FROM PRIVACY TO GENERALIZATION: LINEAR MAX-INFORMATION BOUNDS FOR DIFFERENTIALLY PRIVATE LEARNING ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the relationship between *generalization* and *privacy* remains a challenge in modern machine learning theory, particularly for deep networks that are trained by variants of differentially private stochastic gradient descent (DP-SGD). In this work we make progress on this persistent open problem. First, we derive explicit upper bounds on the approximate max-information of any algorithm that fulfills (ϵ, δ) differential privacy or Rényi differential privacy, thereby going beyond the classical results for pure ϵ differential privacy. Subsequently, we show even stronger guarantees for two common private learning algorithms, output perturbation with the Gaussian mechanism, and streaming DP-SGD, by exploiting the structure of their internal randomization. As an application of our results, we demonstrate how to obtain non-vacuous PAC-Bayes generalization bounds for deep networks, in which the prior distribution is learned by DP-SGD instead of the classical way of choosing it in a data-independent way.

1 INTRODUCTION

The question of how to formally guarantee that learning algorithms generalize from their training set to new data has been at the core of machine learning theory ever since the seminal works on PAC learning in the 1970s and 1980s (Vapnik & Chervonenkis, 1971; Valiant, 1984). In recent years, remarkable progress has been made to extend these generalization guarantees to over-parameterized models, using the PAC-Bayes framework (McAllester, 1999; Seeger, 2002; Catoni, 2007), resulting in non-vacuous bounds for deep neural networks (Dziugaite & Roy, 2017; Zhou et al., 2019; Lotfi et al., 2024). Phrased informally, one of the core insights is that models generalize if they do not *overfit*, i.e. they extract information from the training set but do not *memorize* it. Related considerations arise when learning from datasets that contain sensitive or personal data, e.g. when training *differentially private* (Dwork & Roth, 2014; Ponomareva et al., 2023) models. To guarantee that the original training data cannot be extracted from the trained model, one has to constrain the amount that each data point can influence the model parameters, i.e. preventing *data memorization*.

Given their conceptual similarity, it is not surprising that the intersection of generalization and privacy became a fruitful area for innovation. For example, in classic settings, Dwork et al. (2015); Bassily et al. (2016) established and Jung et al. (2020) refined generalization guarantees for statistical and low-sensitivity queries. Bassily et al. (2014) proved bounds for private risk minimization algorithms in convex settings, and Bombari & Mondelli (2025) and Shi et al. (2026) for private gradient descent for random feature models and two-layer ConvNets, respectively. Specifically in a PAC-Bayes setup, Dziugaite & Roy (2018) demonstrated that privately learned data-dependent distributions can take the place of the usually data-independent prior distributions with only a controllable additive penalty.

Unfortunately, all of the above results are inapplicable or ineffective for modern deep networks, which tend to be over-parameterized, non-convex and trained using stochastic gradient descent, so new theory is needed. One key mismatch between prior work and these networks is the lack of model- and problem-generic results for the privacy notions that these models are trained on, approximate (ϵ, δ)-DP (Dwork et al., 2006), and Rényi DP (Mironov, 2017), which are more amenable to analyzing the privacy guarantees of SGD than pure ϵ-DP. However, the connection between these notions of privacy and the generalization properties of the resulting models is so far not well understood.

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

In this work, we make progress on this question, through the notion of *approximate max-information* (Dwork et al., 2015), which is a statistical measure of how much information a trained model contains about its training set and therefore is highly informative about questions of *memorization* and *generalization*. **In Section 3, we prove general-purpose bounds on the approximate max-information for algorithms that fulfill either (ϵ, δ) approximate DP or Rényi DP.** To our knowledge, these are the first such guarantees that can readily be applied in *practical* private machine learning tasks, in the sense that they hold in a regime where real-world deep learning models are trained, let it be in industry, such as Meta's recent private vision-language model *DP-Cap* (Sander et al., 2024), or Google's large language model *VaultGemma* (Sinha et al., 2025); or academia, such as recent models for private human action recognition (Luo et al., 2024; Nken et al., 2025).

Afterwards, we focus on two specific algorithms, that are particularly popular for machine learning tasks: *output perturbation* using the Gaussian mechanism Dwork et al. (2006), and (streaming) *DP-SGD* (Rajkumar & Agarwal, 2012). **In Section 4, we show that by exploiting the explicit form of randomization invoked by these algorithms, tighter guarantees on the max-information are possible than what would follow just from their privacy properties.**

Finally, in Section 5, we use our results to instantiate PAC-Bayes bounds (McAllester, 1999), with data-dependent prior distributions learned using DP-SGD, following Dziugaite & Roy (2018); Rivasplata et al. (2020), and we demonstrate that the resulting additive penalty terms can be controlled explicitly from the training hyperparameters. Specifically, **we give non-vacuous generalization bounds for deep networks with up to 13 million trainable parameters**, which is a much larger than what was possible with data-independent priors.

2 Background

In this section, we introduce the main concepts and notation. For more details, see Appendix A.

We adopt a standard setup of statistical learning, in which a learning algorithm, $\mathcal{A}$, is given an i.i.d. training set, $S \subset \mathcal{X}^n$, and outputs a model $y \in \mathcal{Y} \subseteq \mathbb{R}^d$ (we identify models and their parametrizations). We generally allow for algorithms to be *stochastic*, i.e. we treat $\mathcal{A}(S)$ as a random variable that takes values in $\mathcal{Y}$, according to the distribution $\mathrm{Law}(\mathcal{A}(S))$.

Differential Privacy. A stochastic algorithm, $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$, is called *approximate differentially private*, (ϵ, δ)-DP, for some $\epsilon > 0$ and $\delta \in [0, 1]$, if

$$\forall O \subset \mathcal{Y} : \quad \mathbb{P}\{\mathcal{A}(S) \in O\} \leq e^{\epsilon}\mathbb{P}\{\mathcal{A}(S') \in O\} + \delta, \tag{1}$$

for all datasets S, S' that are *neighboring*, i.e. identical except for a single data point[1] (Dwork et al., 2006) If an algorithm is (ϵ, δ)-DP for $\delta = 0$, we call it *purely differentially private*, ϵ-DP.

A popular alternative notion of privacy in machine learning contexts is Rényi DP (RDP) (Mironov, 2017). For any $\alpha > 1$ and $\epsilon \geq 0$, the algorithm $\mathcal{A}$ is called (α, ϵ)-*RDP*, if $D_\alpha(\mathcal{A}(S)\|\mathcal{A}(S')) \leq \epsilon$ for any two neighboring datasets S and S', where $D_\alpha(P\|Q) := \frac{1}{\alpha-1}\log \mathbb{E}_{x\sim Q}\big(\frac{P(x)}{Q(x)}\big)^\alpha$ is the Rényi divergence of order α. Note that the relation between both notions of privacy is asymmetric: (α, ϵ)-RDP implies $(\epsilon + \frac{\log(1/\delta)}{\alpha-1})$-DP for any $\delta \in (0, 1)$, but from a single (ϵ, δ)-DP guarantee, no guarantees on RDP follow (Mironov, 2017).

Approximate Max-Information. A statistical quantity, related to privacy, but specifically tailored to studying problems of generalization is the *(approximate) max-information* (Dwork et al., 2015), which measures how much statistical information an algorithm's output, $\mathcal{A}(S) \in \mathcal{Y}$, contains about its input data, $S \in \mathcal{X}^n$. Formally, for any $\beta \in [0, 1)$, we denote by

$$D_\infty^\beta(V\|W) := \log \sup_{O\subseteq\Omega:\Pr\{V\in O\}>\beta} \frac{\mathbb{P}\{V \in O\} - \beta}{\Pr\{W \in O\}} \tag{2}$$

the *approximate max-divergence* of two random variables V and W on a common domain Ω, and we define the approximate max-information

$$I_\infty^\beta\big(\mathcal{A}(S), S\big) := D_\infty^\beta\big((\mathcal{A}(S), S)\|\mathcal{A}(S) \times S\big), \tag{3}$$

[1] In this work, we use the *replace one* neighborhood relation. Adapting our results to other relations, such as *add/remove* would be straightforward and only result in slightly different constants.

where $\mathcal{A}(S) \times S$ denotes the random variable distributed according to $\text{Law}(\mathcal{A}(S)) \times \text{Law}(S)$, i.e. the product of independent marginal. Note that we work with *natural* logarithms, unlike the original definition in Dwork et al. (2015), and so avoid any $\log e$ constants that appear in the original works.

The max-information can act as a tool for studying, e.g., *memorization* or *unlearning* in a principled way (Gupta et al., 2021). Similarly, knowing the max-information of a learning algorithm allows the derivation of bounds on its generalization abilities (Dziugaite & Roy, 2018; Rivasplata et al., 2020). A link between max-information and privacy was established in Dwork et al. (2015, Theorem 8): for any ϵ differentially private algorithm $\mathcal{A}$ that is trained on an i.i.d. dataset of size n, it holds that

$$I_\infty^\beta(\mathcal{A}(S), S) \leq \frac{1}{2} n\epsilon^2 + \epsilon\sqrt{\frac{n}{2}\log\frac{2}{\beta}}. \tag{4}$$

Unfortunately, ϵ-DP is a very strict condition that most practical machine learning algorithms do not fulfill. Therefore, one of the central open questions in the field is how (4) can be extended to more practical relaxed notions of differential privacy. Rogers et al. (2016, Theorem 3.1) discussed a generalization of (4) to (ϵ, δ)-DP algorithms. However their focus was on establishing asymptotic guarantees and holds only for certain combinations of ϵ, δ and β, so the result is not applicable for practical tasks, where constants matter and one would like to choose these parameters freely.

3 From Privacy to Approximate Max-Information

We now present our first set of results: explicit upper bounds on the approximate max-information of differentially private learning algorithms. Specifically, in Theorem 1 we present a new result for (ϵ, δ)-DP that at the same time substantially tightens the classic result (4) for pure ϵ-DP (i.e. $\delta = 0$). In Theorem 2, we establish guarantees on the max-information of Rényi-DP algorithms. Throughout this section, we assume that $\mathcal{A} : \mathcal{X}^n \to \mathbb{R}^d$ is an arbitrary stochastic algorithm, operating on an i.i.d. dataset, S, of size n. Additional assumptions, in particular about $\mathcal{A}$'s privacy properties, will be stated in the individual theorems. The proofs are deferred to Appendix B.

Theorem 1. *Assume that $\mathcal{A}$ is (ϵ, δ)-DP for some $\epsilon \geq 0$ and $\delta \in [0, 1]$.*

1) If $\beta \in [n\delta, 1)$, it holds with $\beta_0 := \beta - n\delta \geq 0$:

$$I_\infty^\beta(\mathcal{A}(S), S) \leq \min\Big\{ n\epsilon,\ \frac{1}{8} n\epsilon^2 + \epsilon\sqrt{\frac{n}{2}\log\frac{1}{\beta_0}} + \frac{n\delta}{\beta_0} c(\epsilon) \Big\}, \tag{5}$$

where $c(\epsilon) = e^{2\epsilon} + \epsilon - 1 + e^{\sqrt{\frac{32\log(1/\beta_0)}{n}}}$ and the second term in the $\min$ only occurs if $\beta_0 > 0$.

2) If $\beta \in (0, n\delta)$, no finite upper bound on $I_\infty^\beta(\mathcal{A}(S), S)$ in terms of n, ϵ, δ and β is possible.

Discussion. Theorem 1 generalizes and improves over the classic result (4). For $\delta = 0$, i.e. pure ϵ-DP, statement 1) offers an improvement by a factor of 4 in the first term, which is often the dominant one. While not a change in rate—which would not be possible, because $O(n\epsilon^2)$ is known to be optimal (Rogers et al., 2016)—it nevertheless can provide a substantial improvement when numerical values of I_∞^β are required, such as for PAC-Bayes generalization guarantees (Dziugaite & Roy, 2017). When $\delta = 0$, it is possible to choose $\beta = 0$, in which case we recover the known upper bound $n\epsilon$ on the *pure* max-information from Dwork et al. (2015, Theorem 7).

For $\delta > 0$, Eq. (5) goes beyond Rogers et al. (2016, Theorem 3.1), in that it provides expressions with explicit constants, that—in particular— converge to the pure-DP result for $\delta \to 0$. In contrast to the earlier work, our bound also does not contain terms of the order $\frac{\delta}{\epsilon}$ and remains valid for $\epsilon \to 0$ without requiring $\delta \to 0$ at the same time.

Statement 2) makes the observation precise that β cannot be too small if the approximate max-information is meant to be controllable, as seen in (Rogers et al., 2016). Specifically, $\beta \geq n\delta$ is necessary to achieve finite bounds on $I_\infty^\beta(\mathcal{A}(S), S)$, and $\beta > n\delta$ for non-trivial ones.

The following result provides an analogous guarantee for Rényi DP algorithms.

Theorem 2. *Assume $\mathcal{A}$ is (α, ϵ)-RDP for some $\alpha > 1$ and $\epsilon > 0$. Then, it holds for any $\beta \in (0, 1)$:*

$$I_\infty^\beta(\mathcal{A}(S), S) \leq \min\Big\{ n\Big(\epsilon + \frac{\log(1/\beta)}{\alpha - 1}\Big),\ v + 2\sqrt{v\log(2/\beta)} + \frac{\log(2/\beta)}{\alpha - 1}, \Big\} \tag{6}$$

with $v = \frac{n}{8}\big(\epsilon + \frac{\log(8/\beta)}{\alpha-1}\big)^2$.

Discussion. Like the bound in Theorem 1, the right hand side of (6) has two parts, which apply in different regimes. The central quantity is the *effective privacy*, $\tilde{\epsilon} := \epsilon + \frac{\log(c/\beta)}{\alpha-1}$, which resembles the conversion rule from RDP to (ϵ, δ)-DP and appears in both parts, only with different constants c. The first part of the bound is simply linear, $n\tilde{\epsilon}$, thereby providing an analog to the first term in (5). The second part of the bound is of the order $O(n\tilde{\epsilon}^2 + \sqrt{n}\tilde{\epsilon})$ and can therefore be much smaller than the first in the high-privacy regime, where n is large and $\tilde{\epsilon} \ll 1$. For large $\tilde{\epsilon}$, whether due to weak privacy (large ϵ) or a very small value of β, the quadratic second term grows quickly, and the first term provides a tighter bound. In contrast to Theorem 1, there is no restriction on the valid range of $\beta > 0$. This is due to the fact that any single (α, ϵ)-RDP guarantee actually ensures (ϵ, δ)-DP guarantees for all values of δ, only with adjusted ϵ, including sufficiently small ones.

An alternative path to prove a bound on the max-information of RDP algorithms would be to rely on the conversion to (ϵ, δ)-DP and evaluate Theorem 1. Typically, this would lead to worse guarantees, though, because of the large last term in (5) that has no equivalent in (6).

4 Approximate Max-Information of Private Learning Algorithms

So far we presented results hold for all algorithms that fulfill certain privacy definitions. Next, we study two specific algorithms for differentially private learning: *output perturbation* (Chaudhuri et al., 2011) using the Gaussian mechanism, and (streaming) *DP-SGD* (Abadi et al., 2016). As our results will show, exploiting the specific structure of randomization inside the algorithms allows us to establish even tighter guarantees, especially in the high-privacy regime that is of most interest in the context of generalization. The proofs of all results can again be found in Appendix B.

4.1 Approximate Max-Information of the Gaussian Mechanism

The *Gaussian mechanism* (Dwork et al., 2006) has central importance in differentially private machine learning. For any function, $\Psi : \mathcal{X}^n \to \mathbb{R}^d$, denote by $\Delta(\Psi) := \sup_{S \sim S'} \|\Psi(S) - \Psi(S')\|_2$ its *sensitivity* for neighboring datasets S, S'. The Gaussian mechanism with noise multiplier σ is defined as $\mathcal{M}_\Psi(S) = \Psi(S) + \sigma\Delta(\Psi)Z$, where $Z \sim \mathcal{N}(0, \mathrm{I}_d)$ is standard Gaussian noise, sampled independently from all other quantities. The privacy properties of the Gaussian mechanism are well studied: $\mathcal{M}_\Psi$ is not ϵ-DP for any ϵ, but for any $\epsilon, \delta \in (0, 1)$ it is (ϵ, δ)-DP as long as $\sigma \geq \sqrt{2\log(1.25/\delta)}/\epsilon$. A tighter analytic characterization that, in particular, also holds for $\epsilon > 1$, can be found in Balle & Wang (2018). Furthermore, $\mathcal{M}_\Psi$ is $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP for every $\alpha > 1$ (Mironov, 2017, Corollary 3).

Because of its simplicity and well-understood properties for high-dimensional problems, the Gaussian mechanism is a common choice for privatizing models after they have already been trained, as long as suitably tight bounds on the learning algorithms' sensitivity are known. The latter is the case, for example, for *regularized logistic regression* (Chaudhuri et al., 2011), or *principal component analysis* (Chaudhuri et al., 2012). The following theorem provides a bound on the approximate max-information of the Gaussian mechanism.

Theorem 3. *Let $\Psi : \mathcal{X}^n \to \mathbb{R}^d$ be a function with sensitivity at most $\zeta > 0$, and let $\mathcal{M}(S) = \Psi(S) + \sigma\zeta Z$ for $Z \sim \mathcal{N}(0, I)$ be the* Gaussian mechanism *with noise multiplier $\sigma > 0$ applied to Ψ. Let $\nu := \frac{n}{4\sigma^2}$, and for any $\beta \in (0, 1)$ write $L := \log(1/\beta)$. If the elements of S are independent, then the* approximate max-information of $\mathcal{M}$ *fulfills,*

$$I_\infty^\beta(\mathcal{M}(S), S) \leq \nu + \inf_{\lambda \in (0, r_\nu)} \frac{1}{\lambda}\Big[L - \log\big(1 - (\lambda + \lambda^2)\nu\big)\Big] \tag{7}$$

with $r_\nu = \sqrt{\frac{1}{4} + \frac{1}{\nu}} - \frac{1}{2}$. In more explicit but less tight forms:

$$\leq 2\nu + \big(\tfrac{1}{2}\nu + \sqrt{\tfrac{1}{4}\nu^2 + \nu}\big)L + \sqrt{2(\nu^2 + 2\nu)L}, \tag{8}$$

and relaxed further, sorted by powers of ν:

$$\leq \big(2 + \sqrt{2L} + L\big)\nu + \big(L + 2\sqrt{L}\big)\sqrt{\nu}. \tag{9}$$

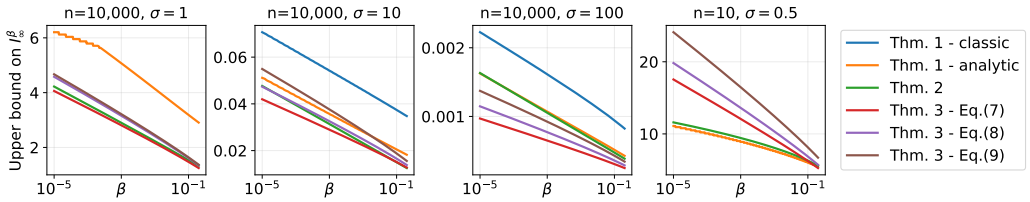

Figure 1: Numeric upper bounds on the approximate max-information of the *Gaussian mechanism* (normalized by $1/n$), as obtained from Theorem 1, Theorem 2, and Theorem 3 for different values of n and σ. In the left three panels, the explicitly derived bound of Theorem 3 is always tightest, and the analytic expressions are usually close in value. Of other bounds, the bound based on Rényi-DP is competitive in case of *low privacy* ($\sigma = 1$), but less so in the *high privacy* regime ($\sigma = 100$). The bound based on (ϵ, δ)-DP is competitive with the RDP-bound for high privacy settings, at least when evaluated by the analytic relation between σ and (ϵ, δ) of (Balle & Wang, 2018). Inferring (ϵ, δ) from the classic formula generally leads to worse results, and is not applicable in the low privacy setting due to its restriction to $\epsilon \in (0, 1)$. The fourth panel shows an case of very low privacy. Here, the (ϵ, δ)-DP and the Rényi-DP bounds are preferable over the explicitly derived ones.

Discussion. First, we observe that the upper bounds (7)–(9) are computable explicitly from known quantities: the noise multiplier σ, the dataset size n, and the desired value of β. Bound (7) requires a one-dimensional optimization over a finite interval for this, which can easily be performed numerically, especially given that it is one-dimensional and strictly convex in λ (Lemma B.5). For theoretical studies, eqs. (8) and (9) provide analytical relaxations, in terms of powers of L, and of ν respectively. In particular, the latter allows insights into the scaling behavior for practical settings: as $\nu = \frac{n}{4\sigma^2}$, eq. (9) is of order $O(n/\sigma^2 + \sqrt{n/\sigma^2})$. This resembles the bounds of the previous section, thinking of σ as inversely proportional to the privacy parameter. Evaluated numerically, however, the specific bounds of Theorem 3 are typically tighter, see our results in Figure 1.

4.2 Approximate Max-Information of Streaming DP-SGD

In *deep learning*, tight bounds on the overall sensitivity of the learning process are typically not known, so the method of choice is not output perturbation, but some variant of *differentially private stochastic gradient descent (DP-SGD)* (Abadi et al., 2016). DP-SGD performs iterative updates of the model parameters, where in each step a bound on the sensitivity is enforced by clipping the length of the contributing gradients to a maximum length. Subsequently, the Gaussian mechanism is employed to privatize the update vector.

Algorithm 1 DP−SGD

input input data batches $B_1, \dots, B_T$, clipping threshold ζ, noise multiplier σ
require subroutine ModelUpdate

$\theta_0 \leftarrow$ initialize model parameters
for $t = 1, \dots, T$ **do**
  $z \leftarrow$ sample from $\mathcal{N}(0, \mathrm{I}_d)$
  $u_t \leftarrow \sum_{x \in B_t} \mathrm{clip}\big(\nabla \ell(x, \theta_{t-1}), \zeta\big) + \sigma\zeta z$ with $\mathrm{clip}(v, \zeta) := \min\{1, \zeta/\|v\|\}v$.
  $\theta_t \leftarrow \mathrm{ModelUpdate}(u_t; \theta_0, \dots, \theta_{t-1})$
  yield θ_t
end for

Algorithm 1 provides pseudo-code, relying on a subroutine ModelUpdate for combining the current (privatized) update vector, u_t with the previous parameters $\theta_1, \dots, \theta_{t-1}$ to produce a new update θ_t. In classic SGD that is $\theta_t \leftarrow \theta_{t-1} - \eta u_t$ for a learning rate η, but our results are not restricted to this choice. Any update rule that depends on the dataset only through the private update vectors can be used, e.g., SGD with momentum and/or weight decay, or Adam (Tang et al., 2024; Hudişteanu et al., 2025), see Appendix A. For simplicity, we refer to the algorithm as DP-SGD.

The following theorem provides a bound on the approximate max-information of DP-SGD in a setting with fixed-size, non-overlapping, batches, as is common in decentralized/federated learning or learning from data streams.

Theorem 4. *Let $\mathcal{A}$ be the* DP-SGD algorithm 1 *with hyperparameters ζ (clipping threshold) and σ (noise multiplier), that is run on T data batches of size m, which contain independent data samples.*

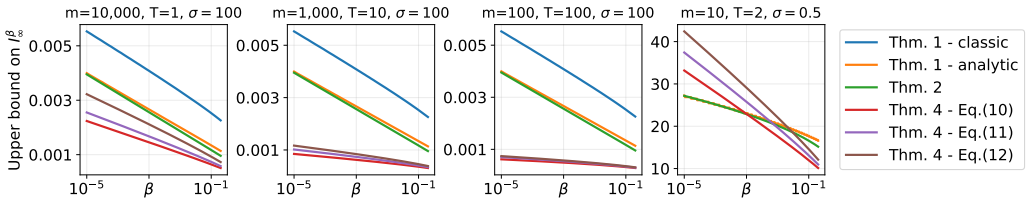

Figure 2: Numeric upper bounds on the approximate max-information of *streaming DP-SGD* (normalized by $1/n$), as obtained from Theorem 1, Theorem 2, and Theorem 4 for different values of m, T, and σ. In the left three panels, the three explicit bounds of Theorem 4 are always the tightest. Increasing the number of update steps at proportionally reduced batchsize improves the bounds of Theorem 4, while the bounds based purely on privacy properties do not benefit. Of the other bounds, the bounds based on Rényi-DP and (ϵ, δ) derived by the analytic rule are comparable and better than using classic conversion rule. The fourth panel shows the limit case of only two update steps with low privacy. Here, the Rényi bound and the (ϵ, δ) bound become preferable for small values of β.

For any $\beta \in (0, 1)$, we write $L = \log(1/\beta)$. Then, the approximate max-information of $\mathcal{A}$ *fulfills*

$$I_\infty^\beta(\mathcal{A}(S), S) \leq 2T\nu + \inf_{\lambda \in (0, r_{4\nu})} \frac{1}{\lambda}\Big[L - \frac{T}{2}\log\big(1 - 4(\lambda + \lambda^2)\nu\big)\Big], \tag{10}$$

where $\nu = \frac{m}{4\sigma^2}$, and $r_{4\nu} = \frac{1}{2}\sqrt{1 + \frac{1}{\nu}} - \frac{1}{2}$. In less tight, but more explicit, form:

$$\leq 4T\nu + 2\big(\nu + \sqrt{\nu^2 + \nu}\big)L + 2\sqrt{(4\nu^2 + 2\nu)TL}. \tag{11}$$

Relaxed further and sorted by powers of ν, it holds that:

$$\leq 4(T + L + \sqrt{TL})\nu + 2(L + \sqrt{2TL})\sqrt{\nu}. \tag{12}$$

Discussion. As in the case of the Gaussian mechanism, the bounds of Theorem 4 can easily be evaluated numerically or analytically, because they depend only on explicitly known quantities, and the optimization problem in (10) is one-dimensional and, in fact, strictly convex. From the explicit expressions one can see that in the typical regime of $T \geq L$ and $m \geq \sigma^2$, the dominant term is of order $\frac{Tm}{\sigma^2}$, which is comparable to the earlier results, as Tm is the total number of data samples used.

For $T = 1$ we can compare directly to Theorem 3 with a half as large value of σ (because a clipping constant ζ implies sensitivity 2ζ), observing that Theorem 4 provides better constants. This is a consequence of the fact that we know DP-SGD's feature function is a sum over individual examples rather than an arbitrary function of sensitivity 2ζ, so we can exploit stronger concentration results.

Another observation is that all three bounds are decreasing under a transformation $(T, m) \mapsto (\alpha T, m/\alpha)$ for any $\alpha > 1$, i.e. the tightest bound could be achieved by running with Tm batches of size 1. In practice, however, this would not be advisable, as DP-SGD adds noise after each update step, so the algorithm's utility would suffer.

5 Tight PAC-Bayes Generalization Bounds from DP-SGD Priors

The bounds on the approximate max-information of the previous sections provide qualitative and quantitative insights into the memorization properties of private learning algorithms. They can also be used to construct tight numerical bounds on the generalization abilities of trained models, regardless of whether they were trained with privacy or not, which we demonstrate using the PAC-Bayes framework.

5.1 Background

PAC-Bayes learning theory allows establishing formal guarantees on the *generalization* abilities of a learning method by providing high-probability upper bounds on the difference between a model's empirical risk (training loss) and its true risk (test loss) (McAllester, 1999; Maurer, 2004).

Specifically, let $\mathcal{H}$ be a model class, and let $\ell : \mathcal{X} \times \mathcal{H} \to [0, 1]$ be a loss function that measures the quality of a model, h, on a data point x. For an i.i.d. training dataset, S, of size n, we denote by

$\widehat{R}(h) := \frac{1}{n}\sum_{x\in S}\ell(x,h)$ the *empirical risk* of h, and by $R(h) := \mathbb{E}\ell(x,h)$ the *true risk*. Then, for any *data-independent* prior, π over $\mathcal{H}$ one can prove that for all $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that, uniformly over all *posterior* distributions ρ over $\mathcal{H}$:

$$\mathrm{kl}(\widehat{R}(\rho)\,\|\,R(\rho)) \leq \frac{\mathrm{KL}(\rho\|\pi) + \log\frac{2\sqrt{n}}{\delta}}{n}, \tag{13}$$

where $\widehat{R}(\rho) = \mathbb{E}_{h\sim\rho}[\widehat{R}(h)]$ and $R(\rho) = \mathbb{E}_{h\sim\rho}[R(h)]$, and $\mathrm{kl}(\hat{p}\|p) = \mathrm{KL}(\mathrm{Ber}(\hat{p})\|\mathrm{Ber}(p)) = \hat{p}\log\frac{\hat{p}}{p} + (1-\hat{p})\log\frac{1-\hat{p}}{1-p}$[2]. Unfortunately, it is often very difficult to find data-agnostic priors π that allow for posteriors ρ with small $\widehat{R}(\rho)$ and $\mathrm{KL}(\rho\|\pi)$ at the same time. Consequently, the resulting generalization guarantees tend to be vacuous in practice, i.e. numerically they do not improve over the trivial bound $R(\rho) \leq 1$.

A way to achieve tighter bounds, specifically to find priors that enable smaller values of $\mathrm{KL}(\rho\|\pi)$, is to extend (13) to allow for *distribution-dependent* or *data-dependent priors* (Catoni, 2007; Lever et al., 2013; Parrado-Hernández et al., 2012). Specifically, Dziugaite & Roy (2018) demonstrated that the prior distributions can be *learned* on the training data itself, if one uses an ϵ-DP algorithm. Unfortunately, as mentioned above, ϵ-DP is a quite restrictive criterion compared to (ϵ,δ)-DP with $\delta > 0$, and so ϵ-DP algorithms are not commonly used for machine learning tasks.

Rivasplata et al. (2020) rephrased the above result to cover all algorithms with small approximate max-information, however, without providing examples of such algorithms besides ϵ-DP ones. The result from Rogers et al. (2016) for (ϵ,δ)-DP algorithms does not suffice in this context, because it lacks the explicit constants required for numeric evaluation. Consequently, the bounds we proved in Sections 3 and 4 are the first candidates for such a procedure.

5.2 Non-vacuous Generalization Bounds for Deep Networks with DP-SGD Priors

We now demonstrate how the max-information framework allows the construction of non-vacuous generalization bounds in a deep learning setting with a completely standard parametrization. Besides our results from Sections 3 and 4, we rely on a variant of (Dziugaite & Roy, 2018, Theorem 4.2) and (Rivasplata et al., 2020, Lemma 7), which we state here for reference in a form compatible with the rest of our notation.

Theorem 5. *For a dataset $S = (X_1,\ldots,X_n)$ of i.i.d. data points, let π be a prior distribution over models, obtained from the output of running an algorithm $\mathcal{A}$ on S. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sampling of S and randomness of $\mathcal{A}$, that uniformly for all posterior distributions ρ, we have:*

$$\mathrm{kl}\big(\widehat{R}(\rho)\|R(\rho)\big) \leq \frac{\mathrm{KL}(\rho\|\pi) + I_\infty^{\delta/2}(\mathcal{A}(S),S) + \log(\frac{4\sqrt{n}}{\delta})}{n}. \tag{14}$$

As an immediate consequence of Theorem 5 (by setting $\rho = \pi$), we obtain a simpler generalization bound that is particularly useful for privately trained models:

Corollary 6. *For a dataset $S = (X_1,\ldots,X_n)$ of i.i.d. data points, let ρ be a distribution over models, obtained from the output of running an algorithm $\mathcal{A}$ on S. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sampling of S and randomness of $\mathcal{A}$, we have:*

$$\mathrm{kl}\big(\widehat{R}(\rho)\|R(\rho)\big) \leq \frac{I_\infty^{\delta/2}(\mathcal{A}(S),S) + \log(\frac{4\sqrt{n}}{\delta})}{n}. \tag{15}$$

Discussion. The generalization bounds hold for any mechanism of deriving distributions over the model set from the algorithms' output. The most common choice is to use isotropic Gaussian distributions centered at the learned parameter vectors, which makes the KL term appearing in the bounds explicitly computable. For other possibilities, see Appendix A. In the subsequent section we show that the resulting bounds can be numerically non-vacuous even for deep networks in their standard weight-matrix parametrization, i.e. without having to rely on discretized, low-rank or compressed parametrizations that several earlier works employed (Arora et al., 2018; Lotfi et al., 2022; Zakerinia et al., 2025) or reserving a subset of the training data just for learning a prior.

[2]More intuitive notions of similarity between $\widehat{R}(\rho)$ and $R(\rho)$, e.g. on their difference, can be derived from (13) using standard relaxations of the KL divergence. See, e.g., Rivasplata et al. (2020) for more details.

Table 1: Numeric upper bounds, $\mathcal{B}$, on the risk, R, for overparameterized deep networks on MNIST, obtained by numerically inverting (15) for the DP models h_0, and (14) for ρ. Reported results and standard deviations are from three runs with different random seeds.

Method	$\widehat{R}(h_0)$	$\widehat{R}(\rho)$	κ	Complexity/n	$\mathcal{B}(h_0)$	$\mathcal{B}(\rho)$
data-indep.	–	$0.104_{\pm 0.003}$	–	$0.136_{\pm 0.003}$	–	$0.325_{\pm 0.006}$
DP-SGD (streaming)	$0.238_{\pm 0.019}$	$0.163_{\pm 0.002}$	1381	$0.034_{\pm 0.001}$	$0.337_{\pm 0.021}$	$0.273_{\pm 0.004}$
DP-SGD (Poisson subs.)	$0.164_{\pm 0.001}$	$0.144_{\pm 0.001}$	1224	$0.025_{\pm 0.001}$	$0.248_{\pm 0.001}$	$0.233_{\pm 0.001}$

5.3 Numerical Experiments

In this section, we illustrate the behavior of our results in Sections 3 and 4 when used in combination with Theorem 5 or Corollary 6. We use the setup of (García-Pérez et al., 2026; Pérez-Ortiz et al., 2021), i.e. optimizing a PAC-Bayes bound over a family of Gaussian posteriors given a Gaussian prior. For details of the experimental setup, see Appendix C.

For MNIST, we use an MLP model with $1.2M$ parameters, and compare three different choices of prior: 1) a data-independent Gaussian prior as in (García-Pérez et al., 2026). 2) a Gaussian prior whose mean is learned by streaming DP-SGD and its max-information is computed based on Theorem 4. 3) a Gaussian prior which its mean is learned by standard multi-epoch DP-SGD with Poisson subsampling, which has better privacy guarantees due to amplification (Zhu & Wang, 2019), and its max-information is computed based on Theorem 2.

Table 1 shows that both private priors achieve tighter bounds than the data-independent prior. Additionally, the comparison between the prior learned from streaming DP-SGD and from DP-SGD with Poisson subsampling highlights the importance of our general results, especially Theorem 2: Even though in the streaming setting, Theorem 4 is tighter than the bound from Theorem 2, when we have full access to the dataset, Poisson subsampling can achieve stronger privacy guarantees, resulting in smaller max-information bounds. Moreover, it allows training for multiple epochs, which is often necessary to obtain good models.

For CIFAR-10, we use three CNN models with approximately $10K$, $5.9M$, and $13.2M$ parameters, and compare the approach with a data-independent Gaussian prior with a prior learned by DP-SGD with Poisson subsampling as described above.

To achieve non-vacuous bounds for large networks, existing approaches either use a compact parameterization to reduce the number of parameters and compute a bound for the compressed network, or use data-splitting, i.e. use part of the samples to train a prior and use the remaining to optimize a PAC-Bayes bound. Although data splitting yields valid bounds, the process that produced the prior is not analyzed, so the bound does not explain why the prior itself does not overfit its training data (Lotfi et al., 2022; García-Pérez et al., 2026). Additionally, they posteriors remain close to their learned priors in deeper architectures (Pérez-Ortiz et al., 2021), so the certificate effectively reduces to a held-out evaluation of the prior (Lotfi et al., 2022). Our goal is instead to certify the full learning process: we learn the prior from the same samples that enter the bound and account for this dependence through its max-information, without any held-out data. The closest work to ours is Dziugaite & Roy (2018), which applies the max-information bound for ϵ-DP to sampling from the Gibbs posterior to learn the prior for MNIST. However, in practice they approximate the Gibbs posterior by SGLD, which the guarantee does not formally cover. To our knowledge, ours is the first approach that simultaneously applies to models with up to $13M$ trainable parameters, in their original parameter space, avoids using data splitting, and computes max-information terms for the algorithm as executed.

In Table 2 we provide the guarantees for the model output of DP-SGD, where the bounds for all parameter sizes are non-vacuous. To the best of our knowledge, these are the first non-vacuous bounds for models of this size without compression or data splitting, for an algorithm used in practice (here, DP-SGD with Poisson subsampling). Additionally, in Table 3 we show that further optimizing the posterior can improve the bounds, and similarly to MNIST, achieve better results compared to using data-independent priors, which cannot provide non-vacuous bounds or bigger networks.

Table 2: Empirical risk and numeric upper bounds for DP models, h_0, on CIFAR-10, with max-information κ. Reported results and standard deviations are over three runs with different random seeds.

# Layers	# Parameters	κ	$\widehat{R}(h_0)$	$\mathcal{B}(h_0)$
4	9,946	757	$0.705_{\pm 0.021}$	$0.780_{\pm 0.019}$
9	5,851,338	342	$0.748_{\pm 0.019}$	$0.797_{\pm 0.017}$
15	13,193,930	757	$0.782_{\pm 0.011}$	$0.848_{\pm 0.009}$

Table 3: Empirical risk $\widehat{R}(\rho)$, max-information κ, total complexity over n , and numeric upper bounds $\mathcal{B}$, for three ConvNet models of different sizes trained on CIFAR-10. Reported results and standard deviations are over three runs with different random seeds.

# Layers	# Parameters	Method	$\widehat{R}(\rho)$	κ	Complexity$/n$	$\mathcal{B}(\rho)$
4	9,946	data-indep.	$0.591_{\pm 0.012}$	–	$0.051_{\pm 0.004}$	$0.738_{\pm 0.007}$
		DP-SGD	$0.534_{\pm 0.011}$	757	$0.074_{\pm 0.004}$	$0.716_{\pm 0.007}$
9	5,851,338	data-indep.	–	–	–	vacuous
		DP-SGD	$0.629_{\pm 0.002}$	342	$0.061_{\pm 0.004}$	$0.782_{\pm 0.003}$
15	13,193,930	data-indep.	–	–	–	vacuous
		DP-SGD	$0.648_{\pm 0.007}$	757	$0.029_{\pm 0.003}$	$0.757_{\pm 0.012}$

6 Conclusion

In this work, we established the first explicit upper bounds on the β-approximate max-information of differentially private learning algorithms in regimes relevant to modern deep learning. Our results provide a principled connection between differentially private deep learning and PAC-Bayes generalization theory, advancing over previous analyses that were restricted to pure differential privacy or restricted problem classes. Leveraging our bounds, we showed that DP-SGD can be used to learn data-dependent priors while retaining generalization guarantees expressible directly in terms of optimization hyperparameters.

Despite these achievements, a number of limitations remain that we see as opportunities for future work. On the technical level, our explicit bounds for DP-SGD were restricted to a setup with fixed-size non-overlapping batches. In real-world settings, DP-SGD can often be run as a batch algorithm, in particular data can be accessed in a random-access fashion. In that case, sampling the data in batches that are random and potentially overlapping is known to improve the privacy guarantees, e.g. by employing *privacy amplification* (Ponomareva et al., 2023). Even in the streaming setting, tighter guarantees might be possible, e.g. through the use of noise that is correlated between update steps rather than independent (Pillutla et al., 2025). While such techniques are already covered in our analysis through the general-purpose results of Section 3 (and in our experiments we do use, e.g., Poisson batch sampling), it would be interesting to see how far the explicit analysis of Section 4 could be extended to these techniques. In particular, the role of batch size versus number of update steps would be an interesting candidate to explore in this context.

On a conceptual level, we would like to emphasize that even though our results apply to practically relevant setups, they should not be misinterpreted as a guide for practitioners. In particular, generalization bounds like the one we present are designed to hold uniformly across all data distributions and mechanisms for determining a model from the training data. As a consequence, they are necessarily conservative and often loose. We rather think of our results as a source of insights into how the core parameters of privacy and generalization interact. For example, the linear scaling of our main results with respect to the dataset size, in combination with the $I_\infty^{\delta/2}/n$ correction term in the PAC-Bayes bounds suggests that—at least by this analysis—the most promising strategy to guarantee generalization is not simply to use more data, but to enforce stronger privacy. Of course, the former is not contradictory to the latter, as more data can allow for larger noise multipliers without having to sacrifice utility, but we leave a formal study of these aspects to future work.

AI USE STATEMENT

In this work, we used generative AI tools for suggesting derivation paths for some of the mathematical statements, as well as for double-checking our proofs, in particular making sure that constants are tracked correctly. We have not used AI tools to formulate research questions or interpret results. Additionally, we used generative AI tools to identify relevant related work, suggest edits to the manuscript for better readability, and implement routine tasks, such as numerically evaluating bounds and data plotting. All AI-assisted work was reviewed manually for correctness and readability. We take responsibility for the final content of this work, including any parts produced by or with the aid of generative AI.

REFERENCES

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

Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2006.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private BERT. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning (ICML)*, 2018.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning (ICML)*, 2018.

Raef Bassily, Abhradeep Thakurta, and Adam Smith. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Symposium on Foundations of Computer Science (FOCS)*, 2014.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan R. Ullman. Algorithmic stability for adaptive data analysis. In *Symposium on Theory of Computing (STOC)*, 2016.

Felix Biggs. A note on the efficient evaluation of pac-bayes bounds. *arXiv preprint arXiv:2209.05188*, 2022.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning (ICML)*, 2015.

Simone Bombari and Marco Mondelli. Privacy for free in the overparameterized regime. *Proceedings of the National Academy of Sciences (PNAS)*, 122(15), 2025.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference (TCC)*, 2016.

Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research (JMLR)*, 12(3), 2011.

Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2012.

Lynn Chua, Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. Scalable DP-SGD: shuffling vs. poisson subsampling. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Lynn Chua, Badih Ghazi, Charlie Harrison, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Amer Sinha, and Chiyuan Zhang. Balls-and-bins sampling for DP-SGD. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 2014.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, 2006.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Uncertainty in Artificial Intelligence (UAI)*, 2017.

Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Symposium on Discrete Algorithms (SODA)*, 2019.

Vitaly Feldman and Moshe Shenfeld. Privacy amplification by random allocation. *arXiv:2502.08202*, 2025.

Diego García-Pérez, Emilio Parrado-Hernandez, and John Shawe-Taylor. Advances in PAC-Bayesian certification of deep neural networks: tighter closed-form inequalities and optimization of bounds on non-differentiable losses. In *Uncertainty in Artificial Intelligence (UAI)*, 2026.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Mihaela Hudişteanu, Nikita P. Kalinin, and Edwige Cyffers. DP-MicroAdam: Private and frugal algorithm for training and fine-tuning. In *EurIPS Workshop on Privacy-Preserving Machine Learning Workshop*, 2025.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2020.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

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

Guy Lever, François Laviolette, and John Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

Sanae Lotfi, Marc Finzi, Sanyam Kapoor, Andres Potapczynski, Micah Goldblum, and Andrew Gordon Wilson. PAC-Bayes compression bounds so tight that they can explain generalization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Sanae Lotfi, Marc Anton Finzi, Yilun Kuang, Tim G. J. Rudner, Micah Goldblum, and Andrew Gordon Wilson. Non-vacuous generalization bounds for large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Zelun Luo, Yuliang Zou, Yijin Yang, Zane Durante, De-An Huang, Zhiding Yu, Chaowei Xiao, Li Fei-Fei, and Animashree Anandkumar. Differentially private video activity recognition. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research (JMLR)*, 18:134:1–134:35, 2017.

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv:0411099*, 2004.

Andreas Maurer, Erfan Mirzaei, and Massimiliano Pontil. Generalization of Gibbs and Langevin Monte Carlo algorithms in the interpolation regime. *arXiv:2510.06028*, 2025.

David A McAllester. Some PAC-Bayesian theorems. In *Conference on Learning Theory (COLT)*, 1999.

Ilya Mironov. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium (CSF)*, 2017.

Allassan Tchangmena A Nken, Susan McKeever, Peter Corcoran, and Ihsan Ullah. Video-DPRP: A differentially private approach for visual privacy-preserving video human activity recognition. In *European Conference on Machine Learning and Data Mining (ECML/PKDD)*, 2025.

Emilio Parrado-Hernández, Amiran Ambroladze, John Shawe-Taylor, and Shiliang Sun. PAC-bayes bounds with data dependent priors. *Journal of Machine Learning Research (JMLR)*, 13:3507–3531, 2012.

María Pérez-Ortiz, Omar Rivasplata, John Shawe-Taylor, and Csaba Szepesvári. Tighter risk certificates for neural networks. *Journal of Machine Learning Research (JMLR)*, 22(227):1–40, 2021.

Krishna Pillutla, Jalaj Upadhyay, Christopher A. Choquette-Choo, Krishnamurthy Dvijotham, Arun Ganesh, Monika Henzinger, Jonathan Katz, Ryan McKenna, H. Brendan McMahan, Keith Rush, Thomas Steinke, and Abhradeep Thakurta. Correlated noise mechanisms for differentially private learning. *arXiv:2506.08201*, 2025.

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

Yu V Prokhorov. An extremal problem in probability theory. *Theory of Probability and its Applications*, 4(2): 201, 1959.

Arun Rajkumar and Shivani Agarwal. A differentially private stochastic gradient descent algorithm for multiparty classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Omar Rivasplata, Vikram M. Tankasali, and Csaba Szepesvári. PAC-Bayes with backprop. *arXiv:1908.07380*, 2019.

Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvári, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Ryan M. Rogers, Aaron Roth, Adam D. Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Symposium on Foundations of Computer Science (FOCS)*, 2016.

Tom Sander, Yaodong Yu, Maziar Sanjabi, Alain Durmus, Yi Ma, Kamalika Chaudhuri, and Chuan Guo. Differentially private representation learning via image captioning. In *International Conference on Machine Learning (ICML)*, 2024.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.

Zhongjie Shi, Puyu Wang, Chenyang Zhang, and Yuan Cao. Towards understanding generalization in DP-GD: A case study in training two-layer cnns. In *Conference on Artificial Intelligence (AAAI)*, 2026.

Amer Sinha, Thomas Mesnard, Ryan McKenna, Daogao Liu, Christopher A. Choquette-Choo, Yangsibo Huang, Da Yu, George Kaissis, Zachary Charles, Ruibo Liu, Lynn Chua, Pritish Kamath, Pasin Manurangsi, Steve He, Chiyuan Zhang, Badih Ghazi, Borja De Balle Pigem, Prem Eruvbetine, Tris Warkentin, Armand Joulin, and Ravi Kumar. VaultGemma: A differentially private gemma model. *arXiv:2510.15001*, 2025.

Makram Talih. A reference-invariant health disparity index based on rényi divergence. *The Annals of Applied Statistics*, 7(2):1217–1243, 2013.

Qiaoyue Tang, Frederick Shpilevskiy, and Mathias Lécuyer. DP-AdamBC: Your DP-Adam is actually DP-SGD (unless you apply bias correction). In *Conference on Artificial Intelligence (AAAI)*, 2024.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Tim van Erven and Peter Harremoës. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7), 2014.

Vladimir N Vapnik and Alexey Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

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

Hossein Zakerinia, Dorsa Ghobadi, and Christoph H Lampert. From low intrinsic dimensionality to non-vacuous generalization bounds in deep multi-task learning. *arXiv preprint arXiv:2501.19067*, 2025.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *International Conference on Learning Representations (ICLR)*, 2019.

Yuqing Zhu and Yu-Xiang Wang. Poisson subsampled rényi differential privacy. In *International Conference on Machine Learning (ICML)*, 2019.

A Extended Background

In this section we provide additional background information on the concepts introduced in Section 2.

A.1 Learning Setup

The learning setup in which we work (input space $\mathcal{X}$, model space $\mathcal{Y}$, loss function $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$) provides a unified notation for a number of standard learning scenarios. Most commonly, in supervised classification, $\mathcal{X}$ would contain input-output pairs, $\mathcal{Y}$, could be the model parameters, e.g. of a neural network, and ℓ measures if the model predicts the correct class label or not. However, the notation also applies more widely. For example, in unsupervised learning, the input set would be unlabeled data, the output could be a set of cluster centroids, and the loss function measures the distance between a data and its closest cluster centroid.

A.2 Differentially-Private Stochastic Gradient Descent (DP-SGD)

Several variants and improvements of DP-SGD and its privacy analysis have been developed.

One aspect under active development is *batch generation*: it has been shown that when data batches are created randomly, e.g. by Poisson sampling (Kasiviswanathan et al., 2011), *balls-and-bins* (Chua et al., 2025), k-out-of-t allocation (Feldman & Shenfeld, 2025), or even just *data shuffling* (Erlingsson et al., 2019), stronger privacy guarantees can be proven than what would be provided by analyzing DP-SGD simply as a concatenation of Gaussian mechanisms. However, these techniques tend to result in batches with different (random) batch sizes, which creates problems with practical implementation and resource utilization. A promising alternative is *Truncated Poisson Sampling* (Chua et al., 2024), which enforces an upper bound on the batch size and has been used, e.g., in training Google's VaultGemma LLM (Sinha et al., 2025).

In orthogonal efforts, it has been shown that the use of correlated instead of uncorrelated noise, equally strong privacy guarantees can often be achieved with less overall added noise and higher model utility (Pillutla et al., 2025).

We believe that an extension of our current analysis, which assumes classic fixed-size disjoint batches and independent noise, to these techniques should be possible, but would require changes to our proof techniques that go beyond the scope of this work.

Our discussion of DP-SGD left the choice of the *ModelUpdate* routine mostly open, because the specific form of the update does not affect either the max-information or the privacy guarantees, since the analysis depends only on the privatized update vectors.

Classically, the update steps of (DP-)SGD, given an update vector of u_t is

- $\theta_t \leftarrow \theta_{t-1} - \eta_t u_t$

for some (constant or time-varying) learning rate η_t (Abadi et al., 2016).

Including *momentum* of strength β and *weight decay* of strength α, changes the rule to

- $\theta_t \leftarrow (1-\alpha)\theta_{t-1} - \eta_t m_t$, where
- $m_t \leftarrow \beta m_{t-1} + (1-\beta)u_t$,

is the momentum term that is also updated in each step.

Finally, DP-Adam (Anil et al., 2022) or variants (Tang et al., 2024) additionally perform a component-wise normalization of the updates, e.g.

- $\theta_t = \theta_{t-1} - \frac{\eta_t}{\sqrt{\hat{v}_t}+\epsilon}\hat{n}_t$, for
- $\hat{n}_t = \frac{m_t}{1-\beta_1^t}$ with $m_t = \beta_1 m_{t-1} + (1-\beta_1)u_t$, and
- $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$ with $v_t = \beta_2 v_{t-1} + (1-\beta_2)u_t^2$.

One can see that in all of them, the model parameters are a deterministic linear (DP-SGD) or non-linear (DP-Adam) combination of the update vectors, so all of them are covered by our analysis.

A.3 PAC-BAYES GENERALIZATION BOUNDS.

Construction of priors and posteriors. At the core of PAC-Bayes theory lie stochastic classifiers, i.e. distributions, π (prior) and ρ (posterior) over the model space from which individual models are sampled. Our work is agnostic to how exactly the distributions are constructed from the output of a learning algorithm, as this step constitutes a postprocessing operations and therefore cannot lead to an increase of the max-information (or a reduction of privacy). In the literature, a number of suitable options exist:

- Most classically, the training step outputs the parameter vector, θ, of one model, from which one constructs a Gaussian distribution centered at y with fixed (often isotropic) covariance matrix. The variance can be treated as a hyper-parameter and chosen by model-selection. The use of a Gaussian form is motivated by the fact that it leads to explicit expression for $\mathrm{KL}(\rho\|\pi)$, which can then serve as part of the objective function for the model training step (Ambroladze et al., 2006; Dziugaite & Roy, 2017).
- In *Bayesian Deep Learning*, the model parameters directly correspond to the parameters of a distribution over models, e.g. by learning not only the mean but also (co)variance term of a Gaussian (Blundell et al., 2015; Rivasplata et al., 2019).
- When the space of model is finite, e.g. because the model coefficients are quantized, one can choose the posterior as a δ-peak at the (quantized) posterior. This way, the bound applies also to deterministic classifiers (Zhou et al., 2019; Lotfi et al., 2022).
- Exploiting the iterative nature of (DP-)SGD, one can the construct the mean and covariance matrix of a Gaussian distribution from its iterates (Maddox et al., 2019), or treat them as approximate samples from the *Gibbs posterior* (Mandt et al., 2017; Maurer et al., 2025).

From *kl* to more interpretable generalization bounds. PAC-Bayes generalization bounds of a form $\mathrm{kl}(\widehat{R}\|R) \leq C/n$ are also known as *fast-rate bounds*, because their convergence speed can be faster than the classical rate $O(\sqrt{1/n})$, specifically when the training risk, $\widehat{R}$, is small (Seeger, 2002; Catoni, 2007). This comes at the disadvantage that the bound on R is only implicit, and the bounding quantity, C/n, lacks immediate interpretability. Two main routes exist to solve this issue.

First, observe that $\mathrm{kl}(q\|p)$ is strictly monotonically increasing in $p \geq q$, so it is invertible in that range. Consequently, from the implicit form $\mathrm{kl}(\widehat{R}\|R) \leq C/n$, we can compute an upper-bound on R as $R \leq \mathrm{kl}^{-1}(\widehat{R}\|C/n)$, where

$$\mathrm{kl}^{-1}(q\|b) = \sup\big\{p \in [q,1] : \mathrm{kl}(q\|p) \leq b\big\}. \tag{16}$$

While kl^{-1} does not have a simple analytic form, we can compute it efficiently with numeric techniques, e.g., using a binary search or Newton's method.

Alternatively, one can use Pinsker's inequality, $\mathrm{kl}(q\|p) \geq 2(q-p)^2$, to derive bounds in the classic *generalization-gap* form, $R \leq \widehat{R} + \sqrt{\frac{C}{2n}}$. Due to its analytic expression, this form can also be used, e.g., as part of the training objective inside of a learning algorithm.

B PROOFS

In the following sections we provide the proofs that were left out of the main body of the manuscript due to space limitations. We present the proofs in a different order to the theorems' appearance in the text, to provide a gradual increase in complexity and make the similarities between proofs clearer.

While the proofs do share structural similarities, there are sufficiently many differences that we present each proof as a self-contained unit, with just the same notation running through.

As a guide, we provide the following structural outline that all of the proofs that follow:

1. The desired approximate max-information bound can be shown using (Dwork et al., 2015, Lemma 18) or a generalization thereof and a tail bound on the information density $\imath(Y;S)$, which is defined as the log-ratio between the joint probability of an input-output pair (S,Y), and the product of their marginal probabilities:
$$\imath(Y;S) := \log \frac{p(Y,S)}{p(Y)p(S)}. \tag{17}$$
2. Decompose $\imath(Y;S)$ (or an upper bound on it) into a sum, over data points/ batch entries, of increments Λ_i adapted to a filtration $\mathcal{F}_{i-1}$ of the preceding steps.
3. Bound the moment generating function $\log \mathbb{E}[\exp(\lambda\Lambda_i) \mid \mathcal{F}_{i-1}]$ of each increment (conditional on the history) using privacy or the mechanism's explicit form. Some of these bounds are stated in terms of previous points.
4. Compose these bounds, starting with the last point and working backwards to the first, resulting in a bound on the moment generating function $\log \mathbb{E}[\exp(\lambda\imath(Y;S))]$ of the information density. Where the per-step bounds depend on the preceding points, we instead we bound the moment generating function of the information density minus the sum of the per-step bounds, and separately show that this sum is small with high probability.
5. Optimize the order λ in this bound, wherever it is valid.

B.1 PROOF OF THEOREM 3

We now provide the proof of Theorem 3, which we restate here for completeness.

Theorem 3. *Let $\Psi : \mathcal{X}^n \to \mathbb{R}^d$ be a function with sensitivity at most $\zeta > 0$, and let $\mathcal{M}(S) = \Psi(S) + \sigma\zeta Z$ for $Z \sim \mathcal{N}(0, I)$ be the* Gaussian mechanism *with noise multiplier $\sigma > 0$ applied to Ψ. Let $\nu := \frac{n}{4\sigma^2}$, and for any $\beta \in (0,1)$ write $L := \log(1/\beta)$. If the elements of S are independent, then the* approximate max-information of $\mathcal{M}$ *fulfills,*

$$I_\infty^\beta(\mathcal{M}(S), S) \leq \nu + \inf_{\lambda \in (0, r_\nu)} \frac{1}{\lambda}\Big[L - \log\big(1 - (\lambda + \lambda^2)\nu\big)\Big] \tag{7}$$

with $r_\nu = \sqrt{\frac{1}{4} + \frac{1}{\nu}} - \frac{1}{2}$. In more explicit but less tight forms:

$$\leq 2\nu + \big(\tfrac{1}{2}\nu + \sqrt{\tfrac{1}{4}\nu^2 + \nu}\big)L + \sqrt{2(\nu^2 + 2\nu)L}, \tag{8}$$

and relaxed further, sorted by powers of ν:

$$\leq \big(2 + \sqrt{2L} + L\big)\nu + \big(L + 2\sqrt{L}\big)\sqrt{\nu}. \tag{9}$$

Proof. Without loss of generality, we can assume $\zeta = 1/\sigma$ and $\mathbb{E}[\Psi(S)] = 0$, because the max-information of an algorithm is not affected by a deterministic rescaling or shift of its outputs.

By Dwork et al. (2015, Lemma 18), adjusted to the fact the we use natural instead of binary logarithms, in order to establish $I_\infty^\beta(\mathcal{A}(S), S) \leq \kappa$, it suffices to prove $\mathbb{P}\{\imath(\mathcal{A}(S); S) \geq \kappa\} \leq \beta$ where

$$\imath(Y;S) := \log \frac{p(Y,S)}{p(Y)p(S)} = \log \frac{p(Y|S)}{p(Y)}. \tag{18}$$

is called the *information density*.

We do so in three steps: first, we establish a computable upper bound on $\imath(Y;S)$ (Lemma B.1). Then, we will derive an upper bound on the moment-generating function of this upper bound, and afterwards relate it to tail bounds of the original information density.

Lemma B.1.

$$\Pr\Big\{\imath(Y;S) \geq \kappa\Big\} \leq \Pr\Big\{g(S,Z) \geq \kappa - \frac{1}{2}\mathbb{E}\|\Psi(S)\|^2\Big\} \tag{19}$$

for

$$g(S,Z) := \big\langle \Psi(S), Z\big\rangle + \frac{1}{2}\|\Psi(S)\|^2 \tag{20}$$

where Z is a standard multivariate Gaussian, independent of all other quantities.

Proof. For any S and $\mathcal{A}$, where the respective expressions are defined, it holds almost surely for any output y:

$$\imath(\mathcal{A}(S)=y;S) = \log\frac{p(\mathcal{A}(S)=y|S)}{\mathbb{E}_{S'}p(\mathcal{A}(S')=y)} \leq \mathbb{E}_{S'}\log\frac{p(\mathcal{A}(S)=y|S)}{p(\mathcal{A}(S')=y|S')} \tag{21}$$

by Jensen's inequality, where S' is an independent copy of S. Inserting $\mathcal{A}(S) = \Psi(S) + Z$, for $Z \sim \mathcal{N}(0,\mathrm{I}_d)$, one has $p(\mathcal{A}(S)=y|S) = (2\pi)^{-d/2}e^{-\frac{1}{2}\|\Psi(S)-y\|^2}$ and $p(\mathcal{A}(S')=y|S') = (2\pi)^{-d/2}e^{-\frac{1}{2}\|\Psi(S')-y\|^2}$, i.e.

$$= \mathbb{E}_{S'}\Big[-\frac{1}{2}\|\Psi(S)\|^2 + \big\langle\Psi(S),y\big\rangle + \frac{1}{2}\|\Psi(S')\|^2 - \big\langle\Psi(S'),y\big\rangle\Big] \tag{22}$$

$$= \mathbb{E}_{S'}\Big[\big\langle\Psi(S)-\Psi(S'),y-\Psi(S)\big\rangle + \frac{1}{2}\|\Psi(S)-\Psi(S')\|^2\Big]. \tag{23}$$

Observing that $\mathbb{E}_{S'}\Psi(S') = \mathbb{E}_S\Psi(S)$ and $\mathbb{E}_{S'}\|\Psi(S)-\Psi(S')\|^2 = \|\Psi(S)\|^2 + \mathbb{E}_S\|\Psi(S)\|^2$, it follows that

$$= \Big[\big\langle\Psi(S),y-\Psi(S)\big\rangle + \frac{1}{2}\|\Psi(S)\|^2 + \frac{1}{2}\mathbb{E}_S\|\Psi(S)\|^2\Big] \tag{24}$$

Now, inserting Y for y and observing that $Y - \Psi(S) = Z$, we obtain that, almost surely

$$\imath(\Psi(S)+Z;S) \leq \big\langle\Psi(S),Z\big\rangle + \frac{1}{2}\|\Psi(S)\|^2 + \frac{1}{2}\mathbb{E}_S\|\Psi(S)\|^2 = g(S,Z) + \frac{1}{2}\mathbb{E}\|\Psi(S)\|^2 \tag{25}$$

which implies the statement of the Lemma. □

We will also use the following classical property of the Gaussian distribution:

Lemma B.2. *Let Z be a multi-variate standard Gaussian, $Z \sim \mathcal{N}(0,I_d)$. Then, for any fixed vector-valued a and scalar b, and it holds for every $t \in \mathbb{R}$*

$$\log\mathbb{E}[e^{t\langle Z,a\rangle+b}] = \log\mathbb{E}[e^{\frac{1}{2}t^2\|a\|^2+b}] \tag{26}$$

Proof. The proof is elementary, because the law of $\langle Z,a\rangle$ is one-dimensional Gaussian with variance parameter $\|a\|^2$, and b is not affected by the expectation. □

Next, we prove some useful lemmas which could be of independent interest, which we have not found in these precise forms in the literature.

Lemma B.3. *Let V be a centered vector-valued random variable whose support has diameter at most c. Then it holds for any vector x and any $0 \leq t < \frac{2}{c^2}$*

$$\log\mathbb{E}e^{t\|x+V\|^2} \leq \frac{1}{2}c^2t + \frac{t}{1-\frac{1}{2}c^2t}\|x\|^2. \tag{27}$$

Proof. Denote the law of V by P. For any x and t, define the tilted probability measure Q as $\frac{dQ}{dP} = \frac{e^{t\|x+V\|_2^2}}{\mathcal{Z}_{x,t}}$ with normalization constant $\mathcal{Z}_{x,t} = \mathbb{E}_P[e^{t\|x+V\|_2^2}]$.

By the definition of the Kullback-Leibler divergence and of Q, we have

$$\log \mathbb{E}_P[e^{t\|x+V\|^2}] = t\mathbb{E}_Q[\|x+V\|^2] - \mathrm{KL}(Q\|P) \tag{28}$$

Next, we derive an upper bound on the right hand side. First, observe that V has the same bounded support of diameter at most c under Q as under P. Consequently,

$$\mathbb{E}_Q\|V - \mathbb{E}_Q[V]\|^2 = \frac{1}{2}\mathbb{E}_{Q\otimes Q}\|V - V'\|^2 \leq \frac{c^2}{2}, \tag{29}$$

and therefore, it holds for any x:

$$\mathbb{E}_Q\|x+V\|^2 \leq \mathbb{E}_Q\|V - \mathbb{E}_Q[V]\|^2 + \|x + \mathbb{E}_Q[V]\|^2 \leq \|x + \mathbb{E}_Q[V]\|^2 + \frac{c^2}{2}. \tag{30}$$

Now, by applying the standard change-of-measure inequality $\mathbb{E}_{v\sim Q}[f(v)] \leq \mathrm{KL}(Q\|P) + \log \mathbb{E}_{v\sim P} e^{f(v)}$ with $f(v) = \langle u, v\rangle$ for some vector u, we obtain

$$\mathrm{KL}(Q\|P) \geq \langle u, \mathbb{E}_Q[V]\rangle - \log \mathbb{E}_P e^{\langle u, V\rangle} \geq \langle u, \mathbb{E}_Q[V]\rangle - \frac{c^2\|u\|^2}{8} \tag{31}$$

where the second inequality is due to Hoeffding's lemma, observing that the range of $\langle u, V\rangle$ has length at most $c\|u\|_2$, so $\log \mathbb{E}_P e^{\langle u,V\rangle} \leq \frac{c^2\|u\|^2}{8}$. Taking supremum over u, it follows:

$$\mathrm{KL}(Q\|P) \geq \sup_u \Big(\langle u, \mathbb{E}_Q[V]\rangle - \frac{c^2\|u\|^2}{8}\Big) = \frac{2\|E_Q[V]\|^2}{c^2}. \tag{32}$$

Inserting (30) and (32) into (28), it follows

$$\log \mathbb{E}_P e^{t\|x+V\|^2} \leq \frac{c^2}{2}t + t\|x + \mathbb{E}_Q[V]\|^2 - \frac{2\|E_Q[V]\|^2}{c^2}. \tag{33}$$

Expanding the norms and completing the square, it follows

$$= \frac{c^2}{2}t + \frac{t}{1-\frac{c^2}{2}t}\|x\|^2 - (\frac{2}{c^2} - t)\left\|\mathbb{E}_Q[V] - \frac{t}{\frac{2}{c^2}-t}x\right\|^2 \tag{34}$$

$$\leq \frac{c^2}{2}t + \frac{t}{1-\frac{c^2}{2}t}\|x\|^2, \tag{35}$$

because $t < \frac{2}{c^2}$ by assumption. This completes the proof of Lemma B.3. □

Proposition B.4. *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function that fulfills a bounded difference inequality with constant $c > 0$:*

$$\forall i \in \{1,\dots,n\} : \sup_{x_1,\dots,x_n,x'\in\mathcal{X}} \Big\| f(x_1,\dots,x_n) - f(x_1,\dots,x_{i-1},x',x_{i+1},\dots,x_n)\Big\| \leq c. \tag{36}$$

Let $X_1,\dots,X_n$ be independent random variables $X_i \in \mathcal{X}$ and write $X = (X_1,\dots,X_n)$. Then, it holds that:

$$\mathbb{E}\|f(X) - \mathbb{E}[f(X)]\|^2 \leq \frac{1}{2}nc^2. \tag{37}$$

For all $0 \leq \lambda < \frac{2}{nc^2}$,

$$\log \mathbb{E}e^{\lambda\|f(X)-\mathbb{E}[f(X)]\|^2} \leq -\log(1 - \frac{1}{2}nc^2\lambda). \tag{38}$$

Proof. To prove the statements, we work iteratively over the steps of a Doob martingale, exploiting exponential tilts from Lemma B.3 along the way. Write $F := f(X) - \mathbb{E}[f(X)]$ and introduce the Doob martingale:

$$M_i := \mathbb{E}[F|\mathcal{F}_i], \qquad D_i := M_i - M_{i-1}, \tag{39}$$

where $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$. By construction, $M_0 = 0$, $M_n = F$, and $\mathbb{E}[D_i|\mathcal{F}_{i-1}] = 0$. Conditioned on any fixed past, $X_1, \dots, X_{i-1}$, set

$$g_i(x) := \mathbb{E}_{X_{i+1},\dots,X_n}\big[f(X_1, \dots, X_{i-1}, x, X_{i+1}, \dots, X_n)\big]. \tag{40}$$

By the bounded difference inequality of f, it follows that $\|g_i(x) - g_i(x')\| \le c$ almost surely for all $x, x' \in \mathcal{X}_i$, so the conditional law of $D_i = g_i(X_i) - \mathbb{E}[g_i(X_i)|\mathcal{F}_{i-1}]$ has diameter at most c. From (29), now under the conditional law, it follows that $\mathbb{E}\|D_i\|^2|\mathcal{F}_{i-1}] \le \frac{c^2}{2}$. Consequently, by the telescoping property of D_i and the orthogonality property of martingales,

$$\mathbb{E}\|f(X) - \mathbb{E}[f(X)]\|^2 = \mathbb{E}\|F\|^2 = \sum_{i=1}^{n} \mathbb{E}\|D_i\|^2 \le \frac{nc^2}{2}, \tag{41}$$

which proves (37).

Conditioned on $\mathcal{F}_{i-1}$, all $M_1, \dots, M_{i-1}$ are determined, so we can invoke Lemma B.3 with $x = M_{i-1}$ and obtain for $v := \frac{c^2}{2}$

$$\log \mathbb{E}\big[e^{t\|M_i\|^2} \;\big|\; \mathcal{F}_{i-1}\big] = \log \mathbb{E}\big[e^{t\|D_i + M_{i-1}\|^2} \;\big|\; \mathcal{F}_{i-1}\big] \le vt + \frac{t}{1-vt}\|M_{i-1}\|^2. \tag{42}$$

In order to obtain a combined bound for any $\lambda \in (0, \frac{2}{nc^2})$, set $\lambda_j := \frac{\lambda}{1-(n-j)v\lambda}$ for $j = 0, \dots, n$. First, observe that $v\lambda_j = \frac{v\lambda}{1-(n-j)v\lambda} < \frac{1}{n-(n-j)} \le 1$ for $j = 1, \dots, n$, because $v\lambda < \frac{2v}{nc^2} = \frac{1}{n}$. Therefore, we can invoke (42) with $t = \lambda_j$ and obtain

$$\log \mathbb{E}\big[e^{\lambda_j\|M_j\|^2}\big] = \log \mathbb{E}\Big[\mathbb{E}\big[e^{\lambda_j\|M_j\|^2} \;\big|\; \mathcal{F}_{j-1}\big]\Big] \le \lambda_j v + \log \mathbb{E}\Big[e^{\lambda_{j-1}\|M_{j-1}\|^2}\Big], \tag{43}$$

where in the last step we have exploited that, by construction, $\frac{\lambda_j}{1-\lambda_j v} = \lambda_{j-1}$ for any $j = 1, \dots, n$. Now, the final identity follows by iterating backwards from $M_n = F$ to $M_0 = 0$, yielding

$$\log \mathbb{E}\big[e^{\lambda\|F\|^2}\big] \le \sum_{j=1}^{n} \lambda_j v = \sum_{k=0}^{n-1} \frac{\lambda v}{1-k\lambda v} \le \int_0^n \frac{\lambda v}{1-\kappa\lambda v} d\kappa = -\log(1-nv\lambda), \tag{44}$$

where the second inequality holds because the function $u \mapsto \frac{\lambda v}{1-u\lambda v}$ is monotonically increasing. Substituting $v = \frac{1}{2}c^2$ completes the proof. □

We now have all components to complete the proof of Theorem 3:

Proof of Theorem 3 (continued). To derive a tail bound on $g(S, Z)$, we study its moment generating function.

$$\log \mathbb{E}e^{\lambda g(S,Z)} = \log \mathbb{E}\big[e^{\lambda(\langle\Psi, Z\rangle + \frac{1}{2}\|\Psi\|^2)}\big] = \log \mathbb{E}\Big[\mathbb{E}\big[e^{\lambda\langle\Psi,Z\rangle + \frac{1}{2}\lambda\|\Psi\|^2}\big] \;\Big|\; \Psi\Big], \tag{45}$$

where the last equality is due to the tower property of expectations. Because Z is independent of all other variables, we can evaluate the conditional expectation by means of Lemma B.2

$$= \log \mathbb{E}e^{\frac{\lambda^2+\lambda}{2}\|\Psi\|^2}. \tag{46}$$

By assumption, Ψ fulfills a bounded difference inequality with parameter $\frac{1}{\sigma}$. Write $\nu = \frac{n}{4\sigma^2}$ and $r_\nu := \sqrt{\frac{1}{4} + \frac{1}{\nu}} - \frac{1}{2}$, such that $\lambda \in (0, r_\nu)$ implies $\frac{1}{2}(\lambda^2 + \lambda) < \frac{1}{2}(r_\nu + 1)r_\nu = \frac{1}{2\nu} = \frac{2}{nc^2}$. Therefore, for $0 \le \lambda < r_\nu$, we can invoke Proposition B.4 with $c = \frac{1}{\sigma}$ and obtain

$$\le -\log(1 - (\lambda^2 + \lambda)\nu). \tag{47}$$

Next, we derive a tail bound of the type required for Lemma B.1 following standard Chernoff arguments, e.g. (Boucheron et al., 2013, Chapter 2). By Markov's inequality applied to $e^{\lambda g(S,Z)}$, one obtains for all $\lambda > 0$:

$$\Pr(g(S,Z) \geq \kappa) \leq e^{-\lambda\kappa}\mathbb{E}[e^{\lambda g(S,Z)}] \tag{48}$$

To ensure that the right hand side remain below some $\beta \in (0,1)$, it suffices that

$$\kappa = \inf_{\lambda>0} \frac{1}{\lambda}\Big(\log \mathbb{E}e^{\lambda g(S,Z)} + L\Big) \tag{49}$$

with $L = \log(1/\beta)$. Inserting (47) and restricting λ to the permitted range, we obtain

$$\leq \inf_{\lambda\in(0,r_\nu)} \frac{1}{\lambda}\Big(-\log(1-(\lambda+\lambda^2)\nu) + L\Big). \tag{50}$$

Furthermore, from the first statement of Proposition B.4 we obtain $\frac{1}{2}\mathbb{E}\|\Psi\|^2 \leq \frac{1}{4}\frac{n}{\sigma^2} = \nu$. Inserting this and (50) into Lemma B.1 yields the first statements of Theorem 3.

For the remaining statements, observe that

$$1-(\lambda+\lambda^2)\nu = (1-a\lambda)(1+b\lambda) \tag{51}$$

for $a = \frac{\nu}{2} + \sqrt{\frac{\nu^2}{4}+\nu}$ and $b = -\frac{\nu}{2} + \sqrt{\frac{\nu^2}{4}+\nu}$, which in particular fulfills $a - b = \nu$ and $ab = \nu$. Therefore,

$$-\log(1-(\lambda+\lambda^2)\nu) - \lambda\nu = -\log(1-a\lambda) - a\lambda - \log(1+b\lambda) + b\lambda. \tag{52}$$

By the standard relations $-\log(1-u) - u \leq \frac{u^2}{2(1-u)}$ and $-\log(1+u)+u \leq \frac{u^2}{2}$, and exploiting $a^2+b^2 = (a-b)^2 + 2ab = \nu^2 + 2\nu$, we obtain

$$\leq \frac{a^2\lambda^2}{2(1-a\lambda)} + \frac{b^2\lambda^2}{2} \leq \frac{(a^2+b^2)\lambda^2}{2(1-a\lambda)} = \frac{(\nu^2+2\nu)\lambda^2}{2(1-a\lambda)}. \tag{53}$$

Therefore, we can upper bound the infimum:

$$\inf_{\lambda\in(0,r_\nu)} \frac{1}{\lambda}\Big(-\log(1-(\lambda+\lambda^2)\nu) + L\Big) \leq \inf_{\lambda\in(0,r_\nu)}\Big[\nu + \frac{(\nu^2+2\nu)\lambda}{2(1-a\lambda)} + \frac{L}{\lambda}\Big] \tag{54}$$

Taking derivative with respect to λ on the right hand side, we find the exact minimizer $\lambda^* = \frac{\sqrt{2L}}{\sqrt{\nu^2+2\nu}+a\sqrt{2L}}$, which indeed fulfills $0 \leq \lambda^* < r_\nu$. Therefore,

$$\leq \nu + \frac{(\nu^2+2\nu)\lambda^*}{2(1-a\lambda^*)} + \frac{L}{\lambda^*} \tag{55}$$

Observing $\frac{1}{\lambda^*} = a + \sqrt{\frac{\nu^2+2\nu}{2L}}$ and $\frac{\lambda^*}{1-a\lambda^*} = \sqrt{\frac{2L}{\nu^2+2\nu}}$, it follows

$$= \nu + \frac{(\nu^2+2\nu)}{2}\sqrt{\frac{2L}{\nu^2+2\nu}} + L\Big(a + \sqrt{\frac{\nu^2+2\nu}{2L}}\Big) \tag{56}$$

$$= \nu + \big(\tfrac{1}{2}\nu + \sqrt{\tfrac{1}{4}\nu^2+\nu}\big)L + \sqrt{2(\nu^2+2\nu)L}. \tag{57}$$

Exploiting $\sqrt{\frac{1}{4}\nu^2+\nu} \leq \frac{1}{2}\nu + \sqrt{\nu}$ and $\sqrt{\nu^2+2\nu} \leq \nu + \sqrt{2\nu}$, we obtain a variant that is less tight but can be sorted by powers of ν:

$$\leq (1+L+\sqrt{2L})\nu + (L+2\sqrt{L})\sqrt{\nu}. \tag{58}$$

This proves the remaining statements of Theorem 3. □

Finally, we also prove the claim that the optimization problem required for evaluation (7) is convex.

Lemma B.5. *For $\nu > 0$, let $r_\nu := \sqrt{\frac{1}{4} + \frac{1}{\nu}} - \frac{1}{2}$. Then*

$$f(\lambda) := \frac{1}{\lambda}\Big[L - \log\big(1 - (\lambda + \lambda^2)\nu\big)\Big] \tag{59}$$

is a smooth convex function on the interval $(0, r_\nu)$.

Proof. The smoothness of f in the given interval is clear, because it is a composition of smooth functions and $\lambda \in (0, r_\nu)$ implies $(\lambda + \lambda^2)\nu < 1$, so f is well-defined in the given interval. To show convexity, we decompose the log terms

$$-\log\big(1 - (\lambda + \lambda^2)\nu\big) = -\log(1 - a\lambda) + -\log(1 + b\lambda) \tag{60}$$

with explicit $a > b > 0$ as for (52) in the proof of Theorem 3. Now, expressing the logarithms by their power series, $-\log(1 - x) = \sum_{k=1}^{\infty} x^k/k$, we obtain

$$= \sum_{k=1}^{\infty} \frac{a^k + (-1)^k b^k}{k} \lambda^k, \tag{61}$$

and therefore

$$f(\lambda) = \frac{L}{\lambda} + \sum_{k=1}^{\infty} \frac{a^k + (-1)^k b^k}{k} \lambda^{k-1}. \tag{62}$$

The second term is strictly convex, because $a > b > 0$ implies that all coefficients $a^k + (-1)^k b^k$ of the power series are strictly positive and $\lambda \in (0, r_\nu)$ in particular ensures $\lambda > 0$. The first term, $\frac{L}{\lambda}$, is convex for all $\lambda > 0$. Therefore, the claim follows. □

B.2 Proof of Theorem 4

We now provide the proof of Theorem 4, which we restate here for completeness.

Theorem 4. *Let $\mathcal{A}$ be the* DP-SGD algorithm 1 *with hyperparameters ζ (clipping threshold) and σ (noise multiplier), that is run on T data batches of size m, which contain independent data samples. For any $\beta \in (0, 1)$, we write $L = \log(1/\beta)$. Then, the* approximate max-information of $\mathcal{A}$ *fulfills*

$$I_\infty^\beta(\mathcal{A}(S), S) \leq 2T\nu + \inf_{\lambda \in (0, r_{4\nu})} \frac{1}{\lambda}\Big[L - \frac{T}{2}\log\big(1 - 4(\lambda + \lambda^2)\nu\big)\Big], \tag{10}$$

where $\nu = \frac{m}{4\sigma^2}$, and $r_{4\nu} = \frac{1}{2}\sqrt{1 + \frac{1}{\nu}} - \frac{1}{2}$. In less tight, but more explicit, form:

$$\leq 4T\nu + 2\big(\nu + \sqrt{\nu^2 + \nu}\big)L + 2\sqrt{(4\nu^2 + 2\nu)TL}. \tag{11}$$

Relaxed further and sorted by powers of ν, it holds that:

$$\leq 4(T + L + \sqrt{TL})\nu + 2(L + \sqrt{2TL})\sqrt{\nu}. \tag{12}$$

Proof. Our proof follows similar steps to the proof of Theorem 3, but now applied repeatedly, as streaming DP-SGD is an iterative application of the Gaussian mechanism. Care is still needed as the update vectors of different batches are not independent of each other, as the model weights change over time. In addition, we are able to improve some constants compared to the Gaussian case, because DP-SGD's feature function is a sum of per-sample bounded vectors over a batch, which allows better control of the concentration that just assuming a function of bounded sensitivity as for Theorem 3.

For easier discussion, we prove this result for the general form DP-SGD that outputs each of its intermediate updates, not just the final model parameters. This is same approach used, for example, for proving its privacy properties. Specifically, this means that DP-SGD outputs its initialization, θ_0, and each update vector u_t, from which the intermediate models, θ_t, are computed. The restricted form of DP-SGD that only outputs the final model is covered by this analysis, too, because it can be obtained by a postprocessing operation (ignore all outputs except the last one), and can therefore not have higher approximate max-information.

As in the previous section, without loss of generality it suffices to prove the result for $\zeta = 1/\sigma$, such that the added noise has unit variance. The behavior for other choices can be recovered by rescaling the loss function (which results in gradients being scaled accordingly) and post-multiplication, which does not influence the max-information. However, we now cannot easily assume a fixed centered feature function by invoking global shift, because Ψ depends on the previously computed model parameters and therefore changes between steps.

For a dataset $S = (B_1, \dots, B_T)$ of independently sampled data batches, each of which consists of m i.i.d. sampled data points, $B_t = (X_1^t, \dots, X_m^t)$, let $Y = (Y_0, Y_1, Y_2, \dots, Y_T)$ with $Y_t = (\Theta_t, U_t)$, denote the output of the (extended) DP-SGD, where $Y_0 = (\Theta_0, \emptyset)$ reflects the potentially independent random initialization. We write $Y_s^t = (Y_s, Y_{s+1}, \dots, Y_t)$ for partial segments of Y, and analogously $y_s^t = (y_s, \dots, y_t)$.

As in the previous section, it suffices to prove $\mathbb{P}\{\imath(S, Y) \geq \kappa\} \leq \beta$ for $\imath(S, Y) = \log \frac{p(Y|S)}{p(Y)}$. Analogously to Section 4.1, we do so by proving a number of intermediate lemmas, which in the end we will combine to establish the statement of Theorem 4.

Lemma B.6. *For any model parameter vector θ, let $\Psi(x_1, \dots, x_m; \theta) = \sum_{i=1}^m \phi(x_i; \theta)$ with $\phi(x; \theta) = \mathit{clip}\big(\nabla \ell(x; \theta), \zeta\big)$. For any $t \in \{1, \dots, T\}$, define random variables $\Psi_t := \Psi(B_t; \Theta_{t-1})$ and $G_t = \Psi_t - \mathbb{E}[\Psi_t | Y_0^{t-1}]$, and write $G = (G_1, \dots, G_T)$. Then, it holds for any $\kappa \geq 0$ that:*

$$\Pr\{\imath(S, Y) > \kappa\} \leq \Pr\Big\{ g(G, Z) + \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}[\|G_t\|^2 | Y_0^{t-1}] > \kappa \Big\} \tag{63}$$

for

$$g(G, Z) := \sum_{t=1}^{T} g_t(G_t, Z_t) \quad \textit{with} \quad g_t(G_t, Z_t) := \big\langle Z_t, G_t \big\rangle + \tfrac{1}{2}\|G_t\|^2. \tag{64}$$

where $Z = (Z_1, \dots, Z_T)$ consists of standard multivariate Gaussians, such that each Z_t is fresh*, i.e. independent of the data, the initialization and the previous noise variables $Z_1, \dots, Z_{t-1}$.*

Proof. Let $\tilde{S} = (\tilde{B}_1, \dots, \tilde{B}_T)$ be an independent copy of S, and denote by $\tilde{Y}$ the output of the algorithm on $\tilde{S}$. By Jensen's inequality applied to $t \mapsto \log \frac{1}{t}$, we have for any output $y = (y_1, \dots, y_T)$:

$$\imath(S, Y = y) \leq \mathbb{E}_{\tilde{S}} \log \frac{p(Y = y|S)}{p(\tilde{Y} = y|\tilde{S})} = \mathbb{E}_{\tilde{S}} \log \prod_{t=1}^{T} \frac{p(Y_t = y_t | Y_0^{t-1} = y_0^{t-1}, S)}{p(\tilde{Y}_t = y_t | \tilde{Y}_0^{t-1} = y_0^{t-1}, \tilde{S})} \tag{65}$$

$$= \sum_{t=1}^{T} \mathbb{E}_{\tilde{S}} \log \frac{p(Y_t = y_t | Y_0^{t-1} = y_0^{t-1}, S)}{p(\tilde{Y}_t = y_t | \tilde{Y}_0^{t-1} = y_0^{t-1}, \tilde{S})}. \tag{66}$$

For each $t \in \{1, \dots, T\}$, DP-SGD computes the model update Θ_t deterministically from the current update vector U_t and the past model parameters Θ_0^{t-1}. Therefore, with $y_t = (u_t, \theta_t)$,

$$= \sum_{t=1}^{T} \mathbb{E}_{\tilde{S}} \log \frac{p(U_t = u_t | Y_0^{t-1} = y_0^{t-1}, S)}{p(\tilde{U}_t = u_t | \tilde{Y}_0^{t-1} = y_0^{t-1}, \tilde{S})}. \tag{67}$$

The model update, U_t, is the result of an application of the Gaussian mechanism, so its distribution, conditioned on any value for Y_0^{t-1} and S, is Gaussian, namely $\mathcal{N}(\Psi(B_t; \Theta_{t-1}), \mathrm{I})$. The analog property holds for $\tilde{U}_t$. Consequently,

$$= \sum_{t=1}^{T} \mathbb{E}_{\tilde{S}} \log \frac{\mathcal{N}(u_t; \Psi(B_t; \theta_{t-1}), \mathrm{I})}{\mathcal{N}(u_t; \Psi(\tilde{B}_t; \theta_{t-1}), \mathrm{I})}, \tag{68}$$

and we can analyze each term in (68) in the same way as we had in the case of a single application of the Gaussian mechanism in the previous section. In particular, because of the independence between batches, we can replace the expectation with respect to $\tilde{S}$ by the expectation with respect to the respective $\tilde{B}_t$.

Analogously to the proof of Lemma B.1, we obtain for any batches B and $\tilde{B}$, any vector $u \in \mathbb{R}^d$ and conditioned on any model parameters $\Theta \in \mathbb{R}^d$ that:

$$\mathbb{E}_{\tilde{B}} \log \frac{\mathcal{N}(u;\, \Psi(B;\Theta), \mathrm{I})}{\mathcal{N}(u;\, \Psi(\tilde{B};\Theta), \mathrm{I})} = \big\langle u - \Psi(B;\Theta), \Psi(B;\Theta) - \mathbb{E}_B[\Psi(B;\Theta)] \big\rangle + \frac{1}{2}\|\Psi(B;\Theta) - \mathbb{E}_B[\Psi(B;\Theta)]\|^2 + \frac{1}{2}\mathbb{E}_B\|\Psi(B;\Theta) - \mathbb{E}_B[\Psi(B;\Theta)]\|^2, \tag{69}$$

where we have used that $\mathbb{E}_B[\Psi(B;\Theta)|\Theta] = \mathbb{E}_{\tilde{B}}[\Psi(\tilde{B};\Theta)|\Theta]$.

We adopt the notation $\Psi_t = \Psi(B_t; \Theta_{t-1})$ and $G_t = \Psi_t - \mathbb{E}[\Psi_t | Y_0^{t-1}]$ from the Lemma statement and insert U_t for u. Then, in continuation of (68), we have:

$$\imath(S, Y) \leq \sum_{t=1}^{T} \big\langle U_t - \Psi_t, G_t \big\rangle + \frac{1}{2}\|G_t\|^2 + \frac{1}{2}\mathbb{E}_S\|G_t\|^2. \tag{70}$$

Observing that each $U_t - \Psi_t(S)$ has the distribution of a standard Gaussian, Z_t, that is independent of all other variables, we obtain the statement of Lemma B.6. □

We also prove refined versions of Lemma B.3 and Proposition B.4 for sums of bounded vectors rather than just functions that fulfill a bounded difference inequality.

Lemma B.7. *Let V be a vector-valued random variable that fulfills $\|V\| \leq r$ almost surely, and let $\bar{V} = V - \mathbb{E}[V]$. Then it holds for any vector x and any $0 \leq t < \frac{1}{2r^2}$*

$$\log \mathbb{E} e^{t\|x+\bar{V}\|^2} \leq r^2 t + \frac{t}{1 - 2r^2 t}\|x\|^2. \tag{71}$$

Note that the support of $\bar{V}$ has at most diameter $2r$, so Lemma B.3 would guarantee $\log \mathbb{E} e^{t\|x+\bar{V}\|^2} \leq 2r^2 t + \frac{t}{1-2r^2 t}\|x\|^2$. In contrast, Lemma B.7 improves the leading term by exploiting that $\bar{V}$ stems from a variable with norm bounded by r.

Proof. The proof mostly follows the same steps as for Lemma B.3, so we only detail where differences occur. Specifically, since we know $\|V\| \leq r$, we obtain

$$\mathbb{E}_Q\|\bar{V} - \mathbb{E}_Q[\bar{V}]\|^2 = \mathbb{E}_Q\|V - \mathbb{E}_Q[V]\|^2 = \mathbb{E}_Q\|V\|^2 - \|\mathbb{E}_Q V\|^2 \leq r^2. \tag{72}$$

This is in contrast to (29), where we had to argue with the diameter instead of the radius, resulting in an upper bound of $\frac{1}{2}c^2 = 2r^2$, i.e. twice as big.

Consequently, we obtain as analog of (30), for all x:

$$\mathbb{E}_Q\|x + \bar{V}\|^2 \leq \|x + \mathbb{E}_Q[\bar{V}]\|^2 + r^2. \tag{73}$$

The other steps cannot be improved as easily, in particular the bound $\mathrm{KL}(Q\|P) \geq \frac{\|\mathbb{E}_Q[\bar{V}]\|^2}{2r^2}$ is the same as (32) for $c = 2r$. Combining the terms and optimizing, it holds for any x:

$$\log \mathbb{E}_P e^{t\|x+\bar{V}\|^2} \leq r^2 t + t\|x + \mathbb{E}_Q[\bar{V}]\|^2 - \frac{\|E_Q[\bar{V}]\|^2}{2r^2}. \tag{74}$$

Expanding the norms and completing the square, it follows

$$\leq r^2 t + \frac{t}{1 - 2r^2 t}\|x\|^2, \tag{75}$$

because $t < \frac{1}{2r^2}$ by assumption. This completes the proof of Lemma B.7. □

As a refined version of Proposition B.4 we obtain

Proposition B.8. *Let $X_1, \ldots, X_m$ be independent random variables that fulfill $\|X_i\| \leq r$ almost surely, and write $X = (X_1, \ldots, X_m)$. Let $f(X) = \sum_{i=1}^m X_i$. Then, for all $0 \leq \lambda < \frac{1}{2mr^2}$, it holds that*

$$\mathbb{E}[\|f(X) - \mathbb{E}[f(X)]\|^2] \leq mr^2 \tag{76}$$

$$\log \mathbb{E} e^{\lambda\|f(X) - \mathbb{E}[f(X)]\|^2} \leq -\frac{1}{2}\log(1 - 2mr^2\lambda). \tag{77}$$

Comparing (76) to the previous (37) with $c = 2r$, one sees an improvement by a factor of two in the leading constant. Similarly, the leading constant of (77) is half of the leading constant in (38), but the terms inside the logarithms remain as they would be for $c = 2r$.

Proof. The proof of (76) is standard. By the additive form of f and the independence condition, it holds that

$$\mathbb{E}[\|f(X) - \mathbb{E}[f(X)]\|^2] = \sum_{i=1}^{m} \mathbb{E}\|X_i - \mathbb{E}[X_i]\|^2 \leq \sum_{i=1}^{m} \mathbb{E}\|X_i\|^2 \leq mr^2. \tag{78}$$

The proof of (77) follows the same steps as for Proposition B.4, except simplified due to the additive structure, and using Lemma B.7 in place of Lemma B.3.

Writing $\bar{X}_j = X_j - \mathbb{E}[X_j]$ and $F_j := \sum_{i=1}^{j} \bar{X}_i$ for $j = 1, \ldots, m$, we obtain for $v = 2r^2$:

$$\log \mathbb{E}\big[e^{t\|F_j\|^2} \;\big|\; F_{j-1}\big] = \log \mathbb{E}\big[e^{t\|\bar{X}_j + F_{j-1}\|^2} \;\big|\; \mathcal{F}_{j-1}\big] \leq \frac{1}{2}vt + \frac{t}{1-vt}\|F_{j-1}\|^2. \tag{79}$$

because F_{j-1} is determined by the conditioning and $\bar{X}_j$ fulfills the conditions of Lemma B.7.

Employing the same iterative construction as in the proof of Proposition B.4, with $\lambda_j = \frac{\lambda}{1-(m-j)v\lambda}$ but now with $v = 2r^2$ as above and for $\lambda \in (0, \frac{1}{2mr^2})$, we obtain

$$\log \mathbb{E}\big[e^{\lambda_j \|F_j\|^2}\big] = \log \mathbb{E}\Big[\mathbb{E}\big[e^{\lambda_j \|F_j\|^2} \;\big|\; F_{j-1}\big]\Big] \leq \frac{1}{2}v\lambda_j + \log \mathbb{E}\big[e^{\lambda_{j-1}\|F_{j-1}\|^2}\big], \tag{80}$$

and ultimately, by iterating backward:

$$\log \mathbb{E}\big[e^{\lambda\|F_m\|^2}\big] \leq \sum_{j=1}^{m} \frac{1}{2}\lambda_j v \leq -\frac{1}{2}\log(1 - mv\lambda), \tag{81}$$

which concludes the proof. □

Proof of Theorem 4 (continued). We now establish a bound on the moment-generating function of $\sum_{t=1}^{T} g_t(G_t, Z_t)$, as introduced in (64), exploiting the tower property of expectations.

For any $1 \leq t \leq T$, let $\mathcal{F}_{t-1} = \sigma(\Theta_0, B_1, \ldots, B_{t-1}, Z_1, \ldots, Z_{t-1})$. Then any partial sum, $\sum_{j=1}^{t-1} g_j(Z_j, G_j)$, is measurable with respect to this filtration, and (B_t, Z_t) are independent of it. Consequently, it holds for the partial sum:

$$\mathbb{E}\Big[e^{\lambda \sum_{j=1}^{t} g_j(Z_j,G_j)}\Big] = \mathbb{E}\Big[e^{\lambda \sum_{j=1}^{t-1} g_j(Z_j,G_j) + \lambda\langle Z_t, G_t\rangle + \frac{\lambda}{2}\|G_t\|^2}\Big] \tag{82}$$

$$= \mathbb{E}\Big[e^{\lambda \sum_{j=1}^{t-1} g_j(Z_j,G_j)} \mathbb{E}\Big[\mathbb{E}\big[e^{\lambda\langle Z_t, G_t\rangle} | B_t, \mathcal{F}_{t-1}\big]\, e^{\frac{\lambda}{2}\|G_t\|^2} \;\Big|\; \mathcal{F}_{t-1}\Big]\Big]. \tag{83}$$

By Lemma B.2, it follows that

$$= \mathbb{E}\Big[e^{\lambda \sum_{j=1}^{t-1} g_j(Z_j,G_j)} \mathbb{E}\big[e^{\frac{\lambda+\lambda^2}{2}\|G_t\|^2} \;\big|\; \mathcal{F}_{t-1}\big]\Big] \tag{84}$$

Remember that $G_t = \Psi(B_t, \Theta_{t-1}) - \mathbb{E}[\Psi(B_t, \Theta_{t-1})|Y_0^{t-1}]$, so conditioned on $\mathcal{F}_{t-1}$, its randomness is only due to B_t, which contains independent elements. Consequently, G_t fulfills the conditions of Proposition B.8 with $r = 1/\sigma$, and we obtain, as long as $\frac{\lambda+\lambda^2}{2} < \frac{\sigma^2}{2m}$:

$$\leq \mathbb{E}\Big[e^{\lambda \sum_{j=1}^{t-1} g_j(Z_j,G_j)} \frac{1}{(1 - \frac{m}{\sigma^2}(\lambda + \lambda^2))^{1/2}}\Big]. \tag{85}$$

Applying the above identity iteratively, we obtain a bound for the complete expression:

$$\mathbb{E}\Big[e^{\lambda \sum_{j=1}^{T} g_j(Z_j,G_j)}\Big] \leq \frac{1}{(1 - \frac{m}{\sigma^2}(\lambda + \lambda^2))^{T/2}}. \tag{86}$$

To obtain tail bounds, we follow the same Chernoff arguments as in the previous section, only with slightly different constants:

$$\Pr(g(G,Z) \geq \kappa) \leq e^{-\lambda\kappa}\mathbb{E}[e^{\lambda g(G,Z)}] \tag{87}$$

With $L = \log(1/\beta)$ for any $\beta \in (0,1)$:

$$\kappa = \inf_{\lambda>0} \frac{1}{\lambda}\Big(\log \mathbb{E}e^{\lambda g(G,Z)} + L\Big) \tag{88}$$

Analogous to (50) we obtain, with $r_{4\nu} = \frac{1}{2}\sqrt{1+\frac{1}{\nu}} - \frac{1}{2}$

$$\leq \inf_{\lambda\in(0,r_{4\nu})} \frac{1}{\lambda}\Big(-\frac{T}{2}\log(1-4(\lambda+\lambda^2)\nu) + L\Big). \tag{89}$$

Furthermore, from Proposition B.8 it follows that $\frac{1}{2}\mathbb{E}[\|G_t\|^2|Y_0^{t-1}] \leq \frac{m}{2\sigma^2} = 2\nu$ for all $t = 1,\dots,T$, so $\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\|G_t\|^2|Y_0^{t-1}] \leq 2T\nu$. Inserting this and (89) into Lemma B.6 yields the first statements of Theorem 4.

For the remaining statements, note that the logarithmic term, $\log\big(1-4(\lambda+\lambda^2)\nu\big)$, has the same form as (50) in case of the Gaussian mechanism, only with ν replaced by 4ν. The additional factor $T/2$ in front of it can be moved outside of the infimum by changing L to $2L/T$. Subsequently, we can reuse the steps from the previous section and obtain

$$\inf_{\lambda\in(0,r_{4\nu})} \frac{1}{\lambda}\Big(-\frac{T}{2}\log(1-4(\lambda+\lambda^2)\nu) + L\Big) \tag{90}$$

$$\leq 2T\nu + 2\big(\nu+\sqrt{\nu^2+\nu}\big)L + 2\sqrt{(4\nu^2+2\nu)TL}, \tag{91}$$

and, simplified further and sorted by powers of ν,

$$\leq (2T+4L+4\sqrt{TL})\nu + 2(L+\sqrt{2TL})\sqrt{\nu}. \tag{92}$$

In combination with $\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\|G_t\|^2|Y_0^{t-1}] \leq 2T\nu$ as above, the remaining statements of Theorem 4 follow. □

Finally, we also formally state the fact that the optimization problem in (10) is strictly convex.

Lemma B.9. *For $\nu > 0$, let $r_{4\nu} := \frac{1}{2}\sqrt{1+\frac{1}{\nu}} - \frac{1}{2}$. Then*

$$f(\lambda) := \frac{1}{\lambda}\Big[L - \frac{T}{2}\log\big(1-4(\lambda+\lambda^2)\nu\big)\Big] \tag{93}$$

is a smooth strictly convex function on the interval $(0, r_{4\nu})$.

Proof. The proof is analogous to Lemma B.5, since neither the switch from ν to 4ν nor the multiplication by $T/2$ does change the convexity properties of the function. □

B.3 Proof of Theorem 2

We first prove the linear bound in Theorem 7, then strengthen it in the strong privacy regime in Theorem 8. Together, these two theorems prove Theorem 2.

Theorem 7 (Max-Information bound for RDP)**.** *Suppose the dataset $S = (X_1,\dots,X_n) \sim \mu^n$ is drawn with i.i.d. entries, and $\mathcal{A}$ is an (α,ϵ)-RDP mechanism for some $\alpha > 1$ and $\epsilon \geq 0$. Then, for every $\beta \in (0,1)$, we have*

$$I_\infty^\beta(\mathcal{A}(S),S) \leq n\left(\epsilon + \frac{\log(1/\beta)}{\alpha-1}\right). \tag{94}$$

Proof. We begin by recalling the Hellinger integral, which is jointly convex in its two arguments (van Erven & Harremoës, 2014):

$$H_\alpha(P\|Q) = \exp\big((\alpha-1)D_\alpha(P\|Q)\big) = \int (p)^\alpha (q)^{1-\alpha}. \tag{95}$$

Define a sequence of random variables $Y_i : \mathcal{X}^i \to \mathcal{Y}$ which take the first i points as input, and draw the remaining $n - i$ from μ i.i.d.:

$$Y_i(x_{1:i}) = \mathcal{A}(S) \mid X_1 = x_1, \ldots, X_i = x_i, \qquad S = (X_1, \ldots, X_n) \sim \mu^n. \tag{96}$$

Pairs of these random variables satisfy two properties concerning their Rényi divergences. Firstly,

$$D_\alpha(Y_i(x_{1:i-1}, x_i) \| Y_i(x_{1:i-1}, x_i')) \le \epsilon, \tag{97}$$

which holds by the joint convexity of H_α, and the fact that $(x_{1:i-1}, x_i, x_{i+1:n})$ and $(x_{1:i-1}, x_i', x_{i+1:n})$ are neighboring datasets for every suffix $x_{i+1:n}$.

Secondly, incremental steps satisfy the same property:

$$D_\alpha(Y_i(x_{1:i-1}, x_i) \| Y_{i-1}(x_{1:i-1})) \le \epsilon, \tag{98}$$

because $Y_{i-1}(x_{1:i-1})$ is the mixture of $Y_i(x_{1:i-1}, x_i')$ over $x_i' \sim \mu$, and H_α is convex in its second argument.

Consider the information-density increment obtained by conditioning on one extra data point:

$$\Lambda_i = \log \frac{p_{Y_i}(Y \mid X_{1:i})}{p_{Y_{i-1}}(Y \mid X_{1:i-1})}; \qquad S = (X_1, \ldots, X_n) \sim \mu^n, \quad Y = \mathcal{A}(S). \tag{99}$$

The (α, ϵ) guarantee controls the Hellinger integral at the order α:

$$\begin{aligned} \mathbb{E} \exp\big((\alpha - 1)\Lambda_i\big) &= \mathbb{E}\left[\mathbb{E}\left[\exp\big((\alpha - 1)\Lambda_i\big) \mid X_{1:i}\right]\right] \\ &= \mathbb{E}\left[H_\alpha(Y_i(X_{1:i}) \| Y_{i-1}(X_{1:i-1}))\right] \\ &\le \mathbb{E}\left[\exp\big((\alpha - 1)\epsilon\big)\right] \\ &= \exp\big((\alpha - 1)\epsilon\big). \end{aligned} \tag{100}$$

The information density for $\mathcal{A}$ is

$$\imath(S, Y) = \log \frac{p_{Y|S}(Y \mid S)}{p_Y(Y)} = \sum_{i=1}^n \Lambda_i \tag{101}$$

Set $\lambda := \frac{\alpha - 1}{n}$, so that $n\lambda = \alpha - 1$, and bound the moments of $\imath(S, Y)$ using the generalized Hölder's inequality:

$$\mathbb{E} \exp\left(\lambda \imath(S, Y)\right) = \mathbb{E} \prod_{i=1}^n \exp\left(\lambda \Lambda_i\right) \le \prod_{i=1}^n \left(\mathbb{E} \exp\left(n\lambda \Lambda_i\right)\right)^{1/n} \tag{102}$$

$$= \prod_{i=1}^n \left(\mathbb{E} \exp\big((\alpha - 1)\Lambda_i\big)\right)^{1/n} \le \exp\big((\alpha - 1)\epsilon\big), \tag{103}$$

where the last inequality is due to (100). We set

$$\kappa := n\epsilon + \frac{n}{\alpha - 1} \log \frac{1}{\beta}, \tag{104}$$

and apply Chernoff's inequality with the same $\lambda = \frac{\alpha - 1}{n}$:

$$\mathbb{P}(\imath(S, Y) > \kappa) \le \exp(-\lambda\kappa)\, \mathbb{E} \exp(\lambda \imath(S, Y)) \le \exp\big(-\lambda\kappa + (\alpha - 1)\epsilon\big) = \beta. \tag{105}$$

An application of Lemma 18 from Dwork et al. (2015) completes the proof of eq. (94).

□

Theorem 8. *Assume that $\mathcal{A}$ is (α, ϵ)-RDP for some $\alpha > 1$ and $\epsilon \ge 0$, and S is a dataset of n i.i.d. samples. Fix $\beta_1, \beta_2 > 0$ with $\beta_1 + \beta_2 = \beta < 1$, write $L_2 := \log(1/\beta_2)$, and define*

$$\psi(u) := \log^2\big(\sqrt{u} + \sqrt{1 + u}\big) = \operatorname{arcsinh}^2 \sqrt{u}, \qquad v_\star := \frac{n}{2(\alpha - 1)^2}\, \psi\left(\frac{e^{(\alpha-1)\epsilon} - 1}{\beta_1}\right). \tag{106}$$

Then

$$I_\infty^\beta(\mathcal{A}(S), S) \le v_\star + 2\sqrt{v_\star L_2} + \frac{L_2}{\alpha - 1}. \tag{107}$$

Moreover,

$$v_\star \le v := \frac{n}{8}\left(\epsilon + \frac{\log(4/\beta_1)}{\alpha - 1}\right)^2, \tag{108}$$

and the right-hand side of (107) *is increasing in* $v_\star$*, so* (107) *remains valid with* v *in place of* $v_\star$*.*

The proof is organized as follows. Lemma B.10 bounds, for a single coordinate, the symmetrized Rényi divergence between the posterior of X_i and the prior; this is the only place the privacy assumption is used. Lemma B.11 then composes the n coordinates into one moment bound for the information density. The proof of the theorem combines both lemmas, excluding a bad event.

Setup and notation Throughout, $Y = \mathcal{A}(S)$, and we write

$$P := P_{S,Y}, \qquad Q := P_S \otimes P_Y = \mu^n \otimes P_Y \tag{109}$$

for the joint distribution of the data and the output and for the product of its marginals, with ψ and $v_\star$ as in (106).

As in the (ϵ, δ)-DP proof, let

$$\mathcal{F}_i := \sigma(Y, X_1, \dots, X_i), \qquad \mathcal{F}_0 := \sigma(Y), \qquad P_i := P_{X_{\le i}, Y}, \qquad P_0 := P_Y. \tag{110}$$

For a fixed history $(y, x_{<i})$, define the posterior-to-prior ratio for datapoint x_i:

$$r_i^{y, x_{<i}}(x) := \frac{dP_i}{d(P_{i-1} \otimes \mu)}(x_{<i}, x, y) = \frac{dP_{X_i | Y=y, X_{<i}=x_{<i}}}{d\mu}(x). \tag{111}$$

When the history is random, write

$$r_i(x) := r_i^{Y, X_{<i}}(x). \tag{112}$$

Thus $r_i(\cdot)$ depends only on $\mathcal{F}_{i-1}$, and

$$r_i(x) > 0 \text{ a.e.}, \qquad \int r_i(x)\mu(dx) = 1. \tag{113}$$

As before, write $H_\alpha(U\|V) := \exp\{(\alpha - 1)D_\alpha(U\|V)\}$ for the Hellinger integral of order α, and, for two probability measures U, V, write

$$\bar{D}_\alpha(U, V) := D_\alpha(U\|V) + D_\alpha(V\|U) \tag{114}$$

for the *symmetrized Rényi divergence* of order α, the analogue of the Jeffreys divergence. A similar construction is given by Talih (2013). For a positive density w with respect to μ, let W be the probability measure with $dW = w\, d\mu$ and put

$$J(w) := \exp\{(\alpha - 1)\bar{D}_\alpha(W, \mu)\} = H_\alpha(W\|\mu)\, H_\alpha(\mu\|W) = \left(\int w^\alpha\, d\mu\right)\left(\int w^{1-\alpha}\, d\mu\right), \tag{115}$$

and $J_i := J(r_i)$. The last equality is immediate from $d\mu/dW = 1/w$. Since the Rényi divergence is nonnegative, $J_i \ge 1$ and each J_i is $\mathcal{F}_{i-1}$-measurable.

The chain rule for the posterior distribution gives

$$\frac{dP}{dQ}(s, y) = \frac{dP_{S|Y=y}}{d\mu^n}(s) = \prod_{i=1}^n r_i^{y, x_{<i}}(x_i). \tag{116}$$

Consequently, for the information density we have:

$$\imath(S, Y) := \log \frac{dP}{dQ}(S, Y) = \sum_{i=1}^n \log r_i(X_i). \tag{117}$$

Symmetrized Rényi divergence of a single coordinate

Lemma B.10. *For every $i \in \{1, \dots, n\}$,*

$$\mathbb{E}_P J_i \le e^{(\alpha-1)\epsilon}. \tag{118}$$

Proof. Fix a coordinate i and a prefix $u = x_{<i}$, and write

$$P_x^u := P_{Y|X_{<i}=u,\, X_i=x} = \int \mathcal{A}\big((u,x,z),\cdot\big)\, \mu^{n-i}(dz), \qquad P^u := \int P_x^u\, \mu(dx), \tag{119}$$

for the conditional distribution of the output given the prefix and the current observation P_x^u, and given the prefix alone, P^u.

By the same logic as in the simpler result, the datasets (u, x, z) and (u, x', z) are neighbors for any fixed suffix z, so

$$H_\alpha(P_x^u \| P_{x'}^u) \le e^{(\alpha-1)\epsilon} \qquad \text{for all } x, x'. \tag{120}$$

holds using the joint convexity of the Hellinger integral and the privacy guarantee on the mechanism.

The posterior-to-prior ratio $r_i^{y,u}(x)$ (111) is a density in $\mathcal{X}$ for fixed y, whereas (120) is a bound on densities in $\mathcal{Y}$ for fixed x. We can convert between these two functions using Bayes' rule, showing that (almost everywhere),

$$r_i^{y,u}(x) = \frac{dP_x^u}{dP^u}(y). \tag{121}$$

Now work conditionally on $X_{<i} = u$, so that the output is distributed as P^u. Substituting (121) into the definition (115) of J_i and exchanging the nonnegative integrals by Tonelli's theorem,

$$\begin{aligned}\mathbb{E}_P[J_i \mid X_{<i} = u] &= \iint \left[\int \left(\frac{dP_x^u}{dP^u}\right)^{\alpha} \left(\frac{dP_{x'}^u}{dP^u}\right)^{1-\alpha} dP^u\right] \mu(dx)\mu(dx') \\ &= \iint H_\alpha(P_x^u \| P_{x'}^u)\, \mu(dx)\mu(dx') \;\le\; e^{(\alpha-1)\epsilon},\end{aligned} \tag{122}$$

the last step by (120). Averaging over the prefix proves (118). □

Moment bound on information density Lemma B.10 controls each coordinate in expectation. The argument requires a moment bound on the information density $\imath(S, Y)$. We show this in terms of variables v_i, defined using ψ as in (106) :

$$v_i := \frac{\psi(J_i - 1)}{2(\alpha-1)^2}, \tag{123}$$

Lemma B.11. *For every $0 < \lambda < \alpha - 1$, writing $b_\lambda := \lambda + \frac{\lambda^2}{1-\lambda/(\alpha-1)}$,*

$$\mathbb{E}_P \exp\left\{\lambda \imath(S,Y) - b_\lambda \sum_{i=1}^n v_i\right\} \le 1. \tag{124}$$

Proof. Fix i and a history $(y, x_{<i})$, and suppress the history in the notation r_i. By (113), r_i is a probability density with respect to μ, positive almost everywhere, and $J(r_i) = J_i$ by definition. Lemma B.16 applied with $w = r_i$ and $M = J_i$, so that $v_w = v_i$, gives

$$\log \int r_i(x)^{1+\lambda} \mu(dx) \le v_i \left(\lambda + \frac{\lambda^2}{1-\lambda/(\alpha-1)}\right). \tag{125}$$

Hence, at every fixed history,

$$\int r_i(x)^{1+\lambda} e^{-b_\lambda v_i} \mu(dx) \le 1. \tag{126}$$

We now integrate the coordinates backwards. Changing measure with $dP/dQ = e^{\imath(S,Y)}$ and using the factorisation and decomposition (116), the left-hand side of (124) equals

$$\int \prod_{i=1}^n \left[r_i(X_i)^{1+\lambda} e^{-b_\lambda v_i}\right] dQ. \tag{127}$$

Under Q, hold $Y = y$ fixed and integrate first over X_n. Every earlier factor is independent of x_n, and v_n depends only on $(y, x_{<n})$. Hence the integral over x_n is

$$\left(\prod_{i=1}^{n-1} r_i(x_i)^{1+\lambda} e^{-b_\lambda v_i}\right) \int r_n(x_n)^{1+\lambda} e^{-b_\lambda v_n} \mu(dx_n), \tag{128}$$

and the last factor is at most one by (126). After replacing it by one, the remaining integral has exactly the same form with $n-1$ factors, and so we repeat the for the remaining coordinates. This proves that (127) is at most one, and integrating over P_Y gives the final result. All integrals can be interchanged as the integrands are non-negative. □

Proof of Theorem 8 Using the two key lemmas above, we prove the main theorem statement.

Bounding $\sum_i v_i$. The bound (124) is useful only where $\sum_i v_i$ is small, so let

$$B := \left\{\sum_{i=1}^{n} v_i > v_\star\right\}. \tag{129}$$

Because ψ is increasing and concave (Lemma B.12), Jensen's inequality gives

$$\sum_{i=1}^{n} v_i = \frac{1}{2(\alpha-1)^2}\sum_{i=1}^{n}\psi(J_i - 1) \le \frac{n}{2(\alpha-1)^2}\psi\left(\frac{1}{n}\sum_{i=1}^{n}(J_i-1)\right), \tag{130}$$

so B is contained in the event that $\frac{1}{n}\sum_i (J_i - 1)$ exceeds $(e^{(\alpha-1)\epsilon} - 1)/\beta_1$. Since $J_i \ge 1$, Lemma B.10 gives $\mathbb{E}_P\left[\frac{1}{n}\sum_i (J_i - 1)\right] \le e^{(\alpha-1)\epsilon} - 1$, so Markov's inequality yields

$$P(B) \le \beta_1. \tag{131}$$

If $\epsilon = 0$ then $J_i = 1$ almost surely for every i, so B is empty and (131) holds trivially.

A tail bound for the information density. We split the tail bound into two parts:

$$P\{\imath(S,Y) > \kappa\} \le P(B) + P\big(B^c \cap \{\imath(S,Y) > \kappa\}\big) \le \beta_1 + P\big(B^c \cap \{\imath(S,Y) > \kappa\}\big). \tag{132}$$

It therefore suffices to choose κ making the second term at most β_2.

Fix $0 < \lambda < \alpha - 1$. A Chernoff inequality bounds that term by $e^{-\lambda\kappa}\,\mathbb{E}_P[e^{\lambda\imath(S,Y)}\mathbf{1}_{B^c}]$. On B^c we have that $\sum_i v_i \le v_\star$, so Lemma B.11 gives

$$\mathbb{E}_P[e^{\lambda\imath(S,Y)}\mathbf{1}_{B^c}] \le e^{b_\lambda v_\star}\mathbb{E}_P \exp\left\{\lambda\imath(S,Y) - b_\lambda\sum_{i=1}^{n} v_i\right\} \le e^{b_\lambda v_\star}. \tag{133}$$

Hence $P(B^c \cap \{\imath(S,Y) > \kappa\}) \le e^{b_\lambda v_\star - \lambda\kappa}$, which is at most β_2 exactly when

$$\kappa \ge \kappa_\lambda := \frac{b_\lambda v_\star + L_2}{\lambda}, \qquad L_2 = \log(1/\beta_2). \tag{134}$$

Taking $\kappa = \kappa_\lambda$ in the display above,

$$P\{\imath(S,Y) > \kappa_\lambda\} \le \beta_1 + \beta_2 = \beta, \qquad 0 < \lambda < \alpha - 1. \tag{135}$$

Conclusion. It remains only to choose λ, which is the routine optimization carried out in Lemma B.13: the resulting value of $\inf_{0<\lambda<\alpha-1}\kappa_\lambda$ is exactly the right-hand side of (107) with $v_\star$ in place of v. Applying (135) at that λ and converting the resulting tail bound on the information density into a max-information bound (Dwork et al., 2015, Lemma 18) therefore proves $I_\infty^\beta(Y,S) \le v_\star + 2\sqrt{v_\star L_2} + L_2/(\alpha-1)$, which is (107) with $v_\star$ in place of v. The remaining claim $v_\star \le v$ is Lemma B.12, and the right-hand side of (107) is increasing in $v_\star$. □

B.3.1 AUXILIARY LEMMAS

Lemma B.12. *The function $\psi(u) = \log^2\big(\sqrt{u} + \sqrt{1+u}\big)$ is increasing and concave on $[0,\infty)$ and satisfies $\psi(u) \le \frac{1}{4}\log^2(4(1+u))$. Consequently $v_\star \le v$, for $v_\star$ as in* (106) *and v as in* (108).

Proof. Differentiating gives $\psi'(u) = \int_0^1 [(1+u)(1+ut^2)]^{-1/2}\,dt$, whose integrand is nonnegative and decreasing in u; hence ψ is increasing and concave. Next, $\sqrt{u} + \sqrt{1+u} \le 2\sqrt{1+u}$ gives $\psi(u) \le \frac{1}{4}\log^2(4(1+u))$. Finally $\beta_1 \le 1$ gives $1 + (e^{(\alpha-1)\epsilon} - 1)/\beta_1 \le e^{(\alpha-1)\epsilon}/\beta_1$, so

$$v_\star \le \frac{n}{8(\alpha-1)^2}\log^2\frac{4e^{(\alpha-1)\epsilon}}{\beta_1} = \frac{n}{8}\left(\epsilon + \frac{\log(4/\beta_1)}{\alpha-1}\right)^2 = v. \tag{136}$$

□

Lemma B.13. *Let $v_\star \ge 0$, and let κ_λ be as in* (134)*. Then*

$$\inf_{0<\lambda<\alpha-1} \kappa_\lambda = v_\star + 2\sqrt{v_\star L_2} + \frac{L_2}{\alpha-1}. \tag{137}$$

Proof. This is elementary calculus. Since $b_\lambda/\lambda = 1 + \lambda/(1 - \lambda/(\alpha-1))$,

$$\kappa_\lambda = v_\star + \frac{(\alpha-1)v_\star\lambda}{(\alpha-1)-\lambda} + \frac{L_2}{\lambda}, \tag{138}$$

which is convex on $(0, \alpha-1)$ with derivative

$$\frac{(\alpha-1)^2 v_\star}{\big((\alpha-1)-\lambda\big)^2} - \frac{L_2}{\lambda^2}. \tag{139}$$

For $v_\star > 0$ this vanishes exactly when $(\alpha-1)\lambda\sqrt{v_\star} = \sqrt{L_2}\,\big((\alpha-1)-\lambda\big)$, that is at

$$\lambda_* = \frac{(\alpha-1)\sqrt{L_2}}{(\alpha-1)\sqrt{v_\star} + \sqrt{L_2}}, \qquad (\alpha-1) - \lambda_* = \frac{(\alpha-1)^2\sqrt{v_\star}}{(\alpha-1)\sqrt{v_\star} + \sqrt{L_2}}. \tag{140}$$

Substituting, the second and third terms of κ_{λ_*} each contribute $\sqrt{v_\star L_2}$, the third leaving an additional $L_2/(\alpha-1)$; this is (137). For $v_\star = 0$ the second term vanishes identically and $L_2/\lambda \downarrow L_2/(\alpha-1)$ as $\lambda \uparrow \alpha - 1$. □

Lemma B.14 (Hoeffding-type variance bound)**.** *Let V be a real random variable, let $h > 0$, and suppose $A := \mathbb{E}e^{hV} < \infty$ and $B := \mathbb{E}e^{-hV} < \infty$. Then $AB \ge 1$, and, with ψ as in* (106)*,*

$$\mathrm{Var}(V) \le \frac{\psi(AB-1)}{h^2}. \tag{141}$$

This is a two-sided analogue of Hoeffding's lemma: rather than bounding the variance of a *bounded* random variable, it bounds the variance of V from a pair of exponential moments, with no boundedness assumption.

Proof. Cauchy–Schwarz gives $AB \ge 1$. Choose $a = (2h)^{-1}\log(A/B)$. Then

$$\mathbb{E}e^{h(V-a)} = \mathbb{E}e^{-h(V-a)} = \sqrt{AB}. \tag{142}$$

The function

$$q(u) := 2\cosh(h\sqrt{u}) = 2\sum_{k=0}^{\infty}\frac{h^{2k}u^k}{(2k)!}, \qquad u \ge 0, \tag{143}$$

is increasing and convex. Therefore Jensen's inequality gives

$$2\sqrt{AB} = \mathbb{E}q((V-a)^2) \ge q(\mathbb{E}(V-a)^2) \ge q(\mathrm{Var}(V)). \tag{144}$$

Writing $\sigma = \sqrt{\mathrm{Var}(V)}$, this reads $e^{h\sigma} + e^{-h\sigma} \le 2\sqrt{AB}$. Solving this quadratic inequality for $e^{h\sigma} \ge 1$ yields $e^{h\sigma} \le \sqrt{AB} + \sqrt{AB-1}$; taking logarithms and squaring proves (141). □

Lemma B.15 (Curvature of the log-Hellinger integral)**.** *Let $\alpha > 1$, let w be a probability density with respect to μ that is positive almost everywhere, let W be the probability measure with $dW = w\,d\mu$, and suppose $J(w) \le M < \infty$, with J as in* (115)*. Write $Z := \log w$, viewed as a random variable under W, and*

$$g(t) := \log \mathbb{E}_W e^{tZ} = \log \int w^{1+t}\,d\mu = \log H_{1+t}(W\|\mu), \qquad -\alpha \le t \le \alpha - 1, \tag{145}$$

the logarithm of the Hellinger integral as a function of its order, equivalently $g(t) = t\,D_{1+t}(W\|\mu)$. Then $M \ge 1$,

$$g(-1) = g(0) = 0, \qquad g'(0) = \mathrm{KL}(W\|\mu), \qquad g(\lambda) = \lambda D_{1+\lambda}(W\|\mu)\ (0 < \lambda < \alpha - 1), \tag{146}$$

and, with ψ as in (106)*,*

$$g''(t) \le \begin{cases} \psi(M-1)/(\alpha-1)^2, & -1 \le t \le 0, \\ \psi(M-1)/(\alpha-1-t)^2, & 0 \le t < \alpha - 1. \end{cases} \tag{147}$$

Proof. By (115) and $J(w) < \infty$, both $\mathbb{E}_W e^{-\alpha Z}$ and $\mathbb{E}_W e^{(\alpha-1)Z}$ are finite. Hölder's inequality makes $g(t) = \log \mathbb{E}_W e^{tZ}$ convex between these endpoints. On each compact subinterval of the interior, for every integer $k \ge 0$, $|Z|^k e^{tZ}$ is bounded by a constant times $e^{-\alpha Z} + e^{(\alpha-1)Z}$. Differentiation under the expectation therefore shows that g is smooth in the interior. It also satisfies

$$g(-1) = g(0) = 0, \qquad g(\alpha - 1) + g(-\alpha) = \log J(w) \le \log M, \tag{148}$$

the first two because $w > 0$ almost everywhere and $\int w\,d\mu = 1$, the third by the definition (115) of J. Convexity implies $g(\alpha-1), g(-\alpha) \ge 0$, hence $M \ge 1$. Two further identifications will be used: $g'(0) = \mathbb{E}_W \log w = \mathrm{KL}(W\|\mu)$, which is the statement that $D_\alpha \to \mathrm{KL}$ as $\alpha \to 1$ (van Erven & Harremoës, 2014, Theorem 5), and $g(\lambda) = \lambda D_{1+\lambda}(W\|\mu)$. This proves (146). Under the distribution tilted by $e^{tZ - g(t)}$,

$$g''(t) = \mathrm{Var}_t(Z). \tag{149}$$

We bound this variance by applying Lemma B.14 to Z under the tilted measure, with a suitable choice of h in each of the two ranges of t.

Under the distribution tilted by $e^{tZ-g(t)}$,

$$\mathbb{E}_t e^{hZ}\,\mathbb{E}_t e^{-hZ} = \exp\{g(t+h) + g(t-h) - 2g(t)\}. \tag{150}$$

For $-1 \le t \le 0$, take $h = \alpha - 1$. Increments of a convex function over a fixed interval length are nondecreasing in their starting point, so

$$\begin{aligned} g(t + \alpha - 1) - g(t) &\le g(\alpha - 1) - g(0) = g(\alpha - 1), \\ g(t - \alpha + 1) - g(t) &\le g(-\alpha) - g(-1) = g(-\alpha). \end{aligned}$$

Their sum is at most $\log M$.

For $0 \le t < \alpha - 1$, take $h = \alpha - 1 - t$. Then $t + h = \alpha - 1$ and $t - h = 2t - \alpha + 1 \in [-\alpha, t]$. A convex function attains its maximum on this interval at an endpoint, and $g(t), g(-\alpha) \ge 0$; hence

$$\begin{aligned} g(t-h) - g(t) &\le \max\{g(-\alpha), g(t)\} - g(t) \le g(-\alpha), \\ g(t+h) - g(t) &= g(\alpha - 1) - g(t) \le g(\alpha - 1). \end{aligned}$$

Again the exponent in (150) is at most $\log M$, so in both cases the two tilted exponential moments $A = \mathbb{E}_t e^{hZ}$, $B = \mathbb{E}_t e^{-hZ}$ satisfy $AB \le M$. By Lemma B.14, monotonicity of ψ (Lemma B.12), and (149),

$$g''(t) = \mathrm{Var}_t(Z) \le \frac{\psi(AB - 1)}{h^2} \le \frac{\psi(M-1)}{h^2}, \tag{151}$$

which is (147) once h is replaced by $\alpha - 1$ or by $\alpha - 1 - t$ in the two respective ranges. □

Lemma B.16. *Let $\alpha > 1$, let w be a probability density with respect to μ that is positive almost everywhere, and let W be the probability measure with $dW = w\,d\mu$. If $J(w) \le M < \infty$, with J as in* (115)*, then $M \ge 1$. Put*

$$v_w := \frac{\psi(M-1)}{2(\alpha-1)^2}. \tag{152}$$

The Kullback–Leibler divergence satisfies

$$0 \le \mathrm{KL}(W\|\mu) \le v_w, \tag{153}$$

and for every $0 < \lambda < \alpha - 1$ *the Rényi divergence exceeds it by at most*

$$\lambda\big[D_{1+\lambda}(W\|\mu) - \mathrm{KL}(W\|\mu)\big] \le \psi(M-1)\left[-\log(1-\lambda/(\alpha-1)) - \lambda/(\alpha-1)\right] \le \frac{v_w\lambda^2}{1-\lambda/(\alpha-1)}. \tag{154}$$

Consequently,

$$\log \int w^{1+\lambda}\, d\mu = \lambda D_{1+\lambda}(W\|\mu) \le v_w\left(\lambda + \frac{\lambda^2}{1-\lambda/(\alpha-1)}\right). \tag{155}$$

The middle expression in (154) is the centred cumulant generating function of a Gamma variable, and the second inequality is the usual relaxation of it: writing $g(t) := \log \int w^{1+t} d\mu$ for the log-Hellinger integral as a function of its order, (154) states that g is *sub-gamma* with variance factor $2v_w$ and scale $1/(\alpha-1)$, in the sense of Boucheron et al. (2013, §2.4). Three cases are worth separating. A one-sided Rényi bound gives only a bound on $g(\alpha-1)$, and hence subexponential tails with no variance factor, which is what Theorem 7 uses; requiring the bound at every order, as zCDP does, gives subgaussian tails (Bun & Steinke, 2016). A single order made two-sided sits between the two: it yields a variance factor v_w that vanishes as $M \to 1$, and this is what turns the linear dependence on n into $\sqrt{n}$. The function $\psi = \mathrm{arcsinh}^2\sqrt{\cdot}$ belongs to the same family as the $\mathrm{arcsinh}$ in Prokhorov's exponential inequality for bounded summands (Prokhorov, 1959).

Proof. Apply Lemma B.15 to w, with g and Z as there; this gives $M \ge 1$, the identifications (146), and the curvature bound (147). It remains to integrate the latter.

Because $g(-1) = g(0) = 0$, integration by parts gives

$$\mathrm{KL}(W\|\mu) = g'(0) = \int_{-1}^{0} (t+1)g''(t)\, dt \le \frac{\psi(M-1)}{2(\alpha-1)^2} = v_w. \tag{156}$$

The lower bound in (153) is the nonnegativity of the Kullback–Leibler divergence.

For $0 < \lambda < \alpha - 1$, Taylor's formula with integral remainder and (147) give

$$\begin{aligned}
\lambda\big[D_{1+\lambda}(W\|\mu) - \mathrm{KL}(W\|\mu)\big] &= g(\lambda) - \lambda g'(0) \\
&= \int_0^\lambda (\lambda - t)g''(t)\, dt \\
&\le \psi(M-1)\int_0^\lambda \frac{\lambda - t}{(\alpha - 1 - t)^2}\, dt \\
&= \psi(M-1)\left[-\log(1-\lambda/(\alpha-1)) - \lambda/(\alpha-1)\right].
\end{aligned}$$

For $0 \le u < 1$,

$$-\log(1-u) - u = \sum_{k=2}^{\infty} \frac{u^k}{k} \le \frac{u^2}{2(1-u)}. \tag{157}$$

Substituting $u = \lambda/(\alpha-1)$ proves (154). Adding $\lambda\mathrm{KL}(W\|\mu) \le \lambda v_w$ to both sides gives $g(\lambda) \le v_w(\lambda + \lambda^2/(1-\lambda/(\alpha-1)))$, which is (155). □

B.4 Proof of Theorem 1

In this section, we prove Theorem 1. First, we state a stronger version of the upper bound for the $\beta \ge n\delta$ case, and then we provide the counterexample for $\beta < n\delta$. Combining these two results prove Theorem 1.

Theorem 9. *Assume that $\mathcal{A}$ is (ϵ, δ)-DP for some $\epsilon > 0$ and $\delta \in [0,1]$, and S is a dataset of n i.i.d. samples. For every $\beta > n\delta$, we have,*

$$I_\infty^\beta(Y,S) \le \min\left\{ n\epsilon, \min_{\lambda>0}\left\{ \frac{n(1+\lambda)\epsilon^2}{8} + \frac{\log(1/\beta_0)}{\lambda} + \frac{n\delta}{\beta_0}c_\lambda \right\}, \right\}, \tag{158}$$

where $c_\lambda := e^{2\epsilon} + \epsilon - 1 + e^{2\lambda\epsilon}$. *In particular, with*

$$\lambda_* := \sqrt{\frac{8\log(1/\beta_0)}{n\epsilon^2}}, \tag{159}$$

we obtain the upper bound stated in Theorem 1:

$$I_\infty^\beta(Y,S) \le \min\left\{n\epsilon, \frac{1}{8}n\epsilon^2 + \epsilon\sqrt{\frac{n}{2}\log\frac{1}{\beta_0}} + \frac{n\delta}{\beta_0}c_{\lambda^*}\right\}. \tag{160}$$

Proof of Theorem 9: Let

$$\mathcal{F}_i := \sigma(Y, X_1, \dots, X_i), \qquad \mathcal{F}_0 := \sigma(Y), \qquad P_i := P_{X_{\le i},Y}, \qquad P_0 := P_Y. \tag{161}$$

For finite measures U, V, their common part $U \wedge V$ has the smaller of their densities relative to any common dominating measure. For a fixed history $(y, x_{<i})$, define the capped posterior-to-prior ratio

$$r_i^{y,x_{<i}}(x) := \frac{d[P_i \wedge e^\epsilon(P_{i-1}\otimes\mu)]}{d(P_{i-1}\otimes\mu)}(x_{<i}, x, y). \tag{162}$$

This retains the absolutely continuous part of the posterior and caps its density at e^ϵ. Choose measurable versions, zero on exceptional histories, so that every fixed-history function satisfies

$$0 \le r_i^{y,x_{<i}} \le e^\epsilon, \qquad \int r_i^{y,x_{<i}}(x)\,\mu(dx) \le 1. \tag{163}$$

When the history is random, write

$$r_i(x) := r_i^{Y,X_{<i}}(x). \tag{164}$$

This random function depends only on $\mathcal{F}_{i-1}$.

The Retained Measure and Retained Information Density. Starting with $\widetilde{P}_0 = P_Y$, retain the capped conditional kernel at each coordinate:

$$\widetilde{P}_i(dx_{\le i}, dy) := \widetilde{P}_{i-1}(dx_{<i}, dy) r_i^{y,x_{<i}}(x_i)\,\mu(dx_i), \qquad \widetilde{P} := \widetilde{P}_n. \tag{165}$$

Multiplying these kernels and dividing by $Q = \mu^{\otimes n} \otimes P_Y$ gives

$$\frac{d\widetilde{P}}{dQ}(x_1^n, y) = \prod_{i=1}^n r_i^{y,x_{<i}}(x_i). \tag{166}$$

Consequently the retained information density is

$$\widetilde{\imath}(S,Y) := \log\frac{d\widetilde{P}}{dQ}(S,Y) = \sum_{i=1}^n \log r_i(X_i) \qquad \widetilde{P}\text{-almost everywhere.} \tag{167}$$

Privacy Excess For any measurable w with $0 \le w \le e^\epsilon$ and $\int w\,d\mu \le 1$, define its privacy excess by

$$\Delta_\epsilon(w) := \iint \big(w(x) - e^\epsilon w(x')\big)_+ \mu(dx)\mu(dx'), \qquad \Delta_i := \Delta_\epsilon(r_i), \tag{168}$$

where $u_+ := \max\{u, 0\}$. The quantity $\Delta_\epsilon(w)$ measures how much the candidate ratios exceed the pure-privacy factor e^ϵ. In particular, Δ_i depends only on $\mathcal{F}_{i-1}$.

The measure $\widetilde{P}$ is not a probability measure and is not normalized, the capping discards some mass, however, in the following lemma we show that $\widetilde{P} \le P$ and $\widetilde{P}(\Omega) \ge 1 - n\delta$. Additionally, we show that for all i, the capped posterior-to-prior ratio r_i has expected privacy excess at most δ.

Lemma B.17. *The retained measure satisfies*

$$\widetilde{P} \le P, \qquad \widetilde{P}(\Omega) \ge 1 - n\delta, \qquad \mathbb{E}_P\Delta_i \le \delta \quad (i = 1, \dots, n). \tag{169}$$

Here Ω *is the whole joint space.*

Proof. Fix a coordinate i and a prefix $u = x_{<i}$. For any measurable set B of outputs, define

$$K_x(B) := \int \mathcal{A}((u, x, v), B)\, \mu^{\otimes(n-i)}(dv), \qquad \bar{K}(B) := \int K_x(B)\, \mu(dx). \tag{170}$$

Thus $K_x(B)$ is the probability of $Y \in B$ when the prefix is u and the current observation is x, averaging over the remaining observations and the algorithm's randomness. The law $\bar{K}$ also averages over the current observation; it is the output law given only the prefix.

Because X_i has prior law μ given the prefix, Bayes' rule is the following identity for the joint law of (X_i, Y):

$$\mu(dx)K_x(dy) = \bar{K}(dy)\, P(X_i \in dx \mid Y = y, X_{<i} = u). \tag{171}$$

Let k_x be the density of the absolutely continuous part of K_x relative to $\bar{K}$. Taking densities in this identity gives

$$\frac{d(P_{X_i|Y=y,X_{<i}=u})_{\mathrm{ac}}}{d\mu}(x) = k_x(y) \qquad (\mu \otimes \bar{K})\text{-almost everywhere.} \tag{172}$$

The subscript ac means the part absolutely continuous with respect to μ. Thus $k_x(y)$ is also the posterior-to-prior ratio for the observation x. The retained ratio r_i caps it:

$$r_i^{y,u}(x) = \min\{k_x(y), e^\epsilon\}. \tag{173}$$

Equivalently, if we clip the output law by setting

$$\widehat{K}_x := K_x \wedge e^\epsilon \bar{K}, \qquad \widehat{K}_x(dy) = r_i^{y,u}(x)\, \bar{K}(dy), \tag{174}$$

we obtain the same retained ratio as in (162). These identities hold almost everywhere and cover singular output laws: the common part discards mass singular to $\bar{K}$.

Privacy and averaging over x' give, for every output event B,

$$K_x(B) \le e^\epsilon K_{x'}(B) + \delta, \qquad K_x(B) \le e^\epsilon \bar{K}(B) + \delta. \tag{175}$$

We use these two inequalities to control the lost mass and the excess.

The measure $\widetilde{P}$ losses at most $n\delta$ mass: The mass removed by the common part is the largest excess over an output event. Hence

$$1 - \widehat{K}_x(\mathcal{Y}) = \sup_B \{K_x(B) - e^\epsilon \bar{K}(B)\} \le \delta, \tag{176}$$

where $\mathcal{Y}$ is the output space. Average over x, then over the prefix, to obtain

$$\int \left[1 - \int r_i^{y,u}(x)\, \mu(dx)\right] P_{i-1}(du, dy) \le \delta. \tag{177}$$

Each retained conditional kernel is dominated by the original one, so induction gives $\widetilde{P}_i \le P_i$. The loss at step i is the left side of (177) integrated against $\widetilde{P}_{i-1}$ instead of P_{i-1}. The integrand is nonnegative, so this loss is also at most δ. Summing the n losses gives $\widetilde{P}(\Omega) \ge 1 - n\delta$, and the domination gives $\widetilde{P} \le P$.

The expected excess is at most δ: Applying the same cap to two measures cannot increase their privacy excess. To see this, take densities a, b, h of $K_x, K_{x'}, e^\epsilon \bar{K}$ relative to their sum. Pointwise,

$$\big(\min\{a, h\} - e^\epsilon \min\{b, h\}\big)_+ \le (a - e^\epsilon b)_+. \tag{178}$$

Indeed, a positive left side forces $b < h$, so the second density is unchanged by the cap, while the first can only decrease. Integrating and using privacy therefore gives

$$\int \big(r_i^{y,u}(x) - e^\epsilon r_i^{y,u}(x')\big)_+ \bar{K}(dy) \le \sup_B \{K_x(B) - e^\epsilon K_{x'}(B)\} \le \delta. \tag{179}$$

Now average over independent $x, x' \sim \mu$. By the definition of Δ_i and Tonelli's theorem, the left side becomes $\mathbb{E}_P[\Delta_i \mid X_{<i} = u]$. Averaging over the prefix proves $\mathbb{E}_P \Delta_i \le \delta$. □

In the following lemma, we prove a moment bound for the retained information density (under $\widetilde{P}$). We would use this moment bound later to prove a tail bound for $\widetilde{\imath}(S,Y)$.

Lemma B.18. *The retained information density $\widetilde{\imath}(S,Y)$ from* (167) *is integrable under $\widetilde{P}$. For every $\lambda > 0$,*

$$\int \exp\left\{\lambda\widetilde{\imath}(S,Y) - \lambda c_\lambda \sum_{i=1}^{n} \Delta_i\right\} d\widetilde{P} \le \exp\left\{\frac{n\lambda(1+\lambda)\epsilon^2}{8}\right\}. \tag{180}$$

Proof. Fix i and $(y, x_{<i})$, and suppress the history in the notation $r_i(x) = r_i^{y,x_{<i}}(x)$. Define the clipped ratio

$$z_i(x) := \max\{r_i(x), e^{-\epsilon}\} = \min\{\max\{k_x(y), e^{-\epsilon}\}, e^{\epsilon}\} \in [e^{-\epsilon}, e^{\epsilon}]. \tag{181}$$

Here k_x is the output likelihood from Lemma B.17; the second equality holds almost everywhere. The lower clipping changes the ratio used inside the logarithm, while mass is still weighted by r_i.

Put $R = \int r_i \, d\mu \le 1$. Give Z weight $r_i(x)\mu(dx)$ at $\log z_i(x)$, and put the missing weight $1 - R$ at zero. Explicitly, for every bounded measurable φ,

$$\mathbb{E}\varphi(Z) = \int r_i(x)\varphi(\log z_i(x))\,\mu(dx) + (1-R)\varphi(0). \tag{182}$$

Thus $Z \in [-\epsilon, \epsilon]$.

Random Variable Z satisfies the assumption needed for the moment bound in Lemma B.19: Specifically we show the assumptions hold with $\Delta = \Delta_i$, i.e. in two parts we show that $\mathbb{E}(Z - Z' - \epsilon)_+ \le \Delta_i$ and $1 - \mathbb{E}e^{-Z} \le \Delta_i$.

Part 1. Let Z' be an independent copy of Z. Pairs involving the added zero have gap at most ϵ, so

$$\mathbb{E}(Z - Z' - \epsilon)_+ = \iint r_i(x) r_i(x') \left(\log \frac{z_i(x)}{e^{\epsilon} z_i(x')}\right)_+ \mu(dx)\mu(dx'). \tag{183}$$

We bound the integrand. Whenever the logarithm is positive, $z_i(x) > e^{\epsilon} z_i(x') \ge 1$, so $z_i(x) = r_i(x)$. Using $(\log u)_+ \le (u-1)_+$, we obtain

$$\begin{aligned} & r_i(x) r_i(x') \left(\log \frac{z_i(x)}{e^{\epsilon} z_i(x')}\right)_+ \\ & \le r_i(x) r_i(x') \left(\frac{z_i(x)}{e^{\epsilon} z_i(x')} - 1\right)_+ \\ & = \frac{r_i(x) r_i(x')}{e^{\epsilon} z_i(x')} \big(r_i(x) - e^{\epsilon} z_i(x')\big)_+ \\ & \le \big(r_i(x) - e^{\epsilon} r_i(x')\big)_+. \end{aligned}$$

For the last step, $r_i(x) \le e^{\epsilon}$ and $r_i(x') \le z_i(x')$ make the prefactor at most one; replacing $z_i(x')$ by the smaller $r_i(x')$ can only increase the positive part. Integrating over x, x' proves

$$\mathbb{E}(Z - Z' - \epsilon)_+ \le \iint \big(r_i(x) - e^{\epsilon} r_i(x')\big)_+ \mu(dx)\mu(dx') = \Delta_i. \tag{184}$$

Part 2. The definition of Z gives

$$\begin{aligned} \mathbb{E}e^{-Z} &= 1 - R + \int \frac{r_i(x)}{z_i(x)}\,\mu(dx) \\ &= 1 - R + \int \min\{e^{\epsilon} r_i(x), 1\}\,\mu(dx). \end{aligned}$$

The second equality follows by considering $r_i(x) \ge e^{-\epsilon}$ and $r_i(x) < e^{-\epsilon}$. For each fixed x', since $R \le 1$,

$$\begin{aligned} R - \min\{e^{\epsilon} r_i(x'), 1\} &\le (R - e^{\epsilon} r_i(x'))_+ \\ &= \left(\int [r_i(x) - e^{\epsilon} r_i(x')]\,\mu(dx)\right)_+ \\ &\le \int \big(r_i(x) - e^{\epsilon} r_i(x')\big)_+ \mu(dx). \end{aligned}$$

The last inequality is Jensen's inequality for the positive-part function. Integrate over x' to conclude

$$1 - \mathbb{E}e^{-Z} = \int \left[R - \min\{e^{\epsilon} r_i(x'), 1\}\right] \mu(dx') \le \Delta_i. \tag{185}$$

Apply the moment bound in Lemma B.19. Since $r_i \le z_i$ and $\lambda > 0$,

$$\begin{aligned}\int r_i^{1+\lambda}\, d\mu \le \int r_i z_i^{\lambda}\, d\mu \le \mathbb{E}e^{\lambda Z} \\ \le \exp\left\{\frac{\lambda(1+\lambda)\epsilon^2}{8} + \lambda c_\lambda \Delta_i\right\}.\end{aligned}$$

The second inequality adds the nonnegative mass $1 - R$ at zero. Thus, at every fixed history,

$$\int r_i(x)^{1+\lambda} e^{-\lambda c_\lambda \Delta_i}\, \mu(dx) \le \exp\left\{\frac{\lambda(1+\lambda)\epsilon^2}{8}\right\}. \tag{186}$$

Integrate the coordinates backwards. The product density (166) rewrites the left side of (180) as

$$\int \prod_{i=1}^{n} \left[r_i(X_i)^{1+\lambda} e^{-\lambda c_\lambda \Delta_i}\right] dQ. \tag{187}$$

Under $Q = \mu^{\otimes n} \otimes P_Y$, we can integrate each coordinate against μ, with $Y = y$ held fixed. Start with X_n. For $i < n$, the i-th factor depends only on $y, x_1, \ldots, x_i$, so it is constant during this integration. Also Δ_n depends only on y, x_1^{n-1}. The integral over x_n is therefore

$$\left(\prod_{i=1}^{n-1} r_i(x_i)^{1+\lambda} e^{-\lambda c_\lambda \Delta_i}\right) \int r_n(x_n)^{1+\lambda} e^{-\lambda c_\lambda \Delta_n}\, \mu(dx_n). \tag{188}$$

By (186), the last integral is at most $\exp\{\lambda(1+\lambda)\epsilon^2/8\}$, uniformly over the fixed history. After replacing it by this constant, the remaining product has exactly the same form with $n-1$ factors.

Repeat for $X_{n-1}, \ldots, X_1$. Each step contributes the same constant, so the full coordinate integral is at most $\exp\{n\lambda(1+\lambda)\epsilon^2/8\}$. This bound is uniform in y, and integrating over P_Y leaves it unchanged. The factors are nonnegative, so Tonelli's theorem justifies the successive integrations. Working backwards ensures that the earlier factors stay fixed; it does not require the factors themselves to be independent.

Finally, (166) and (163) give $0 \le d\widetilde{P}/dQ \le e^{n\epsilon}$. Since $v|\log v|$, with value zero at zero, is bounded on $[0, e^{n\epsilon}]$, the retained information density is integrable under $\widetilde{P}$. □

Proof of Theorem 9: Fix $\lambda > 0$, and set

$$\kappa := \frac{n(1+\lambda)\epsilon^2}{8} + \frac{\log(1/\beta_0)}{\lambda} + \frac{n\delta}{\beta_0} c_\lambda. \tag{189}$$

We show that $\widetilde{P}\{\widetilde{\imath}(S,Y) > \kappa\} \le \beta_0$.

Suppose instead that $E = \{\widetilde{\imath}(S,Y) > \kappa\}$ has retained mass $p = \widetilde{P}(E) > \beta_0$. Define the probability law on this event by

$$R := \frac{\widetilde{P}(\,\cdot \cap E)}{p}. \tag{190}$$

Since $R \le \widetilde{P}/p \le P/p$ and $\Delta_i \ge 0$, Lemma B.17 gives

$$\mathbb{E}_R \sum_{i=1}^{n} \Delta_i \le \frac{1}{p} \sum_{i=1}^{n} \mathbb{E}_P \Delta_i \le \frac{n\delta}{p}. \tag{191}$$

Restrict the moment bound (180) to E and apply Jensen's inequality under R:

$$\begin{aligned}p \exp\left\{\lambda \mathbb{E}_R \widetilde{\imath}(S,Y) - \lambda c_\lambda \mathbb{E}_R \sum_{i=1}^{n} \Delta_i\right\} \le \int_E e^{\lambda \widetilde{\imath}(S,Y) - \lambda c_\lambda \sum_i \Delta_i}\, d\widetilde{P} \\ \le e^{n\lambda(1+\lambda)\epsilon^2/8}.\end{aligned}$$

But $p > \beta_0$, $\mathbb{E}_R \widetilde{\imath}(S,Y) > \kappa$, and $\mathbb{E}_R \sum_i \Delta_i \le n\delta/\beta_0$. Hence the left side strictly exceeds

$$\beta_0 \exp\left\{\lambda\kappa - \frac{\lambda n\delta}{\beta_0} c_\lambda\right\} = e^{n\lambda(1+\lambda)\epsilon^2/8}, \tag{192}$$

where the equality is the definition of κ. This is a contradiction. Integrability follows from Lemma B.18. Thus the retained upper tail has mass at most β_0.

Now consider any measurable event F. Since $\widetilde{P} \le P$, and $\widetilde{P}(\Omega) \ge 1 - n\delta$, we have

$$P(F) \le \widetilde{P}(F) + n\delta \tag{193}$$
$$= n\delta + \widetilde{P}(F \cap E) + \widetilde{P}(F \cap E^c) \tag{194}$$
$$\le n\delta + \beta_0 + e^{\kappa} Q(F) \tag{195}$$

Which, by definition of approximate max-information, shows for $\beta = \beta_0 + n\delta$ we have

$$I_\infty^\beta(Y,S) \le \frac{n(1+\lambda)\epsilon^2}{8} + \frac{\log(1/\beta_0)}{\lambda} + \frac{n\delta}{\beta_0} c_\lambda. \tag{196}$$

The bound $n\epsilon$. Since $r_i(X_i) \le e^\epsilon$ for every coordinate, the factorization (167) gives

$$\widetilde{\imath}(S,Y) = \sum_{i=1}^n \log r_i(X_i) \le n\epsilon \qquad \widetilde{P}\text{-almost everywhere.} \tag{197}$$

Thus $\widetilde{P}\{\widetilde{\imath}(S,Y) > n\epsilon\} = 0$, and for every measurable event F,

$$P(F) \le \widetilde{P}(F) + n\delta \le e^{n\epsilon} Q(F) + n\delta. \tag{198}$$

Consequently, for $\beta \ge n\delta$:

$$I_\infty^\beta(Y,S) \le n\epsilon. \tag{199}$$

Since the bound in Equation (196) holds for every $\lambda > 0$, combining Equation (196) and Equation (199) proves Theorem 9. □

The counterexample for the case $\beta < n\delta$.

Theorem 10. *Let $n \ge 1$, $0 < \delta < 1$, and $0 < \beta < n\delta$. Even for i.i.d. data and $(0,\delta)$ differentially private mechanisms, $I_\infty^\beta(Y,S)$ can be arbitrarily large. Consequently, no finite upper bound depending only on $n, \epsilon, \delta, \beta$ can hold for all (ϵ,δ) differentially private mechanisms throughout this parameter regime.*

Proof. First assume $0 < \delta < 1/n$. Let μ be the uniform distribution on $\{1,\dots,M\}$, and let $S = (X_1,\dots,X_n) \sim \mu^{\otimes n}$. Consider the mechanism

$$Y = \mathcal{A}(S) = \begin{cases} \perp, & \text{with probability } 1 - n\delta, \\ (J, X_J), & \text{with probability } n\delta, \end{cases} \tag{200}$$

where, conditional on disclosure, J is drawn uniformly from $\{1,\dots,n\}$, independently of S.

To verify privacy, consider two datasets differing only in coordinate i. Couple the two runs using the same disclosure decision and index J. Their outputs agree unless disclosure occurs and $J = i$, an event of probability $n\delta/n = \delta$. Thus, for every measurable output event B,

$$\Pr\{\mathcal{A}(s) \in B\} \le \Pr\{\mathcal{A}(s') \in B\} + \delta. \tag{201}$$

Hence $\mathcal{A}$ is $(0,\delta)$ differentially private, and therefore also (ϵ,δ) differentially private for every $\epsilon \ge 0$.

Write $P = P_{S,Y}$ and $Q = P_S \otimes P_Y$, and consider the event

$$E := \{(s,y) : y = (j, s_j) \text{ for some } j \in \{1,\dots,n\}\}. \tag{202}$$

Under P, this event occurs whenever the mechanism discloses an observation, so $P(E) = n\delta$. Under Q, the dataset and output are independent. Conditional on a disclosed pair (j, x), the probability that $X_j = x$ is $1/M$. Therefore,

$$Q(E) = \frac{n\delta}{M}. \tag{203}$$

By the definition of approximate max-information, any admissible upper bound κ must satisfy

$$P(E) \le e^{\kappa} Q(E) + \beta. \tag{204}$$

Since $\beta < n\delta$, it follows that

$$I_\infty^\beta(Y, S) \ge \log \frac{P(E) - \beta}{Q(E)} = \log M + \log\left(1 - \frac{\beta}{n\delta}\right). \tag{205}$$

For fixed n, δ, β, this lower bound tends to infinity as $M \to \infty$, proving the claim.

For $1/n \le \delta < 1$, set $\delta' = \frac{1+\beta}{2n}$, since $0 < \beta < 1$, then $\delta' < 1/n$ and $\beta < n\delta'$. Therefore, the construction above gives a $(0, \delta')$-DP algorithm with arbitrary large $I_\infty^\beta(Y, S)$. Since $\delta' < \frac{1}{n} \le \delta$, the algorithm is also $(0, \delta)$-DP. This extends the conclusion to the remaining range $1/n \le \delta < 1$. □

B.4.1 Auxiliary moment bound

Lemma B.19. *Let $Z \in [-\epsilon, \epsilon]$, let Z' be an independent copy, and let $\Delta \ge 0$.*

Centered bound. *If*

$$\mathbb{E}(Z - Z' - \epsilon)_+ \le \Delta, \tag{206}$$

then, for every $t \in \mathbb{R}$,

$$\log \mathbb{E}e^{t(Z - \mathbb{E}Z)} \le \frac{t^2\epsilon^2}{8} + |t|\big(e^{2\epsilon|t|} - 1\big)\Delta. \tag{207}$$

In particular, if $\mathbb{E}Z = 0$, the left side is $\log \mathbb{E}e^{tZ}$.

Uncentered bound. *If, in addition,*

$$1 - \mathbb{E}e^{-Z} \le \Delta, \tag{208}$$

then, with $c_\lambda := e^{2\epsilon} + \epsilon - 1 + e^{2\lambda\epsilon}$, for every $\lambda > 0$,

$$\log \mathbb{E}e^{\lambda Z} \le \frac{\lambda(1+\lambda)\epsilon^2}{8} + \lambda c_\lambda \Delta. \tag{209}$$

Proof of Lemma B.19. We first prove the centered bound (207) using only the excess-gap assumption. We then use the additional normalization assumption to control the mean and obtain (209).

A variance estimate. For any random variable Z in an interval of length 2ϵ and an independent copy Z', we first show

$$\mathrm{Var}(Z) \le \frac{\epsilon^2}{4} + 4\epsilon\, \mathbb{E}(Z - Z' - \epsilon)_+. \tag{210}$$

Translate the interval to $[0, 2\epsilon]$, and write $V = |Z - Z'|$, $V_\circ = \min\{V, 2\epsilon - V\}$. Independence and symmetry give

$$\begin{aligned} \mathrm{Var}(Z) - 4\epsilon\, \mathbb{E}(Z - Z' - \epsilon)_+ &= \tfrac{1}{2}\mathbb{E}[V^2 - 4\epsilon(V - \epsilon)_+] \\ &= \tfrac{1}{2}\mathbb{E}V_\circ^2 \le \tfrac{\epsilon}{2}\mathbb{E}V_\circ. \end{aligned}$$

For fixed $z, z' \in [0, 2\epsilon]$, the folded gap is

$$\min\{|z - z'|, 2\epsilon - |z - z'|\} = \int_0^\epsilon |\mathbf{1}_{[u,u+\epsilon]}(z) - \mathbf{1}_{[u,u+\epsilon]}(z')|\, du. \tag{211}$$

The integral counts positions of a sliding interval that contains exactly one of the two points. Writing $q_u = \Pr\{Z \in [u, u + \epsilon]\}$, Tonelli's theorem gives

$$\mathbb{E}V_\circ = \int_0^\epsilon 2q_u(1 - q_u)\, du \le \frac{\epsilon}{2}. \tag{212}$$

This proves the variance estimate.

Control the excess under exponential tilting. Return to $Z \in [-\epsilon, \epsilon]$. It suffices to prove the claim for $t \geq 0$: replacing Z by $-Z$ leaves $\mathbb{E}(Z - Z' - \epsilon)_+$ unchanged, by exchangeability. Write $g(s) = \log \mathbb{E}e^{sZ}$, and let $\mathbb{E}_s$ denote expectation under the law tilted by $e^{sZ-g(s)}$. For pairs, both copies are independent under this tilted law.

Put $T = (Z - Z' - \epsilon)_+$ and $h(z') = \mathbb{E}(Z - z' - \epsilon)_+$, with the latter expectation under the original law. Since h is decreasing and $z' \mapsto e^{sz'}$ is increasing for $s \geq 0$, their covariance is nonpositive. Thus

$$\mathbb{E}[e^{sZ'} h(Z')] \leq \mathbb{E}e^{sZ'} \, \mathbb{E}h(Z') \leq e^{g(s)}\Delta. \tag{213}$$

Using $Z \leq \epsilon$ and $g(s) \geq -s\epsilon$, we obtain

$$\begin{aligned} \mathbb{E}_s T &= \frac{\mathbb{E}[e^{s(Z+Z')}T]}{e^{2g(s)}} \\ &\leq \frac{e^{s\epsilon}\mathbb{E}[e^{sZ'}h(Z')]}{e^{2g(s)}} \leq e^{s\epsilon - g(s)}\Delta \leq e^{2s\epsilon}\Delta. \end{aligned}$$

The variance estimate under the tilted law now gives

$$g''(s) = \mathrm{Var}_s(Z) \leq \frac{\epsilon^2}{4} + 4\epsilon e^{2s\epsilon}\Delta. \tag{214}$$

All differentiations are justified by boundedness of Z.

Integrate the variance bound. Since $g(0) = 0$ and $g'(0) = \mathbb{E}Z$,

$$\begin{aligned} \log \mathbb{E}e^{t(Z-\mathbb{E}Z)} &= \int_0^t (t-s)g''(s)\,ds \\ &\leq \frac{t^2\epsilon^2}{8} + \frac{\Delta}{\epsilon}\big(e^{2\epsilon t} - 1 - 2\epsilon t\big). \end{aligned}$$

The trapezoid bound for the convex exponential function gives

$$e^u - 1 = \int_0^u e^v\,dv \leq \frac{u}{2}(1 + e^u), \qquad e^u - 1 - u \leq \frac{u}{2}(e^u - 1), \quad u \geq 0. \tag{215}$$

Apply this with $u = 2\epsilon t$ to obtain (207). The case $t < 0$ follows by the symmetry noted above.

Control the mean using the negative moment. Since $\mathbb{E}e^{-Z} \geq e^{-\epsilon}$, the chord bound for $-\log v$ gives

$$-\log \mathbb{E}e^{-Z} \leq \frac{\epsilon}{1 - e^{-\epsilon}}(1 - \mathbb{E}e^{-Z})_+ \leq (1 + \epsilon)\Delta. \tag{216}$$

Here $\epsilon/(1 - e^{-\epsilon}) \leq 1 + \epsilon$ follows from $e^\epsilon \geq 1 + \epsilon$. Apply (207) at $t = -1$:

$$\begin{aligned} \mathbb{E}Z &\leq -\log \mathbb{E}e^{-Z} + \frac{\epsilon^2}{8} + (e^{2\epsilon} - 1)\Delta \\ &\leq \frac{\epsilon^2}{8} + (e^{2\epsilon} + \epsilon)\Delta. \end{aligned}$$

Control the positive moment. Apply (207) at $t = \lambda$:

$$\begin{aligned} \log \mathbb{E}e^{\lambda Z} &= \lambda \mathbb{E}Z + \log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \\ &\leq \frac{\lambda(1+\lambda)\epsilon^2}{8} + \lambda\big(e^{2\epsilon} + \epsilon + e^{2\lambda\epsilon} - 1\big)\Delta. \end{aligned}$$

This is (209).

□

C EXPERIMENTS

Here we present the details of our model training experiments in Section 5.3.

Implementation. For the experiments, we use Opacus (Yousefpour et al., 2021) for differentially private training and the implementation[3] from (García-Pérez et al., 2026; Pérez-Ortiz et al., 2021) to directly optimize a PAC-Bayes bound for a family of Gaussian distributions. The results for data-independent priors are based on García-Pérez et al. (2026), recomputed from the authors' released `results_multiseed.ipynb` file in their repository.

Datasets. We evaluate the bounds on the MNIST (LeCun et al., 1998) data with $60{,}000$ training samples, and CIFAR-10 (Krizhevsky & Hinton, 2009) $50{,}000$ training samples. All reported guarantees are based on only the training data.

Models. For MNIST, we use a fully-connected MLP with three 600-unit hidden layers and a 10-class output layer wit $1{,}198{,}210$ parameters. For CIFAR-10, we use three different models :1) a four-layer ConvNet with two convolutional layers followed by two fully connected layers with $9{,}946$ parameters, 2) a nine-layer ConvNet with six convolutional layers followed by three fully connected layers with $5{,}851{,}338$ parameters, and 3) a fifteen-layer ConvNet with twelve convolutional layers followed by three fully connected layers with $13{,}193{,}930$ parameters. So, as is typical for real-world deep networks, they are highly overparameterized.

Prior and Posterior distributions. For all experiments, both the prior and posterior are Gaussian distributions. For the baseline, the independent prior is

$$\pi_0 = \mathcal{N}(\mu_0, \tau \mathrm{I}) \tag{217}$$

with a randomly generated μ_0, and τ is chosen from $\{0.001, 0.005, 0.01, 0.02, \ldots, 0.08\}$, using a union-bound argument as described below. The DP-SGD prior is

$$\pi_{DP} = \mathcal{N}(\mu_{DP}, \tau \mathrm{I}), \tag{218}$$

where μ_{DP} is trained with DP-SGD. The posterior is also a Gaussian distribution

$$\rho = \mathcal{N}(\mu_\rho, \Sigma_\rho), \tag{219}$$

with diagonal Σ_ρ. For all experiments, μ_ρ and Σ_ρ are trained by optimizing the corresponding bounds.

Training the DP-SGD prior. For training the prior, we train a deterministic model based on the DP-SGD algorithm in the streaming setting or with Poisson subsampling with multiple epochs.

We use the following hyperparameters: Adam update rule, clipping norm $\zeta \in \{0.01, 0.02, 0.05, 0.1\}$, noise multiplier with $\sigma = 1/\zeta$. We choose the learning rate from $\{0.001, 0.005, 0.01, 0.05, 0.1\}$. For streaming DP-SGD, we use a fixed batch size $m = 5000$ for one epoch, and for Poisson subsampling we use expected batch size $m = 5000$ (at each batch, each sample is included independently with probability $\frac{m}{n}$), and choose the number of epochs from $\{1, 2, 5, 10, 20\}$, i.e. for each experiment at most $4 \cdot 5 \cdot 5 = 100$ tuples of hyperparameters are used.

For streaming data, we compute the max-information using Theorem 4. For Poisson subsampling, we convert the add/remove Rényi DP values computed by Opacus to replacement DP based on Mironov (2017) as described below, use Theorem 8 to compute max-information. We optimized the bound for $\alpha \in \{2, \ldots, 5000\}$.

From add/remove RDP to replace-one RDP. Suppose $\mathcal{A}$ satisfies add/remove RDP with profile $\epsilon_{\mathrm{add}}(\gamma)$, for $\gamma > 1$. For a dataset of n samples, write

$$Y_n(x_{1:n}) = \mathcal{A}(x_1, \ldots, x_n).$$

Let $x_{1:n}$ and $x'_{1:n}$ differ only in coordinate i, and denote their common dataset after deleting that coordinate by

$$x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

[3] `https://github.com/Diegogpcm/pacbayesgradients`

Both $x_{1:n}$ and $x'_{1:n}$ are add/remove neighbors of x_{-i}. Hence, the weak triangle inequality (Mironov, 2017, Corollary 4(1)) gives, for every $\alpha > 1$,

$$\begin{aligned} D_\alpha\big(Y_n(x_{1:n})\|Y_n(x'_{1:n})\big) &\le \frac{2\alpha-1}{2(\alpha-1)} D_{2\alpha}\big(Y_n(x_{1:n})\|\mathcal{A}(x_{-i})\big) \\ &\quad + D_{2\alpha-1}\big(\mathcal{A}(x_{-i})\|Y_n(x'_{1:n})\big) \\ &\le \frac{2\alpha-1}{2(\alpha-1)}\epsilon_{\mathrm{add}}(2\alpha) + \epsilon_{\mathrm{add}}(2\alpha-1). \end{aligned}$$

Thus, for every n, $\mathcal{A}$ satisfies replace-one RDP with profile

$$\epsilon_{\mathrm{rep}}(\alpha) = \frac{2\alpha-1}{2(\alpha-1)}\epsilon_{\mathrm{add}}(2\alpha) + \epsilon_{\mathrm{add}}(2\alpha-1).$$

Computing the certificates. To estimate an upper bound for the training error of the posterior, $\widehat{R}$, instead of using $150{,}000$ Monte Carlo rollouts as in García-Pérez et al. (2026), which requires heavy computations, we use the method described in Biggs (2022), and for each sample we sample different weights. This would save the amount of computation by the order of dataset size, i.e. it needs 150000 forward passes to get the same confidence instead of 150000×50000. Based on the computed empirical risk, we apply the numeric inversion of kl to compute the upper bound on training risk based on a binary search, as discussed in Appendix A.3. Note that for each set of hyperparameters, this computation is done for the final posterior independent of the training procedure.

In the end, given the upper bound on training error and complexity terms, we apply another inversion of kl to compute the final upper bound on the posterior.

Hyperparameter Section via Union Bounds. To apply the bounds, the hyperparameters used to generate the prior must be data-independent. However, in practice, we want to select them from a finite grid based on the resulting certificates. To account for this selection, we employ a union bound argument. Formally, assume we have K_1 possible tuples of DP-SGD hyperparameters, and K_2 possible choices of the prior variance τ. For the i-th DP-SGD hyperparameter tuple, let μ_i be the output of DP-SGD training and τ_j be the j-th possible value for τ. Let $\pi_{i,j} = \mathcal{N}(\mu_i, \tau_j \mathbf{Id})$ be the prior generated by i-th hyperparameter tuple, and j-th value for τ. Let $\kappa_i = I_\infty^{\beta/K_1}(\mu_i(S), S)$ be the corresponding max-information for i-th hyperparameter tuple. By the union bound argument, for any δ, δ', and β, with probability at least $1-(\delta-\delta')$ for all posteriors ρ, and all priors $\pi_{i,j}$, we have

$$R(\rho) \le \mathrm{kl}^{-1}\left(\widehat{R}(\rho)) \;\middle|\; \frac{\mathrm{KL}(\rho\|\pi_{i,j}) + \kappa_i + \log(\frac{2K_1K_2\sqrt{n}}{\delta-\delta'-\beta})}{n}\right), \tag{220}$$

Additionally, for any fixed posterior, with probability $1 - \frac{\delta'}{K_1K_2}$ we have (Biggs, 2022)

$$\widehat{R}(\rho) \le \mathrm{kl}^{-1}\Big(\tilde{R}_N(\rho)\Big|\frac{\log(\frac{K_1K_2}{\delta'})}{N}\Big), \tag{221}$$

where $\tilde{R}_N(\rho)$ is the average training error draws of the posterior ρ as described above.

Combining these two bounds with $\delta = 0.05, \delta' = 0.0025, \beta = 0.045, K_1 = 100, K_2 = 10$, results in the bound values we report in the paper.