithm.

This is no longer the case for many gas stations, especially those owned by a large corporation. PDI Fuel Pricing[1] sells software to gas station managers that helps them set fuel prices more intelligently using data analytics and machine learning. It uses a wide range of data, including historical prices and demand, as well as competitors' prices and claims to "fine-tune your pricing strategy

[1] https://www.pdisoftware.com/fuel-pricing-solutions/

with live competitive insights allowing [managers] to react quickly to market conditions." Needless to say, machine learning algorithms can significantly improve profitability over the heuristic approach.

However, it is not hard to imagine scenarios when a human analyst or the station manager may not be fully convinced by the price prescribed by the algorithm, especially when the algorithm is a black box (typical for many machine learning algorithms) and the prescribed price deviates from human intuition significantly. For example, when the algorithm recommends a price that is much higher than what would have been charged by the station manager, should the algorithmic decision be trusted over human knowledge? On the one hand, the algorithm takes much more quantitative information as input than the human analyst, and the higher price could reflect the rising demand, a pattern in the data missed by the human analyst. On the other hand, human knowledge may have relied on simple rules such as matching the price of another station around the corner. Such price-matching heuristics may have worked well in the past. When the decisions from the algorithm and human knowledge are in conflict, it can be hard for the analyst to make a call.

The problem faced by the station manager in the motivating example is prevalent. Most business owners have realized the importance of the AI revolution and are willing to invest in it to improve business decision-making. However, as AI algorithms become increasingly sophisticated, many firms have no choice but to outsource the standardized components in the decision-making process to commercial AI solutions. These decisions, such as pricing and inventory management, have historically been made through human instincts and experiences. When human knowledge and the decision output by AI algorithms deviate significantly, firms face a similar dilemma to the station manager.

In this paper, we provide a general analytical framework to study practical problems in which humans and AI interact in the decision-making process. Motivated by the gas station example, we consider an AI system prescribing a decision based on past data and some machine learning algorithms. Based on the prescribed decision, the human analyst may set a guardrail using simple rules from the accumulated knowledge, experiences, or expertise. More precisely, human knowledge is translated to a cap or floor, or both of the decision. That is, if the algorithmic decision violates the bounds, the human analyst may override it by clipping it to the imposed cap or floor. For example, the algorithm may recommend a retail price of \$5.10 per gallon. At the same time, human knowledge indicates, "the price can't be higher than \$5.00 per gallon because the station around the corner is only charging \$4.80." As a result, the human analyst may set the final price to \$5.00. Otherwise, if the recommended price is lower than \$5.00, then the algorithmic decision is followed. In this interaction, AI is the main force behind the decision-making, while human knowledge serves as an auxiliary, safeguarding the algorithmic decision from prescribing unreasonably high prices. It is a fair representation of a considerable fraction of human-AI interaction in practice.

With the framework, we aim to answer the following research question: When does human knowledge add value to AI decision-making? Our first result is *negative*: human knowledge does not provide any benefit if (1) the algorithmic decision improves with more data, for example, when the mean squared error with respect to the optimal decision is diminishing, and (2) human knowledge is not improving with more data. This result is somewhat expected: The guardrail prescribed by human knowledge can itself be treated as the pattern extrapolation of past data, albeit a simple and heuristic one. If the algorithm can efficiently recognize and extrapolate the pattern better than the human, as many machine learning algorithms do, then it is unnecessary to augment the algorithmic decision with human knowledge under sufficient data.

The above result may sound intuitive, but it is derived in an ideal situation. While it may be reasonable to assume that human knowledge does not improve constantly with more data, as human brains are generally unable to recognize complex patterns hidden in a large dataset, there are many caveats in applying commercial off-the-shelf AI systems to real-world applications as those algorithms may fail to satisfy condition (1) above. In these cases, human knowledge can be used to augment the algorithmic decision. In this study, we identify three such use cases within our framework and argue that, in these cases, rhetorically, the gas station manager should not completely delegate the pricing decision to the algorithm of PDI Fuel Pricing. The three cases summarize the common pitfalls when making business decisions and trusting the algorithm blindly.

- When the firm is in a *competitive market*, and the algorithm fails to fully take into account the competitors' decision (due to incomplete data or algorithmic design), simple decision rules based on human knowledge, such as price matching, can improve the algorithmic decision. This is not an uncommon setting. For example, PDI Fuel Pricing may not have direct access to the pricing data of other competing local stations unless they subscribe to some service as well. We show that when a competitor sets a price near the Nash equilibrium, using the algorithmic price and matching it to the competitor's price when the algorithmic price is higher can improve the algorithmic decision.
- The algorithm may be susceptible to *model misspecification*. In the pricing context, the algorithm may mistakenly treat the demand function as a linear function and recommend the optimal price based on the misspecified linear demand model. On the other hand, the human analyst may simply observe which price generates the highest profit empirically in the past data without fitting or optimizing a model. This heuristic turns out to be quite robust to model misspecification. We show that human knowledge when used to safeguard the algorithm,

can help mitigate the model misspecification and improve the profitability of the algorithmic decision.

- When the data fed into the algorithm are *contaminated*, possibly due to the reporting or measurement error, then the relative insensitivity of the human knowledge to specific data points turns out to be a robust mechanism. Not surprisingly, the combination of human knowledge and the algorithmic decision can prevent the latter from being misguided by the contaminated data. We provide an analytical condition that characterizes the contamination level for human knowledge to prevail.

In all three cases, instead of an abstract AI system, we materialize the algorithmic decision and study linear regression, which allows us to concretely analyze the trade-off of safeguarding the regression output using simple rules. Linear regression is widely used and is representative of a more complex machine learning algorithm. Such treatment allows us to provide technical conditions under which augmentation by human knowledge can improve the algorithmic decision.

This study contributes to the growing literature on human-AI collaboration. In some applications, it has been shown that AI lacks crucial human strengths such as domain knowledge and common-sense reasoning (Holstein and Aleven 2021, Lake et al. 2017, Miller 2019), which motivates the collaboration between AI and human experts on subjects including chess (Case 2018, Das and Chernova 2020), healthcare (Dai and Singh 2021, Irvin et al. 2019, Patel et al. 2019), criminal justice (Grgić-Hlača et al. 2019, Kleinberg et al. 2018), education (Cheng et al. 2019, Smith et al. 2012), and public services (Binns et al. 2018, Chouldechova et al. 2018). This study is motivated by business problems, and the human-AI interaction is uniquely defined by the context. Below we review the literature closely related to this study.

2. Related Literature

This research is broadly related to two streams of literature: those papers providing conceptual or theoretical frameworks for human-AI collaboration and empirical papers documenting real-world interactions between AI and human analysts. In the first stream, recent literature in computer science aims at the optimal integration of human and AI decisions (Bansal et al. 2021, 2019, Donahue et al. 2022, Gao et al. 2021, Keswani et al. 2021, Madras et al. 2018, Mozannar and Sontag 2020, Raghu et al. 2019, Rastogi et al. 2022, Wilder et al. 2020). On the one hand, Madras et al. (2018) propose a learning-to-defer framework in which the AI can choose to make decision by its own or just pass the task to the downstream human expert. The expert has information unavailable to AI and may make better decisions. Follow-up papers extend the framework to more complex settings, such as multiple experts (Keswani et al. 2021), bandit feedback (Gao et al. 2021), joint optimization of the prediction algorithm and pass function (Mozannar and Sontag 2020, Wilder et al. 2020). On the other hand, Donahue et al. (2022), Rastogi et al. (2022) consider a *weighted average* aggregation of human and AI decisions and show conditions for human-AI complementarity in which the aggregated decision outperforms both individual decisions. Recently, Grand-Clément and Pauphilet (2022) show that in the setting of sequential decision-making, the AI algorithm should be trained differently when a human analyst is involved. Motivated by business applications, our paper differs from this stream of works as we study a particular (not necessarily optimal) way to integrate the algorithmic and human decisions tailored to the application. In the motivating example, the manager does not have access to the internal structure of the algorithm and cannot design a meta-algorithm to optimally instill her own knowledge into the algorithm.

Some recent studies in Operations Management analyze the human-AI interaction in a theoretical framework (Agrawal et al. 2018, 2019, Boyaci et al. 2020, Dai and Singh 2021, de Véricourt and Gurkan 2022, Ibrahim et al. 2021). They focus on modeling the impact of AI-based predictions on the human decision-making process. Boyaci et al. (2020) study the impact of AI predictions on human decision errors and the cognitive effort humans put into their decisions. The human has the cognitive flexibility to attend information from diverse sources but under limited cognitive capacity, while the AI only processes incomplete information but with great accuracy and efficiency. Through a rational inattention model, the authors show that AI prediction improves the overall accuracy of human decisions and reduces cognitive effort. Agrawal et al. (2018) consider the human analyst aiming to maximize the utility which depends on their decision and the uncertain state. The state can be predicted accurately by the AI algorithm. But the human needs to learn the utility function. The authors show that AI prediction generally complements the human effort but could be a substitute in some cases. de Véricourt and Gurkan (2022) consider the human-AI interactions in a sequential setting in which the analyst gradually learns the accuracy of the AI algorithm through a sequence of tasks. Since the analyst can override AI and never actively explores the AI accuracy, the analyst may never know whether AI outperforms herself at the end of the day. The authors provide explanations for the coexistence of AI and humans, even if one actually outperforms the other. Dai and Singh (2021) use a theoretical framework to analyze a physician's decision with regard to whether to use AI when prescribing a treatment. They find that physicians may intentionally avoid using AI, even when AI can help mitigate clinical uncertainty because doing so increases their liability when adverse patient outcomes occur. Our paper differs from these papers in modeling the human decision-making process. In our model, we assume the human analyst aims to directly safeguard the AI decisions using intuition and expertise. We focus on whether the integration improves the raw AI output.

Empirical evidence shows that human knowledge can still improve AI systems, even though the latter have access to big data and computational resources (Campbell and Frei 2011, Karlinsky-Shichor and Netzer 2019, Kesavan and Kushwaha 2020, Liu et al. 2022, Phillips et al. 2015, Sun et al. 2022, Van Donselaar et al. 2010). For example, in the context of inventory replenishment, Van Donselaar et al. (2010) find that store managers often modify the algorithmic recommendation from an automated replenishment system. Kesavan and Kushwaha (2020) use the data from a field experiment to investigate the merchant's modification of the advice from a data-driven central-planning system. They find that the merchant's modification reduces the overall profitability but improves the profit for growth-stage products whose historical data are limited. Liu et al. (2022) conduct a field experiment to compare the inventory replenishment strategies of human buyers and AI algorithms. They find the algorithm outperforms human buyers in terms of reducing out-of-stocks rates and inventory rates. The most related empirical works to our study are Fogliato et al. (2022), Ibrahim et al. (2021). Ibrahim et al. (2021) show how to exploit the human domain knowledge to improve the AI predictions for surgery duration. Particularly, they suggest inputting the human adjustment (so-called private information adjustment in the paper), instead of the human direct forecast, into the prediction algorithm. Their work conveys a message similar to ours: even the human predictions are less accurate than AI, they can still help boost AI performance. Fogliato et al. (2022) investigate human-AI collaboration in the context of child maltreatment hotline screening. Due to the technical glitch caused by incorrect input, the AI may incorrectly predict the risk score in some cases. They find that human analysts are more likely to override AI recommendations when AI makes a mistake. The work shows that humans can augment the algorithmic decision when the algorithm exhibits defects in real-world applications. Our work provides a theoretical framework to complement the empirical evidence provided in the above papers and analyzes the situations when the human augmentation of algorithmic outputs is beneficial.

Another stream of the related empirical literature is "judgmental adjustment of statistical forecasts" (see Arvan et al. 2019, Lawrence et al. 2006 for a review). These studies consider the demand forecast problem in supply chain management. The human analyst is allowed to adjust the forecasts generated by an algorithm. The adjustment can improve the accuracy when the algorithmic forecast is deficient or the human has important domain knowledge that is unavailable to AI (Lawrence et al. 2006). Several empirical studies aim to investigate the effect of the direction and magnitude of the adjustment on accuracy (Baker 2021, Davydenko and Fildes 2013, Fildes et al. 2009). However, the benefit of such adjustments may be highly context-dependent (Khosrowabadi et al. 2022). Although we consider a general decision-making problem, our work can also contribute to this literature by providing an analytical framework to characterize when the adjustment adds value to the algorithmic forecast.

Finally, we notice some recent works focusing on how to design user-friendly AI algorithms which the human analyst can easily understand and follow. Bastani et al. (2019) construct extracted decision trees to interpret complex, black-box AI models and summarize their reasoning process. Applied to the diabetes risk prediction problem, the proposed algorithm produces more accurate interpretations than baseline algorithms. Bastani et al. (2021) propose a reinforcement-learning algorithm for inferring interpretable tips to help workers improve their performance in sequential decision-making tasks. Through a virtual kitchen-management game, they show that the algorithm improves workers' performance. Dietvorst et al. (2018) find that giving the human analyst some control over the AI output can reduce human's aversion to algorithms.

3. An Analytical Framework for Human-Safeguarded Algorithmic Decisions

In the retail fuel example, the gas station manager intends to set prices to maximize the profit. The objective can be viewed more generally as the minimization of the loss in comparison to the optimal price. In this section, we consider a general problem that an analyst intends to minimize a loss function $l(\cdot):\mathbb{R}\to\mathbb{R}$, which measures the loss due to the deviation from the optimal decision x^*, e.g., the profit loss due to making suboptimal operations or pricing decisions. The loss function may represent the operational cost or the expected negative profit. The analyst may not know the form of the loss function exactly and seeks the help from AI algorithms. We do not impose any structure of the loss function but make the following mild assumptions.

Assumption 1. *Suppose the loss function $l(\cdot)$ satisfies: (i) $l(x)$ is nonnegative; (ii) $l(x)$ is quasiconvex with minimizer x^*.*

Part (i) of Assumption 1 is without loss of generality as the loss function can be shifted up by a constant of $|l(x^*)|$. We first give two examples that will serve as running examples throughout the rest of the paper. The two examples are intended to give the context of the loss function and demonstrate the generality of Assumption 1.

Example 1 (Predictive Analytics: Prediction). If a firm intends to forecast a quantity, for example, the demand in the next season, then the firm's problem can be cast as a prediction problem: The goal is to minimize the loss function $l(x)=(x-x^*)^2$, where x^* is the actual value of the quantity of interest.

Example 2 (Prescriptive Analytics: Pricing). Sophisticated algorithms such as online learning has been widely used in pricing (see, e.g., den Boer and Keskin 2022, Keskin et al. 2022). When a new product is launched to the market, the retailer needs to set its price x. The goal of the retailer is to maximize the profit, which is the product of the profit margin $x-c$, where c is the marginal cost and demand, i.e., $\pi(x)=(x-c)f(x)$. Denote x^* by the optimal price. The retailer knows the marginal cost, but does not know the demand function $f(x)$ nor the optimal price. The loss function can be written as $l(x) = \pi(x^*) - \pi(x)$. If the profit function $\pi(\cdot)$ is unimodal, then the loss function satisfies Assumption 1.

We next introduce the algorithmic decision and human knowledge into the framework.

Algorithmic decision. To accommodate a wide range of algorithms, we simply use a generic random variable X_a to represent the decision. The randomness may come from the randomness in the historical data or the randomization of the algorithm itself. The performance of the algorithmic decision is thus evaluated by $\mathbb{E}[l(X_a)]$.

Human knowledge. We focus on human knowledge in the form of a guardrail. That is, the human analyst forms a belief with an upper bound on the optimal decision, based on her domain knowledge and experiences. We use a random variable X_h to denote the upper bound. In Example 2, X_h could be interpreted as a price cap manually imposed by the retailer. Note that unlike the algorithmic decision, X_h is usually not data-dependent and tends to be stable, although we allow it to be random and correlated with X_a. We use the upper bound as a form of domain knowledge due to two reasons. First, it is common for human brains to perceive uncertainty in terms of intervals and worst-case scenarios. The notion is closely related to confidence intervals in statistics that have shaped how human's belief is formed. Second, compared to point estimators, the notion we propose is more flexible and allows for different confidence levels.

To keep the framework general, we do not specify how X_a and X_h are generated. For X_a, it may be output by a machine learning algorithm deployed by the analyst or a black-box commercial software as mentioned in the fuel-pricing example in the introduction. The complexity of the algorithm may vary, e.g., linear regression versus neural networks. The random variable X_a can fully capture the wide range of scenarios. For X_h, although itself may not represent a sensible decision, it may serve as a safeguard distilled from the accumulated knowledge of the human analyst. Depending on the conservativeness and the risk preference of the analyst, X_h may have different values. For instance, in Example 1, X_h may roughly be the upper confidence bound of the targeted quantity with various confidence levels.

Human-safeguarded algorithmic decision. We consider a simple yet pervasive approach to integrate the algorithmic decision and human knowledge. The human analyst safeguards the algorithmic decision by using

$$\hat{X} \triangleq \min\{X_a, X_h\}. \tag{1}$$

This is a rather natural step: the analyst follows the algorithmic decision if the upper bound is not violated; otherwise, the upper bound is used. Consider the example mentioned in the introduction (a special case of Example 2), X_a is the price output by PDI Fuel Pricing; X_h is the price cap imposed by the station manager. The safeguarded algorithmic decision takes the minimum of the two, guaranteeing that the price output by the algorithm does not exceed the price cap. This type of augmentation also captures the interaction between autonomous drones and vehicles and their human overseers (Berger 2022), in which the human overseer needs to step in and override the algorithm when the system encounters an unexpected situation.

Note that neither the algorithm nor the human analyst has access to the optimal decision x^*. If $X_h \geq x^*$ almost surely, i.e., the upper bound provided by the human belief is indeed always larger than the *true* optimal decision, then we can show that the safeguarded decision $\hat{X}$ outperforms the raw algorithmic decision X_a, i.e., $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$. To see this, note that

$$\mathbb{E}[l(\hat{X})] - \mathbb{E}[l(X_a)] = \int_{x^*}^{\infty} \int_{x_h}^{\infty} (l(x_h) - l(x_a)) f(x_a, x_h) dx_a dx_h \leq 0,$$

where $f(\cdot,\cdot)$ represents the joint PDF and the inequality follows from $l(X_a) \geq l(X_h)$ for $X_a \geq X_h \geq x^*$.

The condition $X_h \geq x^*$, however, cannot be guaranteed, because the human analyst does not have precise information about x^*. On one hand, when an unnecessary guardrail $X_h < x^*$ is imposed, the performance of X_a is hurt for $X_a \in [X_h, x^*]$. In other words, if the suggested X_h by the analyst is too aggressive, then $X_h < x^*$ is likely to happen and the human belief ends up clipping the algorithmic output X_a for too many possible scenarios, even though the latter may accurately achieve the true optimal decision x^*. Such an unnecessary guardrail inevitably introduces a significant downward bias and may cause the safeguarded algorithmic decision to be worse. The faulty human knowledge leads to an additional cost to the AI decision. On the other hand, one may argue that X_h can be a sufficiently large number, so that $X_h \geq x^*$ always holds. However, in this case, the human knowledge is almost useless in the process, as it does not provide a meaningful upper bound. The improvement by the human augmentation, if any, is going to be minimal. This is the result of an overly conservative human belief. The observation highlights the impact of aggressive/conservative human augmentation. In the next proposition, we quantify the benefit of human augmentation.

Proposition 1 **(Conditions for beneficial human augmentation)**. *Suppose Assumption 1 holds.*

(i) We can quantify the benefit of human augmentation by

$$\mathbb{E}[l(X_a)] - \mathbb{E}[l(\hat{X})] = \mathbb{E}[(l(X_a) - l(X_h))\mathbb{I}(X_h \leq X_a)]. \tag{2}$$

(ii) A sufficient condition for beneficial augmentation $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$ is

$$\mathbb{E}[l(X_a)\mathbb{I}(X_a > x^*, X_h \leq x^*)] \geq \mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)]. \tag{3}$$

(iii) A necessary condition for beneficial augmentation $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$ *is*

$$\mathbb{E}[l(X_a)\mathbb{I}(X_a \geq x^*)] \geq \mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*, X_a > x^*)]. \tag{4}$$

Next, we interpret the result of Proposition 1. In (2), the benefit of human augmentation depends on the performance of the algorithmic decision $l(X_a)$ and human knowledge $l(X_h)$ on the event that the guardrail takes effect, i.e., $\mathbb{I}(X_h \leq X_a)$. This is intuitive because the analyst counts on their knowledge to improve the algorithmic decision when it looks “unreasonable.” Conditions (3) and (4) are easier to interpret when X_a and X_h are independent, although we allow X_a and X_h to be dependent. For example, suppose the human knowledge X_h is data-independent. Then, (3) and (4) are reduced to, respectively,

$$\mathbb{E}[l(X_a)\mathbb{I}(X_a > x^*)] \geq \mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)]/\mathbb{P}(X_h \leq x^*) \tag{5}$$

$$\mathbb{E}[l(X_a)\mathbb{I}(X_a \geq x^*)] \geq \mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)]\mathbb{P}(X_a > x^*). \tag{6}$$

The left-hand sides of (5) and (6) measure the performance of the algorithmic decision. In particular, (5) says that it is beneficial to safeguard the algorithmic decision when the right-hand side $\mathbb{E}[l(X_h)|X_h \leq x^*]$ is small enough. In other words, the conditional expected loss does not explode when the human makes mistakes and imposes an overly aggressive bound ($X_h \leq x^*$). Moreover, the necessary condition (6), which is weaker than (5), implies not to safeguard the algorithmic decision when the expected loss ($\mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)]$) incurred by human belief is over a certain amount.

It is easier to check whether the human augmentation is beneficial using (5) and (6), than directly comparing (2) with zero. This is because (2) depends on the joint distribution of (X_h, X_a) and the values of $l(X_h)$ and $l(X_a)$. However, in practice, the analyst may have collected the data in the past decision epochs during which one of the algorithmic and the human decisions has been applied and their realized losses have been observed. It may not be the case that the realized $l(X_h)$ and $l(X_a)$ can be observed simultaneously. While the conditions (5) and (6) only require the marginal distributions of X_h and X_a to be observed, which allows the analyst to evaluate whether human augmentation is effective in a data-driven manner.

Furthermore, we show the tightness of the sufficient condition (5) relative to the necessary condition (6). The right-hand sides of both (5) and (6) has the common term $\mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)]$. The residual multipliers $1/\mathbb{P}(X_h \leq x^*)$ and $\mathbb{P}(X_a \geq x^*)$ in (5) and (6) tend to be constant even with increasing data sizes because in the former, human knowledge usually does not scale with big data, while in the latter, for unbiased algorithmic decisions, $\mathbb{P}(X_a \geq x^*) \approx 1/2$. So the sufficient and necessary conditions tend to only differ by a constant factor. In the next example, we show that the sufficient condition in Proposition 1 cannot be improved even when the likelihood $\mathbb{P}(X_h \leq x^*)$ diminishes, and the right-hand side of (5) cannot be relaxed to a constant multiplying $\mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)]$. As a result, the condition tends to be tight.

Example 3 (Tightness of the sufficient condition (5)). Suppose $l(x) = (x - x^*)^2$, $X_a \sim N(x^*, \sigma^2)$, and X_h satisfies

$$\mathbb{P}(X_h = x) = \begin{cases} 1 - \epsilon & \text{if } x = \infty, \\ \frac{3\epsilon}{(x-x^*)^4} & \text{if } x \leq x^* - 1. \end{cases} \tag{7}$$

Then for any $a \geq 1/4$, if $\epsilon \in (0, \sigma^2/(6a))$, $x^* < 1$, and $\sigma^2 < 3/2$, we have $\mathbb{E}[l(X_a)\mathbb{I}(X_a \geq x^*)] \geq a\mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)]$, but $\mathbb{E}[l(\hat{X})] > \mathbb{E}[l(X_a)]$.

Example 3 shows the sufficient condition in Proposition 1 no longer holds if $\mathbb{P}(X_h \leq x^*)$ in (5) is replaced by a constant. To better understand (5) and (6), we show the conditions for Example 1 (predictive analytics: prediction).

Example 4 (Conditions for beneficial augmentation for the prediction problem). Consider historical samples $Z_1, \ldots, Z_n \sim N(x^*, \sigma^2)$. The AI algorithm estimates x^* by the sample mean $X_a = \frac{1}{n}\sum_{i=1}^n Z_i$, which follows the distribution $N(x^*, \sigma^2/n)$. By Proposition 1, if the upper bound derived from the human belief satisfies $\mathbb{E}[(X_h - x^*)^2 | X_h \leq x^*] \leq \sigma^2/(2n)$, then the augmentation improves the algorithmic decision. On the other hand, if $\mathbb{E}[(X_h - x^*)^2\mathbb{I}(X_h \leq x^*)] \geq \sigma^2/n$, then the augmentation is not beneficial.

Proposition 1 provides us with an analytical framework to analyze the benefit of augmentation. Based on the framework, we can show that there is an optimal level of safeguard when the human belief is deterministic.

Corollary 1 **(Optimal safeguard)**. *Suppose $X_h = x_h$ is a constant. Then the benefit of augmentation $\mathbb{E}[l(X_a)] - \mathbb{E}[l(\hat{X})]$ is unimodal in x_h, i.e., it increases when $x_h \leq x^*$ and decreases when $x_h \geq x^*$.*

Corollary 1 holds under the condition that the human belief is deterministic. In this case, although the human analyst imposes an upper bound, it is the best to equate it to the true optimal decision x^*, i.e., a buffer is not necessary. Of course, the corollary cannot provide a guidance for the human analyst to select the optimal bound, because x^* is not accessible. It does show the trade-off between conservative/aggressive guardrails.

Symmetrically, we can derive similar results when the guardrail derived from the human domain knowledge takes the form of a lower bound.

Corollary 2 **(Safeguarded by a lower bound)**. *Consider $\hat{X} = \max\{X_a, X_h\}$. We have (i) $\mathbb{E}[l(X_a)] - \mathbb{E}[l(\hat{X})] = \mathbb{E}[(l(X_a) - l(X_h))\mathbb{I}(X_h \geq X_a)]$. (ii) A sufficient condition for $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$ is*

$$\mathbb{E}[l(X_a)\mathbb{I}(X_a < x^*, X_h \geq x^*)] \geq \mathbb{E}[l(X_h)\mathbb{I}(X_h \geq x^*)]. \tag{8}$$

And (iii) a necessary condition for $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$ is

$$\mathbb{E}[l(X_a)\mathbb{I}(X_a \leq x^*)] \geq \mathbb{E}[l(X_h)\mathbb{I}(X_h \geq x^*, X_a < x^*)]. \tag{9}$$

Next we extend the results by two-sided bounds. In particular, suppose the human analyst imposes both lower and upper bounds on the algorithmic decision. For example, in the motivating example in the introduction, the station manager may propose a range for the retail price: the markup has to be between \$0.10/L and \$0.30/L, regardless of the recommendation of the algorithm. Mathematically, the human belief is translated to an interval $[X_h^l, X_h^u]$. The algorithmic decision X_a is then projected onto the interval, i.e., $\hat{X} = \min\{\max\{X_a, X_h^l\}, X_h^u\}$. Proposition 2 characterizes the benefit of such augmentation.

Proposition 2 **(Benefit of safeguarding using a two-sided bound)**. *Suppose Assumption 1 holds.*

(i) The benefit of the human safeguard by a two-sided bound is

$$\mathbb{E}[l(X_a)] - \mathbb{E}[l(\hat{X})] = \mathbb{E}[(l(X_a) - l(X_h^l))\mathbb{I}(X_a \leq X_h^l)] + \mathbb{E}[(l(X_a) - l(X_h^u))\mathbb{I}(X_a \geq X_h^u)].$$

(ii) A sufficient condition for $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$ is

$$\begin{aligned}\mathbb{E}[l(X_a)\mathbb{I}(X_a \geq x^*, X_h^u \leq x^*)] + \mathbb{E}[l(X_a)\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)]&\\ \geq \mathbb{E}[l(X_h^u)\mathbb{I}(X_h^u \leq x^*)] + \mathbb{E}[l(X_h^l)\mathbb{I}(X_h^l \geq x^*)].&\end{aligned} \tag{10}$$

(iii) A necessary condition for $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$ is

$$\mathbb{E}[l(X_a)] \geq \mathbb{E}[l(X_h^u)\mathbb{I}(X_h^u \leq x^* \leq X_a)] + \mathbb{E}[l(X_h^l)\mathbb{I}(X_a \leq x^* \leq X_h^l)]. \tag{11}$$

If the bounds satisfy $\mathbb{P}(X_h^l \leq x^* \leq X_h^u) = 1$, i.e., they always enclose the actual optimal decision, then the safeguard always improves the algorithmic decision. When this condition fails, (10) and (11) imply conditions to check whether to safeguard the algorithmic decision. Intuitively, the safeguard is beneficial when the loss incurred by the interval not covering x^* is relatively small compared to the loss of the algorithmic decision.

One can see that Proposition 2 reduces to Proposition 1 and Corollary 2 when $X_h^l = -\infty$ or $X_h^u = \infty$. We point out that the conditions for two-side bounds are weaker than the conditions for the one-sided bound. If X_h^u satisfies (3) and X_h^l satisfies (8), then $[X_h^l, X_h^u]$ satisfies (10). But the reverse is not true. So the two-side conditions allow the human to make more mistakes in one side as long as the loss can be compensated by the other.

3.1. Covariate Information

So far, we have considered a simple model that the environment does not provide any covariate information at the specific decision epoch. However, in many data-driven decision-making problems, the analyst may observe additional covariate information W and hence the optimal decision x^*

can depend on such covariate information. Upon observing the covariates, the algorithm outputs a decision $X_a(W)$. In the prediction problem (Example 1), one can think of W as the new input to the prediction algorithm such as weather conditions. In the pricing problem (Example 2), W may represent the available side information about the market to assist the choice of the optimal price. For example, PDI Fuel Pricing would take into account the crude oil price, which is a major cost component, to determine the retail gas price. In this case, the crude oil price is changing over time and can be considered as part of the covariate information.

After receiving the algorithmic recommendation, the human analyst comes up with a bound $[X_h^l(W), X_h^u(W)]$ to safeguard it. That is, $\hat{X}(W) = \min\{\max\{X(W), X_h^l(W)\}, X_h^u(W)\}$. Note that in many cases the human domain knowledge may not be sophisticated enough to adapt to a specific covariate W. In such cases, X_h^l and X_h^u do not depend on W, which is also covered by our framework.

When the covariate information is available, the loss function $l(x,w)$ depends on both the decision and the covariate. We impose the following assumption, in parallel to Assumption 1.

Assumption 2. *Assume the loss function $l(x,w)$ satisfies the following conditions.*

(i) For any decision $x \in \mathbb{R}$ and any covariate $w \in \mathbb{R}^d$, $l(x,w) \geq 0$.

(ii) For any covariate $w \in \mathbb{R}^d$, $l(\cdot, w)$ is quasi-convex with minimizer $x^*(w)$.

Next, we characterize the benefit of human augmentation in the presence of covariate information by generalizing Proposition 1. Note that X_h, X_a, and x^* all depend on (and are correlated with) W. We omit the dependence for the readability.

Proposition 3 **(Benefit of human augmentation with covariate information)**. *Suppose Assumption 2 holds.*

(i) The benefit of human augmentation is

$$\begin{aligned}\mathbb{E}[l(X_a, W)] - \mathbb{E}[l(\hat{X}, W)] = &\ \mathbb{E}[(l(X_a, W) - l(X_h^l, W))\mathbb{I}(X_a \leq X_h^l)] \\ &+ \mathbb{E}[(l(X_a, W) - l(X_h^u, W))\mathbb{I}(X_a \geq X_h^u)]. \end{aligned} \tag{12}$$

(ii) A sufficient condition for $\mathbb{E}[l(\hat{X}, W)] \leq \mathbb{E}[l(X_a, W)]$ is

$$\begin{aligned}&\mathbb{E}[l(X_a, W)\mathbb{I}(X_a \geq x^*, X_h^u \leq x^*)] + \mathbb{E}[l(X_a, W)\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)] \\ &\qquad \geq \mathbb{E}[l(X_h^u, W)\mathbb{I}(X_h^u \leq x^*)] + \mathbb{E}[l(X_h^l, W)\mathbb{I}(X_h^l \geq x^*)]. \end{aligned} \tag{13}$$

(iii) A necessary condition for $\mathbb{E}[l(\hat{X}, W)] \leq \mathbb{E}[l(X_a, W)]$ is

$$\mathbb{E}[l(X_a, W)] \geq \mathbb{E}[l(X_h^u, W)\mathbb{I}(X_h^u \leq x^*, X_a \geq x^*)] + \mathbb{E}[l(X_h^l, W)\mathbb{I}(X_h^l \geq x^*, X_a \leq x^*)]. \tag{14}$$

When W is a constant, Proposition 3 reduces to the special case of Proposition 2. As expected, the conditions in Proposition 3 are more involved, although they follow a similar form to Proposition 2. To explain the intuition, we adopt the following example.

EXAMPLE 5 (LINEAR REGRESSION). Linear regression is a special case of the prediction problem (Example 1) with covariates. Suppose the loss function is $l(x,w)=(x-w^\top\beta)^2$ for some unknown coefficient β. As a result, the optimal decision is $x^*(W)=W^\top\beta$. Using the least squares estimator $\hat{\beta}$, the algorithmic output is $X_a(W)=W^\top\hat{\beta}$. In the necessary condition (14), the left-hand side is the mean-squared error (MSE) of the least-squares estimation, which typically converges to zero at the rate of $1/n$, where n is the sample size. In this case, for the human augmentation to outperform the algorithm, the right-hand side of (14) should diminish at the same or a faster rate. It is only possible if the bounds derived from human belief, X_h^l and X_h^u, have diminishing MSEs $l(X_h^l,W)$ and $l(X_h^u,W)$, or they almost always sandwich the optimal x^*, i.e., $X_h^l\leq x^*\leq X_h^u$. Both of the above two requirements set impractically high bars for the human domain knowledge.

From the example, we see that AI incurs diminishing loss as it gathers more data. In a data-rich environment, it appears that the human domain knowledge is not likely to improve AI. However, Example 5 does not reflect one of the major reasons why human domain knowledge may be helpful: algorithms designed for general purposes sometimes ignore practical factors in the training dataset, such as contamination, model misspecification, and data errors. In the following sections, we provide examples to show that even if AI has a large amount of data, the human knowledge can still play an important role and contribute to the decision-making.

4. Three Use Cases on Beneficial Human Augmentation

In this section, we provide three concrete use cases when the safeguard derived from human knowledge can indeed improve algorithmic decisions even with large data, despite the potential limitation of human knowledge illustrated in Example 5. We first provide a summary of the three use cases as follows:

- In Section 4.1, we consider a pricing problem (Example 2) under competition. We show that when the algorithm fails to take into account the competitive environment the pricing problem resides in, simple human augmentation like price matching can improve the algorithmic decision.
- In Section 4.2, we consider a pricing problem when the algorithm misspecifies the demand function. We show that using empirical observations (in particular, setting a price interval using the historical price range that contains the highest profit in the past) can improve the performance of the algorithm.

- In Section 4.3, we show that in prediction problems (Example 1), the human knowledge can serve as a robust mechanism to limit the damage due to data contamination and thus improve the algorithm's performance.

4.1. Pricing Algorithm under Competition

In this section, consider the pricing problem of a focal firm when there is a competitor in the market. The loss function of the firm under price p, given the price of the competitor p', is the negative revenue under a linear demand function:

$$l(p) = \mathbb{E}[-pd(p,p')] := -p(\alpha - \beta p + \gamma p').$$

We assume $\beta > \gamma > 0$, which is standard in the literature and means that the firm's demand is more sensitive to its own price than its competitor's price. The best response of the firm, given the competitor's price, can be easily solved as:

$$p^* = \arg\min_p l(p) = \frac{\alpha + \gamma p'}{2\beta}.$$

However, this best response requires the knowledge of α, β, γ, which are typically unavailable to the firm. Next we specify how an algorithm may recommend a price based on the historical data.

Algorithmic price. The algorithm attempts to learn the demand function from the historical data. However, the algorithm may not be aware of the presence of the competitor (see, e.g., Cooper et al. 2015). Consider the gas station example in the introduction. To provide the competitors' prices as inputs to PDI Fuel Pricing, the station manager needs to check the prices of nearby gas stations periodically. Even though this is convenient, the resolution of the competitors' prices may be lower than that of the historical prices of the focal station, which have constantly been recorded in its system. To accommodate this realistic setting with data unavailability, we assume that the algorithm attempts to learn a monopolistic demand function

$$\hat{d}(p) = \hat{\alpha} - \hat{\beta} p. \tag{15}$$

This setting of learning a monopolistic demand function under competition has also been studied in Cooper et al. (2015) and Hansen et al. (2021).

The algorithm has access to the historical prices $p_1, \dots, p_n$ and realized demand $d_1, \dots, d_n$. We assume that the demand is generated by

$$d_t = \alpha - \beta p_t + \gamma p'_t + \epsilon_t, \tag{16}$$

for some independent and identically distributed (i.i.d.) noise ϵ_t and $t = 1, \dots, n$. The algorithm uses the ordinary least squares (OLS) to estimate $\hat{\alpha}$ and $\hat{\beta}$. Finally, the algorithm recommends a price maximizing the estimated revenue (16), i.e., $p_a = \hat{\alpha}/2\hat{\beta}$. Note that the historical demand and the algorithmic price p_a depend on the unobserved competitor's prices $p'_1, \dots, p'_n$. To analyze p_a, we assume that

Assumption 3. *The prices* (p_n, p'_n) *are i.i.d. for* $n = 1, 2, \dots$. *Moreover,* $\mathbb{E}[p_n] = \mathbb{E}[p'_n] = \mu$, $\mathrm{Var}(p_n) = \mathrm{Var}(p'_n) = \sigma^2$ *and the correlation of* (p_n, p'_n) *is* $\rho \in [0,1]$.

Assumption 3 can be rather mild if we consider a symmetric duopoly and the past samples are all independent. The condition $\rho \geq 0$ implies that the prices of competing firms are positively correlated. The next result characterizes the asymptotic behavior of p_a.

Lemma 1. *Suppose Assumption 3 holds. The algorithmic price* p_a *converges in probability to*

$$\operatorname*{plim}_{n\to\infty} p_a = \frac{\alpha + \gamma\mu(1-\rho)}{2(\beta - \gamma\rho)}. \tag{17}$$

To understand the algorithmic price with large data ($n \to \infty$), note that the symmetric Nash equilibrium price satisfies $p_{NE} = \arg\max_p \{pd(p, p_{NE})\} = \alpha/(2\beta - \gamma)$. If $\rho = 1$, i.e., the historical prices of the firm and the competitor are perfectly correlated, then $p_a = \alpha/2(\beta - \gamma)$ converges to the collusive price that maximizes the joint revenue of both parties. Clearly such collusive price is higher than p_{NE}. On the other hand, if $\rho = 0$ and $\mu = p_{NE}$, i.e., the historical prices of the firm and the competitor are uncorrelated and centered around the Nash equilibrium, then p_a converges to p_{NE}.

Human safeguard — price matching. We consider price matching, a common competitive strategy for analysts. In particular, after receiving the price recommended by the algorithm, the analyst may intentionally check the competitor's price. If the competitor's price is lower than algorithmic price, then the analyst lowers the algorithmic price to match the competitor's price. That is, the human-safeguarded price is $\hat{p} = \min\{p_a, p'\}$. Such an augmentation strategy is highly relevant for the human analyst: (i) it does not depend on the historical data or the unknown parameters, (ii) is easy to process and explain to human managers, (iii) it takes into account the competitive environment, and (iv) can be used to complement the algorithmic price. Moreover, since price matching specifies a lower bound, it also fits into the general framework in Section 3. In the next theorem, we characterize the condition under which the human augmentation improves the algorithmic price.

Theorem 1. *Suppose Assumption 3 holds and the algorithmic price is given in* (17). *Assume* $\mu \geq p_{NE}$. *If the competitor's price satisfies*

$$p' \geq p_L := \frac{\alpha\beta - 2\alpha\gamma\rho - \beta\gamma(1-\rho)\mu}{2(\beta - \rho\gamma)(\beta - \gamma)} \in (0, p_{NE}),$$

then the revenue of the safeguarded price is higher than that of the algorithmic price, i.e., $\hat{p}d(\hat{p}, p') \geq p_a d(p_a, p')$. *In addition, if* $p' \in (p_L, p_a)$, *then the revenue improvement is strictly positive.*

To understand this theorem, first note that the assumption $\mu \geq p_{NE}$ is mild. It merely states that the historical prices are a convex combination of the equilibrium price and the collusive price, since the latter is higher. The expression for p_L is complicated, but we can show that $p_L \leq p_{NE}$. As long as the competitor's price is not significantly lower than the equilibrium price, the augmentation by price matching improves the algorithmic decision. Price matching is particularly useful when the competitor undercuts the algorithmic price ($p' < p_a$) while not setting a price much lower than the equilibrium price ($p' > p_L$).

4.2. Misspecified Algorithms

Consider the pricing problem in Example 2 in a monopolistic market. The demand function is assumed to be a general non-increasing function $f(p)$, and the unit cost of the product is c. As a result, the loss function faced by the analyst is $l(p) = -(p-c)f(p)$. The optimal price p^* satisfies $p^* = \arg\min_p l(p)$. Since the demand function f is unknown to the analyst, she sets up price experiments to collect data and uses AI algorithms to learn the demand function.

Price experiments. In practice, the analyst cannot charge prices arbitrarily. In fear of consumer backlash, the price experimentation usually takes the form of promotions, such as $10-off coupons. The analyst has done price experimentation at a grid of prices and observed realized demand at those price points. In particular, we consider the following uniform price grid between $[c, \bar{p}]$:

$$p_j = c + j\frac{\bar{p}-c}{n}, \quad j = 0, 1, \ldots, n, \tag{18}$$

where $\bar{p}$ represents the nominal price without any promotion. For each price on the grid, we suppose K noisy demand observations have been collected:

$$f(p_j) + \epsilon_{jk}, \ \forall k = 1, 2, \ldots, K,$$

where ϵ_{jk} is an independent σ-sub-Gausssian noise. For example, the firm may have set the price p_j for K hours and recorded the hourly demand, whose mean is $f(p_j)$ with noise ϵ_{jk}.

Algorithmic decision. In order to find the optimal price p^*, the algorithm needs to learn the demand function $f(p)$. However, because the price experiments are only conducted on a grid, the algorithm typically postulates a model for the demand function and estimate the model parameters. One of the most common models is the linear demand function. That is,

$$\hat{f}(p) = \hat{\alpha} - \hat{\beta}p, \tag{19}$$

where $\hat{\alpha}$ and $\hat{\beta}$ guarantees that $\hat{f}$ is the best linear fit to the points $\{(p_j, f(p_j) + \epsilon_{jk})\}$ for $j = 0, \ldots, n$ and $k = 1, \ldots, K$ in terms of the ℓ_2 error:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \left\{ \sum_{j=0}^{n} \sum_{k=1}^{K} \left(f(p_j) + \epsilon_{jk} - \alpha + \beta p_j\right)^2 \right\}. \tag{20}$$

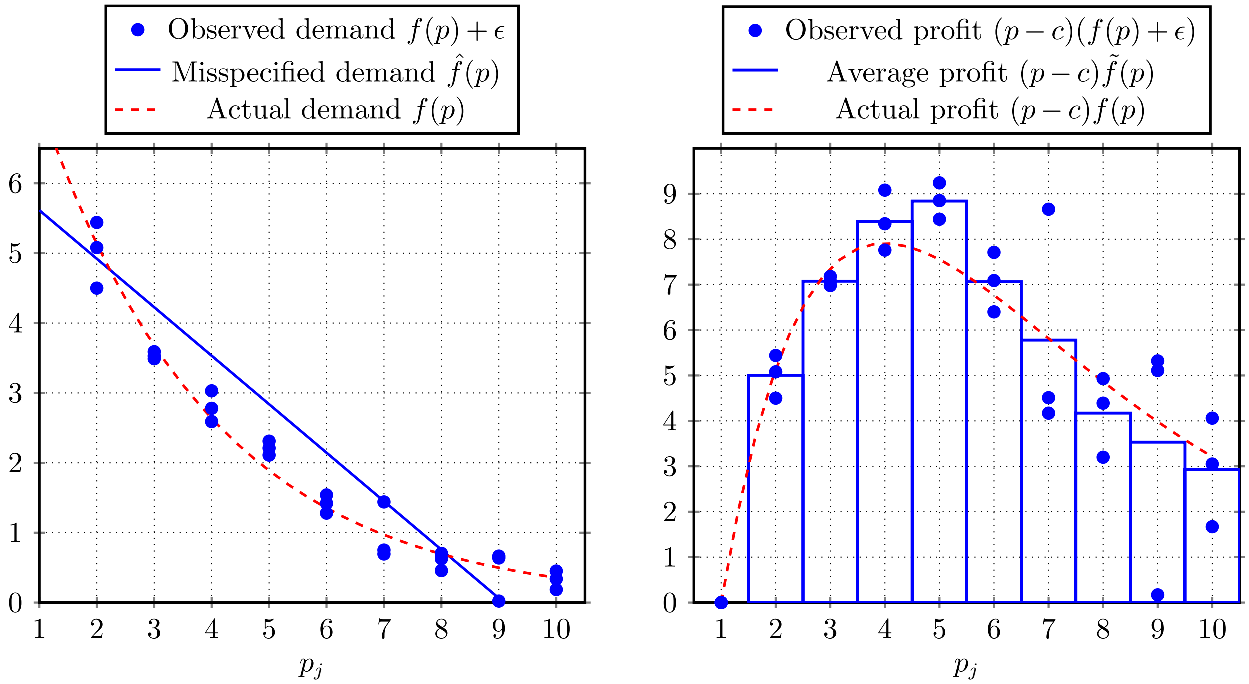

Figure 1 **The left figure shows the misspecified linear model by the algorithm. The actual demand function is $f(p) = 10\exp(-p/3)$ with $c = 1$, $\bar{p} = 10$, $n = 10$, and $K = 3$. The OLS estimator is $\hat{\alpha} = 6.309$, $\hat{\beta} = -0.694$. The right figure shows the human analyst identifying the price that earns the highest empirical profit.**

Eventually, the algorithm outputs an optimized price p_a based on the estimated demand function $\hat{f}(p)$, i.e., $p_a = \arg\max_p (p - c)\hat{f}(p) = \hat{\alpha}/2\hat{\beta} + c/2$.

Although linear demand models have been shown in Cohen et al. (2021) to perform well when the model is misspecified, in our setup, we do not claim that the linear model is necessarily the best algorithmic choice in this scenario. In fact, given the data points $\{(p_j, f(p_j) + \epsilon_{jk})\}$, there may be other choices such as a linear interpolation that can fit the demand function better. Our main goal is to demonstrate a salient feature of a wide range of algorithms based on parametric statistical models—the chosen model may be misspecified and hence the resulting algorithm may be built on a shaky foundation. That is, the relationship between price and demand may not be accurately captured by the postulated class of models. In this case, even with sufficient data, the misspecification cannot be fully remedied. As we shall see in the left of Figure 1, such a misspecification can be clearly pronounced for the linear model.

Human knowledge. How can human knowledge help with the misspecified algorithm? Due to the limitation of the human brain, the human analyst typically does not form a model to process the historical data, and it would be impossible to judge whether the algorithmic price suffers from misspecification. We consider a rather natural and straightforward approach: since the analyst observes the noisy demand on the price grid, it first uses the average to form an estimate of the demand at each price as

$$\tilde{f}(p_j) := f(p_j) + \frac{1}{K}\sum_{k=1}^{K} \epsilon_{jk}. \tag{21}$$

Using this estimate, the optimal price on the grid that generates the highest empirical profit $(p_j - c)\tilde{f}(p_j)$ can be easily calculated. Suppose j^* is the index of one of the optimal prices on the grid:

$$(p_{j^*} - c)\tilde{f}(p_{j^*}) \geq (p_j - c)\tilde{f}(p_j) \quad \forall j = 0, \ldots, n.$$

Taking the demand function in Figure 1 as an example, the human analyst observes the noisy demand on the price grid $\{1, 2, \ldots, 10\}$. Then, she chooses the price point $p_{j^*} = 5$ that gives the highest empirical profit. In this example, the chosen price $p_{j^*} = 5$ does not equal to the true optimal price $p^* = 4$, but the neighborhood $[p_{j^*-1}, p_{j^*+1}] = [4, 6]$ includes p^*. It is easy to see the complementary effects of human knowledge and the algorithm in this example. The optimal price from the human knowledge is empirically validated without any statistical model. However, while the algorithm may suffer from misspecification, it has two strengths unmatched by the human analyst. First, the algorithm aggregates all $(n+1)K$ demand observations while the human analyst takes the average of K demand observations locally. It is well-known that more samples improve the statistical prediction power. Second, the human analyst does not attempt to specify a model to extrapolate the demand function. As a result, only the prices on the grid can be selected, and a discretization error is always born by the price picked by the analyst. For example, if the prices $\{\bar{p} - 20, \bar{p} - 10, \bar{p}\}$ have been experimented, i.e., two types of promotions, \$20 off and \$10 off, in addition to the nominal price $\bar{p}$, have been offered in the past, then p_{j^*} can be suboptimal if the actual optimal price is $\bar{p} - 15$. In this case, the algorithm learns a model that interpolates the price gaps on the grid and remedies the discretization error. The following result characterizes the performances of the two approaches, which allow us to further understand the benefit of human augmentation to the algorithm.

Proposition 4. *Assume that the loss function $l(p)$ is strongly convex with parameter λ (or equivalently, the profit function is λ-concave), i.e.,*

$$l(p) \leq l(p') - l'(p')(p - p') - \frac{\lambda}{2}(p - p')^2, \quad \forall p, p' \in [c, \bar{p}].$$

We then have

(i) [Algorithmic decision] Let p_a^* denote the optimal price for the misspecified linear demand. Given n and K, we have

$$\mathbb{P}(|p_a - p_a^*| \geq \delta) \leq 4\exp(-bnK), \tag{22}$$

where b is a constant independent of n and K.

(ii) [Human knowledge] The probability of the true optimal price not falling into the neighborhood of human's estimated price p_{j^*} satisfies

$$\mathbb{P}\left(p^* \notin [p_{j^*-1}, p_{j^*+1}]\right) \leq 2(n+1)\exp\left(-\frac{K\lambda^2(\bar{p}-c)^4}{32\sigma^2\bar{p}^2n^4}\right). \tag{23}$$

Note that for large samples ($K \to \infty$ or $n \to \infty$), the algorithmic decision does not converge to the true optimal price p^*. Instead, it converges to the optimal price for the misspecified linear demand. For the human knowledge, without misspecification, the price neighborhood $[p_{j^*-1}, p_{j^*+1}]$ on the grid containing p_{j^*} will eventually include the true optimal price as $K \to \infty$. However, compared to the algorithmic decision, the human's error probability can be significantly inflated, which even increases in the number n of price points, reflecting a lack of data efficiency, while the algorithm's error probability decreases in n.

Augmentation by safeguarding. Because of the weaknesses of the finite-sample results in Proposition 4 (i.e., the convergence to the wrong target in (22) by the algorithm and inefficient sample use in (23) by the human), the analyst may decide to integrate both approaches. Based on p_{j^*}, the human analyst imposes a guardrail $[p_{j^*-1}, p_{j^*+1}]$ as in Proposition 2. In other words, the human analyst uses the two neighboring prices of p_{j^*} on the grid to form an interval to regulate the algorithmic output. As a result, the safeguarded algorithmic price is

$$\hat{p} = \max\{\min\{p_a, p_{j^*+1}\}, p_{j^*-1}\}.$$

The following result characterizes the condition under which such an augmentation is beneficial.

Theorem 2. *Assume the profit function $(p-c)f(p)$ is unimodal. The augmentation improves the algorithmic price, i.e., $(\hat{p}-c)f(\hat{p}) \geq (p_a - c)f(p_a)$, if the true optimal price $p^* \in [p_{j^*-1}, p_{j^*+1}]$. In particular, the latter condition always holds when $K \to \infty$.*

Theorem 2 requires the profit function to be unimodal, which is satisfied by most demand functions $f(\cdot)$ (see, e.g., Ziya et al. 2004). As a result, the regime that sees the most benefit of human augmentation is when n is fixed but $K \to \infty$, i.e., the price experimentation is conducted on a few prices for an extended period. This is arguably a common scenario in retailing, due to the infeasibility of frequent price changes. In this case, augmenting the algorithmic price using the bounds distilled from the human knowledge, the augmented price $\hat{p}$ enjoys the benefits of both worlds. Intuitively, when p^* falls in the interval $[p_{j^*-1}, p_{j^*+1}]$ and the algorithmic price is outside the interval, the human safeguard always pulls the algorithmic price toward the actual optimal price and improves the algorithmic recommendation due to the unimodality of the profit function. On the other hand, when the algorithmic price falls into the same interval, indicating the discretization error may exceed the misspecification error (imagine $[p_{j^*-1}, p_{j^*+1}]$ being a wide interval), the guardrail does not take effect and the analyst follows the algorithmic decision. This result confirms the complementary effects of algorithms and human knowledge, in particular, the robustness of simple heuristics against model misspecification.

We next study two commonly used demand function forms of $f(\cdot)$ and characterize the conditions under which the algorithmic price falls outside the interval $[p_{j^*-1}, p_{j^*+1}]$, i.e., when it is strictly improved by the human augmentation. We consider $K \to \infty$ in both examples.

Example 6 (Isoelastic demand). Consider the demand function $f(p) = bp^{-a}$ where $a > 1$ and $b > 0$. It can be shown that the profit function is unimodal, and the optimal price is $p^* = \frac{ac}{a-1}$. We can show that when the nominal price

$$\bar{p} > \frac{\frac{a}{a-1} - \frac{1}{2} - \frac{2}{n}}{\frac{1}{3} - \frac{2}{n}} c, \tag{24}$$

the algorithmic price p_a is outside the interval $[p_{j^*-1}, p_{j^*+1}]$ and thus the human augmentation strictly improves the algorithm. For example, if $a = 2$, $n = 10$, then (24) is equivalent to $\bar{p} > 2.25c$, i.e., when the nominal price is more than 125% higher than the production cost.

Example 7 (Exponential demand). Consider the demand function $f(p) = be^{-ap}$ where $a > 0$. It can be shown that the optimal price is $p^* = 1/a + c$. We can show that when

$$\bar{p} > \frac{\frac{1}{a} + (\frac{1}{2} - \frac{2}{n})c}{\frac{1}{3} - \frac{2}{n}}, \tag{25}$$

the human augmentation strictly improves the algorithm. For example, if $a = 2$, $n = 10$, then (25) is translated to $\bar{p} > 3.75 + 2.25c$.

From both examples, we can see that the misspecification error by the algorithm grows larger relative to the discretization error by the guardrail and hence human's augmentation becomes more beneficial, when $\bar{p}$ is much larger than c, i.e., the interval for the price experiments is wider.

4.3. Data Contamination

In this section, we consider the case when the data can be contaminated, due to outliers, reporting errors, etc. While the possibly contaminated data is fed into algorithms, the error in data also propagates to the output decision. For this reason, Fogliato et al. (2022) advocate humans-in-the-loop, to mitigate the data contamination. Human analysts are less susceptible to data contamination because the human brain cannot process large data sets, which turns out to be a blessing rather than a curse in this case as it makes the human knowledge robust to minor data contamination. Next we provide a formal analysis of human augmentation to the algorithms for this use case.

The application we consider is the linear regression problem in Example 5. Without contamination, the historical data $\{(X_i, W_i)\}_{i=1}^n$ is generated by $X_i = W_i^\top \beta + \epsilon_i$ for some unknown coefficients β and noise ϵ_i. We consider two contamination mechanisms that affect a fraction of samples: contamination in response (Bhatia et al. 2017) and covariates (Loh and Wainwright 2011, McWilliams et al. 2014). Before we provide the formal introduction to the mechanisms, we first explain how the algorithm and human knowledge play a role in the process.

Not able to tell whether the data is contaminated, the algorithm simply applies the OLS estimator to the data[2], as stated in Example 5. For a new covariate W, we denote the prediction from the OLS estimator as $X_a(W)$. For the human analyst, we consider a generic two-sided bounded range $[X_h^l, X_h^u]$. As shown in Section 3, the corresponding safeguarded decision is $\hat{X}(W) = \max\{\min\{X_a(W), X_h^u\}, X_h^l\}$ for the two-sided guardrail. Also recall that the loss function is $l(x,w) = (x - w^\top \beta)^2$.

4.3.1. Contamination in Response We first consider the contamination in the response of the samples. In particular, the observed response X_i is not generated from $W_i^\top \beta + \epsilon_i$, but

$$X_i = W_i^\top \beta + B_i + \epsilon_i \tag{26}$$

for some random variable B_i. Here B_i controls for the degree of contamination: with a high probability, it is zero and the sample is not contaminated. When $B_i \neq 0$, the response of the sample, X_i, deviates from the uncontaminated observation $W_i^\top \beta + \epsilon_i$. Note that B is not to be confused with ϵ, which has a zero mean. We assume $\mathbb{E}[B]$ to be nonzero which means that the contamination has a systematic influence on the estimation.

Contamination in the response is studied in the computer science community, see, e.g., Wright et al. (2009) and Nguyen and Tran (2013) for applications to image recognition. For management applications, this type of contamination may occur due to various reasons: the historical response X_i may be subject to reporting errors, or may be censored in certain periods and some ad hoc imputation methods are used so that the missing data are replaced by their estimated values. Our contamination model can capture both cases.

Not surprisingly, when the historical data is contaminated, the algorithmic output that uses OLS is biased. More precisely, given the i.i.d. historical samples $\{(X_i, W_i)\}_{i=1}^N$ that are generated by (26) and a new covariate w, suppose $X_a(w)$ is the OLS estimator applied to w. We have:

Lemma 2. *Assume the covariance matrix $\Sigma = \mathbb{E}[WW^\top]$ is positive definite. Then the OLS predictor $X_a(w) = w^\top \hat{\beta}$ converges to $w^\top \beta + \mathbb{E}[B]$ in probability for any w as $N \to \infty$. Therefore, $l(X_a(w), w) = \mathbb{E}[B]^2$.*

[2] We acknowledge that there are statistical tests to identify outliers and robust estimators to mitigate data contamination. We do not consider them in the model because they usually require some information about the contamination such as whether the data is contaminated or the contamination mechanism, while in practice, the algorithm is agnostic to such knowledge.

In other words, even with a sufficiently large dataset, the bias caused by the contamination persists. To analyze how the human augmentation may improve the algorithmic result, we consider a simple form of contamination. Let $B = b > 0$ with probability p and $B = 0$ with probability $1 - p$. Therefore, the contamination always leads to upward bias (the downward bias can be formulated similarly), and the parameter b represents the magnitude of the contamination and p represents its propensity. To adjust for the upward bias, the human analyst can impose an upper bound X_h^u and cap the output of the algorithm and lead to the safeguarded prediction as $\hat{X}(W) = \min\{X_a(W), X_h^u\}$. However, because X_h^u is not data-driven, the safeguard runs the risk of overcorrection. The following proposition provides such a condition.

Proposition 5. *Assume the domain of W is a closed and bounded set $\mathcal{W} \in \mathbb{R}^d$, and we take $N \to \infty$. If X_h^u satisfies*

$$X_h^u \geq \max_{w \in \mathcal{W}}\{w^\top \beta\} - pb, \tag{27}$$

then, for all $w \in \mathcal{W}$, we have $l(\hat{X}(w), w) \leq l(X_a(w), w)$.

To interpret the result, on the one hand, in the extreme case of $X_h^u \geq \max_{w \in \mathcal{W}}\{w^\top \beta\} + pb$, we always have $X_h^u \geq X_a$ and $\hat{X} = X_a$. The bound of X_h^u is too conservative, and the safeguard provides no benefit. On the other hand, one can show that the loss function for the algorithmic prediction is simply $l(X_a(w), w) = p^2 b^2$ due to the contamination. If $X_h^u < \max_{w \in \mathcal{W}}\{w^\top \beta\} - pb$, then there exists w such that $l(\hat{X}(w), w) > p^2 b^2$ because the upper bound imposed by the human analyst is too aggressive and outweighs the bias introduced by the contamination. Clearly, condition (27) is easier to satisfy when the contamination gets more severe due to an increased value of p or b.

When the contamination can lead to a bias of either direction, i.e., it is possible that $B > 0$ or $B < 0$, it is safer for the human analyst to set up both bounds X_h^l and X_h^u. That is, once the algorithmic prediction $X_a(w)$ is given, the safeguarded decision is $\hat{X}(w) = \max\{\min\{X_a(w), X_h^u\}, X_h^l\}$. Generalizing Proposition 5, we have:

Theorem 3. *Suppose the domain of W is a closed and bounded set $\mathcal{W} \in \mathbb{R}^d$, and we take $N \to \infty$. If the lower and upper bounds X_h^l, X_h^u satisfy*

$$X_h^u \geq \max_{w \in \mathcal{W}}\{w^\top \beta\} - \left|\mathbb{E}[B]\right|, \quad X_h^l \leq \min_{w \in \mathcal{W}}\{w^\top \beta\} + \left|\mathbb{E}[B]\right|, \tag{28}$$

then for all $w \in \mathcal{W}$, we have $l(\hat{X}(w), w) \leq l(X_a(w), w)$ and $\mathbb{E}[l(\hat{X}(W), W)] \leq \mathbb{E}[l(X_a(W), W)]$.

Note that when the absolute bias ($\left|\mathbb{E}[B]\right|$) is large, the human augmentation of imposing upper and lower bounds tends to be helpful, regardless of the sign of the bias. For example, even when $B > 0$, i.e., the bias is always upward, Theorem 3 states that imposing a lower bound X_h^l is more likely to be beneficial when the bias is large because (28) is more likely to be satisfied. To see the intuition, as the contamination becomes more severe, the algorithmic decision is subject to a larger bias. Hence, it is easier for the human augmentation to outperform the raw algorithmic decision.

4.3.2. Contamination in Covariates In this section, we consider the covariates in the data, W_i, which can be contaminated. This type of contamination is sometimes referred to as the error-in-variable (Loh and Wainwright 2011), which occurs in voting, surveys, and sensor networks. In business applications, there may exist measurement errors in the historical samples of the covariate. For example, when a firm is running a survey to learn consumer sentiment, the design of the survey may lead to biased measurement of the quantity of interest.

We consider the following contamination model: the observed covariate is generated by $W_i = Z_i + U_i$, where Z_i is the actual covariate, and $U_i \in \mathbb{R}^d$ is an error that contaminates the observation, independent of Z_i. The response is generated from $X_i = Z_i^\top \beta + \epsilon_i$. For a new covariate W_0, the algorithm outputs $X_a(W_0)$ using the OLS estimator from the data $\{(X_i, W_i\}_{i=1}^\infty$. (We consider infinite samples in this analysis.)

What differentiates the contamination in covariates from Section 4.3.1 is that even the new covariate W_0 itself may be contaminated. Therefore, the human safeguard can serve for two purposes: it helps to control the contamination in the training data and curtail the potential error in the new covariate based on which the prediction is given. We impose the following technical assumption.

Assumption 4. *The matrix $\Sigma_1 := \mathbb{E}[ZZ^\top]$ is positive definite and $\Sigma_2 := \mathbb{E}[UU^\top]$ is positive semi-definite.*

Next, we show that the contamination usually leads to an inconsistent OLS estimator.

Lemma 3. *Suppose Assumption 4 holds. The OLS estimator $\hat{\beta}$ for (26) converges to $(\mathcal{I} - (\Sigma_1 + \Sigma_2)^{-1}\Sigma_2)\,\beta$ in probability. Furthermore, $\hat{\beta}$ converges to β in probability if and only if $\Sigma_2 \beta = \mathbf{0}$.*

From Lemma 3, we know that the OLS estimator $\hat{\beta}$ does not converge to the true parameter β unless β is in the null space of Σ_2. As a result, the predicted response $X_a(w) = w^\top \hat{\beta}$ given w is usually biased. Note that in this case the bias can be translated to the contamination in response as in Section 4.3.1, and the conditions in Theorem 3 can be similarly applied. In this section, we instead focus on a different angle: even when $\Sigma_2 \beta = \mathbf{0}$ holds and $\hat{\beta}$ converges to β, the algorithm is still not bias-free. This is because the new covariate W_0 may be contaminated. The actual prediction should be $Z_0^\top \beta = (W_0 - U_0)^\top \beta$, while under contamination, even with $\hat{\beta} = \beta$, the prediction is $W_0^\top \beta$.

We consider the two-sided guardrail $[X_h^u, X_h^l]$ for the human augmentation. Note that in this case, the loss functions for the algorithmic and safeguarded outcomes are $l(X_a(W), Z)$ and $l(\hat{X}(W), Z)$, respectively, because the loss only depends on the actual covariate Z, not the observed but potentially contaminated W. We impose the following technical assumption.

Assumption 5. *Assume the domain of Z is a closed and bounded set $\mathcal{Z} \in \mathbb{R}^d$. Assume $\Sigma_2 \beta = \mathbf{0}$ and there exist constants $b, p \in (0, 0.5)$ such that*

$$\mathbb{P}\left(U^\top \beta \geq b\right) \geq p,\ \mathbb{P}\left(U^\top \beta \leq -b\right) \geq p. \tag{29}$$

The compactness of the covariate Z is similar to the assumption in Theorem 3. Equation (29) states that the contamination $U^\top \beta$ is not concentrated at zero, which would make the application of the two-sided range more likely to be beneficial. Next we state our main result.

Theorem 4. *Suppose Assumptions 4 and 5 hold. If the upper and lower bounds X_h^u, X_h^l satisfy*

$$X_h^l \leq \min_{z \in \mathcal{Z}} \{z^\top \beta\} + \sqrt{\frac{p}{1-p}} b, \quad X_h^u \geq \max_{z \in \mathcal{Z}} \{z^\top \beta\} - \sqrt{\frac{p}{1-p}} b, \tag{30}$$

then we have $\mathbb{E}[l(\hat{X}(W), Z)] \leq \mathbb{E}[l(X_a(W), Z)]$.

Comparing to Theorem 3, it is worth pointing out that although the contamination mechanisms differ, the conditions for beneficial human augmentation are surprisingly similar. In particular, when b or p is larger, i.e., the magnitude of the contamination increases, there is more room for human augmentation to be helpful (as conditions (30) are more likely to hold).

5. Conclusion

Motivated by a consulting project on retail fuel pricing, we propose a framework to study the human-AI interaction in which an algorithm first recommends a decision to the human analyst, then the analyst can augment it based on domain knowledge and experience. As far as we know, this is the first study to investigate this type of interaction. With the framework, we investigate when human knowledge adds value to algorithmic decision-making. We demonstrate three common and practical situations in which human knowledge may play a critical role in harnessing and correcting algorithmic decisions, even with large data.

We conclude by discussing potential future directions. First, we may consider more sophisticated yet realistic human augmentation. For example, AI is known to suffer from out-of-distribution issues when the algorithmic decision learned from the training data does not provide much value, and for these instances, human knowledge is particularly useful in correcting. It is desirable to extend our framework and incorporate simple rules to identify such instances and override the algorithmic decision. Second, another important reason for humans to intervene is the consideration for fairness or ethical issues associated with the algorithmic decision. It is a fruitful direction to extend our framework by incorporating these considerations as rules of thumb to guardrail algorithmic outputs. Lastly, in some applications, the adoption choice between the algorithm and the human knowledge needs to be made before the computation of algorithmic decisions because it may incur significant waiting if human correction is conducted after observing the algorithmic decisions. In this case, the human analyst needs to design and commit to a simple rule based on the observed covariate. Our framework may be extended to study this kind of human-AI interaction.

References

Agrawal A, Gans J, Goldfarb A (2018) Prediction, judgment, and complexity: A theory of decision-making and artificial intelligence. *The Economics of Artificial Intelligence: An Agenda*, 89–110 (University of Chicago Press).

Agrawal A, Gans JS, Goldfarb A (2019) Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy* 47:1–6.

Arvan M, Fahimnia B, Reisi M, Siemsen E (2019) Integrating human judgement into quantitative forecasting methods: A review. *Omega* 86:237–252.

Baker J (2021) Maximizing forecast value added through machine learning and "nudges". *Foresight: The International Journal of Applied Forecasting* 60:8–15.

Bansal G, Nushi B, Kamar E, Horvitz E, Weld DS (2021) Is the most accurate AI the best teammate? Optimizing AI for teamwork. *Proceedings of the AAAI Conference on Artificial Intelligence* 35(13):11405–11414.

Bansal G, Nushi B, Kamar E, Weld DS, Lasecki WS, Horvitz E (2019) Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):2429–2437.

Bastani H, Bastani O, Sinchaisri P (2021) Learning best practices: Can machine learning improve human decision-making? *Academy of Management Proceedings* 2021(1):14006.

Bastani O, Bastani H, Kim C (2019) Interpreting blackbox models via model extraction. *Working paper,* https://doi.org/10.48550/arXiv.1705.08504.

Berger P (2022) Autonomous delivery and work drones will still need a human minder. *The Wall Street Journal* https://www.wsj.com/articles/why-autonomous-vehicles-will-still-need-a-human-minder-11667833922.

Bhatia K, Jain P, Kamalaruban P, Kar P (2017) Consistent robust regression. *Advances in Neural Information Processing Systems*, volume 30.

Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14 (New York, NY, USA: Association for Computing Machinery).

Boyaci T, Canyakmaz C, deVericourt F (2020) Human and machine: The impact of machine input on decision-making under cognitive limitations. *Working paper,* http://dx.doi.org/10.2139/ssrn.3740508.

Campbell D, Frei F (2011) Market heterogeneity and local capacity decisions in services. *Manufacturing Service Oper. Management* 13(1):2–19.

Case N (2018) How to become a centaur. *Journal of Design and Science* Https://jods.mitpress.mit.edu/pub/issue3-case.

Cheng HF, Wang R, Zhang Z, O'Connell F, Gray T, Harper FM, Zhu H (2019) Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12 (Association for Computing Machinery).

Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R (2018) A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 134–148 (PMLR).

Cohen MC, Perakis G, Pindyck RS (2021) A simple rule for pricing with limited knowledge of demand. *Management Sci.* 67(3):1608–1621.

Cooper WL, Homem-de Mello T, Kleywegt AJ (2015) Learning and pricing with models that do not explicitly incorporate competition. *Oper. Res.* 63(1):86–103.

Dai T, Singh S (2021) Artificial intelligence on call: The physician's decision of whether to use AI in clinical practice. *Working paper,* http://dx.doi.org/10.2139/ssrn.3987454.

Das D, Chernova S (2020) Leveraging rationales to improve human task performance. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 510–518 (New York, NY, USA: Association for Computing Machinery).

Davydenko A, Fildes R (2013) Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting* 29(3):510–522.

de Véricourt F, Gurkan H (2022) Is your machine better than you? you may never know. *Working paper,* http://dx.doi.org/10.2139/ssrn.4117641.

den Boer AV, Keskin NB (2022) Dynamic pricing with demand learning and reference effects. *Management Sci.* 68(10):7065–7791.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Sci.* 64(3):1155–1170.

Donahue K, Chouldechova A, Kenthapadi K (2022) Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. *Working paper,* https://doi.org/10.48550/arXiv.2202.08821.

Fildes R, Goodwin P, Lawrence M, Nikolopoulos K (2009) Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting* 25(1):3–23.

Fogliato R, De-Arteaga M, Chouldechova A (2022) A case for humans-in-the-loop: Decisions in the presence of misestimated algorithmic scores. *Working paper,* http://dx.doi.org/10.2139/ssrn.4050125.

Gao R, Saar-Tsechansky M, De-Arteaga M, Han L, Lee MK, Lease M (2021) Human-AI collaboration with bandit feedback. *Working paper,* https://doi.org/10.48550/arXiv.2105.10614.

Grand-Clément J, Pauphilet J (2022) The best decisions are not the best advice: Making adherence-aware recommendations. *Working Paper* .

Greene WH (2003) *Econometric Analysis* (Upper Saddle River, NJ: Prentice Hall), 5th ed. edition, ISBN 0131108492.

Grgić-Hlača N, Engel C, Gummadi KP (2019) Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW).

Hansen KT, Misra K, Pai MM (2021) Frontiers: Algorithmic collusion: Supra-competitive prices via independent algorithms. *Marketing Sci.* 40(1):1–12.

Holstein K, Aleven V (2021) Designing for human-AI complementarity in K-12 education. *Working paper,* https://doi.org/10.48550/arXiv.2104.01266.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Sci.* 67(4):2314–2325.

Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, et al. (2019) Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):590–597.

Karlinsky-Shichor Y, Netzer O (2019) Automating the B2B salesperson pricing decisions: Can machines replace humans and when. *Working paper,* http://dx.doi.org/10.2139/ssrn.3368402.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Sci.* 66(11):5182–5190.

Keskin NB, Li Y, Song JS (2022) Data-driven dynamic pricing and ordering with perishable inventory in a changing environment. *Management Sci.* 68(3):1938–1958.

Keswani V, Lease M, Kenthapadi K (2021) Towards unbiased and accurate deferral to multiple experts. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 154–165 (New York, NY, USA: Association for Computing Machinery).

Khosrowabadi N, Hoberg K, Imdahl C (2022) Evaluating human behaviour in response to AI recommendations for judgemental forecasting. *Eur. J. Oper. Res.* 303(3):1151–1167.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2018) Human decisions and machine predictions. *The Quarterly Journal of Economics* 133(1):237–293.

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. *Behavioral and Brain Sciences* 40.

Lawrence M, Goodwin P, O'Connor M, Önkal D (2006) Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting* 22(3):493–518.

Liu J, Lin S, Xin L, Zhang Y (2022) AI vs. human buyers: A study of alibaba's inventory replenishment system. *Working paper,* http://dx.doi.org/10.2139/ssrn.4207171.

Loh PL, Wainwright MJ (2011) High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Advances in neural information processing systems*, volume 24.

Madras D, Pitassi T, Zemel R (2018) Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, volume 31.

McWilliams B, Krummenacher G, Lucic M, Buhmann JM (2014) Fast and robust least squares estimation in corrupted linear models. *Advances in Neural Information Processing Systems*, volume 27.

Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267:1–38.

Mozannar H, Sontag D (2020) Consistent estimators for learning to defer to an expert. *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7076–7087 (PMLR).

Nguyen NH, Tran TD (2013) Robust lasso with missing and grossly corrupted observations. *IEEE Transactions on Information Theory* 59(4):2036–2058.

Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, Rajpurkar P, Amrhein T, Gupta R, Halabi S, et al. (2019) Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2(1):1–10.

Phillips R, Şimşek AS, Van Ryzin G (2015) The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Sci.* 61(8):1741–1759.

Raghu M, Blumer K, Corrado G, Kleinberg J, Obermeyer Z, Mullainathan S (2019) The algorithmic automation problem: Prediction, triage, and human effort. *Working paper,* https://doi.org/10.48550/arXiv.1903.12220.

Rastogi C, Leqi L, Holstein K, Heidari H (2022) A unifying framework for combining complementary strengths of humans and ML toward better predictive decision-making. *Working paper,* https://doi.org/10.48550/arXiv.2204.10806.

Smith VC, Lange A, Huston DR (2012) Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Networks* 16(3):51–61.

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2022) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Sci.* 68(2):846–865.

Van Donselaar KH, Gaur V, Van Woensel T, Broekmeulen RA, Fransoo JC (2010) Ordering behavior in retail stores and implications for automated replenishment. *Management Sci.* 56(5):766–784.

Vershynin R (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press), ISBN 978-1-108-41519-4.

Wilder B, Horvitz E, Kamar E (2020) Learning to complement humans. *Working paper,* https://doi.org/10.48550/arXiv.2005.00582.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT press).

Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2):210–227.

Ziya S, Ayhan H, Foley RD (2004) Relationships among three assumptions in revenue management. *Oper. Res.* 52(5):804–809.

Online Appendix to “Algorithmic Decision-Making Augmented by Human Knowledge”

A. Proofs in Section 3

Proof of Proposition 1. To prove (i), we first write down the expression for $\mathbb{E}[l(\hat{X})]$,

$$\begin{aligned}\mathbb{E}[l(\hat{X})] &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} l\left(\min(x_a,x_h)\right) f(x_a,x_h)\,\mathrm{d}x_a\,\mathrm{d}x_h \\ &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{x_h} l(x_a)f(x_a,x_h)\,\mathrm{d}x_a + \int_{x_h}^{\infty} l(x_h)f(x_a,x_h)\,\mathrm{d}x_a\right)\mathrm{d}x_h,\end{aligned}$$

where $f(x_a,x_h)$ denotes the joint probability density function of X_a, X_h. The last equality follows from separating the integral in X_a by $X_h \le X_a$ and $X_h > X_a$. Then, we have

$$\begin{aligned}\mathbb{E}[l(X_a)] - \mathbb{E}[l(\hat{X})] &= \int_{-\infty}^{\infty}\int_{x_h}^{\infty}\left(l(x_a)-l(x_h)\right)f(x_a,x_h)\,\mathrm{d}x_a\,\mathrm{d}x_h \\ &= \mathbb{E}[(l(X_a)-l(X_h))\,\mathbb{I}(X_h \le X_a)],\end{aligned} \tag{31}$$

which completes the proof of (i).

To prove (ii), we separate the integral in (31) by $X_h \le x^*$ and $X_h > x^*$:

$$\begin{aligned}\mathbb{E}[l(X_a)] - \mathbb{E}[l(\hat{X})] = &\int_{-\infty}^{x^*}\int_{x_h}^{\infty}\left(l(x_a)-l(x_h)\right)f(x_a,x_h)\,\mathrm{d}x_a\,\mathrm{d}x_h \\ &+ \int_{x^*}^{\infty}\int_{x_h}^{\infty}\left(l(x_a)-l(x_h)\right)f(x_a,x_h)\,\mathrm{d}x_a\,\mathrm{d}x_h.\end{aligned} \tag{32}$$

For the first term in (32), we have

$$\begin{aligned}&\int_{-\infty}^{x^*}\int_{x_h}^{\infty}\left(l(x_a)-l(x_h)\right)f(x_a,x_h)\,\mathrm{d}x_a\,\mathrm{d}x_h \\ &\overset{(a)}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\min\{x^*,x_a\}}\left(l(x_a)-l(x_h)\right)f(x_a,x_h)\,\mathrm{d}x_h\,\mathrm{d}x_a \\ &\overset{(b)}{=} \int_{-\infty}^{x^*}\int_{-\infty}^{x_a}\left(l(x_a)-l(x_h)\right)f(x_a,x_h)\,\mathrm{d}x_h\,\mathrm{d}x_a + \int_{x^*}^{\infty}\int_{-\infty}^{x^*}\left(l(x_a)-l(x_h)\right)f(x_a,x_h)\,\mathrm{d}x_h\,\mathrm{d}x_a \\ &= \mathbb{E}[(l(X_a)-l(X_h))\,\mathbb{I}(X_a \le x^*, X_h \le X_a)] + \mathbb{E}[(l(X_a)-l(X_h))\,\mathbb{I}(X_a > x^*, X_h \le x^*)] \\ &= \mathbb{E}[(l(X_a)-l(X_h))\,\mathbb{I}(X_h \le X_a \le x^*)] + \mathbb{E}[(l(X_a)-l(X_h))\,\mathbb{I}(X_h \le x^* < X_a)].,\end{aligned} \tag{33}$$

where (a) follows from changing the order of integration, and (b) follows from separating the integral by $X_a \le x^*$ and $X_a > x^*$.

For the second term in (32), we have

$$\int_{x^*}^{\infty}\int_{x_h}^{\infty}\left(l(x_a)-l(x_h)\right)f(x_a,x_h)\,\mathrm{d}x_a\,\mathrm{d}x_h$$

$$\begin{aligned}
&\overset{(a)}{=} \int_{x^*}^{\infty} \int_{x^*}^{x_a} (l(x_a) - l(x_h))\, f(x_a, x_h)\, \mathrm{d}x_h\, \mathrm{d}x_a \\
&= \mathbb{E}[(l(X_a) - l(X_h))\mathbb{I}(X_a > x^*, x^* < X_h \leq X_a)] \\
&= \mathbb{E}[(l(X_a) - l(X_h))\mathbb{I}(x^* < X_h \leq X_a)], \qquad (34)
\end{aligned}$$

where (a) follows from changing the order of integration. Plugging (33) and (34) into (32), we have

$$\begin{aligned}
(32) &= \mathbb{E}[(l(X_a) - l(X_h))\,\mathbb{I}(X_h \leq X_a \leq x^*)] + \mathbb{E}[(l(X_a) - l(X_h))\,\mathbb{I}(X_h \leq x^* < X_a)] \\
&\quad + \mathbb{E}[(l(X_a) - l(X_h))\mathbb{I}(x^* < X_h \leq X_a)] \qquad (35) \\
&= \mathbb{E}[(l(X_a) - l(x^*))\,\mathbb{I}(X_h \leq X_a \leq x^*)] + \mathbb{E}[(l(x^*) - l(X_h))\,\mathbb{I}(X_h \leq X_a \leq x^*)] \\
&\quad + \mathbb{E}[(l(X_a) - l(x^*))\,\mathbb{I}(X_h \leq x^* < X_a)] + \mathbb{E}[(l(x^*) - l(X_h))\,\mathbb{I}(X_h \leq x^* < X_a)] \\
&\quad + \mathbb{E}[(l(X_a) - l(X_h))\mathbb{I}(x^* < X_h \leq X_a)], \qquad (36)
\end{aligned}$$

where in the last equality, we separate $l(X_a) - l(X_h)$ into $l(X_a) - l(x^*)$ and $l(x^*) - l(X_h)$. By Assumption 1 (ii), we have $l(\cdot) \geq l(x^*)$ and $l(x_1) \geq l(x_2)$ for $x_1 \geq x_2 \geq x^*$. Thus, we have

$$\mathbb{E}[(l(X_a) - l(x^*))\,\mathbb{I}(X_h \leq X_a \leq x^*)] \geq 0, \qquad (37)$$

$$\mathbb{E}[(l(X_a) - l(X_h))\,\mathbb{I}(x^* < X_h \leq X_a)] \geq 0. \qquad (38)$$

Furthermore, since $l(x^*) \leq l(\cdot)$ and $\mathbb{I}(X_h \leq X_a \leq x^*) \leq \mathbb{I}(X_a \leq x^*, X_h \leq x^*)$, we have

$$\mathbb{E}[(l(x^*) - l(X_h))\,\mathbb{I}(X_h \leq X_a \leq x^*)] \geq \mathbb{E}[(l(x^*) - l(X_h))\,\mathbb{I}(X_a \leq x^*, X_h \leq x^*)], \qquad (39)$$

By (37), (38), the first and last term in (36) can be lower-bounded by zero, and the second term in (36) has the lower bound in (39). Thus, plugging (37), (39) and (38) into (36), we have

$$\begin{aligned}
(36) &\geq \mathbb{E}[(l(x^*) - l(X_h))\,\mathbb{I}(X_a \leq x^*, X_h \leq x^*)] + \mathbb{E}[(l(X_a) - l(x^*))\,\mathbb{I}(X_h \leq x^* < X_a)] \\
&\quad + \mathbb{E}[(l(x^*) - l(X_h))\,\mathbb{I}(X_h \leq x^* < X_a)] \\
&\overset{(a)}{=} \mathbb{E}[(l(X_a) - l(x^*))\mathbb{I}(X_h \leq x^* < X_a)] + \mathbb{E}[(l(x^*) - l(X_h))\mathbb{I}(X_h \leq x^*)] \\
&= \mathbb{E}[(l(X_a) - l(x^*))\mathbb{I}(X_h \leq x^* < X_a)] - \mathbb{E}[(l(X_h) - l(x^*))\mathbb{I}(X_h \leq x^*)] \\
&\overset{(b)}{=} \mathbb{E}[l(X_a)\mathbb{I}(X_h \leq x^* < X_a)] - \mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)] + l(x^*)\mathbb{P}(X_a \leq x^*, X_h \leq x^*) \\
&\geq \mathbb{E}[l(X_a)\mathbb{I}(X_h \leq x^* < X_a)] - \mathbb{E}[l(X_h)\mathbb{I}(X_h \leq x^*)], \qquad (40)
\end{aligned}$$

where (a) holds by $\mathbb{I}(X_a \leq x^*, X_h \leq x^*) + \mathbb{I}(X_h \leq x^* < X_a) = \mathbb{I}(X_h \leq x^*)$, (b) holds by $\mathbb{I}(X_h \leq x^*) - \mathbb{I}(X_h \leq x^* < X_a) = \mathbb{I}(X_a \leq x^*, X_h \leq x^*)$ and (c) follows from $l(x^*) \geq 0$ due to Assumption 1 (i). Thus, if $(40) \geq 0$, we have $(32) \geq 0$ and $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$, which completes the proof of (ii).

To prove (iii), note that

$$(35) = \mathbb{E}[(l(X_a) - l(X_h))\,\mathbb{I}(X_h \leq X_a \leq x^*)] + \mathbb{E}[(l(X_a) - l(X_h))\,\mathbb{I}(X_h \leq x^* < X_a)]$$

$$+\mathbb{E}[(l(X_a)-l(x^*))\mathbb{I}(x^*<X_h\le X_a)]+\mathbb{E}[(l(x^*)-l(X_h))\mathbb{I}(x^*<X_h\le X_a)] \tag{41}$$

By Assumption 1 (ii), we have

$$\mathbb{E}[(l(X_a)-l(X_h))\mathbb{I}(X_h\le X_a\le x^*)]\le 0, \tag{42}$$

$$\mathbb{E}[(l(x^*)-l(X_h))\mathbb{I}(x^*<X_h\le X_a)]\le 0. \tag{43}$$

Plugging (42) and (43) into (41), the first and last term in (41) are upper bounded by zero. And we have

$$\begin{aligned}
(36) &\le \mathbb{E}[(l(X_a)-l(X_h))\mathbb{I}(X_h\le x^*<X_a)]+\mathbb{E}[(l(X_a)-l(x^*))\mathbb{I}(x^*<X_h\le X_a)]\\
&\overset{(a)}{=} \mathbb{E}[(l(X_a)-l(x^*)+l(x^*)-l(X_h))\,\mathbb{I}(X_h\le x^*<X_a)]+\mathbb{E}[(l(X_a)-l(x^*))\mathbb{I}(x^*<X_h\le X_a)]\\
&\overset{(b)}{\le} \mathbb{E}[(l(X_a)-l(x^*))\mathbb{I}(X_a\ge x^*)]-\mathbb{E}[(l(X_h)-l(x^*))\mathbb{I}(X_h\le x^*<X_a)]\\
&\overset{(c)}{=} \mathbb{E}[l(X_a)\mathbb{I}(X_a\ge x^*)]-\mathbb{E}[l(X_h)\mathbb{I}(X_h\le x^*<X_a)]-l(x^*)\mathbb{P}((X_a=x^*)\cap(X_h>x^*,X_a>x^*))\\
&\overset{(d)}{\le} \mathbb{E}[l(X_a)\mathbb{I}(X_a\ge x^*)]-\mathbb{E}[l(X_h)\mathbb{I}(X_h\le x^*<X_a)],
\end{aligned} \tag{44}$$

where (a) holds by separating $l(X_a)-l(X_h)$ into $l(X_a)-l(x^*)+l(x^*)-l(X_h)$, (b) follows from $(l(X_a)-l(x^*))\,\mathbb{I}(X_h\le x^*<X_a)+(l(X_a)-l(x^*))\mathbb{I}(x^*<X_h\le X_a)\le(l(X_a)-l(x^*))\mathbb{I}(X_a\ge x^*)$, (c) follows from $\mathbb{I}(X_a\ge x^*)-\mathbb{I}(X_h\le x^*<X_a)=\mathbb{I}((X_a=x^*)\cap(X_h>x^*,X_a>x^*))$, (d) follows from $l(x^*)\ge 0$ due to Assumption 1 (i). Thus, if $(44)\le 0$, we have $(32)<0$ and $\mathbb{E}[l(\hat{X})]>\mathbb{E}[l(X_a)]$. So a necessary condition for $\mathbb{E}[l(\hat{X})]\le\mathbb{E}[l(X_a)]$ is $(44)\ge 0$. Thus, we complete the proof of (iii). □

Proof of Example 3. By the definition of $l(\cdot),X_a,X_h$, we have

$$\mathbb{E}[l(X_a)\mathbb{I}(X_a\ge x^*)]=\sigma^2/2,$$
$$\mathbb{E}[l(X_h)\mathbb{I}(X_h\le x^*)]=\int_{-\infty}^{x^*-1}\frac{3\epsilon}{(x_h-x^*)^2}\,\mathrm{d}x_h=\int_{-\infty}^{-1}\frac{3\epsilon}{x_h^2}\,\mathrm{d}x_h=3\epsilon.$$

So if $\epsilon<\sigma^2/(6a)$, then

$$\mathbb{E}[l(X_a)\mathbb{I}(X_a\ge x^*)]=\sigma^2/2\ge 3a\epsilon=a\mathbb{E}[l(X_h)\mathbb{I}(X_h\le x^*)].$$

Next, we prove $\mathbb{E}[l(\hat{X})]>\mathbb{E}[l(X_a)]$. According to the distribution of X_a,X_h (7), we have

$$\begin{aligned}
&\mathbb{E}[l(\hat{X})]-\mathbb{E}[l(X_a)]\\
&=\mathbb{P}(X_h=\infty)\mathbb{E}[(X_a-x^*)^2]+\int_{-\infty}^{x^*-1}\int_{-\infty}^{\infty}\left(\min\{x_h,x_a\}-x^*\right)^2 f(x_a)\,\mathrm{d}x_a f(x_h)\,\mathrm{d}x_h-\mathbb{E}[(X_a-x^*)^2]\\
&\overset{(a)}{=}\int_{-\infty}^{x^*-1}\int_{-\infty}^{\infty}\left(\min\{x_h,x_a\}-x^*\right)^2 f(x_a)\,\mathrm{d}x_a f(x_h)\,\mathrm{d}x_h-\epsilon\sigma^2\\
&=\int_{-\infty}^{x^*-1}\int_{-\infty}^{x_h}\left(x_a-x^*\right)^2 f(x_a)\,\mathrm{d}x_a f(x_h)\,\mathrm{d}x_h+\int_{-\infty}^{x^*-1}\int_{x_h}^{\infty}\left(x_h-x^*\right)^2 f(x_a)\,\mathrm{d}x_a f(x_h)\,\mathrm{d}x_h-\epsilon\sigma^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{\geq} \int_{-\infty}^{x^*-1} \int_{x_h}^{\infty} (x_h - x^*)^2 f(x_a)\,\mathrm{d}x_a f(x_h)\,\mathrm{d}x_h - \epsilon\sigma^2 \\
&= \int_{-\infty}^{x^*-1} (x_h - x^*)^2 \int_{x_h}^{\infty} f(x_a)\,\mathrm{d}x_a f(x_h)\,\mathrm{d}x_h - \epsilon\sigma^2 \\
&\overset{(c)}{\geq} \frac{1}{2} \int_{-\infty}^{x^*-1} (x_h - x^*)^2 f(x_h)\,\mathrm{d}x_h - \epsilon\sigma^2 \\
&\overset{(d)}{=} \frac{1}{2} \int_{-\infty}^{x^*-1} \frac{3\epsilon}{(x_h - x^*)^2}\,\mathrm{d}x_h - \epsilon\sigma^2 \\
&= \frac{3\epsilon}{2} - \epsilon\sigma^2 \\
&\overset{(e)}{>} 0,
\end{aligned}
$$

where (a) follows from $\mathbb{P}(X_h = \infty) = 1 - \epsilon$, (b) follows from $\int_{-\infty}^{x^*-1} \int_{-\infty}^{x_h} (x_a - x^*)^2 f(x_a)\,\mathrm{d}x_a f(x_h)\,\mathrm{d}x_h \geq 0$, (c) follows from $\int_{x_h}^{\infty} f(x_a)\,\mathrm{d}x_a \geq \int_0^{\infty} f(x_a)\,\mathrm{d}x_a = 1/2$ due to $x_h \leq x^* - 1 < 0$, (d) follows from the distribution of X_h (7), (e) follows from $\sigma^2 < 3/2$ in the condition of Example 3. □

Proof of Corollary 1. We define $D(x_h) \coloneqq \mathbb{E}[l(X_a)] - \mathbb{E}[l(\hat{X})]$ as a function of x_h. By (31), we have

$$
D(x_h) - D(x_h + \Delta) = \int_{x_h}^{x_h+\Delta} (l(x_a) - l(x_h)) f(x_a)\,\mathrm{d}x_a, \tag{45}
$$

for any $\Delta > 0$.

When $x_h \geq x^*$, we have $l(x_a) \geq l(x_h)$ for $x_a \geq x_h$ according to Assumption 1 (ii). Thus, $D(x_h) \geq D(x_h + \Delta)$.

When $x_h < x^*$, there exists a small enough Δ such that $x_h + \Delta < x^*$. And for $x_h \leq x_a \leq x_h + \Delta < x^*$, we have $l(x_a) \leq l(x_h)$ according to Assumption 1 (ii). Thus, $D(x_h) \leq D(x_h + \Delta)$.

In summary, $D(x_h)$ increases as x_h when $x_h < x^*$ and decreases when $x_h \geq x^*$. □

Proof of Proposition 2. We write down the difference of the expected losses:

$$
\begin{aligned}
&\mathbb{E}[l(X_a)] - \mathbb{E}[l(\hat{X})] \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} l(x_a) f(x_a, x_h^l, x_h^u)\,\mathrm{d}x_a\,\mathrm{d}x_h^l\,\mathrm{d}x_h^u - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} l(\hat{x}) f(x_a, x_h^l, x_h^u)\,\mathrm{d}x_a\,\mathrm{d}x_h^l\,\mathrm{d}x_h^u \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{x_h^l} (l(x_a) - l(x_h^l)) f(x_a, x_h^l, x_h^u)\,\mathrm{d}x_a\,\mathrm{d}x_h^l\,\mathrm{d}x_h^u \\
&\quad + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{x_h^u}^{\infty} (l(x_a) - l(x_h^u)) f(x_a, x_h^l, x_h^u)\,\mathrm{d}x_a\,\mathrm{d}x_h^l\,\mathrm{d}x_h^u \\
&= \mathbb{E}[(l(X_a) - l(X_h^l))\mathbb{I}(X_a \leq X_h^l)] + \mathbb{E}[(l(X_a) - l(X_h^u))\mathbb{I}(X_a \geq X_h^u)],
\end{aligned} \tag{46}
$$

where the second equality holds by the definition $\hat{X} = \min\{\max\{X_a, X_h^l\}, X_h^u\}$ and $X_h^u \geq X_h^l$. This proves part (i).

Next, we prove (ii). Separating the first term in (46) by $X_h^l < x^*$ and $X_h^l \geq x^*$, we have

$$
\begin{aligned}
&\mathbb{E}[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \leq X_h^l)] \\
&= \mathbb{E}[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \leq X_h^l < x^*)] + \mathbb{E}[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \leq X_h^l, X_h^l \geq x^*)] \\
&\overset{(a)}{\geq} \mathbb{E}[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \leq X_h^l, X_h^l \geq x^*)] \\
&\overset{(b)}{=} \mathbb{E}\left[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)\right] + \mathbb{E}\left[(l(X_a)-l(X_h^l))\mathbb{I}(X_a > x^*, X_h^l \geq X_a)\right] \\
&\overset{(c)}{=} \mathbb{E}\left[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)\right] + \mathbb{E}\left[(l(X_a)-l(x^*)+l(x^*)-l(X_h^l))\mathbb{I}(X_a > x^*, X_h^l \geq X_a)\right] \\
&\overset{(d)}{\geq} \mathbb{E}\left[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)\right] + \mathbb{E}\left[(l(x^*)-l(X_h^l))\mathbb{I}(X_a > x^*, X_h^l \geq X_a)\right] \\
&\overset{(e)}{\geq} \mathbb{E}\left[(l(X_a)-l(x^*)+l(x^*)-l(X_h^l))\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)\right] + \mathbb{E}\left[(l(x^*)-l(X_h^l))\mathbb{I}(X_a > x^*, X_h^l \geq x^*)\right] \\
&\overset{(f)}{=} \mathbb{E}\left[(l(X_a)-l(x^*))\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)\right] + \mathbb{E}\left[(l(x^*)-l(X_h^l))\mathbb{I}(X_h^l \geq x^*)\right] \\
&\overset{(g)}{=} \mathbb{E}\left[l(X_a)\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)\right] - \mathbb{E}\left[l(X_h^l)\mathbb{I}(X_h^l \geq x^*)\right] + \mathbb{E}[l(x^*)\mathbb{I}(X_a > x^*, X_h^l \geq x^*)] \\
&\overset{(h)}{\geq} \mathbb{E}\left[l(X_a)\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*)\right] - \mathbb{E}\left[l(X_h^l)\mathbb{I}(X_h^l \geq x^*)\right],
\end{aligned}
\tag{47}
$$

where the (a) follows from $(l(X_a)-l(X_h^l))\mathbb{I}(X_a \leq X_h^l < x^*)$ due to Assumption 1 (ii), (b) follows from separating the expectation by $X_a \leq x^*$ and $X_a > x^*$, (c) follows from $l(X_a)-l(X_h^l) = l(X_a) - l(x^*) + l(x^*) - l(X_h^l)$, (d) follows from $\mathbb{E}\left[(l(X_a)-l(x^*))\mathbb{I}(X_a > x^*, X_h^l \geq X_a)\right] \geq 0$ due to Assumption 1 (ii), (e) follows from $\mathbb{I}(X_a \geq x^*, X_h^l \geq X_a) \leq \mathbb{I}(X_a \geq x^*, X_h^l \geq x^*)$, (f) follows from $\mathbb{I}(X_a \leq x^*, X_h^l \geq x^*) + \mathbb{I}(X_a > x^*, X_h^l \geq x^*) = \mathbb{I}(X_h^l \geq x^*)$, (g) follows from $\mathbb{I}(X_h^l \geq x^*) - \mathbb{I}(X_a \leq x^*, X_h^l \geq x^*) = \mathbb{I}(X_a > x^*, X_h^l \geq x^*)$, (h) follows from $l(x^*) \geq 0$ due to Assumption 1 (i).

For the second term in (46), we separate it by $X_h^u \leq x^*$ and $X_h^u > x^*$:

$$
\begin{aligned}
&\mathbb{E}[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \geq X_h^u)] \\
&= \mathbb{E}[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \geq X_h^u > x^*)] + \mathbb{E}[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \geq X_h^u, x^* \geq X_h^u)] \\
&\overset{(a)}{\geq} \mathbb{E}[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \geq X_h^u, x^* \geq X_h^u)] \\
&\overset{(b)}{=} \mathbb{E}\left[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \leq x^*, X_h^u \leq X_a)\right] + \mathbb{E}\left[(l(X_a)-l(X_h^u))\mathbb{I}(X_a > x^*, X_h^u \leq x^*)\right] \\
&= \mathbb{E}\left[(l(X_a)-l(x^*)+l(x^*)-l(X_h^u))\mathbb{I}(X_a \leq x^*, X_h^u \leq X_a)\right] + \mathbb{E}\left[(l(X_a)-l(X_h^u))\mathbb{I}(X_a > x^*, X_h^u \leq x^*)\right] \\
&\overset{(c)}{\geq} \mathbb{E}\left[(l(x^*)-l(X_h^u))\mathbb{I}(X_a \leq x^*, X_h^u \leq X_a)\right] + \mathbb{E}\left[(l(X_a)-l(x^*)+l(x^*)-l(X_h^u))\mathbb{I}(X_a > x^*, X_h^u \leq x^*)\right] \\
&= \mathbb{E}\left[(l(X_a)-l(x^*))\mathbb{I}(X_a \geq x^*, X_h^u \leq x^*)\right] - \mathbb{E}\left[(l(X_h^u)-l(x^*))\mathbb{I}(X_h^u \leq x^*)\right] \\
&= \mathbb{E}\left[l(X_a)\mathbb{I}(X_a \geq x^*, X_h^u \leq x^*)\right] - \mathbb{E}\left[l(X_h^u)\mathbb{I}(X_h^u \leq x^*)\right] + l(x^*)\mathbb{P}(X_a < x^*, X_h \leq x^*) \\
&\overset{(d)}{\geq} \mathbb{E}\left[l(X_a)\mathbb{I}(X_a \geq x^*, X_h^u \leq x^*)\right] - \mathbb{E}\left[l(X_h^u)\mathbb{I}(X_h^u \leq x^*)\right],
\end{aligned}
\tag{48}
$$

where (a) follows from $(l(X_a)-l(X_h^u))\mathbb{I}(X_a \geq X_h^u > x^*) \geq 0$, (b) follows from separating the expectation by $X_a \leq x^*$ and $X_a > x^*$, (b) follows from Assumption 1 (ii), (c) follows from $(l(X_a)-l(x^*)+$

$l(x^*)-l(X_h^u))\mathbb{I}(X_a \le x^*, X_h^u \le X_a) \ge 0$ due to Assumption 1 (ii), (d) follows from $l(x^*) \ge 0$ due to Assumption 1 (i).

Thus, if (47) + (48) > 0, we have (46)> 0 and $\mathbb{E}[l(X_a)] \ge \mathbb{E}[l(\hat{X})]$. Thus, we complete the proof for (ii).

To prove (iii), similar to (47), we separate the first term in (46) by $x^* \le X_a \le X_h^l$, $X_a \le x^* \le X_h^l$, and $X_a \le X_h^l \le x^*$:

$$\begin{align*}
&\mathbb{E}[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \le X_h^l)]\\
&= \mathbb{E}[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \le X_h^l \le x^*)] + \mathbb{E}\left[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \le x^*, X_h^l \ge x^*)\right]\\
&\quad + \mathbb{E}\left[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \ge x^*, X_h^l \ge X_a)\right]\\
&\overset{(a)}{\le} \mathbb{E}[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \le X_h^l \le x^*)] + \mathbb{E}\left[(l(X_a)-l(X_h^l))\mathbb{I}(X_a \le x^*, X_h^l \ge x^*)\right]\\
&\overset{(b)}{=} \mathbb{E}[(l(X_a)-l(x^*)+l(x^*)-l(X_h^l))\mathbb{I}(X_a \le x^*, X_h^l \ge X_a)]\\
&\overset{(c)}{\le} \mathbb{E}[(l(X_a)-l(x^*))\mathbb{I}(X_a \le x^*)] + \mathbb{E}[(l(x^*)-l(X_h^l))\mathbb{I}(X_a \le x^* \le X_h^l)]\\
&= \mathbb{E}[l(X_a)\mathbb{I}(X_a \le x^*)] - \mathbb{E}[l(X_h^l)\mathbb{I}(X_a \le x^* \le X_h^l)] - l(x^*)\mathbb{P}(X_a \le x^*, X_h > x^*)\\
&\overset{(d)}{\le} \mathbb{E}[l(X_a)\mathbb{I}(X_a \le x^*)] - \mathbb{E}[l(X_h^l)\mathbb{I}(X_a \le x^* \le X_h^l)], \tag{49}
\end{align*}$$

where (a) follows from $(l(X_a)-l(X_h^l))\mathbb{I}(X_a \ge x^*, X_h^l \ge X_a) \le 0$ due to Assumption 1 (ii), (b) follows from $\mathbb{I}(X_a \le X_h^l \le x^*) + \mathbb{I}(X_a \le x^*, X_h^l \ge x^*) = \mathbb{I}(X_a \le x^*, X_h^l \ge X_a)$, (c) follows from $l(X_a)-l(x^*) \ge 0$ and $l(X_h)-l(x^*) \ge 0$ due to Assumption 1 (ii) and $\mathbb{I}(X_a \le x^*, X_h^l \ge X_a) \ge \mathbb{I}(X_a \le x^* \le X_h^l)$, (d) follows from $l(x^*) \ge 0$ due to Assumption 1 (i).

Similar to (48), we separate the second term in (46) by $x^* \le X_h^u \le X_a$, $X_h^u \le x^* \le X_a$, and $X_h^u \le X_a \le x^*$:

$$\begin{align*}
&\mathbb{E}[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \ge X_h^u)]\\
&= \mathbb{E}[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \ge X_h^u \ge x^*)] + \mathbb{E}\left[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \le x^*, X_h^u \le X_a)\right] \tag{50}\\
&\quad + \mathbb{E}\left[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \ge x^*, X_h^u \le x^*)\right]\\
&\overset{(a)}{\le} \mathbb{E}[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \ge X_h^u \ge x^*)] + \mathbb{E}\left[(l(X_a)-l(X_h^u))\mathbb{I}(X_a \ge x^*, X_h^u \le x^*)\right]\\
&\overset{(b)}{=} \mathbb{E}[(l(X_a)-l(x^*)+l(x^*)-l(X_h^u))\mathbb{I}(X_a \ge x^*, X_h^u \le X_a)]\\
&\overset{(c)}{\le} \mathbb{E}[(l(X_a)-l(x^*))\mathbb{I}(X_a \ge x^*)] + \mathbb{E}[(l(x^*)-l(X_h^u))\mathbb{I}(X_a \ge x^*, X_h^u \le x^*)]\\
&= \mathbb{E}[l(X_a)\mathbb{I}(X_a \ge x^*)] - \mathbb{E}[l(X_h^u)\mathbb{I}(X_a \ge x^*, X_h^u \le x^*)] - l(x^*)\mathbb{P}(X_a \ge x^*, X_h^u > x^*)\\
&\overset{(d)}{\le} \mathbb{E}[l(X_a)\mathbb{I}(X_a \ge x^*)] - \mathbb{E}[l(X_h^u)\mathbb{I}(X_a \ge x^*, X_h^u \le x^*)], \tag{51}
\end{align*}$$

where (a) follows from $(l(X_a)-l(X_h^u))\mathbb{I}(X_a \le x^*, X_h^u \le X_a) \le 0$ due to Assumption 1 (ii), (b) follows from $\mathbb{I}(X_a \ge X_h^u \ge x^*) + \mathbb{I}(X_a \ge x^*, X_h^u \le x^*) = \mathbb{I}(X_a \ge x^*, X_h^u \le X_a)$, (c) follows from $l(X_a)-l(x^*) \ge$

0, $l(x^*) - l(X_h) \leq 0$ and $\mathbb{I}(X_a \geq x^*, X_h^u \leq X_a) \geq \mathbb{I}(X_a \geq x^*, X_h^u \leq x^*)$, (d) follows from $l(x^*) \geq 0$ due to Assumption 1 (i).

Thus, if (49) + (51) < 0, then (46)< 0 and $\mathbb{E}[l(X_a)] < \mathbb{E}[l(\hat{X})]$. So a necessary condition for $\mathbb{E}[l(\hat{X})] \leq \mathbb{E}[l(X_a)]$ is (46) ≥ 0. We complete the proof for (iii). □

Proof of Proposition 3. The proof basically follows the same argument as in Proposition 2. We take the first part (i) as an example. We write down the difference of the expected losses:

$$\begin{aligned}
&\mathbb{E}[l(X_a, W)] - \mathbb{E}[l(\hat{X}, W)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} l(x_a, w) f(x_a, x_h^l, x_h^u, w)\, \mathrm{d}x_a\, \mathrm{d}x_h^l\, \mathrm{d}x_h^u\, \mathrm{d}w \\
&\quad - \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} l(\hat{x}, w) f(x_a, x_h^l, x_h^u, w)\, \mathrm{d}x_a\, \mathrm{d}x_h^l\, \mathrm{d}x_h^u\, \mathrm{d}w \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{x_h^l} (l(x_a, w) - l(x_h^l, w)) f(x_a, x_h^l, x_h^u, w)\, \mathrm{d}x_a\, \mathrm{d}x_h^l\, \mathrm{d}x_h^u\, \mathrm{d}w \\
&\quad + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{x_h^u}^{\infty} (l(x_a, w) - l(x_h^u, w)) f(x_a, x_h^l, x_h^u, w)\, \mathrm{d}x_a\, \mathrm{d}x_h^l\, \mathrm{d}x_h^u\, \mathrm{d}w \\
&= \mathbb{E}[(l(X_a, W) - l(X_h^l, W))\mathbb{I}(X_a \leq X_h^l)] + \mathbb{E}[(l(X_a, W) - l(X_h^u, W))\mathbb{I}(X_a \geq X_h^u)].
\end{aligned}$$

Thus, we complete the proof for part (i). □

B. Proofs in Section 4.1

Proof of Lemma 1. We reformulate the linear demand function (16) in the matrix form:

$$D = A\theta + \gamma \boldsymbol{p'} + \boldsymbol{\epsilon},$$

where

$$D = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix},\ A = \begin{pmatrix} 1 & -p_1 \\ \vdots & \vdots \\ 1 & -p_n \end{pmatrix},\ \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},\ \boldsymbol{p'} = \begin{pmatrix} p_1' \\ \vdots \\ p_n' \end{pmatrix},\ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

According to the algorithm assumed demand model (15), the OLS estimator is

$$\hat{\theta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (A^\top A)^{-1} A^\top D = \theta + (A^\top A)^{-1} A^\top \boldsymbol{p'}\gamma + (A^\top A)^{-1} A^\top \epsilon. \tag{52}$$

According to Assumption 3, we have

$$\frac{1}{n}\sum_{i=1}^n p_i \xrightarrow{p} \mu,\ \frac{1}{n}\sum_{i=1}^n p_i' \xrightarrow{p} \mu, \tag{53}$$

$$\frac{1}{n}\sum_{i=1}^n p_i^2 \xrightarrow{p} \mu^2 + \sigma^2,\ \frac{1}{n}\sum_{i=1}^n p_i'^2 \xrightarrow{p} \mu^2 + \sigma^2, \tag{54}$$

$$\frac{1}{n}\sum_{i=1}^n p_i p_i' - \mu^2 \xrightarrow{p} \rho\sigma^2, \tag{55}$$

where $\xrightarrow{p}$ denotes convergence in probability, the convergence follows by the weak law of large numbers. By (53), (54), and (55), we have

$$\frac{1}{n}A^\top A = \begin{pmatrix} 1 & -\frac{1}{n}\sum_{i=1}^n p_i \\ -\frac{1}{n}\sum_{i=1}^n p_i & \frac{1}{n}\sum_{i=1}^n p_i^2 \end{pmatrix} \xrightarrow{p} \begin{pmatrix} 1 & -\mu \\ -\mu & \mu^2+\sigma^2 \end{pmatrix},$$

and

$$\frac{1}{n}A^\top \boldsymbol{p'} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n p_i' \\ -\frac{1}{n}\sum_{i=1}^n p_i p_i' \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \mu \\ -(\rho\sigma^2+\mu^2) \end{pmatrix}.$$

And by Slutsky's theorem, we have

$$(\frac{1}{n}A^\top A)^{-1}\frac{1}{n}A^\top \boldsymbol{p'}\gamma \xrightarrow{p} \frac{\gamma}{\sigma^2}\begin{pmatrix} \sigma^2+\mu^2 & \mu \\ \mu & 1 \end{pmatrix}\begin{pmatrix} \mu \\ -\rho\sigma^2-\mu^2 \end{pmatrix} = \frac{\gamma}{\sigma^2}\begin{pmatrix} \mu\sigma^2+\mu^3-\mu\rho\sigma^2-\mu^3 \\ -\rho\sigma^2 \end{pmatrix} = \begin{pmatrix} \gamma\mu(1-\rho) \\ -\gamma\rho. \end{pmatrix}$$

Since ϵ is the random noise, we have $(A^\top A)^{-1}A^\top \epsilon \xrightarrow{p} 0$.

So the OLS estimator in (52) is

$$\hat{\theta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \gamma\mu(1-\rho) \\ -\gamma\rho. \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \alpha+\gamma\mu(1-\rho) \\ \beta-\gamma\rho. \end{pmatrix}$$

Then the optimal price for AI is

$$p_a = \frac{\hat{\alpha}}{2\hat{\beta}} \xrightarrow{p} \frac{\alpha+\gamma\mu(1-\rho)}{2(\beta-\gamma\rho)},$$

which completes the proof for Lemma 1. □

Proof of Theorem 1. We define the revenue function $r(p) := pd(p,p')$ and provide a condition for $\hat{p}$ such that $r(\hat{p}) \geq r(p_a)$. Note that $\hat{p} = \min\{p_a, p'\}$. If $p_a \leq p'$, then $r(\hat{p}) = r(p_a)$. So the condition for $\hat{p}$ boils down to the condition for p' such that $r(p') \geq r(p_a)$ when $p' < p_a$. According to (15), we have

$$\begin{aligned} & r(p') - r(p_a) \\ & = p'(\alpha + (\gamma-\beta)p') - p_a(\alpha - \beta p_a + \gamma p') \\ & = (\gamma-\beta)p'^2 + \left(\alpha - \frac{\gamma}{2}\frac{\alpha+\gamma\mu(1-\rho)}{\beta-\rho\gamma}\right)p' + \frac{\beta}{4}\frac{(\alpha+\gamma\mu(1-\rho))^2}{(\beta-\rho\gamma)^2} - \frac{\alpha}{2}\frac{\alpha+\gamma\mu(1-\rho)}{\beta-\rho\gamma}. \end{aligned} \tag{56}$$

So $r(p') \geq r(p_a)$ is equivalent to (56) ≥ 0. And multiplying (56) by $-4(\beta-\rho\gamma)^2$, we have

$$\begin{aligned} & 4(\beta-\rho\gamma)^2(\beta-\gamma)p'^2 - 2(\beta-\rho\gamma)\left(2\alpha\beta - 2\alpha\rho\gamma - \alpha\gamma - \gamma^2\mu(1-\rho)\right)p' \\ & \quad - (\alpha+\gamma\mu(1-\rho))\left(\beta\gamma\mu(1-\rho) + 2\alpha\rho\gamma - \alpha\beta\right) \leq 0. \end{aligned} \tag{57}$$

We write down the discriminant of the quadratic function (57):

$$\begin{aligned} \Delta = & 4(\beta-\rho\gamma)^2(2\alpha\beta - 2\alpha\gamma\rho - \alpha\gamma - \gamma^2\mu + \gamma^2\mu\rho)^2 \\ & + 16(\beta-\rho\gamma)^2(\beta-\gamma)(\alpha+\gamma\mu(1-\rho))\left(\beta\gamma\mu(1-\rho) + 2\alpha\rho\gamma - \alpha\beta\right), \end{aligned} \tag{58}$$

In the first term of (58), we have

$$\begin{aligned}&(2\alpha\beta-2\alpha\gamma\rho-\alpha\gamma-\gamma^2\mu+\gamma^2\mu\rho)^2\\&=4\alpha^2\beta^2+4\alpha^2\gamma^2\rho^2+\alpha^2\gamma^2+\gamma^4\mu^2+\gamma^4\mu^2\rho^2-8\alpha^2\beta\gamma\rho-4\alpha^2\beta\gamma-4\alpha\beta\gamma^2\mu+4\alpha\beta\gamma^2\mu\rho+4\alpha^2\gamma^2\rho\\&\quad+4\alpha\gamma^3\mu\rho-4\alpha\gamma^3\mu\rho^2+2\alpha\gamma^3\mu-2\alpha\gamma^3\mu\rho-2\gamma^4\mu^2\rho.\end{aligned} \tag{59}$$

In the second term of (58), we have

$$\begin{aligned}&(\beta-\gamma)(\alpha+\gamma\mu(1-\rho))\left(\beta\gamma\mu(1-\rho)+2\alpha\rho\gamma-\alpha\beta\right)\\&=-\alpha^2\beta^2+2\alpha^2\beta\gamma\rho+2\alpha\beta\gamma^2\mu(1-\rho)\rho+\beta^2\gamma^2\mu^2(1-\rho)^2+\alpha^2\beta\gamma-2\alpha^2\gamma^2\rho-2\alpha\gamma^3\mu(1-\rho)\rho-\beta\gamma^3\mu^2(1-\rho)^2.\end{aligned} \tag{60}$$

Plugging (59) and (60) into (58), we have

$$\frac{\Delta}{4(\beta-\rho\gamma)^2}=\gamma^2\left(\alpha(1-2\rho)+(\gamma-2\beta)(1-\rho)\mu\right)^2,$$

and the roots of (57)

$$p=\frac{2\alpha\beta-2\alpha\rho\gamma-\alpha\gamma-\gamma^2\mu(1-\rho)\pm\Big|\gamma\left(\alpha(1-2\rho)+(\gamma-2\beta)(1-\rho)\mu\right)\Big|}{4(\beta-\rho\gamma)(\beta-\gamma)}. \tag{61}$$

Let the function $h(\rho):=\alpha(1-2\rho)+(\gamma-2\beta)(1-\rho)\mu$. Since we assume $\mu\geq p_{NE}=\frac{\alpha}{2\beta-\gamma}$, we have

$$h(0)=\alpha+(\gamma-2\beta)\mu\leq 0,\ h(1)=-\alpha<0.$$

Since $h(\rho)$ is a linear function of ρ, we have $h(\rho)\leq 0$ for any $\rho\in[0,1]$. So the roots of (56) are

$$p_L=\frac{\alpha\beta-2\alpha\gamma\rho-\beta\gamma(1-\rho)\mu}{2(\beta-\rho\gamma)(\beta-\gamma)},\ p_H=\frac{\alpha+\gamma\mu(1-\rho)}{2(\beta-\gamma\rho)}.$$

Note that $p_L<p_H=p_a$. Then the sufficient condition for $(56)\geq 0$ is that $p_L\leq p'\leq p_a$. Especially, if $p_L<p'<p_a$, $r(\hat{p})$ is strictly greater than $r(p_a)$.

Next, we prove that $p_L\leq p_{NE}\leq p_H$. First, we consider

$$\begin{aligned}p_{NE}-p_L&=\frac{\alpha}{2\beta-\gamma}-\frac{\alpha\beta-2\alpha\gamma\rho-\beta\gamma(1-\rho)\mu}{2(\beta-\rho\gamma)(\beta-\gamma)}\\&=\frac{2\beta\mu-\alpha-\gamma\mu+(2\alpha-2\beta\mu+\gamma\mu)\rho}{2(2\beta-\gamma)(\beta-\rho\gamma)(\beta-\gamma)}.\end{aligned} \tag{62}$$

Since $\beta>\gamma$, $\rho\in[0,1]$, we have the denominator in (62) is greater than zero. We claim that the numerator in (62) is greater than zero for all $\rho\in[0,1]$. To see this, let the linear function

$$g(\rho):=2\beta\mu-\alpha-\gamma\mu+(2\alpha-2\beta\mu+\gamma\mu)\rho.$$

We have $g(0)=2\beta\mu-\alpha-\gamma\mu\geq 0$ because of $\mu\geq p_{NE}=\frac{\alpha}{2\beta-\gamma}$. Also, we have $g(1)=\alpha>0$. Thus, for any $\rho\in[0,1]$, we have $g(\rho)\geq 0$. So $(62)\geq 0$ and $p_{NE}\geq p_L$. Next, we have

$$p_H-p_{NE}=\frac{g(\rho)}{2(2\beta-\gamma)(\beta-\rho\gamma)}=\frac{2\beta\mu-\alpha-\gamma\mu+(2\alpha-2\beta\mu+\gamma\mu)\rho}{2(2\beta-\gamma)(\beta-\rho\gamma)}\geq 0. \tag{63}$$

Combining (62) and (63), we have $p_L\leq p_{NE}\leq p_H$. □

C. Proofs in Section 4.2

Proof of Proposition 4. We first prove (i), the finite-sample result for the algorithmic decision. The OLS estimator is

$$\hat{\theta} = (A^\top A)^{-1} A^\top (f(\boldsymbol{p}) + \boldsymbol{\epsilon}), \tag{64}$$

where A is the design matrix

$$A = \begin{pmatrix} A_0 \\ A_1 \\ \vdots \\ A_n \end{pmatrix}_{(n+1)K\times 2}, \quad A_j = \begin{pmatrix} 1 & -p_j \\ 1 & -p_j \\ \vdots & \\ 1 & -p_j \end{pmatrix}_{K\times 2}, \quad f(\boldsymbol{p}) = \begin{pmatrix} f(p_0) \\ f(p_0) \\ \vdots \\ f(p_n) \end{pmatrix}_{(n+1)K\times 1}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{01} \\ \epsilon_{01} \\ \vdots \\ \epsilon_{nK} \end{pmatrix}_{(n+1)K\times 1}.$$

Then, we have

$$\begin{aligned} \frac{1}{(n+1)K} A^\top A &= \frac{1}{(n+1)K} \begin{pmatrix} A_0^\top & A_1^\top \cdots & A_n^\top \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \\ \vdots \\ A_n \end{pmatrix} = \frac{1}{(n+1)K} \sum_{i=0}^n A_i^\top A_i \\ &= \frac{1}{n+1} \sum_{i=0}^n \begin{pmatrix} 1 \\ -p_i \end{pmatrix} \begin{pmatrix} 1 & -p_i \end{pmatrix} = \frac{1}{n+1} \sum_{i=0}^n \begin{pmatrix} 1 & -p_i \\ -p_i & p_i^2 \end{pmatrix}. \end{aligned} \tag{65}$$

According to the price grid (18), we have

$$\frac{1}{n+1} \sum_{i=0}^n p_i = \frac{1}{n+1} \left(c(n+1) + \sum_{i=0}^n i \frac{\bar{p}-c}{n} \right) = \frac{1}{2}(\bar{p}+c), \tag{66}$$

$$\frac{1}{n+1} \sum_{i=0}^n p_i^2 = \frac{1}{n+1} \left(\sum_{i=0}^n c^2 + 2ic\frac{\bar{p}-c}{n} + i^2 \frac{(\bar{p}-c)^2}{n^2} \right) = c\bar{p} + \frac{2n+1}{6n}(\bar{p}-c)^2. \tag{67}$$

Plugging (66) and (67) into (65), we have

$$\begin{aligned} \frac{1}{(n+1)K} A^\top A &= \begin{pmatrix} 1 & -\frac{1}{2}(\bar{p}+c) \\ -\frac{1}{2}(\bar{p}+c) & c\bar{p} + \frac{2n+1}{6n}(\bar{p}-c)^2 \end{pmatrix}, \\ \left(\frac{1}{(n+1)K} A^\top A \right)^{-1} &= \frac{1}{\frac{n+2}{12n}(\bar{p}^2+c^2) - \frac{n+2}{6n}\bar{p}c} \begin{pmatrix} c\bar{p} + \frac{2n+1}{6n}(\bar{p}-c)^2 & \frac{1}{2}(\bar{p}+c) \\ \frac{1}{2}(\bar{p}+c) & 1 \end{pmatrix}, \end{aligned} \tag{68}$$

$$\frac{1}{(n+1)K} A^\top (f(\boldsymbol{p}) + \boldsymbol{\epsilon}) = \begin{pmatrix} \frac{1}{n+1}\sum_{i=0}^n f(p_i) + \frac{1}{(n+1)K}\sum_{i=0}^n \sum_{j=1}^K \epsilon_{ij} \\ -\frac{1}{n+1}\sum_{i=0}^n p_i f(p_i) - \frac{1}{(n+1)K}\sum_{i=0}^n p_i \sum_{j=1}^K \epsilon_{ij} \end{pmatrix} := \begin{pmatrix} c_{1n} \\ -c_{2n} \end{pmatrix}. \tag{69}$$

Plugging (68) and (69) into (64), we have

$$\hat{\theta} = \frac{1}{\frac{n+2}{12n}(\bar{p}^2+c^2) - \frac{n+2}{6n}\bar{p}c} \begin{pmatrix} c\bar{p}c_{1n} + \frac{2n+1}{6n}(\bar{p}-c)^2 c_{1n} - \frac{\bar{p}+c}{2}c_{2n} \\ \frac{\bar{p}+c}{2}c_{1n} - c_{2n} \end{pmatrix},$$

where c_{1n}, c_{2n} are defined in (69). The optimal price prescribed by AI is

$$p_a = \frac{\hat{\alpha}}{2\hat{\beta}} + \frac{c}{2} = \frac{c\bar{p}c_{1n} + \frac{2n+1}{6n}(\bar{p}-c)^2 c_{1n} - \frac{\bar{p}+c}{2}c_{2n}}{(\bar{p}+c)c_{1n} - 2c_{2n}} + \frac{c}{2}. \tag{70}$$

Let c_1, c_2 denote the estimator when $K \to \infty$:

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} := \begin{pmatrix} \frac{1}{n+1}\sum_{i=0}^n f(p_i) \\ \frac{1}{n+1}\sum_{i=0}^n p_i f(p_i) \end{pmatrix}. \tag{71}$$

In the first step, we will show that for small constants $\delta_1, \delta_2 > 0$,

$$|c_{1n} - c_1| \leq \delta_1,\ |c_{2n} - c_2| \leq \frac{\bar{p}+c}{2}\delta_2, \tag{72}$$

with a high probability. In the second step, since the AI price p_a (70) is continuous in c_{1n}, c_{2n}, we have $|p_a - p_a^*| \leq \delta$ with a high probability for a small constant δ.

Step one: Note that $c_{1n} - c_1$ is the average of $(n+1)K$ i.i.d. σ-sub-Gaussian variables:

$$c_{1n} - c_1 = \frac{1}{(n+1)K}\sum_{i=0}^{n}\sum_{j=1}^{K}\epsilon_{ij}.$$

By the concentration inequality (see Proposition 2.6.1 in Vershynin 2018), we have

$$\mathbb{P}\left(|c_{1n} - c_1| \geq \delta_1\right) \leq 2\exp\left(-\frac{\delta_1^2(n+1)K}{2\sigma^2}\right). \tag{73}$$

Similarly, we have

$$c_{2n} - c_2 = \frac{1}{(n+1)K}\sum_{i=0}^{n}p_i\sum_{j=1}^{K}\epsilon_{ij}.$$

By the definition of p_i, we have

$$\sum_{i=0}^{n}p_i^2 = (n+1)\left(c\bar{p} + \frac{2n+1}{6n}(\bar{p}-c)^2\right) \leq \frac{1}{2}(n+1)(\bar{p}+c)^2,$$

where the inequality follows by $\frac{2n+1}{6n} \leq 0.5$ due to $n \geq 1$. Thus, $c_{2n} - c_2$ is $\sqrt{\frac{(\bar{p}+c)^2}{2(n+1)K}}\sigma$-sub-Gaussian and

$$\mathbb{P}\left(|c_{2n} - c_2| \geq \frac{\bar{p}+c}{2}\delta_2\right) \leq 2\exp\left(-\frac{\delta_2^2(n+1)K}{4\sigma^2}\right). \tag{74}$$

Step two: Taking the partial derivative of p_a with respect to c_{1n}, c_{2n}, we have

$$\frac{\partial p_a}{\partial c_{1n}} = \frac{-\left(\frac{1}{6}+\frac{1}{3n}\right)(\bar{p}-c)^2 c_{2n}}{\left((\bar{p}+c)c_{1n} - 2c_{2n}\right)^2} \tag{75}$$

$$\frac{\partial p_a}{\partial c_{2n}} = \frac{\left(\frac{1}{6}+\frac{1}{3n}\right)(\bar{p}-c)^2 c_{1n}}{\left((\bar{p}+c)c_{1n} - 2c_{2n}\right)^2}. \tag{76}$$

We set $\delta_1 < c_1, \delta_2 < \frac{2c_2}{\bar{p}+c}$ in (72) to make sure $c_{1n}, c_{2n} > 0$. Thus, we have $\frac{\partial p_a}{\partial c_{1n}} < 0$ and $\frac{\partial p_a}{\partial c_{2n}} > 0$. On the one hand, when c_{1n}, c_{2n} satisfy (72), the AI price p_a attains the maximum when $c_{1n} = c_1 - \delta_1$, $c_{2n} = c_2 + \frac{\bar{p}+c}{2}\delta_2$. So we have an upper bound for p_a by plugging c_{1n}, c_{2n} into (70):

$$p_a \leq \frac{\left(c\bar{p} + \frac{2n+1}{6n}(\bar{p}-c)^2\right)(c_1 - \delta_1) - \frac{\bar{p}+c}{2}\left(c_2 + \frac{\bar{p}+c}{2}\delta_2\right)}{(\bar{p}+c)(c_1-\delta_1) - 2\left(c_2 + \frac{\bar{p}+c}{2}\delta_2\right)} + \frac{c}{2}. \tag{77}$$

If $K \to \infty$, we have $\delta_1 \to 0$, $\delta_2 \to 0$, and the AI price p_a converging to

$$p_a^* = \frac{c\bar{p}c_1 + \frac{2n+1}{6n}(\bar{p}-c)^2 c_1 - \frac{\bar{p}+c}{2}c_2}{(\bar{p}+c)c_1 - 2c_2} + \frac{c}{2}. \tag{78}$$

Let $A := c\bar{p} + \frac{2n+1}{6n}(\bar{p}-c)^2$, $B := \frac{\bar{p}+c}{2}$. By (77) and (78), we have

$$p_a \leq \frac{Ac_1 - Bc_2 - A\delta_1 - B^2\delta_2}{2Bc_1 - 2c_2 - 2B(\delta_1+\delta_2)} + \frac{c}{2},\ p_a^* = \frac{Ac_1 - Bc_2}{2Bc_1 - 2c_2} + \frac{c}{2}. \tag{79}$$

Thus,

$$p_a - p_a^* \leq \frac{2(A-B^2)(c_2\delta_1 + Bc_1\delta_2)}{(2Bc_1 - 2c_2 - 2B(\delta_1+\delta_2))\,(2Bc_1 - 2c_2)}, \tag{80}$$

where $A - B^2 = \left(\frac{1}{12} + \frac{1}{6n}\right)(\bar{p}-c)^2$. We let

$$\delta_1 = \frac{2(Bc_1-c_2)^2}{4c_2(A-B^2)}\delta = \frac{3\left((\bar{p}+c)c_1/2 - c_2\right)^2}{c_2(1/2+1/n)(\bar{p}-c)^2}\delta,\ \delta_2 = \frac{2(Bc_1-c_2)^2}{4Bc_1(A-B^2)}\delta = \frac{6\left((\bar{p}+c)c_1/2 - c_2\right)^2}{(\bar{p}+c)c_1(1/2+1/n)(\bar{p}-c)^2}\delta. \tag{81}$$

For a constant δ satisfying

$$\delta \leq \frac{c_1c_2(A-B^2)}{(Bc_1+c_2)(Bc_1-c_2)}, \tag{82}$$

we can check $\delta_1 < c_1, \delta_2 < \frac{2c_2}{\bar{p}+c}$ and

$$2B(\delta_1+\delta_2) \leq Bc_1 - c_2. \tag{83}$$

By (82), (80), (81) and (83), we have

$$p_a - p_a^* \leq \delta. \tag{84}$$

On the other hand, when c_{1n}, c_{2n} satisfy (72), the AI price p_a attains the minimum when $c_{1n} = c_1 + \delta_1$, $c_{2n} = c_2 - \frac{\bar{p}+c}{2}\delta_2$. So we have

$$p_a - p_a^* \geq \frac{2(B^2-A)(c_2\delta_1 + Bc_1\delta_2)}{(2Bc_1 - 2c_2 - 2B(\delta_1+\delta_2))\,(2Bc_1 - 2c_2)}. \tag{85}$$

By (82), (85), (81) and (83), we have

$$p_a - p_a^* \geq -\delta. \tag{86}$$

In summary, by (73), (74), we have

$$\mathbb{P}(|p_a - p_a^*| \geq \delta) \leq 2\exp\left(-\frac{\delta_1^2(n+1)K}{2\sigma^2}\right) + 2\exp\left(-\frac{\delta_2^2(n+1)K}{4\sigma^2}\right).$$

Next, we prove (ii), the finite-sample result for the range given by the human analyst. According to the property of strong concavity, we have the revenue $r(p) := (p-c)f(p)$ is unimodal and $r'(p_j) > 0$ for $p_j < p^*$. Thus, we have

$$r(p_{j-1}) \leq r(p_j) - r'(p_j)\frac{\bar{p}}{n} - \frac{\lambda}{2}\left(\frac{\bar{p}-c}{n}\right)^2 < r(p_j) - \frac{\lambda}{2}\left(\frac{\bar{p}-c}{n}\right)^2, \tag{87}$$

for $p_j < p^*$. Similarly, for $p_j \geq p^*$, due to $r'(p_j) < 0$, we have

$$r(p_{j+1}) \leq r(p_j) + r'(p_j)\frac{\bar{p}}{n} - \frac{\lambda}{2}\left(\frac{\bar{p}-c}{n}\right)^2 < r(p_j) - \frac{\lambda}{2}\left(\frac{\bar{p}-c}{n}\right)^2. \tag{88}$$

Let $\Delta := \frac{\lambda}{4}\left(\frac{\bar{p}-c}{n}\right)^2$. Due to (87), (88), we have

$$r(p_{j-1}) + \Delta < r(p_j) - \Delta \text{ for } p_j < p^*,\ r(p_j) - \Delta > r(p_{j+1}) + \Delta \text{ for } p_j \geq p^*. \tag{89}$$

By the concentration inequality of sub-Gaussain variables, we have

$$\begin{aligned}\mathbb{P}(|\hat{r}(p_j) - r(p_j)| \geq \Delta) = 2\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K}\epsilon_{jk} \geq \frac{\Delta}{p_j}\right) \leq 2\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K}\epsilon_{jk} \geq \frac{\Delta}{\bar{p}}\right)\\ \leq 2\exp\left(-\frac{K\Delta^2}{2\sigma^2\bar{p}^2}\right) = 2\exp\left(-\frac{K\lambda^2(\bar{p}-c)^4}{32\sigma^2\bar{p}^2n^4}\right).\end{aligned}$$

Define the good event $G = \{|\hat{r}(p_j) - r(p_j)| < \Delta,\ \forall j \in [0, 1, \ldots, n]\}$. We have

$$\mathbb{P}(G) \geq 1 - 2(n+1)\exp\left(-\frac{K\lambda^2(\bar{p}-c)^4}{32\sigma^2\bar{p}^2n^4}\right).$$

Let $j_1 := \max\{j : p_j < p^*\}$, $j_2 := \min\{j : p_j \geq p^*\}$. Note that $j_2 = j_1 + 1$. By (89) and under event G, we have

$$\begin{aligned}&\hat{r}(p_0) \leq r(p_0) + \Delta < r(p_1) - \Delta \leq \hat{r}(p_1) < \cdots < \hat{r}(p_{j_1}),\\ &\hat{r}(p_{j_2}) \geq r(p_{j_2}) - \Delta > r(p_{j_2+1}) + \Delta \geq \hat{r}(p_{j_2+1}) > \cdots > \hat{r}(p_n).\end{aligned}$$

Thus, the estimated revenue $r(p_j)$ strictly increases first, then strictly decreases. So the optimal index $j^* \in \{j_1, j_2\}$ and $p^* \in [p_{j_1}, p_{j_2}] \subset [p_{j^*-1}, p_{j^*+1}]$. □

Proof of Theorem 2. If $p_a \in [p_{j^*-1}, p_{j^*+1}]$, then $\hat{p} = p_a$ and $(\hat{p} - c)f(\hat{p}) = (p_a - c)f(p_a)$. If $p_a > p_{j^*+1}$, then $p^* \leq \hat{p} = p_{j^*+1} < p_a$ and $(\hat{p} - c)f(\hat{p}) \geq (p_a - c)f(p_a)$, since $(p - c)f(p)$ is unimodal. Also, we have $(\hat{p} - c)f(\hat{p}) \geq (p_a - c)f(p_a)$ when $p_a < p_{j^*-1}$. Therefore, we complete the proof. □

D. Proofs in Section 4.3

Proof of Lemma 2. Since there is a constant term in the covariate W, we define $W = (1\ W'^{\top})^{\top}$ and rewrite the true model and the contaminated model in the following:

$$X = \begin{pmatrix}1 & W'^{\top}\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} + \epsilon,\ X_C = \begin{pmatrix}1 & W'^{\top}\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} + B + \epsilon.$$

We rewrite the contaminated model in the matrix form:

$$\boldsymbol{X_C} = A\beta + \boldsymbol{B} + \boldsymbol{\epsilon},$$

where the vector of response, the design matrix, the vector of contamination and noise are

$$\boldsymbol{X_C} = \begin{pmatrix}X_{C1}\\ \vdots\\ X_{Cn}\end{pmatrix}, A = \begin{pmatrix}1 & W_1'^{\top}\\ \vdots & \vdots\\ 1 & W_n'^{\top}\end{pmatrix}, \boldsymbol{B} = \begin{pmatrix}B_1\\ \vdots\\ B_n\end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix}\epsilon_1\\ \vdots\\ \epsilon_n\end{pmatrix}.$$

The OLS estimator for the contaminated training samples is

$$\hat{\beta} = (A^\top A)^{-1} A^\top \boldsymbol{X_C} = (A^\top A)^{-1} A^\top (A\beta + \boldsymbol{B} + \boldsymbol{\epsilon}) = \beta + \left(\frac{1}{n} A^\top A\right)^{-1} \frac{1}{n} A^\top \boldsymbol{B} + \left(\frac{1}{n} A^\top A\right)^{-1} \frac{1}{n} A^\top \boldsymbol{\epsilon}. \tag{90}$$

By Corollary 3.1 in Wooldridge (2010), we have

$$\left(\frac{1}{n} A^\top A\right)^{-1} \xrightarrow{p} \left(\mathbb{E}[WW^\top]\right)^{-1} = \begin{pmatrix} 1 & \mathbb{E}[W']^\top \\ \mathbb{E}[W'] & \mathbb{E}[W'W'^\top] \end{pmatrix}^{-1}. \tag{91}$$

By the inverses of partitioned matrices (Greene 2003), we have (91)

$$= \begin{pmatrix} 1 + \mathbb{E}[W']^\top \left(\mathbb{E}[W'W'^\top] - \mathbb{E}[W']\mathbb{E}[W']^\top\right)^{-1} \mathbb{E}[W'] & -\mathbb{E}[W']^\top \left(\mathbb{E}[W'W'^\top] - \mathbb{E}[W']\mathbb{E}[W']^\top\right)^{-1} \\ -\left(\mathbb{E}[W'W'^\top] - \mathbb{E}[W']\mathbb{E}[W']^\top\right)^{-1} \mathbb{E}[W'] & \left(\mathbb{E}[W'W'^\top] - \mathbb{E}[W']\mathbb{E}[W']^\top\right)^{-1} \end{pmatrix}. \tag{92}$$

For the last term in (90), we have

$$\frac{1}{n} A^\top \epsilon = \begin{pmatrix} \frac{1}{n}\sum_{t=1}^{n} \epsilon_t \\ \frac{1}{n}\sum_{t=1}^{n} W_t' \epsilon_t \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \mathbb{E}[\epsilon_t] \\ \mathbb{E}[W_t' \epsilon_t] \end{pmatrix} = 0, \tag{93}$$

where $\xrightarrow{p}$ denotes convergence in probability, the convergence follows by the weak law of large numbers and the last equality follows by the independence of ϵ_t and W_t'. Then, by Slutsky's theorem and (91), (92), (93), we have

$$\left(\frac{1}{n} A^\top A\right)^{-1} \frac{1}{n} A^\top \boldsymbol{\epsilon} \xrightarrow{p} 0. \tag{94}$$

Considering the term including $\boldsymbol{B}$ in (90), we have

$$\frac{1}{n} A^\top \boldsymbol{B} = \begin{pmatrix} \frac{1}{n}\sum_{t=1}^{n} B_t \\ \frac{1}{n}\sum_{t=1}^{n} W_t' B_t \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \mathbb{E}[B] \\ \mathbb{E}[W'B] \end{pmatrix} = \begin{pmatrix} 1 \\ \mathbb{E}[W'] \end{pmatrix} \mathbb{E}[B], \tag{95}$$

where the last equality follows by the independence of B_t and W_t'.

By (92) and (95), we have

$$\left(\frac{1}{n} A^\top A\right)^{-1} \frac{1}{n} A^\top \boldsymbol{B} \xrightarrow{p} \begin{pmatrix} 1 \\ \boldsymbol{0} \end{pmatrix} \mathbb{E}[B]. \tag{96}$$

Finally, plugging (94), (96) into (90), we have

$$\hat{\beta} \xrightarrow{p} \beta + \begin{pmatrix} 1 \\ \boldsymbol{0} \end{pmatrix} \mathbb{E}[B], \ X_a(W) = W^\top \hat{\beta} \xrightarrow{p} W^\top \beta + \mathbb{E}[B]. \quad \square$$

Proof of Proposition 5. Recalling that $\mathbb{E}[B] = pb$, $X_a(W) = W^\top \beta + pb$ and the loss function is defined as the square error, i.e., $l(X_a(w), w) = (X_a(w) - x^*(w))^2$. We write down the difference of the losses,

$$\begin{aligned} l(X_a(w), w) - l(\hat{X}(w), w) &= (X_a(w) - w^\top \beta)^2 - (\hat{X}(w) - w^\top \beta)^2 \\ &= \left(p^2 b^2 - (X_h^u - w^\top \beta)^2\right) \mathbb{I}(X_h^u \le w^\top \beta + pb). \end{aligned} \tag{97}$$

Since X_h^u satisfies (27), we have

$$w^\top \beta \le X_h^u + pb, \ \forall \ w \in \mathcal{W}. \tag{98}$$

We will show (97) ≥ 0 by discussing w in the following three cases.

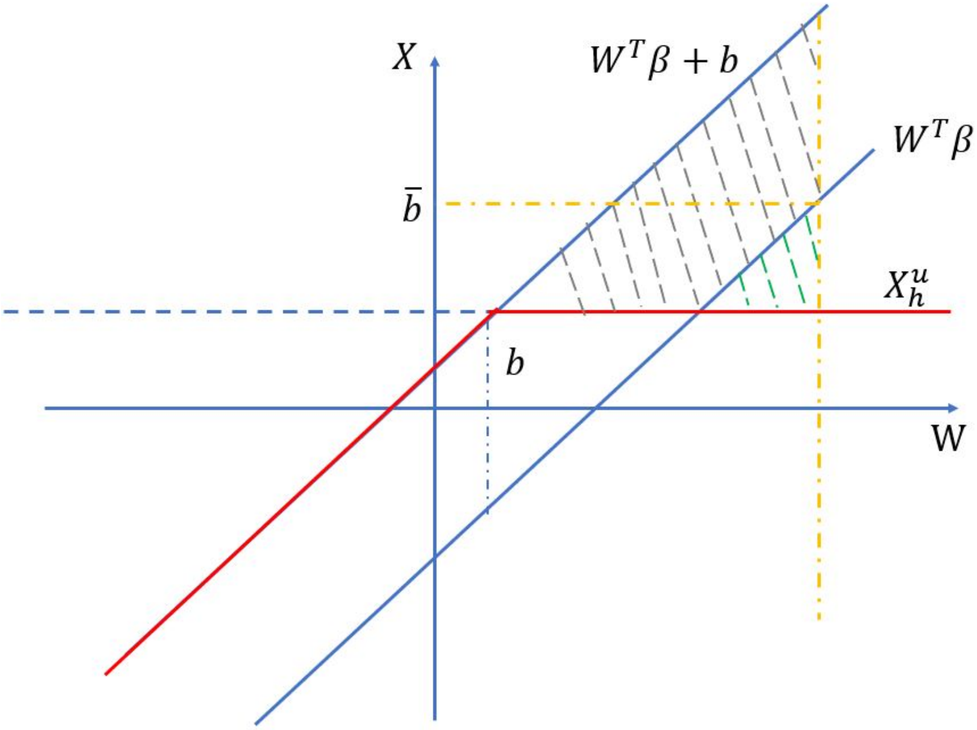

Figure 2 **Intuition behind the proof of Proposition 5.**

(1) For the w satisfying $X_h^u \le w^\top\beta \le X_h^u + pb$, we have $0 \le w^\top\beta - X_h^u \le pb$. Thus, $p^2b^2 - (X_h^u - w^\top\beta)^2 \ge 0$ and $(97) \ge 0$.

(2) For the w satisfying $X_h^u - pb \le w^\top\beta \le X_h^u$, we have $-pb \le w^\top\beta - X_h^u \le 0$. Thus, $p^2b^2 - (X_h^u - w^\top\beta)^2 \ge 0$ and $(97) \ge 0$.

(3) For the w satisfying $w^\top\beta \le X_h^u - pb$, we have $(97) = 0$.

We have proved $(97) \ge 0$ for any $w \in \mathcal{W}$. Thus, we have $l(X_a(w), w) \ge l(\hat{X}(w), w)$ for any $w \in \mathcal{W}$. □

Proof of Theorem 3.

Let $b := \mathbb{E}[B]$, then $X_a(W) = W^\top\beta + b$. We first write down the difference of the square-error losses:

$$\begin{aligned}
&l(X_a(w), w) - l(\hat{X}(w), w)\\
&= (X_a(w) - w^\top\beta)^2 - (\hat{X}(w) - w^\top\beta)^2\\
&\overset{(a)}{=} \left((X_a(w) - w^\top\beta)^2 - (X_h^u - w^\top\beta)^2\right)\mathbb{I}\left(X_h^u \le X_a(w)\right)\\
&\quad + \left((X_a(w) - w^\top\beta)^2 - (X_h^l - w^\top\beta)^2\right)\mathbb{I}\left(X_h^l \ge X_a(w)\right)\\
&\overset{(b)}{=} \left(b^2 - (X_h^u - w^\top\beta)^2\right)\mathbb{I}\left(X_h^u \le w^\top\beta + b\right) + \left(b^2 - (X_h^l - w^\top\beta)^2\right)\mathbb{I}\left(X_h^l \ge w^\top\beta + b\right),
\end{aligned} \tag{99}$$

where (a) follows from the definition of $\hat{X}(w)$, (b) follows from $X_a(W) = W^\top\beta + b$. Since X_h^l, X_h^u satisfies (28), we have for any $w \in \mathcal{W}$,

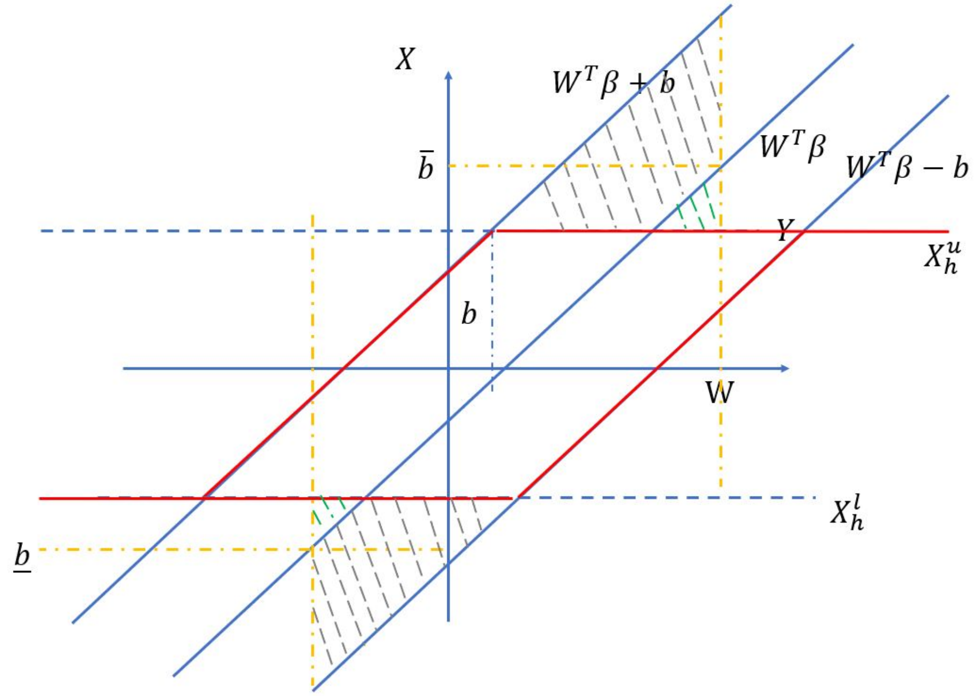

Figure 3 Intuition behind the proof of Theorem 3.

$$X_h^l - b \leq w^\top\beta \leq X_h^u + b, \text{ if } b \geq 0, \tag{100}$$

$$X_h^l + b \leq w^\top\beta \leq X_h^u - b, \text{ if } b < 0. \tag{101}$$

We will show (99) ≥ 0 by discussing the sign of b and range of w in the following cases.

(1) If $b \geq 0$, by (100), we have $w^T\beta - b \leq X_h^u$. For the first term in (99), if X_h^u further satisfies $X_h^u \leq w^T\beta + b$, we have $b^2 - (X_h^u - w^\top\beta)^2 \geq 0$. For the second term in (99), by (100), we have $X_h^l \leq w^T\beta + b$. So we have $(b^2 - (X_h^l - w^\top\beta)^2)\,\mathbb{I}\,(X_h^l \geq w^\top\beta + b) = 0$.

(2) If $b < 0$, we have $X_h^u \geq w^\top\beta + b$ due to (101). Thus, we have

$$\left(b^2 - (X_h^u - w^\top\beta)^2\right)\mathbb{I}\left(X_h^u \leq w^\top\beta + b\right) = 0.$$

Next, we will show $(b^2 - (X_h^l - w^\top\beta)^2)\,\mathbb{I}\,(X_h^l \geq w^\top\beta + b) \geq 0$.

(2.1) For the w satisfying $X_h^l + b \leq w^\top\beta \leq X_h^l$, we have $b \leq w^\top\beta - X_h^l \leq 0$. Thus, $b^2 - (X_h^l - w^\top\beta)^2 \geq 0$ and $(b^2 - (X_h^l - w^\top\beta)^2)\,\mathbb{I}\,(X_h^l \geq w^\top\beta + b) \geq 0$.

(2.2) For the w satisfying $X_h^l \leq w^\top\beta \leq X_h^l - b$, we have $0 \leq w^\top\beta - X_h^l \leq -b$. Thus, $b^2 - (X_h^l - w^\top\beta)^2 \geq 0$ and $(b^2 - (X_h^l - w^\top\beta)^2)\,\mathbb{I}\,(X_h^l \geq w^\top\beta + b) \geq 0$.

(2.3) For the w satisfying $X_h^l - b \leq w^\top\beta$, we have $(b^2 - (X_h^l - w^\top\beta)^2)\,\mathbb{I}\,(X_h^l \geq w^\top\beta + b) = 0$.

Therefore, we have proved (99) ≥ 0 for any $w \in \mathcal{W}$. That is, $l(\hat{X}(w), w) \leq l(X_a(w), w)$ for any $w \in \mathcal{W}$. Taking expectation of w, we have

$$\mathbb{E}[l(X_a(W), W)] - \mathbb{E}[l(\hat{X}(W), W)] = \mathbb{E}\left[(X_a(W) - W^\top\beta)^2\right] - \mathbb{E}\left[(\hat{X}(W) - W^\top\beta)^2\right] \geq 0. \quad \square$$

Proof of Lemma 3. Recall the observed covariate W, contamination error U and the true covariate $Z = W - U$. We have

$$X = Z^\top \beta + \epsilon = W^\top \beta - U^\top \beta + \epsilon. \tag{102}$$

Suppose the design matrix for W, U are A, B. Then the OLS estimator is

$$\hat{\beta} = (A^\top A)^{-1} A^\top (A\beta - B^\top \beta + \boldsymbol{\epsilon}) = \beta - (A^\top A)^{-1} A^\top B^\top \beta + (A^\top A)^{-1} A^\top \boldsymbol{\epsilon}. \tag{103}$$

We have

$$\frac{1}{n} A^\top A \xrightarrow{p} \mathbb{E}[WW^\top] \overset{(a)}{=} \mathbb{E}[ZZ^\top] + \mathbb{E}[UU^\top] \overset{(b)}{=} \Sigma_1 + \Sigma_2, \tag{104}$$

where (a) follows from the independence of W and U, (b) follows from Assumption 4. Also, we have

$$\frac{1}{n} A^\top B \xrightarrow{p} \mathbb{E}[(U+Z)U^\top] = \mathbb{E}[UU^\top] = \Sigma_2, \ (A^\top A)^{-1} A^\top \boldsymbol{\epsilon} \xrightarrow{p} 0. \tag{105}$$

Plugging (104) and (105) into (103), we have

$$\hat{\beta} \xrightarrow{p} \left(\mathcal{I} - (\Sigma_1 + \Sigma_2)^{-1} \Sigma_2\right) \beta.$$

Since Σ_1 is positive-definite and Σ_2 is positive semi-definite, $(\Sigma_1 + p\Sigma_2)^{-1}\Sigma_2 \beta = 0$ if and only if $\Sigma_2 \beta = 0$. □

Proof of Theorem 4. Recall that the OLS estimator $\hat{\beta} = \beta$ and $X_a(Z) = Z^\top \beta + U^\top \beta$. Note that Z is independent of U. We first fix a covariate z and reformulate $\mathbb{E}[l(X_a(z), z)] - \mathbb{E}[l(\hat{X}(z), z)]$, where the expectation is taken with respect to U. That is

$$\begin{aligned}
&\mathbb{E}[l(X_a(z), z)] - \mathbb{E}[l(\hat{X}(z), z)] \\
&= \mathbb{E}\left[(X_a(z) - z^\top \beta)^2\right] - \mathbb{E}\left[(\hat{X}(z) - z^\top \beta)^2\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\left((X_a(z) - z^\top \beta)^2 - (X_h^u - z^\top \beta)^2\right) \mathbb{I}\left(X_h^u \le X_a(z)\right)\right] \\
&\quad + \mathbb{E}\left[\left((X_a(z) - z^\top \beta)^2 - (X_h^l - z^\top \beta)^2\right) \mathbb{I}\left(X_h^l \ge X_a(z)\right)\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\left((U^\top \beta)^2 - (X_h^u - z^\top \beta)^2\right) \mathbb{I}(X_h^u \le z^\top \beta + U^\top \beta)\right] \\
&\quad + \mathbb{E}\left[\left((U^\top \beta)^2 - (X_h^l - z^\top \beta)^2\right) \mathbb{I}(X_h^l \ge z^\top \beta + U^\top \beta)\right] \\
&= \mathbb{E}\left[\left(B^2 - \overline{u}^2(z)\right) \mathbb{I}(\overline{u}(z) \le B)\right] + \mathbb{E}\left[\left(B^2 - \underline{u}^2(z)\right) \mathbb{I}(\underline{u}(z) \ge B)\right],
\end{aligned} \tag{106}$$

where

$$B := U^\top \beta, \ \overline{u}(z) := X_h^u - z^\top \beta, \ \underline{u}(z) := X_h^l - z^\top \beta. \tag{107}$$

Furthermore, $(a), (b)$ follow from the definition of $\hat{X}, X_a$. According to (29), we have

$$\mathbb{P}(B \ge b) \ge p, \ \mathbb{P}(B \le -b) \ge p. \tag{108}$$

According to (30) and $p \leq 0.5$, we have

$$b \overset{(a)}{\geq} \sqrt{\frac{p}{1-p}}b \overset{(b)}{\geq} \max_{z\in\mathcal{Z}}\{z^\top\beta\} - X_h^u \overset{(c)}{\geq} -\overline{u}(z),\ -b \leq -\sqrt{\frac{p}{1-p}}b \leq \min_{z\in\mathcal{Z}}\{z^\top\beta\} - X_h^l \leq -\underline{u}(z),\ \forall z \in \mathcal{Z}, \tag{109}$$

where (a) follows from $p \leq 0.5$, (b) follows from (30), (c) follows from (107). Next, we will show (106) ≥ 0 by discussing the range of z. The second series of inequality can be derived similarly.

(1) For the z satisfying $z^\top\beta \geq X_h^u$, we have $\overline{u}(z) \leq 0$ and $\underline{u}(z) \leq 0$ by (107). Thus, we have

$$\mathbb{E}\left[\left(B^2 - \underline{u}^2(z)\right)\mathbb{I}(\underline{u}(z) \geq B)\right] \geq 0, \tag{110}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left(B^2 - \overline{u}^2(z)\right)\mathbb{I}(\overline{u}(z) \leq B)\right] \\
&\overset{(a)}{=} \mathbb{P}(B \geq b)\left((B^2 - \overline{u}(z)^2)\mathbb{I}(\overline{u}(z) \leq B)\right) + \mathbb{P}(-b \leq B \leq b)\left(B^2 - \overline{u}^2(z)\right)\mathbb{I}(\overline{u}(z) \leq B) \\
&\overset{(b)}{\geq} \mathbb{P}(B \geq b)\left((b^2 - \overline{u}(z)^2)\mathbb{I}(\overline{u}(z) \leq b)\right) + \mathbb{P}(-b \leq B \leq b)\left(B^2 - \overline{u}^2(z)\right)\mathbb{I}(\overline{u}(z) \leq B) \\
&\overset{(c)}{\geq} \mathbb{P}(B \geq b)\left((b^2 - \overline{u}(z)^2)\mathbb{I}(\overline{u}(z) \leq b)\right) - \left(\mathbb{P}(-b \leq B \leq b)\right)\overline{u}^2(z)\mathbb{I}(\overline{u}(z) \leq 0) \\
&\overset{(d)}{\geq} p(b^2 - \overline{u}^2(z)) - (1-2p)\overline{u}^2(z) \\
&= pb^2 - (1-p)\overline{u}^2(z) \\
&= p_1 b^2 - (1-p)(X_h^u - z^\top\beta)^2 \\
&\overset{(e)}{\geq} 0,
\end{aligned} \tag{111}$$

where (a) follows by $\mathbb{P}(B < -b) = 0$ due to $b \geq -\overline{u}(z) \geq -B$ by (109), (b) holds by $(B^2 - \overline{u}(z)^2)\mathbb{I}(\overline{u}(z) \leq B)$ increasing in B when $B \geq b$, (c) holds by $\left(B^2 - \overline{u}^2(z)\right)\mathbb{I}(\overline{u}(z) \leq B) \geq -\overline{u}^2(z)\mathbb{I}(\overline{u}(z) \leq 0)$ when $-b \leq B \leq b$ due to $-b \leq \overline{u}(z) \leq 0$ by (109), (d) follows by (108) and $\overline{u}(z) \leq 0 \leq -\overline{u}(z) \leq b$ by (109), (e) follows by (30). Plugging (110) and (111) into (106), we have (106) ≥ 0.

(2) For the z satisfying $X_h^l \leq z^\top\beta \leq X_h^u$, we have $\overline{u}(z) \geq 0$ and $\underline{u}(z) \leq 0$ by (107). Thus, we have $\mathbb{E}\left[\left(B^2 - \overline{u}^2(z)\right)\mathbb{I}(\overline{u}(z) \leq B)\right] \geq 0$ and $\mathbb{E}\left[(B^2 - \underline{u}^2(z))\,\mathbb{I}(\underline{u}(z) \geq B)\right] \geq 0$. Then, (106) ≥ 0.

(3) For the z satisfying $z^\top\beta \leq X_h^l$, we have $\overline{u}(z) \geq 0$ and $\underline{u}(z) \geq 0$ by (107). Thus, we have

$$\mathbb{E}\left[\left(B^2 - \overline{u}^2(z)\right)\mathbb{I}(\overline{u}(z) \leq B)\right] \geq 0, \tag{112}$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left(B^2 - \underline{u}^2(z)\right)\mathbb{I}(\underline{u}(z) \geq B)\right] \\
&\overset{(a)}{=} \mathbb{P}(B \leq -b)\left((B^2 - \underline{u}(z)^2)\mathbb{I}(\underline{u}(z) \geq B)\right) - \mathbb{P}(-b \leq B \leq b)\left(B^2 - \underline{u}^2(z)\right)\mathbb{I}(\underline{u}(z) \geq B) \\
&\overset{(b)}{\geq} \mathbb{P}(B \leq -b)\left((b^2 - \underline{u}(z)^2)\mathbb{I}(\underline{u}(z) \geq -b)\right) - \left(\mathbb{P}(-b \leq B \leq b)\right)\underline{u}^2(z)\mathbb{I}(\overline{u}(z) \geq 0)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\geq} p(b^2 - \underline{u}^2(z)) - (1-2p)\underline{u}^2(z) \\
&= pb^2 - (1-p)\underline{u}^2(z) \\
&= pb^2 - (1-p)(X_h^l - z^\top \beta)^2 \\
&\overset{(d)}{\geq} 0,
\end{aligned}
\tag{113}
$$

where (a) follows by $\mathbb{P}(B > b) = 0$ due to $-b \leq -\underline{u}(z) \leq -B$ by (109), (b) follows by $(B^2 - \underline{u}(z)^2)\mathbb{I}(\underline{u}(z) \geq -B)$ decreasing in B when $B \leq -b$ and $(B^2 - \underline{u}^2(z))\mathbb{I}(\underline{u}(z) \geq B) \geq -\underline{u}^2(z)\mathbb{I}(\underline{u}(z) \geq 0)$ when $-b \leq B \leq b$ due to $0 \leq \underline{u}(z) \leq b$ by (109), (c) follows by (108) and $\underline{u}(z) \geq 0 \geq -b$, (d) follows by (30). Plugging (112), (113) into (106), we have (106) ≥ 0.

Therefore, we have proved (106) ≥ 0 for any $z \in \mathcal{Z}$. Finally, taking expectation with respect to Z, we have

$$
\mathbb{E}[l(X_a(W), Z)] - \mathbb{E}[l(\hat{X}(W), Z)] = \mathbb{E}\left[(X_a(W) - Z^\top \beta)^2\right] - \mathbb{E}\left[(\hat{X}(W) - Z^\top \beta)^2\right] \geq 0. \quad \square
$$