CS839: Differential Privacy and Learning April 9, 2026

Lecture 22: Score Attack and Fingerprinting

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 these notes we will discuss techniques for proving lower bounds for statistical estimation under approximate differential privacy. Our presentation draws directly from Duchi [2025] and Smith and Ullman [2025]; see Section 5 for discussion of the literature.

Recall from our analysis of FriendlyCore that we gave an (ε, δ)-DP algorithm which, given n i.i.d. samples from a Gaussian $\mathcal{N}(\mu, \mathbb{I})$, if $n \gtrsim \frac{\log(1/\delta)}{\varepsilon}$, with probability at least 9/10 returns an estimate $\tilde{\mu}$ such that

$$\|\tilde{\mu} - \mu\|_2^2 \lesssim \frac{d}{n} + \frac{d^2}{n^2\varepsilon^2} \cdot \log(1/\delta).$$

The first term is required even without privacy: the empirical mean satisfies $\mathbb{E}\,\|\hat{\mu} - \mu\|_2^2 = \frac{d}{n}$, and this is optimal.

Today we will work toward proving a lower bound: if A is (ε, δ)-DP and δ is sufficiently small, then

$$\mathbb{E}\Big[\|A(X) - \mu\|_2^2\Big] \gtrsim \frac{d^2}{\varepsilon^2 n^2}.$$

This establishes that, up to the $\log(1/\delta)$ factor and ignoring the difference between accurate-on-average and accurate-with-high-probability, FriendlyCore is optimal.[1]

We remark that, for this specific result, the techniques developed here are somewhat overkill. There are more direct ways to prove this lower bound. However, the framework here applies in much greater generality and also helps us develop intuition for why such results hold.

1 Membership Inference Attacks

At the core of these proofs is a particular *membership inference attack* (MIA). In a standard MIA, an attacker is given a data point x' and the output of an algorithm $A(X)$, where X is the training data. The attacker's job is to distinguish whether $x' \in X$ (the IN case) or $x' \notin X$ (the OUT case). The is the most common attack used to evaluate privacy in practice, in addition to its use as a theoretical tool.

We will consider MIAs which compute a *test statistic* $T(x', A(X))$ and then threshold, i.e., for some $\tau \in \mathbb{R}$, the attacker returns "IN" if $T(x', A(X)) \geq \tau$ and "OUT" otherwise.

Example 1.1. $A(X)$ may output the weights w of a trained language model, i.e., the weights define a distribution $q_w(x)$ with which we can assign probabilities to text. One baseline MIA use the

[1] Algorithms with better sample complexity are known, although we will not cover them today.

negative loss as a test statistic: $T(x', A(X)) = \log p_w(x')$, since we often expect training examples to have lower loss than examples that weren't seen during training.

Today we will explore the implications of a particular test statistic/MIA for a wide variety of learning tasks. For us, the MIA will only be a tool in the theoretical analysis. We will not actually run it.

2 Setup

We start by assuming a family of distribution $\{p_\theta\}$, where each distribution corresponds to a particular underlying parameter θ. We use the standard negative log-likelihood loss: for a data point x,

$$\ell_\theta(x) = -\log p_\theta(x).$$

We define the *score*

$$\dot{\ell}_\theta(x) = \nabla_\theta \ell_\theta(x)$$

and the *Fisher information*

$$J(\theta) = \mathbb{E}_\theta\Big[\dot{\ell}_\theta(x))\dot{\ell}_\theta(x)^T\Big],$$

where we use the notation $\mathbb{E}_\theta$ to denote the expectation with respect to $x \sim p_\theta$, and omit the subscript entirely when it is unambiguous. We will assume that $\mathbb{E}_\theta[\dot{\ell}_\theta(x)] = 0$, which should be interpreted as: on average over data from the distribution p_θ, the parameter θ has the lowest loss.[2]

At this point, it may unclear how to interpret the score and Fisher information, or to understand why the latter is called "information." That's ok, we will build up the intuition we need.

Example 2.1. Consider the family of d-dimensional Gaussians with identity covariance: $\{\mathcal{N}(\mu, \mathbb{I})\}_{\mu \in \mathbb{R}^d}$. They have density

$$p_\mu(x) = \frac{1}{(2\pi)^{d/2}} \cdot \exp\left(-\frac{1}{2}\|x-\mu\|_2^2\right).$$

Thus the loss is $\ell_\mu(x) = \frac{1}{2}\|x-\mu\|_2^2 - \frac{d}{2}\log(2\pi)$, the score is $\dot{\ell}_\theta(x) = x - \mu$, and the Fisher information is $J(\mu) = \mathbb{E}[(x-\mu)(x-\mu)^T] = \mathbb{I}$.

3 The Score Attack

The *score attack*, introduced in Cai et al. [2023], uses the following test statistic:

$$T_\theta(x', A(X)) = \langle A(X) - \theta, \dot{\ell}_\theta(x')\rangle.$$

Why should this be a useful quantity to consider? On the one hand, when $x' \notin X$, the two sides of the inner product are independent. We will formalize this and show that, in this OUT case, T_θ is mean-zero and unlikely to be too large.

On the other hand, suppose $x' \in X$ and let vector v denote $\dot{\ell}_\theta(x')$. Intuitively, v tells us "if you move your estimate away from θ in the direction of v, you'll get lower loss on x'." This test statistic captures a sense in which the algorithm's empirical estimate differs from the true underlying parameter.

To use this test statistic to prove lower bounds, we execute three steps that we make quantitative below.

[2]Today we will not need to worry about if the loss function is convex.

- Step 1 (IN): when $x \in X$, $T_\theta(x, A(X))$ is "large."
- Step 2 (OUT): when $x' \notin X$, $T_\theta(x', A(X))$ is "small."
- Step 3 (Privacy): For any x and x',

$$T_\theta(x, A(X)) \approx T_\theta(x', A(X)).$$

We will deal with these steps in reverse order.

Step 3 (Privacy) We want to apply the fact that A is DP. Above, we framed this as: for $x \in X$ and $x' \notin X$, we have the following indistinguishability:

$$T_\theta(x, A(X)) \approx_{\varepsilon,\delta} T_\theta(x', A(X)).$$

Equivalently, we can think of x_1 as begin the first element of X and let dataset X' be equal to X except with a fresh draw of the first entry. Then the above is equal to asking

$$T_\theta(x_1, A(X)) \approx_{\varepsilon,\delta} T_\theta(x_1, A(X')).$$

In this formulation, it's easier to see that this holds by postprocessing.

We apply indistinguishability via the following lemma [Steinke and Ullman, 2016, Kamath et al., 2019, Cai et al., 2021, 2023].

Lemma 3.1. *Abbreviate $T = T_\theta(x, A(X))$ and $T' = T_\theta(x', A(X))$ and assume $\varepsilon \in (0, 1]$. For any $B > 0$ we have*

$$\mathbb{E}[T] \le \mathbb{E}[T'] + 2\varepsilon\mathbb{E}[|T'|] + 2\delta B + \int_B^\infty \Pr(|T| > b)\mathrm{d}b.$$

To apply this technique one has to pick an appropriate parameter B and ensure δ is sufficiently small, but today we will simply ignore the last two terms. We will prove that $\mathbb{E}[T'] = 0$ in our next lemma.

Step 2 (OUT case) We establish Step 2 with the following lemma. Observe that it does not use the fact that the algorithm is differentially private.

Lemma 3.2 (OUT Case)**.** *Let $X \sim p_\theta^{\otimes n}$ and let $x' \sim p_\theta$ independently. Then $\mathbb{E}[T_\theta(x', A(X))] = 0$ and*

$$\mathbb{E}\big[\big|T_\theta(x', A(X))\big|\big] \le \sqrt{\mathbb{E}\Big[\|A(X) - \theta\|_2^2\Big]}\sqrt{\|J(\theta)\|_2}.$$

Proof. The test statistic is mean-zero because $\dot\ell_\theta(x')$ is independent of $A(X) - \theta$ and itself mean-zero.

For the second inequality, we work with the square and apply Jensen's inequality:

$$\begin{aligned}
\mathbb{E}_{X,x'}\big[\big|T_\theta(x', A(X))\big|\big]^2 &\le \mathbb{E}_{X,x'}\big[T_\theta(x', A(X))^2\big] \\
&= \mathbb{E}_{X,x'}\Big[(A(X) - \theta)^T \dot\ell_\theta(x')\dot\ell_\theta(x')^T(A(X) - \theta)\Big] \\
&= \mathbb{E}_X\Big[(A(X) - \theta)^T \mathbb{E}_{x'}\Big[\dot\ell_\theta(x')\dot\ell_\theta(x')^T\Big](A(X) - \theta)\Big],
\end{aligned}$$

where we have written out the inner product and pushed the expectation over x' inside (using the fact that X and x' are independent).

Then we recognize that the inner expectation is exactly the Fisher information:

$$\mathbb{E}_{X,x'}\left[\left|T_\theta(x', A(X))\right|\right]^2 \le \mathbb{E}_X\left[(A(X)-\theta)^T J(\theta)(A(X)-\theta)\right].$$

Finally, we apply the fact that for any matrix A and vector v we have $v^T A v \le \|v\|_2^2 \|A\|_2$. Taking square roots on both sides, we have finished the proof. □

Step 1 (IN) To keep the discussion general, we will for now assume this step is already done. (We will see a specific example below, in Section 4). Assume we know some number C such that

$$C \le \sum_{i=1}^n \mathbb{E}[T_\theta(x_i, A(X))],$$

where $x_1, \dots, x_n$ are the n elements of X.

Putting Things Together For independent x', we have

$$\begin{aligned} C \overset{(1)}{\le} \sum_{i=1}^n \mathbb{E}[T_\theta(x_i, A(X))] &\overset{(3)}{\lesssim} 2\varepsilon \sum_{i=1}^n \mathbb{E}[\left|T_\theta(x', A(X))\right|] \\ &= 2\varepsilon n \cdot \mathbb{E}[\left|T_\theta(x', A(X))\right|] \\ &\overset{(2)}{\le} 2\varepsilon n \sqrt{\mathbb{E}\left[\|A(X)-\theta\|_2^2\right]} \sqrt{\|J(\theta)\|_2} \end{aligned}$$

Rearranging, we get the following lower bound on the mean squared error of our DP algorithm:

$$\mathbb{E}\left[\|A(X)-\theta\|_2^2\right] \gtrsim \frac{C^2}{\|J(\theta)\|_2} \cdot \frac{1}{\varepsilon^2 n^2}.$$

To apply this to a specific problem, the main step is to establish Step 1, showing some problem-specific lower bound C. Such a result is usually called a *fingerprinting lemma*. We now turn to establishing such a lemma for Gaussian mean estimation.

4 Fingerprinting For Gaussians

In this section we will follow the presentation of Smith and Ullman [2025] and show that for an accurate Gaussian mean estimator, the average data point leaves its "fingerprints" on the output. We use a Bayesian setup, where μ is drawn from a prior distribution. Recall that, for Gaussians with identity covariance, the score is $\dot{\ell}_\mu(x) = x - \mu$.

Lemma 4.1 (Fingerprinting for Gaussian Means)**.** *Let $\mu \sim \mathcal{N}(0, \mathbb{I}_d)$ and $X = (x_1, \dots, x_n) \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes n}$. Suppose algorithm $A(X)$ satisfies: $\mathbb{E}[\left\|A(X) - \bar{X}\right\|_2^2] \le \frac{d}{16}$. Then*

$$\sum_{i=1}^n \mathbb{E}[\langle A(X) - \mu, x_i - \mu\rangle] \ge \frac{d}{2}.$$

Before the proof, a couple of notes. First, we do not use anything about privacy here. Second, we have made our job easier by assuming that $A(X)$ is somewhat close to the empirical mean; the real version of this argument would assume that $A(X)$ is close to μ, but this requires a more subtle

argument. Third, in this step we do not assume that A is very accurate at all: blindly guessing $A(X) = 0$ already yields error around d.

We need the following fact about Gaussians. Note that this posterior depends only on $\bar{X}$, which is a *sufficient statistic* for the population mean.

Claim 4.2. *Let $\mu \sim \mathcal{N}(0, \mathbb{I}_d)$ and $X = (x_1, \dots, x_n) \sim \mathcal{N}(\mu, \mathbb{I}_d)^{\otimes n}$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n x_i$ be the empirical mean. Conditioned on X, the posterior over μ is*

$$\mu \mid X \sim \mathcal{N}\left(\frac{n}{n+1}\bar{X}, \frac{1}{n+1}\mathbb{I}\right).$$

Proof. By Claim 4.2, we can write $\mu = \frac{n}{n+1}\bar{X} + Z$ where $Z \sim \mathcal{N}(0, \frac{1}{n+1}\mathbb{I})$ is independent of X. This is the conceptual crux of the proof: in this Bayesian setup, we can think of first generating X, computing $A(X)$, and *then* sampling μ independently from the posterior.

Write mechanism $A(X) = \bar{X} + e$, where e is a random vector. By assumption, $\mathbb{E}\,\|e\|_2^2 \le \frac{d}{16}$. (Note that we do not assume e is independent of X or $\bar{X}$; we are not assuming anything about the algorithm being run.)

Dividing by n, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\langle A(X) - \mu, x_i - \mu\rangle] &= \mathbb{E}_i[\mathbb{E}[\langle A(X) - \mu, x_i - \mu\rangle]] \\ &= \mathbb{E}[\langle A(X) - \mu, \bar{X} - \mu\rangle] \\ &= \mathbb{E}[\langle \bar{X} + e - \mu, \bar{X} - \mu\rangle] \\ &= \mathbb{E}[\langle \bar{X} - \mu, \bar{X} - \mu\rangle] + \mathbb{E}[\langle e, \bar{X} - \mu\rangle].\end{aligned}$$

The first term on the right-hand side is $\mathbb{E}\left\|\bar{X} - \mu\right\|_2^2 = \frac{d}{n}$. For the second, we plug in fact that $\mu = \frac{n}{n+1}\bar{X} + Z$ where Z is zero-mean and independent of e:

$$\begin{aligned}\mathbb{E}[\langle e, \bar{X} - \mu\rangle] &= \mathbb{E}[\langle e, \bar{X} - \frac{n}{n+1}\bar{X} + Z\rangle] \\ &= \mathbb{E}[\langle e, \frac{1}{n+1}\bar{X} + Z\rangle] \\ &= \frac{1}{n+1}\mathbb{E}[\langle e, \bar{X}\rangle].\end{aligned}$$

But now we apply Cauchy–Schwarz and have

$$\mathbb{E}[\langle e, \bar{X} - \mu\rangle] \ge -\frac{1}{n+1}\mathbb{E}\,\|e\|_2\,\mathbb{E}\left\|\bar{X}\right\|_2 \ge -\frac{1}{n}\mathbb{E}\,\|e\|_2\,\mathbb{E}\left\|\bar{X}\right\|_2.$$

By assumption $\mathbb{E}\,\|e\|_2 \le \sqrt{d}/4$, and

$$\mathbb{E}\left\|\bar{X}\right\|_2 \le \mathbb{E}[\left\|\bar{X} - \mu\right\|_2 + \|\mu\|_2] = \sqrt{\frac{d}{n}} + \sqrt{d} \le 2\sqrt{d}.$$

Combining everything, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\langle A(X) - \mu, x_i - \mu\rangle] &\ge \mathbb{E}\left\|\bar{X} - \mu\right\|_2^2 - \frac{1}{n}\mathbb{E}\,\|e\|_2\,\mathbb{E}\left\|\bar{X}\right\|_2 \\ &\ge \frac{d}{n} - \frac{d}{2n} = \frac{d}{2n}.\end{aligned}$$

Multiplying by n, we have finished the proof. □

5 Further Reading

It can be difficult to follow the literature on this topic. This approach for lower bounds for differential privacy first appeared in Bun et al. [2014]; the arguments relied on the existence of *fingerprinting codes* developed in cryptography. Over the years such arguments have been used in many places. A lot has been simplified (notice that we did not have to talk about cryptography).

The explicit connections between the score-based test statistics and lower bounds for DP were developed in the last few years over a number of papers [Cai et al., 2021, Kamath et al., 2022, Cai et al., 2023, Portella and Harvey, 2025]. Both Duchi [2025] and Portella and Harvey [2025] give more detailed overviews of these developments.

The calculation for Gaussian fingerprinting is "standard," though I can attest that both assistant professors and LLMs in early 2026 are more than capable of messing it up. See Smith and Ullman [2025] for its direct application to this topic.[3] Kamath et al. [2019] give the entire argument, albeit in a presentation that is more complicated than what we needed in this lecture.

References

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10, 2014.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

T Tony Cai, Yichen Wang, and Linjun Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152*, 2023.

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

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. *Advances in Neural Information Processing Systems*, 35: 24405–24418, 2022.

Victor S Portella and Nicholas JA Harvey. Lower bounds for private estimation of gaussian covariance matrices under all reasonable parameter regimes. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 4640–4667. PMLR, 2025.

Adam Smith and Jonathan Ullman. Privacy in statistics and machine learning, 2025. URL `https://dpcourse.github.io/2025-spring/lecnotes-web/DP-S25-notes-lec-10-adv-composition.pdf`.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2016.

[3] `https://dpcourse.github.io/2025-spring/lecnotes-web/DP-S25-notes-lec-22-lower-bounds-2-MIA.pdf`