Scaling Laws for the Value of Individual Data Points in Machine Learning

Ian Covert [1] [*]   Wenlong Ji [1] [*]   Tatsunori Hashimoto [1]   James Zou [1]

Abstract

Recent works have shown that machine learning models improve at a predictable rate with the total amount of training data, leading to *scaling laws* that describe the relationship between error and dataset size. These scaling laws can help design a model's training dataset, but they typically take an aggregate view of the data by only considering the dataset's size. We introduce a new perspective by investigating scaling behavior for the value of *individual data points*: we find that a data point's contribution to model's performance shrinks predictably with the size of the dataset in a log-linear manner. Interestingly, there is significant variability in the scaling exponent among different data points, indicating that certain points are more valuable in small datasets while others are relatively more useful as a part of large datasets. We provide learning theory to support our scaling law, and we observe empirically that it holds across diverse model classes. We further propose a maximum likelihood estimator and an amortized estimator to efficiently learn the individualized scaling behaviors from a small number of noisy observations per data point. Using our estimators, we provide insights into factors that influence the scaling behavior of different data points. Finally, we demonstrate applications of the individualized scaling laws to data valuation and data subset selection. Overall, our work represents a first step towards understanding and utilizing scaling properties for the value of individual data points.

1. Introduction

Machine learning models for vision and language have improved dramatically in recent years (Radford et al., 2021; Touvron et al., 2023; OpenAI, 2023; Gemini Team et al., 2023), in part due to increasing model sizes, but also due to

using larger amounts of high-quality training data (Gadre et al., 2023). Recent research has found that increasing the amount of training data improves models at a predictable rate, leading to *scaling laws* that describe the relationship between error and dataset size (Kaplan et al., 2020). For example, Hoffmann et al. (2022) introduced the following scaling law to determine Chinchilla's compute-optimal training budget,

$$\text{error} \approx \epsilon + \frac{a}{p^\gamma} + \frac{b}{n^\lambda}, \tag{1}$$

where ϵ represents irreducible error, p is the number of model parameters, n is the number of training examples, and (a, b, ν, λ) are learned constants. These scaling laws help inform how to trade off model and data size, but they take an *aggregate* view of the dataset by not differentiating between training examples; this is limiting because certain data points are more useful than others, particularly for noisy web-scraped datasets.

In the current era of datasets aggregated from heterogeneous sources, it is important to understand how individual data points or data sources affect the behavior of model training. This can give practical guidance on what type of data to prioritize, especially as the dataset grows. It also provides more fundamental insights into how the impact of different data scales with data size—e.g., which points are useful in small datasets but whose value diminishes quickly as the data size increases, and which points are relatively more valuable in large datasets.

Our work aims to take initial steps in this direction: motivated by these questions, we propose *individualized data scaling laws*, which describe how the impact of data scales with the size of the dataset for each training example. Our analysis focuses on a data point's marginal contribution to a trained model's performance, which previous work has noted shrinks with the size of the preceding dataset (Kwon and Zou, 2021); we build on this by showing that the marginal contribution shrinks at a reliable rate, and that this rate varies between data points. The scaling behavior is represented by a simple parametric form, which is inspired by existing aggregate scaling laws (see Section 2). We find support for this phenomenon in certain classical learning settings (e.g., linear regression), and we show that it holds empirically across several model types, including logistic regression, multi-layer perceptrons (MLPs) and support vector

---

[1]Stanford University. Correspondence to: Ian Covert <icovert@stanford.edu>.

machines (SVMs).

Analyzing individual scaling laws can ultimately help improve the dataset, for example by identifying data points that consistently degrade a model's performance, or by surfacing data points whose contributions remain large even as the dataset size grows. Doing so requires fitting scaling parameters for every example in a dataset, which can be computationally costly; for example, it would be intractable to precisely estimate each point's expected marginal contribution at a range of dataset sizes, even though this is a natural way to observe the trend from the scaling law (see Figure 1). We therefore consider how to make the fitting approach efficient in Section 3, where we develop procedures that let us estimate the scaling parameters for entire datasets. These include (1) a statistical approach to estimate the scaling law given a moderate number of sampled marginal contributions, and (2) a neural estimator that amortizes the fitting process across all examples in a dataset.

We conduct experiments to test our approach on several datasets (including tabular, text and image data) and multiple model types. We verify that the scaling law accurately describes the mean contribution at each dataset size: for example, we find that our scaling law explains >0.9 of the variance in marginal contributions between data points (see Table 1). We also test whether it can extrapolate beyond the range where it is fit, which would enable the scaling law to be learned in a less costly dataset range before being used in a larger regime. Finally, we explore applications of the scaling laws to data subset selection, where we find that they identify useful new data points that depend on the current dataset size; and we demonstrate their application to data valuation, which we show is closely related to our scaling analysis (Ghorbani and Zou, 2019).

Our contributions. The main contributions of this work are the following: (1) we propose and find evidence for individualized data scaling laws, (2) we show how to fit the scaling behavior using a small number of noisy observations per data point, (3) we provide qualitative understanding of factors that influence scaling behavior, and (4) we demonstrate that individualized scaling laws can facilitate data valuation and data subset selection. Overall, the scaling behavior of individual data points provides a new tool for understanding and improving training data for machine learning models.

Related work. Scaling laws for deep learning have become well known in recent years (Hestness et al., 2017; Rosenfeld et al., 2019; Kaplan et al., 2020; Hoffmann et al., 2022). They serve several purposes, including reasoning about the trade-offs between increasing training data and model parameters (Kaplan et al., 2020; Hoffmann et al., 2022), predicting the performance of large-scale models (Cherti et al., 2023), and comparing the performance of learning algorithms at manageable scales (Dubois et al., 2023). The most

similar works to ours are Hashimoto (2021) and Rolf et al. (2021), which study how model performance scales when training with multiple data sources; our work instead takes the perspective of studying individual data points, which offers a more granular tool for analyzing the contents of a training dataset.

Separately, another line of work about data valuation focuses on the role of individual data points in improving the model's performance (Ghorbani and Zou, 2019; Kwon and Zou, 2021; Wang and Jia, 2022). These methods typically score training examples based on their *marginal contribution*—how much including them in the training data affects the model's accuracy—by averaging this across many datasets of different sizes. Such methods can be used to identify mislabeled data, filter for high-quality data, upweight helpful examples, and select promising new points for active learning (Ghorbani and Zou, 2019; Kwon and Zou, 2023; Jiang et al., 2023). Our work is similarly framed around each data point's marginal contribution, but we study how the loss improvement scales with the size of the dataset for each training example, and we find that the most helpful training examples can vary with the dataset size.

Aside from these methods focused on each example's marginal contributions, others analyze individual data points using different importance measures (Ilyas et al., 2022; Park et al., 2023; Just et al., 2023). In particular, many works have explored extensions of the classic influence function to nonlinear models like neural networks (Cook and Weisberg, 1980; Koh and Liang, 2017; Grosse et al., 2023; Kwon et al., 2023). Ilyas et al. (2022) briefly discuss the effect of training dataset size on example-to-example influence scores, but to our knowledge, our work is the first to thoroughly explore scaling behavior for individual data points.

2. Individualized data scaling laws

We focus here on supervised learning problems, and as a notational setup, we write individual data points as tuples $z = (x, y)$ where x is an input and y the response variable. We write datasets as $\mathcal{D} = \{z_i\}_{i=1}^k$, and the datasets can have variable sizes that we denote by $|\mathcal{D}|$. When creating datasets of different sizes, we assume access to a single large training pool $\mathcal{D}_t$ from which the smaller datasets $\mathcal{D}$ can be sampled without replacement. We consider a learning algorithm that trains a model on a given dataset $\mathcal{D}$, which we denote by $f_{\mathcal{D}}$, and we consider several options for the model class (e.g., logistic regression, MLPs). Our analysis focuses on the population error $\mathcal{L}(f_{\mathcal{D}})$, for example the cross-entropy loss, and we only consider the effects of the training data (i.e., we do not vary the number of model parameters).

Recall the form of current aggregate scaling laws: if we ignore the number of model parameters, several works are based on the same functional form $\mathcal{L}(f_{\mathcal{D}}) \approx \epsilon + b|\mathcal{D}|^{-\lambda}$

(Rosenfeld et al., 2019; Kaplan et al., 2020; Hashimoto, 2021; Hoffmann et al., 2022). We can view this as a claim about the expected loss given a dataset of size k, where datasets are sampled uniformly at random:

$$\mathbb{E}_{|\mathcal{D}|=k}[\mathcal{L}(f_{\mathcal{D}})] \approx \epsilon + \frac{b}{k^{\lambda}}. \qquad (2)$$

The functional form in Eq. (2) takes an aggregate view by only focusing on the dataset size k, but our goal is to analyze how the loss is affected by individual data points. We therefore focus on each data point's marginal contribution, which is defined as the performance difference between models trained with and without z. We write the marginal contribution $\Delta(z, \mathcal{D})$ as follows, where we expect $\Delta(z, \mathcal{D}) > 0$ in most cases for useful data points:

$$\Delta(z, \mathcal{D}) = \mathcal{L}(f_{\mathcal{D}}) - \mathcal{L}(f_{\mathcal{D} \cup \{z\}}). \qquad (3)$$

This quantity is the basis of many data valuation methods, (Ghorbani and Zou, 2019; Ghorbani et al., 2020; Kwon and Zou, 2021; Wang and Jia, 2022), and its expectation across z is reflected via the derivative of current scaling laws like Eq. (2).[1] Our goal here is to understand how marginal contributions scale as a function of the dataset size $|\mathcal{D}|$. We expect them to shrink monotonically towards zero, but at a rate that may be specific to each data point. We therefore posit the following scaling law for the expectation $\psi_k(z) \equiv \mathbb{E}_{|\mathcal{D}|=k}[\Delta(z, \mathcal{D})]$ across datasets of size k:

$$\psi_k(z) \approx \frac{c(z)}{k^{\alpha(z)}}. \qquad (4)$$

This shares a similar form with aggregate scaling laws, but there is no irreducible term because the contribution should shrink to zero, and we allow both the numerator $c(z)$ and exponent $\alpha(z)$ to depend on the data point. In the remainder of this section, we show that this functional form holds in practice for several model classes (Section 2.1), and we examine support for this parametric form from classical learning settings (Section 2.2).

2.1. Empirical validation of the scaling law

To test if the scaling law in Eq. (4) holds in practice, we first run a simple experiment to visualize the hypothesized behavior. Our approach is to select a range of dataset sizes k and estimate each expectation $\psi_k(z)$ by averaging a large number of sampled contributions $\Delta(z, \mathcal{D})$. We can then plot the results to observe the following linear trend in the log-transformed means and cardinalities k,

$$\log |\psi_k(z)| \approx \log |c(z)| - \alpha(z) \log(k), \qquad (5)$$

where we use the absolute values $|\psi_k(z)|$ and $|c(z)|$ because certain examples have negative contributions. We conduct

---

[1]If the aggregate scaling law $\mathbb{E}_{|\mathcal{D}|=k}[\mathcal{L}(f_{\mathcal{D}})] \approx \epsilon + \frac{b}{k^{\lambda}}$ holds, then we have $\mathbb{E}_{z,|\mathcal{D}|=k}[\Delta(z, \mathcal{D})] \approx \frac{b\lambda}{k^{\lambda+1}}$.
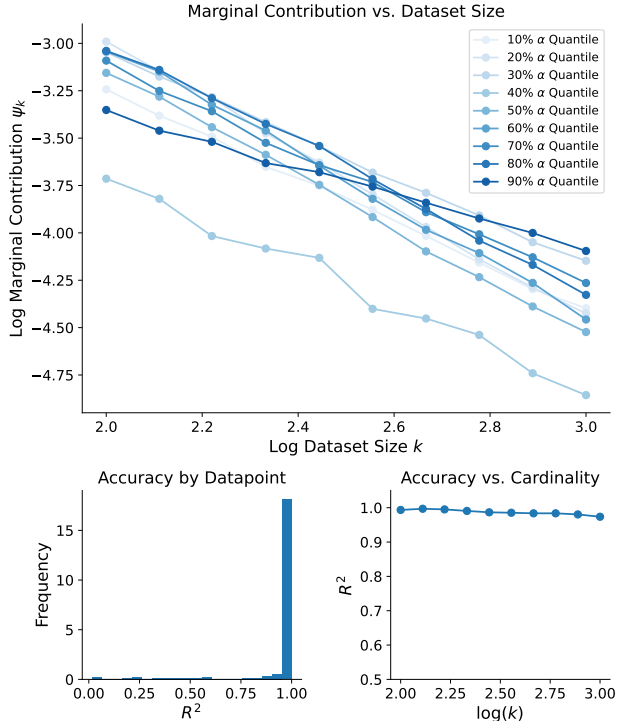

Figure 1. **Individualized scaling laws for logistic regression trained on the IMDB dataset**. **Top:** Marginal contribution vs. the dataset size in log-scale for several data points with a range of scaling exponents $\alpha(z)$. **Left:** Histogram of R^2 scores for linear trend lines fit to each data point in the log-scale. **Right:** Plot of the R^2 score from our scaling law predictions at each cardinality, measured across data points. We achieve an overall $R^2 = 0.987$ for the predictions across all points and dataset sizes.

this experiment for 1000 data points from the IMDB movie review dataset (Maas et al., 2011), where we use 10 log-spaced cardinalities between $k = 100$ and $k = 1000$, and we average 1000 samples $\Delta(z, \mathcal{D})$ to estimate each $\psi_k(z)$.

Table 1. **Overall R^2 score of predicting expected contributions**. The overall R^2 score is measured for linear trends fit in log-space after estimating $\psi_k(z)$ for a range of k values (the same estimation approach used in Figure 1), and it is calculated across data points and cardinalities.

	Logistic Regression	MLP	SVM
MiniBooNE	0.993	0.963	0.929
CIFAR10	0.944	0.942	0.963
IMDB	0.985	0.956	0.946

Figure 1 shows results for logistic regression models trained on frozen BERT embeddings (Devlin et al., 2018). The slope of each line represents the scaling exponent $\alpha(z)$, and the intercept represents the coefficient magnitude $\log |c(z)|$. We observe variability in the estimated slopes, and Figure 1 shows several points corresponding to different quantiles of

the fitted scaling exponent $\alpha(z)$.

To test the scaling law's accuracy, we calculate the goodness-of-fit for each data point's linear trend via the R^2 score, which we fit with ordinary least squares. Figure 1 shows a histogram of these scores, which confirms that the majority of points follow clean log-linear trends. A minority of points do not, but we find that these outliers generally have very small contributions $\log |\psi_k(z)|$ across k values, which may indicate numerical or optimization noise as the loss differences approach zero (see Figure 10). To account for the variability between points, we also calculate the R^2 from the linear trends within each cardinality, where we see that the score is ≈ 1.0 across k values (Figure 1 bottom right). Overall, many data points are accurately described by log-linear trends, but we observe some deviations as k grows and the loss differences shrink towards zero; such issues may be due to noise, or cases where the scaling law does not perfectly fit the data, but it is accurate enough in most cases to provide useful insights.

These observations provide initial evidence for the scaling behavior hypothesized in Eq. (4) and motivate the remainder of our analysis. Because our fitting approach in Figure 1 requires a large number of samples for each z, we develop a more efficient approach to fit the scaling law in Section 3. In our more comprehensive experiments in Section 4, we perform similar analysis of the scaling law's fit using multiple datasets and model classes, and we also discuss implications for data subset selection and data valuation. As a preview, Table 1 shows that the scaling law achieves a high overall R^2 score for three datasets and model classes.

2.2. Theoretical support for the scaling law

We mathematically analyze linear regression and more general M-estimators to provide theoretical support for the existence of the scaling law and give insight into factors that can affect the scaling parameters. The theorems and details are provided in Appendix A, but we summarize the results here. Overall, we find general confirmation for a reliable decay rate in the marginal contributions, and we identify intuitive factors that influence the magnitude of the contribution: for example, for linear regression we find that the relative noise of the response variable is relevant, as is the input vector's leverage score. However, our theory does not make precise predictions about the scaling behavior, because it requires strong regularity conditions and has error terms whose expectations may be non-negligible. These limitations may explain why our empirical results show heterogeneity in the scaling exponent $\alpha(z)$, although it is often concentrated in a range from $[1, 1.5]$ that is consistent with the loose predictions from our theoretical results. We remark that such partial mismatches between theory and practice are also widely observed with aggregate scaling laws (Kaplan et al.,

2020; Hutter, 2021).

3. Efficient scaling law estimation

A key challenge in using individualized scaling laws is fitting them to every example in a dataset. Recall that Section 2.1 used a large number of samples $\Delta(z, \mathcal{D})$ at each cardinality to estimate the expectation $\psi_k(z)$. This lets us visualize a linear trend in the log-log space, but the procedure is intractable to repeat for every example in a large dataset. To address this issue, we now describe two approaches to more efficiently estimate individualized scaling laws.

3.1. Maximum likelihood-based estimation

Our goal is to fit the scaling parameters $c(z), \alpha(z)$ using a limited number of marginal contribution samples $\Delta(z, \mathcal{D})$, which is challenging because the scaling law in Eq. (4) is defined via the expectation $\psi_k(z)$ at each cardinality. We find that there is significant variability in the marginal contributions at each k; in fact, it is common to observe a mix of positive and negative contributions for a single example z (see Figure 2). This means that whereas our procedure in Section 2 relied on noiseless log-transformed mean contributions $\psi_k(z)$, a more efficient procedure must tolerate a noisy view of the scaling trend.

We therefore propose fitting the scaling law in the original loss space, and in a way that carefully handles noise in the sampled contributions, which tends to be largest at smaller cardinalities k. We frame this as a statistical inference problem by modeling the marginal contributions at each k with the following Gaussian distribution:

$$\Delta(z, \mathcal{D}) \mid |\mathcal{D}| = k \sim \mathcal{N}\left(\frac{c(z)}{k^{\alpha(z)}}, \frac{\sigma^2(z)}{k^{\beta(z)}} \right). \quad (6)$$

Notice that the mean for each k is governed by our scaling law from Eq. (4), and the variance is governed by a separate scaling law with parameters $\sigma(z)$ and $\beta(z)$. The variance in marginal contributions also naturally shrinks with k (Kwon and Zou, 2021), and we observe that the variance follows this general scaling behavior (see Figure 11 in Appendix C.1). Fitting the variance parameter $\sigma(z)$ and exponent $\beta(z) > 0$ lets us capture the natural reduction in variability as k grows, and it encourages the learned scaling law to tolerate larger deviations for small k values. This is essential for correctly modeling the mean-related parameters $c(z), \alpha(z)$, but we do not otherwise use the variance-related parameters.[2]

Based on this statistical model, we can fit the parameters by obtaining a set of samples $\Delta(z, \mathcal{D}_i)$ for $i = 1, \ldots, m$ and

---

[2]When fitting their scaling law, Hoffmann et al. (2022) handled outliers by using a Huber loss and downweighting points in the low-compute regime. Directly modeling the variance is an alternative approach that we have not seen in the scaling law literature.
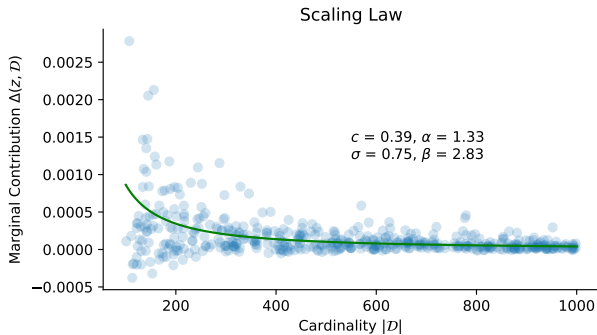
Figure 2. **Likelihood-based scaling law estimator.** The estimator is fit for a single example from the IMDB dataset, and the scaling parameters are fit using $m = 100$ samples.

then minimizing the negative log-likelihood as follows:

$$\underset{c,\alpha,\sigma,\beta}{\arg\min} \ \frac{1}{m} \sum_{i=1}^{m} \text{NLL}(\Delta(z, \mathcal{D}_i), |\mathcal{D}_i|; c, \alpha, \sigma, \beta). \quad (7)$$

The optimization problem is non-convex, but we find that it can be effectively optimized with Adam (Kingma and Ba, 2014) using a moderate number of gradient steps; our exact procedure is described in Appendix B, and it involves using analytic solutions for c, σ to only take gradient steps on α, β. An example of the fitting result is shown in Figure 2, where we see that the learned curve correctly allows larger deviations at smaller k values and fits more precisely when k is large. Our experiments use up to $m = 1000$ samples per data point for the most accurate results, which is an order of magnitude fewer samples than the validation in Section 2 (which used 1000 samples *per cardinality*). Our experiments also test the estimator's convergence and show strong results with as few as $m = 100$ samples (see Section 4), which provides a $100\times$ speedup.

3.2. Amortized scaling law estimation

The likelihood-based approach is a more efficient way to fit individualized scaling laws, but the number of samples m required for each example z is still moderately large (e.g., $m = 100$). The key issue is that each data point is treated independently, and therefore requires its own set of samples $\Delta(z, \mathcal{D})$ to fit the scaling curve. We now consider an alternative approach where we share information across examples z to use fewer samples per data point, and in the process jointly fit the scaling parameters $c(z), \alpha(z)$ for an entire dataset.

Our approach is to replace the objective from Eq. (7), which optimizes over the scaling parameters for a single z, with a combined objective parameterized by a neural network. We use a network $g(z; \theta) \in \mathbb{R}^4$ that predicts the four parameters for the Gaussian distribution in Eq. (6): the network must receive only the data point $z = (x, y)$ as an input, which we implement by passing x to a standard architecture (e.g.,

a MLP), predicting scaling parameters for all classes, and using only those for the relevant class y. We then train the network using the following loss function:

$$\min_{\theta} \ \mathbb{E}_{z, \mathcal{D}}\Big[\text{NLL}\big(\Delta(z, \mathcal{D}), |\mathcal{D}|; g(z; \theta)\big)\Big]. \quad (8)$$

Like the previous objective, this is optimized using a set of samples $\Delta(z, \mathcal{D})$, but this version is equivalent to solving Eq. (7) simultaneously for all data points with parameters predicted by the network $g(z; \theta)$. This general approach is known as *amortized optimization* and is used for a range of machine learning tasks (Amos et al., 2023), but not previously for estimating scaling laws. We can in principle train the network with as few as $m = 1$ contributions per data point, but in our experiments we construct a dataset of relatively few samples $\Delta(z, \mathcal{D})$ per data point (e.g., $m = 10$), and we train until the loss stops improving on a held-out validation set.

4. Experiments

4.1. Evidence for individualized data scaling laws

We begin our more comprehensive experiments by providing further empirical evidence to support the parametric scaling law in Eq. (4).[3] We focus on three model classes: logistic regression, SVMs and MLPs (specifically, two-layer ReLU networks). We evaluate these models on three different datasets: MiniBooNE (Roe et al., 2005), CIFAR-10 (Krizhevsky et al., 2009), and IMDB movie reviews (Maas et al., 2011). To accelerate training and avoid underfitting, we use frozen ResNet-50 (He et al., 2016) and BERT (Devlin et al., 2018) pre-trained embeddings for CIFAR-10 and IMDB, respectively. Full details for the experiments are provided in Appendix C.1.

For each setting mentioned above, we evaluate the model performance by computing the cross-entropy loss on a test dataset of size 1000. Similar to our analysis in Section 2.1, we first test the accuracy of log-linear trendlines by averaging many samples $\Delta(z, \mathcal{D})$ at a range of log-spaced cardinalities. For logistic regression, we use 1000 data points and 1000 samples per k value. For SVMs and MLPs, due to the higher variance in marginal contributions, we use 200 data points and 5000 samples per dataset size k.

Following Figure 1, we show the R^2 from lines fit to each point, and we observe across settings that the R^2 scores are generally very close to 1. Figure 3 shows these results via histograms for each setting. The mass is most concentrated towards 1 for logistic regression, whereas MLPs and SVMs have more points with sub-optimal fits. We attribute this to worse estimation of the mean at each cardinality, and a similar issue mentioned in Section 2.1 with instability as the

---

[3]Code for our experiments is available at https://github.com/iancovert/data-scaling
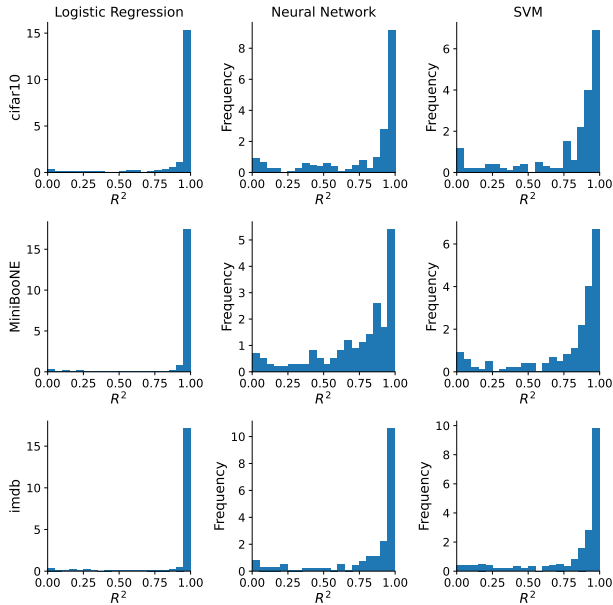
5

Figure 3. **Histogram of R^2 score when fitting the scaling law.** Similar to Figure 1, we find that linear trends in log-space achieve high R^2 scores for most data points, which supports the parametric form in Eq. (4).

Figure 4. **Histogram of estimated $\alpha(z)$.** The estimated values have a mode between 1 and 1.5 and exhibit significant heterogeneity. We exclude points with $R^2 < 0.8$ to ensure the estimated $\alpha(z)$ values are reliably estimated.

loss differences approach zero. The R^2 across data points and cardinalities is reported in Table 1, which reveals that the scaling law overall captures nearly all the variability in expected contributions at each dataset size, despite the existence of some poorly fitting data points.

Next, we examine the distribution of estimated parameters $\alpha(z)$. Analysis of $c(z)$ is deferred to Appendix C. Histograms of the fitted values are shown in Figure 4, where we see that the distribution of $\alpha(z)$ has a mode between 1 and 1.5, which coincides with the range seen in our theoretical results (see Appendix A). The heterogeneity in $\alpha(z)$ is important, because it suggests that different data points have different decay rates for their marginal contributions. A data point with a higher $\alpha(z)$ value may have a larger contribution in a small-sample regime compared with other data points, but due to its fast decay rate, it could have a much smaller contribution in a large-sample regime. We illustrate this phenomenon in Figure 5: for each setting, we select the points with $\alpha(z)$ in 20%, 40%, 60%, and 80% quantiles among all points being evaluated, and we show their marginal contribution against the size of the preceding dataset. The results illustrate that there exist many lines that cross one another, which supports the idea that the most valuable data points depend on the dataset size. In Section 4.4, we test this insight in experiments where we add points to an existing dataset based the scaling law's predictions at different k values.

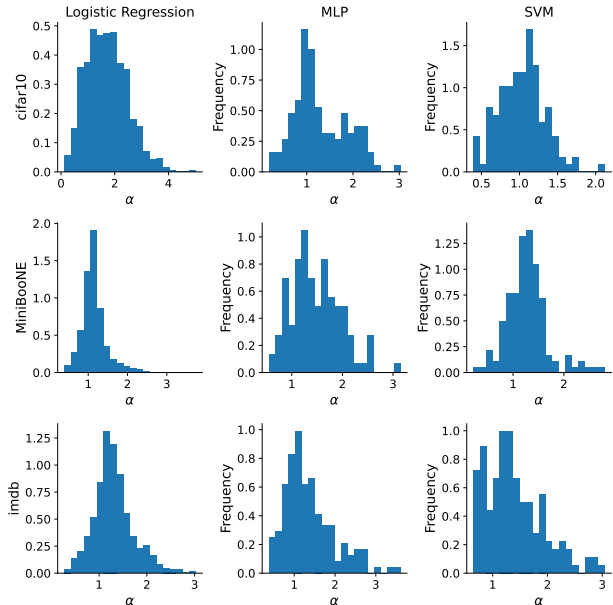A natural question to ask is why different points have dif-

ferent scaling exponents $\alpha(z)$. A comprehensive answer to this question is beyond the scope of our work, but we provide some preliminary ideas. In Figure 6, we show the relationship between the estimated $\alpha(z)$ and the distance to the decision boundary for logistic regression models with two binary classification datasets. There is a strong correlation between $\alpha(z)$ and the distance to the decision boundary: closer points have lower $\alpha(z)$ and further points have higher $\alpha(z)$. A possible explanation is that when the size of the preceding dataset is small, every point will have a significant contribution to learning the decision boundary in a high-dimensional space. However, when the dataset is large, the logistic regression model may have an accurate decision boundary, and only nearby points contribute by helping better calibrate the magnitude of the normal vector; these points may be those with a slower decay rate. However, we also remark that $\alpha(z)$ is not completely determined by this factor, and that there is significant heterogeneity very near the decision boundary, so there are other factors at play. We leave a more thorough analysis of this behavior as a future research direction.

Finally, we also examine the similarity between estimated scaling parameters for different model types. Table 2 shows the Pearson correlation between $\alpha(z)$ estimates for each pair of models, and we observe a positive correlation for two out of three datasets, CIFAR-10 and IMDB, but relatively weak correlations for MiniBooNE. The degree of correlation between models is similar for the estimated $c(z)$ values. The
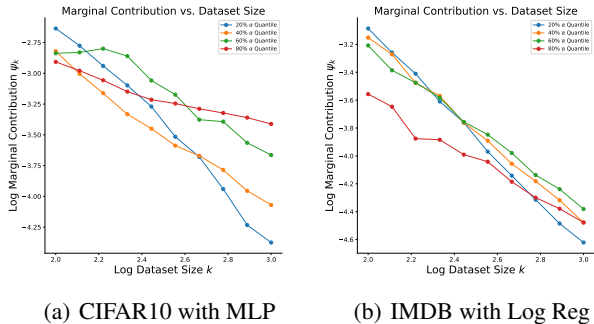
(a) CIFAR10 with MLP   (b) IMDB with Log Reg

Figure 5. **Marginal contribution for points with different** $\alpha(z)$. Similar to Figure 1, we plot the expected contribution $\log |\psi_k(z)|$ against the dataset size $\log k$. Lines with different $\alpha(z)$ cross one another, indicating that the ranking of valuable points depends on the dataset size k.

deviations between models may be attributed to the differing geometry of each function class, where different points can be more influential depending on the shape of the decision boundary. Again, we leave a more complete characterization of this finding to future work.

Table 2. **Correlation between estimated** $\alpha(z)$ **for different models.** The Pearson correlation between estimated scaling exponents shows weak positive similarity for CIFAR-10 and IMDB.

	LogReg/MLP	LogReg/SVM	SVM/MLP
MiniBooNE	-0.01	0.18	-0.04
CIFAR-10	0.38	0.26	0.34
IMDB	0.49	0.50	0.27

4.2. Scaling law estimation accuracy

We now move to testing our more efficient scaling law estimators from Section 3. In evaluating our proposed approaches, we use metrics based on how accurately we predict the marginal contributions at each dataset size. For example, Figure 14 shows that for IMDB with logistic regression, we can accurately predict the expectation $\psi_k(z)$ at multiple dataset sizes from $k = 100$ to $k = 1000$. For a more systematic evaluation, Figure 7 shows the accuracy of our scaling law predictions across dataset sizes for both versions of our likelihood-based estimator, and when fitting with different numbers of samples. Our scaling laws are all fit in the range $k \in [100, 1000]$, so we mark the boundary where our scaling law begins extrapolating to larger ranges. An expanded version of these results is shown in Appendix C.2, where we find that the R^2 score drops when we extrapolate beyond $k = 2500$, but the correlation and rank correlation with the true expectation remain high.

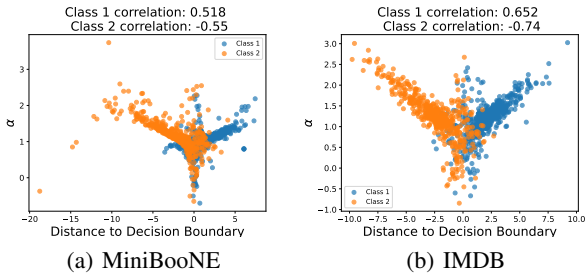The likelihood-based estimator improves as we fit with more

(a) MiniBooNE   (b) IMDB

Figure 6. **Distance to decision boundary vs. scaling exponent for logistic regression model.** We fit a logistic regression model on all points being evaluated and then compute the distance of each example to the decision boundary. The sign of distance is kept to distinguish points from each class. There is a strong correlation between distances and the estimated α values.
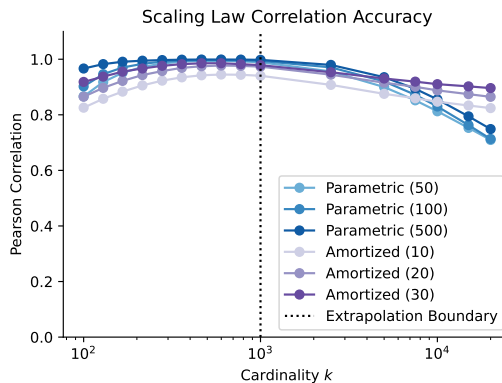

Figure 7. **Scaling law estimation accuracy at multiple cardinalities for IMDB with logistic regression.** We evaluate the scaling law's predictions at cardinalities up to an order of magnitude larger than where it is fit, according to the Pearson correlation with the true expectation $\psi_k(z)$ for each k. The numbers in parentheses are the number of samples used to fit the estimator.

samples, and it provides the highest accuracy across all k value when fit with 500 samples. The amortized estimator is relatively accurate given the small number of samples. The accuracy of all the estimators are consistently high in the fitting range $k \in [100, 1000]$, although the performance suffers slightly at the dataset sizes with the noisiest samples. Interestingly, we see that as we extrapolate to datasets an order of magnitude larger than our fitting range, the amortized estimator tends to be more accurate than the parametric estimator even though it is fit with many fewer samples.

4.3. Application to data valuation

Next, we consider the application of our scaling laws to data valuation. Data valuation is the most direct application of our scaling analysis, because these methods are typically based on a weighted average of the expectations $\psi_k(z)$ at different k values (Ghorbani and Zou, 2019; Kwon and Zou,

Table 3. **Data valuation accuracy with 10 samples per data point.** We report the Pearson correlation of the estimates with ground truth values. Note that the non-amortized likelihood estimator can be unreliable with insufficient samples.

Method	MiniBooNE	CIFAR-10	IMDB
Monte Carlo	0.967	0.857	0.916
Likelihood	0.858	0.077	0.342
Amortized	0.994	0.893	0.930

Table 4. **Data valuation accuracy with 50 samples per data point.** We report the Pearson correlation of the estimates with ground truth values. All the estimators are quite accurate.

Method	MiniBooNE	CIFAR-10	IMDB
Monte Carlo	0.988	0.965	0.981
Likelihood	0.995	0.936	0.985
Amortized	0.996	0.931	0.986

2021; Wang and Jia, 2022), which means we can estimate the scores using our scaling parameters $c(z), \alpha(z)$. In fact, our approach can be viewed as a version of a stratified data valuation estimator (Wu et al., 2023) with the mean at each k estimated via the shared set of scaling parameters. Among the various available methods, we focus on Distributional Data Shapley (Ghorbani et al., 2020), which uses a uniform average across cardinalities. We measure the estimation accuracy relative to ground truth values computed using a large number of sampled marginal contributions, and we compare the accuracy to a standard Monte Carlo estimator (see a more thorough discussion in Appendix C.3). Our goal is to show that our estimators converge to the same value as the conventional estimator, which attests to the scaling law accurately capturing the decay in marginal contributions; and we hope to observe that one or both of our estimators can outperform the standard one when using a small number of marginal contribution samples.

Our full results are in Appendix C.3, but a summary is shown in Table 3 and Table 4 with results calculated for estimates using 10 and 50 samples, respectively. In general, we observe that the amortized estimator achieves the best Pearson correlation with the ground truth valuation scores when using just 10 samples; this reflects its improved sample complexity from sharing information across data points. In contrast, the parametric estimator cannot fit reliably with so few samples. The three estimators tend to perform similarly with larger numbers of samples, with the exception of CIFAR-10, where the Monte Carlo estimator works better. Overall, this validates that our scaling law accurately captures the decay in marginal contributions with k, and offers an alternative approach to calculate data valuation scores.

Table 5. **Accuracy improvement (%) with 20 points added to preceding datasets of size 100.** Each element represents the improvement in test set accuracy with and without the selected points. The last row reports the model's average accuracy before adding new points.

Method	MiniBooNE	CIFAR-10	IMDB
Scaling 1000	0.21 ± 1.03	1.99 ± 0.97	0.88 ± 0.96
Scaling 100	**0.47 ± 1.02**	**3.28 ± 1.09**	**2.22 ± 0.86**
Shapley	0.28 ± 1.02	2.06 ± 0.92	0.77 ± 1.20
Random	0.44 ± 0.91	2.01 ± 1.09	1.00 ± 0.85
Preceding	80.76 ± 1.58	75.50 ± 1.97	78.77 ± 1.64

Table 6. **Accuracy improvement (%) with 20 points added to preceding datasets of size 1000.** The settings are the same as Table 5.

Method	MiniBooNE	CIFAR-10	IMDB
Scaling 1000	0.20 ± 0.39	0.17 ± 0.14	**0.13 ± 0.10**
Scaling 100	0.21 ± 0.34	0.05 ± 0.06	0.06 ± 0.05
Shapley	**0.22 ± 0.36**	**0.18 ± 0.14**	0.11 ± 0.09
Random	0.03 ± 0.20	0.04 ± 0.13	0.03 ± 0.10
Preceding	83.92 ± 0.46	85.38 ± 0.40	85.45 ± 0.35

4.4. Application to point addition

We now apply our scaling law to select high-value examples to add to an existing dataset. Based on our observation in Figure 5, the lines for different points can cross one another, indicating that the relative rankings of data points may vary depending on the size of the dataset. This suggests that the size of the preceding dataset should be taken into account when evaluating which point is most helpful to add. We therefore design a set of experiments inspired by Kwon and Zou (2021) where we compare the addition of high-value data points selected by different methods.

Here, we consider selecting 20 points from a population of 1000 to be added to preceding datasets with two different sizes: 100 and 1000. Each preceding dataset is randomly sampled from the whole training dataset, excluding the pool from which we select the new points, which is also randomly sampled. We use two baseline approaches to select points, random selections from the pool and Distributional Shapley values (Ghorbani et al., 2020) fit with a large number of marginal contribution samples. For our approach, we fit the scaling parameters using the likelihood estimator from Section 3.1 with 500 samples, and we use the parameters to estimate the expected contribution $\psi_k(z) \approx c(z)/k^{\alpha(z)}$. Because we consider two initial dataset sizes, we test our scaling law with two different k values, $k = 100$ and $k = 1000$, which we refer to as *Scaling 100* and *Scaling 1000*.

The results are shown in Tables 5 and 6, where we repeat the experiment 1000 times to report the mean and standard de-

viation in accuracy improvement. According to the results, Scaling 100 achieves the best performance when $k = 100$, as it selects based on the predicted contribution for this dataset size. However, the same set of points performs badly for $k = 1000$, indicating that points with high contribution in the small-sample regime and large-sample regime are different. Similarly, we observe that Scaling 1000 performs well when $k = 1000$, but badly when $k = 100$. Interestingly, we observe that the performance of Shapley and Scaling 1000 are very similar, implying that the Shapley values are closely correlated with contributions in the large sample regime, and may be unreliable when evaluating performance in a small-sample regime. In Appendix C.4, we conduct a similar experiment with larger dataset sizes, which demonstrates the same trend.

In conclusion, selecting new training examples using the scaling law predictions at different dataset sizes can lead to quite different results, and selecting points based on the current dataset size can help more reliably achieve better performance. This provides further support for our analysis of scaling behavior for individual data points, but we caution that a more powerful data selection method should take into account not just the size but the contents of the current dataset; our approach only quantifies the expected improvement when a data point is added to a *random* dataset, and this is a limitation to its current usage for dataset selection.

5. Discussion

This work studies scaling laws for the value of individual data points, and we found evidence for a simple parametric form that holds across datasets and model classes. Our experiments validated the scaling law by visualizing the log-linear trend and testing the scaling law's ability to predict marginal contributions at different dataset sizes, and we found that the scaling law can extrapolate to larger dataset sizes than the range where it is fit. The behavior we study is costly to measure for an entire training dataset, so we also developed procedures to estimate the scaling parameters from a small number of noisy observations per data point.

Our study is a first step towards understanding this phenomenon, and it has certain limitations and suggests several future directions. Fitting individual scaling laws remains somewhat expensive, as we require separate marginal contribution samples for each point; it may be possible to borrow insights from recent data valuation estimators that avoid calculating marginal contributions for each point and rely on a single pool of shared models (Feldman and Zhang, 2020; Wang and Jia, 2022; Li and Yu, 2023). Another limitation is that our method is not specifically designed for dataset selection, and selecting many highly ranked points may lead to negative interaction effects that our scaling law does not take into account; we speculate that our study of scaling

behavior could also be useful for predicting contributions to a specific dataset. Our analysis is restricted to the effect of individual points on a model's error, whereas other recent works have considered how training examples impact specific predictions (Ilyas et al., 2022; Park et al., 2023; Grosse et al., 2023); a follow-up direction is to analyze whether similar scaling behavior holds for these datapoint-to-datapoint relationships. Our experiments additionally focused on relatively simple models and small-scale datasets, so an important direction for future work is studying the same phenomenon for large-scale deep learning models. Finally, we believe that similar techniques that we use here can also be used to study how the value of a fixed *subset* of data scales as the rest of the dataset increases; this would be a natural interpolation of individualized and aggregate scaling.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

References

Amos, B. et al. (2023). Tutorial on amortized optimization. *Foundations and Trends® in Machine Learning*, 16(5):592–732.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Cook, R. D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dubois, Y., Hashimoto, T., and Liang, P. (2023). Evaluating self-supervised learning via risk decomposition. *arXiv preprint arXiv:2302.03068*.

Feldman, V. and Zhang, C. (2020). What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. (2023). Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ghorbani, A., Kim, M., and Zou, J. (2020). A distributional framework for data valuation. In *International Conference on Machine Learning*, pages 3535–3544. PMLR.

Ghorbani, A. and Zou, J. (2019). Data Shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR.

Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., et al. (2023). Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Hashimoto, T. (2021). Model performance scaling with multiple data sources. In *International Conference on Machine Learning*, pages 4107–4116. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hutter, M. (2021). Learning curve theory. *arXiv preprint arXiv:2102.04074*.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., and Madry, A. (2022). Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*.

Jiang, K. F., Liang, W., Zou, J., and Kwon, Y. (2023). OpenDataVal: a unified benchmark for data valuation. *arXiv preprint arXiv:2306.10577*.

Just, H. A., Kang, F., Wang, J. T., Zeng, Y., Ko, M., Jin, M., and Jia, R. (2023). Lava: Data valuation without pre-specified learning algorithms. *arXiv preprint arXiv:2305.00054*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kuchibhotla, A. K. (2018). Deterministic inequalities for smooth M-estimators. *arXiv preprint arXiv:1809.05172*.

Kwon, Y., Wu, E., Wu, K., and Zou, J. (2023). DataInf: Efficiently estimating data influence in lora-tuned llms and diffusion models. *arXiv preprint arXiv:2310.00902*.

Kwon, Y. and Zou, J. (2021). Beta Shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*.

Kwon, Y. and Zou, J. (2023). Data-OOB: Out-of-bag estimate as a simple and efficient data value. *arXiv preprint arXiv:2304.07718*.

Li, W. and Yu, Y. (2023). Robust data valuation with weighted banzhaf values. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

OpenAI (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. (2023). TRAK: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., and Mc-Gregor, G. (2005). Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2-3):577–584.

Rolf, E., Worledge, T. T., Recht, B., and Jordan, M. (2021). Representation matters: Assessing the importance of subgroup allocations in training data. In *International Conference on Machine Learning*, pages 9040–9051. PMLR.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. (2019). A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wang, T. and Jia, R. (2022). Data Banzhaf: A data valuation framework with maximal robustness to learning stochasticity. *arXiv preprint arXiv:2205.15466*.

Wu, M., Jia, R., Lin, C., Huang, W., and Chang, X. (2023). Variance reduced Shapley value estimation for trustworthy data valuation. *Computers & Operations Research*, page 106305.

A. Theoretical support for the individualized scaling law

We now consider theoretical perspectives on our scaling law for some simple models and data distributions. We aim to provide preliminary mathematical support for the existence of the scaling law and give insight into factors that can affect the scaling parameters. A precise characterization requires reasoning about the expectation over $\Delta(z, \mathcal{D})$ for datasets of a fixed size, which we find is a significant technical obstacle. We instead characterize the limiting behavior of $\Delta(z, \mathcal{D})$, which provides some general confirmation of the scaling behavior we observe in practice, and we leave a more rigorous analysis of $\psi_k(z)$ as a future direction.

We begin by considering the case of linear regression, where datasets $\mathcal{D} = \{z_i\}_{i=1}^k$ are generated from the following distribution,

$$(\mathcal{P}) : y_i = x_i^\top \beta^* + \epsilon_i, \, x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma), \, \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

with $\Sigma \in \mathbb{R}^{d \times d}$ assumed to be positive definite. Denote the data point to be added as $z = (x, y)$, the preceding data matrix as $\mathbf{X}_\mathcal{D} = (x_1, \ldots, x_k)^\top \in \mathbb{R}^{n \times d}$, and the ordinary least squares estimator corresponding to $\mathcal{D}$ and $\mathcal{D} \cup \{z\}$ as $\hat{\beta}_\mathcal{D}$ and $\hat{\beta}_{\mathcal{D} \cup \{z\}}$. The loss function is defined as the mean squared error on the population,

$$\mathcal{L}(\beta) = \mathbb{E}_{(x', y') \sim \mathcal{P}}[(y' - \beta^\top x')^2],$$

and the marginal contribution can be written as $\Delta(z, \mathcal{D}) = \mathcal{L}(\hat{\beta}_\mathcal{D}) - \mathcal{L}(\hat{\beta}_{\mathcal{D} \cup \{z\}})$. With this setup, we can characterize the marginal contribution of z as follows.

Theorem A.1. *If we denote the noise in z as $\epsilon = y - x^\top \beta^*$, we have the following expectation with respect to the labels conditioned on the preceding dataset $\mathbf{X}_\mathcal{D}$:*

$$\mathbb{E}[\Delta(z, \mathcal{D}) \mid z, \mathbf{X}_\mathcal{D}] = \frac{2\sigma^2 - \epsilon^2}{k^2} x^\top \Sigma^{-1} x + o_P\left(\frac{1}{k^2}\right).$$

The proof is deferred to Appendix A.1. According to Theorem A.1, marginal contributions are expected to follow the scaling law with a scaling exponent $\alpha(z) = 2$ in the large sample regime. Meanwhile, the coefficient $c(z)$ depends on two factors: the first term $2\sigma^2 - \epsilon^2$ represents the relative noise level of z, where less noise results in a higher contribution, and the threshold for positive/negative contribution is twice the population noise level σ^2. The second term $x^\top \Sigma^{-1} x$ represents the asymptotic leverage score of the point, where a point with a higher score is more influential, and hence amplifies the magnitude of the contribution. Here we only analyze the expectation conditioned on the preceding dataset to give initial intuition about why the scaling law may exist, as the expectation over the preceding dataset is technically hard to analyze. Even if the leading term on the right-hand side does not depend on $\mathbf{X}_\mathcal{D}$, the $o_P(k^{-2})$ term may have a non-negligible expectation. We leave a more complete characterization as a future direction.

Next, we consider the case of M-estimators, a more general family of models that minimize an empirical loss function, including robust least squares and GLMs like logistic regression. Our model in this case has parameters $\theta \in \mathbb{R}^d$, and we consider a per-example loss $\ell(\theta; x, y)$ with a metric $\mathcal{L}(\theta)$ on the model performance (typically an expectation $\mathcal{L}(\theta)$ over a population distribution $\mathcal{Q}$). We assume that the preceding dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^k$ is sampled i.i.d. from $\mathcal{Q}$, and we consider the following M-estimation problem:

$$\hat{\theta}_\mathcal{D} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^k \ell(\theta; x_i, y_i). \tag{9}$$

Given a new training example $z = (x, y)$, we also define the corresponding problem with one additional loss term:

$$\hat{\theta}_{\mathcal{D} \cup \{z\}} = \arg\min_{\theta \in \mathbb{R}^d} \left(\ell(\theta; x, y) + \sum_{i=1}^k \ell(\theta; x_i, y_i) \right). \tag{10}$$

Our goal is to evaluate the difference in model performance from including the single data point z, and the marginal contribution is $\Delta(z, \mathcal{D}) = \mathcal{L}(\hat{\theta}_\mathcal{D}) - \mathcal{L}(\hat{\theta}_{\mathcal{D} \cup \{z\}})$. To characterize this term we define the population minimizer as

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}_{(x', y') \sim \mathcal{Q}}[\ell(\theta; x', y')],$$

and we denote the population Hessian matrix at θ^* as

$$V_{\theta^*} = \nabla^2 \mathbb{E}_{(x',y')\sim\mathcal{Q}}[\ell(\theta^*; x', y')].$$

Based on these terms, we can characterize the marginal contribution as follows.

Theorem A.2. *(Informal)* *Under regularity conditions on $\ell, \mathcal{L}$ and $\mathcal{Q}$,*

(1) if $\nabla\mathcal{L}(\theta^*) \neq 0$, the marginal contribution satisfies

$$\Delta(z, \mathcal{D}) = \frac{1}{k}\nabla\mathcal{L}(\theta^*)^\top V_{\theta^*}^{-1}\nabla\ell(\theta^*; x, y) + O_P\left(\frac{1}{k^{3/2}}\right),$$

(2) if $\nabla\mathcal{L}(\theta^*) = 0$, then $\sqrt{k}\nabla\mathcal{L}(\hat{\theta}_\mathcal{D}) = O_P(1)$ and

$$\Delta(z, \mathcal{D}) = \frac{1}{k^{3/2}}(\sqrt{k}\nabla\mathcal{L}(\hat{\theta}_\mathcal{D}))^\top V_{\theta^*}^{-1}\nabla\ell(\theta^*; x, y) + O_P\left(\frac{1}{k^2}\right).$$

A formal statement of the regularity conditions and the proof are deferred to Appendix A.2. In principle, the model performance should be the loss function evaluated on the population level, i.e., $\mathcal{L}(\theta) = \mathbb{E}_{(x',y')\sim\mathcal{Q}}[\ell(\theta; x', y')]$. In this case, the optimality of θ^* implies $\nabla\mathcal{L}(\theta^*) = 0$. However, the underlying distribution $\mathcal{Q}$ is often intractable, so a common surrogate is the average loss on a held-out validation dataset $\{(x'_j, y'_j)\}_{j=1}^m$, i.e., $\hat{\mathcal{L}}(\theta) = \frac{1}{m}\sum_{j=1}^m \ell(\theta; x'_j, y'_j)$. The minimizer of this finite-sample metric is generally different from the population minimizer θ^*, and we therefore have $\nabla\hat{\mathcal{L}}(\theta^*) \neq 0$. According to Theorem A.2, the asymptotic rate can differ under these two settings. In practice, the test set typically has a comparable size to the training set, hence we should expect the empirical behavior to be mixed. We also remark that even if $\sqrt{k}\nabla\mathcal{L}(\hat{\theta}_\mathcal{D}) = O_P(1)$, its expectation could be non-negligible and the leading term could be $O_P(k^{-2})$ in the second case. This occurs in the linear regression model according to Theorem A.1, but for general M-estimators, it is not clear what the order of this term is after taking the expectation across $\mathcal{D}$.

Overall, our theoretical analysis provides support for a scaling exponent $\alpha(z)$ in the marginal contribution with respect to the dataset size, and some examples of factors that affect the coefficient $c(z)$. However, we emphasize that our theoretical analysis is limited, as it requires strong regularity assumptions on both the models and data distributions. Moreover, our results are limited to single marginal contributions $\Delta(z, \mathcal{D})$, and the expectation of the O_P terms may not be negligible even in the limit. Therefore, these results do not make precise predictions about the exact parameters for the scaling law in practice, which also involves finite sample bias from both training and evaluation. These factors explain why our empirical results show heterogeneity in the scaling parameter $\alpha(z)$, although it is often concentrated in a range from $[1, 1.5]$ that is consistent with the loose predictions from our theoretical results. We remark that such disagreements between theory and practice are also widely observed with aggregate scaling laws (Hutter, 2021).

A.1. Proof of Theorem A.1

Proof. In the following analysis we denote $\mathbf{X}_\mathcal{D} = (X_1, \cdots, X_n)^T \in \mathbb{R}^{n\times p}, \mathbf{y}_\mathcal{D} = (y_1, \cdots, y_n)^T \in \mathbb{R}^n, \epsilon_\mathcal{D} = (\epsilon_1, \cdots, \epsilon_n)^T \in \mathbb{R}^n$ for convenience. Notice that:

$$\mathcal{L}(\beta) = \mathbb{E}_{(x',y')\sim\mathcal{P}}[(x'^T\beta^\star + \epsilon - x'^T\beta)^2] = \sigma^2 + (\beta - \beta^\star)^T\Sigma(\beta - \beta^\star),$$

$$\hat{\beta}_\mathcal{D} - \beta^\star = (\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^T\epsilon_\mathcal{D}, \quad \hat{\beta}_{\mathcal{D}\cup\{n+1\}} - \beta^\star = (\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}(\mathbf{X}_\mathcal{D}^T\epsilon_\mathcal{D} + x\epsilon)$$

Then, in expectation over $\epsilon_\mathcal{D}$ (i.e., expectation conditional on $\mathbf{X}_\mathcal{D}$ and z), we have

$$\begin{aligned}
\mathbb{E}[\Delta(z, \mathcal{D})|z, \mathbf{X}_\mathcal{D}] =& \mathbb{E}_{\epsilon_\mathcal{D}}[\epsilon_\mathcal{D}^T\mathbf{X}_\mathcal{D}(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D})^{-1}\Sigma(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^T\epsilon_\mathcal{D}] \\
& - \mathbb{E}_{\epsilon_\mathcal{D}}[(\mathbf{X}_\mathcal{D}^T\epsilon_\mathcal{D} + x\epsilon)^T(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}\Sigma(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}(\mathbf{X}_\mathcal{D}^T\epsilon_\mathcal{D} + x\epsilon)] \\
=& \sigma^2 \operatorname{tr}(\mathbf{X}_\mathcal{D}(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D})^{-1}\Sigma(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D})^{-1}\mathbf{X}_\mathcal{D}^T) \\
& - \epsilon^2 x^T(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}\Sigma(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}x \\
& - \sigma^2 \operatorname{tr}(\mathbf{X}_\mathcal{D}(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}\Sigma(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}\mathbf{X}_\mathcal{D}^T) \\
=& \sigma^2 \operatorname{tr}((\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D})^{-1}\Sigma - (\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}\Sigma) \\
& - (\epsilon^2 - \sigma^2)x^T(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}\Sigma(\mathbf{X}_\mathcal{D}^T\mathbf{X}_\mathcal{D} + xx^T)^{-1}x
\end{aligned}$$

Applying the Sherman–Morrison formula, we know that

$$(\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}} + xx^T)^{-1} = (\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}})^{-1} - \frac{(\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}})^{-1} xx^T (\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}})^{-1}}{1 + x^T (\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}})^{-1} x}.$$

By the law of large numbers we have $\frac{1}{k} \mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}} \xrightarrow{p} \Sigma$, and therefore by the continuous mapping theorem we have $k(\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}})^{-1} \xrightarrow{p} \Sigma^{-1}$ and $k(\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}} + xx^T)^{-1} \xrightarrow{p} \Sigma^{-1}$. Hence,

$$
\begin{aligned}
\mathbb{E}[\Delta(z, \mathcal{D}) | z, \mathbf{X}_{\mathcal{D}}] =& \frac{\sigma^2 x^T (\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}})^{-1} \Sigma (\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}})^{-1} x}{1 + x^T (\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}})^{-1} x} - (\epsilon^2 - \sigma^2) x^T (\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}} + xx^T)^{-1} \Sigma (\mathbf{X}_{\mathcal{D}}^T \mathbf{X}_{\mathcal{D}} + xx^T)^{-1} x \\
=& \frac{1}{k^2} \frac{\sigma^2 x^T \Sigma^{-1} \Sigma \Sigma^{-1} x + o_P(1)}{1 + o_P(1)} - \frac{1}{k^2} \left((\epsilon^2 - \sigma^2) x^T \Sigma^{-1} \Sigma \Sigma^{-1} x + o_P(1) \right) \\
=& \frac{2\sigma^2 - \epsilon^2}{k^2} x^T \Sigma^{-1} x + o_P \left(\frac{1}{k^2} \right),
\end{aligned}
$$

which finishes the proof. $\qquad\square$

A.2. Proof of Theorem A.2

In this section, we prove the individualized scaling law for M-estimators. For convenience, we index the point to be evaluated as (x_{k+1}, y_{k+1}), and write $\hat{\theta}_{k+1} = \hat{\theta}_{\mathcal{D} \cup \{z\}}, \hat{\theta}_k = \hat{\theta}_{\mathcal{D}}$ in the following analysis. For this general model family, we cannot obtain a closed-form solution as before, and we will focus on the asymptotics instead. A standard result on the asymptotic distribution of M-estimators is the following.

Lemma A.3 (Asymptotic normality of M-estimators). *Assume that the loss function $\ell(\theta; x, y)$ is twice differentiable in θ such that, for every θ_1, θ_2 in a neighborhood of θ^* and a measurable function $M(x, y)$ with $\mathbb{E}_{\mathcal{Q}}[M(x, y)] \leq \infty$,*

$$\|\ell(\theta_1; x, y) - \ell(\theta_2; x, y)\| \leq M(x, y) \|\theta_1 - \theta_2\|.$$

$$\|\nabla \ell(\theta_1; x, y) - \nabla \ell(\theta_2; x, y)\| \leq M(x, y) \|\theta_1 - \theta_2\|.$$

Also assume that $\mathbb{E}_{\mathcal{Q}}[\|\nabla \ell(\theta^*; x, y)\|_2^2] \leq \infty$, $\theta \to \mathbb{E}_{\mathcal{Q}} \nabla \ell(\theta; x, y)$ is differentiable at θ^* with nonsingular derivative matrix V_{θ^*}. If $\hat{\theta}_k \xrightarrow{P} \theta^*$ and $k^{-1} \sum_{i=1}^k \nabla \ell(\hat{\theta}_k; x_i, y_i) = o_p(k^{-1/2})$, then $\sqrt{k}(\hat{\theta}_{\mathcal{D}} - \theta^*)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta^*}^{-1} \mathbb{E}_{\mathcal{Q}}[\nabla \ell(\theta^*; x, y) \nabla \ell(\theta^*; x, y)^T] V_{\theta^*}^{-1}$.

The proof can be found in Van der Vaart (2000) (see Theorem 5.21 therein). Define the empirical Hessian matrix evaluated on the empirical minimizers as

$$\hat{H}_{k+1} := \frac{1}{k+1} \sum_{i=1}^{k+1} \nabla^2 \ell \left(\hat{\theta}_{k+1}; x_i, y_i \right).$$

Also, define for $1 \leq i \leq k+1$,

$$\delta_{i, k+1} := \frac{(k+1)^{-1} \left\| \hat{H}_{k+1}^{-1} \nabla \ell \left(\hat{\theta}_{k+1}; x_i, y_i \right) \right\|_2}{1 - (k+1)^{-1} \left\| \hat{H}_{k+1}^{-1} \nabla^2 \ell \left(\hat{\theta}_{k+1}; x_i, y_i \right) \right\|_{op}}.$$

The main technique we use to prove Theorem A.2 is the following inequality on leave-one-out M-estimators from Kuchibhotla (2018).

Lemma A.4 (Corollary 7.1 from Kuchibhotla 2018). *Assume that $\ell(\theta; x, y)$ is convex and twice differentiable in θ for every (x, y). If $\delta_{i, k+1} \geq 0$ for all $1 \leq i \leq k+1$ and*

$$\max_{1 \leq i \neq j \leq n} C \left(\frac{3}{2} \delta_{i, k+1}; x_j, y_j \right) \leq \frac{4}{3}$$

then

$$\|\hat{\theta}_k - \hat{\theta}_{k+1} - (k+1)^{-1} \hat{H}_{k+1}^{-1} \nabla \ell(\hat{\theta}_{k+1}; x_{k+1}, y_{k+1})\|_2$$

$$\leq \frac{3\delta_{k+1, k+1}}{2} \left[\max_{1 \leq i \leq k} C \left(\frac{3}{2} \delta_{k+1, k+1}; x_i, y_i \right) - 1 + (k+1)^{-1} \left\| \hat{H}_{k+1}^{-1} \nabla^2 \ell(\hat{\theta}_{k+1}; x_{k+1}, y_{k+1}) \right\|_{op} \right].$$

A few more assumptions we need to analyze the marginal contribution are:

Assumption A.5. The metric $\mathcal{L}(\theta)$ is twice differentiable, and $\|\nabla\mathcal{L}(\theta)\|_2$, $\|\nabla^2\mathcal{L}(\theta)\|_{op}$ are bounded by a universal constant $C_1 > 0$.

Assumption A.6. For $u \geq 0$, the function

$$C(u; x, y) = \sup_{\|\theta_1 - \theta_2\|_2 \leq u} \sup_{\|e\|_2 = 1} \frac{e^T \nabla^2 \ell(\theta_1; x, y) e}{e^T \nabla^2 \ell(\theta_2; x, y) e} \tag{11}$$

is differentiable with its derivative bounded by a universal constant C_2 at a neighborhood of $u = 0$ for any (x, y).

Now we are ready to state and prove the formal version of Theorem A.2.

Theorem A.7 (Formal version of Theorem A.2). *Under the conditions of Lemma A.3 and Lemma A.4, assume that Assumption A.5, Assumption A.6 hold, and $\mathbb{E}_{\mathcal{Q}}[\|\nabla^2\ell(\theta; x, y)\|] \leq \infty$, then:*

(1) if $\nabla\mathcal{L}(\theta^*) \neq 0$, the marginal contribution satisfies

$$\Delta(z, \mathcal{D}) = \frac{1}{k}\nabla\mathcal{L}(\hat{\theta}^*)^T V_{\theta^*}^{-1} \nabla\ell(\theta^*; x, y) + O_P(\frac{1}{k^{3/2}}),$$

(2) if $\nabla\mathcal{L}(\theta^*) = 0$, then $\sqrt{k}\nabla\mathcal{L}(\hat{\theta}_k) = O_P(1)$ and

$$\Delta(z, \mathcal{D}) = \frac{1}{k^{3/2}}(\sqrt{k}\nabla\mathcal{L}(\hat{\theta}_k))^T V_{\theta^*}^{-1} \nabla\ell(\theta^*; x, y) + O_P(\frac{1}{k^2}),$$

Proof of Theorem A.2. Applying Taylor's theorem to the marginal contribution, we have

$$\Delta(z, \mathcal{D}) = \mathcal{L}(\hat{\theta}_k) - \mathcal{L}(\hat{\theta}_{k+1}) = \nabla\mathcal{L}(\theta_k)^T(\hat{\theta}_k - \hat{\theta}_{k+1}) + \frac{1}{2}(\hat{\theta}_k - \hat{\theta}_{k+1})^T \nabla^2\mathcal{L}(\xi)(\hat{\theta}_k - \hat{\theta}_{k+1}), \tag{12}$$

where $\xi = \alpha\hat{\theta}_k + (1 - \alpha)\hat{\theta}_{k+1}$ for some $\alpha \in [0, 1]$. Then notice that

$$\|\hat{H}_{k+1} - \frac{1}{k+1}\sum_{i=1}^{k+1}\nabla^2\ell(\theta^*; x_i, y_i)\|_{op} \leq \frac{1}{k+1}\sum_{i=1}^{k+1} M(x_i, y_i)\|(\hat{\theta}_{k+1} - \theta^*)\|_2,$$

from Lemma A.3, we know that $\hat{\theta}_{k+1} - \theta_2^* = O_P(k^{-1/2})$, and by law of large number, $\frac{1}{k+1}\sum_{i=1}^{k+1} M(x_i, y_i) = O_P(1)$, hence $\hat{H}_{k+1} - \frac{1}{k+1}\sum_{i=1}^{k+1}\nabla^2\ell(\theta^*; x_i, y_i) = O_P(k^{-1/2})$. On the other hand, from the central limit theorem,

$$\frac{1}{k+1}\sum_{i=1}^{k+1}\nabla^2\ell(\theta^*; x_i, y_i) - \mathbb{E}\nabla^2\ell(\theta^*; x_i, y_i) = O_P(k^{-1/2}).$$

Use the fact that $\mathbb{E}\nabla^2\ell(\theta^*; x_i, y_i) = \nabla\mathbb{E}\nabla\ell(\theta^*; x_i, y_i) = V_{\theta^*}$, we conclude that $\hat{H}_{k+1} - V_{\theta^*} = O_P(k^{-1/2})$ and hence

$$\hat{H}_{k+1}^{-1} - V_{\theta^*}^{-1} = O_P(k^{-1/2}) \tag{13}$$

with the fact that V_{θ^*} is nonsingular.

Moreover, since $\hat{\theta}_{k+1} - \theta^* = O_P(k^{-1/2})$, and $\nabla\ell, \nabla^2\ell$ are continuous and Lipscitz, we know that

$$\nabla\ell(\hat{\theta}_{k+1}; x_{k+1}, y_{k+1}) - \nabla\ell(\theta^*; x_{k+1}, y_{k+1}) = O_P(k^{-1/2}), \tag{14}$$

and

$$\nabla^2\ell(\hat{\theta}_{k+1}; x_{k+1}, y_{k+1}) - \nabla^2\ell(\theta^*; x_{k+1}, y_{k+1}) = O_P(k^{-1/2}). \tag{15}$$

Based on (13)-(15), we shall conclude that $\delta_{k+1,k+1} = O_P(k^{-1})$, and then use Assumption A.6,

$$\max_{1 \leq i \leq k} C\left(\frac{3}{2}\delta_{k+1,k+1}; x_i, y_i\right) - 1 = O_P(k^{-1}). \tag{16}$$

Combining (13)-(16) with Lemma A.4, we can conclude that

$$\hat{\theta}_k - \hat{\theta}_{k+1} = \frac{1}{k+1} V_{\theta^*}^{-1} \nabla \ell(\theta^*; x_{k+1}, y_{k+1}) + O_P(k^{-3/2}).$$

Then plug it into (12), we know that

$$\Delta(z, \mathcal{D}) = \frac{1}{k+1} \nabla \mathcal{L}(\hat{\theta}_k)^T (V_{\theta^*}^{-1} \nabla \ell(\theta^*; x_{k+1}, y_{k+1}) + O_P(k^{-3/2})) + O_P(k^{-2}). \tag{17}$$

Applying the Delta method to $\nabla \mathcal{L}(\hat{\theta}_k)$, the limiting distribution of $\nabla \mathcal{L}(\hat{\theta}_k)$ is

$$\sqrt{k}(\nabla \mathcal{L}(\hat{\theta}_k) - \nabla \mathcal{L}(\theta^*)) \xrightarrow{d} \mathcal{N}\left(0, \nabla^2 \mathcal{L}(\theta^*) V_{\theta^*}^{-1} \mathbb{E}_{\mathcal{Q}}[\nabla \ell(\theta^*; x, y) \nabla \ell(\theta^*; x, y)^T] V_{\theta^*}^{-1} \nabla^2 \mathcal{L}(\theta^*)\right). \tag{18}$$

If $\nabla \mathcal{L}(\theta^*) \neq 0$, then (17) implies

$$\Delta(z, \mathcal{D}) = k^{-1} \nabla \mathcal{L}(\hat{\theta}^*)^T V_{\theta^*}^{-1} \nabla \ell(\theta^*; x_{k+1}, y_{k+1}) + O_P(k^{-3/2}).$$

And if $\nabla \mathcal{L}(\theta^*) = 0$, then $\sqrt{k} \nabla \mathcal{L}(\hat{\theta}_k) = O_P(1)$ and (17) implies

$$\Delta(z, \mathcal{D}) = k^{-3/2} (\sqrt{k} \nabla \mathcal{L}(\hat{\theta}_k))^T V_{\theta^*}^{-1} \nabla \ell(\theta^*; x_{k+1}, y_{k+1}) + O_P(k^{-2}),$$

which finishes the proof. □

Remark A.8. Although we introduce many assumptions to establish Theorem A.7, many of them are natural and could be verified. For example, for a logistic regression model, the loss function is $\ell(\theta; x, y) = -\theta^T xy + \log(1 + \exp(\theta^T x))$, so the first and second order derivatives are

$$\nabla \ell(\theta; x, y) = -x\left(y - \frac{\exp(\theta^T x)}{1 + \exp(\theta^T x)}\right), \quad \nabla^2 \ell(\theta; x, y) = \frac{\exp(\theta^T x)}{(1 + \exp(\theta^T x))^2} xx^T. \tag{19}$$

Moreover, the third derivative can be written as

$$\nabla^3 \ell(\theta; x, y) = \frac{\exp(\theta^T x)(1 - \exp(\theta^T x))}{(1 + \exp(\theta^T x))^3} x \otimes x \otimes x. \tag{20}$$

Then the Lipschitz condition in Lemma A.3 can be guaranteed if the $\mathbb{E}_{\mathcal{Q}}[\|x\|_2^3] < \infty$. Suppose the metric is either $\mathcal{L}_1(\theta) = \mathbb{E}_{(x', y') \sim \mathcal{Q}}[\ell(\theta; x', y')]$ or $\mathcal{L}_2(\theta) = \frac{1}{m} \sum_{j=1}^m \ell(\theta; x_j', y_j')$, Assumption A.5 and the requirement $\mathbb{E}_{\mathcal{Q}}[\|\nabla^2 \ell(\theta; x, y)\|] \leq \infty$ is also implied by $\mathbb{E}_{\mathcal{Q}}[\|x\|_2^3] < \infty$. For the requirement on function C, the upper bound in Lemma A.4 could be verified by (16). In particular, since $\nabla^2 \ell(\theta; x, y)$ is a rank-one matrix and depends on θ only through the scale, we have

$$\begin{aligned}
C(u; x, y) &= \sup_{\|\theta_1 - \theta_2\|_2 \leq u} \frac{\exp(\theta_1^T x)}{(1 + \exp(\theta_1^T x))^2} \frac{(1 + \exp(\theta_2^T x))^2}{\exp(\theta_2^T x)} \\
&\leq \sup_{\|\theta_1 - \theta_2\|_2 \leq u} \frac{\exp(\theta_1^T x)}{\exp(\theta_2^T x)} \sup_{\|\theta_1 - \theta_2\|_2 \leq u} \frac{\exp(\theta_2^T x)^2}{\exp(\theta_1^T x)^2} \\
&\leq \exp(3u\|x\|_2).
\end{aligned}$$

Hence Assumption A.6 can be fulfilled by assuming $\|x\|_2$ is bounded. We want to highlight that many assumptions can be relaxed, and here we only state a basic version for clarity.

B. Likelihood-based scaling law estimator

Our objective when fitting the likelihood-based estimator is the following loss function,

$$\min_{c,\alpha,\sigma,\beta} \; \mathcal{L}(c,\alpha,\sigma,\beta) = \sum_{i=1}^{m} \mathrm{NLL}(\Delta(z,\mathcal{D}_i), |\mathcal{D}_i|; c,\alpha,\sigma,\beta),$$

where the negative log-likelihood for each marginal contribution is based on Gaussian distributions with certain shared parameters:

$$\mathrm{NLL}(\Delta, k; c,\alpha,\sigma,\beta) = \frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma^2) - \frac{\beta}{2}\log(k) + \frac{(\Delta - ck^{-\alpha})^2}{2\sigma^2 k^{-\beta}}.$$

The objective is non-convex, and we find that directly optimizing over the four parameters leads to unstable results. We therefore use another approach that involves solving analytically for two of the four parameters. Given values for α and β, we can derive optimal values for c and σ in an analytic fashion. Beginning with c, we can see that its optimal value does not depend on σ, so we can write the optimal value $c(\alpha,\beta)$ as follows:

$$c(\alpha,\beta) = \arg\min_{c} \; \sum_{i=1}^{m} k_i^{\beta} \left(\Delta_i - ck_i^{-\alpha}\right)^2 = \frac{\sum_i k_i^{\beta-\alpha}\Delta_i}{\sum_i k_i^{\beta-2\alpha}}.$$

Next, we consider the optimal value for σ^2 given fixed values for α, β and the optimal value $c(\alpha,\beta)$. We solve this problem by reparameterizing in terms of the precision $\tau = 1/\sigma^2$, and we use the solution to define the optimal value $\sigma(\alpha,\beta)$:

$$\sigma(\alpha,\beta) = \left(\arg\min_{\tau} \; \sum_{i=1}^{m} -\log(\tau) + \tau\frac{\left(\Delta_i - c(\alpha,\beta)k_i^{-\alpha}\right)^2}{k^{-\beta}}\right)^{-\frac{1}{2}} = \sqrt{\frac{\sum_i k_i^{\beta}\left(\Delta_i - c(\alpha,\beta)k_i^{-\alpha}\right)^2}{m}}.$$

Using these analytic expressions, we can define our objective solely over the exponents α and β. We therefore simplify the fitting process by optimizing the following loss:

$$\mathcal{L}(\alpha,\beta) = \mathcal{L}\left(c(\alpha,\beta), \alpha, \sigma(\alpha,\beta), \beta\right).$$

We optimize the two-parameter objective with Adam (Kingma and Ba, 2014), and we find that this leads to stable results with a moderate number of gradient steps.

Amortized estimator. For the amortized estimator, we use several approaches to stabilize and improve the model's performance. Because we observe a wide range of values for the $c(z)$ coefficient from the non-amortized estimator, we set up the network to separately predict the sign and log-magnitude of this coefficient, or $\mathrm{sign}(c(z))$ and $\log|c(z)|$. We use gradient clipping, we clamp the predicted β values to have a maximum value of $\exp(2)$ to avoid variance predictions too close to 0 at large k values, and we apply a small prior to the predicted $\alpha(z)$ values via a penalty $(\alpha(z) - 1)^2$.

C. Additional Experimental Results

C.1. Evidence for individualized data scaling laws

Here, we provide experimental details for the results shown in Section 4.1. For SVMs we use the RBF kernel, and for neural networks we set the network width as 32, weight decay as 0.01, and train using the Adam optimizer (Kingma and Ba, 2014) with learning rate 0.01 for 200 epochs. All of our models are run with scikit-learn package in Python. For each setting, we evaluated the model performance on a held-out test dataset of size 1000. In addition, when generating the preceding dataset, we enforced the class balance to avoid degenerate scenarios where certain classes are not represented.

First, we show the estimated $\log|c(z)|$ and the relationship between $\alpha(z)$ and $\log|c(z)|$ in Figure 8. The estimated $c(z)$ values exhibit a significant long-tail behavior in the original space, but we find that it has a well-shaped distribution in log-space. Moreover, we observe a strong positive correlation between $\alpha(z)$ and $\log|c(z)|$. This can be explained by the exponential effect of $\alpha(z)$ on $\psi_k(z)$: if $\alpha(z)$ increases by 1, $c(z)$ needs to increase by a factor of k to keep $\psi_k(z)$ unchanged, therefore it is more reasonable to look at $\log|c(z)|$ values. Similar to $\alpha(z)$, we also observe significant heterogeneity in the $\log|c(z)|$ values.
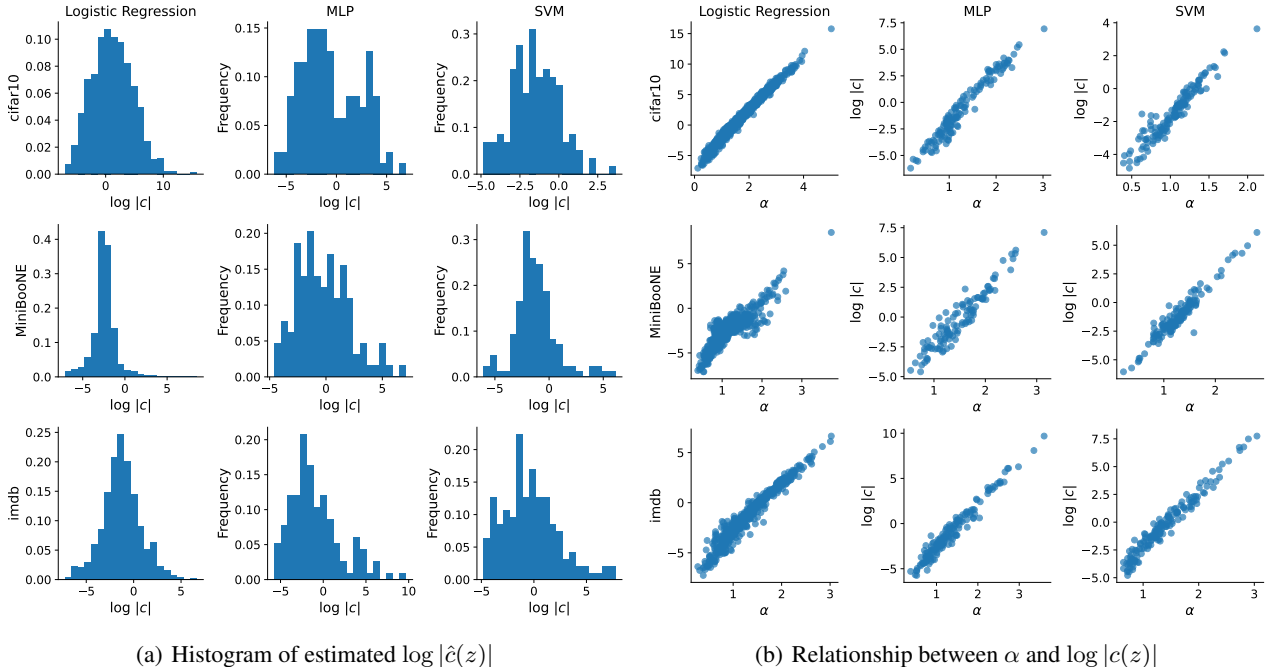

(a) Histogram of estimated $\log|\hat{c}(z)|$ (b) Relationship between α and $\log|c(z)|$

Figure 8. **Histogram of estimated** $\log|c(z)|$ **and the relationship between** $\alpha(z)$ **and** $\log|c(z)|$**:** The estimated $c(z)$ values exhibit a well-shaped distribution in the log scale, and we observe a strong correlation between $\alpha(z)$ and $\log|c(z)|$. Similar to Figure 4, we also exclude points with $R^2 < 0.8$ to make sure the estimates are reliable.

In Figure 3, we observe that most of the R^2 values are very close to 1, but there is still a small portion of points whose R^2 is below 1. We want to highlight that this is largely due to the finite sample error of estimating the mean at each cardinality. In Figure 9, we show the empirical distribution of R^2 using different numbers of samples, and we can observe that the distribution converges to 1 as the sample size increases; we generally expect all data points' R^2 values to be close to 1 when we have an infinite number of samples, which is consistent with our scaling law in Eq. (4). Moreover, in Figure 10, we examine the relationship between R^2 and $\log|\psi_k(z)|$. It is observed that for points with smaller R^2, they also tend to have smaller marginal contributions; this is most apparent when visualizing $\psi_k(z)$ for $k = 100$ and the average of $\psi_k(z)$ in the range from 100 to 1000 (Figure 10 left and bottom). This observation suggests that the poorly-fitting points may not be due to the scaling law not applying, but to estimation noise that is relatively large when viewed in log-space.

Besides the evidence of the scaling law for the marginal contribution, we also provide evidence for the scaling law of the
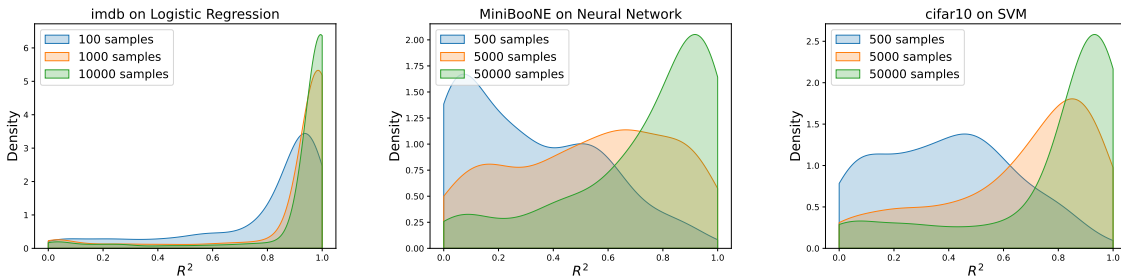
Figure 9. **Density of R^2 score with different numbers of samples.** As the sample size increases, the distribution of R^2 is concentrating towards 1.

marginal contribution's variance, i.e., we aim to show that

$$\text{Var}(\Delta(z, \mathcal{D}) \mid |\mathcal{D}| = k) \approx \frac{\sigma^2(z)}{k^{\beta(z)}}. \tag{21}$$

Similar to our approach in Section 2.1, we can verify this scaling law by taking the following log transform,

$$\log \text{Var}(\Delta(z, \mathcal{D}) \mid |\mathcal{D}| = k) \approx \log \sigma^2(z) - \beta(z) \log k, \tag{22}$$

and evaluating the R^2 score of fitting a simple linear regression in log-space. The experiment results are shown in Figure 11, and we find that the R^2 scores for the variance are very close to 1 for all logistic regression and MLP models. For SVM, the computed R^2 are mostly close to 1, but there exists a portion of points that have a relatively small R^2. A possible reason is that the classifier returned by SVM depends only on the support vectors, hence whether or not the point is a support vector makes the marginal contribution differ significantly. Therefore, although the mean of those marginal contributions aligns well with the scaling law, the variance could be relatively unstable and result in a poor fit.

Next, we examine the distribution of fitted parameters $\beta(z)$ and $\log \sigma^2(z)$. Similar to the $c(z)$ values, here we also observe a long-tail behavior of $\sigma^2(z)$ and hence plot $\log \sigma^2(z)$ instead. As we can observe, the variance of logistic regression models is decaying at a very fast rate, the majority of $\beta(z)$ values lie between 2 and 4. By contrast, for MLPs and SVMs, $\beta(z)$ is mostly between 1 and 2. Therefore, the signal-to-noise ratio of the marginal contribution is much lower for these models, making the estimation of the mean $\psi_k(z)$ at each cardinality more challenging.

Finally, we also show the relationship between distance to decision boundary and $\log |c(z)|$ in Figure 13. Surprisingly, we observe a very different behavior for the logistic regression model on two different datasets. For the MiniBooNE dataset, the heterogeneity of $\log |c(z)|$ mainly comes from the points that are close to the decision boundary, and there is no significant trend when points are far from the decision boundary, as they all concentrate around 0. In contrast, we observe a clear positive correlation between the distance and the estimated $\log |c(z)|$ for the IMDB dataset. This difference implies that the scaling law parameters strongly depend on the data distribution, and they may have fundamentally different behavior when evaluating the same model for different datasets.
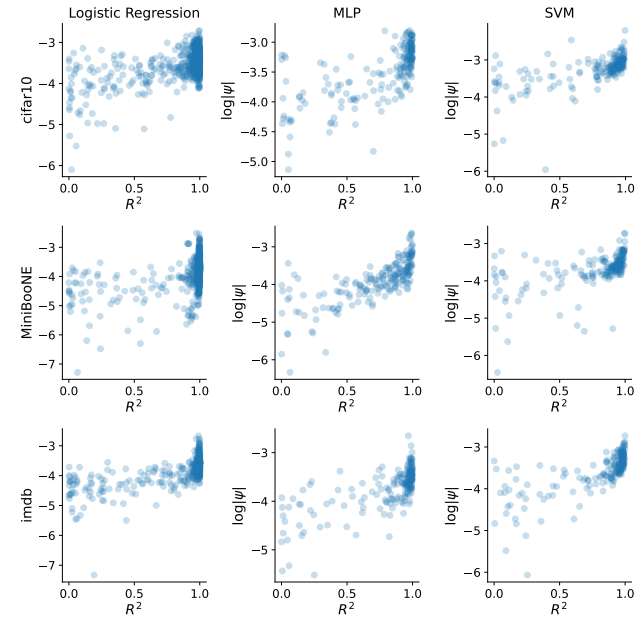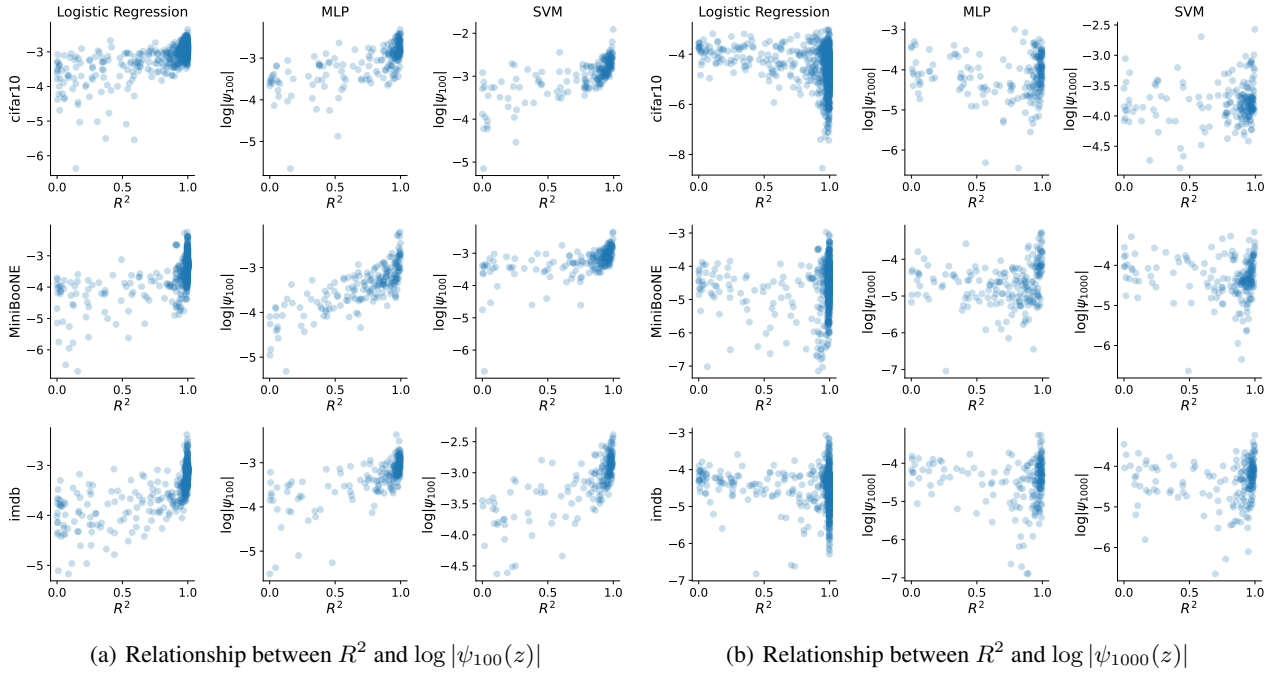
(a) Relationship between R^2 and $\log |\psi_{100}(z)|$

(b) Relationship between R^2 and $\log |\psi_{1000}(z)|$

(c) Relationship between R^2 and $\log |\psi(z)|$, here $\psi(z)$ is defined as the average $\psi_k(z)$ over 10 log-spaced cardinalities between $k = 100$ and $k = 1000$.

Figure 10. **The relationship between R^2 score and the marginal contribution.** We examine the marginal contribution by evaluating $\log |\psi_k(z)|$ at $k = 100$, $k = 1000$, and an average over 10 log-spaced cardinalities between $k = 100$ and $k = 1000$. For points with small R^2, they tend to have small marginal contributions. For $k = 1000$, the marginal contributions of all points are dominated by the sampling error, therefore it is hard to distinguish points with different R^2.
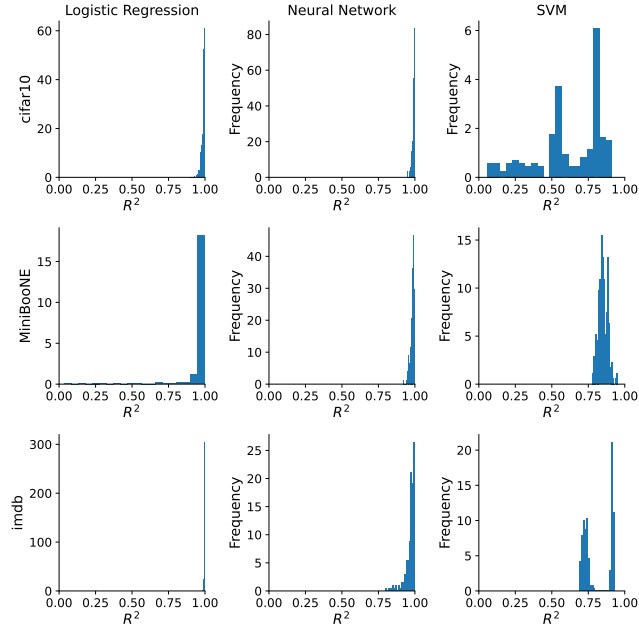
Figure 11. **Histogram of R^2 score of fitting the scaling law fitting on the variance:** for each setting, we compute the variance of sampled $\Delta(z, \mathcal{D})$ with the same size $|\mathcal{D}|$, and then fit a simple linear regression according to (22) to obtain R^2
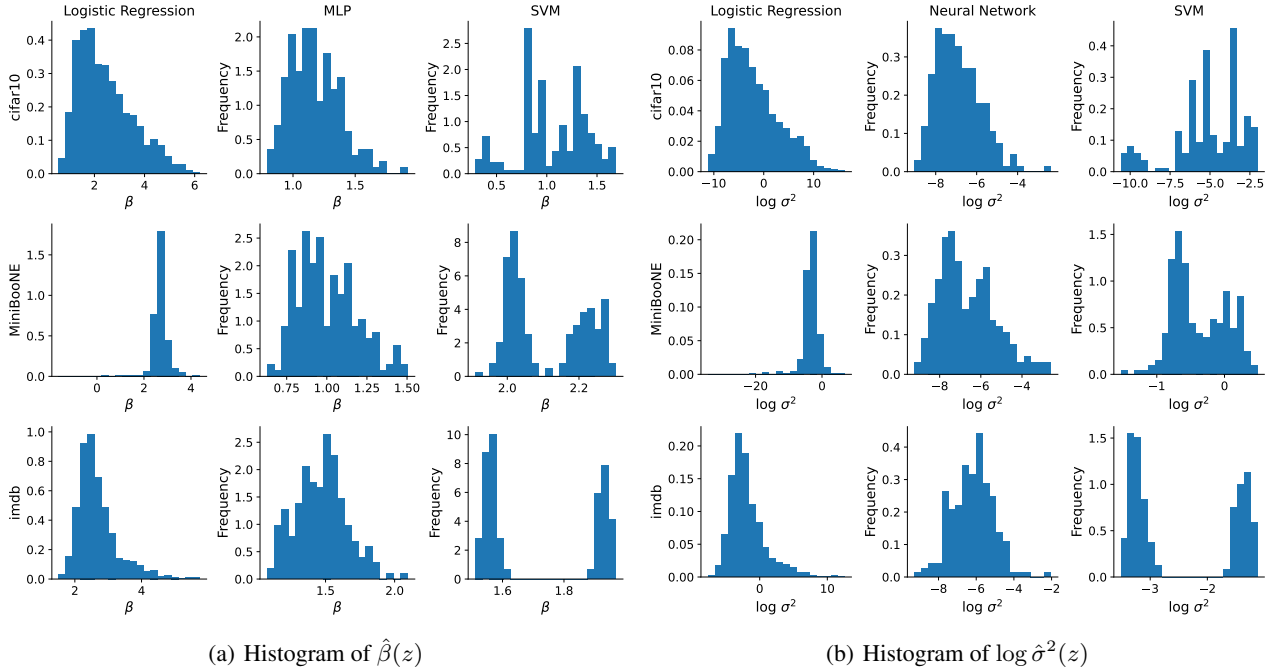
(a) Histogram of $\hat{\beta}(z)$

(b) Histogram of $\log \hat{\sigma}^2(z)$

Figure 12. **Histogram of estimated** $\hat{\beta}(z)$ **and** $\log \hat{\sigma}^2(z)$: the logistic regression model has a much higher β compared with neural networks and SVM.
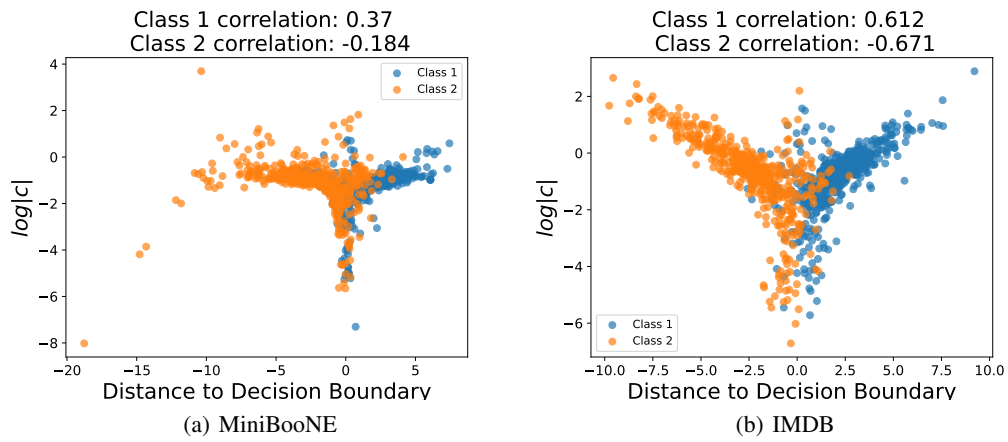
Figure 13. **Distance to decision boundary vs.** $\log |c(z)|$ **for logistic regression model**. We fit a logistic regression model on all points being evaluated and then compute the distance of each data point to the decision boundary. The sign of distance is kept to distinguish points from each class.

C.2. Accuracy of efficient scaling law estimators

Here, we show three additional results related to our scaling law estimators. Like our results in Section 4.2, all are conducted using the IMDB dataset with logistic regression. The first experiment provides intuition for our metrics by showing a scatterplot of the scaling law's predictions at each cardinality, and the true expectation $\psi_k(z)$ for each data point (estimated using 1000 samples). This result is shown in Figure 14, where we observe very strong predictive accuracy for four dataset sizes. We report three measures of the prediction accuracy, where ρ denotes Pearson correlation and τ denotes Spearman correlation. All three metrics are close to 1, particularly for the larger dataset sizes k.
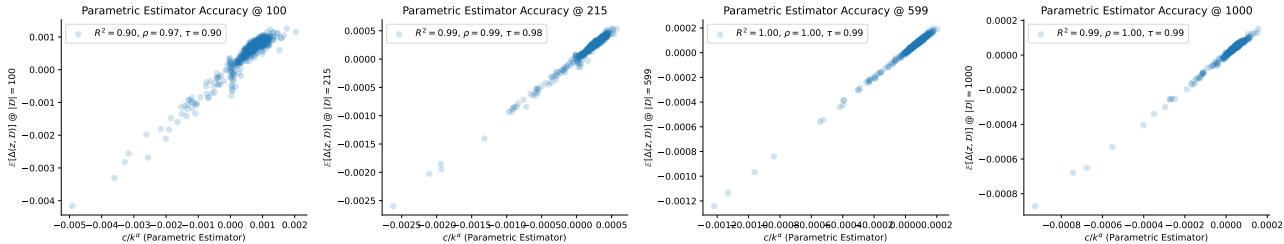

Figure 14. **Scaling law estimation accuracy scatterplots.** For four dataset sizes k, we plot the scaling law's predictions for the marginal contribution against the true expectations $\psi_k(z)$.

The next result is an expanded version of Figure 7, where we show the scaling law's accuracy at different k values with different numbers of samples and according to three metrics: the R^2 score, Pearson correlation, and Spearman correlation. All three metrics are calculated using expectations $\psi_k(z)$ estimated with 1000 samples, similar to the previous result and our validation in Section 2.1. Within the range where the scaling law is fit ($k \in [100, 1000]$), we generally observe high accuracy, particularly at larger cardinalities that have less noise, and when each estimator is fit with more samples. As shown in the main text, we see that the Pearson and Spearman correlation remain relatively high even as we extrapolate to dataset sizes an order of magnitude larger than where the scaling law is fit. However, the R^2 score quickly degrades once we exceed $k = 1000$, which we find is due to the predicted contributions not shrinking as quickly as the empirical ones.
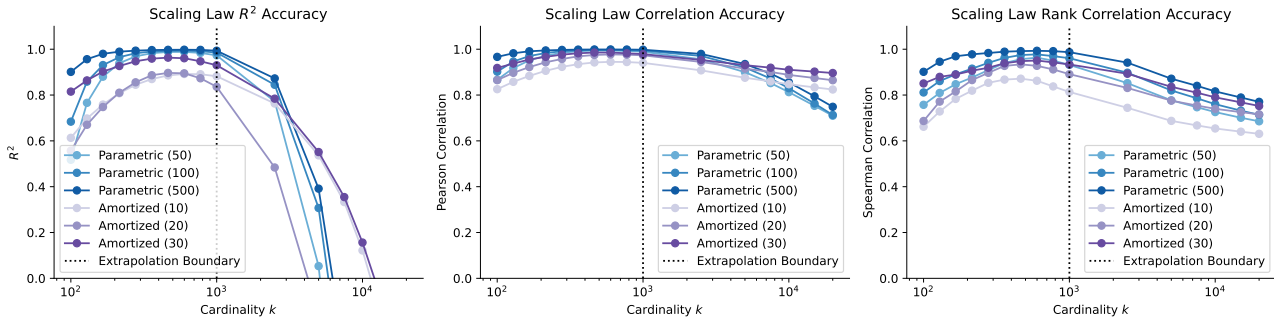

Figure 15. **Scaling law estimation accuracy at multiple cardinalities.** We evaluate the scaling law's predictions at cardinalities up to an order of magnitude larger than where it is fit, and according to three accuracy metrics: R^2, Pearson correlation and Spearman correlation.

Finally, we show a more thorough analysis of the likelihood-based estimator's convergence characteristics. Figure 16 shows the accuracy at each cardinality via a colored line as a function of the number of samples used to fit the estimator. We also separately show results for interpolation ($k \in [100, 1000]$) and extrapolation ($k > 1000$). Overall, we see that for most interpolation cardinalities, the R^2 score converges to 1 within relatively few samples; the more difficult cases in the top row the smallest cardinalities. On the other hand, in the bottom row, we see that the performance has significant room for improvement even after fitting with 100 samples.
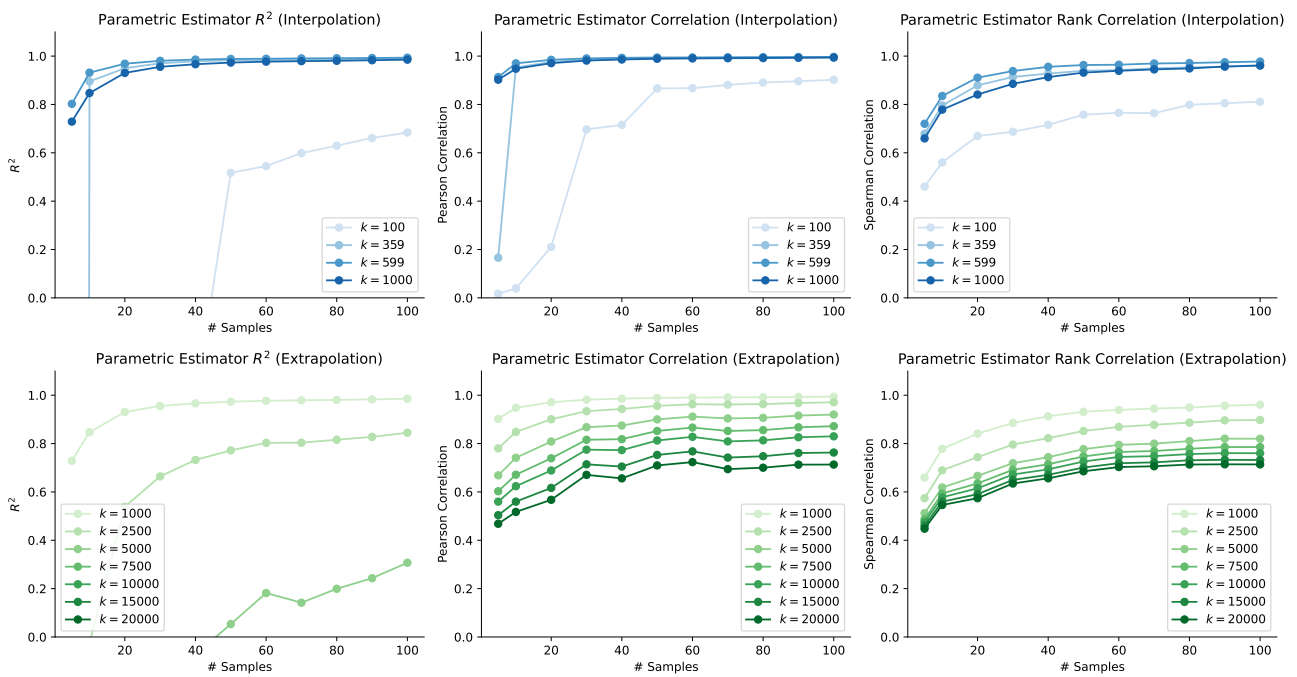
Figure 16. **Likelihood-based scaling law estimator convergence.** We verify that the likelihood-based estimator becomes more accurate as it is fit on more samples, where we separately show cardinalities in the fitting range (top) and beyond the fitting range (bottom).

C.3. Application to data valuation

Here we provide additional results applying individualized scaling laws to data valuation. We first recall the definition of the distributional Shapley value (Ghorbani et al., 2020), which is one of many data valuation scores based on a data point's marginal contributions (Kwon and Zou, 2021; Wang and Jia, 2022; Li and Yu, 2023). Given a maximum cardinality $k_{\max}$, this method defines the valuation score $\psi(z)$ as a uniform average over the expected marginal contributions $\psi_k(z)$ up to $k = k_{\max}$. In practice, we generally also require a minimum cardinality $k_{\min}$ because models cannot be reliably trained without a reasonable amount of data (see the OpenDataVal package for example, Jiang et al. 2023). With this in mind, we can define the valuation score as follows:

$$\psi(z) = \frac{1}{k_{\max} - k_{\min} + 1} \sum_{k=k_{\min}}^{k_{\max}} \psi_k(z).$$

Our experiments involve estimating these scores using a conventional Monte Carlo estimator, which averages marginal contributions $\Delta(z, \mathcal{D})$ with dataset sizes sampled uniformly random, and comparing them to estimates derived from our scaling parameters. For a data point z with parameters $c(z)$ and $\alpha(z)$, we can estimate the expected contribution at cardinality k as $\hat{\psi}_k(z) = c(z)/k^{\alpha(z)}$, and we can average this across k to estimate $\hat{\psi}(z)$. We test this approach when the scaling parameters are estimated using both our likelihood-based estimator from Section 3.1 and the amortized estimator from Section 3.2. As the ground truth for the performance metrics, we use Monte Carlo estimates computed using 10000 sampled marginal contributions, which is chosen to reliably approximate the exact value.

The expanded versions of our main text results are shown in Figure 17 for IMDB, Figure 18 for CIFAR-10, and Figure 19 for MiniBooNE. We also show results for an additional case, the adult census dataset in Figure 20 (Dua and Graff, 2017). All the results are generated by training logistic regression models. We observe that in all four cases, our scaling law-based estimates converge to the same result as the conventional Monte Carlo estimator, which attests to the scaling law accurately capturing the rate of decay of marginal contributions in practice. We observe that the parametric estimator becomes reliable when we use more than 20 samples, although it only outperforms the Monte Carlo estimator for MiniBooNE and the rank correlation metric for IMDB and adult census. On the other hand, the amortized estimator generally provides the best performance in the noisiest regime with just 10 samples, and it is the best estimator for all numbers of samples with the adult census dataset; however, in the case of CIFAR-10 it is surpassed by the non-amortized estimator for larger numbers of samples.
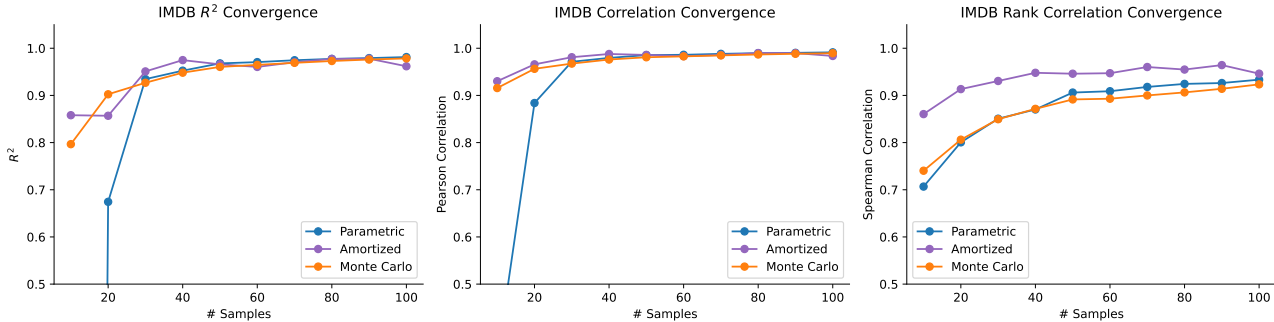

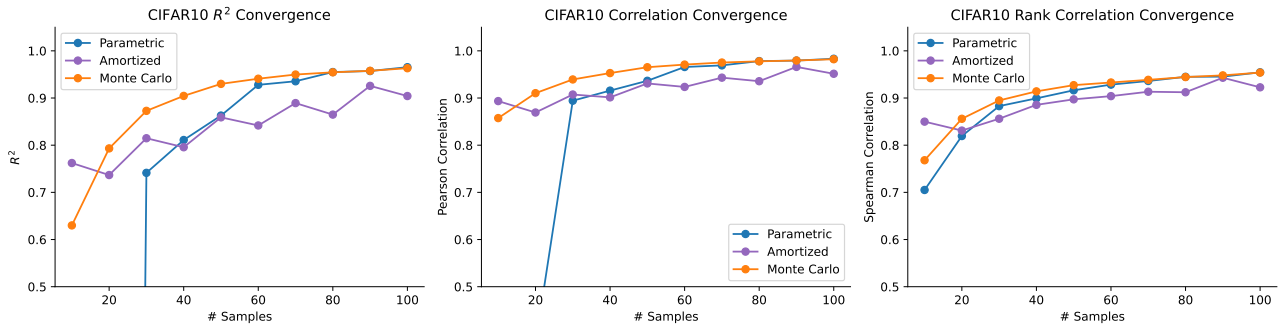*Figure 17.* **Data valuation estimation accuracy for the IMDB dataset with logistic regression**.

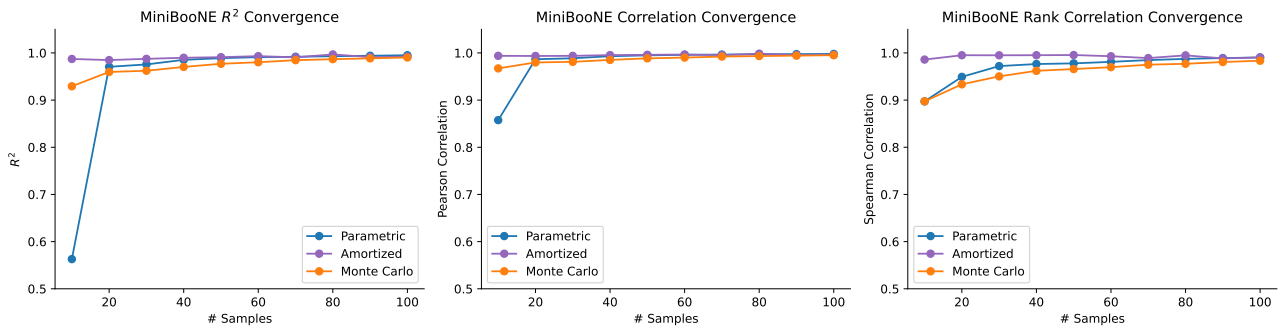*Figure 18.* **Data valuation estimation accuracy for the CIFAR-10 dataset with logistic regression**.

Figure 19. **Data valuation estimation accuracy for the MiniBooNE dataset with logistic regression**.
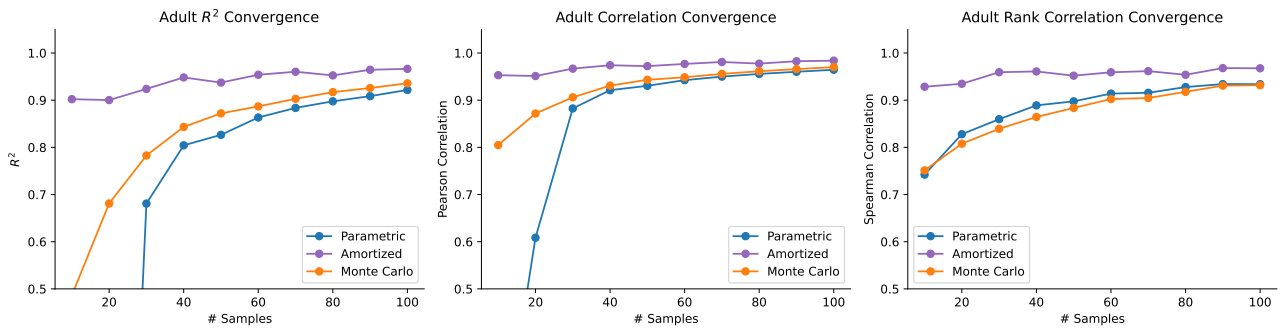

Figure 20. **Data valuation estimation accuracy for the adult census dataset with logistic regression**.

C.4. Application to point addition

Here, we provide additional experimental results on the point addition task. In Section 4.4, we consider adding 20 points to preceding datasets of size 100 and 1000, and here we consider the setting of adding 50 points to preceding datasets of size 200 and 2000. Again, we will predict the expected contribution using the scaling law $\hat{\psi}_k(z) = c(z)/k^{\alpha(z)}$ with fitted parameters $c(z)$ and $\alpha(z)$ at $k = 200$ and $k = 2000$ to select the top points for *Scaling 200* and *Scaling 2000*. The results are shown in Table 7 and Table 8, and the Scaling 200 and Scaling 2000 selections show superior performance on the corresponding smaller and larger datasets, while performing badly in the other regime. Selections based on the Shapley value have worse performance than scaling law selections based on the dataset size. We notice that in Table 7 there is one setting where random selection achieves the best performance, which may be due to interaction effects where the selected points are similar, so the contribution from a whole set may not be as beneficial as expected. In all other settings, we generally observe that random selection has the worst performance.

Table 7. **Accuracy improvement (%) with 50 points added to preceding datasets of size 200.** The settings are the same as Table 5, Scaling 200 achieves the best performance compared to other methods.

Method	MiniBooNE	CIFAR-10	IMDB
Scaling 2000	0.27 ± 0.95	1.51 ± 0.58	0.71 ± 0.68
Scaling 200	0.33 ± 0.98	$\mathbf{2.04 \pm 0.59}$	$\mathbf{1.55 \pm 0.61}$
Shapley	0.28 ± 0.99	1.50 ± 0.58	1.19 ± 0.67
Random	$\mathbf{0.44 \pm 0.68}$	1.16 ± 0.57	0.73 ± 0.53
Preceding	81.99 ± 0.97	80.51 ± 0.97	81.71 ± 0.94

Table 8. **Accuracy improvement (%) with 50 points added to preceding datasets of size 2000.** The settings are the same as Table 5, Scaling 2000 achieves the best performance compared to other methods.

Method	MiniBooNE	CIFAR-10	IMDB
Scaling 2000	$\mathbf{0.15 \pm 0.33}$	$\mathbf{0.16 \pm 0.13}$	$\mathbf{0.09 \pm 0.09}$
Scaling 200	0.14 ± 0.32	0.11 ± 0.10	0.06 ± 0.06
Shapley	0.14 ± 0.32	0.14 ± 0.13	0.08 ± 0.06
Random	0.01 ± 0.17	0.05 ± 0.13	0.03 ± 0.09
Preceding	84.51 ± 0.37	86.39 ± 0.35	86.38 ± 0.24