Opening the Black Box*

A Statistical Theory of the Value of Data

Giovanni Colla Rizzi†

Latest Draft

Abstract

This paper develops a theory of the value of data for prediction. An agent chooses a sample of individuals and a subset of their covariates to collect in order to estimate the parameters of a data-generating process and predict an outcome for a representative individual based on her covariates. The main findings are: (i) covariates collected on the sample exhibit economies of scope, as the value of collecting a covariate on the sample increases when another is observed; (ii) the value of collecting a covariate on the sample is inverted-U-shaped in the sample size; and (iii) the value of collecting a covariate on the representative individual is strictly increasing in the sample size.

These patterns have several policy implications. First, firms that collect different covariates and sell predictions in separate markets may find it profitable to merge even when this reduces welfare, particularly when data are scarce due to, for example, privacy regulation. Second, agreements that allow firms to sell covariate bundles are always procompetitive because they eliminate double marginalization, whereas bundling observations can be anticompetitive when data are abundant because it raises the price of data. Third, when selling data to competing prediction providers, a data seller may profitably exclude one buyer even though this reduces total welfare, especially when competition among providers is fierce and data are abundant.

JEL Classification: C11, D83, L12, O33.

Keywords: Value of Data, Prediction, Economies of Scope, Data Markets.

*I thank Patrick Rey for his guidance, and Jean-Pierre Florens and Doh-Shin Jeon for their suggestions. I also thank Jad Beyhum, Michele Bisceglia, Zhijun Chen, Krishna Dasaratha, Alexandre de Corniere, Eric Gautier, Johannes Gessen, Andrei Hagiu, Johannes Horner, Marco Iansiti, Bruno Jullien, Hiroaki Kaido, Simon Loertscher, Friedrich Lucke, Leonardo Madio, Giovanni Morzenti, Juan Ortner, Christoph Reidl, Andrew Rhodes, David Salant, Maximilian Schaefer, Sara Shahanaghi, Tim Simcoe, Alex Smolin, Emanuele Tarantino, Ehsan Valavi, Marshall Van Alstyne, Davide Viviano, Julian Wright and the participants at the European Association of Industrial Economics Conference (Valencia, 2025), Questrom Digital Platform Seminars, TSE and Boston University.

†Toulouse School of Economics, University of Toulouse Capitole, France. E-mail: giovanni.rizzi@tse-fr.eu

1 Introduction

In digital markets, data drives competitive advantage, notably because it allows firms to make better predictions: Google and Meta use data to predict ad clicks, Amazon and Uber to predict the demand of goods and rides, and Spotify and Netflix to predict choices of songs and movies. Indeed, over the last decade, many digital markets have become dominated by a few firms that control vast amounts of data, attracting the attention of competition authorities. According to the 2021 U.S. House Report, for instance, "data advantages [...] can reinforce dominance and serve as a barrier to entry,"[1] a warning echoed in the EU Data Act proposal[2] and in major policy reports.[3]

These concerns are partly fueled by earlier statements from industry leaders: as Google's CEO declared in 2009, "Scale is the key. We just have so much scale in terms of the data we can bring to bear."[4] More recently, however, technology firms have downplayed policymakers' claims, arguing that data has diminishing returns because of the Law of Large Numbers, and that data concentration reflects the technological gap between firms rather than entry barriers.[5]

To assess these arguments, we must understand whether data exhibit economies of scale and whether there are economies of scope between different varieties of data. I develop a theory of the value of data for an agent predicting a target variable generated by a linear process of infinitely many covariates. I distinguish between the value of additional *observations* (e.g., individuals) and additional *covariates* (e.g., their attributes). Furthermore, I distinguish between the *training covariates* the agent uses to learn the parameters of the process (e.g., tall people tend to like basketball), and the *prediction covariates* she uses to apply the learned parameters to the individual of interest (e.g., Martha is tall, so she probably likes basketball).

Specifically, the agent chooses how many observations and which training and prediction covariates to collect to minimize the out-of-sample prediction error, subject to data collection costs. I first characterize the optimal predictor for a given dataset. Building on this, I derive closed-form expressions for the value of data, showing that returns depend on how the variance is distributed across covariates. Three main insights on the economies of scope and scale of data emerge.

First, there are *economies of scope in training*. Reducing regression noise by observing training covariates has a positive and accelerating effect on the precision of the parameter estimates: when the estimation is already relatively precise (low regression noise), further reductions in noise yield disproportionately large gains in precision. I name this convexity the *House Party Effect*, by analogy to a conversation in a noisy room.[6]

[1] U.S. House of Representatives (2020), Investigation of Competition in Digital Markets, 117-40, pp. 36–38.
[2] European Commission (2023), Data Act proposal, COM(2023) 193 final.
[3] Digital Platforms (2019) and UK Competition and Markets Authority (2019).
[4] Schmidt (2009), "How Google Plans to Stay Ahead in Search," Bloomberg, 2 October.
[5] Varian (2018) and Bajari et al. (2019).
[6] At a loud party with a hundred guests, the departure of five hardly makes a difference in hearing one's

Second, *training covariates and observations are complements when they scarce, but become substitutes when they are abundant.* Observations reduce the estimator variance arising from dimensionality and regression noise. Additional covariates increase dimensionality, which raises the value of observations, but also reduce regression noise, which lowers it. Adding a covariate increases the marginal value of an observation for an agent with few covariates, as the dimensionality effect dominates; the opposite is true for an agent with many covariates, as the noise-reduction effect dominates.

Third, *training data and prediction covariates are complements.* When the agent collects a new training covariates or observations, it increases the precision of the estimates of the parameters of the prediction covariates: there are positive spillovers of parameter estimate precision across training and prediction covariates.

I then allow the agent to select the most valuable covariates, namely those with the highest variance, and show that selection induces diminishing returns.[7] When covariate variances are similar, the House Party Effect dominates, and returns to covariates are increasing. Instead, when variance is concentrated in a few covariates, the declining marginal variance from selection prevails, and returns may be diminishing. Surprisingly, returns to additional covariates may be first decreasing and then increasing, so the agent might need to achieve a minimum scale before she can benefit from economies of scope.

Finally, I show these findings carry significant policy implications in three applications. First, an incumbent (e.g., Google) and an entrant (e.g., Fitbit) with distinct covariates sell to separate prediction markets of a similar size. Even if their demands are unrelated, I show that Google will acquire Fitbit because of economies of scope in covariates, thereby preventing entry even if it is socially desirable, e.g., because of positive knowledge spillovers. This highlights a tension between private and social incentives to data consolidation, all the more so if the number of collected covariates is limited, e.g., because of privacy regulation.

Second, pooling covariates on the same individuals among data brokers selling to prediction firms eliminates double marginalization, generating efficiencies. By contrast, brokers may collude by pooling samples of different individuals, especially when there are large samples and few covariates.

Third, data sellers may profitably conclude exclusive licensing agreements with firms selling predictions, such as the 2024 Reddit–OpenAI deal, to commit not to engage in opportunistic behavior. However, when firms complement sellers' data with their own proprietary data, such agreements may harm social welfare by disincentivizing investment in proprietary data by the excluded prediction firm, all the more so if investment is costly and competition between prediction firms is fierce.

On top of these insights, I exploit the well-known equivalence to show that my Bayesian

interlocutor, but when there are only seven guests, the same reduction makes the conversation perfectly clear, as the room falls silent except for the speakers themselves.

[7] The mechanism is analogous to Ricardo (1817) Law of Diminishing Returns for heterogeneous land quality: as the agent collects the most variable covariates first, the variance of the marginal covariate is diminishing.

agent can be interpreted as a frequentist agent estimating the unknown parameters with a ridge estimator whose penalty parameter is optimal ex-ante under the prior distribution of the training data. In this way, the model delivers a simple non-asymptotic, closed-form rule for tuning regularization as a function of sample size, residual noise, and covariate variances. A key implication is that the optimal overall strength of regularization need not be monotone in the number of regressors: once the model is sufficiently rich and the marginal covariates are informative relative to the residual noise, adding further regressors can reduce the optimal penalty, providing a tractable mechanism consistent with the "double descent" patterns documented in recent machine-learning work.

Related Literature. There is a rich information design literature on the value of data, starting with Bergemann, Bonatti, and Smolin (2018) and continuing with Jones and Tonetti (2020), Bergemann, Bonatti, and Gan (2022), Bergemann and Bonatti (2024), and Acemoğlu et al. (2022). While this literature elegantly connects data value to the choice of a probability distribution by an agent, I link it directly to the realization of a dataset, allowing me to disentangle the statistical forces at play. Methodologically, my work is related to Montiel Olea et al. (2022), Iyer and Ke (2024), and Dasaratha, Ortner, and Zhu (2025), who analyze competition between different models with different covariates. In contrast, I jointly model covariates and observations, and distinguish training from prediction data, which allows me to derive structural non-convexities generating increasing returns and complementarities. Another related paper is Strzalecki (2024) interprets deviations from Bayesian updating through a misspecification parameter that is taken as primitive. In contrast, my model delivers an analogous distortion as "classic" Bayesian behavior under incomplete data collection, providing a microfoundation for this feature and reconnecting it to Bayesian updating.

I also contribute to the broad literature on economies of scope and scale to data. Whereas most models fix covariates and study returns to observations as Bajari et al. (2019) and Goldfarb and Tucker (2011), my framework endogenizes covariate collection. This extension rationalizes empirical findings on complementarities in Schaefer and Sapi (2023) and economies of scope in Carballa-Smichowski et al. (2025b). Schaefer (2025) develops a complementary frequentist approach and shows that the distribution of the covariate eignespectrum shapes returns to scale, coherently with our results. Allcott et al. (2025) run a structural model to estimate returns to scale of additional observations in search. They finding diminishing returns and evidence of limited complementarities across different queries. Radner and Stiglitz (1984) attributes increasing returns to information costs, while I show they may emerge independently of costs.

My work provides microfoundations for two strands of literature that take increasing returns to data as assumptions: the IO literature on platforms and the macroeconomics literature on data as a production input. Prior work explains increasing returns through feedback between data and demand (Hagiu and Wright (2023), Prüfer and Schottmüller (2021), Farboodi

and Veldkamp (2025), Aral, Brynjolfsson, and Wu (2008), and Cong, He, and Yu (2021)) or by assuming complementarities across datasets Carballa-Smichowski et al. (2025a), De Corniere and Taylor (2025), and Calzolari, Cheysson, and Rovatti (2025). I show instead that prediction accuracy alone generates increasing returns due to the statistical structure of data, independent of demand feedback.

The applications of the model contribute more broadly to the IO literature on digital markets. De Corniere and Taylor (2025) shows that the pro- or anti-competitive effect of collecting more data only depends on the supply of data-driven services rather than its demand. Furthermore, Cornière and Taylor (2024) studies data-driven mergers developing a theory of harm of mergers which rely on cross-market effects. Both works complement the applications of my paper as they treat data as an undifferentiated good to which my results readily apply. Several papers on the economics of patents may be applied to datasets. Lerner and Tirole (2004) deals with complementarity/substitutability of patents and the private and social value of commercializing them jointly in pools. Gu, Madio, and Reggiani (2021) also study broker pools in a context without double marginalization and conclude they are anticompetitive when datasets are substitutes. I connect these two papers and microfound the welfare considerations on the statistical properties of datasets, complementing the findings in Nocke, Peitz, and Stahl (2007) by showing another source of potential social benefits of vertical integration and the avoidance of platform fragmentation. Katz and Shapiro (1986) show that exclusive deals are optimal when an inventor sells patents to competing firms, and this results in a suboptimal dissemination of patents; Aghion and Bolton (1987) shows that exclusive deals may serve as a strategic commitment device to deter entry by raising rivals' costs, allowing incumbents to extract rents from buyers at the expense of social welfare. My model develops insights from both papers to nonrival data markets with complementarities and endogenous investment in proprietary data.

Finally, I contribute to the literature on theoretical machine learning by developing a simple framework to study data scaling laws, providing a simple explanation of the phenomenon of double descent, i.e., that maximum likelihood-based algorithms generalize well even when overparametrized as explored in Hastie et al. (2020), Nakkiran et al. (2021), and Belkin et al. (2019).

2 Model Setup

An agent (she) seeks to predict a target variable of a target individual (he). To do so, she can collect (i) training data, consisting of covariates and past target variables for a sample of individuals; and (ii) prediction covariates for the target individual. The agent chooses the size of the training sample and which covariates to use in training and prediction, trading off prediction accuracy and data collection costs.

2.1 Data-Generating Process

The agent must predict a *target variable* $y \in \mathbb{R}$ for a *target individual*. For each individual i, the target variable is determined by

$$y^i = \sum_{k \in \mathbb{N}} \beta_k x_k^i,$$

where $x_k^i \in \mathbb{R}$ is the realization of covariate k for individual i, and β_k is its (unknown) impact on y. I normalize $\text{Var}[y] = 1$.

Parameters Parameters are mutually independent and independent of covariates, with prior distribution[8]

$$\beta_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

The prior variance reflects the agent's uncertainty on the impact of covariate k on the target variable. Because the parameters are shared across individuals, they are learned through across-user learning in the sense of Hagiu and Wright (2023).

Covariates Covariates are mutually independent and i.i.d. across individuals, with prior distribution[9]

$$x_k^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_k^2), \qquad \sigma_k^2 \in (0, 1),$$

where I index covariates in decreasing order of variance,

$$\sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2 \geq \cdots.$$

Because the prior parameter variance is identical across covariates, σ_k^2 captures the amount of information contained in each covariate, i.e., the fraction of the variance of y due to covariate k. Moreover, let

$$S(\mathcal{K}) \equiv \sum_{k \in \mathcal{K}} \sigma_k^2,$$

denote the ***cumulative variance of a covariate set*** $\mathcal{K} \subseteq \mathbb{N}$, which is the fraction of the variance of y due to covariates in $\mathcal{K}$. Note that $S(\varnothing) = 0$ and $S(\mathbb{N}) = 1$.[10]

Notation For future reference, let $\mathcal{K}^c \equiv \mathbb{N} \setminus \mathcal{K}$ denote the complement of $\mathcal{K}$, $\boldsymbol{x}^i \equiv (x_k^i)_{k \in \mathbb{N}}$ denote the covariate vector, and $\boldsymbol{\beta} \equiv (\beta_k)_{k \in \mathbb{N}}$ denote the parameter vector. Furthermore, for any vector $\boldsymbol{v}$ and set $\mathcal{K}$, let $\boldsymbol{v}_\mathcal{K} \equiv (v_k)_{k \in \mathcal{K}}$ denote the subvector of $\boldsymbol{v}$ with indices in $\mathcal{K}$.

[8] Normalizing $\text{Var}[\beta_k] = 1$ is without loss of generality: we may recover any variance τ^2 by $\tilde{\sigma}_k^2 \equiv \tau^2 \sigma_k^2$.

[9] Appendix B relaxes the independence assumption by allowing arbitrary covariance structures. After using Principal Component Analysis to orthogonalize the covariates, all results carry through with the same notation and proofs.

[10] The upper bound follows from the independence of $\{\beta_k\}$ and $\{x_k^i\}$ and from$\text{Var}[y] = 1$.

2.2 Dataset Design and Predictor

I will distinguish between the way the agent collects the data (dataset design choice) and the way she uses the data (predictor choice).

Dataset Design To determine a dataset design, the agent chooses a set of ***training covariates***, $\mathcal{T} \subset \mathbb{N}$, and the size of a sample of n individuals. Their realization is the ***training data***,

$$\boldsymbol{M}_{\mathcal{T}}^{n} \equiv (\boldsymbol{y}, \boldsymbol{X}),$$

where $\boldsymbol{y} \in \mathbb{R}^n$ and $\boldsymbol{X} \equiv \{\boldsymbol{x}_{\mathcal{T}}^i\}_{i=1}^n \in \mathbb{R}^{n \times |\mathcal{T}|}$. For any matrix of covariates $\boldsymbol{X}$, let $\mathscr{P}_X(n, \mathcal{T})$, denote its prior distribution defined under the data-generating process above.

Furthermore, for the target individual, the agent collects a set of ***prediction covariates***, $\mathcal{P} \subset \mathbb{N}$. Their realization is the vector

$$\boldsymbol{x}_{\mathcal{P}} \in \mathbb{R}^{|\mathcal{P}|}.$$

To summarize, the agent chooses a dataset design $(n, \mathcal{T}, \mathcal{P})$ defined as follows:

Definition 1. A ***dataset design*** is a triple $(n, \mathcal{T}, \mathcal{P})$, where $n \in \mathbb{N}$ is the sample size and $\mathcal{T}, \mathcal{P} \subset \mathbb{N}$ are the training and prediction covariate sets. Its ***size*** is $(n, |\mathcal{T}|, |\mathcal{P}|)$, where $|\mathcal{T}|$ is the ***training (covariate) size*** and $|\mathcal{P}|$ is the ***prediction (covariate) size***.

The dataset design is the data collection strategy of the agent and reflects the choice of the experiments, sampling techniques and sensors which generate the data. As the model is linear, the dataset design also determines the number of parameters estimated in learning and, hence, the effective complexity of the training algorithm.

A given design $(n, \mathcal{T}, \mathcal{P})$,will induce an array of random variables, whose realization is a ***dataset*** comprising the training data and prediction covariates:

$$\boldsymbol{D} \equiv (\boldsymbol{M}_{\mathcal{T}}^{n}, \boldsymbol{x}_{\mathcal{P}}),$$

where we let $\mathscr{P}_D(n, \mathcal{T}, \mathcal{P})$, denote the prior distribution of a dataset of design design $(n, \mathcal{T}, \mathcal{P})$,defined under the data-generating process in Section 2.1.

Predictor The agent's predictions will depend on the realization of the dataset. A ***predictor*** is a measurable map $\hat{y} : \mathbb{R}^{n \times (1+|\mathcal{T}|)} \times \mathbb{R}^{|\mathcal{P}|} \to \mathbb{R}$ that maps a dataset $\boldsymbol{D}$ to a prediction

$$\boldsymbol{D} \mapsto \hat{y}(\boldsymbol{D}).$$

Equivalently, the prediction is the realization of a random variable which is a deterministic function of the dataset, i.e. the predictor. The choice of a predictor reflects how the agent plans to use the data it will observes to make a prediction.

2.3 Agent's Problem

Agent Utility Denoting the agent's prediction by $\hat{y} \in \mathbb{R}$, the realized prediction error is $y - \hat{y}$, and the agent suffers a quadratic loss so that

$$u(y, \hat{y}) = -(y - \hat{y})^2 .$$

Therefore, agent is penalized proportionally more for larger prediction errors. Because of its tractability, quadratic loss is a standard assumption in econometrics and machine learning. Furthermore it has several theoretical properties discussed in Brier (1950) and axiomatized in Selten (1998).[11]The agent will attempt to minimize her expected loss based on her knowledge on the data-generating process at the time of prediction. To increase her knowledge, she will choose a design which will produce in a dataset.

Design Cost I assume the cost of a design $(n, \mathcal{T}, \mathcal{P})$ depends only on the size of the sample and the covariate sets, and is additively separable:

$$C(n, \mathcal{T}, \mathcal{P}) = C_n(n) + C_t(|\mathcal{T}|) + C_p(|\mathcal{P}|).$$

This reduced form cost imposes a minimal structure on results, making the qualitative predictions of the model depend exclusively on the improvements in prediction accuracy stemming from data collection rather than from the function form of the cost of design. We assume that the choice of a predictor is costless.

Timing The agent's objective is to maximize expected utility minus dataset design cost. The model unfolds in three steps:

1. ***Dataset Design Problem***: the agent chooses a dataset design $(n, \mathcal{T}, \mathcal{P})$ and pays the associated cost;
2. ***Inference Problem***: given the dataset design, the agent chooses a predictor $\hat{y}(\boldsymbol{D})$; and
3. ***Dataset Realization:*** nature draws a dataset $\boldsymbol{D}$ from the data-generating process $\mathscr{P}$; the prediction and the target variable are realized and losses occur.

As customary we will solve the model backwards. I will give a broad overview of the problem the agent faces, and will defer the solution to Sections 3 and 4.

[11]Namely, among all incentive-compatible scoring rules—that is, rules for which truthful probabilistic forecasts maximize expected scores—the quadratic rule is uniquely characterized (up to a positive linear transformation) by three axioms: symmetry, elongation invariance, and neutrality.

2.3.1 Inference Problem

The agent will choose the predictor so that conditional on the realization of a dataset $\boldsymbol{D}$, maximizes her expected utility, conditional on $\boldsymbol{D}$, i.e.

$$U_{\boldsymbol{D}}(\hat{y}) \equiv \mathbb{E}_{y|\boldsymbol{D}}\left[u\left(y, \hat{y}(\boldsymbol{D})\right) \mid \boldsymbol{D}\right].$$

Therefore, agent chooses the ***optimal predictor***

$$\hat{y}^*(\boldsymbol{D}) = \arg\max_{\hat{y}} U_{\boldsymbol{D}}(\hat{y}).$$

For quadratic loss, $\hat{y}^*(\boldsymbol{D})$ exists and is unique.

2.3.2 Dataset Design Problem

The agent must choose which dataset design allows her to maximize her utility, having as outside option the choice not to collect any data and to make the ex-ante optimal predictor $\hat{y}^*(\varnothing)$. We will define the net ex-ante expected utility she derives from a design $(n, \mathcal{T}, \mathcal{P})$ as the value thereof.

Definition 2. The ***value of a dataset design*** $(n, \mathcal{T}, \mathcal{P})$ is

$$V(n, \mathcal{T}, \mathcal{P}) \equiv \underbrace{\mathbb{E}_{\boldsymbol{D}\sim\mathscr{P}_{\boldsymbol{D}}(n,\mathcal{T},\mathcal{P})}\left[U_{\boldsymbol{D}}(\hat{y})\right]}_{\text{Ex-ante Expected Utility of } (n,\mathcal{T},\mathcal{P})} - \underbrace{\mathbb{E}_{y}\left[U_{\varnothing}(\hat{y})\right]}_{\text{Ex-ante Outside Option}}.$$

The value of a design $(n, \mathcal{T}, \mathcal{P})$ is the maximal increase in ex-ante expected utility that that agent can achieve when using the optimal predictor $\hat{y}^*(\boldsymbol{D})$.

Maximizing the utility minus data collection costs amounts to maximizing the value of a design minus its costs. Therefore, the agent will solve

$$\max_{n,\mathcal{T},\mathcal{P}} \left\{ V(n, \mathcal{T}, \mathcal{P}) - \left(C_n(n) + C_t(|\mathcal{T}|) + C_p(|\mathcal{P}|)\right) \right\},$$

which, I will show, always has a solution. We can distinguish two subproblems.

Covariate Selection Problem Because costs depend only on the dimensions $|\mathcal{T}|$ and $|\mathcal{P}|$, for given a design size (n, t, p), the choice of covariates dose not affect their cost. Therefore, I fix (n, t, p) and let the agent choose

$$\left(\tilde{\mathcal{T}}(n,t,p), \tilde{\mathcal{P}}(n,t,p)\right) \equiv \arg\max_{|\mathcal{T}|\leq t,\ |\mathcal{P}|\leq p} V(n, \mathcal{T}, \mathcal{P}),$$

which, we will show, always exist. The ***optimal constrained covariate sets*** $\tilde{\mathcal{T}}(n,t,p), \tilde{\mathcal{P}}(n,t,p)$ are respectively the best training covariate set of size t and prediction covariate set of size p.

Accuracy Problem Having found the optimal constrained covariate sets, I obtain a micro-founded measure of the increase in prediction accuracy deriving from the best design of a given size (n, t, p) under the optimal predictor.

Definition 3. The ***(surplus prediction) accuracy*** of a design of size (n, t, p) is

$$A(n, t, p) \equiv \max_{\mathcal{T} \leq t, \mathcal{P} \leq p} V(n, \mathcal{T}, \mathcal{P}) = V\left(n, \tilde{\mathcal{T}}(n, t, p), \tilde{\mathcal{P}}(n, t, p)\right).$$

The agent now solves a three-dimensional optimization problem to select the design size which maximizes the difference between the surplus accuracy and the data collection cost:

$$(n^*, t^*, p^*) = \arg\max_{n,t,p} \left\{ A(n, t, p) - \left(C_n(n) + C_t(t) + C_p(p)\right) \right\},$$

which, we will show, always exist. Given the ***optimal design size*** (n^*, t^*, p^*), the ***optimal design*** $(n^*, \mathcal{T}^*, \mathcal{P}^*)$ follows from

$$(\mathcal{T}^*, \mathcal{P}^*) \equiv \left(\tilde{\mathcal{T}}(n^*, t^*, p^*), \tilde{\mathcal{P}}(n^*, t^*, p^*)\right).$$

3 Inference

In this section we characterize the agent's choice of the optimal predictor for a given dataset design $(n, \mathcal{T}, \mathcal{P})$.

3.1 The Optimal Predictor

Before observing the dataset generating by the design, the agent can choose a predictor to specify how to use it to make predictions. Under quadratic loss, the optimal predictor is the posterior mean of the target variable, a well known Bayesian result that motivates the following lemma.

Lemma 1 (Optimal Predictor)**.** *The optimal predictor is*

$$\hat{y}^*(\boldsymbol{D}) = \mathbb{E}[y \mid \boldsymbol{D}] = \boldsymbol{x}'_{\mathcal{P}} \mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{P}} \mid (\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})\right].$$

Because the target variables are independent across individuals conditional on $\boldsymbol{\beta}$, the training data affect the optimal predictor only through the posterior distribution of $\boldsymbol{\beta}$. Furthermore, since covariates are mutually independent and independent from $\boldsymbol{\beta}$, the agent does not condition her predictions on the unobserved covariates of the target individual (those in $\mathbb{N} \setminus \mathcal{P}$), because their prior mean is zero.

Lemma 1 has a straightforward consequence for the value of a dataset design, which will depend on the training information on each parameter, which is thus defined:

Definition 4. The ***(expected) training information on parameter*** k is

$$\tau_k^2(n, \mathcal{T}) \equiv \underbrace{\text{Var}\left[\beta_k\right]}_{=1} - \mathbb{E}_{\mathscr{P}_{(y,X_\mathcal{T})}(n,\mathcal{T})}\left[\text{Var}\left[\beta_k \mid (\boldsymbol{y}, \boldsymbol{X}_\mathcal{T})\right]\right].$$

The training information is the expected reduction of uncertainty on parameter k, which is fixed for the design $(n, \mathcal{T}, \mathcal{P})$. Armed with Definition 4, we can use Lemma 1 to characterize the value of the design $(n, \mathcal{T}, \mathcal{P})$.

Lemma 2. *The value of a dataset design* $(n, \mathcal{T}, \mathcal{P})$ *is*

$$V(n, \mathcal{T}, \mathcal{P}) = \text{Var}_{\mathscr{P}_D(n,\mathcal{T},\mathcal{P})}\left[\hat{y}^*(\boldsymbol{D})\right] = \sum_{k\in\mathcal{P}} \sigma_k^2 \tau_k^2(n, \mathcal{T}).$$

The value of a dataset design is the ex-ante variance of the optimal predictor $\hat{y}^*(\boldsymbol{D})$, taken over the prior distribution of the dataset $\mathscr{P}_D(n, \mathcal{T}, \mathcal{P})$. Intuitively, the variance of the optimal predictor reflects its sensitivity to the realizations of the dataset. Therefore, data designs which yield optimal predictors that "pick up" a large fraction of the variance in the dataset have a higher value. Furthermore, the variance of the optimal predictor is the linear combination of the prediction covariate variance σ_k^2 and the training information on its parameter τ_k^2. As the training information depends on the expected posterior variance, we will characterize the posterior distribution in the following section.

3.2 Posterior Distribution

To characterize the posterior distribution, observe that because individuals are independent conditionally on $\boldsymbol{\beta}$ and covariates are independent, knowing the prediction covariates but not the realization of the target variable gives no information on $\boldsymbol{\beta}$. Hence, the posterior will only depend on the training data $(\boldsymbol{y}, \boldsymbol{X}_\mathcal{T})$.

Recall that, for each individual i,

$$y^i = \sum_{k\in\mathbb{N}} \beta_k x_k^i = \sum_{k\in\mathcal{T}} \beta_k x_k^i + \sum_{k\in\mathbb{N}\setminus\mathcal{T}} \beta_k x_k^i,$$

so that, in matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}_\mathcal{T}\boldsymbol{\beta}_\mathcal{T} + \boldsymbol{\varepsilon}_\mathcal{T}, \qquad \boldsymbol{\varepsilon}_\mathcal{T} \equiv \boldsymbol{X}_{\mathcal{T}^c}\boldsymbol{\beta}_{\mathcal{T}^c}. \tag{1}$$

Because covariates are mutually independent across k and i.i.d. across individuals, and independent of $\boldsymbol{\beta}$, the noise term $\boldsymbol{\varepsilon}_\mathcal{T}$ is independent of $\boldsymbol{X}_\mathcal{T}$ conditional on $\boldsymbol{\beta}$. Moreover, for each fixed parameter vector $\boldsymbol{\beta}$,

$$\varepsilon_\mathcal{T}^i = \sum_{k\in\mathbb{N}\setminus\mathcal{T}} \beta_k x_k^i \mid \boldsymbol{\beta} \sim \mathcal{N}\Big(0,\, v(\boldsymbol{\beta})\Big), \qquad v(\boldsymbol{\beta}) \equiv \sum_{k\in\mathbb{N}\setminus\mathcal{T}} \sigma_k^2 \beta_k^2,$$

and the $\varepsilon^i_{\mathcal{T}}$ are independent across i. Hence, the exact likelihood under the data-generating process in Section 2.1 is

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{X}_{\mathcal{T}} \sim \mathcal{N}\Big(\boldsymbol{X}_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}}, v(\boldsymbol{\beta}_{\mathcal{T}^c})\boldsymbol{I}_n\Big), \qquad v(\boldsymbol{\beta}_{\mathcal{T}^c}) = \sum_{k\in\mathbb{N}\setminus\mathcal{T}} \sigma_k^2\beta_k^2. \tag{2}$$

This differs from the standard conjugate Bayesian linear regression (e.g., see DeGroot (2005) and Berger (1990)), in which the error variance is a fixed constant: here, the variance $v(\boldsymbol{\beta})$ is a nuisance parameter that depends on $\boldsymbol{\beta}_{\mathcal{T}^c}$, so the same parameter vector $\boldsymbol{\beta}$ appears in both the mean and the variance of the likelihood. As a result, the Gaussian prior in Section 2.1 is no longer conjugate to the exact likelihood (2).

Tail variance and moment matching. The following result shows that the residual variance is almost surely deterministic for most covariate sets.

Lemma 3. *For any nested sequence $(\mathcal{T}_m)_{m\geq 1}$ with $\mathcal{T}_m \subset \mathcal{T}_{m+1}$ and $\bigcup_m \mathcal{T}_m = \mathcal{T}$,*

$$\sum_{k\in\mathbb{N}\setminus\mathcal{T}_m} \sigma_k^2\beta_k^2 \xrightarrow[m\to\infty]{a.s.} \mathrm{Var}(\boldsymbol{\varepsilon}_{\mathcal{T}}) = 1 - S(\mathcal{T}_m). \tag{3}$$

For almost every draw of $\boldsymbol{\beta}$ from the prior, the random tail variance $v(\boldsymbol{\beta})$ concentrates around the deterministic tail sum $1-S$ along nested truncations of the training covariate set, i.e. the impact of the residual omitted covariates on y is a white noise with variance $1-S$. This shows that the random tail variance typically lies very close to its deterministic counterpart $1-S$ and provides a natural justification for treating the omitted covariates as homoskedastic noise. As the cumulative variance of omitted covariates is fixed by the design, we will denote it by at

$$S \equiv S(\mathcal{T}).$$

Working Gaussian likelihood. Motivated by Lemma 3, I approximate the exact likelihood (2) by a homoskedastic Gaussian likelihood that matches the conditional mean and the unconditional variance of the data-generating process,

Assumption 1 (Working Gaussian likelihood)**.** *For the purpose of inference, the agent uses the working likelihood*

$$\boldsymbol{y} \mid \boldsymbol{\beta}_{\mathcal{T}}, \boldsymbol{X}_{\mathcal{T}} \sim \mathcal{N}\Big(\boldsymbol{X}_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}}, \big(1-S\big)\boldsymbol{I}_n\Big).$$

This working likelihood coincides with the exact likelihood (2) when $v(\boldsymbol{\beta}) = 1 - S$, and Lemma 3 shows that the discrepancy between $v(\boldsymbol{\beta})$ and $1 - S$ vanishes almost surely along nested refinements of the training covariate set.

Under Assumption 1, the posterior distribution of $\boldsymbol{\beta}_{\mathcal{T}}$ is Gaussian, see DeGroot (2005) and Berger (1990).We will state the classic result in terms of a key variable:

Definition 5. The ***misspecification penalty*** is

$$\lambda \equiv \frac{1-S}{n}.$$

Intuitively, this captures the fraction of variability in the target variable which is due to the regression residual rather than $\boldsymbol{\beta}_{\mathcal{T}}$ rescaled by the sample size n, which the agent uses to average out that residual.

Proposition 1 (Posterior Distribution)**.** *The posterior distribution of the parameter vector is given by:*

- *For untrained parameters,*

$$\boldsymbol{\beta}_{\mathcal{T}^c} \mid (\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}}) \sim \mathcal{N}\left(\mathbf{0}_{|\mathcal{T}^c|}, \boldsymbol{I}_{|\mathcal{T}^c|}\right);$$

- *For trained parameters ,*

$$\boldsymbol{\beta}_{\mathcal{T}} \mid (\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}}) \sim \mathcal{N}\Bigg(\underbrace{\left(\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}} + n\lambda \cdot \boldsymbol{I}_{|\mathcal{T}|}\right)^{-1} \boldsymbol{X}_{\mathcal{T}}'\boldsymbol{y}}_{\equiv \mathbb{E}[\boldsymbol{\beta}_{\mathcal{T}} \mid (\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})]}, \quad \underbrace{\left(\frac{1}{n\lambda}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}} + \boldsymbol{I}_t\right)^{-1}}_{\equiv \mathrm{Var}[\boldsymbol{\beta}_{\mathcal{T}} \mid (\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})]} \Bigg),$$

  with $\boldsymbol{\beta}_{\mathcal{T}}$ independent from $\boldsymbol{\beta}_{\mathcal{T}^c}$.

Because parameters are independent, training $\boldsymbol{\beta}_{\mathcal{T}}$ provides no information on the untrained parameters $\boldsymbol{\beta}_{\mathcal{T}^c}$, whose prior mean is $\mathbf{0}$. The posterior mean of the trained parameters lies between the OLS estimator $\left(\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}}\right)^{-1}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{y}$ and the prior mean $\mathbf{0}_{|\mathcal{T}|}$, and λ reflects the relative weight given to the prior mean. Intuitively, if $\lambda = 0$, then all the variation in the training data is due to the regression parameters $\boldsymbol{\beta}_{\mathcal{T}}$ (i.e., the regression is perfectly specified). Therefore, the posterior mean coincides with the OLS estimator, which is most sensitive to the randomness in the training data, because it maximizes the likelihood. Instead, if $\lambda \to \infty$, then all the variation in the training data is due to the residual (i.e., it is totally uninformative concerning $\boldsymbol{\beta}_{\mathcal{T}}$) so the posterior mean is equal to the prior.

The rescaled penalty $n\lambda = 1 - S$ in the posterior mean is equivalent to the misspecification parameter of Strzalecki (2024). In contrast to his model, where λ reflects an exogenous concern for misspecification, my framework endogenizes λ as the weight a Bayesian agent assigns to the prior because of the omitted parameters. The agent's implicit discounting of the likelihood thus emerges endogenously from the dataset design, rather than exogenously because of a preference for robustness.

Proposition 1 allows us to characterize the training information for any parameter k.

Corollary 1. *The k-training information is*

$$\tau_k^2(\lambda; n, \mathcal{T}) = \left(1 - \mathbb{E}_{X_{\mathcal{T}} \sim \mathcal{P}_{X_{\mathcal{T}}}(n,\mathcal{T})}\left[\left(\frac{1}{n\lambda}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}} + \boldsymbol{I}_{|\mathcal{T}|}\right)^{-1}_{kk}\right]\right)\mathbf{1}(k \in \mathcal{T}).$$

The k-training information is decreasing in the misspecification penalty λ. Furthermore, it is increasing in σ_k^2 because the latter increases the expected variation of the training data in the space spanned by covariate k. However, it is also increasing in σ_j^2 with $j \in \mathcal{T} \setminus \{k\}$ because of higher-order interactions between covariates. Note that if there is no misspecification penalty, i.e., $\lambda = 0$, the agent learns parameter k perfectly and $\tau_k^2(0; n, \mathcal{T}) = 1$; conversely, if the training data is completely uninformative, i.e., $\lambda \to \infty$, the agent learns nothing on the parameters and $\lim_{\lambda\to\infty} \tau_k^2(\lambda; n, \mathcal{T}) = 0$.

4 Dataset Design

4.1 The Value of a Dataset Design

We characterize the value of a generic dataset design $(n, \mathcal{T}, \mathcal{P})$. To do so we use Lemma 2 and Corollary 1, endogenizing the misspecification penalty, because it now depends on the design though

$$\lambda(n, \mathcal{T}) \equiv \frac{1 - S(\mathcal{T})}{n}.$$

Proposition 2 (Value of Dataset Design)**.** *The value of a dataset design* $(n, \mathcal{T}, \mathcal{P})$ *is*

$$V(n, \mathcal{T}, \mathcal{P}) = \sum_{k \in \mathcal{P} \cap \mathcal{T}} \sigma_k^2 \tau_k^2(\lambda(n, \mathcal{T}); n, \mathcal{T}).$$

Intuitively, the value of a design is the linear combination of covariate variances σ_k^2 and the information on its parameter, $\tau_k^2(\lambda(n, \mathcal{T}); n, \mathcal{T})$. Note that adding a covariate j to $\mathcal{T} \setminus \mathcal{P}$ has two positive spillover effects on the value of a covariate $k \in \mathcal{P}$:

1. ***First-order spillovers,*** mediated by the misspecification penalty and deriving from the fact that $\tau_k^2(\lambda; n, \mathcal{T})$ is increasing in λ and that $\lambda(n, \mathcal{T}) < \lambda(n, \mathcal{T} \cup \{j\})$: the reduction in the regression noise from j; and

2. ***Higher-order spillovers,*** which derive from $\tau_k^2(\lambda; n, \mathcal{T} \cup \{j\}) > \tau_k^2(\lambda; n, \mathcal{T})$, meaning that the mixed moments of the empirical variance/covariance matrix help estimation.

Because of these higher-order spillovers, the values of covariates are connected in a complicated web which is analytically intractable. The following result shows that if the training size (i.e., the trained parameter count) grows slower than the sample size, the direct spillovers are negligible, yielding a tractable shrinkage factor approximation $\bar{\tau}_k^2(\lambda(n, \mathcal{T}))$.

Proposition 3 (Training Information Approximation)**.** *The following asymptotic approximation holds*

$$V(n, \mathcal{T}, \mathcal{P}) = \sum_{k \in \mathcal{P} \cap \mathcal{T}} \sigma_k^2 \bar{\tau}_k^2(\lambda(n, \mathcal{T})) + O\left(\sqrt{\tfrac{|\mathcal{T}|}{n}} + \tfrac{|\mathcal{T}|}{n}\right),$$

where

$$\bar{\tau}_k^2(\lambda) \equiv \frac{\sigma_k^2}{\sigma_k^2 + \lambda}.$$

When the training size is small relative to the sample size, all cross-covariate spillovers and dimensionality effects are fully captured by the misspecification penalty. This approximation is typically violated in modern machine-learning architectures such as neural networks or large language models, which feature far more parameters than training tokens and therefore exhibit substantial higher-order spillovers. Therefore, the following assumption amounts to taking a conservative stance on the magnitude of positive spillovers.

Assumption 2 (Low-dimensional Regime)**.** $\frac{|\mathcal{T}|}{n} \to 0$.

Under Assumption 2, the value of a design $(n, \mathcal{P}, \mathcal{T})$, becomes separable in the training covariates, conditional on λ, which absorbs all cross-covariate spillovers.[12] We are now ready to highlight four key properties which drive the returns to data:

- ***Training Covariate Spillovers***: the misspecification penalty $\lambda(n, \mathcal{T})$ is decreasing in $k \in \mathcal{T}$, so adding a training covariate improves estimates of all trained parameters:

$$-\Delta_{\mathcal{T}}^k \lambda(n, \mathcal{T}) \equiv -\left[\lambda(n, \mathcal{T}) - \lambda(n, \mathcal{T} \setminus \{k\})\right] = \frac{\sigma_k^2}{n} > 0;$$

- ***Sampling Spillovers:*** the misspecification penalty $\lambda(n, \mathcal{T})$ is decreasing in n, so adding an observation improves estimates of all trained parameters:

$$-\Delta_n \lambda(n, \mathcal{T}) \equiv -\left[\lambda(n, \mathcal{T}) - \lambda(n-1, \mathcal{T})\right] = \frac{1 - S(\mathcal{T})}{n(n-1)} > 0;$$

- ***Spillover Substitution:*** the sampling spillovers are decreasing in $k \in \mathcal{T}$ and covariate spillovers are decreasing in n:

$$|\Delta_n \lambda(n, \mathcal{T})| - |\Delta_n \lambda(n, \mathcal{T} \setminus \{k\})| = -\frac{\sigma_k^2}{n(n-1)} < 0.$$

- ***House Party Effect:*** the shrinkage factor $\bar{\tau}_k^2(\lambda)$ is convex in the misspecification penalty λ. This means that reductions in λ (which correspond to more precise estimation,

[12] Relaxing this assumption would require random-matrix methods to analyze the regime $t/n \to \gamma > 0$. However, closed-form solutions can then be obtained only under homoskedasticity, i.e. when all covariates have identical variance $\sigma_k^2 = \sigma^2 > 0$ for some fixed σ^2. Because heterogeneity in covariate variances is central to the problem of covariate selection, we maintain the small-t/n approximation.

or equivalently to larger datasets) generate *increasing gains* the value of data as each marginal reduction in noise improves predictive accuracy by more than the previous one:

$$\frac{\partial^2}{\partial \lambda^2} \bar{\tau}_k^2(\lambda) = \frac{2\sigma_k^2}{(\sigma_k^2 + \lambda)^3} > 0.$$

The name "House Party Effect" refers to a useful analogy: two people are having a conversation at a crowded house party. When there are one hundred other people in the room (analogous to one hundred residual covariates contributing to noise), removing five people (observing five covariates/observations) has virtually no effect on how easily the two people can talk (on the precision of estimates). As the crowd thins (as we observe more and more covariates/observations), however, each person who leaves has a stronger impact on the overall noise level, e.g., once only five other people remain, their departure dramatically improves audibility as communication becomes perfectly clear. The same logic applies to the residual variance in the regression model: due to the convexity of $\bar{\tau}_k^2(\lambda)$, when the residual noise λ is large, small reductions in λ have negligible impact; but when residual noise is small, additional reductions sharply improve accuracy. Note that since $\frac{\partial^3}{\partial \lambda^3} \bar{\tau}_k^2(\lambda) < 0$ the House Party Effect strengthens as λ decreases, i.e. $\frac{\partial^2}{\partial \lambda^2} \bar{\tau}_k^2(\lambda)$ increases as n becomes large and $\mathcal{T}$ expands.

4.2 Properties of the Value of a Dataset Design

We will now perform some comparative statics on the value of a dataset design to highlight the contribution of covariates and observations to training and prediction.

For any $\mathcal{P}$, define the ***contribution of prediction covariate*** $k \in \mathcal{P}$ as

$$\Delta_k^{\mathcal{P}}(n, \mathcal{P}, \mathcal{T}) \equiv V(n, \mathcal{P}, \mathcal{T}) - V(n, \mathcal{P} \setminus \{k\}, \mathcal{T}).$$

Proposition 4 (Contribution of Prediction Covariate)**.** *Fix* $(n, \mathcal{T})$*. For any* $\mathcal{P}$ *and any* $k \in \mathcal{P} \cap \mathcal{T}$*,*

$$\Delta_k^{\mathcal{P}}(n, \mathcal{P}, \mathcal{T}) = \sigma_k^2 \bar{\tau}_k^2 \left(\lambda(n, \mathcal{T})\right) > 0,$$

which is independent of $\mathcal{P} \setminus \{k\}$*. Furthermore, it is strictly increasing in* $\mathcal{T}$ *and* n*.*

This result has two consequences for the economics of prediction.

Corollary 2. *There are no spillover effects among training covariates.*

As prediction covariates enter in an additively separable way, the value each of them generates is independent from the collection other prediction covariate

Corollary 3. *Prediction covariates are complements to the training covariates and observations.*

Because of the training covariate spillovers and the sampling spillovers, collecting additional training data reduces the misspecification penalty $\lambda(n, \mathcal{T})$ for all the prediction covariates, meaning that there are complementarities between training and prediction covariates.

The complementarity between prediction covariates and observations matches the empirical results of Schaefer and Sapi (2023) and Lee and Wright (2023), which however do not make the distinction between training and prediction covariates. The findings are also coherent with Wilson (1975): better information can be leveraged across the entire scale of production, so the non-rival nature of information generates complementarities.

We next turn to the training covariates, define the ***contribution of training covariate*** $k \in \mathcal{T}$ as

$$\Delta_k^{\mathcal{T}}(n, \mathcal{P}, \mathcal{T}) \equiv V(n, \mathcal{P}, \mathcal{T}) - V(n, \mathcal{P}, \mathcal{T} \setminus \{k\}).$$

The following result characterizes the contribution of a training covariate.

Proposition 5 (Contribution of Training Covariate)**.** *Fix* $(n, \mathcal{P})$*. For any* $\mathcal{T}$ *and any* $k \in \mathcal{T} \setminus \mathcal{P}$*,*

$$\Delta_k^{\mathcal{T}}(n, \mathcal{P}, \mathcal{T}) = \sum_{m \in \mathcal{P} \cap \mathcal{T}} \sigma_m^2 \int_{-\frac{\sigma_k^2}{n}}^{0} -\partial_\lambda \bar{\tau}_m^2 \left(\lambda(n, \mathcal{T} \setminus \{k\}) + u\right) du > 0,$$

which is increasing in $\mathcal{T} \setminus \{k\}$ *and in* $\mathcal{P}$*.*

The monotonicity in $\mathcal{T} \setminus \{k\}$ follows immediately from the formula: the contribution $\Delta_k^{\mathcal{T}}$ is an integral over a fixed-length interval of u of the function $-\partial_\lambda \bar{\tau}_m^2(\lambda)$ evaluated at $\lambda(n, \mathcal{T} \setminus \{k\}) + u$. Since enlarging $\mathcal{T} \setminus \{k\}$ decreases $\lambda(n, \mathcal{T} \setminus \{k\})$ and $-\partial_\lambda \bar{\tau}_m^2(\lambda)$ is positive and strictly decreasing in λ, the integrand increases pointwise, and therefore the whole integral (and thus $\Delta_k^{\mathcal{T}}$) strictly increases. This finding is novel and has significant economic consequences for the value of data summarized in the following result.

Corollary 4. *There are economies of scope in training.*

This finding gives a theoretical explanation to the surprising convexity in the returns to covariates in the empirical analysis of Carballa-Smichowski et al. (2025b). When sample sizes are fixed, data ownership has a natural tendency to become concentrated: if two agents were to participate in a second-price auction for an additional covariate, the agent with more covariates would have a higher marginal value for that covariate and would therefore acquire it, increasing the asymmetry in data ownership.

Proposition 6 (Complementarity/Substitutability in Training)**.** *Fix* $(n, \mathcal{P})$*. For any* $\mathcal{T}$ *and any* $k \in \mathcal{T} \setminus \mathcal{P}$*,*

$$\frac{d}{dn}\Delta_k^{\mathcal{T}}(n, \mathcal{P}, \mathcal{T}) = \frac{\sigma_k^2}{n^2} \sum_{m \in \mathcal{P} \cap \mathcal{T}} \sigma_m^2 \Bigg(\underbrace{\frac{1 - S(\mathcal{T} \setminus \{k\})}{\sigma_k^2} \int_{-\sigma_k^2/n}^{0} \partial_{\lambda^2} \bar{\tau}_k^2(\lambda(n, \mathcal{T} \setminus \{k\}) + u) du}_{(+)\ \text{HPE} \times \text{Sampling spillovers}} + \underbrace{\partial_\lambda \bar{\tau}_m^2 \left(\lambda(n, \mathcal{T})\right)}_{(-)\ \text{Spillover substitution}} \Bigg) > 0$$

if and only if

$$n < \tilde{n}(\mathcal{T}, \mathcal{P}) \in (0, \infty),$$

which is decreasing in $\mathcal{T}\setminus\{k\}$ and $\mathcal{P}$.

Corollary 5. *Training covariates and observations are complements when they are scarce and substitutes when they are abundant.*

There are two opposite forces: on the one hand, the House Party Effect implies an additional observation raises the value of an additional training covariate, because it increases the marginal value of noise reduction; on the other hand, by spillover substitution the reduction in misspecification penalty brought forth by an additional training covariate is decreasing in the sample size. Both effects strengthen as n and $\mathcal{T}$ expand; initially, the House Party Effect dominates, therefore, training data dimensions are complements, but eventually the spillover substitution will prevail and they become substitutes. This implies that the complementarity between covariates and observations empirically documented in Schaefer and Sapi (2023) and Lee and Wright (2023) may not hold when datasets are large.

We now analyze the impact of a larger sample on the value of an additional observation. For any $n \geq 1$, the ***marginal value of the n-th observation*** is

$$\Delta^n(n, \mathcal{P}, \mathcal{T}) \equiv V(n, \mathcal{P}, \mathcal{T}) - V(n-1, \mathcal{P}, \mathcal{T}).$$

Proposition 7 (Marginal Value of Observation)**.** *Fix $(\mathcal{T}, \mathcal{P})$. For any $\mathcal{T}$ and any $k \in \mathcal{T} \setminus \mathcal{P}$,*

$$\Delta^n(n, \mathcal{P}, \mathcal{T}) = \sum_{m \in \mathcal{P}\cap\mathcal{T}} \sigma_m^2 \int_{-\frac{\lambda(n-1,\mathcal{T})}{n}}^{0} -\partial_\lambda \bar{\tau}_m^2 \left(\lambda(n-1, \mathcal{T}) + u\right) du,$$

which is decreasing in n, increasing in $\mathcal{P}$ and inverted U-shaped in $\mathcal{T}$.

Interestingly, there are two opposite forces driving the marginal value of an observation: on the one hand, the House Party Effect makes additional reductions in misspecification penalty λ more valuable by decreasing the argument of $\partial_\lambda \bar{\tau}_m^2$; on the other, the percentage reduction in λ brought forth by an additional observation is rapidly diminishing, which is reflected in the fact that the length of the integration interval is proportional to $1/n$, an instance of the Law of Large Numbers. The Law of Large Numbers always dominates the House Party Effect, so each additional observation improves predictions less and less.

Corollary 6. *There are diminishing returns to the sample size.*

This finding is consistent with Goldfarb and Tucker (2011), Bajari et al. (2019) and Schaefer and Sapi (2023). When covariate sets are fixed, data ownership has a natural tendency to become more distributed: if two agents were to participate in a second-price auction for an additional observation, the agent with a smaller sample would have a higher marginal value for that observation and would therefore acquire it, reducing the asymmetry in data ownership.

4.3 Constrained Covariate Selection

We will now characterize the optimal dataset design $(n, \mathcal{P}, \mathcal{T})$ under Assumption 2, with an additively separable data collection cost:

$$C_n(n) + C_t(t) + C_p(p) + F.$$

We will therefore solve the Covariate Selection Problem and the Accuracy Production Problem.

Proposition 8 (Optimal Covariate Selection)**.** *The optimal constrained covariate sets are*

$$\tilde{\mathcal{T}}(t) = \bigcup_{k=1}^{t} \sigma_k^2, \qquad ,\tilde{\mathcal{P}}(p) = \bigcup_{k=1}^{p} \sigma_k^2.$$

Recall that we ordered covariates in a sequence of decreasing variance $\{\sigma_k^2\}_{k\in\mathbb{N}}$, so the agent collects the most variable covariates first because $V(n, \mathcal{T}, \mathcal{P})$ is increasing in all σ_k^2, directly for $k \in \mathcal{P}$ and through $\lambda(n, \mathcal{T})$ for $k \in \mathcal{T}$.

We will henceforth abuse notation and denote the misspecification penalty and the cumulative variance of the constrained covariate sets as

$$\lambda(n, t) \equiv \frac{1 - S(t)}{n}, \qquad \text{and} \qquad S(t) \equiv \sum_{k=1}^{t} \sigma_k^2.$$

Using these definitions, we can characterize the accuracy of a dataset design of a given size.

Proposition 9 (Accuracy)**.** *Under Assumption 2, the accuracy of a data design of size* (n, t, p) *is*

$$A(n, p, t) = \sum_{k=1}^{p} \sigma_k^2 \bar{\tau}_k^2 \left(\lambda(n, t)\right).$$

The accuracy reflects the interplay between the economies of scope and the diminishing returns deriving from the optimal covariate selection. The latter is the statistical equivalent of the Law of Diminishing Returns in Ricardo (1817): the agent treats covariates as production factors of heterogeneous quality and each additional covariate marginal productivity (variance) falls as one moves down the ordered list. Therefore there is a ***covariate selection effect*** which implies that $S(t)$ is concave in t. Consequently,

$$\frac{\partial^2}{\partial t^2}\lambda(n, t) < 0$$

meaning that collecting training covariates of higher index (lower variance) brings forth smaller reductions of the misspecification penalty.

4.4 Economies of Scale in Accuracy

Define the ***marginal value of the p-th prediction covariate*** as

$$\Delta^p(n, p, t) \equiv A(n, p, t) - A(n, p-1, t).$$

Proposition 10 (Marginal Value of Prediction Size)**.** *Fix (n, t). For any $p \in [1, t]$,*

$$\Delta^p(n, p, t) = \sigma_p^2 \bar{\tau}_p^2 \left(\lambda^*(n, t)\right) > 0,$$

which is decreasing in p.

Given that the contribution of each covariate is independent from that of any other, only the selection effect is present and therefore the marginal value of the p-th prediction covariate is decreasing.

Corollary 7. *There are diminishing returns to prediction covariate size.*

Similarly, define the ***marginal value of the t-th training covariate*** as

$$\Delta^t(n, p, t) \equiv A(n, p, t) - A(n, p, t-1).$$

Proposition 11 (Marginal Value of Training Size)**.** *Fix (n, p). For any $t > p$,*

$$\Delta^t(n, p, t) = \sum_{m \in \mathcal{P} \cap \mathcal{T}} \sigma_m^2 \int_{-\frac{\sigma_t^2}{n}}^{0} -\partial_\lambda \bar{\tau}_m^2 \left(\lambda(n, t-1) + u\right) du.$$

Now increasing t decreases the interval of integration by $\frac{\sigma_{t+1}^2 - \sigma_t^2}{n}$. Furthermore, λ is convex decreasing in t, so higher ts reduce the misspecification penalty less. These two effects counter the House Party Effect and may induce diminishing returns to covariates if the selection effects are large enough, i.e. if the variance is concentrated in few covariates.

In general, the expression of the accuracy for an arbitrary distribution of variance across covariates σ_k^2 will be complicated. We will focus on the specific case which yields closed form accuracies.

Homoskedastic Covariates Assume there is a set $\mathcal{I} \subset \mathbb{N}$ such that covariates are informative if and only if they are in $\mathcal{I}$,

$$\sigma_k^2 = \sigma^2 \mathbf{1}(k \in \mathcal{I}),$$

so under the normalization $\text{Var}[y] = 1$, $|\mathcal{I}| = 1/\sigma^2$. Under optimal covariates selection,

the agent will observe the informative ones first, so the cumulative variance is

$$S(k) = \min\{k\sigma^2, 1\}.$$

Proposition 12 (Closed Form Accuracy). *If $\sigma_k^2 = \sigma^2 \mathbf{1}(k \in \mathcal{I})$, the accuracy is*

$$A(n, p, t) = \frac{S(\min\{t, p\})}{\frac{1-S(t)}{n\sigma^2} + 1}.$$

We can perform some simple comparative statics.

Corollary 8. $A_{tt}(n, p, t) > 0$.

This implies that there are economies of scale to the training covariate size since there are no selection effects.

Proposition 13. *Provided $t \geq p$,*

$$A_{tn}(n, p, t) > 0 \iff n \leq \hat{n}(t) \equiv \frac{1 - S(t)}{\sigma^2}.$$

Since $\hat{n}(k)$ is decreasing in k, covariates and observations are complements in accuracy when they are scarce and substitutes when they are abundant, as we had uncovered in Proposition 6.

4.5 Comparative Statics

Assuming that an interior solution exists the comparative statics will be given by the substitution/complementarities I have highlighted in this section.

An increase in the cost of prediction covariates $C_p(\cdot)$ will reduce the amount of training covariates and observations as prediction and training are complements.

An increase in the cost of training covariates $C_t(\cdot)$ reduces the optimal number of training covariates and, via the optimality condition linking training and prediction scope, lowers the number of prediction covariates, thereby raising the marginal variance of the last prediction covariate purchased. Moreover, firms with few training covariates and observations —operating in the region where t and n are complementary — optimally cut back on both margins, while firms with abundant t and n — operating in the region where they are substitutes — will reduce training covariates but compensate cost shock on t by expanding their collection of observations.

An increase in the cost of observations $C_n(\cdot)$ has analogous effects: it lowers the optimal number of observations and prediction covariates and reshapes training choices. Firms with scarce t and n, where observations and training covariates are complementary, cut both margins; firms with abundant t and n, where they are substitutes, shift towards relatively more training covariates.

Therefore increases in training costs in the form of $C_n(\cdot)$ or $C_t(\cdot)$, amplify disparities in data collection between data-rich and data-poor agents.

5 Frequentist Interpretation

The reader who is not interested in the statistical properties of the model can safely skip to Section 6.

Ridge Regression Interpretation To give an intuitive interpretation of the results in Section 3, we will draw a connection to a common estimator in frequentist statistics, the ridge estimator. It is well known in the Bayesian statistics literature that optimal estimators of linear models, like that in Proposition 1, have a frequentist counterpart in the ridge estimator defined as:

$$\hat{\boldsymbol{\beta}}_{\mathcal{T}}^{\text{ridge}}\left(\xi;[\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}}]\right) \equiv \arg\min_{\boldsymbol{b}\in\mathbb{R}^t}\left\{\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{T}}\boldsymbol{b}\|_2^2+\xi\|\boldsymbol{b}\|_2^2\right\},$$

where $\xi \geq 0$ is the ridge parameter which penalizes the squared Euclidean distance of $\boldsymbol{b}$ from the origin.[13] Intuitively, the ridge estimator is a form of shrinkage estimator which is used to stabilize parameter estimates when maximum likelihood methods have too great variance. The optimal choice of ξ is the minimizer of the *expected mean squared error (MSE)* as

$$\xi^*(n,\mathcal{T}) \equiv \arg\min_{\xi}\mathbb{E}_{\boldsymbol{M}_{\mathcal{T}}^n}\left[\left\|\hat{\boldsymbol{\beta}}_{\mathcal{T}}^{\text{ridge}}\left(\xi;\boldsymbol{M}_{\mathcal{T}}^n\right)-\boldsymbol{\beta}_{\mathcal{T}}\right\|_2^2\right],$$

where the expectation is taken with respect to the prior distribution of the training data $\boldsymbol{M}_{\mathcal{T}}^n$. We can define the *optimal ridge estimator* as

$$\hat{\boldsymbol{\beta}}_{\mathcal{T}}^*\left(n,\mathcal{T};\boldsymbol{M}_{\mathcal{T}}^n\right) \equiv \hat{\boldsymbol{\beta}}_{\mathcal{T}}^{\text{ridge}}\left(\left(\xi^*(n,\mathcal{T})\right);\boldsymbol{M}_{\mathcal{T}}^n\right).$$

The following result characterizes the optimal ridge parameter ξ^*.

Proposition 14 (Ridge Estimator Interpretation)**.** *The frequentist optimal ridge parameter is the misspecification penalty*

$$\xi^*(n,\mathcal{T}) = \lambda(n,\mathcal{T}) \equiv \frac{1-S(\mathcal{T})}{n}.$$

This implies that there is an equivalence between the posterior mean parameters and the a priori optimal ridge estimator. The posterior mean parameters and the optimal ridge estimator are equivalent

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}^n\right] = \hat{\boldsymbol{\beta}}_{\mathcal{T}}^*\left(n,\mathcal{T};\boldsymbol{M}_{\mathcal{T}}^n\right)$$

[13] See DeGroot (2005) and Berger (1990).

This result is coherent with the implied regularization in a Bayesian linear model characterized in Lindley and Smith (1972). Corollary 5 implies that a Bayesian agent acts like frequentist agent would if she treated the Bayesian prior as the true DGP and selected the ridge penalty that minimizes expected MSE across all possible samples generated from the prior distribution. Therefore, a Bayesian agent endogenously solves a bias–variance trade-off ex ante.

Double Descent Since regularization applies uniformly to all parameters, the aggregate strength of regularization is defined as

$$\Lambda^* \equiv |\mathcal{T}|\lambda^*.$$

Conventional statistical intuition suggests that Λ^* should increase if we add a covariate to $\mathcal{T}$: richer models require stronger regularization to curb estimator variance. However, recent work by Belkin et al. (2019) and Nakkiran et al. (2021) have shown that when t becomes large, the optimal level of overall regularization Λ^* may actually decline—a phenomenon known as double descent. This finding has drawn significant attention as one of the central puzzles in modern machine learning theory. The following result provides a simple and intuitive explanation for why Λ^* can decrease if we add a covariate to the model when $|\mathcal{T}|$ is large.

$$\Lambda^*(n, \mathcal{T} \cup \{k\}) \leq \Lambda^*(n, \mathcal{T}) \iff |\mathcal{T}| \times \sigma_k^2 \geq 1 - S(\mathcal{T} \cup \{k\}),$$

Adding a covariate k to the training set $\mathcal{T}$ reduces the overall rate of regularization Λ^* if and only if dimensionality $|\mathcal{T}|$ and covariate variance σ_k^2 are large relative to the population regression noise $1 - S(\mathcal{T} \cup \{k\})$. Adding a covariate affects the optimal rate of regularization through two opposing forces: an *intensive margin* (positive effect): the inclusion of one additional parameter mechanically increases the total regularization by $\lambda(n, \mathcal{T} \cup \{k\}) = [1-S(\mathcal{T} \cup \{k\})]/n$; an *extensive margin* (negative effect): expanding the set of covariates reduces the optimal regularization applied to each parameter, captured by

$$-t\left[\lambda^*(n, \mathcal{T}) - \lambda^*(n, \mathcal{T} \cup \{k\})\right] = -\frac{|\mathcal{T}|\sigma_k^2}{n} < 0.$$

When the latter prevails adding covariates k to a regression reduces the optimal regularization Λ^*. This decomposition offers a particularly transparent interpretation of double descent—arguably simpler than existing explanations such as Hastie et al. (2020) —and, to the best of my knowledge, the only one that does not rely on high-dimensional asymptotics.

6 Applications

We will first derive some flexible closed forms for the accuracy production function.

6.1 Natural Monopsony and Data-driven Acquisitions

6.1.1 Setup

Demand. For each prediction seller there is a unit mass of captive prediction buyers.[14] Each buyer must choose an action $\hat{y} \in \mathbb{R}$, while the payoff-relevant state is $y \in \mathbb{R}$. Payoffs depend on the squared error $(\hat{y} - y)^2$.

Buyers share a common prior on y as specified in Section 2.1. If they do not buy any prediction, they choose the prior mean and obtain expected utility

$$\bar{u} = -1,$$

using the normalization $\mathrm{Var}(y) = 1$.

Now consider a seller i who, using a dataset $\boldsymbol{D}_i$ defined in Section 2.2, supplies an optimal prediction $\hat{y}_i^*(\boldsymbol{D}_i)$ characterized in Lemma 1. Let $V_i \in [0, 1]$ denote the value of seller i's dataset design, defined as the reduction in expected loss relative to the prior:

$$V_i \equiv 1 - \mathbb{E}\left[(\hat{y}_i^*(\boldsymbol{D}_i) - y)^2\right].$$

If a buyer purchases from seller i at price p_i, her *net* surplus relative to the outside option is

$$u(p_i, V_i) = V_i - p_i.$$

Thus V_i is the buyers' willingness-to-pay for seller i's prediction: each buyer strictly prefers i whenever $p_i < V_i$. We summarize individual and competitive demand by a reduced-form function $D(p_i, V_i)$,which is increasing in V_i and decreasing in p_i.

Supply. There are two potential sellers: an incumbent I ("Big Tech") and an entrant E ("Fitbit"). Both observe the same sample size n. Seller $i \in \{I, E\}$ has access to a set $\mathcal{T}_i$ of training and $\mathcal{P}_i \subseteq \mathcal{T}_i$ prediction covariates, with disjoint covariate sets $\mathcal{T}_I \cap \mathcal{K}_E = \varnothing$.

Given the dataset design value $V(n, \mathcal{T}, \mathcal{P})$ derived earlier, the standalone value of seller i's design is $V_i = V(n, \mathcal{T}_i, \mathcal{P}_i)$.

The entrant E can either: (i) remain active and sell its own predictions, or (ii) accept a take-it-or-leave-it acquisition offer P from I. If I acquires E, it gains access to E's covariates. The merged firm operates with sample size n and combined covariate sets, so post-merger design value is

$$V_{IE} = V(n, \mathcal{T}_I \cup \mathcal{K}_E, \mathcal{P}_I \cup \mathcal{P}_E).$$

Profits. Let $D(p, V)$ denote the demand faced by a monopolist offering a prediction with design value V at price p.

[14] I assume there is no downstream competition to abstract from the usual "killer acquisition" motive: acquisitions are solely motivated by the acquisition of data.

If I acquires E: Only the merged seller operates. Setting price p_{IE}, the incumbent's profit is

$$\Pi_I^{\text{acq}}(P, p_{IE}) = p_{IE} D\big(p_{IE}, V_{IE}\big) - P,$$

while the entrant receives the acquisition payment

$$\Pi_E^{\text{acq}}(P) = P.$$

If there is no acquisition: Both firms are active. As demands are not related because each firm has a unit mass of captive customers, let $D_i(p_i, V_i)$ be seller i's demands. Profits are

$$\Pi_i^{\text{no}}(p_i) = p_i D_i(p_i, V_i).$$

Planner welfare. The planner's objective is

$$W = \Pi_I + \Pi_E + CS - \xi\, \mathbf{1}\big(I \text{ acquires } E\big),$$

where CS is customer surplus and $\xi \geq 0$ is the social cost of reduced entry (e.g., loss of knowledge spillovers).

Timing. The game unfolds in two stages:

1. **Acquisition stage.** The incumbent I offers P to acquire E. The entrant accepts or rejects.
2. **Pricing stage.** If the acquisition occurs, the merged firm chooses p_{IE}. Otherwise, I and E simultaneously set (p_I, p_E).

Payoffs are realized at the end. The solution concept is subgame perfect Nash equilibrium, obtained by backward induction.

6.1.2 Buyer Purchase

Prediction Sale Buyers purchase the prediction if and only if $V \geq p$. The resulting demand function is

$$D(p, V) = \mathbf{1}\left(p \leq V\right).$$

The seller therefore sets

$$p^* = V$$

sells to all buyers, and obtains profit equal to the full informational value of its dataset:

$$\Pi_i^{\text{no}} = V_i.$$

Intuitively, with homogeneous buyers and no frictions, the platform can fully extract the gains from improved prediction.

Acquisition For the entrant E, the standalone outside option is to optimally price its own prediction, earning

$$\Pi_E^{\text{no}} = V_E.$$

To acquire E, it is sufficient that the incumbent I offer at this amount. When I acquires E, it combines covariates and gains an increase of value

$$V_{IE} - V_E$$

compared to its outside option V_E.

Proposition 15. *The incumbent I always acquires the entrant E.*

Intuitively, the economies of scope stemming from Propositions 10 and 11 imply that

$$V_{IE} - \sum_{i \in \{I,E\}} V_i \geq 0.$$

Combining disjoint datasets makes each covariate more informative. This is due to two things: first as far as prediction covariates in $\mathcal{P}_i$ can benefit from the covariate spillovers of the training covariates in $\mathcal{T}_{-i}$ (and vice versa); second, due to the House Party Effect, pushing training covariates from $\mathcal{T}_i$ to $\mathcal{T}_I \cup \mathcal{T}_E$ yields supermodular gains. As the incumbent can fully exploit these complementarities, even if demands are independent it will prevent entry by E. This is the natural-monopsony effect: data-driven economies of scope push data into the hands of a single firm.

6.1.3 Planner's Problem

With homogeneous buyers and full surplus extraction, customer surplus is zero in equilibrium. Social welfare therefore equals industry profit net of any cost of foregone entry and spillovers:

$$W = \begin{cases} \Pi_{IE} - \xi, & \text{if } I \text{ acquires } E, \\ \Pi_I + \Pi_E, & \text{otherwise,} \end{cases}$$

where $\xi \geq 0$ denotes the social cost of eliminating the entrant as an independent innovator (e.g. reduced experimentation, weaker knowledge diffusion, lower dynamic competition).

Proposition 16. *The acquisition is socially desirable if and only if*

$$\tilde{\xi}(n, \boldsymbol{\mathcal{K}}_I, \boldsymbol{\mathcal{K}}_E) \equiv V_{EI} - (V_I + V_E) \geq \xi,$$

To be able to have clsoed form solutions, assume all covariates are equally informative, and let $|\mathcal{K}_I| = k\iota$ and $|\mathcal{K}_E| = k(1-\iota)$ so ι is the fraction of overall covariates $k \in [0,1]$ owned by the incumbent. Using Proposition with $S(x) = x$, we get

$$\tilde{\xi}(n,k,\iota) = \frac{k}{\frac{1-k}{n}+1} - \left(\frac{\iota k}{\frac{1-\iota k}{n}+1} + \frac{(1-\iota)k}{\frac{1-(1-\iota)k}{n}+1} \right)$$

is increasing in k, inverted U-shaped in ι and increasing if and only if $\iota \leq 1/2$, inverted U-shaped in n and increasing if and only $n \leq \tilde{n}(\iota,k)$, which is decreasing in k and increasing in ι if and only if $\iota \leq 1/2$. The result highlights a tension between private and social incentives for acquisition. Privacy worsen this tension. For example, the EU's General Data Protection Regulation (GDPR) emphasizes privacy, but by raising compliance costs it lowers k, thereby dampening economies of scope and reducing the interval $[0, \tilde{\xi}(n,k,\iota)]$ in which the acquisition is socially desirable even though it kills the knowledge spillovers due to entry. Conversely, open-data initiatives raise k and broaden the interval in which economies of scope make acquisition desirable.

6.2 Data Pools

Data owners often form partnerships to pool their datasets and sell access jointly. For example, BMW, Mercedes-Benz, and Audi co-founded the platform *Here Mobility Data Marketplace*, which aggregates GPS, speed, and road-condition data from connected cars. Gu, Madio, and Reggiani (2021) study two data brokers selling datasets that can be either complements or substitutes, showing that pooling is neutral when datasets are complements but collusive when they are substitutes. We extend their analysis by developing a model in the spirit of Lerner and Tirole (2004), building on the fundamental *complements problem* originally noted by Cournot (1838): when complementary goods are sold by independent monopolists, the resulting double marginalization leads to inefficiently high prices.

6.2.1 Setup

Data Owners Consider two data owners, each holding a dataset of identical informativeness. The complete dataset consists of n observations and k covariates, with the training and prediction covariates coinciding. However, the data may be split between the two owners either along the n (sample size) dimension or along the k (covariate) dimension. All parties are symmetrically informed about the informativeness of each dataset.

If the data are split along the sample dimension, each owner holds n observations and all k covariates. Pooling thus doubles the sample size. Conversely, if the data are split along the covariate dimension, each owner holds all n observations but only half the covariates, so pooling expands the covariate space. These two cases capture two distinct sources of comple-

mentarity: statistical precision (more n) and informational richness (more k).

Data Buyers There is a continuum of potential buyers (e.g., prediction firms) who can purchase access to one or both datasets and combine them without cost. Buyers are heterogeneous and indexed by $\theta \in [\underline{\theta}, \bar{\theta}]$, representing their adoption cost or opportunity cost of using the prediction technology.

A buyer of type θ who purchases access to $q \in \{1, 2\}$ datasets obtains a gross surplus

$$U_q = A_q - \theta,$$

where A_q denotes the predictive value of having access to q datasets. Specifically,

$$A_q \equiv \begin{cases} A\left(\frac{2n}{q},\, k\right), & \text{if the data are split along the } n \text{ dimension,} \\ A\left(n,\, \frac{2k}{q}\right), & \text{if the data are split along the } k \text{ dimension.} \end{cases}$$

Since $A(n, k)$ is increasing in both arguments, combining datasets strictly improves predictive accuracy. Intuitively, pooling along the n dimension increases the number of observations available for training, which reduces estimation error, whereas pooling along the k dimension increases the number of predictive features, which broadens the scope of prediction.

The heterogeneity parameter θ is distributed according to

$$G(\theta) = \theta^{\alpha}, \qquad \alpha \in [0, 1],$$

implying a strictly increasing hazard rate $\frac{g(\theta)}{1-G(\theta)}$ with $g = G'$. The parameter α thus governs the curvature of demand: when α is low, heterogeneity is large and demand is relatively inelastic; when α is high, buyers are more homogeneous and demand becomes more elastic.

Given a price P for access to a bundle of $q \in \{1, 2\}$ datasets, only buyers with $A_q - \theta \geq P$ make the purchase. Hence, the corresponding demand function is

$$D(P; A_q) = \Pr(A_q - \theta \geq P) = (A_q - P)^{\alpha}.$$

This formulation implies that the elasticity of demand increases in α: higher α corresponds to a market in which adoption falls faster as price rises. Equivalently, α can be interpreted as the *semi-elasticity of demand*, describing how responsive adoption is to changes in price.

6.2.2 Pooling Price

When data owners form a pool, they coordinate pricing and behave as a single monopolist offering a bundled dataset. This situation mirrors the *pool pricing benchmark* in Section I.b of Lerner and Tirole (2004), where a patent pool chooses the package price P to maximize joint

revenue given demand $D(P - V_2)$. The logic is analogous here: the pooled data bundle yields predictive value V_2, and all buyers face a single posted price P.

Lemma 4 (Optimal Pool Price)**.** *The optimal pool price is*

$$P^* = \frac{A_2}{\alpha + 1}.$$

The pool acts as a monopolist setting the joint-access price that equates marginal revenue to zero, just as in Equation (1) in Lerner and Tirole (2004). The resulting price is decreasing in the demand semi-elasticity α: when α is high, buyers are more sensitive to price, leading the pool to charge less. Conversely, when α is low, demand is inelastic, so the pool can extract a larger fraction of the total value V_2.

Economically, pooling internalizes the complementarity between the dataset, analogous to Cournot's double marginalization problem in complementary goods. When data are sold separately, each owner fails to account for the positive externality that lower prices have on the other's sales. By setting a single joint price, the pool eliminates this inefficiency and behaves as an integrated monopolist.

6.2.3 Fragmentation Price

Fragmented Sample We first characterize when pooling two distinct samples on the same covariates is procompetitive. Section II in Lerner and Tirole (2004) characterizes the unique symmetric equilibrium in the case in which the brokers do not form a pool.

Lemma 5 (Sample Fragmentation Price)**.** *If the brokers have different observations on the same covariates,*

$$p_i = \min\left\{A_2 - V_1, \frac{A_2}{2 + \alpha}\right\},$$

and the buyers will buy from both brokers.

Applying Proposition 1 in Lerner and Tirole (2004) directly yields the following characterization of when fragmented sample pooling is procompetitive.

Lemma 6 (Welfare of Observation Pooling)**.** *A pool of observations is procompetitive if and only if*

$$\alpha > \frac{n}{2(1 - k)}.$$

and the buyers will buy from both brokers.

Fragmented Covariate Sets We now characterize when pooling two distinct covariate sets on the same observations is procompetitive. Section II in Lerner and Tirole (2004) characterizes the unique symmetric equilibrium in the case in which the brokers do not form a pool.

Lemma 7 (Covariate Fragmentation Price)**.** *If the brokers have distinct covariates each B_i prices at*

$$p_i = \frac{A_2}{2+\alpha},$$

and the buyers will buy from both brokers.

Applying Proposition 1 in Lerner and Tirole (2004) directly yields the following result.

Proposition 17. *A pool of buyers with distinct covariates is always procompetitive*

Following Lerner and Tirole (2004), there may be asymmetric equilibria but they result in lower industry profit than the symmetric one so we focus on the symmetric equilibrium.

6.3 Data Exclusivity

Recent exclusive data-licensing agreements—such as Reddit's 2024 deals with OpenAI and Google—highlight broader concerns that proprietary access to datasets may distort competition in AI and prediction markets. Because data are non-rival, a data seller faces a dynamic commitment problem akin to that of a durable-good monopolist: once it has licensed the dataset to one firm, it is tempted to also license it to the rival, eroding the first buyer's advantage. As in Katz and Shapiro (1986) and Aghion and Bolton (1987), exclusivity can serve as a contractual commitment device that mitigates opportunism by the seller but introduces a welfare trade-off: it softens business-stealing while depressing investment by excluded firms when data and proprietary inputs are complements. The model predicts that profitable exclusivity may become socially undesirable when datasets are abundant and product-market rivalry is intense. In such settings, exclusivity amplifies incumbency advantages and can deter entry—offering a micro-foundation for regulatory scrutiny of data-sharing agreements such as the Reddit–OpenAI deal.

6.3.1 Setup

There are three players: a data seller S (e.g., a platform holding user-generated data) and two prediction firms F_1 and F_2. The firms compete to sell predictions to customers whose utility depends on the accuracy of the prediction.

Prediction Buyers. There is a unit mass of customers, divided into:

- a mass $s \in [0,1]$ of *shoppers*, who can buy from either firm;
- a mass $(1-s)/2$ of *captive customers* for each firm, who can only buy from that firm.

The parameter s captures the intensity of competition: when $s = 0$, all customers are captive and firms behave as local monopolists; when $s = 1$, all customers are shoppers and the market

is fully competitive. This structure captures the idea that data quality matters only in relative terms, since shoppers migrate toward the firm offering the more accurate prediction.

Customer utility from firm $i \in \{1, 2\}$ is

$$u_i = A_i - p_i,$$

where A_i denotes the accuracy of the prediction and p_i the price charged. Shoppers buy from the firm with highest net utility $A_i - p_i$; captives always buy from their incumbent firm.

Data Buyers. Each firm i can improve prediction accuracy through:

1. *Licensing* the seller's training dataset, paying a fee f (binary $\ell_i \in \{0, 1\}$);
2. *Proprietary Data Collection* (binary $r_i \in \{0, 1\}$), paying a fixed cost $c > 0$.

Prediction quality is

$$A_i = A(\ell_i, r_i) = \bar{A}\,\mathbf{1}(\ell_i r_i = 1), \qquad \bar{A} \equiv \sum_{k \in \mathcal{P} \cap \mathcal{T}} \sigma_k^2 \tau_k^2(\lambda(n, \mathcal{T})).$$

Pricing and Profits. Firms can price discriminate:

$$\pi_i = p_i^c \frac{1 - s}{2} + p_i^s s D_i^s, \qquad \Pi_i = \pi_i - c r_i - f \ell_i.$$

Data Seller. The seller chooses a uniform license fee f:

$$\Pi_S = f \sum_i \ell_i.$$

Social Planner. Total welfare:

$$W = \Pi_S + \Pi_1 + \Pi_2 + CS.$$

Timing. (1) S sets f; (2) firms choose (ℓ_i, r_i); (3) firms set prices; (4) realization. We solve by backward induction.

6.3.2 Prediction Pricing

Captives:

$$p_i^c = A_i.$$

Lemma 8 (Shopper-price equilibrium). *Fix (A_1, A_2). The unique trembling-hand-perfect equilibrium is*

$$p_i^s = (A_i - A_j)^+,$$

so the higher-accuracy firm extracts the full willingness to pay of shoppers.

Hence equilibrium revenue:

$$\pi(\ell_i, r_i; \ell_j, r_j) = \frac{1-s}{2} A_i + s(A_i - A_j)^+. \tag{??}$$

6.3.3 Data Collection

Assumption 3 (Low Cost). $c < \underline{c} \equiv \frac{1-s}{2}\bar{A}$.

Proposition 18. *Under Assumption 3,*

$$(r_1^*, r_2^*) = (\ell_1, \ell_2).$$

6.3.4 Data Pricing

Lemma 9. *Under Assumption 3,*

$$f_2 = \frac{1-s}{2}\bar{A} - c, \qquad f_1 = \frac{1-s}{2}\bar{A} + s\bar{A} - c.$$

Proposition 19. *Under Assumption 3, the seller prefers exclusivity iff*

$$f_1 > 2f_2 \iff c > c^* \equiv \frac{1-s}{2}\bar{A} - s\bar{A} = \frac{1-3s}{2}\bar{A}.$$

For $s > 1/3$, $c^* < 0$, so exclusivity is always preferred.

6.3.5 Planner's Problem

Proposition 20. *Under Assumption 3,*

$$W^e = \frac{1+s}{2}\bar{A} - c, \qquad W^{ne} = \bar{A} - 2c,$$

and

$$W^e < W^{ne} \iff c < \frac{1-s}{2}\bar{A}.$$

Corollary 9. *Harmful exclusivity occurs iff*

$$c \in \left(c^*, \underline{c}\right) = \left(\frac{1-s}{2}\bar{A} - s\bar{A}, \frac{1-s}{2}\bar{A}\right).$$

Its width is $s\bar{A}$.

6.3.6 What happens above $\underline{c}$?

Assumption 4 (High Cost). $c \in [\underline{c}, \bar{c}]$,

$$\underline{c} = \frac{1-s}{2}\bar{A}, \quad \bar{c} = \frac{1+s}{2}\bar{A}.$$

Lemma 10 (Best responses above $\underline{c}$). *If $\ell_1 + \ell_2 \geq 1$, then*

$$r_i^{BR} = \mathbf{1}\{\ell_i = 1,\ \ell_j + r_j \leq 1\}.$$

If both license, no pure NE exists; a symmetric mixed NE exists.

If both license, the mixed NE investment probability is

$$\xi^* = \frac{\bar{A}(s+1) - 2c}{2\bar{A}s}, \qquad \Pi_i^{mix} = 0.$$

Seller. Under exclusivity:

$$\Pi^{excl} = \frac{1+s}{2}\bar{A} - c, \qquad f_1 = \frac{1+s}{2}\bar{A} - c.$$

Under two licenses: each firm has $\Pi_i^{mix} = 0$ before fees, so the only feasible fee is:

$$f_2 = 0.$$

Thus exclusivity is strictly more profitable.

Social welfare. Under exclusivity:

$$W^e = \frac{1+s}{2}\bar{A} - c.$$

Under non-exclusivity (mixed equilibrium):

$$W^{ne} = \xi^2(\bar{A} - 2c) + 2\xi(1-\xi)\left(\frac{1+s}{2}\bar{A} - c\right) = \frac{c\left(\frac{c}{\bar{A}} - s - 1\right) + \bar{A}\left(\frac{s+1}{2}\right)^2}{s}.$$

One can verify that

$$W^e > W^{ne} \qquad \text{for } c \in (\underline{c}, \bar{c}).$$

Summary. In the low-cost region $c < \underline{c}$ the planner prefers non-exclusivity, while in the high-cost region $c \in [\underline{c}, \bar{c}]$ the planner prefers exclusivity. The seller prefers exclusivity when-

ever $c > c^* = \frac{1-3s}{2}\bar{A}$, so harmful exclusivity arises exactly when

$$c \in (c^*, \underline{c}) = \left(\frac{1-s}{2}\bar{A} - s\bar{A}, \frac{1-s}{2}\bar{A} \right).$$

7 Conclusion

This paper develops a general framework for understanding the value of data in prediction by explicitly modeling covariates. The analysis shows how economies of scope across covariates, interactions between covariates and observations, and complementarities between training and prediction can generate increasing returns, offering a microfoundation for the rich-get-richer effects often observed in data-driven markets.

These forces have direct implications for policy and strategy. Prediction technologies may display natural monopsony characteristics, as concentrating covariates within one firm can raise efficiency. Privacy regulation that fragments data supply may inadvertently reinforce monopsony power, creating a trilemma between privacy, competition, and efficiency. The framework also highlights that not all data pooling agreements are alike: pooling lists of users with the same covariates can be anticompetitive, where as pooling different covariates on similar users raise welfare by eliminating double marginalization. Exclusivity deals, such as those signed between AI labs and data providers, may profitably foreclose entry by depriving rivals of essential complements. For firms, the results imply that prediction entails substantial sunk costs: early on, investment should balance user acquisition and attribute enrichment, while specialization and integration become optimal at a larger scale.

More broadly, the analysis cautions against treating data as homogeneous. Policies promoting open data without regard to dataset composition may miss crucial efficiency margins, whereas access remedies such as FRAND-priced APIs or federated learning preserve economies of scope.

My work opens two natural avenues for future research. The first is empirical. I aim to develop a methodology to test my results on real datasets. While the existing empirical literature[15] provides partial support to my findings, it suffers from two limitations: (i) most studies focus on a single dataset, whereas uncovering general properties requires comparing multiple datasets along common dimensions; and (ii) no existing work systematically tests all the properties identified in my model. Once these empirical properties are validated, my framework could serve as the foundation for a practical formula for data valuation, in the spirit of the Black–Scholes–Merton formula for derivatives.[16] The second avenue is theoretical. Embedding my static model into a dynamic Wald sampling framework would allow me to microfound data-enabled learning and analyze when feedback loops generate convergent

[15] See Bajari et al. 2019; Schaefer and Sapi 2023; Lee and Wright 2023; Yoganarasimhan 2020; Carballa-Smichowski et al. 2025b

[16] See Black and Scholes 1973, Merton 1973

data-collection strategies versus when they diverge.

Finally, the framework invites a broader research agenda: in his seminal critique of central planning, Hayek 1945 emphasized that "knowledge... never exists in concentrated form but solely as the dispersed bits... which all the separate individuals possess". Today, users' online activity transforms such dispersed knowledge into datasets that can be centralized, recombined, and monetized. My analysis shows that statistical properties of prediction create intrinsic incentives for such concentration. The concentration of data in servers controlled by a few large firms raises a broader question: do prediction algorithms substitute for, or complement, the market mechanism? Is the rise of data the panacea to market failures deriving from asymmetric information and search frictions, or is it the first step to the fall of the market? I leave this foundational question open to future research.

References

Acemoğlu, Daron et al. (2022). "Too much data: Prices and inefficiencies in data markets." In: *American Economic Journal: Microeconomics* 14.4, pp. 218–256.

Aghion, Philippe and Patrick Bolton (1987). "Contracts as a Barrier to Entry." In: *The American Economic Review* 77.3, pp. 388–401.

Allcott, Hunt et al. (2025). *Sources of market power in web search: Evidence from a field experiment.* Tech. rep. National Bureau of Economic Research.

Aral, Sinan, Erik Brynjolfsson, and DJ Wu (2008). "Which came first, IT or productivity? The virtuous cycle of investment and use in enterprise systems." In.

Bajari, Patrick et al. (2019). "The impact of big data on firm performance: An empirical investigation." In: *AEA papers and proceedings* 109, pp. 33–37.

Belkin, Mikhail et al. (2019). "Reconciling modern machine-learning practice and the classical bias–variance trade-off." In: *Proceedings of the National Academy of Sciences* 116.32, pp. 15849–15854.

Bergemann, Dirk and Alessandro Bonatti (2024). "Data, Competition, and Digital Platforms." In: *American Economic Review* 114.8, pp. 2553–2595.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan (2022). "The economics of social data." In: *The RAND Journal of Economics* 53.2, pp. 263–296.

Bergemann, Dirk, Alessandro Bonatti, and Alex Smolin (2018). "The Design and Price of Information." In: *American Economic Review* 108.1, pp. 1–48.

Berger, James O. (1990). *Statistical decision theory.* Springer, pp. 277–284.

Black, Fischer and Myron Scholes (1973). "The Pricing of Options and Corporate Liabilities." In: *Journal of Political Economy* 81.3, pp. 637–654.

Brier, Glenn W. (1950). "Verification of Forecasts Expressed in Terms of Probability." In: *Monthly Weather Review* 78.1, pp. 1–3.

Calzolari, Giacomo, Anatole Cheysson, and Riccardo Rovatti (2025). "Machine data: market and analytics." In: *Management Science.*

Carballa-Smichowski, Bruno et al. (2025a). *Data Sharing or Analytics Sharing?* TSE Working Paper 25-1615. Toulouse School of Economics.

Carballa-Smichowski, Bruno et al. (2025b). "Economies of scope in data aggregation: Evidence from health data." In: *Information Economics and Policy* 71, p. 101146.

Cong, Lin William, Zhiguo He, and Changhua Yu (2021). "Data as Capital." In: *Review of Financial Studies* 34.6, pp. 2895–2936.

Cornière, Alexandre de and Greg Taylor (2024). "Data-Driven Mergers." In: *Management Science* 70.9, pp. 6473–6482.

Cournot, Antoine Augustin (1838). *Recherches sur les principes mathématiques de la théorie des richesses.* English translation: *Researches into the Mathematical Principles of the Theory of Wealth*, Macmillan, 1897. Paris: Hachette.

Dasaratha, Krishna, Juan Ortner, and Chengyang Zhu (2025). "Markets for Models." In: *arXiv preprint arXiv:2503.02946*.

De Corniere, Alexandre and Greg Taylor (2025). "Data and Competition: A Simple Framework." In: *Forthcoming, RAND Journal of Economics*.

DeGroot, Morris H. (2005). *Optimal statistical decisions*. John Wiley & Sons.

Digital Platforms, Stigler Committee on (2019). *Final Report*. Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business.

Farboodi, Maryam and Laura Veldkamp (2025). *A model of the Data Economy*. Tech. rep. R&R, Review of Economic Studies.

Goldfarb, Avi and Catherine Tucker (2011). "Privacy Regulation and Online Advertising." In: *Management Science* 57.1, pp. 57–71.

Gu, Yiquan, Leonardo Madio, and Carlo Reggiani (Sept. 2021). "Data brokers co-opetition." In: *Oxford Economic Papers* 74.3, pp. 820–839.

Hagiu, Andrei and Julian Wright (2023). "Data-enabled learning, network effects, and competitive advantage." In: *The RAND Journal of Economics* 54.4, pp. 638–667.

Hastie, Trevor et al. (2020). "Surprises in High-Dimensional Ridgeless Least Squares Interpolation." In.

Hayek, Friedrich A. (1945). "The Use of Knowledge in Society." In: *American Economic Review* 35. Reprinted in F.A. Hayek (ed.), Individualism and Economic Order. London: Routledge and Kegan Paul, pp. 519–530.

Iyer, Ganesh and Tianshu Ke (2024). "Competition and Algorithmic Complexity in Predictive Analytics." In: *Marketing Science* 43.2, pp. 215–233.

Jacot, Arthur, Franck Gabriel, and Clement Hongler (2018). "Neural Tangent Kernel: Convergence and Generalization in Neural Networks." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. New York: Springer-Verlag.

Jones, Charles I. and Christopher Tonetti (2020). "Nonrivalry and the Economics of Data." In: *American Economic Review* 110.9, 2819–58.

Kaplan, Jared et al. (2020). "Scaling laws for neural language models." In: *arXiv preprint arXiv:2001.08361*.

Katz, Michael L. and Carl Shapiro (Aug. 1986). "How to License Intangible Property*." In: *The Quarterly Journal of Economics* 101.3, pp. 567–589.

Lee, Gunhaeng and Julian Wright (2023). "Recommender systems and the Value of User Data." In: *National University of Singapore Working Paper*.

Lerner, Josh and Jean Tirole (2004). "Efficient Patent Pools." In: *American Economic Review* 94.3, 691–711.

Lewis, Patrick et al. (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks." In: *Advances in neural information processing systems* 33, pp. 9459–9474.

Lindley, D. V. and A. F. M. Smith (1972). "Bayes Estimates for the Linear Model." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.1, pp. 1–18.

Liu, Nelson F et al. (2023). "Lost in the middle: How language models use long contexts." In: *arXiv preprint arXiv:2307.03172*.

MacKay, David J. C. (May 1992). "Bayesian Interpolation." In: *Neural Computation* 4.3, pp. 415–447.

Merton, Robert C. (1973). "Theory of Rational Option Pricing." In: *The Bell Journal of Economics and Management Science* 4.1, pp. 141–183.

Montiel Olea, José Luis et al. (Apr. 2022). "Competing Models." In: *The Quarterly Journal of Economics* 137.4, 2419–2457.

Nakkiran, Preetum et al. (2021). "Deep double descent: Where bigger models and more data hurt." In: *Journal of Statistical Mechanics: Theory and Experiment* 2021.12, p. 124003.

Nocke, Volker, Martin Peitz, and Konrad Stahl (Dec. 2007). "Platform Ownership." In: *Journal of the European Economic Association* 5.6, pp. 1130–1160.

Prüfer, Jens and Christoph Schottmüller (2021). "Competing with big data." In: *The Journal of Industrial Economics* 69.4, pp. 967–1008.

Radner, Roy and Joseph Stiglitz (1984). "A Nonconcavity in the Value of Information." In: *Bayesian models in economic theory* 5, pp. 33–52.

Ricardo, D. (1817). *On the Principles of Political Economy and Taxation*. John Murray.

Schaefer, Maximilian (2025). *When Should we Expect Non-Decreasing Returns from Data in Prediction Tasks?*

Schaefer, Maximilian and Geza Sapi (2023). "Complementarities in learning from data: Insights from general search." In: *Information Economics and Policy* 65, p. 101063.

Schmidt, Eric (Oct. 2, 2009). *How Google Plans to Stay Ahead in Search*. Accessed: 2025-11-12. Bloomberg. URL: https://www.bloomberg.com/news/articles/2009-10-02/how-google-plans-to-stay-ahead-in-search.

Selten, Reinhard (1998). "Axiomatic Characterization of the Quadratic Scoring Rule." In: *Experimental Economics* 1.1, pp. 43–62.

Strzalecki, Tomasz (2024). *Variational Bayes and non-Bayesian Updating*.

UK Competition and Markets Authority (2019). *Unlocking Digital Competition: Report of the Digital Competition Expert Panel*. London: Competition and Markets Authority.

Varian, Hal (2018). "Artificial Intelligence, Economics, and Industrial Organization." In: *The Economics of Artificial Intelligence: An Agenda*. NBER Chapters. National Bureau of Economic Research, Inc, pp. 399–419.

Vershynin, Roman (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Wilson, Robert (1975). "Informational Economies of Scale." In: *Bell Journal of Economics* 6.1, pp. 184–195.

Yoganarasimhan, Hema (2020). “Search personalization using machine learning.” In: *Management Science* 66.3, pp. 1045–1070.

A Proofs

Lemma 1 (Optimal Predictor)**.** *The optimal predictor is*

$$\hat{y}^*(\boldsymbol{D}) = \mathbb{E}[y \mid \boldsymbol{D}] = \boldsymbol{x}'_{\mathcal{P}} \mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{P}} \mid (\boldsymbol{y}, \boldsymbol{X}_{\mathcal{T}})\right].$$

Proof. Under squared loss, the Bayes optimal predictor is the conditional mean:

$$\hat{y}^*(\boldsymbol{D}_{\mathcal{T},\mathcal{P}}) = \mathbb{E}[y \mid \boldsymbol{T}, \boldsymbol{x}_{\mathcal{P}}].$$

Write $y = \sum_{k \in \mathcal{K}} \beta_k x_k$. By the law of iterated expectations and independence of $\boldsymbol{x}$ and $\boldsymbol{\beta}$,

$$\mathbb{E}[y \mid \boldsymbol{T}, \boldsymbol{x}_{\mathcal{P}}] = \sum_{k \in \mathcal{K}} \mathbb{E}[\beta_k x_k \mid \boldsymbol{T}, \boldsymbol{x}_{\mathcal{P}}] = \sum_{k \in \mathcal{P}} x_k \, \mathbb{E}[\beta_k \mid \boldsymbol{T}] + \sum_{k \notin \mathcal{P}} \mathbb{E}[\beta_k x_k \mid \boldsymbol{T}, \boldsymbol{x}_{\mathcal{P}}].$$

For $k \notin \mathcal{P}$, x_k is mean zero and independent of $(\boldsymbol{T}, \boldsymbol{x}_{\mathcal{P}}, \boldsymbol{\beta})$, so $\mathbb{E}[\, \beta_k x_k \mid \boldsymbol{T}, \boldsymbol{x}_{\mathcal{P}}] = 0$. Thus

$$\mathbb{E}[y \mid \boldsymbol{T}, \boldsymbol{x}_{\mathcal{P}}] = \sum_{k \in \mathcal{P}} x_k \, \mathbb{E}[\beta_k \mid \boldsymbol{T}].$$

If $k \notin \mathcal{T}$, then β_k is not updated by $\boldsymbol{T}$ and $\mathbb{E}[\beta_k \mid \boldsymbol{T}] = \mathbb{E}[\beta_k] = 0$. Hence, only indices in $\mathcal{T} \cap \mathcal{P}$ contribute, giving

$$\mathbb{E}[y \mid \boldsymbol{T}, \boldsymbol{x}_{\mathcal{P}}] = \boldsymbol{x}'_{\mathcal{P}} \, \mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}} \mid \boldsymbol{T}],$$

which proves the claim. □

Lemma 3. *For any nested sequence $(\mathcal{T}_m)_{m \geq 1}$ with $\mathcal{T}_m \subset \mathcal{T}_{m+1}$ and $\bigcup_m \mathcal{T}_m = \mathcal{T}$,*

$$\sum_{k \in \mathbb{N} \setminus \mathcal{T}_m} \sigma_k^2 \beta_k^2 \xrightarrow[m \to \infty]{a.s.} \operatorname{Var}(\boldsymbol{\varepsilon}_{\mathcal{T}}) = 1 - S(\mathcal{T}_m). \tag{3}$$

Proof. Define $Y_k \equiv \sigma_k^2(\beta_k^2 - 1)$. Then $\mathbb{E}[\, Y_k] = 0$, the Y_k are independent, and

$$\operatorname{Var}(Y_k) = \sigma_k^4 \operatorname{Var}(\beta_k^2) \leq C \, \sigma_k^4 \quad \text{for some finite } C > 0$$

(e.g., $C = 2$ when $\beta_k \sim \mathcal{N}(0,1)$). Because $0 \leq \sigma_k^2 \leq 1$ and $\sum_k \sigma_k^2 = 1$, we have $\sum_k \sigma_k^4 \leq \sum_k \sigma_k^2 = 1 < \infty$, hence

$$\sum_{k=1}^{\infty} \operatorname{Var}(Y_k) \;\leq\; C \sum_{k=1}^{\infty} \sigma_k^4 \;<\; \infty.$$

By Kolmogorov's convergence theorem (a special case of the three-series theorem), the series $\sum_{k=1}^{\infty} Y_k$ converges almost surely. Therefore its *tails* vanish almost surely along any nested complements:

$$R_m \;\equiv\; \sum_{k \in \mathcal{T}_m^c} Y_k \;=\; \sum_{k \in \mathcal{T}_m^c} \sigma_k^2(\beta_k^2 - 1) \xrightarrow[m \to \infty]{a.s.} 0.$$

Now decompose

$$\sum_{k\in\mathcal{T}_m^c}\sigma_k^2\beta_k^2=\sum_{k\in\mathcal{T}_m^c}\sigma_k^2+\sum_{k\in\mathcal{T}_m^c}\sigma_k^2(\beta_k^2-1)=(1-S(\mathcal{T}_m))+R_m,$$

and use $R_m\to 0$ a.s. to conclude. Since $S(\mathcal{T}_m)\uparrow 1$, we also have $\sum_{k\in\mathcal{T}_m^c}\sigma_k^2\beta_k^2\xrightarrow{a.s.}0$ and, when $1-S(\mathcal{T}_m)>0$,

$$\frac{\sum_{k\in\mathcal{T}_m^c}\sigma_k^2\beta_k^2}{1-S(\mathcal{T}_m)}\xrightarrow{a.s.}1.$$

□

Proposition 1 (Posterior Distribution)**.** *The posterior distribution of the parameter vector is given by:*

- *For untrained parameters,*

$$\boldsymbol{\beta}_{\mathcal{T}^c}\mid(\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}})\sim\mathcal{N}\left(\mathbf{0}_{|\mathcal{T}^c|},\boldsymbol{I}_{|\mathcal{T}^c|}\right);$$

- *For trained parameters ,*

$$\boldsymbol{\beta}_{\mathcal{T}}\mid(\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}})\sim\mathcal{N}\Bigg(\underbrace{\left(\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}}+n\lambda\cdot\boldsymbol{I}_{|\mathcal{T}|}\right)^{-1}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{y}}_{\equiv\mathbb{E}[\boldsymbol{\beta}_{\mathcal{T}}|(\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}})]},\quad\underbrace{\left(\frac{1}{n\lambda}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}}+\boldsymbol{I}_t\right)^{-1}}_{\equiv\mathrm{Var}[\boldsymbol{\beta}_{\mathcal{T}}|(\boldsymbol{y},\boldsymbol{X}_{\mathcal{T}})]}\Bigg),$$

  with $\boldsymbol{\beta}_{\mathcal{T}}$ independent from $\boldsymbol{\beta}_{\mathcal{T}^c}$.

Proof. Because the prior is $\boldsymbol{\beta}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_{|\mathbb{N}|})$ and only $\boldsymbol{\beta}_{\mathcal{T}}$ enters the likelihood, the joint posterior factorizes as

$$p(\boldsymbol{\beta}\mid\boldsymbol{T})=p(\boldsymbol{\beta}_{\mathcal{T}}\mid\boldsymbol{T})\,p(\boldsymbol{\beta}_{\mathcal{T}^c}\mid\boldsymbol{T}),$$

with $p(\boldsymbol{\beta}_{\mathcal{T}^c}\mid\boldsymbol{T})=\mathcal{N}(\mathbf{0},\boldsymbol{I}_{|\mathcal{T}^c|})$ since $\boldsymbol{\beta}_{\mathcal{T}^c}$ does not appear in the likelihood and the prior mean is zero. This proves part the second result.

For the first result, the likelihood is

$$\boldsymbol{y}\mid\boldsymbol{\beta}_{\mathcal{T}}\sim\mathcal{N}\left(\boldsymbol{X}_{\mathcal{T}}\boldsymbol{\beta}_{\mathcal{T}},\,u_{\mathcal{T}}^2\boldsymbol{I}_n\right),\quad u_{\mathcal{T}}^2\equiv 1-S(\mathcal{T}),$$

and the prior is $\boldsymbol{\beta}_{\mathcal{T}}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_t)$. Standard derivations in DeGroot (2005) or Berger (1990) show that by conjugacy (or completing the square), the posterior is Gaussian

$$\boldsymbol{\beta}_{\mathcal{T}}\mid\boldsymbol{T}\sim\mathcal{N}(\boldsymbol{\mu}_{\mathcal{T}},\,\Sigma_{\mathcal{T}}),$$

with precision

$$\Sigma_{\mathcal{T}}^{-1}=\frac{1}{u_{\mathcal{T}}^2}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}}+\boldsymbol{I}_t,$$

and mean

$$\boldsymbol{\mu}_{\mathcal{T}} = \Sigma_{\mathcal{T}} \left(\frac{1}{u_{\mathcal{T}}^2} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y} \right) = \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + u_{\mathcal{T}}^2 \boldsymbol{I}_t \right)^{-1} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y}.$$

Therefore, the Bayes estimator (posterior mean) equals the stated ridge form, which proves the first result. □

Proposition 14 (Ridge Estimator Interpretation)**.** *The frequentist optimal ridge parameter is the misspecification penalty*

$$\xi^*(n, \mathcal{T}) = \lambda(n, \mathcal{T}) \equiv \frac{1 - S(\mathcal{T})}{n}.$$

Proof. The ridge objective is strictly convex; its unique minimizer solves the first-order condition:

$$\frac{2}{n} \boldsymbol{X}_{\mathcal{T}}' \left(\boldsymbol{X}_{\mathcal{T}} \hat{\boldsymbol{b}} - \boldsymbol{y} \right) + 2\lambda \hat{\boldsymbol{b}} = \boldsymbol{0}.$$

Hence

$$\left(\tfrac{1}{n} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + \lambda \boldsymbol{I}_t \right) \hat{\boldsymbol{b}} = \tfrac{1}{n} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y}, \quad \text{so} \quad \hat{\boldsymbol{b}} = \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + n\lambda \boldsymbol{I}_t \right)^{-1} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y}.$$

Comparing with the posterior mean from Proposition 1,

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{T}] = \left(\boldsymbol{X}_{\mathcal{T}}' \boldsymbol{X}_{\mathcal{T}} + u_{\mathcal{T}}^2 \boldsymbol{I}_t \right)^{-1} \boldsymbol{X}_{\mathcal{T}}' \boldsymbol{y},$$

we obtain equality when $n\lambda = u_{\mathcal{T}}^2$, i.e. $\lambda = \frac{u_{\mathcal{T}}^2}{n} = \frac{1 - S(\mathcal{T})}{n}$. This proves the claim. □

Lemma 11 (Optimal regularization under a random effects prior)**.** *Consider the linear model*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, u^2 \boldsymbol{I}_n),$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times t}$ is of full column rank and

$$\frac{1}{n} \boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V} \operatorname{diag}(g_1, \ldots, g_t) \boldsymbol{V}'.$$

Suppose the coefficients follow a random-effects prior

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \tau^2 \boldsymbol{I}_t).$$

Then, under squared-error loss, the optimal ridge regularization parameter that minimizes the expected mean-squared error

$$\mathbb{E} \| \hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta} \|^2$$

is given by

$$\lambda^* = \frac{u^2}{n\tau^2}.$$

Proof. The ridge estimator based on the penalized least squares criterion

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \right\}$$

has closed-form solution

$$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{G} + \lambda \boldsymbol{I}_t)^{-1} \frac{1}{n} \boldsymbol{X}'\boldsymbol{y}, \qquad \boldsymbol{G} \equiv \frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V}\,\mathrm{diag}(g_1, \ldots, g_t)\boldsymbol{V}'.$$

In the eigenbasis of $\boldsymbol{G}$, let $\tilde{\boldsymbol{\beta}} = \boldsymbol{V}'\boldsymbol{\beta}$. Using the properties of the Gaussian prior and noise, the posterior mean satisfies

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda \mid \boldsymbol{\beta}] = (\boldsymbol{G} + \lambda \boldsymbol{I}_t)^{-1}\boldsymbol{G}\boldsymbol{\beta}.$$

The mean-squared error can be decomposed as

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|^2 = \sum_{j \in \mathcal{T}} \left[\left(\frac{\lambda}{g_j + \lambda} \right)^2 \mathbb{E}\left[\tilde{\beta}_j^2\right] + \frac{u^2}{n} \frac{g_j}{(g_j + \lambda)^2} \right].$$

Under the random-effects prior $\tilde{\beta}_j \sim \mathcal{N}(0, \tau^2)$, $\mathbb{E}\tilde{\beta}_j^2 = \tau^2$ for all j, so

$$\mathbb{E}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|^2 = \sum_{j \in \mathcal{T}} \frac{\tau^2\lambda^2 + \frac{u^2}{n} g_j}{(g_j + \lambda)^2}.$$

Differentiating with respect to λ gives

$$\frac{\partial}{\partial \lambda}\mathbb{E}\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|^2 = 2 \sum_{j \in \mathcal{T}} \frac{g_j \left(\tau^2\lambda - \frac{u^2}{n} \right)}{(g_j + \lambda)^3}.$$

Setting the derivative to zero yields the first-order condition

$$\sum_{j \in \mathcal{T}} \frac{g_j \left(\tau^2\lambda^* - \frac{u^2}{n} \right)}{(g_j + \lambda^*)^3} = 0.$$

Since all weights $g_j/(g_j + \lambda^*)^3 > 0$, this condition is satisfied if and only if

$$\tau^2\lambda^* - \frac{u^2}{n} = 0,$$

which gives

$$\lambda^* = \frac{u^2}{n\tau^2}.$$

□

Lemma 2. *The value of a dataset design* $(n, \mathcal{T}, \mathcal{P})$ *is*

$$V(n, \mathcal{T}, \mathcal{P}) = \mathrm{Var}_{\mathscr{P}_D(n,\mathcal{T},\mathcal{P})}\left[\hat{y}^*(\boldsymbol{D})\right] = \sum_{k\in\mathcal{P}} \sigma_k^2 \tau_k^2(n, \mathcal{T}).$$

Proof. Under squared loss, the posterior mean minimizes posterior risk, so the ex-ante (expected) value of the dataset equals the prior variance minus the posterior variance of y. Because $y = \boldsymbol{x}_{\mathcal{P}}'\boldsymbol{\beta}_{\mathcal{P}}$ and $\mathcal{P} \subseteq \mathcal{T}$,

$$\hat{y}^*(\boldsymbol{D}^n_{\mathcal{T},\mathcal{P}}) = \boldsymbol{x}'_{\mathcal{P}\cap\mathcal{T}}\mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}} \mid \boldsymbol{T}]\,.$$

Taking the variance over the joint distribution of $\boldsymbol{x}_{\mathcal{P}}$ and $\boldsymbol{T}$ gives

$$\mathrm{Var}\left[\hat{y}^*(\boldsymbol{D}^n_{\mathcal{T},\mathcal{P}})\right] = \mathrm{Var}\left(\boldsymbol{x}'_{\mathcal{P}\cap\mathcal{T}}\mathbb{E}[\boldsymbol{\beta}_{\mathcal{P}\cap\mathcal{T}} \mid \boldsymbol{T}]\right) = \sum_{k\in\mathcal{P}\cap\mathcal{T}} \mathrm{Var}(x_k)\,\mathrm{Var}(\mathbb{E}[\beta_k \mid \boldsymbol{T}])\,,$$

using independence of covariates. Since $\mathrm{Var}(x_k) = \sigma_k^2$ and $\mathbb{E}[\beta_k \mid \boldsymbol{T}] = \hat{\beta}_k^{\mathrm{ridge}}(\lambda(n, \mathcal{T}), \boldsymbol{T})$, we obtain

$$V(n, \mathcal{T}, \mathcal{P}) = \sum_{k\in\mathcal{P}\cap\mathcal{T}} \sigma_k^2\,\mathrm{Var}_T\left[\hat{\beta}_k^{\mathrm{ridge}}(\lambda(n, \mathcal{T}), \boldsymbol{T})\right] = \sum_{k\in\mathcal{P}\cap\mathcal{T}} \sigma_k^2 \tau_k^2(\lambda(n, \mathcal{T})),$$

which proves the result. □

Proposition 3 (Training Information Approximation)**.** *The following asymptotic approximation holds*

$$V(n, \mathcal{T}, \mathcal{P}) = \sum_{k\in\mathcal{P}\cap\mathcal{T}} \sigma_k^2 \bar{\tau}_k^2(\lambda(n, \mathcal{T})) + O\left(\sqrt{\tfrac{|\mathcal{T}|}{n}} + \tfrac{|\mathcal{T}|}{n}\right),$$

where

$$\bar{\tau}_k^2(\lambda) \equiv \frac{\sigma_k^2}{\sigma_k^2 + \lambda}.$$

Proof. Write $\Sigma = \Sigma_{\mathcal{T}}$ and recall $\phi_k(\lambda; n, t, \boldsymbol{\sigma}^2_{\mathcal{T}}) = \lambda\,\mathbb{E}\left[(\boldsymbol{G}_n + \lambda \boldsymbol{I}_t)^{-1}_{kk}\right]$, where $\boldsymbol{G} \equiv \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$. For symmetric $\boldsymbol{A}, \boldsymbol{B} \succeq 0$, the resolvent identity gives

$$(\boldsymbol{A} + \lambda \boldsymbol{I})^{-1} - (\boldsymbol{B} + \lambda \boldsymbol{I})^{-1} = (\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{B} - \boldsymbol{A})(\boldsymbol{B} + \lambda \boldsymbol{I})^{-1}.$$

With $\boldsymbol{A} = \boldsymbol{G}_n$ and $\boldsymbol{B} = \Sigma$, taking (k, k) entries and bounding by operator norms,

$$\left|((\boldsymbol{G}_n + \lambda \boldsymbol{I})^{-1} - (\Sigma + \lambda \boldsymbol{I})^{-1})_{kk}\right| \le \|(\boldsymbol{G}_n + \lambda \boldsymbol{I})^{-1}\|_{\mathrm{op}}\,\|\boldsymbol{G}_n - \Sigma\|_{\mathrm{op}}\,\|(\Sigma + \lambda \boldsymbol{I})^{-1}\|_{\mathrm{op}}.$$

Because $\boldsymbol{G}_n, \Sigma \succeq 0$, we have $\|(\boldsymbol{G}_n + \lambda \boldsymbol{I})^{-1}\|_{\mathrm{op}} \le \lambda^{-1}$ and $\|(\Sigma + \lambda \boldsymbol{I})^{-1}\|_{\mathrm{op}} \le (\sigma^2_{\min} + \lambda)^{-1}$. Hence

$$\left|((\boldsymbol{G}_n + \lambda \boldsymbol{I})^{-1} - (\Sigma + \lambda \boldsymbol{I})^{-1})_{kk}\right| \le \frac{1}{\lambda(\sigma^2_{\min} + \lambda)}\,\|\boldsymbol{G}_n - \Sigma\|_{\mathrm{op}}.$$

Taking expectations and multiplying by λ,

$$\left|\lambda\,\mathbb{E}\left[(\boldsymbol{G}_n+\lambda\boldsymbol{I})^{-1}_{kk}\right]-\lambda\left[(\Sigma+\lambda\boldsymbol{I})^{-1}\right]_{kk}\right|\leq\frac{1}{\sigma^2_{\min}+\lambda}\,\mathbb{E}\|\boldsymbol{G}_n-\Sigma\|_{\text{op}}.$$

Since $[(\Sigma+\lambda\boldsymbol{I})^{-1}]_{kk}=(\sigma_k^2+\lambda)^{-1}$, this becomes

$$\left|\phi_k(\lambda;n,t,\boldsymbol{\sigma}^2_{\mathcal{T}})-\frac{\lambda}{\sigma_k^2+\lambda}\right|\leq\frac{1}{\sigma^2_{\min}+\lambda}\,\mathbb{E}\|\boldsymbol{G}_n-\Sigma\|_{\text{op}},$$

where since covariates are ranked in descending order $\sigma^2_{\min}=\sigma^2_{\max\{\mathcal{T}\}}$.

For Gaussian rows, the sample-covariance bound in Theorem 4.7.1. of Vershynin (2018) shows that there exists an absolute constant $C>0$ such that

$$\mathbb{E}\|\boldsymbol{G}_n-\Sigma\|_{\text{op}}\leq C\,\|\Sigma\|_{\text{op}}\left(\sqrt{\tfrac{t}{n}}+\tfrac{t}{n}\right),$$

and since covariates are ranked in descending order

$$\|\Sigma\|_{\text{op}}\equiv\sup_{\|\boldsymbol{x}\|_2=1}\|\Sigma\boldsymbol{x}\|_2=\sup_{\|x\|_2=1}\boldsymbol{x}'\Sigma\boldsymbol{x}=\max_{k\in\mathcal{T}}\{\sigma_k^2\}=\sigma^2_{\min\{\mathcal{T}\}}.$$

Therefore,

$$\mathbb{E}\|\boldsymbol{G}_n-\Sigma\|_{\text{op}}\leq,$$

which gives the claimed inequality. Putting together,

$$\left|\phi_k(\lambda;n,t,\boldsymbol{\sigma}^2_{\mathcal{T}})-\frac{\lambda}{\sigma_k^2+\lambda}\right|\leq\frac{C\sigma^2_{\min\{\mathcal{T}\}}}{\sigma^2_{\max\{\mathcal{T}\}}+\lambda}\left(\sqrt{\frac{t}{n}}+\frac{t}{n}\right).$$

The $O(\cdot)$ statement follows because $\sigma^2_{\max}$ and $(\sigma^2_{\min}+\lambda)^{-1}$ are uniformly bounded as $\sigma_j^2\in(0,1)$ for all j and $\lambda\in(0,\infty)$.

Proposition 6 (Complementarity/Substitutability in Training). *Fix $(n,\mathcal{P})$. For any $\mathcal{T}$ and any $k\in\mathcal{T}\setminus\mathcal{P}$,*

$$\frac{d}{dn}\Delta_k^{\mathcal{T}}(n,\mathcal{P},\mathcal{T})=\frac{\sigma_k^2}{n^2}\sum_{m\in\mathcal{P}\cap\mathcal{T}}\sigma_m^2\Bigg(\underbrace{\frac{1-S(\mathcal{T}\setminus\{k\})}{\sigma_k^2}\int_{-\sigma_k^2/n}^{0}\partial_{\lambda^2}\bar{\tau}_k^2(\lambda(n,\mathcal{T}\setminus\{k\})+u)\,du}_{(+)\text{ HPE}\times\text{Sampling spillovers}}+\underbrace{\partial_\lambda\bar{\tau}_m^2\left(\lambda(n,\mathcal{T})\right)}_{(-)\text{ Spillover substitution}}\Bigg)>0$$

if and only if

$$n<\tilde{n}(\mathcal{T},\mathcal{P})\in(0,\infty),$$

which is decreasing in $\mathcal{T}\setminus\{k\}$ and $\mathcal{P}$.

Fix $(\mathcal{P}, \mathcal{T})$ and let $k \in \mathcal{T} \setminus \mathcal{P}$. Recall

$$V(n, \mathcal{P}, \mathcal{T}) = \sum_{m \in \mathcal{P} \cap \mathcal{T}} \sigma_m^2 \left[1 - \bar{\phi}_m(\lambda^*(n, \mathcal{T}))\right], \qquad \bar{\phi}_m(\lambda) = \frac{\lambda}{\sigma_m^2 + \lambda}, \qquad \lambda^*(n, \mathcal{T}) = \frac{1 - S(\mathcal{T})}{n}.$$

Since $k \notin \mathcal{P}$, the prediction set does not change when we remove k, so

$$\Delta_k^{\mathcal{T}}(n, \mathcal{P}, \mathcal{T}) \equiv V(n, \mathcal{P}, \mathcal{T}) - V(n, \mathcal{P}, \mathcal{T} \setminus \{k\}) = \sum_{m \in \mathcal{P} \cap \mathcal{T}} \sigma_m^2 \left[\bar{\phi}_m\big(\lambda^*(n, \mathcal{T} \setminus \{k\})\big) - \bar{\phi}_m\big(\lambda^*(n, \mathcal{T})\big)\right].$$

Using

$$\lambda^*(n, \mathcal{T} \setminus \{k\}) = \frac{1 - S(\mathcal{T}) + \sigma_k^2}{n} = \lambda^*(n, \mathcal{T}) + \frac{\sigma_k^2}{n},$$

define

$$\lambda(n) \equiv \lambda^*(n, \mathcal{T}) = \frac{1 - S(\mathcal{T})}{n}, \qquad \delta(n) \equiv \frac{\sigma_k^2}{n},$$

so that $\lambda^*(n, \mathcal{T} \setminus \{k\}) = \lambda + \delta$. Then

$$\Delta_k^{\mathcal{T}}(n, \mathcal{P}, \mathcal{T}) = \sum_{m \in \mathcal{P} \cap \mathcal{T}} \sigma_m^2 \left[\bar{\phi}_m(\lambda + \delta) - \bar{\phi}_m(\lambda)\right]. \tag{4}$$

Step 1: derivative decomposition.

Let

$$F_m(\lambda, \delta) \equiv \sigma_m^2 \left[\bar{\phi}_m(\lambda + \delta) - \bar{\phi}_m(\lambda)\right].$$

By the chain rule,

$$\frac{d}{dn}\Delta_k^{\mathcal{T}}(n, \mathcal{P}, \mathcal{T}) = \sum_{m \in \mathcal{P} \cap \mathcal{T}} \left(\partial_\lambda F_m(\lambda, \delta)\, \lambda' + \partial_\delta F_m(\lambda, \delta)\, \delta'\right),$$

where

$$\lambda' = \frac{d\lambda}{dn} = -\frac{1 - S(\mathcal{T})}{n^2}, \qquad \delta' = \frac{d\delta}{dn} = -\frac{\sigma_k^2}{n^2}.$$

Since

$$\partial_\lambda F_m = \sigma_m^2 \left[\bar{\phi}_m'(\lambda + \delta) - \bar{\phi}_m'(\lambda)\right], \qquad \partial_\delta F_m = \sigma_m^2 \bar{\phi}_m'(\lambda + \delta),$$

writing $[\bar{\phi}_m'(\lambda + \delta) - \bar{\phi}_m'(\lambda)] = \int_\lambda^{\lambda+\delta} \bar{\phi}_m''(u)du$, we obtain using the chain rule

$$\begin{aligned}
\frac{d}{dn}\Delta_k^{\mathcal{T}} &= \sum_{m \in \mathcal{P} \cap \mathcal{T}} \sigma_m^2 \left(\lambda' \int_\lambda^{\lambda+\delta} \bar{\phi}_m''(u)du + \bar{\phi}_m'(\lambda + \delta)\, \delta'\right) \\
&= \sum_{m \in \mathcal{P} \cap \mathcal{T}} \sigma_m^2 \left(\frac{[1 - S(\mathcal{T})]}{n^2} \int_\lambda^{\lambda+\delta} \bar{\phi}_m''(u)du - \frac{\sigma_k^2}{n^2} \bar{\phi}_m'(\lambda + \delta)\right).
\end{aligned} \tag{5}$$

This is the claimed decomposition: a positive term (HPE × observation spillovers) minus a negative term (spillover substitution).

Step 2: signs of the two forces.

We use only concavity of $\bar{\phi}_m$. First, $\bar{\phi}'_m(\lambda) > 0$ and

$$\bar{\phi}''_m(\lambda) = -\frac{2\sigma_m^2}{(\sigma_m^2+\lambda)^3} < 0,$$

so $\bar{\phi}'_m$ is strictly decreasing. Since $\delta > 0$,

$$\bar{\phi}'_m(\lambda) - \bar{\phi}'_m(\lambda+\delta) > 0.$$

Hence the first term in (5) is strictly positive:

$$\frac{\sigma_m^2[1-S(\mathcal{T})]}{n^2}\left[\bar{\phi}'_m(\lambda) - \bar{\phi}'_m(\lambda+\delta)\right] > 0.$$

This captures that as n increases (so λ decreases), concavity of $\bar{\phi}_m$ amplifies the gain from the extra training covariate k (*House Party Effect* interacting with *observation spillovers*).

The second term is strictly negative since $\bar{\phi}'_m(\lambda+\delta) > 0$ and $\sigma_m^2, \sigma_k^2 > 0$:

$$\frac{\sigma_m^2\sigma_k^2}{n^2}\bar{\phi}'_m(\lambda+\delta) > 0 \quad \Rightarrow \quad -\frac{\sigma_m^2\sigma_k^2}{n^2}\bar{\phi}'_m(\lambda+\delta) < 0.$$

This is *spillover substitution*: as n increases, the direct impact $\delta = \sigma_k^2/n$ of covariate k on regularization shrinks.

Therefore

$$\frac{d}{dn}\Delta_k^{\mathcal{T}} > 0 \iff [1-S(\mathcal{T})]\,A(n,\mathcal{T},\mathcal{P}) > \sigma_k^2\,B(n,\mathcal{T},\mathcal{P}),$$

where

$$A(n,\mathcal{T},\mathcal{P}) \equiv \sum_{m\in\mathcal{P}\cap\mathcal{T}} \sigma_m^2\left[\bar{\phi}'_m(\lambda) - \bar{\phi}'_m(\lambda+\delta)\right] > 0, \quad B(n,\mathcal{T},\mathcal{P}) \equiv \sum_{m\in\mathcal{P}\cap\mathcal{T}} \sigma_m^2\bar{\phi}'_m(\lambda+\delta) > 0.$$

Step 3: existence and dependence of the threshold $\tilde{n}(\mathcal{T},\mathcal{P})$.

Define

$$\Psi(n;\mathcal{T},\mathcal{P}) \equiv [1-S(\mathcal{T})]\,A(n,\mathcal{T},\mathcal{P}) - \sigma_k^2\,B(n,\mathcal{T},\mathcal{P}),$$

so that $\frac{d}{dn}\Delta_k^{\mathcal{T}} > 0 \iff \Psi(n;\mathcal{T},\mathcal{P}) > 0$.

For large n, we have $\lambda = \mathcal{O}(1/n)$ and $\delta = \mathcal{O}(1/n)$. Using the Taylor expansion

$$\bar{\phi}'_m(\lambda) = \frac{\sigma_m^2}{(\sigma_m^2+\lambda)^2} = \frac{1}{\sigma_m^2} - \frac{2\lambda}{\sigma_m^4} + \mathcal{O}(\lambda^2),$$

we obtain uniformly over m:

$$\bar{\phi}'_m(\lambda) - \bar{\phi}'_m(\lambda+\delta) = \frac{2\delta}{\sigma_m^4} + \mathcal{O}\left(\frac{1}{n^2}\right),$$
$$\bar{\phi}'_m(\lambda+\delta) = \frac{1}{\sigma_m^2} + \mathcal{O}\left(\frac{1}{n}\right).$$

Hence, as $n \to \infty$,

$$A(n, \mathcal{T}, \mathcal{P}) = 2\delta \sum_{m \in \mathcal{P} \cap \mathcal{T}} \frac{1}{\sigma_m^2} + \mathcal{O}\left(\frac{1}{n^2}\right) = \frac{2\sigma_k^2}{n} \sum_{m \in \mathcal{P} \cap \mathcal{T}} \frac{1}{\sigma_m^2} + \mathcal{O}\left(\frac{1}{n^2}\right),$$

and

$$B(n, \mathcal{T}, \mathcal{P}) = \sum_{m \in \mathcal{P} \cap \mathcal{T}} 1 + \mathcal{O}\left(\frac{1}{n}\right) = |\mathcal{P} \cap \mathcal{T}| + \mathcal{O}\left(\frac{1}{n}\right).$$

Therefore

$$\Psi(n; \mathcal{T}, \mathcal{P}) = \frac{2[1 - S(\mathcal{T})]\sigma_k^2}{n} \sum_{m \in \mathcal{P} \cap \mathcal{T}} \frac{1}{\sigma_m^2} - \sigma_k^2 |\mathcal{P} \cap \mathcal{T}| + \mathcal{O}\left(\frac{1}{n}\right).$$

For large n, the negative constant term dominates, so $\Psi(n; \mathcal{T}, \mathcal{P}) < 0$ and thus $\frac{d}{dn}\Delta_k^{\mathcal{T}} < 0$. For n close to 1 (or small enough), the positive term dominates by continuity of all components, so $\Psi(n; \mathcal{T}, \mathcal{P}) > 0$. Since Ψ is continuous in n, there exists at least one threshold $\tilde{n}(\mathcal{T}, \mathcal{P})$ such that

$$\frac{d}{dn}\Delta_k^{\mathcal{T}} > 0 \iff n < \tilde{n}(\mathcal{T}, \mathcal{P}).$$

Using the leading-order approximation $\Psi(n; \mathcal{T}, \mathcal{P}) \approx 0$ yields

$$\tilde{n}(\mathcal{T}, \mathcal{P}) \approx \frac{2[1 - S(\mathcal{T})]}{|\mathcal{P} \cap \mathcal{T}|} \sum_{m \in \mathcal{P} \cap \mathcal{T}} \frac{1}{\sigma_m^2}.$$

This expression makes the dependence on $\mathcal{T}$ transparent: $\tilde{n}(\mathcal{T}, \mathcal{P})$ is approximately proportional to the *residual variance* $1 - S(\mathcal{T})$ and thus is *decreasing in* $S(\mathcal{T})$: richer training sets (larger $S(\mathcal{T})$) shift the peak to the left.

In summary, the derivative admits the decomposition (5); the first term (HPE × observation spillovers) is positive, the second (spillover substitution) is negative, and their balance induces a unique threshold $\tilde{n}(\mathcal{T}, \mathcal{P})$, decreasing in $S(\mathcal{T})$, such that $\frac{d}{dn}\Delta_k^{\mathcal{T}} > 0$ if and only if $n < \tilde{n}(\mathcal{T}, \mathcal{P})$. □

Proposition 12 (Closed Form Accuracy). *If* $\sigma_k^2 = \sigma^2 \mathbf{1}(k \in \mathcal{I})$, *the accuracy is*

$$A(n, p, t) = \frac{S(\min\{t, p\})}{\frac{1 - S(t)}{n\sigma^2} + 1}.$$

Proof. Let $\mathcal{I} \subset \mathbb{N}$ be the (finite) set of informative covariates. Assume:

$\sigma_k^2 = s^2 > 0$ for all $k \in \mathcal{I}$ and $\sigma_k^2 = 0$ for $k \notin \mathcal{I}$;

$\mathrm{Var}(y) = 1$, so $|\mathcal{I}|s^2 = 1$;

the agent is constrained to observe covariates in a fixed order, so any feasible training and prediction sets are prefixes of $\mathbb{N}$.

For $m \in \mathbb{N}$, let $S(m) \equiv \sum_{k=1}^{m} \sigma_k^2$ denote the cumulative variance, and for $t, p \geq 0$ write

$$S(t) \equiv S(\lfloor t \rfloor), \qquad S(p) \equiv S(\lfloor p \rfloor),$$

viewing t and p as (possibly real) indexes along the ordered list.

Step 1: Optimal sets are prefixes at full capacity.

Under the fixed-order constraint, any feasible training set has the form

$$\mathcal{T} = \{1, \dots, m_T\} \quad \text{with } m_T \leq t,$$

and any prediction set

$$\mathcal{P} = \{1, \dots, m_P\} \quad \text{with } m_P \leq p.$$

Thus $\mathcal{P} \cap \mathcal{T} = \{1, \dots, \min\{m_T, m_P\}\}$.

Each term in V is nonnegative:

$$\sigma_k^2 \left[1 - \bar{\phi}_k(\lambda^*)\right] = \frac{\sigma_k^4}{\sigma_k^2 + \lambda^*} \geq 0.$$

Hence enlarging m_T or m_P weakly increases V. Therefore the supremum is attained at

$$m_T = t, \qquad m_P = p,$$

so we may write

$$A(n, t, p) = V(n, \mathcal{T}, \mathcal{P}_p), \quad \mathcal{T}_t = \{1, \dots, t\}, \ \mathcal{P}_p = \{1, \dots, p\}.$$

Step 2: Identify the intersection mass.

Since only $k \in \mathcal{I}$ are informative (with variance s^2), the total variance of informative covariates in the intersection is

$$\sum_{k \in \mathcal{P}_p \cap \mathcal{T}_t} \sigma_k^2 = S(\min\{t, p\}).$$

Because all informative coordinates have the same variance s^2, the number of informative coordinates in the intersection is

$$M = \frac{S(\min\{t, p\})}{s^2}.$$

Step 3: Compute the contribution of each informative coordinate.

For any informative k,

$$1 - \bar{\phi}_k(\lambda) = 1 - \frac{\lambda}{\sigma_k^2 + \lambda} = \frac{\sigma_k^2}{\sigma_k^2 + \lambda},$$

so

$$\sigma_k^2 \Big[1 - \bar{\phi}_k(\lambda)\Big] = \frac{\sigma_k^4}{\sigma_k^2 + \lambda}.$$

In our problem, $\lambda = \lambda^*(n, \mathcal{T}) = \dfrac{1 - S(t)}{n}$, which does not depend on k. Hence all informative coordinates in $\mathcal{P}_p \cap \mathcal{T}_t$ contribute the same amount, and

$$V(n, \mathcal{T}, \mathcal{P}_p) = M \cdot \frac{s^4}{s^2 + \lambda^*(n, \mathcal{T})}.$$

Substitute $M = S(\min\{t, p\})/s^2$:

$$V(n, \mathcal{T}, \mathcal{P}_p) = \frac{S(\min\{t, p\})}{s^2} \cdot \frac{s^4}{s^2 + \lambda^*(n, \mathcal{T})} = S(\min\{t, p\}) \cdot \frac{s^2}{s^2 + \lambda^*(n, \mathcal{T})}.$$

Step 4: Substitute λ^* and simplify.

We have

$$\lambda^*(n, \mathcal{T}) = \frac{1 - S(t)}{n},$$

so

$$\frac{s^2}{s^2 + \lambda^*(n, \mathcal{T})} = \frac{s^2}{s^2 + \dfrac{1 - S(t)}{n}} = \frac{1}{1 + \dfrac{1 - S(t)}{n s^2}}.$$

Therefore

$$A(n, t, p) = V(n, \mathcal{T}, \mathcal{P}_p) = \frac{S(\min\{t, p\})}{1 + \dfrac{1 - S(t)}{n s^2}},$$

which is the desired expression.

If we reparametrize the sample size in effective units $n' := ns^2$ (so that $s^2 = 1$ in the formula), this becomes

$$A(n', t, p) = \frac{S(\min\{t, p\})}{1 + \dfrac{1 - S(t)}{n'}},$$

as claimed. □

Lemma 4 (Optimal Pool Price). *The optimal pool price is*

$$P^* = \frac{A_2}{\alpha + 1}.$$

Proof. The pool price is

$$P^* = \arg\max_P \{PD\,(P - V_2)\},$$

which can be solved using the first-order condition

$$D(P - V_2) + PD'(P - V_2) = 0.$$

□

Lemma 5 (Sample Fragmentation Price)**.** *If the brokers have different observations on the same covariates,*

$$p_i = \min\left\{A_2 - V_1, \frac{A_2}{2+\alpha}\right\},$$

and the buyers will buy from both brokers.

Proof. We follow Lerner and Tirole (2004). **Demand Margin Binds.** Suppose that the brokers offer prices $\mathcal{P} \equiv (p_1, p_2)$, and wlog $p_1 \leq p_2$. Prediction Sellers decide how many datasets to buy.

$$\mathcal{V}(\mathcal{P}) = \max_{q \in \{1,2\}} \left\{ V_q - p_1 - p_2 \mathbf{1}(\{q = 2\}) \right\}$$

Second, the user adopts the technology if and only if

$$\mathcal{V}(\mathcal{P}) \geq \theta.$$

Lerner and Tirole (2004) demonstrate the existence of a symmetric equilibrium. Individual data sellers solve

$$\hat{p} = \arg\max_{p_i} \left\{ p_i D\left(p_i + \hat{p} - V_2\right) \right\}$$

which has FOC

$$\hat{p}D'\left(2\hat{p} - V_2\right) + D\left(2\hat{p} - V_2\right) = 0$$

which has a unique solution buy hazard-rate monotonicity. It can be seen as selling the whole pool setting total price P and keeping $p_i = P - \hat{p}$ for itself. Therefore

$$\hat{P} = \arg\max_{P} \left\{ \left(P - \hat{p}\right) D\left(P - V_2\right) \right\}.$$

The term $\hat{p}$ can be seen as a marginal cost $\hat{c} = \hat{p}$. In this interpretation when there is the pool $c^* = 0$ so by revealed preference

$$\hat{P} \geq P^*.$$

If demand margin binds in the absence of a pool then the pool reduces price paid by data buyers. This means that if all datasets can increase the price marginally without being excluded, the pool is pro-competitive.

With our CDF G,

$$p_{\text{dem}} = \frac{V_2}{2+\alpha}.$$

Competition Margin Binds. Define the price when the competition margin binds will

be p_{comp} defined by

$$V_2 - 2p_{\text{comp}} = \max_{q \in \{0,1\}} \left\{ V_q - q p_{\text{comp}} \right\}.$$

If $V_1 - p_{\text{comp}} \geq 0$, then $p_{\text{comp}} = V_2 - V_1$. This is consistent because $V_1 - (V_2 - V_1) > 0$ by concavity of $V(n,t)$ in n. Otherwise, if $V_1 - p_{\text{comp}} < 0$, then $p_{\text{comp}} = V_2/2$. This is not consistent because $V_1 - V_2/2 > 0$ by concavity of $V(n,t)$ in n. □

Lemma 6 (Welfare of Observation Pooling). *A pool of observations is procompetitive if and only if*

$$\alpha > \frac{n}{2(1-k)}.$$

and the buyers will buy from both brokers.

Proof. Pools are procompetitive if the demand margin binds i.e.

$$p_{\text{comp}} > p_{\text{dem}} \iff V_2 - V_1 > \frac{V_2}{2+\alpha} \iff \alpha > \alpha_{\text{marg}} \equiv \frac{2V_1 - V_2}{V_2 - V_1}.$$

In this case

$$p_i = \frac{V_2}{2+\alpha}.$$

This implies that observation pooling can be procompetitive when n is not too large and k is not too small, meaning data is relatively abundant and models are relatively complex. Furthermore, as the RHS is increasing in Q, data pools are more likely to be competitive if Q is small meaning if data fragmentation is limited.

Otherwise if the competition margin binds,

$$p_i = V_2 - V_1.$$

the pool is procompetitive if the pool price is lower than the competition price, i.e.

$$P^* < Qz(Q) \iff \frac{V_2}{\alpha+1} < 2(V_2 - V_1) \iff \alpha > \alpha_{\text{comp}} \equiv \frac{V_1 - \frac{V_2}{2}}{V_2 - V_1}.$$

As $\alpha_{\text{marg}} > \alpha_{\text{comp}}$, the relevant threshold is α_{comp} and so a pool of observations is procompetitive if and only if

$$\alpha > \alpha_{\text{comp}} = \frac{V_1 - \frac{V_2}{2}}{V_2 - V_1}.$$

This implies that as k and n increase it becomes less likely that the pool is procompetitive. When data is abundant and demand is inelastic observation pools are anticompetitive. Direct application of Proposition 5 in Lerner and Tirole (2004) implies that the pool is strongly unstable, therefore enforcing independent licensing of datasets will prevent pooling if and only if the pool is welfare-reducing. In this case each data broker will charge $p_{\text{comp}} = V_2 - V_1$. □

Lemma 7 (Covariate Fragmentation Price)**.** *If the brokers have distinct covariates each B_i prices at*

$$p_i = \frac{A_2}{2+\alpha},$$

and the buyers will buy from both brokers.

Proof. If $V_1 - p_{\text{comp}} \geq 0$, then $p_{\text{comp}} = V_2 - V_1$. This is not consistent because $V_1 - (V_2 - V_1) < 0$ by convexity of $V(n,t)$ in t. Otherwise, if $V_1 - p_{\text{comp}} < 0$, then $p_{\text{comp}} = V_2/2$. This is consistent because $V_1 - V_2/2 < 0$ by convexity of $V(n,t)$ in t. Demand margin always binds as

$$p_{\text{comp}} > p_{\text{dem}} \iff \frac{V_2}{2} > \frac{V_2}{2+\alpha}.$$

□

Lemma 8 (Shopper-price equilibrium)**.** *Fix (A_1, A_2). The unique trembling-hand-perfect equilibrium is*

$$p_i^s = (A_i - A_j)^+,$$

so the higher-accuracy firm extracts the full willingness to pay of shoppers.

Proof. We first show that (**??**) is an equilibrium. Consider $v_i > v_j$. If i sets $p_i^s = v_i - v_j$ and j sets $p_j^s = 0$, shoppers are indifferent and (by the refinement) all go to i.[17] Any deviation by i:

If $p_i^s > v_i - v_j$, then $v_i - p_i^s < v_j - p_j^s = v_j$ and i loses all shoppers, strictly reducing its shopper revenue to 0.

If $p_i^s < v_i - v_j$, i still serves all shoppers but leaves revenue on the table; since demand is inelastic at one (all shoppers) at the margin, profit increases by raising p_i^s up to $v_i - v_j$.

Any deviation by j:

If $p_j^s > 0$, then $v_j - p_j^s < v_j = v_i - p_i^s$ and j serves no shoppers with the same zero shopper revenue; under trembling hand, the weakly dominated positive price is eliminated in the limit, selecting $p_j^s = 0$.

If $p_j^s < 0$ is infeasible; if $p_j^s = 0$ already, no profitable deviation exists.

Thus (**??**) is an equilibrium when $v_i > v_j$. The case $v_i = v_j$: for any $p_i^s = p_j^s$, shoppers are indifferent and each firm earns $\sigma p_i^s/2$ from shoppers; any unilateral increase loses all shoppers, any decrease reduces price with the same demand, so $p_i^s = p_j^s = 0$ is the unique trembling-hand limit (positive common prices are not robust to small payoff perturbations). Symmetry covers $v_j > v_i$.

Uniqueness under trembling-hand perfection follows from the standard Bertrand-vertical-differentiation logic: if $v_i > v_j$, any equilibrium must have the higher-quality firm serving all shoppers; then the highest sustainable price for i that keeps all shoppers is $v_i - v_j$, and the lower-quality firm's best response is any price with zero demand and zero revenue, refined to 0. □

[17] Formally, with tiny perturbations (trembles) that give i an ε quality advantage or j an ε higher price with positive probability, the unique limit assigns the shoppers to i.

B Extensions

House i	y^i	x^i_{size}	x^i_{year}	x^i_{dist}	x^i_{sun}	
0	?	x^0_{size}	NA	x^0_{dist}	NA	} Prediction Vector: $\boldsymbol{x}'_{\mathcal{P}}$
1	y^1	x^1_{size}	x^1_{year}	NA	x^1_{sun}	} Training Matrix: $\boldsymbol{M}^{(n)}_{\mathcal{T}}$
2	y^2	x^2_{size}	x^2_{year}	NA	x^2_{sun}	
⋮	⋮	⋮	⋮	⋮	⋮	
n	y^n	x^n_{size}	x^n_{year}	NA	x^n_{sun}	

Table 1: Example of Zillow dataset with prediction covariates $\mathcal{P} = \{\text{size}, \text{dist}\}$ and training covariates $\mathcal{T} = \{\text{size}, \text{year}, \text{sun}\}$, where *size* denotes square meters, *dist* the distance to the nearest supermarket, *year* the construction year, and *sun* the daily sunlight exposure.

B.1 Scope as Model Complexity and LLMs

Scope as Model Complexity. Instead, make no restriction on Σ. Furthermore suppose the firm observes all covariates for all individuals but faces constraints on the number of covariates it can effectively use in the learning and targeting steps. The scope of learning, ℓ, is the number of principal components the firm can use in learning. The scope of targeting, t, is the number of principal components that can be used in targeting. This interpretation captures the *model complexity*, which reflects the higher computing cost deriving from analyzing more covariates.

To reduce the dimensionality whilst extracting the maximum information in the constraints, Jolliffe (2002) shows that the optimal procedure is Principal Component Analysis (PCA). Let the eigendecomposition of the variance/covariance matrix be

$$\Sigma = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}', \qquad \boldsymbol{S} = \operatorname{diag}(s_1 \geq \cdots \geq s_{\bar{\ell}} \geq 0), \qquad \boldsymbol{U} \text{ orthonormal.}$$

Define principal components $\boldsymbol{z}^i \equiv \boldsymbol{x}^i\boldsymbol{U}$. Then

$$\boldsymbol{z}^i \sim \mathcal{N}(0, \Lambda), \qquad z^i_j \text{ are uncorrelated with variances } s_j.$$

Remark 1 (Application to Large Language Models (LLMs)). Although LLMs are trained with cross-entropy loss, near a trained solution their behavior can be well approximated by a linear predictor under squared loss in a suitable linear transformation of the covariates (MacKay (1992); Jacot, Gabriel, and Hongler (2018)). In this local view, our primitives map directly: the scale of learning n corresponds to the amount of training information (e.g., the number of training observations/tokens), the scope of learning ℓ captures the effective number of informative directions used at the learning stage, and the scope of targeting t captures the amount of information observed at the targeting stage for specific instances. Under this mapping, comparative statics in (n, ℓ, t) align with empirical scaling laws for language models (Kaplan

et al. (2020)). Supplying richer information at prediction time corresponds to increasing t via retrieval-augmented inputs (Lewis et al. (2020)), with benefits contingent on relevance and known long-context effects (Liu et al. (2023)).

B.2 Shrinkage Interpretation

We express the Bayes estimator in terms of a generalization of the ordinary least-squares (OLS) estimator — the minimum-norm least-squares (MNLS) estimator, defined as

$$\hat{\boldsymbol{\beta}}_{\mathcal{T}}^{\text{MNLS}} \equiv \left(\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}}\right)^{+}\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{y} = \begin{cases} \hat{\boldsymbol{\beta}}_{\mathcal{T}}^{\text{OLS}}, & \text{if } |\mathcal{T}| \leq n, \\ \min_{\boldsymbol{b}_{\mathcal{T}}} \{\|\boldsymbol{b}_{\mathcal{T}}\|_2 \,:\, \boldsymbol{X}_{\mathcal{T}}\boldsymbol{b}_{\mathcal{T}} = \boldsymbol{y}\}, & \text{if } |\mathcal{T}| > n, \end{cases}$$

where $(\cdot)^{+}$ denotes the Moore–Penrose pseudo-inverse.[18] The MNLS is the estimator that the firm would adopt if the residual variance were approximately zero (i.e., the cumulative signal $S(\mathcal{T}) \approx 1$). It comes in two flavors, depending on whether the number of parameters is greater than the sample size:

- Underparametrized regime ($n \geq |\mathcal{T}|$): the MNLS estimator coincides with the OLS estimator, which is uniquely defined because $\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}}$ is invertible.
- Overparametrized regime ($n < |\mathcal{T}|$): the OLS estimator is not defined because the system $\boldsymbol{X}_{\mathcal{T}}\boldsymbol{b}_{\mathcal{T}} = \boldsymbol{y}$ has infinitely many solutions; the MNLS chooses the solution with the smallest Euclidean norm.

The MNLS is useful because it is well-defined in both regimes and coincides with the maximum-likelihood estimator. The Bayes estimator is a shrinkage of the MNLS estimator towards the prior mean $\mathbf{0}_{|\mathcal{T}|}$

Corollary 10. *The Bayes Estimator is the MNLS estimator with shrinkage:*

$$\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{T}} \mid \boldsymbol{M}_{\mathcal{T}}\right] = \left(\underbrace{(1 - S(\mathcal{T}))}_{\textit{Shrinkage Factor}} \cdot \left(\boldsymbol{X}_{\mathcal{T}}'\boldsymbol{X}_{\mathcal{T}}\right)^{+} + \boldsymbol{I}_{\mathcal{T}}\right)^{-1} \hat{\boldsymbol{\beta}}_{\mathcal{T}}^{\textit{MNLS}}.$$

Because it is the maximum likelihood estimator, the MNLS estimator attributes all the variation in the learning matrix $\boldsymbol{M}_{\mathcal{T}}$ to the parameters $\boldsymbol{\beta}_{\mathcal{T}}$. In reality, a fraction $1 - S(\mathcal{T})$ of the variation in $\boldsymbol{y}$ is residual variance and not due to $\boldsymbol{\beta}_{\mathcal{T}}$. The posterior mean corrects for this by shrinking $\hat{\boldsymbol{\beta}}_{\mathcal{T}}^{\text{MNLS}}$towards the prior mean $\mathbf{0}_{|\mathcal{T}|}$ with a shrinkage factor equal to the residual

[18] For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, the Moore–Penrose pseudo-inverse is the unique matrix $\boldsymbol{A}^{+} \in \mathbb{R}^{m\times n}$ satisfying

$$\boldsymbol{A}\boldsymbol{A}^{+}\boldsymbol{A} = \boldsymbol{A}, \quad \boldsymbol{A}^{+}\boldsymbol{A}\boldsymbol{A}^{+} = \boldsymbol{A}^{+}, \quad (\boldsymbol{A}\boldsymbol{A}^{+})' = \boldsymbol{A}\boldsymbol{A}^{+}, \quad (\boldsymbol{A}^{+}\boldsymbol{A})' = \boldsymbol{A}^{+}\boldsymbol{A}.$$

variance $1 - S(\mathcal{T})$. Adding a new covariate $j \notin \mathcal{T}$ reduces the residual variance by s_j, the variance of x_j, thereby lowering the shrinkage factor and the weight of the prior mean. Hence, the posterior mean moves closer to the MNLS estimator. Hence, covariates lend precision to each other: observing a new variable improves the accuracy of the estimated parameters of the others.

B.3 Double Descent

Corollary 11. *If covariates in $\mathcal{L}$ are highly informative, the Bayes Estimator is equivalent to the ridgeless estimator and the MNLS estimator*

$$\lim_{S(\mathcal{L})\to 1^-} \mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}\right] = \lim_{\lambda\to 0^+} \hat{\boldsymbol{\beta}}_{\mathcal{L}}^{ridge}(\lambda) = \hat{\boldsymbol{\beta}}_{\mathcal{L}}^{MNLS}.$$

In general, sophisticated algorithms are needed to compute or approximate the posterior mean $\mathbb{E}\left[\boldsymbol{\beta}_{\mathcal{L}} \mid \boldsymbol{M}_{\mathcal{L}}\right]$. Instead, the MNLS can be obtained by a simple machine learning algorithm, *gradient descent*. This equivalence therefore shows that once the data is sufficiently rich, even such a rudimentary algorithm approximates the Bayes estimator arbitrarily well. When data is linear-separabler, prediction accuracy is driven almost entirely by data, not by algorithms.

Remark 2. The result also sheds light on a central puzzle in modern statistics and machine learning: the double descent phenomenon first discusssed in Belkin et al. (2019). Classical statistics tells us the prediction error of gradient descent is U-shaped in the number of parameters $|\mathcal{L}|$: with too few parameters the model underfits, while beyond the optimum $|\mathcal{L}|^* \in (0, n)$ prediction error increases due to overfitting, as residual variation $\boldsymbol{\varepsilon}$ is mistakenly attributed to $\boldsymbol{\beta}_{\mathcal{L}}$. However, empirical work shows that expanding $\mathcal{L}$ further can reduce the error again—the second descent in the error. Double descent is not yet fully understood: the dominant explanations rely on intricate properties of high-dimensional geometry (see Hastie et al. (2020)). Our model offers a simpler account that also applies to low-dimensions. As the learning set $\mathcal{L}$ expands, the residual variance $1 - S(\mathcal{L})$ decreases, and the shrinkage operator in the Bayes estimator vanishes. When $S(\mathcal{L}) \approx 1$, the Bayes estimator is arbitrarily close to the MNLS even in finite samples, so gradient descent is approximately optimal.

B.4 Connection with Shannon's Information Theory

Remark 3. Let a real-valued additive white Gaussian residual variance (AWGN) channel be given by

$$y = w + z, \qquad z \sim \mathcal{N}(0, \sigma^2),$$

with an input power conslearnt $\mathbb{E}[w^2] \leq P$. Classical results due to Shannon (1948) show that the mutual information between w and y is[19]

$$I(w; y) = \frac{1}{2} \log_2 \left(1 + \frac{P}{\sigma^2}\right) \quad \text{nats.} \tag{R.1}$$

If the channel is decomposed into independent "frequency" slices indexed by $j \in \mathcal{T}$ that each carry an SNR of

$$\text{SNR}_j = \frac{s_j}{\lambda^*},$$

then (R.1) adds up across slices by orthogonality. The total mutual information revealed by a learning sample of *strength* t is therefore[20]

$$I_{\mathcal{T}}(\lambda^*) = \frac{1}{2} \sum_{j \in \mathcal{T}} \log_2 \left(1 + \frac{s_j}{\lambda^*}\right). \tag{R.2}$$

Equation (R.2) is exactly the functional that appears in our model. Thus the economic value function I study,

$$v(t) = \sum_{j \in \mathcal{T}} \frac{t\, \lambda_j}{1 + t\, \lambda_j},$$

equals

$$v(\mathcal{L}, \mathcal{T}) = 2 \left(\frac{I'_{\mathcal{T}}\left(\lambda^*(\mathcal{L})\right)}{\lambda^*(\mathcal{L})} - I_{\mathcal{T}}\left(\lambda^*(\mathcal{L})\right) \right),$$

linking our "value of accuracy" directly to the canonical Shannon measure of information. Two substantive insights follow:

1. **Capacity-driven diminishing returns.** Because $I''(t) < 0$ by Shannon's law, marginal economic value $v'(t) = 2I'(t)$ must also fall. No additional curvature assumption is needed; the concavity of v is pinned down by fundamental information limits. In policy terms, data economies of scale saturate exactly when further capacity gains are information-theoretically expensive.

[19] See C. E. Shannon, *Bell System Technical Journal*, 1948, eq. (26); or T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., §9.1.

[20] This integral form follows immediately from Gallager, *Information Theory and Reliable Communication*, 1968, Ch. 8, where parallel Gaussian sub-channels are treated.

Table 2: Types of predictions and policy implications

Type of prediction	**Data abundant?**	**Tails thick?**	**Monopoly Remedy**
Genomic risk prediction (health)	No	Yes	Access regulation
Clinical decision support for rare diseases	No	Yes	Access regulation
Credit scoring / SME default probability	No	Yes	Access regulation
Fraud / AML detection	No	Yes	Access regulation
Industrial predictive maintenance (OEM IoT)	No	Yes	Access regulation
Smart grid anomaly detection (critical infra)	No	Yes	Access regulation
Autonomous driving safety edge cases	Yes	Yes	Hybrid
Weather nowcasting for extremes	Yes	Yes	Hybrid
E-commerce CTR / product recommendation	Yes	No	Competition policy
Targeted Ads	Yes	No	Competition policy
Media streaming recommendation	Yes	No	Competition policy
Web search ranking	Yes	No	Competition policy