Bias-Variance Games*

Yiding Feng† Ronen Gradwohl‡ Jason Hartline§ Aleck Johnsen¶
Denis Nekipelov‖

Abstract

Firms increasingly rely on predictive analytics via machine learning algorithms to drive a wide array of managerial decisions. In this paper, we study the effect of competition on the choice of such algorithms, focusing on the tradeoffs between bias and variance in the algorithms' predictions. Absent competition, firms care only about the magnitude of predictive error and not its source. With competition, however, firms prefer to incur error caused by variance over error caused by bias, even at the cost of higher total error.

*This work was initiated during the Special Quarter on Data Science and Online Markets held in the spring of 2018 at Northwestern University when the fifth author was supported as a McCormick Advisory Council Visiting Associate Professor. The first, second, and fourth authors gratefully acknowledge the support of National Science Foundation award number 1718670. The first, third, and fourth authors gratefully acknowledge the support of National Science Foundation award number 1618502.

†Microsoft Research. Email: yidingfeng@microsoft.com.

‡Department of Economics and Business Administration, Ariel University. Email: roneng@ariel.ac.il.

§Department of Computer Science, Northwestern University. Email: hartline@northwestern.edu

¶Department of Computer Science, Northwestern University. Email: aleckjohnsen@u.northwestern.edu

‖Departments of Economics and Computer Science, University of Virginia. Email: denis@virginia.edu

1 Introduction

Firms that engage in electronic commerce increasingly rely on predictive analytics to drive a wide array of managerial decisions, ranging from product recommendations to customer targeting and pricing. Given data, perhaps from past consumer behavior, a firm facing a potential customer will use predictive models or learning algorithms (henceforth algorithms) to anticipate the customer's future behavior and preferences, allowing the firm to better tailor its recommendation, targeting, and pricing decisions. In general, the success of such predictive analyses depends on the effectiveness of the algorithms used. Research on such algorithms has proliferated, and their capabilities have advanced incredibly over the past couple of decades.

The point of departure for this paper is the observation that, in many applications, firms that utilize predictive analytics do so in a competitive environment, and so the efficacy of a firm's analytics depends not only on its own expertise and technology but also on that of its competitors. In this paper we address the question of how the competitive nature of the interaction affects a firm's choice of algorithms. For example, while a particular algorithm may be best for a monopolistic firm targeting a customer, it may be suboptimal in a competitive environment. Furthermore, the optimal choice of algorithm in the competitive environment may depend on the competitors' choices of algorithms.

For a concrete example, consider the increasingly popular box subscription companies that mail personalized monthly boxes of fashion, food, or other products to subscribers. The appeal of these companies lies in their high level of personalization, often achieved by machine learning algorithms (Sinha et al., 2016). Stitch Fix, for example, uses such algorithms to predict each subscriber's fashion taste and sends a box matching this taste (Gaudin, 2016). The better the fit, the more satisfied the subscriber, leading to greater customer acquisition and retention. Of course, Stitch Fix competes with other companies, such as Trunk Club, that also personalize their boxes using learning algorithms. Ultimately, the profitability of such a company will depend not only on how well it manages to predict a customer's fashion taste, but also on the predictions of its competitors.

For a more general example, consider the so-called "Long Tail" marketplaces, which are characterized by huge numbers of goods that individually have low demand but that collectively make up substantial market share. One of the key drivers of Long Tail markets is the ability of firms to connect supply and demand, typically through machine learning algorithms that predict consumers' tastes and match them to products (Anderson, 2006). Often, many firms compete in the same Long Tail market, and in this case their success

depends not only on their own ability to match goods to consumers, but also on the predictive ability of their competitors.

One useful way of analyzing algorithms' predictive ability is by examining the different sources of error they incur. There are two general types of errors: a lack of accuracy—called bias—in which the predictions are not, on average, equal to the true value; and a lack of precision—called variance—in which the predictions are not clustered tightly around their average. The total error of an algorithm can be decomposed into these two kinds of error.

In practice, there are various ways to control the bias and variance of an algorithm. For example, one could allow the algorithm to consider more complex functions to map data onto predictions, such as deeper decision trees or regressions with higher degree functions, which result in lower bias but higher variance. Alternatively, a technique called regularization—intuitively, penalizing predictions that are less smooth—is often used to decrease variance at the expense of higher bias. Finally, increasing the amount of training data decreases variance. Algorithms that predict well are ones that control the tradeoff between bias and variance so as to minimize the total error, regardless of its source.

In this paper we aim to understand how competition affects the optimal way to trade off bias and variance. We show that, holding total error fixed, absent competition there is no preference for variance versus bias. In contrast, in competitive environments it is better to reduce bias at the expense of variance, even when this leads to higher total error. This result holds up under several natural theoretical models of predictive error and in an empirical study. Consequently, training an algorithm in isolation to minimize error does not lead to optimal parameter settings for algorithms in competitive environments. An implication of these results is that, in competitive environments, there is an added benefit to algorithms that consider more complex functions and an added cost to regularization.

Overview of model and results. In this paper we model the interaction of two firms as a game and analyze that game's equilibria. Players' actions are algorithms, and their payoffs in expectation depend both on the error of their chosen algorithm's prediction and on whether or not their algorithm's prediction is better than that of their opponent. The primary analysis of the paper abstracts away from the details of specific algorithms, and instead models an algorithm as a probability distribution over prediction errors. Thus, players' actions correspond to probability distributions and players' action spaces—the sets of possible actions they can choose—correspond to families of probability distributions that range over biases and variances. A canonical example is the set of normal distributions with different means and standard deviations. Finally, we focus on a standard measure of total

error, namely, mean squared error.

Our main theoretical result considers a two-player game where each player's action space consists of normal distributions with the same total mean squared error—namely, they have the same total error but different biases and variances. Absent competition, a player would be indifferent among all these distributions. In contrast, we prove that, in the competitive scenario, each player would prefer error distributions with lower bias (and therefore higher variance), and that this holds regardless of the actual prediction made by the opponent. In game theoretic terms, minimal-bias is an *ex post dominant strategy*. This strong result persists in games with more than two players. It also implies that the unique *Nash equilibrium* of the game is the one in which each players chooses the distribution with minimal bias.

We supplement this theoretical result with numerical analyses that demonstrate the robustness of the theoretical finding. First, we numerically test the robustness of our insight on the strategic preference for reduced bias with non-normal families of distributions, such as Laplace, logistic, and uniform. Our insight persists for many of the variations, but, notably, it fails for uniform distributions. Second, we investigate the dependence of the results on the form of players' payoff functions. In particular, we study variations in the benefit from winning relative to the cost of prediction errors. We find by numerical calculation that the *ex post* preference for lower bias fails to extend. However, we also find that minimal-bias remains a dominant strategy—that is, players prefer lower bias (and higher variance) for every choice of probability distribution by the opponent (although not for every realization of this distribution).

Finally, we conduct an empirical study of our bias-variance game for a family of learning algorithms on a benchmark dataset. In this study, players utilize a particular learning algorithm to predict housing prices, given a particular dataset.[1] Specifically, the players use a ridge regression algorithm, which is a variant of a linear regression that allows for flexibility to control the bias-variance tradeoff using a *regularization parameter*. When there is only one player we show that the optimal choice of regularization parameter is large, but when there are two players, payoffs increase as the parameter is lowered. In other words, in the latter scenario there is a preference for lower bias and higher variance. Thus, the algorithmic optimizations of the non-competitive and competitive settings are qualitatively distinct and result in quite different preferences with respect to the tradeoff between bias and variance.

To give some context for the theoretical results, we provide a few additional observa-

[1]We use the California housing prices data from the 1990 Census, a dataset first utilized by Pace and Barry (1997) and included in the Python Scikit-learn library.

tions about algorithms in competitive situations. Counterintuitively, we show that there are families of distributions and opponent choices for which a player prefers a distribution with higher bias (respectively, higher variance) even while holding variance (respectively, bias) fixed. Nonetheless, we also show that higher bias (with variance fixed) is not beneficial for natural families of distributions, such as normal and Laplace. Moreover, for normal distributions, our main theoretical result (described above) strengthens this conclusion on the harmfulness of higher bias by showing that decreasing bias is beneficial even at the expense of increased variance (holding the total error fixed). The above counterintuitive observation—the possibility that increasing bias can be beneficial—highlights the obstacles that our main theoretical analysis must overcome.

Related literature. The analysis of strategic interactions that involve machine learning algorithms is a newly burgeoning area of study in both economics and computer science. For example, Eliaz and Spiegler (2019) study the interaction of a rational agent and a learning algorithm, and consider the question of whether the agent has an incentive to truthfully report her information to the algorithm. Liang (2019) and Olea et al. (2019) study scenarios in which there are multiple algorithms that compete with one another. Liang (2019) considers games of incomplete information in which the players have data and use algorithms to derive their beliefs. Olea et al. (2019) study a game between agents competing to predict a common variable, and where agents obtain the same data but differ in the algorithms they utilize for prediction. In all these papers, the algorithms under consideration are fixed exogenously. Our paper, in contrast, focuses on the strategic choice of algorithms in competitive environments.

On the computer science side, our study is related to the "dueling algorithms" framework of Immorlica et al. (2011). Within this framework, Ben-Porat and Tennenholtz (2019), building on Ben-Porat and Tennenholtz (2017), study the problem of multiple learners selecting algorithms to make predictions within a particular dataset. They work within the PAC-learning framework of Valiant (1984), and consider equilibria in the game where, for each point in the dataset, a payoff of one is split evenly between all players whose predictions are within a given error tolerance. A key point of difference between this setup and ours is that they consider competition between specific algorithms, such as linear regressors, and study the questions of whether equilibria exist and can be learned. On the other hand, we study the general tradeoff between bias and variance in the equilibrium choice of algorithms.

Organization. Section 2 introduces the model, the one- and two-player games that we study, and some basic properties. Section 3 considers the one-player intuition that reducing

bias or reducing variance is always beneficial, when everything else is held fixed, in the two-player game. Counterintuitively, there are two-player scenarios where a player would want to increase bias or variance. On the other hand, for normally-distributed errors, reducing bias to zero is beneficial. Section 4 contains our main analysis of the two-player game. For normally-distributed error, we show that there is an *ex post* preference for lowering bias at the expense of variance, we show that the natural ridge regression algorithm indeed has normally-distributed error, and we present simulation results for other distributions of error and a variety of utility functions. In Section 5 we then consider the game on a standard benchmark data set with ridge regression, and show that qualitative conclusions of our theoretical analysis continue to hold. Finally, Section 6 concludes with some discussion.

2 Model and Preliminaries

In general, a learning algorithm A takes training set $D = \{(x_1, f(x_1)), \ldots, (x_m, f(x_m))\}$ as input, and outputs a hypothesis $\hat{f}$ that maps features of new data points to a predicted label. Given a new point $(x, f(x))$, an algorithm's prediction of the true label $f(x)$ is denoted $\hat{f}(x)$. In this paper, we will abstract away from the algorithm itself, and will only be concerned with the algorithm's error in prediction, $|f(x) - \hat{f}(x)|$, viewed as a distribution over $\mathbb{R}$. The randomness in this distribution arises both from the algorithm A itself and from the inherent randomness of the dataset and the choice of x. The best prediction is thus one with predictive error of 0.

We will focus on a standard measure of total error, namely, mean squared error. Thus, if the predictive error of an algorithm is a number $|a|$, then the squared error is a^2.

This paper is concerned with players' choices of algorithms. In our abstraction we will thus let each player choose a distribution over errors, from some class of distributions, as follows. Fix a random variable Z with mean 0 and standard deviation 1. A common choice for Z will be normal, a choice we motivate in Section 4.2, but we can also consider uniform, triangle, Laplace, and other distributions. Players will then choose a distribution from a class made up of shifts and spreads of Z. A typical example will be the class

$$\mathcal{X}_Z = \left\{\sigma Z + \mu : \mu^2 + \sigma^2 \geq 1\right\},$$

the set of distributions whose bias squared plus variance is at least 1 (which means their total squared error is at least 1). Each element of this class yields a different distribution over predictive error, and represents a different algorithm whose error has the corresponding distribution.

Some discussion of this modelling assumption is warranted. The class of distributions $\mathcal{X}_Z$ represents the error distributions of *all* algorithms available to a player. In particular, this precludes the possibility that a player chooses an algorithm, observes the realized prediction, and then uses some alteration of that prediction. For a concrete example of what this implies, consider the distribution X with bias $\mu = 1$ and variance $\sigma^2 = 0$, and observe that $X \in \mathcal{X}_Z$ above. Then our assumption precludes the possibility that the player takes the realization of X and subtracts 1 from it. Although this results in a prediction with 0 total error, and is thus a perfect predictor, our view is that if a perfect predictor were available to the player then it would be included in the set $\mathcal{X}_Z$.

2.1 One player

As a benchmark, consider a setting in which there is only one player, and suppose that she chooses an error distribution X. On realization a (that is, a prediction with predictive error a), let the player's utility be $u(a) = 1 - a^2$: a benefit of 1 minus her squared error. This utility function captures the idea that the player obtains positive utility from making a perfect prediction (the benefit of 1), but that this utility decreases with the squared error (the loss of a^2). In the box subscription company example from the introduction, the benefit represents future profits from a particular customer, whereas the loss accounts for the lack of customer retention in case of inaccurate taste predictions. Note that the total utility $1 - a^2$ could be negative; in the example this would represent a firm's failure to recoup the costs of an initial loss leader or promotion. We focus on this utility function for our theoretical analysis, but in Section 4.3 we argue that our results are robust to other specifications of the utility function, and in particular to ones where the chance of a negative payoff is negligible.

Given the utility function $u(a) = 1 - a^2$, a player's expected utility from X is $u(X) = \mathbf{E}[u(X)] = 1 - \mathbf{E}[X^2]$. The following is a simple observation:

Claim 2.1. *If $X = \sigma Z + \mu$, where Z is a random variable with mean 0 and variance 1, then $u(X) = 1 - \mu^2 - \sigma^2$.*

Proof. X is a random variable with variance σ^2 and expected value μ. Since $\sigma^2 = \mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \mathbf{E}[X^2] - \mu^2$, it follows that $\mathbf{E}[X^2] = \sigma^2 + \mu^2$. Thus, $u(X) = 1 - \mathbf{E}[X^2] = 1 - \mu^2 - \sigma^2$. □

This observation implies three corollaries, formalized below. First, for fixed bias, the player prefers minimal variance. Similarly, for fixed variance, the player prefers minimal bias. Finally, the player is indifferent between distributions that have the same bias squared

plus variance. Although these corollaries are straightforward, in Section 3 we will show that, without further assumptions, none of them hold in the competitive setting.

Corollary 2.2. *For any μ, if $\sigma < \sigma'$, then the player prefers $X = (\sigma Z + \mu)$ to $X' = (\sigma' Z + \mu)$.*

Corollary 2.3. *For any σ, if $\mu^2 < \nu^2$, then the player prefers $X = (\sigma Z + \mu)$ to $X' = (\sigma Z + \nu)$.*

Corollary 2.4. *If $X = \sigma Z + \mu$, $X' = \tau Z + \nu$, and $\mu^2 + \sigma^2 = \nu^2 + \tau^2$, then the player is indifferent between X and X'.*

Individual rationality. An additional definition that will be useful is that of individual rationality. Intuitively, a distribution is individually rational if a player derives non-negative utility from choosing it. Formally:

Definition 2.1. *A distribution X satisfies* individual rationality (IR) *if $u(X) \geq 0$.*

A simple observation that follows from Claim 2.1 is that $X = \sigma Z + \mu$ is IR if and only if $\mu^2 + \sigma^2 \leq 1$.

2.2 Two players

Recall that our goal is to analyze the effect of competition on firms' choices of learning algorithms. As described in the introduction, we will model this as a game, and we call this game the *Bias-Variance Game*. In this game, each of two players simultaneously chooses an error distribution. Prediction errors are realized, and the player with lower prediction error, say a_i, obtains utility $1 - a_i^2$, just like the one-player case. The player with higher prediction error obtains utility 0.

This specification of utilities, as well as variants that we discuss in Section 4.3, capture the main competitive force we wish to analyze: the desire of a player both to minimize error (this is the $-a_i^2$ term), and to obtain lower error than her competitor (this is the benefit of 1 from winning).

In our notation, when discussing player $i \in \{1, 2\}$, we will denote by $j = 3 - i$ the identity of the other player. A formal description of the game follows:

Definition 2.2 (The Bias-Variance Game)**.** *Given two classes of distributions, $\mathcal{X}_1$ and $\mathcal{X}_2$, the two player bias-variance game proceeds as follows:*

1. *Each player $k \in \{1, 2\}$ simultaneously chooses a distribution X_k from $\mathcal{X}_k$.*

2. *Each X_k is realized as some a_k.*

3. *Each player i obtains utility $u_i(a_i, a_j) = (1 - a_i^2) \cdot \mathbf{1}\{a_i < a_j\}$. That is, the player $i \in \operatorname{argmin}_k a_k$ wins and obtains utility $u_i(a_i, a_j) = 1 - a_i^2$; the other player, j, loses and obtains utility $u_j(a_j, a_i) = 0$.*

In our theoretical analysis we will primarily consider bias-variance games where $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$ is a family of distributions in which the error of every $X \in \mathcal{X}$ is normalized to $\mu^2 + \sigma^2 = 1$. In our numerical analysis in Section 4.3 we relax this restriction.

Individual rationality. As in the one-player benchmark, we will be interested in player choices that are individually rational. A straightforward result is that individual rationality of the one-player setting implies individual rationality of the two-player setting, as the following proposition demonstrates:

Proposition 2.5. *Bias-variance games with distributions $X \in \mathcal{X}$ satisfying $\mu^2 + \sigma^2 \leq 1$ are individually rational: namely, expected payoffs in the game are non-negative.*

Proof. In the one-player setting player i's expected payoff with realization $a_i \sim X_i$ is $\mathbf{E}[1 - a_i^2] \geq 0$ which satisfies individual rationality. Consider the two-player game where the other player $j = 3 - i$ has realization a_j. When the other player's realization is $|a_j| \leq 1$, the payoff of player i is non-negative for all a_i: for $|a_i| \leq |a_j| \leq 1$ player i wins and has non-negative payoff $1 - a_i^2$ and for $|a_i| > |a_j|$ then player i loses and has payoff 0. When the other player's realization is $|a_j| > 1$, then player i's distribution of payoffs in the two player game dominates his distribution of payoffs in the one-player setting (when player i loses, instead of a negative payoff her payoff is zero). As the latter setting had non-negative expectation, so does the former. □

Solution concepts. We will consider various solution concepts. For our main theoretical result on the preference of variance over bias we will utilize a very strong notion, namely that of ex post dominant strategies. A strategy is ex post dominant for player i if it yields that player the highest utility regardless of the *realized* prediction of the opponent.

Definition 2.3. *A strategy $X_i \in \mathcal{X}_i$ is* ex post dominant *for player i if for all $X_i' \in \mathcal{X}_i$ and all realizations a_j of player j it holds that $u_i(X_i, a_j) \geq u_i(X_i', a_j)$.*

For our numerical and empirical results we will consider two weaker notions. The first, dominant strategies, requires that a strategy be optimal against any strategy of the opponent, but not necessarily against any realization of that strategy:

Definition 2.4. *A strategy $X_i \in \mathcal{X}_i$ is* dominant *for player i if for all $X_i' \in \mathcal{X}_i$ and all $X_j \in \mathcal{X}_j$ it holds that $u_i(X_i, X_j) \geq u_i(X_i', X_j)$. If the inequality is strict for all $X_i' \neq X_i$, then X_i is* strictly dominant.

The second, pure Nash equilibrium, does not require that a strategy be optimal against any strategy of the opponent, but only against that player's own Nash equilibrium strategy:

Definition 2.5. *A strategy profile (X_1, X_2) is a* pure Nash equilibrium *if for each player i and strategy $X_i' \in \mathcal{X}_i$ it holds that $u_i(X_i, X_j) \geq u_i(X_i', X_j)$.*

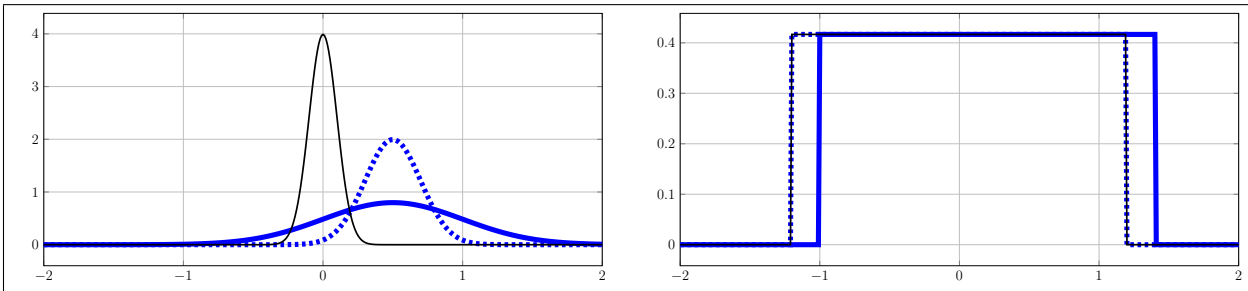
Observe that if X_i is ex post dominant, then it is also dominant. Furthermore, if X_1 and X_2 are dominant for players 1 and 2, respectively, then (X_1, X_2) is a pure Nash equilibrium. Finally, if X_1 and X_2 are strictly dominant then (X_1, X_2) is the *unique* Nash equilibrium.

3 Reducing Bias or Variance, All Else Fixed

We begin our analysis by describing some counterintuitive implications of competition. In particular, we show that the simple corollaries from the one-player benchmark in Section 2.1 no longer hold, and that reducing bias (resp., variance) holding variance (resp., bias) fixed can be harmful.

Example 3.1 (Reducing variance can be harmful; see Figure 1a)**.** *Suppose player 2 plays the distribution $\mathcal{N}(0, \varepsilon)$, where ε is some small number, and player 1 plays the distribution $\mathcal{N}(1/2, 1/2)$. Player 1's strategy is monotone and satisfies IR, and she obtains positive expected utility: Given that she wins, she is likely within ε of 0, and she wins with positive probability. However, if player 1 decreases her variance to 0, she will obtain utility close to 0, since she will hardly ever win (for small enough ε).*

Example 3.2 (Reducing bias can be harmful; see Figure 1b)**.** *Player 2 plays the uniform distribution on the interval $[-1-\varepsilon, 1+\varepsilon]$. For small enough $\varepsilon > 0$ this satisfies IR. Player 1 plays the uniform distribution on the interval $[-1, 1+2\varepsilon]$. Again, for small enough $\varepsilon > 0$ this satisfies IR. Now consider a deviation by Player 1 to the interval $[-1-\varepsilon, 1+\varepsilon]$, a deviation that reduces bias. This is harmful: Before the deviation, she never won when her realization was in $(1+\varepsilon, 1+2\varepsilon]$. After the deviation, however, the only difference is the additional possibility of winning when her realization is in $[-1-\varepsilon, -1)$. But such victories are harmful, as they consist only of negative utilities.*



(a) Example 3.1 (b) Example 3.2

Figure 1: Illustrations for Example 3.1 and Example 3.2. Thin black curves are probability density functions of Player 2's distributions. Thick blue (dashed) curves are probability density functions of Player 1's distributions (after reducing variance or bias).

Unlike Example 3.1, Example 3.2 is somewhat unnatural for our application to machine learning algorithms, in that the error distributions used are uniform. In Section 4.2 we argue that we should expect the distribution of the predictions of learning algorithms to be normal. Is there an example in which reducing bias is harmful, but where the class of distributions is more natural? The following two theorems state that there is not. The first considers a class of distributions that are single-peaked, monotone, and with convex tails,[2] and the second considers normal distributions.

Theorem 3.1. *Let Z be monotonically increasing, convex on $[-\infty, 0]$, and symmetric around 0. Let $X_i = \sigma Z + \mu$ be IR (so as to satisfy $\mu^2 + \sigma^2 \leq 1$) and $X_i' = \sigma Z$. Then $u_i(X_i', c) \geq u_i(X_i, c)$ for any realization c of player j.*

The proof of this theorem and Theorem 3.2, below, are given in Appendix A.

Note that the assumption that X_i is IR is necessary. To see this, consider an X_i that is not IR, for example one in which $\mu = 1000$ and $\sigma = 10$. Suppose also that the opponent's realization is $c = 500$. Observe that, in this case, $u_i(X_i, 500)$ is close to 0, since the probability that i wins is small. However, decreasing μ to 0 leads to an X_i' for which $u_i(X_i', 500) < 0$: player i nearly always wins, in which case her utility will be close to $1 - \mu^2 - \sigma^2 = -99$.

The assumption that Z is single-peaked (which follows from monotonicity and symmetry) is also necessary. To see this, consider a Z with two peaks, for example some infinitesimal perturbation of a Bernoulli distribution. The peaks are at 1 and -1 so that $\mu = 0$. Now, consider the case in which $X_i = Z + 1$, and suppose $c = 1$. Then player i wins whenever

[2] A distribution is defined to have convex tails if its probability density function is convex and decreasing away from the mean.

the realization is close to the lower peak, which happens with probability 1/2, and obtains utility close to 1 conditional on winning. Hence, $u_i(X_i, c) = 1/2$. However, under $X_i' = Z$ the utility conditional on winning is close to 0, since the realized values will be near the peaks 1 and -1.

One drawback of Theorem 3.1 is that it does not apply to normal distributions, since such distributions are not convex on $[-\infty, 0]$. However, as we are particularly interested in normal distributions, we have a version of the theorem that applies specifically to them:

Theorem 3.2. *Let Z be normal; let $X_i = \sigma Z + \mu$ and satisfy $\mu + \sigma \leq 1$; and let $X_i' = \sigma Z$. Then $u_i(X_i', c) \geq u_i(X_i, c)$ for any realization c of player j.*

4 Tradeoff Between Bias and Variance

Thus far, we have argued that simple observations that hold in the one-player case fail to extend to the competitive setting, and that reducing bias or variance may actually be harmful (holding all else fixed). We now turn to our main analysis, which considers the *tradeoff* between bias and variance: if there is only one player, she is indifferent between the two sources of error as long as their sum (or rather, the sum of bias squared and variance) is fixed. In this section, we show that in a competitive setting players are no longer indifferent, and, furthermore, that there is a clear preference for error incurred by variance over error incurred by bias.

We proceed as follows. In Section 4.1 we state our main result about the tradeoff between bias and variance in a two-player bias-variance game. We assume that $\mu^2 + \sigma^2 = 1$, which ensures the individual rationality constraint is satisfied.[3] We fix the random variable Z to be normal, and show that for arbitrary realizations a_j of the opponent player j, reducing bias μ_i while increasing variance to $\sigma_i^2 = 1 - \mu_i^2$ is always strictly beneficial for player i. That is, the minimal-bias strategy is ex post dominant. This result implies that the profile in which both players choose their minimal-bias strategy is the unique Nash equilibrium of the game. We also argue that our result extends to more than two players and to asymmetric strategy classes.

We then turn to the restrictive assumption in the analysis—namely, the normality assumption. In Section 4.2, we motivate the focus on normal distributions by arguing formally

[3] In fact, this assumption implies that the IR constraint binds in the one-player setting—that is, $u(X) = 1 - \mu^2 - \sigma^2 = 0$ for all X. This assumption is necessary for our analysis, but in Section 4.3 we show numerically that it is not necessary for our main result to hold.

that they are a natural distribution for error in machine learning contexts. Then, in Section 4.3, we consider distributions other than normal, as well as variations on the players' utility functions. We perform numerical analyses on these variations, and demonstrate that our insight for normal distributions is robust.

4.1 Preference for Lower Bias in Normal Distribution

In Theorem 3.2 we showed that reducing bias in normal distributions is beneficial if the variance is fixed. Here we give a "stronger" result: that as long as the *total* error is fixed, reducing bias (which increases variance) is beneficial. This is the main result of the paper.

Theorem 4.1. *Let Z be normal with mean 0 and variance 1, and let $X_i = \sigma Z + \mu$ and $X_i' = \tau Z + \nu$, where $\mu^2 + \sigma^2 = \nu^2 + \tau^2 = 1$. If $\nu^2 < \mu^2$ then $u_i(X_i', a) > u_i(X_i, a)$ for any realization $a > 0$ of player j.*

The proof of Theorem 4.1, given in Appendix B, is essentially a straightforward (albeit long and somewhat involved) calculation. The main idea is to write down the expected utility of a player as a function of her chosen bias and some realization of the opponent's strategy. We then calculate the derivative of this expected utility with respect to the bias, and show that it is negative. Thus, increasing bias (and so decreasing variance) leads to lower expected utility.

An immediate corollary of Theorem 4.1 is that the minimal-bias strategy is ex post dominant:

Corollary 4.2. *Let $\bar{\mu} \in [0, 1)$ be an exogenous lower bound on the choice of $|\mu|$. Let Z be normal with mean 0 and variance 1. The minimal-bias strategy Z^* with mean $\bar{\mu}$ and variance $\sqrt{1 - \bar{\mu}^2}$ is ex post strictly dominant within the strategy class $\mathcal{X} = \{\sigma Z + \mu \mid \mu^2 + \sigma^2 = 1 \wedge |\mu| \geq \bar{\mu}\}$.*

Many players and asymmetric strategies Because no-bias is ex post dominant—and so player i prefers no-bias for any realization of the opponent's error—the result of Corollary 4.2 immediately extends to more than two players. Consider the bias-variance game with any number of players, and in which a player's payoff is $1 - a^2$ if her realization a is lower than all others' realizations, and 0 otherwise. Then, within the same strategy class as in Corollary 4.2, no-bias is an ex post dominant strategy. To see this, observe that player i's utility can be maximized with the opponents' errors summarized by ex post error $c = \min_{j \neq i} a_j$. If agent i has lower error then i wins, otherwise i loses. Because Corollary 4.2 implies that no-bias is

dominant regardless of the other's realization, it is, in particular, optimal given realization c.

Another immediate implication of the result that minimal bias is optimal for player i regardless of the *realization* of player j's strategy is that j's strategy class could be different from i's—for example, it could include normal distributions with a different bias-variance tradeoff, or even distributions other than normal.

4.2 Motivation for Normal Distribution

One of the main assumptions we make in our analysis above is that the error of players' algorithms is normally distributed. This assumption is natural in the context of machine learning algorithms, and many commonly-used econometric and machine learning procedures with tuning parameters determining the bias-variance tradeoff have been demonstrated to produce predictions with asymptotically normal error distributions; some examples can be found in Ormoneit and Sen (2002), Hable (2012), and Wager and Athey (2018). To formally describe these results we need a definition:

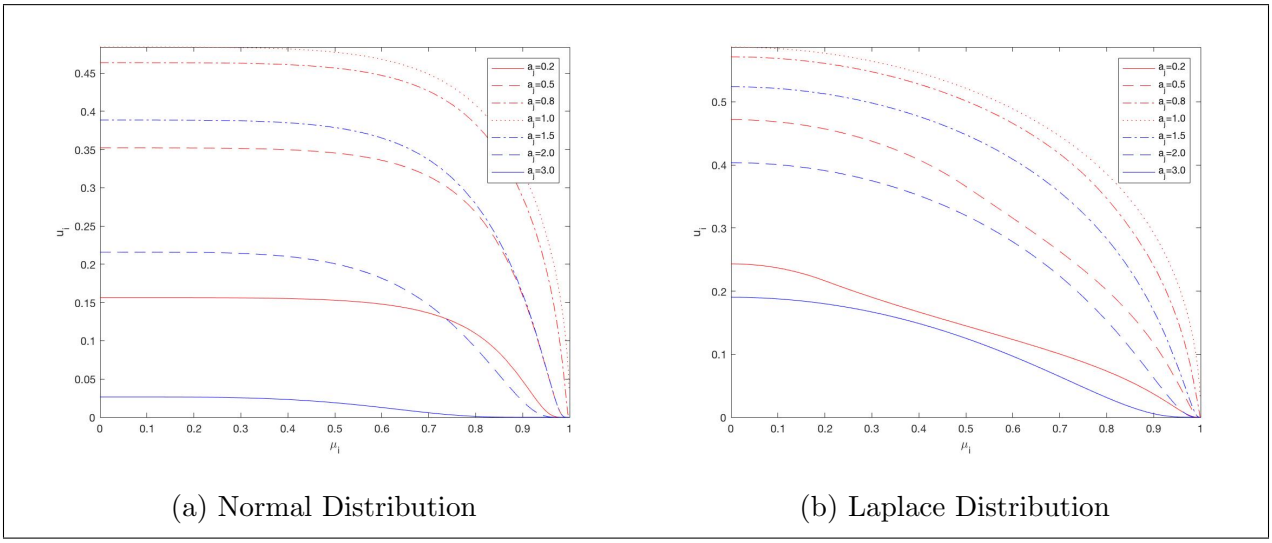
Definition 4.1. *Let $\{\mathcal{X}_1, \mathcal{X}_2, \ldots\}$ be an infinite sequence of random variables. The sequence is* asymptotically normal with asymptotic bias μ and asymptotic variance σ^2 *if $\sqrt{m}(\mathcal{X}_m - \mu)$ converges with m in distribution to $\mathcal{N}(0, \sigma^2)$, the normal distribution with variance σ^2.*[4]

Parameter estimates obtained from the ridge regression—which we employ for our empirical results in Section 5—are known to be asymptotically normal (with m corresponding to the size of the training sample). Ridge regression is equivalently linear regression with a quadratic regularizer. Features x lie in a p dimensional space $\mathbb{R}^p$ and hypotheses are given by weights $w \in \mathbb{R}^p$ as the linear weighted sums of the features, i.e., $w^T x$. The hypothesis selected by the ridge regression on the set of m training points $\{x_i, y_i\}_{i=1}^m$ is $\hat{f}(x) = \widehat{w}^T x$, where the weights $\widehat{w}$ minimize the the following quantity, the regularized empirical risk function with the quadratic regularizer and regularization parameter λ:

$$\widehat{R}(w) = \frac{1}{2m} \sum\nolimits_{i=1}^{m} (y_i - w^T x_i)^2 + \lambda \|w\|_2^2 \tag{1}$$

The regularization parameter λ is responsible for the bias-variance tradeoff in the prediction. The presence of the regularizer reduces the dependence of the vector of predicted

[4]Rearranging, we see that as the sequence converges so $\mathcal{X}_m$ resembles $\mathcal{N}(\mu, \sigma^2/m)$. Asymptotic normality is a consequence of the Central Limit Theorem where the variance converges at a rate of $1/m$ (the standard deviation converges at a rate of $1/\sqrt{m}$); thus, σ^2 is the variance normalized by m which converges to a constant.



(a) Normal Distribution (b) Laplace Distribution

Figure 2: Ex post utility curves of player i against different realization a_j from opponent player j.

weights $\widehat{w}$ on individual observations, making the corresponding predictions more "stable" as the regularization parameter λ increases. At the same time, the increase in λ makes the prediction less data-dependent and, therefore, more biased. In other words, while increasing λ increases the bias of prediction, it decreases the variance.

We summarize the bias-variance tradeoff associated with the ridge regression in the following theorem. While the asymptotic distribution of the prediction from the ridge regression is well-established, we formalize this result to provide a clear practical illustration of the nature of the bias-variance tradeoff in making predictions from data. The proof of this theorem is given in Appendix C.

Theorem 4.3. *Fix an infinite sequence of distinct and bounded features $\{x_1, x_2, \ldots\}$ with labels $f(x_i) = w_0^T x_i + e_i$, where the e_i are i.i.d., bounded, and mean zero. With regularization parameter λ and in the limit of m, the distribution of the prediction error of the ridge regression on the first m data points is asymptotically normal with finite asymptotic bias μ_λ that is a monotonically increasing function of λ and finite asymptotic variance σ_λ^2 that is a monotonically decreasing function of λ.*

4.3 Numerical Results for Other Distributions and Payoffs

In this subsection we illustrate the robustness of Theorem 4.1. Specifically, in the proof of Theorem 4.1, there are two assumptions that enable a clean closed form for ex post utility, and thus simplify our argument: (a) the shape (i.e., density function) of the distributions is

normal; (b) the utility function is $u_i(a_i, a_j) = (1 - a_i^2) \cdot \mathbf{1}\{a_i < a_j\}$, with the assumption that $\mu^2 + \sigma^2 = 1$. The fact that both the benefit from winning and the total error are 1 simplifies the proof of Theorem 4.1.

In this section we vary both the distributions and utility function, and numerically evaluate the players' utilities. The following numerical calculation results suggest that our insight on the preference of variance over bias holds generally in many cases beyond our theoretical assumptions (a) and (b).

Other distributions. We first numerically calculate the ex post utility curves against arbitrary realizations from the opponent player, as well as the expected utility curves against arbitrary strategies of the opponent, on various distributions. As shown in Theorem 4.1, with the normal distribution, the ex post utility curve is always decreasing, for all opponents' realizations. Numerical evaluations with the *Laplace* distribution indicate that its ex post utility curve is also decreasing for all opponents' realizations; see Figure 2, where we plot the ex post utility curves $u_i(\mu_i, a_j)$ holding realization a_j fixed for both normal and Laplace distributions. The x-axis of each figure is player i's bias μ_i, and the y-axis is the ex post utility $u_i(\sqrt{1-\mu_i^2}Z + \mu_i, a_j)$. Another interesting observation from this numerical result is that with normal distributions, the ex post utility is almost flat for $\mu_i \leq 0.5$—i.e., while $\mu_i = 0$ is a dominant strategy, picking any $\mu_i \in [0, 0.5]$ is a "pretty good" strategy.

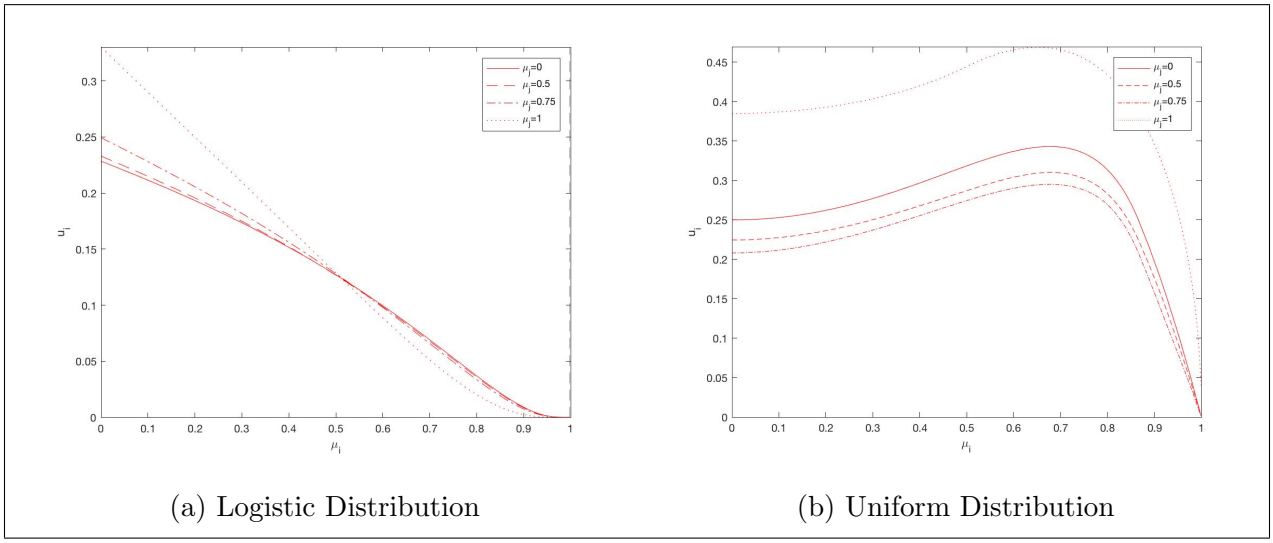
For the *logistic* distribution, the ex post utility curve is no longer decreasing; however, the expected utility curve is decreasing against arbitrary strategies of the opponent. Thus, the no-bias distribution $X_i = Z$ is a dominant (although not ex post dominant) strategy.

Finally, for *uniform* distributions, monotonicity does not hold on either the ex post utility curve or the expected utility curve. In fact, there exists a pure Nash equilibrium with non-zero bias. See Figure 3 for utility plots under logistic and uniform distributions.

These plots and calculations indicate that the preference for variance over bias is robust—and holds for Laplace and logistic distributions—but not universal—as demonstrated by the uniform distribution

Other utility functions. Another assumption that is important for our theoretical analysis is the form of the utility function. More generally, suppose players' utility functions $u_i(a_i, a_j) = (R - a_i^2) \cdot \mathbf{1}\{a_i < a_j\}$ for arbitrary reward $R > 0$, with normal distribution Z. Observe that for very large R, the resulting game is close to a constant-sum game, as the error terms are nearly irrelevant, and that the probability of negative payoffs is negligible.

Our main theoretical result considered the case $R = 1$, but for general reward $R \neq$



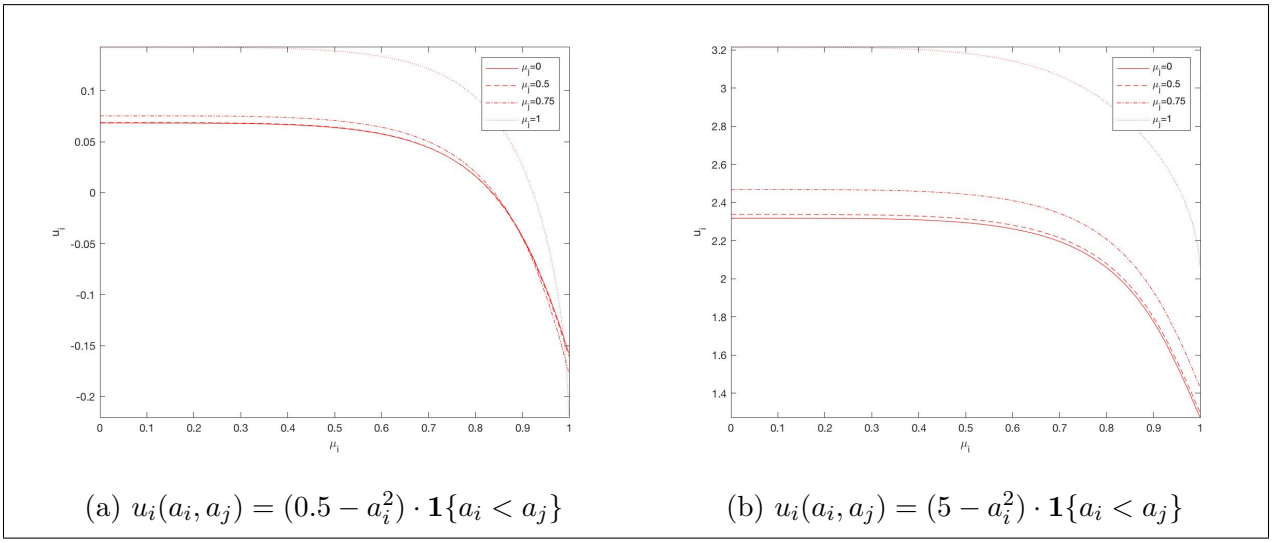
(a) Logistic Distribution (b) Uniform Distribution

Figure 3: Expected utility curves of player i against different strategy μ_j from opponent player j.

1, in which the individual rationality constraint either does not bind or is violated, the monotonicity of ex post utility may not hold. However, numerical calculations demonstrate that the expected utility curve is still decreasing against arbitrary strategies of the opponent, and so no-bias remains a dominant strategy—see Figure 4.

5 Empirical Results

Thus far, we have shown that our insight on the preference of variance over bias provably holds in a particular, restricted setting, and that numerically it appears to hold in broader settings as well. In this section we test our insight in a real dataset. We consider a widely-used benchmark dataset, chosen mostly for convenience and because it is a standard dataset often used to test new ideas in machine learning. In particular, we utilize the California housing prices data from the 1990 Census, a dataset first utilized by Pace and Barry (1997) and included in the Python Scikit-learn library. This dataset is popular amongst machine learning practitioners for testing out new ideas because it is not too small but also not too large. It includes 20,640 observation on 9 featrures, such as number of rooms, median income, etc. From this data we set up a regression problem of predicting the order of magnitude of the median house prices (i.e., its logarithm) from the other features. For features where orders of magnitude may be more relevant than absolute magnitude, we include both the feature and its logarithm.



(a) $u_i(a_i, a_j) = (0.5 - a_i^2) \cdot \mathbf{1}\{a_i < a_j\}$ (b) $u_i(a_i, a_j) = (5 - a_i^2) \cdot \mathbf{1}\{a_i < a_j\}$

Figure 4: Expected utility curves of player i against different strategy μ_j from opponent player j with normal distribution. Rewards are $R = 0.5$ and $R = 5$, respectively.

We analyze a game between two players, each of whom gets roughly half the data and employs a ridge regression in order to predict median housing prices. In housing data various features of houses tend to be highly correlated, which makes the design matrix (covariance matrix of the feature vector) nearly non-invertible. This, in turn, makes the estimated feature weights in the linear regression unstable, largely varying from sample to sample. A common tool used to stabilize the estimated feature weights is linear regression with regularization, namely, ridge regression. . Running a ridge regression involves setting a regularization parameter λ, where $\lambda = 0$ is the standard ordinary least squares (OLS) regression, and increasing λ leads to a model with greater bias but lower variance of the resulting predictions. Thus, we will let the players play the game for various values of λ. Further discussion of ridge regression is given in Section 4.2.

The design of the game is as follows:

1. The dataset is uniformly partitioned with 45% as a training set for player 1, 45% as a training set for player 2, and the remaining 10% as the test set.

2. Players choose regularization parameters λ_1 and λ_2, respectively, and perform ridge regression on their training sets.

3. The expected payoffs of players are given by the bias-variance game (Definition 2.2) evaluated on the test set. On each point in the test set, the player whose prediction is closest to the true label wins and obtains payoff of one minus the squared error of the

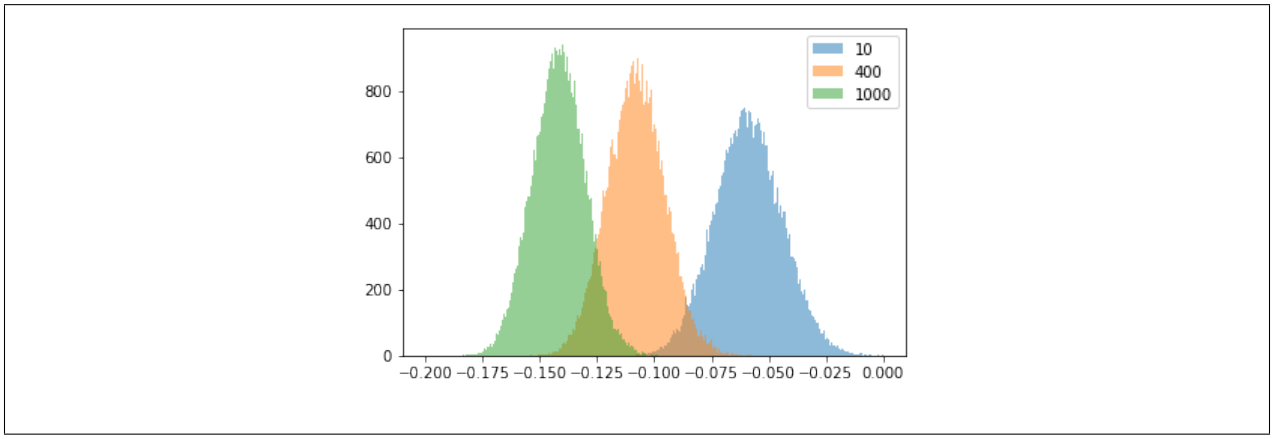


Figure 5: Distribution of predictions for three values of λ.

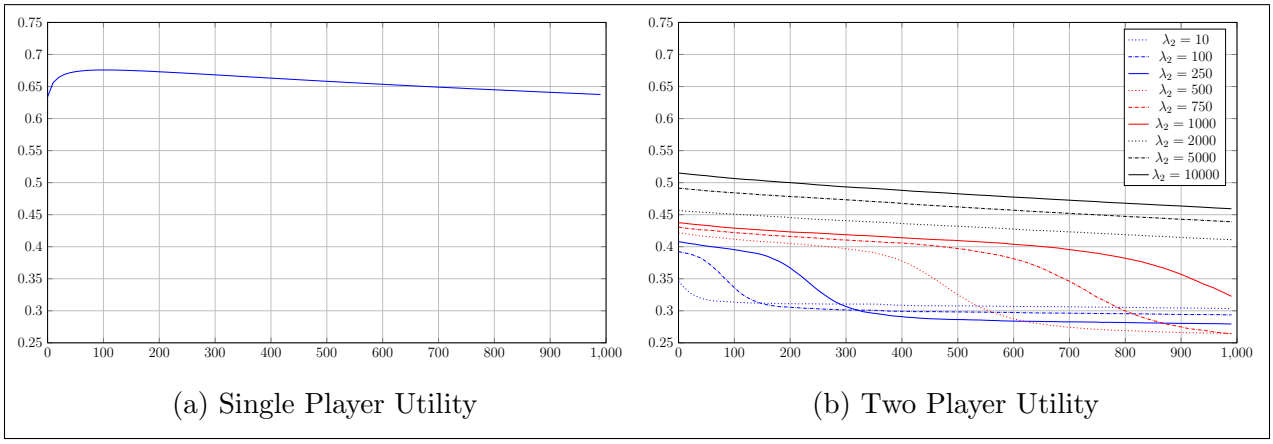
prediction; the other player's payoff is zero. Each player's payoff in the game is the average payoff over the points in the test set.

We used Monte Carlo simulations to approximate expected utilities of the players. We repeat steps 1-3 above by independently drawing training and validation samples 100 times and computing payoffs that result from a given pair of choices of regularization parameters (λ_1, λ_2).

Figure 5 plots the distribution, over random choices of the training data, of the error in prediction on a particular point in the test data for three values of regularization parameter λ. Notice that the distributions appear roughly normal, and that the predictions of the lower λ value have higher variance and lower bias.

We contrast the optimal choice of regularization parameter in the single-player setting with the two-player game. In Figure 6a Player 1's utility in the single-player game is depicted as a function of the player's regularization parameter, ranging from $\lambda_1 = 0$ to $\lambda_1 = 1000$. The optimal parameter choice is $\lambda_1 \approx 100$. In Figure 6b, Player 1's utility in the two-player game is plotted as a function of λ_1 for fixed values of Player 2's regularization parameter λ_2. For all choices of λ_2, Player 1's utility is optimized by selecting $\lambda_1 = 0$. Thus, the conclusions of our theoretical study are corroborated by this empirical ridge regression game.

While we have not done an exhaustive study of variations of this game on many different datasets, the numerous ones that we have considered provide the same qualitative conclusion. Specifically, the optimal single-player regularization parameter in a ridge regression is generally non-zero, while, as long as the benefit of winning is sufficiently large, the two-player best response is to lower the regularization parameter to zero. We view these results as affirming the validity of our insight on the preference of variance over bias in competitive



(a) Single Player Utility

(b) Two Player Utility

Figure 6: The player's utilities under ridge regression in the single-player and two-player environment. In both environments Player 1's utility is plotted as a function of regularizer parameter λ_1. In the two player figure, Player 1's utility is shown for a range of values of Player 2's regularizer parameter λ_2. The single player utility is optimized at $\lambda_1 \approx 100 > 0$; the two-player utility is optimized at $\lambda_1 = 0$ for all λ_2.

settings beyond our theoretical and numerical analyses.

6 Conclusions

In this paper we studied competing machine learning algorithms by abstracting the problem to a distribution-selection game in which bias and variance can be traded off. While outcomes of real learning algorithms can be complex, our bias-variance game is amenable to theoretical analysis, and we formally prove that for normal distributions reducing bias at the expense of variance is an ex post best response. Thus, the no-bias action is ex post dominant.

We next showed that the ridge regression algorithm has normally-distributed error, and so the distribution-selection game with the normal distribution is a reasonable model of real prediction algorithms. We also considered the empirical game on a benchmark data set using ridge regression, where the same qualitative conclusions from our theoretical analysis were demonstrated.

Many aspects of machine learning problems change significantly in competitive environments. For example, in a different context but a similar vein, Mansour et al. (2018) consider the classical bandit model of online learning and study the effect competition between the algorithms on the exploration vs. exploitation tradeoff. These authors show that the presence of competition may lead to the strategic choice of algorithms that do not explore as

much as they would absent competition, and may thus be worse learners. More generally, we believe that there are many more unresolved issues in the intersection of machine learning and competition, and suggest this general area as a fruitful and important one for future study.

References

Anderson, C. (2006). *The long tail: Why the future of business is selling less of more*. Hachette Books.

Ben-Porat, O. and Tennenholtz, M. (2017). Best response regression. In *Advances in Neural Information Processing Systems*, pages 1499–1508.

Ben-Porat, O. and Tennenholtz, M. (2019). Regression equilibrium. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 173–191.

Eliaz, K. and Spiegler, R. (2019). The model selection curse. *American Economic Review: Insights*, 1(2):127–40.

Gaudin, S. (2016). At Stitch Fix, data scientists and A.I. become personal stylists. `https://www.idginsiderpro.com/article/3067264/at-`σ_2`stitch-`σ_2`fix-`σ_2`data-`σ_2`scientists-`σ_2`and-`σ_2`ai-`σ_2`become-`σ_2`personal-`σ_2`stylists.html`. Accessed 2019-12-11.

Hable, R. (2012). Asymptotic normality of support vector machine variants and other regularized kernel methods. *Journal of Multivariate Analysis*, 106:92–117.

Immorlica, N., Kalai, A. T., Lucier, B., Moitra, A., Postlewaite, A., and Tennenholtz, M. (2011). Dueling algorithms. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 215–224. ACM.

Liang, A. (2019). Games of incomplete information played by statisticians. *arXiv preprint arXiv:1910.07018*.

Mansour, Y., Slivkins, A., and Wu, Z. S. (2018). Competing bandits: Learning under competition. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Olea, J. L. M., Ortoleva, P., Pai, M. M., and Prat, A. (2019). Competing models. *arXiv preprint arXiv:1907.03809*.

Ormoneit, D. and Sen, Ś. (2002). Kernel-based reinforcement learning. *Machine learning*, 49(2-3):161–178.

Pace, R. K. and Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297.

Sinha, J. I., Foscht, T., and Fung, T. T. (2016). How data analytics and AI are driving the subscription e-commerce phenomenon. *MIT Sloan Management Review*.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

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

Appendix

A Proofs from Section 3

Theorem 3.1. *Let Z be monotone increasing, convex on $[-\infty, 0]$, and symmetric around 0. Let $X_i = \sigma Z + \mu$ be IR (so as to satisfy $\mu^2 + \sigma^2 \leq 1$) and $X_i' = \sigma Z$. Then $u_i(X_i', c) \geq u_i(X_i, c)$ for any realization c of player j.*

Proof. The proof consists of several cases.

1) $\mu \geq c$: Since $\mu^2 + \sigma^2 \leq 1$, it must be the case that $c \leq 1$. Thus, for any realization in which player i gets non-zero utility, his utility is nonnegative under both X_i and X_i'.

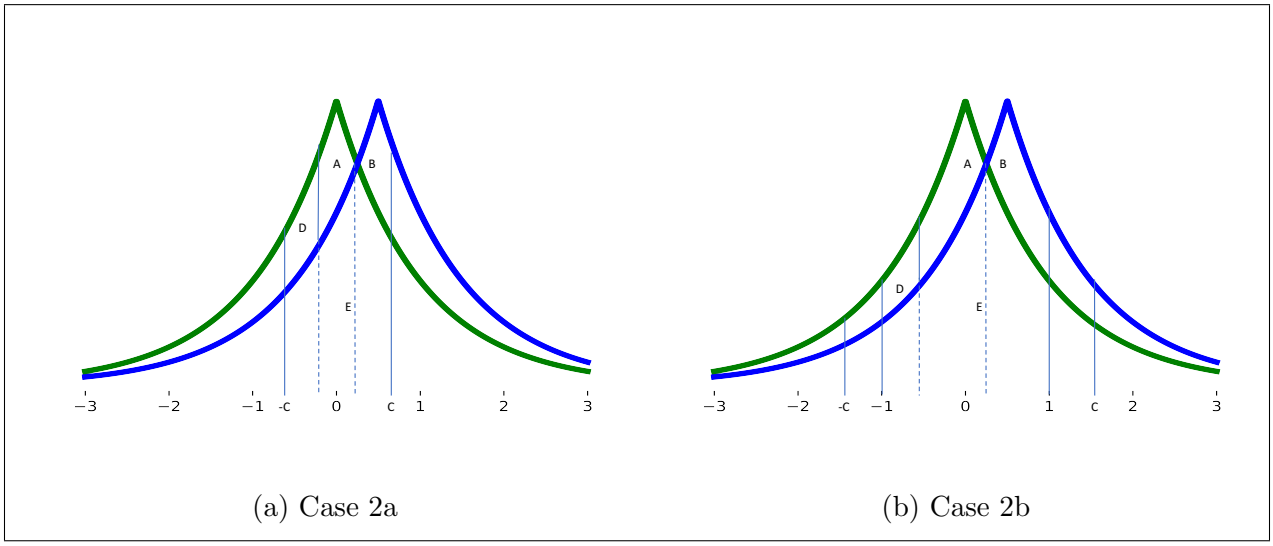
   Consider first the distribution $X_i'' = \sigma Z + c$. Observe that, by monotonicity, for each point $x \in [-c, c]$, the pdf at x under X_i'' is higher than under X_i. Since all such realizations lead to positive utility, $u(X_i'') \geq u(X_i)$.

   Next, consider the comparison between X_i' and X_i''. On the interval $[0, c]$ the distribution X_i' is an inversion of X_i'' with higher probability closer to the origin, and so on this sub-interval X' leads to higher utility. On the interval $[-c, 0]$ the pdf of X_i' dominates that of X_i'', and so also here X_i' leads to higher utility. Thus, $u(X_i') \geq u(X_i'')$, and so $u(X_i') \geq u(X_i)$.

2a) $\mu < c \leq 1$: Consider Figure 7a, in which X' is the green pdf and X is the blue pdf. Area E (from -1 to 1, and below both curves) leads to the same utility for both distributions. Area A (under X') leads to higher utility than area B (under X). And finally, area D (under X') leads to strictly positive utility. Thus, overall, $u(X') \geq u(X)$.

2b) $c > 1$: Consider Figure 7b, in which X' is the green pdf and X is the blue pdf. Area E (from $-c$ to c, and below both curves) leads to the same utility for both distributions. Area A (under X') leads to higher utility than area B (under X). Area D (under X') leads to strictly positive utility.

   It remains to show that the losses under X' due to realizations in $[-c, -1) \cup (1, c]$ are smaller than the losses on the same intervals due to X. To this end, we will consider points $x \in (1, c]$, and show that the sum of the pdfs of X at x and $-x$ is larger than the sum of the pdfs of X' at those same points. Let g be the pdf of X'. Then the sum of the pdfs of X' at points x and $-x$ is $g(x) + g(-x) = 2g(x)$. The sum of the pdfs of X at the



(a) Case 2a (b) Case 2b

Figure 7: Case 2 of Theorem 2.1.

points x and and $-x$ is $g(x+\mu)+g(x-\mu)$. By convexity, $g(x+\mu)+g(x-\mu) \geq 2g(x)$, completing the claim.

□

Theorem 3.2. *Let Z be normal; let $X_i = \sigma Z + \mu$ and satisfy $\mu + \sigma \leq 1$; and let $X_i' = \sigma Z$. Then $u_i(X_i', c) \geq u_i(X_i, c)$ for any realization c of player j.*

Proof. The proof is nearly identical to that of Theorem 3.1, except for case 2b: the upper tail of the normal distribution is convex only from $\mu + \sigma$ onward. So this case must be handled differently.

However, to complete the proof, we actually only need that the pdf g of X' be convex from the point $1-\mu$ and higher. Since $\mu+\sigma \leq 1$ it holds that $1-\mu \geq \sigma$, and so g is convex on $[c-\mu, c+\mu]$ whenever $c \geq 1$, completing the proof. □

B Proof of Theorem 4.1

Theorem 4.1. *Let Z be normal. Let $X_i = \sigma Z + \mu$ and $X_i' = \tau Z + \nu$, where $\mu^2 + \sigma^2 = \nu^2 + \tau^2 = 1$. If $\nu^2 < \mu^2$ then $u_i(X_i', a) > u_i(X_i, a)$ for any realization $a > 0$ of player j.*

Proof. Since players are symmetric, we drop all subscripts in the discussion below. We compute the expected utility $u(X, a)$ of player i when he plays action $X = \sigma Z + \mu$ against the realization a of player j. Without loss of generality, we assume $a \geq 0$. First we characterize the closed form of $\mathbf{E}[u(X, a)]$ for all $\mu \in [0, 1)$; then we argue that $\lim_{\mu \to 1^-} \mathbf{E}[u(X, a)] = 0 =$

$\mathbf{E}[u(X', a)]$, where $X' = 0 \cdot Z + 1$; and finally we show that $\frac{\partial \mathbf{E}[u(X,a)]}{\partial \mu} \leq 0$ for all $\mu \in [0, 1)$, which completes the proof.

We first focus on $\mu \in [0, 1)$. Observe that

$$\mathbf{E}[u(X, a)] = \mathbf{E}\big[(1 - X^2) \cdot \mathbf{1}\{|X| < a\}\big] = \mathbf{Pr}[|X| < a] - \mathbf{E}\big[X^2 \cdot \mathbf{1}\{|X| < a\}\big].$$

We can now evaluate

$$\begin{aligned}\mathbf{Pr}[|X| < a] &= \int_{-a}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{-a-\mu}{\sigma}\right),\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of Z, i.e., the standard normal distribution. Now,

$$\begin{aligned}\mathbf{E}\big[X^2 \cdot \mathbf{1}\{|X| < a\}\big] &= \int_{-a}^{a} x^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx \\ &= -\frac{\mu\sigma}{\sqrt{2\pi}}\left(\exp\left(\frac{-(a-\mu)^2}{2\sigma^2}\right) - \exp\left(\frac{-(a+\mu)^2}{2\sigma^2}\right)\right) \\ &\quad - \frac{a\sigma}{\sqrt{2\pi}}\left(\exp\left(\frac{-(a-\mu)^2}{2\sigma^2}\right) + \exp\left(\frac{-(a+\mu)^2}{2\sigma^2}\right)\right) \\ &\quad + (\sigma^2 + \mu^2)\left(\Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{-a-\mu}{\sigma}\right)\right).\end{aligned}$$

Because of the constraint that $\mu^2 + \sigma^2 = 1$, we can eliminate some terms:

$$\begin{aligned}\mathbf{E}[u(X, a)] &= \frac{\mu\sqrt{1-\mu^2}}{\sqrt{2\pi}}\left(\exp\left(\frac{-(a-\mu)^2}{2-2\mu^2}\right) - \exp\left(\frac{-(a+\mu)^2}{2-2\mu^2}\right)\right) \\ &\quad + \frac{a\sqrt{1-\mu^2}}{\sqrt{2\pi}}\left(\exp\left(\frac{-(a-\mu)^2}{2-2\mu^2}\right) + \exp\left(\frac{-(a+\mu)^2}{2-2\mu^2}\right)\right).\end{aligned}$$

Next we consider $\mathbf{E}[u(X, a)]$ for $\mu = 1$. Note that in this case, agent i's realization is 1 deterministically, which implies that her utility conditioning on winning is zero. Her utility conditioning on losing is also zero by definition. Thus, $\mathbf{E}[u(X, a)] = 0$ for $\mu = 1$. To see that $\lim_{\mu \to 1^-} \mathbf{E}[u(X, a)] = 0$, note that $0 < \exp\left(\frac{-(a-\mu)^2}{2-2\mu^2}\right) \leq 1$ and $0 < \exp\left(\frac{-(a+\mu)^2}{2-2\mu^2}\right) \leq 1$. Thus,

$$\lim_{\mu \to 1^-} \mathbf{E}[u(X, a)] \leq \lim_{\mu \to 1^-} \left(\frac{\mu\sqrt{1-\mu^2}}{\sqrt{2\pi}} + 2\frac{a\sqrt{1-\mu^2}}{\sqrt{2\pi}}\right) = 0$$

and

$$\lim_{\mu \to 1^-} \mathbf{E}[u(X, a)] \geq \lim_{\mu \to 1^-} -\frac{\mu\sqrt{1-\mu^2}}{\sqrt{2\pi}} = 0$$

Invoking the Squeeze Theorem yields $\lim_{\mu \to 1^-} \mathbf{E}[u(X,a)] = 0$. Finally, taking the derivative of $\mathbf{E}[u(X,a)]$ with respect to μ for $\mu \in [0,1)$ yields

$$\begin{aligned}
&\frac{\partial \mathbf{E}[u(X,a)]}{\partial \mu} \\
&= \left[\sqrt{1-\mu^2} - (a+\mu)\frac{\mu}{\sqrt{1-\mu^2}} + \frac{(a^2-\mu^2)(1-a\mu)\sqrt{1-\mu^2}}{(1-\mu^2)^2} \right] \exp\left(-\frac{1}{2}\frac{(a-\mu)^2}{1-\mu^2} \right) \\
&\quad - \left[\sqrt{1-\mu^2} + (a-\mu)\frac{\mu}{\sqrt{1-\mu^2}} + \frac{(a^2-\mu^2)(1+a\mu)\sqrt{1-\mu^2}}{(1-\mu^2)^2} \right] \exp\left(-\frac{1}{2}\frac{(a+\mu)^2}{1-\mu^2} \right).
\end{aligned}$$

To prove Theorem 4.1, it is sufficient to show that this derivative is (strictly) negative for all $\mu \in (0,1)$ and $a \geq 0$. This can be simplified to

$$\begin{aligned}
&[(2\mu^2 - a^2 - 1)(1-a\mu) + 2\mu^2(1-\mu^2)] \exp\left(\frac{a\mu}{1-\mu^2} \right) \\
&- [(2\mu^2 - a^2 - 1)(1+a\mu) + 2\mu^2(1-\mu^2)] \exp\left(\frac{-a\mu}{1-\mu^2} \right) > 0.
\end{aligned} \tag{2}$$

The remaining part of the proof, showing that inequality (2) holds, follows from a long and algebraic calculation that we formalize as Lemma B.1. □

We now show that the derivative of the ex post utility with respect to μ against realization a of the opponent is strictly negative for all $\mu \in (0,1)$ and $a > 0$.

Lemma B.1. *Inequality* (2) *holds for all* $\mu \in (0,1)$ *and* $a > 0$.

Notice that there may be two regimes: (a) the player always gains positive payoff, i.e., $a < 1$; (b) the player sometimes suffers non-positive payoff, i.e., $a \geq 1$. We break Lemma B.1 into these two regimes and show them separately.

Lemma B.2. *Inequality* (2) *holds for all* $\mu \in (0,1)$ *and* $a \in (0,1)$.

Proof. We start with inequality (2), copied here:

$$\begin{aligned}
&[(2\mu^2 - a^2 - 1)(1-a\mu) + 2\mu^2(1-\mu^2)] \exp\left(\frac{a\mu}{1-\mu^2} \right) \\
>\ &[(2\mu^2 - a^2 - 1)(1+a\mu) + 2\mu^2(1-\mu^2)] \exp\left(\frac{-a\mu}{1-\mu^2} \right).
\end{aligned}$$

The proof is a two-case analysis, working backwards from inequality (2) to show that it holds from lemma domain and case assumptions in both cases. Some steps are exact algebraic manipulations ("if and only if" $\Leftrightarrow$ or $\Updownarrow$) and some steps are restrictions to stronger

requirements ("is implied by" $\Leftarrow$ or $\Uparrow$). We use the arrows for visual simplicity to indicate the type of each step. Some steps are additionally marked by $*$ to identify them for use in the proof of a later result.

Each step includes an explanation. Consider the following two cases based on the sign of $(2\mu^2 - a^2 - 1)$.

Case 1: $(2\mu^2 - a^2 - 1) > 0$. To start, multiplying both sides of equation (2) by $\frac{1}{2\mu^2 - a^2 - 1} \cdot \exp\left(\frac{1}{1-\mu^2}\right)$, we get inequality (2) $\Leftrightarrow$

$$\left[(1 - a\mu) + \frac{(2\mu^2(1-\mu^2))}{(2\mu^2 - a^2 - 1)}\right] \exp\left(\frac{1 + a\mu}{1 - \mu^2}\right)$$
$$> \left[(1 + a\mu) + \frac{(2\mu^2(1-\mu^2))}{(2\mu^2 - a^2 - 1)}\right] \exp\left(\frac{1 - a\mu}{1 - \mu^2}\right)$$

($\Uparrow$) Now it is clear that we can drop the term $(2\mu^2 - a^2 - 1)$, because we have $0 < (2\mu^2 - a^2 - 1) < 1$ from $\mu^2 < 1$. The left-hand side decreases by at least as much as the right-hand side. So we get

$$\left[(1 - a\mu) + (2\mu^2(1-\mu^2))\right] \exp\left(\frac{1 + a\mu}{1 - \mu^2}\right)$$
$$> \left[(1 + a\mu) + (2\mu^2(1-\mu^2))\right] \exp\left(\frac{1 - a\mu}{1 - \mu^2}\right)$$

($\Updownarrow$) Replacing the exponential functions with their respective Taylor series, we get

$$\left[(1 - a\mu) + (2\mu^2(1-\mu^2))\right] \left[\sum_{k=0}^{\infty} \frac{((1 + a\mu)/(1-\mu^2))^k}{k!}\right]$$
$$> \left[(1 + a\mu) + (2\mu^2(1-\mu^2))\right] \left[\sum_{k=0}^{\infty} \frac{((1 - a\mu)/(1-\mu^2))^k}{k!}\right]$$

($\Updownarrow$) Pulling out the first two terms of the series, we get

$$\left[(1 - a\mu) + (2\mu^2(1-\mu^2))\right] \left[1 + \frac{1 + a\mu}{1 - \mu^2} + \sum_{k=2}^{\infty} \frac{((1 + a\mu)/(1-\mu^2))^k}{k!}\right]$$
$$> \left[(1 + a\mu) + (2\mu^2(1-\mu^2))\right] \left[1 + \frac{1 - a\mu}{1 - \mu^2} + \sum_{k=2}^{\infty} \frac{((1 - a\mu)/(1-\mu^2))^k}{k!}\right]$$

($\Leftarrow$) Separate the previous line into two inequalities described by **Condition 1a** and **Condition 1b**; if both are true then the combined inequality is true.

Condition 1a:

$$\begin{aligned}&\left[(1-a\mu)+(2\mu^2(1-\mu^2))\right]\cdot\left[1+\frac{1+a\mu}{1-\mu^2}\right]\\ \geq&\left[(1+a\mu)+(2\mu^2(1-\mu^2))\right]\cdot\left[1+\frac{1-a\mu}{1-\mu^2}\right]\end{aligned}$$

($\Updownarrow$) Multiplying through by $(1-\mu^2)$, we get

$$\begin{aligned}&\left[(1-a\mu)+(2\mu^2(1-\mu^2))\right]\cdot\left[1-\mu^2+1+a\mu\right]\\ \geq&\left[(1+a\mu)+(2\mu^2(1-\mu^2))\right]\cdot\left[1-\mu^2+1-a\mu\right]\end{aligned}$$

($\Updownarrow$) Splitting out the terms in the first bracket and canceling $(1-a\mu)(1+a\mu)$, we get

$$\begin{aligned}&(1-a\mu)\left(1-\mu^2\right)+(2\mu^2(1-\mu^2))\left[1-\mu^2+1+a\mu\right]\\ \geq&(1+a\mu)\left(1-\mu^2\right)+(2\mu^2(1-\mu^2))\left[1-\mu^2+1-a\mu\right]\end{aligned}$$

($\Updownarrow$) Further canceling additive constants from both sides, we get

$$\begin{aligned}&(-a\mu)\left(1-\mu^2\right)+(2\mu^2(1-\mu^2))\left[+a\mu\right]\\ \geq&(+a\mu)\left(1-\mu^2\right)+(2\mu^2(1-\mu^2))\left[-a\mu\right]\end{aligned}$$

($\Updownarrow$) Dividing out $a\mu(1-\mu^2)$ and grouping all terms on one side, we get

$2(2\mu^2-1)\geq 0$ ✓ which is finally true, directly from the assumption of Case 1.

Condition 1b:

$$\begin{aligned}&\left[(1-a\mu)+(2\mu^2(1-\mu^2))\right]\left[\sum_{k=2}^{\infty}\frac{((1+a\mu)/(1-\mu^2))^k}{k!}\right]\\ >&\left[(1+a\mu)+(2\mu^2(1-\mu^2))\right]\left[\sum_{k=2}^{\infty}\frac{((1-a\mu)/(1-\mu^2))^k}{k!}\right]\end{aligned}$$

($\Uparrow$) Dropping the $(2\mu^2(1-\mu^2))$ terms – by the left-hand side sum terms dominating for every k – we get

$$\begin{aligned}&\left[(1-a\mu)\right]\left[\sum_{k=2}^{\infty}\frac{((1+a\mu)/(1-\mu^2))^k}{k!}\right]\\ >&\left[(1+a\mu)\right]\left[\sum_{k=2}^{\infty}\frac{((1-a\mu)/(1-\mu^2))^k}{k!}\right]\end{aligned}$$

($\Updownarrow$) Noting $k \geq 2$ and canceling a factor of $(1 - a\mu)(1 + a\mu)$, we get

$$\left[\sum_{k=2}^{\infty} \frac{(1+a\mu)^{k-1}}{(1-\mu^2)^k \cdot k!}\right] > \left[\sum_{k=2}^{\infty} \frac{(1-a\mu)^{k-1}}{(1-\mu^2)^k \cdot k!}\right] \quad \checkmark \text{ which is true for every } k \text{ from } a\mu > 0.$$

Case 2: $(2\mu^2 - a^2 - 1) \leq 0$.

Note that we have $a\mu < 1$ because the lemma's domain has $a < 1$ and $\mu < 1$. Starting anew for **Case 2** from inequality (2), replacing the exponential functions with their respective Taylor Series, we get inequality (2) $\Leftrightarrow$

$$\begin{aligned}
&\left[(2\mu^2 - a^2 - 1) \cdot (1 - a\mu) + (2\mu^2(1 - \mu^2))\right] \left[\sum_{k=0}^{\infty} \frac{((a\mu)/(1-\mu^2))^k}{k!}\right] \\
&> \left[(2\mu^2 - a^2 - 1) \cdot (1 + a\mu) + (2\mu^2(1 - \mu^2))\right] \left[\sum_{k=0}^{\infty} \frac{((-a\mu)/(1-\mu^2))^k}{k!}\right]
\end{aligned}$$

($\Updownarrow$) Pulling out the first two terms of the series, we get

$$\begin{aligned}
&\left[(2\mu^2 - a^2 - 1) \cdot (1 - a\mu) + (2\mu^2(1 - \mu^2))\right] \left[1 + \frac{a\mu}{1-\mu^2} + \sum_{k=2}^{\infty} \frac{((a\mu)/(1-\mu^2))^k}{k!}\right] \\
&> \left[(2\mu^2 - a^2 - 1) \cdot (1 + a\mu) + (2\mu^2(1 - \mu^2))\right] \left[1 - \frac{a\mu}{1-\mu^2} + \sum_{k=2}^{\infty} \frac{((-a\mu)/(1-\mu^2))^k}{k!}\right]
\end{aligned}$$

($\Leftarrow$) Separate the previous line into two inequalities described by **Condition 2a** and **Condition 2b**; if both are true then the combined inequality is true.

Condition 2a:

$$\begin{aligned}
&\left[(2\mu^2 - a^2 - 1) \cdot (1 - a\mu) + (2\mu^2(1 - \mu^2))\right] \left[1 + \frac{a\mu}{1-\mu^2}\right] \\
&> \left[(2\mu^2 - a^2 - 1) \cdot (1 + a\mu) + (2\mu^2(1 - \mu^2))\right] \left[1 - \frac{a\mu}{1-\mu^2}\right]
\end{aligned}$$

($\Updownarrow$) Multiplying through by $(1 - \mu^2)$, we get

$$\begin{aligned}
&\left[(2\mu^2 - a^2 - 1) \cdot (1 - a\mu) + (2\mu^2(1 - \mu^2))\right] \left[1 - \mu^2 + a\mu\right] \\
&> \left[(2\mu^2 - a^2 - 1) \cdot (1 + a\mu) + (2\mu^2(1 - \mu^2))\right] \left[1 - \mu^2 - a\mu\right]
\end{aligned}$$

($\Updownarrow$) Splitting out the terms in the first bracket and canceling the resulting (additively) matching terms, we get

$$\begin{aligned}&\left[(2\mu^2 - a^2 - 1)\cdot(1 - a\mu)\cdot(-\mu^2) + (2\mu^2(1-\mu^2))\cdot(+a\mu)\right]\\ >&\left[(2\mu^2 - a^2 - 1)\cdot(1 + a\mu)\cdot(-\mu^2) + (2\mu^2(1-\mu^2))\cdot(-a\mu)\right]\end{aligned}$$

($\Updownarrow$) Further canceling additively and then moving the -1 factor of $-\mu^2$, we get

$$\begin{aligned}&\left[(1 + a^2 - 2\mu^2)\cdot(-a\mu)\cdot(\mu^2) + (2\mu^2(1-\mu^2))\cdot(+a\mu)\right]\\ >&\left[(1 + a^2 - 2\mu^2)\cdot(+a\mu)\cdot(\mu^2) + (2\mu^2(1-\mu^2))\cdot(-a\mu)\right]\end{aligned}$$

($\Updownarrow$) Combining like-terms on each side and dividing through by 2, we get

$$(2\mu^2(1-\mu^2))\cdot(+a\mu) > (1 + a^2 - 2\mu^2)\cdot(+a\mu)\cdot(\mu^2)$$

($\Updownarrow$) Dividing by $a\mu^3$ and re-organizing, we get

$$(1-\mu^2) + (1-\mu^2) > (1-\mu^2) + (a^2 - \mu^2) \quad \checkmark \text{ which is true because the domain has } a < 1$$

Condition 2b:

$$\begin{aligned}&\left[(2\mu^2 - a^2 - 1)\cdot(1 - a\mu) + (2\mu^2(1-\mu^2))\right]\left[\sum_{k=2}^{\infty}\frac{((a\mu)/(1-\mu^2))^k}{k!}\right]\\ \geq&\left[(2\mu^2 - a^2 - 1)\cdot(1 + a\mu) + (2\mu^2(1-\mu^2))\right]\left[\sum_{k=2}^{\infty}\frac{((-a\mu)/(1-\mu^2))^k}{k!}\right]\end{aligned}$$

($\Leftarrow$) To prove that the inequality of the previous line holds, it is sufficient to show that the next inequality holds for pairs of consecutive terms within its sums $\forall$ even $k \geq 2$; for each fixed, even $k \geq 2$, we require:

$$\begin{aligned}&\left[(2\mu^2 - a^2 - 1)\cdot(1 - a\mu) + (2\mu^2(1-\mu^2))\right]\left[\frac{((a\mu)/(1-\mu^2))^k}{k!} + \frac{((a\mu)/(1-\mu^2))^{k+1}}{(k+1)!}\right]\\ \geq&\left[(2\mu^2 - a^2 - 1)\cdot(1 + a\mu) + (2\mu^2(1-\mu^2))\right]\left[\frac{((a\mu)/(1-\mu^2))^k}{k!} - \frac{((a\mu)/(1-\mu^2))^{k+1}}{(k+1)!}\right]\end{aligned}$$

($\Updownarrow$) Factoring out common terms within the bracket of the Taylor series terms, we get

$$\begin{aligned}&\left[(2\mu^2 - a^2 - 1)\cdot(1 - a\mu) + (2\mu^2(1-\mu^2))\right]\left[\frac{((a\mu)/(1-\mu^2))^k}{k!}\right]\left[1 + \frac{((a\mu)/(1-\mu^2))}{(k+1)}\right]\\ \geq&\left[(2\mu^2 - a^2 - 1)\cdot(1 + a\mu) + (2\mu^2(1-\mu^2))\right]\left[\frac{((a\mu)/(1-\mu^2))^k}{k!}\right]\left[1 - \frac{((a\mu)/(1-\mu^2))}{(k+1)}\right]\end{aligned}$$

($\Updownarrow$) Multiplying through by $\frac{(1-\mu^2)^{k+1}\cdot(k+1)!}{(a\mu)^k}$, we get

$$\begin{aligned}&\left[(2\mu^2-a^2-1)\cdot(1-a\mu)+(2\mu^2(1-\mu^2))\right]\left[(k+1)(1-\mu^2)+a\mu\right]\\ \geq&\left[(2\mu^2-a^2-1)\cdot(1+a\mu)+(2\mu^2(1-\mu^2))\right]\left[(k+1)(1-\mu^2)-a\mu\right]\end{aligned}$$

($\Updownarrow$) Expanding within the second brackets, we get

$$\begin{aligned}&\left[(2\mu^2-a^2-1)\cdot(1-a\mu)+(2\mu^2(1-\mu^2))\right]\left[1-\mu^2+a\mu+k(1-\mu^2)\right]\\ \geq&\left[(2\mu^2-a^2-1)\cdot(1+a\mu)+(2\mu^2(1-\mu^2))\right]\left[1-\mu^2-a\mu+k(1-\mu^2)\right]\end{aligned}$$

($\Uparrow$) Dropping the $[1-\mu^2+a\mu]$ terms because they correspond exactly to an inequality already shown to hold in the sequence of steps to prove **Condition 2a**, we get

$$\begin{aligned}&\left[(2\mu^2-a^2-1)\cdot(1-a\mu)+(2\mu^2(1-\mu^2))\right]\left[k(1-\mu^2)\right]\\ \geq&\left[(2\mu^2-a^2-1)\cdot(1+a\mu)+(2\mu^2(1-\mu^2))\right]\left[k(1-\mu^2)\right]\end{aligned}$$

($\Updownarrow$) Dropping the multiplicative constant and then the additive constant, both of which are non-negative, we get

$$\begin{aligned}&\left[(2\mu^2-a^2-1)\cdot(1-a\mu)\right]\\ \geq&\left[(2\mu^2-a^2-1)\cdot(1+a\mu)\right]\quad\checkmark\text{ by Case 2 assumption }(2\mu^2-a^2-1)\leq 0,\text{ and }0\leq a\mu<1\end{aligned}$$

This last inequality holds because both sides are non-positive and the left-hand-side has weakly smaller magnitude. □

Lemma B.3. *Inequality* (2) *holds for all* $\mu\in(0,1)$ *and* $a\geq 1$.

Proof. Let $f(a,\mu)$ be the left-hand-side in inequality (2), i.e.,

$$\begin{aligned}f(a,\mu)\triangleq&\left[(2\mu^2-a^2-1)(1-a\mu)+2\mu^2(1-\mu^2)\right]\exp\left(\frac{a\mu}{1-\mu^2}\right)\\ &-\left[(2\mu^2-a^2-1)(1+a\mu)+2\mu^2(1-\mu^2)\right]\exp\left(\frac{-a\mu}{1-\mu^2}\right)\end{aligned}$$

Next we show the following inequalities: for all $\mu\in(0,1)$ and $a=1$,

(i) $f(a,\mu)>0$;

(ii) $f_1(a,\mu)\triangleq(1-\mu^2)\frac{\partial}{\partial a}f(a,\mu)\geq 0$;

(iii) $f_2(a,\mu)\triangleq(1-\mu^2)\frac{\partial}{\partial a}f_1(a,\mu)\geq 0$;

and for all $\mu\in(0,1)$ and $a\geq 1$,

(iv) $f_3(a,\mu)\triangleq\frac{1-\mu^2}{\mu^2}\exp\left(\frac{1}{1-\mu^2}\right)\frac{\partial}{\partial a}f_2(a,\mu)\geq 0$.

Combining (i)–(iv) proves the lemma.

Proof of (i). By definition, plugging in $a = 1$, $f(a, \mu)$ is

$$f(1,\mu) = -2(1-\mu^2)(1-\mu-\mu^2)\exp\left(\frac{\mu}{1-\mu^2}\right) + 2(1-\mu^2)(1+\mu-\mu^2)\exp\left(\frac{-\mu}{1-\mu^2}\right)$$

Now consider the Taylor series expansion of $\exp\left(\frac{-\mu}{1-\mu^2}\right)$ and $\exp\left(\frac{\mu}{1-\mu^2}\right)$ in $f(1,\mu)$. We analyze the first term and the remaining terms separately.
The first term of the Taylor series expansion in $f(1,\mu)$ is

$$-2(1-\mu^2)(1-\mu-\mu^2)\exp\left(\frac{\mu}{1-\mu^2}\right) + 2(1-\mu^2)(1+\mu-\mu^2)\exp\left(\frac{-\mu}{1-\mu^2}\right) = 4\mu(1-\mu^2) > 0$$

for all $\mu \in (0,1)$.
The k-th even terms and (k+1)-th odd terms of the Taylor series expansion, for $(k \geq 2)$*, in* $f(1,\mu)$ are

$$\begin{aligned}
&-2(1-\mu^2)(1-\mu-\mu^2)\left(\frac{1}{k!}\left(\frac{\mu}{1-\mu^2}\right)^k + \frac{1}{(k+1)!}\left(\frac{\mu}{1-\mu^2}\right)^{k+1}\right) \\
&\qquad + 2(1-\mu^2)(1+\mu-\mu^2)\left(\frac{1}{k!}\left(\frac{-\mu}{1-\mu^2}\right)^k + \frac{1}{(k+1)!}\left(\frac{-\mu}{1-\mu^2}\right)^{k+1}\right) \\
=&\frac{4\mu^{k+1}k}{(1-\mu^2)^{k-1}(k+1)!} > 0
\end{aligned}$$

for all $\mu \in (0,1)$.

Proof of (ii). By definition,

$$\begin{aligned}
f_1(a,\mu) = &\left[a^3\mu^2 + \mu^3 + a^2\mu(2-3\mu^2) - a(2-3\mu^2+2\mu^4)\right]\exp\left(\frac{a\mu}{1-\mu^2}\right) \\
&- \left[a^3\mu^2 - \mu^3 - a^2\mu(2-3\mu^2) - a(2-3\mu^2+2\mu^4)\right]\exp\left(\frac{-a\mu}{1-\mu^2}\right)
\end{aligned}$$

Plugging in $a = 1$ yields

$$\begin{aligned}
f_1(1,\mu) = &-\left[1-\mu-2\mu^2+\mu^3+\mu^4\right]\exp\left(\frac{\mu}{1-\mu^2}\right) \\
&+ \left[1+\mu-2\mu^2-\mu^3+\mu^4\right]\exp\left(\frac{-\mu}{1-\mu^2}\right)
\end{aligned}$$

Now consider the Taylor series expansion of $\exp\left(\frac{-\mu}{1-\mu^2}\right)$ and $\exp\left(\frac{\mu}{1-\mu^2}\right)$ in $f_1(1,\mu)$. We analyze the first term and the remaining terms separately.

The first term of Taylor series expansion in $f_1(1,\mu)$ is

$$-\left[1-\mu-2\mu^2+\mu^3+\mu^4\right]+\left[1+\mu-2\mu^2-\mu^3+\mu^4\right]=2\mu(1-\mu^2)\geq 0$$

for all $\mu\in[0,1)$.

The k-th even terms and (k+1)-th odd terms of the Taylor series expansion, for $(k\geq 2)$*, in* $f_1(1,\mu)$ are

$$\begin{aligned}&-\left[1-\mu-2\mu^2+\mu^3+\mu^4\right]\left(\frac{1}{k!}\left(\frac{\mu}{1-\mu^2}\right)^k+\frac{1}{(k+1)!}\left(\frac{\mu}{1-\mu^2}\right)^{k+1}\right)\\&\qquad+\left[1+\mu-2\mu^2-\mu^3+\mu^4\right]\left(\frac{1}{k!}\left(\frac{-\mu}{1-\mu^2}\right)^k+\frac{1}{(k+1)!}\left(\frac{-\mu}{1-\mu^2}\right)^{k+1}\right)\\=&\frac{2\mu^{k+1}k}{(1-\mu^2)^{k-1}(k+1)!}\geq 0\end{aligned}$$

for all $\mu\in[0,1)$.

Proof of (iii). By definition,

$$\begin{aligned}f_2(a,\mu)=&\left[-2+5\mu^2+a^3\mu^3-4\mu^4+2\mu^6+a^2\mu^2(5-6\mu^2)+a\mu(2-7\mu^2+4\mu^4)\right]\exp\left(\frac{a\mu}{1-\mu^2}\right)\\&-\left[-2+5\mu^2-a^3\mu^3-4\mu^4+2\mu^6+a^2\mu^2(5-6\mu^2)-a\mu(2-7\mu^2+4\mu^4)\right]\exp\left(\frac{-a\mu}{1-\mu^2}\right)\end{aligned}$$

Plugging in $a=1$ yields

$$\begin{aligned}f_2(1,\mu)=&-\left[1-\mu-5\mu^2+3\mu^3+5\mu^4-2\mu^5-\mu^6\right]\exp\left(\frac{\mu}{1-\mu^2}\right)\\&+\left[1+\mu-5\mu^2-3\mu^3+5\mu^4+2\mu^5-\mu^6\right]\exp\left(\frac{-\mu}{1-\mu^2}\right)\end{aligned}$$

Note that $f_2(1,0)=0$ and we show $f_2(1,\mu)$ is non-increasing in μ below. To see this, consider the partial derivative of $f_2(1,\mu)$ with respect to μ, it is

$$\begin{aligned}\frac{\partial}{\partial\mu}f_2(1,\mu)=&-\left[-11+7\mu+31\mu^2-22\mu^3-24a^4+11\mu^5+6\mu^6\right]\exp\left(\frac{\mu}{1-\mu^2}\right)\frac{\mu}{1-\mu^2}\\&+\left[-11-7\mu+31\mu^2+22\mu^3-24\mu^4-11\mu^5+6\mu^6\right]\exp\left(\frac{-\mu}{1-\mu^2}\right)\frac{\mu}{1-\mu^2}\end{aligned}$$

Multiplying $\frac{\partial}{\partial\mu}f_2(1,\mu)$ by $\frac{1-\mu^2}{\mu}\exp\left(\frac{1}{2-2\mu^2}\right)$, we get

$$\begin{aligned}&-\left[-11+7\mu+31\mu^2-22\mu^3-24a^4+11\mu^5+6\mu^6\right]\exp\left(\frac{1+2\mu}{2-2\mu^2}\right)\\&\quad+\left[-11-7\mu+31\mu^2+22\mu^3-24\mu^4-11\mu^5+6\mu^6\right]\exp\left(\frac{1-2\mu}{2-2\mu^2}\right)\end{aligned}\tag{3}$$

Now consider the Taylor series expansion of $\exp\left(\frac{1+2\mu}{2-2\mu^2}\right)$ and $\exp\left(\frac{1-2\mu}{2-2\mu^2}\right)$ in (3). We analyze the first two terms and the remaining terms separately.

The first and second terms of the Taylor series expansion in (3). It is

$$\begin{aligned}
&-\left[-11+7\mu+31\mu^2-22\mu^3-24\mu^4+11\mu^5+6\mu^6\right]\left[1+\frac{1+2\mu}{2-2\mu^2}\right]\\
&\quad+\left[-11-7\mu+31\mu^2+22\mu^3-24\mu^4-11\mu^5+6\mu^6\right]\left[1+\frac{1-2\mu}{2-2\mu^2}\right]\\
=&\mu+19\mu^3-10\mu^5\geq 0
\end{aligned}$$

for all $\mu\in[0,1)$.

The k-th term of the Taylor series expansion, for $(k\geq 3)$, *in* (3). It is

$$\begin{aligned}
&-\left[-11+7\mu+31\mu^2-22\mu^3-24\mu^4+11\mu^5+6\mu^6\right]\left[\frac{1}{k!}\frac{(1+2\mu)^k}{(2-2\mu^2)^k}\right]\\
&\quad+\left[-11-7\mu+31\mu^2+22\mu^3-24\mu^4-11\mu^5+6\mu^6\right]\left[\frac{1}{k!}\frac{(1-2\mu)^k}{(2-2\mu^2)^k}\right]
\end{aligned}\tag{4}$$

Multiplying by $k!(2-2\mu^2)^k$,

$$\begin{aligned}
&-\left[-11+7\mu+31\mu^2-22\mu^3-24\mu^4+11\mu^5+6\mu^6\right](1+2\mu)(1+2\mu)^{k-1}\\
&\quad+\left[-11-7\mu+31\mu^2+22\mu^3-24\mu^4-11\mu^5+6\mu^6\right](1-2\mu)(1-2\mu)^{k-1}
\end{aligned}$$

Note that $-\left[-11+7\mu+31\mu^2-22\mu^3-24\mu^4+11\mu^5+6\mu^6\right]\geq 0$ and $1+2\mu\geq|1-2\mu|$ for all $\mu\in[0,1)$. Thus, to show (4) is non-negative for all $\mu\in[0,1)$, it is sufficient to argue

$$\begin{aligned}
&-\left[-11+7\mu+31\mu^2-22\mu^3-24\mu^4+11\mu^5+6\mu^6\right](1+2\mu)\\
&\quad\geq\left|\left[-11-7\mu+31\mu^2+22\mu^3-24\mu^4-11\mu^5+6\mu^6\right](1-2\mu)\right|
\end{aligned}$$

which is true for all $\mu\in[0,1)$.

Proof of (iv). By definition,

$$\begin{aligned}
f_3(a,\mu)=&\left[a^3\mu^2+a^2\mu(8-9\mu^2)-\mu(4-7\mu^2+2\mu^4)+a(12-29\mu^2+16\mu^4)\right]\exp\left(\frac{1+a\mu}{1-\mu^2}\right)\\
&-\left[a^3\mu^2-a^2\mu(8-9\mu^2)+\mu(4-7\mu^2+2\mu^4)+a(12-29\mu^2+16\mu^4)\right]\exp\left(\frac{1-a\mu}{1-\mu^2}\right)
\end{aligned}$$

Now consider the Taylor series expansion of $\exp\left(\frac{1+a\mu}{1-\mu^2}\right)$ and $\exp\left(\frac{1-a\mu}{1-\mu^2}\right)$ in $f_3(a,\mu)$. We analyze the first two terms and the remaining terms separately.

The first and second terms of the Taylor series expansion of $f_3(a,\mu)$ are

$$\begin{aligned}
&\left[a^3\mu^2 + a^2\mu(8-9\mu^2) - \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right]\left[1+\frac{1+a\mu}{1-\mu^2}\right] \\
&\quad - \left[a^3\mu^2 - a^2\mu(8-9\mu^2) + \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right]\left[1+\frac{1-a\mu}{1-\mu^2}\right] \\
=&\frac{2\mu}{1-\mu^2}(-8+18\mu^2+a^4\mu^2-11\mu^4+2\mu^6+a^2(28-55\mu^2+25\mu^4)) \geq 0
\end{aligned}$$

which is true for all $\mu \in [0,1)$ and $a \geq 1$.
The k-th term of the Taylor series expansion, for $(k \geq 3)$*, of* $f_3(a,\mu)$. It is

$$\begin{aligned}
&\left[a^3\mu^2 + a^2\mu(8-9\mu^2) - \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right]\left[\frac{1}{k!}\frac{(1+a\mu)^k}{(1-\mu^2)^k}\right] \\
&\quad - \left[a^3\mu^2 - a^2\mu(8-9\mu^2) + \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right]\left[\frac{1}{k!}\frac{(1-a\mu)^k}{(1-\mu^2)^k}\right]
\end{aligned} \tag{5}$$

Multiplying by $k!(1-\mu^2)^k$,

$$\begin{aligned}
&\left[a^3\mu^2 + a^2\mu(8-9\mu^2) - \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right](1+a\mu)(1+a\mu)^{k-1} \\
&\quad - \left[a^3\mu^2 - a^2\mu(8-9\mu^2) + \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right](1-a\mu)(1-a\mu)^{k-1}
\end{aligned}$$

Notice that $\left[a^3\mu^2 + a^2\mu(8-9\mu^2) - \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right](1+a\mu) \geq 0$ and $1+a\mu \geq |1-a\mu|$ for all $\mu \in [0,1)$ and $a \geq 1$. Thus, to show (5) is non-negative for all $\mu \in [0,1)$ and $a \geq 1$, it is sufficient to argue

$$\begin{aligned}
&\left[a^3\mu^2 + a^2\mu(8-9\mu^2) - \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right](1+a\mu) \\
&\quad \geq \left|\left[a^3\mu^2 - a^2\mu(8-9\mu^2) + \mu(4-7\mu^2+2\mu^4) + a(12-29\mu^2+16\mu^4)\right](1-a\mu)\right|
\end{aligned}$$

which is true for all $\mu \in [0,1)$ and $a \geq 1$. □

C Proof of Theorem 4.3

We show that the estimator of ridge regression is asymptotically normal and analyze its bias and asymptotic variance (Definition 4.1). Recall we are running Ridge regression (1) on a training set with m points and each point $(x_i, f(x_i))$ has p-dimensional features $x_i \in \mathbb{R}^p$.

Theorem 4.3 (specified). *Given an infinite sequence of distinct and bounded features* $\{x_1, x_2, \ldots\}$ *with labels* $f(x_i) = w_0^T x_i + e_i$ *where the* e_i *are i.i.d., bounded, mean zero, and*

variance σ_e. With regularization parameter λ and in the limit of m, the distribution of prediction error for any point x, specifically $\hat{f}(x) - \mathbf{E}[f(x)]$, of ridge regression on the first m data points is asymptotically normal with finite average bias $\mu_\lambda(x)$ that is a monotonically increasing function of λ and finite asymptotic variance $\sigma_\lambda^2(x)$ that is a monotonically decreasing function of λ. Specifically,

$$\sqrt{m}\,[\hat{f}(x) - \mathbf{E}[f(x)] - \mu_\lambda(x)] \xrightarrow{d} \mathcal{N}(0,\, \sigma_\lambda^2(x)).$$

The asymptotic bias of the prediction is

$$\mu_\lambda(x) = \lambda\,(A_\infty + \lambda\,I)^{-1}\,w_0^T x,$$

where $A_m = \frac{1}{m}\sum_{i=1}^m x_i x_i^T$, $A_\infty = \lim_{m\to\infty} A_m$, and I is the $p \times p$ identity matrix. The asymptotic variance of the prediction is

$$\sigma_\lambda^2(x) = \sigma_e^2 x^T\,(A_\infty + \lambda\,I)^{-1}\,(A_\infty)\,(A_\infty + \lambda\,I)^{-1}\,x.$$

Proof. We are calculating the distribution of the error of the prediction for point x, i.e., $\hat{f}(x) = \hat{w}^T x$, with the expected label $\mathbf{E}[f(x)] = w_0^T x$. Note that we can express the minimizer of the empirical risk as

$$\begin{aligned}
\hat{w} &= (A_m + \lambda\,I)^{-1}\,\tfrac{1}{m}\sum\nolimits_{i=1}^m f(x_i)\,x_i \\
&= (A_m + \lambda\,I)^{-1}\,\tfrac{1}{m}\sum\nolimits_{i=1}^m (w_0^T x_i + e_i)x_i \\
&= w_0 - (A_m + \lambda\,I)^{-1}\,\lambda w_0 + (A_m + \lambda\,I)^{-1}\,\tfrac{1}{m}\sum\nolimits_{i=1}^m x_i e_i.
\end{aligned}$$

Note that since e_i are mean zero random variables, then

$$\mathbf{E}[\hat{w}] = w_0 - \lambda\,(A_m + \lambda\,I)^{-1}\,w_0.$$

Therefore, for prediction $\hat{y}(x) = \hat{w}^T x$, we can express asymptotic bias as

$$\mu_\lambda(x) = \lambda\,(A_\infty + \lambda\,I)^{-1}\,w_0^T x.$$

By the Central Limit Theorem

$$\tfrac{1}{\sqrt{m}}\sum\nolimits_{i=1}^m x_i e_i \xrightarrow{d} \mathcal{N}(0_p,\, \sigma_e^2 A_\infty),$$

where $\sigma_e^2 = \mathbf{E}[e_i^2]$ and $\mathcal{N}(0_p,\, \sigma_e^2 A_\infty)$ is a p-dimensional normal distribution with the mean at the origin and covariance matrix $\sigma_e^2 A_\infty$. Thus,

$$\sqrt{m}\,[\hat{f}(x) - w_0^T x - \mu_\lambda(x)] \xrightarrow{d} \mathcal{N}(0,\, \sigma_\lambda^2(x)),$$

where the asymptotic variance is

$$\sigma_\lambda^2(x) = \sigma_e^2 x^T \left(A_\infty + \lambda\, I\right)^{-1} A_\infty \left(A_\infty + \lambda\, I\right)^{-1} x.$$

Note that the bias $\mu_\lambda(x)$ is monotonically increasing in λ, and the variance $\sigma_\lambda^2(x)$ is monotonically decreasing in λ. □