The Relative Gaussian Mechanism and its Application to Private Gradient Descent

Hadrien Hendrikx
Centre Inria de l'Univ. Grenoble Alpes
CNRS, LJK, Grenoble, France

Paul Mangold
CMAP, UMR 7641,
École Polytechnique

Aurélien Bellet
Inria, Université de Montpellier,
France

Abstract

The Gaussian Mechanism (GM), which consists in adding Gaussian noise to a vector-valued query before releasing it, is a standard privacy protection mechanism. In particular, given that the query respects some *L2 sensitivity* property (the L2 distance between outputs on any two neighboring inputs is bounded), GM guarantees Rényi Differential Privacy (RDP). Unfortunately, precisely bounding the L2 sensitivity can be hard, thus leading to loose privacy bounds. In this work, we consider a *Relative* L2 sensitivity assumption, in which the bound on the distance between two query outputs may also depend on their norm. Leveraging this assumption, we introduce the *Relative Gaussian Mechanism* (RGM), in which the variance of the noise depends on the norm of the output. We prove tight bounds on the RDP parameters under relative L2 sensitivity, and characterize the privacy loss incurred by using output-dependent noise. In particular, we show that RGM naturally adapts to a latent variable that would control the norm of the output. Finally, we instantiate our framework to show tight guarantees for Private Gradient Descent, a problem that naturally fits our relative L2 sensitivity assumption.

1 INTRODUCTION

Differential Privacy (DP) [Dwork, 2006] is considered the gold standard for protecting privacy, for instance in machine learning. In this framework, a curator has a database x, and would like to answer a query $\mathcal{R}$ on x by releasing an output $\mathcal{R}(x)$. Yet, releasing $\mathcal{R}(x)$ might reveal sensitive information on x. Instead, the curator may use a private algorithm $\mathcal{A}$ to release a sanitized approximation $\mathcal{A}(\mathcal{R})(x)$ of $\mathcal{R}(x)$. To guarantee that the amount of information leaked by releasing $\mathcal{A}(\mathcal{R})(x)$ is limited, DP ensures that the distributions of $\mathcal{A}(\mathcal{R})(x)$ and $\mathcal{A}(\mathcal{R})(y)$ are close for any $y \sim x$, *i.e.*, that is close to x according to a neighboring relation (databases that only differ in one row for instance). Several divergences have been considered to measure the closeness between these two distributions, leading to different variants of DP. Among them, *Rényi-Differential Privacy* (RDP), which is based on the Rényi divergence, has become popular for its mathematical properties [Mironov, 2017].

Definition 1 (Rényi Differential Privacy)**.** *A randomized algorithm $\mathcal{A}$ satisfies (α, ε)-RDP for $\alpha > 1$ and $\varepsilon > 0$ if $\mathcal{D}_\alpha(\mathcal{A}(x)||\mathcal{A}(y)) \leq \varepsilon$ for all pairs of neighboring datasets $x \sim y$, where $\mathcal{D}_\alpha(\mathcal{A}(x)||\mathcal{A}(y))$ is the α-Rényi divergence between $\mathcal{A}(x)$ and $\mathcal{A}(y)$.*

A fundamental building block for designing a private algorithm $\mathcal{A}$ is the *Gaussian Mechanism* (GM_σ) [Dwork et al., 2006, 2014], which adds Gaussian noise to the private value $\mathcal{R}(x)$:

$$\mathrm{GM}_\sigma(\mathcal{R})(x) = \mathcal{R}(x) + \mathcal{N}(0, \sigma^2), \text{ for some } \sigma^2 > 0\,. \quad (1)$$

It is very common (*e.g.,* in machine learning) to compose multiple calls to GM_σ to build iterative algorithms like differentially private gradient descent [Song et al., 2013, Bassily et al., 2014]. RDP is able to tightly track the privacy guarantees of (compositions of) GM_σ, and can be converted into the more classical (ϵ, δ)-DP variant [Mironov, 2017].

The noise scale σ^2 of GM_σ is based on an L2 sensitivity assumption, which guarantees that for any neighboring inputs $x \sim y$, the query $\mathcal{R}$ verifies:

$$\|\mathcal{R}(x) - \mathcal{R}(y)\|^2 \leq R_{\text{abs}}^2 \quad (2)$$

for some $R_{\text{abs}} > 0$. In particular, for $\sigma^2 = \frac{\alpha R_{\text{abs}}^2}{2\varepsilon}$, $\mathrm{GM}_\sigma(\mathcal{R})(x)$ satisfies (α, ε)-RDP. It is thus crucial to

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

estimate the L2 sensitivity precisely to achieve the best possible privacy-utility trade-off. Unfortunately, this R_{abs} constant is often not directly known and difficult to bound tightly. In some cases, the distance between outputs is also highly correlated to the norm of these outputs, and this is the case in particular when the outputs depend on a non-private latent variable.

Consider for instance an institute that would like to assess the mean salary for different jobs in a given company. Individual salaries are sensitive information, but people's job is not secret, and the average salary per job is the desired output. If we were to use the standard Gaussian Mechanism, then we would need an absolute sensitivity bound of the form of (2) (note that other types of noises, such as Laplace, would require similar bounds in other norms, such as $L1$, but the absolute aspect would remain). To do this, the simplest approach is to use a bound on the maximum possible salary across all jobs in the company. However, this is not satisfactory since results for lower-paid jobs would be dominated by noise. An alternative is to restrict the neighboring relation to people that have the same job, which is possible since the job is not private. The problem is that estimating the salary per job (or a bound on it) is exactly what we would like to achieve in the first place. In this case, absolute sensitivity bounds are thus unsatisfactory, and would lead to unnecessarily high, as well as unfair (since the precision would be higher for well-paid jobs) estimates of the mean salary per job. Now consider that we know that by law, there should not be more than 10% variations in salary for a given job in a given company: this corresponds to a *relative* sensitivity assumption. In this case, one is tempted to calibrate the noise to the empirical mean salary for a given job, since we know that all the people with the same job in this company have comparable salaries. In this paper, **we tightly characterize how to scale the noise under this relative sensitivity assumption**, leading to precise and fair estimates of the mean salaries per job. Note that this simple example directly translates for instance to releasing gradients, where the job would be the point at which they are computed and the salary would be their magnitude.

Our contributions are the following: **(i)** We introduce the Relative L2 Sensitivity, which generalizes the L2 sensitivity by allowing the upper bound to depend on the norm of queries. **(ii)** We leverage this assumption to introduce the *Relative Gaussian mechanism* (RGM), in which the noise that we introduce depends on the output that we are about to release. **(iii)** We show tight privacy guarantees for the Relative Gaussian Mechanism. **(iv)** We show how the Relative Gaussian mechanism can be applied for Private Gradient Descent to provide adaptivity to the gradients' magnitude.

We first review related work in Section 2. We then define the Relative L2 Sensitivity in Section 3, and introduce RGM in Section 4. Finally, we instantiate the results for gradient descent on quadratics in Section 5, and present numerical illustrations in Section 6.

2 RELATED WORK

Local and smooth sensitivity. Several classic techniques in the DP literature seek to avoid the calibration of noise to global sensitivity by relying on the notion of *local sensitivity*. The local sensitivity $LS_{\mathcal{R}}(x) = \max_{y:y\sim x} \|\mathcal{R}(x) - \mathcal{R}(y)\|$ of a dataset x measures how much $\mathcal{R}(y)$ can differ from $\mathcal{R}(x)$ for any neighbor y of x, which can be much smaller than the global sensitivity. In general however, calibrating the noise to the local sensitivity does not provide privacy, as two neighboring datasets may have very different local sensitivities. To go around this issue, previous work has proposed approaches based on smoothing the local sensitivity [Nissim et al., 2007]. Of particular relevance to our work, Bun et al. [2018] introduce truncated Concentrated Differential Privacy (tCDP), a privacy notion that is well-suited to analyzing mechanisms with smooth sensitivity. They prove that "Gaussian Smooth Sensitivity" (Gaussian mechanism used with smooth sensitivity) satisfies tCDP. This result proves to be useful in solving problems such as Gap-max, and can be extended to other noise distributions [Bun and Steinke, 2019]. If $\mathcal{R}(x)$ is of dimension 1, the relative sensitivity assumption and the subsequent Gaussian mechanism can be seen as Gaussian Smooth Sensitivity with a special case of smooth upper bound. Yet, we go further in this paper and explore the d-dimensional case, which in particular allows us to consider more complex algorithms such as gradient descent.

Other approaches include refining an absolute sensitivity bound by privately discarding outliers [Tsfadia et al., 2022], going around the lack of global sensitivity bound by contructing a private data-dependent upper bound [Wang, 2018], or proposing and privately testing the validity of a local sensitivity bound before releasing the output [Dwork and Lei, 2009], potentially using distributional assumptions to privately estimate queries whose absolute sensitivity is unbounded [Brunel and Avella-Medina, 2020]. The main drawback of these approaches is that they require a special structure of the problem and release mechanism (which essentially also comes down to a certain smoothness of the local sensitivity). For gradient descent, this would for instance require an absolute bound on individual gradients and is thus not well-suited. Instead, the relative Gaussian Mechanism uses a relative sensitivity assumption which

does not require the absolute boundedness of individual gradients.

Private gradient descent. Differentially private gradient descent (DP-GD) and its stochastic variant (DP-SGD) were first proposed by Song et al. [2013]. These algorithms and further variations have been widely studied as private minimizers of the empirical risk [Song et al., 2013, Bassily et al., 2014, Wang et al., 2017], and of the population risk [Bassily et al., 2019, Feldman et al., 2020]. All these algorithms have been formally shown to achieve the optimal utility derived by Bassily et al. [2014]. The analysis crucially relies on an absolute L2 sensitivity bound on the gradients (typically obtained by assuming the loss function to be Lipschitz) to calibrate the noise. Unfortunately, this often leads to the injection of excessive amounts of noise. Abadi et al. [2016b] proposed a more practical version of DP-SGD (implemented notably in PyTorch Opacus [Yousefpour et al., 2021] and TensorFlow Privacy [Abadi et al., 2016a]) which uses gradient clipping to reduce gradients' L2 sensitivity. Similarly, Asi et al. [2022] reduced this sensitivity using a clipping-like procedure. In both cases, this decrease in L2 sensitivity introduces bias in the computation [Amin et al., 2019]. This phenomenon makes the analysis of clipped algorithms significantly harder [Chen et al., 2020, Yang et al., 2022, Koloskova et al., 2023], and it is difficult to choose a constant clipping threshold without tuning an additional hyperparameter. Pichapati et al. [2019] and Andrew et al. [2021] proposed heuristic methods for chosing clipping thresholds adaptively, although without theoretical guarantees and with limited practical applicability. Our method can reduce the amount of injected noise, while circumventing the difficulty of setting a proper clipping threshold throughout the iterations. Indeed, our relative sensitivity assumption allows the design of a relative Gaussian mechanism where noise naturally adapts to the gradients' norms.

3 RELATIVE L2 SENSITIVITY

As discussed in the introduction, we start by relaxing the restrictive L2 sensitivity assumption.

Definition 2 (Relative L2 sensitivity)**.** *An algorithm $\mathcal{A}$ satisfies Relative L2 sensitivity if there exists constants $\eta > 0$ and $R_{\text{rel}} > 0$ such that for any two neighboring inputs $x \sim y$:*

$$\|\mathcal{R}(x) - \mathcal{R}(y)\|^2 \leq \eta^2 \|\mathcal{R}(x)\|^2 + R_{\text{rel}}^2. \tag{3}$$

Note that by symmetry, this is equivalent to $\|\mathcal{R}(x) - \mathcal{R}(y)\|^2 \leq \eta^2 \min(\|\mathcal{R}(x)\|^2, \|\mathcal{R}(y)\|^2) + R_{\text{rel}}^2$. Besides, we recover the L2 sensitivity for $\eta = 0$.

Examples. This definition is particularly useful when we know *relative* or *multiplicative* bounds on inputs. As discussed earlier, this would be the case when estimating salaries for a given job, if we know that all the people we consider have salaries within 10% of each other (for instance because it is imposed by the law). We would need to know salary estimates for each job to guess the appropriate absolute sensitivity R_{abs}^2 (or use a very imprecise global one for all jobs), whereas knowing the law directly gives us $\eta = 0.1$ and $R_{\text{rel}} = 0$.

In this case, the salaries are directly correlated to a latent variable: the jobs. This is also the case for gradients, whose norm depend on the point at which they are computed. The absolute L2 sensitivity would write $\|\nabla f(\theta) - \nabla f'(\theta)\|^2 \leq R_{\text{abs}}(\theta)^2$, where f and f' are objective functions computed on neighboring datasets. Therefore, we would either need to (i) know $R_{\text{abs}}(\theta)$ for all values of θ, which is a lot of information, or (ii) bound it uniformly, which can be very loose. In contrast, Relative L2 sensitivity can ensure $\|\nabla f(\theta) - \nabla f'(\theta)\|^2 \leq \eta^2 \|\nabla f(\theta)\|^2 + R_{\text{rel}}^2$ with tight absolute (independent of θ) parameters η and R_{rel}, see Section 5 for more details.

Links to local sensitivity. As discussed in the related work section, the motivating idea behind local sensitivity [Nissim et al., 2007] is to set the noise according to the bound on the distance between the specific output we would like to protect and all neighboring ones. This allows much lower noise in general, since some outputs might have small sensitivity. Yet, this does not guarantee differential privacy as the level of noise injected gives information about the input that is released, as two neighboring inputs might have very different local sensitivities.

Note that Definition 2 can be considered as a local sensitivity bound, since it depends on the inputs that we consider. It is stronger however: due to its symmetry, relative L2 sensitivity also ensures that two neighboring inputs also have comparable norms, and so comparable local sensitivities. This guarantees a form of smooth sensitivity [Nissim et al., 2007, Bun et al., 2018], and we can thus expect comparable guarantees. We will see that Definition 2 allows privacy guarantees to hold even though the norm of the input is partly revealed through the noising process. In particular, we will show that Definition 2 can be leveraged to release information privately even when $R_{\text{rel}} \neq 0$. Our framework thus highlights another interesting example in which a form of local sensitivity can be used while still ensuring Differential Privacy, which we will leverage to analyze privatize gradient descent.

4 THE RELATIVE GAUSSIAN MECHANISM

4.1 Mechanism and privacy guarantees

We now present the Relative Gaussian Mechanism ($\text{RGM}_{\gamma,\sigma}$), and derive its privacy guarantees. $\text{RGM}_{\gamma,\sigma}$ extends GM_σ, and leverages relative sensitivity to guarantee privacy while adapting the scale of the noise to the norm of the query.

Definition 3 (Relative Gaussian Mechanism)**.** *Let $\gamma > 0$ and $\sigma > 0$. The Relative Gaussian Mechanism of parameters (γ, σ) is defined as:*

$$\text{RGM}_{\gamma,\sigma}(\mathcal{R})(x) = \mathcal{R}(x) + \mathcal{N}(0, \gamma \left\|\mathcal{R}(x)\right\|^2 + \sigma^2)\,. \quad (4)$$

$\text{RGM}_{\gamma,\sigma}$ generalizes the standard GM_σ, that we recover with $\gamma = 0$. When $\gamma > 0$, it controls to which extent $\|\mathcal{R}(x)\|$ impacts the noise. Note that σ^2 is a baseline noise, which allows to handle inputs where the query's output has small norm. For instance, if $\mathcal{R}(x) = 0$ on some input x, and $\mathcal{R}(y) \neq 0$ on an input $y \sim x$, this baseline noise is necessary to guarantee privacy.

We show that, although $\text{RGM}_{\gamma,\sigma}$ uses the query output to calibrate the noise, it can still guarantee privacy. This perhaps surprising result follows from the relative sensitivity assumption. Intuitively, this assumption ensures that all neighboring outputs have comparable norms, resulting in comparable levels of noise. The next theorem formalizes this intuition, deriving tight privacy guarantees for $\text{RGM}_{\gamma,\sigma}$ on queries that satisfy a relative L2 sensitivity assumption.

Theorem 1 (Privacy guarantees of $\text{RGM}_{\gamma,\sigma}$)**.** *Let $\mathcal{R} : \mathcal{D} \to \mathbb{R}^d$ be a query that verifies (η, R_{rel})-relative L2 sensitivity (Definition 2) for some $\eta > 0$ and $R_{\text{rel}} \geq 0$. Then for $1 \leq \alpha < (1+\eta)^2/(2\eta+\eta^2)$, and $\sigma^2 \geq \gamma\eta^{-2}\left[1-\eta(\alpha-1)\right]R_{\text{rel}}^2$, $\text{RGM}_{\gamma,\sigma}(\mathcal{R})$ satisfies (α,ϵ)-Rényi-DP with*

$$\epsilon = \frac{\alpha\eta^2}{2\gamma} \times \frac{1+\gamma d(2+\eta)^2(1+\eta)^2}{1-\eta(\alpha-1)(2+\eta)}. \quad (5)$$

The proof is mostly technical, we thus defer it to Appendix A. Theorem 1 shows that $\text{RGM}_{\gamma,\sigma}$ can provide meaningful privacy guarantees. For a fixed γ, the guarantee is as strong as η^2 is small. This is in line with the intuition presented above: when η^2 is small, $\mathcal{R}(x)$ and $\mathcal{R}(y)$ (for $x \sim y$) have similar norms, and these norms are less sensitive.

Truncated Concentrated DP (tCDP). Theorem 1 can directly be turned into a $(\eta^2[\gamma^{-1} + d(2+\eta)^2(1+\eta)^2], 1+(2\eta(2+\eta))^{-1})$-tCDP result. This is not surprising as the Relative Gaussian Mechanism can be seen as a multidimensional extension of the Gaussian Smooth Sensitivity mechanism [Bun et al., 2018], which offers comparable guarantees if $\mathcal{R}(x) \in \mathbb{R}$. More details can be found in Appendix A.6.

Scale of the noise. The scale of the noise is controlled by the parameter γ. Indeed, for a fixed γ, our result suggests to set the baseline variance as $\sigma^2 = \gamma\eta^{-2}(1-\eta(\alpha-1))R_{\text{rel}}^2$. As such, small values of γ will lead to small noise addition (both in the baseline and the relative term), but will decrease privacy guarantees. Conversely, higher values of γ require more noise for better privacy guarantees.

Arbitrary privacy guarantees cannot be achieved. Although parameter γ controls the level of the privacy guarantee, not all values of α and ϵ are achievable. This is in stark contrast with the classical GM_σ ($\eta = 0$), where increasing the noise σ always improves privacy. This discrepancy is due to the fact that scaling noise with $\|\mathcal{R}(x)\|$ already releases some information about the input. Sadly, this information cannot be privatized using more baseline noise σ^2 without a priori bounds on $\|\mathcal{R}(x)\|$. Nonetheless, we emphasize that when $\eta \to 0$, all values of α and ϵ are possible.

Theorem 1 implies that $\epsilon \geq 2\alpha\eta^2 d$, where d is the dimension of the output of $\mathcal{R}$. Consequently, $\text{RGM}_{\gamma,\sigma}$ is more likely to give good privacy guarantees on small-dimensional queries. Note that this is tight, as discussed below. To mitigate this issue, one can either (i) restrict the query to a subset of its coordinates, or (ii) adapt the query to decrease the value of η (see discussion in Section 5.3).

Conversion to (ϵ,δ)-DP. Using Proposition 3 of Mironov [2017], we can convert the RDP guarantee given in Theorem 1 to classical DP. For clarity of discussion, we give a closed-form expression of the differential privacy guarantees for the Relative Gaussian Mechanism in Corollary 1. We stress that better guarantees can be obtained by numerically optimizing the bound obtained from Proposition 3 of Mironov [2017], and provide a script to choose the best values of α and γ in the supplementary.

Corollary 1 (Conversion to (ϵ,δ)-DP)**.** *Let $0 \leq \delta \leq 1$. We assume that $\gamma^{-1} \geq 4(2+\eta)^2\log(1/\delta)$ or that $d \geq 4\log(1/\delta)/(1+\eta)^2$, and use the same notations as in Theorem 1. Then, $\text{RGM}_{\gamma,\sigma}$ satisfies (ϵ,δ)-differential privacy with parameter $\epsilon = \chi + 2\sqrt{\chi\log(1/\delta)}$, where $\chi = \frac{\eta^2}{\gamma} + \eta^2(2+\eta)^2(1+\eta)^2 d$.*

We prove this result in Appendix A.5, but a similar one can be deduced using tCDP [Bun et al., 2018, Proposition 6]. While this result does not allow arbitrary privacy guarantee, we stress that meaningful guaran-

tees can still be achieved. For instance, if $\eta = 1e\text{-}3$, $d = 10$, $\delta = 1e\text{-}8$, and $\gamma = 100\eta^2$, the Relative Gaussian mechanism guarantees (ϵ, δ)-DP with $\epsilon \approx 0.86$.

Subsampling. We can also leverage the tCDP result to directly obtain a subsampling result [Bun et al., 2018, Theorem 13]. However, note that the Relative L2 sensitivity assumption should be satisfied for each batch of data. This can be demanding, in particular for gradient descent for which we will see that we will actually use a *local* version of relative sensitivity.

4.2 Privacy Loss and Comparison with GM

Let us consider that we use the relative sensitivity as a local sensitivity to set the noise level for disclosing output $\mathcal{A}(x)$. In this case, guaranteeing $(\alpha, \varepsilon_\star)$-Rényi-DP when releasing output $\mathcal{A}(x)$ requires setting the noise as $\sigma^2_{\text{abs}} = \frac{\alpha}{2\varepsilon_\star}(\eta^2 \|\mathcal{R}(x)\|^2 + R^2_{\text{rel}})$. Unfortunately, as explained before, local sensitivity does *not* guarantee differential privacy. If we were to use the same level of noise in the Relative Gaussian mechanism, this would correspond to $\gamma = \frac{\alpha\eta^2}{2\varepsilon_\star}$, and $\sigma^2 = \gamma\eta^{-2}R^2_{\text{rel}}$. In particular, Theorem 1 tells us that this choice actually guarantees RDP with parameter:

$$\varepsilon = \frac{\varepsilon_\star}{1-\eta(\alpha-1)(2+\eta)} + \frac{\alpha d\eta^2(2+\eta)^2(1+\eta)^2}{2(1-\eta(\alpha-1)(2+\eta))}. \quad (6)$$

The first term corresponds to the target privacy level $\varepsilon_\star$, weighted by a factor which is bounded by 2 as long as $2\eta(2+\eta)(\alpha-1) \leq 1$, and goes to 1 as η decreases (for a fixed α). The second term corresponds to the *privacy loss* incurred by using the norm of the current output to set the noise level. Note that we see from Theorem 1 that this term is independent of γ and σ^2: it corresponds to a baseline loss that is paid for using a local form of sensitivity. We would get rid of this term if all possible queries $\mathcal{R}(x)$ had the same norm, and this norm was public. However, this is a very strong assumption that generally does not hold (or requires very high absolute sensitivity bounds R^2_{abs}). Note that it is tempting to use another output $\mathcal{R}(y)$ to set the noise level, and thus decorrelate the noise level from the specific input that we consider. However, $\mathcal{R}(y)$ would not be independent from $\mathcal{R}(x)$ since $x \sim y$.

This privacy loss term explains why using arbitrary large γ does not lead to arbitrary good privacy guarantees. However, as long as η is small enough compared to α, *the privacy loss is purely additive*. This means that if the dimension d is not too large ($d \leq \gamma^{-1}/36$ for $\eta \leq 1$, more for small η), we are safe using the relative Gaussian mechanism with minimal privacy overhead. Note that the d term comes from the fact that we use Gaussian noise, and other noise distributions might incur other dependencies [Bun and Steinke, 2019].

Standard vs. Relative Gaussian Mechanism. This "privacy loss" point of view allows us to reason about the noise introduced by the Relative Gaussian Mechanism, versus the standard one. Indeed, let us neglect the additive privacy loss term. In this case, as argued in the previous paragraph, the privacy guarantees are comparable to the standard Gaussian mechanism with local sensitivity $\eta^2 \|\mathcal{R}(x)\|^2 + R^2_{\text{rel}}$. In particular, which mechanism yields the best utility (less noise for a given privacy level) depends on which sensitivity bound is the tightest. If $R^2_{\text{abs}} \geq \eta^2 \max_x \|\mathcal{R}(x)\|^2 + R^2_{\text{rel}}$ then the relative Gaussian mechanism is always better, because it will lead to similar guarantees with less noise overall. Otherwise, some outputs might be noised more with one mechanism and less with another. This is highly application-specific, as it is conditioned by the structure of the outputs.

Tightness. One natural question that arises is the tightness of Theorem 1. Due to the parallel with local sensitivity, the first term is tight up to the (usually small) multiplicative factor. The second term is also tight up to a factor 1/2 in the limit of small η, thanks to the tightness of the inequality used to obtain it. We discuss this in Appendix A.4.

5 THE SPECIAL CASE OF GRADIENT DESCENT

An important application of $\text{RGM}_{\gamma,\sigma}$ is private Gradient Descent (GD). In this section, we describe it in the quadratic case, for which we estimate the values of η and R_{rel} and propose a clipping-like procedure.

5.1 RGM for Gradient Descent

We consider a function $f : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}$, where $\mathcal{D}$ is a set of possible datasets. Assume that the gradients of f (w.r.t. its first parameter) verify the relative sensitivity assumption. Given a dataset $D \in \mathcal{D}$, we can then privately minimize function f using the following private gradient descent algorithm, where $\gamma, \sigma > 0$ are parameters of the RGM, and $\tau > 0$ is a step size:

$$\theta_{t+1} = \theta_t - \tau \text{RGM}_{\gamma,\sigma}(\mathcal{R}_{\theta_t})(D), \quad (7)$$
$$\text{where } \mathcal{R}_{\theta_t}(D) = \nabla f(\theta_t; D)\,.$$

We remark that the form of $\text{RGM}_{\gamma,\sigma}$'s noise allows a tight analysis of the utility, as shown below.

Theorem 2. *Let $f : \mathbb{R}^d \times \mathcal{D} \to \mathbb{R}$ be μ-strongly-convex and L-smooth in its first parameter (see,* e.g., *Nesterov et al. [2018]). Let $D \in \mathcal{D}$ be a dataset, and $\theta_\star$ be the minimizer of $f(\cdot; D)$. Assume that f's gradients satisfy (η, R_{rel})-relative sensitivity, and that γ, σ are set as in Theorem 1. Then if $\tau \leq (L+\gamma)^{-1}$, the iterates obtained*

by (7) satisfy, for all $t \geq 0$,

$$\mathbb{E}\left[\|\theta_t - \theta_\star\|^2\right] \leq (1-\tau\mu)^t \|\theta_0 - \theta_\star\|^2 + \frac{\tau\sigma^2}{\mu}. \quad (8)$$

The proof, along with a similar result in the general convex case are in Appendix B.1. It relies on the fact that standard GD proofs already require to bound a $\|\nabla f(\theta_t; D)\|^2$ term by choosing an appropriate step-size, so the norm-scaled noise term can be accounted for in the same way, by only (slightly) decreasing the step size. Contrary to the usual DP-GD, which privatizes gradients using GM_σ, the variance term is $\frac{\tau\sigma^2}{\mu}$, where σ^2 now depends on R_{rel} which can be much smaller than the absolute sensitivity. In the remainder of this section, we exhibit settings in which the gradients verify relative sensitivity.

5.2 Relative sensitivity for linear regression

We now consider the specific case of quadratic objectives. More specifically, f is of the form

$$f(\theta; X, y) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left\|X_i^\top \theta - y_i\right\|^2 + \frac{\mu_{\mathrm{reg}}}{2}\|\theta\|^2, \quad (9)$$

where $X \in \mathbb{R}^{d\times n}$ and $y \in \mathbb{R}^n$. We denote by $(X_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ the i-th data record (*i.e.,* the i-th column of X and i-th element of y). Let $(X', y') \sim (X, y)$ be a dataset that, w.l.o.g, only differs from (X, y) on its first record (X_0, y_0). In the following, we denote $f = f(\cdot; D)$ and $f' = f(\cdot; D')$.

Let $I_d \in \mathbb{R}^{d\times d}$ be the identity matrix and let us denote $A = \frac{1}{n}XX^\top + \mu_{\mathrm{reg}} I_d \in \mathbb{R}^{d\times d}$, $A_i = X_i X_i^\top \in \mathbb{R}^{d\times d}$, $b = \frac{1}{n}Xy$, and $b_i = \frac{1}{n}X_i y_i$ (and similarly for A' and b'). For $\theta \in \mathbb{R}^d$, $\nabla f(\theta) = A\theta - b$ and $\nabla f'(\theta) = A'\theta - b'$. The difference between two gradients writes:

$$\begin{aligned} n^2\|\nabla f(\theta) - \nabla f'(\theta)\|^2 &= \|(A_0 - A_0')\theta - b_0 + b_0'\|^2 \\ &= \left\|(A_0 - A_0')A^{-1}(A\theta - b) + (A_0 - A_0')A^{-1}b - b_0 + b_0'\right\|^2 \\ &\leq 3\left[\left\|A_0A^{-1}\right\|^2 + \left\|A_0'A^{-1}\right\|^2\right]\|\nabla f(\theta)\|^2 \quad (10) \\ &\quad + 3\left\|\nabla f_0(A^{-1}b) - \nabla f_0'(A^{-1}b)\right\|^2, \end{aligned}$$

with $\|A\| = \lambda_{\max}(A)$ being the 2-norm for matrices. This bound hints at relative sensitivity, and we now discuss the corresponding η and R_{rel} terms. We first define $L, \mu > 0$ the bounds on the largest and smallest eigenvalues of all A, *i.e.*, $L \geq \|A\|$, and $\mu \leq \lambda_{\min}(A)$. Then, denote $\kappa = L/\mu$.

General functions. All derivations above consider quadratic objectives. Yet, similar terms with corresponding intuitions can be derived for arbitrary convex functions, as presented in Appendix B.2.

The absolute term R_{rel}. By using the relative framework, we only have to bound the difference between gradients at $A^{-1}b$ rather than all points, the rest being handled by the norm scaling. When an approximation of $A^{-1}b$ is known, this gives much tighter guarantees. Otherwise, this term writes: $\left\|(A_0 - A_0')A^{-1}b - b_0 + b_0'\right\| \leq \left\|(A_0 - A_0')A^{-1}\right\|\|b\| + \|b_0 - b_0'\|$. In the end, one only needs to control the norm of b and $b_0 - b_0'$, which can be done via clipping.

The relative term η. The term in front of the gradient norm can be bounded as:

$$\left\|A_0A^{-1}\right\|^2 = \|X_0\|^2 X_0^\top A^{-2}X_0 \leq \kappa(X_0^\top A^{-1}X_0)^2. \quad (11)$$

While a direct bounding writes $\left\|A_0A^{-1}\right\|^2 \leq \max_i \|A_i\|^2/\mu^2$, we will see that for well-behaved distributions, the bound only depends linearly on κ, or even not at all. This is the case for instance when the data is orthogonal, as shown in the following example.

Orthogonal data. Relative sensitivity can be easily bounded for orthogonal data, *i.e.* if either $X_i^\top X_j = \|X_i\|^2$ or $X_i^\top X_j = 0$. Consider that at least half of the dataset is fixed, and contains all different X_i in equal proportions (so, d^{-1}). In this case, $A \succcurlyeq \frac{1}{2d}\sum_{i=1}^{d} X_iX_i^\top$ so $\|X_i\|^2 X_i^\top A^{-2}X_i \leq 2d$. Note that the relative sensitivity is independent of the scale of each X_i. See Appendix B.3 for more detailed derivations.

However, one can remark that we have made an extra assumption on top of data orthogonality, which is that half of the dataset is assumed to be fixed. This is because if in the dataset X one dimension only has a single data point and this point is removed, then A will not be invertible anymore. This could be fixed through regularization but would lose independence from κ. While this is unavoidable in the worst case, removing a point only mildly affects the covariance matrix A when the data is "well-behaved", leading to small η. Fortunately, there exists a mechanism designed specifically for leveraging such properties, which is Propose-Test-Release [Dwork and Lei, 2009]. We will see in the next section how to instantiate it in our case.

5.3 Enforcing relative L2 sensitivity

Sensitivity bounds are often hard to evaluate, and it is generally desirable to enforce them in practical applications, for instance using gradient clipping to restrict the absolute sensitivity. In the quadratic case, the absolute term R_{rel} can be controlled by clipping the b_i's. Yet, $\left\|(A_0 - A_0')A^{-1}\right\|$ should also be controlled, hopefully without a too strong dependence on the conditioning of A. This can be done by clipping the X_i, *i.e.*, shrinking their norm when it is too large.

Proposition 1 (Clipping). *Let $C \in \mathbb{R}^{d\times d}$, $R_c \in \mathbb{R}$, and let $\tilde{X}$ be the clipped dataset, obtained as $\tilde{X}_i = R_c X_i / \max(R_c, (\|X_i\|^2 X_i^\top C^{-2} X_i)^{\frac{1}{4}})$. If $\tilde{A} = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{X}_i^\top + \mu_{\text{reg}} I_d \succcurlyeq \rho C$ for some $\rho > 0$, then $\tilde{X}$ verifies relative sensitivity with constant $\eta^2 = \frac{6R_c^4}{\rho^2 n^2}$.*

The proof directly follows from (10), (11) and $\tilde{A} \succcurlyeq \rho C$. Note that to guarantee cancellations as in (10), how X_i is clipped should be independent of X (and thus A), which is why we introduce matrix C. This result shows that given a dataset, we can enforce relative sensitivity bounds given sufficient clipping. Interestingly, using $C = I_d$ will just enforce a bound of $\tilde{\kappa}^2$ (the conditioning of the clipped covariance), but any C that better captures the structure of the data will improve this bound. Thus, even loose estimates of the covariance can be used, whether they come from rough private approximations or expert data knowledge.

There are two substantial caveats to this result: ρ is unknown, and $\tilde{A} \succcurlyeq \rho C$ needs to hold regardless of the dataset, and not for the specific dataset that we consider. While these seem to be particularly strong restrictions, if we can privately check that a given value of ρ works for our specific dataset, then we can still guarantee differential privacy thanks to the Propose-Test-Release (PTR) framework [Dwork and Lei, 2009].

Proposition 2. *Let $\rho > 0$, $C \in \mathbb{R}^{d\times d}$. Let Δ be the number of points $\tilde{X}_i$ that need to be changed from $\tilde{X}$ to obtain $\tilde{X}'$ such that $\tilde{A}' \preccurlyeq \rho C$. If $\tilde{A} - \rho C \preccurlyeq 0$ then $\Delta = 0$. Otherwise, $\Delta \geq \Delta_+ = \min\{|I_R|, \sum_{i\in I_R} \tilde{X}_i^\top (\tilde{A} - \rho C)^{-1} \tilde{X}_i \geq n\}$. In particular, for any (ε, δ), PTR writes: (i) Propose a bound ρ. (ii) Compute $\hat{\Delta} = \Delta_+ + \text{Lap}(\varepsilon^{-1})$. (iii) If $\hat{\Delta} \leq -\log(\delta)/\varepsilon$, return $\emptyset$, otherwise, run (multiple instances of) $\text{RGM}_{\gamma,\sigma}$ using ρ to compute the relative sensitivity parameters (which guarantees $(\varepsilon_{\text{RGM}}, \delta_{\text{RGM}})$-DP). This mechanism is $(\varepsilon + \varepsilon_{\text{RGM}}, \delta + \delta_{\text{RGM}})$-DP.*

The proof can be found in Appendix B.4, and more details on PTR can be found, *e.g.,* in Vadhan [2017, Section 3.2]. This proposition means that the PTR framework can be implemented "efficiently": instead of testing all combinations, it is enough to compute and sort the $\tilde{X}_i^\top (\tilde{A} - \rho C)^{-1} \tilde{X}_i$, and check how many must be summed before reaching n. In the end, the procedure to enforce a given η writes:

- Input arbitrary $C \in \mathbb{R}^{d\times d}$, $R_c > 0$. Clip all X_i to obtain $\tilde{X}_i$ so that $\|\tilde{X}_i\|^2 \tilde{X}_i^\top C^{-2} \tilde{X}_i \leq R_c^4$.
- Propose $\rho > 0$. Use PTR to privately test whether $\tilde{A} \succcurlyeq \rho C$ (Proposition 2). If not, abort.
- If it does, then $\tilde{A}^{-2} \preccurlyeq \rho^{-2} C^{-2}$, and so $\eta^2 = \frac{6R_c^4}{\rho^2 n^2}$ can be chosen to run DP-GD with $\text{RGM}_{\gamma,\sigma}$.

Interestingly, our approach combines several methods intended to overcome the problems of local sensitivity. We use the PTR framework to estimate the relative sensitivity parameter η. This is possible because local sensitivity is rather smooth in θ, and we can efficiently bound how far X is from a high local sensitivity dataset.

5.4 Gaussian features

We have just shown that a procedure based on PTR allows to obtain practical differential privacy guarantees for well-behaved datasets. Yet, a major question remains: *how to choose C, R_c and ρ?* First note that ρ could be obtained by a binary search (paying at most log factors), and several values of clipping thresholds R_c could be tried out. A key feature here is that **we only use PTR once at the beginning of training**, and do not need to rerun it at each epoch. Thus, many guesses can be tried out with limited impact on the overall privacy cost, which should be dominated by the actual gradient descent iterations. However, we give in this section a choice of C, R_c and ρ that works with high probability if $X_i \sim \mathcal{N}(0, \Sigma)$.

Proposition 3. *Let $X \in \mathbb{R}^{d\times n}$ such that its columns X_i are drawn* i.i.d. *from $\mathcal{N}(0, \Sigma)$. Let us choose $C = \Sigma$, so that $\tilde{X}_i = R_c X_i / \max(R_c, X_i^\top C^{-1} X_i)^{\frac{1}{2}})$. Let $\nu > 0$, $n \geq 4\log(2d/\nu)/9$ and $\mu_{\text{reg}} = 4\|\Sigma\| R_c^2 \sqrt{\frac{\log(2d/\nu)}{n}}$. Then, with probability at least $1-\nu$, the data sampled is such that*

$$\rho = \frac{1}{2d(2\pi)^{d/2}} \int_{\mathbb{R}^d} \min(\|u\|^2, R_c^2) e^{-\frac{\|u\|^2}{2}} \mathrm{d}u, \qquad (12)$$

is always accepted by the PTR procedure from Proposition 2 as long as $n \geq -\frac{\log(\delta) R_c^2}{2\varepsilon}$, leading to $\eta^2 = \frac{6\kappa_\Sigma R_c^4}{\rho^2 n^2}$, where κ_Σ is the conditioning of Σ.

The proof can be found in Appendix B.5. Note that we clip to enforce the stronger condition $\tilde{X}_i^\top C^{-1} \tilde{X}_i \leq R_c^2$, as this clipping preserves nice properties for Gaussian data. While both kinds of clipping are valid in practice, such a simple analytic expression for ρ might be harder to obtain when clipping for $\|\tilde{X}_i\|^2 \tilde{X}_i^\top C^{-2} \tilde{X}_i \leq R_c^4$. However, we pay a κ_Σ factor in the relative sensitivity constant by doing so (but recover part of it since the proposed ρ is independent of κ_Σ).

6 DISTRIBUTED TRAINING UNDER LOCAL DP

An interesting use-case for the Relative Gaussian Mechanism is distributed training in the local model of DP. Several nodes participate in a global training procedure, minimizing a shared objective. To this end, they periodically exchange (private) gradients, and the key

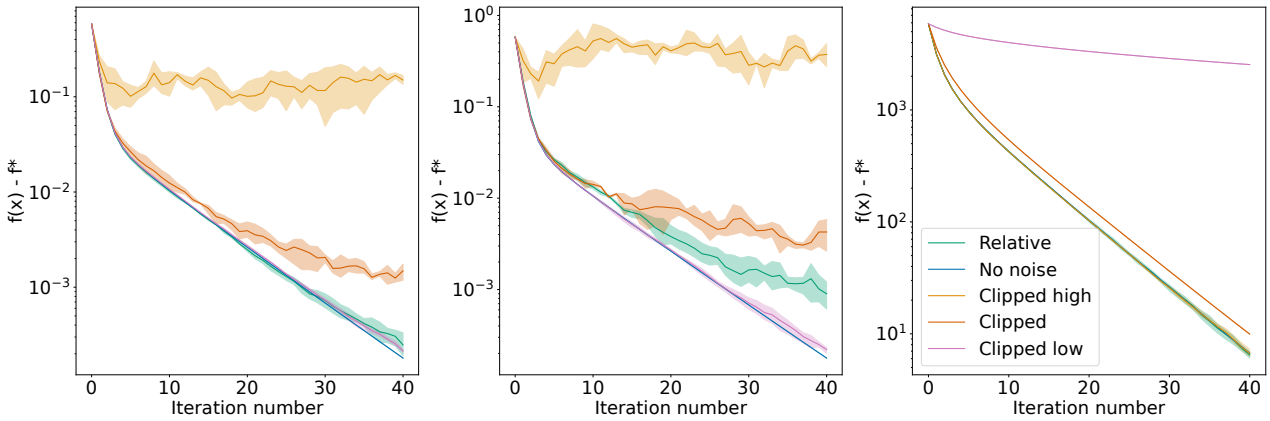

Figure 1: Utility of several private gradient descent algorithms with equivalent RDP guarantees. (Left): 'Random', (Middle): 'label', (Right): 'bias'. Shaded areas are min/max values over 3 runs.

bottleneck is thus *communication*. Yet, privatizing gradients is a *local* procedure, that each node completes on its own. In particular, it is reasonable to assume here that nodes can locally and efficiently estimate (η, R_{rel}) without having to solve the global optimization problem (which would require communication).

We argue that, contrary to the $\text{RGM}_{\gamma,\sigma}$, it is impossible to effectively use the GM_σ without knowledge from other nodes. To illustrate this, consider the simple example where two nodes have respective objectives $f_1(\theta) = \alpha \|\theta\|^2$, and $f_2(\theta) = \beta \|\theta - b\|^2$. There, the sensitive records to keep private are $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ and $\beta \in [\beta_{\min}, \beta_{\max}]$. Both nodes can easily compute their local parameters $\eta_1 = \alpha_{\min}/\alpha_{\max}$, $\eta_2 = \beta_{\min}/\beta_{\max}$, and $R_{\text{rel}} = 0$. Then, they can directly use the $\text{RGM}_{\gamma,\sigma}$. Consider now that nodes use GM_σ with gradient clipping instead, and that for this particular instance, $\alpha = \beta = 1$. In order not to bias the objective, the clipping threshold for node 1 would need to be set to a least $\|b\|$, which is equal to the norm of the gradient of f_1 evaluated at the global optimum. However, node 1 has no knowledge of b, and has thus no way of setting a relevant clipping threshold without exchanging information with node 2. In this simple illustrative example, it would of course be enough to just exchange an approximation of b, which has a reasonable cost. Nonetheless, this highlights that setting a relevant clipping threshold in general requires knowledge of the solution to the global problem, which is generally unavailable as it depends on all nodes' data.

We illustrate this with linear regression experiments on the ijcnn1 dataset [Chang and Lin, 2001], with $N = 141691$, and $d = 22$. We consider ridge linear regression, so f is of the form of (9) where the y_i correspond to the binary classification labels. We set the regularization parameter $\mu_{\text{reg}} = 0.03$, and RDP parameters $\alpha = 2$ and $\varepsilon = 0.1$. In order to avoid having to decide a clipping threshold for GM_σ, we automatically set the threshold as the maximum of the individual stochastic gradients at optimum (to avoid bias). We also run experiments with $c_{\text{high}} = 10c$ ('Clip high') and $c_{\text{low}} = c/10$ ('Clip low'). We compare this to vanilla gradient descent without noise and $\text{RGM}_{\gamma,\sigma}$, where we set R_c just high enough so that no point is actually clipped, and take $C = \tilde{A}$ and $\rho = 1/2$ for PTR, so that the condition on Δ can be reduced to $\Delta = n/2R_c \approx 780$, leading to $\delta = \exp(-\varepsilon\Delta) \approx 10^{-34}$. The results are shown in Figure 1. Code is available in supplementary material, and the precise experimental details can be found in Appendix C. Additional results on a different dataset can be found in Appendix D.

We study 3 different data splits: (i) 'Random' (left plot): the data is split randomly across the two nodes. (ii) 'label' heterogeneity (center): we sample 50 points at random for each node, and then all positive labels are assigned to one node and all negative to the other. (iii) 'bias' (right): we add a bias B to the objective to recreate (with more complex data) the simple example discussed above.

We observe that, although there is generally always a clipping threshold that works well, this threshold is problem-dependent. Small thresholds work well for homogeneous objectives or with label heterogeneity, whereas larger clipping thresholds handle bias heterogeneity better. Therefore, the clipping threshold needs to be tuned, which requires more communication, and incurs additional privacy leaks. On the contrary, the Relative Gaussian mechanism is, in this case, able to deal with heterogeneity without such tuning.

Note that the main contribution of the paper is to give a series of theoretical results that allow to properly instantiate smooth-sensitivity ideas, in particular for a gradient descent setting. The experiments are

meant to illustrate the method and show that it can be competitive with clipping, not to show that the baseline RGM mechanism as presented in this paper gives state-of-the-art results for DP gradient descent. This would require to combine our approach with many additional tweaks (including subsampling), along with heavy tuning, and is out of the scope of this paper. Yet, our results pave the way for such applications by providing a framework that allows to evaluate relative sensitivity parameters by using PTR.

7 CONCLUSION

We introduced the relative L2 sensitivity, a generalization of the usual L2 sensitivity which depends on the norm of the query. We designed the Relative Gaussian mechanism ($\text{RGM}_{\gamma,\sigma}$), a mechanism that exploits this sensitivity assumption to adapt the level of noise to the norm of the query $\mathcal{R}(x)$, and proved tight privacy guarantees. We then applied $\text{RGM}_{\gamma,\sigma}$ to private gradient descent and proposed a framework based on clipping features and PTR to enforce relative L2 sensitivity. An important research direction is to improve the mechanism to reduce the dimension dependence of the privacy guarantees. This might be achieved using relative sensitivity assumptions beyond Gaussian noise [Bun and Steinke, 2019].

Acknowledgements

This work was supported by the Inria Exploratory Action FLAMED and by the French National Research Agency (ANR) through grant ANR-20-CE23-0015 (Project PRIDE), ANR-20-CHIA-0001-01 (Chaire IA CaMeLOt) and ANR 22-PECY-0002 IPOP (Interdisciplinary Project on Privacy) project of the Cybersecurity PEPR. We thank Thomas Steinke and anonymous reviewers of a previous version of this paper for suggesting the use of the PTR mechanism to test that relative sensitivity holds.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016a.

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, October 2016b. Association for Computing Machinery. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Kareem Amin, Alex Kulesza, Andres Munoz, and Sergei Vassilvtiskii. Bounding User Contributions: A Bias-Variance Trade-off in Differential Privacy. In *Proceedings of the 36th International Conference on Machine Learning*, pages 263–271. PMLR, May 2019. URL https://proceedings.mlr.press/v97/amin19a.html. ISSN: 2640-3498.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially Private Learning with Adaptive Clipping. In *Advances in Neural Information Processing Systems*, volume 34, pages 17455–17466. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/91cff01af640a24e7f9f7a5ab407889f-Abstract.html.

Hilal Asi, Karan Chadha, Gary Cheng, and John Duchi. Private optimization in the interpolation regime: faster rates and hardness results. In *Proceedings of the 39th International Conference on Machine Learning*, pages 1025–1045. PMLR, June 2022. URL https://proceedings.mlr.press/v162/asi22a.html.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, October 2014. ISBN 978-1-4799-6517-5. doi: 10.1109/FOCS.2014.56. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6979031.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private Stochastic Convex Optimization with Optimal Rates. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3bd8fdb090f1f5eb66a00c84dbc5ad51-Abstract.html.

Victor-Emmanuel Brunel and Marco Avella-Medina. Propose, test, release: Differentially private estimation with high probability. *arXiv preprint arXiv:2002.08774*, 2020.

Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual*

ACM SIGACT Symposium on Theory of Computing, pages 74–86, 2018.

Chih-Chung Chang and Chih-Jen Lin. Ijcnn 2001 challenge: Generalization ability and text decoding. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 2, pages 1031–1036. IEEE, 2001.

Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding Gradient Clipping in Private SGD: A Geometric Perspective. *arXiv:2006.15429 [cs, math, stat]*, June 2020. URL http://arxiv.org/abs/2006.15429.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407, 2014.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U. Stich. Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees, May 2023. URL http://arxiv.org/abs/2305.01588.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

Venkatadheeraj Pichapati, Ananda Theertha Suresh, Felix X. Yu, Sashank J. Reddi, and Sanjiv Kumar. AdaCliP: Adaptive Clipping for Private SGD. *arXiv:1908.07643 [cs, stat]*, October 2019. URL http://arxiv.org/abs/1908.07643.

Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248, Austin, TX, USA, December 2013. IEEE.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *International Conference on Machine Learning*, pages 21828–21863. PMLR, 2022.

Salil Vadhan. The complexity of differential privacy. *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, pages 347–450, 2017.

Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in neural information processing systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f337d999d9ad116a7b4f3d409fcc6480-Paper.pdf.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 93–103. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/40.pdf.

Xiaodong Yang, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Normalized/Clipped SGD with Perturbation for Differentially Private Non-Convex Optimization, June 2022. URL http://arxiv.org/abs/2206.13033.

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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No] Experiments are small-scale and were performed on a personal laptop.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes] For the ijcnn dataset.

(b) The license information of the assets, if applicable. [No]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

The appendix is organized as follows. Section A contains the full proofs for the general $\mathrm{RGM}_{\gamma,\sigma}$, and in particular Theorem 1. Section B contains the proofs for the results that justify using relative assumptions when minimizing quadratics, and Section C contains the detailed experimental setting, with all the elements needed to reproduce the experiments. The code itself can be found in supplementary material.

A Proofs for the Relative Gaussian Mechanism

A.1 Bounding the noise scale ratio

We start by the following simple lemma, that will allow us to bound the domain of admissible α.

Lemma 1. *If* $\sigma^2 \geq \frac{\gamma(1+\eta^{-1})}{2\eta+\eta^2} R_{\mathrm{rel}}^2$, *then* $(1+\eta)^{-2} \leq \frac{\gamma\|\mathcal{R}(x)\|^2+\sigma^2}{\gamma\|\mathcal{R}(y)\|^2+\sigma^2} \leq (1+\eta)^2$.

Proof.

$$\begin{aligned}
\frac{\gamma\left\|\mathcal{R}(x)\right\|^2+\sigma^2}{\gamma\left\|\mathcal{R}(y)\right\|^2+\sigma^2} &= \frac{\gamma\left\|\mathcal{R}(y)+\mathcal{R}(x)-\mathcal{R}(y)\right\|^2+\sigma^2}{\gamma\left\|\mathcal{R}(y)\right\|^2+\sigma^2} \\
&\leq \frac{\gamma(1+\eta)\left\|\mathcal{R}(y)\right\|^2+\gamma(1+\eta^{-1})\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|^2+\sigma^2}{\gamma\left\|\mathcal{R}(y)\right\|^2+\sigma^2} \\
&\leq \frac{\gamma(1+\eta)\left\|\mathcal{R}(y)\right\|^2+\gamma(1+\eta^{-1})(\eta^2\left\|\mathcal{R}(y)\right\|^2+R_{\mathrm{rel}}^2)+\sigma^2}{\gamma\left\|\mathcal{R}(y)\right\|^2+\sigma^2}.
\end{aligned}$$

The result then follows from using the bound on σ^2 to factor $\gamma\left\|\mathcal{R}(y)\right\|^2+\sigma^2$ in the numerator. The other side (lower bound) is obtained by inverting $\mathcal{R}(x)$ and $\mathcal{R}(y)$. □

A.2 Rényi divergence of two Gaussians

Recall that for $\alpha > 1$ and two distributions, P and Q, the Rényi divergence is

$$\mathcal{D}_\alpha(P||Q) = \frac{1}{\alpha-1}\log\int\frac{P(x)^\alpha}{Q(x)^\alpha}dQ(x) = \frac{1}{\alpha-1}\log\int\frac{P(x)^\alpha}{Q(x)^{\alpha-1}}dx \ .$$

Lemma 2. *Let* P *and* Q *be Gaussian distributions of dimension* d *centered in* μ_1 *and* μ_2 *with variance* $\sigma_1^2 I_d$ *and* $\sigma_2^2 I_d$. *Then, assuming that* $\alpha\sigma_2^2+(1-\alpha)\sigma_1^2 > 0$,

$$\mathcal{D}_\alpha(P||Q) = \frac{\alpha\left\|\mu_1-\mu_2\right\|^2}{2(\alpha\sigma_2^2+(1-\alpha)\sigma_1^2)} + \frac{d}{\alpha-1}\log\left(\frac{\sigma_1^{1-\alpha}\sigma_2^\alpha}{\sqrt{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}}\right) \ . \tag{13}$$

Remark that when $\sigma_1 = \sigma_2$, we recover the divergence of the standard Gaussian mechanism.

Proof.

$$\mathcal{D}_\alpha(P||Q) = \frac{1}{\alpha-1}\log\sqrt{\frac{\sigma_2^{2d(\alpha-1)}}{(2\pi)^d\sigma_1^{2d\alpha}}}\int\exp\left(-\frac{\alpha\left\|u-\mu_1\right\|^2}{2\sigma_1^2}-\frac{(1-\alpha)\left\|u-\mu_2\right\|^2}{2\sigma_2^2}\right)du \ .$$

We first compute the one-dimensional integral

$$\begin{aligned}
&\int_{-\infty}^{+\infty}\exp\left(-\frac{\alpha(u-\mu_1)^2}{2\sigma_1^2}-\frac{(1-\alpha)(u-\mu_2)^2}{2\sigma_2^2}\right) \\
&= \int_{-\infty}^{+\infty}\exp\left(-\left(\frac{\alpha}{2\sigma_1^2}+\frac{1-\alpha}{2\sigma_2^2}\right)u^2+\left(\frac{\alpha\mu_1}{\sigma_1^2}+\frac{(1-\alpha)\mu_2}{\sigma_2^2}\right)u-\frac{\alpha\mu_1^2}{2\sigma_1^2}-\frac{(1-\alpha)\mu_2^2}{2\sigma_2^2}\right) \\
&= \int_{-\infty}^{+\infty}\exp\left(-\left(\frac{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}{2\sigma_1^2\sigma_2^2}\right)u^2+\left(\frac{\alpha\mu_1\sigma_2^2+(1-\alpha)\mu_2\sigma_1^2}{\sigma_1^2\sigma_2^2}\right)u-\frac{\alpha\mu_1^2\sigma_2^2+(1-\alpha)\mu_2^2\sigma_1^2}{2\sigma_1^2\sigma_2^2}\right) \ .
\end{aligned}$$

Now, since $\int_{-\infty}^{+\infty} \exp(-(au^2+bu+c))du = \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a} - c\right)$, we have after simplification, **and assuming** $\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2 > 0$,

$$\begin{aligned}\int_{-\infty}^{+\infty} \exp\left(-\frac{\alpha(u-\mu_1)^2}{2\sigma_1^2} - \frac{(1-\alpha)(u-\mu_2)^2}{2\sigma_2^2}\right) &= \sqrt{\frac{2\pi\sigma_1^2\sigma_2^2}{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}} \exp\left(-\frac{\alpha(1-\alpha)(\mu_1-\mu_2)^2}{2(\alpha\sigma_2^2+(1-\alpha)\sigma_1^2)}\right) \\ &= \frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}} \exp\left(-\frac{\alpha(1-\alpha)(\mu_1-\mu_2)^2}{2(\alpha\sigma_2^2+(1-\alpha)\sigma_1^2)}\right) \quad .\end{aligned}$$

Back to our divergence, we use the 1-dimensional computations on each dimensions and obtain

$$\begin{aligned}\mathcal{D}_\alpha(P||Q) &= \frac{1}{\alpha-1}\log\left(\sqrt{\frac{\sigma_2^{2d(\alpha-1)}}{(2\pi)^d\sigma_1^{2d\alpha}}}\prod_{j=1}^d \frac{\sqrt{2\pi}\sigma_1\sigma_2}{\sqrt{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}}\exp\left(-\frac{\alpha(1-\alpha)(\mu_{1,j}-\mu_{2,j})^2}{2(\alpha\sigma_2^2+(1-\alpha)\sigma_1^2)}\right)\right) \\ &= \frac{1}{\alpha-1}\log\left(\prod_{j=1}^d \frac{\sigma_1^{1-\alpha}\sigma_2^\alpha}{\sqrt{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}}\exp\left(-\frac{\alpha(1-\alpha)(\mu_{1,j}-\mu_{2,j})^2}{2(\alpha\sigma_2^2+(1-\alpha)\sigma_1^2)}\right)\right) \\ &= \frac{1}{\alpha-1}\log\left(\left(\frac{\sigma_1^{1-\alpha}\sigma_2^\alpha}{\sqrt{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}}\right)^d\exp\left(-\frac{\alpha(1-\alpha)\left\|\mu_1-\mu_2\right\|^2}{2(\alpha\sigma_2^2+(1-\alpha)\sigma_1^2)}\right)\right) \\ &= \frac{d}{\alpha-1}\log\left(\frac{\sigma_1^{1-\alpha}\sigma_2^\alpha}{\sqrt{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}}\right) + \frac{\alpha\left\|\mu_1-\mu_2\right\|^2}{2(\alpha\sigma_2^2+(1-\alpha)\sigma_1^2)} \quad .\end{aligned}$$

□

A.3 Privacy guarantees (Theorem 1)

We would now like to apply Lemma 2 where $\mu_1 = \mathcal{R}(x)$, $\sigma_1^2 = \gamma\left\|\mathcal{R}(x)\right\|^2 + \sigma^2$, $\mu_2 = \mathcal{R}(y)$ and $\sigma_2^2 = \gamma\left\|\mathcal{R}(y)\right\|^2 + \sigma^2$.

Verifying the condition on α. To this end, we first need to verify that $\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2 > 0$, or equivalently that $\sigma_2^2/\sigma_1^2 \geq 1 - \alpha^{-1}$. If applicable, Lemma 1 would directly give us that this is true as long as $(1+\eta)^{-2} \geq 1-\alpha^{-1}$, which leads to the bound:

$$\alpha^{-1} \geq 1 - (1+\eta)^{-2} = \frac{\eta(2+\eta)}{(1+\eta)^2}, \tag{14}$$

so in the end:

$$\alpha \leq \frac{(1+\eta)^2}{\eta(2+\eta)}. \tag{15}$$

This is equivalent to $\alpha - 1 \leq \frac{1}{\eta(2+\eta)}$, which is the condition from Theorem 1. In order to apply Lemma 1, we need to verify that

$$\sigma^2 \geq \gamma\frac{1+\eta^{-1}}{2\eta+\eta^2}R_{\text{rel}}^2 = \frac{\gamma R_{\text{rel}}^2}{\eta^2}\frac{1+\eta}{2+\eta}. \tag{16}$$

This is automatically verified for the choice of σ from Theorem 1 since

$$1 - \eta(\alpha-1) \geq 1 - (2+\eta)^{-1} = \frac{1+\eta}{2+\eta}. \tag{17}$$

Bounding the main term in (13). Now that we verified that we can apply Lemma 2, we can use it to bound the divergence. To bound the first term in (13), we start by recalling that, by Young's inequality,

$$\left\|\mathcal{R}(x)\right\|_2^2 \leq (1+\eta)\left\|\mathcal{R}(y)\right\|_2^2 + (1+\eta^{-1})\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 \quad . \tag{18}$$

Using this inequality, we can lower bound the denominator

$$\begin{aligned}
&\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2 && (19)\\
&= \sigma^2 + \gamma\alpha \left\|\mathcal{R}(x)\right\|_2^2 - \gamma(\alpha-1)\left\|\mathcal{R}(y)\right\|_2^2 && (20)\\
&\geq \sigma^2 + \gamma\alpha \left\|\mathcal{R}(y)\right\|_2^2 - \gamma(\alpha-1)(1+\eta)\left\|\mathcal{R}(y)\right\|_2^2 - \gamma(\alpha-1)(1+\eta^{-1})\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 && (21)\\
&= \sigma^2 + (\gamma\alpha - \gamma(\alpha-1)(1+\eta)\left\|\mathcal{R}(y)\right\|_2^2 - \gamma(\alpha-1)(1+\eta^{-1})\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 && (22)\\
&= \sigma^2 + \gamma(1-(\alpha-1)\eta)\left\|\mathcal{R}(y)\right\|_2^2 - \gamma(\alpha-1)(1+\eta^{-1})\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 && (23)
\end{aligned}$$

Now, remark that relative sensitivity gives $\left\|\mathcal{R}(y)\right\|_2^2 \geq \frac{1}{\eta^2}\left(\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 - R^2\right)$, which in turn yields

$$\begin{aligned}
&\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2 && (24)\\
&\geq \sigma^2 + \frac{\gamma}{\eta^2}(1-(\alpha-1)\eta)\left(\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 - R^2\right) - \gamma(\alpha-1)(1+\eta^{-1})\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 && (25)\\
&= \sigma^2 - \frac{\gamma}{\eta^2}(1-(\alpha-1)\eta)R^2 + \left(\frac{\gamma}{\eta^2}(1-(\alpha-1)\eta) - \gamma(\alpha-1)(1+\eta^{-1})\right)\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 \ . && (26)
\end{aligned}$$

We now use the condition on α, which is that $\alpha - 1 = \frac{1-\rho}{\eta(2+\eta)}$ for some $\rho < 1$. If we further assume that $\sigma^2 \geq \frac{\gamma}{\eta^2}(1-(\alpha-1)\eta)R^2$, which is notably the case when $\sigma^2 \geq \frac{\gamma R^2}{\eta^2}$, we obtain

$$\begin{aligned}
\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2 &\geq \left(\frac{\gamma}{\eta^2}(1-(\alpha-1)\eta) - \gamma(\alpha-1)(1+\eta^{-1})\right)\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 && (27)\\
&= \left(\frac{\gamma}{\eta^2} - \frac{\gamma}{\eta}(\alpha-1)\left(2+\eta\right)\right)\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 && (28)\\
&= \frac{\gamma\rho}{\eta^2}\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2 \ . && (29)
\end{aligned}$$

Putting everything back together, we get that for $\eta(2+\eta)(\alpha-1) \leq 1$, and $\sigma^2 \geq \gamma(1-(\alpha-1)\eta)R^2/\eta^2$:

$$\frac{\alpha\left\|\mathcal{R}(x)-\mathcal{R}(y)\right\|_2^2}{2(\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2)} \leq \frac{\alpha\eta^2}{2\gamma\rho} = \frac{\alpha\eta^2}{2\gamma}\frac{1}{1-\eta(2+\eta)(\alpha-1)} \ . \tag{30}$$

Bounding the log term in (13). Remark that, as a consequence of Lemma 1, we have that $\sigma_2^2 = r\sigma_1^2$, for some $r \leq (1+\eta)^2$. In particular, we have

$$\log\left(\frac{\sigma_1^{1-\alpha}\sigma_2^{\alpha}}{\sqrt{\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2}}\right) = \log\left(\frac{r^{\alpha/2}\sigma_1}{\sqrt{1-\alpha+\alpha r}\sigma_1}\right) = \frac{\alpha-1}{2}\log(r) - \frac{1}{2}\log\left(\frac{1+(r-1)\alpha}{r}\right) \ . \tag{31}$$

The two terms can be upper bounded using $\log(1+x) \leq x$ for $x \geq -1$,

$$\frac{\alpha-1}{2}\log(r) \leq \frac{\alpha-1}{2}(r-1) \ , \tag{32}$$

and $\log(1+x) \geq \frac{x}{1+x}$ for $x \geq -1$,

$$\begin{aligned}
-\frac{1}{2}\log\left(\frac{1+(r-1)\alpha}{r}\right) &= -\frac{1}{2}\log\left(1 + \frac{(\alpha-1)(r-1)}{r}\right) && (33)\\
&\leq -\frac{1}{2}\frac{(\alpha-1)(r-1)}{r+(\alpha-1)(r-1)} && (34)\\
&= -\frac{1}{2}\frac{(\alpha-1)(r-1)}{1+\alpha(r-1)} \ . && (35)
\end{aligned}$$

Overall, we obtain

$$\log\left(\frac{\sigma_1^{1-\alpha}\sigma_2^{\alpha}}{\sqrt{\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2}}\right) \leq \frac{(\alpha-1)(r-1)}{2}\left(1 - \frac{1}{1+\alpha(r-1)}\right) = \frac{\alpha(\alpha-1)(r-1)^2}{2(1+\alpha(r-1))} \ , \tag{36}$$

Since $(1+\eta)^{-2} \leq r \leq (1+\eta)^2$, we have that

$$|r-1| \leq \max\left((1+\eta)^2 - 1, 1-(1+\eta)^{-2}\right) = \max\left(\eta(2+\eta), \frac{\eta(2+\eta)}{(1+\eta)^2}\right) \leq \eta(2+\eta), \tag{37}$$

so that in the end:

$$\frac{d}{\alpha-1}\log\left(\frac{\sigma_1^{1-\alpha}\sigma_2^{\alpha}}{\sqrt{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}}\right) \leq \frac{\alpha d\eta^2(2+\eta)^2}{2\left(1-\alpha\frac{\eta(2+\eta)}{(1+\eta)^2}\right)} = \frac{\alpha d\eta^2(2+\eta)^2(1+\eta)^2}{2(1-(\alpha-1)\eta(2+\eta))} \ . \tag{38}$$

A.4 Tightness

As explained in the main text, the link with local sensitivity implies the tightness of the first term in Theorem 1 (the one which depends on γ). We now discuss the tightness of the second term, which comes from the log term in (13). In particular, we write (similarly to the previous section):

$$\log\left(\frac{\sigma_1^{1-\alpha}\sigma_2^{\alpha}}{\sqrt{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}}\right) = \frac{\alpha}{2}\log(1+u) - \frac{1}{2}\log\left(1+\alpha u\right) = g(u) \ , \tag{39}$$

with $u = r-1$. We would like to find an expression for $g(u)$ when $u \approx 0$, which means that $r \approx 1$ and so η is small. Differentiating with respect to u leads to:

$$g'(u) = \frac{\alpha}{2}\left(\frac{1}{1+u} - \frac{1}{1+\alpha u}\right). \tag{40}$$

Note that $g(0) = 0$, and $g'(0) = 0$, so we have to differentiate once again, leading to:

$$g''(u) = \frac{\alpha}{2}\left(\frac{\alpha}{(1+\alpha u)^2} - \frac{1}{(1+u)^2}\right). \tag{41}$$

In particular, $g''(0) = \frac{\alpha(\alpha-1)}{2}$, so when $r \approx 1$,

$$g(r) = \frac{\alpha(\alpha-1)}{4}(r-1)^2 + O((r-1)^3). \tag{42}$$

The leading term is the same expression as we had before, up to a factor $1/2$. In particular, the bounding from the previous subsection is tight up to a factor $1/2$.

Ideally, we would like to use this approximation as $g(r)$. In order to do this, we have to show that $g'''(r-1) \leq 0$ for all (α, η) we consider. Let us differentiate one last time, leading to:

$$g'''(u) = \alpha\left(\frac{1}{(1+u)^3} - \frac{\alpha^2}{(1+\alpha u)^3}\right). \tag{43}$$

One can remark that $g'''(u) \leq 0$ for $u \leq 0$. However, it can be that $g'''(u) > 0$ for some $u > 0$. More specifically, if α is large enough ($\alpha \geq 2$ for instance), then one can show that $g'''(u) \leq 0$ for the range of u that we consider (i.e., $u = r-1 \leq (1+\eta)^2 - 1 = \eta(2+\eta) \leq (\alpha-1)^{-1}$. Yet, this does not hold for all α, and $g'''((\alpha-1)^{-1}) > 0$ for $\alpha = 3/2$ for instance. This is why we keep the result which is off by a factor up to 2 for large α, but works for any value of (α, η) in our range.

A.5 Comparison with differential privacy

Corollary 1. *Assuming that* $\gamma \leq \frac{1}{4(2+\eta)^2\log(1/\delta)}$ *or* $d \geq 4\log(1/\delta)/(1+\eta)^2$*, and that* $0 < \delta \leq 1$*, the Relative Gaussian Mechanism is* (ϵ, δ)*-DP with*

$$\epsilon = \chi + 2\sqrt{\chi\log(1/\delta)}, \qquad \textit{where } \chi = \tfrac{\eta^2}{\gamma} + \eta^2 d(2+\eta)^2(1+\eta)^2. \tag{44}$$

Proof. Applying the standard conversion result from RDP to (ϵ, δ)-DP, we get that the Relative Gaussian Mechanism is (ϵ, δ)-DP for

$$\epsilon = \min_{1 \leq \alpha \leq \frac{(1+\eta)^2}{2\eta+\eta^2}} \left\{ g(\alpha) := \frac{\alpha\eta^2}{2\gamma} \frac{1 + \gamma d(2+\eta)^2(1+\eta)^2}{1 - \eta(2+\eta)(\alpha-1)} + \frac{\log(1/\delta)}{\alpha - 1} \right\}. \tag{45}$$

Restricting to $\alpha \leq 1 + \frac{1}{2\eta(2+\eta)} \leq \frac{(1+\eta)^2}{\eta(2+\eta)}$, we have $\eta(2+\eta)(\alpha-1) \leq 1/2$. We can therefore upper bound $g(\alpha)$ by

$$g(\alpha) \leq \frac{\alpha\eta^2}{\gamma} + \alpha\eta^2 d(2+\eta)^2(1+\eta)^2 + \frac{\log(1/\delta)}{\alpha-1} = \alpha\chi + \frac{\log(1/\delta)}{\alpha-1}. \tag{46}$$

This upper bound is minimal when $\chi = \frac{\log(1/\delta)}{(\alpha-1)^2}$, that is $\alpha = 1 + \sqrt{\frac{\log(1/\delta)}{\chi}}$. Using this value of α, we have

$$\epsilon \leq g(\alpha) \leq \chi + \sqrt{\chi \log(1/\delta)} + \sqrt{\chi \log(1/\delta)} \leq \chi + 2\sqrt{\chi \log(1/\delta)}. \tag{47}$$

Note that we could choose this value of α since either $\gamma \leq \frac{1}{4(2+\eta)^2 \log(1/\delta)}$ or $d \geq 4\log(1/\delta)/(1+\eta)^2$. These inequalities indeed implies that $\alpha = 1 + \sqrt{\frac{\log(1/\delta)}{\chi}} \leq 1 + \frac{1}{2\eta(2+\eta)}$, which is equivalent to

$$\frac{\chi}{4\eta^2(2+\eta)^2} = \frac{1}{4\gamma(2+\eta)^2} + \frac{(1+\eta)^2}{4} d \geq \log(1/\delta). \tag{48}$$

□

A.6 Links with the Gaussian Smooth Sensitivity

Gaussian Smooth Sensitivity [Bun et al., 2018, Section 5] is an analog of the smooth sensitivity framework [Nissim et al., 2007]. The main result writes as follows (changed to be compatible with our notations):

Proposition 4 (Bun et al. [2018]). *For all neighboring $x \sim y$, if $|\mathcal{R}(x) - \mathcal{R}(y)| \leq \Delta_f e^{g(x)/2}$ and $|g(x) - g(x')| \leq \Delta_g$, then if the Gaussian Smooth Sensitivity, GSS, writes $GSS(\mathcal{R})(x) = \mathcal{R}(x) + \mathcal{N}(0, e^{g(x)})$, then GSS satisfies $(\Delta_f^2 + \Delta_g^2, 1/(2\Delta_g))$-tCDP.*

Let us consider $g(x) = \log\left(\gamma \left\|\mathcal{R}(x)\right\|^2 + \sigma^2\right)$. Then, *GSS* is equivalent to $\mathrm{RGM}_{\gamma,\sigma}$, and we have that

$$|g(x) - g(y)| = \log\left(\frac{\gamma \left\|\mathcal{R}(x)\right\|^2 + \sigma^2}{\gamma \left\|\mathcal{R}(x')\right\|^2 + \sigma^2}\right) \tag{49}$$

In particular, since log is an increasing function, we can use Lemma 1 and write:

$$|g(x) - g(y)| \leq \log((1+\eta)^2) \leq 2\eta, \tag{50}$$

so that $\Delta_g = 2\eta$ is a valid bound. Then, if we choose $\sigma^2 = \gamma\eta^{-2}R_{\mathrm{rel}}^2$ (which satisfies the condition in Theorem 1), relative sensitivity gives us:

$$|\mathcal{R}(x) - \mathcal{R}(y)|^2 \leq \eta^2 \left\|\mathcal{R}(x)\right\|^2 + R_{\mathrm{rel}}^2 = \frac{\eta^2}{\gamma}\left(\gamma \left\|\mathcal{R}(x)\right\|^2 + \sigma^2\right) = \frac{\eta^2}{\gamma} e^{g(x)}. \tag{51}$$

In particular, this means that we can take $\Delta_f^2 = \frac{\eta^2}{\gamma}$.

Thus, Proposition 4 ensures that $\mathrm{RGM}_{\gamma,\sigma}$ guarantees $(4\eta^2 + \frac{\eta^2}{\gamma}, 1./(4\eta))$-tCDP.

Alternatively, we can directly transform Theorem 1 into a $(\frac{\eta^2}{\gamma} + \eta^2 d(2+\eta)^2/2], 1 + (2\eta(2+\eta))^{-1})$-tCDP result by eliminating α from the ε bound, using that $\alpha - 1 < (2\eta + \eta^2)^{-1}$, and so $[1 - \eta(\alpha-1)(2+\eta)]^{-1} \leq 2$.

In particular, we see that the guarantees we obtain are very similar for $\eta < 1$, which is the setting for which our results have been optimized (we have used inequalities that are tight for η small, but could also improve them for larger η). Indeed, our bound improves the guarantees obtained by *GSS* by a factor of 2 for small η.

We have seen that if $\mathcal{R}(x) \in \mathbb{R}$ (or equivalently $d = 1$), the results of Theorem 1 can be recovered using Proposition 4. However, Theorem 1 extends these results for $\mathcal{R}(x) \in \mathbb{R}^d$, which is essential to use relative sensitivity for gradient descent.

B Relative L2 sensitivity for releasing gradients

Before we start this section, we introduce a few notions, that will be useful in particular in subsections B.1 and B.2. The first is the notion of Bregman Divergence, which is defined for a function h and for points θ, θ' as:

$$D_h(\theta, \theta') = h(\theta) - h(\theta') - \nabla h(\theta')^\top (\theta - \theta'). \tag{52}$$

One can see that h is L-smooth and μ-strongly convex is equivalent to having for all $\theta, \theta' \in \mathbb{R}^d$:

$$\frac{\mu}{2} \|\theta - \theta'\|^2 \leq D_h(\theta, \theta') \leq \frac{L}{2} \|\theta - \theta'\|^2 . \tag{53}$$

The Bregman divergence also has nice properties w.r.t. convex conjugation. More specifically, the convex conjugate h^* of h is defined as $h^*(\theta) = \arg\max_u \theta^\top u - f(u)$. Then, it holds that:

$$D_h(\theta, \theta') = D_{h^*}(\nabla h(\theta'), \nabla h(\theta)). \tag{54}$$

Bregman divergences are often linked in convex optimization to the notion of *relative* smoothness, which generalizes regular smoothness. In particular, a function f is said to be L_{rel}-smooth with respect to h if for all $\theta, \theta' \in \mathbb{R}^d$,

$$D_f(\theta, \theta') \leq L_{\text{rel}} D_h(\theta, \theta'). \tag{55}$$

This recover standard smoothness when $h = \frac{1}{2}\|\cdot\|^2$, and will play a key role in showing our relative sensitivity assumption for gradient descent with general functions.

B.1 Utility (Proof of Theorem 2)

We provide below the proof of Theorem 2. We also prove that under the same assumptions (but this result can also be used for $\mu = 0$, *i.e.* $f(\cdot; D)$ is convex), we have that:

$$f(\bar{\theta}_t) - f(\theta_\star) \leq \frac{\|\theta_0 - \theta_\star\|^2}{2\tau t} + \frac{\tau \sigma^2}{2}, \tag{56}$$

where $\bar{\theta}_t = \frac{1}{t}\sum_{k=0}^{t-1} \theta_k$.

Proof. We write $f(\theta) = f(\cdot; D)$ for simplicity. In this case, for all $t \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{t+1} - \theta_\star\|^2\right] &\leq \|\theta_t - \theta_\star\|^2 - 2\tau \mathbb{E}\left[g_t^\top (\theta_t - \theta_\star)\right] + \tau^2 \mathbb{E}\left[\|g_t\|^2\right] \\
&= \|\theta_t - \theta_\star\|^2 - 2\tau \nabla f(\theta_t)^\top (\theta_t - \theta_\star) + \tau^2 \left[\|\nabla f(\theta_t)\|^2 + \mathbb{E}\left[\|\xi_t\|^2\right]\right] \\
&= \|\theta_t - \theta_\star\|^2 - 2\tau \nabla f(\theta_t)^\top (\theta_t - \theta_\star) + \tau^2 (1+\gamma) \|\nabla f(\theta_t)\|^2 + \tau^2 \sigma^2.
\end{aligned}$$

We now use the smoothness of f, which ensures that $\|\nabla f(\theta_t)\|^2 \leq 2L D_f(\theta_\star, \theta_t)$, and obtain:

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_\star\|^2\right] \leq \|\theta_t - \theta_\star\|^2 - 2\tau D_f(\theta_t, \theta_\star) - 2\tau(1 - (1+\gamma)\tau L) D_f(\theta_\star, \theta_t) + \tau^2 \sigma^2. \tag{57}$$

We can then use the μ-strong convexity of f, which yields $2D_f(\theta_t, \theta_\star) \geq \mu \|\theta_t - \theta_\star\|^2$, and so:

$$\mathbb{E}\left[\|\theta_{t+1} - \theta_\star\|^2\right] \leq (1 - \tau\mu) \|\theta_t - \theta_\star\|^2 - 2\tau(1 - (1+\gamma)\tau L) \left[f(\theta_t) - f(\theta_\star)\right] + \tau^2 \sigma^2. \tag{58}$$

By taking $\tau \leq [(1+\gamma)L]^{-1}$, using that $D_f \geq 0$ since f is convex, and chaining the inequalities, we obtain:

$$\mathbb{E}\left[\|\theta_t - \theta_\star\|^2\right] \leq (1 - \tau\mu)^t \|\theta_0 - \theta_\star\|^2 + \frac{\tau \sigma^2}{\mu}. \tag{59}$$

In the convex case ($\mu = 0$), we go back from (57), use the same step-size condition, and rewrite it as:

$$f(\theta_t) - f(\theta_\star) \leq \frac{\mathbb{E}\left[\|\theta_t - \theta_\star\|^2 - \|\theta_{t+1} - \theta_\star\|^2\right]}{2\tau} + \frac{\tau \sigma^2}{2}. \tag{60}$$

We now write (telescoping sum):

$$\frac{1}{t}\sum_{k=0}^{t-1} f(\theta_k) - f(\theta_\star) \leq \frac{\|\theta_0 - \theta_\star\|^2}{2\tau t} + \frac{\tau\sigma^2}{2}. \tag{61}$$

Equation (56) is obtained by convexity of f. □

B.2 Bounding relative sensitivity for general functions

Let f, f' be two functions on neighboring datasets (as defined in the main text), such that $f - f' = f_0 - f_0'$. Let f and f' be μ-strongly-convex, f_0 and f_0' be L-smooth, and f and f_0' be $L_{\rm rel}$-relatively smooth *w.r.t.* **both** f and f'. In this case, we have that:

$$\begin{aligned}\|\nabla f(\theta) - \nabla f'(\theta)\|^2 &= \frac{1}{n^2}\|\nabla f_0(\theta) - \nabla f_0'(\theta)\|^2 \\ &= \frac{1}{n^2}\|\nabla f_0(\theta) - \nabla f_0'(\theta_\star) - [\nabla f_0'(\theta) - \nabla f_0'(\theta_\star)] + \nabla f_0(\theta_\star) - \nabla f_0'(\theta_\star)\|^2 \\ &= \frac{3}{n^2}\|\nabla f_0(\theta) - \nabla f_0'(\theta_\star)\|^2 + \frac{3}{n^2}\|\nabla f_0'(\theta) - \nabla f_0'(\theta_\star)\|^2 + \frac{3}{n^2}\|\nabla f_0(\theta_\star) - \nabla f_0'(\theta_\star)\|^2.\end{aligned}$$

Then, using successively the L-smoothness of f_0, the $L_{\rm rel}$-relative smoothness of f_0 *w.r.t.* f, and strong convexity of f, we get:

$$\|\nabla f_0(\theta) - \nabla f_0'(\theta_\star)\|^2 \leq 2LD_{f_0}(\theta,\theta_\star) \leq 2LL_{\rm rel}D_f(\theta,\theta_\star) \leq \frac{LL_{\rm rel}}{\mu}\|\nabla f(\theta)\|^2. \tag{62}$$

We can then use the same bound for the f_0' term, and for making f' appear. Compared to direct bounding without relative smoothness, we have replaced a condition number L/μ by a relative smoothness term $L_{\rm rel}$, which is generally much smaller. We then obtain that $\eta^2 = O(\kappa L_{\rm rel}/n^2)$, much like in Equation (11).

B.3 Proof for orthogonal data

Let us assume that the data is orthogonal, *i.e.* that either $X_i^\top X_j = \|X_i\|^2$ or $X_i^\top X_j = 0$. This is actually slightly stronger, because we also assume that there is no norm spread for a given direction. This could be relaxed so that results would depend on the average norm, but we keep it simple here.

Consider that at least half of the dataset is fixed, and contains all different X_i in equal proportions. This means that the weight of each X_i is d^{-1}, since there are exactly d different ones. Indeed, there cannot be more than d (otherwise the orthogonality constraint would be violated), and if there are less than d we can just restrict to the relevant subspace. In this case, using that half of the data (*w.l.o.g.* is fixed, we have that:

$$A = \frac{1}{n}\sum_{i=1}^n X_iX_i^\top \succcurlyeq \frac{1}{n}\sum_{i=1}^{n/2} X_iX_i^\top \succcurlyeq \frac{1}{2d}\sum_{i=1}^d X_iX_i^\top,$$

where the last line comes from the fact that each X_i is represented n/d times, and we implicitly assume (again, *w.l.o.g.*) that the i first samples are all orthogonal to one another. In particular, for $j \in \{1,\dots,d\}$,

$$\|X_i\|^2 X_i^\top A^{-2} X_i \leq 2d\|X_i\|^2 X_i^\top \left(\sum_{j=1}^d X_jX_j^\top\right)^{-2} X_i = 2d\|X_i\|^2 X_i^\top \left(\frac{X_iX_i^\top}{\|X_i\|^4}\right)^2 X_i = 2d.$$

Note in particular that the relative sensitivity is independent of the scale of each X_i.

B.4 Proof of Proposition 2

Following Vadhan [2017, Section 3.2], the Propose-Test-Release mechanism is based on evaluating Δ, the minimum number of points that we need to change from X to obtain $\tilde{A}'$ such that $\tilde{A}' - \rho C \preccurlyeq 0$. Then, we privately test that $\Delta > 1$ by adding well-calibrated Laplacian noise to it. If the test fails, then this means that the threshold ρ

that we proposed is too low. If it passes, it means that ρ can be used as the sensitivity since the neighbours of ρ also satisfy this threshold.

However, computing Δ might require enumerating all the possibilities to remove points from $\tilde{A}$. We now show that this can be done efficiently in our case.

Here, we only consider different datasets that have **the same number of points** n**.** This technicality can be lifted, but at the expense of complicating the proof by putting a restriction that depends on n. Basically, another way of making $\tilde{A} = \frac{1}{n}XX^\top$ small would be to just add data points that hardly contribute to the covariance while increasing n, and we would have to take this case into account.

Now, let I_R be the indices of points that we remove from X, and let I_R^c be the points that we add instead, then we have that

$$\tilde{A}' = \tilde{A} - \frac{1}{n}\sum_{i\in I_R} X_i X_i^\top + \frac{1}{n}\sum_{i\in I_R^c} X_i X_i^\top \succcurlyeq \tilde{A} - \frac{1}{n}\sum_{i\in I_R} X_i X_i^\top \tag{63}$$

In particular, $\{|I_R|, \tilde{A}' - \rho C \preccurlyeq 0\} \subset \{|I_R|, \tilde{A} - \frac{1}{n}\sum_{i\in I_R} X_i X_i^\top - \rho C \preccurlyeq 0\}$, and so:

$$\Delta = \min\{|I_R|, \tilde{A}' - \rho C \preccurlyeq 0\} \geq \min\{|I_R|, \tilde{A} - \frac{1}{n}\sum_{i\in I_R} X_i X_i^\top - \rho C \preccurlyeq 0\} \tag{64}$$

If $\tilde{A} - \rho C$ is not positive semi-definite then $\Delta = 0$. Otherwise, it has a unique PSD square root Q, so let us define Y such that:

$$\tilde{A} - \rho C = Q^2, \qquad Y_i = Q^{-1}X_i. \tag{65}$$

Then, we have that:

$$E_+ = \left\{I_R, \tilde{A} - \rho C - \frac{1}{n}\sum_{i\in I_R} X_i X_i^\top \succcurlyeq 0\right\} = \left\{I_R, nI_d - \sum_{i\in I_R} Y_i Y_i^\top \succcurlyeq 0\right\} = \left\{I_R, \lambda_{\max}\left(\sum_{i\in I_R} Y_i Y_i^\top\right) \leq n\right\} \tag{66}$$

Let $E_- = \left\{I_R, \lambda_{\max}\left(\sum_{i\in I_R} Y_i Y_i^\top\right) > n\right\}$ be the complement of E_+. Note that at this stage we still have a combinatory problem: there are $\sum_{\ell=1}^{k}\binom{n}{\ell}$ possibilities to test to be sure that $\min\{|I|, I \in E_-\} \geq k$. However, we can introduce a simple inequality at this point, and write that

$$\lambda_{\max}\left(\sum_{i\in I_R} Y_i Y_i^\top\right) \leq \sum_{i\in I_R} \lambda_{\max}(Y_i Y_i^\top) = \sum_{i\in I_R} \|Y_i\|^2. \tag{67}$$

In particular, this means that

$$E_- \subset \{I_R, \sum_{i\in I_R} \|Y_i\|^2 > n\}, \text{ and so } \Delta \geq \min\{|I_R|, \sum_{i\in I_R} \|Y_i\|^2 > n\} \tag{68}$$

It remains to observe that $\|Y_i\|^2 = X_i^\top Q^{-2} X_i = X_i^\top (\tilde{A} - \rho C)^{-1} X_i$, so that we never actually have to compute the matrix square root.

B.5 Proof of Proposition 3

Proof. The proof follows the following plan:

1. The clipped points concentrate, *i.e.*,

$$p(\left\|\tilde{A} - \mathbb{E}\left[\tilde{A}\right]\right\| \geq \mu_{\text{reg}}) \leq \nu \text{ for } n \geq 4\log(2d/\nu)/9, \tag{69}$$

   with $\mu_{\text{reg}} = 4LR_c^2\sqrt{\frac{\log(2d/\nu)}{n}}$. This in particular means that with probability at least $1-\nu$, we have $\tilde{A} \succcurlyeq \mathbb{E}\left[\tilde{A}\right] - \mu_{\text{reg}} I_d$.

2. $\mathbb{E}\left[\tilde{A} - \mu_{\text{reg}} I_d\right] = \bar{\rho}\Sigma$, so that with probability at least $1-\nu$, $\tilde{A} \succcurlyeq \bar{\rho} C$.

3. PTR always succeeds with our choice of ρ, ε, δ.

1 - Concentration of the covariance. For all samples in the clipped dataset, $\tilde{X}_i^\top \Sigma^{-1} \tilde{X}_i \leq R_c^2$ (by definition of clipping), and so $\lambda_{\max}(\tilde{X}_i \tilde{X}_i^\top) \leq R_c^2 L$. Let us define $S_i = \tilde{X}_i \tilde{X}_i^\top / n$. Then,

$$\|S_k - \mathbb{E}[S_k]\| \leq 2R_c^2 L/n, \text{ and } \nu = \sum_{i \in I_c} \mathbb{E}\left[\|S_i - \mathbb{E}[S_i]\|^2\right] \leq \frac{4L^2 R_c^4}{n}. \tag{70}$$

In particular, we can then use Tropp et al. [2015, Corollary 6.1.2] to bound the tails of $\tilde{A} = \sum_{i=1}^n S_i$ as:

$$p(\left\|\tilde{A} - \mathbb{E}\left[\tilde{A}\right]\right\|) \geq t) \leq 2d \exp\left(-\frac{nt^2}{8L^2 R_c^4 + 4LR_c^2 t/3}\right). \tag{71}$$

In particular, if $t \leq 6LR_c^2$ then

$$p(\left\|\tilde{A} - \mathbb{E}\left[\tilde{A}\right]\right\| \geq t) \leq 2d \exp\left(-\frac{nt^2}{16L^2 R_c^4}\right). \tag{72}$$

In this case, we have that

$$p\left(\left\|\tilde{A} - \mathbb{E}\left[\tilde{A}\right]\right\| \geq 4LR_c^2 \sqrt{\frac{\log(2d/\nu)}{n}}\right) \leq \nu \text{ for } n \geq 4\log(2d/\nu)/9. \tag{73}$$

2 - $\mathbb{E}\left[\tilde{A} - \mu_{\text{reg}} I_d\right]$ is proportional to the covariance. We write, using the change of variable $u = \Sigma^{-\frac{1}{2}} x$:

$$\begin{aligned}
\mathbb{E}\left[\tilde{A} - \mu_{\text{reg}} I_d\right] &= \frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}} \int_{\mathbb{R}^d} \left[\mathbb{1}\{x^\top \Sigma^{-1} x \leq R_c^2\} + \mathbb{1}\{x^\top \Sigma^{-1} x \geq R_c^2\} \frac{R_c^2}{x^\top \Sigma^{-1} x}\right] x x^\top e^{-\frac{x^\top \Sigma^{-1} x}{2}} \mathrm{d}x \\
&= \Sigma^{\frac{1}{2}} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left[\mathbb{1}\{\|u\|^2 \leq R_c^2\} + \mathbb{1}\{\|u\|^2 \geq R_c^2\} \frac{R_c^2}{\|u\|^2}\right] u u^\top e^{-\frac{\|u\|^2}{2}} \mathrm{d}u \Sigma^{\frac{1}{2}} \\
&= \Sigma \times \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left[\mathbb{1}\{\|u\|^2 \leq R_c^2\} + \mathbb{1}\{\|u\|^2 \geq R_c^2\} \frac{R_c^2}{\|u\|^2}\right] \frac{\|u\|^2}{d} e^{-\frac{\|u\|^2}{2}} \mathrm{d}u,
\end{aligned}$$

where the last line follows from the fact that the expression within the integral is completely symmetric, so for any u, each direction receives a weight proportional to d^{-1}. In particular, $\mathbb{E}\left[\tilde{A} - \mu_{\text{reg}} I_d\right] = \bar{\rho}\Sigma$ with

$$\bar{\rho} = \frac{1}{d(2\pi)^{d/2}} \int_{\mathbb{R}^d} \min(\|u\|^2, R_c^2) e^{-\frac{\|u\|^2}{2}} \mathrm{d}u, \tag{74}$$

which only depends on R_c, and is close to 1 for instance when $R_c^2 \geq 2d$. We know however that $\bar{\rho} \leq 1$ since it comes from projecting points from a Gaussian distribution of covariance Σ.

3 - PTR always succeeds with this choice of ρ, ϵ, δ. We have shown that with high probability, the regularized clipped covariance matrix is proportional (up to a factor $\bar{\rho} \leq 1$) to the covariance of the data. We now show that in this case, suggesting such a ρ (which we can evaluate) will pass the PTR test for a range of (ϵ, δ) values.

Let $\beta > 0$. If we propose value $\rho = \beta\bar{\rho}$ then $\tilde{A} - \rho C \succcurlyeq (1-\beta)\Sigma$. In particular, using the above derivations and clipping we obtain that for all i,

$$\tilde{X}_i^\top (\tilde{A} - \rho C)^{-1} \tilde{X}_i \leq (1-\beta)^{-1} \tilde{X}_i^\top \Sigma^{-1} \tilde{X}_i \leq (1-\beta)^{-1} R_c^2. \tag{75}$$

Since each term in the summation can be lower bounded, we obtain that

$$\min\{|I_R|, \sum_{i \in I_R} \frac{1}{n} \tilde{X}_i^\top (\tilde{A} - \rho C)^{-1} \tilde{X}_i \geq 1\} \geq \min\{|I_R|, |I_R| \frac{R_c^2}{n(1-\beta)} \geq 1\} = \left\lceil \frac{n(1-\beta)}{R_c^2} \right\rceil \tag{76}$$

In particular, if $n \geq -\frac{\log(\delta)R_c^2}{(1-\beta)\varepsilon}$, we have that

$$\hat{\Delta} = \Delta_+ + Lap(\varepsilon^{-1}) \geq \Delta_+ \geq n(1-\beta)/R_c^2 \geq -\log(\delta)/\varepsilon, \tag{77}$$

and so the PTR always returns a value. Take $\beta = \frac{1}{2}$ to instantiate the theorem.

Note that Proposition 3 uses the condition number of the covariance κ_Σ instead of κ, the condition number of A. This is because κ depends on the specific dataset that we consider, and so we should also run a PTR procedure to release it. Note that in the end, we pay a factor κ_Σ since:

$$\|X_i\|^2 X_i^\top A^{-2} X_i \leq \rho^{-2} \|X_i\|^2 X_i^\top \Sigma^{-2} X_i \leq \kappa_\Sigma \rho^{-2} (X_i^\top \Sigma^{-1} X_i)^2. \tag{78}$$

□

C Details on the experimental setup

In this section, we provide the various details needed to reproduce our results. Note that we also provide code in supplementary material. Our experimental setup assumes that the data is split across 2 nodes. At each step t, node i privately releases its gradient $g_t^{(i)}$, computed at point θ_t. For all methods, we run the following gradient descent algorithm:

$$\theta_{t+1} = \theta_t - \tau \times \frac{1}{2}(g_t^{(1)} + g_t^{(2)}), \tag{79}$$

where τ is the step-size. In order to comply with the guidelines of Theorem 2, we set it as half the maximum of the largest eigenvalue of the local covariance matrices, so $\tau = 0.5/\max(\lambda_{\max}(A^{(1)}), \lambda_{\max}(A^{(2)}))$ which is an approximation for the true largest possible step-size. This is the step-size we use regardless of the method used for noising. Then, all methods only differ in the way they add noise to the gradients. The non-private method thus releases:

$$g_t^{(i)} = \frac{1}{N}\sum_{j=1}^{N} \nabla f_{ij}(\theta) = \frac{1}{N}\sum_{j=1}^{N} A_{ij}\theta - b_{ij}. \tag{80}$$

C.1 Parameters for GD with clipping

For gradient clipping, each node clips all its individual gradients using a clipping threshold $c^{(i)}$. In particular, the noisy gradients write:

$$g_t^{(i)} = \frac{1}{N}\sum_{j=1}^{N} \frac{\nabla f_{ij}(\theta_t) c_i}{\max(c_i, \|\nabla f_{ij}(\theta_t)\|)} + \mathcal{N}(0, \sigma_i^2), \tag{81}$$

where the variance of the noise σ_i^2 is computed as

$$\sigma_i^2 = \frac{\alpha c_i^2}{\varepsilon N^2}, \tag{82}$$

where α and ε are the Rényi-DP parameters. The results of DP-GD heavily depend on how we set the clipping threshold c_i.

One way of setting it is simply to randomly try out some clipping thresholds (potentially with external knowledge), but this is quite inefficient, as each node node needss to release M times more gradients to the other node if we would like to try M thresholds. Instead, we assume in this work that we can estimate the stochastic gradients at the optimum for the function we consider, and choose the smallest threshold such that none of these gradients is clipped, so $c_i = \max_j \|\nabla f_{ij}(\theta_i^\star)\|$. This leverages auxiliary information, and is a very strong baseline in the homogeneous setting, but leads to small clipping thresholds in the heterogeneous setting. Small clipping thresholds induce small noise but potentially large bias. We observe in our experiments that thresholds significantly lower than c_i do not introduce such a large bias still. We conjecture that this is due to the fact that the bias introduced from clipping is small for specific distributions, such as symmetric ones. Also note that per-instance clipping is in general computationally expensive, and in particular required significantly more time in our case.

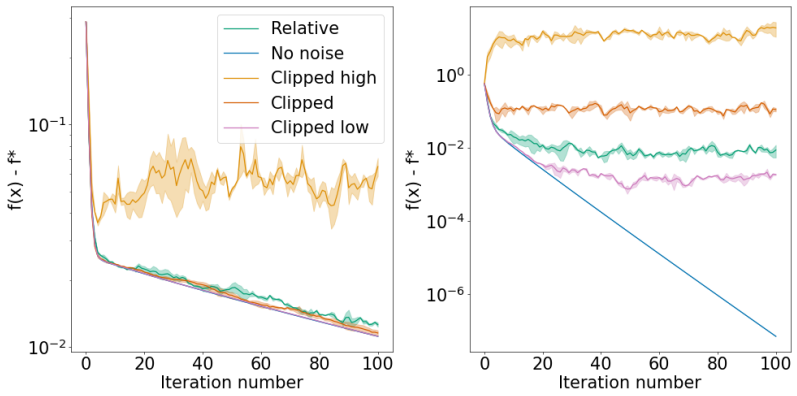

Figure 2: Left: HIGGS dataset, $\varepsilon = 10^{-2}$, Right: ijcnn1 dataset, $\varepsilon = 10^{-3}$

C.2 Parameters for GD with relative sensitivity

In this case, the noisy gradients we release are of the form:

$$g_t^{(i)} = \bar{g}_t{}^{(i)} + \mathcal{N}(0, \gamma_i \|\bar{g}_t{}^{(i)}\|^2 + \sigma_i^2), \text{ where } \bar{g}_t{}^{(i)} = \frac{1}{N}\sum_{j=1}^{N} \tilde{A}_{ij}\theta_t - \tilde{b}_{ij}. \tag{83}$$

The parameters γ_i and σ_i^2 depend on the relative sensitivity parameters η and R_{rel}, as specified in Theorem 1. We set η directly using Equation (11), which we can check directly on our dataset. We could alternatively recover this threshold by privately releasing a rough approximation of the covariance matrix, and then setting R_c and ρ by binary search (or using a small public dataset) to obtain a small η at the cost of logarithmic factors only (multiplying only the PTR procedure, not the whole training).

To set R_{rel}, we use that it is bounded by $\eta \|b_i\| + 2\max \|b_{ij}\|$. Thus, we bound it by computing the max norm of the b_{ij}, which can also be enforced via clipping. Although $\|b_i\| \leq \max \|b_{ij}\|$, we use a separate threshold for $\|b_i\|$ which usually leads to tighter guarantees. Note that we can alternatively approximate the local optimum $\theta_i^\star$ and set $R_{\text{rel}} = \max_{k,\ell} \|\nabla f_{ik}(\theta_i^\star) - \nabla f_{i\ell}(\theta_i^\star)\|$. However, we choose not to in this case because we want something that is *enforceable*, and this bound is harder to enforce via clipping the b_{ij}. The privacy loss term $\alpha\eta^2 d$ is negligible in our case since datasets have moderate dimension. As the examples are mainly intended to be illustrative and the constants have not been optimized for, we omit constant factors when estimating η.

As in the clipping case, this specific procedure of choosing the hyperparameters does not strictly guarantee differential privacy, as it uses local datasets. Yet, in practice, application-specific knowledge can help set these parameters. We choose these favorable parameters to show what different methods can do with appropriate hyperparameters, without explicitly quantifying the privacy loss incurred by having to find these parameters. Besides, *we choose parameters that can be enforced*, so that strict DP is guaranteed if we initially have a rough idea of what parameters should be.

We present experiments on the *ijcnn1* dataset (concatenation of train and test from LibSVM repository[1]) since the estimated η were rather small, and so could illustrate a case in which $\text{RGM}_{\gamma,\sigma}$ is useful. This is also the case for other LibSVM datasets that we tested, such as *HIGGS* or *SUZY*. Other datasets such as *cod-rna* require high levels of regularization to obtain small η, potentially trivializing the initial problem.

D Extra experiments

To show that our results apply more broadly, we include results on the Higgs (truncated at $n = 10^6$, $\varepsilon = 10^{-2}$), in the same setting as Figure 1 (left). We see that similar results to the ijcnn1 datasets are obtained. To compare further, we also try a high privacy regime for ijcnn1 ($\varepsilon = 10^{-3}$), in which we see that the performance of RGM is close to that of clipping with the best (small) threshold.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

Note that although the experiments are for linear regression only, the relative sensitivity bounds can be obtained for more general functions, as detailed in Appendix B.2. Such relative smoothness assumptions can for instance be shown for generalized linear models. Similarly, our theory can handle non-convex functions, as long as the gradients satisfy the relative sensitivity assumption. The difficulty of applying our approach to neural networks is not related to non-convexity per se, but to the complexity of the functions considered in these cases, preventing meaningful relative sensitivity bounds.