Big idea 3: stability and advanced mechanisms

John Duchi
Stanford University

Frejus 2025

Outline

1. Local sensitivities (moduli of continuity)
2. Inverse sensitivity mechanisms
3. Matrix mechanisms and correlated noise

Sensitivity measures

- *global sensitivity* of a statistic

$$\mathsf{GS}_p(f) := \sup \left\{ \left\| f(P_n) - f(P_n') \right\|_p \text{ s.t. } n \left\| P_n - P_n' \right\|_{\mathrm{TV}} \le 1 \right\}$$

- saw that adding noise commensurate with this is sufficient for privacy:

$$M(P_n) = f(P_n) + \mathsf{GS}_p(f) \cdot \mathsf{N}(0, \sigma^2 I)$$

- would rather use *local sensitivity* [10] at P_n:

$$\mathsf{LS}_p(f; P_n) := \sup_{P_n'} \left\{ \left\| f(P_n) - f(P_n') \right\| \text{ s.t. } n \left\| P_n - P_n' \right\|_{\mathrm{TV}} \le 1 \right\}$$

Examples of sensitivities

- mean for $x_i \in [-1, 1]$

- median for $x_i \in [-1, 1]$

- minimizer of empirical loss $\theta(P_n) = \operatorname{argmin} \mathbb{E}_{P_n}[\ell_\theta(X)]$

A nice idea?

- release

$$M(P_n) = f(P_n) + \mathsf{LS}(f; P_n) \cdot \mathsf{Noise}$$

Issue: scale of noise leaks information

- consider medians for samples

$$P_n = \frac{n-1}{2}\mathbf{1}_0 + \frac{n+1}{2}\mathbf{1}_1 \quad \text{and} \quad P_n' = \frac{n+1}{2}\mathbf{1}_0 + \frac{n-1}{2}\mathbf{1}_1$$

Aside: exponential mechanisms

- for a *score* function $s : \Theta \times \mathcal{P}_n \to \mathbb{R}$, where

$$\mathsf{GS}(s) := \sup_{\theta \in \Theta} \mathsf{GS}(s(\theta, \cdot)) < \infty,$$

exponential mechanism [8] releases with density

$$p(\theta) = \exp\left(-\frac{\varepsilon}{2\mathsf{GS}(s)} s(\theta, P_n)\right) / \int \exp\left(-\frac{\varepsilon}{2\mathsf{GS}(s)} s(\theta', P_n)\right) d\mu(\theta')$$

Lemma (McSherry and Talwar [8])

The exponential mechanism is ε-differentially private

A stable statistic

Definition (Asi and Duchi [1])

The *inverse sensitivity* of $\theta : \mathcal{P}_n \to \Theta$ is

$$\mathsf{len}(t; P_n) := \inf \left\{ k \in \mathbb{N} \mid \theta(P_n') = t, \;\; n \left\| P_n - P_n' \right\|_{\mathrm{TV}} \leq k \right\}$$

- number of examples to change to get desired output
- immediate that it is 1-Lipschitz w.r.t. Hamming distance:

$$|\mathsf{len}(t; P_n) - \mathsf{len}(t; P_n')| \leq 1.$$

Inverse sensitivity examples

Example (Inverse sensitivity of the mean)

For data $x_i \in [0, r]$, inverse sensitivity is

$$\mathsf{len}(t; P_n) = \left\lceil \frac{n}{r} \left| \mathbb{E}_{P_n}[X] - t \right| \right\rceil$$

Example (Inverse sensitivity of the median)

$$\mathsf{len}(t; P_n) = \mathrm{card}\left\{ X_i \in [t, \mathsf{Med}(P_n)] \right\}$$

Inverse sensitivity mechanism

Definition (Inverse sensitivity mechanism [1])

Sample according to density

$$p(\theta) = \exp\left(-\frac{\varepsilon}{2}\mathsf{len}(\theta; P_n)\right) / \int \exp\left(-\frac{\varepsilon}{2}\mathsf{len}(t; P_n)\right) d\mu(t)$$

Example (Median sampling)

1. For data $x_i \in [-r, r]$, form shells

$$S_k = \{t \mid \mathsf{len}(t; P_n) = k\} = \left[X_{(\frac{n}{2}-k)}, X_{(\frac{n}{2}-k+1)}\right] \cup \left[X_{(\frac{n}{2}+k)}, X_{(\frac{n}{2}+k+1)}\right]$$

2. Draw $K \in \{0, \ldots, n/2\}$, $\mathbb{P}(K = k) \propto \exp(-\frac{\varepsilon}{2}|S_k|)$
3. Return $\theta \sim \mathsf{Uni}(S_k)$

Accuracy of inverse sensitivity

Heuristic: moving P_n to P'_n changes at most $\mathsf{LS}(\theta; P_n)$

$$\begin{aligned}&\theta(P_n) - \theta(P_n^{(k)}) \\ &= \theta(P_n) - \theta(P_n^{(1)}) + \theta(P_n^{(1)}) - \theta(P_n^{(2)}) + \cdots + \theta(P_n^{(k-1)}) - \theta(P_n^{(k)})\end{aligned}$$

i.e.

$$|\theta(P_n) - \theta(P_n^{(k)})| \underset{\text{sort of}}{\le} O(1)k \cdot \mathsf{LS}(\theta; P_n)$$

Idea: unlikely to select distance $k \gg \frac{1}{\varepsilon}$

Heuristic accuracy of inverse sensitivity

- use heuristic

$$\mathsf{len}(t; P_n) \approx \frac{|t - \theta(P_n)|}{\mathsf{LS}(\theta, P_n)}$$

- density of inverse sensitivity

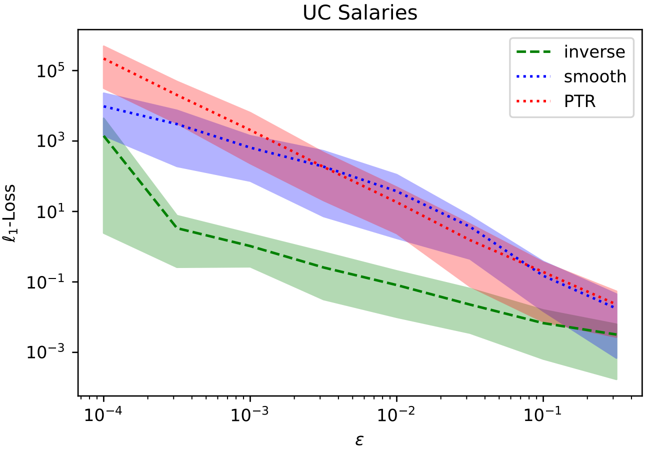
$$p(t) \propto \exp\left(-\frac{\varepsilon}{2}\mathsf{len}(t; P_n)\right) \approx \exp\left(-\frac{\varepsilon}{2}|t - \theta(P_n)|\right)$$

- So $M(P_n) \dot{\sim} \theta(P_n) + \frac{2}{\varepsilon}\mathsf{LS}(\theta, P_n) \cdot \mathsf{Lap}(1)$

Example (Median behavior)

If X_i have density f, expect median θ to have $\mathsf{LS}(\theta, P_n) \asymp \frac{1}{nf(\mathsf{Med}(P_n))}$

Implementation performance



Error in median salary for University of California school system (over 100,000 salaries) [1]

Optimality of inverse sensitivity

Intuition: If $d_{\mathsf{ham}}(P_n, P_n^{(k)})$, *cannot* test ε-differentially private mechanisms

$$M(P_n) \text{ vs. } M(P_n^{(k)}) \text{ if } k \leq \frac{1}{\varepsilon}$$

Define $\mathsf{LS}^k(P_n) = \sup\{|\theta(P_n') - \theta(P_n)| \text{ s.t. } d_{\mathsf{ham}}(P_n', P_n) \leq k\}$.

Proposition (Asi and Duchi [1])

For any statistic θ and sample P_n, there exists P_n' with $d_{\mathsf{ham}}(P_n, P_n') \leq \frac{1}{\varepsilon}$ such that

$$\max_{P \in \{P_n, P_n'\}} \mathbb{E}[|M(P) - \theta(P)|] \gtrsim \mathsf{LS}^{\frac{1}{\varepsilon}}(P_n)$$

Further work and questions inverse sensitivity

- Johnson and Shmatikov's *distance score* instantiates mechanism [6]
- connecting Robustness and Privacy: estimators can be robust if and only if they are private (see Hopkins et al. [5] and Asi et al. [2])
- general (1-dimensional) optimal mechanism, but implementation leaves many open questions:

Example (Asi et al. [3])

In a statistical model $Y \sim P_\theta(\cdot \mid X)$, optimally estimate a single coordinate θ_1?

Linear queries

(abstract) linear query problem: data in n observations

$$X = \begin{bmatrix} \cdots & x_1^T & \cdots \\ & \vdots & \\ \cdots & x_n^T & \cdots \end{bmatrix} = \begin{bmatrix} \vdots & & \vdots \\ x^{(1)} & \cdots & x^{(d)} \\ \vdots & & \vdots \end{bmatrix} \in \mathbb{R}^{n \times d}$$

and *query matrix* of m queries $a_i \in \mathbb{R}^n$,

$$A = \begin{bmatrix} \cdots & a_1^T & \cdots \\ & \vdots & \\ \cdots & a_m^T & \cdots \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Goal: accurately provide

$$AX \in \mathbb{R}^{m \times d}$$

Linear query examples

Example (Running sums)

Take A all ones below and on diagonal:

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \quad \text{then} \quad AX = \begin{bmatrix} x_1^T \\ (x_1 + x_2)^T \\ (x_1 + x_2 + x_3)^T \\ \vdots \\ (x_1 + \cdots + x_n)^T \end{bmatrix}$$

Example

For $a = \mathbf{1}$ and $X = [x^{(1)} \ \cdots \ x^{(d)}] \in \{0, 1\}^{n \times d}$

$$\langle a, x^{(j)} \rangle = \langle \mathbf{1}, x^{(j)} \rangle$$

counts individuals with feature j

Gaussian noise addition

release

$$Y = A(X + Z) \quad \text{for} \quad Z_{ij} \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$$

alternative: factorize $A = BC$, (cf. [9, 7]) release

$$Y_{\mathsf{fac}} = B(CX + Z) \quad \text{for} \quad Z_{ij} \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \tau^2)$$

Errors:

$$\mathbb{E}\left[\|Y - AX\|_{\text{Fr}}^2\right] = \sigma^2 \|A\|_{\text{Fr}}^2 \quad \text{vs} \quad \mathbb{E}\left[\|Y_{\mathsf{fac}} - AX\|_{\text{Fr}}^2\right] = \tau^2 \|B\|_{\text{Fr}}^2$$

Gaussian noise addition

for $Z_{ij} \stackrel{\text{iid}}{\sim} \mathsf{N}(0, \frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ release

$$Y = A(X + \sigma Z) \quad \text{or} \quad Y_{\mathsf{fac}} = B(CX + \tau Z)$$

Errors:

$$\mathbb{E}\left[\|Y - AX\|_{\text{Fr}}^2\right] = \sigma^2 \|A\|_{\text{Fr}}^2 \quad \text{vs} \quad \mathbb{E}\left[\|Y_{\mathsf{fac}} - AX\|_{\text{Fr}}^2\right] = \tau^2 \|B\|_{\text{Fr}}^2$$

Lemma (cf. Pillutla et al. [11])

For $C = [c_1 \ \cdots \ c_n]$, to achieve same level of privacy, set

$$\sigma^2 \propto \sup_{x \in \mathcal{X}} \|x\|_2^2 \quad \textit{and} \quad \tau^2 \propto \sup_{x \in \mathcal{X}} \max_i \|c_i\|_2^2 \|x\|_2^2$$

Error control

- assume data $x_i \in \mathbb{R}^d$, $\|x_i\|_2 \leq 1$

Lemma
Frobenius error of

$$Y = AX + Z \quad \textit{vs} \quad Y_{\mathsf{fac}} = B(CX + Z)$$

for same privacy level

$$\mathbb{E}\left[\|Y - AX\|_{\mathrm{Fr}}^2\right] \propto \|A\|_{\mathrm{Fr}}^2 \quad \textit{vs} \quad \mathbb{E}\left[\|Y_{\mathsf{fac}} - AX\|_{\mathrm{Fr}}^2\right] \propto \|B\|_{\mathrm{Fr}}^2 \|C\|_{1\to2}^2$$

Maximum error control

- assume data $x_i \in \mathbb{R}^d$, $\|x_i\|_2 \leq 1$
- consider maximum row norm

$$\|G\|_{2\to\infty} = \max_i \|g_i\|_2 \quad \text{for} \quad G = [g_1 \ \cdots \ g_m]^T.$$

Lemma

maximum row error of $Y = AX + Z$ versus $Y_{\mathsf{fac}} = B(CX + Z)$ for same privacy level is

$$\frac{\mathbb{E}\left[\|Y - AX\|_{2\to\infty}^2\right]}{\log m} \propto \|A\|_{2\to\infty}^2$$

versus

$$\frac{\mathbb{E}\left[\|Y_{\mathsf{fac}} - AX\|_{2\to\infty}^2\right]}{\log m} \propto \|B\|_{2\to\infty}^2 \|C\|_{1\to2}^2$$

The optimal matrix mechanism problem

$$\begin{array}{ll} \text{minimize} & \|B\|_{\mathrm{Fr}} \|C\|_{1\to 2} \\ \text{subject to} & A = BC \end{array} \quad \text{or} \quad \begin{array}{ll} \text{minimize} & \|B\|_{2\to\infty} \|C\|_{1\to 2} \\ \text{subject to} & A = BC \end{array}$$

Some issues

- typically hard to solve these problems
- unclear if improvement is that big? (but trivial example: $A = \mathbf{1}\mathbf{1}^T$)

Running sums

- special case that A is all ones on and below diagonal
- important in *online gradient* methods [11]:

$$\theta_{k+1} = \theta_k - \eta g_k = -\eta \sum_{i=1}^{k} g_i$$

Theorem

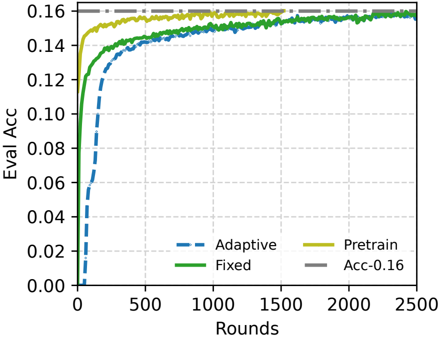
There is a factorization of the running sum matrix $A = BC$ with $\|B\|_{2\to\infty} = O(1)\log n$ and $\|C\|_{1\to 2} = O(1)\log n$

(NB: this is suboptimal, and $O(1)\log n$ possible [4].)

Demonstration of factorization of running sums

Experimental evidence

- "We are happy to announce that all the next word prediction neural network LMs in Gboard now have DP guarantees, and all future launches of Gboard neural network LMs will require DP guarantees" [12]



Lan	Acc.$(+)$	Pop$(\cdot 10^6)$
en-US	.11%	13
pt-BR	.29%	16.6
en-IN	.4%	7.7

[1] H. Asi and J. Duchi. Near instance-optimality in differential privacy. *arXiv:2005.10630 [cs.CR]*, 2020.

[2] H. Asi, J. Ullman, and L. Zakynthinou. From robustness to privacy and back. *arXiv:2302.01855 [cs.LG]*, 2023.

[3] H. Asi, J. Duchi, and K. Talwar. On privately estimating a single parameter. *arXiv:2503.17252v1 [cs.LG]*, 2025.

[4] H. Fichtenberger, M. Henzinger, and J. Upadhyay. Constant matters: fine-grained analysis error bound on differentially private continual observation. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[5] S. B. Hopkins, G. Kamath, and M. Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the Fifty-Fourth Annual ACM Symposium on the Theory of Computing*, pages 1406–1417, 2022.

[6] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1079–1087, 2013.

[7] C. Li, G. Miklau, M. Hay, A. McGregor, and V. Rastogi. The matrix mechanism: optimizing linear counting queries under differential privacy. *The VLDB Journal*, 24:757–781, 2015.

[8] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual Symposium on Foundations of Computer Science*, 2007.

[9] A. Nikolov, K. Talwar, and L. Zhang. The geometry of differential privacy: the sparse and approximate case. In *Proceedings of the Forty-Fifth Annual ACM Symposium on the Theory of Computing*, 2013.

[10] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on the Theory of Computing*, 2007.

[11] K. Pillutla, J. Upadhyay, C. A. Choquette-Choo, K. Dvijotham, A. Ganesh, M. Henzinger, J. Katz, R. McKenna, H. B. McMahan, K. Rush, T. Steinke, and A. Thakurta. Correlated noise mechanisms for differentially private learning: A tutorial on mathematical foundations and practical aspects. *arXiv:2506.08201*, 2025.

[12] Z. Xu, Y. Zhang, G. Andrew, C. Choquette, P. Kairouz, B. Mcmahan, J. Rosenstock, and Y. Zhang. Federated learning of gboard language models with differential privacy. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Industry Track)*, pages 629–639, 2023.