Bayesian Regression Markets

Thomas Falconer FALCO@DTU.DK
Department of Wind and Energy Systems
Technical University of Denmark
Elektrovej, 325, 105
Kgs. Lyngby, 2800, Denmark

Jalal Kazempour JALAL@DTU.DK
Department of Wind and Energy Systems
Technical University of Denmark
Elektrovej, 325, 105
Kgs. Lyngby, 2800, Denmark

Pierre Pinson P.PINSON@IMPERIAL.AC.UK
Dyson School of Design Engineering
Imperial College London
London, SW7 2DB, United Kingdom

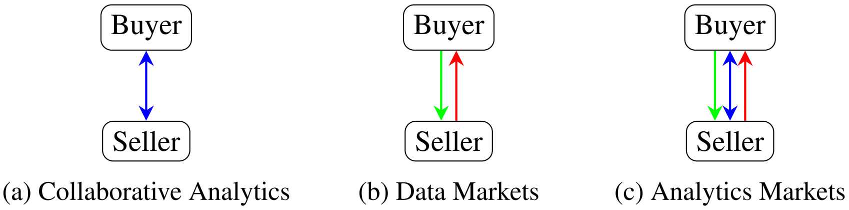
Abstract

Machine learning tasks are vulnerable to the quality of data used as input. Yet, it is often challenging for firms to obtain adequate datasets, with them being naturally distributed amongst owners, that in practice, may be competitors in a downstream market and reluctant to share information. Focusing on supervised learning for regression tasks, we develop a *regression market* to provide a monetary incentive for data sharing. Our proposed mechanism adopts a Bayesian framework, allowing us to consider a more general class of regression tasks. We present a thorough exploration of the market properties, and show that similar proposals in current literature expose the market agents to sizeable financial risks, which can be mitigated in our setup.

Keywords: regression, bayesian inference, collaborative analytics, data markets, game theory

1 Introduction

As machine learning models continue to demand more data, practitioners often concentrate on challenges associated with data processing, feature selection and engineering, in addition to model building and validation, to optimize performance. These efforts are typically based on the assumption that data is readily available via some central authority, yet in practice, datasets are inherently distributed amongst owners with heterogeneous characteristics (e.g., privacy preferences). This has motivated several developments in the field of collaborative analytics, also known as federated learning (Figure 1a), where models are trained on local servers without the need for data centralization, thereby preserving privacy and distributing the computational burden (Kairouz et al., 2019). However, this method for data sharing is *incentive-free*, relying on the critical assumption that owners are willing to collaborate (i.e., by sharing their private information) altruistically. This strong assumption may be violated if owners are competitors in a downstream market environment (Gal-Or, 1985). Consequently, a fruitful area of research has emerged that proposes to instead *commoditize* data within a market-based framework, where compensation (e.g., remuneration) can be used as an incentive for collaboration (Bergemann and Bonatti, 2019).



(a) Collaborative Analytics (b) Data Markets (c) Analytics Markets

Figure 1: Schematic illustration of existing frameworks for data sharing with multiple buyers and sellers, where each figure depicts a building block consisting of a single interaction. The blue, red and green arrows indicate computational, information and monetary transactions between the buyer and the seller, respectively.[1]

Information economics has been well established in game theory literature since the 1980s (Gal-Or, 1985), with early works focusing on incentive-free data sharing, both publicly (Morris and Shin, 2002) and within local information channels (Dahleh et al., 2016). Over recent years, interest in data monetization has grown rapidly, for which the first proposals considered *data markets* (Figure 1b), allowing buyers to purchase raw data (i.e., features) from sellers through bilateral transactions (Rasouli and Jordan, 2021). Whilst a seemingly practical way to acquire data from others, the value of a feature to the buyer depends on the *analytics task* at hand, hence pricing raw data in these markets is difficult (Cong et al., 2022), especially considering privacy-preservation (Acemoglu et al., 2022). Early game theoretic notions of informational efficiency in financial markets (Hayek, 1986) also laid the groundwork for the closely related topic of *prediction markets*, which are designed to aggregate information about the likelihood of future events. These markets establish financial securities, with their payout contingent on the outcome of the events and trading prices construed as collective forecasts (Chen and Pennock, 2010). Traders effectively wager on repeatedly occurring (Bottazzi and Giachini, 2019) or one-shot (Manski, 2006) events.

Prediction markets are widely embraced within the machine learning community, with the common goal of forecasting future outcomes and familiar use of convex analysis (Frongillo and Waggoner, 2018). Recent works propose these markets as an effective means of crowdsourcing rows of data for machine learning by providing incentives to update a centralized hypothesis (Abernethy and Frongillo, 2011). So-called *machine learning markets* extend this concept to complex multivariate prediction problems, where algorithms can compete using strategies that are either dynamic (Jahedpari et al., 2017) or fixed, using betting functions (Barbu et al., 2012; Lay and Barbu, 2012), utility functions (Storkey, 2011), or static risk measures (Hu and Storkey, 2014), with the possibility of bounded regret for the owner of the task (Chen and Vaughan, 2010; Abernethy et al., 2013). The connection to data markets is clear; both provide incentives for agents to disclose relevant private information. That being said, prediction markets are in impractical for the general case of information sharing as support agents need to decide which tasks to make predictions for, yet the relevance of a dataset to a particular task is unknown *a priori*, especially when considering the inevitability of overlapping information.

Instead, by recognizing that the commodity need not be data itself, ideas from both data markets and collaborative analytics can be combined to form *analytics markets* (Figure 1c), real-time mechanisms that match features to analytics tasks based on the enhanced capabilities provided (Agarwal et al., 2019). As in Pinson et al. (2022a), we focus on applications in which the analytics task de-

[1] Adapted from Pinson et al. (2022b).

scribes a regression model along with the process for inference used for training (i.e., our attention centers on *regression markets*). This builds upon current literature concerning data elicitation from strategic (Dekel et al., 2010) and privacy-sensitive (Cummings et al., 2015) owners. In this context, owners of regression models seek to enhance predictive performance, for which they have a private valuation (e.g., their value of forecast accuracy in a downstream decision-making process). Their public bids, which may not equal private valuations, are then used to set the price. Sellers propose their own data as features and are remunerated based on their marginal contributions to the improved model-fitting. The market revenue is therefore a function of both the market price and the overall enhancement in predictive performance.

In this work, we develop a regression market framework that enables Bayesian methods, allowing us to consider a more general class of regression tasks compared with the frequentist frameworks of previous works that treat parameters as fixed quantities (Pinson et al., 2022a). Neglecting parameter estimation uncertainty can yield overly confident predictions that miss-represent the true level of variability in the data. In contrast, Bayesian analysis offers a principled framework for modelling parameter uncertainty that subsumes many frequentist regression methods, providing the buyer with richer and more nuanced information about future outcomes. Treating parameters as random variables acknowledges that for a given regression task, the value of a particular feature, and hence the remuneration to its owner, is itself subject to uncertainty. Conceptually, this is a paradigm shift in the field of regression markets and opens new avenues of research into uncertainty-aware mechanism designs for data sharing.

We also provide a thorough exploration of the market properties. In previous works (e.g., Agarwal et al., 2019, Pinson et al., 2022a), such properties are only considered in expectation, however we demonstrate that many can be violated in a single-shot of the market, exposing agents to considerable financial risks, especially when a limited number of observations are available. We propose an alternative framework to value features based on the expected information gain provided as opposed to simply the impact on the objective value, which has the propensity to mitigate these risks entirely, despite being asymptotically equivalent.

The remainder of the paper is structured as follows: Section 2 introduces the market agents and the design of our proposed market mechanism. Section 3 assesses the theoretical market properties of our proposal and presents methods for mitigating financial risks exhibited by the agents. Section 4 and Section 5 illustrate our findings through a set of simulation-based and real-world case studies, respectively. Finally, Section 6 gathers our conclusions and perspectives for future work.

2 Market Setup

We begin by presenting a general mechanism design for the analytics market, which is intended to be hosted on a platform capable of handling both the analytical (e.g., parameter inference) and market-based (e.g., revenue allocation) components together in tandem. The market comprises multiple agents—we define a *transaction* as an exchange between a single *central agent* (i.e., a buyer) and multiple *support agents* (i.e., sellers), at a particular point in time, whereby the central agent seeks to enhance the predictive performance of a *regression task*, for which the support agents propose their own data as input features.[2]

[2] Whilst this definition preserves the capacity for parallel transactions, we assume each is independent, thereby disregarding data exclusivity, wherein the same data can only be sold a finite number of times (Cao et al., 2017), as well as any externalities this may exert (Agarwal et al., 2020).

Although there may be multiple sellers, the enhancement in performance received by the central agent is a function of the complete set of information available. We hence view a transaction as being between the central agent and a *single* monopolistic support agent, a single agent with access to the complete set of features with only one item for sale, specifically the available *loss reduction*. The private valuation is assumed to equal the public bid (i.e., the valuation for a marginal improvement in model fitting). The central agent is allocated the full performance enhancement offered by the monopolistic support agent, and the payment collected is a function of these values. One can view this as a specification of the mechanism proposed in Agarwal et al. (2019), where the monopolistic support agent offers several possible performance enhancements, each representing varying degrees of obfuscation of the true data, characterized by the discrepancy between the bid of the central agent and the market price. Since we assume the market price is exogenous, our work is concerned specifically with the regression analysis and subsequent revenue allocation, as opposed to the pricing mechanism.

2.1 Market Agents

Let $\mathcal{A}$ denote the set of market agents, one of which $c \in \mathcal{A}$ is the central agent seeking to enhance their forecasts. The remaining agents $a \in \mathcal{A}_{-c}$ are support agents that propose data as input features, whereby $\mathcal{A}_{-c} = \mathcal{A} \setminus \{c\}$. The central agent is characterized by their interest in a particular stochastic process $\{Y_t\}$, defined as a set of successive random variables Y_t indexed over discretized time steps t. Eventually, a time-series $\{y_t\}$ is observed, comprising realizations from $\{Y_t\}$ (i.e., one per time step). Instead of assuming that a particular characteristic of Y_t is sought (e.g., the expected value, a specific quantile, etc.), we rather model the entire distribution, albeit conditioned on the observed data; the characteristic extracted by the central agent is simply treated as some downstream decision-making process.

We write $\mathbf{x}_{\mathcal{I},t}$ as the vector of input features at time t, indexed by the ordered set $\mathcal{I}$. Each agent $a \in \mathcal{A}$ owns a subset $\mathcal{I}_a \subseteq \mathcal{I}$ of indices, such that the features are distributed as follows: the central agent c owns the subset $\mathcal{I}_c \subset \mathcal{I}$. Each support agent $a \in \mathcal{A}_{-c}$ also owns a subset of features, with indices $\mathcal{I}_a \subset \mathcal{I}$, such that $|\mathcal{I}_c| + \sum_{a \in \mathcal{A}_{-c}} |\mathcal{I}_a| = |\mathcal{I}|$. We write $\mathcal{I}_{-c}$ as the set of indices for features owned only by the support agents. Since the data is observed at successive time steps, we let $\mathbf{x}_t = [x_{1,t}, \dots, x_{|\mathcal{I}|,t}]^\top$ be the vector of values for all features at time t. When only a particular subset of features $C \subseteq \mathcal{I}$ is used, we add an index for the set itself, such that the vector of values for features in C at time t is denoted by $\mathbf{x}_{C,t}$. We write $\mathcal{D}_{C,t} = \{\mathbf{x}_{C,t'}, y_{t'}\}_{\forall t' \leq t}$ to be the set of input-output pairs for a particular subset of features observed over a set of discrete time indices $t' \in \{1, \dots, t\}$ up until time t.

2.2 Regression Task

To instigate a transaction, the central agent first posts a regression task to the market platform, which describes the particular model for which they seek to enhance predictive performance. We consider the problem of interpolating through data (i.e., the observations $\{y_t\}$) under the assumption that the target signal is subject to noise, whilst the input features are noise-free. Let us define an interpolant as a mapping f between a subset of features $\mathbf{x}_{C,t}$ and a real-valued scalar, which may represent the expected value of the target signal conditioned on the inputs such that

$$f : \mathbf{x}_{C,t} \in \mathbb{R}^{|C|} \mapsto \mathbb{E}[Y_t \,|\, \mathbf{x}_{C,t}] \in \mathbb{R}, \quad \forall t, \ \forall C. \tag{1}$$

We focus solely on parametric regression, and further limit ourselves to functions that can be expressed as linear in their coefficients, with a view to preserve convexity and later guarantee certain market properties. We obtain a rich class of models by considering linear combinations of a fixed set of nonlinear functions (i.e., basis functions). Let $\mathbf{w}_C \in \mathbb{R}^{|C|}$ be a vector of coefficients that is used to parameterize the mapping in (1), which, for notational brevity, we assume to be part of a general set of free parameters Θ_C that shall be inferred from data. We write $\varphi(\mathbf{x}_{C,t})$ to be the vector of basis functions specified by the central agent, such that the linear interpolant is given by

$$f(\mathbf{x}_{C,t}, \mathbf{w}_C) = \mathbf{w}_C^\top \varphi(\mathbf{x}_{C,t}), \quad \forall t,\ \forall C, \tag{2}$$

where we assume that the vector of basis functions under consideration invariably incorporates a dummy basis function (i.e., $\psi_0(\mathbf{x}_{C,t}) = 1,\ \forall t$) which is included as part of the feature set owned by the central agent.

Remark 1 *In general, the central agent need not own any feature themselves and thereby crowdsource their predictions, in which case only the dummy term is provided and all predictive performance is supplied by the features owned by support agents.*

We model the target variable as a deviation from the deterministic mapping in (2) under a zero-mean additive noise process, the parameters of which are also held in Θ_C. Bayesian inference treats the parameters as random variables and aims to infer their distribution by incorporating prior beliefs, which are updated as new data is observed. Let $h \in \mathcal{H}$ be a hypothesis, a set of fixed assumptions that restricts the space of possible regression models, comprising the vector of basis functions, as well as the functional forms of two probability distributions: both the prior (i.e., plausible parameter values) and the likelihood (i.e., the probability of the data conditioned on the parameters). The regression task posted to the market platform by the central agent at time t is therefore fully described by a hypothesis and the observed data.

2.3 Market Clearing

We suppose each support agent is willing to accept any nonnegative payment if their data is deemed useful. However, we acknowledge that support agents may prefer to condition their participation on a minimum payment to, for instance, reflect privacy costs (Acquisti et al., 2016). Certain features in the market may also be irrelevant for the regression task, hence we assume the following.

Assumption 1 *Given the specified hypothesis, the market operator is tasked with selecting relevant features (e.g., by means of cross-validation), such that only those that reduce the expected value of the loss function are considered.*

As our problem is convex, any additional feature cannot increase the in-sample loss in expectation. Therefore we refer here to the out-of-sample loss, evaluated for instance with cross-validation, since by submitting a feature to the market a support agent provides the market operator with training data. For discussions on conventional feature selection problems, cross-validation in a Bayesian context and methods for marginal likelihood optimization, the reader is referred to Guyon and Elisseeff (2003), Watanabe and Opper (2010) and Fong and Holmes (2020), respectively.

Remark 2 *The assumption of an ex-ante feature selection process is merely to align with standard practices in machine learning operations (MLOps). One could also consider feature selection using either endogenous or ex-post processes, by adopting informative priors (Han et al., 2022b) or* *applying post-processing steps (Liu, 2020), respectively, albeit at a cost to the market. In particular, such approaches may undermine potential improvements in predictive performance or lead to skewed payments as a price to pay for avoiding ex-ante feature selection.*

Once the entire set of required market inputs have been received, the market operator is tasked with clearing the market. This procedure involves several steps, namely parameter inference, performance evaluation, payment collection and revenue allocation.

2.3.1 Parameter Inference

Based on all of the observations up until time t, we can summarize our updated beliefs regarding the parameters through the posterior distribution, which, by virtue of Bayes theorem, is proportional to the product of the likelihood and the prior such that

$$p(\Theta_C|\mathcal{D}_{C,t}) \propto p(\mathcal{D}_{C,t}|\Theta_C)p(\Theta_C), \quad \forall t,\ \forall C. \tag{3}$$

For an arbitrary choice of prior, the posterior may not be available in closed-form, necessitating methods for approximate Bayesian inference (e.g., Monte-Carlo integration) to be employed. However, for a known functional form of the likelihood, priors that are conjugate can result in posteriors with tractable, well-known densities. It may be appropriate to allow the moments of this distribution to vary in time, thereby accounting for nonstationarities in any of the underlying processes that can lead to concept drift. In a Bayesian treatment of linear regression, batch inference can be viewed as a specification of this more general *online learning* problem, whereby the parameters are updated in a recursive manner. To see this, we re-write the expression in (3) as a series of sequential updates such that

$$\begin{aligned} p(\Theta_C|\mathcal{D}_{C,t}) &\propto p(\mathcal{D}_{C,t}|\Theta_C)p(\Theta_C|\mathcal{D}_{C,t-1}), && \forall t,\ \forall C, && \text{(4a)} \\ &= p(\mathcal{D}_{C,t}|\Theta_C)\left[p(\Theta_C)\prod_{t'<t} p(\mathcal{D}_{C,t'}|\Theta_C)\right], && \forall t,\ \forall C. && \text{(4b)} \end{aligned}$$

To place greater weight on more recent data, we can augment this update step to use exponential forgetting, where the importance given to past information decreases exponentially. This generally translates to the idea of likelihood flattening, whereby we reformulate (4b) as a trade-off between the posterior at the previous time step and the original prior (i.e., before any data had been observed), thereby emulating a loss in belief with respect to the historic estimates (Peterka, 1981). This trade-off between the two distributions can be framed as the problem of finding the probability density function with minimum expected Kullback–Leibler (KL) divergence (i.e., relative entropy) between them (Kulhavỳ and Zarrop, 1993), which has a unique solution enabling us to replace the prior at time t in (4a) with the following:

$$\begin{aligned} p(\Theta_C|\mathcal{D}_{C,t-1},\tau) &= \underset{p^*}{\text{argmin}}\ \tau\, D_{\text{KL}}\left(p^* \,\|\, p(\Theta_C\,|\mathcal{D}_{C,t-1})\right) + (1-\tau)\, D_{\text{KL}}\left(p^* \,\|\, p(\Theta_C)\right), && \forall t,\ \forall C, && \text{(5a)} \\ &\propto p(\Theta_C|\mathcal{D}_{C,t-1})^{\tau}\, p(\Theta_C)^{1-\tau}, && \forall t,\ \forall C, && \text{(5b)} \end{aligned}$$

where the variable p^* denotes the resultant density function, $D_{\text{KL}}(\cdot \,\|\, \cdot) \in \mathbb{R}_+$ is the KL divergence and the parameter $\tau \in [0,1]$ is analogous to the forgetting factor in time-weighted Least-Squares fitting (Vahidi et al., 2005). Observe that, as $\tau \mapsto 1$, the prior information available at time t becomes identical to the posterior information at the previous time step as in (4b), emulating batch learning, whereas when $\tau = 0$, the previous information is *forgotten* and we resort to the original (i.e., flat) prior. For convenience, we treat τ as a time-invariant hyperparameter, however for a full Bayesian treatment one could also infer its value jointly, together with Θ_C.

2.3.2 Performance Evaluation

Given observations up until time t, we can evaluate the performance of a specific subset of features by making a prediction for a time step t^*, conditioned on the observed input features. For now, we consider the general case where t^* is an arbitrary time step to account for both in-sample (i.e., $t^* \leq t$) and out-of-sample (i.e., $t^* > t$) situations. In Bayesian regression analyses, a *prediction* is typically defined to be the computation of the posterior predictive distribution, derived by integrating out the parameters using the convolution of the likelihood with the posterior, given by

$$p(y_{t^*}|\mathbf{x}_{C,t^*}, \mathcal{D}_{C,t}) = \int p(y_{t^*}|\mathbf{x}_{C,t^*}, \mathcal{D}_{C,t}, \Theta_C) p(\Theta_C|\mathcal{D}_{C,t}) d\Theta_C, \quad \forall C, \tag{6}$$

which for brevity we hereafter omit the training dataset and write as $p(y_{t^*}|\mathbf{x}_{C,t^*})$. In order to evaluate predictive performance, we define a loss function ℓ. If a model describing a particular characteristic of Y_t is sought, then this loss function could be set as a direct function of the residuals (i.e., by extracting the corresponding point from the predictive distribution). However, as we intend to provide the entire predictive distribution, we can generally define ℓ as a function of the predictive density (i.e., $\ell_{C,t^*} : p(y_{t^*}|\mathbf{x}_{C,t^*}) \mapsto \mathbb{R}$), assuming the following.

Assumption 2 *The mapping ℓ is a negatively-oriented strictly proper scoring rule. Accordingly, it holds that: (i) for any two models, the one that provides the more accurate description of the data will render a lower score; and (ii) the lowest score is uniquely obtained when the prediction converges to the true distribution.*

In an online setup with exponential forgetting, evaluating ℓ at each time step can be perceived as a recursive and adaptive time-varying estimator of its expected value; adaptive in the sense that a greater weight is placed on more recent data. Hence, the in-sample estimate of $\mathbb{E}[\ell]$ for a particular subset of features at time t with respect to (6) can be described by the following recursion:

$$\mathbb{E}[\ell_C]_t = (1 - \tau)\, \ell_{C,t} + \tau\, \mathbb{E}[\ell_C]_{t-1}, \quad \forall t,\ \forall C, \tag{7}$$

where to consider the case of out-of-sample evaluation (e.g., if t is the next available time step) we simply replace $\mathcal{D}_{C,t}$ in (7) with the most recent set of observations, $\mathcal{D}_{C,t-1}$.

2.3.3 Payment Collection

As well as a regression task, our market requires the central agent to post to the platform their public bid, denoted by $\lambda \in \mathbb{R}_+$, which represents an exogenous linear mapping between a unit improvement in ℓ and the corresponding downstream monetary reward that would be earned, thereby determining the market price. We do acknowledge the weakness of this linearity assumption, as in practice, λ may be, for instance, a logarithmic function of the central agent's revenue (i.e., further reductions in ℓ may provide diminishing returns), albeit with exponential costs for the support agents. Nevertheless, we leave it as future work to explore the optimal functional form of λ. The market revenue at time t is equal to the payment collected from the central agent, denoted $\pi_{c,t}$, which is a function of λ, as well as the overall improvement in the objective, such that

$$\pi_{c,t} = \lambda \left(\mathbb{E}[\ell_{\mathcal{I}_c}]_t - \mathbb{E}[\ell_{\mathcal{I}}]_t \right), \quad \forall t. \tag{8}$$

2.3.4 Revenue Allocation

Once the market has been cleared, the natural question that follows is: *how can we fairly allocate market revenue amongst support agents?* To answer this question, several auction-based setups have been proposed, considering topics such as privacy (Koutsopoulos et al., 2015), data exclusivity (Cao et al., 2017) and negative externalities exhibited by the market agents (Agarwal et al., 2020). Other methods bear upon interoperability in machine learning, adopting widely adopted solution concepts (namely, semivalues) for the problem of attribution in cooperative game theory to allocate revenue amongst support agents directly (Dubey et al., 1981). The benefit of this approach being that these solution concepts are generally characterized by a collection of axioms that yield desirable market properties by design (Ghorbani and Zou, 2019), specifically: symmetry, efficiency, null-player and additivity. For a definition of these axioms, the reader is referred to Chalkiadakis et al. (2011).

If we frame features as players and their interactions as a cooperative game, the semivalue of a feature can be defined as its expected marginal contribution towards a set of other features, weighted solely based on the size of the sets. For many applications, the semivalue of choice is the *Shapley value* (Shapley, 1997), the unique value that satisfies all of the four axioms stated above. Given the set $\mathcal{I}_{-c}$ of indices corresponding to features owned by the support agents, let $v : \mathcal{C} \in \mathcal{P}(\mathcal{I}_{-c}) \mapsto \mathbb{R}$ be a characteristic function that maps the power set $\mathcal{P}(\mathcal{I}_{-c})$ of all features with indices in $\mathcal{I}_{-c}$ to a real-valued scalar, where the set $\mathcal{C}$ denotes a coalition in the cooperative game. The Shapley value is given by

$$\phi_{i,t} = \sum_{\mathcal{C} \in \mathcal{P}(\mathcal{I}_{-c} \setminus \{i\})} \frac{|\mathcal{C}|!(|\mathcal{I}_{-c}| - |\mathcal{C}| - 1)!}{|\mathcal{I}_{-c}|!} m_{i,t}(\mathcal{C} \cup \{\mathcal{I}_c\}), \quad \forall i \in \mathcal{I}_{-c}, \forall t, \tag{9}$$

where $m_{i,t}(\cdot)$ is the marginal contribution, often defined as $m_{i,t}(\cdot) = v_t(\cdot) - v_t(\cdot \cup \{i\})$ in relation to the characteristic function.[3]

The Shapley value is then used to allocate market revenue. Given our estimator of the expected loss varies with time, attributions are time-varying too, as well as the market revenue in (8). In line with (7), the expected Shapley value at time t is given by

$$\mathbb{E}[\phi_i]_t = (1 - \tau)\phi_{i,t} + \tau \mathbb{E}[\phi_i]_{t-1}, \quad \forall i \in \mathcal{I}_{-c}, \forall t. \tag{10}$$

Then, as we evaluate (10) for each feature, the overall payment received by each support agent is simply given by

$$\pi_{a,t} = \sum_{i \in \mathcal{I}_a} \lambda \, \mathbb{E}[\phi_i]_t, \quad \forall a \in \mathcal{A}_{-c}, \forall t. \tag{11}$$

[3] The weight in this discrete expectation assigned to each coalition is defined as such to avoid unnecessary calculations of the marginal contribution of the i-th feature to permutations of the same coalition, which would have equal value by virtue of the symmetry axiom. For instance, $m_{i,t}(\{j\}) \equiv m_{j,t}(\{i\})$, $\forall (i, j) \in \mathcal{I}_{-c}$, $i \neq j$, thus it is computationally favourable to avoid making this calculation twice.

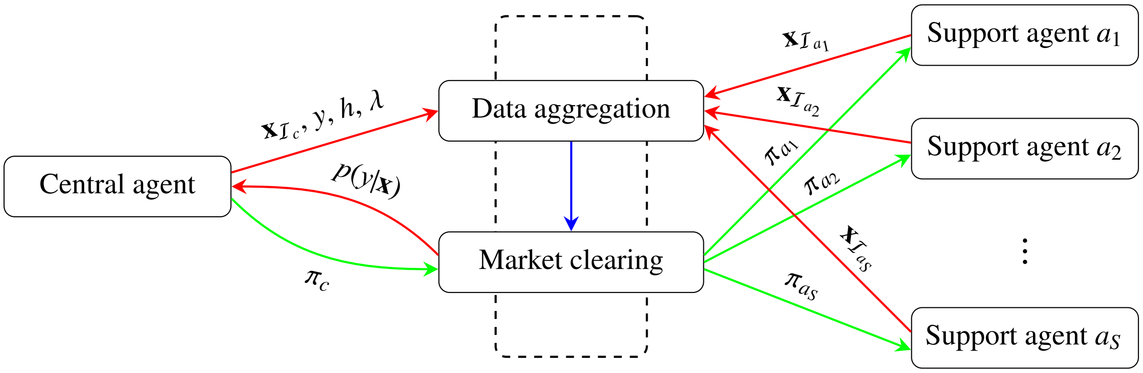


Figure 2: Overview of in-sample market platform operations at time t, with the total number of support agents given by $S = |\mathcal{A}_{-c}|$. The time index t is omitted for brevity. Recall that the blue, red and green arrows indicate computational, information and monetary transactions, respectively.

2.4 Market Dynamics

In practice, MLOps pipelines are typically divided into in-sample (i.e., training) and out-of-sample (i.e., testing) stages. The first stage involves Bayesian inference using observed input-output pairs, whilst at the second stage, a trained model is used for genuine forecasting on previously unseen data, testing its capacity to generalize beyond the training set. We hence adopt the two-stage regression market model proposed in Pinson et al. (2022a), that is, the value of a feature is assessed based on marginal contributions to both the in-sample and out-of-sample estimates of $\mathbb{E}[\ell]$, albeit in separate transactions.

We round off this section with a summary of the key market platform operations—see Figure 2 for a graphical overview. At each time step t, the steps in Algorithm 1 occur until termination (i.e., when a specified time step T has been reached). The out-of-sample market operations are similar, the only differences being that the posterior is not updated before predictions are made, and only when the observation of the target signal arrives is the revenue collected and allocated.

Algorithm 1: In-sample online regression market

```
Input: $\mathbf{x}_{\mathcal{I}_c,t} \in \mathbb{R}^{|\mathcal{I}_c|}$, $y_t \in \mathbb{R}$, $\lambda \in \mathbb{R}_+$, $h \in \mathcal{H}$
Output: $p(y_t|\mathbf{x}_t)$, $[\pi_{a,t} : a \in \mathcal{A}_{-c}]$
while $t \leq T$ do
    $p(\Theta|\mathcal{D}_t, \tau) = p(\mathcal{D}_t|\Theta_{\mathcal{C}})p(\Theta|\mathcal{D}_{t-1}, \tau)$ ;        /* Update posterior */
    $p(y_t|\mathbf{x}_t) = \int p(y_t|\mathbf{x}_t, \Theta)p(\Theta|\mathcal{D}_t)d\Theta$
    $\pi_{c,t} = \lambda(\mathbb{E}[\ell_{\mathcal{I}_c}]_t - \mathbb{E}[\ell_{\mathcal{I}}]_t)$ ;     /* Revenue extracted from central agent */
    for $a \in \mathcal{A}_{-c}$ do
        for $i \in \mathcal{I}_a$ do
            $\mathbb{E}[\phi_i]_t = (1-\tau)\sum_{\mathcal{C}\in\mathcal{P}(\mathcal{I}_{-c}\setminus\{i\})} \frac{|\mathcal{C}|!(|\mathcal{I}_{-c}|-|\mathcal{C}|-1)!}{|\mathcal{I}_{-c}|!} m_{i,t}(\mathcal{C}\cup\mathcal{I}_c) + \tau\mathbb{E}[\phi_i]_{t-1}$
            $\pi_{a,t} = \pi_{a,t} + \lambda\mathbb{E}[\phi_i]_t$ ;     /* Revenue allocated to support agents */
        end
    end
end
```

Remark 3 *In the proposed mechanism above, the market price is solely dependent on the valuation of the central agent, λ, demonstrating that the value of data need not be an intrinsic property, but rather derived from its contribution to the particular analytics task at hand. This indeed assumes features are interchangeable provided the same inferences can be drawn from either, but moreover, the inevitable overlapping information between features would render parameterized valuations for each subset computationally intractable.*

3 Market Properties

The remaining design decision that will affect the market properties relates to the choice of characteristic function, $v_t(\cdot)$, used to value a coalition of features. The Shapley value has indeed emerged as the *de facto* tool for interpreting predictions from complex machine learning models (Guyon and Elisseeff, 2010; Sundararajan and Najmi, 2020; Tsai et al., 2023), yet its application in probabilistic contexts is not yet as well studied, as comparing models outputs as probability distributions is less straightforward that scalars. In general, at a given time t we can set the characteristic function to be equal to the current estimate of the expected loss, described in (7), such that $v_t(C) = \mathbb{E}[\ell_C]_t$, hence the design decision is not the characteristic function itself *per se*, but the particular functional form of ℓ, which recall maps the predictive density to a real value.

In this section we introduce the following market designs: (i) $\mathcal{M}^{\text{MLE}}_{\text{NLL}}$—a frequentist framework based on maximum likelihood estimation (MLE) which values features using the negative logarithm of the likelihood (NLL), (ii) $\mathcal{M}^{\text{BLR}}_{\text{NLL}}$—the analogue of $\mathcal{M}^{\text{MLE}}_{\text{NLL}}$ now in a Bayesian linear regression (BLR) framework, and (iii) $\mathcal{M}^{\text{BLR}}_{\text{KL}-v}$ and $\mathcal{M}^{\text{BLR}}_{\text{KL}-m}$—BLR frameworks that instead value features based on the information gain they provide, measured using the KL divergence.

3.1 Likelihood-based Designs

In Pinson et al. (2022a), a maximum likelihood framework is adopted, treating parameters as fixed, albeit unknown, quantities. The characteristic function can then simply be set to the expected value of the NLL. We denote this frequentist market design by $\mathcal{M}^{\text{MLE}}_{\text{NLL}}$. In the following, we shall analyze the market properties obtained by extending this idea to its Bayesian analogue.

In Bayesian regression we have access to the posterior distribution, from which revenue allocations derived using any random sample could be considered plausible with respect to the frequentist design. To attain the most nuanced representation of uncertainty, we instead provide the predictive distribution derived by marginalizing over the entire space of parameters. A reasonable candidate for the characteristic function is therefore again the NLL, which now incorporates the uncertainty in the parameter estimates in (6), such that

$$\ell_{C,t} = -\log(p(y_t \,|\, \mathbf{x}_{C,t})), \quad \forall t, \ \forall C, \tag{12}$$

with $\mathcal{M}^{\text{BLR}}_{\text{NLL}}$ denoting the corresponding market design. In order for Assumption 2 to be satisfied, the predictive density must be log-concave. Whilst many common distributions are indeed log-concave and could be utilized easily in a maximum likelihood framework, in order to avoid approximation errors in general Bayesian inference, we require the posterior to be available in closed-form, thus the prior and posterior should be conjugate. Therefore, as well as for mathematical convenience, to adhere to this we further assume the following, and leave exploration of alternative hypotheses to future work.

Assumption 3 *The hypothesis space $\mathcal{H}$ comprises only Gaussian likelihood functions along with a conjugate uninformative Gaussian prior.*

Remark 4 *Whilst Assumption 3 assumption is restrictive, it is in fact common in practice (i.e., it is a byproduct of simply using mean-squared error in frequentist regression methods), and still permits non-Gaussian data generating processes, but merely induces misspecifiations in such a case.*

It is worth highlighting the tangible benefits to the central agent by transitioning from frequentist to Bayesian regression analyses. For instance, the additional element of predictive uncertainty provides richer and more nuanced information about future outcomes. In addition, maximum likelihood estimation also has a tendency to render implausible overparameterized models that generalize poorly to out-of-sample analyses. This is especially true when the number of training observations is limited, since increasing model complexity inevitably results in overfitting. In contrast, Bayesian methods inherently embody *Occam's razor* (i.e., a proclivity towards simplicity) by exploiting prior knowledge that induces regularization without the need for ad-hoc penalty terms, thereby facilitating well-calibrated uncertainty estimates using training data alone.

We now explore the key properties of the likelihood-based Bayesian regression market. These properties are derived from the axioms that characterize the semivalue, all four of which are satisfied by the Shapley value. We first present the properties that we refer to as *universal*, those which are guaranteed to be satisfied under all circumstances.

Theorem 1 *Likelihood-based Bayesian regression markets of this kind yield the following universal market properties:*

1. *Symmetry—two features $x_{i,t}$ and $x_{j,t}$ with equal marginal contribution to any coalition receive the same attribution, that is, $\forall \mathcal{C} \in \mathcal{I}_{-c} \setminus \{i, j\} : v_t(\mathcal{C} \cup \mathcal{I}_c \cup \{i\}) \equiv v_t(\mathcal{C} \cup \mathcal{I}_c \cup \{j\}) \mapsto \phi_{i,t} \equiv \phi_{j,t},\ \forall (i, j) \in \mathcal{I}_{-c},\ i \neq j,\ \forall t$.*
2. *Linearity—for any two features $x_{i,t}$ and $x_{j,t}$, their joint contribution to a particular coalition of other features is equal to the sum of their marginal contribution, that is, $v_t(\mathcal{C} \cup \mathcal{I}_c \cup \{i\}) + v_t(\mathcal{C} \cup \mathcal{I}_c \cup \{j\}) = v_t(\mathcal{C} \cup \mathcal{I}_c \cup \{i, j\}),\ \forall \mathcal{C} \in \mathcal{I}_{-c} \setminus \{i, j\},\ \forall t$.*
3. *Budget balance—the payment of the central agent is equal to the sum of revenues received by the support agents, that is, $\pi_{c,t} \equiv \sum_{a \in \mathcal{A}_{-c}} \pi_{a,t},\ \forall t$.*

Proof *Omitted since each universal property follows directly from the semivalue axioms satisfied by the Shapley value.* ■

With symmetry, attributions are invariant to permutation of indices, equivalent to the anonymity property in Lambert et al. (2008), whilst linearity ensures that revenues remain consistent regardless of whether the features are offered individually or as a bundle, removing any incentive to strategically package features. Budget balance is a byproduct of the efficiency axiom, that total attribution allocated to all features should sum to the value of the grand coalition, that is, $v_t(\mathcal{I}) = \sum_{i \in \mathcal{I}_{-c}} \phi_{i,t},\ \forall t$. Accordingly, given the definitions in (8) and (11), it holds universally that the total sum of the revenues of each support agent equals the payment collected from the central agent.

In addition to these universally held market properties, our likelihood-based Bayesian regression market further obtains a collection of properties that we hereafter refer to as *asymptotic*, those which can only be guaranteed up to sampling uncertainty.

Theorem 2 *Likelihood-based Bayesian regression markets of this kind yield the following asymptotic market properties:*

1. *Individual rationality—support agents have a weak preference for participating in the market rather than not participating, that is, $\pi_{a,t} \geq 0, \forall a \in \mathcal{A}_{-c}, \ \forall t$.*
2. *Zero-element—a support agent that provides no feature, or only provides features with zero marginal contribution to all coalitions of other features should receive no payment, that is, $\forall C \in \mathcal{I}_{-c} : v_t(C \cup \mathcal{I}_c \cup \{i\}) \equiv v_t(C \cup \mathcal{I}_c), \ \forall i \in \mathcal{I}_a \mapsto \pi_a = 0, \ \forall t$.*
3. *Truthfulness—support agents receive their maximum potential payment when reporting their true data, that is, $v_t(C \cup \mathcal{I}_c; \mathbf{x}_{C \cup \mathcal{I}_c, t}) \geq v_t(C'; \mathbf{x}_{C \cup \mathcal{I}_c} + \boldsymbol{\eta}_t), \ \forall C \in \mathcal{I}_{-c}, \forall i \in C_{-b}, \ \forall t$, where $\boldsymbol{\eta}_t$ represents noise added to the original feature.*

Proof *Individual rationality follows directly from Assumption 1 and zero-elemet follows directly from the null-player axiom of semivalues satisfied by the Shapley value. For a proof of truthfulness, see Appendix A.* ■

In practice, only an in-sample estimate of the posterior moments are available. We assume that the specified hypothesis is such that as more data is observed, the posterior distribution converges to the Dirac measure around the maximum likelihood estimate of the parameter values almost surely, that is

$$D_{\mathrm{KL}}(p(\Theta_C | \mathcal{D}_{C,t} \| \delta(\Theta_C^*)) \xrightarrow{t} 0, \quad \forall C, \tag{13}$$

where $\delta(\cdot)$ is the probability density function of the Dirac delta distribution and Θ^* is the maximum likelihood estimate of the parameters.

Remark 5 *This assumption implies asymptotic consistency of well-specified models. Although in practice model misspecification is inevitable, concentration around the maximum likelihood estimate is sufficient to guarantee the properties in Theorem 2 hold up to sampling uncertainty.*

Given (13), individual rationality proceeds from Assumption 1, as given $\phi_{i,t} \geq 0, \ \forall i \in \mathcal{I}_{-c}, \ \forall t$, it follows from definitions (8) and (11) that payments can only be nonnegative in expectation. Similarly, the zero-element property, inherited from the null-player axiom, holds by design—if no feature is reported to the market then trivially no revenue is allocated, and if instead the true coefficient associated with a feature is zero, so too would be the associated revenue. Truthfulness ensures incentive compatibility, such that there is an incentive for support agents to report their true feature data. We assume that if a support agent is to provide an untruthful report of their data, they do so through the addition of centred noise with finite variance. Noise added to a particular feature is uncorrelated with noise added to any other, and conditionally independent of the target given the feature.

Corollary 1 *Following Assumptions 2 and 3 the revenue of each of the support agents exhibits a unique maximum when each reports their true feature data.*

Proof *See Appendix A.* ■

Remark 6 *Even in expectation, Theorem 2 can only be guaranteed in-sample, and may not generalize to the out-of-sample market stage. This issue pertains to the rich field of generalization in machine learning, for which bounds can typically only be attained under strict assumptions about the data generating processes (Mohri et al., 2018). We leave a thorough examination of the generalization characteristics of these market properties to future work.*

Lastly, we acknowledge properties of similar markets proposed in related works. For instance, Lambert et al. (2008) introduce *normality* in the context of wagering mechanisms, which would hold universally in our setup if features are independent. The same authors also introduce *sybilproofness* and *monotonicity*, which are not deemed relevant to our setup. Another property frequently discussed in literature is that of *robustness to replication*, which states that no support agent should be able to increase their revenue by replicating their data. Whilst several mechanism designs have been proposed to satisfy this property (e.g., Agarwal et al. 2019, Ohrimenko et al. 2019, Han et al. 2022a), its satisfaction generally comes at a cost, for instance Agarwal et al. (2019) sacrifice budget balance. Therefore, data replication remains an open challenge; we leave exploration of this topic in relation to our setup as future work.

3.2 Information-based Designs

Since the properties in Theorem 2 can only be guaranteed in expectation, it is likely that they will be violated in a single-shot of the market. Whilst violation of these properties would have no impact to the central agent with respect to predictive performance, support agents would be exposed to considerable financial risks, especially when a limited number of observations, as sub-optimal estimates of the parameters could distort allocations. This issue would be exacerbated out-of-sample, for which the in-sample estimate of the posterior may be less efficient.

To alleviate these risks, we explore alternative methods for valuing coalitions of features. Our approach is inspired by recent works concerned with multi-class classification—model outputs are instead discrete probability distributions. In this setting, Covert et al. (2020) demonstrate that models can be compared using relative mutual information. However this requires explicit computation of the joint distribution over the observed data, which may be intractable when dealing with continuous distributions, demanding expensive approximation (Kraskov et al., 2004). Rather than focusing on predictive performance, in the work of Agussurja et al. (2022) multiple data owners seek to perform joint inference of a set of parameters. Each subset of features is valued using the information gain on the *true* parameters, measured by the KL divergence of the posterior from a common prior. This is not immediately applicable to our setup, as we instead compensate support agents based on their contribution to overall predictive performance. Instead, we can make use of the information gain by considering the predictive densities, which encapsulate the value of the features in relation to predictive performance. In the following, we derive two methods for employing the KL divergence in our setup, demonstrating the implications on the market properties for each.

3.2.1 Marginal Contribution

We can express the marginal contribution of a feature to a coalition as the additional information that it provides, that is, the KL divergence between the predictive distribution *with* and *without* the particular feature, such that

$$m_{i,t}(C) = \mathbb{E}[D_{\mathrm{KL}}(p(y_t|\mathbf{x}_{C\cup\{i\},t})\|p(y_t|\mathbf{x}_{C,t}))], \quad \forall i, \ \forall C. \tag{14}$$

with $\mathcal{M}^{\mathrm{BLR}}_{\mathrm{KL}-m}$ denoting the corresponding market design. We remove the conventional characteristic function altogether and replace it with a function that maps the predictive density of both coalitions to a real-valued scalar. With Assumption 3, we can express the KL divergence as the expected value of the logarithm of the Radon-Nikodym derivative, since any two univariate Gaussian distributions satisfy absolute continuity.

Corollary 2 *The definition in (14) yields revenue allocations asymptotically equivalent to those obtained using the likelihood-based design.*

Proof *See Appendix B.* ■

Despite this asymptotic equivalence, the impact of using the KL divergence as described in (14) becomes apparent when the number of observations is limited; the resultant revenue allocations will be less volatile, reducing risk exposure of the support agents. This results from the fact that the KL divergence accounts only for the relative entropy, considering the overall information held within the distributions rather than the specific observations of the target signal, which can be distorted by outliers.

Theorem 3 *Replacing the marginal contribution with the definition in (14) alters the market properties in Theorems 1 and 2 as follows: individual rationality becomes a universally held property at the expense of budget balance violation, whilst the remaining properties are unchanged.*

Proof *Individual rationality follows directly from Gibbs' inequality. For a proof of the loss of budget balance, see Appendix C.* ■

Although using the KL divergence in such a way yields universal individual rationality by design, reducing the definition of marginal contribution to a single inseparable expression removes the telescoping sum structure of the original Shapley value. This leads to a violation of the efficiency axiom and hence budget balance. For brevity, we omit a proof for the remaining properties by virtue of similarity to Theorems 1 and 2. Although budget balance is violated, the universal satisfaction of individual rationality theoretically removes the most severe financial risks exhibited by the support agents, as they are guaranteed a nonnegative revenue. We see this as a similar trade-off exhibited in Agarwal et al. (2019) in pursuit of robustness to replication, that is, the addition of financial security is simply paid for by the market.

3.2.2 Characteristic Function

In some cases, violating budget balance may be impractical. If so, the KL divergence can instead be used in a manner that more closely resembles that presented in Agussurja et al. (2022), however instead of considering the posterior distribution, we set the common prior to be the predictive distribution of the central agent, such that

$$v_t(C) = \mathbb{E}[D_{\mathrm{KL}}(p(y_t|\mathbf{x}_{C,t})\|p(y_t|\mathbf{x}_{\mathcal{I}_c,t}))], \quad \forall i,\ \forall C. \tag{15}$$

Now we have instead only modified the characteristic function, with $\mathcal{M}^{\mathrm{BLR}}_{\mathrm{KL}-v}$ denoting the corresponding market design. The marginal contribution is still $m_{i,t}(\cdot) = v_t(\cdot) - v_t(\cdot \cup \{i\})$.

Theorem 4 *Valuing a coalition as in (15) yields revenue allocations asymptotically equivalent to those obtained using the likelihood-based design.*

Proof *See Appendix D* ■

Corollary 3 *Replacing the characteristic function of likelihood-based design with the definition (15) preserves the market properties in Theorems 1 and 2.*

Proof *Omitted due to similarity to that for these theorems.* ■

Since the telescoping sum structure of the Shapley value remains, budget balance is reinstated as a universal property. However, individually rationality reduces back to an asymptotic property. This follows from the fact that the marginal contribution now involves subtraction of expected KL divergences, for which Gibbs' inequality no longer applies. Nevertheless, this design should still provide us with less volatile allocations relative to $\mathcal{M}_{\text{NLL}}^{\text{MLE}}$ when a limited number of observations are available. Hence, one should still expect reductions in risk exposure for the support agents, the extent to which will be studied through a series of simulation studies in Section 4.

3.3 Summary of Market Designs

Apart from the transition to Bayesian regression analyses, the proposed market designs differ solely in their marginal contribution formulations. Both the payment of the central agent *and* the revenue allocations are impacted by the difference in the functional form of $m_{i,t}$, the extent of which will also be studied in Section 4. To end this section, we provide a summary of the different formulations in Table 1.

Market Design	Formulation of marginal contribution: $m_{i,t}(C)$, $\forall t$, $\forall i$, $\forall C$
(i) $\mathcal{M}_{\text{NLL}}^{\text{MLE}}$	$\mathbb{E}[-\log(p(y_t \vert \mathbf{x}_{C,t}, \Theta^*_C))] - \mathbb{E}[-\log(p(y_t \vert \mathbf{x}_{C\cup\{i\},t}, \Theta^*_{C\cup\{i\}}))]$
(ii) $\mathcal{M}_{\text{NLL}}^{\text{BLR}}$	$\mathbb{E}[-\log(p(y_t \vert \mathbf{x}_{C,t}))] - \mathbb{E}[-\log(p(y_t \vert \mathbf{x}_{C\cup\{i\},t}))]$
(iii) $\mathcal{M}_{\text{KL}-m}^{\text{BLR}}$	$\mathbb{E}[D_{\text{KL}}(p(y_t \vert \mathbf{x}_{C\cup\{i\},t}) \Vert p(y_t \vert \mathbf{x}_{C,t}))]$
(iv) $\mathcal{M}_{\text{KL}-v}^{\text{BLR}}$	$\mathbb{E}[D_{\text{KL}}(p(y_t \vert \mathbf{x}_{C,t}) \Vert p(y_t \vert \mathbf{x}_{\mathcal{I}_c,t}))] - \mathbb{E}[D_{\text{KL}}(p(y_t \vert \mathbf{x}_{C\cup\{i\},t}) \Vert p(y_t \vert \mathbf{x}_{\mathcal{I}_c,t}))]$

Table 1: Marginal contribution formulation for each of the market designs introduced in Section 3.

4 Simulation Studies

To illustrate our findings, we now present a collection of scenarios and simulation-based case studies.[4] To emphasize the versatility of our proposed Bayesian regression market, we devote particular attention to four distinct setups, each portraying an additional layer of complexity to emulate real-world intricacies. We note that these setups provide simplified representations the real world for the purpose of demonstration. We explore compounding effects of likelihood misspecification, specifically with respect to both the interpolated function and the intrinsic noise in the target signal.

In each of the simulation-based case studies, the central agent seeks to model a target variable Y_t using their own feature $x_{1,t}$ and the relevant features available in the market, each owned by a unique support agent, namely $x_{2,t}$ and $x_{3,t}$. The likelihood is an independent Gaussian stochastic process with finite precision ξ_{Y_t}. The linear interpolant for the grand coalition is $f(\mathbf{x}_t, \mathbf{w}) = w_0 + w_1 x_{1,t} + w_2 x_{2,t} + w_3 x_{3,t}$, $\forall t$. The various setups differ solely in the model of the likelihood as follows: (i) *Baseline*—the likelihood is well specified with respect to the *true* data generating process, given by $p(y_t|\mathbf{x}_t, \mathbf{w}) = \mathcal{N}(f(\mathbf{x}_t, \mathbf{w}), \xi_{Y_t})$; (ii) *Interpolant*—the interpolant is misspecified such that we write the *true* mean of the likelihood as $f(\mathbf{x}_t, \mathbf{w}) = \mathbf{w}^\top (\mathbf{x}_t \odot \mathbf{x}_t)$, $\forall t$, where $\odot$ denotes

[4] Our code is publicly available at: https://github.com/tdfalc/regression-markets

the Hadamard product; (iii) *Noise*—further to the misspecified interpolant, the Gaussian noise assumption is incorrect, with the *true* process given by a Student's t-distribution with two degrees of freedom; and (iv) *Heteroskedasticity*—the non-Gaussian noise is heteroskedastic, such that at each time step it is multiplied by $x^2_{2,t}$.

4.1 In-sample Market

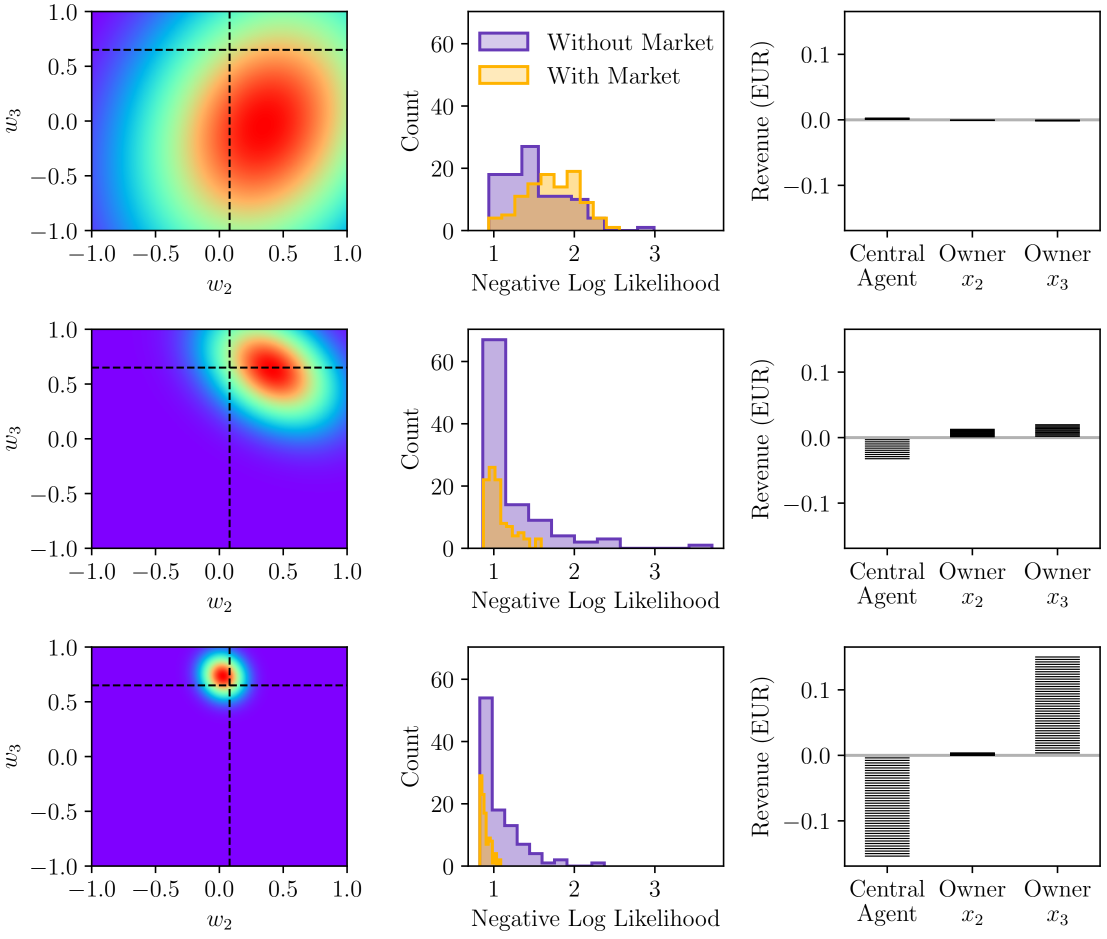
We first demonstrate the link between online Bayesian inference and the eventual revenue allocation, using the in-sample stage of the $\mathcal{M}^{\text{BLR}}_{\text{MLE}}$ market design as case study. We emulate batch inference (i.e., $\tau = 1$) for simplicity and consider only the *Baseline* setup. We assume the *true* coefficients to be $\mathbf{w} = [-0.11, 0.31, 0.08, 0.65]^\top$, and the noise precision to be constant for all time steps, treated as a hyperparameter with $\xi_{Y_t} = 1.23$, $\forall t$. We further set the valuation of the central agent to $\lambda = 0.01$ EUR per time step and per unit improvement in ℓ. We run the market for increasing sample sizes, specifically 4, 10 and 40, recording posterior moments, predictive performance and market revenue allocations for each. The results are shown in Figure 3.

In Figure 3a, we see that as one would expect, increasing the number of observations improves the estimation of the posterior, eventually centering around the *true* coefficient values. In Figure 3b, we present the NLL distribution for the in-sample observations. As the sample size increases, the improved posterior facilitates better capturing of the additional information provided by the features of the support agents, resulting in considerably enhanced predictive performance. The central agent indeed must pay for such improvements, highlighted by the additional revenue earned by the support agents, presented in Figure 3c. In the case of only 4 samples, we see that the predictive performance in fact decreases with the additional features, yielding small but negative revenues for the support agents. This emphasizes the importance of Assumption 1, as a prior feature selection process could remove these features so that individual rationality is preserved.

4.2 Uncertainty Quantification

Next we illustrate our four considered setups, highlighting the benefit to the central agent of merely facilitating Bayesian regression analyses. We set the *true* parameters to $\mathbf{w} = [-0.1, 0.3, 0.8, -0.4]^\top$ and $\xi_{Y_t} = 0.5$, $\forall t$. We again emulate batch inference and run a Monte-Carlo simulation whereby we clear the market 10^3 times for several different sample sizes and record the expected NLL for 10^3 out-of-sample observations for each. This is carried out using both maximum likelihood estimation and Bayesian regression analyses.

Figure 4 shows the empirical average of the percentage improvement in the objective value for the Bayesian regression model. Observe that the improvement is most significant across all setups when the sample size is relatively small, as the additional piece of uncertainty in the parameter estimates plays a greater role in the predictive distribution, increasing the predictive likelihood. Then, as the sample size increases, the parameter estimates converge in accordance with (13). Furthermore, as the additional layers of complexity are introduced, the benefit of incorporating parameter uncertainty increases considerably. These improvements attained by converting to a Bayesian framework indicate a better calibration of uncertainty, enriching the information used by the central agent for risk-informed decision-making downstream, compared with the deterministic proposals of Agarwal et al. (2019) and Pinson et al. (2022a).



(a) Posterior Updates (b) Predictive Performance (c) Revenue

Figure 3: In-sample market with increasing batch size. The dashed lines in (a) highlight the *true* coefficient values. The histogram in (b) shows the in-sample NLL distribution. The horizontal bars in (c) are the cumulative revenues given the value of each datapoint provided.

4.3 Convergence Analysis

Now we present an empirical study of the in-sample asymptotic convergence for our various market designs. Let the *true* coefficeints and noise precision be given by $\mathbf{w} = [-0.1, 0.8, 0.7, -0.9]^\top$ and $\xi_{Y_t} = 1.0$, $\forall t$, respectively, focusing here solely on the *Baseline* setup, since asymptotic convergence is irrespective of the *true* data generating processes, but rather the set of modelling assumptions. A similar Monte-Carlo simulation is performed, recording the in-sample Shapley values for each run, the results of which are presented in Figure 5 and Figure 6.

Looking first at Figure 5, we see that with a small sample size, the frequentist market design assigns a larger contribution to the features compared with those using Bayesian regression, however these values indeed converge asymptotically in align with the theory. This discrepancy is likely due to the greater reduction in the in-sample objective provided by the maximum likelihood estimate,

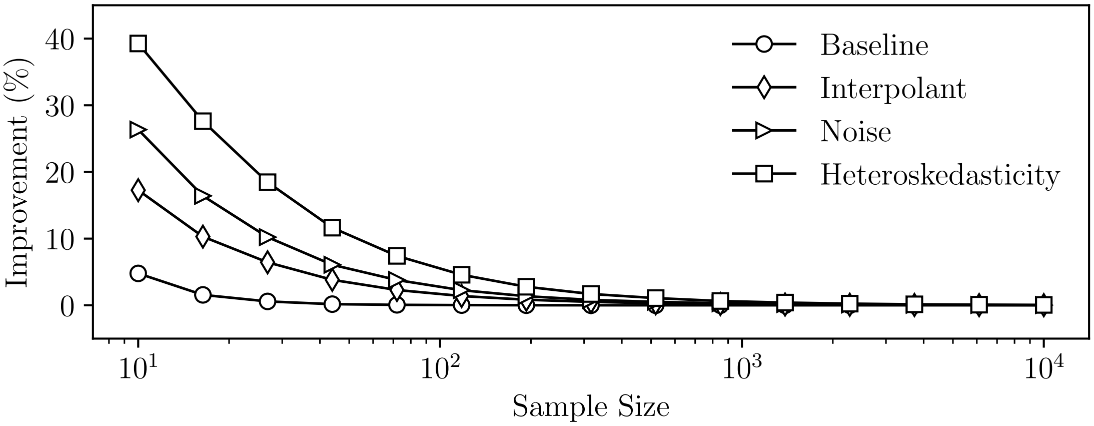


Figure 4: Empirical average of the percentage improvement in the NLL ratio for BLR relative to MLE, plotted as a function of sample size.

which is of course prone to overfitting. In Figure 6, we can observe that the $\mathcal{M}^{\text{BLR}}_{\text{KL}-m}$ market renders a surplus in revenue when the sample size is small. This demonstrates the trade-off incurred by virtue of the now universally held individual rationality property—budget balance is no longer guaranteed, even during the in-sample stage. This problem indeed resolves with increasing number of observations as the Shapley values converge.

4.4 Risk Exposure

We now turn our attention to the finances of the support agents, which we assess by computing both the expected value of the revenue, $\int \pi_{a,t} p(\pi_{a,t})\, d\pi_{a,t}$, $\forall t$, and the expected shortfall (i.e., conditional value at risk), $-1/\alpha \int_{\pi_{a,t} \leq q_\alpha(\pi_{a,t})} \pi_{a,t} p(\pi_{a,t})\, d\pi_{a,t}$, $\forall t$, for all $a \in \mathcal{A}_{-c}$, where $q_\alpha(\cdot)$ is the quantile with nominal level $\alpha \in (0, 1)$. We present empirical estimations of these financial metrics for a case study where we again clear the market for a new sample of data 10^3 times and record the revenue of each support agent, with the *true* coefficients set to $\mathbf{w} = [0.1, -0.5, 0.0, 0.7]^\top$, with noise precision $\xi_{Y_t} = 0.67$, $\forall t$. We additionally set $\lambda = 0.03$ EUR per time step and per unit improvement in ℓ for the both in-sample and out-of-sample stages. We use a simple sub-sampling method to derive the corresponding two-sided confidence intervals of both the expected value and expected shortfall of the revenue with a 95% confidence level. We run this simulation for each market design, as well as each misspecification setup, with 10^3 in-sample and out-of-sample observations.

In Figure 7, we plot the revenue of the support agent who owns $x_{2,t}$. Oobserve that the expected value of the revenue is relatively consistent across all market designs for each setup. However, for each additional layer of complexity, the expected shortfall is positive for the $\mathcal{M}^{\text{BLR}}_{\text{NLL}}$ market, increasing by almost two orders of magnitude in the latter setups. For the market designs that use the KL divergence, the expected shortfall remains somewhat constant around zero, highlighting the sizeable reductions in risk exposure compared to using NLL. As before, the $\mathcal{M}^{\text{MLE}}_{\text{NLL}}$ market design seemingly performs better than its Bayesian counterpart in-sample, however we see this indeed does not generalize out-of-sample.

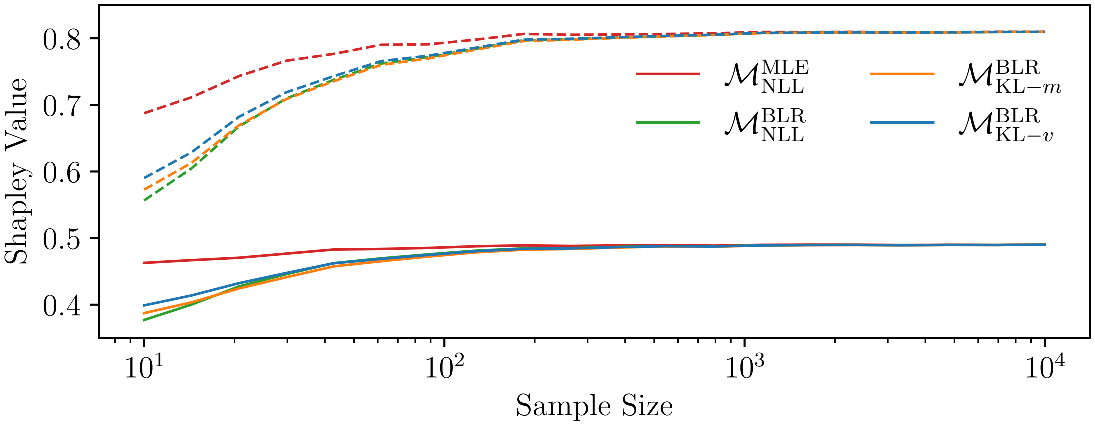


Figure 5: Empirical average of the expected Shapley values for each market design, plotted as a function of sample size. Dashed and solid lines correspond to features $x_{2,t}$ and $x_{3,t}$, respectively.

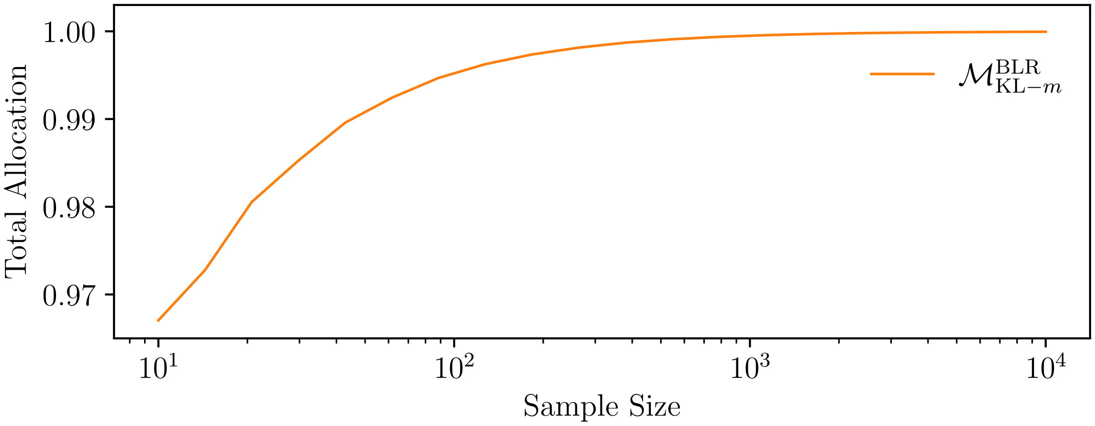


Figure 6: Empirical average of the expected total allocation for the $\mathcal{M}_{\text{KL}-m}^{\text{BLR}}$ market design. Only this design is plotted since, in the remaining markets, budget balance is a universal property by design hence the total allocation is always 1.

For this small sample size, parameter estimates are more likely to be sub-optimal, and hence the predictive likelihood is more volatile. In consequence, even the expected value of the out-of-sample revenue becomes increasingly negative with additional layers of complexity for the likelihood-based market designs, meaning that the individual rationality property is violated even in expectation. The out-of-sample revenues in the $\mathcal{M}_{\text{NLL}}^{\text{MLE}}$ market become worse than for $\mathcal{M}_{\text{NLL}}^{\text{BLR}}$, demonstrating that the superior performance of this market design in-sample was indeed due to overfitting. In contrast, the expected value of the revenue is relatively consistent with those in-sample for both KL divergence-based markets.

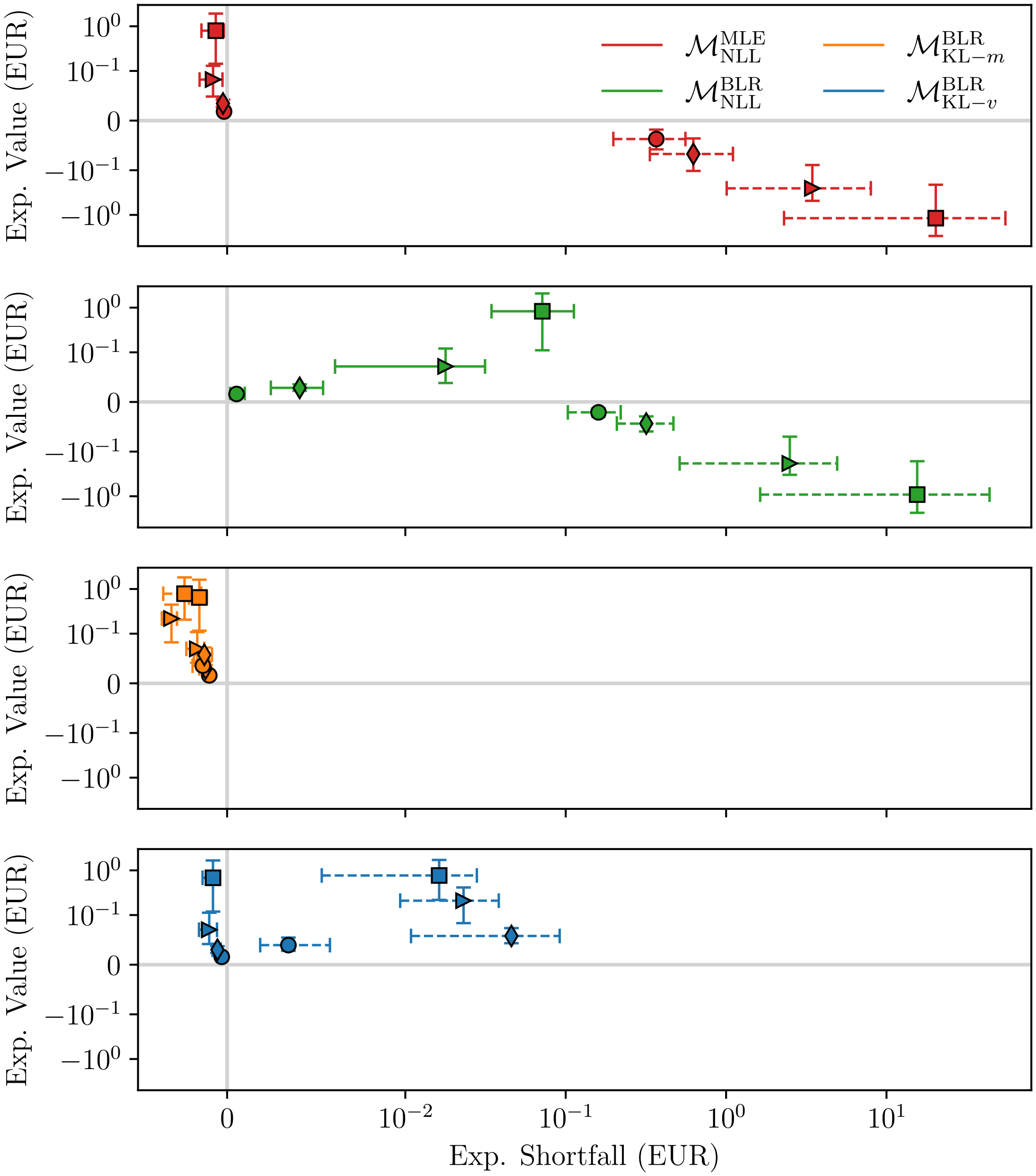


Figure 7: Two-sided confidence intervals with a 95% confidence level for both the expected value and expected shortfall of the revenue received by the owner of $x_{2,t}$, with quantile parameter $\alpha = 0.05$, for each setup, namely *Baseline* (○), *Interpolant* (◇), *Noise* (▷) and *Heteroskedasticity* (□). Both in-sample (solid) and out-of-sample (dashed) metrics are plotted.

The expected shortfall for the likelihood-based markets increased by several orders of magnitude in the out-of-sample stage. Interestingly, one can now observe the consequence of re-instating budget balance by modifying the characteristic function instead of the marginal contribution. Specif-

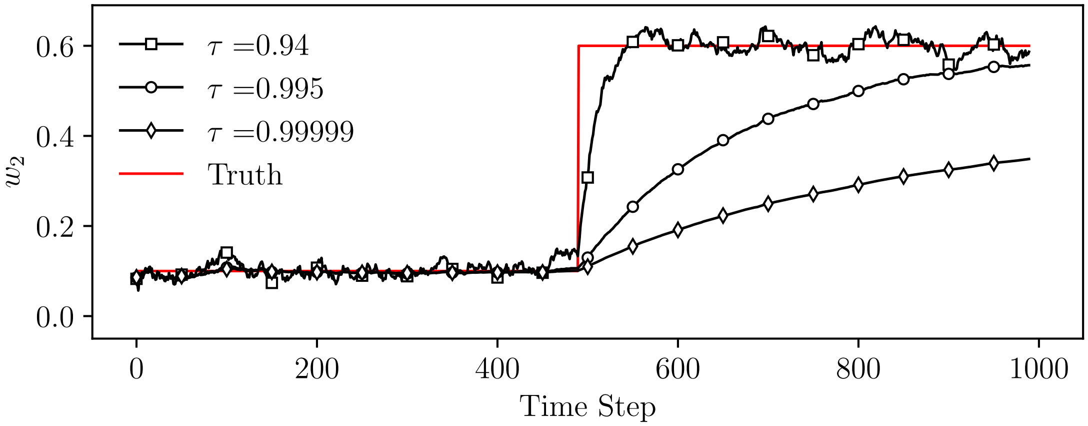


Figure 8: Temporal evolution of the empirical average of the estimated value for w_2. The estimates for the remaining parameters are omitted for clarity.

ically, whilst for the $\mathcal{M}^{\text{BLR}}_{\text{KL}-m}$ market design individual rationality is held universally, the expected shortfall in $\mathcal{M}^{\text{BLR}}_{\text{KL}-v}$ has drifted positive. Nevertheless, the risks are generally considerably less compared with $\mathcal{M}^{\text{BLR}}_{\text{NLL}}$, suggesting that there is still merit to this approach. These results demonstrate that valuing features based on the the information provided is able mitigate these risks entirely, despite being asymptotically equivalent to the approach proposed in Agarwal et al. (2019) and Pinson et al. (2022a).

4.5 Nonstationary Processes

So far we have assumed only batch inference (i.e., $\tau = 1$), however in Section 2.3 we showed theoretically that this is merely a specification of the more general online Bayesian inference problem, which facilitates time-varying posterior moments. For the final simulation-based case study, let us consider a nonstationary data generating process, wherein the parameters initially take on the values $\mathbf{w} = [0.0, -0.2, 0.1, 0.3]^\top$, with noise precision $\xi_{Y_t} = 0.98$, $\forall t$.

For simplicity, we only let the coefficient associated with $x_{2,t}$ vary with time, with the remaining kept constant. To illustrate the effect of likelihood flattening, we consider cases where w_2 exhibits a discontinuity, representing a more complex processes to capture with respect to its stationary analogue. We carry out a Monte-Carlo simulation whereby we record the empirical average of the parameter estimates at each time step with various values for τ, the results of which are presented in Figure 8. Of course, for the previous time-invariant cases, there would be no advantage of using likelihood flattening since the coefficients are stationary. For the more complex cases, as $\tau \mapsto 1$, our posterior beliefs decay more gradually, but as τ is reduced, we are able to better track the coefficient values, albeit with increased variance due to the fact that more weight is given to the flat prior.

We run a Monte-Carlo simulation whereby we fix $\tau = 0.94$ in order to better track the coefficient, albeit at the expense of increased variance. We re-run the entire market clearing procedure 10^3 times, each time tracking the temporal evolution of market revenue over 10^3 time steps. We set $\lambda = 0.95$ EUR per time step and per unit improvement in ℓ for both stages. As before, we carry out

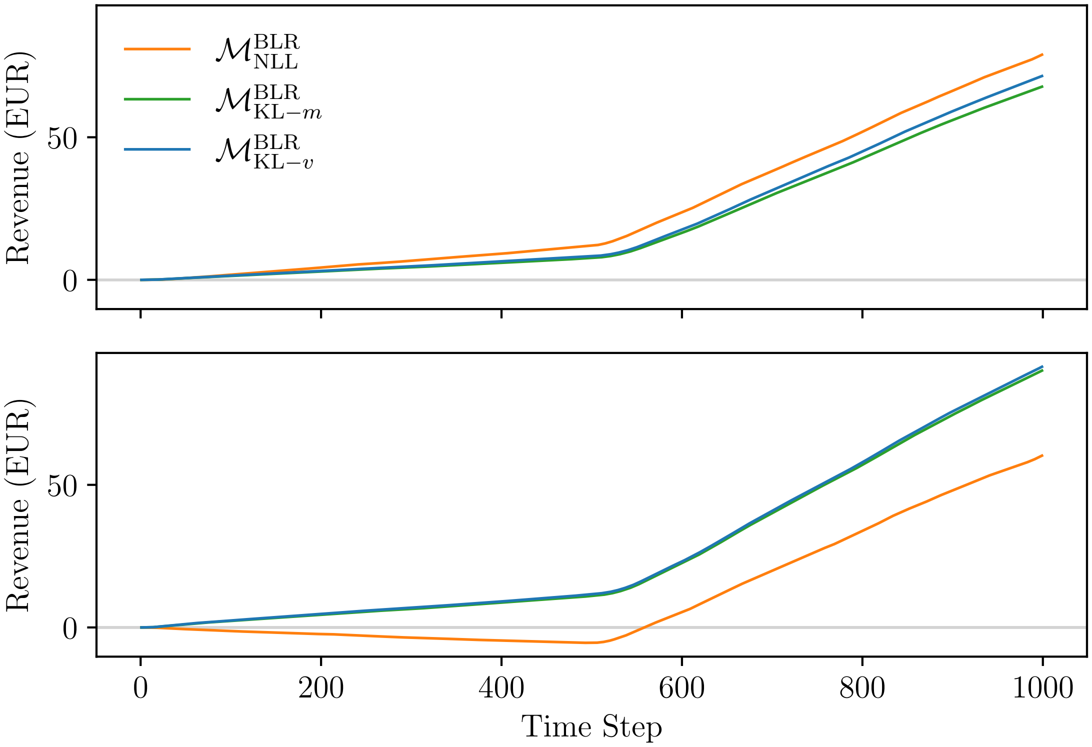


Figure 9: Cumulative empirical averages of the expected value of the revenue, with quantile parameter $\alpha = 0.05$, for the first support agent considering the *Baseline* setup.

this simulation for each of the proposed Bayesian regression market designs, considering only the *Baseline* setup, the results of which are shown in Figure 9.

Given the use of likelihood flattening, each of the market designs are able to capture the step-change in the *true* coefficient value. However, the extent of flattening required reduces the effective window size of observations, emulating a consistently small sample size even as more observations arrive. As a result, the likelhood-based market exhibits poor generalization. In fact, even though the true coefficient is $w_2 = 0.1$ before the step change, the expected value of the revenue is less than 0, resulting in a negative cumulative revenue in the first half of the simulation, and hence the overall revenue earned by the agent is considerably less in this market. The greater revenue in-sample is likely again due to overfitting. In contrast, we see that both the expected value and expected shortfall of the revenue in both $\mathcal{M}^{\text{BLR}}_{\text{KL}-m}$ and $\mathcal{M}^{\text{BLR}}_{\text{KL}-v}$ markets remain relatively consistent to those observed in-sample. This highlights that valuing features using information gain has the propensity to mitigate financial risks entirely, even out-of-sample where the majority of risk exposure manifests, compared with likelihood-based market designs in Agarwal et al. (2019) and Pinson et al. (2022a) which subject agents to sizeable losses.

Note that here we have investigated concept drift with respect to the model parameters. Specifically, we show that our fully online setup is as robust as possible to such drift with the parameter estimates being updated at the closest time step to the next, rather than training the model once and making predictions for several time steps (e.g., in batch or rolling window inference). This indeed has a greater computational cost, but exploring this trade-off is outwith the scope of this work. In addition, what we have not explored is concept drift with respect to the model hyperparameters (e.g., the basis function vector, the noise process, etc.). This of course is a key problem in real forecasting applications, however in our setup it is the central agent who has discretion over these hyperparameters. An interesting direction for future work would be to entrust the market operator also with such model selection, in which case it would be up to the market to account for hyperparameter drift as well.

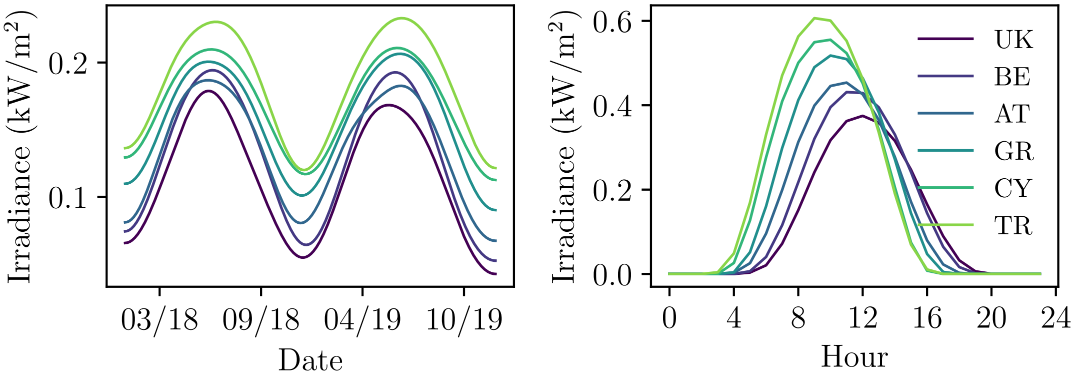
5 Real-world Application

We complete our experimental analysis by verifying the applicability of our proposal to real-world applications. We make use of an open source dataset, namely the *Pan-European Climate Database*, as detailed in Koivisto and Leon (2022). This dataset consists of hourly average solar irradiance values by country in Europe, obtained by simulating the output from south-facing solar photovoltaic (PV) modules across several intra-country regions. Although this data is not exactly *real*, it effectively captures the spatio-temporal aspects of solar irradiance across the continent, with the benefit of not being contaminated with spurious data points, as can often be the case with real-world datasets.

Suppose that the electricity system operator in each country seeks to forecast its own country's average generation from solar PV modules, with a view to estimate electricity demand and determine balancing resource requirements. For illustration, we consider six countries, namely United Kingdom (UK), Belgium (FR), Austria (AT), Greece (GR), Cyprus (CY) and Turkey (TR), each of which is assumed to enter the regression market to enhance their respective forecasts. For simplicity, we focus on a 1-hour forecasting horizon (i.e., nowcasting) using only linear basis functions, though both longer latency periods and more complex models could be considered.

We extract data that spans a two-year period from the start of 2018 to the end of 2019, with an hourly resolution. Suppose that each of the six countries takes turn in assuming the role of the central agent in parallel transactions. We use a simple *Auto-Regressive with eXogenous input* model with a maximum of one lag for each feature. For solar energy, forecasting with lags simultaneously captures temporal correlations at particular locations and any indirect spatial correlations between neighboring locations, resulting from the natural development of cloud coverage and the rotation of the sun. We present the rolling average of the raw irradiance values in Figure 10a, which highlights the seasonality of generation, peaking during the summer months as expected. Similarly, by plotting the hourly average irradiance in Figure 10b, one can observe the spatial correlations such that at any given time, the actual generation in the more Easterly countries could be indicative of what is to come in Western Europe later in the day.

For each forecast, we model the likelihood as an independent Gaussian stochastic process with finite precision, similar to the setup in Section 4. We consider an online setting such that over the entire two-year period, at each time step (i.e., one hour interval), when a new observation of the target signal is collected, the forecast issued at the previous time step is used for out-of-sample market clearing, whilst at the same time, the posterior is updated and the in-sample market is cleared, and a forecast for the next time step is subsequently made. We set $\tau = 0.998$ and assume the valuation of each central agent to be $\lambda = 50$ EUR and $\lambda = 150$ EUR per time step and per unit improvement in ℓ for the in-sample and out-of-sample stages, respectively, to reflect costs of balancing. With each



(a) Rolling Average

(b) Hourly Average

Figure 10: The rolling average and hourly average solar irradiance observed in each country during the two-year time period of 2018–2019.

country set as the central agent, we record the predictive performance and cumulative revenues of the remaining countries across both stages over the entire two-year span using the $\mathcal{M}^{\mathrm{BLR}}_{\mathrm{KL}-\nu}$ market design.

Let us first consider the improvements in predictive performance exhibited by each of the countries when assuming the role of the central agent. We present the average quarterly results in Table 2. In general, we observe a seasonality in the loss equivalent to that of the irradiance itself, such that smaller enhancements in predictive performance are exhibited during the end two quarters, since there is less potential for improving predictive performance when irradiance is low. Both United Kingdom and Greece receive the greatest improvements, with Cyprus and Turkey the smallest, the latter of which is likely due to the fact that these countries are further East, thus less able to exploit the spatial correlations depicted in Figure 10b. We also note that the distribution of performance improvements amongst the countries is fairly similar between the in-sample and out-of-sample stages, which suggests any nonstationarities, as well as the time-varying objective estimates, are smooth, and hence the in-sample posterior is a relatively efficient estimator for use out-of-sample in the next time step.

In Figure 11, we present the smoothed evolution of the revenues across both the in-sample and out-of-sample market stages. As with the loss estimates, the allocation is by no means constant with time, such that the revenues of each agent are typically lower over the winter months and increase throughout the rest of the year. The value of each observation therefore also reflects the seasonality observed in the generation from solar PV modules. We also see the spatio-temporal dynamics of solar irradiance, as countries to the East of the central agent, particularly those nearby or with high nominal generation, contribute most to the uplift. The revenues received by the remaining countries when either Cyprus or Turkey assume the role of the central agent are relatively small, in accordance with the results in Table 2. Lastly, we note that the revenues earned by some of the countries over the entire two-year period are substantial, for instance with Greece as the central agent, the system operator in Cyprus earns approximately 1.2×10^6 EUR, representing an average unit value of around 70 EUR per observation shared.

Country	In-sample				Out-of-sample			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
UK	0.40	2.24	2.24	0.37	0.39	2.32	2.45	0.36
BE	0.34	1.58	1.59	0.51	0.33	1.60	1.61	0.50
AT	0.66	1.77	1.47	0.72	0.65	1.81	1.49	0.72
GR	0.73	2.11	2.40	0.82	0.74	2.15	2.44	0.81
CY	0.44	1.05	1.20	0.56	0.43	1.05	1.21	0.55
TR	0.43	1.00	1.35	0.65	0.42	1.00	1.36	0.64

Table 2: Fractional improvement in the NLL ratio as a result of participating in the regression market, averaged over each calendar quarter, for both in-sample and out-of-sample market stages.

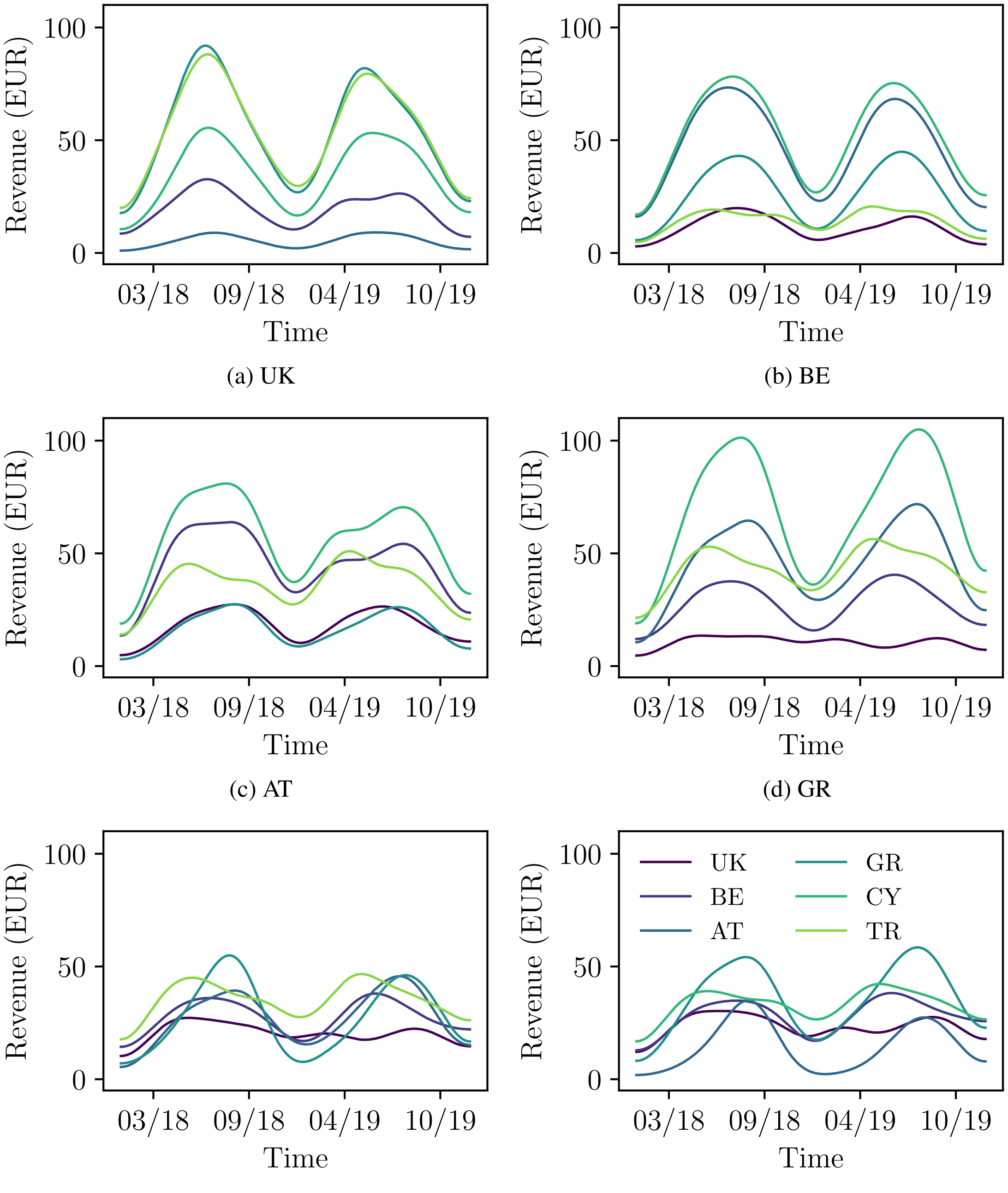
6 Conclusions

Firms that employ predictive analytics (e.g., machine learning) often lack access to adequate sources of data. Whilst sharing data amongst others could bring potential advantages, many firms remain hesitant to do so, predominantly due to privacy concerns and the fear of losing a competitive edge, rather than practical complexities involved in establishing data sharing pipelines. Analytics markets, or in our case *regression markets*, offer a possible solution to this, wherein data is commoditized with respect to the particular analytics task at hand, providing incentives for information exchange through remuneration.

In this paper, we proposed a mechanism design for a regression market that facilitates a generalized approach to forecasting, one based on Bayesian regression analyses. As a result, we provide the buyer with richer and more nuanced information about future outcomes, offering better calibration of uncertainty to be used for risk-informed decision-making downstream. We first introduced what we posed as the Bayesian analogue of recent frequentist-based proposals, but showed that this market design, akin to those in current literature, exposes the buyer to considerable financial risks, especially when a limited number of observations are available or when the data generating processes are nonstationary. In these settings, sub-optimal estimates of the posterior distribution led to sizeable expected losses, especially during the out-of-sample market stage, for which the in-sample estimates of the posterior moments are less efficient.

To mitigate these risks, we posed to re-formulate the value of a feature in terms of the information gain it provides. In particular, we derived two alternative definitions of the marginal contribution of a feature towards a set of other features using the Kullback–Leibler divergence, the first of which could guarentee individual rationality universally (i.e., no support agents would be allocated negative revenue). However, there is of course no free lunch, as this was at the expense of budget balance. Nevertheless, we showed that in both cases using the KL divergence was able to provide more robust revenue allocations by alleviating the financial risks that the support agents were exposed to, even at the out-of-sample market stage.

Possible directions for immediate future work could include extending the concepts of our proposal to a broader class of machine learning models, such as (i) non-convex regression, which will have implications on market property guarantees; (ii) non-Gaussian hypotheses, which may require approximation bounds; and lastly (iii) alternative modelling paradigms, for instance, classification, unsupervised learning, or data-driven optimization problems in general.



(a) UK

(b) BE

(c) AT

(d) GR

(e) CY

(f) TR

Figure 11: Smoothed evolution of total revenue per time step made by each of the six countries to the remaining five whilst assuming the role of the central agent.

On a broader note, there are still many unanswered questions in relation to the complexities of treating data as a commodity. For instance, in practice, datasets cover different spatial and temporal horizons, and may become (un-)available to the market at different times. Accordingly, aggregating real-world datasets in an online fashion may not be straightforward and may require revision of fundamental concepts in online learning and mechanism design. Additionally, much of the current literature relies on the assumption that the valuation of the central agent is both linear and easily conceivable with respect to the loss function, which may not be true if the downstream decision-making process is complex or in the face of externalities, for instance, whether or not competing firms also get access to the data may affect the valuation. Support agents may also have reservations to share their information, for instance due to privacy concerns or conflicts of interest. This, as well as physical costs of collecting and storing data, may require a minimum revenue threshold to be established. Lastly, if firms that share data are indeed competitors in a downstream market, one may be interested in if, by providing better use of information, the analytics market is beneficial to social welfare, and if those that lose competitive advantage by sharing their information are adequately compensated.

Appendix A. Truthfulness Property (Theorem 2)

We provide a proof of the truthfulness property provided by Theorem 2, which describes the asymptotic market properties.

We model the target signal, $\{Y_t\}$, as a deviation from the deterministic linear interpolant in (2) under the following centred additive noise process

$$p(\mathcal{D}_{C,t}|\Theta_C) = \prod_{t' \leq t} \mathcal{N}(f(\mathbf{x}_{C,t}, \mathbf{w}_C), \xi_{Y_t}^{-1}), \quad \forall t. \tag{16}$$

Without loss of generality, consider $\xi_{Y_t} = \xi, \forall t$ to be a hyperparameter, such that the parameters to be inferred from data are only the regression coefficients (i.e., $\Theta_C = \{\mathbf{w}_C\}$). As the prior is assumed to be uninformative, we avoid imposing any specific assumptions or biases on the parameter estimate. Accordingly, we set the conjugate prior to be a zero-mean isotropic Gaussian distribution with infinitely broad variance, that is, $p(\mathbf{w}_C|\xi) = \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{I})$ with $\gamma \mapsto 0$ and $\mathbf{I} \in \mathbb{R}^{|C|\times|C|}$ is the identity matrix. Let $\mathbf{m}_{C,t} \in \mathbb{R}^{|C|}$ and $\mathbf{K}_{C,t} \in \mathbb{R}^{|C|\times|C|}$ denote the mean vector and the covariance matrix of the posterior at a particular time step t, respectively. As the posterior is Gaussian, we indeed know that its mode coincides with its mean, and since we can write the logarithm of the posterior as the sum of both the logarithm of the likelihood and the logarithm of the prior, the posterior mean reduces to the maximum likelihood estimate, given by

$$\hat{\mathbf{m}}_{C,t} = \underset{\mathbf{w}_C}{\text{argmax}} \ \log\left(p(\mathcal{D}_{C,t}|\mathbf{w}_C, \xi)\right) + \log\left(p(\mathbf{w}_C|\xi)\right), \quad \forall t, \tag{17a}$$

$$= \underset{\mathbf{w}_C}{\text{argmin}} \sum_{t' \leq t} (y_{t'} - f(\mathbf{x}_{C,t'}, \mathbf{w}_C))^2, \quad \forall t, \tag{17b}$$

as would the estimated noise precision if inferred from data. Now suppose the data for one or more of the features is reported untruthfully, such that for any particular time step t, the vector of features is, $\tilde{\mathbf{x}}_t = \mathbf{x}_t + \boldsymbol{\eta}_t$, where $p(\boldsymbol{\eta}_t) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ is diagonal, the elements of which corresponding to truthfully reported features are zero, such that only untruthful features are subject to noise. For brevity, and without loss of generality, we consider only the set of linear basis functions. Substituting the new feature vector into the linear interpolant, the optimization problem in (17b) is augmented to minimize the expected sum-of-squares error, where the expectation is taken over the random noise,

$$
\begin{aligned}
\hat{\mathbf{m}}_{C,t} &= \underset{\mathbf{w}_C}{\operatorname{argmin}}\ \mathbb{E}\left[\sum_{t'\leq t} \left(y_{t'} - f(\tilde{\mathbf{x}}_{C,t'}, \mathbf{w}_C)\right)^2\right], \quad \forall t, && \text{(18a)}\\
&= \underset{\mathbf{w}_C}{\operatorname{argmin}} \sum_{t'\leq t} \Big((y_{t'} - \mathbf{w}_C^\top \mathbf{x}_{C,t'})^2 - 2\mathbf{w}_C^\top \mathbb{E}\left[\boldsymbol{\eta}_{C,t'}\right](y_{t'} - \mathbf{w}_C^\top \mathbf{x}_{C,t'}) + \mathbf{w}_C^\top \mathbb{E}\left[\boldsymbol{\eta}_{C,t'}\boldsymbol{\eta}_{C,t'}^\top\right]\mathbf{w}_C \Big), \ \forall t, && \text{(18b)}\\
&= \underset{\mathbf{w}_C}{\operatorname{argmin}} \sum_{t'\leq t} \left(y_{t'} - \mathbf{w}_C^\top \mathbf{x}_{C,t}\right)^2 + \mathbf{w}_C^\top \boldsymbol{\Sigma} \mathbf{w}_C, \quad \forall t, && \text{(18c)}
\end{aligned}
$$

where (18c) is derived from (18b) by recalling that $\mathbb{E}[\boldsymbol{\eta}_t] = \mathbf{0}$ and hence $\mathbb{E}[\boldsymbol{\eta}_t \boldsymbol{\eta}_t^\top] = \boldsymbol{\Sigma}$. For the special case whereby equal noise is added to all features, such that $\boldsymbol{\Sigma} = \beta \mathbf{I}$, for some constant β, the second term in (18c) reduces to $\mathbf{w}_C^\top \boldsymbol{\Sigma} \mathbf{w}_C = \beta \|\mathbf{w}_C\|_2^2$, which results in a vector of coefficient analogous to that obtained using Ridge regression (Hoerl and Kennard, 1970), with shrinkage penalty $\beta = \gamma/\xi$. Moreover, since $\boldsymbol{\Sigma}$ is a diagonal matrix, agents are not able to behave spitefully by adding noise to their features in effort to reduce the payments of others. In general though, we observe that the likelihood is maximized as $\gamma \mapsto 0$ and any addition of noise will create an endogeneity bias and reduce the predictive likelihood and subsequently the revenues, given the definition in (11), thereby completing the proof.

■

Appendix B. Proof of Corollary 2

Lemma B.1 *The difference between the expected value of the quadratic loss of two maximum likelihood linear regression models, both with a Gaussian likelihood function, is equivalent to the expected squared difference in their interpolated functions.*

Proof [Lemma B.1] In a maximum likelihood linear regression setting with a Gaussian likelihood, the objective is proportional to the expected value of the quadratic loss which we can decompose as follows:

$$
\begin{aligned}
\mathbb{E}\left[(f_{C,t} - y_t)^2\right] &= var\left(f_{C,t} - y_t\right) + \mathbb{E}\left[f_{C,t} - y_t\right]^2, && \text{(19a)}\\
&= var\left(f_{C,t}\right) + var\left(y_t\right) - 2\, cov\left(f_{C,t}, y_t\right), && \text{(19b)}\\
&= var\left(y_t\right) - var\left(f_{C,t}\right). && \text{(19c)}
\end{aligned}
$$

where for brevity, we write $f_{C,t} = f(\mathbf{x}_{C,t}, \mathbf{w}_C)$, $\forall t$. We can write the expression in (19c) since the covariance between the prediction and the target is equal to the variance of the prediction itself since the estimators are unbiased. If we let $C_i = C \cup \{i\}$ denote the addition of the i-th feature index to a particular coalition for all $C \subseteq \mathcal{I}_{-c}$, we get that

$$
\begin{aligned}
\mathbb{E}\left[(f_{C,t} - y_t)^2\right] - \mathbb{E}\left[(f_{C_i,t} - y_t)^2\right] &= var\left(f_{C_i,t}\right) - var\left(f_{C,t}\right), && \text{(20a)}\\
&= \mathbb{E}\left[(f_{C_i,t})^2\right] - \mathbb{E}\left[f_{C_i,t}\right]^2 - \mathbb{E}\left[(f_{C,t})^2\right] + \mathbb{E}\left[f_{C,t}\right]^2, && \text{(20b)}\\
&= \mathbb{E}\left[(f_{C_i,t})^2\right] - \mathbb{E}\left[(f_{C,t})^2\right], && \text{(20c)}\\
&= \mathbb{E}\left[(f_{C_i,t})^2\right] + \mathbb{E}\left[(f_{C,t})^2\right] - 2\mathbb{E}\left[(f_{C,t})^2\right], && \text{(20d)}\\
&= \mathbb{E}\left[(f_{C_i,t})^2\right] + \mathbb{E}\left[(f_{C,t})^2\right] - 2\left(var\left(f_{C_i,t}\right) + \mathbb{E}\left[f_{C_i,t}\right]^2\right). && \text{(20e)}
\end{aligned}
$$

Note that we can ignore the covariance $cov\left(f_{C,t}, (f_{C_i,t} - f_{C,t})\right)$ since the prediction is not correlated with the residuals. Hence, the variance term can be re-written as follows:

$$\begin{aligned} var\left(f_{C_i,t}\right) &= var\left(f_{C,t} + (f_{C_i,t} - f_{C,t})\right), && \text{(21a)} \\ &= var\left(f_{C,t}\right) + var\left(f_{C_i,t} - f_{C,t}\right), && \text{(21b)} \\ &= 2var\left(f_{C,t}\right) + var\left(f_{C_i,t}\right) - 2cov\left(f_{C_i,t}, f_{C,t}\right), && \text{(21c)} \end{aligned}$$

and we hence get that $var\left(f_{C,t}\right) = cov\left(f_{C_i,t}, f_{C,t}\right)$. Given that upon standardization, $\mathbb{E}\left[f_{C,t}\right] = 0$ and $\mathbb{E}\left[f_{C_i,t}\right] = 0$, we can re-write the last term in (20e) as follows:

$$\begin{aligned} var\left(f_{C_i,t}\right) + \mathbb{E}\left[f_{C_i,t}\right]^2 &= cov\left(f_{C_i,t}, f_C\right) + \mathbb{E}\left[f_{C_i,t}\right]\mathbb{E}\left[f_{C,t}\right], && \text{(22a)} \\ &= \mathbb{E}\left[f_{C_i,t} f_{C,t}\right]. && \text{(22b)} \end{aligned}$$

Therefore, the difference in expected values of the quadratic loss reduces to the following, which completes the proof:

$$\begin{aligned} \mathbb{E}\left[(f_{C,t} - y_t)^2\right] - \mathbb{E}\left[(f_{C_i,t} - y_t)^2\right] &= \mathbb{E}\left[(f_{C_i,t})^2\right] + \mathbb{E}\left[(f_{C,t})^2\right] - 2\mathbb{E}\left[f_{C_i,t} f_{C,t}\right], && \text{(23a)} \\ &= \mathbb{E}\left[(f_{C_i,t} - f_{C,t})^2\right]. && \text{(23b)} \end{aligned}$$

■

Lemma B.2 *Under Assumption 3, the expected KL divergence between two predictive distributions is asymptotically equivalent to the expected difference in their predictive means.*

Proof [Lemma B.2]. Following the notation as in Appendix A, we can write the general expression for the posterior predictive distribution in (6) as

$$p(y_t \,|\, \mathbf{x}_{C,t}) = \mathcal{N}(f(\mathbf{x}_{C,t}, \mathbf{m}_{C,t}), \xi_{C,t}), \quad \forall t, \qquad \text{(24)}$$

where $\xi_{C,t}$ is the precision (i.e., inverse variance) comprising the finite precision of the intrinsic noise and the uncertainty the coefficients such that, $1/\xi_{C,t} = 1/\xi + \mathbf{x}_{C,t}^\top \mathbf{S}_{C,t} \mathbf{x}_{C,t},\ \forall t$.

Let $Z_t = Y_t - f_{C_d,t},\ \forall t$. Since the predictive distribution is a univariate Gaussian, the logarithm of the likelihood ratio can be written as follows:

$$\begin{aligned} \log\left(\frac{p(y_t \,|\, \mathbf{x}_{C_i,t})}{p(y_t \,|\, \mathbf{x}_{C,t})}\right) &= \frac{1}{2}\left(\frac{\frac{\sqrt{\xi_{C,t}}}{\sqrt{2\pi}} \exp\left(\frac{1}{2}\xi_{C,t}(Y_t - f_{C,t})^2\right)}{\frac{\sqrt{\xi_{C_i,t}}}{\sqrt{2\pi}} \exp\left(\frac{1}{2}\xi_{C_i,t}(Y_t - f_{C_i,t})^2\right)}\right) && \text{(25a)} \\ &= \frac{1}{2}\left(\log\left(\frac{\xi_{C_i,t}}{\xi_{C,t}}\right) - \xi_{C_i,t}(Y_t - f_{C_i,t})^2 - \xi_{C,t}(Y_t - f_{C,t})^2\right), && \text{(25b)} \\ &= \frac{1}{2}\log\left(\frac{\xi_{C_i,t}}{\xi_{C,t}}\right) - \frac{1}{2}(\xi_{C_i,t} - \xi_{C,t})Z_t^2 - \xi_{C,t}\Delta_{f_{C_i,t}} Z_t + \frac{1}{2}\xi_{C,t}\left(\Delta_{f_{C_i,t}}\right)^2, && \text{(25c)} \end{aligned}$$

where the expression, $\Delta_{f_{C_i,t}} = f_{C,t} - f_{C_i,t},\ \forall t$, denotes the difference in means.

The KL divergence is given by the expectation of (25c) with respect to the predictive distribution that includes the i-th feature. Whilst the expression is not linear in Z_t, recall that in general $\mathbb{E}[Z_t^2] \triangleq var\left[Z_t\right] + \mathbb{E}[Z_t]^2$. The variance and expected value, of Z_t reduces to $var\left[Z_t\right] = var\left[Y_t - f_{C,t}\right] = 1/\xi_{C,t}$ and $\mathbb{E}[Z_t] = \mathbb{E}[Y_t - f_{C_d,t}] = 0$, respectively.

Given (13), we get that $\xi_{C_i,t} \xrightarrow{t} \xi_{C,t}$, hence we can derive the KL divergence as follows, which completes the proof:

$$\begin{aligned}
\mathbb{E}\left[D_{\mathrm{KL}}\left(p(y_t|\mathbf{x}_{C_i,t})\|p(y_t|\mathbf{x}_{C,t})\right)\right] &= \mathbb{E}\left[\int_{Y_t} p(y_t\,|\,\mathbf{x}_{C_i,t})\,\log\left\{\frac{p(y_t\,|\,\mathbf{x}_{C_i,t})}{p(y_t\,|\,\mathbf{x}_{C,t})}\right\}dy_t\right], && \text{(26a)}\\
&= \mathbb{E}\left[\frac{1}{2}\left(\log\left\{\frac{\xi_{C_i,t}}{\xi_{C,t}}\right\} - 1 + \frac{\xi_{C,t}}{\xi_{C_i,t}} + \xi_{C,t}\Delta^2_{f_{C_i,t}}\right)\right], && \text{(26b)}\\
&\xrightarrow{t} \frac{\xi}{2}\,\mathbb{E}\left[\left(f_{C_i,t} - f_{C,t}\right)^2\right],\ \text{almost surely.} && \text{(26c)}
\end{aligned}$$

■

Proof [Corollary 2]. The marginal contribution derived using the expected NLL is equivalent to that obtained using the quadratic loss, and hence given by an in-sample estimate of the following,

$$m_{i,t}(C) = \frac{\xi^2}{2}\left(\mathbb{E}\left[(f_{C,t} - y_t)^2\right] - \mathbb{E}\left[(f_{C_i,t} - y_t)^2\right]\right). \tag{27}$$

Combining the results from Lemmas B.1 and B.2, we can see that (14) $\xrightarrow{t}$ (27), and therefore, since all other terms within the definition in (9) remain unchanged, the Shapley values, and therefore the payments, will converge, thereby completing the proof. ■

Appendix C. Proof of Budget Balance Violation (Theorem 3)

We provide here a proof that budget balance is violated under the definition of marginal contribution in (14), as described in Theorem 4.

Recall that budget balance proceeds from the semivalue axiom *efficiency*, which in our context translate to: the total attribution allocated to all features should sum to the value of the grand coalition, that is, $v_t(\mathcal{I}_c) - v_t(\mathcal{I}) = \sum_{i\in\mathcal{I}_{-c}} \phi_{i,t},\ \forall t$. Using (9) we can expand this definition to reveal the telescoping sum structure of the Shapley value such that

$$\begin{aligned}
\sum_{i\in\mathcal{I}_{-c}} \phi_{i,t} = &\sum_{i\in\mathcal{I}_{-c}}\ \sum_{C\in\mathcal{P}(\mathcal{I}_{-c}\backslash\{i\})} \frac{|C|!(|\mathcal{I}_{-c}| - |C| - 1)!}{|\mathcal{I}_{-c}|!}\ (C\cup\mathcal{I}_c) - v_t(C\cup\mathcal{I}_c\cup\{i\}))\,, \quad \forall t, && \text{(28a)}\\
=&|\mathcal{I}_{-c}|\ \underbrace{\frac{0!(|\mathcal{I}_{-c}| - 1)!}{|\mathcal{I}_{-c}|!}v_t(\mathcal{I}_c)}_{\substack{\text{The value of the \textbf{central agent}}\\ \text{coalition appears } |\mathcal{I}_{-c}|\\ \text{times.}}} \ - \ |\mathcal{I}_{-c}|\underbrace{\frac{(|\mathcal{I}_{-c}| - 1)!1!}{|\mathcal{I}_{-c}|!}v_t(\mathcal{I})}_{\substack{\text{The value of the \textbf{grand}}\\ \text{coalition appears } |\mathcal{I}_{-c}|\\ \text{times.}}} && \text{(28b)}\\
&+ \sum_{\substack{C\in\mathcal{P}(\mathcal{I}_{-c})\\ C\neq\emptyset}} (|\mathcal{I}_{-c}| - |C|)\underbrace{\left(\frac{|C|!(|\mathcal{I}_{-c}| - |C| - 1)!}{|\mathcal{I}_{-c}|!}\right)v_t(C\cup\mathcal{I}_c)}_{\substack{\text{The value of the coalition } C \text{ appears}\\ |\mathcal{I}_{-c}| - |C| \text{ times with a \textbf{positive}}\\ \text{sign, i.e., once per agent \textbf{in} } C.}}\\
&- \sum_{\substack{C\in\mathcal{P}(\mathcal{I}_{-c})\\ C\neq\emptyset}} |C|\underbrace{\left(\frac{(|C| - 1)!(|\mathcal{I}_{-c}| - |C|)!}{|\mathcal{I}_{-c}|!}\right)v_t(C\cup\mathcal{I}_c),}_{\substack{\text{The value of the coalition } C \text{ appears}\\ |C| \text{ times with a \textbf{negative} sign, i.e.,}\\ \text{once per agent \textbf{not in} } C.}} \quad \forall t,\\
=&v_t(\mathcal{I}_{)}) - v_t(\mathcal{I}), \quad \forall t. && \text{(28c)}
\end{aligned}$$

However, if we replace the expression for the marginal contribution with the definition in (14), we can similarly derive an expression for the value of the grand coalition, given by

$$\sum_{i \in \mathcal{I}_{-c}} \phi_{i,t} = |\mathcal{I}_{-c}| \frac{(|\mathcal{I}_{-c}| - 1)!}{|\mathcal{I}_{-c}|!} \mathbb{E} \left[D_{\mathrm{KL}} \left(p(y_t | \mathbf{x}_{\mathcal{I},t}, \mathcal{D}_{\mathcal{I},t-1}) \| p(y_t | \mathbf{x}_{\mathcal{I}_c,t}, \mathcal{D}_{\mathcal{I}_c,t-1}) \right) \right]. \tag{29}$$

This, however, holds true if and only if the KL divergence satisfies the triangle inequality with equality. Yet, we know that statistical divergence metrics do not satisfy the triangle inequality, that is, for any probability densities, X, Y and Z, we get that $D_{\mathrm{KL}}(X\|Z) \nleq D_{\mathrm{KL}}(X\|Y) + D_{\mathrm{KL}}(Y\|Z)$. Hence, by contradiction we can prove that using the KL divergence in this manner violates the efficiency axiom, and subsequently budget balance cannot be guaranteed. ■

Appendix D. Proof of Corollary 3

Proof [Corollary 3]. In (15), we re-defined the valuation of a coalition to incorporate the KL divergence as below, which in Appendix B, was shown to converge to,

$$v_t(C) = \mathbb{E} \left[D_{\mathrm{KL}} \left(p(y_t | \mathbf{x}_{C,t}) \| p(y_t | \mathbf{x}_{\mathcal{I}_c,t}) \right) \right], \tag{30a}$$

$$\xrightarrow{t} \frac{\xi^2}{2} \left(\mathbb{E} \left[(f_{\mathcal{I}_c,t} - y_t)^2 \right] - \mathbb{E} \left[(f_{C,t} - y_t)^2 \right] \right), \text{ almost surely.} \tag{30b}$$

Therefore, the marginal contribution of a feature to a coalition converges to the following:

$$m_{i,t}(C) \xrightarrow{t} \frac{\xi^2}{2} \left(\mathbb{E} \left[(f_{\mathcal{I}_c,t} - y_t)^2 \right] - \mathbb{E} \left[(f_{C_i,t} - y_t)^2 \right] - \mathbb{E} \left[(f_{\mathcal{I}_c,t} - y_t)^2 \right] - \mathbb{E} \left[(f_{C,t} - y_t)^2 \right] \right), \tag{31a}$$

$$= \frac{\xi^2}{2} \left(\mathbb{E} \left[(f_{C,t} - y_t)^2 \right] - \mathbb{E} \left[(f_{C_i,t} - y_t)^2 \right] \right), \tag{31b}$$

$$= (27). \tag{31c}$$

As in Appendix B, since all other terms within the definition in (9) remain unchanged, the Shapley values, and therefore the payments, will converge, thereby completing the proof. ■

References

Jacob Abernethy, Yiling Chen, and Jennifer Wortman Vaughan. Efficient market making via convex optimization, and a connection to online learning. *ACM Transactions on Economics and Computation (TEAC)*, 1(2):1–39, 2013.

Jacob D Abernethy and Rafael Frongillo. A collaborative mechanism for crowdsourcing prediction problems. *Advances in neural information processing systems*, 24, 2011.

Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, 14(4):218–256, 2022.

Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of Economic Literature*, 54(2):442–492, 2016.

Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726, 2019.

Anish Agarwal, Munther Dahleh, Thibaut Horel, and Maryann Rui. Towards data auctions with externalities, 2020. URL https://arxiv.org/abs/2003.08345.

Lucas Agussurja, Xinyi Xu, and Bryan Kian Hsiang Low. On the convergence of the shapley value in parametric Bayesian learning games. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 180–196. PMLR, 17–23 Jul 2022.

Adrian Barbu, Nathan Lay, and Shie Mannor. An introduction to artificial prediction markets for classification. *Journal of Machine Learning Research*, 13(7), 2012.

Dirk Bergemann and Alessandro Bonatti. Markets for information: An introduction. *Annual Review of Economics*, 11(1):85–107, 2019.

Giulio Bottazzi and Daniele Giachini. Far from the madding crowd: Collective wisdom in prediction markets. *Quantitative Finance*, 19(9):1461–1471, 2019.

Xuanyu Cao, Yan Chen, and K. J. Ray Liu. Data trading with multiple owners, collectors, and users: An iterative auction mechanism. *IEEE Transactions on Signal and Information Processing over Networks*, 3(2):268–281, 2017.

Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6): 1–168, 2011.

Yiling Chen and David M Pennock. Designing markets for prediction. *AI Magazine*, 31(4):42–52, 2010.

Yiling Chen and Jennifer Wortman Vaughan. A new understanding of prediction markets via no-regret learning. In *Proceedings of the 11th ACM conference on Electronic commerce*, pages 189–198, 2010.

Zicun Cong, Xuan Luo, Jian Pei, Feida Zhu, and Yong Zhang. Data pricing in machine learning pipelines. *Knowledge and Information Systems*, 64(16):1417–1455, 2022.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223, 2020.

Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. Truthful linear regression. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of the 28th Conference on Learning Theory*, pages 448–483, Paris, France, 2015.

Munther A Dahleh, Alireza Tahbaz-Salehi, John N Tsitsiklis, and Spyros I Zoumpoulis. Coordination with local information. *Operations Research*, 64(3):622–637, 2016.

Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

Pradeep Dubey, Abraham Neyman, and Robert James Weber. Value theory without efficiency. *Mathematics of Operations Research*, 6(1):122–128, 1981.

Edwin Fong and Chris C Holmes. On the marginal likelihood and cross-validation. *Biometrika*, 107 (2):489–496, 2020.

Rafael Frongillo and Bo Waggoner. Bounded-loss private prediction markets. *Advances in Neural Information Processing Systems*, 31, 2018.

Esther Gal-Or. Information sharing in oligopoly. *Econometrica*, 53(2):329–343, 1985.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR, 09–15 Jun 2019.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(1):1157–1182, 2003.

Isabelle Guyon and André Elisseeff. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(1):1–18, 2010.

Dongge Han, Michael Wooldridge, Alex Rogers, Olga Ohrimenko, and Sebastian Tschiatschek. Replication robust payoff allocation in submodular cooperative games. *IEEE Transactions on Artificial Intelligence*, 2022a.

Liyang Han, Pierre Pinson, and Jalal Kazempour. Trading data for wind power forecasting: A regression market with lasso regularization. *Electric Power Systems Research*, 212:108442, 2022b.

FA Hayek. The use of knowledge in society. *The Economic Nature of the Firm: A Reader*, pages 66–71, 1986.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Jinli Hu and Amos Storkey. Multi-period trading prediction markets with connections to machine learning. In *International Conference on Machine Learning*, pages 1773–1781. PMLR, 2014.

Fatemeh Jahedpari, Talal Rahwan, Sattar Hashemi, Tomasz P Michalak, Marina De Vos, Julian Padget, and Wei Lee Woon. Online prediction via continuous artificial prediction markets. *IEEE Intelligent Systems*, 32(1):61–68, 2017.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning, 2019. URL https://arxiv.org/abs/1912.04977.

Matti Juhani Koivisto and Juan Pablo Murcia Leon. Solar pv generation time series (pecd 2021 update), 2022. URL https://data.dtu.dk/articles/dataset/Solar_PV_generation_time_series_PECD_2021_update_/19727239.

Iordanis Koutsopoulos, Aristides Gionis, and Maria Halkidi. Auctioning data for learning. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 706–713, 2015.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

R Kulhavỳ and Martin B Zarrop. On a general concept of forgetting. *International Journal of Control*, 58(4):905–924, 1993.

Nicolas S Lambert, John Langford, Jennifer Wortman, Yiling Chen, Daniel Reeves, Yoav Shoham, and David M Penno k. Self-financed wagering mechanisms for forecasting. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 170–179, 2008.

Nathan Lay and Adrian Barbu. The artificial regression market. *arXiv preprint arXiv:1204.4154*, 2012.

Jinfei Liu. Absolute shapley value, 2020.

Charles F Manski. Interpreting the predictions of prediction markets. *economics letters*, 91(3): 425–429, 2006.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Stephen Morris and Hyun Song Shin. Social value of public information. *american economic review*, 92(5):1521–1534, 2002.

Olga Ohrimenko, Shruti Tople, and Sebastian Tschiatschek. Collaborative machine learning markets with data-replication-robust payments, 2019. URL https://arxiv.org/abs/1911.09052.

Václav Peterka. Bayesian approach to system identification. In *Trends and Progress in System identification*, pages 239–304. Elsevier, 1981.

Pierre Pinson, Liyang Han, and Jalal Kazempour. Regression markets and application to energy forecasting. *TOP*, 30(3):533–573, 2022a.

Pierre Pinson et al. To share or not to share? the future of collaborative forecasting. *Foresight: The International Journal of Applied Forecasting*, pages 8–15, 2022b.

Mohammad Rasouli and Michael I Jordan. Data sharing markets, 2021. URL https://arxiv.org/abs/2107.08630.

Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997.

Amos Storkey. Machine learning markets. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 716–724. JMLR Workshop and Conference Proceedings, 2011.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR, 13–18 Jul 2020.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Ardalan Vahidi, Anna Stefanopoulou, and Huei Peng. Recursive least squares with forgetting for online estimation of vehicle mass and road grade: theory and experiments. *Vehicle System Dynamics*, 43(1):31–55, 2005.

Sumio Watanabe and Manfred Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.