ONLINE APPENDIX

Machine Data: market and analytics

Giacomo Calzolari, Anatole Cheysson, Riccardo Rovatti
August 12, 2024

B1 Equilibrium existence

In this appendix, we prove the existence of a Nash equilibrium in the case of the free-analytics. The existence in the other cases discussed in the paper follows similar arguments. The proof requires three preliminary Lemmas. In the case of free analytics, producers maximize the following problem:

$$\max_{d_i} \alpha_i y(d) - \gamma_i d_i \tag{53}$$

Lemma 1. *If a data profile d^* is solution to the maximization then $\forall i$, d_i is either such that*

1. $y_i''(d) \leq 0$
2. $d_i = 0$

Proof. Let d_i^* be such that $d_i^* > 0$ and $y_i''(d) > 0$. Then, either:

1. $\alpha_i y_i' d \geq \gamma_i$. Then d_i^* is not a solution because profits increase when increasing data provision. It is beneficial to increase data provision until $\alpha_i y_i'(d) = \gamma_i$, which implies $y_i''(d) \leq 0$ by the properties of $y(d)$.
2. $\alpha_i y_i'(d) < \gamma_i$. Then, profits can be increased by decreasing data provision until $d_i = 0$.

Hence it cannot be that d_i^* be such that $d_i^* > 0$ and $y_i''(d) > 0$. □

Lemma 2. *For each producer i, there are only two candidate best-responses to each data profile d_{-i}, either $d_i = 0$ or d_i $s.t$ $\alpha_i y_i'(d) = \gamma_i$*

Proof. Either the equilibrium lies in the concave space of $y(.)$ or at $d_i = 0$. If the candidate equilibrium lies in the concave space, it should be such that the first derivative of the objective function is equal to 0. □

Denote as $Br_i(d_{-i})$ the positive response of i to d_{-i}. Additionally, denote $d_{-i}^{alt} < d_{-i}$ when each data in d_{-i}^{alt} is at least weakly inferior to those in d_{-i} and one is strictly inferior.

Lemma 3. *Let $\Pi_i(d_i, d_{-i}) = \alpha_i y(d_i, d_{-i}) - \gamma_i d_i$, If*

$$\Pi_i(0, d_{-i}) \geq \Pi_i(Br_i(d_{-i}), d_{-i}) \tag{54}$$

Then, $\forall d_{-i}^{alt} < d_{-i}$ and $\forall d_i$:

$$\Pi_i(0, d_{-i}^{alt}) > \Pi_i(d_i, d_{-i}^{alt}) \tag{55}$$

Proof. Let $\Pi(0, d_{-i}) \geq \Pi(Br_i(d_{-i}), d_{-i})$, then:

$$\alpha_i y(0, d_{-i}) \geq \alpha_i y(Br_i(d_{-i}), d_{-i}) - \gamma_i Br_i(d_{-i}) > \alpha_i y(Br_i(d_{-i}^{alt}), d_{-i}) - \gamma_i Br_i(d_{-i}^{alt}) \iff \tag{56}$$

$$\alpha_i y(0, d_{-i}^{alt}) \geq \alpha_i y(Br_i(d_{-i}^{alt}), d_{-i}) - \gamma_i Br_i(d_{-i}^{alt}) - \alpha_i y(0, d_{-i}) + \alpha_i y(0, d_{-i}^{alt}) \tag{57}$$

Now we show that $\alpha_i y(Br_i(d_{-i}^{alt}), d_{-i}) - \gamma_i Br_i(d_{-i}^{alt}) - \alpha_i y(0, d_{-i}) + \alpha_i y(0, d_{-i}^{alt}) \geq \Pi_i(Br_i(d_{-i}^{alt}, d_{-i}))$:

$$\alpha_i y(Br_i(d_{-i}^{alt}), d_{-i}) - \gamma_i Br_i(d_{-i}^{alt}) - \alpha_i y(0, d_{-i}) + \alpha_i y(0, d_{-i}^{alt}) \geq \Pi_i(Br_i(d_{-i}^{alt}, d_{-i})) \iff$$
$$\int_{d_{-i}^{alt}}^{d_{-i}} \alpha_i \left[\frac{\partial y(Br_i(d_{-i}^{alt}, x)}{\partial x} - \frac{\partial y(0, x)}{\partial x} \right] dx \geq 0$$

Which we know is true thanks to the Scope Property. In other terms moving from d_{-i} to d_{-i}^{alt} implies a stronger reduction of $y(Br_i(d_{-i}^{alt}), d_{-i})$ than of $y(0, d_{-i})$ because of the positive cross-derivative of the function $y(.)$. □

For the continuity of the best response of producers when not providing any data: a reduction in other players' data provisions does not change the best response of a producer not providing any data.

We now combine these results and show that an equilibrium with a free-analytics always exists. Given Lemma 2 and 3, we know that a producer is either providing a level of data such that $y_i'(d) = \frac{\gamma_i}{\alpha_i}$ or is not providing any data. With $I \in \mathbb{N}$ producers, each producer belongs to one of two possible sets in any equilibrium. Set $\mathcal{N}$ is the set of producers not providing any data. Set $\mathcal{J}$ is the set of producers providing a strictly positive amount of data.
Consider the candidate equilibrium d^* such that $\forall i,\ d_i^*$ is such that $y_i'(d) = \frac{\gamma_i}{\alpha_i}$. Then either (i) each producer i is playing its best response, which implies d^* is an equilibrium, (ii) or some producers would be better off providing no data.
If d^* is not an equilibrium, some producers of set $\mathcal{I}$ move to set $\mathcal{N}$. Data provision in set $\mathcal{I}$ are re-adjusted, and d^{**} is the new candidate equilibrium. Then, as before, either (i) each producer i is playing its best response, so that d^{**} is an equilibrium, (ii) or some producers would be better off providing no data. If d^{**} is not an equilibrium, some producers of set $\mathcal{I}$ move to set $\mathcal{N}$. Importantly, by continuity of the best response function when not providing any data (Lemma 4), the movement of producers from set $\mathcal{I}$ to set $\mathcal{N}$ does not change the best response of producers in set $\mathcal{N}$.

This method can be iterated until either (ii) all producers are in set $\mathcal{N}$, (ii) no producer in set $\mathcal{I}$ would be better off not providing any data. After a finite number of iterations, one of these two cases will be reached. In both cases, Lemma 3 ensures that all producers in set $\mathcal{N}$ are playing their best responses. The former case ensures a Nash equilibrium is reached since all players are in set $\mathcal{N}$. In the latter case, as no producer in set $\mathcal{I}$ would be better off not providing data, all producers are playing their best responses.

B2 On the properties of the analytics' value

When discussing properties of Machine Learning tools, the features of Scale and Scope (embedded in our function u discusses in section A1) are usually assumed with a generic reference to common practitioners' experience (e.g., Duch-Brown et al. (2017) and the recent surveys in Computer Science Meng et al. (2020)). However, it isn't easy to find neat accounts of these intuitive properties isn't easy. The problem in developing

a complete theory of this feature is that, nowadays, machine learning models are highly non-linear and result from complicated and expensive training procedures often designed by trial and error. Nevertheless, we provide a direct account of these important properties in this section. We think that, although specific, the analysis in this appendix provides some useful insights.

In the first subsection, we show the Scale property and that the function υ in section A1 is, first, convex, then concave and bounded. We also show that aggregating data and making the resulting analytics available to producers results in a higher utility with respect to a situation in which each producer uses local data to compute local MDA.

In a second subsection, the Scope property so that when multiple producers contribute data with a sufficient *diversity*, the value of the analytics is larger than what can be obtained from the same aggregated amount of data from a single producer. This gives ground to the features of the aggregating operator $\oplus$ in appendix A1.

We also obtain a byproduct from this analysis, which is of value even if we do not directly exploit it in the paper (at least so far). In particular, we define the notion of *complexity*, which is the number of scalar quantities (e.g., sensor readings, configuration settings, etc.) used to characterize every single piece of data and, in other terms, the dimensionality of each data point in the data sets. We show that once the number of features in data is enough to allow classification, increasing the complexity of the data negatively affects the value that one can squeeze out of a given amount of data.

The approach that we use in the following two subsections is to define suitably simplified classification models on which the effect of training can be theoretically anticipated, either exactly or for the worst-case scenarios.

Then, when we want to assess the effect of statistically diverse contributions to the data set, we run Monte Carlo simulations varying the actual data points to see how performance varies with the characteristics of the data set.

We assume that producers' goods come in units and that, due to the fabrication process, there is a certain probability that a unit is defective.

The purpose of the analytics is to identify the good units. It does so by exploiting the fact that the producers have a common technological basis (e.g., manufacturing machines employ the same kind of electrical motor, units are assembled using the same welding process, etc.). Thus, even though the final products may be different, each unit is characterized by the same D numerical features (e.g., sensor readings acquired during production, measurements from final quality inspection, etc.) that we will indicate with $x_1, \ldots, x_D$ and compound into the D-dimensional vector $x \in X \subset [0,1]^D$, where X indicates the whole range that we assume to be uniformly spanned by the production.

The difference between producers is modeled by assuming that the i-th one produces units corresponding to a proper subset $X_i \subset X$.

We assume that the subsets X_i in which the producers generate data are such that $V(X_i) = v$ for some v such that $\Delta = \sqrt[D-1]{v}$ has an integer inverse $1/\Delta$ and

$$X_i = \left[\bigtimes_{k=1}^{D-1} \left[(\xi_{i,k} - 1)\, \Delta, \xi_{i,k} \Delta \right] \right] \times [0,1]$$

for some choice of the $D-1$ integers $1 \leq \xi_{i,1}, \xi_{i,2}, \ldots, \xi_{i,D-1} \leq 1/\Delta$.

To each data point x there is a label $y \in \{-1, +1\}$ whose negative value indicates a defective unit.

We assume that the truth is $y = \operatorname{sgn}(x_D - 1/2)$, thus implicitly considering some change of coordinates used to transform the original data into X in which discriminating between the two classes is trivial.

Trivial as it may be, discrimination must be learned from samples, and thus, we have to define a model with some adjustable parameters and identify how these parameters are learned.

We use a simple 1-neuron piece-wise linear model,

$$y = \mathrm{sgn}\left(x_D - \max_{k=1,\ldots,D-1}\{\alpha_k x_k\} - \max_{k=1,\ldots,D-1}\{\beta_k x_k\} - \frac{1}{2}\right) \tag{58}$$

that mimics the behavior of a *neuron* with excitation and inhibition weights aggregated with a $\max$ instead of a sum, as it has been recently proposed to allow efficient complexity reduction of complex neural networks (Prono et al., 2022a,b).

Once set, the parameters identify a piecewise affine manifold,

$$x_D = g_{\alpha,\beta}(x_2,\ldots,x_{D-1}) = \max_{k=1,\ldots,D-1}\{\alpha_k x_k\} + \max_{k=1,\ldots,D-1}\{\beta_k x_k\} - \frac{1}{2} \tag{59}$$

that separates the points that the model marks as positive (above the manifold) from the points that the model marks as negative (below the manifold). The optimum value for the parameters is $\alpha_k = \beta_k = 0$.

If this is not the case, all the data points such that $1/2 \leq x_D \leq g_{\alpha,\beta}(x_1,\ldots,x_{D-1})$ are false negatives, while all the data points such that $g_{\alpha,\beta}(x_1,\ldots,x_{D-1}) \leq x_D \leq 1/2$ are false negatives.

Given data points in $\hat{X} \subset X$, one may identify two extreme manifolds. One is characterized by the parameters,

$$\begin{aligned} \hat{A}_k &= \min_{x\in\hat{X}\wedge y>0}\left\{\frac{x_D - 1/2}{x_k}\right\} \\ \hat{B}_k &= 0 \end{aligned} \tag{60}$$

and the other by the parameters

$$\begin{aligned} \hat{a}_k &= 0 \\ \hat{b}_k &= \min_{x\in\hat{X}\wedge y<0}\left\{\frac{1/2 - x_D}{x_k}\right\} \end{aligned} \tag{61}$$

Some intuition on our toy setting and the role of the two above manifolds can be obtained from Figure 3. In that Figure $D = 3$ and samples for classification training are marked as points. Blue points have $x_3 > 1/2$ and correspond to working units while red points have $x_3 < 1/2$ and correspond to broken units. Give a trainign set $\hat{X}$ one may identify the yellow manifold $x_3 = g_{\hat{A},\hat{B}}(x_1,x_2)$ and the green manifold $x_3 = g_{\hat{a},\hat{b}}(x_1,x_2)$. These two manifolds have a double significance. First, any other manifold $x_3 = g_{\hat{\alpha},\hat{\beta}}(x_1,x_2)$ produced by our model, no matter how determined by the training algorithm, is such that $g_{\hat{a},\hat{b}}(x_1,x_2) \leq g_{\hat{\alpha},\hat{\beta}}(x_1,x_2) \leq g_{\hat{A},\hat{B}}(x_1,x_2)$. Second, given the training set and the model (58), $\mathrm{sgn}\left(x_3 - g_{\hat{A},\hat{B}}(x_1,x_2)\right)$ is the decision rule yielding the least false-working error rate with the largest false-broken error rate, and $\mathrm{sgn}\left(x_3 - g_{\hat{a},\hat{b}}(x_1,x_2)\right)$ is the decision rule that yields the least false-broken error rate with the largest false-working error rate.

B2.1 Scale Property

Assume that defective units cannot be sold nor repaired and must be identified and discarded, as the cost of selling a potentially defective unit is too high for the producers.

To be on the safe side, each producer will inevitably discard some good units, thus waiving part of its revenues, by adopting a decision rule like $\mathrm{sgn}\left(x_3 - g_{\hat{A},\hat{B}}(x_1,x_2)\right)$ (the yellow one in Figure 3) whose parameters depend on the amount and localization of data from which they are computed.

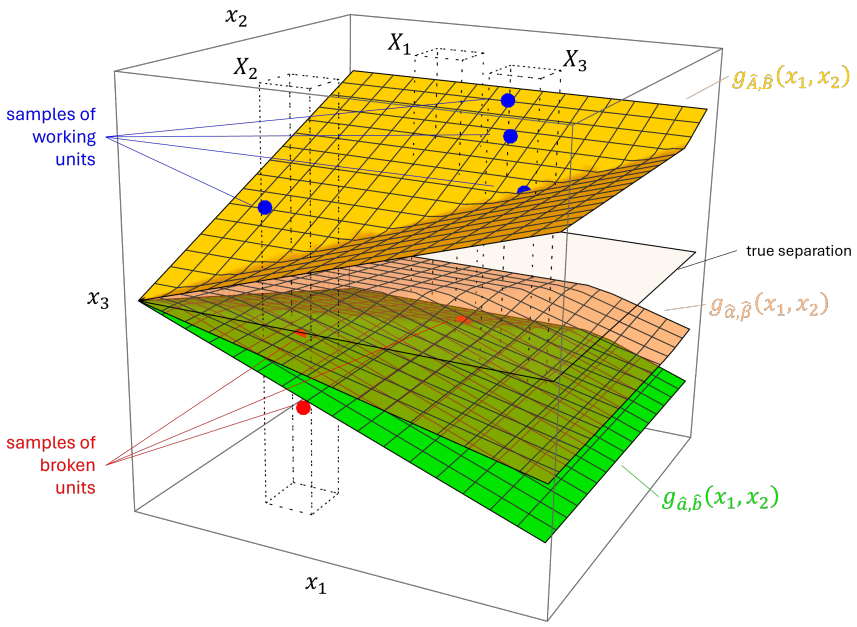


Figure 3: An example of the simplified setting with $D = 3$, $v = 1/64$ and $P = 3$ producers of data. Positive (blue) data points and negative (red) data points determine the separation manifold (in yellow) of the worst-false-broken classifier and the separation manifold (in green) of the worst-false-working classifier. All classifiers that can be obtained from our model have separation manifolds like the pink one, i.e., between the worst-false-working and the worst-false working ones. The ideal separation plane is also shown along with the subregions X_i (dashed parallelepipeds) within which each of the three producers generates its data points.

Assuming that production is uniformly distributed in the feature space, the number of units sold by the i-th producer is proportional to the volume of the convex *revenue polytope* defined by

$$\begin{aligned} 0 \leq \quad x_k \quad &\leq 1 \quad k = 1, \ldots, D-1 \\ \hat{A}_k x_k \leq \quad x_D \quad &\leq 1 \quad k = 1, \ldots, D-1 \end{aligned}$$

that in Figure 3 is the part of the data cube above the yellow manifold, intersected with X_i

The i-th producer may train a model on its own data $\hat{X}_i \subset X_i$, using (60) to compute the model parameters $\hat{A}_k^i$, and obtaining the revenue polytope $\hat{R}_i$. Alternatively, it may use the analytics provided by the aggregator and computed with the data $\hat{X} = \bigcup_{i=1}^{P} \hat{X}_i$, using (60) to compute the model parameters $\hat{A}_k$, and obtaining the revenue polytope $\hat{R}$.

Since (60) takes a minimum over the available data and since $\hat{X}_i \subset \hat{X}$ we have $\hat{A}_k^i \geq \hat{A}_k$ for any i and k and thus $\hat{R}_i \subseteq \hat{R}$. Actual revenues are proportional to volumes $V\left(X_i \cap \hat{R}\right) \geq V\left(X_i \cap \hat{R}_i\right)$ and thus

$$\sum_{i=1}^{P} V\left(X_i \cap \hat{R}\right) \geq \sum_{i=1}^{P} V\left(X_i \cap \hat{R}_i\right) \tag{62}$$

that is equivalent to saying that the value of the analytics based on the aggregated data (left-hand side of (62)) is larger than the sum of the values of the analytics based on separate datasets (right-hand side of (62)).

Consider now a sequence of datasets $\hat{X}^{(t)}$ of increasing size, such that $\hat{X}^{(1)} \subset \hat{X}^{(2)} \subset \cdots \subset X$ the application of (60) yields a corresponding sequence of revenue polytopes $\hat{R}^{(1)} \subseteq \hat{R}^{(2)} \subseteq \cdots \subseteq [0,1] \times [0,1] \times \cdots \times [1/2, 1]$.

Hence, $V\left(X \cap \hat{R}^{(1)}\right) \leq V\left(X \cap \hat{R}^{(2)}\right) \leq \cdots \leq V\left(X \cap [0,1] \times [0,1] \times \cdots \times [1/2, 1]\right)$ but we have $\lim_{t \to \infty} V\left(X \cap \hat{H}^{(t)}\right) = V\left(X \cap [0,1] \times [0,1] \times \cdots \times [1/2, 1]\right)$. Hence, data-dependent revenues are increasing with the number of samples and have an upper bound that is also their limit. This implies that their trend must be asymptotically concave.

All the above shows that global utility increases when producers cooperate, but it is ultimately concave and bounded.

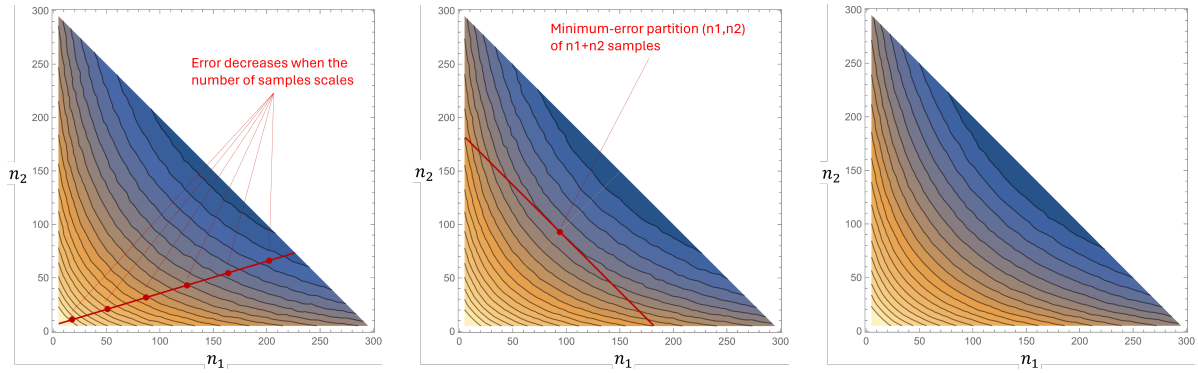
A further piece of information can be obtained by noting that the above calculations are based on the *a priori* availability of a model and its training strategy. Hence, they fail to consider what happens at the beginning of the design of a data-driven application. In real-world applications, the first available data lots are commonly used to set up and tune the ingestion stage (i.e., the data processing pipeline that acquires and transforms raw, incomplete, possibly incoherent data into normalized quantities that can be fed into machine learning blocks), the architecture of the trainable blocks (layers, connections, substructure, etc.) and the training strategy (algorithm, losses, etc.). Accurate, valuable information is obtained from data only after this setup phase is over. Thus, the first data lots have an (apparent) marginal utility much lower than those of data lots that enter a smoothed processing pipeline. This causes the function υ to be convex for small arguments, i.e., when the first data are acquired and used to set up the analytics.

B2.2 Scope Property

The scope property of data utility is more complex as it has to do not only with the quantitative increase of data but also with their distribution in the data space.

This prevents a straightforward analytical derivation like the one exemplifying the scale property in the previous section.

Yet, we may observe the emergence of the scope property in our simple model by Montecarlo simulation.



a) b) c)

Figure 4: Contour plots of the relationship between the contributions n_1 and n_2 of two producers and the (logarithm of the) loss of the toy classifier for $D = 2$ (a), $D = 3$ (b), $D = 4$ (c). The convexity of the contours quantifies the scope effect.

Consider $D = 2, 3, 4$, $n = 10, 15, \ldots, 300$, $v = 1/64$, $P = 4$ producers, and different values for the data contributions n_1, n_2, n_3, n_4.

Different data contributions are obtained by dividing the dataset into $\ell = n/5$ lots of 5 data points each. These lots are then assigned to the P producers considering all possible distinguished partitions of ℓ, i.e., all the set of integers $\ell_1 \geq \cdots \geq \ell_P \geq 0$ such that $\ell_1 + \cdots + \ell_P = \ell$, and then setting $n_i = 5\ell_i$ for $i = 1, \ldots, P$.

For each of the resulting P-tuple, $n_1, \ldots, n_P$, our worst-case classifier's training and performance evaluation is repeated for 10^5 trials. In each trial, P random sets of indexes $0 \leq \xi_{i,1}, \ldots \xi_{i,D-1} \leq M$ identifying X_i are generated. In each X_i, n_i labeled samples are drawn at random. Based on all the generated samples, the worst-false-negative and worst-false-positive classifiers are computed along with their error rate. The largest between the maximum false negative rate and the maximum false positive rate is used to quantify the absolute worst-case performance.

To appreciate changes in an asymptotically saturating utility we reverse our point of view and concentrate on losses, i.e., in terms of volumes, on the volume of the *loss polytope* between the separating manifold of the classifier and the true discriminating plane $x_3 = 0$. Such a volume is $1/2$ minus the volume of the revenue polytope and tends to 0 when performance increases. Hence, its trend can be effectively inspected by logarithmic plots.

Following this principle and to cope with the fact that performance depends on data and thus is itself a random variable, we consider the logarithm of the empirical average over the 10^5 trials of the loss.

Figure 4 is obtained by selecting the P-tuples in which only n_1 and n_2 are positive. This allows us to plot the logarithmic average loss against n_1, n_2 in the $P = 2$ case as a sub-case of the $P = 4$ case.

The scale effect manifests as the fact that any straight line passing through the origin as in Figure 4-a) (along which one sees contributions with a constant ratio n_1/n_2 with increasing size of the overall dataset $N = n_1 + n_2$) intersect iso-performance lines with progressively lower losses.

Yet, the convexity of the same iso-performance lines reveals the effect of scope. Moving along an iso-scale line $n_1 + n_2 = n = \text{constant}$ as in Figure 4-b), the performance consistently improves as one approaches the even distribution of the data set between the two producers $n_1 = n_2 = n/2$.

To assess whether this scope effect holds with $P > 2$ we should agree on how to measure the *evenness* of a partition of n among more than 2 producer. Among the many ways of measuring *evenness*, we choose scaled Shannon entropy, i.e., in the case of P producers,

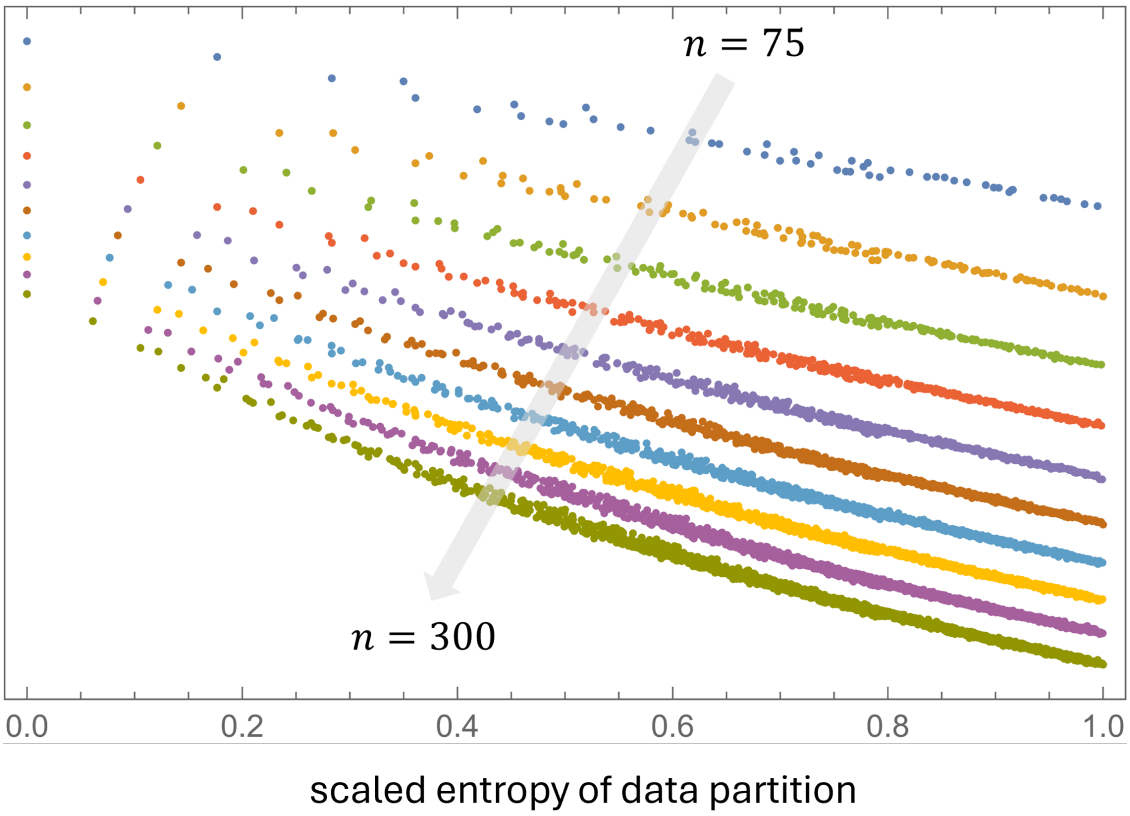


Figure 5: Logarithmic worst false-working performance plotted against the scaled entropy of the distribution of n data points among $P = 4$ producers.

$$E(n_1, \ldots, n_P) = -\frac{1}{\log P} \sum_{i=1}^{P} \frac{n_i}{n} \log \frac{n_i}{n}$$

whatever the basis of the logarithm.

The scaled entropy is minimum for $E(n, 0, \ldots, 0) = 0$ and is is maximum for $E(n/P, \ldots, n/P) = 1$. This is clearly what we want, though the behavior in intermediate configurations depends on the fact that Shannon devised his entropy to quantify the amount of information emitted by a source with P symbols, each with probability n_i/n.

Despite this somehow unrelated origin, scaled entropy seems to interpret quite well the *evenness* on which the scope effect hinges. Figure 5 shows that the logarithmic loss correlates negatively with scaled entropy (and, of course, with n due to the scale effect). Hence, data sets aggregating a substantially equal number of data from each producer yield more utility than equivalent-scale datasets in which most of the data are contributed by a few producers.

Finally, Figure 6 shows the effect of data dimensionality by plotting the logarithm of the worst-case error against the data contribution of $P = 2$ producers working with data of increasing dimensionality $D = 2, 3, 4$.

Note that, given a certain n_1 and n_2 (and thus fixing the effect of scope and scale), as D increases also, the worst-case error increases, showing that higher dimensional models are harder to train.

Hence, the scale property is confirmed. At the same time, the positive effect of aggregating a data set from (possibly evenly contributing) sources exhibiting diversity emerges naturally, as does the effect of

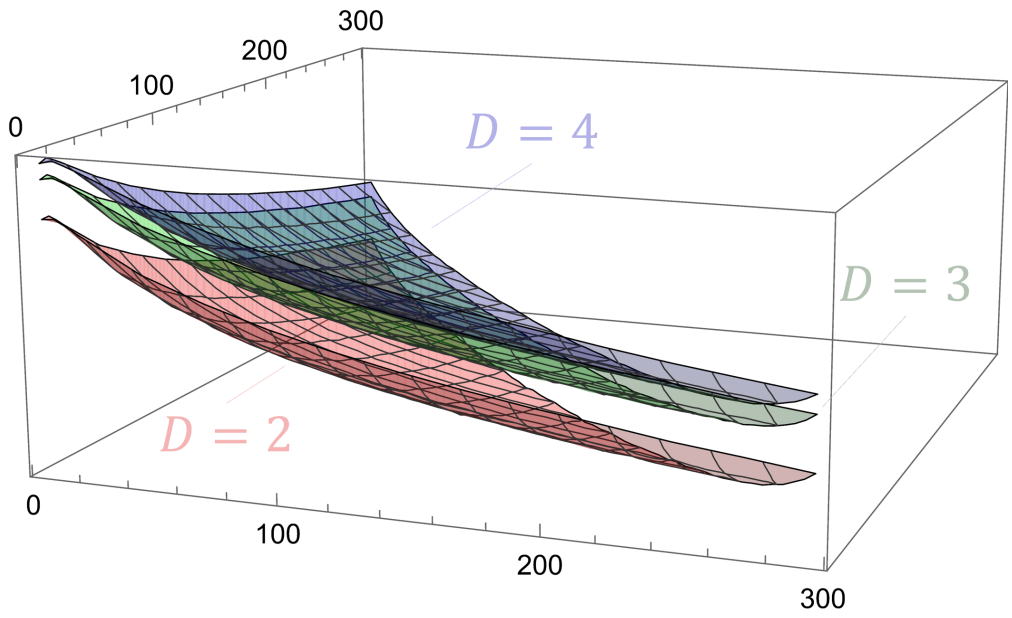


Figure 6: The logarithmic worst false-working performance plotted against the contributions of $P = 2$ producers working with data in a D-dimensional space for $D = 2, 3, 4$.

using a fixed-size data set to train models in high-dimensional settings.

B2.3 Synergy Property

To model Multi-Homing scenarios, we need to define how the indications from more than one independent classifier trained on potentially different data sets can be merged by a producer to reach its final decision on whether to discard a unit or not. To do so, we rely on the concept of *score*, i.e., a pre-decision continuous quantity computed by the classification model and considered a proxy of the likelihood of the outcome.

Almost all real-world models that output a finite valued decision (like the binary one we consider here) generate scores that are then fed into thresholding mechanisms. Given the simplicity of such mechanisms, the intelligence of this architecture is in the stages that compute the scores and ensure properties like '*the higher this score, the more likely is that the best decision is that*'.

In our simple case, the quantity $s = x_D - g_{\alpha,\beta}(x_2, \ldots, x_{D-1})$ is a natural candidate for a score that is then thresholded into the decision $\mathrm{sgn}\,(s)$ (58). In fact, if $s > 0$, then the larger its value, the more likely it is that $x_D > 0$, and thus that the unit is actually working when it is classified so. The reverse is also true as, if $s < 0$, then the smaller its value, the more likely it is that $x_D < 0$, and thus the unit is actually broken when it is classified so.

The synergy between classifiers is modeled by the sum of the corresponding scores, which is then thresholded. Thus, agreeing classifiers (scores of the same signs) reinforce each other while disagreeing classifiers (scores with different signs) compromise a lower-confidence decision.

Summing scores also automatically solves the problem of ties that affects decisions taken starting from an even number of binary classifiers.

Assuming that more than one model is available, the h-th one is described by the discriminating manifold $x_3 = g_{\alpha^{[h]},\beta^{[h]}}(x_1, x_2)$ for $h = 1, \ldots, H$.

If a producer has access to H models, for any new unit can compute the scores $s^{[h]} = x_3 - g_{\alpha^{[h]},\beta^{[h]}}(x_1, x_2)$, the overall score $s = \sum_{h=1}^{H} s^{[h]}$, and the final decision $\mathrm{sgn}\,(s)$.

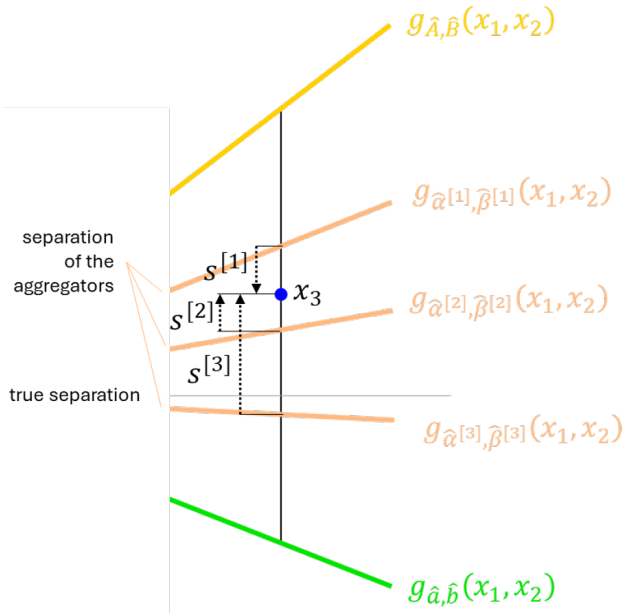


Figure 7: The mechanism to merge the output from 3 classifiers relying on the analytics of independent aggregators working on the same data set. Individual scores are summed to obtain the final score on which the decision is made.

The quality of the result depends both on the data and on the training procedure, as the same data may give rise to different models. In our case, given a data set $\hat{X}$, (59) may feature any manifold function between the lower and upper bound manifold.

The situation is exemplified in Figure 7, which focuses on a given input (x_1, x_2, x_3) when $H = 3$. Given x_1 and x_2, the h-th separation manifold sets $g_{\hat{\alpha}^{[h]},\hat{\beta}^{[h]}}(x_1, x_2)$ between $g_{\hat{a},\hat{b}}(x_1, x_2)$ and $g_{\hat{A},\hat{B}}(x_1, x_2)$.

The scores $s^{[h]}$ are the differences between x_3 and such separation manifolds. Note that the individual classifiers would disagree as $s^{[1]} < 0$ while $s^{[3]} > s^{[2]} > 0$. Yet, the confidence of the wrong classifier $s^{[1]} < 0$ is overcome by the sum of the other two confidences, and the final decision is for a working unit that is correct.

In general, the overall score is $s = \sum_{h=1}^{H} s^{[h]} = \sum_{h=1}^{H} g_{\hat{\alpha}^{[h]},\hat{\beta}^{[h]}}(x_1, x_2) - Hx_3$.

For a given data set, the upper and lower separation manifolds can be computed by (60) and (61). Each new unit to classify fixes x_1 and x_2 and thus the vertical segment in Figure 7, as well as x_3 on the line containing it. If we assume that the actual separation manifolds of the H aggregators are uniformly and independently distributed in the interval $[g_{\hat{a},\hat{b}}(x_1, x_2), g_{\hat{A},\hat{B}}(x_1, x_2)]$, the statistics of the score can be obtained in closed form by scaling and shifting a standard Irwin-Hall (Johnson et al., 1995, Section 26.9) distribution. With this, one may easily compute the false-working probability. Such a probability can be sampled in a Monte Carlo simulation that draws different data sets and units to estimate the average performance.

We provide empirical evidence in two different scenarios, both with $P = 2$ producers. In the first symmetric scenario, there are $H = 1, 2, 3, 4$ aggregators, and both producers use all of them. Each producer decides how much data to share and provides to all the aggregators, receiving H analytics merged in the final classification. In the second asymmetric scenario, there is only one shared aggregator to which the two producers provide data. Yet, the producer providing the largest quantity of data also processes them locally and has access to two analytics: the one coming from the aggregator and its own.

Simulations in the first scenario allow us to confirm that scale, scope, and dimensionality properties

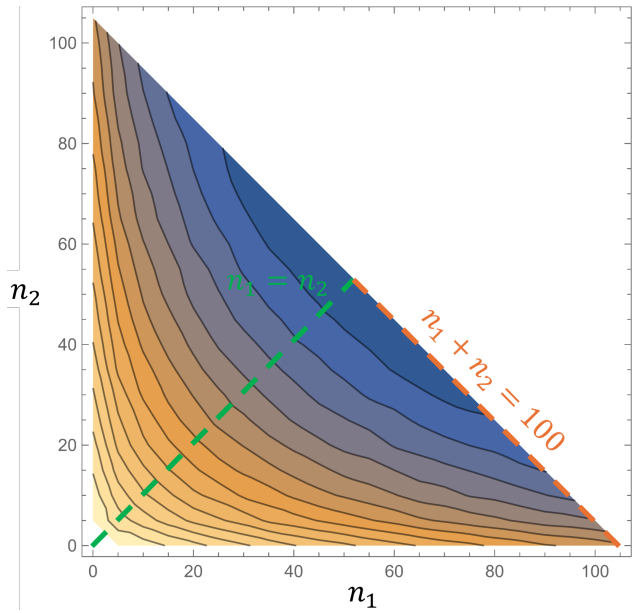


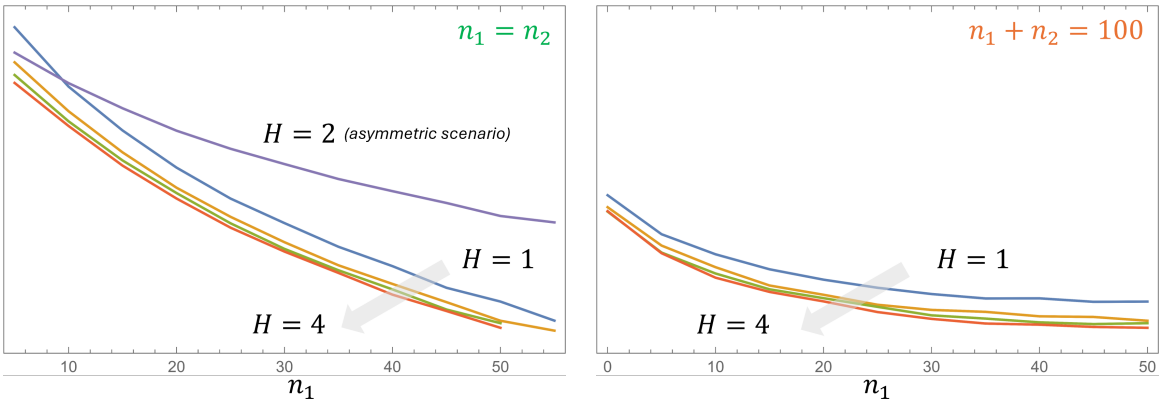
Figure 8: Contour plots of the relationship between the contributions n_1 and n_2 of two producers and the (logarithm of the) loss of the toy classifier for $D = 4$ and $H = 4$.

are maintained even in a multi-homing configuration. For example, Figure 8 reports the iso-performance contours for the most complex case we could address, i.e., $D = 4$ and $H = 3$. The decreasing trend and the convexity of the iso-performance lines we discussed for Figure 4 replicates in this multi-home configuration.

Furthermore, the data for different values of H show that increasing the number of homes can be beneficial, though the marginal benefit rapidly decreases. To see this, consider performance surfaces like the one in Figure 8 for $H = 1, 2, 3, 4$ and plot their section along the dashed lines highlighted in the Figure, i.e., for $n_1 = n_2$ (the best scaling direction) and for $n_1 + n_2 = 100$ (the largest data set). Figure 9 shows the resulting loss profile that lowers as H increases through reducing marginals. Since Figure 9-a) and Figure 9-b) share the same vertical range, the profiles of Figure 9-b) span a smaller loss region corresponding to the largest possible data set.

Let's now consider the second asymmetric scenario. We may plot the corresponding loss in the $n_1 = n_2$ case, which is the most favorable since the aggregator's analytics is computed in the optimal conditions given the total available samples. This yields the topmost line in Figure 9-a) reporting the loss when the aggregator's analytics is merged with the local analytics. Though it is implicitly a $H = 2$-home scenario, the loss of the asymmetric configuration is larger than the symmetric $H = 2$ scenario but also of the symmetric $H = 1$ scenario in which only the aggregator's analytics is used. This is because the local analytics relies on fewer samples (half of the total available) and thus is worse than the aggregator's analytics. When the two are merged, the resulting classifiers are worse than the one relying only on the aggregator's analytics.

The fact that the asymmetric case has $H > 1$ gives some improvement over the $H = 1$ case only for a minimal amount of data for which the synergy effect, even if unbalanced, is larger than the scope and scale effects. The symmetric and asymmetric $H = 2$ converges for $n_1 = n_2 \to 0$ as, in that case, the quality of the aggregator's analytics becomes comparable with the one of the local analytics.


$n_1 = n_2$
$H = 2$ (asymmetric scenario)
$H = 1$
$H = 4$
10
20
30
40
50
n_1
$n_1 + n_2 = 100$
$H = 1$
$H = 4$
0
10
20
30
40
50
n_1

a) b)

Figure 9