Towards Replication-Robust Analytics Markets

Anonymous Authors[1]

Abstract

Many industries rely on data-driven analytics, yet useful datasets are often distributed amongst market competitors that are reluctant to collaborate and share information. Recent literature proposes *analytics markets* to provide monetary incentives for data sharing, however many of these market designs are vulnerable to malicious forms of *replication*—whereby agents replicate their data and act under multiple identities to increase revenue. We develop a *replication-robust* analytics market, centering on supervised learning for regression. To allocate revenue, we use a Shapley value-based attribution policy, framing the features of agents as players and their interactions as a characteristic function game. We show that there are different ways to describe such a game, each with causal nuances that affect robustness to replication. Our proposal is validated using a real-world wind power forecasting case study.

1. Introduction

It is often the case that, when faced with an analytics task, a firm could benefit from using the data of others. For example, rival distributors of similar retail goods could improve supply forecasts by sharing sales data, hotel owners might find value in data from airline companies for anticipating demand, hospitals could reduce socio-economic biases from diagnostic support systems by sharing patient information, and so forth. In our work, we consider the example of renewable energy producers. Specifically, wind power generators exhibit uncertain levels of production and thus require forecasts to competitively participate in electricity markets, their revenue being a function of predictive performance. By harnessing data that is distributed (i.e., both geographically and by ownership) these agents can leverage spatio-temporal correlations between sites to improve their forecasts (Tastu et al., 2013). However, in practice, such altruistic sharing of information amongst market competitors would likely be hindered by privacy concerns or perceived conflicts of interest. Data can instead be *commoditized* within a market-based framework, where remuneration provides an incentive for data sharing (Bergemann & Bonatti, 2019).

Analytics markets are a subset of such frameworks, where data of distributed agents is used to enhance an analytics task without the need to directly transfer raw data to competing agents, through the use of a central market platform (which may additionally ensure privacy preservation) (Pinson et al., 2022). Market revenue is then a function of the enhanced capabilities provided, and the value this brings to the owner of the task. For *fair* allocation of revenue, each dataset owned by a distributed agent should be remunerated based on its marginal contribution to the enhancement of the task (e.g., improved forecast accuracy). However, this can be challenging to quantify when these datasets are correlated. For instance, if datasets are valued sequentially, correlations can reduce social welfare, with agents eventually selling their data for less than their initial valuation since their information becomes redundant (Acemoglu et al., 2022). Whilst this is not the case in our proposed analytics market (i.e., valuation occurs in parallel, hence one agent cannot intentionally undercut another), the value of overlapping information is inherently combinatorial.

To address this, recent works have proposed to borrow concepts from cooperative game theory, framing the features as players and their interactions as a characteristic function game (Ghorbani & Zou, 2019). For many practitioners, the Shapley value (Shapley, 1997) is the solution concept of choice for such a game, allocating each player its expected marginal contribution towards a set of other players, satisfying a collection of axioms that yield several desirable market properties by design, namely individual rationality, zero-element and truthfulness, symmetry, linearity and budget balance, as demonstrated in Agarwal et al. (2019). That being said, a key limitation of this approach is that there is an incentive for agents to replicate their data and act under multiple identities, rendering grossly undesirable revenue allocations. This incentive arises from the fundamental nature data—it can be replicated at zero marginal cost. Whilst several attempts have been made to remedy this issue (e.g., Agarwal et al., 2019, Ohrimenko et al., 2019, Han et al.,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2023), doing so typically involves a trade-off.

The contributions of our paper are as follows: (i) we develop an analytics market with Shapley value-based attribution that is *replication-robust* under a more favourable definition compared to previous works; (ii) we propose a general market design that subsumes many of the existing proposals in the literature and explore the intricacies of the different ways in which an analytics task can be represented as a cooperative game; (iii) we demonstrate that each has causal nuances that can determine robustness to replication of the market; and finally (iv) we apply our work on a real-world case study—out of many potential applications, we choose to study wind power forecasting due to data availability, the known value of sharing distributed data, and the fact it is a sandbox that can be easily shared and used by others.

2. Preliminaries

We define an *analytics task* as a regression model to be used for forecasting, such that our focus is on so-called *regression markets* (Pinson et al., 2022). This setting builds upon prior work on data acquisition from both strategic (Dekel et al., 2010) and privacy-conscious (Cummings et al., 2015) agents. The owner of the regression model is characterized by their private valuation for a marginal improvement in predictive performance, which sets the price for the distributed agents, whom in turn propose their own data as features and are eventually remunerated based on their relative marginal contributions. We write this valuation as $\lambda \in \mathbb{R}_+$, the value of which we assume to have been learnt through some preliminary analyses.

Market Agents The set $\mathcal{A}$ denotes the market agents, one of which $c \in \mathcal{A}$ is the *central agent* seeking to improve their forecasts, whilst the remaining agents $a \in \mathcal{A}_{-c}$ are *support agents*, whom propose their own data as features, whereby $\mathcal{A}_{-c} = \mathcal{A} \setminus \{c\}$. Let $y_t \in \mathbb{R}_+$ be the target signal recorded by the central agent at time t, a sample from the stochastic process $\{Y_t\}_{\forall t}$. We write $\boldsymbol{x}_{\mathcal{I},t}$ as the vector of all 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. For each subset of features $\mathcal{C} \subseteq \mathcal{I}$ we write $\mathcal{D}_{\mathcal{C},t} = \{\boldsymbol{x}_{\mathcal{C},t'}, y_{t'}\}_{\forall t' \leq t}$ to be the set of observations up until time t.

Regression Framework To model the target signal, Y_t, we use a parametric Bayesian regression framework, formulating the likelihood as a deviation from a deterministic mapping under an independent Gaussian noise process the variance of which is treated as a hyperparameter. The mapping, f, is a linear interpolant parameterized by a vector of coefficients, $\boldsymbol{w}$, and represents the conditional expectation of the target signal, such that the expectation of the likelihood corresponding to the *grand coalition* (i.e., using all available input features) at any particular time step can be decomposed as follows:

$$f(\boldsymbol{x}_t, \boldsymbol{w}) = \underbrace{w_0 + \sum_{i \in |\mathcal{I}_c|} w_i x_{i,t}}_{\text{Terms belonging to the central agent.}} + \underbrace{\sum_{a \in \mathcal{A}_{-c}} \sum_{j \in |\mathcal{I}_a|} w_j x_{j,t}}_{\text{Terms belonging to the support agents.}} . \quad (1)$$

Market Clearing Once the data has been collected and the valuation of the central agent is revealed, the market is then cleared. We consider a two-stage (i.e., in-sample and out-of-sample) batch market, as in Pinson et al. (2022). We do, however, relax the assumption that features are independent, but still assume that any redundant features owned by the support agents (i.e., those highly correlated with the central agent's features) are removed via the detailed feature selection process. An important step in the market clearing procedure is parameter inference—to mitigate bias we opt for a centred isotropic (i.e., uninformative) Gaussian prior, which is conjugate for our likelihood, resulting in a tractable Gaussian posterior which summarizes our updated beliefs, which, for a particular subset of features is given by

$$p(\boldsymbol{w}_{\mathcal{C}}|\mathcal{D}_{\mathcal{C},t}) \propto p(\boldsymbol{x}_{\mathcal{C},t}, y_t|\boldsymbol{w}_{\mathcal{C}}) p(\boldsymbol{w}_{\mathcal{C}}|\mathcal{D}_{\mathcal{C},t-1}), \quad \forall t, \quad (2)$$

where recall $\mathcal{D}_{\mathcal{C},t}$ is the set of input-output pairs observed up until time t, for all $\mathcal{C} \subseteq \mathcal{I}$. The market revenue is then a function of the exogenous valuation, λ, and the extent to which model-fitting is improved, which we measure using the negative logarithm of the predictive density (i.e., the convolution of the likelihood with the posterior), denoted by $\ell_t = -\log[p(y_t|\boldsymbol{x}_t)]$, $\forall t$, where for a batch of observations, the market revenue is $\pi = \lambda(\mathbb{E}[\ell_t]_{\mathcal{I}_c} - \mathbb{E}[\ell_t]_{\mathcal{I}})$.

Revenue Allocation To allocate market revenue amongst support agents, we define an attribution policy based on the Shapley value. We let $v : \mathcal{C} \in \mathcal{P}(\mathcal{I}) \mapsto \mathbb{R}$ be a characteristic function that maps the power set of indices of all the features to a real-valued scalar—the set $\mathcal{C}$ represents a coalition in the cooperative game. If we let Θ be the set of all possible permutation of indices in $\mathcal{I}_{-c}$, the Shapley value is $\phi_i = 1/|\mathcal{I}_{-c}|! \sum_{\theta \in \Theta} \Delta_i(\theta)$, $\forall i \in \mathcal{I}_{-c}$, where $\Delta_i(\theta) = v(\mathcal{I}_c \cup \{j : j \prec_\theta i\}) - v(\mathcal{I}_c \cup \{j : j \preceq_\theta i\})$, where $j \prec_\theta i$ denotes that j precedes i in permutation θ. With this attribution policy, the revenue of each support agent can be written as $\pi_a = \sum_{i \in \mathcal{I}_a} \lambda \mathbb{E}[\phi_i]$, $\forall a \in \mathcal{A}_{-c}$. Observe that calculating Shapley values requires evaluating the objective function using subsets of features, which is not that straightforward in general—once trained, machine learning models typically require an input vector containing a value for each feature to avoid matrix dimension mismatch. As a result, the characteristic function must *lift* the original objective to simulate removal of features (Merrill et al., 2019).

Recall that our objective function, ℓ, relates to the mapping $f : \mathbb{R}^{|\mathcal{I}|} \mapsto \mathbb{R}$ described in (1), and is therefore itself only defined in $\mathbb{R}^{|\mathcal{I}|}$. To calculate the Shapley values, a value for each of the $2^{|\mathcal{I}|}$ subsets of input features is needed. Accordingly, we lift the objective function to the space of all subsets of features by formulating the characteristic function mapping as $v(\mathcal{C}) : \mathbb{R}^{|\mathcal{I}|} \times 2^{|\mathcal{I}|} \mapsto \mathbb{R}, \forall \mathcal{C}$. Hence, for the grand coalition, $v(\mathcal{I}) = \mathbb{E}[\ell_{\mathcal{I},t} | \boldsymbol{X}_t = \boldsymbol{x}_t]$, where $\boldsymbol{X}_t$ is the multivariate random variable from which the features are perceived to be sampled. For a particular feature, the Shapley value is therefore not generally well-defined, since there exists many methods to formulate such a lift (Sundararajan & Najmi, 2020). In the next section we explore the following: (i) how to compute such a lift within a linear regression setup, (ii) what implications different lifts have in relation to causality, and (iii) how this subsequently affects the market revenue allocations.

3. Lift Formulations

Commonly adopted lifts can broadly be categorized as either *observational* or *interventional*, differing only in the functional form of the characteristic function that underpins the cooperative game. The former is typically found in work pertinent to regression markets (e.g, Agarwal et al. 2019), with the latter used an approximation for the former for interpreting model predictions (Lundberg & Lee, 2017). The observational lift uses the *observational conditional expectation*, the expectation of the objective over the conditional density of the out-of-coalition features, given that those in the coalition take on their observed values, such that

$$v^{\text{obs}}(\mathcal{C}) = \int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\mathcal{C},t}, \boldsymbol{x}_{\overline{\mathcal{C}},t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | \boldsymbol{x}_{\mathcal{C},t}) d\boldsymbol{x}_{\overline{\mathcal{C}},t}, \quad (3)$$

where $\overline{\mathcal{C}} = \mathcal{I} \setminus \mathcal{C}$ denotes the out-of-coalition features.

The interventional lift uses the *interventional conditional expectation*, whereby features in the coalition are manually fixed to their observed values, intentionally manipulating the data generating process, which we express mathematically using Pearl's *do*-calculus (Pearl, 2012), such that

$$v^{\text{int}}(\mathcal{C}) = \int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\mathcal{C},t}, \boldsymbol{x}_{\overline{\mathcal{C}},t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | do(\boldsymbol{x}_{\mathcal{C},t})) d\boldsymbol{x}_{\overline{\mathcal{C}},t}. \quad (4)$$

The difference between (3) and (4) is that in the latter, dependence between the out-of-coalition features and those within the coalition is broken. In theory, *observing* $\boldsymbol{X}_{\mathcal{C},t} = \boldsymbol{x}_{\mathcal{C},t}$ would change the distribution of $\boldsymbol{X}_{\overline{\mathcal{C}},t}$ if the random variables were connected through latent effects. However, by *intervening* on a coalition, the distribution of these out-of-coalition features is unaffected. To illustrate this, consider two random variables, X and Y, with the causal relationship in Figure 1. If we observe $X = x$ the observational conditional distribution describes: *the distribution of Y given that X is observed to take on the value x*, which we normally write as $p(y|x) = p(x, y)/p(x)$. The interventional conditional distribution describes instead: *the distribution of Y given we artificially set the value of X to x*, denoted $p(y|do(x))$, obtained by assuming that Y is distributed by the original data generating process. Graphically, interventions remove all edges going into the corresponding variable. Consequently, we get that, $p(y|do(x)) = p(y|x)$ but $p(x|do(y)) = p(x)$. This means that the distribution of y under the *intervention* $X = x$ is equivalent to y *conditioned* on $X = x$, yet for $Y = y$, x and y become disconnected, hence x has no effect on y, which is simply sampled from its marginal distribution.

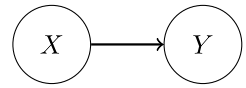


Figure 1. Causal graph indicating a direct effect between two random variables, X and Y.

Typically, the choice of which lift to use is driven by their relative computational expenditure (Lundberg & Lee, 2017)—evaluating the conditional expectation of the objective function is intractable in general, requiring complex and costly methods for approximation (Covert et al., 2021), whereas cheap and relatively simple algorithms exist to *intervene* on the features (Sundararajan & Najmi, 2020). Whilst the most suitable method for evaluating the conditional expectation is widely disputed (Chen et al., 2022), one such method merely requires training separate models for each subset of features; if each model is optimal with respect to the objective, then this is equivalent to marginalizing out features using their conditional distribution (Covert et al., 2021). Similarly, one can evaluate the interventional conditional expectation of the objective function for linear regression models by imputing, or even removing completely, the features not present in a coalition. We note that, both of these lifts preserve the axioms of the original Shapley value, and subsequently the market properties provided, albeit in expectation.

Causal Nuances With independent features, both lift formulations are in fact equivalent. Specifically, Janzing et al. (2020) showed that by distinguishing between the *true* features and those actually used as *input* to the model, as in our example we get that $p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | do(\boldsymbol{x}_{\mathcal{C},t})) = p(\boldsymbol{x}_{\overline{\mathcal{C}},t})$. We can then calculate (4) from (3) by simply replacing $p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | \boldsymbol{x}_{\mathcal{C},t})$ with the marginal distribution, which would be equivalent for independent features.

Theorem 3.1. *Marginal contributions derived using the observational conditional expectation formulation for $v(\cdot)$ as defined in (3) can be decomposed into indirect and direct causal effects.*

Proof. Following (3), the marginal contribution of the i-th feature for a single permutation $\theta \in \Theta$ derived using the observational lift can be written as

$$
\begin{aligned}
\Delta_i^{\text{obs}}(\theta) &= v(\underline{\mathcal{C}}) - v(\underline{\mathcal{C}} \cup i), \\
&= \underbrace{\int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\underline{\mathcal{C}},t}, \boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t} | \boldsymbol{x}_{\underline{\mathcal{C}},t}) d\boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t} - \int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\underline{\mathcal{C}} \cup i,t}, \boldsymbol{x}_{\overline{\mathcal{C}},t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | \boldsymbol{x}_{\underline{\mathcal{C}} \cup i,t}) d\boldsymbol{x}_{\overline{\mathcal{C}},t},}_{\text{Total effect}} \\
&= \underbrace{\int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\underline{\mathcal{C}},t}, \boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t} | \boldsymbol{x}_{\underline{\mathcal{C}},t}) d\boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t} - \int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\underline{\mathcal{C}} \cup i,t}, \boldsymbol{x}_{\overline{\mathcal{C}},t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | \boldsymbol{x}_{\underline{\mathcal{C}},t}) d\boldsymbol{x}_{\overline{\mathcal{C}},t}}_{\text{Direct effect}} \\
&\quad \underbrace{+ \int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\underline{\mathcal{C}} \cup i,t}, \boldsymbol{x}_{\overline{\mathcal{C}},t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | \boldsymbol{x}_{\underline{\mathcal{C}},t}) d\boldsymbol{x}_{\overline{\mathcal{C}},t} - \int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\underline{\mathcal{C}} \cup i,t}, \boldsymbol{x}_{\overline{\mathcal{C}},t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | \boldsymbol{x}_{\underline{\mathcal{C}} \cup i,t}) d\boldsymbol{x}_{\overline{\mathcal{C}},t},}_{\text{Indirect effect}}
\end{aligned}
$$

where $\underline{\mathcal{C}} = \{j : j \prec_\theta i\}$ and $\overline{\mathcal{C}} = \{j : j \succ_\theta i\}$. This decomposition measures the difference in the loss function when: (i) the value of the i-th feature is observed and the distribution of the remaining out-of-coalition features is unchanged (i.e., direct effect); and (ii) the distribution of the other out-of-coalition features does changed as a result of observing the i-th feature (i.e., indirect effect). $\square$

Following Theorem 3.1, by replacing the condition by observation with the marginal distribution as in (3), we eliminate the indirect effect entirely. Hence, using the interventional lift removes consideration of causal effects *between features*, and subsequently any root causes with strong *indirect effects* (Heskes et al., 2020). As a result, this lift is more effective at crediting features on which the regression model has an explicit algebraic dependence. In contrast, the observational lift attributes features in proportion to indirect effects (Aas et al., 2021).

To illustrate this, consider the following example, adapted from Janzing et al. (2020) to fit our context. Let $x_{1,t}, x_{2,t} \in \{0,1\}$ be two binary features such that $p(x_{1,t}, x_{2,t}) = 1/2$ if $x_{1,t} = x_{2,t}$, otherwise $p(x_{1,t}, x_{2,t}) = 0$. If $p(y_t|\boldsymbol{x}_t) = \mathcal{N}(x_{1,t}, 1)$ and $y_t = 1$, the expected value of the loss function simplifies to: $\mathbb{E}[\ell_t | x_{1,t}, x_{2,t}] = \log(\sqrt{2\pi}) + 1/2(x_{1,t} - 1)^2$. The following results are obtained:

(i) Observational lift

$$v(\varnothing) = \log(\sqrt{2\pi}) + 1/4, \tag{5a}$$

$$v(\{1\}) = \log(\sqrt{2\pi}) + (1 - x_{1,t})^2, \tag{5b}$$

$$v(\{2\}) = \log(\sqrt{2\pi}) + (1 - x_{2,t})^2, \tag{5c}$$

$$v(\{1,2\}) = \log(\sqrt{2\pi}) + (1 - x_{1,t})^2, \tag{5d}$$

which gives,

$$
\begin{aligned}
\mathbb{E}[\phi_2] &\propto \mathbb{E}[(5a) - (5c) + (5b) - (5d)] \\
&= 1/4 - (1 - x_{1,t})^2 = \mathbb{E}[\phi_1],
\end{aligned}
$$

(ii) Interventional lift

$$v(\varnothing) = \log(\sqrt{2\pi}) + 1/4, \tag{6a}$$

$$v(\{1\}) = \log(\sqrt{2\pi}) + (1 - x_{1,t})^2, \tag{6b}$$

$$v(\{2\}) = \log(\sqrt{2\pi}) + 1/4, \tag{6c}$$

$$v(\{1,2\}) = \log(\sqrt{2\pi}) + (1 - x_{1,t})^2, \tag{6d}$$

which gives,

$$
\begin{aligned}
\mathbb{E}[\phi_2] &\propto \mathbb{E}[(6a) - (6b) + (6c) - (6d)] \\
&= 0 \neq \mathbb{E}[\phi_1].
\end{aligned}
$$

We see that in (5), these features are given equal attribution, which some works argue to be illogical as features not explicitly used by the model have the possibility of receiving non-zero allocation (Sundararajan & Najmi, 2020), whereas in (6), $\phi_i \neq 0$ intuitively implies that the model depends on $x_{i,t}$. Whilst this dispute has been used as an argument to reject the general use of Shapley values for model interoperability in machine learning (Kumar et al., 2020) and that Lundberg & Lee (2017) were mistaken to only convey (4) as a cheap approximation of (3) (Janzing et al., 2020), the choice between the observational and interventional lifts can in fact be viewed as conditional on as to whether one wants to be *true to the data* or *true to the model*, respectively (Chen et al., 2020), meaning the trade-offs of each approach can be seen as context-specific.

Interpreting Payments We can explore this conjecture by considering how the revenues of the support agents may differ depending on the choice of lift. We know that the predictive performance of the regression model out-of-sample is contingent upon the availability of features that were used during training, which, in practice, requires data of the support agents to be streamed continuously in a timely fashion, particularly for an online setup (Pinson et al., 2022). If the stream of any of these features were interrupted, the efficacy of the forecast may drop, the extent to which would relate not to any root causes or indirect effects regarding the data generating process, but rather solely the magnitude of direct effects. Ergo, in a market with an attribution policy based on the interventional Shapley value, larger payments would be made to support agents that own features to which the predictive performance of the model is most sensitive.

This provides an incentive for investment in efforts to decrease the chance of their data being unavailable, resembling availability payments in electricity markets, whereby assets are remunerated for being available in times of need. For the

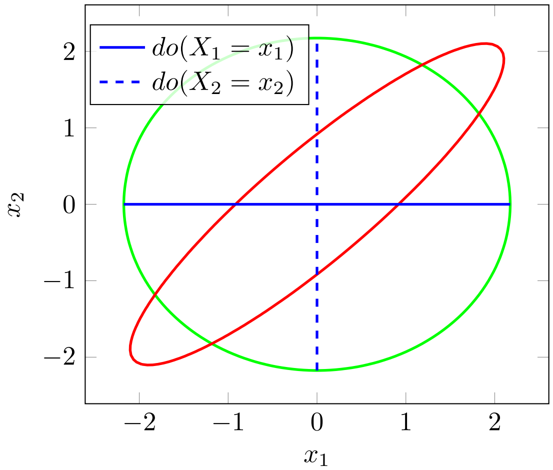


Figure 2. Interventions yielding points outwith the true data manifold. The green and red lines represent the level set within which 0.99 quantile of the training data when features (i.e., X_1 and X_2) are independent and correlated, respectively. The blue lines represent the data extrapolated as a result of intervening on X_1 (solid) and X_2 (dashed).

observational lift, it would instead be unclear as to whether comparatively larger payments in the regression market are consequential of features having a sizeable impact on predictive performance, or merely a result of indirect effects through those that do. It could however be argued that the latter outcome is more *fair*, accounting for the fact that a feature having only indirect effects does not necessarily diminish its propensity to increase predictive performance in the absence of its counterpart with a direct effect, albeit only if the model is re-trained. At this point, this trade-off merely highlights that the choice of lift could yield counter-intuitive allocations if not considered carefully.

Risk Implications When features are not independent, unlike the observational lift, conditioning by intervention leads to the possibility of model evaluation on points outwith the true data manifold (Frye et al., 2020). This can visualized with the simple illustration in Figure 2. If independent, intervening on either feature yields samples that remain within the original data manifold. However, if features are correlated, there is a possibility for extrapolating far beyond the training distribution, where the model is not trained and behaviour is unknown. We now consider what impact this may have on the market outcomes.

We know that if multicollinearity exists, the variance of the coefficients is inflated, which can distort the estimated mean when the number of in-sample observations is limited. The posterior variance of the i-th coefficient can be written as $var(w_i) = \kappa_i/\xi|\mathcal{D}_t|$, where ξ is the intrinsic noise precision of the target and κ_i is the variance inflation factor, given by $\kappa_i = \mathbf{e}_i^\top (\sum_{t\leq t} \boldsymbol{x}_t^\top \boldsymbol{x}_t)^{-1}\mathbf{e}_i,\ \forall i \in \mathcal{I}$, where $\mathbf{e}_i$ is the i-th basis vector. Although $\kappa_i \geq 1$, it has no upper bound, such that $\kappa_i \mapsto \infty,\ \forall i$, with increasing collinearity.

From a variance-decomposition perspective, the expected Shapley value of the i-th feature can be shown to be equivalent to the variance in the target signal that it explains, such that, $\mathbb{E}[\phi_i] = \mathbb{E}[w_i]^2\, var(X_i)$, approximating the behaviour of the interventional Shapley value when features are correlated (Owen & Prieur, 2017). As the posterior distribution is Gaussian, the Shapley value for each feature will follow a noncentral Chi-squared distribution with one degree of freedom. For a particular feature, we can write the probability density function of the distribution of the Shapley value in closed-form as $p(\phi_i)/(var(X_i)var(w_i)) = \sum_{k=0}^{\infty}(1/k!)e^{\eta/2}(\eta/2)^k)\chi^2(1+2k),\ \forall i$, where the noncentral Chi-squared distribution is seen to simply be given by a Poisson-weighted mixture of central Chi-squared distributions, $\chi^2(\cdot)$, with noncentrality $\eta = \mathbb{E}[w_i]^2/var(w_i)$. Since we know the moment generating function for such a mixture, we derive the second moment as follows: $var(\phi_i) = 2var(w_i)\left(2\mathbb{E}[w_i]^2 + var(w_i)\right)\left(var(X_i)\right)^2,\ \forall i.$

This implies that the variance of the attribution, and subsequently the revenue, for any given feature is a quadratic function of the variance of the corresponding coefficient, thus the variance inflation induced by multicollinearity. That being said, this problem indeed vanishes with increasing sample size, as $var(w_i) \mapsto 0,\ \forall i$ (Qazaz et al., 1997). If only a limited number of in-sample observations are available, distorted revenues could in theory be remedied using *zero-Shapley* or *absolute-Shapley* proposed in Liu (2020), or restricting model evaluations to the data manifold (Taufiq et al., 2023). We leave a thorough investigation of these remedies in relation to analytics markets to future work.

4. Robustness To Replication

Although it is natural for datasets to contain some amount of overlapping information, in our analytics market such redundancy may also arise as a result of replication. The fact that data can be freely replicated differentiates it from traditional commodities—a motive for reassessing fundamental mechanism design concepts (Aiello et al., 2001). For example, implementing a simple second price auction becomes impractical unless sellers somehow limit the number of replications, which may in turn curtail revenue.

Definition 4.1. A *replicate* of the i-th feature is defined as $x'_{i,t} = x_{i,t} + \eta$, where η represents centred noise with finite variance, conditionally independent of the target given the feature.

In our context, replication can be seen as strategic behaviour. Specifically, under Definition 4.1, markets that harness the observational lift described in (3) in fact provide a monetary incentive for support agents to replicate their data and act under multiple identities. To see this, consider the causal graph in Figure 3. Suppose that $x_{1,t}$ and $x_{2,t}$ are identical features, such that $w_1 = w_2$, each owned by a separate support agent, a_1 and a_2, respectively. Considering Theorem 3.1 and the example in Section 3, the payment to each support agent before any replication is made will be $\pi/2$, where π is the total market revenue. Now suppose a_2 replicates their feature k times and for simplicity assume $var(\eta) = 0$. With the same logic, the revenues of agents a_1 and a_2 will be $\pi/(2+k)$ and $\sum_{1+k} \pi/(2+k) = \pi(1+k)/(2+k)$, respectively. Hence there is an incentive for agents to simply replicate their data infinitely many times so as to maximize revenue, which is undesirable in practice.

Definition 4.2. Let $\boldsymbol{x}_t^+$ denote the original vector of features augmented to include any additional replicates, with an analogous index set, $\mathcal{I}^+$. According to Agarwal et al. (2019), a market is *replication-robust* if $\pi_a^+ \leq \pi_a, \forall a \in \mathcal{A}_{-c}$, where π_a^+ is the new revenue derived using $\boldsymbol{x}_t^+$ instead.

Since allocation policies based on the observational lift violate this definition, the authors propose *Robust-Shapley* described by $\phi_i^{\text{robust}} = \phi_i \exp(-\gamma \sum_{j \in \mathcal{I}_{-c} \setminus \{i\}} s(X_{i,t}, X_{j,t}))$, with $s(\cdot,\cdot)$ a similarity metric (e.g., cosine similarity). This method penalizes similar features so as to remove the incentive for replication, satisfying Definition 4.2. However, not only replicated features are penalized, but also those with naturally occurring correlations between features. This yields a loss of budget balance, the extent to which depends on the chosen similarly metric and the value of γ. A similar result is obtained by Han et al. (2023) by considering the general class of semivalues to which the Shapley value belongs (Dubey et al., 1981). It is shown that the way in which a semivalue weights coalition sizes has an affect on the resultant properties, and that the Banzhaf value (Lehrer, 1988) is in fact replication-robust by design (i.e., with respect to Definition 4.2), along with many other semivalues, albeit still penalizing naturally occurring correlations. In our view, what has lacked acknowledgement so far is that Definition 4.2 leaves the market susceptible to spiteful agents (i.e., those who seek to minimize the revenue of other agents while maximizing their own profits), thus we refer to this definition as *weakly* robust.

Proposition 4.3. *Analytics markets that adopt a Shapley-value based attribution policy based on the interventional lift instead yield a stricter notion of being replication-robust, such that* $\pi_a^+ \equiv \pi_a, \forall a \in \mathcal{A}_{-c}$.

Proof. Under Definition 4.1, each replicate in $\boldsymbol{x}_t^+$ will only induce an indirect effect on the target. However from Theorem 3.1, we know that the interventional lift only captures direct effects. Therefore, for each of the replicates, we write

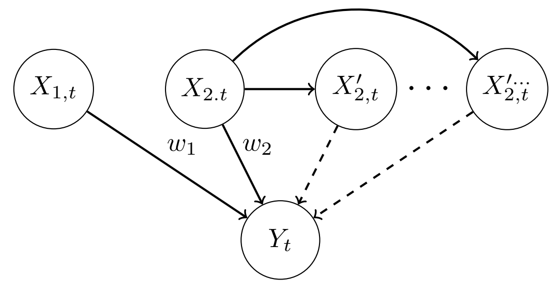


Figure 3. Causal graph indicating direct effects (solid lines) and indirect effects (dashed lines) induced by replicating $X_{2,t}$. The prime superscript denotes a replicated feature.

the marginal contribution for a single permutation $\theta \in \Theta$ as

$$\begin{aligned}
\Delta_i^{\text{int}}(\theta) &= \int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\underline{\mathcal{C}},t}, \boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t} | \boldsymbol{x}_{\underline{\mathcal{C}},t}) d\boldsymbol{x}_{\overline{\mathcal{C}} \cup i,t} \\
&\quad - \int \mathbb{E}\left[\ell_t | \boldsymbol{x}_{\underline{\mathcal{C}} \cup i,t}, \boldsymbol{x}_{\overline{\mathcal{C}},t}\right] p(\boldsymbol{x}_{\overline{\mathcal{C}},t} | \boldsymbol{x}_{\underline{\mathcal{C}},t}) d\boldsymbol{x}_{\overline{\mathcal{C}},t}, \\
&= 0, \quad \forall i \in \mathcal{I}_{-c}^+ \setminus \mathcal{I}_{-c},
\end{aligned}$$

and therefore $\phi_i \propto \sum_{\theta \in \Theta} \Delta_i(\theta) = 0$ for each of the replicates. For the original features, any direct effects will remain unchanged, as visualized in Figure 3. This leads to

$$\pi_a^+ = \sum_{i \in \mathcal{I}_a} \lambda \mathbb{E}[\phi_i] + \sum_{j \in \mathcal{J}_a} \lambda \mathbb{E}[\cancel{\phi_j}], \quad = \pi_a, \quad \forall a \in \mathcal{A}_{-c}$$

where $\mathcal{J}_a = \mathcal{I}_a^+ \setminus \mathcal{I}_a$. □

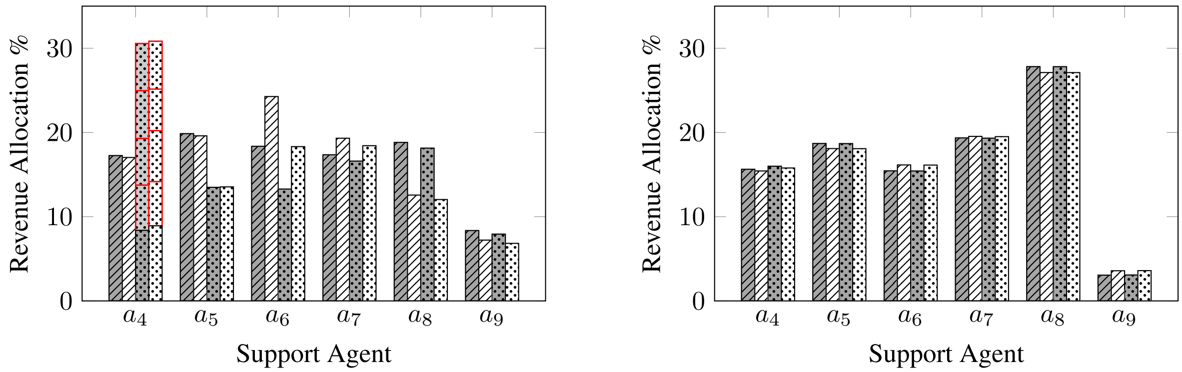
This proposition shows that by using the inverventional lift, the Shapley value-based attribution policy, and hence the regression market, is *strictly* robust to both replication and spiteful agents by design.

5. Experimental Analysis

We now validate our main findings using a real-world case study.[1] We use an open source dataset to aid reproduction of our work, namely the Wind Integration National Dataset (WIND) Toolkit, detailed in Draxl et al. (2015). Our setup is a stylised electricity market setup where agents—in our case, wind producers—are required to notify the system operator of their expected electricity generation in a forward stage, one hour ahead of delivery, for which they receive a fixed price per unit. In real-time, they receive a penalty for deviations from the scheduled production, thus their downstream revenue is an explicit function of forecast accuracy.

Methodology The dataset comprises wind power measurements simulated for 9 wind farms in South Carolina,

[1]Our code has been made publicly available at: `https://github.com/tdfalc/regression-markets`



(a) *Observational*: The revenue earned by a_4 is increased due to indirect effects induced by the replicates.

(b) *Interventional*: The revenue earned by a_4 remains unchanged by accounting only for direct effects.

Figure 4. Revenue allocations for each support agent for both (a) observational and (b) interventional lifts, when agent a_4 is honest (//) and malicious (∘), by replicating their feature. The gray and white bars correspond to in-sample and out-of-sample market stages, respectively. The revenue split amongst replicates is depicted by the stacked bars highlighted in red.

Table 1. Agents and corresponding site characteristics considered in South Carolina (USA). C_{f} denotes the capacity factor and P the nominal capacity.

AGENT	ID.	C_{F} (%)	P (MW)
a_1	4456	34.11	1.75
a_2	4754	35.75	2.96
a_3	4934	36.21	3.38
a_4	4090	26.60	16.11
a_5	4341	28.47	37.98
a_6	4715	27.37	30.06
a_7	5730	34.23	2.53
a_8	5733	34.41	2.60
a_9	5947	34.67	1.24

all located within 150 km of each other—see Table 1 for a characteristic overview. Whilst this data is not exactly *real*, it captures the spatio-temporal aspects of wind power production, with the benefit of remaining free from any spurious measurements. Measurements are available from 2007 to 2013, with an hourly granularity. For simplicity, we let a_1 be the central agent, however in practice each could assume this role in parallel.

We use the regression framework described in Section 2. We employ an *Auto-Regressive with eXogenous input* model, such that each agent is assumed to own a single feature, namely a 1-hour lag of their power measurement. For wind power forecasting, the lag not only captures the temporal correlations of the production at a specific site, but also indirectly encompasses the dependencies amongst neighboring sites owing to the natural progression of wind patterns. We are interested in assessing market outcomes rather than competing with state-of-the-art forecasting methods, so we consider only a very short-term lead time (i.e., 1-hour ahead), thereby permitting a fairly simple time-series analysis. Nevertheless, our mechanism readily allows more complex models for those aiming to capture specific intricacies of wind power production, for instance the bounded extremities of the power curve (Pinson, 2012).

We perform a pre-screening, such that given the redundancy between the lagged measurements of a_2 and a_3 with that of a_1, we remove them from the market in line with our assumptions. The first 50% of observations are used to clear the in-sample regression market and fit the regression model, whilst the latter half is used for the out-of-sample market. We clear both markets considering each agent is honest, that is, they each provide a single report of their true data. Next, we re-clear the markets, but this time assume agent a_4 is malicious, replicating their data, thereby submitting multiple separate features to the market in effort to increase their revenue.

Results We start by setting the number of replicates $k = 4$, and $\lambda = 0.5$ USD per time step and per unit improvement in ℓ, for both in-sample and out-of-sample market stages. However, we are primarily interested in the revenue allocation rather than the magnitude—see Pinson et al. (2022) for a complete analysis of the monetary incentive to each agent participating in the market. Overall the in-sample and out-of-sample objectives improved by 10.6% and 13.3% respectively with the help of the support agents. In Figure 4, we plot the resultant allocation for each agent with and without the malicious behavior of agent a_4, for both lifts. When this agent is honest, we observe that the observational lift spreads credit relatively evenly amongst most features, suggesting that many of them have similar indirect effects on the target. The interventional lift favours agent a_8, which, as expected, owns the features with the greatest

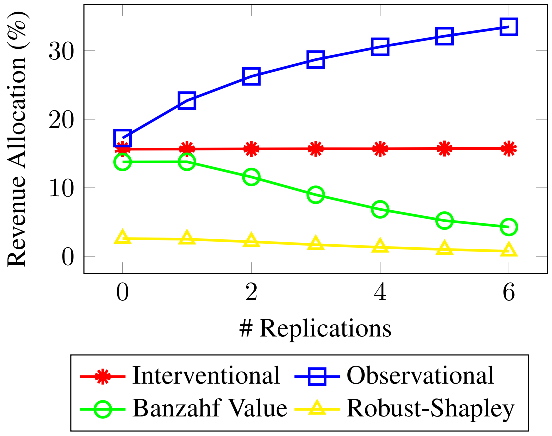


Figure 5. Revenue allocation of agent a_4 with increasing number of replicates.

spatial correlation with the target. In this market, most of the additional revenue of agent a_8 appears to be lost from agent a_9 compared with the observational lift, suggesting that whilst these features are correlated, it is agent a_8 with the greatest direct effect.

When agent a_4 replicates their data, with the observational lift we see agents a_5 to a_8 earn less, whilst agent a_4 earns considerably more. This demonstrates that this conditional expectation indeed spreads revenue proportionally amongst indirect effects, of which there are now four more due to the replicates, and consequently the malicious agent out-earns the others. In contrast, since the interventional lift only attributes direct effects, each replicate is allocated zero revenue, hence the malicious agent is no better off than before. Lastly, we observe that in both cases, the market outcomes were relatively consistent between the in-sample and out-of-sample stages, likely due to the large batch size considered, combined with limited nonstationarities within the data.

To compare our proposal with those in related works, in Figure 5 we plot the allocation of agent a_4 with increasing number of replicates. Here, *Robust-Shapley* and *Banzahf Value* refer to both the penalization approach of Agarwal et al. (2019) and the use of an alternative semivalue in Han et al. (2023), respectively. With the observational lift, the proportion of revenue obtained increases with the number of replicates as expected, as a greater number of indirect effects are owned by the agent. With *Robust-Shapley*, the allocation indeed decreases with the number of replicates, demonstrating this approach is *weakly* replication-robust, but is considerably less compared with the other approaches since natural similarities are also penalized. The authors argue this is an incentive for provision of unique information, but this allows agents to be spiteful. The *Banzahf Value* is strictly replicaiton-robust for $k = 0$, but only weakly for $k \geq 1$. Lastly, unlike these approaches, our proposed interventional lift remains strictly replication-robust throughout as expected, with agent a_4 not able to benefit from replication their feature, without penalizing the other agents.

6. Conclusions

Many analytics tasks akin to the one presented here could benefit from distributed data, however convincing firms to share information, even with assurances of privacy protection, poses a challenge. Rather than relying on data altruism, there have been several proposals of market mechanisms (e.g., analytics markets) to provide incentives for data sharing through monetary compensation, many of which adopt Shapley value-based attribution policies. Nevertheless, there are a number of open issues that remain before such mechanisms can be used in practice, one of which is vulnerability to replication.

In this work, we introduced a general framework for Shapley value-based regression markets that subsumes these existing proposals. We demonstrated that there are different ways to formulate this cooperative game and provided a full causal picture for each formulation, as well as an insight into how each influences the market outcomes. Conventional use of the observational lift to value a coalition is the source of the vulnerability to replication, which many works have tried to remedy through penalization methods, which enable only weak robustness. Our main contribution is the alternative use of the interventional lift, which we have proved to be robust to replication by design, even under a more favourable definition of *strict* robustness.

From a causal perspective, the interventional lift has other potential benefits, including revenue allocations that better represent the reliance of the model on each feature, providing an incentive for timely and reliable data streams for useful features. There is of course, no free lunch, as using the interventional conditional expectation can yield undesirable payments when feautres are highly correlated and the number of observations is low. Nevertheless, future work could examine the extent to which the mentioned remedies mitigate this issue, as well as their impact on the market outcomes. Ultimately, when it comes to data valuation, the Shapley value is not without its limitations—it is not generally well-defined in a machine learning context and requires strict assumptions, not to mention its computational complexity. Perhaps this should also incite future work into alternative mechanism designs, for example those based on non-cooperative game theory instead.

References

Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021. ISSN 0004-3702.

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

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

Aiello, B., Ishai, Y., and Reingold, O. Priced oblivious transfer: How to sell digital goods. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 119–135. Springer, 2001.

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

Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. True to the model or true to the data?, 2020.

Chen, H., Lundberg, S. M., and Lee, S.-I. Explaining a series of models by propagating shapley values. *Nature Communications*, 13(1):4512, 2022.

Covert, I. C., Lundberg, S., and Lee, S.-I. Explaining by removing: A unified framework for model explanation. *The Journal of Machine Learning Research*, 22(1):9477–9566, 2021.

Cummings, R., Ioannidis, S., and Ligett, K. Truthful linear regression. In Grünwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of the 28th Conference on Learning Theory*, pp. 448–483, Paris, France, 2015.

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

Draxl, C., Clifton, A., Hodge, B.-M., and McCaa, J. The wind integration national dataset (wind) toolkit. *Applied Energy*, 151:355–366, 2015.

Dubey, P., Neyman, A., and Weber, R. J. Value theory without efficiency. *Mathematics of Operations Research*, 6(1):122–128, 1981.

Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., and Feige, I. Shapley explainability on the data manifold, 2020.

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2242–2251, 09–15 Jun 2019.

Han, D., Wooldridge, M., Rogers, A., Ohrimenko, O., and Tschiatschek, S. Replication robust payoff allocation in submodular cooperative games. *IEEE Transactions on Artificial Intelligence*, 4(5):1114–1128, 2023.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in Neural Information Processing Systems*, 33:4778–4789, 2020.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500, 2020.

Lehrer, E. An axiomatization of the banzhaf value. *International Journal of Game Theory*, 17:89–99, 1988.

Liu, J. Absolute shapley value, 2020.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Merrill, J., Ward, G., Kamkar, S., Budzik, J., and Merrill, D. Generalized integrated gradients: A practical method for explaining diverse ensembles, 2019.

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

Owen, A. B. and Prieur, C. On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017.

Pearl, J. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pp. 3–11, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

Pinson, P. Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(4):555–576, 2012.

Pinson, P., Han, L., and Kazempour, J. Regression markets and application to energy forecasting. *TOP*, 30(3):533–573, 2022.

Qazaz, C. S., Williams, C. K., and Bishop, C. M. An upper bound on the bayesian error bars for generalized linear regression. In *Mathematics of Neural Networks: Models, Algorithms and Applications*, pp. 295–299. Springer, 1997.

Shapley, L. S. A value for n-person games. *Classics in Game Theory*, 69, 1997.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International Conference on Machine Learning*, pp. 9269–9278, 2020.

Tastu, J., Pinson, P., Trombe, P.-J., and Madsen, H. Probabilistic forecasts of wind power generation accounting for geographically dispersed information. *IEEE Transactions on Smart Grid*, 5(1):480–489, 2013.

Taufiq, M. F., Blöbaum, P., and Minorics, L. Manifold restricted interventional shapley values. In *International Conference on Artificial Intelligence and Statistics*, pp. 5079–5106. PMLR, 2023.