CS839: Differential Privacy and Learning March 26, 2026

Lecture 20: Robustness and Privacy

Instructor: Gavin Brown *Scribe:* Gavin Brown

Disclaimer: This document is intended as an informal supplement to in-class note-taking. It has not been given the level of scrutiny expected in polished lecture notes, let alone that reserved for peer-reviewed publications.

In this lecture we will discuss relationships between differential privacy and robust statistics. The latter is a field that studies how statistical inference can be performed in a way that is insensitive to misspecification or data corruptions, even adversarial ones. We will use the following notion of robustness.

Definition 0.1. Fix a distribution P and population quantity $\mu(P)$. Algorithm A is a (τ, β, α)-*robust estimator* for $\mu(P)$ if with probability $1-\beta$ over the draw of $x \sim P^{\otimes n}$ for all x' that differ from x in at most $n\tau$ entries we have

$$\big\|A(x') - \mu(P)\big\| \leq \alpha.$$

A few remarks. The notation $x \sim P^{\otimes n}$ means that $x = (x_1, \ldots, x_n)$ is drawn from the n-fold product distribution over P, i.e., each $x_i \sim P$ independently. We will assume the robust estimator is deterministic. We use the notation $\mu(P)$ but there is no assumption that this is the mean of the distribution. This definition, and the rest of the lecture, will apply to any norm $\|\cdot\|$. However, it is important that the norm does not depend on the underlying distribution.

Connections between robustness and privacy were recognized almost immediately after the introduction of DP, but only recently have we started to understand the depth of these connections.

This lecture will limit itself in scope to pure differential privacy and information-theoretic results (i.e., we will not discuss computation). Our presentation mostly follows that of Asi et al. [2023].

1 The (Smoothed) Inverse Sensitivity Mechanism

1.1 Review: Inverse Sensitivity

Before discussing robustness further, we recall the main DP mechanism we will use today.

Definition 1.1. For a function $f : \mathcal{X}^n \to \mathcal{Y}$, the *inverse sensitivity mechanism* A_{inv} samples from

$$p(y) \propto \exp\left(-\frac{\varepsilon}{2} \cdot \text{len}_f(y; x)\right)$$

where

$$\text{len}_f(y; x) = \inf_{x'} \left\{ d_H(x, x') \mid f(x') = y \right\}$$

Here $d_H(x, x')$ is Hamming distance.

Important questions we will ignore:

- How can we compute $\text{len}_f(y; x)$?
- How can we sample from $p(y)$?

Before stating the utility guarantees, recall *local sensitivity* and the local sensitivity at distance k:

$$\text{LS}_f(x) = \sup_{x' \sim x} \left\| f(x) - f(x') \right\|$$

$$\text{LS}_f^k(x) = \sup_{\substack{x' \\ d_H(x,x') \le k}} \left\| f(x) - f(x') \right\|.$$

We saw the latter quantity in our analysis of the inverse sensitivity mechanism on Homework 2. It captures: for a given dataset x, if I allow you you change k inputs, how much can you function value? This notion has another name, which we will use today.

Definition 1.2. The *local modulus of continuity* of f at x is, for $k \in \mathbb{N}$,

$$\omega_f(x; k) = \sup_{\substack{x' \\ d_H(x,x') \le k}} \left\| f(x) - f(x') \right\|.$$

Thus $\omega_f(x; 1) = \text{LS}_f(x)$. The local modulus terminology is older and used outside of privacy contexts, often with metrics other than Hamming distance. See Duchi [2025], Chapter 13 for examples and discussion.

Let $\text{diam}(\mathcal{Y}) = \sup_{u,v \in \mathcal{Y}} \|u - v\|$.

Theorem 1.3 (Utility, Discrete)**.** *For $f : \mathcal{X}^n \to \mathcal{Y}$, on any input x, with probability $1 - \beta$ A_{inv} returns $\tilde{y}$ such that*

$$\|\tilde{y} - f(x)\| \le \omega_f\left(x; \frac{2}{\varepsilon} \cdot \log\left(\frac{2|\mathcal{Y}|\text{diam}(\mathcal{Y})}{\beta\varepsilon}\right)\right).$$

This is vacuous when $|\mathcal{Y}| = +\infty$, which is what we usually care about for parameter estimation.

Here are two options to proceed:

- Can assume more about the distribution or function f.
- Can modify the mechanism.

The first is sometimes appropriate, but today we aim for generality and proceed with the second.

1.2 Smoothed Inverse Sensitivity

Our main result today will be an application of the following mechanism.

Definition 1.4. For a function $f : \mathcal{X}^n \to \mathcal{Y}$ and $\rho \ge 0$, the *smoothed inverse sensitivity mechanism* A_{inv}^ρ samples from

$$p(y) \propto \exp\left(-\frac{\varepsilon}{2} \cdot \text{len}_f^\rho(y; x)\right)$$

where

$$\text{len}_f^\rho(y; x) = \inf_{x'} \left\{ d_H(x, x') \mid \left\| f(x') - y \right\| \le \rho \right\}$$

Theorem 1.5 (Utility, Smoothed Inverse)**.** *For $f : \mathcal{X}^n \to \mathcal{Y} \subseteq B(0, R)$, on any input x, with probability $1 - \beta$ A^{ρ}_{inv} returns $\tilde{y}$ such that*

$$\|\tilde{y} - f(x)\| \leq \omega_f\left(x; \frac{2}{\varepsilon} \cdot (d \log(R/\rho + 1) + \log(1/\beta))\right).$$

The proof is very similar to our analysis on Homework 2. For intuition, we can connect it with Theorem 1.3: the volume of $B(0, R)$ scales with R^d, so $d \log R$ stands in for $\log |\mathcal{Y}|$.

2 Privacy Induces Robustness

The output of DP algorithms cannot change much if you change a single point. We have already seen how this extends to changing multiple points.

Lemma 2.1 (Group Privacy)**.** *Suppose A is ε-DP. If datasets x and x' satisfy $d_H(x, x') \leq k$, then $A(x) \approx_{k\varepsilon} A(x')$.*

That is, for any event E, $\Pr[A(x) \in E] \leq e^{k\varepsilon} \Pr[A(x') \in E]$.

Theorem 2.2. *Suppose A is ε-DP and satisfies: for $x \sim P^{\otimes n}$, with probability at least $1 - \beta$ we have $\|A(x) - \mu(P)\| \leq \alpha$.*

For any $\gamma \in (0, 1)$, let $\tau = \frac{\log(1/\gamma)}{n\varepsilon}$. A is $(\tau, \beta/\gamma, \alpha)$-robust.

Proof. Let $W = \{y \mid \|y - u\| > \alpha\}$ be the bad event. For $x \sim P^{\otimes n}$ and x' with $d_H(x, x') \leq \tau n$:

$$\begin{aligned}
\Pr[A(x') \in W] &\leq e^{\tau n \varepsilon} \Pr[A(x) \in W] \\
&= \exp\left(\frac{\log(1/\gamma)}{n\varepsilon} \cdot n\varepsilon\right) \cdot \Pr[A(x) \in W] \\
&= \frac{1}{\gamma} \cdot \Pr[A(x) \in W].
\end{aligned}$$

Finally, $\Pr[A(x) \in W] \leq \beta$ by assumption. □

See Georgiev and Hopkins [2022b] for the generalization to approximate DP and discussion of the implications. We give a taste of them here. Setting $\gamma = \frac{1}{e}$ shows that DP algorithms are robust to changing $\frac{1}{\varepsilon}$ points with only a constant-factor increase in failure probability.

However, this is a rather weak level of robustness; the standard in the robustness literature is a tolerance to changing $\Omega(n)$ points. This can be achieved by applying the theorem with $\gamma = e^{-c\varepsilon n}$ for some constant c. Such a small γ means our bound on the failure probability is vacuous unless β itself was already less than $e^{-c\varepsilon n}$. Thus the theorem says that pure DP algorithms which succeed with extremely high probability enjoy quantitatively strong robustness guarantees.

3 Robustness Implies Privacy

The following is from concurrent works of Hopkins, Kamath, Majid, and Narayanan [2023] and Asi, Ullman, and Zakynthinou [2023]. Although we omit further discussion, the former notably shows when this transformation can be accomplished in a way that preserves polynomial running time.

Theorem 3.1. *Let $A : \mathcal{X}^n \to \mathcal{Y} \subseteq B(0, R)$ be a (τ, β, α)-robust estimator for $\mu(P)$. Assume*

$$n \geq \frac{2}{\tau\varepsilon}(d \log(R/\alpha + 1) + \log(1/\beta)).$$

Consider applying A^{ρ}_{inv} to function $A : \mathcal{X}^n \to \mathcal{Y}$ with $\rho = \alpha$. With probability at least $1 - 2\beta$, this returns $\tilde{y}$ such that

$$\|\tilde{y} - \mu\| \leq 4\alpha.$$

Proof. Set $k = \frac{2}{\varepsilon}(d \log(R/\rho+1)+\log(1/\beta))$, so A is robust against changing $n\tau \geq k$ points. Applying the triangle inequality and the accuracy guarantees of Theorem 1.5, we have

$$\begin{aligned}
\left\|A^{\rho}_{\text{inv}}(x) - \mu\right\| &\leq \|A(x) - \mu\| + \left\|A^{\rho}_{\text{inv}}(x) - A(x)\right\| \\
&\leq \|A(x) - \mu\| + \omega_A(x; k) + \alpha \\
&= \|A(x) - \mu\| + \sup_{\substack{x' \\ d_H(x,x') \leq k}} \left\|A(x) - A(x')\right\| + \alpha,
\end{aligned}$$

plugging in the definition of ω_A in the last line. This holds except with probability at most β.

Also with probability at most β, for all x' within distance k of x we have $\|A(x) - \mu\| \leq \alpha$. Thus, applying the triangle inequality again we have

$$\sup_{\substack{x' \\ d_H(x,x') \leq k}} \left\|A(x) - A(x')\right\| \leq \sup_{\substack{x' \\ d_H(x,x') \leq k}} \|A(x) - \mu\| + \left\|A(x') - \mu\right\| \leq 2\alpha.$$

With a union bound over the two probability-β bad events, we are done. □

4 Further Reading

Despite our informal takeaways of "DP ⇒ robust" and "robust ⇒ DP", research on this topic is far from done. Things are only getting started!

The fact that DP algorithms are automatically robust to adversarial corruptions is, at least in some form, folklore. See Georgiev and Hopkins [2022a] for a clear statement of this and an investigation of its implications for information-computation gaps.

Similarly, many have recognized that robust algorithms are very useful in DP algorithm design; see Dwork and Lei [2009] for an early and influential work. Prior to the work of Hopkins et al. [2023] and Asi et al. [2023], there were attempts to produce frameworks formalizing this relationship [Avella-Medina, 2021, Liu et al., 2022]. While they delivered strong results for specific problems, they fall short of the generality seen in today's lecture.

Robust statistics has been studied for decades, but recent years witnessed the development of efficient algorithmic techniques. The prospect of drawing on such techniques is one reason the robustness-to-privacy transformation is so exciting. The textbook of Diakonikolas and Kane [2023] covers recent advances. For a nice supplement, including connections to DP, see Kamath [2024].

References

Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. In *International Conference on Machine Learning*, pages 1121–1146. PMLR, 2023.

Marco Avella-Medina. Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983, 2021.

Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

John Duchi. Statistics and information theory. Lecture notes, Stanford University, 2025. `https://web.stanford.edu/class/ee377/lecture-notes.pdf`.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st ACM Symposium on Theory of Computing*, STOC '09, pages 371–380. ACM, 2009.

Kristian Georgiev and Samuel B. Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. 2022a. doi: 10.48550/ARXIV.2211.00724. URL `https://arxiv.org/abs/2211.00724`.

Kristian Georgiev and Samuel B Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. *arXiv preprint arXiv:2211.00724*, 2022b.

Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 497–506, 2023.

Gautam Kamath. The broader landscape of robustness in algorithmic statistics. *arXiv preprint arXiv:2412.02670*, 2024.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. PMLR, 2022.