17 The Gaussian Mechanism and Approximate Differential Privacy

This section is based heavily on the course Privacy in Statistics and Machine Learning by Adam Smith and Jonathan Ullman, primarily Lecture 8 (Spring 2025) [SU21b].

The Laplace mechanism calibrates its noise to ℓ_1-sensitivity. For high-dimensional statistics, we can often get better accuracy by calibrating to ℓ_2-sensitivity instead. Gaussian noise lets us do this, but requires a slight relaxation of the privacy definition.

17.1 Approximate Differential Privacy

Pure differential privacy requires a multiplicative bound for every event, even extremely unlikely ones. We will allow a small additive slack δ.

Definition 17.1. A randomized algorithm $\mathcal{A}\colon \mathcal{X}^n \to \mathcal{R}$ is (ε, δ)-*differentially private* if for every pair of neighboring datasets S, S' and every event $R \subseteq \mathcal{R}$,

$$\mathbb{P}[\mathcal{A}(S) \in R] \le e^{\varepsilon}\, \mathbb{P}[\mathcal{A}(S') \in R] + \delta.$$

We call this *approximate differential privacy*, and call the case $\delta = 0$ *pure differential privacy*. The parameter δ needs to be small: an algorithm that independently publishes each record with probability δ already satisfies $(0, \delta)$-DP, but exposes $n\delta$ records on average. This motivates choosing $\delta \ll 1/n$.

The basic properties extend naturally. Postprocessing preserves (ε, δ)-DP, and adaptive composition of mechanisms with parameters $(\varepsilon_t, \delta_t)$ gives $(\sum_t \varepsilon_t, \sum_t \delta_t)$-DP, under the same conditions as Lemma 15.5. Group privacy for k changed records gives parameters $(k\varepsilon, \delta \sum_{j=0}^{k-1} e^{j\varepsilon})$, by chaining the defining inequality.

The hypothesis testing view. The argument from Theorem 15.7 now gives

$$\mathrm{TPR} \le \min\{e^{\varepsilon}\mathrm{FPR} + \delta,\ 1 - e^{-\varepsilon}(1 - \mathrm{FPR} - \delta),\ 1\},$$
$$\mathrm{TPR} \ge \max\{e^{-\varepsilon}(\mathrm{FPR} - \delta),\ 1 - e^{\varepsilon}(1 - \mathrm{FPR}) - \delta,\ 0\}.$$

As before, these constraints for all testers and all neighboring pairs are equivalent to the definition. They also hold in our earlier i.i.d. membership game by the same replacement argument. In particular, an attack may now have TPR as large as δ even when FPR $= 0$. Figure 14 uses $\delta = 0.05$ to make the widening of the testing region visible.

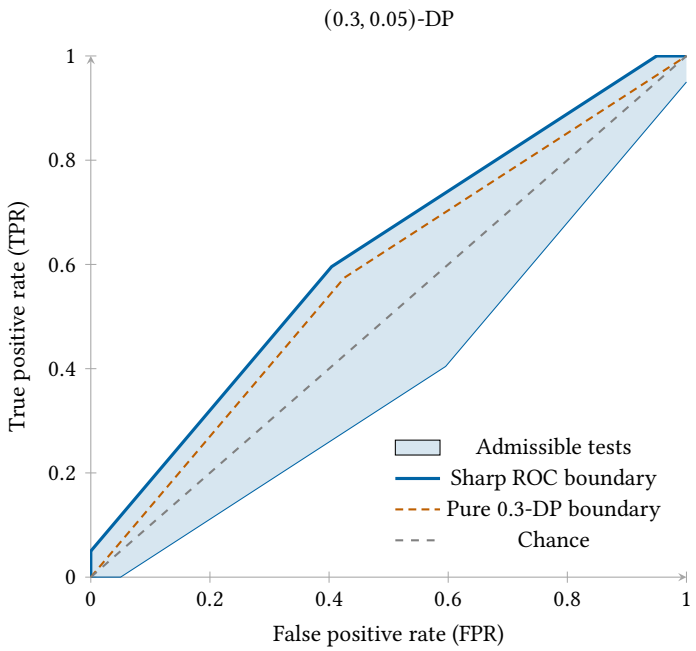


Figure 14: The hypothesis testing region for $(0.3, 0.05)$-DP, with the pure 0.3-DP upper boundary shown for comparison. Every test lies in the shaded region. The blue curve is the sharp upper ROC boundary, attainable by a mechanism that reveals its input bit with probability 0.05 and otherwise uses 0.3-randomized response, with the two branches distinguished in the output. The extra slack allows TPR 0.05 at FPR 0.

17.2 The Gaussian Mechanism

For a fixed function $f\colon \mathcal{X}^n \to \mathbb{R}^d$, define its *global ℓ_2-sensitivity* by

$$\Delta_2(f) := \sup_{S,S' \text{ neighboring}} \|f(S) - f(S')\|_2.$$

The norm inequalities give $\Delta_2(f) \le \Delta_1(f) \le \sqrt{d}\,\Delta_2(f)$. In one dimension the two sensitivities coincide, but in higher dimensions they can differ by a factor of $\sqrt{d}$.

The Gaussian mechanism releases $f(S) + W$ with $W \sim \mathcal{N}(0, \sigma^2 I_d)$. Equivalently, it adds independent $\mathcal{N}(0, \sigma^2)$ noise to each coordinate. Unlike Laplace noise, shifted Gaussians with different means have an unbounded density ratio, so this mechanism cannot satisfy pure DP for a nonconstant f. Approximate DP lets us control the probability of the outputs where the ratio is large.

Algorithm 10: The Gaussian mechanism

Input: A dataset $S \in \mathcal{X}^n$, a fixed function $f\colon \mathcal{X}^n \to \mathbb{R}^d$, a fixed bound $\Delta \geq \Delta_2(f)$ with $\Delta > 0$, and parameters $0 < \varepsilon \leq 1$ and $\delta \in (0,1)$.
Choose

$$\sigma \geq \frac{\Delta}{\varepsilon}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right).$$

Output: $\mathcal{A}(S) = f(S) + W$, where $W \sim \mathcal{N}(0, \sigma^2 I_d)$.

We use a slightly conservative choice of constants to keep the proof simple. The important dependence is $\sigma = O(\Delta\sqrt{\log(2/\delta)}/\varepsilon)$.

Theorem 17.2. *Algorithm 10 is (ε, δ)-differentially private. Its error satisfies*

$$\sqrt{\mathbb{E}[\|\mathcal{A}(S) - f(S)\|_2^2]} = \sigma\sqrt{d}.$$

For every $\beta \in (0,1)$, with probability at least $1 - \beta$,

$$\|\mathcal{A}(S) - f(S)\|_\infty \leq \sigma\sqrt{2\log\frac{2d}{\beta}}.$$

In particular, in one dimension the root mean squared error is σ.

Proof. Fix an ordered pair of neighboring datasets S, S' and let $p_S, p_{S'}$ be their output densities. Define the *privacy loss*

$$L(y) := \log\frac{p_S(y)}{p_{S'}(y)}.$$

It suffices to show that $\mathbb{P}[L(\mathcal{A}(S)) > \varepsilon] \leq \delta$. Indeed, splitting any event R according to whether $L(y) \leq \varepsilon$ gives

$$\mathbb{P}[\mathcal{A}(S) \in R] \leq e^\varepsilon\,\mathbb{P}[\mathcal{A}(S') \in R] + \mathbb{P}[L(\mathcal{A}(S)) > \varepsilon].$$

This is a sufficient condition for approximate DP; the exceptional outputs may depend on the neighboring pair.

Let $h = f(S) - f(S')$ and substitute $y = f(S) + W$. The Gaussian density gives

$$L(f(S) + W) = \frac{\|W + h\|_2^2 - \|W\|_2^2}{2\sigma^2} = \frac{\langle W\,,\,h\rangle}{\sigma^2} + \frac{\|h\|_2^2}{2\sigma^2}.$$

Since $\langle W\,,\,h\rangle \sim \mathcal{N}(0, \sigma^2\|h\|_2^2)$, the privacy loss is distributed as $rG + r^2/2$, where $r = \|h\|_2/\sigma$ and $G \sim \mathcal{N}(0,1)$. Set $q = \sqrt{2\log(1/\delta)}$. The one-sided Gaussian tail bound gives $\mathbb{P}[G > q] \leq \delta$. Our choice of σ ensures $r \leq \varepsilon/(1+q)$, so on the event $G \leq q$,

$$L(f(S) + W) \leq rq + \frac{r^2}{2} \leq \frac{\varepsilon q}{1+q} + \frac{\varepsilon^2}{2(1+q)^2} \leq \varepsilon,$$

using $\varepsilon \leq 1$. Thus the exceptional event has probability at most δ, proving privacy. The argument also covers $h = 0$, when the privacy loss is identically zero.

For accuracy, $\mathbb{E}[\|W\|_2^2] = \sum_{j=1}^{d} \mathbb{E}[W_j^2] = d\sigma^2$. Finally, $\mathbb{P}[|W_j| > t\sigma] \le 2e^{-t^2/2}$, and a union bound over the d coordinates gives the stated maximum error. □

When does this improve accuracy? Suppose $\mathcal{X} = [0,1]^d$ and we release the vector of empirical means $f(S) = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $\Delta_1(f) = d/n$ and $\Delta_2(f) = \sqrt{d}/n$. The root mean squared error in each coordinate is

$$\text{Laplace:} \quad \frac{\sqrt{2}\,d}{\varepsilon n}, \qquad \text{Gaussian:} \quad O\left(\frac{\sqrt{d\log(2/\delta)}}{\varepsilon n}\right).$$

Thus, when $\log(2/\delta) \ll d$, the Gaussian mechanism gives substantially less error by using the smaller ℓ_2-sensitivity. In one dimension there is no such sensitivity advantage, and this Gaussian calibration incurs an extra factor of order $\sqrt{\log(2/\delta)}$ compared with Laplace noise. The benefit comes from the geometry of the high-dimensional statistic and the relaxation to approximate DP.

References

[SU21b] Adam Smith and Jonathan Ullman. *Lecture 9: Approximate Differential Privacy*. Privacy in Statistics and Machine Learning, Spring 2021. Linked as the notes for Lecture 8 in the Spring 2025 course schedule. 2021. (Link).