Gaussian certified unlearning in high dimensions

A hypothesis testing approach

Aaradhya Pandey
joint with Arnab Auddy, Haolin Zou, Arian Maleki, Sanjeev Kulkarni

Operations Research and Financial Engineering,
Princeton University

October 18, 2025

What is machine unlearning?

Formulation of **privacy** using ROC curves

Formulation of **accuracy** using fresh samples

A Newton-Raphson based unlearning procedure

A glimpse at some of the results

What is machine unlearning?

Machine Unlearning: Motivation

- **Why care about machine unlearning?** Companies collect user data to train their ML models. Users may later request that their personal records be deleted.
- **The resulting concern and motivation:** Does the deployed model still encode information about the removed data? Re-training from scratch for every deletion request is prohibitively expensive, motivating the field of machine unlearning.
- **The goal of machine unlearning:** It is to address the problem of **efficiently** removing the influence of individual data points from trained models.
  - **Privacy for users:** It enables them to exercise their right to be forgotten.
  - **Accuracy for company:** It retains the generalization capabilities of the model.

What is machine learning?

A mathematical framework of machine learning for GLM under ERM learning

- **Dataset that company receives:** We assume that company receives the dataset $\mathscr{D}_n := \{z_i = (x_i, y_i)\}_{i=1}^n$ with features $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ the response.
- **Generalized linear model assumption on the dataset:** We assume that the data points $\{z_i\}_{i=1}^n$ are IID from $P_{\beta^*}(x,y) = p(x)q(y|x^T\beta^*)$ with unknown $\beta^* \in \mathbb{R}^p$.
- **Empirical risk minimization framework of learning:** The goal of a learning procedure A is to learn β^* from $\mathscr{D}_n$. We choose to find it by optimizing a loss L.

$$\textbf{RERM: } \hat{\beta} = A(\mathscr{D}_n) := \underset{\beta \in \mathbb{R}^p}{\arg\min} \quad L(\beta) := \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \ell(y_i, x_i^\top \beta) + \lambda r(\beta).$$

- **Comments:** Popular choices of the **loss** $l(y,z)$ includes $-\log q(y|z)$ and the **regularizer** includes ridge ($r(\beta) = \|\beta\|_2^2$) or Lasso ($r(\beta) = \|\beta\|_1$).

A closer look at (a version of) the problem of machine unlearning

- **Relearning:** Given dataset $\mathscr{D}_n = \{z_i\}_{i=1}^n$ and a trained model $\hat{\beta} = A(\mathscr{D}_n)$ some subset $\mathscr{M} \subset [n]$ of users want their data $\mathscr{D}_{\mathscr{M}} := \{z_i\}_{i \in \mathscr{M}}$ to be removed. An ideal unlearning procedure is to retrain A from scratch on the remaining dataset $\mathscr{D}_{\setminus \mathscr{M}} := \mathscr{D}_n \setminus \mathscr{D}_{\mathscr{M}}$.

$$\hat{\beta}_{\setminus \mathscr{M}} := A(\mathscr{D}_{\setminus \mathscr{M}}) := \arg\min_{\beta \in \mathbb{R}^p} \quad L_{\setminus \mathscr{M}}(\beta) := \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \notin \mathscr{M}} \ell(y_i, x_i^T \beta) + \lambda r(\beta).$$

- **Unlearning:** We want to avoid full retraining from scratch $A(\mathscr{D}_{\setminus \mathscr{M}})$, but also want to obscure residual information $\mathscr{D}_{\mathscr{M}}$. We hope to construct a randomization procedure $\bar{A}$.

Unlearning procedure: $\quad \tilde{\beta}_{\setminus \mathscr{M}} := \bar{A}(\hat{\beta}, \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}_n), b)$

- **Comments:** We assume $\bar{A}$ has access to removal requests $\mathscr{D}_{\mathscr{M}}$, trained model $\hat{\beta}$ and some auxiliary information $T(\mathscr{D}_n)$ such as gradient or Hessian of loos function L on $\mathscr{D}_n$, at $\hat{\beta}$. b is a noise independent of $\mathscr{D}_n$ to be added during the unlearning step.

Desiderata for our unlearning procedure $\bar{A}$

We need **efficient** procedure $\bar{A}$ protecting user **privacy** and preserving model **accuracy**.

- **Efficiency:** We need $\bar{A}$ to be far more efficient than retraining from scratch $A(\mathscr{D}_{\setminus \mathscr{M}})$.
- **Privacy:** We need the two distributions to be 'indistinguishable' to an adversary.

Relearned: $\bar{A}(\hat{\beta}_{\setminus \mathscr{M}}, \emptyset, T(\mathscr{D}_{\setminus \mathscr{M}}), b)$[1] vs. **Unlearned:** $\bar{A}(\hat{\beta}, \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}), b)$

- **Accuracy:** We need the unlearned output $\bar{A}(\hat{\beta}, \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}), b)$ to have the same generalization capabilities as that of $\hat{\beta}_{\setminus \mathscr{M}}$ on a fresh sample from the population P_{β^*}.

Now we precisely describe the **privacy** and **accuracy** requirements for the unlearning procedure $\bar{A}$.

[1] Ideally this should have been just $\hat{\beta}_{\setminus \mathscr{M}}$ but that requires the original algorithm A to be randomized!

Formulation of **privacy** using ROC curves

Hypothesis testing and Receiver Operating Characteristic curves I

Definition (Trade-off function)

Given two probability distributions P, Q on a measurable space $(\mathscr{W}, \mathscr{F}_{\mathscr{W}})$, we define the *trade-off function* as the map $T(P,Q) : [0,1] \to [0,1]$ as

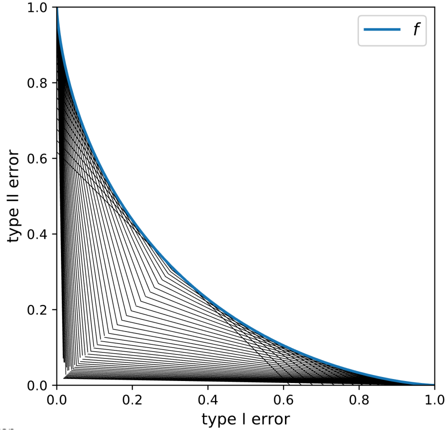
$$T(P,Q)(\alpha) := \inf_{\varphi}\left\{\beta_\varphi := \mathbb{E}_Q[1-\varphi] \;\middle|\; \alpha_\varphi := \mathbb{E}_P[\varphi] \leq \alpha,\ \varphi : \mathscr{W} \to [0,1] \text{ measurable}\right\}.$$

- In words, for any given type I error α, the trade-off function returns the smallest possible value of type II error β_φ over all possible test functions $\varphi : \mathscr{W} \to [0,1]$.
- **Neyman-Pearson lemma:** Optimal choice of φ is given by a likelihood ratio test[2].

$$\varphi(x) = 1\left(\log\frac{dQ}{dP} \geq z_\alpha\right) \text{ such that } P\left(\log\frac{dQ}{dP} \geq z_\alpha)\right) = \alpha$$

[2] Sometimes we need to consider randomized likelihood ratio test!

Blackwell ordering and ROC curves II



- **TOF:** A function $f : [0,1] \to [0,1]$ is a trade-off curve $T(P,Q)$ if and only if it is convex, continuous, non-increasing, and $f(\alpha) \leq 1-\alpha$.
- **Blackwell ordering :** If $T_{P_0,Q_0}(\alpha) \geq T_{P_1,Q_1}(\alpha)$ for all α, then (P_1,Q_1) is easier to distinguish than (P_0,Q_0).
- **Complete indistinguishability:** $f(\alpha) = 1-\alpha$ means random guess ROC with Bernoulli(α).

Blackwell's theorem and ROC curves III

Theorem (Equivalence of Blackwell informativeness and post-processing)

Let P_1, Q_1 be probability measures on Y_1 and P_0, Q_0 be probability measures on Y_0. The following two statements (Blackwell informativeness and post-processing) are equivalent:

1. ***Blackwell ordering:*** $T(P_0, Q_0)(\alpha) \geq T(P_1, Q_1)(\alpha)$ *for all* $\alpha \in [0, 1]$.
2. ***Post-processing:*** $\exists$ *a Markov Kernel* $\mathrm{R}: Y_1 \to Y_0$ *such that* $(P_0, Q_0) = (\mathrm{R}(P_1), \mathrm{R}(Q_1))$.

- **Gaussian trade-off function:** Let Φ be the Gaussian CDF. Then we have $G_\varepsilon(\alpha) := T(N(0,1), N(\varepsilon, 1))(\alpha) = \Phi(\Phi^{-1}(1-\alpha) - \varepsilon)$ for $\varepsilon \geq 0$.
- **Gaussian comparison:** $T(P, Q) \geq G_\varepsilon$ means it is at least as difficult to distinguish the pair (P, Q) than it is to a pair of Normals with one having a shifted mean.

Back to machine unlearning – Desiderata for unlearning procedure $\bar{A}$

- Recall that we need the two distributions to be 'indistinguishable' to an adversary.

  Relearned: $\mathscr{P}_{\mathrm{re}} := \bar{A}(\hat{\beta}_{\setminus \mathscr{M}}, \emptyset, T(\mathscr{D}_{\setminus \mathscr{M}}), b)$ vs. **Unlearned:** $\mathscr{P}_{\mathrm{un}} := \bar{A}(\hat{\beta}, \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}), b)$

Definition (f-certifiability [Pandey *et al.*, 2025])

Given $\phi > 0, m \in [n]$ and $\mathscr{P}_{\mathrm{re}}, \mathscr{P}_{\mathrm{un}}$ as defined above (13), and a trade-off curve $f : [0,1] \to [0,1]$, we say unlearning algorithm $\bar{A}$ satisfies ϕ-probabilistically certified data removal property with respect to f if the following holds (with high probability).

$$\mathbb{P}\left[\inf_{|\mathscr{M}| \le m} \min(T(\mathscr{P}_{\mathrm{re}}, \mathscr{P}_{\mathrm{un}})(\alpha), T(\mathscr{P}_{\mathrm{un}}, \mathscr{P}_{\mathrm{re}})(\alpha)) \ge f(\alpha) \quad \text{for all } \alpha \in [0,1]\right] \ge 1 - \phi$$

where the probability $\mathbb{P}$ is solely over the randomness of the data $\mathscr{D}$.

Formulation of **accuracy** using fresh samples

Generalization error divergence

- Recall that we need the unlearned output $\bar{A}(\hat{\beta}, \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}), b)$ to have the same generalization capabilities as that of $\hat{\beta}_{\setminus \mathscr{M}}$ on a fresh sample from the population P_{β^*}.
- Without such a criterion, an unlearning procedure $\bar{A}$ could output pure noise-achieving perfect user privacy at the cost of severely degraded model performance.

Definition (Generalization Error Divergence [Zou *et al.*, 2025])

Given IID dataset $\mathscr{D}_n = \{(x_i, y_i)\}_{i=1}^n$ from GLM 6 for training A, and a fresh IID sample (x_0, y_0) let $\ell(y|x^T\beta)$ be a measure of error between y and $x^T\beta$. Then we define the **GED** of the learning-unlearning pair $(A, \bar{A})$ on $\mathscr{D}_n$ with data removal requests $\mathscr{D}_{\mathscr{M}}$ as:

$$\mathrm{GED}_\ell(A, \bar{A}; \mathscr{M}, \mathscr{D}_n) := \mathbb{E}\left(\left[\left|\ell(y_0|x_0^T A(\mathscr{D}_{\setminus \mathscr{M}})) - \ell(y_0|x_0^T \bar{A}(A(\mathscr{D}_n), \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}_n), b))\right|\right] \middle| \mathscr{D}_n\right),$$

where we condition on the randomness of the data set $\mathscr{D}_n$ and average over the randomness of the unlearning algorithm $\bar{A}$, as well as that of the test data point (x_0, y_0).

A Newton-Raphson based unlearning procedure

Our unlearning procedure $\bar{A}(\hat{\beta}, \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}_n), b)$

- **Approximation:** Starting from $\hat{\beta}$, we run one step of the Newton method to obtain:

$$\hat{\beta}^{(1)}_{\setminus \mathscr{M}} = \hat{\beta} - G(L_{\setminus \mathscr{M}})^{-1}(\hat{\beta}) \nabla L_{\setminus \mathscr{M}}(\hat{\beta}), \tag{1}$$

  where $G(L_{\setminus \mathscr{M}})$ is the Hessian of $L_{\setminus \mathscr{M}}$ defined in (7).

- **Randomization:** Note that since $\hat{\beta}^{(1)}_{\setminus \mathscr{M}}$ differs from $\hat{\beta}_{\setminus \mathscr{M}}$, the difference between the two vectors may reveal information about the data to be removed, $\mathscr{D}_{\mathscr{M}}$. Hence, a standard practice is to hide the data by adding random noise,

$$\tilde{\beta}_{\setminus \mathscr{M}} = \bar{A}(\hat{\beta}, \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}_n), b) := \hat{\beta}^{(1)}_{\setminus \mathscr{M}} + b. \tag{2}$$

  We choose $b \sim N(0, \sigma^2 I_p)$. The choice of σ is to balance privacy with accuracy.

The things that we want to achieve

The main aim of an unlearning procedure $\bar{A}(\hat{\beta}, \mathscr{D}_{\mathscr{M}}, T(\mathscr{D}_n), b)$ is to achieve

- **Privacy:** Under the high-dimensional setting $n, p \to \infty$ and $n/p \to \gamma$, how should we set the value of σ to ensure that $\tilde{\beta}_{\setminus \mathscr{M}}$ satisfies $f-$ certifiability with $f = G_\varepsilon$?
- **Accuracy:** Can we make the generalization error divergence of $\tilde{\beta}_{\setminus \mathscr{M}}$ go to zero as $n, p \to \infty$ while $n/p \to \gamma$, given the choice of σ as above.

Our results I– Gaussian certifiability in high dimensions

Theorem (ε-Gaussian certifiability [Pandey *et al.*, 2025])

Under some mild assumptions[a] on l, and r as well as Gaussianity assumptions on the data x there exist $C_1(n), C_2(n) = O(polylog(n))$ for which the randomized one-step Newton unlearning (2) procedure when used with a perturbation vector $b \sim N\left(0, \frac{r^2}{\varepsilon^2} I_p\right)$, achieves ϕ_n,-Gaussian certifiability with

$$r = C_1(n)\sqrt{\frac{C_2(n)m^3}{2\lambda \nu n}}, \quad \phi_n = nq_n^{(y)} + 8n^{-3} + ne^{-p/2} + 2e^{-p} \to 0.$$

[a] These are separability, convexity, smoothness, polynomial boundedness assumptions on l, and r.

Our results II- Vanishing generalization gap after one step

Theorem (Vanishing change in model accuracy [Pandey *et al.*, 2025])

Consider the unlearning estimator defined in (2) *with the noise variance set according to above Theorem 5, along with assuming the same setting. Then, with probability at least* $1-(n+1)q_n^{(y)}-14n^{-3}-ne^{-p/2}-2e^{-p}-e^{-(1-\log(2))p}$,

$$\mathrm{GED}(\tilde{\beta}_{\setminus \mathcal{M}}, \hat{\beta}_{\setminus \mathcal{M}}) \leq C_1(n)\sqrt{C_2(n)}\left(\frac{1}{\varepsilon}+\frac{1}{\sqrt{p}}\right)\sqrt{\frac{m^3(m+2)}{\lambda \nu n}}\cdot polylog(n).$$

Overall message

- If we set the variance of the Gaussian noise as suggested by the certifiability Theorem 5, the unlearning algorithm that is based on one-Step of the Newton method offers:

$$\mathrm{GED}(\tilde{\beta}_{\setminus \mathcal{M}}, \hat{\beta}_{\setminus \mathcal{M}}) = o_p(1) \text{ if } m = o(n^{\frac{1}{4}-\alpha}) \text{ for arbitrary } \alpha > 0$$

- Both theorems are valid in high-dimensional settings where $n, p \to \infty$, while $n/p \to \gamma$.
- **Why care?** [Zou *et al.*, 2025] introduced the high-dimensional setting into the machine unlearning literature. It showed that under a ‘different notion’ of certifiability even for removing a single data point, at least two Newton steps are required to ensure $\mathrm{GED}(\tilde{\beta}_{\mathcal{M}}, \hat{\beta}_{\setminus \mathcal{M}}) = o_p(1)$.
- The sharp contrast between their conclusion and ours highlights the subtle interplay between perturbation methods, certifiability definitions, and prediction accuracy.

Trade off functions and (ε, δ) Differential privacy

Definition ((ε, δ) differential privacy)

A randomized algorithm M that takes as input a dataset consisting of individuals is (ε, δ) differentially private (DP) if for any pair of datasets S, S' that differ in the record of a single individual, and any event E, (when $\delta = 0$, the guarantee is simply called ε-DP.)

$$\mathbb{P}[M(S) \in E] \leqslant \mathrm{e}^{\varepsilon} \mathbb{P}\left[M\left(S'\right) \in E\right] + \delta.$$

Definition (Trade off function)

For two probability distributions P and Q on a space $(\mathscr{X}, \mathscr{F})$, define the trade-off function $T(P,Q) : [0,1] \to [0,1]$ with the infimum taken over all (measurable) rejection rules.

$$T(P,Q)(\alpha) = \inf\left\{\beta_\varphi := \mathbb{E}_Q[1-\varphi] : \alpha_\varphi := \mathbb{E}_P[\varphi] \leq \alpha, \varphi : (\mathscr{X}, \mathscr{F}) \to [0,1] \text{ Borel}\right\}$$

Trade off functions and Neyman-Pearson lemma

Proposition

A function $f : [0,1] \to [0,1]$ is a trade-off function (for some distributions P, Q on $\mathscr{X}$) if and only if f is convex, continuous, non-increasing, and $f(x) \leqslant 1 - x$ for $x \in [0,1]$.

Theorem (Neyman-Pearson lemma)

Let P and Q be probability distributions on Ω with densities p and q, respectively. A test $\varphi : \Omega \to [0,1]$ achieves $T(P,Q)(\alpha)$ if and only if there are two constants $h \in [0, +\infty]$ and $c \in [0,1]$ such that φ has the form (with $\mathbb{E}_P[\varphi] = \alpha$)

$$\varphi(\omega) = \begin{cases} 1, & \text{if } p(\omega) > hq(\omega) \\ c, & \text{if } p(\omega) = hq(\omega) \\ 0, & \text{if } p(\omega) < hq(\omega) \end{cases}.$$

Trade off functions and f-differential privacy

Definition (f-differential privacy)

Let f be a trade-off function. A mechanism M is said to be f-differentially private if

$$T\left(M(S), M\left(S'\right)\right) \geqslant f \text{ for all neighboring datasets S and } S'.$$

Proposition

A mechanism M is (ε, δ) DP if and only if M is $f_{\varepsilon,\delta} - DP$.

$$f_{\varepsilon,\delta}(\alpha) = \max\left\{0, 1-\delta-\mathrm{e}^{\varepsilon}\alpha, \mathrm{e}^{-\varepsilon}(1-\delta-\alpha)\right\}$$

Properties of Trade off functions

Proposition

Let a mechanism M be f-DP. Then, M is f^{S}-DP with $f^{\mathrm{S}} = \max\left\{f, f^{-1}\right\}$, where

$$f^{-1}(\alpha) := \inf\{t \in [0,1] : f(t) \leqslant \alpha\} \text{ for } \alpha \in [0,1] \text{ with } f^{\mathrm{S}} = \left(f^{\mathrm{S}}\right)^{-1}$$

$$\operatorname{epi}(f) := \{(\alpha, \beta) \mid \alpha \in [0,1], f(\alpha) \leqslant \beta \leqslant 1 - \alpha\}$$
$$\operatorname{epi}(f) = \left\{\left(\alpha_\varphi, \beta_\varphi\right) \mid \varphi : \Omega \to [0,1] \text{ measurable}, \alpha_\varphi + \beta_\varphi \leqslant 1\right\}.$$

Lemma

f and $\operatorname{epi}(f)$ are equivalent and if $f = T(P,Q)$, then $f^{-1} = T(Q,P)$.

Examples of Trade off functions and Blackwell ordering

Proposition

Let $\xi \sim P$ on $\mathbb{R}$ with density p, CDF $F : \mathbb{R} \to [0,1]$, quantile $F^{-1} : [0,1] \to [-\infty, +\infty]$. Then $T(\xi, t+\xi)(\alpha) = F\left(F^{-1}(1-\alpha) - t\right) \forall t > 0$ if and only if p is log-concave.

Proposition

For any two distributions P and Q, we have $T(R(P), R(Q)) \geqslant T(P,Q)$. As a consequence if a mechanism M is f-DP, then its post-processing $R \circ M$ is also f-DP.

Theorem (Blackwell's informativeness theorem)

Let P, Q be distributions on Y and P', Q' be distributions on Z. TFAE

- $T(P,Q) \leqslant T(P', Q')$.
- $\exists$ *a Kernel* $R : Y \to Z$ *such that* $(R(P), R(Q)) = (P', Q')$.

Primal Dual perspective of Trade off functions

Proposition

Let I be an arbitrary index set associated with $\varepsilon_i \in [0,\infty)$ and $\delta_i \in [0,1]$ for $i \in I$. A mechanism is $(\varepsilon_i, \delta_i)$-DP for all $i \in I$ if and only if it is f-DP with $f = \sup_{i\in I} f_{\varepsilon_i,\delta_i}$

$$g^*(y) = \sup_{-\infty<x<\infty} yx - g(x) \text{ for } g : \mathbb{R} \to \mathbb{R}$$

$f : [0,1] \to [0,1]$ we set $f(x) = \infty$ for $x \in (-\infty, 0) \cup (1,\infty)$ (supremum is over $0 \leqslant x \leqslant 1$).

Proposition (Envelope theorem)

For a symmetric trade-off function f, a mechanism is f-DP if and only if it is $(\varepsilon, \delta(\varepsilon)) - DP$ for all $\varepsilon \geqslant 0$ with $\delta(\varepsilon) = 1 + f^*(-e^{\varepsilon})$. As a consequence a mechanism is μ GDP if and only if it is $(\varepsilon, \delta(\varepsilon)) - DP$ for all $\varepsilon \geqslant 0$, where

$$\delta(\varepsilon) = \Phi\left(-\frac{\varepsilon}{\mu} + \frac{\mu}{2}\right) - e^{\varepsilon}\Phi\left(-\frac{\varepsilon}{\mu} - \frac{\mu}{2}\right).$$

Group privacy

Theorem (Group-privacy lift for f-DP)

For distributions P, Q, R on $\mathcal{X}$, if $\bar{T}(P,Q) \le \bar{f}$, and $\bar{T}(Q,R) \le \bar{g}$ then $\bar{T}(P,R) \le \bar{g} \circ \bar{f}$. As a consequence if a mechanism M satisfies $\bar{f}$-DP, then it satisfies $\left(\bar{f}^{\circ k}\right)$-DP with respect to groups of size $k \in \mathbb{N}$. In particular, μ-GDP implies $k\mu$-GDP for groups of size k.

Proposition (Laplace limit of group-privacy for ε-DP)

Fix $\mu > 0$ and set $\varepsilon = \mu/k$. As $k \to \infty$, $\bar{f}_{\varepsilon,0}^{\circ k} \to \bar{T}\big(L(0,1), L(\mu,1)\big)$ uniformly on $[0,1]$.

$$T_{L(0,1),L(\mu,1)}(\alpha) = \begin{cases} 1 - e^{\mu}\alpha, & 0 \le \alpha < e^{-\mu}/2, \\ e^{-\mu}/4\alpha, & e^{-\mu}/2 \le \alpha \le \frac{1}{2}, \\ e^{-\mu}(1-\alpha), & \frac{1}{2} < \alpha \le 1. \end{cases}$$

Composition and limit theorem

Represenation of functionals satisfying data-processing inequalities

Proposition

If $D(R(P),R(Q)) \leq D(P,Q)$ for probability distributions P,Q and Markov kernels R, then there exists a functional $l_D : \mathscr{F} \to \mathbb{R}$ such that $D(P,Q) = l_D(T(P,Q))$.

References I

[Pandey *et al.*, 2025] A. Pandey, A. Auddy, H. Zou, A. Maleki, and S. Kulkarni.
Gaussian Certified Unlearning in High Dimensions: A hypothesis testing approach.

[Dong *et al.*, 2022] J. Dong, A. Roth, and W. J. Su.
Gaussian Differential Privacy, *J. R. Stat. Soc. Series B* **84**(1), 337, 2022.

[Zou *et al.*, 2025] H. Zou, A. Auddy, Y. Kwon, K. Rahnama Rad, and A. Maleki.
Certified Data Removal Under High-dimensional Settings.
arXiv preprint `arXiv:2505.07640`, 2025.