Open Problem: Data Selection for Regression Tasks

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Department of Computer Science, Purdue University

Shay Moran SMORAN@TECHNION.AC.IL
Departments of Mathematics, Computer Science, and Data and Decision Sciences, Technion and Google Research

Alexander Shlimovich ASHLIMOVICH@CAMPUS.TECHNION.AC.IL
Department of Mathematics, Technion

Amir Yehudayoff YEHUDAYOFF@TECHNION.AC.IL
Department of Mathematics, Technion and Department of Computer Science, Copenhagen University

Editors: Nika Haghtalab and Ankur Moitra

Abstract

This note proposes a set of open problems concerning data selection in regression tasks. The central question is: given a natural learning rule $\mathcal{A}$ and a selection budget n, how well can $\mathcal{A}$ perform when trained on n examples selected from a larger dataset? We present concrete instances of this question in basic regression settings, including mean estimation and linear regression.

Keywords: Data Selection, Empirical Risk Minimizers, Compression Schemes, Machine Teaching, Active Learning.

1. Introduction

Machine learning advances along two main axes: *data* and *algorithms*. Much of the theoretical focus, however, has been on optimizing the algorithmic side, typically assuming that data is drawn i.i.d. from a fixed distribution. As a result, the role of data has received comparatively less theoretical attention.

Nevertheless, the selection of data is no less central. In practice, better data is often key to better performance, and the choice of data can be as consequential as the choice of learning algorithm. Moreover, resource constraints frequently limit how much data can be labeled, stored, or processed—raising the question of which examples should be prioritized for training.

Inspired by these challenges, we propose the following basic theoretical question: given a dataset D of N examples, a selection budget $n \ll N$, and a fixed natural learning rule A—how well can A perform when trained only on n examples from D, in terms of the loss on the full dataset?

Overarching Question: *Given a natural learning rule A and a selection budget n, how well can A perform on the full dataset when trained on only n selected examples?*

This formulation isolates an information-theoretic aspect of data selection, connecting it to classical topics such as sample compression (Littlestone and Warmuth, 1986), subsampling and active learning (Cohn et al., 1994; Hanneke, 2014), and coreset theory (Feldman, 2020; Maalouf et al., 2024). Yet even in simple settings like mean estimation or linear regression, the achievable performance under optimal subset selection remains poorly understood.

2. Mean Estimation

We begin with a simple and basic instance of the data selection problem in mean estimation. Let $D = \{z_1, \ldots, z_N\} \subseteq \mathbb{R}^d$ be a multiset, and define the squared loss of an hypothesis $h \in \mathbb{R}^d$ by: $L_D(h) = \frac{1}{N}\sum_{i=1}^{N} \|h - z_i\|_2^2$. The unique minimizer of this loss is the mean $h^\star = \frac{1}{N}\sum_{i=1}^{N} z_i$. Accordingly, we consider the learning rule that returns the average of its input dataset. Specialized to this setting, our main question becomes: how well can the average of n selected points approximate the performance of the true mean? Define:

$$L_D^\star = \min_{h \in \mathbb{R}} L_D(h), \qquad L_D^\star(n) = \min_{z_1,\ldots,z_n \in D} L_D\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right).$$

The quantity $L_D^\star(n)$ represents the best loss one can achieve on the full dataset by averaging only n selected examples.

Question 1 (Mean Estimation in $\mathbb{R}^d$) *What is the value of*

$$\mathcal{F}(d, n) = \sup_{D \subseteq \mathbb{R}^d} \frac{L_D^\star(n)}{L_D^\star},$$

for fixed $d > 1$ and $n > 1$?

The 1D case is resolved by Hanneke, Moran, Shlimovich, and Yehudayoff (2025), but the behavior in higher dimensions remains open. A summary of known bounds for various values of d and n is provided in the following table. Except for the case $d = n = 2$, all bounds are due to Hanneke et al. (2025). Our derivation of the $d = n = 2$ bound is somewhat tedious and computational—though it relies only on basic classical geometry—and we are not aware of a simple proof. Finding one may be of interest, as it could shed light on more general cases.

$d \backslash n$	1	2	3	n	$n \to \infty$
1	2	$\frac{4}{3}$	$\frac{6}{5}$	$\frac{2n}{2n-1}$	1
2	2	$\frac{4}{3}$	?	?	1
3	2	?	?	?	1
$d \to \infty$	2	$\frac{3}{2}$	$\frac{4}{3}$	$\frac{n+1}{n}$	undefined

3. Weighted Data Selection

We next consider a convex relaxation of the data selection problem, in which the learner may assign non-negative weights to selected examples. This weighted setting is also studied in the context of *coresets* (Feldman, 2020; Lucic et al., 2018), and often enables stronger approximation guarantees using fewer examples. Formally, let $\mathcal{Z}$ denote the example space, $\mathcal{W}$ the parameter space, and let A

be a fixed learning rule that returns a parameter $w \in \mathcal{W}$ minimizing the loss with respect to a given (convex) combination of individual losses. Given a dataset $D \subseteq \mathcal{Z}$, define

$$L_D^\star(n; A) = \inf_{\substack{z_1,\dots,z_n \in D \\ F \in \mathrm{conv}(\ell_{z_1},\dots,\ell_{z_n})}} L_D\left(A(F)\right),$$

where $\ell_z \colon \mathcal{W} \to \mathbb{R}_{\geq 0}$ is the loss associated with $z \in D$, and the full-dataset loss is defined by $L_D(w) = \frac{1}{|D|} \sum_{z \in D} \ell_z(w)$. In this formulation, the learner chooses a convex combination F of n example losses and then applies A to minimize this combination. The goal remains to achieve good performance (low loss) on the full dataset using only these n weighted examples.

3.1. Linear Regression

Consider linear regression with squared loss:

$$\ell_{(x,y)}(w) = (w^\top x - y)^2, \quad \text{for } (x, y) \in \mathbb{R}^d \times \mathbb{R}.$$

Assume the learning rule $A^\star$ is the *min-norm ERM*, which selects the empirical risk minimizer of smallest Euclidean norm. Hanneke et al. (2025) proved the following result in this context:

Theorem 1 (Linear Regression (Hanneke et al., 2025)) *Let $A^\star$ be the min-norm ERM. Then:*

$$\sup_D \frac{L_D^\star(n; A^\star)}{L_D^\star} = \begin{cases} 1 & \text{if } n \geq 2d, \\ d+1 & \text{if } n = d, \\ \infty & \text{if } n < d. \end{cases}$$

Question 2 (Weighted Data Selection Gap) *What is the value of*

$$\sup_D \frac{L_D^\star(n; A^\star)}{L_D^\star}$$

for $d < n < 2d$?

Theorem 1 implies that this quantity lies in the interval $(1, d+1]$. In the case $n = 2d - 1$, we believe we have an argument showing that the value is exactly $1 + \frac{1}{d}$, though the argument is somewhat ad hoc and does not clearly extend to other values of n.

The case of $n > 2d$ in Theorem 1 relies on a generalization of Carathéodory's Theorem due to Steinitz Steinitz (1916), while the bound for $n = d$ relies on determinantal point processes Derezinski and Warmuth (2017). The difference between these techniques suggests that resolving the intermediate regime may require new ideas that bridge these two approaches; see Hanneke et al. (2025) for further details.

4. Vector-Valued Linear Regression

We conclude by formulating a unified setting that generalizes both mean estimation and linear regression. Let $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}^m$, and consider linear predictors $W : \mathbb{R}^d \to \mathbb{R}^m$ with squared loss:

$$\ell_z(W) = \|Wx - y\|^2,$$

where $z = (x, y) \in \mathcal{Z}$. Given a dataset $D = \{z_i\}_{i=1}^N \subseteq \mathcal{Z}$, define the average loss over D as

$$L_D(W) = \frac{1}{N}\sum_{i=1}^{N} \ell_{z_i}(W), \quad \text{and let } L_D^\star = \min_W L_D(W).$$

This setting recovers: (i) mean estimation when[1] $d = 0$, and standard linear regression when $m = 1$.

Throughout this section, we consider the empirical risk minimizer (ERM) A that selects the minimizer with the smallest Frobenius norm (i.e., the minimum-norm solution). However, the questions posed below are of broader interest, including for other forms of ERM selection.

As before, we consider two variants of data selection:

Unweighted selection. Select n examples from D, apply the algorithm A, and evaluate its performance on the full dataset:

$$L_D^\star(n) = \min_{z_1,\ldots,z_n \in D} L_D\left(A\left(\frac{1}{n}\sum_{i=1}^{n} \ell_{z_i}\right)\right).$$

Weighted selection. Select n examples from D, form a convex combination of the corresponding losses, apply A, and evaluate on the full dataset:

$$L_D^\star(n; \text{weighted}) = \min_{\substack{z_1,\ldots,z_n \in D \\ F \in \text{conv}(\ell_{z_1},\ldots,\ell_{z_n})}} L_D\left(A(F)\right).$$

Question 3 (Unweighted General Linear Prediction) *Given d, m, n, what is the value of*

$$\mathcal{F}(d, m, n) = \sup_{D \subseteq \mathbb{R}^d \times \mathbb{R}^m} \frac{L_D^\star(n)}{L_D^\star}?$$

Question 4 (Weighted General Linear Prediction) *Given d, m, n, what is the value of*

$$\mathcal{F}_{\textit{weighted}}(d, m, n) = \sup_{D \subseteq \mathbb{R}^d \times \mathbb{R}^m} \frac{L_D^\star(n; \textit{weighted})}{L_D^\star}?$$

In particular, what is the minimal $n = n(d, m)$ such that this ratio equals 1?

1. By interpreting $\mathbb{R}^0$ as a singleton $\{1\}$, we get that $d = 0$ corresponds to constant predictors $W : \mathbb{R}^0 \to \mathbb{R}^m$, recovering mean estimation.

References

D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

Michal Derezinski and Manfred K. K Warmuth. Unbiased estimates for linear regression via volume sampling. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/54e36c5ff5f6a1802925ca009f3ebb68-Paper.pdf.

Dan Feldman. Introduction to core-sets: an updated survey, 2020. URL https://arxiv.org/abs/2011.09384.

S. Hanneke. Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2–3):131–309, 2014.

Steve Hanneke, Shay Moran, Alexander Shlimovich, and Amir Yehudayoff. Data selection for erms, 2025. URL https://arxiv.org/abs/2504.14572.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. *Unpublished manuscript*, 1986.

Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training gaussian mixture models at scale via coresets. *Journal of Machine Learning Research*, 18(160):1–25, 2018. URL http://jmlr.org/papers/v18/15-506.html.

Alaa Maalouf, Gilad Eini, Ben Mussay, Dan Feldman, and Margarita Osadchy. A unified approach to coreset learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):6893–6905, May 2024. ISSN 2162-237X. doi: 10.1109/TNNLS.2022.3213169. Publisher Copyright: © 2012 IEEE.

E. Steinitz. Bedingt konvergente reihen und konvexe systeme. (schluß.). *Journal für die reine und angewandte Mathematik*, 146:1–52, 1916. URL http://eudml.org/doc/149441.