

---

# Bounding Training Data Reconstruction in DP-SGD

---

**Jamie Hayes\***  
Google DeepMind  
jamhay@google.com

**Saeed Mahloujifar\*†**  
Meta AI  
saeedm@meta.com

**Borja Balle**  
Google DeepMind  
bballe@google.com

## Abstract

Differentially private training offers a protection which is usually interpreted as a guarantee against membership inference attacks. By proxy, this guarantee extends to other threats like reconstruction attacks attempting to extract complete training examples. Recent works provide evidence that if one does not need to protect against membership attacks but *instead* only wants to protect against training data reconstruction, then utility of private models can be improved because less noise is required to protect against these more ambitious attacks. We investigate this further in the context of DP-SGD, a standard algorithm for private deep learning, and provide an upper bound on the success of any reconstruction attack against DP-SGD together with an attack that empirically matches the predictions of our bound. Together, these two results open the door to fine-grained investigations on how to set the privacy parameters of DP-SGD in practice to protect against reconstruction attacks. Finally, we use our methods to demonstrate that different settings of the DP-SGD parameters leading to the same DP guarantees can result in significantly different success rates for reconstruction, indicating that the DP guarantee alone might not be a good proxy for controlling the protection against reconstruction attacks.

## 1 Introduction

Machine learning models can and do leak training data (Ippolito et al., 2022; Kandpal et al., 2022; Carlini et al., 2019; Yeom et al., 2018; Song & Shmatikov, 2019; Tirumala et al., 2022). If the training data contains private or sensitive information, this can lead to information leakage via a variety of different privacy attacks (Balle et al., 2022; Carlini et al., 2022; Fredrikson et al., 2015; Carlini et al., 2021). Perhaps the most commonly studied privacy attack, membership inference (Homer et al., 2008; Shokri et al., 2017), aims to infer if a sample was included in the training set, which can lead to a privacy violation if inclusion in the training set is in and of itself sensitive. Membership inference leaks a single bit of information about a sample – whether that sample was or was not in the training set – and so any mitigation against this attack also defends against attacks that aim to reconstruct more information about a sample, such as training data reconstruction attacks (Balle et al., 2022; Carlini et al., 2021; Zhu et al., 2019).

Differential privacy (DP) (Dwork et al., 2006) provides an effective mitigation that provably bounds the success of *any* membership inference attack, and so consequently any training data reconstruction attack. The strength of this mitigation is controlled by a privacy parameter  $\epsilon$  which, informally, represents the number of bits that can be leaked about a training data sample, and so  $\epsilon$  must be small to guarantee the failure of a membership inference attack (Sablayrolles et al., 2019; Nasr et al., 2021). Unfortunately, training machine learning models with DP in the small  $\epsilon$  regime usually produces models that perform significantly worse than their non-private counterpart (Tramèr & Boneh, 2021; De et al., 2022).

---

\*Equal contribution.

†Work done at Princeton University.

Membership inference may not always be the privacy attack that is of most concern. For example, in social networks, participation is usually public; recovering privately shared photos or messages from a model trained on social network data is the privacy violation. These kinds of attacks are referred to as training data reconstruction attacks, and have been successfully demonstrated against a number of machine learning models including language models (Carlini et al., 2021; Mireshghallah et al., 2022), generative models (Somepalli et al., 2022; Carlini et al., 2023), and image classifiers (Balle et al., 2022; Haim et al., 2022). Recent work (Bhowmick et al., 2018; Balle et al., 2022; Guo et al., 2022a,b; Stock et al., 2022) has begun to provide evidence that if one is willing to forgo protection against membership inference, then the  $\epsilon$  regime that protects against training data reconstruction is far larger, as predicted by the intuitive reasoning that successful reconstruction requires a significant number of bits about an individual example to be leaked by the model. This also has the benefit that models trained to protect against training data reconstruction *but not* membership inference do not suffer as large a drop in performance, as less noise is added during training. Yet, the implications of choosing a concrete  $\epsilon$  for a particular application remain unclear since the success of reconstruction attacks can vary greatly depending on the details of the threat model, the strength of the attack, and the criteria of what constitutes a successful reconstruction.

In this paper we re-visit the question of training data reconstruction against image classification models trained with DP-SGD (Song et al., 2013; Abadi et al., 2016), the workhorse of differentially private deep learning. We choose to concentrate our analysis on DP-SGD because state-of-the-art results are almost exclusively obtained with DP-SGD or other privatized optimizers (De et al., 2022; Cattan et al., 2022; Mehta et al., 2022). Our investigation focuses on attacks performed under a strong threat model where the adversary has access to intermediate gradients and knowledge of all the training data except the target of the reconstruction attacks. This threat model is consistent with the privacy analysis of DP-SGD (Abadi et al., 2016) and the informed adversary implicit in the definition of differential privacy (Nasr et al., 2021; Balle et al., 2022), and implies that conclusions about the impossibility of attacks in this model will transfer to weaker, more realistic threat models involving real-world attackers. Our investigation focuses on three main questions. 1) How do variations in the threat model (e.g. access to gradients and side knowledge available to the adversary) affect the success of reconstruction attacks? 2) Is it possible to obtain upper bounds on reconstruction success for DP-SGD that match the best known attacks and thus provide actionable insights into how to tune the privacy parameters in practice? 3) Does the standard DP parameter  $\epsilon$  provide enough information to characterize vulnerability against reconstruction attacks?

Our contributions are summarized as follows:

- We illustrate how changes in the threat model for reconstruction attacks against image classification models can significantly influence their success by comparing attacks with access to the final model parameters, access to intermediate gradients, and access to prior information.
- We obtain a tight upper bound on the success of any reconstruction attack against DP-SGD with access to intermediate gradients and prior information. Tightness is shown by providing an attack whose reconstruction success closely matches our bound’s predictions.
- We provide evidence that the DP parameter  $\epsilon$  is not sufficient to capture the success of reconstruction attacks on DP-SGD by showing that different configurations of DP-SGD’s hyperparameters leading to the same DP guarantee lead to different rates of reconstruction success.

## 2 Training Data Reconstruction in DP-SGD

We start by introducing DP-SGD, the algorithm we study throughout the paper, and then discuss reconstruction attacks with access to either only the final trained model, or all intermediate gradients. Then we empirically compare both attacks and show that gradient-based attacks are more powerful than model-based attacks, and we identify a significant gap between the success of the best attack and a known lower bound. As a result, we propose the problem of closing the gap between theoretical bounds and empirical reconstruction attacks.

**Differential privacy and DP-SGD.** Differential privacy (Dwork et al., 2006) formalizes the idea that data analysis algorithms whose output does not overly depend on any individual input data point

can provide reasonable privacy protections. Formally, we say that a randomized mechanism  $M$  satisfies  $(\epsilon, \delta)$ -differential privacy (DP) if, for any two datasets  $D, D'$  that differ by one point, and any subset  $S$  of the output space we have  $P[M(D) \in S] \leq e^\epsilon P[M(D') \in S] + \delta$ . Informally, this means that DP mechanisms bound evidence an adversary can collect (after observing the output) about whether the point where  $D$  and  $D'$  differ was used in the analysis. For the bound to provide a meaningful protection against an adversary interested in this membership question it is necessary to take  $\epsilon \approx 1$  and  $\delta \leq 1/|D|$ .

A differentially private version of stochastic gradient descent useful for training ML models can be obtained by bounding the influence of any individual sample in the trained model and masking it with noise. The resulting algorithm is called DP-SGD (Abadi et al., 2016) and proceeds by iteratively updating parameters with a privatized gradient descent step. Given a sampling probability  $q$ , current model parameters  $\theta$  and a loss function  $\ell(\theta, \cdot)$ , the privatized gradient  $g$  is obtained by first creating a mini-batch  $B$  including each point in the training dataset with probability  $q$ , summing the  $L_2$ -clipped gradients<sup>3</sup> for each point in  $B$ , and adding Gaussian noise with standard deviation  $\sigma \cdot C$  to all coordinates of the gradient:  $g \leftarrow \sum_{z \in B} \text{clip}_C(\nabla_{\theta} \ell(\theta, z)) + \mathcal{N}(0, C^2 \sigma^2 I)$ . Running DP-SGD for

$T$  training steps yields an  $(\epsilon, \delta)$ -DP mechanism with  $\epsilon \approx \frac{q\sqrt{T \log(1/\delta)}}{\sigma}$  (Abadi et al., 2016) – in practice, tighter numerical bounds on  $\epsilon$  are often used (Gopi et al., 2021; Google DP Team, 2022). For analytic purposes it is often useful to consider alternatives to  $(\epsilon, \delta)$ -DP to capture the differences between distributions  $M(D)$  and  $M(D')$ . Rényi differential privacy (RDP) (Mironov, 2017) is one such alternative often used in the context of DP-SGD. It states that the mechanism is  $(\alpha, \epsilon)$ -RDP for  $\alpha > 1$  and  $\epsilon \geq 0$  if  $\mathbb{E}_{W \sim M(D')}$   $\left[ \left( \frac{\mathbb{P}[M(D)=W]}{\mathbb{P}[M(D')=W]} \right)^\alpha \right] \leq e^{(\alpha-1)\epsilon}$ . In particular,  $T$  iterations of full-batch DP-SGD (i.e.  $q = 1$ ) with noise multiplier  $\sigma$  give  $(\alpha, \frac{\alpha T}{2\sigma^2})$ -RDP for every  $\alpha > 1$ .

Ultimately, we are interested in understanding how the privacy guarantee of DP-SGD affects *reconstruction attacks* and, in particular, whether  $\epsilon \gg 1$  still provides some protection against these more ambitious attacks. The first step is to understand what the most idoneous threat model to investigating this question is, and then to instantiate a powerful attack in that model.

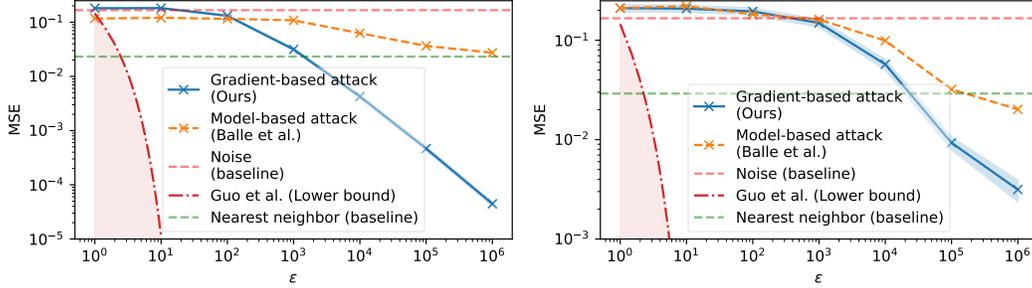
## 2.1 Comparing reconstruction attacks under intermediate and final model access

In the case of DP-SGD, the privacy guarantee for the final model is obtained by analyzing the privacy loss incurred by *releasing the  $T$  private gradients* used by the algorithm. Thus, the guarantee applies both to the intermediate gradients and the final model (by virtue of the post-processing property of DP). It has been shown through membership inference attacks that the guarantee obtained through this analysis for the *collection of intermediate gradients* can be numerically tight (Nasr et al., 2021). However, in some specific settings amenable to mathematical treatment, it has also been shown that the final model produced by DP-SGD can enjoy a stronger DP guarantee than the collection of all intermediate gradients (Ye & Shokri, 2022; Altschuler & Talwar, 2022).

Although these observations apply the standard membership formulation of DP, they motivate an important question for trying to understand the implications of DP guarantees for reconstruction attacks: *how does access to intermediate gradients affect the success of reconstruction attacks against DP-SGD?* We investigate this question by introducing and comparing the success of *model-based* and *gradient-based* attacks. We assume the adversary receives the output of a DP mechanism  $M(D)$  and all the input data in  $D$  except for one point  $z^*$ ; the adversary’s goal is then to produce a reconstruction  $\hat{z}$  of the unknown point. This adversary is referred to as the *informed adversary* in Balle et al. (2022).

**Model-based training data reconstruction.** Suppose we consider the output of DP-SGD to be only the final model  $\theta = M(D)$ . Under the informed adversary threat model, Balle et al. (2022) propose a reconstruction attack in which the adversary uses their knowledge of  $M$  and  $D_- = D \setminus \{z^*\}$  to train a *reconstructor neural network* (RecoNN) capable of mapping model parameters  $\theta$  to reconstructed points  $\hat{z}$ . The training dataset for this network consists of model-point pairs  $(\theta^i, z_i)$  where the  $z_i$  are auxiliary points representing side knowledge the adversary might possess about the distribution of the target point  $z^*$ , and the so-called shadow model  $\theta^i$  is obtained by applying  $M$  to the dataset  $D_- \cup \{z_i\}$  obtained by replacing  $z^*$  by  $z_i$  in the original dataset. Despite its computational cost, this attack is effective in reconstructing complete training examples from image classification models

<sup>3</sup>The clipping operation is defined as  $\text{clip}_C(v) = v / \max\{1, \|v\|/C\}$ .



(a) Average MSE between targets and reconstructions on MNIST. (b) Average MSE between targets and reconstructions on CIFAR-10.

Figure 1: Comparison of model-based and gradient-base reconstruction attacks. We run the attacks over 1K different images for both MNIST and CIFAR-10 for a range of  $\epsilon$ . In Figure 1a and Figure 1b, we plot the average MSE between reconstruction and target images for both attacks. To help calibrate a reader’s interpretation of good and bad reconstructions we plot a nearest neighbor (NN) baseline, marking the average  $L_2$ -distance between NN images over the entire dataset, and a baseline corresponding to the average distance between random uniform vectors. We also plot a lower bound for MSE given by Guo et al. (2022a). We give qualitative examples of reconstructions in Appendix B.

trained *without* DP on MNIST and CIFAR-10, but fails to produce correct reconstructions when the model is trained with any reasonable setting of DP parameters.

**Gradient-based training data reconstruction.** Suppose now that the adversary gets access to all the intermediate gradients  $(g_1, \dots, g_T) = M(D)$  produced by DP-SGD when training model  $\theta$ . We can instantiate a reconstruction attack for this scenario by leveraging gradient inversion attacks found in the federated learning literature (Yin et al., 2021; Huang et al., 2021; Jeon et al., 2021; Jin et al., 2021; Zhu et al., 2019; Geiping et al., 2020). In particular, we use the gradients produced by DP-SGD to construct a loss function  $\mathcal{L}(z)$  for which the target point  $z^*$  would be optimal if the gradients contained no noise, and then optimize it to obtain our candidate reconstruction. An important difference between our attack and previous works in the federated learning setting is that we use gradients from multiple steps to perform the attack and only attempt to recover a single point.

More formally, we start by removing from each privatized gradient  $g_t$  the gradients of the known points in  $D$ . For simplicity, in this section we only consider the full batch case ( $q = 1$ ) where every training point in  $D$  is included in every gradient step. Thus, we set  $\bar{g}_t = g_t - \sum_{z \in D} \text{clip}_C(\nabla_{\theta} \ell(\theta_t, z))$ , where  $\theta_t$  are the model parameters at step  $t$ . Note these can be inferred from the gradients and knowledge of the model initialization, which we also assume is given to the adversary. Similar (but not identical) to Geiping et al. (2020), we use the loss  $\mathcal{L}(z) = \sum_{t=1}^T \mathcal{L}_t(z)$  with

$$\mathcal{L}_t(z) = -\langle \text{clip}_C(\nabla_{\theta} \ell(\theta_t, z)), \bar{g}_t \rangle + \|\text{clip}_C(\nabla_{\theta} \ell(\theta_t, z)) - \bar{g}_t\|_1. \quad (1)$$

This loss was selected after exhaustively trying all other losses suggested in the aforementioned literature; empirically, Equation (1) performed best.

**Comparing reconstruction attacks.** We now compare the success of model-based and gradient-based reconstruction attacks against classification models trained with DP-SGD on MNIST and CIFAR-10. We refer to Appendix A for experimental details.

Following Balle et al. (2022), we report the mean squared error (MSE) between the reconstruction and target as a function of  $\epsilon$  for both attacks. Results are shown in Figure 1, where we include two baselines to help calibrate how MSE corresponds to good and bad reconstructions. Firstly, we include a threshold representing the average distance between target points and their nearest neighbor in the remaining of the dataset. An MSE below this line indicates a near-perfect reconstruction. Secondly, we include a threshold representing the average distance between random uniform vectors of the same data dimensionality. An MSE above or close to this line indicates a poor reconstruction.

We make the following observations. First, the gradient-based attack outperforms the model-based attack by orders of magnitude at larger  $\epsilon$  values. Second, the best attack starts to perform well on

MNIST and CIFAR-10 between  $10^2 < \epsilon < 10^3$  and  $10^3 < \epsilon < 10^4$  respectively. This indicates that attack success is affected by the complexity of the underlying data, including its dimension and geometry. Finally, the attacks give an upper bound for reconstruction error (MSE), however it is not clear if this bound is tight (i.e., whether a better attack could reduce MSE). Guo et al. (2022a) report a *lower bound* for MSE of the form  $1/4(e^\epsilon - 1)$ , which is very far from the upper bound.

The experiment above illustrates how a change in threat model can make a significant difference in the success of reconstruction attacks. In particular, it shows that the attack from Balle et al. (2022) is far from optimal, potentially because of its lack of information about intermediate gradients. Thus, while optimal attacks for membership inference are known (Nasr et al., 2021) and can be used to empirically evaluate the strength of DP guarantees (e.g. for auditing purposes (Tramèr et al., 2022)), the situation is far less clear in the case of reconstruction attacks. Gradient-based attacks improve over model-based attacks, but *are they (near) optimal?* Optimality of attacks is important for applications where one would like to calibrate the privacy parameters of DP-SGD to provide a demonstrable, pre-specified amount of protection against any reconstruction attacks.

### 3 Bounding Training Data Reconstruction

We will now formalize the reconstruction problem further and then provide bounds on the success probability of reconstruction attacks against DP-SGD. We will also develop improvements to the reconstruction attack introduced in Section 2 from which to benchmark the tightness of our bound.

#### 3.1 Threat models for reconstruction

Balle et al. (2022) introduce a formal definition for reconstruction that attempts to capture the success probability of any reconstruction attack against a given mechanism. The definition involves an informed adversary with knowledge of the mechanism  $M$ , the fixed part of the training dataset  $D_-$ , and a prior distribution  $\pi$  from which the target  $z^*$  is sampled – the prior encodes the adversary’s side knowledge about the target.

**Definition 1.** Let  $\pi$  be a prior over target points and  $\rho$  a reconstruction error function. A randomized mechanism  $M$  is  $(\eta, \gamma)$ -ReRo (reconstruction robust) with respect to  $\pi$  and  $\rho$  if for any fixed dataset  $D_-$  and any reconstruction attack  $R$  we have  $\mathbb{P}_{Z \sim \pi, w \sim M(D_- \cup \{Z\})}[\rho(Z, R(w)) \leq \eta] \leq \gamma$ .

Note that the output of the mechanism does not necessarily need to be final model parameters – indeed, the definition also applies to attacks operating on intermediate gradients when those are included in the output of the mechanism. Balle et al. (2022) also proved that any  $(\alpha, \epsilon)$ -RDP mechanism is  $(\eta, \gamma)$ -ReRo with  $\gamma = (\kappa_{\pi, \rho}(\eta) \cdot e^\epsilon)^{\frac{\alpha-1}{\alpha}}$ , where  $\kappa_{\pi, \rho}(\eta) = \sup_{z_0 \in \mathcal{Z}} \mathbb{P}_{Z \sim \pi}[\rho(Z, z_0) \leq \eta]$ . In particular, using the RDP guarantees of DP-SGD, one obtains that in the full-batch case, running the algorithm for  $T$  iterations with noise multiplier  $\sigma$  is  $(\eta, \gamma)$ -ReRo with  $\gamma$  bounded by

$$\exp \left( - \max \left\{ 0, \sqrt{\log \frac{1}{\kappa_{\pi, \rho}(\eta)}} - \sqrt{\frac{T}{2\sigma^2}} \right\}^2 \right). \quad (2)$$

The quantity  $\kappa_{\pi, \rho}(\eta)$  can be thought of as the prior probability of reconstruction by an adversary who outputs a fixed candidate  $z_0$  based only on their knowledge of the prior (i.e. without observing the output of the mechanism).

**Instantiating threat models within the  $(\eta, \gamma)$ -ReRo framework** In Section 2.1, we described two variants of a reconstruction attack, one where the adversary has access to intermediate model updates (gradient-based reconstruction) and one where the adversary has access to the final model only (model-based reconstruction). In the language of  $(\eta, \gamma)$ -ReRo, both the gradient-based and model-based reconstruction attacks introduced in Section 2 take as arguments:  $\pi$  – prior information about the target point,  $D_-$  – the training dataset excluding the target point,  $w$  – the output of the mechanism  $M$ , and side information about DP-SGD – such as hyperparameters used in training and how to the model was initialized. The model-based reconstruction attack assumes that  $w = \theta_T$ , the parameters of the final model, whereas for gradient-based attacks,  $w = (g_1, \dots, g_T)$ , and so the adversary has access to all intermediate privatized model parameter updates.

The gradient-based attack optimizing Equation (1) does not make use of any potential prior knowledge the adversary might have about the target point  $z^*$ , beyond input bounds (the input is bound between  $[0, 1]$ ) and the dimensionality of target ( $28 \times 28$  for an MNIST digit). On the other hand, the model-based attack makes use of a prior in the form of the auxiliary points  $z_i$  used in the construction of shadow models; these points represent the adversary’s knowledge about the distribution from which the target points  $z^*$  is sampled.

Going forward in our investigation we will assume the adversary has access to more “reliable” side knowledge: their prior is a uniform distribution over a finite set of candidate points  $\{z_1, \dots, z_n\}$ , one of which corresponds to the true target point. This setting is favorable towards the adversary: the points in the prior represent a shortlist of candidates the adversary managed to narrow down using side knowledge about the target. In this case it is also reasonable to use the error function  $\rho(z, z') = \mathbb{I}[z \neq z']$  since the adversary’s goal is to identify which point from the prior is the target. As DP assumes the adversary knows (and can influence) all but one record in the training set, the assumption that the adversary has prior knowledge about the target is aligned with the DP threat model. The main distinction between membership inference and reconstruction with a uniform prior is that in the former the adversary (implicitly) managed to narrow down the target point to two choices, while in the latter they managed to narrow down the target to  $n$  choices. This enables us to smoothly interpolate between the exceedingly strong DP threat model of membership inference (where the goal is to infer a single bit) and a relaxed setting where the adversary’s side knowledge is less refined: here  $n$  controls the amount of prior knowledge the adversary is privy to, and requires the adversary to infer  $\log(n)$  bits to achieve a successful reconstruction.

The discrete prior setup, we argue, provides a better alignment between the success of reconstruction attacks and what actually constitutes privacy leakage. In particular, it allows us to move away from (approximate) verbatim reconstruction as modelled, e.g., by an  $\ell_2$  reconstruction criteria success, and model more interesting situations. For example, if the target image contains a car, an attacker might be interested in the digits of the license plate, not the pixels of the image of the car and its background. Thus, if the license plate contains 4 digits, the attacker’s goal is to determine which of the possible 10,000 combinations was present in the in the training image.

In Appendix I, we also conduct experiments where the adversary has less background knowledge about the target point, and so the prior probability of reconstruction is *extremely* small (e.g.  $\mathcal{O}\left(\frac{1}{256}\right)^{32 \times 32 \times 3}$ ) for a CIFAR-10 image).

### 3.2 Reconstruction Robustness of DP-SGD

**ReRo bounds from blow-up functions** We now state a novel reconstruction robustness bound that, instead of using the (R)DP guarantees of the mechanism, is directly expressed in terms of its output distributions. The bound depends on the *blow-up function* between two distributions  $\mu$  and  $\nu$ :

$$\mathcal{B}_\kappa(\mu, \nu) = \sup\{\mathbb{P}_\mu[E] : E \text{ s.t. } \mathbb{P}_\nu[E] \leq \kappa\} . \quad (3)$$

In particular, let  $\nu_{D_\cdot} = M(D_\cdot)$  denote the output distribution of the mechanism on a fixed dataset  $D_\cdot$ , and  $\mu_{D_\cdot z} = M(D_\cdot \cup \{z\})$  for any potential target  $z$ . Then we have the following (see Appendix M.1 for the full statement).

**Theorem 2 (Informal).** *Fix  $\pi$  and  $\rho$ . Suppose that for every fixed dataset  $D_\cdot$  there exists a pair of distributions  $\mu_{D_\cdot}^*, \nu_{D_\cdot}^*$  such that  $\sup_{z \in \text{supp}(\pi)} \mathcal{B}_\kappa(\mu_{D_\cdot z}, \nu_{D_\cdot}) \leq \mathcal{B}_\kappa(\mu_{D_\cdot}^*, \nu_{D_\cdot}^*)$  for all  $\kappa \in [0, 1]$ . Then  $M$  is  $(\eta, \gamma)$ -ReRo with  $\gamma = \sup_{D_\cdot} \mathcal{B}_{\kappa_{\pi, \rho}(\eta)}(\mu_{D_\cdot}^*, \nu_{D_\cdot}^*)$ .*

This result basically says that the probability of successful reconstruction for an adversary that does observe the output of the mechanism can be bounded by the maximum probability under  $\mu_{D_\cdot}^*$  over all events that have probability  $\kappa_{\pi, \rho}(\eta)$  under  $\nu_{D_\cdot}^*$ , when we take the worst-case setting over all fixed datasets  $D_\cdot$  and all target points  $z$  in the support of the prior. If  $M$  satisfies  $(\epsilon, 0)$ -DP (i.e.  $(\infty, \epsilon)$ -RDP), then  $\mathbb{P}_{\nu_{D_\cdot}}[E] \leq e^\epsilon \mathbb{P}_{\mu_{D_\cdot z}}[E]$  for any event  $E$ , in which case Theorem 2 gives with  $\gamma = \kappa_{\pi, \rho}(\eta)e^\epsilon$ . This recovers the case  $\alpha = \infty$  of the bound from Balle et al. (2022) stated above.

**Remark 3.** *The blow-up function is tightly related to the notion of trade-off function defined in Dong et al. (2019). Precisely, the trade-off function between two probability distributions  $\mu$  and  $\nu$  is defined as  $\mathcal{T}_\kappa(\mu, \nu) = \inf\{\mathbb{P}_\mu[E] : E \text{ s.t. } \mathbb{P}_\nu[E] \leq \kappa\}$ . The trade-off function is usually introduced in terms of Type I and Type II errors; it defines the smallest Type II error achievable given constraint on the maximum Type I error, and we have the following relationship:  $\mathcal{T}_\kappa(\mu, \nu) = 1 - \mathcal{B}_\kappa(\mu, \nu)$ .*

**ReRo for DP-SGD** Next we apply this bound directly to DP-SGD without an intermediate appeal to its RDP guarantees; in Section 4.1 we will empirically demonstrate that in practice this new bound is vastly superior to the bound in Balle et al. (2022).

Let  $M$  be the mechanism that outputs all the intermediate gradients produced by running DP-SGD for  $T$  steps with noise multiplier  $\sigma$  and sampling rate  $q$ . Our analysis relies on the observation from Mahlouljifar et al. (2022) that comparing the output distributions of  $M$  on two neighbouring datasets can be reduced to comparing the  $T$ -dimensional Gaussian distribution  $\nu_{T,\sigma} = \mathcal{N}(0, \sigma^2 I)$  with the Gaussian mixture  $\mu_{T,\sigma,q} = \sum_{w \in \{0,1\}^T} \mathbb{P}[B(q, T) = w] \mathcal{N}(w, \sigma^2 I)$ , where  $B(q, T)$  is a random binary vector obtained by sampling each coordinate independently from a Bernoulli distribution with success probability  $q$ .

**Corollary 4.** *For any prior  $\pi$  and error function  $\rho$ ,  $M$  is  $(\eta, \gamma)$ -ReRo with  $\gamma = \mathcal{B}_{\kappa_{\pi,\rho}(\eta)}(\mu_{T,\sigma,q}, \nu_{T,\sigma})$ .*

We defer proofs and discussion to Appendix M. We note that the concurrent work of Guo et al. (2022b) also provides an upper bound on the success of training data reconstruction under DP. We compare this to our own upper bound in Appendix H.

### 3.3 Upper bound estimation via reconstruction robustness

To estimate the bound in Corollary 4 in practice we first estimate the baseline probability  $\kappa = \kappa_{\pi,\rho}(\eta)$  and then the blow-up  $\gamma = \mathcal{B}_{\kappa}(\mu_{T,\sigma,q}, \nu_{T,\sigma})$ .

**Estimating  $\kappa$ .** All our experiments deal with a uniform discrete prior  $\pi = \frac{1}{n} \sum_i \mathbf{1}_{z_i}$  over a finite set of points  $\{z_1, \dots, z_n\}$ . In this case, to compute  $\kappa$  it suffices to find the  $\rho$ -ball of radius  $\eta$  containing the largest number of points from the prior. If one is further interested in exact reconstruction (i.e.  $\rho(z, z') = \mathbb{I}[z \neq z']$  and  $\eta < 1$ ) then we immediately have  $\kappa = 1/n$ . Although this already covers the setting from our experiments, we show in Appendix L that more generally it is possible to estimate  $\kappa$  only using samples from the prior. We use this to calculate  $\kappa$  for some non-uniform settings where we relax the condition to be on a close reconstruction instead of exact reconstruction.

**Estimating  $\gamma$ .** Note that it is easy to both sample and evaluate the density of  $\mu = \mu_{T,\sigma,q}$  and  $\nu = \nu_{T,\sigma}$ . Any time these two operations are feasible we can estimate  $\mathcal{B}_{\kappa}(\mu, \nu)$  using a non-parametric approach to find the event  $E$  that maximizes Equation (3). The key idea is to observe that, as long as  $\mu/\nu < \infty$ ,<sup>4</sup> then a change of measure gives  $\mathbb{P}_{\mu}[E] = \mathbb{E}_{W \sim \nu} \left[ \frac{\mu(W)}{\nu(W)} \mathbb{I}[W \in E] \right]$ . For fixed  $E$  this expectation can be approximated using samples from  $\nu$ . Furthermore, since we are interested in the event of probability less than  $\kappa$  under  $\nu$  that maximizes the expectation, we can take  $N$  samples from  $\nu$  and keep the  $\kappa N$  samples that maximize the ratio  $\mu/\nu$  to obtain a suitable candidate for the maximum event. This motivates the Monte-Carlo approximation of  $\gamma$  in Algorithm 1.

**Proposition 5.** *For  $\mu = \mu_{T,\sigma,q}$  and  $\nu = \nu_{T,\sigma}$  we have  $\lim_{N \rightarrow \infty} \hat{\gamma} = \mathcal{B}_{\kappa}(\mu, \nu)$  almost surely.*

---

#### Algorithm 1 Estimating $\gamma$

---

**Input:** Prior probability  $\kappa$ , number of samples  $N$ , sampling and density evaluation access to  $\nu$  and  $\mu$ .

**Output:** Estimate  $\hat{\gamma}$  of reconstruction robustness probability.

Sample  $w_1, \dots, w_N$  independently from  $\nu$

Calculate the ratios  $r_i = \mu(w_i)/\nu(w_i)$

Sort the ratios in decreasing order:  $r'_1 \geq$

$\dots \geq r'_N$

Let  $N' = \lceil \kappa N \rceil$  and return  $\hat{\gamma} =$

$\frac{1}{N'} \sum_{i=1}^{N'} r'_i$

---



---

#### Algorithm 2 Prior-aware attack

---

**Input:** Discrete prior  $\pi = \{z_1, \dots, z_n\}$ , Model parameters  $\{\theta_1, \theta_1, \dots, \theta_T\}$ , Privatized gradients (with known gradients subtracted)  $\{\bar{g}_1, \dots, \bar{g}_T\}$

Observations:  $\mathcal{O} \leftarrow \{\}$

**Output:** Reconstruction guess  $\hat{z} \in \pi$

**for**  $i \in [1, 2, \dots, n]$  **do**

$\mathcal{O}[i] \leftarrow \sum_{t=1}^T \langle \text{clip}_C(\nabla_{\theta} \ell(\theta_t, z_i)), \bar{g}_t \rangle$

**end for**

$\hat{i} \leftarrow \arg \max \mathcal{O}$

**return**  $\hat{z} \leftarrow \pi[\hat{i}]$

---

<sup>4</sup>This is satisfied by  $\mu = \mu_{T,\sigma,q}$  and  $\nu = \nu_{T,\sigma}$ .

**Estimation guarantees** Although Proposition 4 is only asymptotic, we can provide confidence intervals around this estimate using two types of inequalities. Algorithm 1 has two points of error, the gap between the empirical quantile and the population quantile, and the error from mean estimation  $\hat{\gamma}$ . The first source of error can be bounded using the DKW inequality (Dvoretzky et al., 1956) which provides a uniform concentration for each quantile. In particular, we can show that the error of quantile estimation is at most  $2e^{-2n\eta^2}$ , where  $n$  is the number of samples and  $\eta$  is the error in calculation of the quantile. Specifically, with a million samples, we can make sure that with probability 0.999 the error of quantile estimation is less than 0.002, and we can make this smaller by increasing the number of samples. We can account for the second source of error with Bennett’s inequality, that leverages the bounded variance of the estimate. In all our bounds, we can show that the error of this part is also less than 0.01 with probability 0.999. We provide an analysis on the computational cost of estimating  $\gamma$  in Appendix A.

### 3.4 Lower bound estimation via a prior-aware attack

We now present a prior-aware attack whose success can be directly compared to the reconstruction upper bound from Corollary 4.

In addition to the uniform prior  $\pi = \frac{1}{n} \sum_i \mathbf{1}_{z_i}$  from which we assume the target is sampled from, our prior-aware attack has access to the same information as the gradient-based attack from Section 2: all the privatized gradients (and therefore, intermediate models) produced by DP-SGD, and knowledge of the fixed dataset  $D$  that can be used to remove all the known datapoints from the privatized gradients. The attack is given in Algorithm 2. For simplicity, in this section we present and experiment with the version of the attack corresponding to the full-batch setting in DP-SGD (i.e.  $q = 1$ ). We present an extension of the algorithm to the mini-batch setting in Section 4.3.

The rationale for the attack is as follows. Suppose, for simplicity, that all the gradients are clipped so they have norm exactly  $C$ . If  $z^*$  is the target sampled from the prior, then  $\bar{g}_t \sim \text{clip}_C(\nabla_{\theta} \ell(\theta_t, z^*)) + \mathcal{N}(0, C^2 \sigma^2 I)$ . Then the inner products  $\langle \text{clip}_C(\nabla_{\theta} \ell(\theta_t, z_i)), \bar{g}_t \rangle$  follow a distribution  $\mathcal{N}(C^2, C^4 \sigma^2)$  if  $z_i = z^*$  and  $\mathcal{N}(A, C^4 \sigma^2)$  for some  $A < C^2$  otherwise (assuming no two gradients are in the exact same direction). In particular,  $A \approx 0$  if the gradients of the different  $z_i$  in the prior are mutually orthogonal. Thus, finding the  $z_i$  that maximizes this sum of inner products is likely to produce a good guess for the target point. This attack gives a lower bound on the probability of successful reconstruction, which we can compare with the upper bound estimate derived in Section 3.3.

## 4 Experimental Evaluation

We now evaluate both our upper bounds for reconstruction success and our empirical privacy attacks (which gives us lower bounds on reconstruction success). We show that our attack has a success probability nearly identical to the bound given by our theory. We conclude this section by inspecting how different variations on our threat models change both the upper bound and the optimality of our new attack.

### 4.1 Attack and Bound Evaluation on CIFAR-10

We now evaluate our upper bound (Corollary 4) and prior-aware reconstruction attack (Algorithm 2) on full-batch DP-SGD, and compare it to the gradient-based (prior-oblivious) attack optimizing Equation (1) and the RDP upper bound obtained from previous work (Equation (2)). To perform our evaluation on CIFAR-10 models with relatively high accuracy despite being trained with DP-SGD, we follow De et al. (2022) and use DP-SGD to fine-tune the last layer of a WideResNet model (Zagoruyko & Komodakis, 2016) pre-trained on ImageNet. In this section and throughout the paper, we use a prior with uniform support over 10 randomly selected points unless we specify the contrary. Further experimental details are deferred to Appendix A.

The results of our evaluation are presented in Figure 2, which reports the upper and lower bounds on the probability of successful reconstruction produced by the four methods at three different values of  $(\epsilon, 10^{-5})$ -DP. The first important observation is that our new ReRo bound for DP-SGD is a significant improvement over the RDP-based bound obtained in Balle et al. (2022), which gives a trivial bound for  $\epsilon \in \{10, 100\}$  in this setting. On the other hand, we also observe that the prior-aware attack substantially improves over the prior-oblivious

gradient-based attack<sup>5</sup>, which Section 2 already demonstrated is stronger than previous model-based attacks – a more thorough investigation of how the gradient-based attack compares to the prior-aware attack is presented in Appendix D for varying sizes of the prior distribution. We will use the prior-aware attack exclusively in following experiments due to its superiority over the gradient-based attack. Finally, we observe that our best empirical attack and theoretical bound are very close to each other in all settings. We conclude there are settings where the upper bound is nearly tight and the prior-based attack is nearly optimal.

Now that we have established that our attacks work well on highly accurate large models, our remaining experiments will be performed on MNIST which are significantly more efficient to run.

## 4.2 Effects of the prior size

We now investigate how the bound and prior-aware attack are affected by the size of the prior in Figure 3. As expected, both the upper and lower bound to the probability that we can correctly infer the target point decreases as the prior size increases. We also observe a widening gap between the upper bound and attack for larger values of  $\epsilon$  and prior size. However, this bound is still relatively tight for  $\epsilon < 10$  and a prior size up to  $2^{10}$ . Further experiments regarding how the prior affects reconstruction are given in Appendix E.

## 4.3 Effect of DP-SGD Hyperparameters

Given that we have developed an attack that is close to optimal in a full-batch setting, we are now ready to inspect the effect DP-SGD hyper-parameters controlling its privacy guarantees have on reconstruction success. First, we observe how the size of mini-batch affects reconstruction success, before measuring how reconstruction success changes with differing hyper-parameters at a fixed value of  $\epsilon$ . Further experiments that measure the effect of the clipping norm,  $C$ , are deferred to Appendix K.

**Effect of sampling rate  $q$ .** All of our experiments so far have been in the full-batch ( $q = 1$ ) setting. We now measure how mini-batching affects both the upper and lower bounds by reducing the data sampling probability  $q$  from 1 to 0.02 (see Appendix A for additional experimental details).

The attack presented in Algorithm 2 assumes that the gradient of  $z^*$  is present in every privatized gradient. Since this is no longer true in mini-batch DP-SGD, we design an improved attack by factoring in the knowledge that only a fraction  $q$  of privatized gradients will contain the gradient of the target  $z^*$ . This is achieved by, for fixed  $z_i$ , collecting the inner-products  $\langle \text{clip}_C(\nabla_{\theta} \ell(\theta_t, z_i)), \bar{g}_t \rangle$  across all  $T$  iterations,

<sup>5</sup>To measure the success probability of the gradient-based attack we take the output of the optimization process, find the closest point in the support of the prior, and declare success if that was the actual target. This process is averaged over 10,000 randomly constructed priors and samples from each prior.

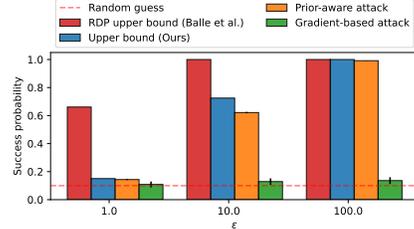


Figure 2: Comparison of success of reconstruction attacks against reconstruction upper bounds on CIFAR-10 with test accuracy at  $\epsilon = 1, 10$ , and  $100$  equal to 61.88%, 89.09%, and 91.27%, respectively. Note, the best model we could train at  $\epsilon = \infty$  with our experimental settings reaches 93.91%.

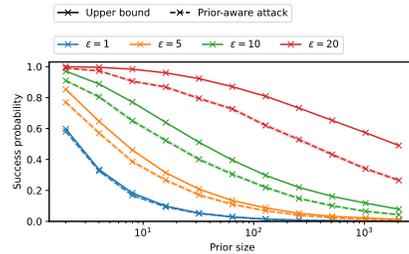


Figure 3: How the size of the prior  $\pi$  affects reconstruction success probability for a range of different values of  $\epsilon$ .

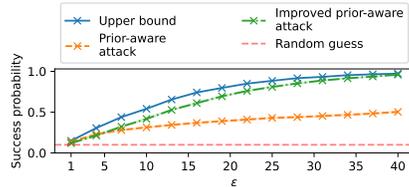


Figure 4: We compare how the attack from Algorithm 2 compares with the improved prior-aware attack when  $q = 0.02$ .

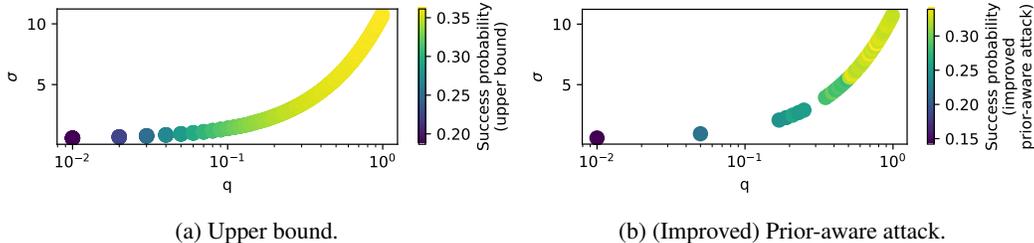


Figure 5: We plot  $\sigma$  against  $q$  at  $(4, 10^{-5})$ -DP at a fixed  $T = 100$  and the corresponding upper bound (Figure 5a) and (improved) prior-aware attack (Figure 5b). Reconstruction is less successful at smaller  $q$  in comparison to large  $q$  for the same value of  $\epsilon$ .

and computing the score using only the top  $qT$  values. Pseudo-code for this improved prior-aware attack variant is given in Appendix F.

Results are shown in Figure 4, where we observe the improved attack is fairly tight to the upper bound, and also a large difference between the previous and improved attack.

**Effect of DP-SGD hyperparameters at a fixed  $\epsilon$ .** In our final experiment, we fix  $\epsilon$  and investigate if and how the reconstruction upper bound and lower bound (through the improved prior-aware attack) change with different DP-SGD hyperparameters. In Figure 5, we fix  $\epsilon = 4$  and run DP-SGD for  $T = 100$  iterations with  $C = 1$  while varying  $q \in [0.01, 0.99]$ . In each setting we tune  $\sigma$  to satisfy  $(4, 10^{-5})$ -DP. For experimental efficiency, we use fewer values of  $q$  for the attack than the theoretical bound. Surprisingly, we find that both the upper bound and lower bound *change* with different hyperparameter configurations. For example, at  $q = 0.01$  the upper bound on successful reconstruction is  $\approx 0.20$  (and the attack is  $\approx 0.15$ ), while at  $q = 0.99$  it is  $\approx 0.35$  (and the attack is  $\approx 0.32$ ). Generally, an increase in  $T$  or  $q$  increases the upper bound. This suggests that an  $(\epsilon, \delta)$ -DP guarantee alone is not able to fully specify the probability that a reconstruction attack is successful. We conduct a more extensive evaluation for different  $\epsilon$  and prior sizes in Appendix K.2.

## 5 Conclusion

In this work, we investigated training data reconstruction bounds on DP-SGD. We developed upper bounds on the success that an adversary can reconstruct a training point  $z^*$  under DP-SGD. In contrast to prior work that develop reconstruction upper bounds based on DP or RDP analysis of DP-SGD, our upper bounds are directly based on the parameters of the algorithm. We also developed new reconstruction attacks, specifically against DP-SGD, that obtain lower bounds on the probability of successful reconstruction, and observe they are close to our upper bounds. Our experiments show that both our upper and lower bounds are superior to previously proposed bounds for reconstruction in DP-SGD. Our investigations further showed that the  $\epsilon$  parameter in DP (or RDP) cannot solely explain robustness to reconstruction attacks. In other words, one can have the same value of  $\epsilon$  for two different algorithms while achieving stronger robustness against reconstruction attacks in one over the other. This suggest that a more fine-grained analysis of algorithms against specific privacy attacks can lead to superior privacy guarantees and also opens up the possibility of better hyperparameter selection for specific privacy concerns.

## 6 Acknowledgements

The authors would like to thank Leonard Berrada and Taylan Cemgil for their thoughtful feedback on an earlier version of this work.

## References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

- Altschuler, J. M. and Talwar, K. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *arXiv preprint arXiv:2205.13710*, 2022.
- Balle, B., Cherubin, G., and Hayes, J. Reconstructing training data with informed adversaries. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, 2022.
- Bhowmick, A., Duchi, J., Freudiger, J., Kapoor, G., and Rogers, R. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.
- Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284, 2019.
- Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.
- Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022.
- Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Schwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.
- Cattan, Y., Choquette-Choo, C. A., Papernot, N., and Thakurta, A. Fine-tuning with differential privacy necessitates an additional hyperparameter search. *arXiv preprint arXiv:2210.02156*, 2022.
- De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.
- Dong, J., Roth, A., and Su, W. J. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.
- Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.
- Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, 2006.
- Feldman, V. and Zrnic, T. Individual privacy accounting via a renyi filter. *Advances in Neural Information Processing Systems*, 34:28080–28091, 2021.
- Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.
- Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- Google DP Team. Privacy loss distributions. [https://github.com/google/differential-privacy/blob/main/common\\_docs/Privacy\\_Loss\\_Distributions.pdf](https://github.com/google/differential-privacy/blob/main/common_docs/Privacy_Loss_Distributions.pdf), 2022.
- Gopi, S., Lee, Y. T., and Wutschitz, L. Numerical composition of differential privacy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- Guo, C., Karrer, B., Chaudhuri, K., and van der Maaten, L. Bounding training data reconstruction in private (deep) learning. *arXiv preprint arXiv:2201.12383*, 2022a.
- Guo, C., Sablayrolles, A., and Sanjabi, M. Analyzing privacy leakage in machine learning via multiple hypothesis testing: A lesson from fano. *arXiv preprint arXiv:2210.13662*, 2022b.

- Haim, N., Vardi, G., Yehudai, G., Shamir, O., and Irani, M. Reconstructing training data from trained neural networks. *arXiv preprint arXiv:2206.0775*, 2022.
- Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8): e1000167, 2008.
- Huang, Y., Gupta, S., Song, Z., Li, K., and Arora, S. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34:7232–7241, 2021.
- Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette-Choo, C. A., and Carlini, N. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*, 2022.
- Jeon, J., Lee, K., Oh, S., Ok, J., et al. Gradient inversion with generative image prior. *Advances in Neural Information Processing Systems*, 34:29898–29908, 2021.
- Jin, X., Chen, P.-Y., Hsu, C.-Y., Yu, C.-M., and Chen, T. Cafe: Catastrophic data leakage in vertical federated learning. *Advances in Neural Information Processing Systems*, 34:994–1006, 2021.
- Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. *arXiv preprint arXiv:2202.06539*, 2022.
- Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer, 2005.
- Ligett, K., Neel, S., Roth, A., Waggoner, B., and Wu, S. Z. Accuracy first: Selecting a differential privacy level for accuracy constrained erm. *Advances in Neural Information Processing Systems*, 30, 2017.
- Mahloujifar, S., Sablayrolles, A., Cormode, G., and Jha, S. Optimal membership inference bounds for adaptive composition of sampled gaussian mechanisms. *arXiv preprint arXiv:2204.06106*, 2022.
- Mehta, H., Thakurta, A., Kurakin, A., and Cutkosky, A. Large scale transfer learning for differentially private image classification. *arXiv preprint arXiv:2205.02973*, 2022.
- Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., and Shokri, R. Quantifying privacy risks of masked language models using membership inference attacks. *arXiv preprint arXiv:2203.03929*, 2022.
- Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.
- Nasr, M., Song, S., Thakurta, A., Papemoti, N., and Carlini, N. Adversary instantiation: Lower bounds for differentially private machine learning. In *Symposium on Security and Privacy (S&P)*, 2021.
- Redberg, R. and Wang, Y.-X. Privately publishable per-instance privacy. *Advances in Neural Information Processing Systems*, 34:17335–17346, 2021.
- Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., and Jégou, H. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.
- Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.
- Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022.
- Song, C. and Shmatikov, V. Overlearning reveals sensitive attributes. *arXiv preprint arXiv:1905.11742*, 2019.

- Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.
- Stock, P., Shilov, I., Mironov, I., and Sablayrolles, A. Defending against reconstruction attacks with rényi differential privacy. *arXiv preprint arXiv:2202.07623*, 2022.
- Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. *arXiv preprint arXiv:2205.10770*, 2022.
- Tramèr, F. and Boneh, D. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations (ICLR)*, 2021.
- Tramèr, F., Terzis, A., Steinke, T., Song, S., Jagielski, M., and Carlini, N. Debugging differential privacy: A case study for privacy auditing. *arXiv preprint arXiv:2202.12219*, 2022.
- Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., and Qi, H. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 2512–2520. IEEE, 2019.
- Ye, J. and Shokri, R. Differentially private learning needs hidden state (or much faster convergence). *arXiv preprint arXiv:2203.05363*, 2022.
- Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268–282. IEEE, 2018.
- Yin, H., Mallya, A., Vahdat, A., Alvarez, J. M., Kautz, J., and Molchanov, P. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16337–16346, 2021.
- Yu, D., Kamath, G., Kulkarni, J., Yin, J., Liu, T.-Y., and Zhang, H. Per-instance privacy accounting for differentially private stochastic gradient descent. *arXiv preprint arXiv:2206.02617*, 2022.
- Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.
- Zhao, B., Mopuri, K. R., and Bilen, H. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.
- Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.
- Zhu, Y., Dong, J., and Wang, Y.-X. Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pp. 4782–4817. PMLR, 2022.

Table 1: Hyperparameter settings for each experiment. The set of examples that make up the prior ( $\pi$ ), including the target ( $z^*$ ), and the other examples in the training set ( $D_-$ ) are always drawn from the same data distribution, except for the experiment performed in Figure 9.

Dataset	Experiment	Clipping norm $C$	Sampling probability $q$	Update steps $T$	Model architecture $\theta$	Training dataset size ( $ D_-  + 1$ )	Prior size
CIFAR-10	Figure 1b, Figure 6b, Figure 12b	1	1	100	WRN-28-10	500	-
	Figure 2	1	1	100	WRN-28-10	500	10
	Figure 1a, Figure 6a, Figure 12a	0.1	1	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	1,000	-
MNIST	Figure 3	0.1	1	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	1,000	$2^1 - 2^{11}$
	Figure 4	0.1	0.02	1,000	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	500	10
	Figure 5	1	0.01-0.99	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	1,000	10
	Figure 7	0.1	1	100	MLP (784 $\rightarrow$ 10, 100, 1000 $\rightarrow$ 10)	1,000	10
	Figure 8	0.1	1	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	1,000	$2^1, 2^3, 2^7$
	Figure 9	0.1	1	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	1,000	10
	Figure 10	0.1	1	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	5, 129, 1,000	10
	Figure 11	1.0	1	-	-	-	10
	Figure 13	0.1, 1	1	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	1,000	10
	Figure 14	1	0.01-0.99	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	1,000	2, 10, 100
	Figure 15	1	0.01-0.99	100	MLP (784 $\rightarrow$ 10 $\rightarrow$ 10)	1,000	2, 10, 100

## A Experimental details

We detail the experimental settings used throughout the paper, and specific hyperparameters used for the various attacks we investigate. The exact configurations for each experiment are given in Table 1. We vary many experimental hyperparameters to investigate their effect on reconstruction, however, the default setting is described next.

For MNIST experiments we use a two layer MLP with hidden width 10 and eLU activations. The attacks we design in this work perform equally well on all common activation functions, however it is well known that the model-based attack (Balle et al., 2022) performs poorly on piece-wise linear activations like ReLU. We set  $|D_-| = 999$  (and so the training set size is  $|D_- \cup \{z^*\}| = 1,000$ ) and train with full-batch DP-SGD for  $T = 100$  steps. For each  $\epsilon$ , we select the learning rate by sweeping over a range of values between 0.001 and 100; we do not use any momentum in optimization. We set  $C = 0.1$ ,  $\delta = 10^{-5}$  and adjust the noise scale  $\sigma$  for a given target  $\epsilon$ . The accuracy of this model is over 90% for  $\forall \epsilon \geq 10$ , however we emphasize that our experiments on MNIST are meant to primarily investigate the tightness of our reconstruction upper bounds. We set the size of the prior  $\pi$  to ten, meaning the baseline probability of successful reconstruction is 10%.

For the CIFAR-10 dataset, we use a Wide-ResNet (Zagoruyko & Komodakis, 2016) model with 28 layers and width factor 10 (denoted as WRN-28-10), group normalization, and eLU activations. We align with the set-up of De et al. (2022), who fine-tune a WRN-28-10 model from ImageNet to CIFAR-10. However, because the model-based attack is highly expensive, we only fine-tune the final layer. We set  $|D_-| = 499$  (and so the training set size is  $|D_- \cup \{z^*\}| = 500$ ) and train with full-batch DP-SGD for  $T = 100$  steps; again we sweep over the choice of learning rate for each value of  $\epsilon$ . We set  $C = 1$ ,  $\delta = 10^{-5}$  and adjust the noise scale  $\sigma$  for a given target  $\epsilon$ . The accuracy of this model is over 89% for  $\forall \epsilon \geq 10$ , which is close to the state-of-the-art results given by De et al. (2022), who achieve 94.2% with the same fine-tuning setting at  $\epsilon = 8$  (with a substantially larger training set size). Again, we set the size of the prior  $\pi$  to ten, meaning the baseline probability of successful reconstruction is 10%.

For the gradient-based and model-based attack we generate 1,000 reconstructions and for prior-aware attack experiments we generate 10,000 reconstructions from which we estimate a lower bound for probability of successful reconstruction. That is, for experiments in Section 2 repeat the attack 1,000 times for targets randomly sampled from base dataset (MNIST or CIFAR-10), and for all other experiments we repeat the attack 10,000 times for targets randomly sampled from the prior, which is itself sampled from the base dataset (MNIST or CIFAR-10). We now give experimental details specific to the various attacks used throughout the paper. Note that for attack results, we report 95% confidence intervals around our lower bound estimate, however, in many cases these intervals are so tight it renders them invisible to the eye.

**Model-based attack details.** For the model-based attack given by Balle et al. (2022), we train  $40K$  shadow models, and as stated above, construct a test set by training a further 1,000 models on 1,000 different targets (and  $D_-$ ) from which we evaluate our reconstructions. We use the same architecture for the RecoNN network and optimization hyperparameters as described in the MNIST and CIFAR-10 experiments in Balle et al. (2022), and refer the interested reader there for details.

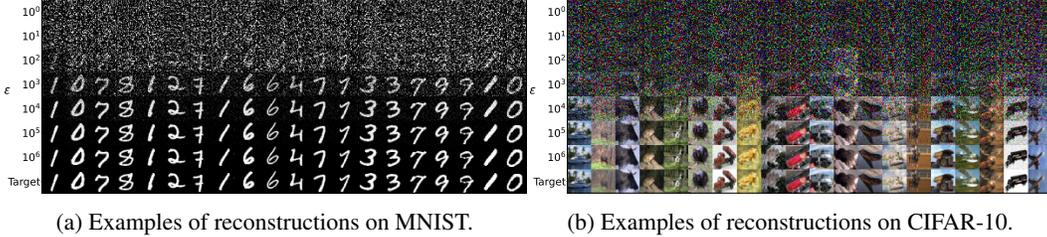


Figure 6: We give qualitative examples of reconstructions in Figure 6a and Figure 6b for the gradient-based reconstruction attack described in Section 2.

**Gradient-based attack details.** Our optimization hyperparameters are the same for both MNIST and CIFAR-10. We initialize a  $\hat{z}$  from uniform noise and optimize it with respect to the loss given in Equation (1) for 1M steps of gradient descent with a learning rate of 0.01. We found that the loss occasionally diverges and it is useful to have random restarts of the optimization process; we set the number of random restarts to five. Note we assume that the label of  $z^*$  is known to the adversary. This is a standard assumption in training data reconstruction attacks on federated learning, as Zhao et al. (2020) demonstrated the label of the target can be inferred given access to gradients. If we did not make this assumption, we can run the attack by exhaustively searching over all possible labels. For the datasets we consider, this would increase the cost of the attack by a factor of ten. We evaluate the attack using the same 1,000 targets used to evaluate the model-based attack.

**Prior-aware attack details.** The prior-aware attacks given in Algorithm 2 (and in Algorithm 3) have no specific hyper-parameters that need to be set. As stated, the attack proceeds by summing the inner-product defined in Section 3.4 over all training steps for each sample in the prior and selecting the sample that maximizes this sum as the reconstruction. One practical note is that we found it useful to normalize privatized gradients such that the privatized gradient containing the target will be sampled from a Gaussian with unit mean instead of  $C^2$ , which will be sensitive to choice of  $C$  and can lead to numerical precision issues.

**Estimating  $\gamma$  details.** As described in Section 3,  $\nu$  is instantiated as  $\mathcal{N}(0, \sigma^2 I)$ , a  $T$ -dimensional isotropic Gaussian distribution with zero mean, and  $\mu$  is given by  $\sum_{w \in \{0,1\}^T} p(w) \mathcal{N}(w, \sigma^2 I)$ , a mixture of  $T$ -dimensional isotropic Gaussian distributions with means in  $\{0, 1\}^T$  sampled according to  $B(q, T)$ . Throughout all experiments, we use 1M independent Gaussian samples to compute the estimation of  $\gamma$  given by the procedure in Algorithm 1, and because we use a discrete prior of size  $|\pi|$ , the base probability of reconstruction success,  $\kappa$ , is given as  $1/|\pi|$ . Estimating  $\gamma$  is cheap; setting  $T = 1$  and using a 2.3 GHz 8-Core Intel Core i9 CPU it takes 0.002s to estimate with 10,000 samples. This increases to 0.065s, 0.196s, and 2.084s with 100,000, 1M, and 10M samples. The estimate sensitivity is also small; for the experimental parameters used in Figure 2 at  $\epsilon = 10$ , over 1,000 different calls to Algorithm 1 the standard deviation of  $\gamma$  estimates is  $< 0.019$  using 10,000 samples, and  $< 0.0017$  using 1M samples.

## B Visualization of reconstruction attacks on MNIST and CIFAR-10

In Figure 6, we give a selection of examples for the gradient-based reconstruction attack presented in Section 2 and plotted in Figure 1.

## C Does the model size make a difference to the prior-aware attack?

Our results on MNIST and CIFAR-10 suggest that the model size does not impact the tightness of our reconstruction attack (lower bound on probability of success); the MLP model used for MNIST has 7,960 trainable parameters, while the WRN-28-10 model used for CIFAR-10 has 36.5M. We systematically evaluate the impact of the model size on our prior-aware attack by increasing the size of the MLP hidden layer by factors of ten, creating models with 7,960, 79,600, and 796,000 parameters. Results are given in Figure 7, where we observe almost no difference in terms of attack success between the different model sizes.

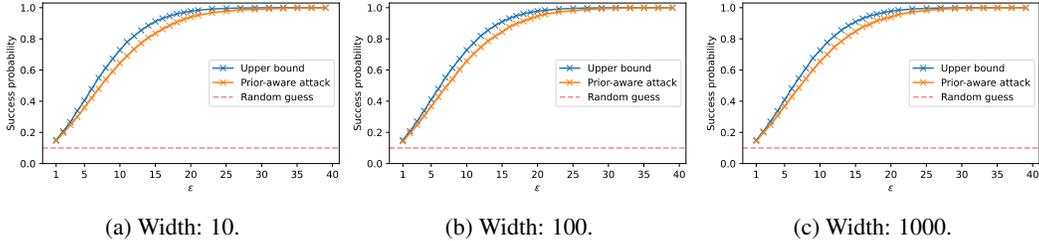


Figure 7: Comparison of model sizes on reconstruction by varying the hidden layer width in a two layer MLP.

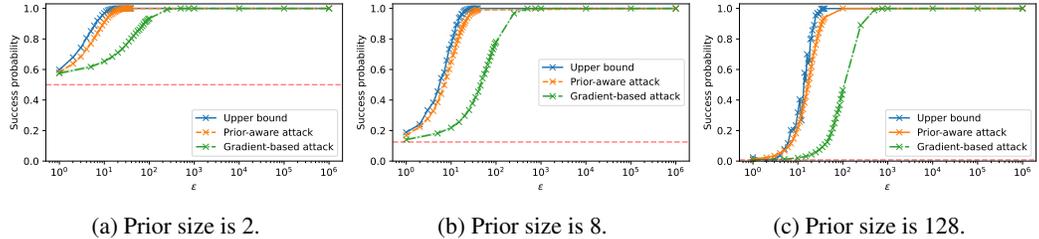


Figure 8: Comparison of prior-aware and gradient-based attack for different prior sizes.

## D Comparing the gradient-based attack with the prior-aware attack

Our experiments have mainly been concerned with measuring how DP affects an adversary’s ability to infer which point was included in training, given that they have access to all possible points that could have been included, in the form of a discrete prior. This experimental set-up departs from Figure 1, where we assumed the adversary does not have access to a prior set, and so cannot run the prior-aware attack as described in Algorithm 2. Following on from results in Section 4.1, we transform these gradient-based attack experimental findings into a probability of successful reconstruction by running a post-processing conversion, allowing us to measure how the assumption of adversarial access to the discrete prior affects reconstruction success. We run the post-processing conversion in the following way: Given a target sample  $z^*$  and a reconstruction  $\hat{z}$  found through optimizing the gradient based loss in Equation (1), we construct a prior consisting of  $z^*$  and  $n - 1$  randomly selected points from the MNIST dataset, where  $n = 10$ . We then measure the  $L_2$  distance between  $\hat{z}$  and every point in this constructed prior, and assign reconstruction a success if the smallest distance is with respect to  $z^*$ . For each target  $z^*$ , we repeat this procedure 1,000 times, with different random selections of size  $n - 1$ , and overall report the average reconstruction success over 1,000 different targets.

This allows us to compare the gradient-based attack (which is prior “unaware”) directly to our prior-aware attack. Results are shown in Figure 8, where we vary the size of the prior between 2, 8, and 128. In all cases, we see an order of magnitude difference between the gradient-based and prior-aware attack in terms of reconstruction success. This suggests that if we assume the adversary does not have prior knowledge of the possible set of target points, the minimum value of  $\epsilon$  necessary to protect against reconstruction attacks increases.

## E Effects of the threat model and prior distribution on reconstruction

The ability to reconstruct a training data point will naturally depend on the threat model in which the security game is instantiated. So far, we have limited our investigation to align with the standard adversary assumptions in the DP threat model. We have also limited ourselves to a setting where the prior is sampled from the same base distribution as  $D$ . These choices will change the performance of our attack, which is what we measure next.

**Prior type.** We measure how the choice of prior affects reconstruction in Figure 9. We train models when the prior is from the same distribution as the rest of the training set (MNIST), and when the prior is sampled random noise. Note, because the target point  $z^*$  is included in the prior, this means

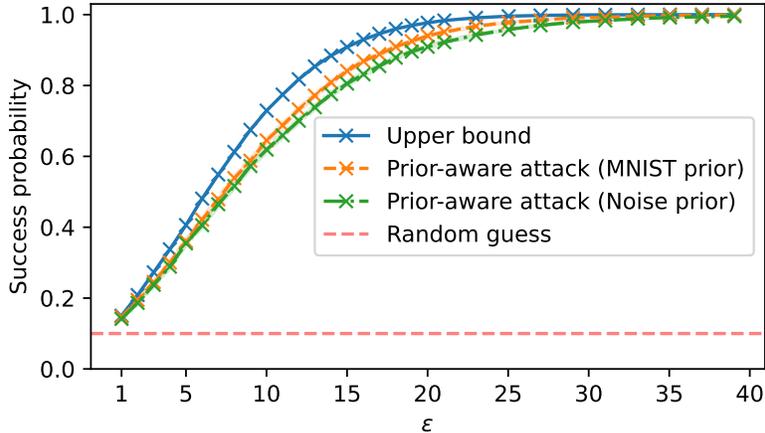


Figure 9: Comparison of how the choice of prior,  $\pi$ , affects reconstruction success. The prior is selected from a set of examples sampled from MNIST or uniform noise (that has the same intra-sample distance statistics as the MNIST prior).

we measure how reconstruction success changes when we change the distribution the target was sampled from. One may expect that the choice of prior to make a difference to reconstruction success if the attack relies on distinguishability between  $D$  and  $z^*$  with respect to some function operating on points and model parameters (e.g. the difference in loss between points in  $D$  and  $z^*$ ). However, we see that there is little difference between the two; both are close to the upper bound.

On reflection, this is expected as our objective is simply the sum of samples from a Gaussian, and so the choice of prior may impact our probability of correct inference if this choice affects the probability that a point will be clipped, or if points in the prior have correlated gradients. We explore how different values of clipping,  $C$ , can change reconstruction success probability in Appendix K.

**Knowledge of batch gradients.** The DP threat model assumes the adversary has knowledge of the gradients of all samples other than the target  $z^*$ . Here, we measure how important this assumption is to our attack. We compare the prior-aware attack (which maximizes  $\sum_{t=1}^T \langle \text{clip}_C(\nabla_{\theta_t} \ell(z_i)), \bar{g}_t \rangle$ ) against the attack that selects the  $z_i$  maximizing  $\sum_{t=1}^T \langle \text{clip}_C(\nabla_{\theta_t} \ell(z_i)), g_t \rangle$ , where the adversary does not subtract the known gradients from the objective.

In Figure 10, we compare, in a full-batch setting, when  $|D_-|$  is small (set to 4), and see the attack does perform worse when we do not deduct known gradients. However, the effect is more pronounced as  $|D_-|$  becomes larger, the attack completely fails when setting it to 128. This is somewhat expected, as with a larger number of samples in a batch it is highly likely there are gradients correlated with the  $z^*$  target gradient, masking out its individual contribution and introducing noise into the attack objective.

## F Improved prior-aware attack algorithm

As explained in Section 4.3, the prior-aware attack in Algorithm 2 does not account for the variance introduced into the attack objective in mini-batch DP-SGD, and so we design a more efficient attack specifically for the mini-batch setting. We give the pseudo-code for this improved prior-aware attack in Algorithm 3.

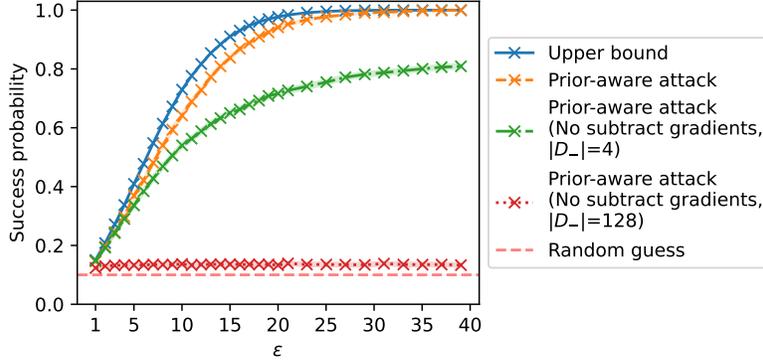


Figure 10: In line with the DP threat model, our attack in Algorithm 2 assumes the adversary can subtract known gradients from the privatized gradient. We measure what effect removing this assumption has on reconstruction success probability. When the size of the training set is small, removing this assumption has a minor effect, while reconstruction success drops to random with a larger training set size.

---

### Algorithm 3 Improved prior-aware attack

---

**Input:** Discrete prior  $\pi = \{z_1, \dots, z_n\}$ , Model parameters  $\{\theta_1, \theta_1, \dots, \theta_T\}$ , Privatized gradients (with known gradients subtracted)  $\{\bar{g}_1, \dots, \bar{g}_T\}$ , sampling probability  $q$ , function that takes the top  $qT$  values from a set of observed gradients  $top_{qT}$   
**Observations:**  $\mathcal{O} \leftarrow \{\}$   
**Output:** Reconstruction guess  $\hat{z} \in \pi$   
**for**  $i \in [1, 2, \dots, n]$  **do**  
     $\mathcal{R} \leftarrow \{\}$   
    **for**  $t \in [1, 2, \dots, T]$  **do**  
         $\mathcal{R}[t] \leftarrow \langle \text{clip}_C(\nabla_{\theta_t} \ell(\theta_t, z_i)), \bar{g}_t \rangle$   
    **end for**  
     $\mathcal{R} \leftarrow top_{qT}(\mathcal{R})$   
     $\mathcal{O}[i] \leftarrow sum(\mathcal{R})$   
**end for**  
 $\hat{i} \leftarrow \arg \max \mathcal{O}$   
**return**  $\hat{z} \leftarrow \pi[\hat{i}]$

---

## G Alternative variant of the prior-aware attack

Here, we state an alternative attack that uses the log-likelihood to find out which point in the prior set is used for training. Assume we have  $T$  steps with clipping threshold  $C = 1$ , noise  $\sigma$ , and the sampling rate is  $q$ .

Let  $\bar{g}_1, \dots, \bar{g}_T$  be the observed gradients minus the gradient of the examples that are known to be in the batch and let  $l_1, \dots, l_T$  be the  $\ell_2$  norms of these gradients.

For each example  $z$  in the prior set let  $g_1^z, \dots, g_T^z$  be the clipped gradient of the example on the intermediate model. Also let  $l_1^z, \dots, l_T^z$  be the  $\ell_2$  norms of  $(\bar{g}_1 - g_1^z), \dots, (\bar{g}_T - g_T^z)$ .

Now we describe the optimal attack based on  $l_i^z$ . For each example  $z$ , calculate the following:

$s_z = \sum_{i \in [T]} \ln(1 - q + qe^{-\frac{(l_i^z)^2 + l_i^2}{2\sigma^2}})$ . It is easy to observe that this is the log probability of outputting the steps conditioned on  $z$  being used in the training set. Then since the prior is uniform over the prior set, we can choose the  $z$  with maximum  $s_z$  and report that as the example in the batch.

In fact, this attack could be extended to the non-uniform prior by choosing the example that maximizes  $s_z \cdot p_z$ , where  $p_z$  is the original probability of  $z$ .

## H Comparison with Guo et al. (2022b)

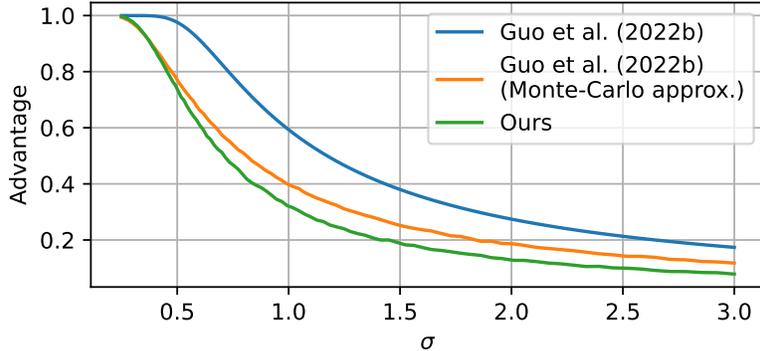


Figure 11: Comparison of our upper bound on advantage (Equation (4)) with Guo et al. (2022b) as function of  $\sigma$  for a uniform prior of size ten. We use a single step of DP-SGD with no mini-batch subsampling, and use 100,000 samples for Monte-Carlo approximation.

Table 2: Comparison of our upper bound on advantage (Equation (4)) with Guo et al. (2022b) and the Guo et al. (2022b) Monte-Carlo approximation (abbreviated to MC) as function of  $\sigma$  for a uniform prior size of ten and one hundred.

Prior size	Method	Advantage upper bound					
		$\sigma$					
		0.5	1	1.5	2	2.5	3
10	Guo et al. (2022b)	0.976	0.593	0.380	0.274	0.213	0.174
	Guo et al. (2022b) (MC)	0.771	0.397	0.257	0.184	0.144	0.118
	Ours	<b>0.737</b>	<b>0.322</b>	<b>0.189</b>	<b>0.128</b>	<b>0.099</b>	<b>0.080</b>
100	Guo et al. (2022b)	0.861	0.346	0.195	0.131	0.097	0.076
	Guo et al. (2022b) (MC)	0.549	0.210	0.120	0.081	0.062	0.049
	Ours	<b>0.362</b>	<b>0.077</b>	<b>0.035</b>	<b>0.024</b>	<b>0.018</b>	<b>0.012</b>

Recently, Guo et al. (2022b) have analyzed reconstruction of discrete training data. They note that DP bounds the mutual information shared between training data and learned parameters, and use Fano’s inequality to convert this into a bound on reconstruction success. In particular, they define the advantage of the adversary as

$$\text{Adv} := \frac{p_{\text{adversary success}} - p_{\pi}^{\max}}{1 - p_{\pi}^{\max}} \in [0, 1]. \quad (4)$$

where  $p_{\pi}^{\max}$  is the maximum sampling probability from the prior,  $\pi$ , and  $p_{\text{adversary success}}$  is the probability that the adversary is successful at inferring which point in the prior was included in training. They then bound the advantage by lower bounding the adversary’s error  $t := 1 - p_{\text{adversary success}}$  and by appealing to Fano’s inequality they show this can be done by finding the smallest  $t \in [0, 1]$  satisfying

$$\begin{aligned} f(t) := & H(\pi) - I(\pi; w) + t \log t + (1 - t) \log(1 - t) \\ & - t \log(|\pi| - 1) \leq 0, \end{aligned} \quad (5)$$

where  $w$  is output of the private mechanism,  $H(\pi)$  is the entropy of the prior, and  $I(\pi; w)$  is the mutual information between the prior and output of the private mechanism. For an  $(\alpha, \epsilon)$ -RDP mechanism,  $I(\pi; w) \leq \epsilon$ , and so  $I(\pi; w)$  can be replaced by  $\epsilon$  in Equation (5). However, Guo et al. (2022b) show that for the Gaussian mechanism, this can be improved upon either by using a Monte-Carlo approximation of  $I(\pi; w)$  — this involves approximating the KL divergence between a Gaussian and a Gaussian mixture — or by showing that  $I(\pi; w) \leq -\sum_{i=1}^{|\pi|} p_{\pi}^i \log \left( p_{\pi}^i + (1 - p_{\pi}^i) \exp \left( \frac{-\Delta^2}{2\sigma^2} \right) \right)$ ,

where  $\Delta$  is the sensitivity of the mechanism, and  $p_\pi^i$  is the probability of selecting the  $i$ th element from the prior. We use a uniform prior in all our experiments and so  $H(\pi) = -\log(\frac{1}{|\pi|})$  and  $p_\pi^i = p_\pi^{\max} = \frac{1}{|\pi|}$ .

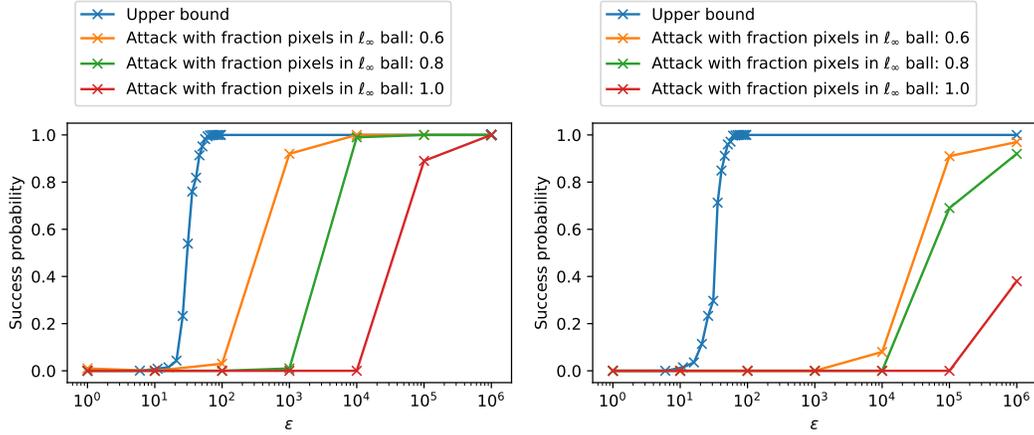
We convert our bound on success probability to advantage and compare with the Guo et al. (2022b) upper bound (and its Monte-Carlo approximation) in Figure 11 and Table 2, and note our bound is tighter.

## I Experiments with *very* small priors

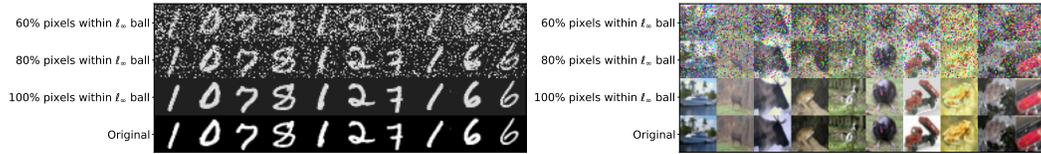
Our experiments in Section 3 and Section 4.3 were conducted with an adversary who has side information about the target point. Here, we reduce the amount of background knowledge the adversary has about the target, and measure how this affects the reconstruction upper bound and attack success.

We do this in the following set-up: Given a target  $z$ , we initialize our reconstruction from uniform noise and optimize with the gradient-based reconstruction attack introduced in Section 2 to produce  $\hat{z}$ . We mark  $\hat{z}$  as a successful reconstruction of  $z$  if  $\frac{1}{d} \sum_{i=1}^d \mathbb{I}[|z[i] - \hat{z}[i]| < \delta] \geq \rho$ , where  $\rho \in [0, 1]$ ,  $d$  is the data dimensionality, and we set  $\delta = \frac{32}{255}$  in our experiments. If  $\rho = 1$  this means we mark the reconstruction as successful if  $\|\hat{z} - z\|_\infty < \delta$ , and for  $\rho < 1$ , then at least a fraction  $\rho$  values in  $\hat{z}$  must be within an  $\ell_\infty$  ball of radius  $\delta$  from  $z$ . Under the assumption the adversary has no background knowledge of the target point, with  $\delta = \frac{32}{255}$  and a uniform prior, the prior probability of reconstruction is given by  $(2 \times 32/256)^{d\rho}$  — if  $\rho = 1$ , for MNIST and CIFAR-10, this means the prior probability of a successful reconstruction is  $9.66 \times 10^{-473}$  and  $2.96 \times 10^{-1850}$ , respectively.

We plot the reconstruction upper bound compared to the attack success for different values of  $\rho$  in Figure 12. We also visualize the quality of reconstructions for different values of  $\rho$ . Even for  $\rho = 0.6$ , where 40% of the reconstruction pixels can take any value, and the remaining 60% are within an absolute value of  $\frac{32}{255}$  from the target, one can easily identify that the reconstructions look visually similar to the target.



(a) Comparison of reconstruction success under a *very* small prior for MNIST, where we judge a reconstruction as successful if at least  $\rho$  pixels are within an absolute distance of  $\frac{32}{255}$  of the target. (b) Comparison of reconstruction success under a *very* small prior for CIFAR-10, where we judge a reconstruction as successful if at least  $\rho$  pixels are within an absolute distance of  $\frac{32}{255}$  of the target.



(c) MNIST examples of reconstructions where at least  $\rho$  pixels are within an absolute distance of  $\frac{32}{255}$  of the target. (d) CIFAR-10 examples of reconstructions where at least  $\rho$  pixels are within an absolute distance of  $\frac{32}{255}$  of the target.

Figure 12: Comparison of reconstruction success under a *very* small prior. The prior probability of success for MNIST and CIFAR-10 are  $9.66 \times 10^{-473}$  and  $2.96 \times 10^{-1850}$ , respectively.

## J Discussion on related work

Here, we give a more detailed discussion of relevant related work over what is surfaced in Section 1 and Section 2.

**DP and reconstruction.** By construction, differential privacy bounds the success of a membership inference attack, where the aim is to infer if a point  $z$  was in or out of the training set. While the connection between membership inference and DP is well understood, less is known about the relationship between training data reconstruction attacks and DP. A number of recent works have begun to remedy this in the context of models trained with DP-SGD by studying the value of  $\epsilon$  required to thwart training data reconstruction attacks (Bhowmick et al., 2018; Balle et al., 2022; Guo et al., 2022a,b; Stock et al., 2022). Of course, because differential privacy bounds membership inference, it will also bound ones ability to reconstruct training data; if one cannot determine if  $z$  was used in training, they will not be able to reconstruct that point. These works are interested in both formalizing training data reconstruction attacks, and quantifying the necessary  $\epsilon$  required to bound its success. Most of these works share a common finding – the  $\epsilon$  value needed for this bound is much larger than the value required to protect against membership inference attacks ( $< 10$  in practice). If all other parameters in  $q\sqrt{T\log(\frac{1}{\delta})}/\epsilon$  remain fixed, one can see that a larger value of  $\epsilon$  reduces the scale of noise we add to gradients, which in turn results in models that achieve smaller generalization error than models trained with DP-SGD that protect against membership inference.

The claim that a protection against membership inference attacks also protects against training data reconstruction attacks glosses over many subtleties. For example, if  $z$  was not included in training it could still have a non-zero probability of reconstruction if samples that are close to  $z$  were included in training. Balle et al. (2022) take the approach of formalizing training reconstruction attacks in a

Bayesian framework, where they compute a prior probability of reconstruction, and then find how much more information an adversary gains by observing the output of DP-SGD.

Balle et al. (2022) use an average-case definition of reconstruction over the output of a randomized mechanism. In contrast, Bhowmick et al. (2018) define a worst-case formalization, asking when should an adversary not be able to reconstruct a point of interest regardless of the output of the mechanism. Unfortunately, such worst-case guarantees are not attainable under DP-relaxations like  $(\epsilon, \delta)$ -DP and RDP, because the privacy loss is not bounded; there is a small probability that the privacy loss will be high.

Stock et al. (2022) focus on bounding reconstruction for language tasks. They use the probability preservation guarantee from RDP to derive reconstruction bounds, showing that the length of a secret within a piece of text itself provides privacy. They translate this to show a smaller amount of DP noise is required to protect longer secrets.

While Balle et al. (2022) propose a Bayesian formalization for reconstruction error, Guo et al. (2022a) propose a frequentist definition. They show that if  $M$  is  $(2, \epsilon)$ -RDP, then the reconstruction MSE is lower bounded by  $\sum_{i=1}^d \text{diam}_i(\mathcal{Z})^2 / 4d(e^\epsilon - 1)$ , where  $\text{diam}_i(\mathcal{Z})$  is the diameter of the space  $\mathcal{Z}$  in the  $i$ th dimension.

**Gradient inversion attacks.** The early works of Wang et al. (2019) and Zhu et al. (2019) showed that one can invert single image representation from gradients of a deep neural network. Zhu et al. (2019) actually went beyond this and showed one can jointly reconstruct both the image and label representation. The idea is that given a target point  $z$ , a loss function  $\ell$ , and an observed gradient (wrt to model parameters  $\theta$ )  $g_z = \nabla_\theta \ell(\theta, z)$ , to construct a  $\hat{z}$  such that  $\hat{z} = \arg \min_{z'} \|g_{z'} - g_z\|$ . The expectation is that images that have similar gradients will be visually similar. By optimizing the above objective with gradient descent, Zhu et al. (2019) showed that one can construct visually accurate reconstruction on standard image benchmark datasets like CIFAR-10.

Jeon et al. (2021); Yin et al. (2021); Jin et al. (2021); Huang et al. (2021); Geiping et al. (2020) proposed a number of improvements over the reconstruction algorithm used in Zhu et al. (2019): they showed how to reconstruct multiple training points in batched gradient descent, how to optimize against batch normalization statistics, and incorporate priors into the optimization procedure, amongst other improvements.

The aforementioned attacks assumed an adversary has access to gradients through intermediate model updates. Balle et al. (2022) instead investigate reconstruction attacks when adversary can only observe a model after it has finished training, and propose attacks against (parametric) ML models under this threat model. However, the attack they construct is computationally demanding as it involves retraining thousands of models. This computational bottleneck is also a factor in Haim et al. (2022), who also investigate training data reconstruction attacks where the adversary has access only to final model parameters.

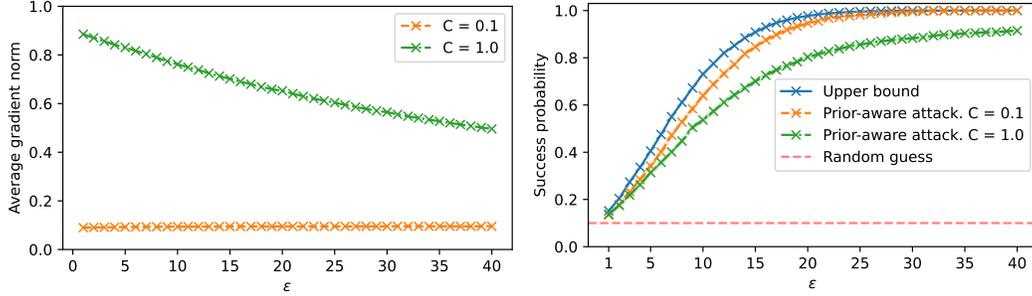
## K More experiments on the effect of DP-SGD hyperparameters

We extend on our investigation into the effect that DP-SGD hyperparameters have on reconstruction. We begin by varying the clipping norm parameter,  $C$ , and measure the effect on reconstruction. Following this, we replicate our results from Section 4.3 (the effect hyperparameters have on reconstruction at a fixed  $\epsilon$ ) across different values of  $\epsilon$  and prior sizes,  $|\pi|$ .

### K.1 Effect of clipping norm

If we look again at our attack set-up in Algorithm 2, we see that in essence we are either summing a set of samples only from a Gaussian centred at zero or a Gaussian centred at  $C^2$ . If the gradient of the target point is not clipped, then this will reduce the sum of gradients when the target is included in a batch, as the Gaussian will be centred at a value smaller than  $C^2$ . This will increase the probability that the objective is not maximized by the target point.

We demonstrate how this changes the reconstruction success probability by training a model for 100 steps with a clipping norm of 0.1 or 1, and measuring the average gradient norm of all samples over all steps. Results are shown in Figure 13. We see at  $C = 0.1$ , our attack is tight to the upper bound,



(a) Average gradient norm (over all samples and steps) for different values of  $\epsilon$  at  $C = 0.1$  and  $C = 1$ . (b) Reconstruction success probability for different values of  $\epsilon$  at  $C = 0.1$  and  $C = 1$ .

Figure 13: Comparison of how reconstruction success is changes with the clipping norm,  $C$ . We see that if examples have a gradient norm smaller than  $C$ , and so are not clipped, reconstruction success probability becomes smaller.

and the average gradient norm is 0.1 for all values of  $\epsilon$ ; all individual gradients are clipped. When  $C = 1$ , the average gradient norm decreases from 0.9 at  $\epsilon = 1$  to 0.5 at  $\epsilon = 40$ , and we see a larger gap between upper and lower bounds. The fact that some gradients may not be clipped is not taken into account by our theory used to compute upper bounds, and so we conjecture that the reduction in reconstruction success is a real effect rather than a weakness of our attack.

We note that these findings chime with work on individual privacy accounting (Feldman & Zrnic, 2021; Yu et al., 2022; Ligett et al., 2017; Redberg & Wang, 2021). An individual sample’s privacy loss is often much smaller than what is accounted for by DP bounds. These works use the gradient norm of an individual sample to measure the true privacy loss, the claim is that if the gradient norm is smaller than the clipping norm, the amount of noise added is too large, as the DP accountant assumes all samples are clipped. Our experiments support the claim that there is a disparity in privacy loss between samples whose gradients are and are not clipped.

## K.2 More results on the effect of DP-SGD hyperparameters at a fixed $\epsilon$

In Section 4.3, we demonstrated that the success of a reconstruction attack cannot be captured only by the  $(\epsilon, \delta)$  guarantee, when  $\epsilon = 4$  and the size of the prior,  $\pi$ , is set to ten. We now observe how these results change across different  $\epsilon$  and  $|\pi|$ , where we again fix the number of updates to  $T = 100$ ,  $C = 1$ , vary  $q \in [0.01, 0.99]$ , and adjust  $\sigma$  accordingly.

Firstly, in Figure 14, we measure the upper *and* lower bound ((improved) prior-aware attack) on the probability of successful reconstruction across different  $q$ . In all settings, we observe smaller reconstruction success at smaller  $q$ , where the largest fluctuations in reconstruction success are for larger values of  $\epsilon$ . We visualise this in another way by plotting  $\sigma$  against  $q$  and report the upper bound in Figure 15. Note that the color ranges in Figure 15 are independent across subfigures.

## L Estimating $\kappa$ from samples

Here, we discuss how to estimate the base probability of reconstruction success,  $\kappa$ , if the adversary can only sample from the prior distribution.

Let  $\hat{\pi}$  be the empirical distribution obtained by taking  $N$  independent samples from the prior and  $\hat{\kappa} = \kappa_{\hat{\pi}, \rho}(\eta)$  be the corresponding parameter for this discrete approximation to  $\pi$  – this can be computed using the methods sketched in Section 3. Then we have the following concentration bound.

**Proposition 6.** *With probability  $1 - e^{-N\tau^2\kappa/2}$  we have*

$$\kappa \leq \frac{\hat{\kappa}}{1 - \tau} .$$

The proof is given in Appendix M.

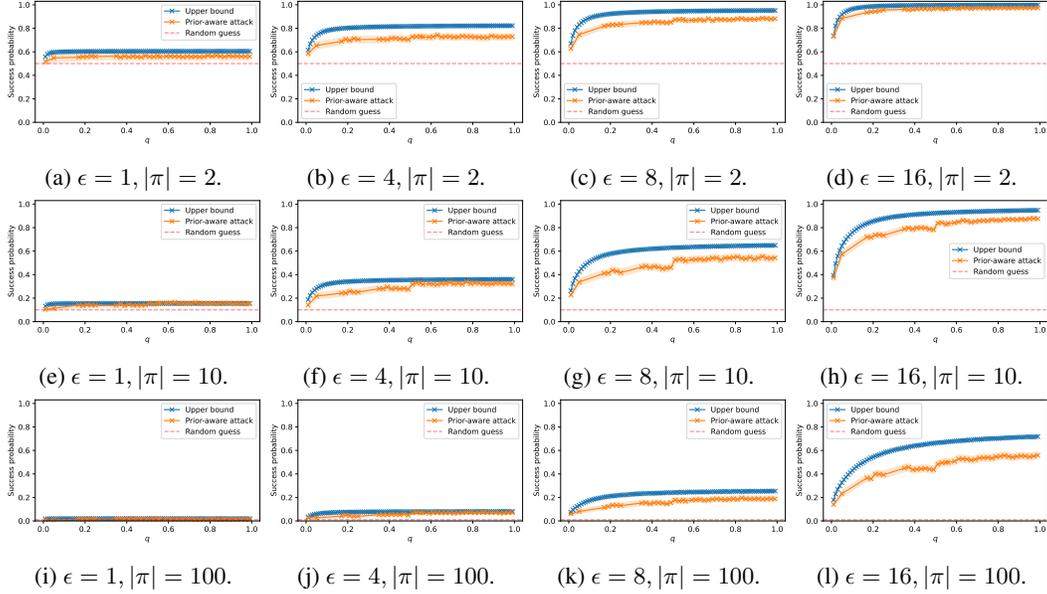


Figure 14: How the upper bound and (improved) prior-aware attack change as a function of  $q$  at a fixed value of  $\epsilon$  and prior size,  $|\pi|$ . The amount of privacy leaked through a reconstruction at a fixed value of  $\epsilon$  can change with different  $q$ .

## M Proofs

Throughout the proofs we make some of our notation more succinct for convenience. For a probability distribution  $\omega$  we write  $\omega(E) = \mathbb{P}_\omega[E]$ , and rewrite  $\mathcal{B}_\kappa(\mu, \nu) = \sup\{\mathbb{P}_\mu[E] : E \text{ s.t. } \mathbb{P}_\nu[E] \leq \kappa\}$  as  $\sup_{\nu(E) \leq \kappa} \mu(E)$ . Given a distribution  $\omega$  and function  $\phi$  taking values in  $[0, 1]$  we also write  $\omega(\phi) = \mathbb{E}_{X \sim \omega}[\phi(X)]$ .

### M.1 Proof of Theorem 2

We say that a pair of distributions  $(\mu, \nu)$  is *testable* if for all  $\kappa \in [0, 1]$  we have

$$\inf_{\nu(\phi) \leq \kappa} (1 - \mu(\phi)) = \inf_{\nu(E) \leq \kappa} (1 - \mu(E)) ,$$

where the infimum on the left is over all  $[0, 1]$ -valued measurable functions and the one on the right is over measurable events (i.e.  $\{0, 1\}$ -valued functions). The Neyman-Pearson lemma (see e.g. Lehmann & Romano (2005)) implies that this condition is satisfied whenever the statistical hypothesis problem of distinguishing between  $\mu$  and  $\nu$  admits a uniformly most powerful test. For example, this is the case for distributions on  $\mathbb{R}^d$  where the density ratio  $\mu/\nu$  is a continuous function.

**Theorem 7** (Formal version of Theorem 2). *Fix  $\pi$  and  $\rho$ . Suppose that for every fixed dataset  $D$ . there exists a distribution  $\mu_{D_\cdot}$  such that  $\sup_{z \in \text{supp}(\pi)} \mathcal{B}_\kappa(\mu_{D_\cdot}, \nu_{D_\cdot}) \leq \mathcal{B}_\kappa(\mu_{D_\cdot}, \nu_{D_\cdot})$  for all  $\kappa \in [0, 1]$ . If the pair  $(\mu, \nu)$  is testable, then  $M$  is  $(\eta, \gamma)$ -ReRo with*

$$\gamma = \sup_{D_\cdot} \sup_{\nu_{D_\cdot}(E) \leq \kappa_{\pi, \rho}(\eta)} \mu_{D_\cdot}(E) .$$

The following lemma from Dong et al. (2019) will be useful.

**Lemma 8.** *For any  $\mu$  and  $\nu$ , the function  $\kappa \mapsto \inf_{\nu(\phi) \leq \kappa} (1 - \mu(\phi))$  is convex in  $[0, 1]$ .*

**Lemma 9.** *For any testable pair  $(\mu, \nu)$ , the function  $\kappa \mapsto \sup_{\nu(E) \leq \kappa} \mu(E)$  is concave.*

*Proof.* By the testability assumption we have

$$\begin{aligned}
\sup_{\nu(E) \leq \kappa} \mu(E) &= \sup_{\nu(E) \leq \kappa} \mu(E) \\
&= \sup_{\nu(E) \leq \kappa} (1 - \mu(\bar{E})) \\
&= 1 - \inf_{\nu(E) \leq \kappa} \mu(\bar{E}) \\
&= 1 - \inf_{\nu(E) \leq \kappa} (1 - \mu(E)) \\
&= 1 - \inf_{\nu(\phi) \leq \kappa} (1 - \mu(\phi)) .
\end{aligned}$$

Concavity now follows from Lemma 8.  $\square$

*Proof of Theorem 7.* Fix  $D$ . and let  $\kappa = \kappa_{\pi, \rho}(\eta)$  throughout. Let also  $\nu = \nu_{D_\cdot}$ ,  $\mu_z = \mu_{D_z}$ ,  $\nu^* = \nu_{D_\cdot}^*$  and  $\mu^* = \mu_{D_\cdot}^*$ .

Expanding the probability of successful reconstruction, we get:

$$\begin{aligned}
\mathbb{P}_{Z \sim \pi, W \sim M(D_\cdot \cup \{Z\})}[\rho(Z, R(W)) \leq \eta] &= \mathbb{E}_{Z \sim \pi} \mathbb{P}_{W \sim M(D_\cdot \cup \{Z\})}[\rho(Z, R(W)) \leq \eta] \\
&= \mathbb{E}_{Z \sim \pi} \mathbb{E}_{W \sim M(D_\cdot \cup \{Z\})} \mathbb{I}[\rho(Z, R(W)) \leq \eta] \\
&= \mathbb{E}_{Z \sim \pi} \mathbb{E}_{W \sim \mu_Z} \mathbb{I}[\rho(Z, R(W)) \leq \eta] \\
&= \mathbb{E}_{Z \sim \pi} \mathbb{E}_{W \sim \nu} \left[ \frac{\mu_Z(W)}{\nu(W)} \mathbb{I}[\rho(Z, R(w)) \leq \eta] \right] .
\end{aligned}$$

Now fix  $z \in \text{supp}(\pi)$  and let  $\kappa_z = \mathbb{P}_{W \sim \nu}[\rho(z, R(W)) \leq \eta]$ . Using the assumption on  $\mu^*$  we get:

$$\begin{aligned}
\mathbb{E}_{W \sim \nu} \left[ \frac{\mu_z(W)}{\nu(W)} \mathbb{I}[\rho(z, R(w)) \leq \eta] \right] &\leq \sup_{\nu(E) \leq \kappa_z} \mathbb{E}_{W \sim \nu} \left[ \frac{\mu_z(W)}{\nu(W)} \mathbb{I}[W \in E] \right] && \text{(By definition of } \kappa) \\
&= \sup_{\nu(E) \leq \kappa_z} \mathbb{E}_{W \sim \mu_z} [\mathbb{I}[W \in E]] \\
&= \sup_{\nu(E) \leq \kappa_z} \mu_z(E) \\
&\leq \sup_{\nu^*(E) \leq \kappa_z} \mu^*(E) . && \text{(By definition of } \mu^* \text{ and } \nu^*)
\end{aligned}$$

Finally, using Lemma 9 and Jensen's inequality on the following gives the result:

$$\begin{aligned}
\mathbb{E}_{Z \sim \pi}[\kappa_Z] &= \mathbb{E}_{Z \sim \pi} \mathbb{P}_{W \sim \nu}[\rho(Z, R(W)) \leq \eta] \\
&= \mathbb{E}_{W \sim \nu} \mathbb{P}_{Z \sim \pi}[\rho(Z, R(W)) \leq \eta] \\
&\leq \mathbb{E}_{W \sim \nu} \kappa \\
&= \kappa .
\end{aligned}$$

$\square$

## M.2 Proof of Corollary 4

Here we prove Corollary 4. We will use the following shorthand notation for convenience:  $\mu = \mathcal{N}(B(T, q), \sigma^2 I)$  and  $\nu = \mathcal{N}(0, \sigma^2 I)$ . To prove our result, we use the notion of  $TV_a$ .

**Definition 10** (Mahloujifar et al. (2022)). *For two probability distributions  $\omega_1(\cdot)$  and  $\omega_2(\cdot)$ ,  $TV_a$  is defined as*

$$TV_a(\omega_1, \omega_2) = \int |\omega_1(x) - a \cdot \omega_2(x)| dx.$$

Now we state the following lemma borrowed from Mahloujifar et al. (2022).

**Lemma 11** (Theorem 6 in Mahloujifar et al. (2022)). *Let  $\nu_{D_\cdot}$ ,  $\mu_{D_z}$  be the output distribution of DP-SGD applied to  $D$ . and  $D_z$  respectively, with noise multiplier  $\sigma$ , sampling rate  $q$ . Then we have*

$$TV_a(\nu_{D_\cdot}, \mu_{D_z}) \leq TV_a(\nu, \mu) .$$

Now, we state the following lemma that connects  $TV_a$  to blow-up function.

**Lemma 12** (Lemma 21 in Zhu et al. (2022)). *For any pair of distributions  $\omega_1, \omega_2$  we have*

$$\sup_{\omega_1(E) \leq \kappa} \omega_2(E) = \inf_{a > 1} \min \left\{ 0, a \cdot \kappa + \frac{TV_a(\omega_1, \omega_2) + 1 - a}{2}, \frac{2\kappa + TV_a(\omega_1, \omega_2) + a - 1}{2a} \right\}$$

Since  $TV_a(\nu_{D_-}, \mu_{D_z})$  is bounded by  $TV_a(\nu, \mu)$  for all  $a$ , therefore we have

$$\sup_{\nu_{D_-}(E) \leq \kappa} \mu_{D_z}(E) \leq \sup_{\nu(E) \leq \kappa} \mu(E) .$$

### M.3 Proof of Proposition 6

Recall  $\kappa = \sup_{z_0 \in \mathcal{Z}} \mathbb{P}_{Z \sim \pi}[\rho(Z, z_0) \leq \eta]$  and  $\hat{\kappa} = \sup_{z_0 \in \mathcal{Z}} \mathbb{P}_{Z \sim \hat{\pi}}[\rho(Z, z_0) \leq \eta]$ . Let  $\kappa_z = \mathbb{P}_{Z \sim \pi}[\rho(Z, z) \leq \eta]$  and  $\hat{\kappa}_z = \mathbb{P}_{Z \sim \hat{\pi}}[\rho(Z, z) \leq \eta]$ . Note  $\hat{\kappa}_z$  is the sum of  $N$  i.i.d. Bernoulli random variables and  $\mathbb{E}_{\hat{\pi}}[\hat{\kappa}_z] = \kappa_z$ . Then, using a multiplicative Chernoff bound, we see that for a fixed  $z$  the following holds with probability at least  $1 - e^{-N\tau_z^2 \kappa/2}$ :

$$\kappa_z \leq \frac{\hat{\kappa}_z}{1 - \tau} .$$

Applying this to  $z^* = \arg \sup_{z_0 \in \mathcal{Z}} \mathbb{P}_{Z \sim \pi}[\rho(Z, z_0) \leq \eta]$  we get that the following holds with probability at least  $1 - e^{-N\tau^2 \kappa/2}$ :

$$\kappa = \kappa_{z^*} \leq \frac{\hat{\kappa}_{z^*}}{1 - \tau} \leq \frac{\hat{\kappa}}{1 - \tau} .$$

### M.4 Proof of Proposition 5

Let  $z = \frac{(r'_{N'} + r'_{N'-1})}{2}$ . Let  $E_1$  be the event that  $|\mathbb{P}[r > z] - \kappa| \geq \tau$ . By applying Chernoff-Hoeffding bound we have  $\mathbb{P}[E_1] \leq 2e^{-2N\tau^2}$ . Now note that since  $\mu$  is a Gaussian mixture, we can write  $\mu = \sum_{i \in [2T]} a_i \mu_i$  where each  $\mu_i$  is a Gaussian  $\mathcal{N}(c_i, \sigma)$  where  $|c_i|_2 \leq \sqrt{T}$ . Now let  $r_i = \mu_i(W)/\nu(W)$ . By holder, we have  $\mathbb{E}[r^2] \leq \sum a_i \mathbb{E}[r_i^2]$ . We also now that  $\mathbb{E}[r_i^2] \leq e^T$ , therefore,  $\mathbb{E}[r^2] \leq e^T$ . Now let  $E_2$  be the event that  $|\mathbb{E}[r \cdot I(r > z)] - \gamma'| \geq \tau$ . Since the second moment of  $r$  is bounded, the probability of  $E_2$  goes to zero as  $N$  increases. Therefore, almost surely we have

$$\sup_{\nu(E) \leq \kappa - \tau} \mu(E) - \tau \leq \lim_{N \rightarrow \infty} \gamma' \leq \sup_{\nu(E) \leq \kappa + \tau} \mu(E) + \tau.$$

Now by pushing  $\tau$  to 0 and using the fact that  $\mu$  and  $\nu$  are smooth we have

$$\lim_{N \rightarrow \infty} \gamma' = \lim_{\tau \rightarrow 0} \sup_{\nu(E) \leq \kappa + \tau} \mu(E) + \tau = \sup_{\nu(E) \leq \kappa} \mu(E).$$

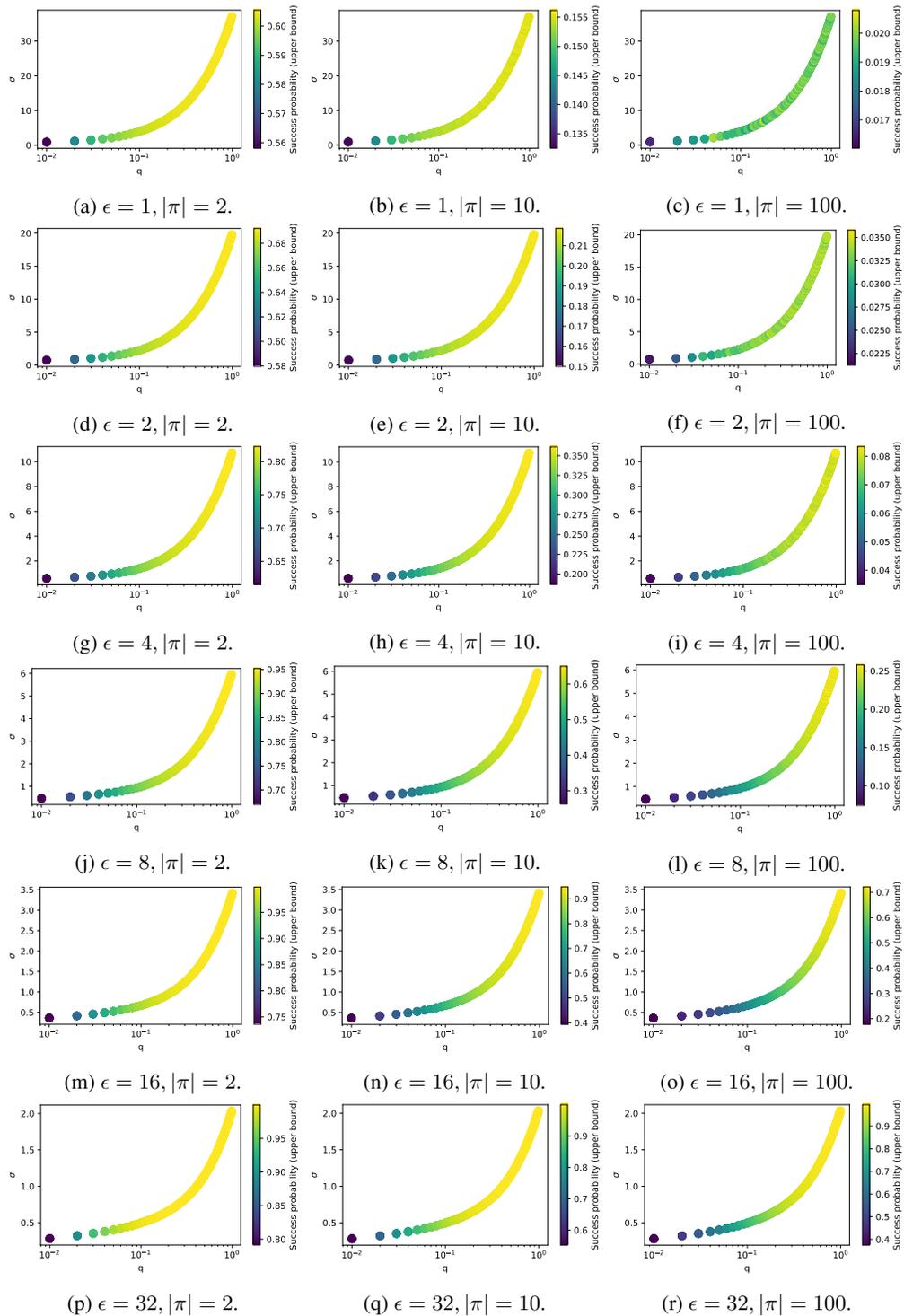


Figure 15: How the upper bound changes as a function of  $q$  and  $\sigma$  at a fixed value of  $\epsilon$  and prior size,  $|\pi|$ , and setting  $T = 100$ . The probability of a successful reconstruction can vary widely with different values of  $q$ . For example, at  $\epsilon = 32$  and  $|\pi| = 100$ , at  $q = 0.01$  the upper bound is 0.4 and at  $q = 0.99$  it is 1. Note that the color ranges are independent across subfigures.