Table 1: Hyperparameter settings for each experiment. The set of examples that make up the prior (π), including the target (z^*), and the other examples in the training set ($D_{\text{-}}$) 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 θ	Training dataset size ($\lvert D_{\text{-}}\rvert + 1$)	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
MNIST	Figure 1a, Figure 6a, Figure 12a	0.1	1	100	MLP ($784 \to 10 \to 10$)	1,000	-
	Figure 3	0.1	1	100	MLP ($784 \to 10 \to 10$)	1,000	$2^1 - 2^{11}$
	Figure 4	0.1	0.02	1,000	MLP ($784 \to 10 \to 10$)	500	10
	Figure 5	1	0.01-0.99	100	MLP ($784 \to 10 \to 10$)	1,000	10
	Figure 7	0.1	1	100	MLP ($784 \to 10, 100, 1000 \to 10$)	1,000	10
	Figure 8	0.1	1	100	MLP ($784 \to 10 \to 10$)	1,000	$2^1, 2^3, 2^7$
	Figure 9	0.1	1	100	MLP ($784 \to 10 \to 10$)	1,000	10
	Figure 10	0.1	1	100	MLP ($784 \to 10 \to 10$)	5, 129, 1,000	10
	Figure 11	1.0	1	-	-	-	10
	Figure 13	0.1, 1	1	100	MLP ($784 \to 10 \to 10$)	1,000	10
	Figure 14	1	0.01-0.99	100	MLP ($784 \to 10 \to 10$)	1,000	2, 10, 100
	Figure 15	1	0.01-0.99	100	MLP ($784 \to 10 \to 10$)	1,000	2, 10, 100

A Experimental details

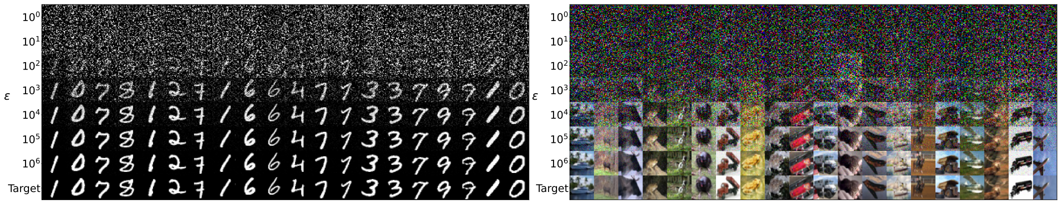
We detail the experimental settings used throughout the paper, and specific hyperparameters used for the various attacks we investigate. The exact configurations for each experiment are given in Table 1. We vary many experimental hyperparameters to investigate their effect on reconstruction, however, 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 $|D_{\text{-}}| = 999$ (and so the training set size is $|D_{\text{-}} \cup \{z^*\}| = 1,000$) and train with full-batch DP-SGD for $T = 100$ steps. For each ϵ, 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 σ for a given target ϵ. 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 π 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 $|D_{\text{-}}| = 499$ (and so the training set size is $|D_{\text{-}} \cup \{z^*\}| = 500$) and train with full-batch DP-SGD for $T = 100$ steps; again we sweep over the choice of learning rate for each value of ϵ. We set $C = 1$, $\delta = 10^{-5}$ and adjust the noise scale σ for a given target ϵ. 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 π 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 $D_{\text{-}}$) 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.



(a) Examples of reconstructions on MNIST. (b) Examples of reconstructions on CIFAR-10.

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 be 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 3.3 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.

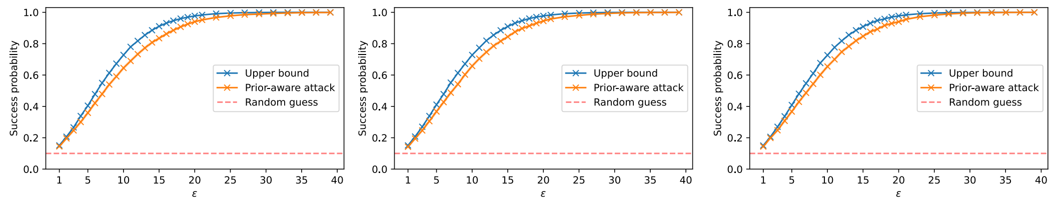
Estimating γ details. As described in Section 3, ν is instantiated as $\mathcal{N}(0, \sigma^2 I)$, a T-dimensional isotropic Gaussian distribution with zero mean, and μ is given by $\sum_{w \in \{0,1\}^T} p(w)\mathcal{N}(w, \sigma^2 I)$, a mixture of T-dimensional isotropic Gaussian distributions with 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 γ given by the procedure in Algorithm 1, and because we use a discrete prior of size $|\pi|$, the base probability of reconstruction success, κ, is given as ${}^1\!/_{|\pi|}$.

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.

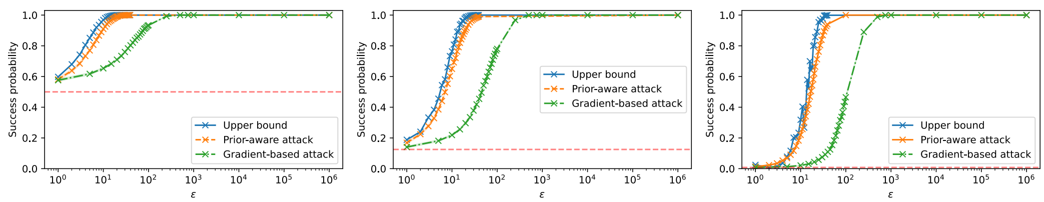
C Does the model size make a difference to the prior-aware attack?

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.



(a) Width: 10. (b) Width: 100. (c) Width: 1000.

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



(a) Prior size is 2. (b) Prior size is 8. (c) Prior size is 128.

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 3.4, 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 ϵ 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

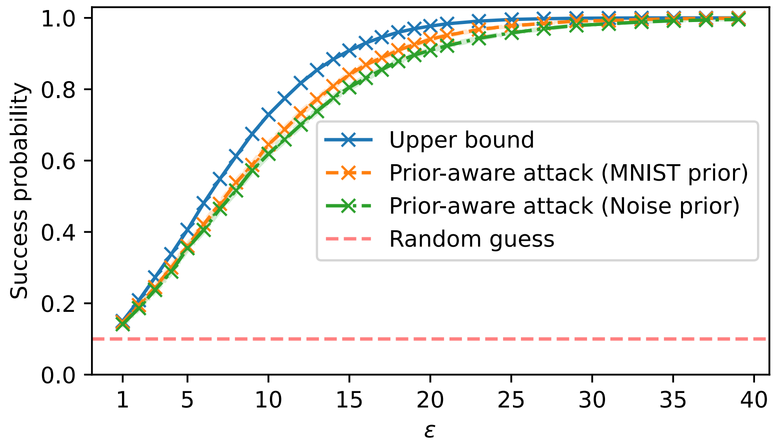


Figure 9: Comparison of how the choice of prior, π, affects reconstruction success. The prior is selected from a set of examples sampled from MNIST or uniform noise (that has the same intra-sample distance statistics as the MNIST prior).

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 $D_{\text{-}}$ and z^* with respect to some function operating on points and model parameters (e.g. the difference in loss between points in $D_{\text{-}}$ 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 L.

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 $|D_{\text{-}}|$ 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 $|D_{\text{-}}|$ 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, 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.

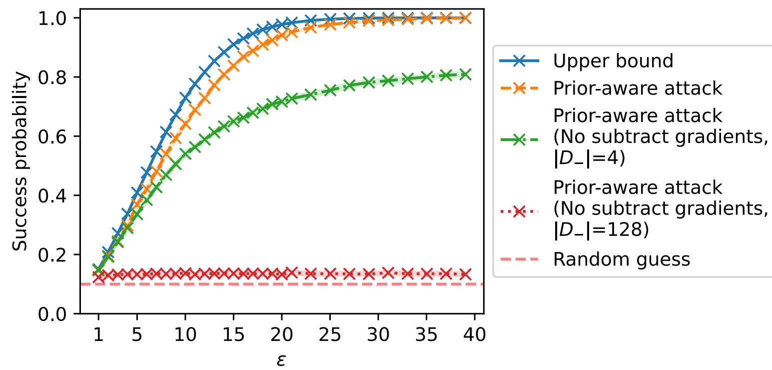


Figure 10: In line with the DP threat model, our attack in Algorithm 2 assumes the adversary can subtract known gradients from the privatized gradient. We measure what effect removing this assumption has on reconstruction success probability. When the size of the training set is small, removing this 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**
        $\mathcal{R}[t] \leftarrow \langle \mathrm{clip}_C(\nabla_{\theta_t} \ell(\theta_t, z_i)), \bar{g}_t \rangle$
    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 σ, 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 ℓ_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 ℓ_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)

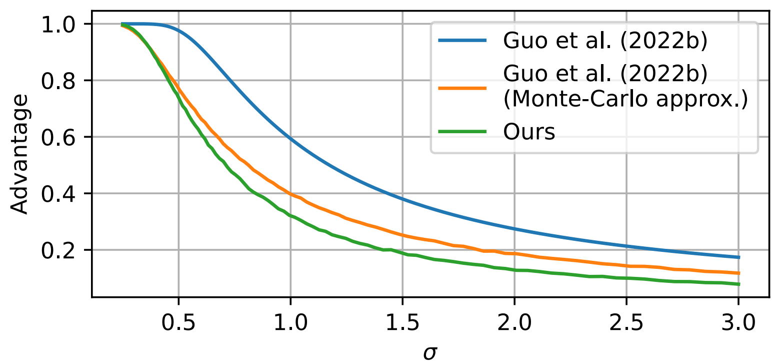


Figure 11: Comparison of our upper bound on advantage (Equation (4)) with Guo et al. (2022b) as function of σ for a uniform prior of size ten. We use a single step of DP-SGD with no mini-batch subsampling, and use $100,000$ samples for Monte-Carlo approximation.

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 σ for a uniform prior size of ten and one hundred.

Prior size	Method	Advantage upper bound					
		σ					
		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	**0.737**	**0.322**	**0.189**	**0.128**	**0.099**	**0.080**
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	**0.362**	**0.077**	**0.035**	**0.024**	**0.018**	**0.012**

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

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

where $p_\pi^{\max}$ is the maximum sampling probability from the prior, π, 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

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

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 (α, ϵ)-RDP mechanism, $I(\pi; w) \leq \epsilon$, and so $I(\pi; w)$ can be replaced by ϵ in Equation (5). However, Guo et al. (2022b) show that for the Gaussian mechanism, this can 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 Δ is the sensitivity of the mechanism, and p_π^i is the probability of selecting the ith element from the prior. We use a uniform prior in all our experiments and so $H(\pi) = -\log(\frac{1}{|\pi|})$ and $p_\pi^i = p_\pi^{\max} = \frac{1}{|\pi|}$.

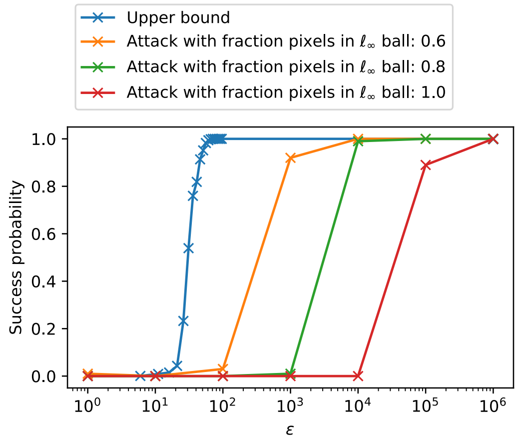
We convert our bound on success probability to advantage and compare with the Guo et al. (2022b) upper bound (and its Monte-Carlo approximation) in Figure 11 and Table 2, and note our bound is tighter.

I Experiments with *very* small priors (aka. experiments where the adversary has no background knowledge about the target)

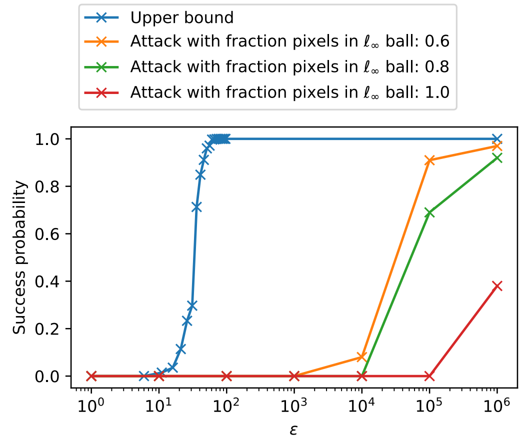
Our experiments in Section 3 and Section 4 were conducted with an adversary who has side information about the target point. Here, we reduce the amount of background knowledge the adversary has about the target, and measure how this affects the reconstruction upper bound and attack success.

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 ρ values in $\hat{z}$ must be within an ℓ_∞ ball of radius δ 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 9.66×10^{-473} and 2.96×10^{-1850}, respectively.

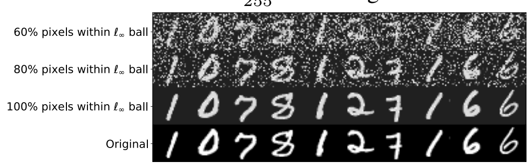
We plot the reconstruction upper bound compared to the attack success for different values of ρ in Figure 12. We also visualize the quality of reconstructions for different values of ρ. Even for $\rho = 0.6$, where 40% of the reconstruction pixels can take any value, and the remaining 60% are within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



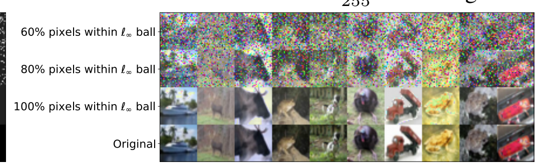
(a) Comparison of reconstruction success under a *very* small prior for MNIST, where we judge a reconstruction as successful if at least ρ pixels are within an absolute distance of $\frac{32}{255}$ of the target.



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



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



(d) CIFAR-10 examples of reconstructions where at least ρ 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 9.66×10^{-473} and 2.96×10^{-1850}, respectively.

J Estimating κ from samples

Here, we discuss how to estimate the base probability of reconstruction success, κ, 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 π – this can be computed using the methods sketched in Section 3. Then we have the following concentration bound.

Proposition 5. *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.

K 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 ϵ

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 ϵ required to bound its success. Most of these works share a common finding – the ϵ 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 $q\sqrt{T\log(\frac{1}{\delta})}/\varepsilon$ remain fixed, one can see that a larger value of ϵ 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 (ϵ, δ)-DP and RDP, because the privacy loss is not bounded; there is a small probability that the privacy loss will be high.

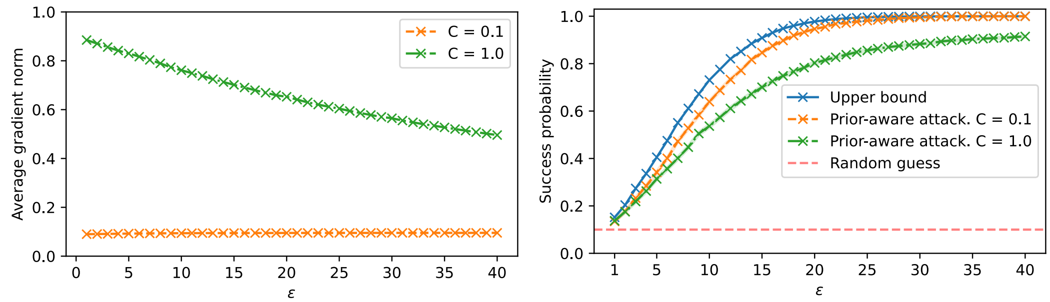
Stock et al. (2022) focus on bounding reconstruction for language tasks. They use the probability preservation guarantee from RDP to derive reconstruction bounds, showing that the length of a secret within a piece of text itself provides privacy. They translate this to show a smaller amount of DP noise 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 ith 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 ℓ, and an observed gradient (wrt to model parameters θ) $g_z = \nabla_\theta \ell(\theta, z)$, 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.



(a) Average gradient norm (over all samples and steps) for different values of ϵ at $C = 0.1$ and $C = 1$.

(b) Reconstruction success probability for different values of ϵ 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.

L 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 (the effect hyperparameters have on reconstruction at a fixed ϵ) across different values of ϵ and prior sizes, $|\pi|$.

L.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.

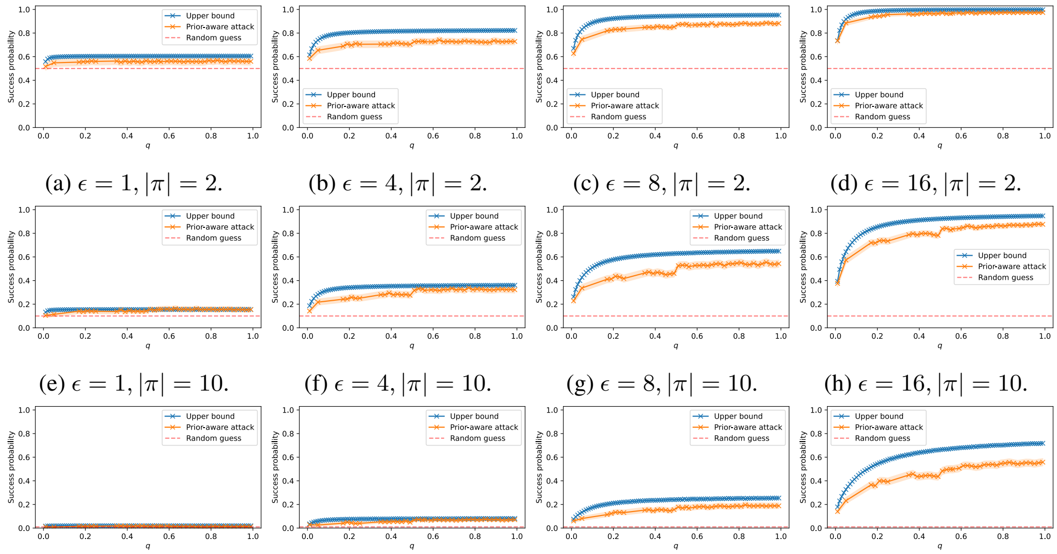
We demonstrate how this changes the reconstruction success probability by training a model for 100 steps with a clipping norm of 0.1 or 1, and measuring the average gradient norm of all samples over all steps. Results are shown in Figure 13. We see at $C = 0.1$, our attack is tight to the upper bound,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

L.2 More results on the effect of DP-SGD hyperparameters at a fixed ϵ

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



(a) $\epsilon = 1, |\pi| = 2$. (b) $\epsilon = 4, |\pi| = 2$. (c) $\epsilon = 8, |\pi| = 2$. (d) $\epsilon = 16, |\pi| = 2$.

(e) $\epsilon = 1, |\pi| = 10$. (f) $\epsilon = 4, |\pi| = 10$. (g) $\epsilon = 8, |\pi| = 10$. (h) $\epsilon = 16, |\pi| = 10$.

(i) $\epsilon = 1, |\pi| = 100$. (j) $\epsilon = 4, |\pi| = 100$. (k) $\epsilon = 8, |\pi| = 100$. (l) $\epsilon = 16, |\pi| = 100$.

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

larger values of ϵ. We visualise this in another way by plotting σ against q and report the upper bound in Figure 15. Note that the color ranges in Figure 15 are independent across subfigures.

M Proofs

Throughout the proofs we make some of our notation more succinct for convenience. For a probability distribution ω 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 ω and function ϕ 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 (μ, ν) 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 μ and ν admits a uniformly most powerful test. For example, this is the case for distributions on $\mathbb{R}^d$ where the density ratio μ/ν is a continuous function.

Theorem 6 (Formal version of Theorem 2)**.** *Fix π and ρ. Suppose that for every fixed dataset $D_{\text{-}}$ there exists a distribution $\mu_{D_{\text{-}}}$ such that $\sup_{z \in supp(\pi)} \mathcal{B}_\kappa(\mu_{D_z}, \nu_{D_{\text{-}}}) \leq \mathcal{B}_\kappa(\mu_{D_{\text{-}}}, \nu_{D_{\text{-}}})$ for all $\kappa \in [0, 1]$. If the pair (μ, ν) is testable, then M is (η, γ)-ReRo with*

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

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

Lemma 7. *For any μ and ν, the function $\kappa \mapsto \inf_{\nu(\phi) \leq \kappa}(1 - \mu(\phi))$ is convex in $[0, 1]$.*

Lemma 8. *For any testable pair (μ, ν), the function $\kappa \mapsto \sup_{\nu(E) \leq \kappa} \mu(E)$ is concave.*

Proof. By the testability assumption we have

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

Concavity now follows from Lemma 7. □

Proof of Theorem 6. Fix $D_{\text{-}}$ and let $\kappa = \kappa_{\pi,\rho}(\eta)$ throughout. Let also $\nu = \nu_{D_{\text{-}}}$, $\mu_z = \mu_{D_z}$, $\nu^* = \nu^*_{D_{\text{-}}}$ and $\mu^* = \mu^*_{D_{\text{-}}}$.

Expanding the probability of successful reconstruction, we get:

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

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

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

Finally, using Lemma 8 and Jensen's inequality on the following gives the result:

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

□

M.2 Proof of Corollary 3

Here we prove Corollary 3. 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 9 (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 10 (Theorem 6 in Mahloujifar et al. (2022))**.** *Let $\nu_{D_{\text{-}}}$, μ_{D_z} be the output distribution of DP-SGD applied to $D_{\text{-}}$ and D_z respectively, with noise multiplier σ, sampling rate q. Then we have*

$$TV_a(\nu_{D_{\text{-}}},\mu_{D_z}) \le TV_a(\nu,\mu)\ .$$

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

Lemma 11 (Lemma 21 in Zhu et al. (2022).). *For any pair of distributions ω_1, ω_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 $TV_a(\nu_{D_\cdot}, \mu_{D_z})$ is bounded by $TV_a(\nu,\mu)$ for all a, therefore we have

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

M.3 Proof of Proposition 5

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 $1 - e^{-N\tau^2\kappa/2}$:

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

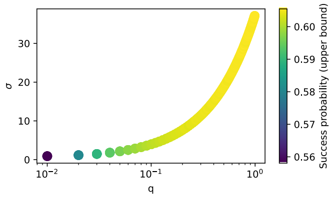
M.4 Proof of Proposition 4

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-Hoefding bound we have $\mathbb{P}[E_1] \leq 2e^{-2N\tau^2}$. Now note that since μ is a Gaussian mixture, we can write $\mu = \sum_{i\in[2^T]} a_i\mu_i$ where each μ_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

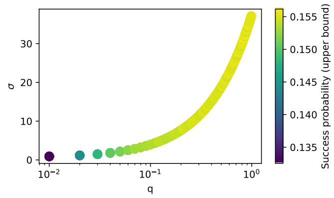
$$\sup_{\nu(E)\leq\kappa-\tau} \mu(E) - \tau \leq \lim_{N\to\infty} \gamma' \leq \sup_{\nu(E)\leq\kappa+\tau} \mu(E) + \tau.$$

Now by pushing τ to 0 and using the fact that μ and ν are smooth we have

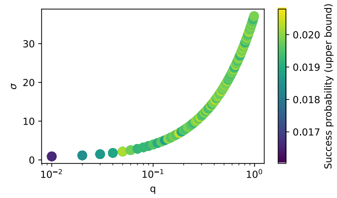
$$\lim_{N\to\infty} \gamma' = \lim_{\tau\to 0} \sup_{\nu(E)\leq\kappa+\tau} \mu(E) + \tau = \sup_{\nu(E)\leq\kappa} \mu(E).$$



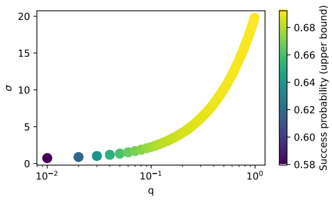
(a) $\epsilon = 1, |\pi| = 2$.



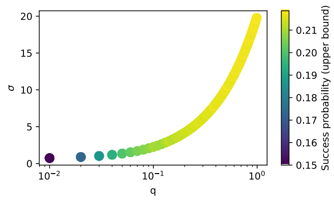
(b) $\epsilon = 1, |\pi| = 10$.



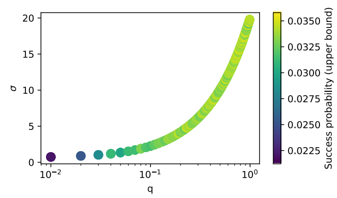
(c) $\epsilon = 1, |\pi| = 100$.



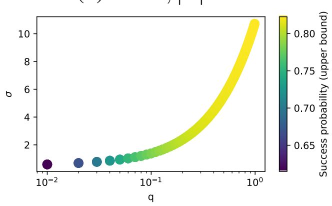
(d) $\epsilon = 2, |\pi| = 2$.



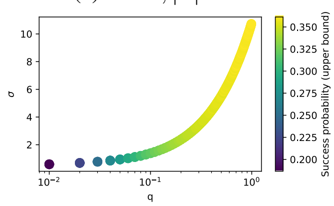
(e) $\epsilon = 2, |\pi| = 10$.



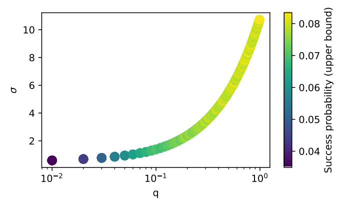
(f) $\epsilon = 2, |\pi| = 100$.



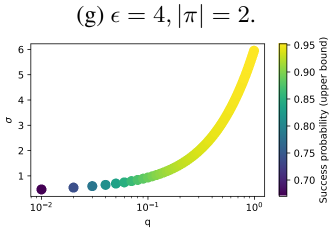
(g) $\epsilon = 4, |\pi| = 2$.



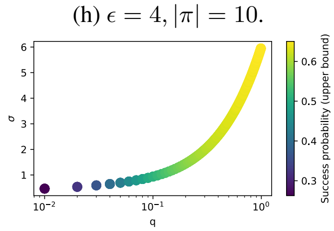
(h) $\epsilon = 4, |\pi| = 10$.



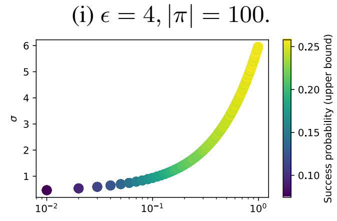
(i) $\epsilon = 4, |\pi| = 100$.



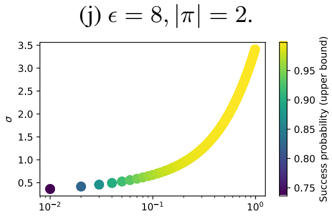
(j) $\epsilon = 8, |\pi| = 2$.



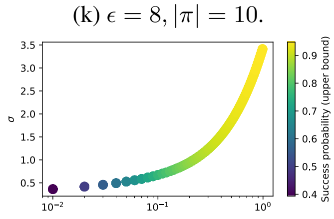
(k) $\epsilon = 8, |\pi| = 10$.



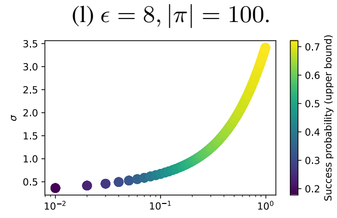
(l) $\epsilon = 8, |\pi| = 100$.



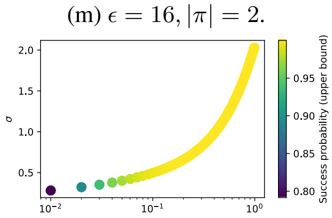
(m) $\epsilon = 16, |\pi| = 2$.



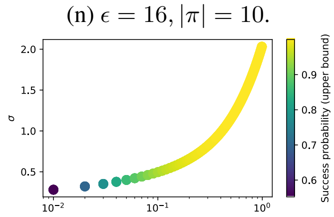
(n) $\epsilon = 16, |\pi| = 10$.



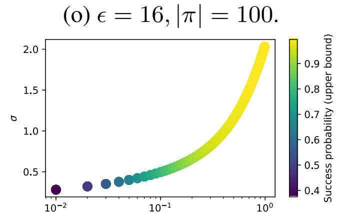
(o) $\epsilon = 16, |\pi| = 100$.



(p) $\epsilon = 32, |\pi| = 2$.



(q) $\epsilon = 32, |\pi| = 10$.



(r) $\epsilon = 32, |\pi| = 100$.

Figure 15: How the upper bound changes as a function of q and σ at a fixed value of ϵ 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.