FROM PIXELS TO PRIVACY-SENSITIVE FEATURES: AN INFORMATION-THEORETIC FRAMEWORK FOR RECONSTRUCTION BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

What can an adversary recover from a model trained with DP-SGD? Its (ε, δ) guarantee does not answer this question directly. We derive a distortion bound that can be computed before training and applies even to an adversary who observes the parameter trajectory during training, knows every other training record and the population distribution from which the training records are drawn. The analyst may choose the representation in which error is measured, allowing the bound to target the aspects of reconstruction that are privacy-sensitive in a given application. Our information-theoretic framework separates two contributions: the information about a record revealed by training, measured through conditional mutual information, and the uncertainty already present in the data distribution, measured through source entropy. Rate–distortion theory combines them into a lower bound on the expected error of every reconstruction attack. For DP-SGD, the leakage bound depends only on the sampling rates and noise multipliers, while the source entropy is lower-bounded from a sample with finite-sample confidence. An orthogonal block construction keeps the resulting bound informative for high-dimensional images. We instantiate the framework in pixel space and in LPIPS feature space, obtaining, to our knowledge, the first lower bound on reconstruction distortion measured directly by LPIPS. On a range of image datasets the bounds establish non-trivial unavoidable expected reconstruction error within heuristically recommended privacy-budget ranges and beyond. Experiments compare the bounds with strong reconstruction attacks and identify the principal sources of the remaining slack.

1 INTRODUCTION

Models trained on sensitive data can reveal individual training examples through gradients, model parameters, or query access (Zhu et al., 2019; Geiping et al., 2020; Haim et al., 2022; Carlini et al., 2020). The standard defence is differentially private stochastic gradient descent (DP-SGD), which clips per-example gradients, adds Gaussian noise, and reports its guarantee through (ε, δ) (Abadi et al., 2016). These parameters provide a general privacy guarantee, but they do not directly answer the operational question considered here: *How accurately could an adversary reconstruct a training*

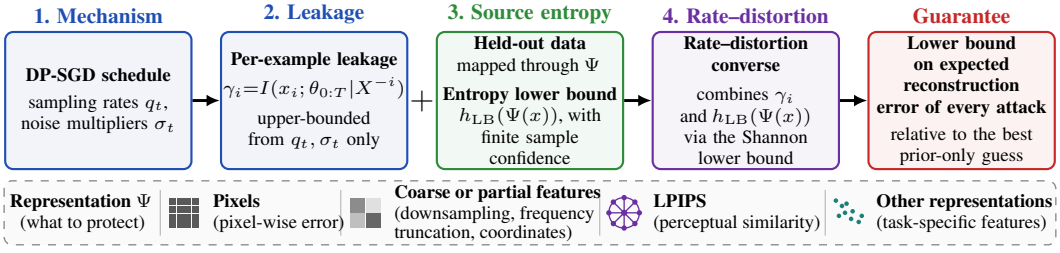


Figure 1: **Modular information-theoretic framework.** Mechanism leakage and a held-out entropy bound in a chosen representation Ψ combine through rate–distortion into a lower bound on every attack; each module can be tightened or replaced independently.

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.

example? This question is especially relevant when participation in the training set is already known but the content of an individual record remains sensitive, or when useful model accuracy requires a large privacy budget, ε. For orientation we use the tiers of Ponomareva et al. (2023): strong ($\varepsilon \leq 1$), reasonable ($1 < \varepsilon \leq 10$) and weak ($\varepsilon > 10$); the upper two are justified by common practice and empirical attacks rather than by a guarantee.

Reconstruction difficulty depends partly on what can be inferred from the training-data distribution before the trained model and its trajectory are observed. Even without seeing the training mechanism, an adversary who knows the training-data distribution can mount an effective guessing attack when that distribution has low entropy. We therefore measure reconstruction error relative to the best prior-only reconstruction. The resulting reconstruction ratio equals one when training reveals nothing useful beyond the training-data distribution and decreases in value as more accurate reconstruction becomes possible.

We consider an adversary who observes the noisy parameter trajectory during training, knows every other training record, knows the population distribution from which the training records are drawn, and may use any, possibly randomised, reconstruction rule.

Our framework separates the problem into a mechanism-side bound on the information revealed by training and a source-side bound on the uncertainty already present in the population training-data distribution. Rate–distortion theory combines them into a lower bound on the expected error of every reconstruction rule, biased or not; Figure 1 summarises this modular construction. For DP-SGD, the mechanism-side bound is computed before training from the sampling rates and noise multipliers, without using a reported (ε, δ) guarantee or analysing a trained model; it depends on neither the architecture nor the parameter count and can be added across training runs. The source-side entropy is lower-bounded from a finite sample with explicit confidence. The representation Ψ determines which reconstruction differences matter: the identity map gives pixel-space error, LPIPS features give perceptual error, and other representations can target coarse, partial, or task-specific recovery. An orthogonal block construction extends the calculation to high-dimensional representations while preserving squared reconstruction error exactly.

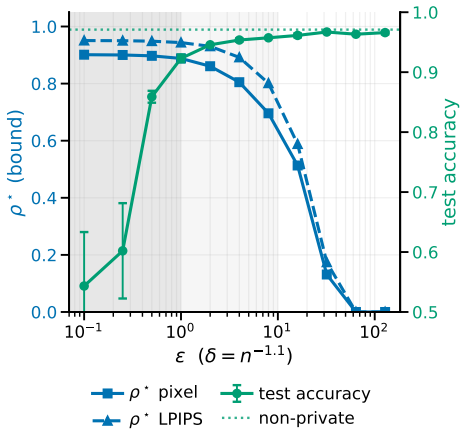


Figure 2: **CelebA: 95%-confidence reconstruction bounds and accuracy.** Pixel and LPIPS bounds $\rho^\star$ (left axis) and test accuracy (mean $\pm$ s.d. over three seeds; right axis), against ε at $\delta = n_{\text{train}}^{-1.1}$. The ratio $\rho^\star$ measures reconstruction error relative to the optimal prior-only guess: larger is more protective, and $\rho^\star = 1$ means no improvement over that guess. Shading marks the privacy tiers of Ponomareva et al. (2023); other datasets in Figure 5.

Our contributions include:

A modular information-theoretic framework. We combine mechanism leakage, source uncertainty, and an information-to-distortion converse to bound every attack under a strong informed-adversary model (Theorem 3.6); each component can be tightened or replaced independently.

Knowledge of training-data distribution. We assume that the adversary has complete knowledge of the population training-data distribution and can sample from it. In our bound it enters through a differential entropy that a finite sample-set can be used to bound from below with explicit confidence (Proposition 5.1).

Finite-sample, representation-flexible DP-SGD bounds. The mechanism-side input is computable before training; the source-side input is controlled from finite reference data. Choosing Ψ adapts the guarantee to different reconstruction harms; with LPIPS, this gives, to our knowledge, the first reconstruction lower bound measured directly by LPIPS.

A priori and retrospective analysis. The framework can select a DP-SGD noise multiplier to meet a prescribed reconstruction-error bound, or analyse the reconstruction-privacy of a published training process retrospectively from its reported schedule and a source sample.

Experiments and slack analysis. Across MNIST, CIFAR-10, CelebA and ImageNette, the bounds remain active in 'reasonable' and 'weak' ($\varepsilon > 10$) privacy levels. We compare attack performance with our lower bound and diagnose the gap between guaranteed and achieved reconstruction error.

2 Related Work

Reconstruction robustness. Balle et al. (2022) bound the probability that an informed adversary reconstructs a target within distortion η; subsequent work sharpens the accounting and extends it to settings including multiple targets (Hayes et al., 2023; Kaissis et al., 2023; Kulynych et al., 2026; Swanberg et al., 2025; Guerra-Balboa et al., 2026). Existing instantiations specify the required prior-only success probability analytically or estimate it without finite-sample coverage. We instead estimate a source-entropy term with explicit confidence. Threshold guarantees suit secrets for which near-exact recovery is the harm, whereas our expected distortion captures harmful partial reconstruction of images. Further comparison appears in Table 1, Section A.

Expected-error bounds. Estimation-theoretic work lower-bounds reconstruction error using Fisher information and related inequalities (Hannun et al., 2021; Guo et al., 2022; Chaudhuri et al., 2024; Maeng et al., 2023). The deployable Cramér–Rao and Hammersley–Chapman–Robbins bounds in this line apply to unbiased reconstructors and therefore exclude attacks exploiting the source prior. Our bound covers arbitrary reconstructors and is normalised by the optimal prior-only error; see Section A for a detailed comparison.

Information-theoretic antecedents. Conditional mutual information has been used to quantify information about individual samples revealed by learning algorithms (Raginsky et al., 2016; Feldman & Steinke, 2018; Cuff & Yu, 2016), including through sequential decompositions for noisy iterative algorithms (Pensia et al., 2018). Distortion-based secrecy originates in rate–distortion theory (Schieler & Cuff, 2013). Tan et al. (2024) consider the ambient entropy-power inequality (related to Theorem 3.6) using it as motivation for minimising one of the terms it contains. This is geared toward a defence against (rather than lower bound for) reconstruction in federated learning, and neither estimates the entropy nor addresses its looseness in high dimension.

3 Method

Our framework has three components. We first specify the reconstruction adversary and the representation in which error is measured. We then convert an upper bound on per-record information leakage into a lower bound on reconstruction distortion. Finally, we derive a blockwise form that remains useful when the chosen representation is high-dimensional.

3.1 Setting and threat model

We first formalise what the adversary observes, what side information it possesses, and how reconstruction quality is measured. Let $X = (x_1, \ldots, x_n)$ be a training dataset of n samples in $\mathbb{R}^d$, write X^{-i} for X with the i-th row removed, and let $\mathcal{A}$ be a possibly stochastic training algorithm producing the full released parameter trajectory

$$\theta := \theta_{0:T} = (\theta_0, \ldots, \theta_T)$$

from X. Our running instance is DP-SGD (Abadi et al., 2016), but the bound sees the mechanism only through the scalar γ_i of Definition 3.3; DP-SGD enters solely in Section 4.

Assumption 3.1 (Finite-precision data)**.** *Each training sample x_i takes values in a finite subset $\mathcal{Q} \subset \mathbb{R}^d$, obtained by representing each of the d components at b_x bits of precision. Concretely, each component takes values in a set of at most 2^{b_x} levels with spacing $\Delta = 2^{-b_x} \cdot c$ for some scale parameter $c > 0$.*

Assumption 3.2 (I.I.D. source)**.** *The rows of X are drawn i.i.d. from a per-sample distribution P_x on $\mathcal{Q}$, so the dataset has law $P_X = P_x^{\otimes n}$.*

These are all the reconstruction bound requires. The accountant (Section 4) and entropy estimator (Section 5) state any additional assumptions where they arise. Assumption 3.2 is violated by preprocessing computed across the training set and then applied back to it; see Section F.

Threat model. We adopt the adversary against which DP-SGD's own guarantee is proved (Abadi et al., 2016). The adversary knows the architecture, $\mathcal{A}$ and its hyperparameters, every non-target training record X^{-i}, the training set size n, and the population distribution P_x from which training

records are drawn. It observes the full released parameter trajectory $\theta_{0:T}$, including every intermediate noisy update, and may exploit all of this information through a measurable, possibly randomised reconstruction rule

$$\hat{x}_i = g(\theta_{0:T}, X^{-i}).$$

A bound against it also covers weaker adversaries, such as one observing only θ_T or lacking X^{-i}. Expectations are over $X \sim P_x^{\otimes n}$ and all randomness of training, reconstruction and dither, so the guarantee is averaged over the source, as in existing reconstruction guarantees formulated relative to a prior.

Definition 3.3 (Per-sample CMI)**.** *The information (which we measure in bits) about the target revealed by the full released parameter trajectory, conditional on all non-target records, is*

$$\gamma_i := I(x_i; \theta_{0:T} \mid X^{-i}).$$

The quantity γ_i measures what the trajectory carries about x_i once the rest of the dataset is known. Under Assumption 3.2, the joint distribution of $(x_i, \theta_{0:T}, X^{-i})$ is the same for every i. Because DP-SGD treats every index alike under Poisson subsampling, γ_i does not depend on i; we write γ for the common value and retain the index only in per-sample statements.

Lemma 3.4 (Per-sample leakage chain)**.** *Under Assumption 3.2, fix $m \geq 1$ and let $\hat{u}_i = g(\theta_{0:T}, X^{-i}) \in \mathbb{R}^m$ for a measurable, possibly randomised g whose internal randomness is independent of x_i given $(\theta_{0:T}, X^{-i})$. Then*

$$I(x_i; \hat{u}_i) \leq I(x_i; \theta_{0:T}, X^{-i}) = I(x_i; \theta_{0:T} \mid X^{-i}) =: \gamma_i.$$

Proof sketch. The inequality is the data-processing inequality along $x_i \to (\theta_{0:T}, X^{-i}) \to \hat{u}_i$. The equality follows from the chain rule $I(x_i; \theta_{0:T}, X^{-i}) = I(x_i; X^{-i}) + I(x_i; \theta_{0:T} \mid X^{-i})$, whose first term vanishes under Assumption 3.2. Full proof in Section B.1. □

Hence any reconstruction carries at most γ_i bits of mutual information about the target.

3.2 From information leakage to distortion

We now convert the information budget γ_i into unavoidable reconstruction error. The private object is x_i, the representation $\Psi(x_i)$ specifies the aspect of that record whose reconstruction is to be controlled, and a dithered version of $\Psi(x_i)$ is an analytical device that permits a differential-entropy bound for finite-precision data.

The rate–distortion bound requires a continuous source entropy, whereas Assumption 3.1 makes x_i discrete. We therefore add independent dither after applying Ψ. This makes differential entropy well defined, and the dither's contribution to squared error can later be subtracted exactly. The construction is internal to the analysis: the model trains on the raw X, and neither Ψ nor the dither is applied to the training data or observed by the adversary.

Definition 3.5 (Dithered representation)**.** *Call $\Psi : \mathbb{R}^d \to \mathbb{R}^m$* admissible *if, when random, its randomness is independent of X, of $\mathcal{A}$, and of the attack, with statements then holding conditionally on its draw; deterministic maps qualify. Fix a* dither scale *$s > 0$. The* dithered representation *of x_i under Ψ is*

$$z_i^{\Psi} := \Psi(x_i), \qquad \tilde{z}_i^{\Psi} := z_i^{\Psi} + W_i^{(m)}, \qquad W_i^{(m)} \overset{\text{iid}}{\sim} \mathcal{N}(0, s^2 I_m), \tag{1}$$

where the $W_i^{(m)}$ are drawn independently across i and of everything above. Thus z_i^{Ψ}, $\tilde{z}_i^{\Psi}$, and $\hat{z}_i^{\Psi}$ denote respectively the target representation, its dithered form, and the adversary's output, all in $\mathbb{R}^m$; the ambient case $\Psi = \mathrm{id}$ is written $\tilde{x}_i$ and $\hat{x}_i$.

The bound holds for every $s > 0$. The dither contribution is removed exactly at a cost that grows with s, while larger s reduces the bias of the entropy estimator supporting the bound; see Section E.1. For every Ψ, the dithered representation has a density and finite entropy (Section B.2). Write $\Sigma_\Psi := \mathrm{Cov}(\Psi(x_i))$, and abbreviate $\Sigma := \mathrm{Cov}(x_i)$.

Distortion–rate function. For an $\mathbb{R}^m$-valued source Y under squared error, the *distortion–rate function*, $D_Y(R)$, satisfies the following *Shannon lower bound* (SLB) (Koch, 2016):

$$D_Y(R) := \inf\{\mathbb{E}\|Y - \hat{Y}\|^2 : I(Y; \hat{Y}) \leq R\} \geq \frac{m}{2\pi e} 2^{2(h(Y)-R)/m}. \tag{2}$$

Equality holds at every R when Y is isotropic Gaussian (Cover & Thomas, 2006).

The reconstruction ratio. Absolute error is difficult to compare across datasets and representations. We therefore normalise by the optimal reconstruction that ignores the released parameter trajectory and uses only the source distribution. For a constant guess $a \in \mathbb{R}^m$,

$$\mathbb{E}\|\Psi(x_i) - a\|^2 = \operatorname{tr}\Sigma_\Psi + \|\mathbb{E}\Psi(x_i) - a\|^2,$$

so the optimal prior-only reconstruction is the prior mean $a = \mathbb{E}\Psi(x_i)$, with error $\operatorname{tr}\Sigma_\Psi$. We report each guarantee through the *reconstruction ratio*

$$\rho_\Psi(g) := \frac{\sqrt{\mathbb{E}\|z_i^\Psi - \hat{z}_i^\Psi\|^2}}{\mathrm{RMSE}_{\mathrm{prior}}(\Psi)}, \qquad \mathrm{RMSE}_{\mathrm{prior}}(\Psi) := \sqrt{\operatorname{tr}\Sigma_\Psi}, \qquad \nu_\Psi^2 := \frac{ms^2}{\operatorname{tr}\Sigma_\Psi}, \tag{3}$$

dropping the subscript Ψ in the ambient case. Since the prior mean is always available, $\inf_g \rho_\Psi(g) \leq 1$, and a valid bound never exceeds one (Remark C.2).

Choosing the representation. The map Ψ specifies which aspects of the record are treated as privacy-sensitive. Taking $\Psi = \mathrm{id}$ measures pixel-space reconstruction; a coordinate selection measures recovery of a region, channel, or attribute; and downsampling or frequency truncation measures coarse-scale recovery. Taking $\Psi = \Psi_{\mathrm{LPIPS}}$, the concatenated unit-normalised weighted feature maps of LPIPS (Zhang et al., 2018), makes $\|z_i^\Psi - \hat{z}_i^\Psi\|^2$ the perceptual distortion itself. More generally, any measurable map fixed independently of the protected data and training run, including a frozen learned encoder, is admissible. The leakage bound does not change with Ψ; only the source entropy and prior-only reconstruction error in the chosen representation must be recomputed.

We can now combine any valid leakage bound with the source entropy in the chosen representation.

Theorem 3.6 (Reconstruction lower bound under a representation)**.** *Let Assumptions 3.1–3.2 hold, let $\Psi : \mathbb{R}^d \to \mathbb{R}^m$ be admissible, and let $s > 0$. Every possibly randomised $\hat{z}_i^\Psi = g(\theta_{0:T}, X^{-i})$ taking values in $\mathbb{R}^m$ satisfies*

$$\mathbb{E}\big[\|z_i^\Psi - \hat{z}_i^\Psi\|^2\big] \geq \Big(\frac{m}{2\pi e}\, 2^{2(h(\tilde{z}_i^\Psi) - \gamma_i)/m} - ms^2\Big)_+, \tag{4}$$

equivalently, on the ratio scale of equation 3,

$$\rho_\Psi(g)^2 \geq \left(\frac{m}{2\pi e \operatorname{tr}\Sigma_\Psi} 2^{2(h(\tilde{z}_i^\Psi) - \gamma_i)/m} - \nu_\Psi^2\right)_+, \qquad \nu_\Psi^2 := \frac{ms^2}{\operatorname{tr}\Sigma_\Psi}. \tag{5}$$

The lower bound increases with uncertainty in the represented source and decreases with the information revealed by training.

Proof sketch. By Lemma 3.4 and the data-processing inequality along $\tilde{z}_i^\Psi \to x_i \to \hat{z}_i^\Psi$, $I(\tilde{z}_i^\Psi; \hat{z}_i^\Psi) \leq \gamma_i$. The dithered representation, unlike the discrete source, has a density and finite entropy, so the SLB controls $\mathbb{E}\|\tilde{z}_i^\Psi - \hat{z}_i^\Psi\|^2$ at rate γ_i. Because the dither is centred and independent of $z_i^\Psi - \hat{z}_i^\Psi$, its error ms^2 can then be subtracted exactly. Full proof in Section B. □

3.3 Orthogonal block tightening

Theorem 3.6 is valid in any dimension, but can be weak and difficult to estimate for high-dimensional, non-Gaussian image sources. We therefore rotate the chosen representation and partition it into low-dimensional orthogonal blocks. Squared reconstruction error is preserved exactly, while each entropy calculation is performed in only k dimensions.

Let $U \in \mathbb{R}^{m\times m}$ be orthogonal, let $N_k := m/k$ for $k \mid m$, let $P_b \in \mathbb{R}^{k\times m}$ select block $b \in \{1, \ldots, N_k\}$, and set $\Pi_b := \sqrt{N_k} P_b U$. Each block carries independent dither of scale s as in Definition 3.5, so the entropies below depend on s as well as k; we suppress s in the notation. Define the block-averaged entropy

$$\bar{h}_U(k) := \frac{1}{N_k} \sum_{b=1}^{N_k} h\big(\tilde{z}_i^{\Pi_b \Psi}\big), \qquad \tilde{z}_i^{\Pi_b \Psi} := \Pi_b \Psi(x_i) + W_{i,b}^{(k)}. \tag{6}$$

Block size introduces a trade-off: smaller blocks tend to carry more entropy per coordinate but each pays the full leakage γ_i, a factor $2^{-2\gamma_i/k}$, while larger blocks dilute this charge at the cost of lower entropy. The following corollary optimises k over a pre-specified grid.

Corollary 3.7 (Orthogonal-block instantiation). *Let Assumptions 3.1–3.2 hold, let $\Psi : \mathbb{R}^d \to \mathbb{R}^m$ be admissible with $\operatorname{tr}\Sigma_\Psi > 0$, let k divide m, and let $U \in \mathbb{R}^{m\times m}$ be orthogonal, with Π_b and $\bar{h}_U(k)$ as in equation 6. For every $s > 0$ and every possibly randomised $\hat{z}_i^\Psi = g(\theta_{0:T}, X^{-i}) \in \mathbb{R}^m$, writing now $\nu_\Psi^2 := ks^2/\operatorname{tr}\Sigma_\Psi$,*

$$\rho_\Psi(g)^2 \geq \left(\frac{k}{2\pi e \operatorname{tr}\Sigma_\Psi} 2^{2(\bar{h}_U(k)-\gamma_i)/k} - \nu_\Psi^2 \right)_+, \tag{7}$$

and hence, over any grid $\mathcal{G}$ of pairs (k, ν_Ψ^2) with $k \mid m$,

$$\rho_\Psi(g) \geq \rho_\Psi^\star := \max_{(k,\nu_\Psi^2)\in\mathcal{G}} \left(\frac{k}{2\pi e \operatorname{tr}\Sigma_\Psi} 2^{2(\bar{h}_U(k)-\gamma_i)/k} - \nu_\Psi^2 \right)_+^{1/2}. \tag{8}$$

The ambient case $\Psi = \mathrm{id}$ and $m = d$ gives the pixel bound $\rho^\star$; taking $\Psi = \Psi_{\mathrm{LPIPS}}$ gives the perceptual bound $\rho^\star_{\Psi_{\mathrm{LPIPS}}}$.

Proof sketch. Each $\Pi_b \circ \Psi$ is admissible, so Theorem 3.6 applies blockwise to $\Pi_b \hat{z}_i^\Psi$. The maps Π_b preserve squared norm on average, so averaging the N_k block bounds returns the error in $\mathbb{R}^m$. Jensen's inequality for convex $t \mapsto 2^{2t/k}$ replaces the average of block exponentials by the exponential of $\bar{h}_U(k)$, while the dither, removed exactly per block, contributes ks^2 once. Each pair (k, ν_Ψ^2) is separately valid, hence so is the maximum over the pre-specified grid. Proof in Section C. □

We take $U = m^{-1/2} H_m V^\top$, where V is an eigenbasis of Σ_Ψ and H_m is a Hadamard matrix, so that every block carries the same variance $\operatorname{tr}\Sigma_\Psi$ and every coordinate mixes all eigendirections equally; see Remark C.1. Halving k typically raises the entropy per coordinate but doubles the leakage charge γ/k; doubling k has the opposite effect. The grid therefore selects the scale at which the entropy gain outweighs the per-coordinate leakage penalty, and the dither cost ν^2 is optimised jointly with k.

4 The a priori accountant

We next supply the mechanism-side input to Theorem 3.6. For Poisson-subsampled DP-SGD, the schedule $(q_t, \sigma_t)_{t=1}^T$ upper-bounds the per-sample leakage $\gamma = I(x_i; \theta_{0:T} \mid X^{-i})$. The calculation uses neither the trained parameters nor the realised gradients, and does not take a reported (ε, δ) guarantee as input. It is independent of the parameter dimension p in value and cost.

At step t of DP-SGD (Abadi et al., 2016), a Poisson-subsampled minibatch B_t includes each index independently with probability q_t. Per-sample gradients are clipped to L_2-norm C,

$$\bar{g}_t(x_j) := \nabla_\theta \ell(\theta_{t-1}; x_j) \min\left\{1, \frac{C}{\|\nabla_\theta \ell(\theta_{t-1}; x_j)\|}\right\},$$

and, with learning rate η_t, noise multiplier σ_t, and expected batch size $b_t = q_t n$, the released parameter trajectory evolves as

$$\theta_t = \theta_{t-1} - \frac{\eta_t}{b_t}\left(\sum_{j\in B_t} \bar{g}_t(x_j) + \sigma_t C\, \xi_t\right), \qquad \xi_t \sim \mathcal{N}(0, I_p). \tag{9}$$

We write $\mathcal{H}_{t-1} := \theta_{0:t-1}$. We take $C = 1$ throughout: rescaling the update by C leaves mutual information invariant, so the bound, which depends only on (q_t, σ_t), holds for any clipping norm.

Theorem 4.1 (Dominating-pair leakage bound). *For DP-SGD with updates equation 9, hyperparameters $(q_t, \sigma_t)_{t=1}^T$, and data-independent initialisation,*

$$\gamma \leq \frac{1}{\ln 2} \sum_{t=1}^T K_t^{\mathrm{PRV}}, \qquad K_t^{\mathrm{PRV}} := D_{\mathrm{KL}}\big(P_t^{(1)} \,\|\, Q_t^{(1)}\big), \tag{10}$$

where

$$P_t^{(1)} := (1-q_t)\,\mathcal{N}(0, \sigma_t^2) + q_t\,\mathcal{N}(1, \sigma_t^2), \qquad Q_t^{(1)} := \mathcal{N}(0, \sigma_t^2).$$

The pair $(P_t^{(1)}, Q_t^{(1)})$ is the standard one-dimensional dominating pair for the Poisson-subsampled Gaussian mechanism with sampling rate q_t, noise multiplier σ_t, and unit sensitivity (Mironov et al., 2019; Zhu et al., 2022). Moreover, K_t^{PRV} is the expectation under $P_t^{(1)}$ of the privacy-loss random variable $\ln(\mathrm{d}P_t^{(1)}/\mathrm{d}Q_t^{(1)})$ used by the PRV accountant (Gopi et al., 2021).

Proof sketch. The chain rule decomposes trajectory leakage into per-step conditional mutual informations. Conditional on the past and non-target records, each update is a Bernoulli-gated Gaussian channel whose divergence is bounded by the standard one-dimensional dominating pair. Summing these bounds and converting nats to bits gives equation 10; see Section D. □

Thus the reconstruction analysis consumes DP-SGD through a single scalar information budget, obtained by summing schedule-dependent per-step charges. The factor $1/\ln 2$ converts the KL divergences in equation 10 from nats to the bits used in Theorem 3.6. Each K_t^{PRV} is a one-dimensional integral, evaluated to an error negligible relative to the finite-sample margins of Section 5 (Section G.4). The accountant can therefore be computed before training, or retrospectively from reported sampling rates and noise multipliers.

5 A CONFIDENCE BOUND ON THE REPRESENTATION ENTROPY

Section 4 supplies the mechanism-side leakage bound. We now obtain the source-side entropy bound required by Corollary 3.7 from finite reference data. Because equation 7 is increasing in the entropy it uses, a lower confidence bound on $\bar{h}_U(k)$ gives a valid, conservative reconstruction bound.

By Definition 3.5, each $\tilde{z}_i^{\Pi_b\Psi}$ is a finitely supported source convolved with Gaussian dither of a chosen scale s. We use the plug-in estimator of Goldfeld et al. (2020), namely the entropy of the empirical mixture. For a sample $x_{1:n}$, define

$$\hat{f}_n(y;\Pi_b) := \frac{1}{n}\sum_{j=1}^{n}\phi_s\big(y-\Pi_b\Psi(x_j)\big), \qquad \hat{h}_n(\Pi_b) := h\big(\hat{f}_n(\cdot;\Pi_b)\big), \qquad \hat{h}_n := \frac{1}{N_k}\sum_{b=1}^{N_k}\hat{h}_n(\Pi_b), \tag{11}$$

where ϕ_s is the k-dimensional dither density. We evaluate this entropy using the Monte Carlo integrator of Goldfeld et al. (2020) with n_{MC} dither replicates per centre and denote the result by $\hat{h}_{n,\mathrm{MC}}$. The estimator is minimax rate-optimal for this estimand, and its bias is downward (Goldfeld et al., 2020, Eq. (70)), which is the conservative direction required here. The finite-sample margins below are new. Throughout, n denotes the size of the sample used for entropy estimation, which need not be the training set (see Section F).

The confidence statement is simultaneous over the candidate block sizes and dither scales, so the final pair may be selected after estimation without invalidating coverage.

Proposition 5.1 (Confidence bound on the representation entropy)**.** *Let Assumptions 3.1–3.2 hold. Fix $\alpha \in (0,1)$, an admissible $\Psi : \mathbb{R}^d \to \mathbb{R}^m$ with $\operatorname{tr}\Sigma_\Psi > 0$, and the grid $\mathcal{G}$ of equation 8. Let U and the reference trace $\operatorname{tr}\hat{\Sigma}_\Psi$ be fixed using held-out data, independently of the estimation sample $x_{1:n}$, as is Ψ, and for each $(k,\nu^2) \in \mathcal{G}$ set $s = \nu(\operatorname{tr}\hat{\Sigma}_\Psi/k)^{\frac{1}{2}}$. For the closed-form margins $t_{\mathrm{MC}}, t_{\mathrm{MD}} > 0$ defined in Section E.3, and $\hat{h}_{n,\mathrm{MC}}$, t_{MC} and $\bar{h}_U(k)$ evaluated at the dither scale s of each pair, with probability at least $1-\alpha$ over the i.i.d. estimation sample $x_{1:n} \sim P_x$ and Monte Carlo noise, given $D_\perp$,*

$$\bar{h}_{\mathrm{LB}}(k) := \hat{h}_{n,\mathrm{MC}} - t_{\mathrm{MC}} - t_{\mathrm{MD}} \le \bar{h}_U(k) \qquad \text{for all } (k,\nu^2) \in \mathcal{G}. \tag{12}$$

On the same event, every reconstruction admitted by the threat model in Section 3.1 satisfies

$$\mathbb{E}\big[\|z_i^\Psi - \hat{z}_i^\Psi\|^2\big] \ge \max_{(k,\nu^2)\in\mathcal{G}}\left(\frac{k}{2\pi e}2^{2(\bar{h}_{\mathrm{LB}}(k)-\gamma)/k} - ks^2\right)_+, \tag{13}$$

and, if in addition $\operatorname{tr}\hat{\Sigma}_\Psi \ge \operatorname{tr}\Sigma_\Psi$,

$$\rho_\Psi(g) \ge \max_{(k,\nu^2)\in\mathcal{G}}\left(\frac{k}{2\pi e\,\operatorname{tr}\hat{\Sigma}_\Psi}2^{2(\bar{h}_{\mathrm{LB}}(k)-\gamma)/k} - \nu^2\right)_+^{1/2}. \tag{14}$$

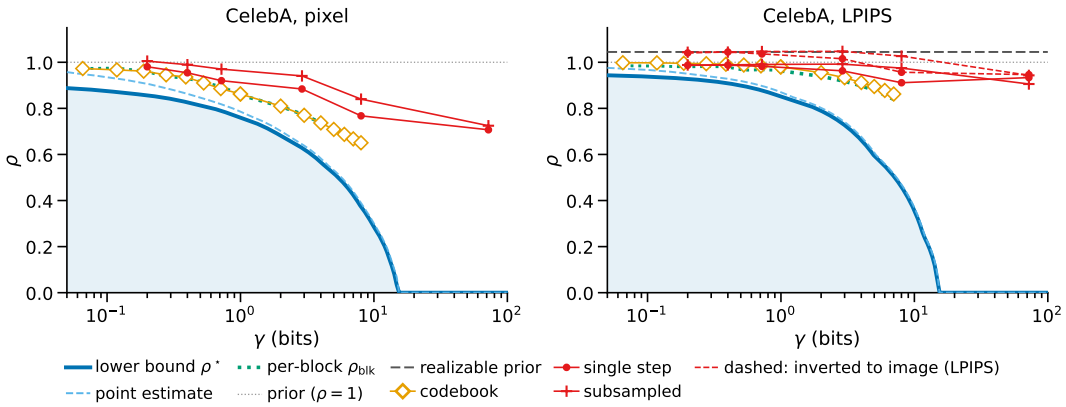


Figure 3: **CelebA: bounds, attacks and slack decomposition.** Pixel (left) and LPIPS (right) reconstruction ratios at leakage γ. The bound $\rho^\star$, block-specific bound ρ_{blk}, achieved codebook-release error ρ_{cb}, and attack errors against a single full-batch noisy-gradient release and a 100-step subsampled DP-SGD trajectory are shown. The large $\rho^\star$-to-ρ_{blk} gap is caused by charging every block the full leakage; the close ρ_{blk}-to-ρ_{cb} match shows that little slack remains in the rate–distortion step. Details are in Section G.6.

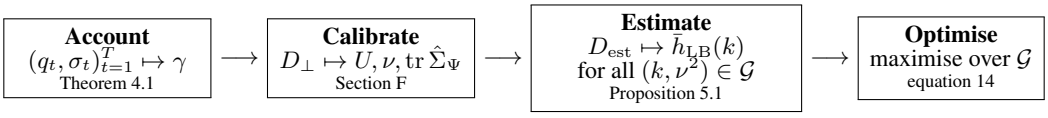
Proof sketch. Cantelli's inequality controls the Monte Carlo error, McDiarmid's inequality controls the sampling error, and the downward plug-in bias requires no margin. A union bound over $\mathcal{G}$ permits post-estimation selection of (k, ν^2); see Section E. □

Both entropy margins are finite-sample (Section E.3). For normalisation we replace $\operatorname{tr}\Sigma$ by its held-out estimate $\operatorname{tr}\hat{\Sigma}$; dividing instead by its finite-sample upper confidence bound lowers $\rho^\star$ by at most 0.03, except on ImageNette, whose $D_\perp$ is small (Table 3). The underlying absolute-distortion bound does not involve $\operatorname{tr}\Sigma$ and keeps its finite-sample coverage.

6 Experiments

We ask two questions: what reconstruction protection does a realistic DP-SGD run guarantee at a given utility cost, and what explains the gap between this guarantee and a strong reconstruction attack? Figure 1 gives the conceptual framework; the pipeline below shows how we compute its components in both experiments.

Computing the bound.



Datasets and representations. We evaluate equation 14, the finite-sample instantiation of Corollary 3.7 obtained using Proposition 5.1. The main text reports representative pixel-space and LPIPS results, while the appendix extends the evaluation across MNIST, CIFAR-10, CelebA and ImageNette (Sections G.1 and G.3). Both representations use the same schedule leakage γ; only their source entropy and prior-only error differ. Deployment results are reported against (ε, δ), with $\delta = n_{\mathrm{train}}^{-1.1}$. Further experimental and data-handling details appear in Sections F and G.

Protection and utility. For each target ε, we calibrate σ, compute the resulting leakage γ and bound $\rho^\star_\Psi$, and train a model under the same DP-SGD schedule. We denote the pixel and LPIPS bounds by $\rho^\star_{\mathrm{px}}$ and $\rho^\star_{\Psi_{\mathrm{LPIPS}}}$. Thus Figures 2 and 5 pair the reconstruction bound computable before training with the test accuracy subsequently achieved by each schedule. On CelebA, $\rho^\star_{\mathrm{px}}$ remains above 0.5 up to $\varepsilon = 16$ ($\rho^\star_{\mathrm{px}} = 0.513$), where accuracy is $96.1 \pm 0.3\%$, one point below the non-private baseline; $\rho^\star_{\Psi_{\mathrm{LPIPS}}}$ is higher there ($\rho^\star_{\Psi_{\mathrm{LPIPS}}} = 0.588$). At $\varepsilon = 8$, $\rho^\star_{\mathrm{px}}$ lies between 0.68 and

0.75 across all four datasets. The bounds remain non-trivial for $\varepsilon > 10$, but weaken as leakage approaches the source uncertainty certifiable from finite data. We report against ε for familiarity; equal (ε, δ) does not imply equal γ or equal reconstruction protection.

Attack comparison. We use a single reconstruction attack throughout: the prior-aware attack of Hayes et al. (2023), applied to the linear model of Schwethelm et al. (2024), whose per-example gradient is the input (Section G.5). The attack uses all non-target records and released updates and estimates the source prior from a shadow sample excluding the target. At matched leakage γ, we apply the attack to a one-step full-batch release, $(q, T) = (1, 1)$, and a 100-step subsampled trajectory, $(q, T) = (0.02, 100)$. These are the one-step and multi-step series in Figures 3 and 7. Both mean errors remain above the same bound; the subsampled release is harder to exploit because target participation is hidden. Individual targets may fall below the bound because it bounds population-average expected error (Figure 6). Numerical results appear in Table 6.

Where does the slack arise? Figure 3 gives a clear answer: the principal identifiable looseness comes from charging every block the full leakage, while little comes from the rate–distortion step. To show this, at leakage $R = \gamma$ we compare DP-SGD with a reconstruction-optimised codebook release that indexes the nearest of 2^R representative points; each bit released by this codebook is much more helpful for reconstruction than bits leaked from DP-SGD. We need two codebooks, one for rates above one bit and one for sub-bit rates (Section G.6). We evaluate the chain: $\rho^\star \leq \rho_{\mathrm{blk}} \leq \rho_{\mathrm{cb}} \leq \rho_{\mathrm{dpsgd}}$. Here $\rho^\star$ is the bound, ρ_{blk} charges each block only its actual leakage, and ρ_{cb} and ρ_{dpsgd} are attack errors against the codebook and DP-SGD releases. The first gap is 0.10–0.33 in ρ, whereas the second is at most 0.03, showing that the rate–distortion step is close to tight after correcting the block charge. The remaining gap is mechanism-specific, though it may also include accountant and attack looseness. Separately, we also introduce some slack by taking a confidence bound; what we lose by not using the entropy point estimate is shown in Figure 3. See Section G.6 for details.

7 Conclusion

We introduced a modular framework that lower-bounds population-average reconstruction error under an informed adversary. For DP-SGD it combines a schedule-level leakage bound with a finite-sample source-entropy bound in an analyst-chosen representation. Across the image datasets studied, the pixel and LPIPS bounds remain non-trivial across commonly used privacy regimes, including settings with $\varepsilon > 10$.

Limitations. The guarantee is in expectation under the source and does not preclude more accurate reconstruction of particular records or subpopulations. The bound has looseness, nearly all looseness can be attributed to the full leakage penalty being charged to every block and not accounting for the fact that some released bits are more helpful for reconstruction than others. Observing the full noisy trajectory directly covers trajectory-exposing settings but is conservative when only the final model is released. The theory also excludes cross-record preprocessing and auxiliary information beyond the stated threat model.

Future work. The modular construction allows mechanism-specific leakage bounds, source-specific converses, other distortion measures, and representations targeting particular harms. Conditioning leakage and source uncertainty on auxiliary information could model better-informed adversaries, while joint source and leakage terms may yield group-reconstruction guarantees. Finally, using unreleased iterates or model-specific gradient structure could tighten a deployment-specific guarantee at the cost of architecture independence.

AI USE STATEMENT

In this work, we only used generative AI tools (LLM-based chat and coding assistants) for the following tasks: assisting with the mathematical development, providing feedback on research methodology, as a writing aid and writing and running code. The mathematical assistance consisted of checking proofs for errors, suggesting intermediate steps in proofs, making proofs more formal, keeping notation consistent and general brainstorming. All mathematical statements were thoroughly checked multiple times and line-by-line by the authors. As a writing aid it was used to help shorten text and increase clarity. We used AI to help write much of the code. The authors checked the code by eye and employed sanity checks, plots and tests during development to make sure that the code is valid. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is intended as a tool to help practitioners choose and assess privacy protections, by bounding how accurately a training record can be reconstructed from a model trained with DP-SGD. To test the bound we run reconstruction attacks, adapted from the prior-aware attack of Hayes et al. (2023), against a deliberately simple model on public benchmark datasets.

REPRODUCIBILITY STATEMENT

Algorithm 1 gives the full procedure for computing the bound, and Section F specifies the data splits it uses. Section G gives all remaining experimental details: dataset preprocessing, the parameter grid and estimator settings, the architectures, hyperparameters and noise calibration for the DP-SGD runs, and the construction of the attack and the codebook mechanism. All datasets are public. Anonymised code for all experiments is provided in the supplementary material `https://anonymous.4open.science/r/Pixels-to-Privacy-Sensitive-Features`

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1138–1156. IEEE, 2022.

Thomas B. Berrett, Richard J. Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *The Annals of Statistics*, 47(1):288–318, 2019. doi: 10.1214/18-AOS1688.

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *USENIX Security Symposium*, 2020. URL `https://api.semanticscholar.org/CorpusID:229156229`.

Kamalika Chaudhuri, Chuan Guo, Laurens van der Maaten, Saeed Mahloujifar, and Mark Tygert. Guarantees of confidentiality via hammersley-chapman-robbins bounds. *Trans. Mach. Learn. Res.*, 2024, 2024. URL `https://api.semanticscholar.org/CorpusID:268876369`.

Thomas M Cover and Joy A Thomas. Elements of information theory. 2006.

Paul W. Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016. URL https://api.semanticscholar.org/CorpusID:9204999.

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

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12(3):793–815, 1984. doi: 10.1214/aos/1176346703.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet (eds.), *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 535–544. PMLR, 2018. URL https://proceedings.mlr.press/v75/feldman18a.html.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? *ArXiv*, abs/2003.14053, 2020. URL https://api.semanticscholar.org/CorpusID:214728347.

Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020.

Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.

Patricia Guerra-Balboa, Annika Sauer, Héber H Arcolezi, and Thorsten Strufe. Understanding disclosure risk in differential privacy with applications to noise calibration and auditing (extended version). *arXiv preprint arXiv:2603.12142*, 2026.

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*, pp. 8056–8071. PMLR, 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*, pp. 11998–12011. PMLR, 2023.

Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35:22911–22924, 2022.

Awni Y. Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. *ArXiv*, abs/2102.11673, 2021. URL https://api.semanticscholar.org/CorpusID:232013768.

Jamie Hayes, Borja Balle, and Saeed Mahloujifar. Bounding training data reconstruction in dp-sgd. *Advances in neural information processing systems*, 36:78696–78722, 2023.

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

Tobias Koch. The shannon lower bound is asymptotically tight. *IEEE Transactions on Information Theory*, 62(11):6155–6161, 2016.

L. F. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23(2):95–101, 1987.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Bogdan Kulynych, Juan Gomez, Georgios Kaissis, Jamie Hayes, Borja Balle, Flavio Calmon, and Jean Raisaro. Unifying re-identification, attribute inference, and data reconstruction risks in differential privacy. *Advances in Neural Information Processing Systems*, 38:111967–112003, 2026.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Kiwan Maeng, Chuan Guo, Sanjay Kariyappa, and G Edward Suh. Bounding the invertibility of privacy-preserving instance encoding using fisher information. *Advances in Neural Information Processing Systems*, 36:51904–51925, 2023.

Mikko I Malinen and Pasi Fränti. Balanced k-means for clustering. In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*, pp. 32–41. Springer, 2014.

Tudor Manole and Aaditya Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 69(7):4641–4658, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550. IEEE, 2018.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2025.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pp. 26–30. IEEE, 2016.

Curt Schieler and Paul W. Cuff. Rate-distortion theory for secrecy systems. *2013 IEEE International Symposium on Information Theory*, pp. 2219–2223, 2013. URL `https://api.semanticscholar.org/CorpusID:6970765`.

Kristian Schwethelm, Johannes Kaiser, Moritz Knolle, Sarah Lockfisch, Daniel Rueckert, and Alexander Ziller. Visual privacy auditing with diffusion models. *arXiv preprint arXiv:2403.07588*, 2024.

Marika Swanberg, Meenatchi Sundaram Muthu Selva Annamalai, Jamie Hayes, Borja Balle, and Adam Smith. Beyond the worst case: Extending differential privacy guarantees to realistic adversaries. *arXiv preprint arXiv:2507.08158*, 2025.

Qi Tan, Qi Li, Yi Zhao, Zhuotao Liu, Xiaobing Guo, and Ke Xu. Defending against data reconstruction attacks in federated learning: An information theory approach. In *33rd USENIX security symposium (USENIX security 24)*, pp. 325–342, 2024.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Hanshen Xiao and Srinivas Devadas. Pac privacy: Automatic privacy measurement and control of data processing. In *Annual International Cryptology Conference*, pp. 611–644. Springer, 2023.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:195316471.

Yuqing Zhu, Jinshuo Dong, and Yu-Xiang Wang. Optimal accounting of differential privacy via characteristic function. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 4782–4817. PMLR, 2022. URL https://proceedings.mlr.press/v151/zhu22c.html.

A Extended related work

Here we provide further details on related reconstruction bound work, this literature is helpfully summarized by Table 1 which organises each distinct lineage (ReRo and expected distortion) in order of publication date.

The information quantities in the ReRo literature. Two threshold entries consume an information quantity, and neither is γ_i. Xiao & Devadas (2023) consume the dataset-level $I(X;\mathcal{M}(X))$, computed after the fact by simulating the mechanism, and in their individual form a conditional mutual information given one other user's record rather than the leave-one-out conditioning on all of X^{-i} that the informed adversary requires. Guo et al. (2023) consume the Rényi mutual information $I_\alpha(X;Y)$ bounded by the RDP parameter, together with the Shannon entropy of a supplied probability vector, spent through a generalised Fano inequality on the probability of exact recovery rather than on distortion.

Expected-error bounds. Both bounds of Guo et al. (2022) carry an explicit bias-variance trade-off, but their deployable forms, $\mathrm{MSE} \geq \sum_i \mathrm{diam}_i(\mathcal{Z})^2/(4d(e^\varepsilon - 1))$ from RDP, and $\mathrm{MSE} \geq d/\operatorname{tr} I_h(z)$ from Fisher information, are the unbiased case, and the consequences of this assumption are visible in their own experiments. On $[0,1]^{784}$ an adversary that ignores the model and predicts the midpoint of each coordinate attains $\mathrm{MSE} \leq 1/4$, well inside the $\mathrm{MSE} = 1$ line they draw as "perfect privacy". Their per-sample Fisher bounds all exceed 1, so the accounting assures every training sample as safe. Conversely, the informed-adversary attack of Balle et al. (2022) falls below the Fisher bound on some samples, which they attribute to the attack incurring "a high bias when the sample is hard to reconstruct".

The van Trees bound of Maeng et al. (2023) is the closest in form to ours as its prior distribution is estimated from a sample, and it needs a smooth density, so they perturb the images with Gaussian noise before score matching. The differences are in the direction of the estimate and where the smoothing lives. Theirs is a point estimate, which they note "can be incorrect" and should be read "as a guideline", and which is breached by their own attack when that perturbation is small, whereas ours is a one-sided confidence bound whose estimator is biased in the safe direction. Their perturbation is applied to the inputs themselves, where our dither is internal to the analysis and removed exactly (Theorem 3.6).

B PROOF OF THE RECONSTRUCTION LOWER BOUND

B.1 PROOF OF LEMMA 3.4

Proof. The first inequality is the Data Processing Inequality (DPI) on $x_i \to (\theta, X^{-i}) \to \hat{u}_i$, valid since $\hat{u}_i$ depends on x_i only through (θ, X^{-i}) and randomness independent of x_i given it. For the equality, recall the chain rule for mutual information (Cover & Thomas, 2006, Theorem 2.5.2): for any random variables A, B, C we have $I(A;\, B, C) = I(A;\, C) + I(A;\, B \mid C)$. Applying this with $A = x_i$, $B = \theta$, and $C = X^{-i}$ gives $I(x_i;\, \theta, X^{-i}) = I(x_i;\, X^{-i}) + I(x_i;\, \theta \mid X^{-i})$, and $I(x_i;\, X^{-i}) = 0$ under Assumption 3.2 leaves $I(x_i;\, \theta, X^{-i}) = I(x_i;\, \theta \mid X^{-i}) =: \gamma_i$. □

B.2 REGULARITY OF THE DITHERED REPRESENTATION

Lemma B.1 (Regularity of the dithered representation)**.** *Let Assumption 3.1 hold, let $\Psi : \mathbb{R}^d \to \mathbb{R}^m$ be admissible (conditioning on its draw where Ψ is random), let $s > 0$, and let $\tilde{z}_i^{\,\Psi} = \Psi(x_i) + W_i^{(m)}$ be as in Definition 3.5. Write ϕ_s for the $\mathcal{N}(0, s^2 I_m)$ density. Then*

(i) $\tilde{z}_i^{\,\Psi}$ has density $f_\Psi(y) = \sum_{v \in \mathcal{Q}} \Pr(x_i = v)\, \phi_s\big(y - \Psi(v)\big)$, a finite Gaussian mixture;

(ii) $\mathrm{Cov}(\tilde{z}_i^{\,\Psi}) = \Sigma_\Psi + s^2 I_m$, and $\mathbb{E}\|\tilde{z}_i^{\,\Psi}\|^2 < \infty$;

(iii) $h(\tilde{z}_i^{\,\Psi})$ is finite and $h(\tilde{z}_i^{\,\Psi}) = h(W^{(m)}) + I\big(\Psi(x_i);\, \tilde{z}_i^{\,\Psi}\big) \geq h(W^{(m)}) = \frac{m}{2} \log_2(2\pi e s^2)$.

Proof. (i). With probability $\Pr(x_i = v)$ the representation is $\Psi(v) + W_i^{(m)}$, whose density is $\phi_s(\cdot - \Psi(v))$, so $\tilde{z}_i^{\,\Psi}$ has the mixture density f_Ψ.

(ii). Covariances of independent summands add, and $\mathrm{Cov}(W_i^{(m)}) = s^2 I_m$. Expanding $\mathbb{E}\|\Psi(x_i) + W_i^{(m)}\|^2$, the cross term vanishes since $W_i^{(m)}$ is centred and independent of $\Psi(x_i)$, and $\mathbb{E}\|W_i^{(m)}\|^2 = m s^2$; $\mathbb{E}\|\Psi(x_i)\|^2$ is finite as $\Psi(x_i)$ takes finitely many values.

(iii). Each $\phi_s(\cdot - \Psi(v))$ is at most $(2\pi s^2)^{-m/2}$ and the weights in (i) sum to one, so $f_\Psi \leq (2\pi s^2)^{-m/2}$ everywhere; hence $-\log_2 f_\Psi \geq \frac{m}{2} \log_2(2\pi s^2)$ wherever $f_\Psi > 0$ and $h(\tilde{z}_i^{\,\Psi}) \geq \frac{m}{2} \log_2(2\pi s^2) > -\infty$. Gaussian maximum entropy at the covariance of (ii) (Cover & Thomas, 2006, Thm. 8.6.5) gives $h(\tilde{z}_i^{\,\Psi}) < \infty$. Since $\Psi(x_i)$ is discrete and $W_i^{(m)}$ is Gaussian and independent of it, $h(\tilde{z}_i^{\,\Psi}) = h(W^{(m)}) + I(\Psi(x_i); \tilde{z}_i^{\,\Psi})$ (Goldfeld et al., 2020, eq. (53)), and the inequality follows from non-negativity of mutual information. □

B.3 PROOF OF THEOREM 3.6

Proof of theorem 3.6. If $\mathbb{E}\big[\|z_i^{\,\Psi} - \hat{z}_i^{\,\Psi}\|^2\big] = \infty$ the bound holds trivially, so assume throughout that it is finite. Since $z_i^{\,\Psi} = \Psi(x_i)$ is bounded ($\mathcal{Q}$ is finite by Assumption 3.1), this gives $\mathbb{E}\|\hat{z}_i^{\,\Psi}\|^2 < \infty$, and hence $\mathbb{E}\|z_i^{\,\Psi} - \hat{z}_i^{\,\Psi}\| < \infty$ by Cauchy–Schwarz.

Fix an index $i \in \{1, \ldots, n\}$ and, where Ψ is random, work conditionally on its draw.

Step 1: Rate bound. For any (possibly randomised) reconstruction $\hat{z}_i^{\,\Psi} = g(\theta, X^{-i})$, applying Lemma 3.4 at $\hat{u}_i = \hat{z}_i^{\,\Psi}$ gives

$$I(x_i;\, \hat{z}_i^{\,\Psi}) \;\leq\; \gamma_i. \tag{15}$$

Step 2: Markov chain through the dithered representation. Recall $\tilde{z}_i^{\,\Psi} = \Psi(x_i) + W_i^{(m)}$ (Definition 3.5). Conditionally on x_i, the variable $\tilde{z}_i^{\,\Psi}$ is a function of $W_i^{(m)}$ alone (working, where Ψ is random, conditionally on its draw, which admissibility permits since that draw is independent of the data, of $\mathcal{A}$, and of the attack), while $\hat{z}_i^{\,\Psi} = g(\theta, X^{-i})$ is a function of (θ, X^{-i}) and the attack's internal randomness. By Definition 3.5 the dither $W_i^{(m)}$ is drawn independently of all of these jointly, so

$\hat{z}_i^{\Psi} \perp\!\!\!\perp \tilde{z}_i^{\Psi} \mid x_i$, i.e. $\tilde{z}_i^{\Psi} \to x_i \to \hat{z}_i^{\Psi}$ is a Markov chain, and the data-processing inequality (Cover & Thomas, 2006, Theorem 2.8.1) with equation 15 gives

$$I(\tilde{z}_i^{\Psi};\, \hat{z}_i^{\Psi}) \;\le\; I(x_i;\, \hat{z}_i^{\Psi}) \;\le\; \gamma_i. \tag{16}$$

Now we have bounded the mutual information between the two terms whose expected MSE we would like to bound, we can use the distortion-rate bound from equation 2,

$$\mathbb{E}\big[\|\tilde{z}_i^{\Psi} - \hat{z}_i^{\Psi}\|^2\big] \;\ge\; D_{\tilde{z}_i^{\Psi}}(\gamma_i). \tag{17}$$

Step 3: Shannon lower bound. By Lemma B.1, $\tilde{z}_i^{\Psi}$ is absolutely continuous on $\mathbb{R}^m$ with finite differential entropy and finite second moment, so the SLB applies at rate $R = \gamma_i$:

$$D_{\tilde{z}_i^{\Psi}}(\gamma_i) \;\ge\; \frac{m}{2\pi e}\, 2^{\,2(h(\tilde{z}_i^{\Psi}) - \gamma_i)/m}. \tag{18}$$

Step 4: Exact removal of the dither. Since $\tilde{z}_i^{\Psi} - \hat{z}_i^{\Psi} = (z_i^{\Psi} - \hat{z}_i^{\Psi}) + W_i^{(m)}$, expanding the square gives

$$\mathbb{E}\big[\|\tilde{z}_i^{\Psi} - \hat{z}_i^{\Psi}\|^2\big] \;=\; \mathbb{E}\big[\|z_i^{\Psi} - \hat{z}_i^{\Psi}\|^2\big] \;+\; 2\,\mathbb{E}\big[\langle z_i^{\Psi} - \hat{z}_i^{\Psi},\, W_i^{(m)}\rangle\big] \;+\; \mathbb{E}\big[\|W_i^{(m)}\|^2\big].$$

By Definition 3.5 the dither $W_i^{(m)}$ is independent of $z_i^{\Psi} - \hat{z}_i^{\Psi}$ and has zero mean. So, expanding coordinatewise gives $\mathbb{E}\langle z_i^{\Psi} - \hat{z}_i^{\Psi},\, W_i^{(m)}\rangle = \sum_{j=1}^{m} \mathbb{E}[(z_i^{\Psi} - \hat{z}_i^{\Psi})_j]\,\mathbb{E}[W_{i,j}^{(m)}] = 0$, the factorisation valid since $\mathbb{E}\|z_i^{\Psi} - \hat{z}_i^{\Psi}\| < \infty$ by the assumption at the start of the proof. The last term is ms^2 exactly. Hence

$$\mathbb{E}\big[\|z_i^{\Psi} - \hat{z}_i^{\Psi}\|^2\big] \;=\; \mathbb{E}\big[\|\tilde{z}_i^{\Psi} - \hat{z}_i^{\Psi}\|^2\big] \;-\; ms^2, \tag{19}$$

Substituting equation 17 and equation 18 into equation 19, and noting the left-hand side is non-negative, proves equation 4.

Ratio form. Assume $\operatorname{tr}\Sigma_{\Psi} > 0$, i.e. $\Psi(x_i)$ is not a.s. constant, so that ρ_{Ψ} of equation 3 is defined. Dividing the inequality of Step 4 — before the positive part is taken — through by $\mathrm{RMSE}_{\mathrm{prior}}(\Psi)^2 = \operatorname{tr}\Sigma_{\Psi} > 0$ and using $\nu_{\Psi}^2 = ms^2/\operatorname{tr}\Sigma_{\Psi}$ from equation 3 gives

$$\rho_{\Psi}(g)^2 \;\ge\; \frac{m}{2\pi e\,\operatorname{tr}\Sigma_{\Psi}}\, 2^{\,2(h(\tilde{z}_i^{\Psi}) - \gamma_i)/m} \;-\; \nu_{\Psi}^2,$$

the left-hand side being $\rho_{\Psi}(g)^2$ by equation 3 and the reduction at the start of the proof. Since $\rho_{\Psi}(g)^2 \ge 0$, the two combine to give the positive part in equation 5.

□

C Proof of Corollary 3.7

Proof of Corollary 3.7. Fix an orthogonal matrix U. Fix a pair $(k, \nu_{\Psi}^2) \in \mathcal{G}$, where k divides m, and let $s = \nu_{\Psi}\sqrt{\operatorname{tr}\Sigma_{\Psi}/k}$ be the corresponding dither scale.

Step 1: Per-block bound. Given any reconstruction $\hat{z}_i^{\Psi} = g(\theta, X^{-i}) \in \mathbb{R}^m$, the induced block estimate $\hat{z}_i^{\Pi_b\Psi} := \Pi_b \hat{z}_i^{\Psi}$ relies only on (θ, X^{-i}) and the attack's internal randomness. The composite $\Pi_b \circ \Psi : \mathbb{R}^d \to \mathbb{R}^k$ is admissible in the sense of Definition 3.5: Π_b is deterministic, so the only randomness in $\Pi_b \circ \Psi$ is that of Ψ, which admissibility of Ψ already places independent of X, of $\mathcal{A}$ and of the attack. Theorem 3.6 therefore applies at the representation $\Pi_b \circ \Psi$ with m replaced by k, at the same leakage γ_i:

$$\mathbb{E}\big[\|\Pi_b(\Psi(x_i) - \hat{z}_i^{\Psi})\|^2\big] \;\ge\; \frac{k}{2\pi e}\, 2^{\,2(h(\tilde{z}_i^{\Pi_b\Psi}) - \gamma_i)/k} \;-\; ks^2, \qquad \tilde{z}_i^{\Pi_b\Psi} = \Pi_b\Psi(x_i) + W_{i,b}^{(k)}. \tag{20}$$

These are N_k separate applications of the theorem, each block picking fresh i.i.d. dither, so no joint condition across the $W_{i,b}^{(k)}$ is required. They are combined below by summation alone.

Step 2: Transfer to $\mathbb{R}^m$. The P_b group the m rows of U into N_k disjoint blocks of k, so by Pythagoras $\sum_b \|P_b u\|^2 = \|u\|^2$ for every $u \in \mathbb{R}^m$. Taking $u = Uv$ and using $\|Uv\| = \|v\|$, as U is orthogonal,

$$\frac{1}{N_k}\sum_{b=1}^{N_k} \|\Pi_b v\|^2 \;=\; \sum_{b=1}^{N_k} \|P_b U v\|^2 \;=\; \|v\|^2 \qquad \text{for every } v \in \mathbb{R}^m, \tag{21}$$

the factor N_k in $\Pi_b = \sqrt{N_k}\, P_b U$ being what turns the partition sum into an average. Applying equation 21 at $v = \Psi(x_i) - \hat{z}_i^{\,\Psi}$ and taking expectations,

$$\mathbb{E}\big[\|\Psi(x_i) - \hat{z}_i^{\,\Psi}\|^2\big] \;=\; \frac{1}{N_k}\sum_{b=1}^{N_k} \mathbb{E}\big[\|\Pi_b(\Psi(x_i) - \hat{z}_i^{\,\Psi})\|^2\big]. \tag{22}$$

Step 3: Averaging the block bounds. Substituting equation 20 into equation 22 and pulling out the factors common to all blocks,

$$\mathbb{E}\big[\|\Psi(x_i) - \hat{z}_i^{\,\Psi}\|^2\big] \;\geq\; \frac{k}{2\pi e}\, 2^{-2\gamma_i/k}\, \frac{1}{N_k}\sum_{b=1}^{N_k} 2^{\,2h(\tilde{z}_i^{\,\Pi_b\Psi})/k} \;-\; ks^2 \;\geq\; \frac{k}{2\pi e}\, 2^{\,2(\bar{h}_U(k)-\gamma_i)/k} \;-\; ks^2,$$

the last step by Jensen on the convex map $t \mapsto 2^{2t/k}$ applied to the uniform average over blocks, $\bar{h}_U(k)$ by equation 6.

Step 4: Ratio form. Dividing by $\mathrm{RMSE}_{\mathrm{prior}}(\Psi)^2 = \operatorname{tr}\Sigma_\Psi$ and using $\nu_\Psi^2 = ks^2/\operatorname{tr}\Sigma_\Psi$ gives

$$\rho_\Psi(g)^2 \;\geq\; \frac{k}{2\pi e\, \operatorname{tr}\Sigma_\Psi}\, 2^{\,2(\bar{h}_U(k)-\gamma_i)/k} \;-\; \nu_\Psi^2,$$

and since $\rho_\Psi(g)^2 \geq 0$ as well, the two combine to give the positive part in equation 7. The pair (k, ν_Ψ^2) was arbitrary and is chosen by us, not the adversary: g is fixed first, so every pair in $\mathcal{G}$ bounds the same $\rho_\Psi(g)^2$ and the largest of those bounds holds too, proving equation 8. □

Remark C.1 (Choosing the rotation)**.** *Every orthogonal U is valid; the choice affects only $\bar{h}_U(k)$, hence tightness. We write $\Psi = \mathrm{id}$; for general Ψ replace d, Σ by m, Σ_Ψ. By equation 21 the block variances $v_b := \operatorname{tr}\Sigma_{\Pi_b}$ average to $\operatorname{tr}\Sigma$ for every U, and Gaussian maximum entropy (Cover & Thomas, 2006, Thm. 8.6.5) caps each block entropy by $f(v_b) := \frac{k}{2}\log_2\big(2\pi e\,(v_b + ks^2)/k\big)$. Since f is concave, Jensen's inequality gives $N_k^{-1}\sum_b f(v_b) \leq f\big(N_k^{-1}\sum_b v_b\big) = f(\operatorname{tr}\Sigma)$, with equality exactly when every $v_b = \operatorname{tr}\Sigma$. We take $U := d^{-1/2} H_d V^\top$. By the spectral theorem for real symmetric matrices, $\Sigma = V\Lambda V^\top$ with V orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, so $V^\top x_i$ has uncorrelated coordinates with variances $\lambda_1, \ldots, \lambda_d$. The Hadamard matrix H_d satisfies $H_d H_d^\top = dI_d$ with every entry ± 1, so $d^{-1/2}H_d$ is orthogonal and mixes the uncorrelated coordinates with equal weight $1/d$, giving every output coordinate variance $\operatorname{tr}\Sigma/d$ and hence every block $v_b = \operatorname{tr}\Sigma$, for every $k \mid d$ at once. The equal-weight mixing also places each coordinate in the regime where projections of a high-dimensional source are near-Gaussian (Diaconis & Freedman, 1984), which the SLB rewards; we measure rather than assume this, through the saturation value of Remark C.2. V is computed on data held out from the estimation sample, as Proposition 5.1 requires. In our experiments V is the eigenbasis for the MNIST, CIFAR-10 and CelebA pixel bounds. For the ImageNette pixel bound ($|D_\perp| = 1{,}000 < d$) and for every LPIPS bound we take instead $V^\top = D$, a fixed random diagonal sign flip, which needs no data; the block variances then equal $\operatorname{tr}\Sigma$ only on average over the signs.*

Hadamard matrices exist only at orders 1, 2 and multiples of 4, and are known at every such order we use, by Sylvester's doubling, Paley's construction and Kronecker products of these (Section G.2). If no Hadamard matrix of order d is available we zero-pad to the nearest order at which one is: the padded coordinates are deterministically zero, so a bound for the padded representation transfers to the original, at a cost in tightness set by the padding fraction.

Remark C.2 (The bound never exceeds the prior baseline)**.** *Corollary 3.7 holds for every reconstruction g, and the prior-mean guess is one with $\rho_\Psi(g) = 1$ by equation 3, so $\rho_\Psi^\star \leq 1$ for every U, k and s: the bound never claims more protection than releasing nothing affords, and the same holds for the reported bound when it is normalised by an upper bound on $\operatorname{tr}\Sigma$, since $\bar{h}_{\mathrm{LB}}(k) \leq \bar{h}_U(k)$*

only lowers it. Equality at $\gamma = 0$ requires $2^{2\bar{h}_U(k)/k} = 2\pi e\,(\operatorname{tr}\Sigma + ks^2)/k$, which by Gaussian maximum entropy (Cover & Thomas, 2006, Thm. 8.6.5) holds only when every block is isotropic Gaussian with variance $\operatorname{tr}\Sigma$; the saturation value below 1 in Figures 2 and 5 (0.87–0.91 in pixel space across the four datasets and 0.92–0.96 in LPIPS space across the three, at $\gamma = 0$) is what the source's departure from that ideal costs, together with the dither and the estimation margins.

D PROOFS FOR THE A PRIORI ACCOUNTANT

The proof of Theorem 4.1 has three ingredients, presented in the order they are used: the chain rule splits γ over steps (Section D.1); each step is reduced to a Bernoulli-gated Gaussian channel (Section D.2); and the channel's leakage is bounded by the one-dimensional reduction of Mironov et al. (2019). Section D.3 assembles them into the final result.

D.1 TRAJECTORY CHAIN RULE

Lemma D.1 (Trajectory chain rule)**.** *With data-independent initialisation, the per-sample CMI decomposes exactly as*

$$\gamma_i \;=\; \sum_{t=1}^{T} I(x_i;\, \theta_t \mid \mathcal{H}_{t-1}, X^{-i}). \tag{23}$$

Proof. Applying the chain rule for mutual information to $\theta = \theta_{0:T}$ one iterate at a time,

$$\gamma_i \;=\; I(x_i;\, \theta_{0:T} \mid X^{-i}) \;=\; I(x_i;\, \theta_0 \mid X^{-i}) \;+\; \sum_{t=1}^{T} I(x_i;\, \theta_t \mid \mathcal{H}_{t-1}, X^{-i}).$$

The first term vanishes because θ_0 is drawn independently of X, hence of x_i given X^{-i}. □

D.2 PER-STEP CHANNEL REDUCTION

Lemma D.2 (Per-step channel reduction)**.** *Fix a step t with $\eta_t \neq 0$ and let $\Omega := (\mathcal{H}_{t-1},\, X^{-i},\, B_t \setminus \{i\})$ collect the past, the other records, and the other minibatch indices. Let $G := \bar{g}_t(x_i)$ be the target's clipped gradient and*

$$Y := \mathbf{1}_t\, G + \sigma_t \xi_t, \qquad \mathbf{1}_t := \mathbf{1}[i \in B_t] \sim \text{Bernoulli}(q_t),\; \xi_t \sim \mathcal{N}(0, I_p), \tag{24}$$

the Bernoulli-gated Gaussian channel *with input G. Then*

$$I(x_i;\, \theta_t \mid \mathcal{H}_{t-1}, X^{-i}) \;\le\; I(x_i;\, \theta_t \mid \Omega) \;=\; I(G;\, Y \mid \Omega). \tag{25}$$

Proof. Equality. Given Ω, the iterate θ_{t-1} and the other members' gradient sum $S := \sum_{j \in B_t \setminus \{i\}} \bar{g}_t(x_j)$ are determined, so the update of equation 9 reads

$$\theta_t \;=\; \theta_{t-1} - \tfrac{\eta_t}{b_t}\big(Y + S\big), \qquad \text{equivalently} \qquad Y \;=\; -\tfrac{b_t}{\eta_t}\big(\theta_t - \theta_{t-1}\big) - S,$$

so θ_t is a function of (Y, Ω) and Y a function of (θ_t, Ω). Expanding $I(x_i;\, Y, \theta_t \mid \Omega)$ by the chain rule (Cover & Thomas, 2006, Thm. 2.5.2) in both orders,

$$I(x_i;\, \theta_t \mid \Omega) + \underbrace{I(x_i;\, Y \mid \theta_t, \Omega)}_{=\,0} \;=\; I(x_i;\, Y \mid \Omega) + \underbrace{I(x_i;\, \theta_t \mid Y, \Omega)}_{=\,0},$$

each conditional term vanishing because the variable it measures is determined by its conditioning. Hence $I(x_i; \theta_t \mid \Omega) = I(x_i; Y \mid \Omega)$. Next, $G = \bar{g}_t(x_i)$ is a function of (x_i, Ω), the gradient being evaluated at θ_{t-1}, so

$$I(x_i;\, Y \mid \Omega) \;=\; I(x_i, G;\, Y \mid \Omega) \;=\; I(G;\, Y \mid \Omega) + I(x_i;\, Y \mid G, \Omega),$$

and the last term vanishes: $Y = \mathbf{1}_t G + \sigma_t \xi_t$ with $(\mathbf{1}_t, \xi_t)$ drawn independently of (x_i, Ω), so given (G, Ω) it is independent of x_i.

Inequality. Poisson subsampling draws the inclusion indicators independently of the data and the past, so $B_t \setminus \{i\} \perp x_i \mid (\mathcal{H}_{t-1}, X^{-i})$. Expanding $I(x_i;\, \theta_t, B_t \setminus \{i\} \mid \mathcal{H}_{t-1}, X^{-i})$ by the chain rule in both orders,

$$\begin{aligned}&\underbrace{I(x_i;\, B_t \setminus \{i\} \mid \mathcal{H}_{t-1}, X^{-i})}_{=0} + I(x_i;\, \theta_t \mid \Omega)\\ &= I(x_i;\, \theta_t \mid \mathcal{H}_{t-1}, X^{-i}) + \underbrace{I(x_i;\, B_t \setminus \{i\} \mid \theta_t, \mathcal{H}_{t-1}, X^{-i})}_{\geq 0},\end{aligned}$$

which is the inequality in equation 25. □

D.3 Proof of Theorem 4.1

Proof. By Lemmas D.1 and D.2,

$$\gamma \;=\; \sum_{t=1}^{T} I(x_i;\, \theta_t \mid \mathcal{H}_{t-1}, X^{-i}) \;\leq\; \sum_{t=1}^{T} I(G_t;\, Y_t \mid \Omega_t),$$

so it suffices to show $I(G_t; Y_t \mid \Omega_t) \leq K_t^{\mathrm{PRV}}$ at each step, with K_t^{PRV} as defined in equation 10. Fix t and drop it from the subscripts of G, Y and Ω.

Per-input output distribution. Recall $P_{Y|z} = (1 - q_t)P_0 + q_t P_z$ from equation 10. Conditionally on $G = z$ and $\Omega = \omega$, the only randomness remaining in $Y = \mathbf{1}_t G + \sigma_t \xi_t$ is the gate and the noise, which are independent of (x_i, Ω) — Poisson subsampling draws $\mathbf{1}_t$ independently of the data, the past and the other indices, and ξ_t is fresh — so

$$P_{Y|G=z,\,\Omega=\omega} \;=\; P_{Y|z} \qquad \text{for every } z \text{ and } \omega :$$

with probability $1 - q_t$ the target is excluded and Y is pure noise, P_0; with probability q_t it is included and Y is noise about the clipped gradient, P_z. The input G is a function of (x_i, Ω) and so is correlated with Ω, but the channel from G to Y is not: only the input distribution varies with ω. In particular P_0 is the distribution of Y when the target is not sampled.

Dropping the conditioning. Since $P_{Y|G,\Omega} = P_{Y|G}$, we have $Y \perp \Omega \mid G$, so expanding $I(G, \Omega;\, Y)$ by the chain rule in both orders,

$$I(G;\, Y) + \underbrace{I(\Omega;\, Y \mid G)}_{=0} \;=\; I(G;\, Y \mid \Omega) + \underbrace{I(\Omega;\, Y)}_{\geq 0},$$

hence $I(G; Y \mid \Omega) \leq I(G; Y)$: the right-hand side is the mutual information of the channel equation 24 at input distribution P_G, the marginal distribution of the target's clipped gradient, and conditioning on Ω only removes what Y already reveals about Ω.

Golden formula. For distributions P_X, $P_{Y|X}$, Q_Y, write $D_{\mathrm{KL}}(P_{Y|X} \,\|\, Q_Y \mid P_X) := \mathbb{E}_{x \sim P_X}\big[D_{\mathrm{KL}}(P_{Y|X=x} \,\|\, Q_Y)\big]$ for the conditional divergence. By the golden formula (Polyanskiy & Wu, 2025, Thm. 4.1), for any Q_Y,

$$D_{\mathrm{KL}}(P_{Y|G} \,\|\, Q_Y \mid P_G) \;=\; I(G;\, Y) \;+\; D_{\mathrm{KL}}(P_Y \,\|\, Q_Y) \;\geq\; I(G;\, Y). \tag{26}$$

Leave-one-out reference. Take $Q_Y = P_0$ in equation 26. Unpacking the conditional divergence by its definition and using $P_{Y|G=z} = P_{Y|z}$,

$$I(G;\, Y) \;\leq\; D_{\mathrm{KL}}(P_{Y|G} \,\|\, P_0 \mid P_G) \;=\; \mathbb{E}_{z \sim P_G}\big[D_{\mathrm{KL}}(P_{Y|z} \,\|\, P_0)\big] \;\leq\; \sup_{\|z\| \leq 1} D_{\mathrm{KL}}(P_{Y|z} \,\|\, P_0), \tag{27}$$

the last step because every value G can take has norm at most one (equation 9).

Worst case over clipped inputs. Write $\mathcal{M}$ for the Poisson-subsampled Gaussian mechanism at rate q_t, noise σ_t and unit sensitivity. Mironov et al. (2019, Thm. 4) show that for every dataset S, every record x with $\|\bar{g}_t(x)\| \leq 1$, and every order $\alpha \geq 1$,

$$D_\alpha\big(\mathcal{M}(S \cup \{x\}) \,\big\|\, \mathcal{M}(S)\big) \;\leq\; D_\alpha\big(P_t^{(1)} \,\big\|\, Q_t^{(1)}\big),$$

the one-dimensional pair of equation 10.[1] Take $S = \emptyset$ and x with clipped gradient z: $\mathcal{M}(\emptyset)$ is pure noise, P_0, and $\mathcal{M}(\{x\})$ includes the record with probability q_t, giving $P_{Y|z}$. The other records of the actual training set do not appear because Lemma D.2 has already conditioned them away through Ω. At $\alpha = 1$ the right-hand side is K_t^{PRV}, so the supremum in equation 27 is at most K_t^{PRV}, hence

$$I(G; Y \mid \Omega) \;\le\; I(G; Y) \;\le\; K_t^{\mathrm{PRV}}, \tag{28}$$

with both sides in nats; dividing by $\ln 2$ to convert to bits and summing over t completes the proof.

□

E Details on the representation entropy confidence bound

Section E.1 sets out what the bound needs from an entropy estimator and why the plug-in is the only candidate that supplies it; Sections E.2 and E.3 then verify each requirement in turn, and Section E.4 chains them into Proposition 5.1.

E.1 What the bound needs from the entropy estimator

Corollary 3.7 consumes $\bar{h}_U(k)$ through the exponent $2^{\,2(\bar{h}_U(k)-\gamma)/k}$, maximised over a pair (k, ν^2) chosen from $\mathcal{G}$ after the data is seen. We supply it through a lower confidence bound $\bar{h}_{\mathrm{LB}}(k) \le \bar{h}_U(k)$, and the bound pays for the shortfall $\bar{h}_U(k) - \bar{h}_{\mathrm{LB}}(k)$ exactly as it pays for leakage, since a shortfall of t bits costs the bounded ratio a factor $2^{-2t/k}$. The shortfall has two parts, whatever the point estimate is biased by and whatever is subtracted from it to make the bound hold at the stated confidence. What the bound needs is that both be controlled in a single one-sided, finite-sample statement, namely that $\bar{h}_{\mathrm{LB}}(k) \le \bar{h}_U(k)$ for every $(k, \nu^2) \in \mathcal{G}$ with probability at least $1 - \alpha$. Only the lower side is used, and the union bound over $\mathcal{G}$ in Proposition 5.1 needs a per-pair tail probability at the n we have rather than in the limit.

Most entropy estimators do not supply this. Nearest-neighbour estimators (Kozachenko & Leonenko, 1987; Berrett et al., 2019) have a bias depending on the unknown density, so no finite-sample statement is available in either direction, and variational estimators of the mutual information into which $h(\tilde{z}^{\,\Pi_b})$ decomposes (Lemma B.1(iii)) are lower bounds in expectation only. The finite-sample error bounds that do exist for the smoothed setting, the risk bound of Goldfeld et al. (2020, Thm. 2) and the confidence sequence of Manole & Ramdas (2023, Cor. 21), are two-sided, and a two-sided bound must control the estimator's bias, which costs a prefactor exponential in the dimension and makes the bound vacuous at our sample sizes. We need only a lower bound, and the plug-in's bias is downward, so it needs no control at all; what remains is the fluctuation, which Lemma E.6 bounds at a constant that does not grow with the dimension.

The plug-in of Goldfeld et al. (2020) meets the requirement. Its error about $\bar{h}_U(k)$ splits into a bias and a fluctuation. The bias is downward at every n (Goldfeld et al., 2020, eq. (70)), so it needs no margin, and this is what places the plug-in outside the two-sided barrier above, since it is one-sided by construction. The fluctuation is a function of the n independent samples, the rotation and dither scale having been fixed on disjoint data (Section F), and changing one sample moves it by at most $H_b(1/n)$ regardless of k (Corollary E.3), so McDiarmid gives a finite-sample margin of $\tilde{O}(n^{-1/2})$ bits carrying no k (Lemma E.6). The Monte Carlo integrator we evaluate it with has known conditional mean and variance given the sample (Lemma E.4), so its error is given a margin the same way (Lemma E.5). It is capped, however: $\hat{h}_n \le h(W^{(k)}) + \log_2 n$ by Lemma E.1.

E.2 Bounded differences for the plug-in

Goldfeld et al. (2020) do not provide a lower confidence bound, they bound $\mathbb{E}[\hat{h}_n]$ and have no reason to control its fluctuation about that mean. Here we derive the constant needed for McDiarmid's inequality. It rests on a standard result about mixtures which we will re-derive here first for the reader's benefit.

[1] Theorem 4 of Mironov et al. (2019) bounds the Rényi divergence in both directions, $D_\alpha(\mathcal{M}(S) \,\|\, \mathcal{M}(S \cup \{x\}))$ and $D_\alpha(\mathcal{M}(S \cup \{x\}) \,\|\, \mathcal{M}(S))$; we need only the latter. The theorem is stated for Rényi order α, and the KL divergence is the case $\alpha = 1$, which their Definition 1 includes by taking the limit $\alpha \to 1$; nothing in their proof requires $\alpha > 1$.

Lemma E.1 (Entropy of a finite mixture). *Let $f_1, \dots, f_r$ be densities on $\mathbb{R}^k$ with finite differential entropy, and let λ be a probability vector on $\{1, \dots, r\}$. Then*

$$\sum_{i=1}^{r} \lambda_i \, h(f_i) \;\le\; h\Big(\sum_{i=1}^{r} \lambda_i f_i\Big) \;\le\; \sum_{i=1}^{r} \lambda_i \, h(f_i) + H(\lambda), \tag{29}$$

with $H(\lambda) := -\sum_i \lambda_i \log_2 \lambda_i$ the Shannon entropy of the weights.

Proof. Let $A \sim \lambda$ and, given $A = i$, let $B \sim f_i$, so that B has the mixture density and $h(B \mid A) = \sum_i \lambda_i h(f_i)$. The left inequality is $h(B \mid A) \le h(B)$, conditioning reducing entropy (Cover & Thomas, 2006, Thm. 8.6.1). For the right, $h(B) = h(B \mid A) + I(B; A)$ and $I(B; A) \le H(A) = H(\lambda)$. That is, knowing which component a draw came from saves at most $H(\lambda)$ bits. □

Lemma E.2 (Bounded differences under one relocation). *Fix $s > 0$ and $\Pi : \mathbb{R}^d \to \mathbb{R}^k$. Let $x_{1:n}$ and $x'_{1:n}$ be two samples that agree except at one index j, and let $\hat{f}_n$ and $\hat{f}'_n$ be the corresponding mixtures equation 11: each is the equal-weight average of the n Gaussians $\phi_s(\cdot - \Pi x_i)$, respectively $\phi_s(\cdot - \Pi x'_i)$. Then*

$$\big|h(\hat{f}_n) - h(\hat{f}'_n)\big| \;\le\; H_b(1/n),$$

with $H_b(p) := -p \log_2 p - (1-p)\log_2(1-p)$ the binary entropy in bits.

Proof. Let $g := \frac{1}{n-1} \sum_{i \ne j} \phi_s(\cdot - \Pi x_i)$ be the equal-weight mixture of the $n-1$ Gaussians the two share. Then

$$\hat{f}_n = \tfrac{n-1}{n}\, g + \tfrac{1}{n}\, \phi_s(\cdot - \Pi x_j), \qquad \hat{f}'_n = \tfrac{n-1}{n}\, g + \tfrac{1}{n}\, \phi_s(\cdot - \Pi x'_j),$$

two-component mixtures with the same first component and the same weights $\lambda = (\frac{n-1}{n}, \frac{1}{n})$. Since a Gaussian's entropy depends only on its covariance, $h(\phi_s(\cdot - \Pi x_j)) = h(\phi_s(\cdot - \Pi x'_j)) = h(W^{(k)})$, so the weighted component entropies coincide for the two mixtures, $\sum_i \lambda_i h(f_i) = \frac{n-1}{n} h(g) + \frac{1}{n} h(W^{(k)}) =: L$ in both cases. Lemma E.1 then places both $h(\hat{f}_n)$ and $h(\hat{f}'_n)$ in $[L,\, L + H(\lambda)]$, and $H(\lambda) = H_b(1/n)$. □

Corollary E.3 (Bounded differences for the block average). *Fix $s > 0$ and a finite family $\Pi_1, \dots, \Pi_{N_k}$ of maps $\mathbb{R}^d \to \mathbb{R}^k$, all independent of $x_{1:n}$. The map $x_{1:n} \mapsto \hat{h}_n$ of equation 11 satisfies the bounded-differences condition with constant $H_b(1/n)$ in each coordinate: for every j and every $x'_j \in \mathcal{Q}$,*

$$\big|\hat{h}_n(x_1, \dots, x_j, \dots, x_n) - \hat{h}_n(x_1, \dots, x'_j, \dots, x_n)\big| \;\le\; H_b(1/n).$$

The constant does not depend on k, N_k, s, or the Π_b.

Proof. Changing x_j relocates exactly one centre of every $\hat{f}_n(\cdot\,; \Pi_b)$, so by Lemma E.2 each $\hat{h}_n(\Pi_b)$ moves by at most $H_b(1/n)$, and an average of quantities each moving by at most c moves by at most c. □

E.3 The two margins

Fix $(k, \nu^2) \in \mathcal{G}$, which fixes $s = \nu\sqrt{\operatorname{tr}\Sigma/k}$. Establishing equation 12 involves three quantities: the target $\bar{h}_U(k)$, a fixed property of the source; the plug-in $\hat{h}_n$ of equation 11, the exact block-averaged entropy of the empirical mixtures, which is random through the sample $x_{1:n}$ and has no closed form; and $\hat{h}_{n,\mathrm{MC}}$, the Monte Carlo approximation to $\hat{h}_n$ that we actually compute, which is random through the sample and through the Monte Carlo noise. Writing

$$\hat{h}_{n,\mathrm{MC}} - \bar{h}_U(k) \;=\; \big(\hat{h}_{n,\mathrm{MC}} - \hat{h}_n\big) + \big(\hat{h}_n - \mathbb{E}[\hat{h}_n]\big) + \big(\mathbb{E}[\hat{h}_n] - \bar{h}_U(k)\big), \tag{30}$$

the third term is a deterministic quantity and is non-positive for every n: it is the downward bias of the plug-in (Goldfeld et al., 2020, eq. (70)), averaged over blocks. The first two are random and each is given a margin, t_{MC} and t_{MD}, that it exceeds with probability at most $\alpha_1/|\mathcal{G}|$ and $\alpha_2/|\mathcal{G}|$ respectively, where $\alpha_1 + \alpha_2 = \alpha$. The two are separated by conditioning on the sample: with

$x_{1:n}$ fixed, $\hat{h}_n$ is a constant and the first term is pure integration error, whose conditional mean and variance are known (Lemma E.4). Since a conditional probability bounded by $\alpha_1/|\mathcal{G}|$ for every sample averages to an unconditional probability bounded by $\alpha_1/|\mathcal{G}|$, the two margins combine by a union bound (Section E.4).

The rotation U and the scale s are computed from the held-out sample $D_\perp$ and treated as fixed constants throughout; every probability below is over the estimation sample and the Monte Carlo noise only. This is what makes $\hat{h}_n$ a function of the independent $x_1, \dots, x_n$ alone, as Lemma E.6 requires.

The integrator. Write $y_{b,i} := \Pi_b x_i$ for the projected sample points, so $\hat{f}_n(\cdot\,; \Pi_b)$ is the equal-weight mixture of the Gaussians $\phi_s(\cdot - y_{b,i})$, $i \in [n]$. A draw from it is $y_{b,J} + W$ with $J \sim \mathrm{Unif}[n]$ and $W \sim \mathcal{N}(0, s^2 I_k)$ independent, so

$$h(\hat{f}_n(\cdot\,; \Pi_b)) \;=\; -\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\log_2 \hat{f}_n(y_{b,i} + W; \Pi_b)\big],$$

and we estimate each of the n expectations by n_{MC} independent draws of W:

$$\hat{h}_{n,\mathrm{MC}}(\Pi_b) \;:=\; -\frac{1}{n\, n_{\mathrm{MC}}} \sum_{i=1}^{n}\sum_{j=1}^{n_{\mathrm{MC}}} \log_2 \hat{f}_n\big(y_{b,i} + W_{ij,b};\, \Pi_b\big), \qquad W_{ij,b} \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, s^2 I_k), \tag{31}$$

with the $W_{ij,b}$ independent of each other and of the sample, and $\hat{h}_{n,\mathrm{MC}} := N_k^{-1} \sum_b \hat{h}_{n,\mathrm{MC}}(\Pi_b)$. This is the integrator of Goldfeld et al. (2020, eqs. (36a)–(36b)), which uses natural logarithms; equation 31 is their estimator divided by $\ln 2$.

Lemma E.4 (Integrator: conditional mean and variance)**.** *Conditionally on the sample $x_{1:n}$,*

(i) $\mathbb{E}[\hat{h}_{n,\mathrm{MC}}(\Pi_b) \mid x_{1:n}] = h(\hat{f}_n(\cdot\,; \Pi_b))$ *for every b, and hence* $\mathbb{E}[\hat{h}_{n,\mathrm{MC}} \mid x_{1:n}] = \hat{h}_n$*;*

(ii) for every b,

$$\mathrm{Var}[\hat{h}_{n,\mathrm{MC}}(\Pi_b) \mid x_{1:n}] \;\le\; \frac{c_{\mathrm{MC}}(\Pi_b)}{(\ln 2)^2\, n\, n_{\mathrm{MC}}}\ \textit{bits}^2, \tag{32}$$

where, with $\hat{m}_b := \frac{1}{n}\sum_{i=1}^{n} \|y_{b,i}\|^2$ the mean squared norm of the component means,

$$c_{\mathrm{MC}}(\Pi_b) \;:=\; \frac{9ks^2 + 8\big(2 + s\sqrt{k}\big)\hat{m}_b + 3\big(11 s\sqrt{k} + 1\big)\sqrt{\hat{m}_b}}{s^2}; \tag{33}$$

(iii) $\mathrm{Var}[\hat{h}_{n,\mathrm{MC}} \mid x_{1:n}] \;\le\; \dfrac{1}{(\ln 2)^2\, N_k^2\, n\, n_{\mathrm{MC}}} \displaystyle\sum_{b=1}^{N_k} c_{\mathrm{MC}}(\Pi_b)\ \textit{bits}^2.$

Proof. Given the sample, $\hat{f}_n(\cdot\,; \Pi_b)$ is a fixed Gaussian mixture with component means $y_{b,1}, \dots, y_{b,n}$, the setting of Goldfeld et al. (2020, Sec. III-D). Their eq. (77) is the unbiasedness of their estimator; ours is theirs divided by $\ln 2$, so (i) holds for it too. Their Theorem 5(ii) bounds the mean squared error of their estimator about $h(\hat{f}_n(\cdot\,; \Pi_b))$ by $c_{\mathrm{MC}}(\Pi_b)/(n\, n_{\mathrm{MC}})$ in nats^2, where their m is the mean squared norm of the component means, here $\hat{m}_b$. Mean squared error is variance plus squared bias, and the bias is zero by (i), so this bounds the variance; dividing by $(\ln 2)^2$ converts it to bits^2 for our estimator, giving (ii). For (iii), the noise $W_{ij,b}$ is drawn independently for each block, so given the sample the N_k terms $\hat{h}_{n,\mathrm{MC}}(\Pi_b)$ are independent; the variance of their average is $N_k^{-2} \sum_b \mathrm{Var}[\hat{h}_{n,\mathrm{MC}}(\Pi_b) \mid x_{1:n}]$, and (ii) bounds each summand. □

Lemma E.5 (Monte Carlo margin)**.** *Let σ^2_{MC} denote the right-hand side of Lemma E.4(iii). Conditionally on the sample, for every $t > 0$,*

$$\Pr\big[\hat{h}_{n,\mathrm{MC}} > \hat{h}_n + t \;\big|\; x_{1:n}\big] \;\le\; \frac{\sigma^2_{\mathrm{MC}}}{\sigma^2_{\mathrm{MC}} + t^2}.$$

In particular, with

$$t_{\mathrm{MC}} \;:=\; \sigma_{\mathrm{MC}} \sqrt{|\mathcal{G}|/\alpha_1 - 1}, \tag{34}$$

the value of t at which the bound equals $\alpha_1/|\mathcal{G}|$, we have $\hat{h}_{n,\mathrm{MC}} \le \hat{h}_n + t_{\mathrm{MC}}$ with probability at least $1 - \alpha_1/|\mathcal{G}|$.

Proof. Given the sample, $\hat{h}_{n,\mathrm{MC}}$ has mean $\hat{h}_n$ and variance at most σ^2_{MC} by Lemma E.4. Cantelli's inequality, $\Pr[X - \mathbb{E}X > t] \le \mathrm{Var}X/(\mathrm{Var}X + t^2)$, gives the tail bound; its right-hand side is increasing in $\mathrm{Var}X$, so the variance may be replaced by its upper bound. Setting it equal to $\alpha_1/|\mathcal{G}|$ and solving for t gives equation 34. Writing $A := \{\hat{h}_{n,\mathrm{MC}} > \hat{h}_n + t_{\mathrm{MC}}\}$, we have shown $\Pr[A \mid x_{1:n}] \le \alpha_1/|\mathcal{G}|$ for every realisation of the sample, and $\Pr[A] = \mathbb{E}[\Pr[A \mid x_{1:n}]]$, so $\Pr[A] \le \alpha_1/|\mathcal{G}|$ as well. Note that t_{MC} depends on the sample through $\hat{m}_b$, which is why we condition. □

Lemma E.6 (Sampling margin)**.** *For every $t > 0$,*

$$\Pr\big[\hat{h}_n > \mathbb{E}[\hat{h}_n] + t\big] \;\le\; \exp\Big(-\frac{2t^2}{n\,H_b(1/n)^2}\Big).$$

In particular, with

$$t_{\mathrm{MD}} \;:=\; H_b(1/n)\sqrt{\tfrac{n}{2}\ln(|\mathcal{G}|/\alpha_2)}, \tag{35}$$

the value of t at which the bound equals $\alpha_2/|\mathcal{G}|$, we have $\hat{h}_n \le \mathbb{E}[\hat{h}_n] + t_{\mathrm{MD}}$ with probability at least $1 - \alpha_2/|\mathcal{G}|$.

Proof. The x_i are independent (Assumption 3.2) and changing any one of them moves $\hat{h}_n$ by at most $c := H_b(1/n)$ (Corollary E.3), the Π_b and s being fixed independently of the sample. McDiarmid's inequality (McDiarmid, 1989) gives $\Pr[\hat{h}_n - \mathbb{E}\hat{h}_n > t] \le \exp(-2t^2/(nc^2))$; setting this equal to $\alpha_2/|\mathcal{G}|$ and solving gives equation 35. □

E.4 Proof of Proposition 5.1

Proof. Fix a pair $(k, \nu^2) \in \mathcal{G}$. By Lemmas E.5 and E.6 the events $\{\hat{h}_{n,\mathrm{MC}} > \hat{h}_n + t_{\mathrm{MC}}\}$ and $\{\hat{h}_n > \mathbb{E}[\hat{h}_n] + t_{\mathrm{MD}}\}$ have probability at most $\alpha_1/|\mathcal{G}|$ and $\alpha_2/|\mathcal{G}|$ respectively, so with probability at least $1 - (\alpha_1 + \alpha_2)/|\mathcal{G}|$ neither occurs, and then

$$\bar{h}_{\mathrm{LB}}(k) \;=\; \hat{h}_{n,\mathrm{MC}} - t_{\mathrm{MC}} - t_{\mathrm{MD}} \;\le\; \hat{h}_n - t_{\mathrm{MD}} \;\le\; \mathbb{E}\big[\hat{h}_n\big] \;\le\; \bar{h}_U(k).$$

The last inequality is deterministic and holds for every n: it is the downward bias of the plug-in, $\mathbb{E}[\hat{h}_n(\Pi_b)] \le h(\tilde{z}^{\,\Pi_b})$ (Goldfeld et al., 2020, eq. (70)), applied to each block and averaged. A union bound over the grid gives

$$\Pr\Big[\exists\,(k,\nu^2) \in \mathcal{G} :\; \bar{h}_{\mathrm{LB}}(k) > \bar{h}_U(k)\Big] \;\le\; \sum_{(k,\nu^2)\in\mathcal{G}} \Big(\frac{\alpha_1}{|\mathcal{G}|} + \frac{\alpha_2}{|\mathcal{G}|}\Big) \;=\; \alpha,$$

so equation 12 holds simultaneously for every pair with probability at least $1 - \alpha$, over the sample and the Monte Carlo noise, conditionally on $D_\perp$. The union bound over $\mathcal{G}$ is necessary: equation 8 selects the pair (k, ν^2) that gives the largest bound, and that selection depends on the data through $\bar{h}_{\mathrm{LB}}(k)$. A guarantee holding at any single pre-specified pair with probability $1 - \alpha$ would not cover the pair chosen afterwards, whereas equation 12 holding for every pair at once does, whichever is selected. The bound equation 13 follows since the absolute form of equation 7 (Step 3 of the proof of Corollary 3.7) is nondecreasing in the entropy, so substituting $\bar{h}_{\mathrm{LB}}(k)$ can only lower it; dividing by $\mathrm{tr}\,\hat{\Sigma}_\Psi \ge \mathrm{tr}\,\Sigma_\Psi$ and using $ks^2/\,\mathrm{tr}\,\hat{\Sigma}_\Psi = \nu^2$ gives equation 14. □

F Data handling

The bound consumes four inputs.

- The **leakage**, γ, read from the DP-SGD hyperparameters.
- The **block rotation**, U, and the **dither scale**, s, which we fix from a covariance estimate.

- The **entropy confidence bound**, $\bar{h}_{\mathrm{LB}}$, on the representation.

The attack experiment adds the components an adversary fits. Each has its own conditions on the data it may use, Section F.1 lists those conditions. Section F.2 turns them into labelled partitions (Table 2) and Section F.3 gives the layout we recommend to a practitioner, which is also the setup we use for the privacy-spectrum experiments. Section F.4 gives the layout of the attack experiment. Figure 4 draws both and Algorithm 1 gives the order of computation.

F.1 What each quantity requires

The protected records. Write D_{train} for the records DP-SGD is run on. These are the X of Section 3.1. Two conditions follow from the assumptions.

Req 1. D_{train} is an i.i.d. sample from P_x (Assumption 3.2), which Lemma 3.4 uses to discharge $I(x_i; X^{-i})$. Any preprocessing computed across D_{train} and applied back to it (a dataset mean, a whitening transform, deduplication, class rebalancing, a fitted encoder) couples x_i to X^{-i} and breaks the argument at its first step. Per-sample maps, per-example augmentation, and any transform whose constants are fixed on data disjoint from D_{train} preserve it.

Req 2. Every trajectory computed on D_{train} that the adversary may observe composes into γ (Lemma D.1). A hyperparameter search of k runs on the protected records is therefore charged the leakage of k runs, not one; since γ reads only (q_t, σ_t), the search is simply one longer trajectory. Alternatively the search is run on records outside the protected set, in which case we aren't required to account its leakage.

The leakage γ. Needs no data. Theorem 4.1 requires Poisson subsampling and a data-independent initialisation, together with Req 2.

The rotation U. We take $U = d^{-1/2} H_d V^\top$ (Remark C.1). The Hadamard factor is a fixed matrix, so the only part of U that touches data is the eigenbasis V of Σ, which we estimate. Corollary 3.7 holds for every orthogonal U, so U may be computed from any data, but Lemma E.2 requires the Π_b to be fixed independently of the estimation sample, so:

Req 3. the sample on which V is estimated is disjoint from the estimation sample.

V buys tightness, not validity: every orthogonal U is admissible (Corollary 3.7), so where no such sample can be spared one takes $V = I_d$, giving $U = d^{-1/2} H_d$ requiring no data. We use the data-free variant $U = d^{-1/2} H_d D$, with D a fixed random diagonal sign flip, for the ImageNette pixel bound and for every LPIPS bound (Remark C.1).

The dither scale and the reference scale. The estimator convolves the sample with a Gaussian of bandwidth $s = \nu \sqrt{\operatorname{tr} \hat{\Sigma}/k}$, so s is set by an estimate of $\operatorname{tr} \Sigma$.

Req 4. The sample on which $\operatorname{tr} \Sigma$ is estimated is disjoint from the estimation sample. Lemma E.6 treats s as a constant and lets one record of the estimation sample move; if s were computed from that sample, moving one record would also move the bandwidth, and the bounded-differences constant of Lemma E.2 no longer applies.

The same estimate is the reference scale $\mathrm{RMSE}^2_{\mathrm{prior}} = \operatorname{tr} \Sigma$ of ρ. The guarantee is the absolute bound equation 4, in which $\operatorname{tr} \Sigma$ does not appear and the ratio only rescales it - so a point estimate suffices. The attack-bound comparison of Figures 3 and 7 is unaffected in any case, both sides being divided by the same number. A practitioner who wants a confidence bound on $\rho^\star$, and so provably at most one, divides instead by a one-sided upper confidence bound on $\operatorname{tr} \Sigma$ (for example using empirical-Bernstein (Section G.2)).

The entropy bound $\bar{h}_{\mathrm{LB}}$. $\bar{h}_U(k)$ is a property of the source P_x, and Proposition 5.1 lower-bounds it from any i.i.d. sample of P_x on which U and s were not computed.

Req 5. The estimation sample is i.i.d. from P_x and disjoint from the sample of Req 3. Both margins scale as $n^{-1/2}$ (Section E.3), so the sample should be as large as possible.

Nothing requires the estimation sample to be disjoint from D_{train}. Under Req 1 the training set is itself an i.i.d. sample of P_x, and using it estimates the entropy of exactly the source the protected records are drawn from, with the most data. Coverage is then over the random draw of the training set, not conditional on the realised one. Any other sample must be a draw from P_x as well, since Proposition 5.1 bounds the entropy of whatever distribution the sample comes from.

Validation and test. Neither is an input to any estimand.

Req 6. As is ordinarily practiced, one can't use the same data to both train a model and evaluate its performance.

Req 7. Records on which accuracy is reported are disjoint from both data used to train the model and data used to evaluate the model in order to for example pick hyperparameters.

Neither set is trained on, so neither carries γ and the bound makes no claim about it.

The attack. The bound quantifies over reconstructions $g(\theta, X^{-i})$ with internal randomness and knowledge of P_x (Section 3.1); an experiment testing it must keep the attack inside that class.

Req 8. Everything the attack fits or draws from — here the candidate set and the centring mean — excludes the target. An attack that has seen its target is not a function of (θ, X^{-i}) and can fall below the bound without the bound being wrong. X^{-i} itself is known to the adversary, so the shadow pool may include it.

Req 9. The attacked architecture, including the centring mean of its input (Section G.5), is fixed on records disjoint from the attacked set $X = X^{-i} \cup \{x_i\}$. The bound holds for every architecture at a given (q, σ, T), so a fixed input map is on the same footing as the linear layer itself; but one fit on the attacked records is preprocessing across X and violates Req 1.

Req 10. The targets are i.i.d. from P_x, so that the mean error over them estimates the expectation the bound controls, and are disjoint from the sample U is computed on (Req 3, admissibility) and from the shadow pool (Req 8). They need not be disjoint from the estimation sample. Proposition 5.1 is a statement about the population entropy $\bar{h}_U(k)$, holding on an event of probability $1-\alpha$ over the estimation sample. On that event the bound holds for the expected error of every g, whichever records the targets are. The attack's mean error over the targets is an estimate of that expectation, and its expectation does not depend on whether the targets also appear in the estimation sample. Overlap makes the two random quantities dependent, but changes neither the probability of the coverage event nor the expected attack error, so the comparison in Figures 3 and 7 is valid either way.

F.2 The partitions

Table 2 names a partition for each role above and records, for each, what it supplies, the requirement that constrains it, which partitions it must be disjoint from, and the property its size must have. A $\star$ marks a disjointness the guarantee or the threat model requires. The rest are kept for a clean joint statement and could be relaxed. Two partitions are shared by both experiments: $D_\perp$, which fixes U and s, and D_{est}, which supplies the entropy bound.

F.3 Deployment, and the privacy-spectrum runs

A practitioner makes two splits of the available data, independently of each other. For the bound, they draw a sample $D_\perp$ of at least d records, on which they compute the covariance $\hat{\Sigma}$, whose eigenvectors give V and whose trace gives $\operatorname{tr}\hat{\Sigma}$, and a disjoint estimation sample D_{est}, as large as possible. For training, they split the data into D_{train}, D_{val} and D_{test} as in any training pipeline. Only $D_\perp$ and D_{est} must be disjoint (Req 3, Req 4). D_{val} is used to evaluate performance during training for purposes such as hyperparameter selection and D_{test} is used to report final utility achieved.

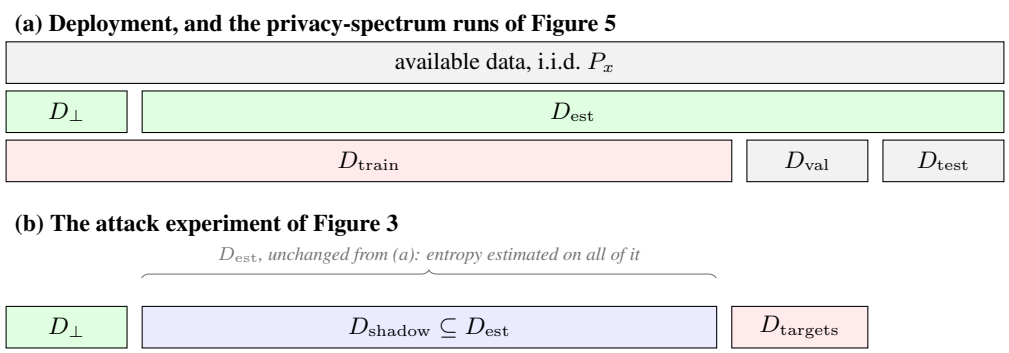


Figure 4: How the available data is partitioned. Green blocks supply the bound, red blocks are the protected records trained on, and blue blocks are data the attack uses. In **(a)** the bound's split into $D_\perp$ and D_{est} (top row) and the training split (bottom row) are drawn independently and may overlap; only $D_\perp$ and D_{est} must be disjoint. $D_\perp$ and D_{est} are shared, so the rotation and the entropy estimate are the same in the deployment **(a)** and the attack experiment **(b)**; the targets are held out from both.

Algorithm 1 Reconstruction lower bound for a DP-SGD deployment.

Require: $(q_t, \sigma_t)_{t=1}^T$ for every run charged to D_{train}; samples $D_\perp$ and D_{est} (i.i.d. from P_x; D_{est} may be D_{train}); grid $\mathcal{G}$; confidence α

1: $\gamma \leftarrow \frac{1}{\ln 2} \sum_{\text{runs}} \sum_{t=1}^T K_t^{\text{PRV}}$ ▷ Theorem 4.1; Req 2
2: On $D_\perp$: fix, for each Ψ, $U = m^{-1/2} H_m V^\top$, the dither scales $s_{k,\nu}$, and $\operatorname{tr} \hat{\Sigma}_\Psi$ ▷ Req 3, Req 4
3: **for** $\Psi \in \{\text{id}, \Psi_{\text{LPIPS}}\}$ and $(k, \nu^2) \in \mathcal{G}$ **do**
4: On D_{est}: $\bar{h}_{\text{LB}}(k) \leftarrow \hat{h}_{n,\text{MC}} - t_{\text{MC}} - t_{\text{MD}}$ ▷ Proposition 5.1; Req 5
5: **end for**
6: $\rho^\star_\Psi \leftarrow$ equation 14
7: Train by Poisson-subsampled DP-SGD on D_{train}; to target a given $\rho^\star$, invert steps 1 and 6 for σ

The hyperparameter search is either charged through γ (by summing Theorem 4.1 over every run), moved onto records outside the protected set, or done privately (such as on trusted private compute). The learning rate and architecture do not enter γ, so a search over them at fixed (q_t, σ_t) adds one trajectory's leakage per run and nothing more. This layout is also that of the privacy-spectrum runs (Figure 5), so Figure 4(a) is both the recommendation and what we do: $D_\perp$ and D_{est} are the splits of the attack experiment (Section F.4), drawn independently of the utility runs' split into D_{train}, D_{val} and D_{test}. As is common in the literature (Ponomareva et al., 2023), and to keep the experimental results simple to interpret, we assume hyperparameter selection leaks nothing. Where no $D_\perp$ can be spared, U is taken data-free and $\operatorname{tr} \Sigma$ is replaced by a declared upper bound, at a cost in tightness alone.[2]

F.4 The attack experiment

The attack experiment reuses the rotation, the dither scales and the entropy bounds of (a) unchanged, so the two experiments differ only in the leakage their mechanism incurs. It uses two further sets. $D_{\text{shadow}} \subseteq D_{\text{est}}$ is the attack's candidate set and supplies its centring mean (Req 8, Req 9). D_{targets} is 100 records from a held-out slice disjoint from D_{shadow}, D_{est} and $D_\perp$ (Req 10).

[2] Any declared value keeps the absolute bound equation 4 valid, since the dither scale it sets may be any positive number; the ratio bound stays valid provided the declared value is an upper bound on $\operatorname{tr} \Sigma$, since dividing by a larger number can only lower it. For 8-bit pixels, $\operatorname{tr} \Sigma \le d \cdot 255^2/4$ always.

G EXPERIMENTAL DETAILS

Sections G.1 to G.3 apply to both experiments; Section G.4 to Figures 2 and 5; Sections G.5 to G.7 to Figures 3 and 7.

G.1 DATASETS

Dataset	$C \times H \times W$	d	$\operatorname{tr}\Sigma$	$\mathrm{RMSE}_{\mathrm{prior}}/\sqrt{d}$	n_{est}	$\lvert D_\perp \rvert$
MNIST	$1\times28\times28$	784	3.43×10^6	66.2	59k	10k
CIFAR-10	$3\times32\times32$	3072	1.24×10^7	63.6	49k	10k
CelebA	$3\times64\times64$	12,288	5.91×10^7	69.3	44k	117k
ImageNette	$3\times128\times128$	49,152	2.46×10^8	70.7	10k	1k

The deployment runs span $d = 784$ to 49,152, greyscale and color. The CelebA dataset serves as a particularly interesting dataset as recovery where one is able to ascertain the identity of the individual in the samples constitutes a natural privacy threat that one can validate with their own eyes. We deploy our bound and evaluate the utility achieved by DP-SGD across all the datasets above. For ImageNette, $D_\perp$ supplies only $\operatorname{tr}\hat{\Sigma}$; no eigenbasis is estimated on it (Remark C.1).

The attack experiment uses MNIST and CelebA, a low-dimensional greyscale source and a colour face source. On CelebA, recovering a recognisable face is the natural privacy harm, so it is the case in which the perceptual bound is the one that matters. Each configuration is attacked on the same 100 held-out targets at every leakage (Section F.4).

CelebA (Liu et al., 2015) uses the binary *Male* attribute as its label. Networks consume images rescaled to $[0, 1]$ (CIFAR-10 utility training additionally standardises by fixed, public per-channel constants); entropy is computed on the raw 8-bit grid, so Assumption 3.1 holds with $b_x = 8$, $c = 256$ and $\Delta = 1$. No preprocessing couples samples: every normalisation uses constants fixed independently of X, so Assumption 3.2 holds (Section F).

G.2 THE ENTROPY BOUND

The splits are those of Section F.3, with sizes in Section G.1.

Reference scale. We normalise ρ by the point estimate $\operatorname{tr}\hat{\Sigma}_\Psi$ from $D_\perp$ (Table 3). This does not affect the guarantee: the absolute bound equation 4 does not involve $\operatorname{tr}\Sigma_\Psi$, and any other normaliser c simply rescales $\rho^\star$ by $(\operatorname{tr}\hat{\Sigma}_\Psi/c)^{1/2}$. To gauge the estimation error we also give a 95% upper confidence bound from the empirical Bernstein inequality of Maurer & Pontil (2009), which requires i.i.d. variables with mean $\operatorname{tr}\Sigma_\Psi$ and a range fixed in advance. Pairing the records of $D_\perp$ at random supplies these: for independent $x, x' \sim P_x$, $Y := \frac{1}{2}\|\Psi(x) - \Psi(x')\|^2$ has mean exactly $\operatorname{tr}\Sigma_\Psi$ and, unlike $\|\Psi(x) - \mathbb{E}\Psi(x)\|^2$, needs no estimate of the mean. Its range is $[0, B/2]$, with B the largest possible value of $\|\Psi(x) - \Psi(x')\|^2$. In pixel space $B = 255^2 d$. In LPIPS space this squared distance is the LPIPS distance, a sum over layers ℓ of the spatial average of $\sum_c w_{\ell c}(a_c - b_c)^2$, where $w_{\ell c} \geq 0$ is the learned LPIPS weight on channel c of layer ℓ and a, b are the two images' normalised features at one location. These features are nonnegative with norm at most one, so $\|a - b\|^2 \leq 2$, each layer contributes at most $2\max_c w_{\ell c}$, and $B = 2\sum_\ell \max_c w_{\ell c}$. Empirical Bernstein is much tighter here than Hoeffding, whose width scales as $B/\sqrt{m}$ for m pairs: B is attained only by two opposite-extreme records, so Y varies far less than its range, and empirical Bernstein uses the sample variance of Y and pays for B only at rate B/m. The resulting bound exceeds the point estimate by at most 5.8% on MNIST, CIFAR-10 and CelebA, and by 16.6% on ImageNette, where $|D_\perp|$ is small. Normalising by it instead would lower every reported bound by at most 0.026, or 0.064 on ImageNette, and the bound would then hold with probability at least 90% by a union bound with Proposition 5.1.

Grid. We take $\mathcal{G} = K \times N$, with block sizes K the divisors of d up to 32 in pixel space ($\{1, 2, 4, 7, 8, 14, 16, 28\}$ on MNIST, $\{1, 2, 3, 4, 6, 8, 12, 16, 24, 32\}$ on the others) and $K = \{1, 2, 4, 8, 16, 32\}$ in LPIPS space, and dither levels $N = \{0.2, 0.4, 0.8, 1.6, 3.2, 6.4\}$, so $|\mathcal{G}| = 48$

(MNIST pixel), 60 (other pixel) and 36 (LPIPS), both fixed before any entropy is estimated, as Section E.4 requires. In pixel space K contains only divisors of d, and in LPIPS space only divisors of the padded dimension 32,768 (see Blocks below). The selected pair often lies on the boundary of the grid (Table 5): $k^\star = \max K$ at $\varepsilon \geq 8$ in pixel space on MNIST, CIFAR-10 and ImageNette (and at $\varepsilon = 16$ on CelebA), and at almost every leakage in LPIPS space; $\nu^{2\star}$ reaches an end of N in a few places (CelebA pixel: 0.2 at $\varepsilon \leq 1$ and $\varepsilon = 32$; CelebA LPIPS: 6.4 at $\varepsilon \leq 0.5$; MNIST LPIPS: 6.4 at $\varepsilon = 2$).

Blocks. We never store the rotation as a dense matrix. Instead we keep the Hadamard factor as a Kronecker product of small Hadamard matrices, so applying U costs one small matrix product per factor. In pixel space we use $H_{784} = H_{28} \otimes H_{28}$, $H_{3072} = H_{12} \otimes H_{64} \otimes H_4$, $H_{12288} = H_{12} \otimes H_{64} \otimes H_{16}$ and $H_{49152} = H_{12} \otimes H_{64} \otimes H_{64}$, with the power-of-two factors from Sylvester's doubling and the others from Paley's construction. These are Hadamard matrices of order exactly d, so no zero-padding is needed; padding adds dimensions that carry no entropy and loosens the bound. In LPIPS space we zero-pad the $m = 31{,}872$ features to $32{,}768 = 2^{15}$ and rotate by $U = 32768^{-1/2} H_{32768} D$, with $H_{32768} = H_{64} \otimes H_{64} \otimes H_8$ and D a fixed random diagonal sign flip, so no eigenbasis is estimated. The padded coordinates are deterministically zero, so by Remark C.1 the bound transfers to the unpadded features, at a cost in tightness set by the padding fraction (2.7%), and the block variances equal $\operatorname{tr} \Sigma_{\Psi_{\mathrm{LPIPS}}}$ only on average over the signs.

Estimator and margins. The plug-in evaluates $\log_2 \hat{f}_n$ at each Monte Carlo point exactly, summing over all n_{est} centres with a numerically stable log-sum-exp. We draw $n_{\mathrm{MC}} = 2$ dither replicates per centre at $k = 1$ and $n_{\mathrm{MC}} = \max(k, 4)$ at block size $k \geq 2$. We set $\alpha = 0.05$, split it evenly between the two margins, and divide each by $|\mathcal{G}|$, separately for each representation. Each bound therefore holds at confidence $1 - \alpha$ on its own, and the pixel and LPIPS bounds hold jointly at $1 - 2\alpha$. Both margins are in bits of the block entropy $\bar{h}_U(k)$, the entropy of one k-dimensional block averaged over the blocks, so neither grows with d. Their cost to the final bound is the gap between $\rho^\star$ and its point-estimate version (thin dashed in Figures 3 and 7), which drops both margins and maximises over $\mathcal{G}$ again: at most 0.075 in pixel space and 0.047 in LPIPS space at $\varepsilon = 1$ (0.083 in pixel space on ImageNette), and 0.01–0.035 at $\varepsilon = 8$ and 16 (up to 0.046 at $\varepsilon = 32$, where $\rho^\star$ itself is only 0.13–0.19). The selected pairs $(k^\star, \nu^{2\star})$ are listed in Table 5.

G.3 The LPIPS representation

Ψ_{LPIPS} is the frozen LPIPS network (Zhang et al., 2018) with its AlexNet backbone (Krizhevsky et al., 2012). We concatenate its normalised feature maps, each rescaled by the square root of the LPIPS weights and the spatial size, so that $\|\Psi_{\mathrm{LPIPS}}(x) - \Psi_{\mathrm{LPIPS}}(x')\|^2$ is the LPIPS distance, giving $m = 31{,}872$ coordinates. Images are resized to 64×64 first, so the perceptual bound measures LPIPS at that resolution.

Ψ_{LPIPS} is a fixed public map, so it is admissible (Definition 3.5) provided its training images are disjoint from the records we protect. We do not report an LPIPS bound on ImageNette, whose images are drawn from the ImageNet training set on which the LPIPS features were trained.

The data-blind baseline. The prior mean $\mathbb{E}[\Psi_{\mathrm{LPIPS}}(x)]$ is the best data-blind guess in feature space, but it is not the feature vector of any image, so an attacker who must output an image cannot return it. For any image $\hat{x}$ chosen without seeing the release,

$$\mathbb{E}\|\Psi_{\mathrm{LPIPS}}(x) - \Psi_{\mathrm{LPIPS}}(\hat{x})\|^2 \;=\; \operatorname{tr} \Sigma_{\Psi_{\mathrm{LPIPS}}} + \|\Psi_{\mathrm{LPIPS}}(\hat{x}) - \mathbb{E}\Psi_{\mathrm{LPIPS}}(x)\|^2.$$

The first term does not depend on $\hat{x}$, so the optimal data-blind image is the one whose features lie closest to the prior mean. We find it by searching over images directly. Because the objective is not convex, we run the search from 16 different starting images, drawn from D_{shadow}, and keep the best result. From each, we run 400 steps of Adam on the pixels of $\hat{x}$ to minimise $\|\Psi_{\mathrm{LPIPS}}(\hat{x}) - \hat{\mu}_{\Psi_{\mathrm{LPIPS}}}\|^2$ plus a total-variation penalty of weight 10^{-4}, and we keep the final image whose features are closest to $\hat{\mu}_{\Psi_{\mathrm{LPIPS}}}$. Here $\hat{\mu}_{\Psi_{\mathrm{LPIPS}}}$ is the mean feature vector over 2,000 images of D_{shadow}, and the pixels are kept in the valid range by the reparametrisation $\hat{x} = \frac{1}{2}(\tanh u + 1)$. The gradient passes through the frozen network, so each step moves the image towards features closer to the mean. Descent may stop at a local minimum, so the error we report may slightly exceed that of the optimal data-blind image.

Because the second term is positive for every image, this baseline has error above $\operatorname{tr}\Sigma_{\Psi_{\mathrm{LPIPS}}}$. We normalise $\rho^\star_{\Psi_{\mathrm{LPIPS}}}$ by $\sqrt{\operatorname{tr}\Sigma_{\Psi_{\mathrm{LPIPS}}}}$, as in equation 3, which keeps $\rho^\star \le 1$. On that scale the image baseline sits slightly above 1 (1.131 on MNIST and 1.045 on CelebA), and we draw it as the grey dashed line in the LPIPS panels of Figures 3 and 7.

G.4 The privacy-spectrum experiment

The bound. We report results against ε, the quantity practitioners compare, rather than σ or γ. For each target ε we calibrate the noise multiplier σ with the PRV accountant (Gopi et al., 2021; Yousefpour et al., 2021) at $\delta = n_{\mathrm{train}}^{-1.1}$, with n_{train} the training-set size. Since q and σ are constant, every step has the same charge K^{PRV} in Theorem 4.1, so $\gamma = TK^{\mathrm{PRV}}/\ln 2$ needs a single one-dimensional integral per configuration, which we evaluate with a 256-node quadrature rule. Against a reference adaptive quadrature its error is at most 6.7×10^{-9} bits in γ over all configurations, six orders of magnitude below the smallest entropy margin per dimension (0.0054 bits), so it has no visible effect on $\rho^\star$.

Configuration. Our configuration follows common DP-SGD practice (Ponomareva et al., 2023; De et al., 2022) in its main ingredients: large Poisson-subsampled batches trained for many steps, GroupNorm, augmentation multiplicity and weight averaging (below). Hyperparameters are not tuned per privacy level: expected batch size $b = 1024$ for 100 epochs, giving $T = 4{,}900$ steps at $q = 0.0204$ (MNIST), $T = 4{,}400$ at $q = 0.0227$ (CIFAR-10, CelebA) and $T = 1{,}000$ at $q = 0.1$ (ImageNette), with a data-independent initialisation as Lemma D.1 requires. We clip at $C = 1$ on every dataset; neither γ nor $\rho^\star$ depends on C (Section 4).

Utility. We train a LeNet-5 (LeCun et al., 1998) on MNIST ($\sim$62K parameters), a Wide-ResNet-16-4 on CIFAR-10 ($\sim$2.7M) and a ResNet-18 on CelebA and ImageNette ($\sim$11M). All use GroupNorm (Wu & He, 2018) rather than BatchNorm, which would make each record's gradient depend on the others in its batch and break Lemma D.2. Training uses Opacus (Yousefpour et al., 2021) at the configuration above, with SGD at momentum 0.9 and cross-entropy loss. We use augmentation multiplicity 4 on CIFAR-10, CelebA and ImageNette (1, i.e. none, on MNIST) (De et al., 2022; Ponomareva et al., 2023), meaning that four augmentations of each example are drawn and their gradients averaged before clipping, so every record still contributes a single clipped gradient. We also keep an exponential moving average of the iterates during training (decay 0.999) and use those averaged weights when measuring accuracy, while training itself continues from the raw iterates (De et al., 2022; Ponomareva et al., 2023). Averaging cancels much of the noise DP-SGD injects, since that noise is independent across steps, and it is post-processing of released iterates, so it costs no privacy. The learning rate is fixed per dataset (0.5 on MNIST, 1.0 on CIFAR-10, 0.25 on CelebA and ImageNette) and reused at every privacy level, so the accuracies we report are attainable but not tuned per level. Every privacy level is trained with 3 seeds and we report the mean and standard deviation of test accuracy. The non-private baseline is the same architecture and optimiser trained without clipping or noise, with a single seed and its own learning rate, selected on D_{val} (0.05 on MNIST, CIFAR-10 and ImageNette, 0.1 on CelebA), and accuracy is reported on D_{test}.

Hyperparameter tuning. Any search over the clip norm or the learning rate reads the training data, so strictly their leakage composes with that of the final run (Lemma D.1) (assuming that an adversary can observe the noisy trajectories of these training runs). As is standard in the DP-SGD literature (Ponomareva et al., 2023), we do not include it, and the reported ε and $\rho^\star$ are those of the final run alone. Including it requires no new theory. Since γ reads only the schedule (q_t, σ_t), a search of k runs is simply a longer trajectory and adds the leakage of k runs, so a practitioner can either sum Theorem 4.1 over every run or move the search onto data outside the protected set (Section F.3).

G.5 The attack on DP-SGD

Model. We attack the model of Schwethelm et al. (2024): a single linear layer whose scalar output is the loss, $\ell(W; u) = Wu$, so that the per-example gradient is the input u itself, whatever the value of W. Nothing of the record is lost to the architecture, and since Theorem 4.1 holds for every model and loss at a given (q_t, σ_t), it is the model on which an attack can come closest to the bound. The

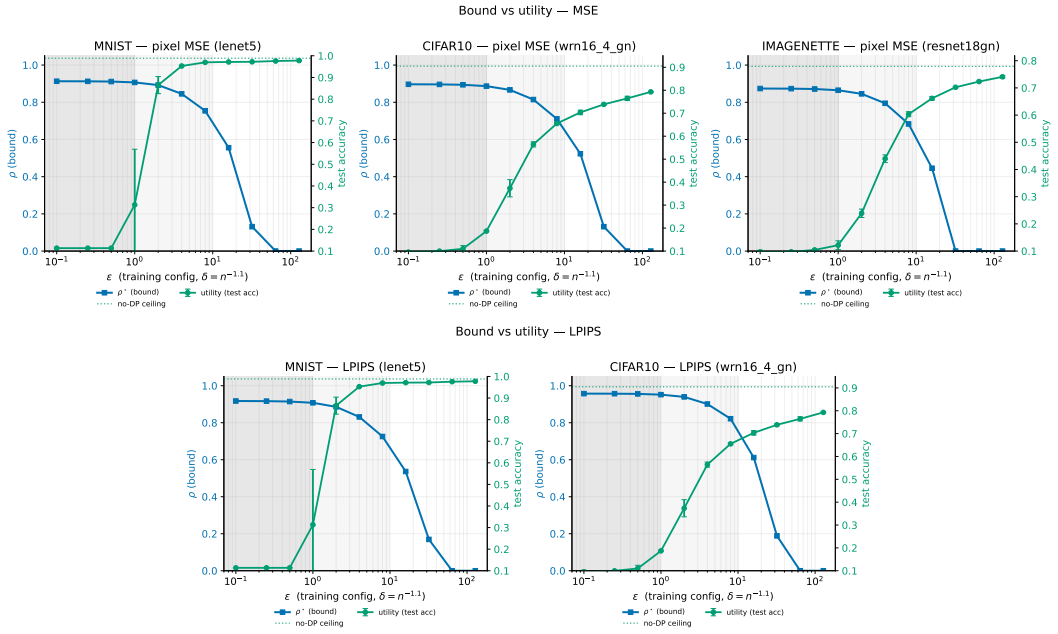


Figure 5: What a deployed DP-SGD run guarantees across the privacy spectrum, on MNIST, CIFAR-10 and ImageNette (CelebA is in Figure 2). At each target ε (expected batch 1024, 100 epochs, $C = 1$, $\delta = n_{\text{train}}^{-1.1}$): the pixel reconstruction bounds and, on MNIST and CIFAR-10, the perceptual bounds (left axis), each a 95% lower bound holding against every adversary of Section 3.1 and every architecture at that configuration, and the test accuracy the same run attains (right axis, mean $\pm$ standard deviation over 3 seeds; non-private baseline as the dotted horizontal line). Shaded: the $\varepsilon \leq 1$ (dark) and $1 < \varepsilon \leq 10$ (light) regimes of Ponomareva et al. (2023); $\varepsilon > 10$ is unshaded.

input is centred with a mean fixed on D_{shadow}: $u = x - \bar{x}$ in pixel space and $u = \Psi_{\text{LPIPS}}(x) - \bar{\Psi}$ in feature space, where the frozen feature map of Section G.3 is placed before the layer so that the attack estimates the features directly. The informed adversary knows the other records and which of them were sampled, so it subtracts their clipped gradients from the released sum. What remains of step t is

$$Y_t \;=\; \mathbf{1}_t\, a(u)\, u \;+\; \sigma C\, \xi_t, \tag{36}$$

the target's clipped gradient $a(u)u$, with $a(u) = \min\{1, C/\|u\|\}$ the clip factor, present or absent according to the target's participation $\mathbf{1}_t \sim \text{Bernoulli}(q)$, plus the mechanism's noise $\sigma C\, \xi_t$ with $\xi_t \sim \mathcal{N}(0, I)$. The adversary does not observe $\mathbf{1}_t$ ($\mathbf{1}_t \equiv 1$ at $q = 1$). We take $C = 0.3$, small enough that every gradient is clipped, as Hayes et al. (2023) do for the same reason (their App. K.1).

Attack. We use the prior-aware attack of Hayes et al. (2023), in the likelihood form of their Appendix G. The adversary holds candidates $z_1, \ldots, z_n$ and, for each, computes the probability of the releases if the target were z_j. Under that hypothesis each release is either $c_j = a(z_j)z_j$ plus noise, if the target was sampled at that step, or noise alone, so the probability is a product over steps of two-way mixtures,

$$p(Y_{1:T} \mid z_j) \;\propto\; \prod_{t=1}^{T} \Big[(1-q) + q\, e^{L_{tj}}\Big], \qquad L_{tj} = \frac{2\langle Y_t, c_j\rangle - \|c_j\|^2}{2\sigma^2 C^2}, \tag{37}$$

which is their score s_z with a general clip norm in place of their $C = 1$. Participation is integrated out exactly, step by step, and the clip factor of every candidate is known, so the score is in closed form for any q. We depart from Hayes et al. (2023) in two ways. Their adversary is given a shortlist that contains the target (their Section 3.4); ours holds the shadow sample D_{shadow} (58,000 records on MNIST, 44,000 on CelebA), which does not, so it knows the source only through a sample of it and is weaker than the adversary of Section 3.1. And where they report the candidate with the largest score (their Appendix G), we report the posterior mean

$$\hat{u} \;=\; \sum_{j=1}^{n} w_j\, z_j, \qquad w_j \;=\; \frac{p(Y_{1:T} \mid z_j)}{\sum_{k=1}^{n} p(Y_{1:T} \mid z_k)}, \tag{38}$$

the candidates averaged by how well each explains the releases, since the bound controls squared error and the posterior mean minimises it under the candidate prior. The reconstruction is $\bar{x} + \hat{u}$ in pixel space and $\bar{\Psi} + \hat{u}$ in feature space, and costs one matrix product per target.

Evaluation. Both configurations, $(q, T) = (1, 1)$ and $(0.02, 100)$, are run on the same 100 held-out targets (Section F.4), each attacked once, with σ solved so that Theorem 4.1 charges the target γ. A marker is the ratio equation 3 with the expectation replaced by the mean over targets, divided by the $\sqrt{\text{tr}\,\hat{\Sigma}}$ of Table 3, the same scale as the bound; in the LPIPS panels it is a feature-space error, which is what the bound controls, and the dashed lines show, for both configurations, the error after inverting $\hat{u}$ to an image by gradient descent on $\|\Psi_{\text{LPIPS}}(\hat{x}) - \hat{z}\|^2$. Two features of Figures 3 and 7 are worth stating. Subsampling weakens the attack at matched γ by up to 0.08 in ρ (paired differences -0.03 to $+0.08$; it is slightly stronger only on CelebA LPIPS at 72 bits), the cost of not knowing when the target participated, as Hayes et al. (2023) also observe. And at 72 bits the CelebA markers stop at 0.71–0.73 (pixel) and 0.91–0.93 (LPIPS), over the two configurations. The attack can only output a weighted average of shadow records, so it can get no closer to a held-out face than such an average can, and D_{shadow} holds no near duplicate of a held-out face. This limits the attack, not the bound, which depends on the source's entropy rather than on any sample of it, and is zero at that leakage in any case.

G.6 Tightness at fixed leakage

The bound equation 14 sees the mechanism only through γ, so it holds for every release that leaks at most γ bits about the record, whether or not that release is a model. It can therefore be tight at a given γ only against the release whose bits buy the most reconstruction, which is not DP-SGD. We construct one that is close to it, a vector quantiser, and use it to locate the slack of the bound. At each leakage we compute

$$\rho^\star \;\le\; \underbrace{\rho_{\text{blk}}}_{\text{Proposition G.1}} \;\le\; \underbrace{\rho_{\text{cb}}}_{\text{attack on the quantiser}} \;\le\; \underbrace{\rho_{\text{dpsgd}}}_{\text{attack on DP-SGD}}, \tag{39}$$

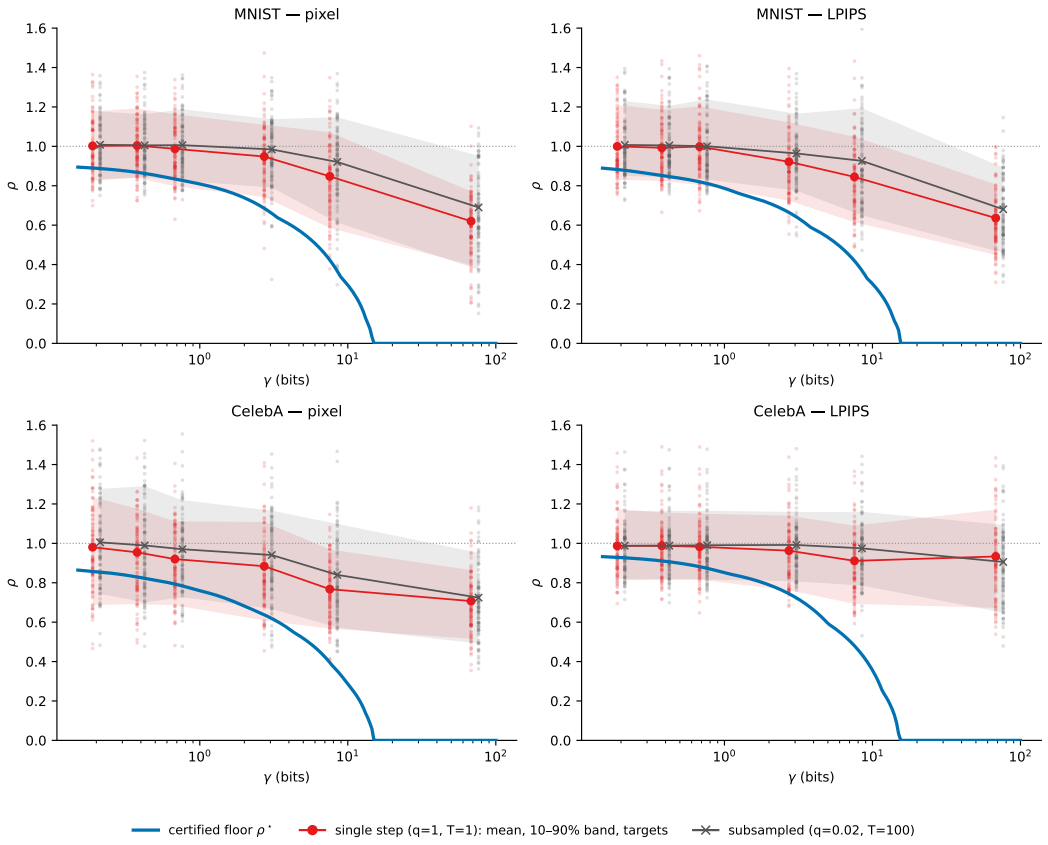


Figure 6: Per-target spread behind Figures 3 and 7, for the single-step (circles) and subsampled (crosses) attack. Lines: the mean over the 100 targets; shaded: the 10–90th percentile band of the per-target error at each leakage; faint points: every target. The bound is on expected error. At low leakage a sizeable fraction of individual targets fall below it (9–40% at $\gamma \leq 0.72$ bits, 1–11% at 2.89 bits, at most 3% at 8.01 bits), which the bound permits.

where ρ_{blk} relaxes the bound while remaining a lower bound on every attack of the quantiser (Table 7). Across the four panels of Figures 3 and 7 the first gap is 0.10–0.33 in ρ and the second at most 0.03.

The codebook mechanism. The distortion–rate function is the least expected squared error that a release of R bits permits. It is approached by quantisers, which release the index of the codeword nearest to the record and so spend every released bit on reducing squared error (Cover & Thomas, 2006, Ch. 10). We use a single-record quantiser of this kind. No model is trained: the mechanism releases the index $Q \in \{1, \dots, M\}$ of the codeword nearest to $\Psi(x_i)$, among $M = 2^R$ codewords in the space of Ψ. The codewords are built in two steps. First, k-means on D_{shadow} gives M centroids. Second, we assign D_{est} to these centroids with at most $\lceil n_{\mathrm{est}}/M \rceil$ records per cell (a single balanced step in the spirit of Malinen & Fränti (2014)), and take the cell means as the codewords. The size cap spreads the codewords so that every cell stays populated, which estimating the per-block leakage below requires. The codebook is fixed independently of the held-out targets and is known to the adversary; for LPIPS the same construction is applied to $\Psi_{\mathrm{LPIPS}}(x)$. Every bound below holds for any codebook, and a worse codebook only widens the measured gaps, so its optimality is not at issue. Since Q is a deterministic function of x_i with M values, independent of X^{-i}, the release leaks $I(x_i; Q \mid X^{-i}) = H(Q) \le R$ bits. We charge $\gamma = R$, which slightly overstates the leakage (the realised $H(Q)$ is within 0.07 bits of R) and so only lowers $\rho^\star$. Below one bit, which a deterministic codebook cannot reach, we take the two-cell codebook and pass Q through a binary symmetric channel: the release Y equals Q with probability $p \in [\frac{1}{2}, 1]$ and the other index otherwise, independently of everything else. Because Y depends on x_i only through Q, it leaks exactly $\gamma = I(Q;Y) = H_b\big(f_0 p + f_1(1-p)\big) - H_b(p)$ bits (Cover & Thomas, 2006, Ch. 7), where f_0, f_1 are the cell frequencies, and every $\gamma \in (0, H_b(f_0)]$ is reached by choice of p.

The codebook attack. The adversary sees the release Z ($Z = Q$ for the deterministic codebook, $Z = Y$ for the randomised one) and the mechanism does not involve X^{-i}, so its reconstruction is a function $g(Z)$. Under squared error the optimal g is the conditional mean $\mathbb{E}[\Psi(x_i) \mid Z]$ (Bishop, 2006). For the deterministic codebook this is the mean of the released cell, and for the randomised codebook it is the average of the two cell means, each weighted by the probability that it is the true cell given the release. We estimate the cell means and frequencies on D_{est} and report the error on targets held out from D_{shadow} and D_{est}. The estimate is a function of the release and of records disjoint from the target, so it is a valid reconstruction of Section 3.1, and estimation error can only raise its error above the optimum, which widens rather than narrows the measured gap.

The bounds. In the proof of Corollary 3.7 the mechanism enters in exactly one place: the rate at which each block's Shannon lower bound is taken, every block being charged the whole of γ. Write $\tilde{z}_b$ for the dithered block $\tilde{z}_i^{\,\Pi_b \Psi}$ and $\gamma_b := I(\tilde{z}_b; \theta \mid X^{-i})$ for the leakage block b receives, which for the quantiser is $I(\tilde{z}_b; Q) = h(\tilde{z}_b) - \sum_c \Pr(Q = c)\, h(\tilde{z}_b \mid Q = c)$. Both bounds in equation 39 have the form

$$\rho^2 \;\ge\; \frac{k}{2\pi e \,\operatorname{tr}\Sigma_\Psi} \cdot \frac{1}{N_k} \sum_{b=1}^{N_k} F(b) \;-\; \nu^2, \tag{40}$$

with $F_{\mathrm{cert}} = 2^{\,2(h(\tilde{z}_b) - \gamma)/k}$ the bound before its Jensen step and $F_{\mathrm{blk}} = 2^{\,2(h(\tilde{z}_b) - \gamma_b)/k}$ charging each block only the leakage it received. The difference is the *per-block charge*.

Proposition G.1 (Per-block bound)**.** *In the setting of Corollary 3.7, $\gamma_b \le \gamma_i$ for every b, and equation 40 holds with F_{blk}. In particular $F_{\mathrm{cert}} \le F_{\mathrm{blk}}$ at any common (k, ν^2).*

Proof. $\tilde{z}_b \to x_i \to (\theta, X^{-i})$ is a Markov chain (Definition 3.5), so $\gamma_b \le \gamma_i$ by data processing, and $F_{\mathrm{cert}} \le F_{\mathrm{blk}}$ follows. The proof of Theorem 3.6 at the representation $\Pi_b \circ \Psi$ uses the leakage only in Step 1, as $I(\tilde{z}_b; \Pi_b \hat{z}_i^{\,\Psi}) \le \gamma_i$, and data processing along $\tilde{z}_b \to (\theta, X^{-i}) \to \Pi_b \hat{z}_i^{\,\Psi}$ gives the sharper $I(\tilde{z}_b; \Pi_b \hat{z}_i^{\,\Psi}) \le \gamma_b$. Steps 2–4 then give $\mathbb{E}\|\Pi_b(\Psi(x_i) - \hat{z}_i^{\,\Psi})\|^2 \ge \frac{k}{2\pi e} 2^{2(h(\tilde{z}_b) - \gamma_b)/k} - k s^2$ for every block, and averaging over b by equation 21 and dividing by $\operatorname{tr}\Sigma_\Psi$ gives equation 40 with F_{blk}. □

Why the relaxed bounds use $k = 1$. The per-block bound is evaluated at $k = 1$. Its exponent carries γ_b rather than γ, so nothing pushes it towards large k, and one-dimensional plug-ins suffice for $\gamma_b = h(\tilde{z}_b) - \sum_c \Pr(Q = c)\, h(\tilde{z}_b \mid Q = c)$, which is why every cell must hold enough records.

Interpretation. The chain splits the distance from the bound to the DP-SGD attack into three gaps. The first, from $\rho^\star$ to ρ_{blk}, is 0.10–0.33 across the four panels of Figures 3 and 7. It is the cost of charging every block the full γ. Removing it would require either a per-block leakage bound for the mechanism or a lower bound on the entropy of the whole record, and the plug-in caps the latter at $\log_2 n$ bits above the dither entropy (Section E.1). The second, from ρ_{blk} to ρ_{cb}, is at most 0.03 (Table 8, last column), and slightly negative, within standard error, on CelebA pixels at $R = 3$–5. Since ρ_{blk} lower-bounds every attack on the codebook, this small gap means both that the Shannon lower bound is close to tight for these sources and that our codebook attack is close to optimal. The third, from ρ_{cb} to ρ_{dpsgd} (Tables 6 and 8), is specific to DP-SGD. The leakage γ fixes how many bits are released, not which: the codebook spends its bits on a cell index chosen for reconstruction, whereas DP-SGD spends them on a noisy clipped gradient spread over every direction, so the same γ buys far less reconstruction. This gap is the margin by which Figures 2 and 5 are conservative for DP-SGD. It may also include slack in the accountant, since γ upper-bounds DP-SGD's true leakage, and the suboptimality of our attack, neither of which the chain can separate from the mechanism.

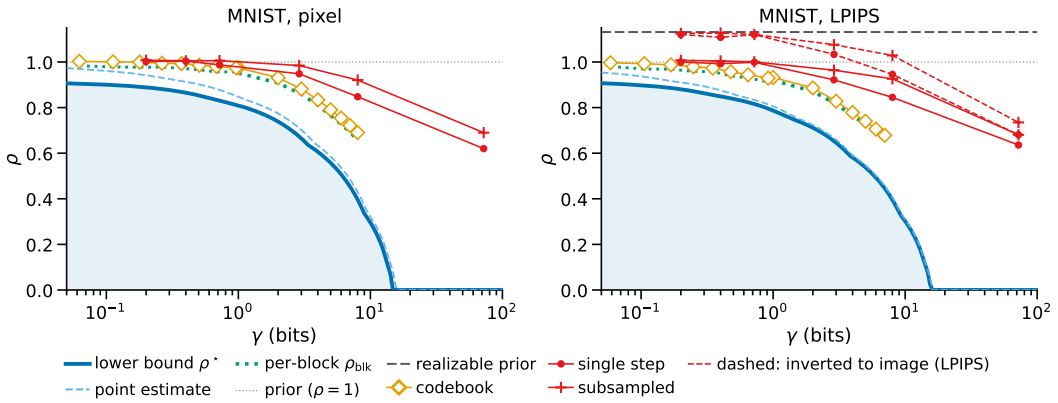


Figure 7: Tightness at fixed leakage on MNIST, in pixel (left) and LPIPS (right) error, as in Figure 3: the chain equation 39, with terms as in Table 7. No attack crosses the bound. ρ_{blk} is drawn while every codebook cell holds 20 estimation records. Dotted: the prior at $\rho = 1$. On MNIST the attack reaches lower ρ than on CelebA at high leakage because the shadow sample contains near-duplicates of many held-out digits (Figures 8 and 9).

Table 1: The reconstruction-guarantee lineage in order of first public release, split by bounded metric: probability guarantees ($\Pr$) above the double rule, expected-error guarantees ($\mathbb{E}$) below. *Threat Model* gives background knowledge (**Informed**: adversary knows all training records but the target; **Uninformed**: no background dataset knowledge; **Target-informed**: adds target-specific auxiliary data) and access (**Final release**: the mechanism's final output only; **All iterates**: intermediate updates or gradients). *Prior Integration* indicates whether a prior over the target enters the guarantee itself, and *Prior Assumption and Source* whether it is assumed (**Analytic**) or approximated from data (**Empirical**); structural assumptions on the estimator or data space are noted in the same cell. *Baseline success* unites disparate notations for the maximum probability of guessing the target from prior knowledge alone; Guerra-Balboa et al. (2026) instead use the collision probability $\kappa_\pi = \Pr_{Z,Z'\sim\pi}[Z = Z']$, which coincides with a baseline success only under a uniform prior. $\Delta\Pr$ denotes a guarantee stated as an advantage — the increase in success attributable to the target's participation — rather than a raw probability. Experimental cells list representative settings, not an exhaustive inventory. Abbreviations: LDP/RDP/f-DP (Local, Rényi, and f-differential privacy), CMI (conditional mutual information), LCB (lower confidence bound), HCRB (Hammersley–Chapman–Robbins bound), CRB (Cramér–Rao bound), FIL (Fisher information leakage), dFIL (diagonal FIL), RR (randomized response).

	Threat Model	Mechanism Leakage Bound	Bounded Object (Metric)	Prior Integration into Bound	Prior Assumption and Source	Experimental Instantiation
Bhowmick et al. (2018)	Uninformed, all iterates	ε-LDP: e^ε scaling	$\Pr[\text{error} \le \eta]$	Baseline success p	Analytic: Uniform/Zipfian	MNIST, CIFAR-10, Flickr, Reddit: Accuracy gap
Balle et al. (2022)	Informed, final release	RDP/ε-DP: e^ε scaling	$\Pr[\text{error} \le \eta]$	Baseline success κ	Analytic: Uniform/Gaussian	NA
Xiao & Devadas (2023)	Uninformed, final release	PAC Privacy: $I(X;\mathcal{M}(X)+B)$	$\Pr[\text{error} \le \eta]$	Baseline success δ	Analytic: Uniform population	CIFAR-10, MNIST: Noise magnitude
Guo et al. (2023)	Informed, final release	RDP + Fano: Generalized Fano	$\Pr[\text{exact}]$	Baseline success p^*, Entropy $H(p)$	Analytic: Prob. vector p	Synthetic (RR, Gaussian), IWPC, GPT-2: Advantage, MSE
Hayes et al. (2023)	Informed, all iterates	ReRo: Blow-up function	$\Pr[\text{exact}]$	Baseline success κ	Analytic: Uniform shortlist	MNIST, CIFAR-10: Success probability
Kaissis et al. (2023)	Informed, final release	f-DP: Closed-form / Edgeworth	$\Pr[\text{exact}]$	Baseline success κ	Analytic: Uniform $1/n$ prior	Numerical: success bound vs. κ at published DP-SGD hyperparameters
Swanberg et al. (2025)	Informed (n targets), final release	Pure/f-DP: e^ε tail	$\Pr[\text{count of hits}]$	Baseline success π	Analytic: Product distribution $\mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_n$	LLM canaries, tabular: Advantage, Success
Kulynych et al. (2026)	Informed, final release	f-DP: Trade-off (Thm. 3.1); TV (Thm. 3.3)	$\Pr[\text{attack success}]$	Baseline success $base$	Analytic: Scalar worst-case	US Census, GPT-2: Advantage, Accuracy
Guerra-Balboa et al. (2026)	Target-informed, final release	TV (Thm. 4.2); density difference (Thm. 4.3)	$\Delta\Pr[\text{error} \le \eta]$	Collision probability κ_π	Uniform/Data-estimated frequencies	MNIST, Adult, Porto: RAD
Guo et al. (2022)	Informed, final release	RDP: HCRB (Thm. 1); FIL: CRB (Thm. 2)	$\mathbb{E}[\text{MSE}]$	No prior; domain geometry / Fisher info	Analytic: bounded $\mathcal{Z}$; regularity; non-vacuous only for unbiased attacks	MNIST, CIFAR-10: MSE lower bound; dFIL vs. reconstruction
This paper	Informed, all iterates	CMI: KL of dominating pair	$\mathbb{E}[\text{MSE}]$	Differential entropy	Empirical: LCB from sample	MNIST–ImageNette: MSE, LPIPS

Table 2: The partitions of Figure 4, and what constrains each. The requirement column refers to Section F.1; ⋆ marks a disjointness the guarantee or the threat model requires, the rest are kept for a clean joint statement and could be relaxed.

Split	Supplies	Requirement	Disjoint from	Size and property
shared — the source side, computed once per dataset				
$D_\perp$	$U = d^{-1/2} H_d V^\top$, $\operatorname{tr} \hat{\Sigma}$, dither scales	Req 3, Req 4	$D^\star_{\text{est}}$, $D^\star_{\text{targets}}$	$\gtrsim d$, for a full-rank $\hat{\Sigma}$; may be replaced by a data-free U and a declared $\operatorname{tr} \Sigma$
D_{est}	$\bar{h}_{\text{LB}}$	Req 5	$D^\star_\perp$	largest available; i.i.d. from P_x; margins $\propto n^{-1/2}$
(a) deployment, and the privacy-spectrum runs				
D_{train}	every charged DP-SGD run	Req 1, Req 2	D_{val}, D_{test}	i.i.d. from P_x; no cross-record preprocessing; may overlap $D_\perp$ and D_{est}
D_{val}	selection among runs	Req 6	D_{train}	small
D_{test}	reported accuracy	Req 7	D_{train}, D_{val}	small
(b) the attack experiment				
D_{shadow}	candidate set, centring mean	Req 8, Req 9	$D^\star_{\text{targets}}$	most of D_{est}
D_{targets}	the reconstructed records	Req 10	$D^\star_\perp$, $D^\star_{\text{shadow}}$	100; i.i.d. from P_x; enough for the standard error of the mean; may lie in D_{est}

Table 3: The reference scale and how far its finite-sample one-sided 95% upper confidence bound (Section G.2) lies above it, relative to the point estimate, estimated on $D_\perp$ (sizes in Section G.1). $\text{RMSE}_{\text{prior}}/\sqrt{d}$ is the per-coordinate root-mean-squared error of the prior-mean guess, in grey levels for pixels and feature units for Ψ_{LPIPS} ($m = 31{,}872$).

	pixel ($\Psi = \text{id}$)			perceptual ($\Psi = \Psi_{\text{LPIPS}}$)		
Dataset	$\operatorname{tr} \Sigma$	UCB margin	$\text{RMSE}_{\text{prior}}/\sqrt{d}$	$\operatorname{tr} \Sigma_{\Psi_{\text{LPIPS}}}$	UCB margin	$\text{RMSE}_{\text{prior}}/\sqrt{m}$
MNIST	3.43×10^6	2.5%	66.2	0.1338	5.8%	2.05×10^{-3}
CIFAR-10	1.24×10^7	3.2%	63.6	0.2463	3.3%	2.78×10^{-3}
CelebA	5.91×10^7	0.5%	69.3	0.1445	0.6%	2.13×10^{-3}
ImageNette	2.46×10^8	16.6%	70.7	—	—	—

Table 4: Values behind Figures 2 and 5: test accuracy (mean over 3 seeds) and the pixel and perceptual bounds at each target ε, each a 95% lower bound. †: the $\varepsilon > 10$ band of Ponomareva et al. (2023).

	MNIST			CIFAR-10			CelebA			ImageNette	
ε	acc.	$\rho^\star_{\text{px}}$	$\rho^\star_{\Psi_{\text{LPIPS}}}$	acc.	$\rho^\star_{\text{px}}$	$\rho^\star_{\Psi_{\text{LPIPS}}}$	acc.	$\rho^\star_{\text{px}}$	$\rho^\star_{\Psi_{\text{LPIPS}}}$	acc.	$\rho^\star_{\text{px}}$
1	0.313	0.907	0.908	0.187	0.887	0.952	0.923	0.888	0.944	0.122	0.865
2	0.865	0.892	0.886	0.373	0.867	0.940	0.946	0.861	0.931	0.239	0.845
4	0.953	0.845	0.831	0.564	0.815	0.901	0.954	0.805	0.892	0.440	0.795
8	0.970	0.754	0.726	0.655	0.711	0.822	0.957	0.696	0.801	0.603	0.683
$16^\dagger$	0.971	0.556	0.536	0.703	0.523	0.612	0.961	0.513	0.588	0.662	0.445
$32^\dagger$	0.972	0.131	0.169	0.739	0.131	0.188	0.967	0.131	0.174	0.702	0.000
$64^\dagger$	0.976	0.000	0.000	0.765	0.000	0.000	0.963	0.000	0.000	0.723	0.000
non-priv.	0.989	—	—	0.905	—	—	0.971	—	—	0.779	—

Table 5: Mechanism internals behind Figures 2 and 5: the leakage each configuration incurs, the noise multiplier achieving it, and the binding pair. A dash: the bound is zero at that leakage; n/a: LPIPS is not evaluated on ImageNette. $k^\star$ grows with γ overall up to single grid steps on CelebA.

Dataset	ε	γ (bits)	σ	$k^\star$ (pix / Ψ_{LPIPS})		$\nu^{2\star}$ (pix / Ψ_{LPIPS})	
MNIST	1	0.048	5.574	14	16	1.6	3.2
	2	0.169	3.028	16	32	1.6	6.4
	4	0.570	1.741	16	32	1.6	3.2
	8	1.785	1.110	28	32	1.6	1.6
	16	5.042	0.796	28	32	0.8	0.8
	32	13.150	0.617	28	32	0.4	0.4
	64	32.883	0.497	–	–	–	–
CIFAR-10	1	0.049	5.838	6	16	0.8	1.6
	2	0.171	3.167	6	32	0.4	3.2
	4	0.577	1.815	12	32	0.8	3.2
	8	1.806	1.147	32	32	1.6	1.6
	16	5.108	0.816	32	32	0.8	0.8
	32	13.317	0.629	32	32	0.4	0.4
	64	33.222	0.504	–	–	–	–
CelebA	1	0.049	5.838	3	16	0.2	1.6
	2	0.171	3.167	6	32	0.4	3.2
	4	0.577	1.815	8	32	0.4	3.2
	8	1.806	1.147	16	32	0.8	1.6
	16	5.108	0.816	32	32	0.8	0.8
	32	13.317	0.629	24	32	0.2	0.4
	64	33.222	0.504	–	–	–	–
ImageNette	1	0.060	10.969	24	n/a	3.2	n/a
	2	0.207	5.941	24	n/a	3.2	n/a
	4	0.682	3.313	32	n/a	3.2	n/a
	8	2.111	1.948	32	n/a	1.6	n/a
	16	6.010	1.242	32	n/a	0.8	n/a
	32	15.682	0.863	–	–	–	–
	64	38.345	0.639	–	–	–	–

Table 6: The attack against the bound: ρ over the 100 targets, bootstrap standard errors between 0.013 and 0.024.

	γ (bits)	0.72	2.89	8.01	72.13
MNIST, pixel	$q = 1$	0.987	0.948	0.848	0.620
	$(q, T) = (0.02, 100)$	1.006	0.984	0.921	0.690
	bound	0.831	0.673	0.397	0
MNIST, LPIPS	$q = 1$	0.999	0.922	0.845	0.636
	$(q, T) = (0.02, 100)$	1.001	0.964	0.926	0.681
	bound	0.816	0.653	0.396	0
CelebA, pixel	$q = 1$	0.920	0.884	0.768	0.707
	$(q, T) = (0.02, 100)$	0.970	0.941	0.840	0.725
	bound	0.789	0.628	0.373	0
CelebA, LPIPS	$q = 1$	0.983	0.963	0.911	0.933
	$(q, T) = (0.02, 100)$	0.991	0.992	0.975	0.906
	bound	0.878	0.734	0.459	0

Table 7: The chain equation 39: what each term is, how it is evaluated, and what the gap between consecutive terms measures.

	$F(b)$ in equation 40	Lower-bounds	How evaluated	Gap to the row above
$\rho^\star$	$F_{\mathrm{cert}} = 2^{\,2(h(\tilde{z}_b)-\gamma)/k}$	every attack on every release leaking $\leq \gamma$	Proposition 5.1 at $\gamma = R$, with its grid and margins	—
ρ_{blk}	$F_{\mathrm{blk}} = 2^{\,2(h(\tilde{z}_b)-\gamma_b)/k}$	every attack on this release	$k = 1$, $\nu^2 = 0.2$, one-dimensional plug-ins equation 31 on D_{est} for $h(\tilde{z}_b)$ and each $h(\tilde{z}_b \mid Q = c)$; point estimate; every cell must hold 20 records, so $R \leq 8$ (pixel) and $R \leq 7$ (LPIPS)	per-block charge
ρ_{cb}	—	—	cell-mean decoder; error on held-out targets, bootstrap standard error	residual slack of the Shannon lower bound at rate γ_b, plus sampling error of the targets
ρ_{dpsgd}	—	—	posterior mean under the empirical shadow prior (Section G.5) on one full-batch step, at $\gamma = 1/(2 \ln 2\, \sigma^2)$ bits, and on $T = 100$ steps at $q = 0.02$	what DP-SGD fails to convert into reconstruction

Table 8: Values behind Figures 3 and 7, one block per panel. Gaps are computed before rounding.

R	$\rho^\star$	ρ_{blk}	ρ_{cb} (SE)	$\rho_{\mathrm{blk}} - \rho^\star$	$\rho_{\mathrm{cb}} - \rho_{\mathrm{blk}}$
MNIST, pixel					
1	0.809	0.951	0.975 (0.005)	0.142	0.023
2	0.739	0.908	0.931 (0.005)	0.169	0.023
3	0.665	0.859	0.882 (0.005)	0.194	0.023
4	0.606	0.811	0.834 (0.005)	0.205	0.023
5	0.558	0.765	0.790 (0.005)	0.207	0.026
6	0.507	0.728	0.754 (0.005)	0.220	0.027
7	0.454	0.692	0.721 (0.005)	0.238	0.029
8	0.397	0.658	0.691 (0.005)	0.261	0.032
CelebA, pixel					
1	0.760	0.858	0.862 (0.006)	0.098	0.004
2	0.682	0.809	0.810 (0.005)	0.127	0.001
3	0.622	0.772	0.770 (0.005)	0.149	−0.002
4	0.566	0.737	0.737 (0.005)	0.171	0.000
5	0.518	0.711	0.709 (0.005)	0.193	−0.002
6	0.472	0.687	0.687 (0.005)	0.215	0.000
7	0.424	0.662	0.667 (0.005)	0.239	0.005
8	0.374	0.641	0.651 (0.004)	0.267	0.010
MNIST, LPIPS					
1	0.787	0.920	0.932 (0.004)	0.133	0.012
2	0.712	0.871	0.886 (0.004)	0.159	0.015
3	0.646	0.815	0.828 (0.005)	0.169	0.013
4	0.581	0.765	0.778 (0.005)	0.184	0.013
5	0.538	0.725	0.740 (0.005)	0.186	0.015
6	0.493	0.688	0.706 (0.004)	0.195	0.017
7	0.446	0.657	0.678 (0.004)	0.211	0.021
CelebA, LPIPS					
1	0.851	0.962	0.980 (0.005)	0.112	0.017
2	0.789	0.939	0.955 (0.005)	0.150	0.016
3	0.727	0.917	0.933 (0.005)	0.190	0.016
4	0.662	0.898	0.912 (0.005)	0.235	0.015
5	0.593	0.878	0.895 (0.005)	0.285	0.017
6	0.550	0.857	0.878 (0.005)	0.307	0.020
7	0.506	0.839	0.862 (0.005)	0.333	0.023

G.7 ADDITIONAL FIGURES



Figure 8: **Attack reconstructions at each leakage, CelebA.** As Figure 9, for CelebA in pixel space (top) and LPIPS space (bottom). D_{shadow} holds no near duplicate of a held-out face, so the pixel ρ levels off near 0.71 beyond $\gamma \approx 70$ bits.

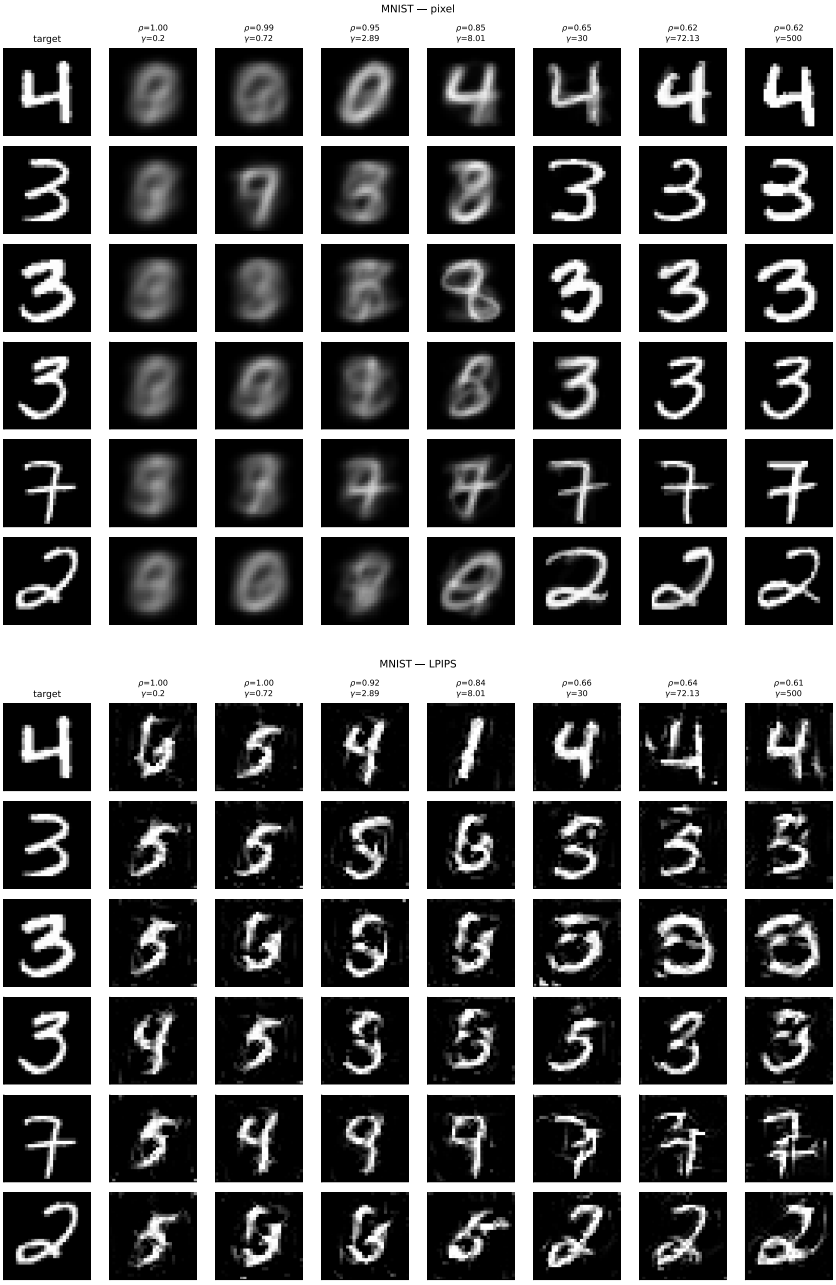


Figure 9: **Attack reconstructions at each leakage, MNIST.** Outputs of the single-step attack of Section G.5 on 6 fixed held-out targets at 7 leakages, $\gamma \in \{0.2, 0.72, 2.89, 8.01, 30, 72.13, 500\}$ bits, in pixel space (top) and LPIPS space (bottom). Each column shows what reconstruction at its value of ρ looks like in practice, so the bounds of Figures 2 and 5 can be read as images rather than numbers. In each panel the first column is the target and each further column is one leakage, headed by ρ, the mean over all 100 targets at that leakage. The LPIPS panel shows the feature estimate inverted to an image, headed by the feature-space ρ. In pixel space most reconstructed digits are wrong at $\rho = 0.85$ ($\gamma = 8$ bits) and all six are right at $\rho = 0.65$ ($\gamma = 30$ bits). The attack outputs a weighted mix of shadow images, none of them close to a held-out target, so ρ levels off near 0.62 beyond $\gamma \approx 30$ bits.