Importance Sampling for Nonlinear Models

Fred Roosta

School of Mathematics and Physics
University of Queensland

Prakash P. Rajmohan and Fred Roosta (2025). "Importance Sampling for Nonlinear Models". In: *Forty-second International Conference on Machine Learning (ICML)*. URL: https://openreview.net/forum?id=fY7KjywLK1

Optimization Problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(f_i(\boldsymbol{\theta})) \right\},$$

- ℓ is a loss function
- $f_i : \mathbb{R}^p \to \mathbb{R}$ is a (nonlinear) model evaluated at the i^{th} data point
- *Large datasets* and *underparameterized*: $n \gg p$

Example: Linear Regression

In simple linear regression:

- $f_i(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{x}_i \rangle - y_i$
- $\ell(t) = t^2$ (squared loss)

$$\mathcal{L}(\boldsymbol{\theta}) = \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|^2$$

where

- $\mathbf{X} \in \mathbb{R}^{n \times p}$ has rows $\mathbf{x}_i^\mathsf{T}$
- $\mathbf{y} \in \mathbb{R}^n$ has entries y_i

Large Datasets

Large datasets ($n \gg 1$) come with computational challenges:

- training, and/or
- post-training diagnostic analysis

$$\textbf{Linear Settings} \implies \textbf{RandNLA}$$

RandNLA: Successes

- Major advances in core matrix problems:
  - matrix multiplication
  - least-squares / least-absolute deviation
  - low-rank approximation
  - and many more ...
- Broad toolkit developed:
  - oblivious methods (sketching, projections),
  - non-oblivious methods (leverage, row-norm sampling).
- Provides strong **approximation guarantees**: for any small ε,

$$\mathcal{L}(\boldsymbol{\theta}^\star_{\mathcal{S}}) \leq \mathcal{L}(\boldsymbol{\theta}^\star) + \mathcal{O}(\varepsilon)$$

$\boldsymbol{\theta}^\star$: optimal parameter from full dataset

$\boldsymbol{\theta}^\star_{\mathcal{S}}$: optimal parameter from subsampled dataset

RandNLA: Limitations

- Hard to extend RandNLA ideas to truly **nonlinear** objectives.
- Existing nonlinear methods are mostly **ad hoc** and limited[1]
- Lack of a **general framework** to extend RandNLA toolkits (e.g., leverage score sampling) to nonlinear structures with strong guarantees.

[1] (Avron, Kapralov, Musco, Musco, Velingker, and Zandieh, 2017; Avron, Kapralov, Musco, Musco, Velingker, and Zandieh, 2019; Erdélyi, Musco, and Musco, 2020; Gajjar and Musco, 2021)

Revisit: Linear Regression

Consider linear least squares:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|^2 = \sum_{i=1}^{n} \overbrace{\frac{1}{2}(\langle \mathbf{x}_i, \boldsymbol{\theta}\rangle - y_i)^2}^{\ell_i(\boldsymbol{\theta})}$$

- **Hessians:**

$$\nabla^2 \ell_i(\boldsymbol{\theta}) = \mathbf{x}_i \mathbf{x}_i^\mathsf{T}, \qquad \nabla^2 \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\mathsf{T} = \mathbf{X}^\mathsf{T}\mathbf{X}$$

- Suppose $\mathbf{X} \in \mathbb{R}^{n \times p}$ has full column rank.
- i^{th} **leverage score:**

$$\tau_i = \mathbf{x}_i^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_i = \text{Trace}((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T)$$

Going Beyond Linear Regression

- Assume each $\ell_i(\boldsymbol{\theta})$ is convex
- Nonlinear "Leverage" score (?):

$$\tau_i(\boldsymbol{\theta}) \triangleq \text{Trace}\Big([\nabla^2 \mathcal{L}(\boldsymbol{\theta})]^\dagger \nabla^2 \ell_i(\boldsymbol{\theta})\Big)$$

- Intuition: Consider the Riemannian manifold $\mathcal{M}$ equipped with metric

$$g_{\boldsymbol{\theta}}(\mathbf{v}, \mathbf{v}) = \mathbf{v}^T \nabla^2 \mathcal{L}(\boldsymbol{\theta})\, \mathbf{v}, \quad \mathbf{v} \in T_{\boldsymbol{\theta}}(\mathcal{M}).$$

  So, $\tau_i(\boldsymbol{\theta})$ measures how much removing individual components ℓ_i changes curvature (locally)

- Properties:

$$0 \leq \tau_i(\boldsymbol{\theta}) \leq \text{Rank}(\nabla^2 \ell_i(\boldsymbol{\theta})), \quad \text{and} \quad \sum_{i=1}^{n} \tau_i(\boldsymbol{\theta}) = \text{Rank}(\nabla^2 \mathcal{L}(\boldsymbol{\theta})).$$

- This did **not** lead us to guarantees of the kind:

$$\mathcal{L}(\boldsymbol{\theta}^\star_{\mathcal{S}}) \leq \mathcal{L}(\boldsymbol{\theta}^\star) + \mathcal{O}(\varepsilon).$$

Revisit: Linear Regression (Take 2)

Consider linear least squares:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|^2 = \frac{1}{2}\sum_{i=1}^{n} \overbrace{\left(\langle \mathbf{x}_i, \boldsymbol{\theta}\rangle - y_i\right)^2}^{f_i^2(\boldsymbol{\theta})}$$

- With

$$\widehat{\mathbf{x}} = [\mathbf{x} \mid -y], \quad \widehat{\boldsymbol{\theta}} = [\boldsymbol{\theta}; 1]$$

  we can write

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n} \left\langle \widehat{\mathbf{x}}_i, \widehat{\boldsymbol{\theta}} \right\rangle^2 = \frac{1}{2}\left\langle \widehat{\mathbf{X}}\widehat{\boldsymbol{\theta}}, \widehat{\mathbf{X}}\widehat{\boldsymbol{\theta}} \right\rangle$$

- The entire objective is expressed via an **inner product**.
- This structure enables the use of linear **subspace embeddings**:

$$(1-\varepsilon)\widehat{\mathbf{X}}^\intercal\widehat{\mathbf{X}} \preceq \widehat{\mathbf{X}}_{\mathcal{S}}^\intercal\widehat{\mathbf{X}}_{\mathcal{S}} \preceq (1+\varepsilon)\widehat{\mathbf{X}}^\intercal\widehat{\mathbf{X}}$$

Key Idea: A Nonlinear Inner-Product Analogue

- Nonlinear models lose the inner-product structure.
- We reintroduce an *inner-product-like* form via a new operator.
- This operator captures the nonlinearity while maintaining a linear-like structure.
- We call it the **nonlinear adjoint operator**.

Deriving the Nonlinear Adjoint

- Let $f : \mathbb{R}^p \to \mathbb{R}$ be continuously differentiable.
- Mean value theorem gives

$$f(\boldsymbol{\theta}) = f(\mathbf{0}) + \int_0^1 \left\langle \frac{\partial}{\partial \boldsymbol{\theta}} f(t\boldsymbol{\theta}), \boldsymbol{\theta} \right\rangle \mathrm{d}t.$$

- Equivalently,

$$f(\boldsymbol{\theta}) = f(\mathbf{0}) + \left\langle \boldsymbol{\theta}, \int_0^1 \frac{\partial}{\partial \boldsymbol{\theta}} f(t\boldsymbol{\theta}) \mathrm{d}t \right\rangle.$$

- This looks like a dual pairing.

Definition: Nonlinear Adjoint Operator

Definition (Adjoint Operator)

Let $f : \mathbb{R}^p \to \mathbb{R}$. The adjoint operator is

$$\mathbf{f}^\star(\boldsymbol{\theta}) \triangleq \int_0^1 \frac{\partial}{\partial \boldsymbol{\theta}} f(t\boldsymbol{\theta}) \mathrm{d}t,$$

whenever $t \mapsto f(t\boldsymbol{\theta})$ is absolutely continuous.

- Absolute continuity allows a.e. differentiability.
- Extends beyond smooth settings.

Inner-Product Representation

- With the adjoint operator,

$$f(\boldsymbol{\theta}) = \left\langle \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{f}}^{\star}(\boldsymbol{\theta}) \right\rangle,$$

where

$$\widehat{\mathbf{f}}^{\star}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{f}^{\star}(\boldsymbol{\theta}) \\ f(\mathbf{0}) \end{bmatrix}, \qquad \widehat{\boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{\theta} \\ 1 \end{bmatrix}.$$

- If $f(\mathbf{0}) = 0$, this becomes $f(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{f}^{\star}(\boldsymbol{\theta}) \rangle$.
- A nonlinear analogue of Riesz representation[2].

[2] Continuous linear functionals can be identified with inner products against a unique representer in the Hilbert space.

Computing the Adjoint Operator

- In general, $\mathbf{f}^\star(\boldsymbol{\theta})$ involves an integral.
- Numerical approximation is possible if needed.
- Many ML models yield closed forms.

Proposition

Let $f = g \circ h$, where

- $g : \mathbb{R} \to \mathbb{R}$,
- $h : \mathbb{R}^p \to \mathbb{R}$ *is positively homogeneous of degree* α.

Then

$$\mathbf{f}^\star(\boldsymbol{\theta}) = \begin{cases} \left(\dfrac{g(h(\boldsymbol{\theta})) - g(0)}{\alpha\, h(\boldsymbol{\theta})} \right) \dfrac{\partial}{\partial \boldsymbol{\theta}} h(\boldsymbol{\theta}), & h(\boldsymbol{\theta}) \neq 0, \\ \left(\dfrac{g'(0)}{\alpha} \right) \dfrac{\partial}{\partial \boldsymbol{\theta}} h(\boldsymbol{\theta}), & h(\boldsymbol{\theta}) = 0. \end{cases}$$

Example: Generalized Linear Predictors

- Generalized linear predictor models (single-index models):

$$f(\boldsymbol{\theta}) = \phi(\langle \boldsymbol{\theta}, \mathbf{x} \rangle), \quad \mathbf{x} \in \mathbb{R}^p.$$

- Here $f = g \circ h$ with $g(t) = \phi(t)$ and $h(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle$.
- Homogeneity degree: $\alpha = 1$.
- Adjoint:

$$\mathbf{f}^\star(\boldsymbol{\theta}) = \left(\frac{\phi(\langle \boldsymbol{\theta}, \mathbf{x} \rangle) - \phi(0)}{\langle \boldsymbol{\theta}, \mathbf{x} \rangle} \right) \mathbf{x}.$$

- Lifted adjoint:

$$\widehat{\mathbf{f}}^\star(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{f}^\star(\boldsymbol{\theta}) \\ \phi(0) \end{bmatrix}.$$

- Scales $\mathbf{x}$ by the local nonlinearity of ϕ.

Example: Logistic Regression

- Logistic function:

$$\phi(t) = \frac{1}{1 + e^{-t}}.$$

- Adjoint:

$$\mathbf{f}^{\star}(\boldsymbol{\theta}) = -\left(\frac{\tanh(\langle \boldsymbol{\theta}, \mathbf{x} \rangle / 2)}{2 \langle \boldsymbol{\theta}, \mathbf{x} \rangle}\right) \mathbf{x}.$$

- Smooth, bounded scaling vs. linear regression's constant adjoint.

Example: ReLU Neural Networks (Single Neuron)

- Let
$$r(\boldsymbol{\theta}) = \phi(\psi(\boldsymbol{\theta})), \quad \psi(\boldsymbol{\theta}) = a \cdot \max\{\langle \mathbf{b}, \mathbf{x} \rangle, 0\},$$
with $\boldsymbol{\theta} = [a, \mathbf{b}]$.

- ψ has homogeneity degree $\alpha = 2$.

- Adjoint:
$$\mathbf{r}^\star(\boldsymbol{\theta}) = \frac{\phi(a \max\{\langle \mathbf{b}, \mathbf{x} \rangle, 0\}) - \phi(0)}{2a \max\{\langle \mathbf{b}, \mathbf{x} \rangle, 0\}} \begin{bmatrix} \max\{\langle \mathbf{b}, \mathbf{x} \rangle, 0\} \\ a\,\mathbf{x}\, \mathbb{1}(\langle \mathbf{b}, \mathbf{x} \rangle > 0) \end{bmatrix}.$$

- For $\phi(z) = z$:
$$\mathbf{r}^\star(\boldsymbol{\theta}) = \frac{1}{2} \begin{bmatrix} \max\{\langle \mathbf{b}, \mathbf{x} \rangle, 0\} \\ a\,\mathbf{x}\, \mathbb{1}(\langle \mathbf{b}, \mathbf{x} \rangle > 0) \end{bmatrix}.$$

ReLU Networks and Additivity

- Two-layer RelU Network (m hidden neurons and a single output):

$$f(\boldsymbol{\theta}) = \sum_{j=1}^{m} \phi(a_j \max\{\langle \mathbf{b}_j, \mathbf{x} \rangle, 0\}).$$

- Adjoint is additive:

$$\mathbf{f}^\star(\boldsymbol{\theta}) = \sum_{j=1}^{m} \mathbf{e}_j \otimes \mathbf{r}^\star(\boldsymbol{\theta}_j).$$

- Lifted adjoint:

$$\widehat{\mathbf{f}}^\star(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{f}^\star(\boldsymbol{\theta}) \\ m\,\phi(0) \end{bmatrix}.$$

From Linear to Nonlinear Least Squares

Consider nonlinear least-squares:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(f_i(\boldsymbol{\theta})\right)^2,$$

Since $f_i(\boldsymbol{\theta}) = \langle \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{f}}_i^{\star}(\boldsymbol{\theta}) \rangle$, we can write

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \langle \widehat{\boldsymbol{\theta}}, \widehat{\mathbf{f}}_i^{\star}(\boldsymbol{\theta}) \rangle^2 = \left\| \widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})\, \widehat{\boldsymbol{\theta}} \right\|^2,$$

where

$$\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta}) \triangleq \begin{bmatrix} \widehat{\mathbf{f}}_1^{\star}(\boldsymbol{\theta}) & \widehat{\mathbf{f}}_2^{\star}(\boldsymbol{\theta}) & \cdots & \widehat{\mathbf{f}}_n^{\star}(\boldsymbol{\theta}) \end{bmatrix}^{\mathsf{T}}.$$

is the **nonlinear dual matrix**.

This mirrors the linear form $\|\mathbf{X}\boldsymbol{\theta}\|^2$, with $\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})$ acting as the data matrix.

Key point: Tools from RandNLA can now be used.

Nonlinear Importance Sampling

We can define importance scores from the rows of $\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})$.

- **Nonlinear Leverage Scores**:

$$\tau_i(\boldsymbol{\theta}) \triangleq \frac{\langle \mathbf{e}_i, \widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})\big[\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})\big]^{\dagger}\mathbf{e}_i\rangle}{\mathrm{Rank}(\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta}))}.$$

- **Nonlinear Row-norm Scores**:

$$\tau_i(\boldsymbol{\theta}) \triangleq \frac{\|\widehat{\mathbf{f}}_i^{\star}(\boldsymbol{\theta})\|_2^2}{\|\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})\|_{\mathrm{F}}^2}.$$

Both score types define valid distributions:

$$\tau_i(\boldsymbol{\theta}) \geq 0, \qquad \sum_{i=1}^{n} \tau_i(\boldsymbol{\theta}) = 1.$$

Remark: For linear models $f(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle$, these scores reduce to the classical ones.

Sampling and Approximate Loss

Let $\mathcal{S}$ be a multiset of s indices sampled with replacement using $\{\tau_i(\boldsymbol{\theta})\}$.

Define the sampled loss

$$\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) \triangleq \sum_{i \in \mathcal{S}} \frac{(f_i(\boldsymbol{\theta}))^2}{s\,\tau_i(\boldsymbol{\theta})}.$$

This estimator is unbiased:

$$\mathbb{E}\,\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}).$$

Let $\widehat{\mathbf{F}}^{\star}_{\mathcal{S}}(\boldsymbol{\theta})$ be the sampled/rescaled rows of $\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})$.

Then

$$\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) = \left\| \widehat{\mathbf{F}}^{\star}_{\mathcal{S}}(\boldsymbol{\theta})\,\widehat{\boldsymbol{\theta}} \right\|^2.$$

Approximation Guarantee

For fixed $\boldsymbol{\theta}$, if

$$s = \widetilde{\mathcal{O}}\left(p/\varepsilon^2\right),$$

then with probability at least $1-\delta$, for all $\mathbf{v} \in \mathbb{R}^{p+1}$,

$$(1-\varepsilon)\,\|\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})\mathbf{v}\|^2 \leq \|\widehat{\mathbf{F}}^{\star}_{\mathcal{S}}(\boldsymbol{\theta})\mathbf{v}\|^2 \leq (1+\varepsilon)\,\|\widehat{\mathbf{F}}^{\star}(\boldsymbol{\theta})\mathbf{v}\|^2$$

As a direct consequence, with probability at least $1-\delta$,

$$(1-\varepsilon)\,\mathcal{L}(\boldsymbol{\theta}) \leq \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) \leq (1+\varepsilon)\,\mathcal{L}(\boldsymbol{\theta})$$

What Is Missing?

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \ \mathcal{L}(\boldsymbol{\theta})$$

$$\boldsymbol{\theta}^\star_{\mathcal{S}} = \arg\min_{\boldsymbol{\theta}} \ \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta})$$

$$\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}^\star_{\mathcal{S}}) \leq \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}^\star) \leq (1+\varepsilon)\mathcal{L}(\boldsymbol{\theta}^\star)$$

But this is *not yet* the desired guarantee. Our goal is

$$\mathcal{L}(\boldsymbol{\theta}^\star_{\mathcal{S}}) \leq \mathcal{L}(\boldsymbol{\theta}^\star) + \mathcal{O}(\varepsilon)\,.$$

Two fundamental obstacles remain:

1. Importance scores depend on $\boldsymbol{\theta}^\star$, which is unknown.
2. The bound involves $\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}^\star_{\mathcal{S}})$, not $\mathcal{L}(\boldsymbol{\theta}^\star_{\mathcal{S}})$.

Challenge 1: Unknown Sampling Scores

The inequality

$$\mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}^{\star}_{\mathcal{S}}) \leq (1+\varepsilon)\mathcal{L}(\boldsymbol{\theta}^{\star}),$$

requires evaluating

$$\tau_i(\boldsymbol{\theta}^{\star}),$$

which is infeasible since $\boldsymbol{\theta}^{\star}$ is unknown.

Strategy: approximate nonlinear scores by *parameter-independent* quantities.

Near-optimal Sampling in RandNLA

From RandNLA: if

$$\beta\,\tau_i \leq \hat{\tau}_i,$$

for some $\beta \in (0, 1)$, then subspace embedding holds with sample size

$$\widetilde{\mathcal{O}}\left(p/(\beta\varepsilon^2)\right).$$

If we can approximate *nonlinear* scores similarly, guarantees carry over.

For specific model classes, the structure of the nonlinear adjoint operator allows parameter-independent upper bounds on nonlinear scores.

Example: Generalized Linear Predictors

Generalized linear predictor models $f(\boldsymbol{\theta}) = \phi(\langle \boldsymbol{\theta}, \mathbf{x} \rangle)$ with ϕ satisfying

$$\ell \leq \frac{\phi^2(t)}{t^2} \leq u \quad \text{for } t \in \mathcal{T}.$$

Then, for full-rank $\mathbf{X}$ and $\mathbf{F}^\star(\boldsymbol{\theta})$,

$$\ell\, \mathbf{X}^\mathsf{T}\mathbf{X} \preceq \left[\mathbf{F}^\star(\boldsymbol{\theta})\right]^\mathsf{T}\mathbf{F}^\star(\boldsymbol{\theta}) \preceq u\, \mathbf{X}^\mathsf{T}\mathbf{X}.$$

As a consequence,

$$\tau_i(\boldsymbol{\theta}) \leq \frac{u}{\ell}\, \tau_i,$$

where τ_i is the *linear* leverage score from $\widehat{\mathbf{X}}$.

Example: ReLU Neural Networks

Now consider a one-hidden-layer ReLU network with row-norm sampling.

Assume the activation satisfies

$$c_1 \leq \frac{(\phi(t) - \phi(0))^2}{t^2} \leq c_2 \quad \text{for } t \in \mathcal{T}.$$

Under mild assumptions on $\boldsymbol{\theta}^\star$, define a compact set $\mathcal{C}$ containing it.

For all $\boldsymbol{\theta} \in \mathcal{C}$,

$$c_1 \ell \|\mathbf{x}_i\|^2 \leq \|\mathbf{f}_i^\star(\boldsymbol{\theta})\|^2 \leq c_2 u \|\mathbf{x}_i\|^2,$$

where ℓ and u are used in the definition of $\mathcal{C}$.

Hence,

$$\tau_i(\boldsymbol{\theta}) \leq \frac{\max\{c_2 u, 1\}}{\min\{c_1 \ell, 1\}} \tau_i,$$

where τ_i is a row-norm score from $\widehat{\mathbf{X}}$.

Challenge 2: Relating Sampled and Full Loss

To complete the argument, we need a lower bound of the form

$$\mathcal{L}(\boldsymbol{\theta}^\star_{\mathcal{S}}) \lesssim \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}^\star_{\mathcal{S}})$$

This requires uniform control over $\boldsymbol{\theta}$.

Consider a ball $\mathcal{B}^\star_R$ containing $\boldsymbol{\theta}^\star$. Construct an ε-net $\mathcal{N}_\varepsilon \subset \mathcal{B}^\star_R$ such that:

$$\forall \boldsymbol{\theta} \in \mathcal{B}^\star_R, \quad \exists \boldsymbol{\theta}' \in \mathcal{N}_\varepsilon \text{ with } \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \leq \varepsilon R.$$

The net size satisfies $|\mathcal{N}_\varepsilon| = \Omega(1/\varepsilon^p)$. By choosing

$$\delta' = \delta/|\mathcal{N}_\varepsilon|,$$

a union bound ensures that

$$(1-\varepsilon)\mathcal{L}(\boldsymbol{\theta}) \leq \mathcal{L}_{\mathcal{S}}(\boldsymbol{\theta}) \leq (1+\varepsilon)\mathcal{L}(\boldsymbol{\theta})$$

holds *simultaneously* for all $\boldsymbol{\theta} \in \mathcal{N}_\varepsilon$ with probability $1-\delta$.

Challenge 2: Relating Sampled and Full Loss

Assume $\mathcal{L}_S$ is continuous on $\mathcal{B}_R^\star$. By compactness, $\mathcal{L}_S$ is Lipschitz:

$$|\mathcal{L}_S(\boldsymbol{\theta}) - \mathcal{L}_S(\boldsymbol{\theta}')| \leq L(f, \mathbf{X}, R)\, \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

This allows us to move from the net to arbitrary $\boldsymbol{\theta}$.

Using Lipschitz continuity and the net approximation,

$$\mathcal{L}_S(\boldsymbol{\theta}_S^\star) \geq (1-\varepsilon)\mathcal{L}(\boldsymbol{\theta}_S^\star) - \varepsilon(2-\varepsilon) R\, L(f, \mathbf{X}, R).$$

Putting it all together, we obtain

$$\mathcal{L}(\boldsymbol{\theta}_S^\star) \leq \mathcal{L}(\boldsymbol{\theta}^\star) + \frac{\varepsilon}{1-\varepsilon}\Big(\mathcal{L}(\boldsymbol{\theta}^\star) + (2-\varepsilon) R\, L(f, \mathbf{X}, R)\Big).$$

Main Result

Theorem (Main Result)

Suppose $\boldsymbol{\theta}^\star \in \mathcal{C}$ *and nonlinear scores satisfy*

$$\beta\, \tau_i(\boldsymbol{\theta}) \leq \tau_i \quad \forall \boldsymbol{\theta} \in \mathcal{C}.$$

If

$$s = \widetilde{\mathcal{O}}\left(p^2/(\beta\varepsilon^2)\right),$$

then, with probability at least $1-\delta$,

$$\mathcal{L}(\boldsymbol{\theta}^\star_{\mathcal{S}}) \leq \mathcal{L}(\boldsymbol{\theta}^\star) + \mathcal{O}(\varepsilon).$$

Discussion

- Recent work by Gajjar, Musco, and Hegde (2023) and Gajjar, Tai, Xingyu, Hegde, Musco, and Li (2024) studies leverage-score-based sampling for single-index models $f(\boldsymbol{\theta}) = \phi(\langle \boldsymbol{\theta}, \mathbf{x} \rangle)$.
- If ϕ is L-Lipschitz, Gajjar, Musco, and Hegde (2023) show that

$$\widetilde{\mathcal{O}}(p^2/\varepsilon^4)$$

samples suffice to guarantee

$$\mathcal{L}(\boldsymbol{\theta}^\star_{\mathcal{S}}) \leq C\, \mathcal{L}(\boldsymbol{\theta}^\star) + \mathcal{O}(\varepsilon),$$

for some constant $C > 1$.

- Gajjar, Tai, Xingyu, Hegde, Musco, and Li (2024) improve the sample complexity to

$$\widetilde{\mathcal{O}}(p/\varepsilon^2),$$

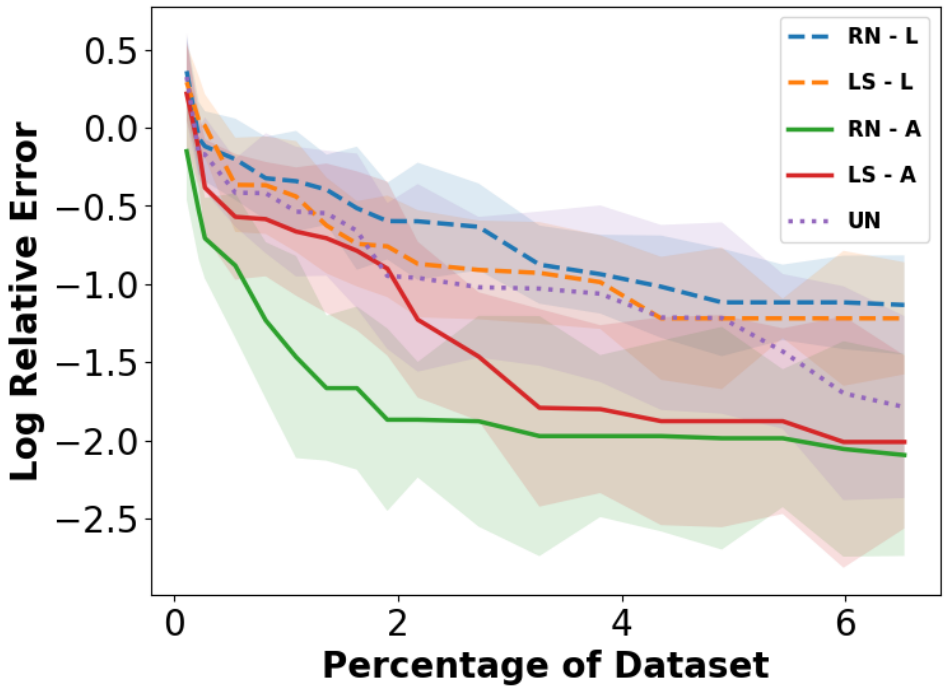
matching linear rates up to logarithmic factors.

- **However:** The constant C can be large (exceeding 10^3).

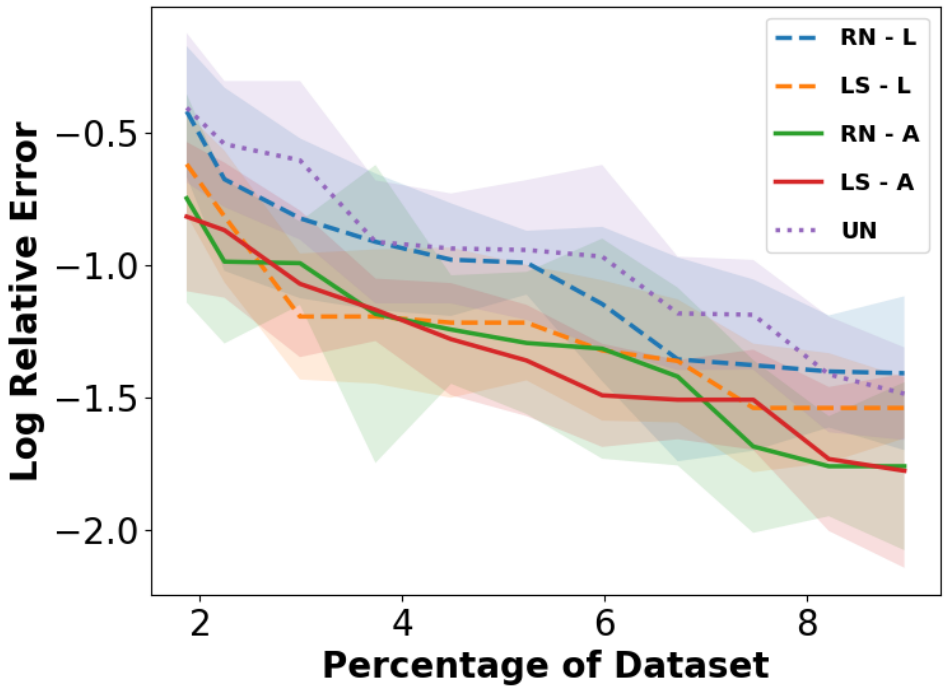
Discussion

	This work	Gajjar et al. (2024)
Applicability	Beyond single-index models	Single-index models
ε **dependence**	Same	Same
p **dependence**	$\mathcal{O}(p^2)$	$\mathcal{O}(p)$
Subproblem	Constrained	Unconstrained
Guarantee constant C	$C = 1$	Potentially large ($C \gg 1$)

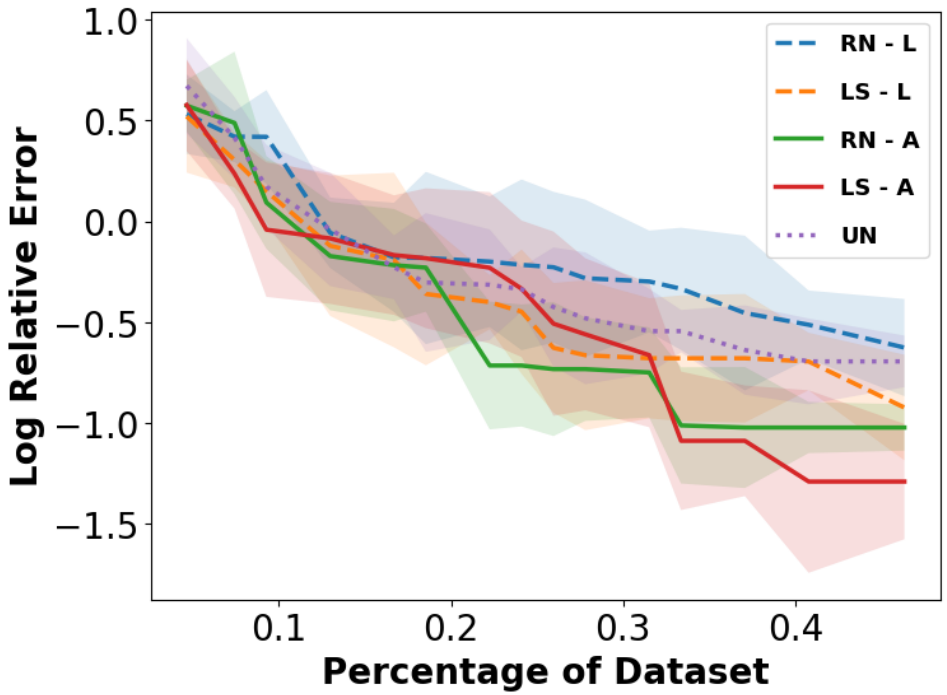
California Housing Dataset



Medical Insurance Dataset



Diamonds Dataset



SVHN - High Leverage (1 vs 0)

SVHN - Low Leverage (1 vs 0)

SVHN - High Leverage (1 vs 7)

SVHN - Low Leverage (1 vs 7)

NOTMNIST - High Leverage (A vs B)

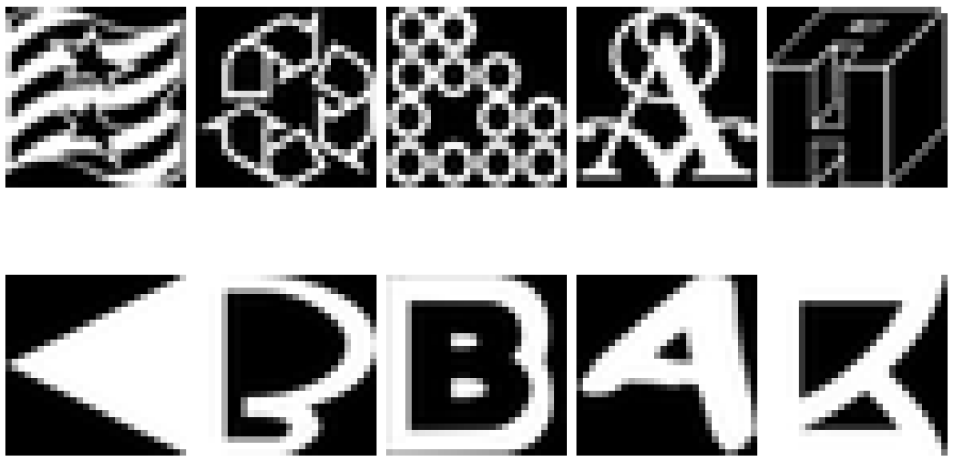


Figure: Ambiguous or atypical character.

NOTMNIST - Low Leverage (A vs B)

Figure: Canonical, easy-to-classify sample.

NOTMNIST - High Leverage (B vs D)

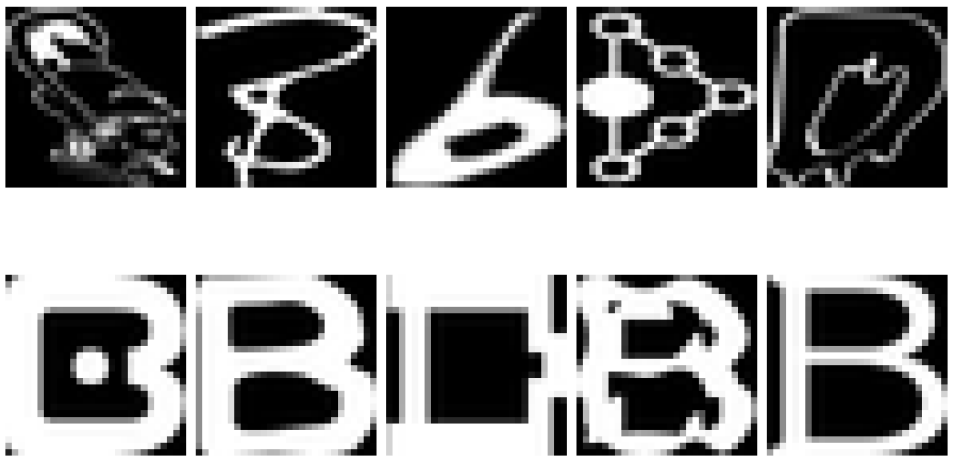


Figure: High leverage outlier or mislabeled example.

NOTMNIST - Low Leverage (B vs D)

Figure: Low-information, prototypical sample.

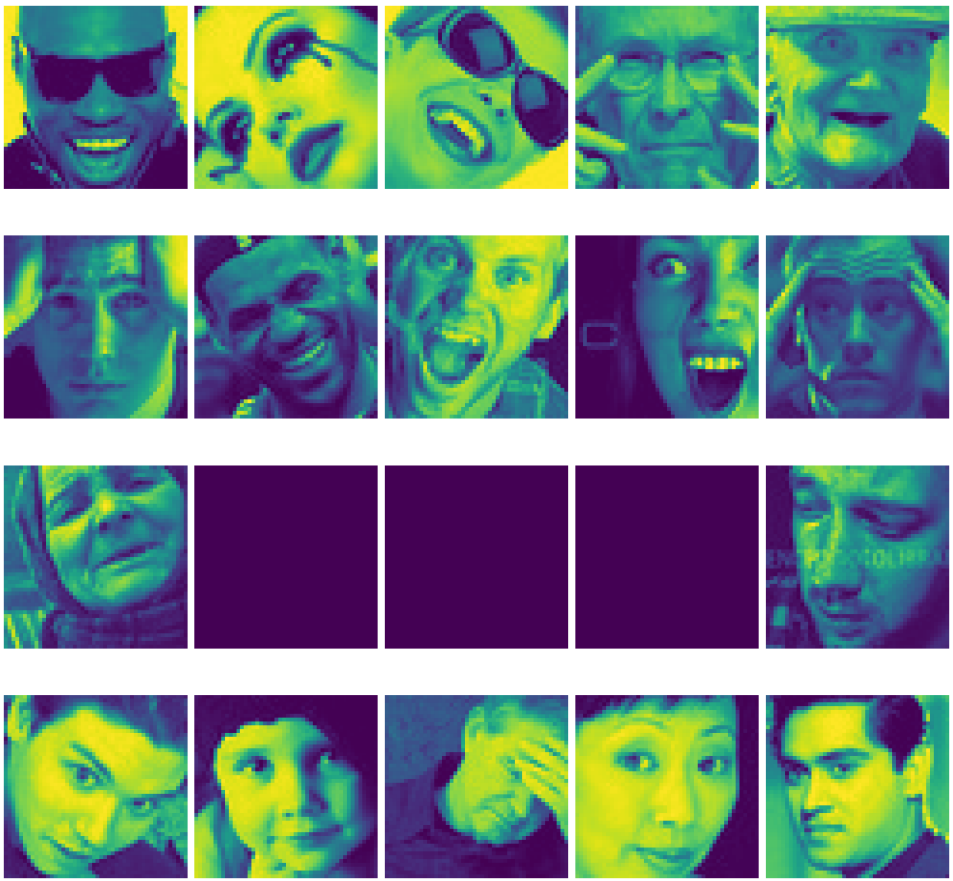
QuickDraw - High Nonlinear Leverage

Figure: Rare drawing style or ambiguous structure.

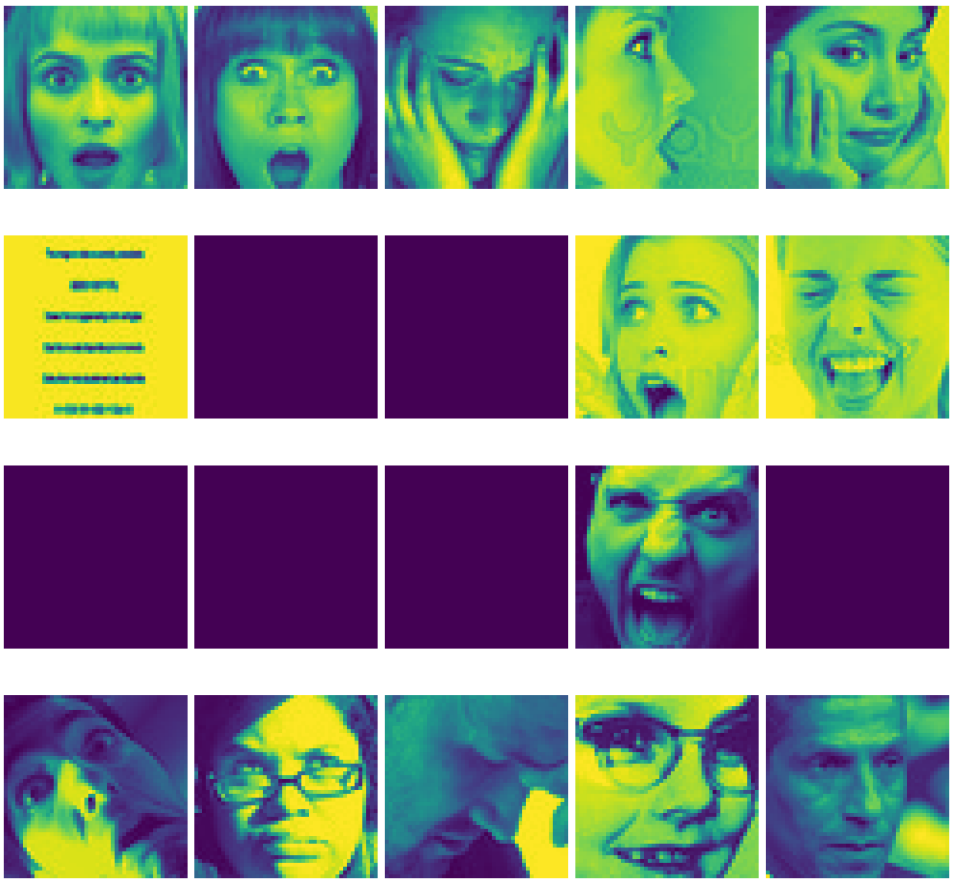
QuickDraw - Low Nonlinear Leverage

Figure: Simple and unambiguous sketch.

FER - High Nonlinear Leverage



FER - Low Nonlinear Leverage




THANK
YOU


Avron, Haim, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh (2017). "Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees". In: *International conference on machine learning*. PMLR, pp. 253–262.


— (2019). "A universal sampling method for reconstructing signals with simple fourier transforms". In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1051–1063.


Erdélyi, Tamás, Cameron Musco, and Christopher Musco (2020). "Fourier sparse leverage scores and approximate kernel learning". In: *Advances in Neural Information Processing Systems* 33, pp. 109–122.


Gajjar, Aarshvi and Cameron Musco (2021). "Subspace embeddings under nonlinear transformations". In: *Algorithmic Learning Theory*. PMLR, pp. 656–672.


Gajjar, Aarshvi, Christopher Musco, and Chinmay Hegde (2023). "Active learning for single neuron models with lipschitz non-linearities". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 4101–4113.

 Gajjar, Aarshvi, Wai Ming Tai, Xu Xingyu, Chinmay Hegde, Christopher Musco, and Yi Li (2024). "Agnostic active learning of single index models with linear sample complexity". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR, pp. 1715–1754.

 Rajmohan, Prakash P. and Fred Roosta (2025). "Importance Sampling for Nonlinear Models". In: *Forty-second International Conference on Machine Learning (ICML)*. URL: https://openreview.net/forum?id=fY7KjywLK1.