A Theory of the Value of Data in Prediction*

Giovanni Colla Rizzi†

May 26, 2025

Abstract

We develop a Bayesian model in which data scale (n) and feature number(k)—the resources that turn raw data into actionable insights—jointly shape predictive performance and market outcomes. Data expansions boost accuracy but face diminishing returns, whereas more sophisticated analytics yield superadditive gains that eventually plateau. Critically, data and algorithms serve as complements at low levels but substitute for each other once one dimension becomes large. Applying these insights to dynamic platform and data monopsony settings reveals that incumbents leverage extensive data stocks yet can lose ground if rivals outpace them in algorithmic sophistication. Monopsonists under-acquire data, distorting social welfare, when they invest in technology to reduce reliance on costly new observations. Taken together, our findings challenge the "data-equals-power" narrative by underscoring how diminishing returns, technological advancements, and strategic data purchasing affect competition. Policy implications include the need for targeted interventions—such as data-sharing mandates in early-stage (complementary) regimes and support for algorithmic innovation that can spur entry—in order to balance innovation incentives and prevent entrenched data-driven dominance.

*I thank Patrick Rey for his clairvoyant guidance. I thank Pierre Azoulay, Jad Beyhum, Michele Bisceglia, Jin Chuqing, Jacques Crémer, Jean-Pierre Florens, Eric Gautier, Andrei Hagiu, Johannes Hörner, Doh-Shin Jeon, Bruno Jullien, Hiroaki Kaido, Giacomo Lanzani, Simon Loertscher, Friedrich Lucke, Edvin Villmones Mæhre, Pablo Mileni Munari, Giovanni Morzenti, Nour Meddahi, David Salant, Enrico Mattia Salonia, Tim Simcoe, Alex Smolin, Ehsan Valavi, Marshall Van Alstyne, Davide Viviano, Julian Wright, Wuenxuan Xu, and the participants in the TSE I.O. Workshop for their valuable comments.

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

1 Introduction

In recent years, the idea that "data is the new oil" has gained currency in both public discourse and academic research. The analogy suggests that large-scale data acts as an indispensable resource in the digital economy—concentrating market power in the hands of data-rich firms, raising barriers to entry, and arguably necessitating new regulatory frameworks. Proponents of this view often emphasize that assembling vast user datasets confers a decisive advantage for machine learning algorithms, making data ownership the key to monopolizing predictive accuracy. This perspective has informed debates in antitrust, competition policy, and privacy regulation, with proposals ranging from forced data sharing ("essential facilities") to data portability mandates.

Despite its intuitive appeal, however, recent studies and industry trends indicate that the relationship between data and algorithmic performance is more nuanced than a simplistic "data = power" view implies. Economists and data scientists have increasingly stressed that data is subject to diminishing returns and crucially dependent on complementary inputs such as algorithmic sophistication. As a result, expanding the sheer volume of data is not always sufficient to maintain a durable edge in prediction. A variety of real-world examples illustrate how improvements in algorithmic design can partially substitute for data volume. Firms that once relied on massive datasets (e.g., Google Translate in its early statistical machine translation days) gradually reduced their dependence on specialized data when they adopted neural architectures that could transfer knowledge across languages. Similar patterns of data–technology substitution have emerged in industries ranging from predictive maintenance to targeted advertising. Cutting-edge Large Language Model DeepSeek claims to have been trained with a fraction of the data used to train ChatGPT

Motivated by these tensions, this paper develops a simple theoretical model clarifying how two distinct factors combine to shape predictive accuracy: *data scale*, denoted by n, and *algorithmic capital*, denoted by k. Concretely, n represents the quantity (or breadth) of observations available, while k reflects the technological capacity to extract information from each data point—an investment in more sophisticated models, richer features, or advanced analytics. This distinction helps us illuminate precisely when more data and better models act as complements and when they substitute for one another. In a Bayesian linear-regression setup, we derive three robust empirical patterns:

1. **Diminishing returns to data scale (n).** In line with the Law of Large Numbers and standard estimation theory, each additional observation yields progressively smaller incremental gains in predictive accuracy once model parameters are sufficiently well-identified.

2. **Increasing-then-decreasing returns to feature number(k).** Investing in better technology initially has superadditive effects—each improvement in k boosts the marginal value of existing features—but eventually plateaus, underscoring that while advanced analytics can be transformative, it too faces diminishing returns at high levels.

3. **Complementarity when both n and k are low, substitutability when they are high.** At early stages, expansions in data scale and feature numberreinforce each other, but once either dimension is sufficiently large, further growth in one can reduce the marginal returns to investing in the other.

These findings echo observed industry practices. At early stages, leading technology firms often exert substantial effort to accumulate vast amounts of user data (high n), capitalizing on scale-driven feedback loops. As their feature number(k) matures, however, these same firms pivot from brute-force data collection to more selective methods—developing, for instance, sophisticated deep-learning models or leveraging transfer learning. Our model formalizes how such a *Data–Technology Substitution Threshold* emerges naturally from the interplay of data scale (n) and feature number(k).

We develop several applications:

- **Prediction Monopoly and Data Valuation:** A firm selling a horizontally differentiated good benefits from predictive accuracy to better match products to consumer preferences. Our model shows that the firm's willingness to purchase data (n) depends on the trade-off between data scale and feature number(k). Specifically, we find that firms facing high per-unit data costs prioritize algorithmic improvements over additional sample collection, reinforcing the substitutability of data and technology once the dataset is already large. This implies that, in data-rich environments, further investments in collecting observations may be less attractive, while smaller firms or new entrants, who are still in the complementarity phase, benefit from acquiring additional consumer data. From a policy perspective, this insight suggests that data-sharing mandates may be most effective in industries where scale and capital are still

in the complementary region. Conversely, in mature markets where substitutability dominates, forcing incumbents to share data may yield diminishing competitive benefits.

- **Platform Competition and Entry Barriers:** In a dynamic setting where platforms compete over time, our model predicts that incumbents with a large stock of historical data (n) enjoy a significant advantage, especially when predictive technology (k) is in a complementary phase. However, once technology matures, new entrants can compete effectively by leveraging algorithmic innovations rather than relying solely on catching up in data acquisition. This nuance challenges the common narrative that "big data" permanently locks in incumbents. Instead, it shows how technological progress can erode incumbency advantages over time. Policymakers concerned with platform competition could thus encourage algorithmic innovation (e.g., R&D tax credits or support for AI research), rather than focusing exclusively on data-sharing regulations.

- **Data Monopsony and incentives to invest in algorithms**: When firms acquire data from users or third-party suppliers, they may act as monopsonists in data markets. Our framework shows that under certain conditions, firms under-purchase data because the marginal value of additional data diminishes at high n. If data scale and feature numbereventually become substitutes, firms' reliance on data alone declines, potentially reducing compensation for data suppliers. These findings highlight a possible justification for policies that protect or augment data suppliers' bargaining positions. If a firm's monopsony power leads to inefficiently low data acquisition, requiring fair compensation or revenue-sharing could bring outcomes closer to the social optimum. Conversely, if the firm and data remain in a region of strong complementarity, ensuring open access to large user datasets may be more critical to achieving efficiency.

These insights speak directly to debates in antitrust and digital-market competition. Regulators often worry that the "new oil" of massive user data confers insurmountable advantages to incumbents, but our results suggest this advantage may be constrained if (i) diminishing returns kick in, and (ii) competitive firms develop superior algorithms to glean similar insights from fewer observations. From this standpoint, proposals such as data portability

mandates—designed to help smaller firms catch up on "big data"—may be most valuable in the early stages of technology adoption, when across-individual data and within-individual granularity are still complements. Once technology matures, large swaths of additional user data may become less pivotal than algorithmic innovation.

A second policy dimension concerns data-minimization regulations such as the General Data Protection Regulation (GDPR). Our framework indicates that restricting the volume of data available might reduce incumbents' reliance on brute-force scale, potentially increasing the returns to investing in richer but more privacy-friendly data sources, or in high-quality modeling techniques. Paradoxically, firms with advanced algorithms could be *less* impacted by data minimization, because at high levels of technological sophistication, data and technology become substitutes. Our theoretical model thus underscores the importance of carefully calibrating privacy and competition policies to the context of actual data returns, rather than applying broad-brush assumptions about data's "oil-like" properties.

The rest of the paper is structured as follows. Section 3 presents our baseline univariate regression model and derives the three stylized facts about returns to across-individual and within-individual information. Section 4 generalizes the analysis to a multi-covariate context, connects it to standard econometric methods (e.g. ridge regression), and examines how the value of data changes in high-dimensional settings. Section 6 explores the model's implications in scenarios of monopoly pricing, data monopsony, and platform competition, highlighting how sample size and feature depth interact to shape market outcomes. Section 7 concludes with policy remarks on data sharing, regulation, and the evolving role of algorithmic sophistication in modern digital markets.

By emphasizing that data alone need not be destiny—and that algorithmic capacity mediates the benefits of data at scale—this paper contributes to a more nuanced, evidence-based framework for understanding data's role in competition and innovation. In doing so, we hope to inform ongoing debates around "data as the new oil," shedding light on when data truly constitutes a binding constraint and when technological advances can (and do) reduce dependence on sheer volume.

2 Relevant Literature

The paper is at the intersection of four strands of literature:

The Value of Data. This paper characterizes separately how scale (the number of observations) and scope (the number of covariates) determine the value of data. Empirical and theoretical work by Varian (2018), Bajari et al. (2019), and Schaefer and Sapi (2023) finds that expanding a dataset's breadth of users yields diminishing marginal returns. Schaefer and Sapi (2023) also show empirically that complementarities exist between the number of observations and the number of covariates. Carballa Smichowski et al. (2022) provide empirical evidence of economies of scope, where adding covariates improves predictions with increasing returns. This paper:

1) Formalizes these empirical findings in a coherent theoretical framework. 2) Generalizes these properties, showing that complementarities between dimensions are specific to small datasets, while in large datasets, observations and covariates are substitutes. 3) Provides clear statistical explanations demonstrating that returns to observations and covariates follow predictable patterns determined by statistical laws. Specifically, economies of scope arise because when the number of covariates varies, the level of misspecification is endogenous.

Data and Platform Competition. The analysis shows how prediction generates value in recommendation systems where platforms compete to offer recommendations to users positioned at an unknown location on the Hotelling line. It demonstrates that data collection is a strategic substitute, analogous to quantity competition in Cournot (1838). The paper relates to Hagiu and Wright (2021), who distinguish between across-user and within-user data and introduce the concept of a data-quality feedback loop. In their framework, multi-sided platforms leverage data to strengthen network effects: larger datasets improve predictions, increasing platform quality, attracting more users, and further expanding the dataset. Similarly, Prüfer and Schottmüller (2021) show that firms with large user bases benefit from a reinforcing feedback loop, where additional user data reduces per-user investment costs in product improvements, making it harder for smaller competitors to catch up.

Bayesian Statistics and Random Matrix Theory. Methodologically, this paper develops a Bayesian linear regression model with a variable number of regressors under quadratic loss. Bayesian linear regression is a well-established approach covered in DeGroot (2005)

and Berger (1990), where the value of data corresponds to the reduction in posterior variance relative to prior variance. This framework enables a transparent decomposition of how user-level covariates and training observations contribute to reducing predictive uncertainty. This paper innovates by:

1. Allowing the regression residual to be endogenous, depending on the number of covariates observed—a key factor in determining economies of scope.

2. Addressing the technical issue that the posterior variance of Bayesian regression coefficients is random, as it depends on the data itself. Using high-dimensional asymptotics from Marčenko and Pastur (1967), popularized in theoretical machine learning applications by Hastie et al. (2020), the paper derives simple expressions for the value of data as a function of the number of observations and covariates in large-dimensional datasets. Applying these tools in a Bayesian framework is particularly revealing, as ridge regression naturally emerges as a regularization technique to account for data noise.

3 A Simple Model of Prediction

In this section we will develop a reduced form model of prediction. We characterize the value of the information contained in a training dataset and we show how it depends on across-individual information (i.e. , the amount of data or the sample size) and within-individual information (i.e., the level of complexity in our prediction technology or the amount of information we have on every sample). We show that even modeling learning using the most basic econometric model, univariate linear regression, we can establish three stylized facts on the returns to scale of datasets: there are decreasing returns to across-individual information, increasing returns to within-individual information, complementarities between across- and within-individual information when information is scarce and substitutability when information is abundant.

3.1 Setup

Prediction Problem A decision-maker M must predict a target variable $y \in \mathbb{R}$ for a continuum of individuals indexed by $i \in \mathcal{I}$, where the total mass of individuals is normalized

to one. Denoting by $\hat{y}_i$ the prediction for individual i, M incurs in a quadratic loss given by:

$$L\left(y_i, \hat{y}_i\right) \equiv \int_{\mathcal{I}} \left(y_i - \hat{y}_i\right)^2 di.$$

We assume that y_i is i.i.d. across individuals with mean 0 and variance $\sigma^2 \geq 0$. The parameter σ^2 reflects the *difficulty* of the prediction problem.

Data-Generating Process On each individual $i \in \mathcal{I}$, M observes a covariate x_i, which is i.i.d. with mean 0 and variance $S \in [0, 1)$, which is the signal in the data. The relationship between x_i and y_i is:

$$y_i = \beta x_i + \epsilon_i,$$

where we assume:

1. β is unknown Gaussian and common across individuals with prior mean 0 and variance σ^2.
2. β, x_i, and ϵ_i are mutually independent.
3. ϵ_i is an independent individual-specific noise, with mean 0 and variance $\sigma^2\left(1 - S\right)$.

As $\mathbb{E}[\beta] = 0$, it follows that $\mathbb{E}\left[y_i | x_i\right] = 0$. Therefore, in the absence of any additional knowledge about β, the optimal prediction is trivially the prior mean $\hat{y}_i = 0$. This implies that M must first acquire information about β to make use of x_i for prediction.

Learning M can purchase a training dataset $\{(x_j, y_j)\}_{j=1}^n$ from the same population to update beliefs about β, with $n \in \mathbb{N}_*$. The value of information (VoI) in $\{(x_j, y_j)\}_{j=1}^n$ depends on the improvement in predictive accuracy M can achieve by training thereon:

Definition 1 (Value of Information). *Let $\hat{y}_n^*$ denote the optimal predictor after observing n data points, and let $\hat{y}^* = 0$ denote the optimal predictor without data. The Value of Information ("VoI") of a dataset $\{(x_j, y_j)\}_{j=1}^n$ is:*

$$VoI\left(\{(x_j, y_j)\}_{j=1}^n\right) = \int_{\mathcal{I}} \mathbb{E}_{y_i}\left[\left(y_i - \hat{y}^*(x_i)\right)^2 - \left(y_i - \hat{y}_n^*(x_i)\right)^2\right] di.$$

3.2 The Value of Data

The following result characterizes the value of information in large samples. Intuitively, this approximation is valid in modern machine learning models which are trained with large samples of data.

Theorem 1 (Asymptotic Value of Information)**.** *As $n \to \infty$, the VoI converges asymptotically to a function of n and k:*

$$VoI\left(\{(x_j, y_j)\}_{j=1}^n\right) = S\left(\mathbb{V}\left[\beta\right] - \mathbb{V}\left[\beta|\{(x_j, y_j)\}_{j=1}^n\right]\right) \sim V(n, k),$$

where

$$V(n, S) \equiv \underbrace{S}_{\text{Specification}} \cdot \underbrace{\frac{\sigma^2}{1 + \frac{1}{n}\left(\frac{1}{S} - 1\right)}}_{\text{Estimation}}.$$

Conceptually the VoI is the product of two terms:

1. Specification term*: this captures the knowledge about target individuals, i.e., how informative the covariate x_i is about y_i. A higher S increases the maximum amount of information that could theoretically be extracted from each x_i with an infinite number of samples.*

2. Estimation term*: this captures the knowledge about β. This reflects the level of understanding of the relationship between x_i and y_i. how much additional precision is gained from increasing the sample size n. As n grows, this term approaches 1, meaning β is fully learned.*

Across-individual Learning *The VoI depends on n only though the estimation term: as individuals are i.i.d. the only information relevant to predicting y_i for the target individuals from different individuals in $\{(x_j, y_j)\}_{j=1}^n$ is that which pertains to the estimation of β. $V(n, S)$ is increasing and concave in n*

Corollary 1 (Decreasing Returns to Scale)**.** *The marginal value of data*

$$mv(n; k) \equiv \frac{\partial V(n, S)}{\partial n} = \frac{\sigma^2 (1 - S)}{\left(n + \frac{1}{S} - 1\right)^2} \quad \text{is decreasing in } n.$$

This is because the estimation term is increasing and concave in n. This follows from the Law of Large Numbers: as n increases, the estimation of β becomes more precise, but each additional observation reduces uncertainty on β by less than the previous ones. Thus, VoI exhibits diminishing

returns in sample size. This is coherent with the results in Bajari et al. (2019) and Schäfer et al. (2018).

Within-individual Learning *The VoI is increasing and convex in k as the technology has two positive effects on the value of data, the former of which is linear.*

$$\frac{\partial V(n;S)}{\partial S} \equiv \underbrace{S \cdot \frac{\sigma^2}{1+\frac{\frac{1-S}{S}-1}{n}}}_{\text{Specification Effect}} + \underbrace{\frac{\sigma^2}{\left(n+\frac{1}{S}-1\right)^2}(1-S)}_{\text{Estimation Effect}} > 0.$$

A better technology k has two effects on $V(n;S)$:

1. Specification effect *(SE): a higher k increases knowledge about target individuals as x_i explains a greater fraction of the variance in y_i.*

2. Estimation effect *(EE): a higher k improves the estimation of β because the fraction of variance in $(y_i, x_i)_{i=1}^n$ due to noise $\frac{1-S}{S}$ decreases, meaning that updates on β are more efficient.*

The double derivative of $V(n;k)$ is strictly positive, as the SE is increasing in k and this effect always dominates the EE, which is inverted-U shaped in k.

Corollary 2 (Economies of Scope)**.** *The marginal value of additional within-individual information is S-shaped:*

$$\frac{\partial^2 V(n;S)}{\partial^2 S} \equiv \underbrace{\frac{\partial SE}{\partial S}}_{>0} + \underbrace{\frac{\partial EE}{\partial S}}_{>0} > 0.$$

This is consistent with the economies of scope to data: as information on individual is contextual, additional units of information allow to "place into context" previously collected information, increasing the latter's usefulness. This has a clear statistical meaning: as S increases, the fraction of variability in the data which is due to noise decreases; therefore M will increase its reliance on the empirical estimates and reduce the influence of the prior. This compounding effect implies that there are increasing returns to within-individual learning. However, the decreasing returns to k reflected in α contrast these increasing returns. Note also that the scope for increasing returns is decreasing in n. This implies that firms with less data benefit from larger economies of scope. Note that these results are coherent with the findings of Carballa Smichowski et al. (2022) who find that there are S-shaped returns to additional of covariates in a prediction model using health data.

Across- and within-individual Learning *It is interesting to explore how better knowledge a across individuals affect the value of learning within-individual and vice verse. To study this context vs specialisation problem we study the cross derivative* $\frac{\partial^2 V(n;k)}{\partial k \partial n}$ *or equivalently we analyze how the marginal value of data* $mv(n;k)$ *changes as within-user learning* k *increases:*

Corollary 3. *The marginal value of data is increasing in the within-individual information* k *if and only*

$$\frac{\partial mv(n,S)}{\partial S} = \frac{\partial^2 V(n,S)}{\partial S \partial n} = \underbrace{\frac{\partial SE}{\partial n}}_{>0,} + \underbrace{\frac{\partial EE}{\partial n}}_{\geq 0 \iff n \leq \frac{1-S}{S}} \geq 0 \iff n \leq \tilde{n}(S) \equiv 1 + \frac{1}{S}\left(\frac{2}{S} - 3\right).$$

Equivalently, across- and within-individual information are complements when they are small and substitutes when they are large.

The marginal value of data is inverted U-shaped in within user learning A. The following proposition draws conclusions for the marginal value of data as a function of the amount of data n*. To understand this, let's break it down the impact of* n *on the specification and estimation effects of* A *on* $V(n,S)$*:*

1. *SE is increasing: as higher* n *means M has a better knowledge about* β *the relationship between* x_i *and* y_i*, this increases the marginal value of gaining extra knowledge on the target individuals by increasing* S*. This effect is decreasing in* n *due to diminishing returns, and therefore dominates for small* n*.*

2. *EE is increasing if* $n \leq \frac{1-S}{S}$ *and decreasing otherwise. Intuitively, for small* n*, each additional data point carries a lot of new information about* β*. Hence, a rise in* S *(which increases the signal to noise ratio of the dataset) strengthens that information gain. However, for large* n *the learning about* β *is already very precise (diminishing returns to adding even more data). Even though* k *makes each data point more informative, once* n *is large, adding yet another data point has a smaller incremental contribution in reducing the uncertainty about* β*.*

This implies that SE drives a complementarity between data and technology but the EE entails that they are complements when data is scarce and technology rudimentary nut substitutes when data is abundant and the technology is sophisticated.

This finding is consistent with Figure 7 in Schaefer and Sapi (2023), which compares the marginal value of n *as a function of* k *for words in different deciles of number of searches* (n)*, and shows that*

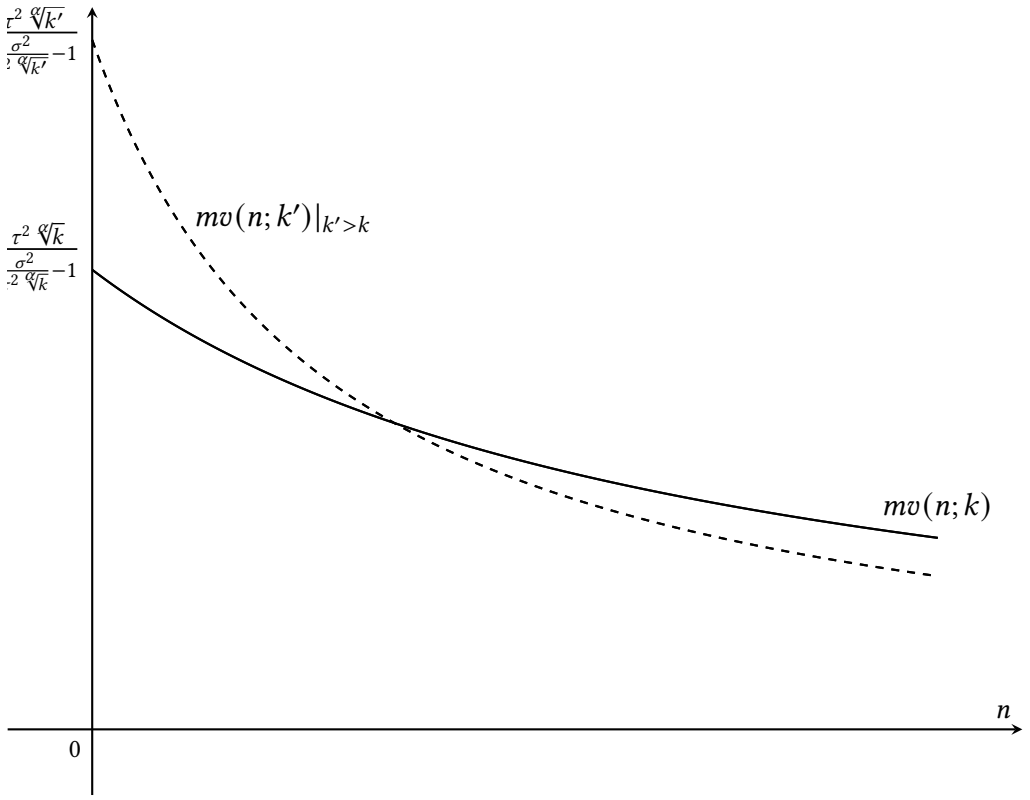

Figure 1: This figure illustrates the effect of an increase in k on the marginal value of data. There exists a unique interior point such that is increasing in for and decreasing for . Equivalently, an increase in makes rotate clockwise around $\tilde{n}(k)$. The solid curve represents the marginal value of data $mv(n;k)$, while the dashed curve corresponds to the marginal value function after an increase in signal strength, $k' > k$

the marginal value of n is increasing in k for words with a short search history (small n), but inverted U-shaped in k for words with a long search history (large n), a finding which confirms the insight that the dimensions of data are complements for small datasets and substitutes for large ones. This finding suggests a general pattern:

- *In the early stages of technological development, firms depend heavily on gathering vast amounts of data.*
- *As predictive algorithms and modeling capabilities improve, technology becomes a substitute for data, enabling firms to achieve comparable (or even superior) predictive performance with less incremental data.*
- *This shift allows mature firms to focus less on data volume and more on the quality of models and aggregation of different data sources.*

This theoretical insight into the Data-Technology Substitution Threshold *not only helps explain these industry trends but also has implications for data regulation. Firms that possess cutting-edge*

modeling technology will need far less data to maintain competitive performance, potentially reducing their vulnerability to data minimization regulations *like GDPR. Conversely, firms with* less sophisticated algorithms *will continue to depend heavily on expansive data collection to achieve competitive prediction performance.*

In summary the preceding study establishes three stylized facts about across- and within-individual learning.

Fact 1. *There are decreasing returns across-user learning*

Fact 2. *There are S-shaped or decreasing returns to within-user learning.*

Fact 3. *Across- and within-individual learning are complements when they are scarce and substitutes when they are large.*

4 Ridge Regression

A drawback of the reduced model explored above is that it abstracts form issues of dimensionality in estimation. Specifically, it assumes that the platform can increase the variance in the data, whilst keeping constant the number of covariates used in the regression. A more realistic model would model the choices of a platform which must choose how many covariates to sample knowing each one has a given variance. Such a model should take into account that increasing the number of covariates increases the number of linear parameters to be estimated, which comes with additional noise in estimation. We identify across-user learning as the number of observations in a dataset and within-user learning as the number of covariates per user. We therefore assume that the individual is take from a different decision-making process:

Data-Generating Process On each individual $i \in \mathcal{I}$, M observes a vector of individual covariates $z_i \in \mathbb{R}^Z$, with $Z \in \mathbb{N}_*$, which are i.i.d. with mean 0 and variance $1/Z$. The relationship between z_i and y_i is:[1]

$$y_i = \boldsymbol{\beta}' z_i,$$

[2] where we assume:

[1] The normalization of covariate variance ensures that the variance of vector z is 1 and the variance of y_i does not depend on Z.

[2] We denote all vectors as column vectors.

1. $\boldsymbol{\beta} \in \mathbb{R}^Z$ is an unknown vector of coefficients common across individuals, i.i.d. Gaussians with prior mean 0 and variance σ^2.[3]

2. $\boldsymbol{\beta}$ and $\boldsymbol{z}_i$ are mutually independent.

Data. *M* collects two types of data:

1. For each individual $i \in \mathcal{I}$, a vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^{ZS}$ which is a subvector of $\boldsymbol{z}_i$, where $S \in [0, 1)$ is the fraction of observed covariates. We denote the collection of covariate vectors $(\boldsymbol{x}_i)_{i\in\mathcal{I}}$. Without loss of generality, we assume they are the first components of $\boldsymbol{z}_i$ so that the latter can be partitioned as

$$\boldsymbol{z}_i' = (\boldsymbol{x}_i', \boldsymbol{u}_i'),$$

   where $\boldsymbol{u}_i \in \mathbb{R}^{(1-k)Z}$ are the unobserved covariates. The coefficients associated with these covariates are correspondingly partitioned as $\boldsymbol{\beta}' = (\boldsymbol{\beta}_x', \boldsymbol{\beta}_u')$.

2. Realizations of y and $\boldsymbol{x}$ for individuals $i = 1, ..., nZ$ taken from a population identical to $\mathcal{I}$, with $n \in \mathbb{N}_*$ [4]

$$(\boldsymbol{y}, \boldsymbol{X}) \equiv \{(y_i, \boldsymbol{x}_i)\}_{i=1}^{nZ} \in \mathbb{R}^{nZ\times(1+kZ)}.$$

   We will denote by $\boldsymbol{D} \equiv ((\boldsymbol{x}_i)_{i\in\mathcal{I}}, (\boldsymbol{y}, \boldsymbol{X})) \in \mathcal{D} \equiv \mathbb{R}^{kZ} \times \mathbb{R}^{nZ\times(1+kZ)}$ the dataset observed by *M*.

In Section section 4.1 we will characterize the optimal predictor and in Section section 4.2 we will characterize the value of data as a function of n and k.

[3] In this section we assume that the decision maker has a mean zero prior on all coefficients $m(t) = 0$ and the same level of uncertainty on all the coefficients, theat is $v(t) = 1$. This assumption is can be given both a statistical and a information-theoretical interpretation. From an information-theoretic perspective, the principle of maximum entropy suggests that, in the absence of further information about the relative importance of the individual coefficients in the model, the least informative prior is one that treats all coefficients equally. In this case, setting $v(t) = 1$ implies that the variance is uniformly distributed across all coefficients, reflecting no prior preference or bias regarding the importance of any particular covariate. Statistically, this assumption corresponds to an isotropic prior, meaning that the coefficients are equally uncertain, which is a common choice in high-dimensional Bayesian regression settings. Additionally, this assumption leads to asymptotic efficiency in learning, as it ensures that no single covariate is over- or under-weighted, promoting an even contribution from each covariate as the model is estimated. Therefore, the assumption that $v(t) = 1$ is a reasonable choice as it maximizes the uncertainty about the model parameters in a way that is unbiased and computationally tractable.

[4] We are assuming that prediction and training data have the same covariates. This is a reasonable assumption as the choice is typically technological (e.g. how many data sensors to build into an app).

4.1 Prediction Problem

The problem of prediction with quadratic loss is well-studied in the Bayesian decision theory literature and its solution is the following result which can be found in DeGroot (2005):

Lemma 1. *The optimal predictor is* $\hat{y} : \mathcal{D} \to \mathbb{R}$ *such that*

$$\hat{y}^*(\boldsymbol{D}) = \mathbb{E}[Y|\boldsymbol{D}].$$

To minimize the expected quadratic loss, M predicts the most likely value of Y given the data. The posterior mean optimally trades off the influence of the prior and the data based on their respective variance. Straightforward application of Lemma 1 to Definition 1 yields the following result:

Corollary 4. *The VoI in* $\boldsymbol{D}$ *is the reduction of posterior variance*

$$VoI(\boldsymbol{D}) \equiv \int_{i \in \mathcal{I}} \mathbb{V}[Y] - \mathbb{V}[Y|\boldsymbol{D}]\, di.$$

To characterize VoI $(\boldsymbol{D})$ we characterize the posterior distribution of $Y|\boldsymbol{D}$ in the following result:

Lemma 2. *For all* $i \in \mathcal{I}$*,the posterior distribution of* $Y|\boldsymbol{D}$ *is*

$$Y_i|\boldsymbol{D} \sim \mathcal{N}\left(s_i^*(\boldsymbol{D}), \mathbb{V}\left[s_i^*(\boldsymbol{D})\right] + \sigma^2(1-S)\right),$$

and the optimal predictor conditional on $\boldsymbol{D}$*is*

$$s_i^*(\boldsymbol{D}) = \boldsymbol{x}_i' \mathbb{E}[\boldsymbol{\beta}_x|(\boldsymbol{y}, \boldsymbol{X})],$$
$$\mathbb{V}\left[s_i^*(\boldsymbol{D})\right] = \boldsymbol{x}_i' \mathbb{V}[\boldsymbol{\beta}_x|(\boldsymbol{y}, \boldsymbol{X})]\, \boldsymbol{x}_i.$$

2 shows that the posterior mean is a weighted average of $\boldsymbol{x}_i$, the covariates observed on i, with weights equal to the posterior mean of coefficients $\mathbb{E}[\boldsymbol{\beta}_x|(\boldsymbol{y}, \boldsymbol{X})]$. The independence of $\boldsymbol{z}_i$ across individuals implies that the information in $(\boldsymbol{y}, \boldsymbol{X})$ enters the prediction of Y_i only through the belief on $\boldsymbol{\beta}_x$. Prediction occurs in two steps, first M updates the prior on $\boldsymbol{\beta}_x$ based on $(\boldsymbol{y}, \boldsymbol{X})$, and the uses $\boldsymbol{x}_i$ to personalize the prediction on i based on the updated beliefs on $\boldsymbol{\beta}_x$. By Proposition 2 and Corollary 4, to characterize $V(\boldsymbol{D})$ we need only characterize

the posterior distribution $\boldsymbol{\beta}_x|\,(\boldsymbol{y}, \boldsymbol{X})$.

Proposition 1. *The posterior distribution of* $\boldsymbol{\beta}_x|\,(\boldsymbol{y}, \boldsymbol{X})$ *is*

$$\boldsymbol{\beta}_x|\,(\boldsymbol{y}, \boldsymbol{X}) \sim \mathcal{N}\left(\boldsymbol{t}_x^*\,(\boldsymbol{y}, \boldsymbol{X}), \sigma^2\left(\boldsymbol{I}_{kZ} + \frac{1}{1-S} \cdot \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right),$$

where $\boldsymbol{t}_x : \mathbb{R}^{nZ\times(1+kZ)} \to \mathbb{R}^{kZ}$ *is an estimator defined by*

$$\boldsymbol{t}_x^*\,(\boldsymbol{y}, \boldsymbol{X}) \equiv \left((1-S)\cdot \boldsymbol{I}_{kZ} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1}(\boldsymbol{X}'\boldsymbol{y})\,.$$

Proposition 1 characterizes the posterior distribution of the coefficient vector $\boldsymbol{\beta}_x$ after observing training data $(\boldsymbol{y}, \boldsymbol{X})$, which determines how the observed covariates $\boldsymbol{x}$ relate to the location Y. The posterior mean $\boldsymbol{t}_x^*(\boldsymbol{y}, \boldsymbol{X})$ represents a weighted least squares estimator that balances prior knowledge with empirical evidence from training data. The dependence on $1 - S$ reflects how much uncertainty remains after observing a fraction x of the covariates: when $1 - S$ is large, a substantial portion of variance is still unexplained, increasing the weight of the prior and thus amplifying shrinkage. This mechanism ensures that coefficient estimates are not overly influenced by noise in the data, stabilizing predictions by integrating prior knowledge with empirical observations in a structured way. It is well known that the optimal estimator in a Bayesian linear regression model is the ridge regression estimator with a specifically chosen regularization parameter. The following section shows how the estimator characterized in Proposition 1 can be seen as a generalized ridge regression estimator which closely maps estimators which are convergence points of techniques used in machine learning. This section can therefore be skipped by readers who are not interested in the statistical underpinnings of the paper's results.

4.1.1 Ridge Regression

Ridge regression is a technique used in statistical and econometric modeling to address invertibility issues in regression analysis. When there is a large number of features compared to observations, ordinary least squares (OLS) estimates are unstable because the inverse of $\boldsymbol{X}'\boldsymbol{X}$ is close to being non defined and therefore estimator variance is high. Ridge regression introduces a penalty term that shrinks the estimated coefficients toward zero, thereby reducing overfitting and improving predictive performance. This technique helps improve

out-of-sample predictions by trading off bias for lower variance.[5] The following result illuminates the connection between our results and the ridge estimator.

Corollary 5. *The estimator* $\boldsymbol{t}_x^*(\boldsymbol{y}, \boldsymbol{X})$ *is the unique solution to*

$$\min_{\boldsymbol{t}_x} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}_x||_2^2 + \lambda \boldsymbol{t}_x' \boldsymbol{t}_x \right\},$$

where $\lambda = \lambda(k) \equiv 1 - S$.

The result in Proposition 1 can therefore be interpreted as a ridge regression estimator, where regularization depends the unexplained variance fraction $1 - S$ which depends on the dimensionality of the data. This structure closely resembles modern machine learning algorithms, such as adaptive regularization techniques used in neural networks, gradient-boosted trees and data-driven shrinkage. Thus, Proposition 1 formalizes a Bayesian framework that mirrors the principles of adaptive regularization in real-world predictive algorithms, capturing both prior beliefs and observed information to optimize predictions.

Finally, it is important to note that as more covariates are observed (higher k), a lower unexplained variance $1 - S$ leads to less aggressive shrinkage as k increases. Intuitively, as the model gains access to more informative covariates, the weight placed on the observed data increases, reducing the reliance on prior regularization. This means that for small k, where much of the variance in Y remains unexplained, regularization plays a stronger role in controlling the estimator's variance. Conversely, as k grows, the estimation relies more on observed data, leading to weaker shrinkage and a greater responsiveness to the training sample. This feature suggests that there can be increasing returns to increasing the fraction of covariates collected k: by reducing the reliance on the prior, adding a marginal covariates increases the value of inframarginal covariates *ceteris paribus*.

[5] Mathematically, ridge regression solves the following optimization problem:

$$\min_{\boldsymbol{t}_x} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}_x||_2^2 + \lambda \boldsymbol{t}_x' \boldsymbol{t}_x \right\},$$

where $\lambda \geq 0$ is a tuning parameter that controls the strength of regularization. When $\lambda = 0$, ridge regression reduces to OLS, while larger values of λ increase the shrinkage effect.

4.2 The Joint Value of Data

Theorem 2. *The VoI* $(\boldsymbol{D})$ *is to* $V(n,k)$

$$VoI(\boldsymbol{D}) = S - \left(1 - Tr\left[\left(\boldsymbol{I}_{kZ} + \frac{1}{1-S} \cdot \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]\right). \tag{4.1}$$

The following result exploits the properties of large random matrices using the convergence results in Marčenko and Pastur (1967).

Theorem 3. *As* $Z \to \infty$*,* $VoI(\boldsymbol{D})$ *converges to* $V(n,k)$

$$VoI(\boldsymbol{D}) \to V(n,S) \equiv \frac{\sigma^2}{2}(n+1)\left(1 - \sqrt{1 - \frac{4nS}{(n+1)^2}}\right). \tag{4.2}$$

Theorem #.1 shows that as the number of covariates Z grows, the value of information in a dataset converges to a deterministic function $V(n,k)$, depending only on dataset dimensions rather than specific data realizations. This suggests that in large datasets, individual data points contribute negligibly to overall predictive improvement, making data valuation predictable. This challenges Arrow's Information Paradox (see Arrow (1962)), which states that information's value is unknown until acquired, but once acquired, it has effectively been obtained for free—creating a fundamental obstacle to information markets. The theorem implies that when data is sufficiently granular, its value depends only on its size rather than specific content. The result therefore provides a theoretical foundation for data valuation based on statistical properties of datasets.

The following corollary highlights that the results in Corollary 1 apply also to the multivariate case.

Corollary 6 (Decreasing Returns to Scale)**.** *The marginal value of data*

$$mv(n;k) \equiv \frac{\partial V(n;k)}{\partial n} \quad \textit{is decreasing in } n.$$

The following corollary highlights that the results in Corollary 2 apply also to the multivariate case.

Corollary 7 (Economies of Scope)**.** *The marginal value of additional within-individual infor-*

mation is increasing:

$$\frac{\partial^2 V(n;S)}{\partial^2 S} > 0.$$

The following corollary highlights that the results in Corollary 3 apply also to the multivariate case.

Corollary 8. *The marginal value of data is increasing in the within-individual information k if and only*

$$\frac{\partial mv(n;k)}{\partial k} \geq 0 \iff n \leq \tilde{n}(k) \equiv \frac{1}{2k-1}.$$

Equivalently, across- and within-individual information are complements when they are small and substitutes when they are large

4.3 Multiple Datasets

We now develop a more general model in which P has j datasets, each of which has $n_j Z$ observations and $F(k_j)Z$ non overlapping covariates so that

$$\boldsymbol{D}_j \equiv \left(\left(\boldsymbol{x}_j\right)_{i\in\mathcal{I}}, \left(\boldsymbol{y}_j, \boldsymbol{X}_j\right)\right) = \left\{\left(y_{ij}, \boldsymbol{x}_{ij}\right)\right\}_{i=1}^{n_j Z} \in \mathbb{R}^{n_j Z \times (1+F(k_j)Z)},$$

with $\boldsymbol{x}_j \equiv (x_h)_{h\in j}$ the vector consisting of covariates in j and $\boldsymbol{\beta}_j$ as the corresponding vector of coefficients. We assume there is no overlap of covariates across datasets. We therefore have a collection of datasets $\boldsymbol{D} \equiv \left(\boldsymbol{D}_j\right)_{j=1}^{d}$.

Definition 1. The **contribution of a dataset** $\boldsymbol{D}_j$ to a collection of datasets $\boldsymbol{D}$ is a function defined by:

$$\Delta_j(\boldsymbol{D}_j, \boldsymbol{D}) \equiv V(\boldsymbol{D}) - V(\boldsymbol{D} \setminus \boldsymbol{D}_j).$$

Lemma (1) is unchanged and can applied directly to the new definition of $\boldsymbol{D}$. The following result plays the role of Proposition (2).

Proposition 2. *The posterior distribution of $Y|\boldsymbol{D}$ is*

$$Y|\boldsymbol{D} \sim \mathcal{N}\left(s^*(\boldsymbol{D}), \mathbb{V}\left[s^*(\boldsymbol{D})\right] + \sigma^2\left(1-k\right)\right),$$

where

$$s^*(\boldsymbol{D}) = \sum_{j=1}^{d} \boldsymbol{x}_j' \mathbb{E}\left[\boldsymbol{\beta}_j \middle| (\boldsymbol{y}_j, \boldsymbol{X}_j)_{j=1}^{d}\right],$$

$$\mathbb{V}\left[s^*(\boldsymbol{D})\right] = \sum_{j=1}^{d} \boldsymbol{x}_j' \mathbb{V}\left[\boldsymbol{\beta}_j \middle| (\boldsymbol{y}_j, \boldsymbol{X}_j)_{j=1}^{d}\right] \boldsymbol{x}_j. \tag{4.3}$$

The following proposition characterizes the posterior distribution of coefficients after the observation of a collection of non-overlapping datasets.

Proposition 3. *The posterior distribution of* $\boldsymbol{\beta}_j | (\boldsymbol{y}_j, \boldsymbol{X}_j)_{j=1}^{d}$ *is*

$$\boldsymbol{\beta}_j | (\boldsymbol{y}_j, \boldsymbol{X}_j)_{j=1}^{d} \sim \mathcal{N}\left(\left(\left(1 - F(k_j)\right) \cdot \boldsymbol{I}_{kZ} + \boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1} \boldsymbol{X}_j'\boldsymbol{y}_j, \sigma^2 \cdot \left(\boldsymbol{I}_{kZ} + \frac{1}{1 - F(k_j)} \cdot \boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1}\right).$$

The proof relies on the prior independence of covariates and coefficients, which ensures coefficients are affected exclusively by the covariate which they refer to. The following proposition exploits independence. The posterior of $\boldsymbol{\beta}_j$ the coefficients of covariates in dataset j is not affected by the realization of covariates outside that dataset. Therefore the variance is the sum of the contributions of each dataset and each dataset's contribution is its value, a result contained in the following theorem.

Corollary 9. *The value of a collection of non-overlapping datasets is the sum of the contribution of each dataset*

$$V(\boldsymbol{D}) = \sum_{j=1}^{d} V(\boldsymbol{D}_j),$$

which implies the contribution of each dataset is equal to its value $\Delta_j(\boldsymbol{D}_j, \boldsymbol{D}) = V(\boldsymbol{D}_j)$.

Therefore the application of Theorem naturally yields the following result

Theorem 4. *When* $Z \to \infty$*, the value* $V(\boldsymbol{D})$ *converges to a deterministic function* $v(\boldsymbol{k}, \boldsymbol{n}) : \mathbb{R}^d \times \mathbb{R}^d$*:*

$$V(\boldsymbol{D}) \to v(\boldsymbol{k}, \boldsymbol{n}) \equiv \sum_{j=1}^{d} v(k_j, n_j).$$

5 Applications

5.1 A monopoly problem

Netflix has a continuum of movies it can show to consumers, and movies are ranked from the least comedic to the most comedic, with $\hat{y} \in \mathbb{R}$ reflecting their inherent comediness, so that $\hat{y} = -\infty$ is a total tragedy and $\hat{y} = +\infty$ is utterly comnedic. Each consumer has an unknown preferred variety $y \in \mathbb{R}$ which will give her utility $v \geq 0$, and suffers a quadratic disutility proportional to a scalar $t \geq 0$ for deviations from y so that the utility from purchasing variety y for price $p \geq 0$ is

$$u_y(y, p) = v - t\,(y - \hat{y})^2 - p.$$

Suppose there is a population of consumers of unit mass $\mathcal{I} = [0, 1]$ in which y is distributed according to some distribution with mean 0 and variance σ^2. The expected utility is

$$U(p) \equiv \mathbb{E}_y\left[u_y(p)\right] = v - t\mathbb{E}_y\left[(y - \hat{y})^2\right] - p.$$

The users have as outside option going to the local cinema which will show the mean movie $y = 0$. Doing so will yield

$$u = v - t\sigma^2$$

So Netflix can set a subscription price

$$p = t\left(\sigma^2 - \mathbb{E}_y\left[(y - \hat{y})^2\right]\right)$$

and make a revenue of

$$R = t\left(\sigma^2 - \mathbb{E}_y\left[(y - \hat{y})^2\right]\right).$$

Therefore the increase in profit from collecting a dataset of n observations is precisely

$$R = tV(n; k).$$

Assume M pays a fixed cost c per observation. Then the problem of Netflix will be

$$\max_{c,k} tV(n;k) - cn - rk.$$

The parameter c is the cost of acquiring/maintaining data on many users. It can be a storage cost The paramer r is the cost of collecting a lot of data on one user. It can stem from regulation like GDPR

quite trivially, n^* is decreasing in c and inverted U-shaped in r as quite trivially, k^* is decreasing in k and inverted U-shaped in c.

Corollary 10. *The amount of data purchased $n^*(r)$ is increasing in r if $c \geq \bar{c}(k) \in [0, \infty)$ which is increasing in k and decreasing otherwise.*

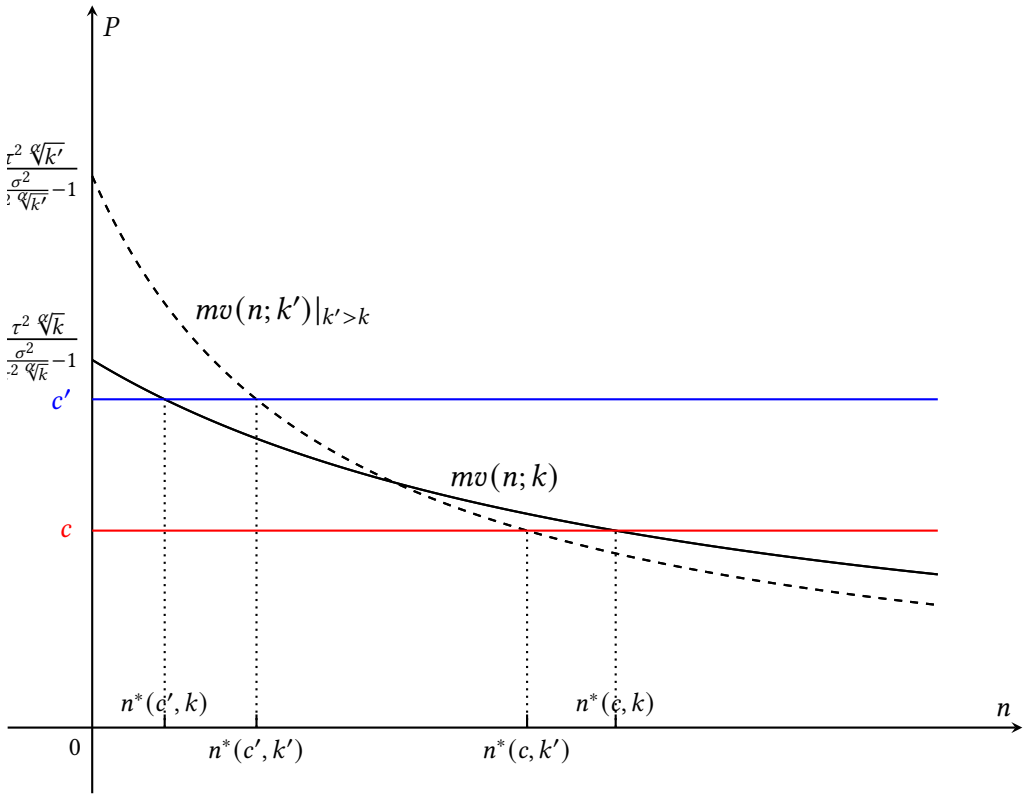

Figure 2: This figure illustrates the effect of an increase in signal strength k on the optimal data acquisition decision. The solid curve represents the marginal value of data $\text{mv}(n;k)$, while the dashed curve corresponds to the marginal value function after an increase in signal strength, $k' > k$. As the signal improves, the optimal number of observations purchased increases, shifting the equilibrium data quantity from $n^*(c,k)$ to $n^*(c,k')$ and from $n^*(c',k)$ to $n^*(c',k')$.

5.2 Dynamic model

We consider a two-period game in which two platforms, an incumbent (I) and an entrant (E), compete to sell predictions to consumers of a horizontally differentiated good. Each consumer has unit demand and can purchase from only one platform in each period. We let $v^j \geq 0$ represent the standalone quality of platform $j \in \{I, E\}$, and let n_t^j denote the number of users who have previously purchased from platform j by period $t \in \{1, 2\}$. The price charged by platform j in period t is $p_t^j \in \mathbb{R}$; it may be negative (i.e., a subsidy). The expected (indirect) utility of a representative consumer who buys from platform j in period t is

$$U_t^j \;=\; v^j + V\big(n_t^j; S\big) \;-\; \sigma^2 \;-\; p_t^j,$$

where $V(\cdot; S)$ is the surplus from predictive accuracy (increasing in n_t^j), σ^2 captures variance costs, and S reflects the level of predictive-analytics technology.

There is a unit mass of consumers arriving each period. The first period has duration 1 and the second period has duration δ, with $\delta \in (0, 1)$. The incumbent starts with a stock of historical data, n, which corresponds to the number of past periods in which it engaged in sales. Firms face no production costs, so if a firm is active in both periods, its total profit is:

$$p_1^j \;+\; \delta\, p_2^j.$$

To ensure nontrivial competition, we assume that

$$\Delta \;\equiv\; v^E - v^I \;\geq\; 0,$$

so that the entrant's standalone quality is not strictly lower than the incumbent's.

5.2.1 Equilibrium Analysis

Proposition 4 (Free Reentry)**.** *There exists a unique threshold* $\hat{\Delta}(\delta, n, k) \geq 0$ *such that E sells in* both *periods if and only if*

$$\Delta \geq \Delta^*(\delta, n, \lambda, S) \equiv \frac{V(n\lambda, S) + \delta\left(V\left((n+1)\lambda, S\right) + V(n\lambda, S) - V(\lambda, S)\right)}{1 + 2\delta}.$$

It is increasing in n and decreasing in δ. If n is large it is inverted U-shaped in S and

λ, otherwise if n is small it is increasing in both. Higher n (the incumbent's data stock) pushes up this threshold because the incumbent's advantage from accumulated data is more difficult to surmount, forcing the entrant to rely on a larger standalone quality edge. The parameter $\delta \in (0,1)$ measures the length of the second period: as δ decreases, the second period becomes 'shorter,' reducing the entrant's incentive to incur first-period competitive costs, thereby decreasing the likelihood of long-term competition. The function $V(\cdot;S)$ is increasing and concave in n and increasing and convex in S. Hence, when n is large, further improvements in S or λ do not increase $V(\cdot;S)$ as significantly, making the threshold behave in an inverted-U shape in S and λ. When n is small, the gains from collecting data in the first period are more pronounced, so Δ^* becomes increasing in S and λ.

5.2.2 Welfare Analysis

Proposition 5 (Social Optimum)**.** *A benevolent social planner would make E sell in both periods if and only if*

$$\Delta \geq \Delta^w \equiv \frac{V(n\lambda, S) + \delta\left(V\left((n+1)\lambda, S\right) - V(\lambda, S)\right)}{1+\delta}.$$

It is inverted U-shaped in S and λ, increasing in n and decreasing in δ. Because $V(\cdot;S)$ is concave in n, the marginal gains of having a larger user base diminish at high n. When n is large and S grows, the function $V(\cdot;S)$ increases at a decreasing rate in n but at an increasing rate in S, thus the overall shape in (S,λ) becomes inverted U-shaped. By contrast, for small n, the direct effect of additional data is more pronounced, so a higher S or λ strictly boosts welfare gains from allowing the entrant to compete in both periods. Finally, since $\delta < 1$, a longer second period (larger δ) makes it relatively more valuable to foster competition over the entire horizon, but higher δ also increases the discount factor for immediate costs/benefits, rendering Δ^w decreasing in δ.

Corollary 11. *Define the excess in incumbency advantage*

$$\Psi(\delta, n, \lambda, S) = \Delta^*(\delta, n, \lambda, S) - \Delta^w(\delta, n, \lambda, S) = \frac{\delta^2\left(V(n\lambda, S) - V((n+1)\lambda, S) + V(\lambda, S)\right)}{(1+\delta)(1+2\delta)}.$$

Increasing in λ, S, δ and n. It represents the 'excess' advantage that an incumbent enjoys (or, equivalently, the additional standalone quality advantage the entrant must have) in order to justify entering two-period competition, relative to what would be socially efficient. Because $V(\cdot;S)$ is increasing and concave in n, additional data from either period may not

always translate into proportionate gains if n is already large. However, the fraction λ of users who generate usable data, along with higher S or a longer second period (δ), can amplify these data-driven benefits. The result is that $\Psi(\delta, n, \lambda, S)$ increases in all parameters (δ, n, λ, S), indicating that as the value of data grows in the marketplace, the gap between private and social incentives to admit new entrants can widen.

Our analysis highlights how an incumbent's initial stock of data, n, confers a persistent advantage that may exceed the socially optimal level. This excess incumbency advantage is magnified when a larger fraction λ of users contributes data, when the predictive-analytics technology S is more powerful, or when the second period is sufficiently long (higher δ). In practice, such dynamics can lead to market structures where new entrants find it increasingly difficult to break in, even if it would be welfare-enhancing to do so.

One policy approach suggested by these findings is to mitigate entrenched advantages through data-portability or data-sharing requirements. For instance, measures akin to the European Union's *General Data Protection Regulation* (GDPR) facilitate user mobility by allowing consumers to transfer their historical data to alternative platforms. The proposed *Digital Markets Act* (DMA) in the EU also discusses obligations for 'gatekeeper' platforms to ensure interoperability and data-sharing, which could help entrants close the gap in predictive accuracy. Similar policies are emerging in various jurisdictions, such as open-banking initiatives that require incumbent financial institutions to share consumer banking data with licensed challengers. By reducing the incumbent's data advantage, such interventions effectively lower Δ^* toward the socially optimal threshold Δ^w.

Another potential remedy involves promoting collaboration among incumbents, entrants, and public institutions to create open data pools or standardized data formats. These arrangements can reduce duplication of data-gathering efforts, cut entry costs, and foster competition on overall service quality rather than on raw data. In some cases, competition authorities might consider mandating structured data access for qualified entrants, subject to privacy and security safeguards. These policy instruments aim to realign private incentives with social optima, ensuring that competition is neither stifled by excessive data-based barriers nor distorted by free-riding concerns. Overall, our model underscores that data-based network effects can be powerful and self-reinforcing, demanding careful regulatory scrutiny to safeguard dynamic competition and innovation.

5.3 Data Monopsony

We assume M is a monopsonist and faces an upward sloping supply curve for data $P(n)$, meaning acquiring more data becomes more costly.[6] Therefore M must solve:

$$\max_{n,k} \Pi(n) \equiv V(n,k) - cnP(n) - rkW(k).$$

This model can be solved analogously to a monopsony model, where M is analogous to a monopsonist purchasing labor, but instead, here, it is acquiring data.

Theorem 5. *Provided second-order conditions hold, the monopsonist's problem has a unique solution (n^*, k^*)characterized by*

$$\begin{aligned} \frac{\partial}{\partial n} V(n^*, k^*) &= c\left(P(n^*) + n^* P'(n^*)\right) \\ \frac{\partial}{\partial S} V\left(n^*, S(k^*)\right) s(k^*) &= r\left(W(k^*) + k^* W'(k^*)\right). \end{aligned}$$

A social planner would maximize

$$\max_{n,k} W(n,k) = V(n,k) - c\int_0^n P(s)ds - r\int_0^k W(s)ds.$$

Theorem 6. *Provided second-order conditions hold, the planner's problem has a unique solution (n^*, k^*)characterized by*

$$\begin{aligned} \frac{\partial}{\partial n} V\left(n^{opt}, k^{opt}\right) &= cP(n^{opt}) \\ \frac{\partial}{\partial S} V\left(n^{opt}, S\left(k^{opt}\right)\right) s(k^{opt}) &= rW(k^{opt}) \end{aligned}$$

The following corollary captures the policy conclusions.

Corollary 12. *The monopsonist always purchases less data than optimal as $n^*(\eta) < n^{opt}(\eta)$. Furthermore, the monopolist underinvests in k if $n^*(\eta) \leq \tilde{n}\left(k^*(\eta)\right)$, equivalently, there exists a level*

[6] To microfound the supply function $P(n)$, one could suppose that the data has to be purchased (directly through monetary transfers or indirectly through transfers in utility) from a population of N potential users who have different outside options. Concretely, one may think of users having to spend some time on an app developed by M in order to generate the data, and the opportunity cost of time to be some θ distributed according to some CDF $F(\cdot)$. Assuming quasilinear utilities a user of type θ will earn a utility of $U_\theta = p - \theta$,from using the app developed by M. Only the users of type θ such that $\theta \leq p$ will use the app so the demand for the app will therefore be $n = NF(p)$. Assuming each user generates one unit of data the supply of data will be equal to $P(n) \equiv F^{-1}\left(\frac{n}{N}\right)$.

of technological efficiency $\tilde{\eta}(\sigma^2)$ increasing in σ^2such that there is underinvestment in technology if and only if $\eta \leq \tilde{\eta}(\sigma^2)$.

This is the standard result in monopsony, where the buyer does not internalize the revenue that the data suppliers earns. Therefore it will buy less data than socially optimal. If in this optimum n and k are complements there will also be an underinvesment in technology k: the firm does not internalize the full positive impact that technology has on the value of data as part of it goes to data suppliers. Conversely, if k and n are substitutes the firm will overinvest in k: effectively it uses technology to drive down the marginal value of data to reduce the compensation it must pay to data suppliers.

This suggests that if it is cheap to collect more data on each user (η is large), monopolists will overinvest in technology to drive down the compensation they pay data suppliers.

5.3.1 Stackelberg

A firm collects too much data to deter entry.

5.3.2 Data Shring

Two firms, each with a number of users n_i and tehcnilogy A_i. Suppose they can share users and p is price paid by 1.

$$(V(n_1+n_2,k_1)-p-V(n_1,k_1))^{\gamma}\,(V(n_1+n_2,k_2)+p-V(n_2,k_2))^{1-\gamma}$$

The Nash price is

$$p^* = V\left(n_1+n_2,k_1\right)-V\left(n_1,k_1\right)-\gamma\left(V\left(n_1+n_2,k_1\right)+V\left(n_1+n_2,k_2\right)-V\left(n_1,k_1\right)-V\left(n_2,k_2\right)\right)\geq 0$$

$$\iff \gamma < \frac{V\left(n_1+n_2,k_1\right)-V\left(n_1,k_1\right)}{V\left(n_1+n_2,k_1\right)+V\left(n_1+n_2,k_2\right)-V\left(n_1,k_1\right)+V\left(n_2,k_2\right)}$$

So

$$\Pi_1 = V\left(n_1,k_1\right)+\gamma\left(V\left(n_1+n_2,k_1\right)+V\left(n_1+n_2,k_2\right)-V\left(n_1,k_1\right)-V\left(n_2,k_2\right)\right)$$

hence

$$\frac{\partial \Pi_1}{\partial k_2} = \gamma \left[V_A(n_1 + n_2, k_2) - V_A\left(n_2, k_2\right) \right]$$
$$\frac{\partial \Pi_1}{\partial n_2} = \gamma \left[V_n(n_1 + n_2, k_1) + V_n(n_1 + n_2, k_2) - V_n\left(n_2, k_2\right) \right]$$

5.3.3 Cournot

Let us now consider Z firms

$$\frac{\partial}{\partial n} V(n_i^*, k_i^*) = P\left(n_i^* + \sum_{j \neq i} n_j\right) + n_i^* P'\left(n_i^* + \sum_{j \neq i} n_j\right)$$
$$\frac{\partial}{\partial k} V(n_i^*, k_i^*) = \frac{1}{\eta}$$

It is well known we can rewrite it as

$$\frac{\partial}{\partial n} V(n^*, k^*) = P\left(Zn^*\right) + n^* P'\left(Zn^*\right)$$
$$\frac{\partial}{\partial k} V(n^*, k^*) = \frac{1}{\eta}$$

A planner would set

$$\max \sum_{i=1}^{Z} \left(V(n_i, k_i) - \frac{k_i}{\eta} \right) - \int_0^{\sum_{i=1}^{Z} n_i} P(s) ds$$

$$\frac{\partial}{\partial n} V\left(n^{\text{opt}}, k^{\text{opt}}\right) = P\left(Zn^{\text{opt}}\right)$$
$$\frac{\partial}{\partial k} V\left(n^{\text{opt}}, k^{\text{opt}}\right) = \frac{1}{\eta}$$

As$n^*$is decreasing in Z so if complements (underinvestment), the underinvestment becomes worse as k decreases in Z. If substitutes there is overinvestment also overinvestment becomes wors as k increases.

5.4 Perfect competition

Let us now assume $\eta \to \infty$ and that $k = \bar{k}$, as $V(n,k)$ is always increasing in k.

$$V(n,A) - nP(nZ) = 0$$

therefore

$$Z = \frac{F\left(\frac{V(n,k)}{n}\right)}{n}$$

Assume that there is a common level of technology A in firms competing for data in market where the prediction difficulty is σ^2. Under perfect competition denoting by

$$mv(n;A) \equiv \frac{\partial V(n;A)}{\partial n} = \frac{\sigma^2 - A}{\left(\frac{\sigma^2}{A} + n - 1\right)^2} \geq 0 \iff p \leq p_{\text{choke}}(A) \equiv \frac{A}{\frac{\sigma^2}{A} - 1}.$$

Note that the choke price is increasing in A and decreasing in σ^2. This implies that the market will be active only in markets where the prediction problem is not too hard and if and only if the technology A is high enough. In markets with high σ^2 and low A, the cost of acquiring enough data to be competitive is prohibitive, creating barriers to entry.

The data demand curve is

$$mv(D;A) = p \iff D(p,A) = \sqrt{\frac{\sigma^2 - A}{p}} - \left(\frac{\sigma^2}{A} - 1\right).$$

The elasticity of demand is ϵ_D

$$\varepsilon_D(A) \equiv -\frac{pD'(p)}{D(p)} = \frac{1}{2\left(1 - \sqrt{\frac{p}{A}\left(\frac{\sigma^2}{A} - 1\right)}\right)}.$$

It is increasing in σ^2 implying and decreasing in A implying that an equal percentage increase in the price of data reduces data acquisition by a larger percentage in markets where predictions are more difficult and where there is worse technology. **This implies several things: an optimal subsidy directed to incentivizing data collection (conversely a tax directed to reduce it) will be most effective in markets where the demand for data is elastic meaning the difficulty σ^2 is high (health, financial services) and the**

level of technology A is low. As the level of technology A becomes larger, the effect of taxes and subsidies on data collection will become smaller.

The competitive equilibrium is

$$D(p, A) = S(p),$$

where η is a privacy externality. The comparative statics of the equilibirum $n^*(A)$ will depend on $\frac{\partial mv(n;A)}{\partial A} = \frac{\partial^2 V(n;A)}{\partial n \partial A}$ which implies that n^* will be inverted U-shaped in A and U-shaped in σ^2.

Suppose there is a tax on data (for instance GDPR)

$$D(p) = S(p - t).$$

Let us study by how much the price paid by firms increases if t increases and call it the incidence $\rho \equiv \frac{dp}{dt}$. Implicit differentiation yields

$$D'(p)\rho = (\rho - 1)S'(p - t) \iff \rho(A) = \frac{1}{1 + \frac{\varepsilon_D(A,\sigma^2)}{\varepsilon_S}} \text{ is increasing in } A \text{ and decreasing in } \eta \text{ and } \sigma^2.$$

The fraction of an increase of a tax/subsidy on data passed on to firms (respectively paid for by data owners) will be larger (smaller) if technology A is large and the prediction σ^2 is small.The model predicts that stricter data protection laws (analogous to raising the cost of data collection per user, c) push platforms into a regime where the value of each additional user increases. This is because platforms need to offset the loss of granular data per user by increasing sample size. As a result, platforms will be willing to offer higher compensation or more favorable terms to attract new users.

Interpretation: Data minimization rules (as found in GDPR and CCPA) do not necessarily reduce the economic value of data. Instead, they shift the source of value from deep profiling (within-user data) to broader participation (across-user data). This shift could lead to more competitive user compensation markets, where platforms actively bid for access to user data.

Policy takeaway: Regulators could frame data minimization not only as a privacy-enhancing measure but also as a mechanism to foster competition for user data and re-

distribute some of the data's value back to consumers.

The model captures how technological sophistication (higher x) changes the relative importance of adding more users (higher n). This has significant regulatory implications:

- In early stages (low x), platforms are data-hungry and heavily reliant on user participation. Regulations limiting individual-level data collection shift the focus to participation incentives, fostering a competitive market for user data.
- As technology improves (higher x), platforms become better at extracting value from sparse data. This reduces the marginal value of additional users, making platforms less reliant on broad participation.

Over time, the importance of data volume declines relative to data quality and modeling sophistication. This shift may reduce the effectiveness of data portability regulations (which assume data volume drives competitive advantage) and highlight the importance of ensuring fair access to algorithmic innovations. The policy takeaway is that regulators should anticipate the shift from data-centric competition (focused on collecting more data) to model-centric competition (focused on better algorithms). This suggests that policies focused on algorithmic transparency and fair access to machine learning infrastructure may become more important than policies focused on raw data access.

5.5 Ad intermediation by a Monopsonist

Suppose that there is a unit mass of potential users of a platform P. Each user $i \in \mathcal{I} \equiv [0,1]$ consumes online content through P app and has an unknown preferred variety of content $y_k \in \mathbb{R}$ which gives her a stand alone utility $u \geq 0$ which is a monetary transfer by P. We assume each user is located at an unknown point $y_k \in \mathbb{R}$ on an extended Hotelling line which describes the possible content varieties, and must pay a quadratic cost for distance traveled to the variety of content they consume, scaled by a transportation cost $t_k > 0$. The *utility from content* of a user with preferred variety y_k consuming $\hat{y}_k$ when n consumer use the app offered by P is therefore

$$U_k = u - t_k(y_k - \hat{y}_k)^2 - c(\theta) + e(n),$$

where $c(\theta, n) \geq 0$ is the opportunity cost which is increasing in the type of the user

$\theta \in \Theta$ (the cost of privacy and/or the cost of time wasted) and decreasing in the number of users who use the platform n. There are two main reasons for which a user's choice of using an app might have a negative externality on non-users: on the one hand, there is a fear-of-missing-out (FOMO) effect that makes it more costly for individuals not to use aqn app the greater the number of their peers who use; this effect has been widely documented in social media cite XXX; on the other hand there are privacy externalities which imply that the privacy value of not using an app is decresaing in the number of user who use it.

Suppose that users receive ads for a horizontally differentiated good. Each user has an unknown preferred variety of good $y_a \in \mathbb{R}$, which gives her a stand alone utility $v \geq 0$ which is exogenous. We assume each user is located at an unknown point $y_a \in \mathbb{R}$ on an extended Hotelling line which describes the possible consumption good varieties, and must pay a quadratic cost for distance traveled to the variety of good they consume, scaled by a transportation cost $t_a > 0$. The *utility from consumption* of a user with preferred good y_a consuming $\hat{y}_a$ is therefore

$$U_a = v - t_a(y_a - \hat{y}_a)^2 - p,$$

where we assume a zero outside option. We assume that z_i and a_i are mutually independent and i.i.d. across individuals with mean 0 and variance $\sigma^2 \geq 0$. The variance parameter σ^2 is a measure of the difficulty of the prediction problem. For each individual $i \in \mathcal{I}$, P observes a covariate x_i. The relationship between x_i and the target variable$y \in \{y_k, y_a\}$ is given by:

$$y_i = \beta x_i + \epsilon_i,$$

as per Eq. XXX

5.5.1 Advertiser Problem

The *expected utility from consumption* of a user with preferred good advertised by y_a consuming $\hat{y}$ is therefore

$$U_a = v - t_a\left(\sigma^2 - V\left(n; S\right)\right) - p.$$

If P collects n samples, the advertising firm can therefore set

$$p(n) = v - t_a\left(\sigma^2 - V\left(n; S\right)\right),$$

and leave the user/consumer with nothing. If n users use P, the advertiser therefore makes

$$\pi_A = \begin{cases} (1-f)p(n) & \text{if use } P \\ p(0) & \text{if sell directly} \end{cases}$$

where f is an ad valorem fee charged by platform. The platform can therefore charge

$$f^* = 1 - \frac{p(0)}{p(n)} = \frac{V(n;S)}{v/t_a - \left(\sigma^2 - V(n;S)\right)}.$$

The advertiser will net a profit of $\pi_A = v - t_a\sigma^2$, which is its outside option on the users of P and the same non non-users.

Therefore the platform makes per user revenue

$$R(n) = t_a V(n;S)$$

5.5.2 User Problem

As users get no utility from the app. The *expected utility from content* of a user with preferred content variety y_k consuming $\hat{y}_a$ is therefore

$$U_k = u - t_k\left(\sigma^2 - V(n;S)\right) - c(\theta)\left(1 - e(n)\right).$$

Note that

$$u - t_k\left(\sigma^2 - V(\theta;k)\right) - c(\theta)\left(1 - e(\theta)\right) \text{ is concave in } \theta.$$

Therefore if $u \geq t_k\left(\sigma^2 - V(\theta)\right) + c(\theta,\theta)$ (which will always be the case because otherwise P would have no users and therefore make no profit).

Proposition 6. *There is a cutoff $\hat{\theta}$ such that $\theta \leq \hat{\theta}$ use the platform and the other users do not. Especially*

$$u = t_k\left(\sigma^2 - V(\theta;k)\right) + c(\theta)\left(1 - e(\theta)\right).$$

Trivially this cutoff is increasing in $p_z$and k and decreasing in t_z. Better technology induces more users to enter for the same price. Therefore there is a Supply function for data

$$S(n) = t_z \left(\sigma^2 - V\big(n; S\big)\right) + c(n)\left(1 - e(n)\right).$$

5.5.3 Platform Problem

The platform solves

$$\max_n \Pi(n) \equiv R(n) - nS(n)$$

Therefore

$$r(n) = S(n) + ns(n)$$

The social planner maximizes total welfare,

$$\begin{aligned} W(n) &= R(n) - \int_0^n \left(t_k(\sigma^2 - V(n;k)) + c(\theta)\right) d\theta - \int_n^1 e(n) d\theta \\ &= R(n) - \int_0^n \left(S(n) - c(n) + e(n) + c(\theta)\right) d\theta - (1-n)e(n) \\ &= R(n) - n\left(S(n) - c(n)\right) - \int_0^n c(\theta) d\theta - e(n) \end{aligned}$$

First-Order Condition (FOC). Differentiating with respect to n:

$$\frac{dW(n)}{dn} = r(n) - \left(S(n) - c(n)\right) - n\left(s(n) - \dot{c}(n)\right) - c(n) - \dot{e}(n) = 0. \tag{5.1}$$

Therefore

$$n\dot{c}(n) - \dot{e}(n) \leq 0$$

As n is inverted U-shaped in k

5.6 Hotelling Model

Model Setup. Consider two platforms, P_1 and P_2, located at the endpoints of a unit Hotelling line $[0, 1]$, on which a continuum of consumers of total unit mass is uniformly distributed. Each consumer is identified by their location $\beta \in [0, 1]$ and must choose to buy from P_1 or P_2. The consumer's location β determines a transportation cost $\tau \geq 0$, so that if

a consumer at β purchases from P_1 (at location 0), she incurs a travel cost $\tau\,\beta^2$. If instead she purchases from P_2 (at location 1), she incurs travel cost $\tau\,(\beta-1)^2$.

Additionally, each consumer has a "favorite variety" of the good $Y \sim \mathcal{N}(0, \sigma^2)$ which has a data-generating process described in Eq. (??)-(??) and is independent of β. However, the consumer does not know Y. Each platform P_i observes some data $\boldsymbol{D}_i$ of dimensions (n_i, k_i), which determines a *targeting value* $v_i \geq 0$ as per Eq. (4.2).

A representative consumer's gross utility from buying the good of her favorite variety is $u \geq 0$. If she buys from P_i at price p_i, her expected mismatch loss is $t\,\mathbb{E}\big[(Y-\hat{s}(\mathbf{D}_i))^2\big]$, where $t \geq 0$ indicates how "picky" (mismatch-sensitive) consumers are. Hence, the net utility of purchasing from P_i is:

$$U(p_1,p_2) \;=\; \begin{cases} u - p_1 \;-\; t\,\mathbb{E}\big[(Y-\hat{s}(\boldsymbol{D}_1))^2\big] \;-\; \tau\,\beta^2, & \text{if buying from } P_1, \\ u - p_2 \;-\; t\,\mathbb{E}\big[(Y-\hat{s}(\boldsymbol{D}_2))^2\big] \;-\; \tau\,(\beta-1)^2, & \text{if buying from } P_2. \end{cases}$$

Let $D_i(p_i)$ be the fraction of consumers who choose P_i. Then P_i's profit is

$$\Pi_i(p_i) \;=\; p_i\,D_i(p_i).$$

We solve for a Nash equilibrium in prices (p_1, p_2), taking each platform's data value (v_1, v_2) as given. Denote the *equilibrium* profit of P_i, when each platform's data value is (v_i, v_j), by

$$\hat{\Pi}_i(v_i,v_j) \;=\; \Pi_i\big(p_i^*(v_i,v_j)\big).$$

Proposition #5 (Nash Equilibrium in Prices and Demands). *In the interior solution where* $\big|v_i - v_j\big| < \frac{3\tau}{t}$, *the unique Nash equilibrium in prices is*

$$p_i^*(v_i,v_j) \;=\; \tau \;+\; \frac{t\big(v_i - v_j\big)}{3\,\tau},$$

and the associated market shares are

$$D_i(v_i,v_j) \;=\; \tfrac{1}{2} \;+\; \frac{t\,\big(v_i - v_j\big)}{6\,\tau}.$$

Hence, P_i's equilibrium profit is

$$\hat{\Pi}_i(v_i, v_j) \;=\; \frac{\left(\tau + \frac{t\,(v_i - v_j)}{3}\right)^2}{2\,\tau}.$$

Proof. A consumer at location β chooses P_1 if

$$u - p_1 - t\,\mathbb{E}\big[(Y - \hat{s}(\mathbf{D}_1))^2\big] - \tau\,\beta^2 \;\geq\; u - p_2 - t\,\mathbb{E}\big[(Y - \hat{s}(\mathbf{D}_2))^2\big] - \tau\,(\beta - 1)^2.$$

Simplifying yields

$$\beta \;\leq\; \frac{1}{2} + \frac{p_2 - p_1 + t\,(v_1 - v_2)}{2\,\tau}.$$

Hence,

$$D_1(p_1, p_2) \;=\; \frac{1}{2} + \frac{p_2 - p_1 + t\,(v_1 - v_2)}{2\,\tau}.$$

Each platform maximizes $\Pi_i(p_i; p_j) = p_i\,D_i(p_i, p_j)$. The first-order condition gives

$$p_i(p_j) \;=\; \frac{\tau}{2} + \frac{p_j + t\,(v_i - v_j)}{2}.$$

Solving simultaneously yields

$$p_i^* \;=\; \tau + \frac{t\,(v_i - v_j)}{3\,\tau}.$$

Substituting p_i^* back into D_i and Π_i yields the expressions in the proposition. The equilibrium is interior if $|v_i - v_j| < \frac{3\tau}{t}$, which guarantees strictly positive demand for both platforms. □

Proposition (5.6), together with Theorem (3), which establishes that v_i is increasing in both dimensions, implies the following result on the strategic behaviour of the platforms in when collecting data.

Corollary 13. *Data collection choices are strategic substitutes,*

$$\frac{\partial^2 \hat{\Pi}_i\big(v(n_i, k_i),\, v(n_j, k_j)\big)}{\partial x_i\,\partial y_j} = -\frac{t^2 v_x(n_i, k_i) v_y(n_j, k_j)}{9\tau} < 0,$$

for $x, y \in \{n, k\}$.

When one platform invests in more or better data, the other platform's incentive to invest

shrinks. If one platform takes the lead in data gathering (thus boosting its targeting quality), the rival sees less benefit from investing in its own data. This dynamic can create a disparity in data capabilities rather than both firms investing aggressively. Because the return to the second investor falls, one platform may end up with a significant advantage in data/targeting precision. Over time, this can reinforce that platform's competitive position—especially if the initial "data lead" snowballs into higher margins and even more data investment. Furthermore, just like in the Cournot equilibrium, total industry data investment may be lower than if both firms had incentives to match each other's data expansions. This finding suggests that regulators should be concerned that a platform starting with a data advantage can entrench its position by deterring rivals' investment, strengthening the case for scrutinizing whether initial data advantages can lock out effective competition over time.

Let us now imagine that a new dataset Δ of dimensions n and k is made available which we assume contains covariates unobserved by either platform. Let us assume that the individuals in the dataset are a subset of the users of both platforms, and are identified so that platforms can perform a merge operations to study the covariates of each individual conjointly. How much would each platform be willing to pay for it? This will be the subject of our next exercise.

6 Applications

In this section, we will apply the model to analyze several scenarios under which data affects competition between platforms. In the applications, k can also represent the sophistication or capacity of the AI/algorithm that processes the data.

6.1 Contextual Ads

Suppose now that there are D data brokers each holding distinct covariates. Each data broker $d \in 1, ..., D$has data on the same nZ individuals but comprising $\frac{kZ}{d}$ distinct covariates so that the total amount of covariates is kZ. As d has a monopoly on its covariates, it can achieve a profit of $v(n, \frac{k}{D})$ by basing targeting on its own covariates, namely doing contextual advertising. However there is an ad tech that can aggregate their data. The outside option of the data brokers is selling directly and earning $v(n, \frac{k}{D})$. Straightforward application of Theorem 4 leads to the following definition of the surplus deriving from data

aggregation:

Definition 2. The Aggregation Surplus of k covariates siloed in D datasets on n observations is

$$S(n,k,D) \;=\; v(n,k) \;-\; D\,v\Big(n,\tfrac{k}{D}\Big).$$

The following result analyzes how the value of aggregation depends on the dimensions of the data and the level of fragmentation D.

Proposition 7. *The Aggregation Surplus $S(n,k,D)$ is increasing in D, increasing in k, and exhibits an inverted-U shape in both n and k.*

Proof. First, for a fixed n and D, convexity of $v(\cdot,\cdot)$ in k implies that distributing k among many data brokers, each with $\frac{k}{D}$ covariates, is (weakly) less efficient than having all k covariates combined. Formally,

$$\frac{\partial}{\partial k}\Big[v(n,k) - D\,v\Big(n,\tfrac{k}{D}\Big)\Big] \;\geq\; 0,$$

so $S(n,k,D)$ is increasing in k. Similarly, for a fixed n and k, splitting the same total k across more data brokers raises total standalone usage only sub-linearly under the usual convexity of v in its second argument; hence

$$\frac{\partial}{\partial D}\Big[v(n,k) \;-\; D\,v\Big(n,\tfrac{k}{D}\Big)\Big] \;\geq\; 0,$$

so $S(n,k,D)$ is increasing in D. To see the inverted-U shape in n and k, note that for small (n,k), v is supermodular, so increments in n increase returns to combining data. However, once (n,k) become sufficiently large, v switches to submodularity, making further increments in n reduce additional benefits from data combination. This change from supermodular to submodular behavior in v explains why S follows an inverted-U pattern in both n. □

Assume now that P that bargains bilaterally with each of the data brokers following a Nash-in-Nash as per Collard-Wexler et al. (2019). We can state the following result:

Proposition 8. *For a generic data broker $d \in \{1,...,D\}$, the Nash-in-Nash price will solve*

$$\max_{p}\; \Big[v(n,k) - p \;-\; v\Big(n,\tfrac{(D-1)\,k}{D}\Big)\Big]^{\gamma}\;\Big[p - v\Big(n,\tfrac{k}{D}\Big)\Big]^{1-\gamma},$$

where $\gamma \in [0,1]$ is P's bargaining power. Then the equilibrium price is

$$p^*(n,k,D,\gamma) \;=\; \gamma\, v\Big(n, \tfrac{k}{D}\Big) \;+\; (1-\gamma)\Big[v(n,k) \;-\; v\Big(n, \tfrac{(D-1)\,k}{D}\Big)\Big],$$

which is increasing in k and n and decreasing in D and γ.

Proof. If P obtains all data brokers' data, it secures $v(n,k)$. If it misses developer d, it secures $v\big(n, \frac{(D-1)\,k}{D}\big)$. Data broker d's outside option is $v\big(n, \frac{k}{D}\big)$. P's disagreement point is $v\Big(n, \frac{(D-1)\,k}{D}\Big)$ therefore its net surplus if it pays p for the data of d is

$$U_P(p) \;=\; v(n,k) \;-\; p \;-\; v\Big(n, \tfrac{(D-1)\,k}{D}\Big).$$

The disagreement point of d is $v\Big(n, \frac{k}{D}\Big)$ her net payoff is

$$U_d(p) \;=\; p \;-\; v\Big(n, \tfrac{k}{D}\Big).$$

The Nash product is

$$\big[U_P(p)\big]^{\gamma}\,\big[U_d(p)\big]^{1-\gamma}.$$

Taking logs, differentiating with respect to p, and setting the derivative to zero yields

$$\gamma\,\frac{-1}{U_P(p)} \;+\; (1-\gamma)\,\frac{1}{U_d(p)} \;=\; 0.$$

Rearranging gives

$$\gamma\Big[p - v\Big(n, \tfrac{k}{D}\Big)\Big] \;=\; (1-\gamma)\Big[v(n,k) - p - v\Big(n, \tfrac{(D-1)\,k}{D}\Big)\Big].$$

Solving for p yields

$$p^*(n,k,D,\gamma) \;=\; \gamma\, v\Big(n, \tfrac{k}{D}\Big) \;+\; (1-\gamma)\Big[v(n,k) \;-\; v\Big(n, \tfrac{(D-1)\,k}{D}\Big)\Big].$$

□

Corollary 14. *The total data broker surplus is*

$$R^*(n,k,D,\gamma) \;=\; D \cdot p^*(n,k,D,\gamma),$$

which is U-shaped in D.

The profit of P is

$$\Pi^*(n,k,D,\gamma) \;=\; v(n,k) \;-\; R^*(n,k,D,\gamma),$$

which is inverted U-shaped in D, k and n.

$$\Pi^*(n,k,D,\gamma) \;\geq\; 0 \iff \gamma \;\geq\; \widehat{\gamma}(n,k,D) \;=\; \frac{\frac{D-1}{D}\,v(n,k) \;-\; v\!\left(n,\frac{(D-1)\,k}{D}\right)}{v(n,k) - \; v\!\left(n,\frac{k}{D}\right) \;-\; v\!\left(n,\frac{(D-1)\,k}{D}\right)} \geq \frac{1}{2}.$$

Proof. From $\Pi^*(n,k,D,\gamma) = v(n,k) - D\,p^*(n,k,D,\gamma)$, substituting the expression for p^* from Proposition 8 and rearranging the inequality $\Pi^* \geq 0$ yields

$$\gamma\left[v\!\left(n,\tfrac{k}{D}\right) + v\!\left(n,\tfrac{(D-1)\,k}{D}\right) - v(n,k)\right] \;\geq\; \frac{D-1}{D}\,v(n,k) \;-\; v\!\left(n,\tfrac{(D-1)\,k}{D}\right).$$

Solving for γ gives $\widehat{\gamma}(n,k,D)$. □

The function $\widehat{\gamma}(n,k,D)$ is increasing in k and D, because as more covariates are collected or data becomes more fragmented, the incremental synergy from combining an additional piece of data increases (because of convexity), meaning P's outside option decreases and d's outside option increases, pushing up the minimum bargaining power P must have in order to stay profitable. An analogous reasoning explains why $\widehat{\gamma}(n,k,D)$ is inverted-U shape in n, for the same reason as in Proposition 7, namely that Π^* is tied to the incremental surplus $S(n,k,D)$ and inherits the supermodular-for-small versus submodular-for-large behavior of $v(n,k)$ in its two arguments.

The U-shaped value of aggregation suggests that there is indeed a peak scale at which the platform can extract he highest marginal value from its aggregation services, a finding which is analogous to the S-shaped returns suggested in Posner and Weyl (2018) and Tirole (2020). These findings have several implications for competition policy and antitrust. A more fragmented upstream data brokerage sector can fundamentally impair innovation by downstream ad tech platforms.

Proposition 9. *As $D \to \infty$, the total broker profit tends to*

$$\lim_{D\to\infty} R^*(n,k,D,\gamma) = \gamma\frac{nk}{n+1} + (1-\gamma)\frac{nk}{\sqrt{(n+1)^2-4nk}},$$

and

$$\lim_{D\to\infty} \widehat{\gamma}(n,k,D) = \frac{1}{\sqrt{1-\frac{4kn}{(n+1)^2}}+1}.$$

7 Conclusion

By emphasizing that data alone need not be destiny—and that feature numbermediates the benefits of data at scale—this paper offers a more nuanced, evidence-based lens for assessing data's role in shaping competition and innovation. Our theoretical framework clarifies when accumulating more data confers unique advantages, when further improvements in feature numberserve as a partial substitute, and how market dynamics shift as firms evolve from data scarcity to abundance.

These insights help refine ongoing debates around digital regulation. Policies premised on "data equals power" may overstate the value of additional observations once a platform has already crossed into a high-n region. Meanwhile, measures that foster algorithmic innovation—such as supporting AI research or encouraging portability for smaller data sets—could be pivotal in bolstering competition. The analysis also suggests that data monopsony can distort data prices or quantities, prompting potential remedies such as collective bargaining rights for users or forced data-sharing arrangements.

Ultimately, the interplay of data scale (n) and feature number(k) dictates whether incumbents preserve an unassailable lead or face renewed competition from more agile, tech-savvy entrants. Understanding this interplay is crucial for policymakers seeking to balance innovation incentives with protections against data-driven market power.

References

Kenneth J. Arrow. The economic implications of learning by doing. *The Review of Economic Studies*, 29(3):155–173, 1962. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2295952.

Z. Bai and J.W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer New York, 2009. ISBN 9781441906618. URL https://books.google.fr/books?id=kd-o5Qdm7ngC.

Patrick Bajari, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki. The impact of big data on firm performance: An empirical investigation. *AEA papers and proceedings*, 109: 33–37, 2019.

Bruno Carballa Smichowski, Néstor Duch, Seyit Höcük, Pradeep Kumar, Bertin Martens, Joris Mulder, and Patricia Prufer. Economies of scope in data aggregation: evidence from health data. 11 2022.

Allan Collard-Wexler, Gautam Gowrisankaran, and Robin Lee. "nash-in-nash" bargaining: A microfoundation for applied work. *Journal of Political Economy*, 127(1): 163 – 195, 2019. URL https://EconPapers.repec.org/RePEc:ucp:jpolec:doi:10.1086/700729.

Morris H. DeGroot. Optimal statistical decisions. 2005.

Vladimir A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

Eric A. Posner and Eric Glen Weyl. *Radical Markets*. Princeton University Press, Princeton, 2018. ISBN 9780691196978. doi: doi:10.1515/9780691196978. URL https://doi.org/10.1515/9780691196978.

Maximilian Schaefer and Geza Sapi. Complementarities in learning from data: Insights from general search. *Information Economics and Policy*, 65:101063, 2023. ISSN 0167-6245. doi: https://doi.org/10.1016/j.infoecopol.2023.101063. URL https://www.sciencedirect.com/science/article/pii/S0167624523000483.

Maximilian Schäfer, Geza Sapi, and Szabolcs Lorincz. The effect of big data on recommendation quality. the example of internet search. *SSRN Electron. J.*, 2018.

Ryan Tibshirani. High-dimensional regression: Ridge. lecture notes of advanced topics in statistical learning, spring 2023. 2023.

Jean Tirole. Competition and the industrial challenge for the digital age. *IFS Deaton Review on Inequalities*, 2020.

Appendix

The optimal predictor trivially corresponds to the posterior mean:

$$\hat{y}_n^* = \mathbb{E}_Y \left[Y \middle| x, (\boldsymbol{x}, \boldsymbol{y})_n \right].$$

The value of information is therefore the reduction of posterior variance of $\boldsymbol{y}_m$ which is characterized in the following result.

Lemma 3. *The Value of Information of $(\boldsymbol{x}, \boldsymbol{y})_n$ for a decision maker observing covariates x is given by:*

$$VoI_{(\boldsymbol{x},\boldsymbol{y})_n}(x) = \frac{x^2}{1 + \frac{\sigma_\epsilon^2(S;\sigma^2)}{\sum_{i=1}^n x_i^2}}.$$

[7]

Proof. The optimal prior predictor (before observing data) is:

$$\hat{y} = 0$$

with mean squared error (MSE):

$$\mathbb{E}[(y_0 - \hat{y})^2 \mid x] = x + \sigma_\epsilon^2$$

After observing n data points, the platform forms an estimate $\hat{\beta}_n$. The posterior variance of β is given by precision weighting:

$$\frac{\sigma^2}{\mathbb{V}(\beta \mid y_1, \ldots, y_n, x_1, \ldots, x_n)} = 1 + \frac{\sum_{i=1}^n x_i^2}{\sigma_\epsilon^2}$$

The posterior prediction for y given x is:

$$\hat{y}_n = \mathbb{E}[y \mid x, \text{data}] = x \cdot \mathbb{E}[\beta \mid \text{data}]$$

[7] Observe that the vector of target variables y does not affect the VoI. The reason is that y gives information about where β is (the location/mean), but not about the precision of that estimate once the x's are fixed. Once you condition on x, y doesn't change how "spread out" your posterior beliefs about β — it just shifts the mean.

The corresponding posterior MSE is:

$$\mathbb{E}[(y_0 - \hat{y}_n)^2 \mid x] = x \cdot \mathbb{V}(\beta \mid \text{data}) + \sigma_\epsilon^2$$

The value of information is then:

$$\text{VoI}_n(x) = \sigma^2 x + \sigma_\epsilon^2 - \left(x \cdot \mathbb{V}(\beta \mid \text{data}) + \sigma_\epsilon^2\right)$$

which simplifies to:

$$\text{VoI}_n(x) = x \cdot \left(\sigma^2 - \mathbb{V}(\beta \mid \text{data})\right)$$

Substituting the expression for $\mathbb{V}(\beta \mid \text{data})$,

$$\text{VoI}_n(x) = \sigma^2 x \cdot \left(1 - \left(1 + \frac{\sum_{i=1}^n x_i^2}{\sigma_\epsilon^2}\right)^{-1}\right)$$

This completes the proof. □

Theorem 7. *The expected Value of Information for large n for technological level x is:*

$$\mathbb{E}_{\{x_i\}_{i=0}^n}[VoI_n(x)] \sim v(k,n) = \frac{k\sigma^2}{1 + \frac{\frac{1}{k}-1}{n}}$$

Proof. It is sufficient to observe $\mathbb{E}\left[\sum_{i=1}^n x_i^2\right] = n \cdot S$, using the LLN we can put this into the formula

$$\lim_{n\to\infty} \text{VoI}_{n,k}(x) = \sigma^2 k \cdot \left(1 - \frac{1}{1 + n \cdot \frac{k}{1-k}}\right)$$

then we take the expectation $\mathbb{E}x = S$ and we are done. □

Proposition 10. *Let $\boldsymbol{D} \equiv (\boldsymbol{x}, (\boldsymbol{y}, \boldsymbol{X}))$ be the data described above. Then the posterior distribution of $Y \mid \boldsymbol{D}$ is*

$$Y \,\big|\, \boldsymbol{D} \;\sim\; \mathcal{N}\Big(s^*(\boldsymbol{D}),\, \mathbb{V}\big[s^*(\boldsymbol{D})\big] + \sigma^2(1-k)\Big),$$

where

$$s^*(\boldsymbol{D}) \;=\; \boldsymbol{x}'\,\mathbb{E}\big[\boldsymbol{\beta}_x \,\big|\, (\boldsymbol{y}, \boldsymbol{X})\big], \quad \mathbb{V}\big[s^*(\boldsymbol{D})\big] \;=\; \boldsymbol{x}'\,\mathbb{V}\big[\boldsymbol{\beta}_x \,\big|\, (\boldsymbol{y}, \boldsymbol{X})\big]\boldsymbol{x}.$$

Proof. Recall that $z_i = \begin{pmatrix} x_i \\ u_i \end{pmatrix}$ and $\beta = \begin{pmatrix} \beta_x \\ \beta_u \end{pmatrix}$ so that for the individual of interest,

$$Y \;=\; z'\beta \;=\; x'\beta_x \;+\; u'\beta_u.$$

Since we observe only the "observed" covariates x (but not the "unobserved" part u), we decompose the posterior distribution of Y using:

$$Y \;=\; \underbrace{x'\beta_x}_{\text{term (A)}} \;+\; \underbrace{u'\beta_u}_{\text{term (B)}}.$$

Term (A). The data set (y, X) (with X partitioned analogously into its x_i parts) identifies the posterior distribution of β_x. Because prior and likelihood are both Gaussian (and x is independent of u), standard Bayesian linear regression results imply:

$$\beta_x \,\big|\, (y, X) \;\sim\; \mathcal{N}\Big(\mathbb{E}\big[\beta_x \,\big|\, (y, X)\big],\, \mathbb{V}\big[\beta_x \,\big|\, (y, X)\big]\Big).$$

Hence, the random variable $x'\beta_x \mid (y, X, x)$ is normally distributed with mean $x'\,\mathbb{E}[\beta_x \mid (y, X)]$ and variance $x'\,\mathbb{V}[\beta_x \mid (y, X)]x$.

Term (B). Because the unobserved u and its associated coefficients β_u are independent of x (and not identified by the data), the posterior for β_u remains the same as its prior $\mathcal{N}(0, \sigma^2\mathbf{I})$, and u itself is also independent (with mean 0). Thus, the contribution $u'\beta_u$ still has mean 0 and variance $\sigma^2(1-k)$ (the factor $1-k$ reflects that u is a $(1-k)Z$-dimensional subvector of z, whose total variance of $z'\beta$ was σ^2).

Since **term (A)** and **term (B)** are independent Gaussian random variables, their sum is Gaussian with mean equal to the sum of means and variance equal to the sum of variances. Hence,

$$Y \,\big|\, D \;=\; \big(x'\beta_x + u'\beta_u\big) \,\big|\, \big(x, (y, X)\big) \;\sim\; \mathcal{N}\Big(x'\,\mathbb{E}\big[\beta_x \,\big|\, (y, X)\big],\; x'\,\mathbb{V}\big[\beta_x \,\big|\, (y, X)\big]x \;+\; \sigma^2(1-k)\Big).$$

Equivalently, letting

$$s^*(D) \;\equiv\; x'\,\mathbb{E}\big[\beta_x \,\big|\, (y, X)\big], \quad \mathbb{V}\big[s^*(D)\big] \;\equiv\; x'\,\mathbb{V}\big[\beta_x \,\big|\, (y, X)\big]x,$$

we obtain the stated normal distribution:

$$Y \,|\, \boldsymbol{D} \;\sim\; \mathcal{N}\Big(s^*(\boldsymbol{D}),\, \mathbb{V}\big[s^*(\boldsymbol{D})\big] + \sigma^2(1-k)\Big).$$

□

Proposition 11. *The posterior distribution of* $\boldsymbol{\beta}_x|\,(\boldsymbol{y}, \boldsymbol{X})$ *is*

$$\boldsymbol{\beta}_x|\,(\boldsymbol{y}, \boldsymbol{X}) \sim \mathcal{N}\left(\left(1-k+\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y},\, \sigma^2\left(1+\frac{1}{1-k}\cdot\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

We base our proof on the classic treatment by DeGroot (2005) , generalized to allow an endogenous noise term. We can rewrite the DGP as $y = \boldsymbol{x}'\boldsymbol{\beta}_x + \varepsilon$ where $\varepsilon \equiv \boldsymbol{u}'\boldsymbol{\beta}_u \sim \mathcal{N}(0, (1-V(x)))$. Hence, the likelihood is

$$y|\boldsymbol{\beta}_x \sim \mathcal{N}\left(\boldsymbol{x}'\boldsymbol{\beta}_x, (1-V(x))\right).$$

Hence, using the Bayes rule, we express the posterior as a function of the prior $p(\boldsymbol{\beta}_x)$, the likelihood $\mathscr{L}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}_x)$ and the evidence $p(\boldsymbol{y}|\boldsymbol{X})$:

$$p(\boldsymbol{\beta}_x|\boldsymbol{X}, \boldsymbol{y}) = \frac{\mathscr{L}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}_x)p(\boldsymbol{\beta}_x)}{p(\boldsymbol{y}|\boldsymbol{X})}, \tag{7.1}$$

where the prior and the likelihood are known to be Gaussians:

$$p(\boldsymbol{\beta}_x) = \sqrt{\frac{1}{2\pi \mathrm{tr}(\boldsymbol{V}_x)}}e^{-\frac{\boldsymbol{\beta}_x'\boldsymbol{V}_x\boldsymbol{\beta}_x}{2}}, \tag{7.2}$$

$$\mathscr{L}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}_x) = \left(\frac{1}{2\pi\,(1-V(x))}\right)^{\frac{N}{2}} e^{-\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x)'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x)}{2(1-V(x))}}. \tag{7.3}$$

As we are interested in computing the posterior of $\boldsymbol{\beta}_x$, we want to isolate the terms in the product of Eq. (7.2) and (7.3) that depend on it. Define the Maximum Likelihood Estimator implicitly as $\tilde{\boldsymbol{\beta}}_x$, to avoid invertibility issues, as $\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}}_x = \boldsymbol{X}'\boldsymbol{y}$. We can rewrite the numerator of the exponent of the exponential in (7.3) as

$$\begin{aligned}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x)'\,(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}_x) &= \boldsymbol{y}'\boldsymbol{y} + (\boldsymbol{\beta}_x-\tilde{\boldsymbol{\beta}}_x)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}_x-\tilde{\boldsymbol{\beta}}_x) - \tilde{\boldsymbol{\beta}}_x'\boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}}_x \\ &\propto (\boldsymbol{\beta}_x-\tilde{\boldsymbol{\beta}}_x)'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta}_x-\tilde{\boldsymbol{\beta}}_x),\end{aligned}$$

given that we only care about the terms that depend on $\boldsymbol{\beta}_x$. We can therefore rewrite the part of the product of Eq. (7.2) and (7.3) which depends on $\boldsymbol{\beta}_x$ as

$$\boldsymbol{\beta}_x' \boldsymbol{V}_x \boldsymbol{\beta}_x + \frac{(\boldsymbol{\beta}_x - \tilde{\boldsymbol{\beta}}_x)' \boldsymbol{X}' \boldsymbol{X} (\boldsymbol{\beta}_x - \tilde{\boldsymbol{\beta}}_x)}{(1 - V(x))}.$$

define a ridge estimator as

$$\hat{\boldsymbol{\beta}}_x = \left((1 - V(x)) \cdot \boldsymbol{V}_x^{-1} + \boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{X}\tilde{\boldsymbol{\beta}}_x.$$

We now want to isolate the term depending on $\boldsymbol{\beta}_x$ to find the posterior

$$(\boldsymbol{\beta}_x - \hat{\boldsymbol{\beta}}_x)' \left(\boldsymbol{V}_x^{-1} + \frac{1}{(1 - V(x))} \cdot \boldsymbol{X}'\boldsymbol{X}\right) (\boldsymbol{\beta}_x - \hat{\boldsymbol{\beta}}_x) + \frac{\tilde{\boldsymbol{\beta}}_x' \boldsymbol{X}' \boldsymbol{X} \tilde{\boldsymbol{\beta}}_x}{(1 - V(x))} - \hat{\boldsymbol{\beta}}_x' \left(\boldsymbol{V}_x^{-1} + \frac{1}{(1 - V(x))} \cdot \boldsymbol{X}'\boldsymbol{X}\right) \hat{\boldsymbol{\beta}}_x.$$

Therefore,

$$\pi(\boldsymbol{\beta}_x | \boldsymbol{y}) \propto \exp\left\{-\frac{\sigma^2}{2} (\boldsymbol{\beta}_x - \hat{\boldsymbol{\beta}}_x)' \left(\boldsymbol{V}_x^{-1} + \frac{1}{(1 - V(x))} \cdot \boldsymbol{X}'\boldsymbol{X}\right) (\boldsymbol{\beta}_x - \hat{\boldsymbol{\beta}}_x)\right\}.$$

We can deduce

$$\boldsymbol{\beta}_x | \boldsymbol{y}, \boldsymbol{X} \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_x, \sigma^2 \cdot \left(\boldsymbol{V}_x^{-1} + \frac{1}{(1 - V(x))} \cdot \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

Corollary 15. *The estimator $\boldsymbol{t}_x(\boldsymbol{y}, \boldsymbol{X})$ is the unique solution to the optimization problem*

$$\min_{\boldsymbol{t}_x} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}_x\|_2^2 + (1 - V_Z(x)) \sum_{j=1}^{kZ} \frac{t_j^2}{v\left(\frac{j}{Z}\right)} \right\}. \tag{7.4}$$

Proof. We prove this by explicitly solving the optimization problem and showing that the resulting estimator matches the Bayesian posterior mean derived in Proposition 2.

The given optimization problem is a ridge regression problem with a weighted penalty. The objective function to minimize is:

$$L(\boldsymbol{t}_x) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}_x\|_2^2 + (1 - V_Z(x)) \sum_{j=1}^{kZ} \frac{t_j^2}{v\left(\frac{j}{Z}\right)}. \tag{7.5}$$

To find the optimal solution, we differentiate $L(\boldsymbol{t}_x)$ with respect to $\boldsymbol{t}_x$ and set the deriva-

tive to zero:

$$\frac{\partial L}{\partial \boldsymbol{t}_x} = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{t}_x) + 21 - V_Z(x)\boldsymbol{V}_x^{-1}\boldsymbol{t}_x = 0. \tag{7.6}$$

Rearranging, we obtain the normal equation:

$$(\boldsymbol{X}'\boldsymbol{X} + (1 - V_Z(x)) \cdot \boldsymbol{V}_x^{-1})\boldsymbol{t}_x = \boldsymbol{X}'\boldsymbol{y}. \tag{7.7}$$

Solving for $\boldsymbol{t}_x$ yields:

$$\boldsymbol{t}_x^*(\boldsymbol{y}, \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X} + (1 - V_Z(x)) \cdot \boldsymbol{V}_x^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{7.8}$$

which matches the Bayesian posterior mean of $\boldsymbol{\beta}_x$ From Proposition 2. This completes the proof. □

Proposition 12. *The following convergence holds as $Z \to \infty$*

$$\mathbb{V}\left[s^*(\boldsymbol{D})\right] \to \frac{\sigma_k^2}{2}\left(k\left(2 - \frac{n}{k} - \sqrt{\left(\frac{n+1}{k}\right)^2 - \frac{4n}{k}}\right) - 1\right).$$

Concentration for Quadratic Forms

We start with the following proposition from Boucheron *et al.* ?, Example 2.12:

Proposition. *Let $x_1, \ldots, x_{kZ}$ be kZ independent, zero-mean normal random variables with $\mathbb{E}[\,x_j^2\,] = \frac{1}{Z}$ for $j = 1, \ldots, kZ$. For any matrix $\boldsymbol{A} \in \mathbb{R}^{kZ \times kZ}$ and any $\xi > 0$,*

$$\mathbb{P}\Big(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}] \;>\; \tfrac{2}{Z}\Big(\|\boldsymbol{A}\|_{HS}\sqrt{\xi} + \|\boldsymbol{A}\|_2^2\,\xi\Big)\Big) \;\le\; e^{-\xi}.$$

This result tells us that, conditionally on $\boldsymbol{A}$, the quadratic form $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ concentrates sharply about its mean $\mathbb{E}\big[\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\big]$. In our setting, $\boldsymbol{x}$ is a Gaussian vector of dimension kZ with variance $\mathbb{E}[\,x_j^2\,] = 1/Z$, and

$$\boldsymbol{A} \;\equiv\; \mathbb{V}\big[\boldsymbol{\beta}_x \,\big|\, (\boldsymbol{y}, \boldsymbol{X})\big].$$

By conditioning on $\boldsymbol{X}$ and applying the above proposition, one obtains that

$$\boldsymbol{x}'\,\mathbb{V}\big[\boldsymbol{\beta}_x \mid (\boldsymbol{y}, \boldsymbol{X})\big]\boldsymbol{x} \;\xrightarrow{p}\; \mathbb{E}\Big[\boldsymbol{x}'\,\mathbb{V}\big[\boldsymbol{\beta}_x \mid (\boldsymbol{y}, \boldsymbol{X})\big]\boldsymbol{x}\Big] \quad \text{as } Z \to \infty.$$

That is, we may replace the random quadratic form with its conditional (hence also uncon-

ditional) expectation as Z grows large.

Rewriting the Variance as a Trace and Inverting

It is a well-known property of quadratic forms that for any vector $\boldsymbol{w}$ (with mean $\boldsymbol{\mu}$ and covariance Ω) and any matrix Λ,

$$\mathbb{E}\left[\boldsymbol{w}^T \Lambda \boldsymbol{w}\right] \;=\; \operatorname{tr}\left[\Lambda\, \Omega\right] \;+\; \boldsymbol{\mu}^T\, \Lambda\, \boldsymbol{\mu}.$$

Since $\boldsymbol{x}$ is zero-mean and independent of $\boldsymbol{X}$, the second term vanishes, and we simply get

$$\mathbb{E}\Big[\boldsymbol{x}'\, \mathbb{V}\left[\boldsymbol{\beta}_x \mid (\boldsymbol{y}, \boldsymbol{X})\right] \boldsymbol{x}\Big] \;=\; \operatorname{tr}\Big[\mathbb{V}\left[\boldsymbol{\beta}_x \mid (\boldsymbol{y}, \boldsymbol{X})\right] \operatorname{Cov}(\boldsymbol{x})\Big].$$

But by construction, $\operatorname{Cov}(\boldsymbol{x}) = \frac{1}{Z}\mathbf{I}_{kZ}$. Hence

$$\boldsymbol{x}'\, \mathbb{V}\left[\boldsymbol{\beta}_x \mid (\boldsymbol{y}, \boldsymbol{X})\right] \boldsymbol{x} \;=\; \frac{1}{Z} \operatorname{tr}\Big[\mathbb{V}\left[\boldsymbol{\beta}_x \mid (\boldsymbol{y}, \boldsymbol{X})\right]\Big].$$

Therefore,

$$\frac{1}{Z} \operatorname{tr}\Big[\mathbb{V}\left[\boldsymbol{\beta}_x \mid (\boldsymbol{y}, \boldsymbol{X})\right]\Big]. \tag{7.9}$$

By Proposition 1, we know that

$$\mathbb{V}\left[\boldsymbol{\beta}_x \mid (\boldsymbol{y}, \boldsymbol{X})\right] \;= \sigma^2 \cdot \left(\boldsymbol{V}_x^{-1} \;+\; \tfrac{1}{1 - V_Z(x)}\, \boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Hence the quantity in (7.9) is

$$\frac{\sigma^2}{Z} \operatorname{tr}\left[\left(\boldsymbol{V}_x^{-1} \;+\; \tfrac{1}{1 - V_Z(x)}\, \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right].$$

We factor out $\frac{1 - V_Z(x)}{n}$ from the inverse, yielding

$$\frac{1 - V_Z(x)\sigma^2}{nZ} \operatorname{tr}\left[\left(\tfrac{1 - V_Z(x)}{n}\, \boldsymbol{V}_x^{-1} \;+\; \tfrac{1}{n}\, \boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right].$$

Note that the emprical varianec copvariacne matrix $\hat{\Sigma} \equiv \frac{1}{nZ}\, \boldsymbol{X}'\boldsymbol{X}$, so $\frac{1}{n}\, \boldsymbol{X}'\boldsymbol{X} = Z\hat{\Sigma}$ is the standardized empirical variance covariance matrix whose expectation is the identity matrix $\boldsymbol{I}_{kZ}$.

7.1 Spectral Decomposition

Denote by $\lambda_1, \ldots, \lambda_{kZ}$ the eigenvalues of $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$. Since the prior $\boldsymbol{\beta}_x$ has *diagonal covariance* $\boldsymbol{V}_x$ with diagonal entries f, one finds that

$$\boldsymbol{V}_x^{-1} \quad \text{is diagonal with } \left[f\right]^{-1} \text{ on the diagonal.}$$

Hence, in block form,

$$\left(\frac{1-V_Z(x)}{n} \cdot \boldsymbol{V}_x^{-1} + \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

leads to terms of the form

$$\frac{1}{\frac{1-V_Z(x)}{nv\left(\frac{j}{Z}\right)} + \lambda_j}.$$

Thus, we can write

$$\frac{\bar{V}_Z(x)}{nZ} \sum_{j=1}^{kZ} \frac{1}{\frac{1-V_Z(x)}{nv\left(\frac{j}{Z}\right)} + \lambda_j}.$$

Decoupling the index j from the eigenvalues λ_j

The subtlety is that we have

$$\sum_{j=1}^{kZ} \frac{1}{\frac{1-V_Z(x)}{nv\left(\frac{j}{Z}\right)} + \lambda_j},$$

which might appear to couple the index j (which specifies v) with the jth eigenvalue λ_j. However, in *i.i.d. Gaussian* designs, the kZ eigenvalues are *exchangeable*, so there is no fundamental pairing of coordinate j with eigenvalue λ_j. A permutation argument, together with continuity bounds on $v(\cdot)$, shows that any re–labeling $\lambda_{\pi(1)}, \ldots, \lambda_{\pi(kZ)}$ yields the *same* asymptotic distribution. Therefore we can treat $\left\{v\left(\frac{j}{Z}\right)\right\}$ and $\{\lambda_j\}$ "independently" in the large limit.

More precisely, one shows by bounding that, for suitable permutation π,

$$\sum_{j=1}^{kZ} \frac{1}{\frac{1-V_Z(x)}{nv\left(\frac{j}{Z}\right)} + \lambda_{(j)}} - \sum_{j=1}^{kZ} \frac{1}{\frac{1-V_Z(x)}{nv\left(\frac{j}{Z}\right)} + \lambda_{\pi(j)}} \xrightarrow[Z\to\infty]{p} 0,$$

where $\lambda_{(j)}$ denotes the jth ordered eigenvalue. Hence we can "separate" j from λ_j in the large–Z limit.

Double integral limit

As a consequence, the normalized sum

$$\frac{1}{kZ}\sum_{j=1}^{kZ}\frac{1}{\frac{1-V_Z(x)}{nv\left(\frac{j}{Z}\right)}+\lambda_j}$$

becomes a *two–dimensional* Riemann sum in the variables $(t = \frac{j}{Z}, \lambda_j)$, where $t \in [0, x]$ (via uniform partition as j runs $1, \ldots, kZ$) and λ_j are distributed according to $G_{\frac{1}{n}X'X}$, the standardized empirical spectral distribution (SESD) of $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$. By the MP convergence and bounded convergence arguments, we obtain

$$\frac{1}{kZ}\sum_{j=1}^{kZ}\frac{1}{\frac{1-V_Z(x)}{nv\left(\frac{j}{Z}\right)}+\lambda_j}\quad\xrightarrow[Z\to\infty]{}\quad\int_0^x\left[\int_0^\infty\frac{dG_{\frac{1}{n}X'X}(\lambda)}{\frac{1-V(x)}{nv(t)}+\lambda}\right]\frac{dt}{x}=\int_0^\infty\int_0^x\frac{1}{\frac{1-V(x)}{nv(t)}+\lambda}\frac{dt}{x}dG_{\frac{1}{n}X'X}(\lambda).$$

Multiplying by the appropriate factor $\frac{1-V_Z(x)}{nZ}\cdot kZ$ yields exactly the final limit for the trace. Using$1 - V_Z(x) \to (1 - V(x))$ that establishes

$$\frac{\sigma^2}{Z}\mathrm{tr}\left[\left(\boldsymbol{V}_x^{-1}+\tfrac{1}{1-V_Z(x)}\,\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]\quad\xrightarrow[Z\to\infty]{}\quad\frac{(1-V(x))\,\sigma^2}{n}\int_0^x\left[\int_0^\infty\frac{dG_{\frac{1}{n}X'X}(\lambda)}{\frac{(1-V(x))}{nv(t)}+\lambda}\right]dt$$

which can be simplified to

$$\frac{(1-V(x))}{n}\int_0^x m_{G_{\frac{1}{n}X'X}}\left(-\tfrac{(1-V(x))}{nv(t)}\right)dt$$

where$m_{G_{\frac{1}{n}X'X}}(z)$ is the Stieltjes transform of the SESD. Now we can use the following classic result which is a reformulation of Marčenko and Pastur (1967) and Bai and Silverstein (2009) found in Tibshirani (2023). The SESD $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$ is the empirical covariance matrix of a i.i.d. standard normal matrix $\boldsymbol{W} = \sqrt{Z}\cdot\boldsymbol{X}$ as $\frac{1}{nZ}\boldsymbol{W}'\boldsymbol{W} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$. Therefore $m_{G_{\frac{1}{n}X'X}}$ is the same as $m_{G_{\frac{1}{nZ}W'W}}$, which is governed by a Marchenko Pastur law.

Theorem. *Let* $\{\mathbf{W}_n\}_{n\geq 1}$ *be a sequence of random matrices, where*

$$\mathbf{W}_n \in \mathbb{R}^{n\times x}$$

has i.i.d. entries $W_{ij}^{(n)} \sim \mathcal{N}(0,1)$ *(mean 0, variance 1). Suppose*

$$\lim_{n\to\infty} \frac{x}{n} \;=\; \gamma \;\in\; (0,\infty).$$

Consider the sample covariance matrix

$$\frac{1}{n}\,\mathbf{W}_n'\,\mathbf{W}_n \quad \in\; \mathbb{R}^{k\times k}.$$

For each n*, let* G_n *be its* empirical spectral distribution (ESD)*, i.e. the probability distribution that places mass* $\frac{1}{x}$ *at each of the* x *eigenvalues of* $\frac{1}{n}\,\mathbf{W}_n'\,\mathbf{W}_n$*.*

Then, as $n \to \infty$*, the ESD* G_n *converges weakly (almost surely) to the* ***Marchenko–Pastur distribution*** $F_{\mathrm{MP},\gamma}$ *with parameter* γ*. The limit distribution* $F_{\mathrm{MP},\gamma}$ *is supported on*

$$\left[\,\lambda_-,\,\lambda_+\right] \;=\; \left[\,(1-\sqrt{\gamma})^2,\,(1+\sqrt{\gamma})^2\right]$$

and has density

$$f_{\mathrm{MP},\gamma}(\lambda) \;=\; \frac{1}{2\pi\,\gamma\,\lambda}\,\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}\;\;\mathbf{1}_{[\lambda_-,\,\lambda_+]}(\lambda).$$

Equivalently, its Stieltjes transform $m_{\mathrm{MP},\gamma}(z)$ *satisfies the well-known functional equation*

$$-\frac{1}{m_{\mathrm{MP},\gamma}(z)} \;=\; z \;-\; \frac{\gamma}{1+m_{\mathrm{MP},\gamma}(z)}.$$

The theorem essentially says that the spectral bulk of $\frac{1}{n}\mathbf{W}_n'\mathbf{W}_n$ concentrates near the MP curve, whose support length grows with γ. Therefore, the value of data is

$$\sigma^2\left[v(x)-\frac{(1-V(x))}{n}\int_0^x m_{G_{\frac{1}{n}X'X}}\left(-\tfrac{(1-V(x))}{nv(t)}\right)dt\right]=\sigma^2\int_0^x v(t)\left[1-\frac{(1-V(x))}{nv(t)}m_{MP,\gamma}\left(-\tfrac{(1-V(x))}{nv(t)}\right)\right]dt$$

where

$$m_{MP,\gamma}(z)=\frac{\gamma+\sqrt{\gamma^2-2\gamma(z+1)+(z-1)^2}-z-1}{2z}.$$

$$m_{G_{\frac{1}{n}X'X}}(t)\xrightarrow{d} m(t)$$

This limiting distribution can be identified with its Stieltjes transform m_F, which can be described as follows:

$$m_F(t) + \frac{1}{z} = \frac{n}{x}\Big(v_G(t) + \frac{1}{t}\Big), \tag{7.10}$$

where $v_G(z)$ is the unique solution of the nonlinear equation:

$$-\frac{1}{v_G(t)} = t - \frac{x/n}{1+v_G(t)}. \tag{7.11}$$

In our application $\gamma = \frac{x}{n}$,and $H(s) = \delta\left(\frac{1}{Z}\right)$ where $\delta(\cdot)$ is the Dirac delta function. Therefore

$$-\frac{1}{v_G(z)} = z - \frac{\gamma}{Z+v_G(z)}.$$

$$v(n,x) = \int_0^x \frac{v(t)}{x}\left(1 - \sqrt{1 - \frac{4nx}{\left(n + \frac{1-\int_0^x v(s)ds}{v(t)} + x\right)^2}}\right)\left(n + \frac{1-\int_0^x v(s)ds}{v(t)} + x\right)dt$$

Variational Bayes

We will here show that analogous results can be derive using a differnet approach under which the decision make chooses a posterior satisfying certain conditions (variational Bayes approach). Assume we have a parameter space $\boldsymbol{\Theta} = \mathbb{R}^Z$ and a prior $p \in \Delta(\boldsymbol{\Theta})$. Let $\mathcal{D} = \mathbb{R}^{nK(1+xK)}$ be the set of possible signal realizations (training datasets). We denote by $\mathscr{L}(\boldsymbol{D} \mid \boldsymbol{\beta})$ the likelihood of data $\boldsymbol{D} \in \mathcal{D}$ given $\boldsymbol{\beta} \in \boldsymbol{\Theta}$. Suppose the agent's posterior belief upon observing $\boldsymbol{D}$ is $q_D \in \Delta(\boldsymbol{\Theta})$. Standard Bayesian updating says:

$$q_D(\boldsymbol{\beta}) = \frac{\mathscr{L}(\boldsymbol{D} \mid \boldsymbol{\beta})\, p(\boldsymbol{\beta})}{\sum_{\boldsymbol{\beta}' \in \boldsymbol{\Theta}} \mathscr{L}(\boldsymbol{D} \mid \boldsymbol{\beta}')\, p(\boldsymbol{\beta}')}.$$

Recall the Kullback-Leibler (KL) divergence

$$D(q \,\|\, p) = \sum_{\boldsymbol{\beta} \in \boldsymbol{\Theta}} q(\boldsymbol{\beta}) \ln \frac{q(\boldsymbol{\beta})}{p(\boldsymbol{\beta})},$$

which is always nonnegative. A well-known variational characterization of Bayesian up-

dating is:

$$\arg\min_{q\in\Delta(\boldsymbol{\Theta})}\left\{D(q\,\|\,p)\;-\;\sum_{\boldsymbol{\beta}\in\boldsymbol{\Theta}}q(\boldsymbol{\beta})\ln\mathscr{L}(\boldsymbol{D}\mid\boldsymbol{\beta})\right\}\;=\;\text{(Bayes rule solution)}.$$

In other words, the posterior q solves

$$\min_{q\in\Delta(\boldsymbol{\Theta})}D(q\,\|\,p)\;-\;\sum_{\boldsymbol{\beta}\in\boldsymbol{\Theta}}q(\boldsymbol{\beta})\ln\mathscr{L}(\boldsymbol{D}\mid\boldsymbol{\beta}).$$

7.1.1 Revisiting the Benchmark

Recall that

$$\boldsymbol{D}\;=\;(\boldsymbol{X},\boldsymbol{y}),$$

where $\boldsymbol{y}\in\mathbb{R}^{nZ}$ and $\boldsymbol{X}\in\mathbb{R}^{n\times(kZ)}$.

Proposition 13. *The posterior distribution of $\boldsymbol{\beta}|\boldsymbol{D}$ characterized in Proposition 1 is the solution to*

$$\min_{q\in\Delta(\boldsymbol{\Theta})}\;D(q\,\|\,p)\;+\;\frac{1}{(1-V(x))}\sum_{\boldsymbol{\beta}\in\boldsymbol{\Theta}}q(\boldsymbol{\beta})\,\frac{(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}_x)'(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}_x)}{2}.$$

Proof. The likelihood factorizes as:

$$\mathscr{L}(\boldsymbol{D}\mid\boldsymbol{\beta})\;\equiv\;\mathscr{L}(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\beta}_x)\,p(\boldsymbol{X}),$$

using the assumption $\boldsymbol{X}\;\perp\;\boldsymbol{\beta}_x$. Because $p(\boldsymbol{X})$ does not depend on $q(\cdot)$, the Bayesian update is equivalently given by

$$\min_{q\in\Delta(\boldsymbol{\Theta})}D(q\,\|\,p)\;-\;\sum_{\boldsymbol{\beta}\in\boldsymbol{\Theta}}q(\boldsymbol{\beta})\;\ln\mathscr{L}\big(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\beta}_x\big).$$

If the likelihood is Gaussian with variance $(1-V(x))$, namely

$$\mathscr{L}\big(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\beta}_x\big)\;=\;\left(\sqrt{\tfrac{1}{2\pi\,(1-V(x))}}\right)^{nZ}\exp\!\left(-\,\tfrac{(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}_x)'(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}_x)}{2\,(1-V(x))}\right),$$

then up to constants independent of q, the objective becomes

$$\min_{q\in\Delta(\Theta)} D(q \,\|\, p) \;+\; \frac{1}{1-V(x)} \sum_{\boldsymbol{\beta}\in\boldsymbol{\Theta}} q(\boldsymbol{\beta}) \, \frac{(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}_x)'(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}_x)}{2}.$$

□

Now suppose the agent relies on a different prior described by CVDF $G(\cdot)$. We will say that the agent behaves more confidently compared to $F(\cdot)$ at x if and only if $G(x) \geq v(x)$. Therefore if x covariates are collected the neglected variance is $(1 - G(x))$. The agent's update is thus given by minimizing:

$$\min_{q_w\in\Delta(\Theta)} D(q_w \,\|\, p) \;+\; \frac{\lambda}{(1-G(x))} \sum_{\boldsymbol{\beta}\in\boldsymbol{\Theta}} q_w(\boldsymbol{\beta}) \, \frac{(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}_x)'(\boldsymbol{y}-\boldsymbol{X}\,\boldsymbol{\beta}_x)}{2}, \tag{7.12}$$

where λ is a "misspecification penalty" that weights the agent's empirical term. The following proposition highlights that there is a clean interpretation of Proposition 13.

Corollary 16. *The solution to Problem 7.12 satisifies*

$$q_w \propto p(\boldsymbol{\beta}) \left[f\left((\boldsymbol{w}, \boldsymbol{X}) \,|\boldsymbol{\beta} \right) \right]^{\lambda}.$$

Therefore λ can be interperted as a parameter which weighs the reliance on the data to account for misspecification.

Proposition 14. *A decision maker with prior CDF $G(\cdot)$ will have the posterior characterized in Proposition 1 if and only if*

$$\lambda^*(x) \equiv \frac{1-G(x)}{1-V(x)}.$$

Consider an agent updating their belief about a prediction problem based on an alternative prior $G(\cdot)$, which differs from the original prior $F(\cdot)$. The key result states that the relative weight placed on empirical data versus prior beliefs is given by $\lambda^*(x) = \frac{1-G(x)}{1-V(x)}$. If $G(x) < V(x)$, then $\lambda^*(x) > 1$, meaning the agent discounts the prior more and relies more on empirical observations due to a greater perceived level of uncertainty. Conversely, if $G(x) > V(x)$, then $\lambda^*(x) < 1$, indicating the agent places more trust in their prior and behaves more confidently. This formulation suggests that different initial beliefs can lead to

identical posterior distributions if the updating process adjusts for the prior's misspecification, making it difficult to distinguish between Bayesian and non-Bayesian learning solely based on observed updates. This formulation highlights a key source of *observational equivalence:* an econometrician observing only the agent's posterior inference may be unable to distinguish whether deviations from standard Bayesian updating stem from prior heterogeneity or from a non-Bayesian adjustment rule that corrects for model misspecification.

We define the elastivicty of residual variance $\bar{H}(\cdot)$ at covariate x as

$$\varepsilon_{\bar{H}}(x) = |\frac{xh(x)}{\bar{H}(x)}|,$$

which captures the percentage decrease in unexplained variance in response to a percentage increase in the fraction of covariates observed. Essentially this measure captures how rapidly uncertainty shrinks as they observe more data, i.e. how diminishing she expects the returns to new covariates to be. If $\varepsilon_G(x) \geq \varepsilon_V(x)$,a marginal increase in the amount of data observed x will increase the level of overconfidence of $G(\cdot)$ relative to $V(\cdot)$.

Corollary 17. *$\lambda^*(x)$ is increasing in x if and only if $\varepsilon_G(x) \geq \varepsilon_V(x)$.*

If $\varepsilon_G(x) \geq \varepsilon_V(x)$, then $\lambda^*(x)$ is increasing in x, meaning the discrepancy in confidence between the two priors grows as more data is observed. This implies that if an agent with $G(\cdot)$ becomes more overconfident relative to $V(\cdot)$ as they observe additional covariates, the penalty needed to align their inference with a standard Bayesian approach must increase. In practice, this means that an agent relying on a miscalibrated prior may appear to increasingly deviate from Bayesian updating as their dataset expands.

These results highlight that observed deviations from standard Bayesian learning can emerge due to prior misspecification rather than fundamentally different learning processes. If an agent's prior systematically underestimates or overestimates uncertainty, their updates may compensate in a way that mimics alternative updating rules. Furthermore, as more covariates are observed, the elasticity of residual variance determines whether the agent's level of confidence relative to a standard Bayesian increases or decreases. This is crucial in settings where an econometrician aims to infer the nature of an agent's belief system, as differences in observed behavior may stem from adjustments to model misspecification rather than intrinsic deviations from Bayesian rationality.

Proposition 15. *There exists a unique threshold $\Delta^*(\delta, n, A) \geq 0$ such that E enters and sells for*

both periods if and only if $\Delta \geq \Delta_{reentry}(A)$. *Especially,*

$$\Delta^*(\delta, n, A) \equiv \frac{V(n, A) + \delta\left(V(n+1, A) - V(1, A)\right)}{\delta + 1}.$$

Proof. Suppose E has sold in Period 1. In Period 2, user gets

$$U = \begin{cases} v_E + V(\lambda, S) - \sigma^2 - p_E^2 & \text{if use } E \\ v_I + V(n\lambda, S) - \sigma^2 - p_I^2 & \text{if use } I \end{cases}$$

Therefore E can therefore make a profit of at most

$$\bar{p}_2^E = \Delta - V\left(n\lambda, S\right) + V\left(\lambda, S\right).$$

Suppose I has sold in Period 1. In Period 2, user gets

$$U = \begin{cases} v_E - \sigma^2 - p_E^2 & \text{if use } E \\ v_I + V\left((n+1)\lambda, S\right)) - \sigma^2 - p_I^2 & \text{if use } I \end{cases}$$

I can therefore make a profit of

$$\bar{p}_2^I = V\left((n+1)\lambda, S\right) - \Delta.$$

In period 1, the minimal price $X \in \{I, E\}$ is willing to charge is

$$\underline{p}_1^X + \delta \bar{p}_2^X = 0 \iff \underline{p}_1^X = -\delta \bar{p}_2^X$$

What is the maximum price E can charge in Period 1 to win over consumers if I charges its minimal price $\{\underline{p}\}_1^I$? Consumers get

$$U = \begin{cases} v_E - \sigma^2 - p_E^1 & \text{if buy from } E \\ v_I + V(n\lambda, S) - \sigma^2 - p_I^1 & \text{if buy from } I \end{cases}$$

$$\bar{p}_1^E \equiv \Delta - V\left(n\lambda, S\right) + \underline{p}_1^I.$$

Therefore E will enter the market if and only if

$$\bar{p}_1^E \geq \underline{\{p\}}_1^E \iff \Delta - V(n\lambda, S) \geq \delta\left(\bar{p}_2^I - \bar{p}_2^E\right) = \delta\left(V((n+1)\lambda, S) - 2\Delta + V(n\lambda, S) - V(\lambda, S)\right)$$

$$\Delta \geq \Delta^*(\delta, n, \lambda, S) \equiv \frac{V(n\lambda, S) + \delta\left(V((n+1)\lambda, S) + V(n\lambda, S) - V(\lambda, S)\right)}{1 + 2\delta}$$

If E wins the price in Period 1 will be the highest price E can charge whilst assuring I will not be active

$$p_1^* = \bar{p}_1^E = \Delta - V(n;k) + \underline{p}_1^I = \Delta - V(n;k) - \frac{V(n+1;k) - \Delta}{\frac{1}{\delta} - 1} < 0$$

$$p_2^* = \bar{p}_2^E = \Delta - V(n;k) + V(1;k)$$

And

$$CS_1^I = \left(v_I + V(n;k) - \sigma^2 + \frac{V(n+1;k) - \Delta}{\frac{1}{\delta} - 1}\right)\delta$$

$$CS_2^E = \left(v_I + V(n;k) - \sigma^2\right)(1 - \delta)$$

So surplus if E wins,

$$CS_E = \left(\frac{V(n+1;k) - \Delta}{\frac{1}{\delta} - 1}\right)\delta + v_I + V(n;k) - \sigma^2$$

What is the maximum price I can charge in Period 1 to win over consumers if E charges its minimal price $\underline{p}_1^E$ (i.e. deterring E's entry)?

$$\bar{p}_1^I \equiv V(n;k) - \Delta + \underline{p}_1^E = \frac{V(n,k) - \Delta + \delta V(1,k)}{1 - \delta}.$$

Hence if I wins

$$p_1^* = \underline{p}_1^I = -\frac{V(n+1;k) - \Delta}{\frac{1}{\delta} - 1} < 0$$

$$p_2^* = \bar{p}_2^I = V(n+1;k) - \Delta > 0$$

And

$$CS_1^I = \left(v_I - \sigma^2 + V(n;k) + \frac{V(n+1;k) - \Delta}{\frac{1}{\delta} - 1} \right) \delta$$

$$CS_2^I = \left(v_E - \sigma^2 \right) (1 - \delta)$$

Suppose I wins the competition,

$$CS_I = \left(v_I + V(n;k) + \frac{V(n+1;k) - \Delta}{\frac{1}{\delta} - 1} \right) \delta + v_E(1 - \delta) - \sigma^2$$

□