Sharper Bounds for ℓ_p Sensitivity Sampling

Taisuke Yasuda and David P. Woodruff

Carnegie Mellon University
Computer Science Department

ICML
International Conference
On Machine Learning

Sampling for Efficient Machine Learning

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

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

Sum over n training examples

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

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

Approximate the objective fn up to a $(1+\varepsilon)$ factor for every $\mathbf{x} \in X$

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

Sampling for Efficient Machine Learning

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

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

Sum over n training examples

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

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

Approximate the objective fn up to a $(1 + \varepsilon)$ factor for every $\mathbf{x} \in X$

Question. How small can the sample S be to achieve the above guarantee?

Sensitivity Sampling

- **Sampling**: we seek a <u>subset of training examples</u> $S \subseteq [n]$ and <u>weights</u> w_i for $i \in S$ s.t.

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

- Classic technique for achieving ☝️: **sensitivity sampling**
  - [Langberg-Shulman 2010, Feldman-Langberg 2011]
  - Define **sensitivity scores**:

    for each training example $i \in [n]$, define $\sigma_i = \sup_{\mathbf{x} \in X} \frac{f_i(\mathbf{x})}{f(\mathbf{x})} = \sup_{\mathbf{x} \in X} \frac{f_i(\mathbf{x})}{\sum_{j=1}^{n} f_j(\mathbf{x})}$
  - Sample i-th example with probability proportional to the sensitivity scores

Sensitivity Sampling

- **Prior work**: sensitivity sampling is very effective!
  - Provable guarantees for a wide class of ERM problems

Theorem [FL11]. Sensitivity sampling gives $(1+\varepsilon)$-approximations 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

Question. What about for ℓ_p linear regression? How well does sensitivity sampling perform for this problem?

- ℓ_p linear regression: let $\mathbf{A}$ be an $n \times d$ design matrix, let $\mathbf{b}$ be an n-dimensional target vector

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

Sensitivity Sampling

ℓ_p linear regression

- Sensitivity sampling immediately applies!

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

- Sampling bound [Feldman-Langberg 2011] bound:

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

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

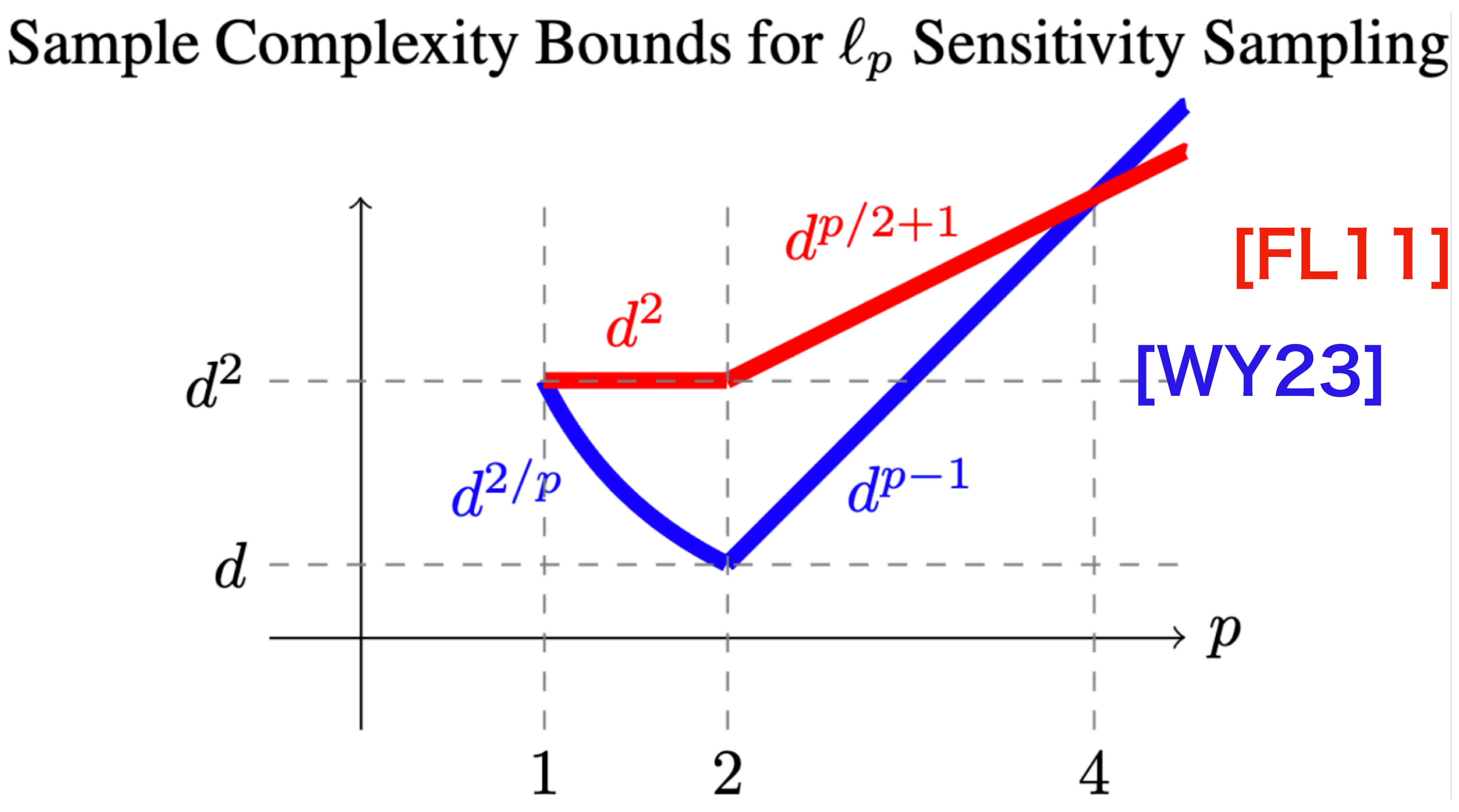
Question. How small can the sample S be with sensitivity sampling for ℓ_p linear regression?

Our Results

Theorem [WY23]. For ℓ_p linear regression, sensitivity sampling gives $(1+\varepsilon)$-approximations with

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

- **Remarks**
  - Analysis of [FL11] is loose - we can do better!
  - Upper bound is nearly tight for $p \leq 2$; there exist matrices $\mathbf{A}$ that require $\Omega(\mathfrak{S}^{2/p})$ samples



Our Results

- **Techniques**
  - We have a tight analysis for $p = 2$, how can we make use of this?
  - Key idea: relate ℓ_p sensitivity scores to ℓ_2 sensitivity scores

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

- **Applications**
  - Sampling algorithms for ℓ_p linear regression on low sensitivity instances
    - Low rank + sparse, polynomial feature maps, etc⋯
  - ℓ_p polynomial regression with noise

Conclusion

- **Summary**
  - We give a sharper analysis of sensitivity sampling, a classic sampling technique, for ℓ_p linear regression

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

- **Open Directions**
  - Can guarantees for sensitivity sampling be improved in other settings?
- **Poster**:
  - Thursday 1:30 pm - 3:00 pm, Exhibit Hall 1 #336
  - Come chat!

Thank you!!