CS839: Differential Privacy and Learning April 21, 2026

Lecture 25: Privacy Wrappers

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.

Suppose we have a function $f : \mathcal{X}^n \to \mathbb{R}$ and a dataset $x \in \mathcal{X}^n$ and we want to privately approximate $f(x)$. How can we do this? Most techniques rely on our knowledge of certain properties of f such as its global sensitivity. Today we'll talk about a group of approaches which allow us to privately estimate f with little or no prior knowledge about it. For general functions these approaches are computationally inefficient, but in some cases they can be made efficient and, more importantly, they help us understand what is possible to learn privately.

In this lecture we will abuse notation in an attempt to make things consistent. For datasets $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $x' = (x'_1, \ldots, x'_n) \in \mathcal{X}^n$, their Hamming distance is

$$d_H(x, x') = \sum_{i=1}^{n} \mathbb{1}\{x_i \neq x'_i\}.$$

We interpret this as "the number of swaps needed to turn x' into x." We will use the same notation for datasets $x, x' \in \mathcal{X}^*$ which are multisets[1] of possibly different sizes. Write

$$d_H(x, x') = \min\{|s| : x \setminus s =$$

1 Review: Inverse Sensitivity Mechanism

We've seen before the *inverse sensitivity mechanism*, which for a dataset $x \in \mathcal{X}^n$ and function $f : \mathcal{X}^n \to \mathcal{Y}$ with $|\mathcal{Y}| \leq \infty$ releases

$$p(y) \propto \exp\left(-\frac{\varepsilon}{2} \cdot \text{len}_f(y; x)\right) \tag{1}$$

where

$$\text{len}_f(y; x) = \min_{x'} \{d_H(x, x') \mid f(x') = y)\}. \tag{2}$$

We've already seen the privacy guarantees in this class, but we'll prove it again for a refresher.

Claim 1.1. *The inverse sensitivity mechanism is ε-DP.*

Proof. Fix output y and adjacent x, x'. Then we can apply the triangle inequality:

$$\begin{aligned}
\text{len}_f(y; x) &= \min_{x''} \{d_H(x, x'') \mid f(x'') = y)\} \\
&\leq \min_{x''} \{d_H(x, x') + d_H(x', x'') \mid f(x'') = y)\} \\
&\leq 1 + \text{len}_f(y; x').
\end{aligned}$$

[1] That is, unordered sets where elements may be repeated.

Thus the inverse sensitivity mechanism is an instance of the exponential mechanism whose score function has global sensitivity at most one. □

A canonical instantiation of the inverse sensitivity mechanism is our friend who showed up last lecture: the exponential mechanism for the median [McSherry, 2009]:

$$p(y) \propto \exp\left(-\frac{\varepsilon}{2}\left|\frac{n}{2} - \#\{i : x_i \le y\}\right|\right),$$

which counts the number of data points one would have to change to make y a median.

We've also seen utility guarantees for this mechanism, but we'll state them today in a slightly different form.

Definition 1.2. The *k-neighborhood* of x is $\mathcal{N}_k(x) = \{x' : d_H(x, x') \le k\}$.

Theorem 1.3. *For $k = \frac{2}{\varepsilon}\log(\frac{2|\mathcal{Y}|^2}{\beta\varepsilon})$, with probability at least $1-\beta$ the inverse sensitivity mechanism returns $\tilde{y}$ such that*

$$\min_{x' \in \mathcal{N}_k(x)} f(x') \le \tilde{y} \le \max_{x' \in \mathcal{N}_k(x)} f(x').$$

This is a very nice guarantee, very simple to set up, and in a strong sense it's asymptotically optimal: our approximation on $f(x)$ must look somewhat like our approximation on $f(x')$ for any x' with $d_H(x, x') \lesssim \frac{1}{\varepsilon}$. However, there are two major drawbacks.

- It is unclear how to compute len_f when $\mathcal{X}$ is infinite, even by brute force.
- In many problems there is an asymmetry between addition and removal: when you can add points, you can add outliers.

To elaborate on the second point, suppose I want to compute the mean of some Gaussian data, drawn i.i.d. from $\mathcal{N}(\mu, 1)$ with unknown μ. Then for any dataset x the 1-neighborhood of x contains datasets with empirical means arbitrarily far away from the mean of x. This is for any dataset: the fact that x is well-concentrated about μ does not matter.

However, if we consider datasets x' obtained by *deleting* entries of x, then we are much more constrained. If you take a typical dataset x drawn from a Gaussian $\mathcal{N}(\mu, \sigma^2)$, then by deleting k elements you can expect to move the empirical mean by no more than $\approx \sigma k$. In the next section, we will discuss a method that competes with this removals-only baseline.

2 Privacy Wrappers

We introduce notation for the datasets that can be obtained by deletions only.

Definition 2.1. The *k-down-neighborhood* of x is $\mathcal{N}_k^{\downarrow}(x) = \{x' : x' \subseteq x, |x'| \ge |x| - k\}$.

The following is a recent result of Linder, Raskhodnikova, Smith, and Steinke [2025].

Theorem 2.2. *There is an (ε, δ)-DP algorithm that, given any $f : \mathcal{X}^n \to \mathcal{Y}$, for $k = \frac{\log(1/\delta)}{\varepsilon} \cdot 2^{O(\log^* |\mathcal{Y}|)}$, with probability 1 returns $\tilde{y}$ such that*

$$\min_{x' \in \mathcal{N}_k^{\downarrow}(x)} f(x') \le \tilde{y} \le \max_{x' \in \mathcal{N}_k^{\downarrow}(x)} f(x').$$

The algorithm accesses f via queries on datasets in $\mathcal{N}_k^{\downarrow}(x)$.

Ignoring the time needed to compute f, the algorithm's running time is dominated by the large number of queries it makes: $\binom{n}{k} \approx n^{O(\log(1/\delta)/\varepsilon)}$. This is called a *privacy wrapper* because it is DP for any function f. Linder et al. [2025] also give a pure-DP version that succeeds with probability $1-\beta$ and achieves $k = O(\frac{1}{\varepsilon}\log(|\mathcal{Y}|/\beta))$; since $\mathcal{N}_k^{\downarrow}(x) \subset \mathcal{N}_k(x)$, this is never worse than the guarantees in Theorem 1.3. The algorithm is much more complicated than the inverse sensitivity mechanism (we will sketch some main ideas below), but it addresses the addition/removal drawback as well as the computability question: the algorithm only ever operates on data points it has at hand.

The inverse sensitivity guarantees can be phrased in terms of *local sensitivity*, introduced in Lecture 8:

$$\mathrm{LS}_f(x) = \max_{\substack{x' \\ \text{s.t. } x \sim x'}} |f(x) - f(x')|.$$

Analogously, the guarantees of Theorem 2.2 can be given in terms of *down sensitivity*:

$$\mathrm{DS}_f(x) = \max_{\substack{x' \\ \text{s.t. } x' \subseteq x \\ |x'|=|x|-1}} |f(x) - f(x')|.$$

3 Analysis: Monotonicity and Shifted Inverse

Linder et al. [2025] build upon the results of Fang et al. [2022], who gave an algorithm which approximates *monotone* functions in a similar way.

Definition 3.1. A function $f : \mathcal{X}^* \to \mathcal{Y}$ is *monotone* if $x \subseteq x'$ implies $f(x) \le f(x')$.

Theorem 3.2. *For any monotone $f : \mathcal{X}^* \to \mathcal{Y}$, there is an algorithm which is ε-DP and, for $k = O(\frac{1}{\varepsilon}\log(|\mathcal{Y}|/\beta))$, with probability $1-\beta$ returns $\tilde{y}$ such that*

$$\min_{x' \in \mathcal{N}_k^{\downarrow}(x)} f(x') \le \tilde{y} \le f(x).$$

This is not a privacy wrapper (since it may fail to be DP for non-monotone functions) but it contains many of the key ideas needed later. Fang et al. [2022] use a mechanism, which they call the *shifted inverse sensitivity mechanism*, which uses the following loss function:

$$\mathrm{len}_f^{\downarrow}(y, x) = \min\{|x \setminus x'| : x' \subseteq x, f(x') \le y\}. \tag{3}$$

This is a one-sided version of Equation (2): Hamming distance measures the number of swaps needed to turn tuple x into tuple x', while $|x \setminus x'|$ measures the number of removals needed to turn x into x', where both are multisets.[2] The following is a key argument of Fang et al. [2022].

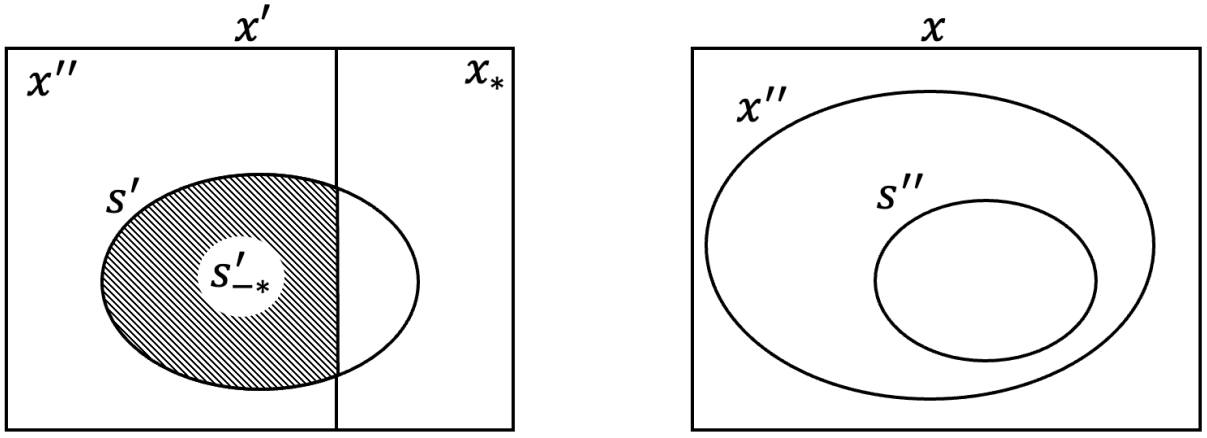
Claim 3.3. *For monotone f, any $y \in \mathcal{Y}$ and $x \sim x'$, $|\mathrm{len}_f^{\downarrow}(y, x) - \mathrm{len}_f^{\downarrow}(y, x')| \le 1$.*

We present the (rather slick) proof for completeness; our presentation is taken directly from Linder et al. [2025].

Proof. Fix $y \in \mathcal{Y}$ and $x, x' \in \mathcal{X}^*$ which differ in one addition or removal. We prove two subclaims. For any $x'' \in \mathcal{X}$:

1. if $x'' \subseteq x'$ and f is monotone then $\mathrm{len}_f^{\downarrow}(y, x'') \le \mathrm{len}_f^{\downarrow}(y, x')$, and
2. if $x'' \subset x$ then $\mathrm{len}_f^{\downarrow}(y, x) \le \mathrm{len}_f^{\downarrow}(y, x'') + |x \setminus x''|$.

[2] That is, x and x' are unordered sets which may contain duplicates. The set difference operation respects duplication: $\{1, 1, 2\} \setminus \{1\} = \{1, 2\}$.



(a) (b)

Figure 1: Figure accompanying the proof of Claim 3.3. (a) For subclaim 1, x'' and x_* partition x'. While the witness s' may span both, shaded s'_{-*} is contained in x''. (b) For subclaim 2, the witness s'' is contained in x'', which is contained in x.

Subclaim 1 Let $x_* = x' \setminus x''$, so $x' = x'' \cup x_*$ and $x'' \cap x_* = \emptyset$. Let $s' \subset x'$ be a set which witnesses the value of $\text{len}_f^{\downarrow}(y, x')$, i.e., $\text{len}_f^{\downarrow}(y, x') = |x' \setminus s'|$ and $f(s') \le y$. Let $s'_{-*} = s' \setminus x_*$. We have $f(s'_{-*}) \le f(s') \le y$ by monotonicity. Thus

$$\begin{aligned}
\text{len}_f^{\downarrow}(y, x'') &= \min\{|x'' \setminus s''| : s'' \subset x'', f(s'') \le y\} \\
&\le |x'' \setminus s'_{-*}| \\
&\le |x' \setminus s'| \\
&= \text{len}_f^{\downarrow}(y, x')..
\end{aligned}$$

Subclaim 2 Let $s'' \subset x''$ be a set which witnesses the value of $\text{len}_f^{\downarrow}(y, x'')$, i.e., $\text{len}_f^{\downarrow}(y, x'') = |x'' \setminus s''|$ and $f(s'') \le y$. Then

$$\begin{aligned}
\text{len}_f^{\downarrow}(y, x) &= \min\{|x \setminus s| : s \subset x, f(s) \le y\} \\
&\le |x \setminus s''| \\
&= |x'' \setminus s''| + |x \setminus x''| \\
&= \text{len}_f^{\downarrow}(y, x'').
\end{aligned}$$

We apply these claims to $x'' = x \cap x'$, as

$$\text{len}_f^{\downarrow}(y, x) \le \text{len}_f^{\downarrow}(y, x'') + |x \setminus x''| \le \text{len}_f^{\downarrow}(y, x') + 1.$$

Since $\text{len}_f^{\downarrow}(y, x') \le \text{len}_f^{\downarrow}(y, x) + 1$ is symmetric, we are done. □

Fang et al. [2022] take this loss function and run the exponential mechanism to find $\tilde{y}$ approximately minimizing the loss

$$\ell(y, x) := |\text{len}_f^{\downarrow}(y, x) - k|.$$

With this, we have almost proved Theorem 3.2: we know there is one value of y achieving small $\ell(y, x)$, since $\text{len}_f^{\downarrow}(f(x), x) = 0$, so the exponential mechanism will reliably return a $\tilde{y}$ which is not much worse. (Unfortunately, there are some annoying but minor technical details; see Steinke [2023] for more discussion).

3.1 Medians, Inverse Sensitivity, and Shifting

The exponential mechanism we've seen for the median uses the loss

$$\ell_{\text{median}}(y; x) = \left| \frac{n}{2} - \#\{i : x_i \le y\} \right|.$$

This is exactly "the number of points we need to change in x to make y a median," thus inverse sensitivity. Here is another view: it is a shifted version of the maximum. We can use

$$\ell_{\max}(y; x) = \#\{i : x_i \le y\}$$

as a measure of how well y approximates the maximum of x. This is clearly a monotone function. It measures "how many points we need to remove from x to make y greater than the maximum." We can thus regard the median (or in fact any quantile) as a shifted version of the maximum.

Fang et al. [2022] use the loss function

$$\text{len}_f^{\downarrow}(y, x) = \min\{|x \setminus x'| : x' \subseteq x, f(x') \le y\}$$

and run the exponential mechanism to minimize $|\text{len}_f^{\downarrow}(y, x) - k|$ for some parameter k. This is very familiar to us! Our loss function for the max is

$$\begin{aligned} \ell_{\max}(y; x) &= \#\{i : x_i \le y\} \\ &= \min\{|x \setminus x'| : x' \subseteq x, \max x' \le y\}. \end{aligned}$$

And we know how to shift it.

3.2 Monotonization and Generalized Interior Points

At a very high level, we get to Theorem 2.2 with two additional steps.

- For an arbitrary function f, we monotonize it. Roughly, we run Fang et al. [2022]'s mechanism on the function
$$M_{\tilde{k}}[f](x) \approx \max\{f(x') : x' \subseteq x, |x'| \ge |x| - \tilde{k}\}.$$
Here $\tilde{k}$ needs to be taken to be random, but it lives on the same scale as k in the theorem statement.
- We also replace the exponential mechanism, which we suggested above can be viewed as finding an approximate median, with a "generalized interior point" method. This is where the $2^{\log^*(|\mathcal{Y}|)}$ comes in.

This is far from a proof, but it at least names the key ingredients.

4 Further Reading

There are many open directions on this topic, the most pressing of which is: what is the analog of this discussion in multiple dimensions?

For more discussion of down sensitivity and interior points, see the blog post Steinke [2023]. Another open direction is understanding when these techniques can be made efficient; see Steinke and Steinke [2025], Brown et al. [2026] for work in this direction.

We previously discussed the strong connections between robust algorithms and private algorithms. We can now draw more parallels. The guarantees of Theorem 2.2 depend on f's behavior on all large subsets of x. Compare with standard *stability condition* from algorithmic robust statistics.

Definition 4.1 (Diakonikolas and Kane [2023], Definition 2.1)**.** A finite set $S \subset \mathbb{R}^d$ is (α, β)-stable with respect to a vector μ if for every unit vector v and every $S' \subseteq S$ with $|S'| \geq (1-\alpha)|S|$ we have

1. $\left|\frac{1}{|S'|}\sum_{x\in S'} v \cdot (x-\mu)\right| \leq \beta$, and
2. $\left|\frac{1}{|S'|}\sum_{x\in S'} (v \cdot (x-\mu))^2 - 1\right| \leq \beta^2/\alpha$.

The question of "how much can dropping a few point change my estimate?" also arises explicitly in a topic called *robustness auditing* [Broderick et al., 2020].

References

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: when can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999*, 2020.

Gavin Brown, Ephraim Linder, Mahbod Majid, and Vikrant Singhal. Privately estimating monotone statistics in polynomial time. *arXiv preprint arXiv:2605.27912*, 2026.

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

Juanru Fang, Wei Dong, and Ke Yi. Shifted inverse: A general mechanism for monotonic functions under user differential privacy. In *Proceedings of the SIGSAC Conference on Computer and Communications Security, CCS*, pages 1009–1022. ACM, 2022. doi: 10.1145/3548606.3560567. URL https://doi.org/10.1145/3548606.3560567.

Ephraim Linder, Sofya Raskhodnikova, Adam Smith, and Thomas Steinke. Privately evaluating untrusted black-box functions. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, page 2350–2361, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718247. URL https://doi.org/10.1145/3717823.3718247.

Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

Günter F Steinke and Thomas Steinke. Privately estimating black-box statistics. *arXiv preprint arXiv:2510.00322*, 2025.

Thomas Steinke. Beyond local sensitivity via down sensitivity, 2023. URL https://differentialprivacy.org/down-sensitivity/.