# Geometric Data Valuation via Leverage Scores

**Rodrigo Mendoza-Smith**
Isotropic
rms@isotropic.sh

## Abstract

Shapley data valuation provides a principled, axiomatic framework for assigning importance to individual datapoints, and has gained traction in dataset curation, pruning, and pricing. However, it is a combinatorial measure that requires evaluating marginal utility across all subsets of the data, making it computationally infeasible at scale. We propose a geometric alternative based on statistical leverage scores, which quantify each datapoint's structural influence in the representation space by measuring how much it extends the span of the dataset and contributes to the effective dimensionality of the training problem. We show that our scores satisfy the dummy, efficiency, and symmetry axioms of Shapley valuation and that extending them to *ridge leverage scores* yields strictly positive marginal gains that connect naturally to classical A- and D-optimal design criteria. We further show that training on a leverage-sampled subset produces a model whose parameters and predictive risk are within $O(\varepsilon)$ of the full-data optimum, thereby providing a rigorous link between data valuation and downstream decision quality. Finally, we conduct an active learning experiment in which we empirically demonstrate that ridge-leverage sampling outperforms standard baselines without requiring access gradients or backward passes.

As machine learning systems increasingly rely on specialised data, understanding the *value* of individual datapoints has become a central challenge. Quantifying this value supports numerous downstream tasks like identifying mislabeled or redundant examples [5], constructing compact and informative training subsets [15], allocating incentives in federated settings [22], and building fair and efficient data markets [1, 9]. A growing body of work has addressed this challenge through *data valuation*, estimating each point's contribution to model performance. Among the most theoretically grounded approaches is *data Shapley*, which defines value through the Shapley axioms of cooperative game theory [5, 8], and has inspired a range of algorithms [8, 13, 12, 28, 25]. While these methods are based on appealing axiomatic guarantees, they are often computationally expensive, require model retraining, or need access to model's weights and gradients. Moreover, most work has focused on data-quality control during pre-training [5, 8, 11], with little attention to settings where data is costly or arrives under uncertainty and decisions must be made about which datapoints to select or acquire.

In this work, we take a geometric, model-agnostic perspective on data valuation grounded in the structure of the dataset itself. Specifically, we propose using *statistical leverage scores*, a well-established concept in numerical linear algebra (NLA) [4] to assess the *structural importance* of datapoints. Intuitively, high-leverage points span unique directions in feature space and are therefore valuable, while low-leverage points are often redundant. Recent work has used leverage scores to estimate Shapley values [17, 26] through sampling. In contrast, our approach treats leverage scores as direct geometric surrogates for Shapley data valuation. Our contributions are threefold: (i) we adapt leverage scores to data valuation and show that, under certain conditions, our geometric proxy satisfies the core Shapley axioms; (ii) we extend this valuation to *ridge leverage scores* [3, 16, 4] to mitigate dimensional saturation and connect with classical A- and D-optimal design criteria; and (iii) we provide theoretical guarantees and empirical validation, proving that leverage-based sampling yields

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

$\varepsilon$-close decision quality to the full-data optimum and achieves strong performance in a small-scale active learning experiment without requiring gradients, labels, or quadratic computation.

## 1 Leverage scores as proxies for Shapley value

**Shapley value and data valuation** Let $n \in \mathbb{N}$, and define $[n] := \{1, \ldots, n\}$. In the game theory literature, the Shapley value [21] quantifies the expected marginal contribution of a player $i \in [n]$ in a cooperative game with $n$ players, as determined by an utility function $U : 2^{[n]} \to \mathbb{R}$ that assigns to each coalition $S \subseteq [n]$ of players the total value or payoff that the members of $S$ can achieve by cooperating. This is,

$$\phi_U(i) = \mathbb{E}_{S \sim 2^{[n]\setminus\{i\}}} \left[ U(S \cup \{i\}) - U(S) \right], \tag{1}$$

In the Data Shapley framework [5], training is modelled as a cooperative game where each datapoint is a player, and the utility function is defined by the model's performance on a validation set. Under this formulation, the Shapley value $\phi_U(i)$ assigns each datapoint its expected marginal contribution under all data permutations as seen in (1). It has been shown that for a given utility function $U$, the Shapley value $\phi_U(i) \in \mathbb{R}$ is the only solution that satisfies:

$$\text{Symmetry:} \quad \text{If } U(S \cup \{i\}) = U(S \cup \{j\}) \quad \forall S \subset [n] \setminus \{i, j\}, \text{ then } \phi_U(i) = \phi_U(j) \tag{2}$$

$$\text{Efficiency:} \quad \phi_U(1) + \cdots + \phi_U(n) = U([n]) - U(\emptyset) \tag{3}$$

$$\text{Dummy:} \quad \text{If } U(S \cup \{i\}) = U(S) \quad \forall S \subset [n] \setminus \{i\}, \text{ then } \phi_U(i) = 0 \tag{4}$$

$$\text{Linearity:} \quad \phi_{\alpha U + \beta V}(i) = \alpha \phi_U(i) + \beta \phi_V(i), \quad \forall \alpha, \beta \in \mathbb{R} \tag{5}$$

*Symmetry* ensures that datapoints with equivalent contributions are valued equally. It enforces fairness in valuation when the contribution depends purely on content or structure, rather than dataset composition. *Efficiency* is critical in seettings where valuations are interpreted as prices, rewards , or payouts such as in data markets or collaborative training. It guarantees that the value distribution reflects the full contribution of the dataset without over- or under-counting. *Dummy* is useful in the context of dataset construction as it allows us to systematically identify and eliminate redundant or uninformative examples. *Linearity* guarantees that data valuations are additive across different tasks or objective functions, making it easier to aggregate valuations across tasks or adapt to changing utility functions.

**A non-linear geometric proxy based on leverage scores** Let $\mathcal{D} = \{\mathbf{x}_i, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ be a dataset, and let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix where each row $\mathbf{x}_i^\top$ of $\mathbf{X}$ corresponds to a datapoint in $\mathcal{D}$. We assume that $\mathbf{X}$ has full column rank. The *leverage score* $\ell_i$ of the $i$-th datapoint $\mathbf{x}_i^\top$ in dataset $\mathbf{X}$ is defined as the $i$-th diagonal entry of the projection (hat) matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$,

$$\ell_i = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i. \tag{6}$$

In numerical linear algebra, leverage scores are used to evaluate the sensitivity of least-squares problems and guide importance sampling in randomized matrix algorithms. Here, we use them to build a *geometric proxy for data Shapley values*, capturing how much each datapoint extends the span of the dataset in representation space. We define a normalized leverage-based value function:

$$\pi_i = \frac{\ell_i}{\sum_{j=1}^{n} \ell_j}. \tag{7}$$

Our first result is showing that (7) is a *non-linear* proxy to Data Shapley values.

**Theorem 1** (Shapley axioms)**.** *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix with rows $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$ and define*

$$\phi_U(i) := \pi_i \ \ \forall i \in [n], \tag{8}$$

*Then, if $\operatorname{rank}(\mathbf{X}) = d$, $\phi_U$ satisfies the symmetry* (2)*, efficiency* (3)*, and dummy* (4) *axioms of data Shapley for $U(S) := \operatorname{span}\{\mathbf{x}_i : i \in S\}$ for all $S \subset [n]$.*

A proof is given in Appendix A. We note that our leverage valuation scores do not generally satisfy linearity. While this may appear to be a limitation, linearity is not essential in applications where the value of a datapoint depends on its marginal contribution to the structural diversity of the dataset. A more severe limitation of (7), however, is that of *dimensional saturation*: because the scores measure value in terms of the structural diversity contributed to the span of the dataset, once the span reaches the ambient dimension $d$, any additional datapoint has zero marginal value.

**Mitigating Dimensional Saturation** This saturation at $\operatorname{rank}(\mathbf{X}_S) = d$ is the price we pay for the simplicity of (7). In practice, however, we want scores that continue to capture variance reduction and predictive improvement even once the span is full. A natural way to achieve this is through *ridge leverage* [3], which regularizes (6) and (7) as

$$\ell_i^{(\lambda)} \;=\; \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{x}_i, \qquad \pi_i^{(\lambda)} \;=\; \frac{\ell_i^{(\lambda)}}{\sum_{j=1}^n \ell_j^{(\lambda)}}. \tag{9}$$

Note that for any $\lambda > 0$, ridge leverage $\ell_i^{(\lambda)} \in (0, 1)$ and the statistical dimension $k_\lambda = \sum_{i=1}^n \ell_i^{(\lambda)}$ lies strictly between 0 and $d$. As new datapoints are added, $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$ contracts, reducing but never eliminating marginal gains; thus even after $\operatorname{rank}(\mathbf{X}_S) = d$ additional examples retain nonzero value. This connects ridge leverage to classical criteria from optimal experimental design: the marginal gains under both A- and D-optimality can be written directly in terms of $\ell^{(\lambda)}(\mathbf{x})$. Indeed, letting $\mathbf{A} = \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} \succ 0$, standard matrix identities[1] yield

$$\text{(D-optimality)} \qquad \log\det(\mathbf{A} + \mathbf{x}\mathbf{x}^\top) - \log\det(\mathbf{A}) = \log\big(1 + \ell^{(\lambda)}(\mathbf{x})\big) \;>\; 0,$$

$$\text{(A-optimality)} \qquad \operatorname{Tr}\big((\mathbf{A} + \mathbf{x}\mathbf{x}^\top)^{-1}\big) - \operatorname{Tr}(\mathbf{A}^{-1}) = -\frac{\|\mathbf{A}^{-1}\mathbf{x}\|_2^2}{1 + \ell^{(\lambda)}(\mathbf{x})} \;<\; 0.$$

Normalizing $\pi_i^{(\lambda)} = \ell_i^{(\lambda)}/k_\lambda$ yields a valuation that connects ridge leverage directly to classical design criteria. In particular, the marginal gain in D-optimality is $\log(1 + \ell^{(\lambda)}(\mathbf{x}))$, and the marginal gain in A-optimality is likewise a monotone function of $\ell^{(\lambda)}(\mathbf{x})$. Thus ridge leverage scores govern the size of these improvements, ensuring that every nonzero datapoint contributes positively. This softens the hard-$d$ saturation of the span utility and reflects the practical reality that additional data can still reduce variance and improve accuracy even after the feature space is fully spanned.

**Proposition 2** (Shapley axioms for Ridge leverage)**.** *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be full column rank. For $\lambda > 0$ and $i \in [n]$ let $\ell_i^{(\lambda)}$ and $\pi_i^{(\lambda)}$ be as in* (9)*. Then, $\pi^{(\lambda)}$ satisfies the symmetry* (2) *and efficiency* (3) *properties of Data Shapley.*

The proof is analogous to that of Theorem 1. Note, however, that in general ridge leverage does not satisfy the Dummy axiom: for $\lambda > 0$ every nonzero datapoint yields strictly positive marginal gain under ridge-based utilities such as

$$U_{\mathrm{D}}(S) = \log\det(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I}), \qquad U_{\mathrm{A}}(S) = -\operatorname{Tr}((\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}),$$

so the exact Shapley value for these $U$ vanishes only when $\mathbf{x}_i = \mathbf{0}$. Our normalized ridge leverage $\pi^{(\lambda)}$ should therefore be viewed as a *geometric surrogate* for Shapley: it preserves efficiency and a natural notion of symmetry, and it recovers linearity in structured regimes, while deliberately departing from Dummy in the same way as ridge-based experimental design. This departure is in fact desirable: it ensures that redundant datapoints beyond rank $d$ are still assigned positive value, which is consistent with their role in reducing estimation variance and improving downstream decision quality.

From an NLA perspective, leverage scores are motivated by their geometric properties and their role in randomized least-squares algorithms. From an Operations Research (OR) perspective, however, the key question is how such valuations affect the *quality of downstream decisions*, i.e., the fitted model $\hat{\theta}$ and its predictive risk. We therefore ask: if we subsample training data according to leverage-based valuations, how close is the resulting model to the one trained on the full dataset? Using ridge regression as a tractable proxy, we prove that our valuation satisfies $\varepsilon$-close decision quality bounds. Mathematically, our analysis draws on well-known ingredients from compressed-sensing and matrix concentration [24, 23, 2] and randomized NLA [3, 16, 4].

**Theorem 3** ($\varepsilon$-close to the full-data ridge solution)**.** *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix with rows $\mathbf{x}_i^\top$, let $\mathbf{y} \in \mathbb{R}^n$, and define*

$$\mathbf{A} := \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d, \qquad \mathbf{b} := \mathbf{X}^\top \mathbf{y}, \qquad R(\theta) := \tfrac{1}{2}\|\mathbf{X}\theta - \mathbf{y}\|_2^2 + \tfrac{\lambda}{2}\|\theta\|_2^2.$$

[1] Matrix determinant lemma: $\det(\mathbf{A} + \mathbf{u}\mathbf{v}^\top) = \det(\mathbf{A})(1 + \mathbf{v}^\top \mathbf{A}^{-1}\mathbf{u})$ with $\mathbf{u} = \mathbf{v} = \mathbf{x}$; Sherman–Morrison: $(\mathbf{A} + \mathbf{x}\mathbf{x}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^\top\mathbf{A}^{-1}}{1 + \mathbf{x}^\top\mathbf{A}^{-1}\mathbf{x}}$, followed by a trace.

*Let* $\theta^\star := \operatorname{argmin}_\theta R(\theta)$. *Sample* $m$ *i.i.d. indices with probabilities* $p_i := \ell_i^{(\lambda)}/k_\lambda$, *set weights* $\mathbf{W}_{tt} = (mp_{i_t})^{-1/2}$, *matrices* $\widetilde{\mathbf{X}} = \mathbf{WSX}$ *and* $\widetilde{\mathbf{y}} = \mathbf{WSy}$, *and let*

$$\mathbf{A}_S := \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} + \lambda \mathbf{I}_d, \qquad \mathbf{b}_S := \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{y}}, \qquad \widehat{\theta} := \mathbf{A}_S^{-1} \mathbf{b}_S.$$

*Fix* $\varepsilon \in (0, \frac{1}{2})$ *and* $\delta \in (0,1)$, *and assume* $\mathbf{y} = \mathbf{X}\theta_{\text{lin}}$ *for some* $\theta_{\text{lin}} \in \mathbb{R}^d$. *If* $m \geq C \frac{k_\lambda + \log(2d/\delta)}{\varepsilon^2}$, *then with probability at least* $1 - \delta$,

$$(1-\varepsilon)\mathbf{A} \preceq \mathbf{A}_S \preceq (1+\varepsilon)\mathbf{A} \quad \textit{and} \quad \|\mathbf{b}_S - \mathbf{b}\|_{\mathbf{A}^{-1}} \leq \varepsilon \|\theta_{\text{lin}}\|_{\mathbf{A}}. \tag{A}$$

*Consequently,*

$$\|\widehat{\theta} - \theta^\star\|_{\mathbf{A}} \leq 4\varepsilon \|\theta_{\text{lin}}\|_{\mathbf{A}}, \quad \textit{and} \quad R(\widehat{\theta}) - R(\theta^\star) \leq 8\varepsilon^2 \|\theta_{\text{lin}}\|_{\mathbf{A}}^2. \tag{Q}$$

The same techniques underlying Theorem 3 can also be extended beyond the noiseless linear model and to derive guarantees in the setting where the labels are contaminated by sub-Gaussian noise. We leave a careful treatment of these extensions to future work.

## 2 An Active Learning experiment

To illustrate ridge leverage, we designed a small-scale Active Learning (AL) experiment on MNIST [14] using a 3-layer MLP (784→256→64→10 neurons) with six selection strategies: (1) ridge leverage with adaptive regularization $\lambda = 0.01 \times \text{Tr}(\mathbf{X}^\top \mathbf{X})/64$, and scores computed on 64-dimensional learned embeddings from the penultimate layer; (2) K-center [19], which uses greedy selection based on distances to the nearest center; (3) Margin [20], which selects samples with the smallest difference between the top-2 predicted probabilities; (4) Entropy [20], which selects samples with the highest Shannon entropy; (5) Expected Gradient Length (EGL) [7], which selects samples with the largest expected gradient length; and (6) a random uniform baseline. Other, more sophisticated active learning strategies exist, such as BALD [6], BatchBALD [10], ActiveMatch [29], LESS [27], TRAK [18], but we do not consider these in our experiments. Each experiment began with 100 randomly labeled samples and performed 20 rounds of deterministic pretraining for fair comparison, followed by 40 active learning rounds selecting 5 samples per round. We ran 5 independent trials (seeds 0–4) and evaluated performance using test accuracy [2] As shown in Figure 1, ridge-leverage sampling provides a clear advantage over standard active learning baselines as soon as the pretraining phase is completed. By the end of the acquisition process, ridge-leverage attains the highest mean test accuracy ($0.846 \pm 0.006$) while maintaining low variability across runs without requiring access to gradients, labelled data, or quadratic computation. These findings support our theoretical intuition that ridge leverage selects samples that contribute to the model's stability and generalization, making it an effective and robust strategy for data-efficient learning.

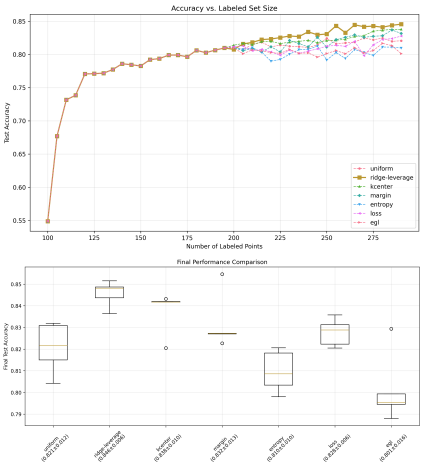


Figure 1: (Top) Test accuracy versus number of labeled samples for six AL strategies on MNIST. (Bottom) Final test accuracy after 40 acquisition rounds.

## 3 Conclusion

In this work we introduced a geometric perspective on data valuation based on ridge leverage scores. Our main contribution is to show that these scores constitute a principled measure of the influence of individual datapoints and yield $\mathcal{O}(\varepsilon)$-close guarantees for a specific risk model. These theoretical developments position ridge-leverage scores as a sound and tractable alternative to data Shapley values. We further demonstrated the applicability of ridge leverage to *active learning*, where it performs competitively with standard selection strategies. Fully developing these scores into a selector that competes with the state-of-the-art warrants separate, dedicated study so we leave this as future work.

[2] All code and experiments are available at `https://github.com/rodrgo/geosh`.

## A Proof of Theorem 1

*Proof.* Let $\mathbf{H} := \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ be the projection matrix onto $\mathrm{col}(\mathbf{X})$. Its diagonal entries are the leverage scores $\ell_i = \mathbf{H}_{ii}$, and by construction $\pi_i = \ell_i / \sum_j \ell_j$.

*Efficiency:* Since $\mathbf{H}$ is a projection of rank $d$, $\mathrm{Tr}(\mathbf{H}) = d$. Thus

$$\sum_{i=1}^{n} \pi_i = \sum_{i=1}^{n} \frac{\ell_i}{d} = \frac{1}{d} \cdot \sum_{i=1}^{n} \ell_i = \frac{d}{d} = 1,$$

so the total value is fully distributed.

*Dummy:* Suppose datapoint $\mathbf{x}_i$ satisfies the condition that for all subsets $S \subseteq [n] \setminus \{i\}$, the span of $\mathbf{X}_S \cup \{\mathbf{x}_i\}$ is equal to the span of $\mathbf{X}_S$. That is, adding $\mathbf{x}_i$ does not increase the dimension of the span or change the subspace. In this case, $\mathbf{x}_i \in \mathrm{col}(\mathbf{X}_S)$ for all $S$, and so its projection onto the column space is already covered by other datapoints. This implies:

$$\ell_i = \mathbf{x}_i^\top (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i = 0,$$

and hence $\pi_i = 0$. Therefore, if $\mathbf{x}_i$ contributes no additional utility (as measured via subspace expansion), its assigned value is zero, satisfying the dummy property.

*Symmetry:* Suppose datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$ satisfy the subset symmetry condition stated in the theorem. The leverage score $\ell_i$ is given by:

$$\ell_i = \mathbf{x}_i^\top (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i,$$

which measures the squared projection of $\mathbf{x}_i$ onto the column space of $\mathbf{X}$.

Now, consider the effect of removing one of the two symmetric datapoints (say, $\mathbf{x}_j$) from $\mathbf{X}$. Since the symmetry assumption guarantees that the span of the dataset is unchanged regardless of whether $\mathbf{x}_i$ or $\mathbf{x}_j$ is included, the projection matrix $\mathbf{H}$ remains invariant under swapping $\mathbf{x}_i$ and $\mathbf{x}_j$. In particular, the projections of $\mathbf{x}_i$ and $\mathbf{x}_j$ onto the same column space yield the same squared norm:

$$\ell_i = \|P\mathbf{x}_i\|^2 = \|P\mathbf{x}_j\|^2 = \ell_j,$$

and hence

$$\pi_i = \frac{\ell_i}{\sum_{k=1}^{n} \ell_k} = \frac{\ell_j}{\sum_{k=1}^{n} \ell_k} = \pi_j.$$

□

## B Proof of Theorem (3)

First, we rephrase Theorem 1.1 in [24] and provide a bound on scalar factors as a Lemma.

**Theorem 4** (Matrix Chernoff [24])**.** *Let $\{\mathbf{Y}_t\}_{t=1}^m$ be independent random self-adjoint matrices in $\mathbb{R}^{d\times d}$ such that*

$$\mathbf{Y}_t \succeq 0 \qquad \textit{and} \qquad \lambda_{\max}(\mathbf{Y}_t) \le R \quad \textit{almost surely.}$$

*Define*

$$\mathbf{M} \;:=\; \mathbb{E}\left[\sum_{t=1}^{m} \mathbf{Y}_t\right], \qquad \mu_{\min} \;:=\; \lambda_{\min}(\mathbf{M}), \quad \mu_{\max} \;:=\; \lambda_{\max}(\mathbf{M}).$$

*Then the following bounds hold:*

$$\Pr\Big[\lambda_{\max}\Big(\sum_{t=1}^{m} \mathbf{Y}_t\Big) \;\ge\; (1+\varepsilon)\,\mu_{\max}\Big] \;\le\; d \cdot \left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu_{\max}/R}, \qquad \textit{for all } \varepsilon \ge 0,$$

$$\Pr\Big[\lambda_{\min}\Big(\sum_{t=1}^{m} \mathbf{Y}_t\Big) \;\le\; (1-\varepsilon)\,\mu_{\min}\Big] \;\le\; d \cdot \left[\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right]^{\mu_{\min}/R}, \qquad \textit{for } \varepsilon \in [0,1].$$

Now, we state a lemma that will prove useful to our argument.

**Lemma 5** (Bounds on scalar factors)**.** *For* $\varepsilon \in (0, \frac{1}{2})$*,*

$$\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}} \;\le\; \exp\Big(-\frac{\varepsilon^2}{3}\Big), \qquad \frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}} \;\le\; \exp\Big(-\frac{\varepsilon^2}{2}\Big).$$

*Proof.* Let $f(\varepsilon) = \varepsilon - (1+\varepsilon)\log(1+\varepsilon)$ and $g(\varepsilon) = -\varepsilon - (1-\varepsilon)\log(1-\varepsilon)$. Use Taylor series on $\log(1+\varepsilon)$ and $\log(1-\varepsilon)$ and the result follows. □

Finally, we prove a lemma that links $\theta^\star$ and $\theta_{\mathrm{lin}}$.

**Lemma 6** (Ridge contraction in the $\|\cdot\|_{\mathbf{A}}$ norm)**.** *Let* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *and* $\lambda > 0$*, and define* $\mathbf{A} := \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d$*. Assume* $\mathbf{y} = \mathbf{X}\theta_{\mathrm{lin}}$ *for some* $\theta_{\mathrm{lin}} \in \mathbb{R}^d$*, and let*

$$\theta^\star := \arg\min_{\theta} \tfrac{1}{2}\|\mathbf{X}\theta - \mathbf{y}\|_2^2 + \tfrac{\lambda}{2}\|\theta\|_2^2.$$

*Then*

$$\theta^\star = \mathbf{A}^{-1}\mathbf{X}^\top\mathbf{X}\,\theta_{\mathrm{lin}} = (\mathbf{I}_d - \lambda\mathbf{A}^{-1})\,\theta_{\mathrm{lin}} \qquad \textit{and} \qquad \|\theta^\star\|_{\mathbf{A}} \le \|\theta_{\mathrm{lin}}\|_{\mathbf{A}}.$$

*Proof.* First-order optimality gives $\mathbf{A}\theta^\star = \mathbf{X}^\top\mathbf{y} = \mathbf{X}^\top\mathbf{X}\,\theta_{\mathrm{lin}}$, hence

$$\theta^\star \;=\; (\mathbf{I}_d - \lambda\mathbf{A}^{-1})\,\theta_{\mathrm{lin}}. \tag{10}$$

Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be an SVD with singular values $\sigma_1, \dots, \sigma_r > 0$ (and $\sigma_j = 0$ for $j > r$). Then

$$\mathbf{A} = \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d = \mathbf{V}\,(\mathbf{\Sigma}^\top\mathbf{\Sigma} + \lambda\mathbf{I}_d)\,\mathbf{V}^\top, \qquad \mathbf{A}^{1/2} = \mathbf{V}(\mathbf{\Sigma}^\top\mathbf{\Sigma} + \lambda\mathbf{I}_d)^{1/2}\mathbf{V}^\top.$$

Work in the $\mathbf{V}$-basis: $\tilde\theta_{\mathrm{lin}} := \mathbf{V}^\top\theta_{\mathrm{lin}}$ and $\tilde\theta^\star := \mathbf{V}^\top\theta^\star$. By (10),

$$\tilde\theta^\star = \Big(\mathbf{I}_d - \lambda(\mathbf{\Sigma}^\top\mathbf{\Sigma} + \lambda\mathbf{I}_d)^{-1}\Big)\tilde\theta_{\mathrm{lin}},$$

and therefore

$$\|\theta^\star\|_{\mathbf{A}} = \big\|(\mathbf{\Sigma}^\top\mathbf{\Sigma} + \lambda\mathbf{I}_d)^{1/2}\tilde\theta^\star\big\|_2 = \big\|(\mathbf{\Sigma}^\top\mathbf{\Sigma} + \lambda\mathbf{I}_d)^{1/2}\big(\mathbf{I}_d - \lambda(\mathbf{\Sigma}^\top\mathbf{\Sigma} + \lambda\mathbf{I}_d)^{-1}\big)\tilde\theta_{\mathrm{lin}}\big\|_2.$$

Thus, each coordinate $j$ of $\tilde\theta_{\mathrm{lin}}$ is multiplied by

$$h_j := \sqrt{\sigma_j^2 + \lambda}\,\Big(1 - \tfrac{\lambda}{\sigma_j^2+\lambda}\Big) = \sqrt{\sigma_j^2 + \lambda}\,\tfrac{\sigma_j^2}{\sigma_j^2+\lambda} = \frac{\sigma_j^2}{\sqrt{\sigma_j^2 + \lambda}}.$$

In comparison, the multiplier for $\|\theta_{\mathrm{lin}}\|_{\mathbf{A}} = \|\mathbf{A}^{1/2}\theta_{\mathrm{lin}}\|_2$ is

$$a_j := \sqrt{\sigma_j^2 + \lambda},$$

because, using the orthogonal invariance of the Euclidean norm,

$$\|\theta_{\mathrm{lin}}\|_{\mathbf{A}} = \|\mathbf{A}^{1/2}\theta_{\mathrm{lin}}\|_2 = \big\|\mathbf{V}(\mathbf{\Sigma}^\top\mathbf{\Sigma}+\lambda\mathbf{I}_d)^{1/2}\mathbf{V}^\top\theta_{\mathrm{lin}}\big\|_2 = \big\|(\mathbf{\Sigma}^\top\mathbf{\Sigma}+\lambda\mathbf{I}_d)^{1/2}\tilde\theta_{\mathrm{lin}}\big\|_2 = \Big(\sum_{j=1}^{d} a_j^2\,\tilde\theta_{\mathrm{lin},j}^2\Big)^{1/2}.$$

Now compare the *squared* norms coordinatewise:

$$\|\theta^\star\|_{\mathbf{A}}^2 = \sum_{j=1}^{d} h_j^2\,\tilde\theta_{\mathrm{lin},j}^2 = \sum_{j=1}^{d} \frac{\sigma_j^4}{\sigma_j^2+\lambda}\,\tilde\theta_{\mathrm{lin},j}^2, \qquad \|\theta_{\mathrm{lin}}\|_{\mathbf{A}}^2 = \sum_{j=1}^{d} a_j^2\,\tilde\theta_{\mathrm{lin},j}^2 = \sum_{j=1}^{d} (\sigma_j^2+\lambda)\,\tilde\theta_{\mathrm{lin},j}^2.$$

For each $j$,

$$\frac{h_j^2}{a_j^2} = \frac{\sigma_j^4/(\sigma_j^2+\lambda)}{\sigma_j^2+\lambda} = \frac{\sigma_j^4}{(\sigma_j^2+\lambda)^2} = \Big(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\Big)^2 \le 1,$$

with equality only as $\lambda \to 0$. In particular, if $\sigma_j = 0$ then $h_j = 0$ while $a_j = \sqrt{\lambda} > 0$. Hence each summand in $\|\theta^\star\|_{\mathbf{A}}^2$ is no larger than the corresponding summand in $\|\theta_{\mathrm{lin}}\|_{\mathbf{A}}^2$, and summing over $j$ gives

$$\|\theta^\star\|_{\mathbf{A}} \;\le\; \|\theta_{\mathrm{lin}}\|_{\mathbf{A}}.$$

□

Now we're ready to prove Theorem (3).

*Proof.* **First part of (A):** Define, for each sampled index $i_t$,

$$\mathbf{Y}_t \;:=\; \frac{1}{m\,p_{i_t}}\,\mathbf{A}^{-1/2}\,\mathbf{x}_{i_t}\mathbf{x}_{i_t}^\top\,\mathbf{A}^{-1/2} \;\in \mathbb{R}^{d\times d}, \qquad t=1,\dots,m.$$

Then $\mathbf{Y}_t \succeq 0$ and the $\mathbf{Y}_t$ are independent. Moreover, letting $\mathbf{M} := \mathbf{A}^{-1/2}\mathbf{X}^\top\mathbf{X}\mathbf{A}^{-1/2}$ we obtain

$$\mathbb{E}\left[\sum_{t=1}^m \mathbf{Y}_t\right] = \sum_{i=1}^n p_i \cdot \frac{1}{mp_i}\,\mathbf{A}^{-1/2}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{A}^{-1/2} = \frac{1}{m}\,\mathbf{A}^{-1/2}\mathbf{X}^\top\mathbf{X}\mathbf{A}^{-1/2} = \frac{1}{m}\,\mathbf{M}.$$

Under ridge–leverage sampling $p_i = \ell_i^{(\lambda)}/k_\lambda$ with $\ell_i^{(\lambda)} = \mathbf{x}_i^\top\mathbf{A}^{-1}\mathbf{x}_i$, each summand has the uniform spectral bound

$$\|\mathbf{Y}_t\|_2 = \frac{1}{mp_{i_t}}\,\|\mathbf{A}^{-1/2}\mathbf{x}_{i_t}\|_2^2 = \frac{1}{mp_{i_t}}\,\ell_{i_t}^{(\lambda)} = \frac{k_\lambda}{m}$$

Now, let $R := k_\lambda/m$. Since $\mathbf{X}^\top\mathbf{X} \preceq \mathbf{A} = \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d$, we have $0 \preceq \mathbf{M} \preceq \mathbf{I}_d$, hence $\lambda_{\max}(\mathbf{M}) \le 1$ and $\lambda_{\min}(\mathbf{M}) \in [0,1]$. Using Theorem 4 with Lemma 5 and the fact that $\lambda_{\max}(\mathbf{M}) \le 1$, and $R = k_\lambda/m$, we obtain

$$\Pr\Big[\,\lambda_{\max}\Big(\sum_{t=1}^m \mathbf{Y}_t\Big) \;\ge\; (1+\varepsilon)\,\lambda_{\max}(\mathbf{M})\,\Big] \;\le\; d\cdot\exp\Big(-\frac{m\,\varepsilon^2}{3k_\lambda}\Big),$$

$$\Pr\Big[\,\lambda_{\min}\Big(\sum_{t=1}^m \mathbf{Y}_t\Big) \;\le\; (1-\varepsilon)\,\lambda_{\min}(\mathbf{M})\,\Big] \;\le\; d\cdot\exp\Big(-\frac{m\,\varepsilon^2}{2k_\lambda}\Big).$$

Finally, note that the event $\|\sum_{t=1}^m \mathbf{Y}_t - \mathbf{M}\|_2 \;\ge\; \varepsilon$ implies that either

$$\lambda_{\max}\Big(\textstyle\sum_{t=1}^m \mathbf{Y}_t\Big) \ge (1+\varepsilon)\lambda_{\max}(\mathbf{M}) \;\text{ or }\; \lambda_{\min}\Big(\textstyle\sum_{t=1}^m \mathbf{Y}_t\Big) \le (1-\varepsilon)\lambda_{\min}(\mathbf{M}).$$

Using a union bound on the event $\|\sum_{t=1}^m \mathbf{Y}_t - \mathbf{M}\|_2 \;\ge\; \varepsilon$ gives

$$\Pr\Big[\,\big\|\textstyle\sum_{t=1}^m \mathbf{Y}_t - \mathbf{M}\big\|_2 \ge \varepsilon\,\Big] \;\le\; d\,\exp\Big(-\tfrac{m\,\varepsilon^2}{3k_\lambda}\Big) \;+\; d\,\exp\Big(-\tfrac{m\,\varepsilon^2}{2k_\lambda}\Big) \;\le\; 2d\cdot\exp\Big(-c\,\tfrac{m\,\varepsilon^2}{k_\lambda}\Big),$$

for some absolute $c \in (0, \frac{1}{2}]$ and all $\varepsilon \in (0, \frac{1}{2})$. Therefore, if

$$m \;\ge\; C\,\frac{k_\lambda + \log(2d/\delta)}{\varepsilon^2},$$

for a suitably large constant $C > 0$, the deviation event occurs with probability at most $\delta/2$, so with probability at least $1 - \delta/2$,

$$\Big\|\textstyle\sum_{t=1}^m \mathbf{Y}_t - \mathbf{M}\Big\|_2 \le \varepsilon.$$

Since $\mathbf{M} \succeq 0$, this implies

$$(1-\varepsilon)\,\mathbf{M} \;\preceq\; \sum_{t=1}^m \mathbf{Y}_t \;\preceq\; (1+\varepsilon)\,\mathbf{M}.$$

Undoing the $\mathbf{A}^{-1/2}$ normalization gives

$$(1-\varepsilon)\,\mathbf{X}^\top\mathbf{X} \;\preceq\; \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} \;\preceq\; (1+\varepsilon)\,\mathbf{X}^\top\mathbf{X},$$

and adding the ridge term $\lambda\mathbf{I}_d$ yields

$$(1-\varepsilon)\,\mathbf{A} \;\preceq\; \mathbf{A}_S \;\preceq\; (1+\varepsilon)\,\mathbf{A}.$$

**Second part of (A).** By first-order optimality of ridge regression,

$$\mathbf{A}\,\theta^\star = \mathbf{b}, \qquad \text{where} \quad \mathbf{A} = \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_d, \quad \mathbf{b} = \mathbf{X}^\top\mathbf{y}.$$

Under the realizable assumption $\mathbf{y} = \mathbf{X}\theta_{\text{lin}}$, we have

$$\mathbf{b} = \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top (\mathbf{X}\theta_{\text{lin}}) = \mathbf{X}^\top \mathbf{X}\, \theta_{\text{lin}}.$$

For the sampled system, recall that $\widetilde{\mathbf{y}} = \mathbf{WSy}$ and $\widetilde{\mathbf{X}} = \mathbf{WSX}$, so substituting the same realizable model gives

$$\mathbf{b}_S = \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{y}} = (\mathbf{WSX})^\top (\mathbf{WSy}) = \mathbf{X}^\top \mathbf{S}^\top \mathbf{W}^2 \mathbf{SX}\, \theta_{\text{lin}} = \widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}\, \theta_{\text{lin}}.$$

Hence,

$$\mathbf{b}_S - \mathbf{b} = (\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} - \mathbf{X}^\top \mathbf{X})\, \theta_{\text{lin}}.$$

We now bound this deviation in the $\mathbf{A}^{-1}$-norm:

$$\|\mathbf{b}_S - \mathbf{b}\|_{\mathbf{A}^{-1}} = \left\|\mathbf{A}^{-1/2}(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} - \mathbf{X}^\top \mathbf{X})\mathbf{A}^{-1/2}\, \mathbf{A}^{1/2}\theta_{\text{lin}}\right\|_2 \leq \left\|\mathbf{A}^{-1/2}(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} - \mathbf{X}^\top \mathbf{X})\mathbf{A}^{-1/2}\right\|_2 \|\theta_{\text{lin}}\|_{\mathbf{A}}.$$

By the spectral approximation established in the proof for the first part of (A),

$$\left\|\mathbf{A}^{-1/2}(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}} - \mathbf{X}^\top \mathbf{X})\mathbf{A}^{-1/2}\right\|_2 \;\leq\; \varepsilon \quad \text{with probability at least } 1 - \tfrac{\delta}{2}.$$

Combining these inequalities yields

$$\|\mathbf{b}_S - \mathbf{b}\|_{\mathbf{A}^{-1}} \;\leq\; \varepsilon\, \|\theta_{\text{lin}}\|_{\mathbf{A}},$$

which completes the proof of the second inequality in (A).

**First inequality in (Q):** Let $\Delta := \mathbf{A}_S - \mathbf{A}$ and $e := \mathbf{b}_S - \mathbf{b}$. Using $\mathbf{A}\theta^\star = \mathbf{b}$ and $\mathbf{A}_S\widehat{\theta} = \mathbf{b}_S$,

$$\widehat{\theta} - \theta^\star = \mathbf{A}_S^{-1}\big(\mathbf{b}_S - \mathbf{A}_S\theta^\star\big) = \mathbf{A}_S^{-1}\big(e - \Delta\theta^\star\big).$$

Taking the $\mathbf{A}$-norm and inserting $\mathbf{A}^{1/2}\mathbf{A}^{-1/2}$,

$$\|\widehat{\theta} - \theta^\star\|_{\mathbf{A}} \leq \|\mathbf{A}^{1/2}\mathbf{A}_S^{-1}\mathbf{A}^{1/2}\|_2 \left(\|e\|_{\mathbf{A}^{-1}} + \|\Delta\theta^\star\|_{\mathbf{A}^{-1}}\right).$$

From the spectral part of (A), $\mathbf{A}_S \succeq (1-\varepsilon)\mathbf{A}$, hence

$$\|\mathbf{A}^{1/2}\mathbf{A}_S^{-1}\mathbf{A}^{1/2}\|_2 \leq \frac{1}{1-\varepsilon}.$$

Also $-\varepsilon\mathbf{A} \preceq \Delta \preceq \varepsilon\mathbf{A}$, so

$$\|\Delta\theta^\star\|_{\mathbf{A}^{-1}} = \|\mathbf{A}^{-1/2}\Delta\mathbf{A}^{-1/2}\, \mathbf{A}^{1/2}\theta^\star\|_2 \leq \varepsilon\, \|\theta^\star\|_{\mathbf{A}}.$$

Combining this with the second part of (A), $\|e\|_{\mathbf{A}^{-1}} \leq \varepsilon\|\theta_{\text{lin}}\|_{\mathbf{A}}$, and Lemma 6 (which gives $\|\theta^\star\|_{\mathbf{A}} \leq \|\theta_{\text{lin}}\|_{\mathbf{A}}$), we obtain

$$\|\widehat{\theta} - \theta^\star\|_{\mathbf{A}} \leq \frac{1}{1-\varepsilon}\left(\varepsilon\|\theta_{\text{lin}}\|_{\mathbf{A}} + \varepsilon\|\theta^\star\|_{\mathbf{A}}\right) \leq \frac{2\varepsilon}{1-\varepsilon}\, \|\theta_{\text{lin}}\|_{\mathbf{A}} \leq 4\varepsilon\, \|\theta_{\text{lin}}\|_{\mathbf{A}},$$

since $\varepsilon < \frac{1}{2}$.

**Second inequality in (Q):** For the quadratic ridge objective,

$$R(\theta) - R(\theta^\star) = \tfrac{1}{2}\|\theta - \theta^\star\|_{\mathbf{A}}^2.$$

Hence,

$$R(\widehat{\theta}) - R(\theta^\star) = \tfrac{1}{2}\|\widehat{\theta} - \theta^\star\|_{\mathbf{A}}^2 \;\leq\; \tfrac{1}{2}\,(4\varepsilon)^2\, \|\theta_{\text{lin}}\|_{\mathbf{A}}^2 \;=\; 8\,\varepsilon^2\, \|\theta_{\text{lin}}\|_{\mathbf{A}}^2,$$

where we used the first inequality in (Q) to bound $\|\widehat{\theta} - \theta^\star\|_{\mathbf{A}} \leq 4\varepsilon\|\theta_{\text{lin}}\|_{\mathbf{A}}$.

□

## References


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

[2] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[3] Michael B Cohen, Cameron Musco, and Christopher Musco. Ridge leverage scores for low-rank approximation. *arXiv preprint arXiv:1511.07263*, 6, 2015.

[4] Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

[5] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

[6] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[7] Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. Active learning for speech recognition: the power of gradients. *arXiv preprint arXiv:1612.03226*, 2016.

[8] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J Spanos, and Dawn Song. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*, 2019.

[9] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

[10] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

[11] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[12] Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*, 2021.

[13] Yongchan Kwon and James Zou. Data-oob: Out-of-bag estimate as a simple and efficient data value. In *International conference on machine learning*, pages 18135–18152. PMLR, 2023.

[14] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[15] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020.

[16] Cameron Musco and Christopher Musco. Recursive sampling for the nystrom method. *Advances in neural information processing systems*, 30, 2017.

[17] Christopher Musco and R Teal Witter. Provably accurate shapley value estimation via leverage score sampling. *arXiv preprint arXiv:2410.01917*, 2024.

[18] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

[19] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[20] Burr Settles. Active learning literature survey. 2009.

[21] Lloyd S Shapley et al. A value for n-person games. 1953.

[22] Tianshu Song, Yongxin Tong, and Shuyue Wei. Profit allocation for federated learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2577–2586. IEEE, 2019.

[23] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[24] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[25] Jiachen T Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. *arXiv preprint arXiv:2406.11011*, 2024.

[26] R Teal Witter, Yurong Liu, and Christopher Musco. Regression-adjusted monte carlo estimators for shapley values and probabilistic values. *arXiv preprint arXiv:2506.11849*, 2025.

[27] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.

[28] Xinyi Xu, Lingjuan Lyu, Xingjun Ma, Chenglin Miao, Chuan Sheng Foo, and Bryan Kian Hsiang Low. Gradient driven rewards to guarantee fairness in collaborative machine learning. *Advances in Neural Information Processing Systems*, 34:16104–16117, 2021.

[29] Xinkai Yuan, Zilinghan Li, and Gaoang Wang. Activematch: End-to-end semi-supervised active representation learning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1136–1140. IEEE, 2022.