Sharper Bounds for ℓ_p Sensitivity Sampling

David P. Woodruff and Taisuke Yasuda

Carnegie Mellon University
Computer Science Department

Sampling for Efficient Machine Learning

- **Empirical risk minimization**: minimize $f: X \to \mathbb{R}_{\geq 0}$ of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} f_i(\mathbf{x})$$

- **Sampling**: we seek a subset $S \subseteq [n]$ and weights w_i for $i \in S$ s.t.

$$\text{for all } \mathbf{x} \in X, \qquad \sum_{i \in S} w_i \cdot f_i(\mathbf{x}) = (1 \pm \varepsilon) \sum_{i=1}^{n} f_i(\mathbf{x}) \qquad (1)$$

Approximate the objective fn for every $\mathbf{x} \in X$

- **Why sample?**
  - Reduce training/inference resources (time, memory, communication)
  - Reduce number of labels needed
  - Preserves sparsity and structure

Question. How small can the sample S be to achieve the guarantee (1)?

Sensitivity Sampling

- Classic technique for achieving (1) : **sensitivity sampling**
  - [Langberg-Schulman 2010, Feldman-Langberg 2011]
  - Define **sensitivity scores**:

$$\sigma_i = \sup_{\mathbf{x} \in X} \frac{f_i(\mathbf{x})}{f(\mathbf{x})} = \sup_{\mathbf{x} \in X} \frac{f_i(\mathbf{x})}{\sum_{j=1}^{n} f_j(\mathbf{x})}$$

  This can often be approximated efficiently

  - **Idea**: Sample the ith example, $i \in [n]$ with probability p_i proportional to the sensitivity scores
    - Probability $p_i = \min\{1, \sigma_i/\alpha\}$, weight $p_i = 1/w_i$
- **Prior work**: sensitivity sampling is very effective!
  - Provable guarantees for a wide class of ERM problems

Theorem [FL11]. Sensitivity sampling gives guarantee (1) with $|S| = \tilde{O}\left(\varepsilon^{-2}\mathfrak{S}d\right)$, for VC dimension d and total sensitivity $\mathfrak{S} = \sum_{i=1}^{n} \sigma_i$

  - Nearly optimal sampling guarantees for least squares regression

Sensitivity Sampling for ℓ_p Linear Regression

- ℓ_p **linear regression**:

$$f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_p^p = \sum_{i=1}^{n} |\langle \mathbf{a}_i, \mathbf{x} \rangle - \mathbf{b}_i|^p$$

  - $\mathbf{A}$ is an $n \times d$ design matrix, $\mathbf{b}$ is an n-dim target vector
    - WLOG assume $\mathbf{b} = 0$
- Sensitivity sampling immediately applies!

$$\text{VC dimension } d\text{, total sensitivity } \mathfrak{S} \leq \begin{cases} d^{p/2} & p > 2 \\ d & p \leq 2 \end{cases}$$

  - Sampling bound [FL11]:

$$|S| = \tilde{O}\left(\varepsilon^{-2}\mathfrak{S}d\right) \leq \begin{cases} \tilde{O}\left(\varepsilon^{-2}d^{p/2+1}\right) & p > 2 \\ \tilde{O}\left(\varepsilon^{-2}d^{2}\right) & p \leq 2 \end{cases}$$

  - But we know this bound is loose for $p = 2$!
    - [Drineas-Mahoney-Muthukrishnan 2006]
    - $|S| = \tilde{O}(\varepsilon^{-2}d)$ for $p = 2$

Question. How small can the sample S be with sensitivity sampling for ℓ_p linear regression?

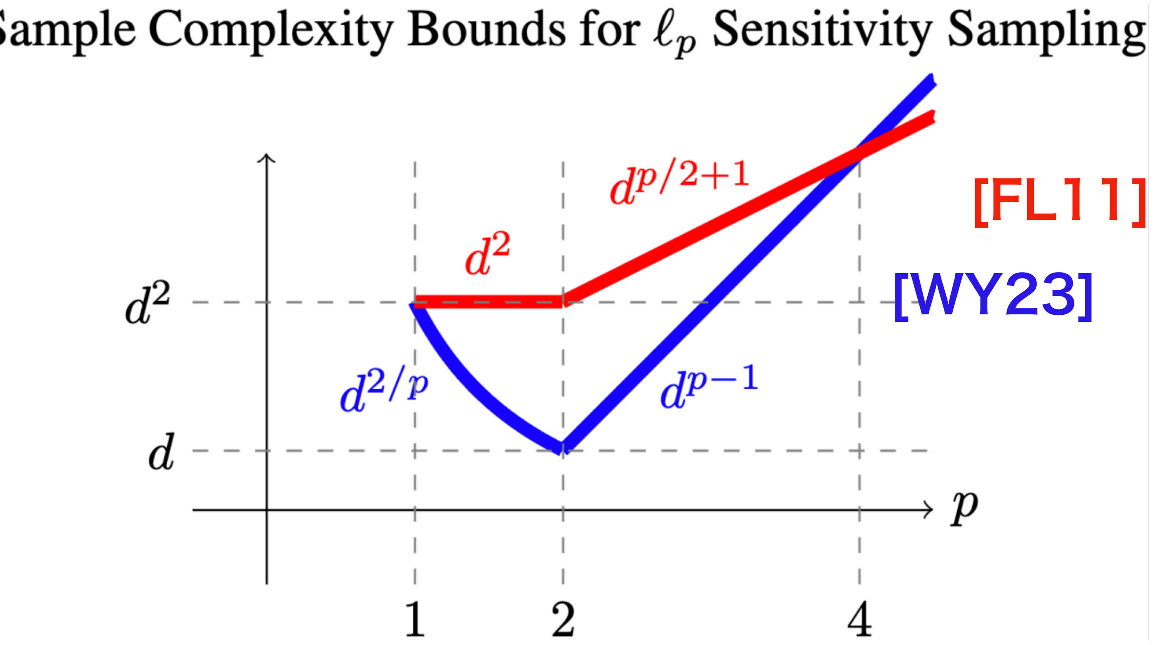
Our Results

Theorem [WY23]. For ℓ_p linear regression, sensitivity sampling gives guarantee (1) with

$$|S| = \begin{cases} \tilde{O}\left(\varepsilon^{-2}\mathfrak{S}^{2-2/p}\right) & p > 2 \\ \tilde{O}\left(\varepsilon^{-2}\mathfrak{S}^{2/p}\right) & p \leq 2 \end{cases} \leq \begin{cases} \tilde{O}\left(\varepsilon^{-2}d^{p-1}\right) & p > 2 \\ \tilde{O}\left(\varepsilon^{-2}d^{2/p}\right) & p \leq 2 \end{cases}$$

- The analysis of [FL11] is loose
- Upper bound is nearly tight for $p \leq 2$; there exist matrices $\mathbf{A}$ that require $\Omega(\mathfrak{S}^{2/p})$ samples

Sample Complexity Bounds for ℓ_p Sensitivity Sampling



Comparison to Lewis Weights

- Sample complexity comparison between
  - Lewis weight sampling [Cohen-Peng 2015]
  - Sensitivity sampling with small $\mathfrak{S}$ ($d^{p/2}$ for $p < 2$, d for $p > 2$)
    - Low rank + sparse, polynomial feature maps, etc...
  - Sensitivity sampling with large $\mathfrak{S}$ (d for $p < 2$, $d^{p/2}$ for $p > 2$)

	Lewis weights	Sensitivity, small $\mathfrak{S}$	Sensitivity, large $\mathfrak{S}$
$p < 2$	$\tilde{O}(\varepsilon^{-2}d)$	$\tilde{O}(\varepsilon^{-2}d)$	$\tilde{O}(\varepsilon^{-2}d^{2/p})$
$p > 2$	$\tilde{O}(\varepsilon^{-2}d^{p/2})$	$\tilde{O}(\varepsilon^{-2}d^{2-2/p})$	$\tilde{O}(\varepsilon^{-2}d^{p-1})$

Sharpest known bounds for inputs with small total sensitivity $\mathfrak{S}$

Applications: noisy ℓ_p polynomial regression

Techniques

- [Feldman-Langberg 2011] analysis of sensitivity sampling
  - Prove that (1) holds for a fixed $\mathbf{x} \in X$ with high probability
  - Union bound over a fine discretization of X
- [Bourgain-Lindenstrauss-Milman 1989] analysis of Lewis weights
  - Improve the union bound via **chaining arguments**
  - Relies on special structure of Lewis weights
- **Key question**: can ℓ_p sensitivity sampling use similar chaining arguments, without using the special structure of Lewis weights?
  - Yes! Lewis weights are a way to use ℓ_2 sensitivity sampling (aka leverage score sampling) for ℓ_p sampling
  - ℓ_p sensitivity sampling is also related to ℓ_2 sensitivity sampling:

Lemma [WY23]. ℓ_p sensitivities are within a $n^{p/2-1}$ factor away from the ℓ_2 sensitivities.

Open Directions

- Are there better sampling algorithms when $\mathfrak{S}$ is small?
- Can guarantees for sensitivity sampling be improved in other settings?