A SHARP TRANSITION IN DATA RECONSTRUCTION UNDER DIFFERENTIAL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Data reconstruction attacks have empirically been successful in recovering training samples from learned models, raising privacy concerns and motivating defenses with guarantees that remain valid against future threats. While differential privacy (DP) provides formal protection, choosing the privacy budget remains a challenge: small budgets severely reduce utility, but it is hard to quantify how large the budget can be without allowing accurate reconstruction. In this work, we study informed attackers who aim to reconstruct a single d-dimensional training sample from a ρ-zero-concentrated DP model, knowing all other training data. Our main contribution is to establish a sharp transition at $\rho \asymp d$ for data reconstruction: on the one hand, we derive entropy-based lower bounds for any private mechanism and any attack, characterizing a set of target priors for which reconstruction is information-theoretically impossible for $\rho \ll d$; on the other hand, we analyze a simple attack on private linear regression with output perturbation, showing that reconstruction is practically feasible for $\rho \gg d$. Remarkably, the transition moves to $\rho \asymp s$ for data lying in an s-dimensional subspace, demonstrating that the privacy budget guaranteeing adequate protection must be assessed in terms of the effective dimension of the data. We validate our findings via experiments on synthetic data and natural images (CIFAR-10, ImageNet).

1 INTRODUCTION

Easy access to vast amounts of training data is a key ingredient in the success of modern deep learning, naturally encouraging the collection of data that may contain sensitive personal information or copyrighted material. Standard training methods often memorize such training data, making it possible for attackers to recover information about their content from the final model (Carlini et al., 2021; Nasr et al., 2025; Cooper et al., 2026). For example, membership inference attacks are designed to determine whether a given target sample x_t was included in the training set (Shokri et al., 2017; Nasr et al., 2019; Carlini et al., 2022), while *data reconstruction* aims to recover the full x_t from scratch (Haim et al., 2022; Buzaglo et al., 2023).

A way to address these risks is via *differential privacy* (DP) (Dwork et al., 2006), which has become the standard paradigm for provably safeguarding models from leaking private information e.g. through unintended memorization. DP quantifies data protection through a numerical parameter (ρ in Definition 1), which bounds the impact a single data point can have on the output of the algorithm. However, the noise required by DP training negatively affects performance (Abadi et al., 2016), thus limiting its adoption. In particular, this performance cost is generally noticeable for constant-order (i.e., not dependent on sample size or data dimensionality) privacy budgets $\rho \asymp 1$, a regime that is also the usual focus in the literature (De et al., 2022; McKenna et al., 2025)[1]. This scaling provides non-trivial trade-offs for membership inference attacks (Dong et al., 2022; Mahloujifar et al., 2022; Kairouz et al., 2015): informally, attacks are possible when $\rho \gg 1$, and impossible when $\rho \ll 1$.

On the other hand, much larger privacy budgets yield negligible utility losses in some domains and still protect models from reconstruction attacks (Bhowmick et al., 2019; Stock et al., 2022; Ziller et al., 2024), motivating interest in the regime $\rho \gg 1$ (Cyffers, 2025). In particular, Balle et al.

[1]The cited work often focuses on (ε, δ)-DP with $\varepsilon \asymp 1$. Given the conversion from ρ-zCDP to (ε, δ)-DP (Bun & Steinke, 2016), we interpret this regime as $\rho \asymp 1$, neglecting the logarithmic dependence on δ. As an example, a zCDP guarantee with $\rho \simeq 1.05$ implies the $(8, 10^{-5})$-DP guarantee reported by De et al. (2022).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2022) have shown that, for suitable priors on $x_t \in \mathbb{R}^d$, $\rho \ll d$ suffices to make reconstruction statistically impossible. However, for general priors, their bound relies on small-ball probabilities which may be hard to compute in practice (e.g., for natural language data). Most importantly, it is unclear whether the result is tight, in the sense that reconstruction attacks are possible for $\rho \gg d$.

Our work takes a step towards understanding when the privacy budget sharply characterizes the feasibility of reconstruction from the final model parameters, identifying a sharp transition at $\rho \asymp d$ in *DP linear regression*. We focus on informed attackers with access to the training set (except for the target example), as well as the hyper-parameters of the *output perturbation* algorithm used to train the model. More specifically, our contributions are the following:

1. We derive a lower bound on the expected reconstruction error in terms of the target's differential entropy $h(x_t)$, valid for any ρ-zCDP mechanism and any attack (Theorem 1). This result implies that accurate reconstruction is impossible as long as $\rho \ll h(x_t)$. For the target distributions considered by Balle et al. (2022), or for distributions with independent entries, we have $h(x_t) \asymp d$, demonstrating the impossibility of reconstruction for $\rho \ll d$.
2. We characterize the performance of an explicit reconstruction attack on a DP model trained via output perturbation (Theorem 2). For a natural range of regularization parameters λ and residual clipping thresholds C_{clip}, the attack succeeds with high probability whenever $\rho \gg d$, providing a tight converse to Theorem 1.
3. We extend our results to data distributions with lower *effective dimension*, in particular to data supported on an s-dimensional subspace of $\mathbb{R}^d$. We obtain an analogous entropy-based lower bound (Theorem 3) and derive a reconstruction attack that exploits knowledge of the subspace (Theorem 4). Together, these results show that the relevant transition for data reconstruction in this setting occurs at $\rho \asymp s$.

We validate our findings through numerical experiments on synthetic data (Figs. 2 and 5a) and natural images from CIFAR-10 and ImageNet (Figs. 3, 4 and 5b), demonstrating the wider generality of our results beyond the technical assumptions. Overall, our work identifies data dimensionality as a relevant scale for assessing privacy budgets, providing a concrete basis for studying the risks and potential utility benefits of larger budgets.

2 Related work

Data reconstruction from non-private models. Large models are known to memorize individual training data within their parameters. Such data can be extracted through appropriate prompting in generative models (Carlini et al., 2023; Nasr et al., 2025; Cooper et al., 2026), or *reconstructed* directly from the parameters of classifiers (Haim et al., 2022; Buzaglo et al., 2023; Oz et al., 2024; Shen et al., 2026). In the theoretical literature, the problem of fully reconstructing the training set has been tackled in specific settings: the heuristic method of Haim et al. (2022) relies on the implicit bias of gradient descent in homogeneous networks (Lyu & Li, 2020; Ji & Telgarsky, 2020); Loo et al. (2024) and Iurada et al. (2026) show that reconstruction is possible in, respectively, infinite-width shallow networks and sufficiently over-parameterized models. If the attacker knows the rest of the dataset, an effective reconstruction method against a generalized linear model, trained without the clipping or noise required for privacy, has also been expressed in closed form by Balle et al. (2022).

Attacks on private models. For models trained with DP, the feasibility of membership inference attacks (MIAs) is closely linked to the privacy budget (Yeom et al., 2018; Kairouz et al., 2015; Mahloujifar et al., 2022), and there is a connection to privacy auditing, as these results mean that successful membership attacks set lower bounds on the privacy guarantee of an algorithm (Jagielski et al., 2020; Nasr et al., 2023; Steinke et al., 2023). Reconstruction attacks have also been linked to privacy budgets, primarily through reconstruction robustness (Balle et al., 2022). This framework has been extended to give stronger bounds for specific mechanisms (Hayes et al., 2023; Kaissis et al., 2023), to use f-DP (Kulynych et al., 2025), and to reconstruct multiple data points (Swanberg et al., 2026). Additional related work has focused on recovering one batch from DP-SGD (Liu et al., 2025), specializing reconstruction robustness to discrete alphabets (Guo et al., 2023), or lower bounding the reconstruction error through the bias and variance of the attack (Guo et al., 2022), though requiring the privacy budget to be at most logarithmic in the diameter of the target prior to rule out reconstruction. Our work combines impossibility guarantees based on the entropy of the target with a provably successful attack, thus establishing a sharp transition for data reconstruction.

Large privacy budgets. A classical convention in the DP literature is to take the privacy budget to be a moderately small constant (Dwork & Roth, 2014). This is consistent with theoretical analyses showing that MIAs fail when $\rho \ll 1$ (Bun & Steinke, 2016; Kairouz et al., 2015; Dong et al., 2022). When membership is not a sensitive attribute, much larger (possibly dimension-dependent) budgets provide meaningful protection while better preserving utility. This has been shown in the locally private setting (Bhowmick et al., 2019; Sarwate & Sankar, 2014; Kalantari et al., 2018), in language models (Stock et al., 2022), and in medical imaging (Ziller et al., 2024), with Aerni et al. (2024) even finding that budgets well above 10^3 can outperform other heuristic defenses against membership attacks. This evidence also partly motivates the recent position taken by Cyffers (2025), who argues that larger budgets should not be dismissed in favor of uncertified protection methods. Our results provide a theoretical foundation for this perspective, identifying the data's effective dimension as the scale for assessing how large privacy budgets can be while still protecting against accurate reconstruction.

3 PRELIMINARIES

Notation. We set $[n] := \{1, \dots, n\}$. Given a vector v, let $\|\cdot\|_2$ denote its Euclidean norm. Given a matrix $A \in \mathbb{R}^{n\times m}$, let $\|A\|_{\mathrm{op}}$ be its operator (spectral) norm. Let $h(a) = -\int p_a(x) \log p_a(x) \mathrm{d}x$ denote the differential entropy of a random variable a. We use the letters $x, \mathcal{X}$ to denote feature vectors and the space of features; $y, \mathcal{Y}$ for labels; and $z, \mathcal{Z}$ for feature-label pairs. Complexity notations O ($\lesssim$), Ω ($\gtrsim$) and Θ ($\asymp$) are meant for large input dimension d and sample size n.

Linear regression. Let $Z = (X, Y)$ be a labeled training dataset, where $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n\times d}$ contains the training data (features) on its rows and $Y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$ contains the corresponding labels. We assume features to have bounded norm $\|x\|_2 \le R$ and input-label pairs to be sampled i.i.d. from a joint distribution P_{XY}. We consider the *linear regression* model

$$y_i = x_i^\top \theta_{\mathrm{true}} + \nu_i, \tag{1}$$

where $\theta_{\mathrm{true}} \in \mathbb{R}^d$ and ν_i is independent label noise. The goal of DP linear regression is to output an estimate that guarantees a privacy budget and minimizes the *test risk*, defined for $\theta \in \mathbb{R}^d$ as

$$\mathcal{P}(\theta) = \frac{1}{2}\mathbb{E}_{(x,y)\sim P_{XY}}\left[\left(x^\top\theta - y\right)^2\right] \tag{2}$$

Differential privacy (DP) and output perturbation. We say that the datasets Z and Z' are *adjacent* if they differ in a single input-label pair. We quantify DP via ρ-zero-concentrated DP (ρ-zCDP) which uses the α-Rényi divergence between two probability distributions P, Q:

$$D_\alpha(P\|Q) := \frac{1}{\alpha - 1}\log \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^\alpha \mathrm{d}Q, \qquad \alpha > 1. \tag{3}$$

Definition 1 (ρ-zCDP (Bun & Steinke, 2016))**.** *A randomized algorithm $\mathcal{M}$ satisfies ρ-zCDP if, for every adjacent pair of datasets Z, Z', and every $\alpha > 1$,*

$$D_\alpha\big(\mathcal{M}(Z)\|\mathcal{M}(Z')\big) \le \rho\alpha. \tag{4}$$

Fix a regularization parameter $\lambda > 0$, and a residual clipping threshold $C_{\mathrm{clip}} > 0$. Let $\ell_{C_{\mathrm{clip}}}$ be the Huber loss, i.e., a square loss linearized to be C_{clip}-Lipschitz:

$$\ell_{C_{\mathrm{clip}}}(a) := \begin{cases} a^2/2, & |a| \le C_{\mathrm{clip}}, \\ C_{\mathrm{clip}}|a| - C_{\mathrm{clip}}^2/2, & |a| > C_{\mathrm{clip}}. \end{cases} \tag{5}$$

Writing $Z = (X, Y)$, output perturbation first computes the regularized empirical-risk minimizer

$$\theta^*(Z) := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d}\left\{\frac{1}{n}\sum_{i=1}^n \ell_{C_{\mathrm{clip}}}\big(x_i^\top\theta - y_i\big) + \frac{\lambda}{2}\|\theta\|_2^2\right\}, \tag{6}$$

and then releases

$$\tilde\theta = \theta^*(Z) + b_{\mathrm{DP}}, \qquad b_{\mathrm{DP}} \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d). \tag{7}$$

Proposition 1. *The output perturbation algorithm defined by* (6)-(7) *is ρ-zCDP provided*

$$\sigma_{\mathrm{DP}}^2 = \frac{2R^2C_{\mathrm{clip}}^2}{\lambda^2\rho n^2}. \tag{8}$$

This result follows standard arguments on the Gaussian mechanism applied to minimization algorithms with bounded sensitivity (Mironov, 2017), and its proof is deferred to Appendix B. Intuitively, larger λ (or smaller C_{clip}) implies that the effect of a single training sample on the resulting $\theta^*(Z)$ is smaller. Then, to guarantee privacy, Gaussian noise is added proportionally to $1/\rho$ and $R^2 = \sup_x \|x\|_2^2$. We will refer to C_{clip} and λ as the hyper-parameters of output perturbation.

Reconstruction threat model. We follow the same threat model as Balle et al. (2022). A trusted processor has a dataset $Z = (X, Y)$, decomposed into a target point $z_t = (x_t, y_t)$ and the rest of the dataset, $Z_- = (X_-, y_-)$. For convenience, we associate $t \in [n]$ with an index in the data. The processor releases the output $\mathcal{M}(Z) \in \Theta$ of a ρ-zCDP mechanism on the dataset. The attacker then develops a function $\mathcal{A} : \mathcal{Z}^{n-1} \times \Theta \to \mathcal{X}$ to reconstruct the target feature $x_t \in \mathcal{X}$ from $\mathcal{M}(Z)$ and all other datapoints Z_-. We note that the attacker also has knowledge of the mechanism's hyperparameters λ and C_{clip}. The resulting reconstruction is $\widehat{x} := \mathcal{A}(Z_-, \mathcal{M}(Z))$.

4 Main results

4.1 Information-theoretic lower bound

We first show a lower bound on the expected error an attacker must incur when reconstructing the target point, regardless of (X_-, y_-), the ρ-zCDP mechanism, or the reconstruction method.

Theorem 1. *Let the target feature x_t be independent of the rest of the dataset $Z_- = (X_-, y_-)$. Let $\mathcal{A} : \mathcal{Z}^{n-1} \times \Theta \to \mathcal{X}$ be an attack, taking as input Z_- and the output $\mathcal{M}(Z_- \cup \{z_t\})$ of a ρ-zCDP mechanism, and producing as output an estimate $\widehat{x}$ of x_t. Then,*

$$\frac{\mathbb{E}\|x_t - \widehat{x}\|_2^2}{\mathbb{E}\|x_t\|_2^2} \geq \frac{1}{2\pi e} \exp\left(\frac{2}{d}\left(h(x_t') - \rho\right)\right), \tag{9}$$

with $x_t' := \frac{\sqrt{d}}{\sqrt{\mathbb{E}\|x_t\|_2^2}} x_t$.

We note that the bound in (9) is scale-invariant, in the sense that neither the LHS or RHS depend on the norm of x_t. Since the LHS of (9) is 1 for the all-0 estimator and 0 for $\hat{x} = x_t$, we conclude that any attack fails whenever ρ is small compared to the differential entropy of the (normalized) point to be reconstructed. This dependence on $h(x_t')$ is natural: the differential entropy quantifies the prior uncertainty in the target, with normalization removing its dependency on the scale of the distribution, and ρ bounds the information that the mechanism can reveal.

When $h(x_t') = \Theta(d)$, Theorem 1 implies that, if $\rho \ll d$, reconstructing x_t accurately is information-theoretically impossible. This is the case e.g. for isotropic Gaussians, for the uniform distribution on a ball, and for distributions with independent entries, each with marginal entropy bounded away from 0. Moreover, while (9) bounds normalized mean squared error in terms of $h(x_t')$, similar bounds can be obtained for *(i)* the per-coordinate error $\mathbb{E}\|x_t - \widehat{x}\|_2^2/d$ in terms of $h(x_t)$, and for *(ii)* different norms/orders (e.g., $\mathbb{E}\|x_t - \widehat{x}\|_2$) whenever the max-entropy distribution under bounded distortion is computable. The proof of Theorem 1 is in Appendix C, with a sketch given below.

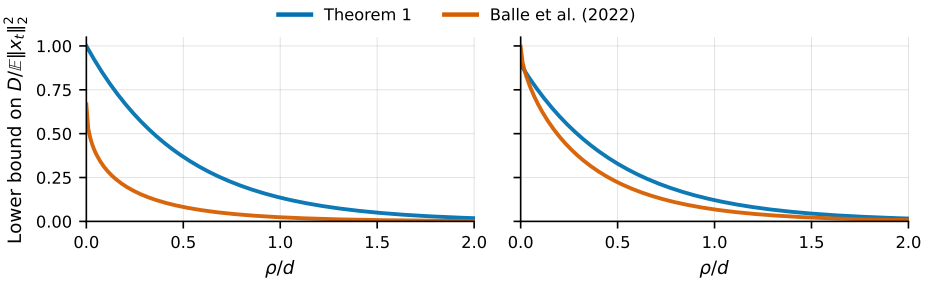
Proof sketch. Let us consider normalized features $\mathbb{E}[\|x_t\|_2^2] = d$ for simplicity, so that $x_t' = x_t$. The idea is to combine an upper bound on the mutual information coming from the privacy guarantee with a rate-distortion inequality. First, in Lemma 2, we show that $I(z_t; \mathcal{M}(Z_- \cup \{z_t\})) \leq \rho$. This follows by taking the Rényi order $\alpha \downarrow 1$ in the zCDP condition and using convexity of the KL divergence to bound mutual information. As shown by Bun & Steinke (2016), an application of the data processing inequality, combined with the independence of x_t and Z_-, then implies

$$I(x_t; \widehat{x}) \leq \rho. \tag{10}$$

Writing $D = \mathbb{E}\|x_t - \widehat{x}\|_2^2$ and using that the Gaussian distribution maximizes the differential entropy when the second moment is constrained, in Lemma 3 we show that

$$h(x_t) - \frac{d}{2}\log\left(\frac{2\pi e D}{d}\right) \leq I(x_t; \widehat{x}). \tag{11}$$

Combining (10)-(11) yields the claimed lower bound. □



(a) Gaussian (b) Uniform in ball radius $\sqrt{d}$

Figure 1: Lower bounds on the normalized mean squared error $D/\mathbb{E}\|x_t\|_2^2$, with $D = \mathbb{E}\|x_t - \widehat{x}\|_2^2$. We take $d = 20$ and let the distribution of x_t be standard Gaussian (left) or uniform on the ball of radius $\sqrt{d}$ (right), following the setting of Propositions 6-7 of Balle et al. (2022). The guarantees of Balle et al. (2022) are translated into bounds on the mean squared error using the Gaussian Chernoff and uniform-ball volume bounds, see Appendix F for details.

Comparison with prior bounds. Balle et al. (2022, Corollary 4) bound the probability of reconstruction within a radius r through the small-ball probability $\sup_v \mathbb{P}(\|x_t - v\|_2 \le r)$. In contrast, our bound is on the mean squared error and the distribution of x_t enters through its differential entropy. We note that the examples of a uniform and Gaussian distribution in Propositions 6-7 of Balle et al. (2022) also show that $\rho \ll d$ guarantees the impossibility of reconstructing training data. However, Fig. 1 demonstrates that the bound of Theorem 1 is tighter for the Gaussian case and also for the uniform case unless ρ/d approaches 0. The improvement is particularly noticeable for moderate values of ρ/d, i.e., when a strong privacy guarantee is required.

4.2 Algorithmic upper bound

We next design an attack on the linear model in (1) privatized via output perturbation (see (6)), and give provable guarantees on its reconstruction error.

Assumption 1. *Each feature datapoint $x \in \mathbb{R}^d$ in X is independently and identically drawn from a $O(1)$-subGaussian distribution with bounded norm $\|x\|_2 \le R$ for $R = \Theta(\sqrt{d})$. Furthermore, we assume $\|x_t\|_2 = \Theta(\sqrt{d})$ and $\nu_i \sim \mathcal{N}(0, \zeta^2)$.*

In words, we require the data to have well-behaved tails, as commonly done in related work (Brown et al., 2024; Iurada et al., 2026). We also note that this assumption can be relaxed to requiring $\|X_-{}^\top X_-\|_{\mathrm{op}} = O(n + d)$. The normalization $R = \Theta(\sqrt{d})$ is chosen for convenience, and it corresponds to taking the entries of x of constant order. The Gaussian assumption on ν_i is also chosen for simplicity, as it can be relaxed to requiring sufficient anti-concentration of the noise.

Note that the derivative of the Huber loss is $\psi_{C_{\mathrm{clip}}}(a) := \ell'_{C_{\mathrm{clip}}}(a) = \mathrm{sign}(a)\min\{|a|, C_{\mathrm{clip}}\}$. Given the non-target data, define the leave-one-out gradient

$$F_-(\theta) := \sum_{i\in[n], i\neq t} x_i \psi_{C_{\mathrm{clip}}}(x_i^\top \theta - y_i) + n\lambda\theta. \tag{12}$$

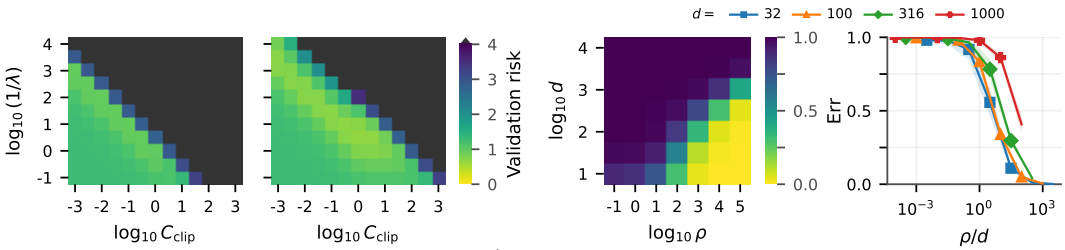
For $\lambda > 0$, the regularized objective in the RHS of (6) is strongly convex, and it has a unique minimizer θ^*. The first-order optimality condition thus gives

$$-F_-(\theta^*) = x_t\, \psi_{C_{\mathrm{clip}}}(x_t^\top \theta^* - y_t). \tag{13}$$

Then, given access to θ^*, through the knowledge of X_-, y_-, C_{clip} and λ, one can compute $F_-(\theta^*)$. This quantity, assuming $\psi_{C_{\mathrm{clip}}}(x_t^\top \theta^* - y_t) \neq 0$, is an exact estimator of the direction of x_t, due to (13). As the attacker has access to the perturbed weights $\tilde{\theta}$ (and not to θ^*), it is natural to pick $F_-(\tilde{\theta})$ to estimate the direction of x_t. The performance of this reconstruction attack is characterized below.

Theorem 2. *Let Assumption 1 hold, and let $\lambda > 0$, $C_{\mathrm{clip}} > 0$ and $\rho > 0$ be the parameters of the algorithm chosen by the learner. The attacker computes*

$$\widehat{x} := R\,\frac{F_-(\tilde{\theta})}{\|F_-(\tilde{\theta})\|_2}. \tag{14}$$



(a) Hyperparameter search at $\rho = 10$ and $\rho = 10^4$ (b) Reconstruction error relative to ρ and d

Figure 2: Hyper-parameter selection and reconstruction, for synthetic data sampled uniformly on the sphere. The first two panels show validation mean squared error (MSE) over a hyper-parameter grid at $n = d = 1000$, $\zeta = 0.5$, for $\rho = 10$ and $\rho = 10^4$. The third and fourth panel show normalized reconstruction error after hyper-parameter selection within the range $\lambda \geq 0.1$ and $C_{\text{clip}} \leq 0.1\zeta$.

Let $\eta \in (0,1)$ be a failure probability and $C, c > 0$ be absolute constants. Then, with probability at least $1 - \eta - 2e^{-cn} - 2e^{-cd}$,

$$\frac{\min_{\tau\in[-1,1]} \|x_t - \tau\widehat{x}\|_2^2}{\|x_t\|_2^2} \leq C \left[\max\left\{1, \frac{C_{\text{clip}}}{\zeta\eta}\left(1 + \frac{d}{n\lambda}\right)\right\}\left(1 + \frac{1}{\lambda}\right)\left(1 + \frac{d}{n}\right)\right]^2 \frac{d}{\rho}. \tag{15}$$

Theorem 2 states that, for large enough privacy parameter ρ, the estimate $F_-(\tilde{\theta})$ is approximately aligned to the target x_t. We note that the remaining sign and scale ambiguities may be easily resolvable in practice: the attacker has only two signs to consider (which would reconstruct the original image or its negative), and prior knowledge of the typical sample norm is a natural choice of scale. Formally, we remark that it is also possible to extend the lower bound in Theorem 1 to the metric considered here, without losing the dependence on the dimension.

Consider now the case where η is a small constant (e.g., $\eta = 0.01$), the label noise is of constant order $\zeta = \Theta(1)$ and the number of training samples satisfies $n = \Omega(d)$, often needed in linear regression to achieve non-trivial test loss (Wainwright, 2019; Mourtada, 2022). Then, the smallest ρ allowing successful reconstruction depends on the hyper-parameters C_{clip}, λ used by output perturbation. Fig. 2a shows that small clipping constants $C_{\text{clip}} = O(1)$ and a sufficiently large regularization $\lambda = \Theta(1)$ empirically yield good performance and, under this scaling, (15) reads

$$\frac{\min_{\tau\in[-1,1]} \|x_t - \tau\widehat{x}\|_2^2}{\|x_t\|_2^2} \lesssim \frac{d}{\rho}. \tag{16}$$

This matches the scaling of the impossibility result of Theorem 1, and the sharp transition in data reconstruction at $\rho \asymp d$ is clearly displayed in Fig. 2b. We also remark that the findings in Fig. 2(a) on C_{clip} are in analogy with the evidence that small gradient clipping constants yield good downstream performance in private gradient methods (Li et al., 2022; De et al., 2022; Bombari et al., 2026). The proof of Theorem 2 is in Appendix D, with a sketch given below.

Proof sketch. The idea is to separate the signal carried by the target point from the perturbation induced by the privacy noise. Writing $a = \psi_{C_{\text{clip}}}(x_t^\top \theta^* - y_t)$, the optimality condition (13) gives

$$-F_-(\tilde{\theta}) = ax_t + w, \qquad w := F_-(\theta^*) - F_-(\tilde{\theta}).$$

First, to lower-bound $|a|$, we show that the target residual at θ^* is close to that at the leave-one-out estimator θ_- defined in (31). Conditional on (x_t, Z_-), the latter residual is Gaussian with variance ζ^2. Gaussian anti-concentration, combined with Lemma 5 , then yields

$$|a| \gtrsim \min\left\{C_{\text{clip}}, \frac{\zeta\eta}{1 + R^2/(n\lambda)}\right\}$$

with probability at least $1 - \eta$; see (36). Next, the same lemma controls the amplification of the privacy noise through F_-. Combining concentration of the Gram matrix ${X_-}^\top X_-$ and of the norm of the Gaussian noise gives, with high probability,

$$\|w\|_2 \lesssim \sigma_{\text{DP}}\sqrt{d}\,(n\lambda + n + d),$$

as established in (38). Finally, Lemma 4 gives $\min_{\tau\in[-1,1]} \|x_t - \tau\widehat{x}\|_2^2 \leq 4\|w\|_2^2/a^2$. Substituting σ_{DP} from Proposition 1, using $R^2 = O(d)$, and taking a union bound yields the final guarantee. □

4.3 EXTENSION TO LOW-DIMENSIONAL DATA

The bounds above show that $\rho \asymp d$ is a sharp threshold for reconstructing data in dimension d privatized via a ρ-zCDP mechanism: Theorem 1 implies that reconstruction is information-theoretically impossible for $\rho \ll d$, and Theorem 2 exhibits a successful attack for $\rho \gg d$. However, the impossibility result, and therefore the tightness of this threshold, relies on the entropy of x_t to scale with the number of dimensions, i.e., $h(x_t) \asymp d$, which may not be the case for data in practical applications. To model that, we now focus on data lying in a lower dimensional subspace of $\mathbb{R}^d$.

More precisely, let P be a fixed orthogonal projection of rank $s < d$, and let $U \in \mathbb{R}^{d\times s}$ have orthonormal columns spanning $\mathrm{Im}(P)$, so that $P = UU^\top$. Define

$$h_s(x_t) := h(U^\top x_t),$$

where h is taken with respect to the Lebesgue measure on $\mathbb{R}^s$. Note that this definition does not depend on the specific choice of orthonormal basis, and its purpose is to avoid computing the differential entropy of Px_t in the original ambient space $\mathbb{R}^d$, where its distribution would be singular.

The result below (proved in Appendix E) extends the lower bound of Theorem 1.

Theorem 3. *In the same setting as Theorem 1, for any rank-s projection P satisfying the conditions above, any reconstruction of x_t from a ρ-zCDP mechanism obeys:*

$$\frac{\mathbb{E}\|x_t - \widehat{x}\|_2^2}{\mathbb{E}\|Px_t\|_2^2} \geq \frac{1}{2\pi e}\exp\left(\frac{2}{s}(h_s(x'_s) - \rho)\right), \tag{17}$$

with $x'_s := \frac{\sqrt{s}}{\sqrt{\mathbb{E}\|Px_t\|_2^2}}Px_t$.

We note that the lower bound in (17) is analogous to that in (9) upon replacing *(i)* $\mathbb{E}\|x_t\|_2^2$ with $\mathbb{E}\|Px_t\|_2^2$ in the denominator of the LHS, and *(ii)* $h(x'_t)$ with $h_s(x'_s)$ on the RHS. Thus, if the point to be reconstructed x_t lies close to the span of $\mathrm{Im}(P)$ (i.e., $\mathbb{E}\|x_t\|_2^2 \approx \mathbb{E}\|Px_t\|_2^2$), then any attack fails whenever ρ is small compared to the differential entropy $h_s(x'_s)$ computed on that span. Furthermore, as this span has dimension s, we would typically have $h_s(x'_s) = \Theta(s)$, implying the impossibility of data reconstruction whenever $\rho \ll s$.

To complement this lower bound, the result below (also proved in Appendix E) characterizes the performance of a variant of the attack in Theorem 2 that seeks to reconstruct an s-dimensional projection of the target.

Theorem 4. *Consider the same setting as Theorem 2, but with $R = \Theta(\sqrt{s})$ and $\|x_t\|_2 = \Theta(\sqrt{s})$ replacing the corresponding $\Theta(\sqrt{d})$ assumptions. Let $P \in \mathbb{R}^{d\times d}$ be a rank-s projector, known to the attacker, such that $x_i \in \mathrm{Im}(P)$ for all $i \in [n]$. Let*

$$\hat{x}_P := R\frac{PF_-(\tilde{\theta})}{\|PF_-(\tilde{\theta})\|_2}. \tag{18}$$

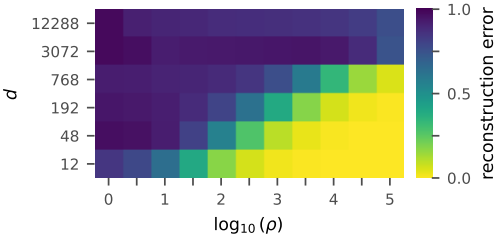
Then, if $n = \Omega(s), \lambda = \Omega(1), C_{\mathrm{clip}} = O(\zeta)$, we have

$$\frac{\min_{\tau\in[-1,1]}\|x_t - \tau\hat{x}_P\|_2^2}{\|x_t\|_2^2} \lesssim \frac{s}{\rho}, \tag{19}$$

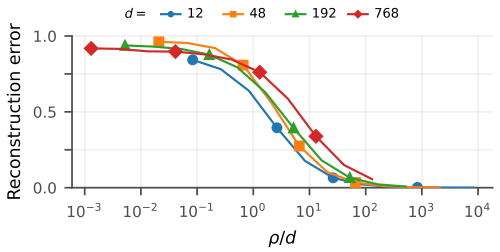
with probability at least 0.99.

While the result of Theorem 3 is scale-invariant, Theorem 4 considers R and $\|x_t\|_2$ of order $\sqrt{s}$.[2] Modulo this scale difference, the upper bound in (19) is analogous to that in (16) upon replacing *(i)* the estimator $\widehat{x}$ with its projected version $\hat{x}_P$, and *(ii)* the dimension d of the ambient space with the dimension s of the subspace containing the data. In words, Theorem 4 shows that an attack having knowledge of the subspace of dimension s in which the data lies, is successful whenever $\rho \gg s$ and, combined with Theorem 3, this establishes $\rho \asymp s$ as a sharp threshold for data reconstruction.

[2] If one wishes to keep those quantities of order $\sqrt{d}$, we note that the same result would hold upon taking λ of order d/s and assuming that, for any feature $x \in \mathbb{R}^d$, $\sqrt{s/d}x$ is $O(1)$-subGaussian.



(a) ImageNet reconstruction heatmap



(b) ImageNet, ρ/d

Figure 3: Reconstruction of ImageNet data (elephants vs. pandas). Hyperparameter search for mean validation MSE within $\lambda \geq 0.1$ and $C_{\text{clip}} \leq 0.01$. The panels show squared relative error against (ρ, d) and against ρ/d.



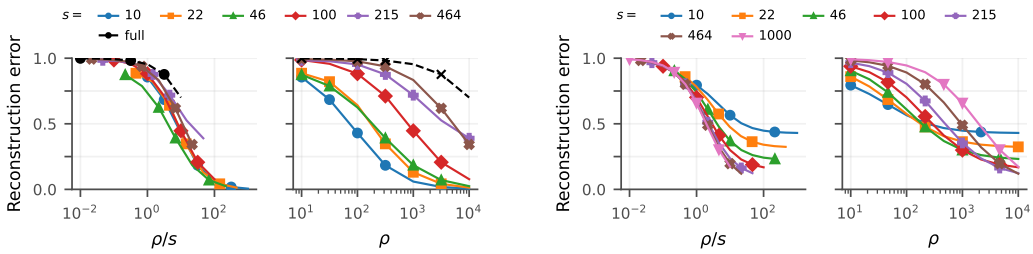
Figure 4: Reconstructions of four target images from ImageNet (elephants vs. pandas) as the privacy budget grows, with $\lambda = 1$ and $C_{\text{clip}} = 0.125$. The number inside each reconstruction is its reconstruction error. Here, $d = 12{,}288$ (64×64 RGB images). Additional plots appear in Figs. 7 and 8.

5 Experiments

We study the attacks of Sections 4.2-4.3 for data sampled uniformly on the d-dimensional sphere, synthetic data concentrated near an s-dimensional subspace, and binary regression on data taken from ImageNet (elephants vs. pandas) and CIFAR-10 (frog vs. trucks). Details on datasets construction, learning algorithm, and attacks are given in Appendix G. We report the squared reconstruction error (see e.g. the LHS of (19)), with the sign ambiguity resolved.

Synthetic d-dimensional data. In Fig. 2a, we investigate how different choices of C_{clip} and λ affect the performance of the output perturbation algorithm in (7). The heatmaps show that, for both values of ρ, the best performance is achieved for *(i)* sufficiently small values of C_{clip}, and *(ii)* an appropriate ratio C_{clip}/λ. In fact, the lowest validation risk occurs on a top-left to bottom-right diagonal of the heatmap. Then, for any output perturbation algorithm, we take the values of $\lambda \geq 0.1$ and $C_{\text{clip}} \leq 0.1\zeta$ that minimize the validation risk. In the left panel of Fig. 2b, we plot the reconstruction error as a function of ρ and d, and identify a clear diagonal boundary $\rho/d \asymp 1$ between successful and unsuccessful attacks, validating our result in (16). In the right panel of Fig. 2b, we plot the same reconstruction error as a function of ρ/d, and show that curves corresponding to different values of d (in different colors) collapse onto each other under this scaling.

Natural images. In Fig. 3, we consider ImageNet data, and perform the same experiments as those described in Fig. 2 for synthetic data. Here, we tune the value of d by down-scaling images to a lower resolution, and we optimize the output perturbation algorithm across the hyper-parameters ranges $\lambda \geq 0.1$ and $C_{\text{clip}} \leq 0.01$. In Fig. 3a, we plot the reconstruction error as a function of ρ and d, and identify the same clear diagonal boundary as in Fig. 2 for the attack success across different image resolutions. In Fig. 3b, we then plot the same reconstruction error as a function ρ/d, noting that curves corresponding to different values of d collapse onto each other. This provides evidence that our proposed threshold $\rho \asymp d$ for data reconstruction persists in natural images as well. Furthermore, in Fig. 4 we give a visual presentation of some target images (first row) and the corresponding reconstructions for increasing values of ρ/d. Additional visual reconstructions appear in Figs. 7 and 8 (Appendix G.4).



(a) Synthetic rank-s data

(b) CIFAR-10 (frog vs. truck)

Figure 5: Reconstruction error for projected attacks on (a) synthetic rank-s data with $d = 1000$, and (b) CIFAR-10 (frog vs. truck) data with $d = 3{,}072$ (32×32 RGB images). Each pair of plots uses ρ/s (left) and ρ (right) on the x-axis, and different colors for different values of the rank s.

Low-dimensional data. In Fig. 5, we numerically investigate the conclusions of Theorems 3 and 4. More precisely, in Figure 5a, we fix the ambient dimension at $d = 1000$, and generate synthetic data concentrated near an s-dimensional subspace. The construction is detailed in Appendix G.1. We test the reconstruction attack in (18), estimating the rank-s PCA projector P from X_-, and report its performance as a function of ρ/s (left panel) and ρ (right panel). The reconstruction error curves with respect to ρ/s overlap for different values of s, showing a transition at the level of $\rho \asymp s$. In contrast, the curves are clearly separated when plotted against ρ. This validates our bound in (19), and it shows that the same privacy budget ρ can result in different guarantees in terms of data reconstruction, depending on the effective dimension of the data.

Next, in Fig. 5b, we consider a similar class of experiments, training our DP linear regression model on CIFAR-10 data. We perform the reconstruction attack in (18), defining P as the projector over the s principal components of the data distribution, estimated from X_-. We again plot the reconstruction error both in terms of ρ/s (left panel) and ρ (right panel). As for the synthetic data, the curves are separated when plotted against ρ, and they collapse onto each other when plotted against ρ/s. However, differently from the synthetic data, when s is too small (e.g., $s = 10$), the error exhibits a plateau even at large ρ/s, capturing the fact that low-rank subspaces do not capture all of the signal of the target. Similar results for ImageNet are reported in Fig. 6 (Appendix G.3).

6 Conclusions

In this work, we study how differential privacy quantitatively determines the feasibility of successful reconstruction attacks. First, in Theorem 1, we prove that any ρ-zCDP mechanism prevents data reconstruction as long as the privacy budget is sufficiently smaller than the entropy of the target, i.e., $\rho \ll h(x_t)$. When $h(x_t) = \Theta(d)$ (which holds, e.g., for distributions with independent entries), this provides the sufficient condition $\rho \ll d$ to protect from reconstruction attacks. Then, in Theorem 2, we show that this condition is also necessary: there exists an attack on a linear model trained with output perturbation that achieves arbitrary accuracy as $\rho \gg d$. Taken together, these results identify $\rho \asymp d$ as a sharp transition in data reconstruction. We next show that this threshold shifts according to the *effective* dimension of the data: when the data lies on an s-dimensional subspace known to the attacker, the transition moves to $\rho \asymp s$ (Theorems 3 and 4). This highlights the role of the data prior in assessing the effective guarantees given by a privacy budget.

A first natural research direction for future work is to extend Theorem 2 to other algorithms (e.g., DP gradient descent or objective perturbation) and beyond the linear model assumption, as well as expand our results on low-dimensional data to non-linear subspaces. A second exciting avenue is to consider less powerful attackers, without knowledge of all remaining training data. Information-theoretic lower bounds in that setting would complement the recent work by Swanberg et al. (2026) on adversary-aware privacy guarantees. We suspect that matching those bounds with attacks would require a different learning model: when multiple training examples are unknown, the optimality conditions of a linear model generally leave their individual values under-determined (Runkel et al., 2025). Thus, the success of the attack will likely depend on the capacity of the learning model, in agreement with the results of Iurada et al. (2026) in the context of non-private learning.

AI USE STATEMENT

In this work, we used generative AI tools to implement straightforward parts of the code (e.g., Matplotlib code for generating figures), fill in few details of proofs for which we had already identified the key methods and arguments, and improve the completeness of the related work section. We did not use generative AI tools to develop the paper's core ideas, define the problem setting, or prove the main theoretical results. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

We do not believe this paper raises any ethical concerns, as it studies data reconstruction in a theoretical setting and aims to better understand the protection provided by differential privacy. Although reconstruction methods can potentially be misused, our analysis considers a strong informed-adversary model, and our experiments only use established public datasets, with no private, sensitive data.

REPRODUCIBILITY STATEMENT

Our theoretical results are proved in Appendices B to E. To aid readers in reproducing our experiments, we detail our setup in Appendix G, describing our datasets and procedure. Full hyperparameters are given in Appendix G.5 for all experimental runs used in this paper.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2016.

Michael Aerni, Jie Zhang, and Florian Tramèr. Evaluations of machine learning privacy defenses are misleading. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing Training Data with Informed Adversaries. In *IEEE Symposium on Security and Privacy (SP)*, May 2022. doi: 10.1109/SP46214.2022.9833677.

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv:1812.00984*, 2019.

Simone Bombari and Marco Mondelli. Privacy for free in the overparameterized regime. *Proceedings of the National Academy of Sciences*, 2025.

Simone Bombari, Jialei Luo, Inbar Seroussi, and Marco Mondelli. High-dimensional private linear regression with optimal rates. *arXiv:2505.16329*, 2026.

Gavin R Brown, Krishnamurthy Dj Dvijotham, Georgina Evans, Daogao Liu, Adam Smith, and Abhradeep Guha Thakurta. Private gradient descent for linear regression: Tighter error bounds and instance-specific uncertainty estimation. In *International Conference on Machine Learning (ICML)*, 2024.

Mark Bun and Thomas Steinke. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In *Theory of Cryptography*, Lecture Notes in Computer Science. Springer, 2016. doi: 10.1007/978-3-662-53641-4_24.

Gon Buzaglo, Niv Haim, Gilad Yehudai, Gal Vardi, Yakir Oz, Yaniv Nikankin, and Michal Irani. Deconstructing data reconstruction: Multiclass, weight decay and general losses. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (SP)*, 2022.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

A Feder Cooper, Mark A Lemley, Allison Casasola, Ahmed Ahmed, Aaron Gokaslan, Amy B Cyphert, Christopher De Sa, Daniel E Ho, and Percy Liang. Extracting memorized pieces of (copyrighted) books from open-weight language models. In *Conference on Language Modeling*, 2026.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2001.

Edwige Cyffers. Setting ε is not the issue in differential privacy. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv:2204.13650*, 2022.

Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2022.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Lecture Notes in Computer Science, Berlin, Heidelberg, 2006. Springer.

Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding Training Data Reconstruction in Private (Deep) Learning. In *International Conference on Machine Learning (ICML)*, 2022.

Chuan Guo, Alexandre Sablayrolles, and Maziar Sanjabi. Analyzing privacy leakage in machine learning via multiple hypothesis testing: A lesson from Fano. In *International Conference on Machine Learning (ICML)*, 2023.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing Training Data from Trained Neural Networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Jamie Hayes, Saeed Mahloujifar, and Borja Balle. Bounding training data reconstruction in DP-SGD. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Leonardo Iurada, Simone Bombari, Tatiana Tommasi, and Marco Mondelli. A law of data reconstruction for random features (and beyond). In *International Conference on Learning Representations (ICLR)*, 2026.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private SGD? In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning (ICML)*, 2015.

Georgios Kaissis, Jamie Hayes, Alexander Ziller, and Daniel Rueckert. Bounding data reconstruction attacks with the hypothesis testing interpretation of differential privacy. *arXiv:2307.03928*, 2023.

Kousha Kalantari, Lalitha Sankar, and Anand D. Sarwate. Robust Privacy-Utility Tradeoffs Under Differential Privacy and Hamming Distortion. *IEEE Transactions on Information Forensics and Security*, 2018.

Antti Koskela and Tejas Kulkarni. Practical differentially private hyperparameter tuning with subsampling. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Bogdan Kulynych, Juan Felipe Gomez, Georgios Kaissis, Jamie Hayes, Borja Balle, Flavio P. Calmon, and Jean Louis Raisaro. Unifying Re-Identification, Attribute Inference, and Data Reconstruction Risks in Differential Privacy. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022.

Sheng Liu, Zihan Wang, Yuxiao Chen, and Qi Lei. Data Reconstruction Attacks and Defenses: A Systematic Evaluation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Noel Loo, Ramin Hasani, Mathias Lechner, Alexander Amini, and Daniela Rus. Understanding reconstruction attacks with the neural tangent kernel and dataset distillation. In *International Conference on Learning Representations (ICLR)*, 2024.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Saeed Mahloujifar, Alexandre Sablayrolles, Graham Cormode, and Somesh Jha. Optimal membership inference bounds for adaptive composition of sampled Gaussian mechanisms. *arXiv:2204.06106*, 2022.

Ryan McKenna, Yangsibo Huang, Amer Sinha, Borja Balle, Zachary Charles, Christopher A. Choquette-Choo, Badih Ghazi, Georgios Kaissis, Ravi Kumar, Ruibo Liu, Da Yu, and Chiyuan Zhang. Scaling laws for differentially private language models. In *International Conference on Machine Learning (ICML)*, 2025.

Ilya Mironov. Renyi Differential Privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, 2017.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 2022.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy (SP)*, 2019.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *USENIX Security Symposium*, 2023.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Yakir Oz, Gilad Yehudai, Gal Vardi, Itai Antebi, Michal Irani, and Niv Haim. Reconstructing training data from real world models trained with transfer learning. *arXiv:2407.15845*, 2024.

Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. In *International Conference on Learning Representations (ICLR)*, 2022.

Christina Runkel, Kanchana Vaishnavi Gandikota, Jonas Geiping, Carola-Bibiane Schönlieb, and Michael Moeller. Training data reconstruction: Privacy due to uncertainty? In *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025.

Anand D. Sarwate and Lalitha Sankar. A rate-distortion perspective on local differential privacy. In *Allerton Conference on Communication, Control, and Computing (Allerton)*, 2014.

Yujie Shen, Zihan Wang, Jian Qian, and Qi Lei. Data Reconstruction: Identifiability and Optimization with Sample Splitting. In *International Conference on Machine Learning (ICML)*, 2026.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, 2017.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Pierre Stock, Igor Shilov, Ilya Mironov, and Alexandre Sablayrolles. Defending against reconstruction attacks with rényi differential privacy. *arXiv:2202.07623*, 2022.

Marika Swanberg, Meenatchi Sundaram Muthu Selva Annamalai, Jamie Hayes, Borja Balle, and Adam Smith. A unified framework for adversary-aware differential privacy bounds. *arXiv:2507.08158*, 2026.

Tim van Erven and Peter Harremoës. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 2014.

Roman Vershynin. *Introduction to the Non-Asymptotic Analysis of Random Matrices*. Cambridge University Press, 2012.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, 2018.

Alexander Ziller, Tamara T. Mueller, Simon Stieger, Leonhard F. Feiner, Johannes Brandt, Rickmer Braren, Daniel Rueckert, and Georgios Kaissis. Reconciling privacy and accuracy in AI for medical imaging. *Nature Machine Intelligence*, 2024.

A Additional notation

Given a symmetric matrix A, we denote with $\mathrm{tr}(A)$ its trace, and by $s_{\min}(A)$ ($s_{\max}(A)$) its smallest (largest) singular value. Given a symmetric matrix A, we denote by $\lambda_{\min}(A)$ ($\lambda_{\max}(A)$) its smallest (largest) eigenvalue. For square, symmetric matrices A, B, we write $A \preceq B$ if $B - A$ is positive semi-definite (p.s.d.), where A is p.s.d. if $x^\top A x \geq 0$ for all $x \in \mathbb{R}^n$. We use the notation $[b]_+ = \max\{b, 0\}$ for $b \in \mathbb{R}$.

B Sensitivity and privacy of output perturbation

Lemma 1 (Sensitivity of (6))**.** *Fix $\lambda, R, C_{\mathrm{clip}} > 0$. Let Z, Z' be adjacent datasets of size n, with every feature vector in either dataset having norm at most R. Then, the minimizer in* (6) *satisfies*

$$\|\theta^*(Z) - \theta^*(Z')\|_2 \leq \frac{2RC_{\mathrm{clip}}}{\lambda n}. \tag{20}$$

Proof. Without loss of generality, let Z and Z' differ in their first example, and write

$$\theta := \theta^*(Z), \qquad \theta' := \theta^*(Z'), \qquad r := \theta - \theta'.$$

Define

$$g_i(\vartheta) := x_i \ell'_{C_{\mathrm{clip}}}(x_i^\top \vartheta - y_i), \qquad g_1'(\vartheta) := x_1' \ell'_{C_{\mathrm{clip}}}((x_1')^\top \vartheta - y_1').$$

The first-order optimality conditions for the two strongly convex objectives are

$$0 = \frac{1}{n}\left(g_1(\theta) + \sum_{i=2}^{n} g_i(\theta)\right) + \lambda\theta, \qquad 0 = \frac{1}{n}\left(g_1'(\theta') + \sum_{i=2}^{n} g_i(\theta')\right) + \lambda\theta'.$$

Subtracting these equations and taking the inner product with r gives

$$\lambda\|r\|_2^2 = -\frac{1}{n} r^\top \left(\sum_{i=2}^{n} \big(g_i(\theta) - g_i(\theta')\big)\right) - \frac{1}{n} r^\top \left(g_1(\theta) - g_1'(\theta')\right). \tag{21}$$

The first term on the RHS is non-positive. In fact, $\sum_{i=2}^{n} \ell_{C_{\text{clip}}}(x_i^\top \vartheta - y_i)$ is convex in ϑ, so its gradient is monotone, i.e.,

$$r^\top \left(\sum_{i=2}^{n} \big(g_i(\theta) - g_i(\theta')\big)\right) \geq 0.$$

Since $|\ell'_{C_{\text{clip}}}(\cdot)| \leq C_{\text{clip}}$ and $\|x_i\|_2 \leq R$, every per-example gradient satisfies

$$\|g_i(\vartheta)\|_2 \leq RC_{\text{clip}}.$$

Then, plugging this result into (21), Cauchy-Schwarz and the triangle inequality yield

$$\begin{aligned}\lambda\|r\|_2^2 &\leq \frac{\|r\|_2}{n}\left(\|g_1(\theta)\|_2 + \|g_1'(\theta')\|_2\right) \\ &\leq \frac{2RC_{\text{clip}}}{n}\|r\|_2.\end{aligned}$$

If $r = 0$, the claim is immediate. Otherwise, dividing by $\lambda\|r\|_2$ proves

$$\|r\|_2 = \|\theta^*(Z) - \theta^*(Z')\|_2 \leq \frac{2RC_{\text{clip}}}{\lambda n},$$

which gives the desired result. □

Proposition 1. *The output perturbation algorithm defined by* (6)-(7) *is ρ-zCDP provided*

$$\sigma_{\text{DP}}^2 = \frac{2R^2C_{\text{clip}}^2}{\lambda^2\rho n^2}. \tag{8}$$

Proof. By Lemma 1, for adjacent Z, Z',

$$\|\theta^*(Z) - \theta^*(Z')\|_2 \leq \Delta_2 := \frac{2RC_{\text{clip}}}{\lambda n}.$$

For every $\alpha > 1$, the Rényi divergence between Gaussians with means μ, μ' and common covariance $\sigma_{\text{DP}}^2 I_d$ is given by (see (Mironov, 2017, Proposition 7))

$$D_\alpha\big(\mathcal{N}(\mu, \sigma_{\text{DP}}^2 I_d)\big\|\mathcal{N}(\mu', \sigma_{\text{DP}}^2 I_d)\big) = \frac{\alpha\|\mu - \mu'\|_2^2}{2\sigma_{\text{DP}}^2}.$$

Taking $\mu = \theta^*(Z)$ and $\mu' = \theta^*(Z')$, the stated lower bound on σ_{DP}^2 gives

$$D_\alpha\big(\tilde{\theta}(Z)\|\tilde{\theta}(Z')\big) \leq \frac{\alpha\Delta_2^2}{2\sigma_{\text{DP}}^2} \leq \rho\alpha.$$

The claim follows from Definition 1. □

Note, of course, that larger variances than the value stated also satisfy ρ-zCDP given the same hyperparameters, but lead to greater-than-necessary utility loss; we therefore assume the learner chooses the smallest variance to satisfy their desired privacy bound.

C PROOF OF THE INFORMATION-THEORETIC LOWER BOUND

We start with a lemma using arguments from Bun & Steinke (2016).

Lemma 2 (Mutual information bound from ρ-zCDP)**.** *Suppose $\mathcal{M}$ is a ρ-zCDP mechanism, applied to a dataset $Z = Z_- \cup \{z_t\}$, where z_t is a random variable and Z_- is fixed. Then, we have*

$$I(z_t; \mathcal{M}(Z_- \cup \{z_t\})) \leq \rho, \tag{22}$$

where $I(\cdot;\cdot)$ denotes the mutual information in the probability space of z_t and of the private mechanism.

Proof. As $\mathcal{M}$ is ρ-zCDP, due to Definition 1, we have that

$$D_\alpha(\mathcal{M}(Z) \| \mathcal{M}(Z')) \leq \rho \cdot \alpha,$$

for all $\alpha > 1$ and all neighboring datasets Z, Z' where $Z = Z_- \cup \{z\}$, $Z' = Z_- \cup \{z'\}$, for arbitrary $z, z' \in \mathcal{Z}$. By continuity of the Rényi divergence as $\alpha \to 1$ (on $\alpha \in [0, \infty]$ when $D_\alpha(P\|Q) < \infty$) (van Erven & Harremoës, 2014, Theorem 7), the above condition gives

$$D_{\mathrm{KL}}(\mathcal{M}(Z) \| \mathcal{M}(Z')) = D_1(\mathcal{M}(Z) \| \mathcal{M}(Z')) \leq \rho. \tag{23}$$

For shorthand, we introduce the notation $\tilde{\theta} = \mathcal{M}(Z_- \cup \{z_t\})$, θ for a value taken by $\tilde{\theta}$, and z for a value taken by z_t. Then, we have

$$\begin{aligned}
I(z_t; \tilde{\theta}) &= \int p_{z_t}(z) p_{\tilde{\theta}|z_t}(\theta|z) \log \frac{p_{\tilde{\theta}|z_t}(\theta|z)}{p_{\tilde{\theta}}(\theta)} \, \mathrm{d}z \, \mathrm{d}\theta \\
&= \int p_{z_t}(z) \left(\int p_{\tilde{\theta}|z_t}(\theta|z) \log \frac{p_{\tilde{\theta}|z_t}(\theta|z)}{p_{\tilde{\theta}}(\theta)} \, \mathrm{d}\theta \right) \mathrm{d}z \\
&= \int p_{z_t}(z) D_{\mathrm{KL}}((\tilde{\theta} \mid z_t = z) \,\|\, \tilde{\theta}) \, \mathrm{d}z \\
&= \int p_{z_t}(z) D_{\mathrm{KL}}((\tilde{\theta} \mid z_t = z) \,\|\, \mathbb{E}_{z_t}[\tilde{\theta} \mid z_t]) \, \mathrm{d}z \\
&\leq \int p_{z_t}(z) p_{z_t}(z') \left(\int p_{\tilde{\theta}|z_t}(\theta|z) \log \frac{p_{\tilde{\theta}|z_t}(\theta|z)}{p_{\tilde{\theta}|z_t}(\theta|z')} \, \mathrm{d}\theta \right) \mathrm{d}z' \, \mathrm{d}z \\
&= \int p_{z_t}(z) p_{z_t}(z') D_{\mathrm{KL}}(\mathcal{M}(Z) \| \mathcal{M}(Z')) \, \mathrm{d}z' \, \mathrm{d}z \\
&\leq \int p_{z_t}(z) p_{z_t}(z') \rho \, \mathrm{d}z' \, \mathrm{d}z, \\
&= \rho,
\end{aligned}$$

where the fifth line holds as the KL divergence is convex in its second argument (van Erven & Harremoës, 2014, Theorem 12) and the seventh line follows from (23). □

We can now turn to a brief discussion on rate-distortion and distortion-rate functions, before we develop a theorem combining this with the mutual information bound. For more details, we refer the interested reader to the classical textbook (Cover & Thomas, 2001). Let S be a random variable in $\mathbb{R}^k$, for an arbitrary dimension $k \geq 1$, and consider a (possibly random) function $\mathcal{C} : \mathbb{R}^k \to \mathbb{R}^k$ which we will refer to as *channel*. We define its rate and distortion by

$$\mathcal{R}(\mathcal{C}) := I(S; \mathcal{C}(S)), \qquad \mathcal{D}(\mathcal{C}) := \mathbb{E}\big[\|S - \mathcal{C}(S)\|_2^2\big].$$

Then, we define the *rate-distortion* and *distortion-rate* functions as

$$\mathfrak{r}(D) := \inf_{\mathcal{C}: \mathcal{D}(\mathcal{C}) \leq D} \mathcal{R}(\mathcal{C}), \qquad \mathfrak{d}(\mathsf{R}) := \inf_{\mathcal{C}: \mathcal{R}(\mathcal{C}) \leq \mathsf{R}} \mathcal{D}(\mathcal{C}). \tag{24}$$

We will use this notation in the remaining part of the section.

Lemma 3 (Rate/distortion bounds for the squared ℓ_2 distortion)**.** *Let $D > 0$ and $\mathsf{R} \geq 0$. Then, the rate-distortion and distortion-rate functions defined in* (24) *satisfy*

$$\mathfrak{r}(D) \geq \left[h(S) - \frac{k}{2} \log \left(2\pi e \frac{D}{k} \right) \right]_+, \tag{25}$$

$$\mathfrak{d}(\mathsf{R}) \geq \frac{k}{2\pi e} \exp \left(\frac{2}{k} (h(S) - \mathsf{R}) \right). \tag{26}$$

Proof. We have

$$\begin{aligned}
\mathfrak{r}(D) &= \inf_{\mathcal{C}: \mathbb{E}\|\mathcal{C}(S) - S\|_2^2 \leq D} I(S; \mathcal{C}(S)) \\
&= h(S) - \sup_{\mathcal{C}: \mathbb{E}\|\mathcal{C}(S) - S\|_2^2 \leq D} h(S|\mathcal{C}(S)) \\
&= h(S) - \sup_{\mathcal{C}: \mathbb{E}\|\mathcal{C}(S) - S\|_2^2 \leq D} h(S - \mathcal{C}(S)|\mathcal{C}(S)) \\
&\geq h(S) - \sup_{\mathcal{C}: \mathbb{E}\|\mathcal{C}(S) - S\|_2^2 \leq D} h(S - \mathcal{C}(S)) \\
&= h(S) - \sup_{\Delta: \mathbb{E}\|\Delta\|_2^2 \leq D} h(\Delta) \\
&= h(S) - \frac{k}{2} \log \left(2\pi e \frac{D}{k} \right).
\end{aligned}$$

Here, the second line uses $I(S; \mathcal{C}(S)) = h(S) - h(S|\mathcal{C}(S))$; the third line uses that the conditional differential entropy is translation-invariant; the fourth line uses that conditioning decreases entropy; in the fifth line we take $\Delta = S - \mathcal{C}(S)$ and use that the space of all random variables Δ with $\|\Delta\|_2^2 \leq D$ is equivalent to $\{S - \mathcal{C}(S) : \mathbb{E}\|\mathcal{C}(S) - S\|_2^2 \leq D\}$; the sixth line uses that, under this distortion condition, the max-entropy distribution is Gaussian (Cover & Thomas, 2001, Theorem 9.6.5).

Combining this with $\mathfrak{r}(D) \geq 0$ gives the first bound. For any channel with $\mathcal{R}(\mathcal{C}) \leq \mathsf{R}$, we have

$$h(S) - \frac{k}{2} \log \left(2\pi e \frac{\mathcal{D}(\mathcal{C})}{k} \right) \leq \mathfrak{r}(\mathcal{D}(\mathcal{C})) \leq \mathcal{R}(\mathcal{C}) \leq \mathsf{R}.$$

Rearranging and taking the infimum over such channels gives

$$\mathfrak{d}(\mathsf{R}) \geq \frac{k}{2\pi e} \exp \left(\frac{2}{k} (h(S) - \mathsf{R}) \right).$$

□

Theorem 1. *Let the target feature x_t be independent of the rest of the dataset $Z_- = (X_-, y_-)$. Let $\mathcal{A} : \mathcal{Z}^{n-1} \times \Theta \to \mathcal{X}$ be an attack, taking as input Z_- and the output $\mathcal{M}(Z_- \cup \{z_t\})$ of a ρ-zCDP mechanism, and producing as output an estimate $\widehat{x}$ of x_t. Then,*

$$\frac{\mathbb{E}\|x_t - \widehat{x}\|_2^2}{\mathbb{E}\|x_t\|_2^2} \geq \frac{1}{2\pi e} \exp \left(\frac{2}{d} \left(h(x_t') - \rho \right) \right), \tag{9}$$

with $x_t' := \frac{\sqrt{d}}{\sqrt{\mathbb{E}\|x_t\|_2^2}} x_t$.

Note that we assume that the differential entropy is well-defined, i.e. that x_t has a density with respect to the Lebesgue measure on $\mathbb{R}^d$, has finite differential entropy, and $0 < \mathbb{E}\|x_t\|_2^2 < \infty$.

Proof. Fix Z_- and write $z_t = (x_t, y_t)$. Consider the shorthand $\theta(z) = \mathcal{M}(Z_- \cup \{z\})$. The target pair z_t, the resulting learned weights $\theta(z_t)$, and the reconstruction $\widehat{x} = \mathcal{A}(Z_-, \theta(z_t))$ form a Markov chain

$$z_t \to \theta(z_t) \to \widehat{x}.$$

Since x_t is the feature component of z_t, we have

$$I(x_t; \widehat{x}) \leq I(z_t; \widehat{x}) \leq I(z_t; \theta(z_t)) \leq \rho,$$

where the first two steps follow from the data-processing inequality, and the last step holds due to Lemma 2, since $\mathcal{M}$ is ρ-zCDP. All information quantities here are evaluated conditional on the fixed Z_-.

Now, define the induced channel $\mathcal{C}(x_t) = \widehat{x}$, averaging over the conditional distribution of y_t and the randomness of the mechanism and attack. We apply the rate-distortion definitions to the feature variable x_t, taking $S = x_t$ and $k = d$ in Lemma 3. Since $\mathcal{R}(\mathcal{C}) = I(x_t; \widehat{x}) \leq \rho$, using the definition of distortion-rate function $\mathfrak{d}$ in (24) and Lemma 3, we have

$$\mathbb{E}\|x_t - \widehat{x}\|_2^2 = \mathcal{D}(\mathcal{C}) \geq \mathfrak{d}(\rho) \geq \frac{d}{2\pi e} \exp\left(\frac{2}{d}(h(x_t) - \rho)\right).$$

Finally, to remove the effect of the scale of x_t, we apply the fact that $h(AX) = h(X) + \log|\det(A)|$ (Cover & Thomas, 2001, Theorem 9.6.4) with $x_t' = \sqrt{\frac{d}{\mathbb{E}\|x_t\|_2^2}} x_t$, and divide by $\mathbb{E}\|x_t\|_2^2$ to obtain (9). □

D Proof of the algorithmic upper bound

Lemma 4. *Let $x_t \in \mathbb{R}^d$, and let $q \in \mathbb{R}^d$ be such that $q = ax_t + w \neq 0$, for some real number $a \neq 0$, and some $w \in \mathbb{R}^d$. Defining $\widehat{x} = \|x_t\|_2 \frac{q}{\|q\|_2}$, we have*

$$\min_{\tau \in [-1,1]} \|x_t - \tau \widehat{x}\|_2^2 \leq \frac{4\|w\|_2^2}{a^2}.$$

Proof. If $x_t = 0$, choose $\tau = 0$. Otherwise, set $R_t = \|x_t\|_2 > 0$ and choose $\tau = \operatorname{sign}(a) \in [-1, 1]$. Then

$$\begin{aligned}
\frac{1}{R_t}\|x_t - \tau\widehat{x}\|_2 &= \left\| \frac{q}{\|q\|_2} - \frac{ax_t}{\|ax_t\|_2} \right\|_2 \\
&\leq \left\| \frac{q}{\|q\|_2} - \frac{q}{\|ax_t\|_2} \right\|_2 + \left\| \frac{q - ax_t}{\|ax_t\|_2} \right\|_2 \\
&= \frac{\big|\|ax_t\|_2 - \|q\|_2\big|}{\|ax_t\|_2} + \frac{\|w\|_2}{\|ax_t\|_2} \\
&\leq \frac{2\|w\|_2}{|a|R_t}.
\end{aligned}$$

The last inequality follows from the reverse triangle inequality and $q = ax_t + w$. Multiplying by R_t, squaring, and minimizing over $\tau \in [-1, 1]$ gives the desired result. □

Lemma 5 (A property of F_-)**.** *Let F_- be defined as in* (12)*. Then, for any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$F_-(\theta_1) - F_-(\theta_2) = M(\theta_1 - \theta_2), \tag{27}$$

for some positive semi definite matrix M that satisfies

$$n\lambda I \preceq M \preceq {X_-}^\top X_- + n\lambda I. \tag{28}$$

Proof. Define a diagonal matrix $D \in \mathbb{R}^{(n-1)\times(n-1)}$ with entries

$$D_{ii} := \begin{cases} \dfrac{\psi_{C_{\text{clip}}}(x_i^\top \theta_1 - y_i) - \psi_{C_{\text{clip}}}(x_i^\top \theta_2 - y_i)}{x_i^\top(\theta_1 - \theta_2)}, & \text{if } x_i^\top(\theta_1 - \theta_2) \neq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where we consider non-target points $X_- = \{x_i\}_{i \in [n-1]}$. Then, (12) yields

$$F_-(\theta_1) - F_-(\theta_2) = \sum_{i \in [n-1]} x_i D_{ii} x_i^\top (\theta_1 - \theta_2) + \lambda n(\theta_1 - \theta_2) = M(\theta_1 - \theta_2),$$

with

$$M = {X_-}^\top D X_- + n\lambda I. \tag{29}$$

Since the gradient of the clipped loss function $\psi_{C_{\text{clip}}}$ is non-decreasing and 1-Lipschitz, we have $0 \leq D_{ii} \leq 1$, which implies that ${X_-}^\top D X_-$ is p.s.d. and such that ${X_-}^\top D X_- \preceq {X_-}^\top X_-$. □

Theorem 2. *Let Assumption 1 hold, and let $\lambda > 0$, $C_{\text{clip}} > 0$ and $\rho > 0$ be the parameters of the algorithm chosen by the learner. The attacker computes*

$$\widehat{x} := R \, \frac{F_-(\tilde{\theta})}{\|F_-(\tilde{\theta})\|_2}. \tag{14}$$

Let $\eta \in (0,1)$ be a failure probability and $C, c > 0$ be absolute constants. Then, with probability at least $1 - \eta - 2e^{-cn} - 2e^{-cd}$,

$$\frac{\min_{\tau \in [-1,1]} \|x_t - \tau \widehat{x}\|_2^2}{\|x_t\|_2^2} \leq C \left[\max\left\{ 1, \frac{C_{\text{clip}}}{\zeta \eta} \left(1 + \frac{d}{n\lambda}\right) \right\} \left(1 + \frac{1}{\lambda}\right) \left(1 + \frac{d}{n}\right) \right]^2 \frac{d}{\rho}. \tag{15}$$

Proof. Consider the shorthand for the residuals

$$r_t(\theta) := x_t^\top \theta - y_t.$$

As noted in (13), we have that

$$-F_-(\theta^*) = x_t \, \psi_{C_{\text{clip}}}(x_t^\top \theta^* - y_t).$$

On the perturbed noisy weights $\tilde{\theta}$, we have

$$-F_-(\tilde{\theta}) = x_t \, \psi_{C_{\text{clip}}}(r_t(\theta^*)) - (F_-(\tilde{\theta}) - F_-(\theta^*)). \tag{30}$$

Firstly, we lower bound the absolute value of $\psi_{C_{\text{clip}}}(r_t(\theta^*))$. A useful tool is the optimal solution to the same learning problem on all datapoints except the target, i.e., just on X_-, with a slightly altered regularization parameter $\frac{n}{n-1}\lambda$:

$$\theta_- := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n-1} \sum_{i \in [n], i \neq t} \ell_{C_{\text{clip}}}(x_i^\top \theta - y_i) + \frac{1}{2} \frac{n\lambda}{(n-1)} \|\theta\|_2^2 \right\}. \tag{31}$$

This objective is strongly convex, so its minimizer is unique, and its gradient is $F_-/(n-1)$. Hence $F_-(\theta_-) = 0$. Due to Lemma 5, we have

$$\theta^* - \theta_- = M^{-1}(F_-(\theta^*) - F_-(\theta_-)) = M^{-1} F_-(\theta^*),$$

where the second step holds as $F_-(\theta_-) = 0$. Note that $M \in \mathbb{R}^{d \times d}$ is a p.s.d. matrix such that $\left\|M^{-1}\right\|_{\text{op}} \leq (n\lambda)^{-1}$. Then, we have

$$\begin{aligned}
r_t(\theta^*) &= x_t^\top \theta^* - y_t && (32)\\
&= x_t^\top (\theta^* - \theta_-) + \underbrace{x_t^\top \theta_- - y_t}_{:= r_t(\theta_-)} && (33)\\
&= x_t^\top M^{-1} F_-(\theta^*) + r_t(\theta_-) && (34)\\
&= -x_t^\top M^{-1} x_t \psi_{C_{\text{clip}}}(r_t(\theta^*)) + r_t(\theta_-). && (35)
\end{aligned}$$

Since $r_t(\theta^*)$ and $\psi_{C_{\text{clip}}}(r_t(\theta^*))$ have the same sign, and $0 \leq x_t^\top M^{-1} x_t \leq \frac{R^2}{n\lambda}$, we have

$$|r_t(\theta_-)| = |r_t(\theta^*)| + x_t^\top M^{-1} x_t |\psi_{C_{\text{clip}}}(r_t(\theta^*))| \leq \left(1 + \frac{R^2}{n\lambda}\right) |r_t(\theta^*)|.$$

Thus, for any $u \geq 0$, we have

$$\mathbb{P}(|r_t(\theta^*)| \leq u) \leq \mathbb{P}\left(|r_t(\theta_-)| \leq \left(1 + \frac{R^2}{n\lambda}\right) u \right).$$

The residual $r_t(\theta_-)$ decomposes as $r_t(\theta_-) = x_t^\top (\theta_- - \theta_{\text{true}}) - \nu_t$. Conditional on (x_t, Z_-), the first term is fixed and $\nu_t \sim \mathcal{N}(0, \zeta^2)$ remains independent. Its density is bounded by $1/(\zeta\sqrt{2\pi})$, so

$$\mathbb{P}\left(|r_t(\theta_-)| \leq \zeta \eta \sqrt{\pi/2} \right) \leq \eta.$$

Together, this gives

$$\mathbb{P}(|r_t(\theta^*)| \le u) \le \eta, \qquad u := \frac{\zeta\eta\sqrt{\pi/2}}{1 + R^2/(n\lambda)}. \tag{36}$$

Next, we upper bound the term $\|F_-(\tilde{\theta}) - F_-(\theta^*)\|_2$. Using Lemma 5, since $\tilde{\theta} - \theta^* = b_{\mathrm{DP}}$ by definition (see (7)), we can write

$$F_-(\tilde{\theta}) - F_-(\theta^*) = \underbrace{(X_-{}^\top D X_- + n\lambda I)}_{M_{\mathrm{dp}}} b_{\mathrm{DP}},$$

where $\|M_{\mathrm{dp}}\|_{\mathrm{op}} \le \left\|X_-{}^\top X_-\right\|_{\mathrm{op}} + \lambda n$. Due to Assumption 1, we can apply Remark 5.40 in Vershynin (2012) with deviation parameter $t = \sqrt{n+d}$ to get

$$\left\|X_-{}^\top X_-\right\|_{\mathrm{op}} \le (n-1)\left\|\mathbb{E}[xx^\top]\right\|_{\mathrm{op}} + C(n+d),$$

with probability at least $1 - 2e^{-c_1(n+d)}$, for absolute constants $C, c_1 > 0$. Then, again by Assumption 1, we have that $\left\|\mathbb{E}[xx^\top]\right\|_{\mathrm{op}} \le C$ (see e.g. the argument in Lemma C.1 of Bombari & Mondelli (2025)), which readily gives, after enlarging C,

$$\|M_{\mathrm{dp}}\|_{\mathrm{op}} \le n\lambda + C(n+d). \tag{37}$$

Theorem 3.1.1 in Vershynin (2018), applied to the independent Gaussian coordinates of $b_{\mathrm{DP}}/\sigma_{\mathrm{DP}}$, yields $\|b_{\mathrm{DP}}\|_2 = O(\sqrt{d}\sigma_{\mathrm{DP}})$ with probability at least $1 - 2e^{-c_2 d}$, for an absolute constant $c_2 > 0$. Taking the intersection with the previous event, whose total failure probability is at most $2e^{-c_1(n+d)} + 2e^{-c_2 d}$, gives

$$\|F_-(\tilde{\theta}) - F_-(\theta^*)\|_2 = \|M_{\mathrm{dp}} b_{\mathrm{DP}}\|_2 = O\left(\sigma_{\mathrm{DP}}\sqrt{d}\,(n\lambda + n + d)\right). \tag{38}$$

Then, merging (36) and (38) in (30), and using Lemma 4, gives, with probability at least $1 - \eta - 2e^{-c_1(n+d)} - 2e^{-c_2 d}$,

$$\frac{\min_{\tau\in[-1,1]}\|x_t - \tau\widehat{x}\|_2^2}{R^2} \le C_1 \frac{n^2\left(\lambda + (1+d/n)\right)^2 \sigma_{\mathrm{DP}}^2 d}{R^2 \min\{u^2, C_{\mathrm{clip}}^2\}}, \tag{39}$$

where we used $|\psi_{C_{\mathrm{clip}}}(r)|^2 = \min\{r^2, C_{\mathrm{clip}}^2\}$ and $|r_t(\theta^*)| > u$ on the signal event.

Recalling u from above and σ_{DP} from Proposition 1:

$$u := \frac{\zeta\eta\sqrt{\pi/2}}{1 + R^2/(n\lambda)}, \qquad \sigma_{\mathrm{DP}} = \frac{\sqrt{2}RC_{\mathrm{clip}}}{\lambda\sqrt{\rho}n},$$

the bound becomes, for an absolute constant $C_2 > 0$,

$$\frac{\min_{\tau\in[-1,1]}\|x_t - \tau\widehat{x}\|_2^2}{\|x_t\|_2^2} \le C_2 \frac{R^2}{\|x_t\|_2^2} \max\left\{1, \frac{2C_{\mathrm{clip}}^2}{\pi\zeta^2\eta^2}\left(1 + \frac{R^2}{n\lambda}\right)^2\right\}\left(1 + \frac{C}{\lambda}\right)^2\left(1 + \frac{d}{n}\right)^2 \frac{d}{\rho}.$$

Here we used

$$1 + \frac{C(1+d/n)}{\lambda} \le \left(1 + \frac{C}{\lambda}\right)\left(1 + \frac{d}{n}\right).$$

By Assumption 1, $\|x_t\|_2^2 = \Theta(d)$ and $R^2 = \Theta(d)$, so the ratio displayed in the bound is $O(1)$. Absorbing it into the absolute constant proves (15). A union bound over the Gram-matrix event, the privacy-noise event, and the signal anti-concentration event gives total failure probability at most $\eta + 2e^{-c_1(n+d)} + 2e^{-c_2 d}$. □

E PROOFS FOR THE EXTENSION TO LOW-DIMENSIONAL DATA

Theorem 3. *In the same setting as Theorem 1, for any rank-s projection P satisfying the conditions above, any reconstruction of x_t from a ρ-zCDP mechanism obeys:*

$$\frac{\mathbb{E}\|x_t - \widehat{x}\|_2^2}{\mathbb{E}\|Px_t\|_2^2} \geq \frac{1}{2\pi e} \exp\left(\frac{2}{s}(h_s(x'_s) - \rho)\right), \tag{17}$$

with $x'_s := \frac{\sqrt{s}}{\sqrt{\mathbb{E}\|Px_t\|_2^2}} Px_t$.

Proof. Let $U \in \mathbb{R}^{d\times s}$ have orthonormal columns spanning $\mathrm{Im}(P)$, so that $P = UU^\top$ and $U^\top U = I_s$. Set $x_\| = Px_t$ and $\widehat{x}_\| = P\widehat{x}$, both in $\mathrm{Im}(P) \subseteq \mathbb{R}^d$, and let $v = U^\top x_t$ and $\widehat{v} = U^\top \widehat{x}$ be their coordinates in $\mathbb{R}^s$. Thus $x_\| = Uv$, $\widehat{x}_\| = U\widehat{v}$, and the entropy notation in the theorem means $h_s(x_t) = h(v)$.

Fix Z_-. As in the proof of Theorem 1, the data-processing inequality and Lemma 2, applied to $z_t = (x_t, y_t)$, give

$$I(v; \widehat{v}) \leq I(x_t; \widehat{x}) \leq \rho.$$

Let us apply Lemma 3 with $S = v$ and $k = s$, to the source v and reconstruction $\widehat{v}$ in $\mathbb{R}^s$. Since U preserves Euclidean norms, we obtain

$$\mathbb{E}\|x_\| - \widehat{x}_\|\|_2^2 = \mathbb{E}\|v - \widehat{v}\|_2^2 \geq \frac{s}{2\pi e} \exp\left(\frac{2}{s}(h_s(x_t) - \rho)\right). \tag{40}$$

Letting $x_\perp = (I - P)x_t \in \mathrm{Im}(P)^\perp$, we have

$$\|x_t - \widehat{x}\|_2^2 = \|x_\| - P\widehat{x}\|_2^2 + \|x_\perp - (I - P)\widehat{x}\|_2^2 \tag{41}$$

$$\geq \|x_\| - \widehat{x}_\|\|_2^2. \tag{42}$$

Taking expectations, dividing by $\mathbb{E}\|Px_t\|_2^2$ and using $h_s(x'_s) = h_s(x_t) + \frac{s}{2}\log\big(s/\mathbb{E}\|Px_t\|_2^2\big)$ (again, via an application of Theorem 9.6.4 in Cover & Thomas (2001)) gives (17). □

Theorem 4. *Consider the same setting as Theorem 2, but with $R = \Theta(\sqrt{s})$ and $\|x_t\|_2 = \Theta(\sqrt{s})$ replacing the corresponding $\Theta(\sqrt{d})$ assumptions. Let $P \in \mathbb{R}^{d\times d}$ be a rank-s projector, known to the attacker, such that $x_i \in \mathrm{Im}(P)$ for all $i \in [n]$. Let*

$$\hat{x}_P := R\frac{PF_-(\tilde{\theta})}{\|PF_-(\tilde{\theta})\|_2}. \tag{18}$$

Then, if $n = \Omega(s), \lambda = \Omega(1), C_{\mathrm{clip}} = O(\zeta)$, we have

$$\frac{\min_{\tau\in[-1,1]} \|x_t - \tau\hat{x}_P\|_2^2}{\|x_t\|_2^2} \lesssim \frac{s}{\rho}, \tag{19}$$

with probability at least 0.99.

Proof. Consider the shorthand $a := \psi_{C_{\mathrm{clip}}}(x_t^\top\theta^* - y_t)$. Recall that $b_{\mathrm{DP}} = \tilde{\theta} - \theta^*$ is the Gaussian noise defined in (7) and $M_{\mathrm{dp}} = X_-{}^\top DX_- + n\lambda I$ is the matrix constructed in the proof of Theorem 2. Projecting both sides of (30) over P gives

$$-PF_-(\tilde{\theta}) = aPx_t - PM_{\mathrm{dp}}b_{\mathrm{DP}} = ax_t - PM_{\mathrm{dp}}b_{\mathrm{DP}}. \tag{43}$$

Notice that we have

$$PM_{\mathrm{dp}}b_{\mathrm{DP}} = PM_{\mathrm{dp}}Pb_{\mathrm{DP}} + PM_{\mathrm{dp}}(I - P)b_{\mathrm{DP}} = PM_{\mathrm{dp}}Pb_{\mathrm{DP}},$$

where we use that $X_-(I - P) = 0$, due to $x_i \in \mathrm{Im}(P)$ for all $i \in [n]$.

Let $U \in \mathbb{R}^{d\times s}$ have orthonormal columns spanning $\mathrm{Im}(P)$, and set $Z_- := X_-U$. The rows of Z_- are independent $O(1)$-subGaussian vectors in $\mathbb{R}^s$. Applying the concentration argument used for (37), now in dimension s, gives

$$\left\|X_-{}^\top X_-\right\|_{\mathrm{op}} = \left\|Z_-^\top Z_-\right\|_{\mathrm{op}} \leq C(n + s)$$

with probability at least $1 - 2e^{-c_2 n}$. Consequently,

$$\|PM_{\text{dp}}\|_{\text{op}} \le \|M_{\text{dp}}\|_{\text{op}} \le n\lambda + C(n+s).$$

Then, we have that

$$\|PM_{\text{dp}} b_{\text{DP}}\|_2 \le \|PM_{\text{dp}}\|_{\text{op}} \|Pb_{\text{DP}}\|_2 = O\left(((1+\lambda)n + s)\,\sigma_{\text{DP}}\sqrt{s}\right),$$

with probability at least $1-2e^{-c_2 n}-2e^{-c_3 s}$, where the last equality follows from a similar argument as the one used to obtain (38).

Then, the argument follows the same path as the one for the proof of Theorem 2, after (38) and using Lemma 4. This yields, with probability at least $1 - \eta - 2e^{-c_2 n} - 2e^{-c_3 s}$,

$$\frac{\min_{\tau\in[-1,1]} \|x_t - \tau\widehat{x}\|_2^2}{\|x_t\|_2^2} \le C_1 \frac{R^2}{\|x_t\|_2^2} \frac{n^2\left(\lambda + (1+s/n)\right)^2 \sigma_{\text{DP}}^2 s}{R^2 \min\{u^2, C_{\text{clip}}^2\}},$$

where u is defined as a function of η in (36). Apart from the factor $R^2/\|x_t\|_2^2$, the RHS is the bound in (39) with d replaced by s. Furthermore, $\|x_t\|_2^2 = \Theta(s)$ and $R^2 = \Theta(s)$, so that $R^2/\|x_t\|_2^2 = O(1)$. Following the same substitution as at the end of the proof of Theorem 2, with $n = \Omega(s)$, $\lambda = \Omega(1)$, and $C_{\text{clip}} = O(\zeta\eta)$, we obtain the desired result by taking η to be a sufficiently small absolute constant. □

F Comparison with Balle et al. (2022)

Corollary 4 in (Balle et al., 2022) bounds reconstruction success in terms of the small-ball probability $\kappa(r) := \sup_{v\in\mathbb{R}^d} \mathbb{P}(\|x_t - v\|_2 \le r)$. Writing $\bar{\kappa}(r) \in (0,1]$ for an upper bound on $\kappa(r)$, we convert their guarantee to expected squared error by integrating the corresponding tail bound:

$$D := \mathbb{E}\|x_t - \widehat{x}\|_2^2 \ge \int_0^\infty 2r\left[1 - \exp\left(-\left[\sqrt{\log(1/\bar{\kappa}(r))} - \sqrt{\rho}\right]_+^2\right)\right] dr. \tag{44}$$

This follows from $D = \int_0^\infty 2r\mathbb{P}(\|x_t - \widehat{x}\|_2 > r)\,dr$.

For $x_t \sim \mathcal{N}(0, I_d)$, we use the Chernoff bound from their proof of Proposition 7,

$$\bar{\kappa}(r) = \begin{cases} \exp\left[\dfrac{d}{2}\left(1 - \dfrac{r^2}{d} + \log\dfrac{r^2}{d}\right)\right], & 0 < r \le \sqrt{d}, \\ 1, & r > \sqrt{d}. \end{cases}$$

For the uniform distribution on the ball of radius $\sqrt{d}$, their Proposition 6 proof gives, after rescaling, $\kappa(r) = \min\{(r/\sqrt{d})^d, 1\}$. Thus, our comparison uses the small-ball bounds in their proofs, retaining their radius dependence before the asymptotic simplifications in the statements of the propositions.

G Experimental details and additional numerical results

This appendix describes the datasets (Appendix G.1) and the experimental procedure (Appendix G.2), and then provides more details for the figures in the main text. We provide further visual reconstructions for the full-dimensional attack (Appendix G.4) and the remaining experiments on the low-dimensional setting (Appendix G.3). Parameter configurations are given in Appendix G.5.

G.1 Datasets

Synthetic data. We draw $n = 1000$ training examples independently and uniformly from the sphere of radius $R := \sqrt{d}$. We draw θ_{true} uniformly from the unit sphere and generate labels according to Eq. (1). Validation and test sets of 10,000 examples are drawn from the same distribution.

Synthetic rank-s data. For the low-dimensional experiments, we design a distribution that has a subspace with a strong signal, and a weak component on its complement. For $s < d$, each example is initially generated as $x = x_{\|} + x_{\perp}$, where the two components are independent and uniform on spheres in complementary coordinate blocks of size s and $d - s$, with respective radii $\sqrt{s}$ and $\tau_b\sqrt{s/(1-\tau_b{}^2)}$. Then, $\mathbb{E}[x_{\|}x_{\|}{}^{\top}] = I$ on the signal block, $\mathbb{E}[x_{\perp} \mid x_{\|}] = 0$, and every example has norm $\|x\|_2 = \sqrt{s/(1-\tau_b{}^2)}$, with $\|x_{\perp}\|_2/\|x\|_2 = \tau_b$. The experiments in Fig. 5a use $\tau_b = 0.01$, so the data lies near, rather than exactly in, the signal subspace. A fixed Householder reflection is applied to every split, mixing the coordinate blocks without changing these norms. The attacker estimates the rank-s projector from the leading right singular vectors of X_{-}; the true signal subspace is not supplied to the attack. The feature clipping radius R is set 0.01 above the sample norm to avoid clipping due to rounding. We use 1000 training examples and 2000 examples in each of the validation and test sets, with θ_{true} and labels generated as in the preceding paragraph.

CIFAR-10. We form a binary frog-versus-truck problem from classes 6 and 9. The official training data for those classes are split into 8000 training and 2000 validation examples, and the official test split provides 2000 test examples.

ImageNet. We form binary problems from pairs of ImageNet-21K classes, specifically African elephant (`n02504458`) versus giant panda (`n02510455`). The pooled images of each class are shuffled and split $80/10/10$, giving 2160 training and 270 validation and test examples. Images are resized by downscaling to produce data of varying dimension.

Image preprocessing. RGB pixel values are mapped to $[0, 1]$ and centered as follows: for each dataset and resolution, we subtract a fixed public mean image, fitted using $10{,}000$ training images from classes outside the private binary pair. The same mean is reused across the training, validation, and test splits. Each centered feature vector is then rescaled to norm $\sqrt{d}$, with $R = \sqrt{d}$. Images retain their binary class labels, and label noise is not added.

G.2 Experimental procedure

For each run, data is generated as described above. The learner's private linear regression is implemented using a damped Newton solver on the clipped, regularized Huber objective in (6), running until convergence.

The private release is implemented by adding Gaussian noise as in (7), with the variance set by (8).

Replicates and aggregation. Each quantitative synthetic setting attacks one target for each of three independent seeds, hence three target attacks in total. Each quantitative image setting attacks one target from each binary class for each of three seeds, hence six target attacks in total. Hyperparameters are selected using the mean validation risk over these replicates. Each plotted heatmap cell or curve point is then the mean reconstruction error over the same target attacks. The visual reconstruction panels are not averaged: each displays four targets (two from each class) for a single random seed. The numbers of seeds and targets per class used for each figure are listed in Appendix G.5.

Privacy of hyperparameter selection. Our experiments compare output-perturbation mechanisms to illustrate the theoretical scaling laws, rather than evaluate an end-to-end private model-selection procedure. We therefore select hyperparameters using validation performance without accounting for the privacy cost of selection. The reported ρ guarantees apply to each trained mechanism, not the complete process including hyperparameter selection. Naturally, a real-world deployment on sensitive data would require private hyperparameter selection, for example using the methods of Papernot & Steinke (2022); Koskela & Kulkarni (2023), and would give guarantees that covered the entire process.

Image visualisation. For display of centered reconstructions, we choose the sign to give a non-negative inner product with the centered, normalized target. We then multiply by the target's saved inverse normalization scale, add back the public mean image, and clip RGB values to $[0, 1]$. We use the sign and scale of the target for visualization only.

G.3 Projected reconstruction: remaining figures

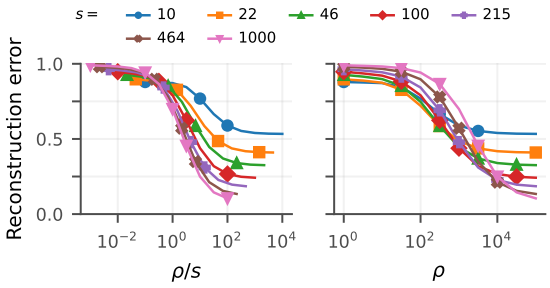
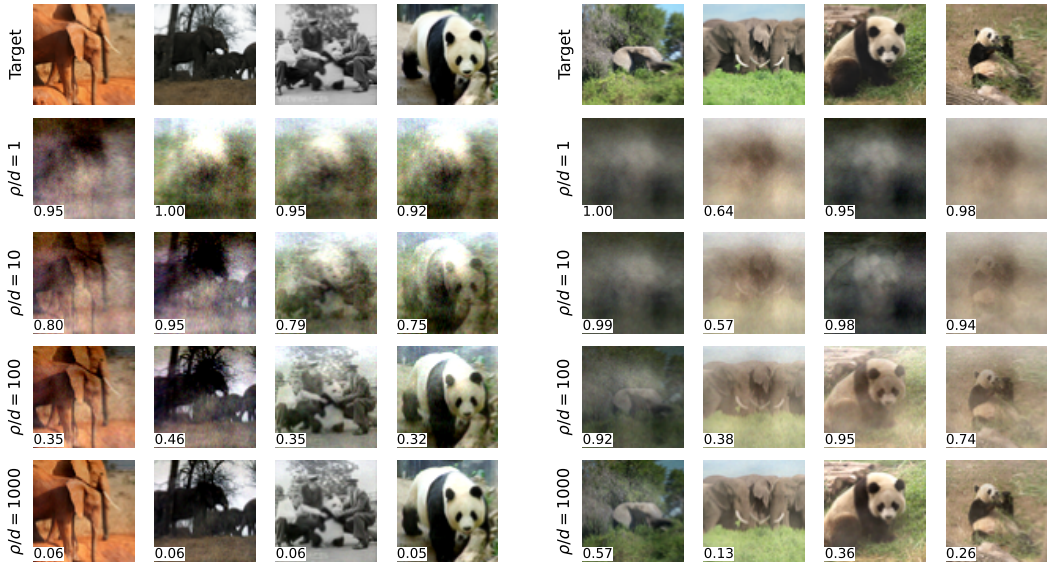


Figure 6: Projected reconstruction error on ImageNet (elephant vs. panda) data with $d = 12{,}288$. The projected attack uses a rank-s PCA subspace estimated from the known data. The left panel plots the projected attacks against ρ/s; the right panel plots them against ρ. Colours indicate the projection rank s.

G.4 Further visual reconstructions

In addition to Fig. 4, we present several more examples of reconstruction of ImageNet data in Fig. 7; and similarly for CIFAR-10 in Fig. 8.

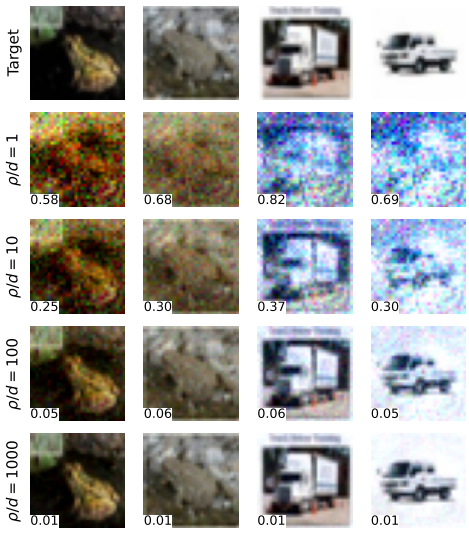


(a) (b)

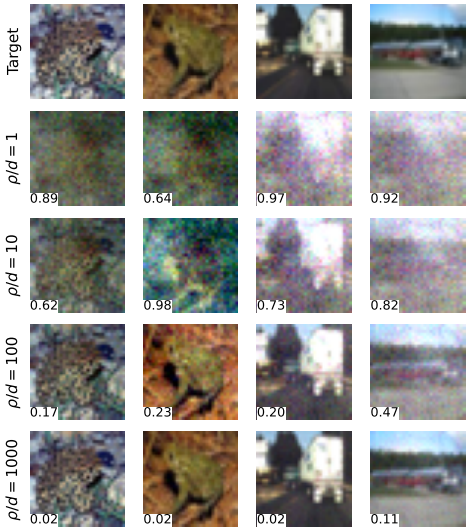
Figure 7: Additional ImageNet elephant-versus-panda reconstructions, presented as in Fig. 4.



(a)



(b)



(c)

Figure 8: CIFAR-10 frog-versus-truck reconstructions for all three target seeds, presented as in Fig. 4.

G.5 RUN CONFIGURATIONS PER FIGURE

Each table below lists the configurations for each figure. If a parameter had a large number of values, we summarize this as the endpoints of the range, the number of values, and the increment, if it was a geometric range.

Setting	Synthetic
Training examples n	1,000
Validation examples	10,000
Test examples	10,000
Dimension d	10–10,000 (7)
Feature norm $\|x\|_2$	$\sqrt{d}$
Feature clipping R	$\sqrt{d}$
Label noise ζ	0.5
Privacy budget ρ	0.1–10^5 (7, $\times 10$)
Regularization λ	10^{-4}–10 (11, $\times\sqrt{10}$)
Residual clipping C_{clip}	10^{-3}–1,000 (13, $\times\sqrt{10}$)
# of seeds	3

Table 1: Configuration for Figure 2.

Setting	ImageNet (elephant/panda)
Training examples n	2,160
Validation examples	270
Test examples	270
Dimension d	12–12,288 (6, $\times 4$)
Feature norm $\|x\|_2$	$\sqrt{d}$
Feature clipping R	$\sqrt{d}$
Privacy budget ρ	1–10^5 (11, $\times\sqrt{10}$)
Regularization λ	10^{-4}–1 (9, $\times\sqrt{10}$)
Residual clipping C_{clip}	10^{-3}–10 (9, $\times\sqrt{10}$)
# of seeds	3
Targets per class	1

Table 2: Configuration for Figure 3. Hyperparameter selection used validation risk within $\lambda \geq 0.1$ and $C_{\text{clip}} \leq 0.01$.

Setting	ImageNet (elephant/panda)
Training examples n	2,160
Validation examples	270
Test examples	270
Dimension d	12,288
Feature norm $\|x\|_2$	$\sqrt{d}$
Feature clipping R	$\sqrt{d}$
Privacy budget ρ	1.23×10^3–1.23×10^8 (6, $\times 10$)
Regularization λ	1
Residual clipping C_{clip}	0.125
# of seeds	1
Targets per class	2

Table 3: Configuration for Figure 4.

Setting	Synthetic rank-s signal	CIFAR-10 (frog/truck)
Training examples n	1,000	8,000
Validation examples	2,000	2,000
Test examples	2,000	2,000
Dimension d	1,000	3,072
Feature norm $\|x\|_2$	—	$\sqrt{d}$
Feature clipping R	3.17–31.6 (7)	$\sqrt{d}$
Label noise ζ	0.1	—
Privacy budget ρ	10–10,000 (7, $\times\sqrt{10}$)	10–10,000 (10, $\times 2.15$)
Regularization λ	10^{-4}–1 (9, $\times\sqrt{10}$)	$3.07, 307, 3{,}072$
Residual clipping C_{clip}	10^{-4}–1 (9, $\times\sqrt{10}$)	0.1
Attacker PCA rank s	10–1,000 (7)	10–1,000 (7)
Signal rank of the data	10–1,000 (7)	—
Bulk-to-signal ratio τ_b	0.01	—
# of seeds	3	3
Targets per class	—	1

Table 4: Configuration for Figure 5. Hyperparameter selection used validation risk within $\lambda \geq 0.1$ and $C_{\text{clip}} \leq 0.1\zeta$ for Synthetic rank-s signal; $\lambda \geq 10$ and $C_{\text{clip}} \leq 0.1$ for CIFAR-10 (frog/truck).

Setting	ImageNet (elephant/panda)
Training examples n	2,160
Validation examples	270
Test examples	270
Dimension d	12,288
Feature norm $\|x\|_2$	$\sqrt{d}$
Feature clipping R	$\sqrt{d}$
Privacy budget ρ	1–10^5 (11, $\times\sqrt{10}$)
Regularization λ	1–1,000 (7, $\times\sqrt{10}$)
Residual clipping C_{clip}	$0.01, 0.1$
Attacker PCA rank s	10–1,000 (7)
# of seeds	3
Targets per class	1

Table 5: Configuration for Figure 6. Hyperparameter selection used validation risk within $\lambda \geq 10$ and $C_{\text{clip}} \leq 0.01$.

Setting	ImageNet (elephant/panda)
Training examples n	2,160
Validation examples	270
Test examples	270
Dimension d	12,288
Feature norm $\|x\|_2$	$\sqrt{d}$
Feature clipping R	$\sqrt{d}$
Privacy budget ρ	1.23×10^3–1.23×10^8 (6, $\times 10$)
Regularization λ	1
Residual clipping C_{clip}	0.125
# of seeds	2
Targets per class	2

Table 6: Configuration for Figure 7.

Setting	CIFAR-10 (frog/truck)
Training examples n	8,000
Validation examples	2,000
Test examples	2,000
Dimension d	3,072
Feature norm $\|x\|_2$	$\sqrt{d}$
Feature clipping R	$\sqrt{d}$
Privacy budget ρ	307–3.07×10^7 (6, $\times 10$)
Regularization λ	1
Residual clipping C_{clip}	0.125
# of seeds	3
Targets per class	2

Table 7: Configuration for Figure 8.