Private Gradient Descent for Linear Regression: Tighter Error Bounds and Instance-Specific Uncertainty Estimation

Gavin Brown [1] **Krishnamurthy (Dj) Dvijotham** [2] **Georgina Evans** [2] **Daogao Liu** [1] **Adam Smith** [3 2]
Abhradeep Thakurta [2]

Abstract

We provide an improved analysis of standard differentially private gradient descent for linear regression under the squared error loss. Under modest assumptions on the input, we characterize the distribution of the iterate at each time step.

Our analysis leads to new results on the algorithm's accuracy: for a proper fixed choice of hyperparameters, the sample complexity depends only linearly on the dimension of the data. This matches the dimension-dependence of the (non-private) ordinary least squares estimator as well as that of recent private algorithms that rely on sophisticated adaptive gradient-clipping schemes (Varshney et al., 2022; Liu et al., 2023).

Our analysis of the iterates' distribution also allows us to construct confidence intervals for the empirical optimizer which adapt automatically to the variance of the algorithm on a particular data set. We validate our theorems through experiments on synthetic data.

1. Introduction

Machine learning models trained on personal data are now ubiquitous—keyboard prediction models (Xu et al., 2023), sentence completion in email (Li et al., 2022), and photo labeling (Kurakin et al., 2022), for example. Training with *differential privacy* (Dwork et al., 2006) gives a strong guarantee that the model parameters reveal little about any single individual. However, differentially private algorithms necessarily introduce some distortion into the training process. Understanding the most accurate and efficient training procedures remains an important open question, with an extensive line of research dating back 15 years (Kasiviswanathan et al., 2008; Chaudhuri et al., 2011).

The distortion introduced for privacy is complex to characterize; recent work has thus also investigated how to provide confidence intervals and other inferential tools that allow the model's user to correctly interpret its parameters. Confidence intervals on parameters are critical for applications of regression in the social and natural sciences, where they serve to evaluate effects' significance.

In this paper, we address these problems for a fundamental statistical learning problem: least-squares linear regression. Specifically, we give a new analysis of a widely studied differentially private algorithm, *noisy gradient descent* (DP-GD). This algorithm repeatedly computes the (full) gradient at a point, adds Gaussian noise, and updates the iterate using the noisy gradient.

Our central technical tool is a new characterization of the distribution of the iterates of private gradient descent. Under the assumption that the algorithm does not clip any gradients, we show that the distribution at any time step can be written as a Gaussian distribution about the empirical minimizer (plus a small bias term). All together, the iterates are drawn from a Gaussian process. We apply this characterization in two ways: we derive tighter error bounds and prove finite-sample coverage guarantees for natural confidence intervals constructions.

Our main result shows that the algorithm converges to a nontrivial solution—that is, an estimate whose distance from the true parameters is a $o(1)$ fraction of the parameter space's diameter—using only $n = \widetilde{\Theta}(p)$ samples (omitting dependency on the privacy parameters) when the features and errors are distributed according to a Gaussian. Until recently, all private algorithms for linear regression required $\Omega(p^{3/2})$ samples to achieve nontrivial bounds. This includes previous analyses of gradient descent (Cai et al., 2021). Three recent papers have broken this barrier: the exponential-time approach of Liu et al. (2022) and the efficient algorithms of Varshney et al. (2022) and subsequently

[1]Paul G Allen School of Computer Science and Engineering, University of Washington. Part of this work was done while G.B. was at Boston University [2]Google DeepMind [3]Department of Computer Science, Boston University. Correspondence to: Gavin Brown <grbrown@cs.washington.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Liu et al. (2023).[1] The latter two algorithms are based on variants of private gradient descent that use adaptive clipping frameworks that complicate both the privacy analysis and implementation. We discuss these approaches in more detail in Related Work.

Our characterization of the iterates' distribution suggests that confidence interval constructions for Gaussian processes should apply to our setting. We confirm this, in both theory and practice: we formally analyze and empirically test methods for computing instance-specific confidence intervals (that is, tailored to the variability of the algorithm on this particular data set). These intervals convey useful information about the noise for privacy: in our experiments, their width is roughly comparable to that of the textbook nonprivate confidence intervals for the population parameter. Even beyond the settings of our formal analysis, we give general heuristics that achieve good coverage experimentally. These confidence interval constructions come at no cost to privacy—they use the variability among iterates (and their correlation structure) to estimate the variability of their mean. While prior works (Shejwalkar et al., 2022; Rabanser et al., 2023) discuss methods for generating confidence intervals based on intermediate iterates from DP gradient descent, they fail to provide any meaningful coverage guarantees.

We perform extensive experiments on synthetic data, isolating the effects of dimension and gradient clipping. To the best of our knowledge, these are the first experiments which show privacy "for free" in high-dimensional (that is, with $p \approx n$ and p large) private linear regression. We also demonstrate the practicality of our confidence interval constructions.

1.1. Our Results

Formal Guarantees for Accuracy Our theoretical results are simplest to state in the following distributional setting.

Definition 1.1 (Generative Setting)**.** *Let $\theta^* \in \mathbb{R}^p$ be the true regression parameter satisfying $\|\theta^*\| \leq 1$. For each $i \in [n]$, let the covariate $\mathbf{x}_i$ be drawn i.i.d. from $\mathcal{N}(0, \mathbb{I}_p)$ and the response $y_i \leftarrow \mathbf{x}_i^\dagger \theta^* + \xi_i$, for $\xi_i \sim \mathcal{N}(0, \sigma^2)$.*

We emphasize that the assumption of random-design Gaussian data allows for clean theorem statements but is not strictly necessary. Formally, we establish a set of deterministic conditions on the input dataset (Condition 2.4) under which our algorithm provably performs well. These conditions are satisfied with high probability by data arising from the generative setting above. In Appendix B we present experiments on other distributions.

Theorem 1.2 (Informal)**.** *Assume we are in the generative setting (Definition 1.1). Assume $n = \widetilde{\Omega}(p)$. Set clipping threshold $\gamma = \widetilde{\Theta}(\sigma\sqrt{p})$, step size $\eta = O(1)$, and number of steps $T = \widetilde{O}(1)$. With high probability the final iterate θ_T of Algorithm 1 satisfies*

$$\|\theta_T - \theta^*\| \leq \widetilde{O}\left(\sqrt{\frac{\sigma^2 p}{n}} + \frac{\sigma p}{\sqrt{\rho} n}\right).$$

Here the parameters are set so that Algorithm 1 satisfies ρ-zCDP (see Section 2 and Appendix A.1).

Theorem 1.2 follows from combining Theorem 2.7, which analyzes the accuracy of DP-GD relative to the ordinary least squares (OLS) solution $\hat{\theta}$, with Claim A.12, which comes from the work of Hsu et al. (2011) and says that $\hat{\theta}$ is close to θ^*.

Formal Guarantees for Confidence Intervals From our theoretical results, we expect any two iterates sufficiently separated in time to be well-approximated by independent draws from the stationary distribution, which (we show) is a Gaussian centered at the empirical minimizer: $\mathcal{N}(\hat{\theta}, c \cdot \mathbf{A})$, where $c \in \mathbb{R}$ depends on the algorithm's hyperparameters (e.g., privacy budget) and fixed matrix $\mathbf{A}$ is approximately spherical; it depends only on the step size and the covariance of the data. We can control the quality of this approximation and provide provable, non-asymptotic guarantees for natural confidence interval constructions.

As with our accuracy guarantees, these results do not rely directly on the assumption of Gaussian covariates and noise. Instead, they hold for any dataset satisfying Condition 2.4.

Experiments Our empirical results provide a strong complement to our formal results. They validate our theorems by showing that the predicted behavior occurs at practical sample sizes and privacy budgets. For example, Figure 1 demonstrates how our algorithm's error depends only linearly on the dimension, which mirrors the behavior of OLS and stands in contrast to algorithms requiring $\Omega(p^{3/2})$ examples.

Our empirical results also investigate settings not considered by our theorems. We demonstrate the accuracy of DP-GD and the validity of our confidence intervals constructions on distributions beyond those in Definition 1.1. One phenomenon that affects our methods is gradient clipping, which introduces bias in settings with irregular data (such as outliers). In such settings, DP-GD can be viewed as optimizing a different objective function (roughly, a Huberized loss) and that the confidence intervals we generate capture the minimum of this smoothed objective.

[1] Appearing after the submission of this paper, the approach of Brown et al. (2024) uses different techniques but also requires only $\widetilde{\theta}(p)$ samples for accurate estimation.

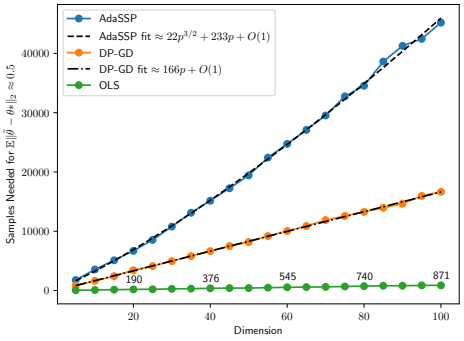


Figure 1: Iso-accuracy lines for three algorithms. Run on data from a well-specified linear model, both DP-GD and OLS require n to grow linearly with the dimension. We compare with AdaSSP (Wang, 2018), a popular algorithm that requires $n = \Omega(p^{3/2})$ examples (Kamath et al., 2022; Narayanan, 2023). Points represent the number of samples needed to achieve error $1/2$ in expectation. In our setting, an error of 1 is trivial. Selected points for OLS receive labels.

1.2. Techniques

The privacy analysis of DP-GD is standard and we thus focus our technical efforts on utility: that the algorithm has low error and the confidence interval constructions are valid. At a high level, our analysis follows a standard formula from the privacy literature: we identify a deterministic set of conditions on input datasets and show that, when these conditions are met, "good things happen."

In slightly more detail, we consider Condition 2.4, which asks that the dataset satisfies a number of deterministic concentration properties. When this condition holds, we show that DP-GD is unlikely to clip any gradients. This "no-clipping" event sets up our accuracy and confidence-interval analyses.

The final conceptual step in our theoretical analysis is to show that the deterministic conditions are meaningful. We show that the conditions are satisfied with high probability by data drawn from the Gaussian linear model in Definition 1.1.

Convergence Without Clipping Our first technical tool, presented in Section 2.2, characterizes the distribution of the iterates of DP-GD. We observe that, on any step where there is no clipping, the iterates satisfy the linear recurrence relation

$$\theta_t \leftarrow \eta \cdot \Sigma\hat{\theta} + (\mathbb{I} - \eta\Sigma)\theta_{t-1} + \eta \cdot \mathbf{z}_{t-1},$$

where $\mathbf{z}_t$ is the noise added at time t and $\Sigma = \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}$ is the covariance. We solve this recurrence and collect the noise terms, expressing θ_t as a Gaussian centered at $\hat{\theta}$ plus an exponentially decaying bias term:

$$\theta_t = \hat{\theta} + (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta}) + \eta \cdot \mathbf{z}'_t,$$

where $\mathbf{z}'_t$ is a Gaussian random variable that depends on $\mathbf{z}_0, \ldots, \mathbf{z}_{t-1}$.

Clipping Is Unlikely With this tool, we now aim to show that clipping is unlikely under the assumption that the data satisfies Condition 2.4. One subitem, Condition 2.4.ii, requires the covariates to have ℓ_2 norm bounded by roughly $\sqrt{p}$. When this is true, a standard application of Cauchy–Schwarz implies that gradients are bounded by $\widetilde{O}(p)$, so setting this as our clipping threshold γ allows us ensure no clipping happens with high probability.

Our main theoretical result improves upon this step. We show that the gradients are bounded by roughly $\sqrt{p}$, ignoring logarithmic factors and dependence on σ. Using $\gamma \approx \sqrt{p}$ as the clipping threshold allows us to add, at each iteration, privacy noise $\mathcal{N}(0, c^2\mathbb{I})$ for $c^2 \approx \frac{p}{n^2}$ (ignoring privacy parameters). We show such a bound holds with high probability over the randomness of the algorithm. Since we have assumed the norm of $\mathbf{x}_i$ is roughly $\sqrt{p}$, to control the norm of the gradient $\mathbf{x}_i(y_i - \mathbf{x}_i^\dagger\theta_t)$ it suffices to bound the absolute residual by $O(1)$. We can show that this holds at initialization, using the fact that our initial iterate θ_0 has no dependence on the data.

We also expect the bound to hold after convergence. For an informal argument, consider a single update away from θ^*, where the gradient should nearly vanish. We would move to $\theta_1 \leftarrow \theta^* + \mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(0, c^2\mathbb{I})$. Since $c^2 \approx \frac{p}{n^2}$, we expect $\|\mathbf{z}\| \approx c\sqrt{p} = O(1)$. Plugging this in, we see

$$\begin{aligned}|y_i - \mathbf{x}_i^\dagger\theta_1| &= |(\mathbf{x}_i^\dagger\theta^* + \xi_i) - \mathbf{x}_i^\dagger(\theta^* + \mathbf{z})| \\ &= |\xi_i - \mathbf{x}_i^\dagger\mathbf{z}|.\end{aligned}$$

Since ξ_i is drawn from a known distribution and $\mathbf{x}_i$ is independent of $\mathbf{z}$, we can bound this residual.

The proof of Lemma 2.5 formalizes these heuristics and shows that the same bound applies over every gradient step.

Confidence Intervals The distributional form we show for the iterates suggests several natural methods for constructing confidence intervals for the parameter being estimated. In this work, we focus on confidence intervals for the *empirical* minimizer—that is, the regression vector that minimizes the loss on the data set. Thus, our confidence intervals capture the uncertainty introduced by our privacy mechanism. (It also makes sense to give confidence intervals for a population-level minimizer in settings where the data represent a random sample—we focus on the empirical parameter for simplicity.) We consider confidence

intervals for a single coordinate of the parameter, since this is the most common use case.

We consider two approaches: one based on running the entire algorithm repeatedly and the other based on estimating the in-sequence variance of the stream of iterates. For this latter approach, we consider both a simple checkpointing strategy as well as a more data-efficient averaging strategy from the empirical process literature. These are discussed in Section 4.2.

Gaussian Data with Gaussian Errors Satisfy Condition 2.4 The analysis described above operates under the assumption that the input satisfies Condition 2.4, a deterministic set of conditions. To better interpret the results and compare with prior work, we show that this condition is satisfied with high probability by data drawn from a well-specified Gaussian linear model. This proof requires a collection of standard concentration inequalities.

1.3. Limitations and Future Work

Our work has several limitations, each of which presents natural directions for further exploration. First, some parts of our theoretical analysis require that the data be well-conditioned and have Gaussian-like concentration properties. Aside the exponential-time approach of Liu et al. (2022), all private algorithms achieving $O(p)$ sample complexity incur a polynomial dependence on the condition number of the covariates (Varshney et al., 2022; Liu et al., 2023). Removing this dependence remains a notable open problem.[2]

We construct confidence intervals for the empirical minimizer of the loss function after clipping, which in general differs from the least-squares estimator (see Section 2.5) and the population parameter. The former limitation is inherent (since the exact OLS solution has unbounded sensitivity) but an extension of our methods to population quantities would likely be useful.

Our experiments use synthetic data sets. A wider study of how these methods adapt to practical regression tasks is an important project but beyond the scope of this paper.

1.4. Related Work

Private Linear Regression Under the assumption that the covariates are drawn from a subgaussian distribution and the responses arise from a linear model, the exponential-time approach of Liu et al. (2022) achieves nearly optimal error, matching a lower bound of Cai et al. (2021). In the remainder of this subsection, we survey a number of efficient approaches. Table 1 in appendix A summarizes the approaches and their dimension-dependence in our setting.

Sufficient Statistics A standard approach for private regression is *sufficient statistics perturbation* (see, e.g., Vu & Slavkovic, 2009; Foulds et al., 2016; Sheffet, 2017; 2019). One algorithm which stands out for its practical accuracy and theoretical guarantees is the *AdaSSP* algorithm of Wang (2018), which relies on prior bounds for the covariates and labels.

To overcome the reliance on this prior knowledge, Milionis et al. (2022) build on algorithms of Kamath et al. (2019) to give theoretical guarantees for linear regression with unbounded covariates. More recent variations on SSP include Tang et al. (2023), who use AdaSSP inside a boosting routine, and Ferrando & Sheldon (2024), who post-process tables of two-way marginals produced by a private query-answering mechanism.

As the dimension of the problem grows, these approaches suffer high error: accurate private estimation of $\mathbf{X}^\dagger\mathbf{X}$, which is necessary for SSP to succeed, requires $n = \widetilde{\Omega}(p^{3/2})$ examples (Dwork et al., 2014; Kamath et al., 2022; Narayanan, 2023).

Optimization An alternative approach is to view regression as an optimization problem, seeking a parameter vector that minimizes the empirical error. Such algorithms form a cornerstone of the differential privacy literature.

Under assumptions similar to ours, the approach of Cai et al. (2021) achieves a near-optimal statistical rate via full-batch private gradient descent, where sensitivity of the gradients is controlled via projecting parameters to a bounded set. Their analysis, however, only applies when $n = \Omega(p^{3/2})$. Avella-Medina et al. (2021) gave a general convergence analysis for private M-estimators (including Huber regression) but did not explicitly track the dimension dependence. Their approach bears similarities to gradient clipping, as we discuss in Section 2.5. The work of Varshney et al. (2022) gave the first efficient algorithm for private linear regression requiring only $n = \widetilde{O}(p)$ examples. Their approach uses differentially private stochastic gradient descent and an adaptive gradient-clipping scheme based on private quantiles. Later, Liu et al. (2023) gave a robust and private algorithm using adaptive clipping and full-batch gradient descent, improving upon the sample complexity of Varshney et al. (2022) (by improving the dependence on the condition number of the design matrix). Our approach is most similar to that of Liu et al. (2023); while they rely on adaptive clipping and strong *resilience* properties of the input data, our algorithm and analysis are simpler.

A natural alternative strategy for private linear regression

[2] Appearing after the submission of this paper, the approach of Brown et al. (2024) removes this dependence through completely different techniques.

is to first clip the covariates and the responses and then run noisy *projected* gradient descent, projecting each iterate into a constraint set. While this results in ℓ_2-bounded gradient, the sample complexity of such an algorithm becomes $n = \Omega(p^{3/2})$ (Cai et al., 2021).

Sample-and-Aggregate In connection with robust statistics, a line of work gives private regression algorithms based on finding approximate medians (Dwork & Lei, 2009; Alabi et al., 2022; Sarathy & Vadhan, 2022; Knop & Steinke, 2022; Amin et al., 2022). Informally, the algorithms solve the linear regression problem (or a robust variant) on multiple splits of the data and apply a consensus-based DP method (e.g., propose-test-release (Dwork & Lei, 2009), or the exponential mechanism (McSherry & Talwar, 2007)) to choose the regression coefficients. To the best of our knowledge, this class of approaches cannot achieve sample complexity $n = o(p^2)$, as it requires $\Omega(p)$ samples per split and $\Omega(p)$ splits for nontrivial private aggregation.

Private Confidence Intervals Some approaches generate confidence intervals using bootstrapping and related approaches (Brawner & Honaker, 2018; Wang et al., 2022; Covington et al., 2021).

Several approaches arise naturally from sufficient statistics perturbation. Sheffet (2017) gave valid confidence intervals for linear regression. The parametric bootstrap (Ferrando et al., 2022) is also a natural choice in this setting where we already work under strong distribution assumptions.

Other approaches stem from the geometry of optimization landscape (Wang et al., 2019; Avella-Medina et al., 2021); see also the non-private work of (Chen et al., 2020).

In a recent line of work, (Shejwalkar et al., 2022; Rabanser et al., 2023) use multiple checkpoints from a single run of DP-GD (Bassily et al., 2014; Abadi et al., 2016) to provide confidence intervals for predictions. While there is some algorithmic similarity to our procedures, these works do not provide rigorous parameter confidence interval estimates.

2. Private Gradient Descent for Regression

Notation We use lowercase bold for vectors and uppercase bold for matrices, so $(\mathbf{X}, \mathbf{y})$ is a data set and $(\mathbf{x}_i, y_i)$ a single observation. Special quantities receive Greek letters: $\theta \in \mathbb{R}^p$ denotes a regression vector and $\Sigma = \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}$, the empirical covariance. We "clip" vectors in the standard way: $\texttt{CLIP}_\gamma(\mathbf{x}) = \mathbf{x} \cdot \min\{1, \gamma/\|\mathbf{x}\|\}$.

2.1. Algorithm

Algorithm 1 is differentially private gradient descent. Similar to the more complex linear regression algorithm of Liu et al. (2023), it controls the sensitivity by clipping individual gradients: a full-batch version of the widely used private *stochastic* gradient descent (Abadi et al., 2016). This is in contrast to approaches which rely on projecting the parameters to a convex set (see, e.g., Bassily et al., 2014).

We present our privacy guarantee with *(zero-)concentrated differential privacy* (zCDP) (Dwork & Rothblum, 2016; Bun & Steinke, 2016). For the definition of zCDP and basic properties, including conversion to (ε, δ)-DP, see Appendix A.1. The guarantee comes directly from composing the Gaussian mechanism.

Lemma 2.1. *For any noise variance $\lambda^2 \geq \frac{2T\gamma^2}{\rho n^2}$, Algorithm 1 satisfies ρ-zCDP.*

Algorithm 1 DP-GD, $\mathcal{A}(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, \theta_0)$

Input: data $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$; clipping threshold $\gamma > 0$; noise scale $\lambda > 0$; step size $\eta > 0$; number of iterations $T \in \mathbb{N}$; initial vector $\theta_0 \in \mathbb{R}^p$
for $t = 1, \ldots, T$ **do**
  $\bar{\mathbf{g}}_t \leftarrow \frac{1}{n}\sum_{i=1}^n \texttt{CLIP}_\gamma(-\mathbf{x}_i(y_i - \mathbf{x}_i^\dagger\theta_{t-1}))$
  Draw $\mathbf{z}_t \sim \mathcal{N}(0, \lambda^2\mathbb{I})$
  $\theta_t \leftarrow \theta_{t-1} - \eta \cdot \bar{\mathbf{g}}_t + \eta \cdot \mathbf{z}_t$
end for
Output: $\theta_1, \ldots, \theta_T$.

2.2. Convergence, With and Without Clipping

Throughout the paper, we deal with the exact distribution over the iterates of DP-GD. As we show, if we remove the clipping step (i.e., set $\gamma = +\infty$), this distribution is exactly Gaussian. This algorithm appears in many contexts under different names, including *Noisy GD* and the *(Unadjusted) Langevin Algorithm*. Clipping ensures privacy, even if a particular execution of the algorithm does not clip any gradients. We might hope to condition on the event that Algorithm 1 clips no gradients. However, conditioning changes the output distribution. We require a more careful approach.

We consider a *coupling* between two executions of Algorithm 1: one with clipping enforced and the other with $\gamma = +\infty$. Of course, if Algorithm 1 receives an input where clipping occurs with high probability, then the output distributions of the two executions may differ greatly. When clipping in the first case is unlikely, we can connect the output distributions between the two executions.

Lemma 2.2 (Coupling DP-GD without Clipping)**.** *Fix data set $\mathbf{X}, \mathbf{y}$ and hyperparameters γ, λ, η, T, and θ_0. Define random variables $\mathcal{O}_\gamma$ and $\mathcal{O}_\infty$ as*

$$\mathcal{O}_\gamma = \mathcal{A}(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, \theta_0)$$
$$\mathcal{O}_\infty = \mathcal{A}(\mathbf{X}, \mathbf{y}; \infty, \lambda, \eta, T, \theta_0).$$

If Algorithm 1 on input $(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, \theta_0)$ *clips nothing with probability* $1 - \beta$*, then* $TV(\mathcal{O}_\gamma, \mathcal{O}_\infty) \leq \beta$.

Appendix C contains details on couplings and the simple proof of this lemma, which uses the coupling induced by sharing randomness across runs of the algorithm.

The following claim characterizes the output of Algorithm 1 with $\gamma = +\infty$. We see the distribution is Gaussian and centered at the empirical minimizer plus a bias term that goes to zero quickly as t grows. We define a matrix $\mathbf{D} = (\mathbb{I} - \eta\Sigma)^2$, where Σ is the empirical covariance matrix and η the step size.

The proof, which appears in Appendix C only relies on the fact that the loss function is quadratic.

Lemma 2.3. *Fix a data set* $(\mathbf{X}, \mathbf{y})$ *and step size* η*, noise scale* λ*, number of iterations* T*, and initial vector* θ_0*. Define matrices* $\Sigma = \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}$ *and* $\mathbf{D} = (\mathbb{I} - \eta\Sigma)^2$*; assume both are invertible. Let* $\hat{\theta} = (\mathbf{X}^\dagger\mathbf{X})^{-1}\mathbf{X}^\dagger\mathbf{y}$ *be the least squares solution. Consider* $\mathcal{A}(\mathbf{X}, \mathbf{y}; \infty, \lambda, \eta, T, \theta_0)$*, i.e., Algorithm 1 without clipping. For any* $t \in [T]$*, we have*

$$\theta_t = \hat{\theta} + (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta}) + \eta \cdot \mathbf{z}'_t$$

for $\mathbf{z}'_t \sim \mathcal{N}(0, \lambda^2\mathbf{A}^{(t)})$ *with* $\mathbf{A}^{(t)} = (\mathbb{I} - \mathbf{D})^{-1}(\mathbb{I} - \mathbf{D}^t)$.

Note that $\mathbf{z}'_t$ depends on all the noise vectors up to time t. We must take care when applying this lemma across the same run, as $\mathbf{z}'_{t_1}$ and $\mathbf{z}'_{t_2}$ are dependent random variables.

2.3. Conditions that Ensure No Clipping

To apply Lemma 2.3, we need to reason about when no gradients are clipped. To that end, we now define a deterministic "No-Clipping Condition" under which DP-GD clips no gradients with high probability. This is a condition on data sets $(\mathbf{X}, \mathbf{y})$ that is defined in terms of a few hyperparameters. We will often leave these hyperparameters implicit, saying "data set $(\mathbf{X}, \mathbf{y})$ satisfies the No-Clipping Condition" instead of "data set $(\mathbf{X}, \mathbf{y})$ satisfies the No-Clipping Condition with values σ, η, T, and c_0."

Later, we will show that the No-Clipping Condition is satisfied with high probability under distributional assumptions. Conditions (i)-(iii) follow from standard concentration statements. Conditions (iv) and (v) are less transparent; they capture notions of independence between θ^*, $\{\mathbf{x}_i\}$, and $\{y_i - \mathbf{x}_i^\dagger\theta^*\}$.

Condition 2.4 (No-Clipping Condition)**.** *Let* σ, η *and* c_0 *be nonnegative real values and let* T *be a natural number. Let* $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^p$ *be a data set. Define* $\Sigma = \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}$*. There exists a* $\theta \in \mathbb{R}^p$ *such that, for all* $i \in [n]$ *and* $t \in [T]$*,*

(i) $\frac{1}{2}\mathbb{I} \preceq \Sigma \preceq 2\mathbb{I}$, $\|\mathbb{I} - \eta\Sigma\| \leq \frac{7}{8}$,

(ii) $\|\mathbf{x}_i\| \leq c_0\sqrt{p}$,

(iii) $|y_i - \mathbf{x}_i^\dagger\theta| \leq c_0\sigma$,

(iv) $|\mathbf{x}_i^\dagger(\mathbb{I} - \eta\Sigma)^t\theta| \leq c_0$ *and*

(v) $|\sum_{j=1}^n (y_j - \mathbf{x}_j^\dagger\theta) \cdot \mathbf{x}_i^\dagger A_t\mathbf{x}_j| \leq c_0\sigma\sqrt{np}$*, where* $A_t = (\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\Sigma^{-1}$.

The definition requires the existence of some θ with certain properties. Informally, the reader should think of this as θ^*, the "true" parameter. Crucially, however, the definition does not require the existence of an underlying distribution. We prove Lemma 2.5 in Appendix C.

Lemma 2.5 (No Clipping Occurs)**.** *Fix nonnegative real numbers* σ, η*, and* c_0*. Fix natural number* T*. Assume data set* $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$ *satisfies the No-Clipping Condition (Condition 2.4) with values* (σ, η, c_0, T)*.*

Fix nonnegative real numbers γ *and* λ*. Consider running Algorithm 1, i.e.,* $\mathcal{A}(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, 0)$ *with* $\theta_0 = 0$ *the initial point. Assume* $\gamma \geq 4c_0^2\sigma\sqrt{p}$*. For any* $\beta \in (0, 1)$*, if* $\frac{\gamma}{\eta\lambda} \geq 64c_0^2 p\sqrt{\ln 2nT/\beta}$*, then with probability* $1 - \beta$ *Algorithm 1 clips no gradients.*

2.4. Analyzing Algorithm 1

Recall the distributional setting we discussed in the introduction: there is some true parameter θ^* with $\|\theta^*\| \leq 1$ and observations are generated by drawing covariate $\mathbf{x}_i \sim \mathcal{N}(0, \mathbb{I})$ and setting response $y_i \leftarrow \mathbf{x}_i^\dagger\theta^* + \xi_i$ for $\xi_i \sim \mathcal{N}(0, \sigma^2)$. In this section, we first show that data sets generated in this way satisfy the No-Clipping Condition with high probability. This implies that Algorithm 1, with high probability, does not clip any gradients.

Lemma 2.6. *Fix data set size* n *and data dimension* $p \geq 2$*. Fix* θ^* *with* $\|\theta^*\| \leq 1$*, let covariates* $\mathbf{x}_i$ *be drawn i.i.d. from* $\mathcal{N}(0, \mathbb{I})$*, and responses* $y_i = \mathbf{x}_i^\dagger\theta^* + \xi_i$ *for* $\xi_i \sim \mathcal{N}(0, \sigma^2)$*. Fix the step size* $\eta = \frac{1}{4}$ *and a natural number* T*. There exists a constant* c *such that, for any* $\beta \in (0, 1)$*, if* $n \geq c(p + \ln 1/\beta)$ *then with probability at least* $1 - \beta$ *data set* (X, y) *satisfies the No-Clipping Condition (Condition 2.4) with* $c_0 = 12\ln^{1.5}(5nT/\beta)$*.*

We combine this statement with Lemma 2.5, which says that clipping is unlikely under the No-Clipping Condition, and Lemma 2.3, which characterizes the output distribution of DP-GD when there is no clipping. We prove this theorem in Appendix C.

Theorem 2.7 (Main Accuracy Claim). *Fix $\theta^* \in \mathbb{R}^p$ with $p \geq 2$ and $\|\theta^*\| \leq 1$, let n covariates $\mathbf{x}_i$ be drawn i.i.d. from $\mathcal{N}(0, \mathbb{I})$ and responses $y_i = \mathbf{x}_i^\dagger \theta^* + \xi_i$ for $\xi_i \sim \mathcal{N}(0, \sigma^2)$ for some fixed σ.*

Fix $\rho \geq 0$ and $\beta \in (0, 1)$. Consider running Algorithm 1, i.e., $\mathcal{A}(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, 0)$ with step size $\eta = \frac{1}{4}$, initial point $\theta_0 = 0$, and, for some absolute constant c,

$$T = c \log \frac{n\rho}{p}, \quad \lambda^2 = \frac{2T\gamma^2}{\rho n^2}, \text{ and}$$
$$\gamma = c\sigma\sqrt{p} \log^3 \left(\frac{nT}{\beta} \right).$$

Recall $\hat{\theta}$ the OLS solution. If $n \geq c\left(p + \sqrt{p} \log^4 \rho/\beta\right)$, then with probability at least $1 - \beta$ Algorithm 1 returns a final iterate θ_T such that, for some constant c',

$$\|\hat{\theta} - \theta_T\| \leq c' \ln^4(n\rho/\beta p) \cdot \frac{\sigma p}{\sqrt{\rho} n}.$$

Recall from Lemma 2.1 that this setting of λ is exactly what we need to achieve ρ-zCDP for Algorithm 1.

2.5. Characterizing the Effect of Clipping

In Sections 2 and 3.2, we analyze DP-GD when there is no clipping. However, the optimization problem remains well-specified, even under significant clipping

Song et al. (2020) showed that for generalized linear losses, the post-clipping gradients correspond to a different convex loss, which for linear regression relates to the well-studied *Huber loss*. For parameter $B > 0$, define

$$\ell_B(\theta; \mathbf{x}, y) = \begin{cases} \frac{1}{2}(y_i - \langle \mathbf{x}_i, \theta \rangle)^2, & \text{if } |y_i - \langle \mathbf{x}_i, \theta \rangle| \leq B \\ B\left(|y_i - \langle \mathbf{x}_i, \theta \rangle| - \frac{1}{2}B\right), & \text{otherwise.} \end{cases}$$

In our setting, the individual loss $\frac{1}{2}(y_i - \langle \mathbf{x}_i, \theta \rangle)^2$ after clipping corresponds to the Huber loss with per-datum parameter $B_i = \frac{\gamma}{\|\mathbf{x}_i\|}$. Compare with Avella-Medina et al. (2021), who perform private gradient descent with a loss similar to $\ell_B(\theta; \mathbf{x}, y) \cdot \min\left\{1, \frac{2}{\|\mathbf{x}\|^2}\right\}$ for fixed B.

3. Constructing Confidence Intervals

In this section we present three methods for per-coordinate confidence intervals and provide coverage guarantees for two.

3.1. Methods for Confidence Intervals

Each construction creates a list $\theta^{(1)}, \ldots, \theta^{(m)}$ of parameter estimates and computes the sample mean $\bar{\theta} = \frac{1}{m}\sum_{\ell=1}^m \theta^{(\ell)}$ and, for every $j \in [p]$, the sample variance $\hat{\sigma}_j^2 = \frac{1}{m-1}\sum_{\ell=1}^m (\theta_j^{(\ell)} - \bar{\theta}_j)^2$. The confidence interval is constructed as if the iterates came from a Gaussian distribution with unknown mean and variance:

$$\bar{\theta}_j \pm t_{\alpha/2, m-1} \cdot \frac{\hat{\sigma}_j}{\sqrt{m}},$$

where $t_{\alpha, m-1}$ denotes the α-th percentile of the Student's t distribution with $m - 1$ degrees of freedom.

The methods, then, differ in how they produce the estimates.

Independent Runs We run Algorithm 1 m times independently and take $\theta^{(\ell)}$ to be the final iterate of the ℓ-th run.

This estimator is simple and, since the estimates are independent, easy to analyze. However, this method requires paying for m burn-in periods.

Checkpoints We run Algorithm 1 for mT time steps and take $\theta^{(\ell)}$ to be the ℓT-th iterate.

Our theorems show that if the checkpoints are sufficiently separated then they are essentially independent. While this approach pays the burn-in cost only once, it disregards information by using only a subset of the iterates.

All Iterates/Batched Means The empirical variance of batched means from a single run of mT time steps. Formally, we separate the iterates of Algorithm 1 into m disjoint batches each of length T and set $\theta^{(\ell)}$ to the empirical mean of the ℓ-th batch.

This approach may make better use of the privacy budget but poses practical challenges. The batch size needs to be large relative to the autocorrelation, but we also require several batches (as a rule of thumb, at least 10).

In the next subsection, we provide formal guarantees for the first two methods. We note that our experiments use methods that are slightly more practical but less amenable to analysis, for instance replacing "final iterate" with an average of the final few iterates. For further discussion of these details, see Section 4 and Appendix B.

3.2. Formal Guarantees

To highlight the relevant aspects, in this section we state guarantees for DP-GD without clipping and under the assumption that the input data satisfies the No-Clipping Condition. We can replace these restrictions with a distributional assumption, as in Section 2.

Theorem 3.1 (Coverage). *Fix a data set $(\mathbf{X}, \mathbf{y})$, step size η, and noise scale λ. Fix $\alpha, \beta \in (0, 1)$ and integers m, T. Assume the data satisfies Condition 2.4 with values (σ, η, c_0, mT). Let $\hat{\theta}$ be the OLS solution.*

There exists a constant c such that, if $T \geq c \log \frac{\sigma m p}{\eta \lambda \beta}$, then Equation (3.1) is a $1-\alpha-\beta$ confidence interval[3] for $\hat{\theta}_j$ when the parameter estimates $\{\theta^{(\ell)}\}_{\ell=1}^m$ are produced in either of the following ways:

- *Independent Runs: repeat DP-GD m times independently, each with no clipping and T time steps. Let $\theta^{(\ell)}$ be the final iterate of each run.*
- *Checkpoints: run DP-GD with no clipping for mT steps. Let $\theta^{(\ell)}$ be the ℓT-th iterate.*

4. Experiments

We perform experiments to confirm and complement our theoretical results. Unless otherwise mentioned, we generate data in $p = 10$ dimensions by drawing θ^* randomly from the unit sphere, $\mathbf{x}_i$ from $\mathcal{N}(0, \mathbb{I})$, and $y_i = \mathbf{x}_i^\dagger \theta^* + \xi_i$, where $\xi_i \sim \mathcal{N}(0,1)$. The hyperparameters for gradient descent are: clipping threshold $\gamma = 5\sqrt{p}$, number of steps $T = 10$, step size $\eta = \frac{1}{3}$, and privacy parameters $\rho = 0.015$ unless otherwise stated.[4] We include more details and results in Appendix B.

4.1. Error, Dimension, and a Bias-Variance Tradeoff

We highlight how our algorithm's error depends linearly on p, in contrast to standard approaches that require $p^{3/2}$ examples. Prior algorithms with formal accuracy analysis demonstrating this linear dependence limit experiments to modest ($p \approx 10$) dimensions (Varshney et al., 2022; Liu et al., 2023). We show that our approach can achieve privacy "for free:" at reasonable sample sizes, the error from sampling dominates the error due to privacy.

We first explore the accuracy of DP-GD and how it depends on the dimension and sample size. Figure 1, presented in the introduction, shows the "iso-accuracy" lines of DP-GD: as the dimension grows, how many samples are needed to maintain a fixed level of error? These lines represent expected ℓ_2 error equal to $\frac{1}{2}$; in our setting, an error of one is trivial. (See Appendix B for further details.) These results demonstrate that the error of DP-GD is constant when p/n is constant. We compare with OLS and the well-known AdaSSP algorithm (Wang, 2018). In contrast to the other two approaches, the number of samples AdaSSP needs grows with $p^{3/2}$ in this plot.

Figure 2 fixes the dimension and allows the sample size to grow. We see that the error due to privacy noise (i.e., $\|\theta_T - \hat{\theta}\|$) falls off with $1/n$, while the error due to sampling (i.e., $\|\hat{\theta} - \theta^*\|$) falls off with $1/\sqrt{n}$. At larger sample sizes, the non-private error dominates. This experiment has a larger privacy budget to highlight the effect; see Figure 8 in Appendix B for additional results.

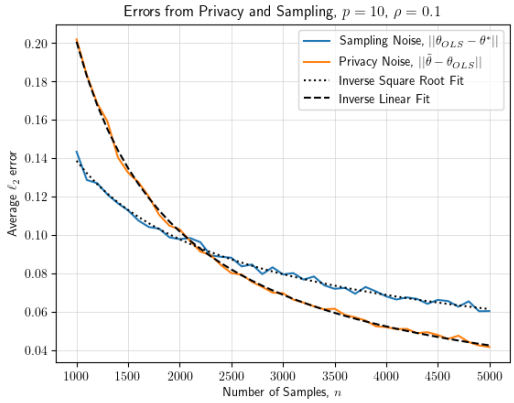


Figure 2: The "cost of privacy:" fixing the dimension and allowing the sample size to grow, we see how the error due to sampling dominates the error from privacy. Each point is averaged over 100 independent trials.

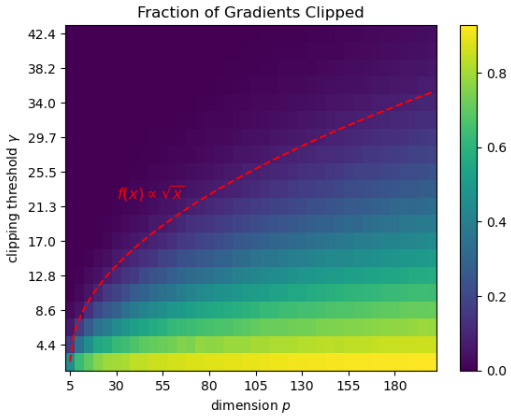


Figure 3: We see the fraction of gradients clipped over a grid on dimension and clipping threshold. As the theory predicts, we see low clipping with $\gamma = \Omega(\sqrt{p})$.

The clipping threshold in these experiments was fixed to $\gamma = 5\sqrt{p}$, as guided by our theory that no clipping occurs when $\gamma \gtrsim \sqrt{p}$. Figure 3 validates this theoretical result across dimensions: as the dimension grows, clipping thresholds much larger than $\sqrt{p}$ induce negligible clipping. We then move deliberately beyond this no-clipping regime in Figure 4, revealing a bias-variance tradeoff and highlighting how the lowest error may occur under significant clipping.

[3] That is, with probability at least $1-\alpha-\beta$ the interval contains the parameter of interest.

[4] This corresponds to an (ε, δ)-DP guarantee with $\varepsilon = 0.925$ and $\delta = 10^{-6}$, see Claim A.6.

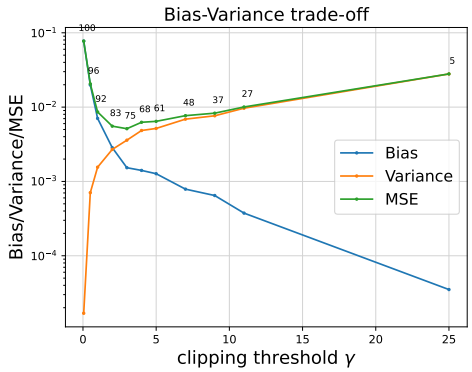


Figure 4: We plot the error, (squared) bias, and variance of DP-GD as we change the clipping threshold. The numeric labels give the percentage of all gradients clipped. Low thresholds cause high clipping and bias, while high thresholds have little clipping but high variance.

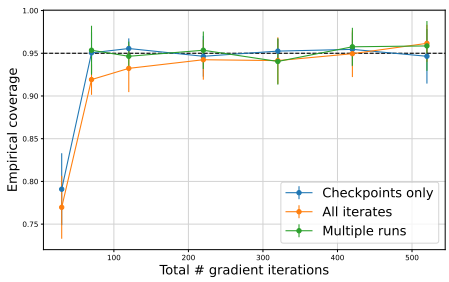


Figure 5: Average empirical coverage across co-ordinates over 100 algorithm runs. Error bars reflect the 95-percentiles of coverage across coordinates.

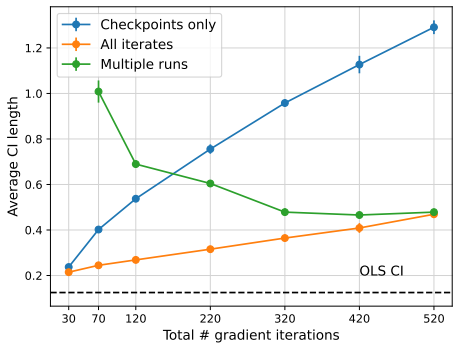


Figure 6: Average confidence interval length for each construction, as the total number of gradient iterations increases. Error bars reflect 95-th percentiles. For comparison, the dashed line shows the width of standard nonprivate OLS confidence intervals for the population quantity.

4.2. Confidence Intervals

Finally, we evaluate the three confidence interval constructions from Section 3, comparing their empirical coverage and interval width. We vary the total number of gradient iterations to clarify the regimes where each method performs well. Using the notation from Section 3, we use $m = 10$ runs/checkpoints/batches and vary T. We place the total number of gradient updates (i.e., the product mT) on the x axis. For more details, see Appendix B.

Figure 5 shows the constructions' coverage properties as a function of the total number of gradients. Figure 6 shows the average length of the confidence intervals. We see that all produce valid confidence intervals but the relative efficiency differs. With fewer iterations the burn-in period is proportionally longer, so running the algorithm multiple times yields wider confidence intervals. In contrast, running the algorithm longer induces more auto-correlation between the iterates which means a larger batch-size is required to obtain valid intervals from the batched means approach. The checkpoints approach has a poor dependence on the number of iterations since a large fraction are disregarded.

Acknowledgements

We thanks our anonymous reviewers for their feedback.

While at the University of Washington, G.B. was supported by NSF Award 2019844. Part of this work was completed while G.B. was at Boston University. A.S. and G.B., while at Boston University, were supported in part by NSF awards CCF-1763786 and CNS-2120667, Faculty Awards from Google and Apple, and US Census Bureau cooperative agreement CB16ADR0160001.

Impact Statement

Our work fits into the large body of literature on the design and analysis of differentially private algorithms. Since our main contributions lie in the evaluation of a well-known algorithm, and since this evaluation comes in the form of theoretical analysis and experiments on synthetic data, we do not discuss any specific social impacts.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS'16)*, pp. 308–318, 2016.

Alabi, D., McMillan, A., Sarathy, J., Smith, A., and Vadhan, S. Differentially private simple linear regression. *Proceedings on Privacy Enhancing Technologies*, 2022.

Amin, K., Joseph, M., Ribero, M., and Vassilvitskii, S. Easy differentially private linear regression. *arXiv preprint arXiv:2208.07353*, 2022.

Avella-Medina, M., Bradshaw, C., and Loh, P.-L. Differentially private inference via noisy optimization. *arXiv preprint arXiv:2103.11003*, 2021.

Bassily, R., Smith, A., and Thakurta, A. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS)*, pp. 464–473, 2014.

Brawner, T. and Honaker, J. Bootstrap inference and differential privacy: Standard errors for free. 2018.

Brown, G., Hayase, J., Hopkins, S., Kong, W., Liu, X., Oh, S., Perdomo, J. C., and Smith, A. Insufficient statistics perturbation: Stable estimators for private least squares. *arXiv preprint arXiv:2404.15409*, 2024.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

Cai, T. T., Wang, Y., and Zhang, L. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5): 2825–2850, 2021.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

Chen, X., Lee, J. D., Tong, X. T., and Zhang, Y. Statistical inference for model parameters in stochastic gradient descent. *Annals of Statistics*, 48(1):251–273, 2020.

Covington, C., He, X., Honaker, J., and Kamath, G. Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*, 2021.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380, 2009.

Dwork, C. and Rothblum, G. N. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, pp. 265–284, 2006. URL `http://dx.doi.org/10.1007/11681878_14`.

Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 11–20, 2014.

Ferrando, C. and Sheldon, D. Private regression via data-dependent sufficient statistic perturbation. *arXiv preprint arXiv:2405.15002*, 2024.

Ferrando, C., Wang, S., and Sheldon, D. Parametric bootstrap for differentially private confidence intervals. In *International Conference on Artificial Intelligence and Statistics*, pp. 1598–1618. PMLR, 2022.

Foulds, J., Geumlek, J., Welling, M., and Chaudhuri, K. On the theory and practice of privacy-preserving bayesian data analysis. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 192–201, 2016.

Hsu, D., Kakade, S. M., and Zhang, T. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 6, 2011.

Kamath, G., Li, J., Singhal, V., and Ullman, J. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pp. 1853–1902. PMLR, 2019.

Kamath, G., Mouzakis, A., and Singhal, V. New lower bounds for private estimation and a generalized fingerprinting lemma. *Advances in Neural Information Processing Systems*, 35:24405–24418, 2022.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. D. What can we learn privately? In *49th Annual IEEE Symp. on Foundations of Computer Science (FOCS)*, pp. 531–540, 2008.

Knop, A. and Steinke, T. Differentially private linear regression via medians. 2022.

Kurakin, A., Song, S., Chien, S., Geambasu, R., Terzis, A., and Thakurta, A. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Li, X., Tramèr, F., Kulkarni, J., and Hashimoto, T. Differentially private deep learning can be effective with self-supervised models. https://differentialprivacy.org/dp-fine-tuning/, 2022. URL https://differentialprivacy.org/dp-fine-tuning/.

Liu, X., Kong, W., and Oh, S. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pp. 1167–1246. PMLR, 2022.

Liu, X., Jain, P., Kong, W., Oh, S., and Suggala, A. S. Near optimal private and robust linear regression. *arXiv preprint arXiv:2301.13273*, 2023.

McSherry, F. and Talwar, K. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pp. 94–103. IEEE, 2007.

Milionis, J., Kalavasis, A., Fotakis, D., and Ioannidis, S. Differentially private regression with unbounded covariates. In *International Conference on Artificial Intelligence and Statistics*, pp. 3242–3273. PMLR, 2022.

Narayanan, S. Better and simpler fingerprinting lower bounds for differentially private estimation. 2023.

Rabanser, S., Thudi, A., Thakurta, A., Dvijotham, K., and Papernot, N. Training private models that know what they don't know. *arXiv preprint arXiv:2305.18393*, 2023.

Sarathy, J. and Vadhan, S. Analyzing the differentially private theil-sen estimator for simple linear regression. *arXiv preprint arXiv:2207.13289*, 2022.

Sheffet, O. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pp. 3105–3114. PMLR, 2017.

Sheffet, O. Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, pp. 789–827. PMLR, 2019.

Shejwalkar, V., Ganesh, A., Mathews, R., Thakkar, O., and Thakurta, A. Recycling scraps: Improving private learning by leveraging intermediate checkpoints. *arXiv preprint arXiv:2210.01864*, 2022.

Song, S., Thakkar, O., and Thakurta, A. Characterizing private clipped gradient descent on convex generalized linear problems. *arXiv preprint arXiv:2006.06783*, 2020.

Tang, S., Aydore, S., Kearns, M., Rho, S., Roth, A., Wang, Y., Wang, Y.-X., and Wu, Z. S. Improved differentially private regression via gradient boosting. *arXiv preprint arXiv:2303.03451*, 2023.

Varshney, P., Thakurta, A., and Jain, P. (nearly) optimal private linear regression via adaptive clipping. *arXiv preprint arXiv:2207.04686*, 2022.

Vu, D. and Slavkovic, A. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 138–143. IEEE, 2009.

Wang, Y., Kifer, D., and Lee, J. Differentially private confidence intervals for empirical risk minimization. *Journal of Privacy and Confidentiality*, 9(1), 2019.

Wang, Y.-X. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *Conference on Uncertainty in Artificial Intelligence*, 2018. URL https://api.semanticscholar.org/CorpusID:3740480.

Wang, Z., Cheng, G., and Awan, J. Differentially private bootstrap: New privacy analysis and inference strategies. *arXiv preprint arXiv:2210.06140*, 2022.

Xu, Z., Zhang, Y., Andrew, G., Choquette-Choo, C. A., Kairouz, P., McMahan, H. B., Rosenstock, J., and Zhang, Y. Federated learning of gboard language models with differential privacy. *arXiv preprint arXiv:2305.18465*, 2023.

A. Preliminaries

Table 1: A high-level view of various approaches for private linear regression and their dimension-dependence in the setting of Definition 1.1.

Approach	Examples	Samples Required
SSP	Sheffet (2017); Wang (2018)	$\Omega(p^{3/2})$
HPTR (exp. time)	Liu et al. (2022)	$\Omega(p)$
Sample-and-Aggregate	Knop & Steinke (2022); Amin et al. (2022)	$\Omega(p^2)$
Gradient Descent (DP-GD)	Bassily et al. (2014); Cai et al. (2021)	$\Omega(p^{3/2})$
DP-GD w/Adaptive Clipping	Varshney et al. (2022); Liu et al. (2023)	$\Omega(p)$
Gradient Descent (DP-GD)	**Our Work**	$\Omega(p)$

A.1. Differential Privacy

We present definitions and basic facts about differential privacy.

Definition A.1 (Approximate Differential Privacy)**.** *For $\varepsilon \geq 1$, and $\delta \in (0,1)$, a mechanism A mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-differential privacy if, for all $x, x' \in \mathcal{X}^n$ that differ in one entry and all measurable events $E \subseteq \mathcal{Y}$,*

$$\mathbf{Pr}[M(x) \in E] \leq e^{\varepsilon} \mathbf{Pr}[M(x') \in E] + \delta.$$

The guarantees we present in this work are in terms of *(zero) concentrated differential privacy* (Dwork & Rothblum, 2016; Bun & Steinke, 2016), a variant of differential privacy that allows us to cleanly express the privacy guarantees of DP-GD.

Definition A.2 (ρ-zCDP)**.** *A mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies* ρ-zero concentrated differential privacy *(ρ-zCDP) if for all $x, x' \in \mathcal{X}$ that differ in one entry and all $\alpha \in (1, \infty)$, we have $D_\alpha(M(x) \| M(x')) \leq \rho\alpha$.*

Definition A.3 (Rényi Divergence)**.** *For two distributions p, w and $\alpha \geq 1$, the* α-Rényi divergence *is $D_\alpha(p\|q) = \frac{1}{\alpha - 1} \log \mathbf{E}_{y \sim q} \left[\left(\frac{p(y)}{q(y)} \right)^\alpha \right]$.*

The privacy guarantees for DP-GD follow composition plus the privacy guarantee for the standard multivariate Gaussian mechanism. zCDP allows us to cleanly express both.

Claim A.4 (Composition)**.** *Suppose mechanism M satisfies ρ-zCDP and mechanism M' satisfies ρ'-zCDP. Then (M, M') satisfies $(\rho + \rho')$-zCDP.*

Claim A.5 (Gaussian Mechanism)**.** *Let $q : \mathcal{X}^n \to \mathcal{Y}$ satisfy, for all $x, x' \in \mathcal{X}^n$ that differ in one entry, $\|q(x) - q(x')\|_2 \leq \Delta$. Then, for any $\lambda > 0$, the mechanism $M(x) = \mathcal{N}(q(x), \lambda^2 \mathbb{I})$ is $\frac{\Delta^2}{2\lambda^2}$-zCDP.*

We apply Claim A.5 to the average of gradients, where each gradient norm is clipped to γ. The straightforward sensitivity analysis shows that we can set $\Delta = \frac{2\gamma}{n}$.

A mechanism satisfying ρ-zCDP also satisfies (ε, δ)-differential privacy. In fact, it provides a continuum of such guarantees: one for every $\delta \in (0, 1)$.

Claim A.6. *If M satisfies ρ-zCDP, then M satisfies $(\rho + 2\sqrt{\rho \log 1/\delta}, \delta)$-differential privacy for any $\delta > 0$.*

A.2. Linear Algebra

Fact A.7 (Matrix Geometric Series)**.** *Let $\mathbf{T}$ be an invertible matrix. Then $\sum_{j=0}^{n-1} \mathbf{T}^n = (\mathbb{I} - \mathbf{T})^{-1}(\mathbb{I} - \mathbf{T}^n)$.*

A.3. Concentration Inequalities

Claim A.8. *If $x \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma^2 > 0$, then $\mathbf{Pr}\left[|x| \geq \sigma\sqrt{2 \ln 2/\beta}\right] \leq \beta$.*

Claim A.9. *Fix the number of dimensions p and a PSD matrix Σ. Let $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$. For any $\beta \in (0,1)$, we have*

$$\mathbf{Pr}\left[\|\mathbf{x}\| \geq \sqrt{\mathrm{tr}(\Sigma)} + \sqrt{2\|\Sigma\| \log 1/\beta}\right] \leq \beta.$$

Claim A.10 (Concentration of Covariance)**.** *Fix $\beta \in (0,1)$. Draw independent $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(0, \mathbb{I})$ and let $Z = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\dagger$. There exists a constant c such that, if $n \geq c(p + \log 1/\beta)$, then with probability at least $1-\beta$ we have $\frac{1}{2}\mathbb{I} \preceq Z \preceq 2\mathbb{I}$.*

Lemma A.11. *Let X be a uniform distribution on the unit sphere $\mathbb{S}^{p-1}$, and z be any fixed unit vector. Then we know the inner product $\langle X, z\rangle$ is sub-exponential. Specifically, we have*

$$\mathbf{Pr}[|\langle X, z\rangle| \geq t] \leq e^{-\frac{\sqrt{p}-1}{4}t}.$$

The classic analysis of least squares under fixed design establishes the convergence of the OLS estimator to the true parameter. A nearly identical result holds under random design. We state the result for the family of distributions we consider, but (Hsu et al., 2011) prove it for a much broader family of distributions.

Claim A.12 (Theorem 1 of (Hsu et al., 2011))**.** *Let θ^* satisfy $\|\theta^*\| \leq 1$. Draw covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ i.i.d. from $\mathcal{N}(0, \mathbb{I}_p)$ and let $y_i = \mathbf{x}_i^\dagger \theta^* + \xi_i$ for $\xi_i \sim \mathcal{N}(0, \sigma^2)$. Let $\hat{\theta}$ be the OLS estimate. There exists constants c, c' such that, if $n \geq c(p + \ln 1/\beta)$, then with probability at least $1-\beta$ we have*

$$\|\theta^* - \hat{\theta}\|^2 \leq \frac{c'\sigma^2\,(p + \ln 1/\beta)}{n}.$$

B. Experimental Details and Additional Results

Details on Iso-Accuracy Plots Figures 1 and 9 show the number of samples needed to achieve expected error $\frac{1}{2}$. Formally, a dot for one algorithm (e.g., DP-GD) at (p^*, n^*) means that we ran $m = 50$ independent trials in p^* dimensions with n^* examples and observed that the average error was approximately one half:

$$\left|\frac{1}{2} - \frac{1}{m}\sum_{i=1}^m \|\theta_i^* - \widetilde{\theta}_i\|_2\right| \leq \frac{1}{100}.$$

Here $\theta_i^* \in \mathbb{R}^p$ represents the true parameter in the i-th trial, and $\widetilde{\theta}_i$ the respective estimate produced by DP-GD. In these experiments θ^* is chosen randomly from the unit ball, so an error of one is trivial. For each value of p and each algorithm, we find the corresponding value of n via binary search.

Details on Confidence Interval Experiments Our experiments hold fixed the burn-in period to the first 20 iterates. We set m, the number of algorithm runs/checkpoints/batches to 10 while varying the total number of gradient iterations. (This is not always possible for multiple runs while holding fixed the total number of iterations, in which case we omit this approach from the figures.) We vary the total number of iterations in this way to clarify the regimes where each method performs well.

In Figure 7 we show how the rate of gradient clipping impacts ℓ_2-error, and how this interacts with the total number of gradient iterations. One key insight is that error stays relatively low even when clipping 50% of all gradients. As the number of gradient iterations grows, the tolerance to gradient clipping also seems to increase. Although our theoretical accuracy analysis proceeds by showing that no gradients are clipped, these experiments demonstrate that this stringent requirement is not necessary in practice.

Experiments with Anisotropic Data Our primary experiments are conducted on data sets where the covariates are drawn from a standard multivariate Gaussian distribution. We now move beyond this isotropic setting. In these experiments, for each dataset we first draw a random covariance matrix in the following way. We generate a random diagonal matrix Λ with $\Lambda_{1,1} = 2$, $\Lambda_{2,2} = 1$, and, for all $3 \leq i \leq p$, $\Lambda_{i,i} \sim \mathrm{Unif}([1,2])$ independently. We then set the covariance $\Sigma = U\Lambda U^T$, where U is a uniformly random rotation matrix. The covariates are then sampled i.i.d. from $\mathcal{N}(0, \Sigma)$; the remainder of the process is identical to the previous experiments.

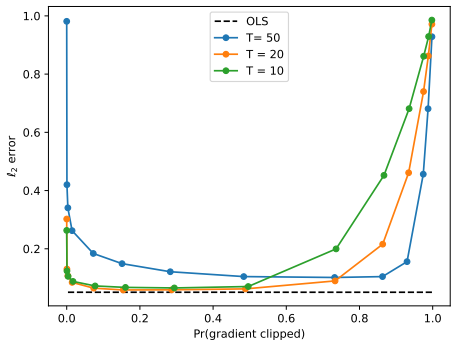
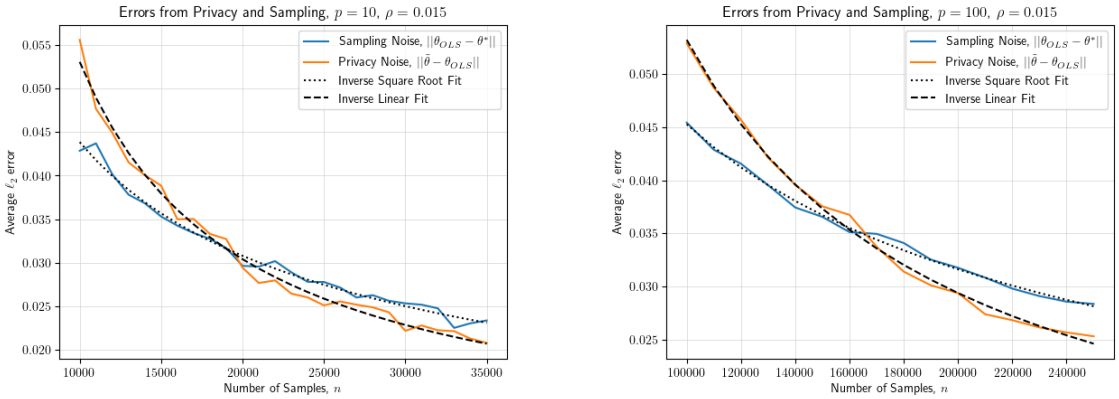


Figure 7: The estimation error plotted against the proportion of gradients clipped. The right-hand side corresponds to bias from overly aggressive clipping. The left-hand side corresponds to variance from overly conservative clipping, which causes higher levels of noise for privacy.



(a) Experiment with $p = 10$

(b) Experiment with $p = 100$

Figure 8: As in Figure 2, we fix the dimension and let the sample size grow, seeing that the sampling error dominates the noise from privacy at reasonable sample sizes. These experiments are conducted with $\rho = 0.015$, our standard setting. (a) uses $p = 10$ and repeats each trial 100 times. (b) uses $p = 100$ and repeats each trial 20 times, to reduce running time.

C. Deferred Proofs

C.1. Clipping, Coupling, and Accuracy

Before proving Lemma 2.2, we define coupling and show how it relates to total variation distance.

Definition C.1 (Coupling)**.** *Let p and q be distributions over a space $\mathcal{X}$. A pair of random variables (X, Y) is called a* coupling *of (p, q) if, for all $x \in \mathcal{X}$, $\mathbf{Pr}[X = x] = p(x)$ and $\mathbf{Pr}[Y = x] = q(x)$.*

Note that X and Y will not, in general, be independent.

Claim C.2 (Coupling and TV Distance)**.** *Let (X, Y) be a coupling of (p, q). Then $TV(p, q) \leq \mathbf{Pr}[X \neq Y]$.*

Lemma C.3 (Restatement of Lemma 2.2)**.** *Fix data set $\mathbf{X}, \mathbf{y}$ and hyperparameters γ, λ, η, T, and θ_0. Define random variables $\mathcal{O}_\gamma$ and $\mathcal{O}_\infty$ as*

$$\mathcal{O}_\gamma = \mathcal{A}(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, \theta_0)$$
$$\mathcal{O}_\infty = \mathcal{A}(\mathbf{X}, \mathbf{y}; \infty, \lambda, \eta, T, \theta_0).$$

If Algorithm 1 on input $(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, \theta_0)$ clips nothing with probability $1 - \beta$, then $TV(\mathcal{O}_\gamma, \mathcal{O}_\infty) \leq \beta$.

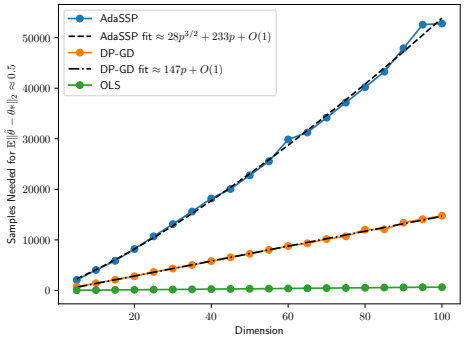


Figure 9: We reproduce Figure 1 with anisotropic data. We plot iso-accuracy lines as p grows. Run on data from a well-specified linear model, both DP-GD and OLS require a number of samples that grow linearly with the dimension. We compare with AdaSSP (Wang, 2018), a popular algorithm that requires $n = \Omega(p^{3/2})$ examples.

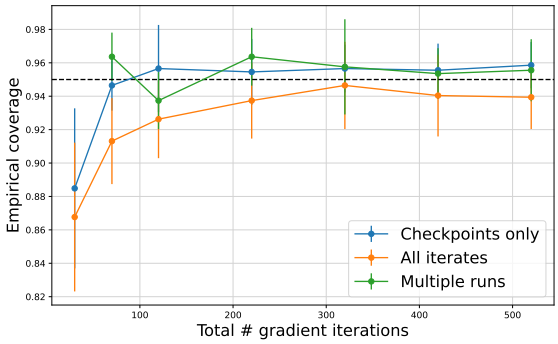


Figure 10: We reproduce Figure 5 with anisotropic data. Average empirical coverage across co-ordinates over 100 algorithm runs. Error bars reflect the 95-percentiles of coverage across coordinates.

Proof. We use the coupling induced by sharing randomness across the execution of the two algorithms. Let $\mathbf{z}_1, \ldots, \mathbf{z}_T$ be drawn i.i.d. from $\mathcal{N}(0, \lambda^2 \mathbb{I})$. Note that this is the only randomness used by either algorithm.

When the $\mathbf{z}_i$ draws cause $\mathcal{A}(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T)$ to not clip, the two algorithms return the same output. Thus the probability two runs return *different* outputs is at most β, which yields our bound on the total variation distance. □

We now prove our statement about the distribution of DP-GD without clipping. The statement we prove also establishes a slightly more complicated expression for the noise.

Lemma C.4 (Expanded Statement of Lemma 2.3)**.** *Fix a data set $(\mathbf{X}, \mathbf{y})$ and step size η, noise scale λ, number of iterations T, and initial vector θ_0. Define matrices $\Sigma = \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}$ and $\mathbf{D} = (\mathbb{I} - \eta\Sigma)^2$; assume both are invertible. Let $\hat{\theta} = (\mathbf{X}^\dagger\mathbf{X})^{-1}\mathbf{X}^\dagger\mathbf{y}$ be the least squares solution. Consider $\mathcal{A}(\mathbf{X}, \mathbf{y}; \infty, \lambda, \eta, T, \theta_0)$, i.e., Algorithm 1 without clipping. For any $t \in [T]$, we have*

$$\theta_t = \hat{\theta} + (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta}) + \eta \sum_{i=1}^{t} (\mathbb{I} - \eta\Sigma)^{i-1}\mathbf{z}^{t-i}.$$

This is equal in distribution to

$$\theta_t = \hat{\theta} + (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta}) + \eta \cdot \mathbf{z}'_t$$

for $\mathbf{z}'_t \sim \mathcal{N}(0, \lambda^2\mathbf{A}^{(t)})$ with $\mathbf{A}^{(t)} = (\mathbb{I} - \mathbf{D})^{-1}(\mathbb{I} - \mathbf{D}^t)$.

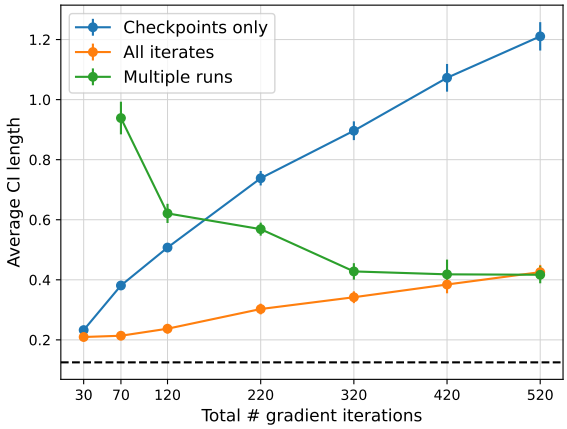


Figure 11: We reproduce Figure 6 with anisotropic data. Average confidence interval length across co-ordinates for each confidence interval algorithm described in Section 3 as the total number of gradient iterations increases. Error bar reflect the 95-percentiles of coordinates CI length.

Proof. Algorithm 1's update step looks like

$$\theta_{t+1} \leftarrow \theta_t - \eta \cdot \bar{\mathbf{g}}_t + \eta \cdot \mathbf{z}_t,$$

where $\mathbf{z}_t \sim \mathcal{N}(0, \lambda^2 \mathbb{I})$. Since there is no clipping, we have a closed form for $\bar{\mathbf{g}}_t$:

$$\begin{aligned}\bar{\mathbf{g}}_t &= \frac{1}{n}\sum_{i=1}^{n} -\mathbf{x}_i(y_i - \mathbf{x}_i^\dagger \theta_t) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\dagger \theta_t - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i \\ &= \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}\theta_t - \frac{1}{n}\mathbf{X}^\dagger\mathbf{y}.\end{aligned}$$

We simplify further and apply the fact that $\hat{\theta} = (\mathbf{X}^\dagger\mathbf{X})^{-1}\mathbf{X}^\dagger\mathbf{y}$. We also plug in $\Sigma = \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}$:

$$\bar{\mathbf{g}}_t = \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}\theta_t - \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}\hat{\theta} = \Sigma\theta_t - \Sigma\hat{\theta}.$$

Plugging this into Equation (C.1), the update formula, we have

$$\theta_{t+1} \leftarrow \eta \cdot \Sigma\hat{\theta} + (\mathbb{I} - \eta\Sigma)\theta_t + \eta \cdot \mathbf{z}_t.$$

Solving the recursion, we arrive at a formula for θ_{t+1}:

$$\theta_{t+1} = (\mathbb{I} - \eta\Sigma)^t\theta_0 + \sum_{i=1}^{t}(\mathbb{I} - \eta\Sigma)^{i-1}\eta\Sigma\hat{\theta} + \sum_{i=1}^{t}(\mathbb{I} - \eta\Sigma)^{i-1}\eta \cdot \mathbf{z}_{t-i}.$$

To simplify the second term in Equation (C.1), apply the formula for matrix geometric series (Fact A.7):

$$\sum_{i=1}^{t}(\mathbb{I} - \eta\Sigma)^{i-1} = (\eta\Sigma)^{-1}(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t).$$

Thus the second term in Equation (C.1) is $(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\hat{\theta} = \hat{\theta} - (\mathbb{I} - \eta\Sigma)^t\hat{\theta}$. The first two terms together are $\hat{\theta} + (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta})$. This establishes the first part of the claim.

The final term in Equation (C.1), corresponding to the noise for privacy, is slightly more involved. We use the independence of the noise and the fact that Gaussianity is preserved under summation. Continuing from the third term of Equation (C.1) and abusing notation to substitute distributions for random variables, we have

$$\begin{aligned}
&\sum_{i=1}^{t}(\mathbb{I}-\eta\Sigma)^{i-1}\eta\cdot\mathbf{z}_{t-i}\\
&=\sum_{i=1}^{t}(\mathbb{I}-\eta\Sigma)^{i-1}\eta\cdot\mathcal{N}(0,\lambda^2\mathbb{I})\\
&=\eta\lambda\sum_{i=1}^{t}\mathcal{N}(0,(\mathbb{I}-\eta\Sigma)^{2(i-1)})\\
&=\eta\lambda\cdot\mathcal{N}\left(0,\sum_{i=1}^{t}(\mathbb{I}-\eta\Sigma)^{2(i-1)}\right)\\
&=\eta\lambda\cdot\mathcal{N}\left(0,(\mathbb{I}-(\mathbb{I}-\eta\Sigma)^2)^{-1}(\mathbb{I}-(\mathbb{I}-\eta\Sigma)^{2t})\right),
\end{aligned}$$

applying the formula for matrix geometric series (Fact A.7) in the last line. This concludes the proof. □

Lemma C.5 (Restatement of Lemma 2.5)**.** *Fix nonnegative real numbers σ, η, and c_0. Fix natural number T. Assume data set $(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$ satisfies the No-Clipping Condition (Condition 2.4) with values (σ, η, c_0, T).*

Fix nonnegative real numbers γ and λ. Consider running Algorithm 1, i.e., $\mathcal{A}(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, 0)$ with $\theta_0 = 0$ the initial point. Assume $\gamma \geq 4c_0^2\sigma\sqrt{p}$. For any $\beta \in (0, 1)$, if $\frac{\gamma}{\eta\lambda} \geq 64c_0^2 p\sqrt{\ln 2nT/\beta}$, then with probability $1-\beta$ Algorithm 1 clips no gradients.

Proof. We prove the claim by strong induction, relying on items (i-v) of the No-Clipping Condition (Condition 2.4). Before beginning the induction, we prove a high-probability statement about the noise added for privacy.

Setup: Noise for Privacy Recall that, at time t, we add independent noise $\mathbf{z}_t \sim \mathcal{N}(0, \lambda^2\mathbb{I})$. We show that, with probability at least $1-\beta$, we have for all $i \in [n]$ and $t_1, t_2 \in [T]$

$$|\mathbf{x}_i^\dagger(\mathbb{I}-\eta\Sigma)^{t_1}\mathbf{z}_{t_2}| \leq 2c_0\lambda\sqrt{p}\cdot\left(\frac{7}{8}\right)^{t_1}\sqrt{\ln 2nT/\beta}.$$

For any fixed covariates, we have

$$\mathbf{x}_i^\dagger(\mathbb{I}-\eta\Sigma)^{t_1}\mathbf{z}_{t_2} \sim \lambda\cdot\mathcal{N}\left(0, \|(\mathbb{I}-\eta\Sigma)^{t_1}\mathbf{x}_i\|^2\right).$$

With probability at least $1-\beta$ (simultaneously over all i, t_1, and t_2) we have

$$|\mathbf{x}_i^\dagger(\mathbb{I}-\eta\Sigma)^{t_1}\mathbf{z}_{t_2}| \leq 2\lambda\|(\mathbb{I}-\eta\Sigma)^{t_1}\mathbf{x}_i\|\cdot\sqrt{\ln 2nT/\beta}.$$

This holds independently of the value of the norm, so Conditions (i) and (ii) (and Cauchy–Schwarz repeatedly) proves Equation (C.1).

We now proceed to the induction. Let $g_{i,t} = -\mathbf{x}_i(y_i - \mathbf{x}_i^\dagger\theta_t)$ be the gradient of the loss on point i with respect to parameter θ_t. Observe that $\|g_{i,t}\| = |y_i - \mathbf{x}_i^\dagger\theta_t|\cdot\|\mathbf{x}_i\|$ and, furthermore, that Condition (ii) promises $\|\mathbf{x}_i\| \leq c_0\sqrt{p}$. Thus, to show that the norm of the gradient is less than γ we will show that the absolute residual is less than $\frac{\gamma}{c_0\sqrt{p}}$.

Base Case: for $t = 1$, we start from $\theta_0 = 0$, which means our residual is just $|y_i|$. By the triangle inequality and Conditions (iii) and (iv), we have

$$\begin{aligned}
|y_i| = |y_i - \mathbf{x}_i^\dagger\theta^* + \mathbf{x}_i^\dagger\theta^*| &\leq |y_i - \mathbf{x}_i^\dagger\theta^*| + |\mathbf{x}_i^\dagger\theta^*|\\
&\leq c_0\sigma + c_0.
\end{aligned}$$

This is less than $\frac{\gamma}{c_0\sqrt{p}}$ by assumption.

Induction Step: Consider time $t+1$ and assume that no gradients have been clipped from the start through time t. From Lemma C.4, we have a formula for the value of θ_t:

$$\theta_t = \hat{\theta} + (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta}) + \eta \mathbf{z}'_t.$$

Plugging this equation into the formula for the residual at time t, we have (after adding and subtracting identical terms)

$$\begin{aligned}
|y_i - \mathbf{x}_i^\dagger \theta_t| &= |y_i - \mathbf{x}_i^\dagger \theta_t + \mathbf{x}_i^\dagger \theta - \mathbf{x}_i^\dagger \theta| \\
&= |\mathbf{x}_i^\dagger(\theta - \theta_t) + (y_i - \mathbf{x}_i^\dagger \theta)| \\
&= |\mathbf{x}_i^\dagger \left(\theta - \hat{\theta} - (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta}) - \eta \mathbf{z}'_t\right) + (y_i - \mathbf{x}_i^\dagger \theta)| \\
&= |\mathbf{x}_i^\dagger \left(\theta - \hat{\theta} - (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta} + \theta - \theta) - \eta \mathbf{z}'_t\right) + (y_i - \mathbf{x}_i^\dagger \theta)|,
\end{aligned}$$

where in the final line we added and subtracted θ. We distribute terms and apply the triangle inequality, arriving at

$$|y_i - \mathbf{x}_i^\dagger \theta_t| \leq \underbrace{|\mathbf{x}_i^\dagger \left(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t\right)(\theta - \hat{\theta})|}_{\text{(i)}} + \underbrace{|\mathbf{x}_i^\dagger (\mathbb{I} - \eta\Sigma)^t(\theta_0 - \theta)|}_{\text{(ii)}} + \underbrace{|\eta \cdot \mathbf{x}_i^\dagger \mathbf{z}'_t|}_{\text{(iii)}} + \underbrace{|y_i - \mathbf{x}_i^\dagger \theta|}_{\text{(iv)}}.$$

We will show that each of the four terms in Equation (C.1) is at most $\frac{\gamma}{4} \cdot \frac{1}{c_0\sqrt{p}}$ with high probability. Combined with Condition (ii), this establishes that $\|g_{i,t}\| = |y_i - \mathbf{x}_i^\dagger \theta_t| \cdot \|\mathbf{x}_i\| \leq \gamma$, as we desire.

Term (i): We decompose the vector $(\theta - \hat{\theta})$ (see Hsu et al., 2011, Lemma 1). As in Condition (v), define matrix $A_t = (\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\Sigma^{-1}$. We have

$$\begin{aligned}
|\mathbf{x}_i^\dagger \left(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t\right)(\theta - \hat{\theta})| &= |\mathbf{x}_i^\dagger \left(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t\right) \cdot \frac{1}{n} \sum_j \Sigma^{-1}(y_j - \mathbf{x}_j^\dagger \theta)\mathbf{x}_j| \\
&= \frac{1}{n}|\sum_j (y_j - \mathbf{x}_j^\dagger \theta) \cdot \mathbf{x}_i^\dagger A_t \mathbf{x}_j| \\
&\leq \frac{1}{n} \cdot c_0 \sigma \sqrt{np},
\end{aligned}$$

applying Condition (v) in the last line. Since $n \geq p$ by assumption, term (i) is at most $c_0\sigma$, which is less than $\frac{\gamma}{4} \cdot \frac{1}{c_0\sqrt{p}}$ by assumption.

Term (ii): Since $\theta_0 = 0$, Condition (iv) directly says that term (ii) is at most c_0, which is less than $\frac{\gamma}{4} \cdot \frac{1}{c_0\sqrt{p}}$ by assumption.

Term (iii): Push $\mathbf{x}_i^\dagger$ inside the sum and apply the triangle inequality:

$$\begin{aligned}
|\eta \cdot \mathbf{x}_i^\dagger \mathbb{Z}_t| &= \eta \cdot |\sum_{\ell=1}^{t} \mathbf{x}_i^\dagger (\mathbb{I} - \eta\Sigma)^{\ell-1} \mathbf{z}_{t-\ell}| \\
&\leq \eta \sum_{\ell=1}^{t} |\mathbf{x}_i^\dagger (\mathbb{I} - \eta\Sigma)^{\ell-1} \mathbf{z}_{t-\ell}|.
\end{aligned}$$

By Equation (C.1) (which relies on Conditions (i) and (ii)), we have an upper bound on each of these terms that holds with probability at least $1 - \beta$. Plugging this in, we have

$$\begin{aligned}
\|\eta \cdot \mathbf{x}_i^\dagger \mathbf{z}'_t\| &\leq \eta \sum_{\ell=1}^{t} 2c_0\lambda\sqrt{p}\,(7/8)^{\ell-1} \sqrt{\ln 2nT/\beta} \\
&\leq 2c_0\eta\lambda\sqrt{p}\sqrt{\ln 2nT/\beta} \sum_{\ell=1}^{t} (7/8)^{\ell-1}.
\end{aligned}$$

where in the second line we have pulled out the terms that do not depend on ℓ. Because it is a geometric series, the sum is at most 8. Rearranging, we see that term (iii) is at most $\frac{\gamma}{4} \cdot \frac{1}{c_0\sqrt{p}}$ when

$$\frac{\gamma}{\eta\lambda} \geq 64c_0^2 p\sqrt{\ln 2nT/\beta},$$

which is exactly what we assumed.

Term (iv): Condition (iii) says that term (iv) is at most σc_0, which is less than $\frac{\gamma}{4} \cdot \frac{1}{c_0\sqrt{p}}$ by assumption. □

Lemma C.6. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be drawn i.i.d. from $\mathcal{N}(0, \mathbb{I})$. Let η be a real number and t an integer. Let $\Sigma = \frac{1}{n}\sum_i \mathbf{x}_i\mathbf{x}_i^\dagger$. For any i, the distribution of $(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i$ is spherically symmetric.*

Proof. Let $\mathbf{v}$ be a vector and Π an orthogonal rotation matrix. We will show that

$$\mathbf{Pr}\left[(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i = \mathbf{v}\right] = \mathbf{Pr}\left[(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i = \Pi\mathbf{v}\right],$$

using the spherical symmetry of the covariates' distribution. We write out

$$\begin{aligned}\mathbf{Pr}\left[(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i = \mathbf{v}\right] &= \mathbf{Pr}\left[\Pi(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i = \Pi\mathbf{v}\right] \\ &= \mathbf{Pr}\left[\Pi(\mathbb{I} - \eta\Sigma)\cdots(\mathbb{I} - \eta\Sigma)\mathbf{x}_i = \Pi\mathbf{v}\right] \\ &= \mathbf{Pr}\left[\Pi\left(\Pi^\dagger\Pi\right)(\mathbb{I} - \eta\Sigma)(\Pi^\dagger\Pi)\cdots(\mathbb{I} - \eta\Sigma)(\Pi^\dagger\Pi)\mathbf{x}_i = \Pi\mathbf{v}\right],\end{aligned}$$

inserting $\mathbb{I} = \Pi^\dagger\Pi$ between each term. We cancel and rearrange:

$$\mathbf{Pr}\left[(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i = \mathbf{v}\right] = \mathbf{Pr}\left[(\mathbb{I} - \eta(\Pi\Sigma\Pi^\dagger))\cdots(\mathbb{I} - \eta(\Pi\Sigma\Pi^\dagger))(\Pi\mathbf{x}_i) = \Pi\mathbf{v}\right].$$

Of course, we have $\Pi\Sigma\Pi^\dagger = \Pi\left(\frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\dagger\right)\Pi^\dagger = \frac{1}{n}\sum_i(\Pi\mathbf{x}_i)(\Pi\mathbf{x}_i)^\dagger$. Therefore, by the rotation invariance of the Gaussian distribution, we arrive at $\mathbf{Pr}\left[(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i = \mathbf{v}\right] = \mathbf{Pr}\left[(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i = \Pi\mathbf{v}\right]$. □

Lemma C.7 (Restatement of Lemma 2.6)**.** *Fix data set size n and data dimension $p \geq 2$. Fix θ^* with $\|\theta^*\| \leq 1$, let covariates $\mathbf{x}_i$ be drawn i.i.d. from $\mathcal{N}(0, \mathbb{I})$, and responses $y_i = \mathbf{x}_i^\dagger\theta^* + \xi_i$ for $\xi_i \sim \mathcal{N}(0, \sigma^2)$. Fix the step size $\eta = \frac{1}{4}$ and a natural number T. There exists a constant c such that, for any $\beta \in (0, 1)$, if $n \geq c(p + \ln 1/\beta)$ then with probability at least $1 - \beta$ data set (X, y) satisfies the No-Clipping Condition (Condition 2.4) with $c_0 = 12\ln^{1.5}(5nT/\beta)$.*

Proof. the No-Clipping Condition has five sub-conditions. We prove each holds with probability at least $1 - \beta/5$; a union bound finishes the proof. To establish Conditions (iii), (iv), and (v), we take $\theta \leftarrow \theta^*$.

Condition (i) By Claim A.10, with probability at least $1 - \beta/5$ we have $\frac{1}{2}\mathbb{I} \preceq \Sigma \preceq 2\mathbb{I}$. Applying $\eta = 1/4$ establishes Condition (i).

Condition (ii) By Claim A.9, for any fixed $\mathbf{x}_i$ we have $\|\mathbf{x}_i\| \leq \sqrt{p} + \sqrt{2\ln 5n/\beta}$ with probability at least $1 - \beta/(5n)$. Let $c_1 = (1 + \sqrt{2/p}\ln 5n/\beta)$, so we have $\|\mathbf{x}_i\| \leq c_1\sqrt{p}$; this is at most $3\sqrt{\ln 5n/\beta}$ and, furthermore, no greater than c_0. A union bound over all n covariates means this condition holds with probability at least $1 - \beta/5$.

Condition (iii) Since $\xi_i \sim \mathcal{N}(0, \sigma^2)$, we apply Claim A.8: with probability at least $1 - \beta/5$, for all $i \in [n]$ we have $|\xi_i| \leq \sigma\sqrt{2\ln 10n/\beta}$.

Condition (iv) Fix i and t. Observe that the distribution of $(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i$ is spherically symmetric and independent of θ^*. (Lemma C.6, below, contains a rigorous proof of this statement.) Thus we can write it $(\mathbb{I} - \eta\Sigma)^t\mathbf{x}_i = \lambda_{i,t}\mathbf{u}_{i,t}$ for $\lambda_{i,t} \in \mathbb{R}$ and $\mathbf{u}_{i,t} \in \mathbb{S}^{p-1}$. By Lemma A.11, then, with probability at least $1 - \beta/(5nT)$ we have

$$\begin{aligned}|\mathbf{x}_i^\dagger(\mathbb{I} - \eta\Sigma)^t\theta^*| &= \lambda_{i,t} \cdot |\mathbf{u}_{i,t}^\dagger\theta^*| \\ &\leq \lambda_{i,t} \cdot \frac{4\ln 5nT/\beta}{\sqrt{p} - 1}.\end{aligned}$$

Conditions (i) and (ii) imply $\lambda_{i,t} = \|(\mathbb{I} - \eta)^t\mathbf{x}_i\| \leq c_0\sqrt{p}$, so we arrive at $|\mathbf{x}_i^\dagger(\mathbb{I} - \eta\Sigma)^t\theta^*| \leq \frac{4c_0\ln 5nT/\beta}{1 - p^{-1/2}}$, which is at most $16c_0\ln 5nT/\beta$ for $p \geq 2$. A union bound over all n, T implies this holds with probability at least $1 - \beta/5$.

Condition (v) We want to bound $|\sum_j (y_j - \mathbf{x}_j^\dagger \theta^*) \cdot \mathbf{x}_i^\dagger A_t \mathbf{x}_j|$ for $A_t = (\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\Sigma^{-1}$. Recall that $\xi_j = y_i - \mathbf{x}_j^\dagger \theta^*$ by definition. For any fixed value of the covariates, the variances sum: we have

$$\sum_j \xi_j \cdot \mathbf{x}_i^\dagger A_t \mathbf{x}_j \sim \mathcal{N}\left(0, \sigma^2 \sum_j (\mathbf{x}_i^\dagger A_t \mathbf{x}_j)^2\right).$$

Developing the square, we have

$$\begin{aligned}\sum_j (\mathbf{x}_i^\dagger A_t \mathbf{x}_j)^2 &= \sum_j \mathbf{x}_i^\dagger A_t \mathbf{x}_j \mathbf{x}_j^\dagger A_t^T \mathbf{x}_i \\ &= n \cdot \mathbf{x}_i^\dagger A_t \left(\frac{1}{n}\sum_j \mathbf{x}_j \mathbf{x}_j^\dagger\right) A_t^T \mathbf{x}_i \\ &= n \cdot \mathbf{x}_i^\dagger A_t \Sigma A_t^T \mathbf{x}_i.\end{aligned}$$

Plugging in the definition of A_t, we cancel and apply Cauchy–Schwarz:

$$\begin{aligned}\sum_j (\mathbf{x}_i^\dagger A_t \mathbf{x}_j)^2 &= n \cdot \mathbf{x}_i^\dagger (\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\Sigma^{-1}\Sigma\Sigma^{-1}(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\mathbf{x}_i \\ &= n \cdot \mathbf{x}_i^\dagger (\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\Sigma^{-1}(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\mathbf{x}_i \\ &\le n \cdot \|(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\Sigma^{-1}(\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\| \cdot \|\mathbf{x}_i\|^2 \\ &\le n \cdot \|\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t)\|^2 \cdot \|\Sigma^{-1}\| \cdot \|\mathbf{x}_i\|^2.\end{aligned}$$

Our previous assumptions imply that $\|\mathbb{I} - (\mathbb{I} - \eta\Sigma)^t\| \le \|\mathbb{I}\| + \|\mathbb{I} - \eta\Sigma\| \le 2$ and $\|\Sigma^{-1}\| \le 2$, so we arrive at

$$\sum_j (\mathbf{x}_i^\dagger A_t \mathbf{x}_j)^2 \le 8c_1^2 np,$$

plugging in $\|\mathbf{x}_i\| \le c_1\sqrt{p}$ from Condition (ii).

Thus, with probability at least $1 - \beta/5$, for all i and t we have

$$\begin{aligned}|\sum_j \xi_j \cdot \mathbf{x}_i^\dagger A_t \mathbf{x}_j| &\le \sqrt{\sigma^2 \sum_j (\mathbf{x}_i^\dagger A_t \mathbf{x}_j)^2} \cdot \sqrt{2\ln 5nT/\beta} \\ &\le \sqrt{\sigma^2 8c_1^2 np} \cdot \sqrt{2\ln 5nT/\beta}.\end{aligned}$$

Plugging in $c_1 \le 3\sqrt{\ln 5n/\beta}$ concludes the proof. □

Theorem C.8 (Restatement of Theorem 2.7, Main Accuracy Claim)**.** *Fix $\theta^* \in \mathbb{R}^p$ with $p \ge 2$ and $\|\theta^*\| \le 1$, let n covariates $\mathbf{x}_i$ be drawn i.i.d. from $\mathcal{N}(0, \mathbb{I})$ and responses $y_i = \mathbf{x}_i^\dagger \theta^* + \xi_i$ for $\xi_i \sim \mathcal{N}(0, \sigma^2)$ for some fixed σ.*

Fix $\rho \ge 0$ and $\beta \in (0, 1)$. Consider running Algorithm 1, i.e., $\mathcal{A}(\mathbf{X}, \mathbf{y}; \gamma, \lambda, \eta, T, 0)$ with step size $\eta = \frac{1}{4}$, initial point $\theta_0 = 0$, and, for some absolute constant c,

$$\begin{aligned}T &= c\log\frac{n\rho}{p}, \quad \lambda^2 = \frac{2T\gamma^2}{\rho n^2}, \text{ and} \\ \gamma &= c\sigma\sqrt{p}\log^3\left(\frac{nT}{\beta}\right).\end{aligned}$$

Recall $\hat{\theta}$ the OLS solution. If $n \ge c\left(p + \sqrt{p}\log^4\rho/\beta\right)$, then with probability at least $1 - \beta$ Algorithm 1 returns a final iterate θ_T such that, for some constant c',

$$\|\hat{\theta} - \theta_T\| \le c' \ln^4(n\rho/\beta p) \cdot \frac{\sigma p}{\sqrt{\rho} n}.$$

Proof. By Lemma 2.6, with probability at least $1 - \beta/4$ we have that data set $(\mathbf{X}, \mathbf{y})$ satisfies the No-Clipping Condition with $c_0 = 12\ln^{1.5}(20nT/\beta)$. This uses the assumption that $n \geq c(p + \ln 1/\beta)$.

Under this condition, Lemma 2.5 says that Algorithm 1 does not clip with probability at least $1 - \beta/4$. This uses the assumptions that

$$\gamma \geq 4c_0^2\sigma\sqrt{p} = 576\ln^3(20nT/\beta) \times \sigma\sqrt{p}$$
$$\frac{\gamma}{\eta\lambda} \geq 64c_0^2 p\sqrt{\ln(2nT/\beta)}.$$

The first is satisfied by construction. To see that the second is satisfied, plug in $\eta = \frac{1}{4}$, $\lambda = \frac{\sqrt{2T}\gamma}{\sqrt{\rho}n}$, and $T = c\log\frac{n\rho}{p}$ to turn this into a lower bound on n: omitting constants, it suffices to take

$$n \gtrsim \sqrt{p}\log^4(n\rho/\beta p).$$

This is satisfied when $n \geq c\sqrt{p}\log^4(\rho/\beta)$ for some sufficiently large constant c.

This implies, via Lemma 2.2, that the total variation distance between the output of Algorithm 1 and the same algorithm without clipping (i.e., $\gamma = +\infty$) is at most $\beta/2$. Thus, if we prove an error bound for Algorithm 1 with $\gamma = \infty$ that holds with probability at least $1 - \beta/2$, then the same guarantee holds for the clipped version of Algorithm 1 with probability at least $1 - \beta$.

Lemma 2.3 gives us the explicit distribution for the algorithm's final iterate θ_T. We take the ℓ_2 norm and apply the triangle inequality:

$$\|\theta_T - \hat{\theta}\|_2 \leq \|(\mathbb{I} - \eta\Sigma)^T\hat{\theta}\|_2 + \|\eta \cdot \mathbf{z}'_T\|_2.$$

Here $\mathbf{z}'_T$ is a noise term: defining $\mathbf{D} = (\mathbb{I} - \eta\Sigma)^2$ as in Lemma 2.3, we have $\mathbf{z}'_T \sim \mathcal{N}(0, \lambda^2\mathbf{A}^{(T)})$ for

$$\mathbf{A} = (\mathbb{I} - \mathbf{D})^{-1}(\mathbb{I} - \mathbf{D}^T).$$

The No-Clipping Condition gives us upper and lower bounds on $\mathbf{D}$: in particular, we have $\|\mathbf{A}\| \leq 16$. Thus, with probability at least $1 - \beta/4$, we have $\|\eta \cdot \mathbf{z}'_T\| \leq \eta\lambda\sqrt{32p\ln 4/\beta}$.

The second term in Equation (C.1) decays exponentially with T. By Cauchy–Schwarz, it is at most $\|(\mathbb{I} - \eta\Sigma)^T\| \cdot \|\hat{\theta}\| \leq (7/8)^T\|\hat{\theta}\|$, applying Condition 2.4 to bound the operator norm. We bound the norm of $\hat{\theta}$ based on its distance to θ^*:

$$\begin{aligned}\|\hat{\theta}\| = \|\hat{\theta} - \theta^* + \theta^*\| &\leq \|\theta^*\| + \|\hat{\theta} - \theta^*\| \\ &= \|\theta^*\| + \|\frac{1}{n}\sum_j \Sigma^{-1}\xi_j\mathbf{x}_j\|.\end{aligned}$$

The first norm is at most 1 by assumption; a rough bound suffices for the second:

$$\|\frac{1}{n}\sum_j \Sigma^{-1}\xi_j\mathbf{x}_j\| \leq \frac{1}{n}\sum_j \|\Sigma^{-1}\| \cdot |y_j - \mathbf{x}_i^\dagger\theta^*| \cdot \|\mathbf{x}_j\| \leq 2c_0^2\sigma\sqrt{p},$$

applying our assumptions.

Plugging these bounds back into Equation (C.1) and plugging in our expressions for η, λ, and γ, we have (omitting constants)

$$\begin{aligned}\|\theta_T - \hat{\theta}\| &\lesssim \sigma\sqrt{p}\log^3\left(\frac{nT}{\beta}\right)\left(\frac{7}{8}\right)^T + \eta\lambda\sqrt{p\log 1/\beta} \\ &\lesssim \sigma\sqrt{p}\log^3\left(\frac{nT}{\beta}\right)\left(\frac{7}{8}\right)^T + \frac{\sigma p\sqrt{T}}{\sqrt{\rho}n} \cdot \log^{3.5}(nT/\beta).\end{aligned}$$

The first term is dominated by the second when $T \geq c\log(n\rho/p)$. Substituting in this value of T finishes the proof. □

C.2. Confidence Intervals (Proof of Theorem 3.1)

C.2.1. Confidence Intervals for Independent Draws

A fundamental task in statistical inference is to produce confidence intervals for the mean given independent samples from a normal distribution with unknown mean and variance. Claim C.9 reproduces this basic fact. After that, we observe that confidence intervals produced in this way are valid for samples from distributions that are close in TV distance.

Claim C.9. *Let $\{\theta_i\}_{i\in[m]}$ be m samples drawn i.i.d. from $\mathcal{N}(\mu,\Sigma)$, for any $\mu\in\mathbb{R}^p$ and $\Sigma\in\mathbb{R}^{p\times p}$. For any $j\in[p]$, let $\bar{\theta}_j = \frac{1}{m}\sum_{i=1}^m (\theta_i)_j$ be the sample mean and $\hat{\sigma}_j^2 = \frac{1}{m-1}\sum_{i=1}^m \left((\theta_i)_j - \bar{\theta}_j\right)^2$ the sample variance. Then $(\bar{\theta})_j \pm t_{\alpha/2,m-1}\cdot \frac{\hat{\sigma}_j}{\sqrt{m}}$, is a $1-\alpha$ confidence interval. Here $t_{\alpha,m-1}$ denotes the α percentile for the student's t distribution with $m-1$ degrees of freedom.*

Claim C.10. *Suppose mechanism $\mathcal{M}:\mathbb{R}^m\to\mathbb{R}^2$, when given samples $z_1,\dots,z_m$ drawn i.i.d. from a distribution $\mathcal{N}(\mu,\sigma^2)$, produces a $1-\alpha$ confidence interval for μ. Let q be a distribution over $\mathbb{R}^m$. If $(z_i',\dots,z_m')\sim q$, then $\mathcal{M}(z_1',\dots,z_m')$ produces a $1-\alpha-\mathrm{TV}(q,\mathcal{N}(\mu,\sigma^2)^{\otimes m})$ confidence interval for μ.*

C.2.2. Analyzing Independent Runs and Checkpoints

Lemma 2.3 tells us that the t-th iterate of DP-GD without clipping has the distribution

$$\theta_t = \hat{\theta} + (\mathbb{I}-\eta\Sigma)^t(\theta_0-\hat{\theta}) + \mathbf{z}_t',$$

where $\mathbf{z}_t'\sim\mathcal{N}(0,\eta^2\lambda^2\mathbf{A}^{(t)})$ for $\mathbf{A}^{(t)} = (\mathbb{I}-\mathbf{D})^{-1}(\mathbb{I}-\mathbf{D}^t)$ and $\mathbf{D} = (\mathbb{I}-\eta\Sigma)^2$. Recall η the step size and $\Sigma = \frac{1}{n}\mathbf{X}^\dagger\mathbf{X}$ the empirical covariance. For appropriate $\mathbf{D}$, as t grows this approaches the distribution $\mathcal{N}(\hat{\theta},\eta^2\lambda^2\mathbf{A}^{(\infty)})$, where $\mathbf{A}^{(\infty)} = (\mathbb{I}-\mathbf{D})^{-1}$.

In this section, we give guarantees for two algorithms that constructing confidence intervals: independent runs and checkpoints. We show that each produces a sequence of m vectors which, as a whole, is close in total variation distance to m independent draws from $\mathcal{N}(\hat{\theta},\eta^2\lambda^2\mathbf{A}^{(\infty)})$. We call this latter the "ideal" case, as it corresponds to Claim C.9. The other two cases are "independent runs" and "checkpoints." Thus, we consider three sets of random variables: $\theta_1^{(1)},\dots,\theta_m^{(1)}$ are drawn i.i.d. from $\mathcal{N}(\hat{\theta},\eta^2\lambda^2\mathbf{A}^{(\infty)})$; $\theta_1^{(2)},\dots,\theta_m^{(2)}$, where $\theta_\ell^{(2)}$ is t-th iterate of an independent run of DP-GD without clipping; and $\theta_1^{(3)},\dots,\theta_m^{(3)}$, where $\theta_\ell^{(2)}$ is the ℓt-th iterate of a single run of DP-GD without clipping. We interpret each of these as a draw from a multivariate Gaussian distribution in mp dimensions: we define the concatenated vector $\theta^{(1)} = \left[\theta_1^{(1)}\middle|\cdots\middle|\theta_m^{(1)}\right]$ and let $\mu^{(1)}\in\mathbb{R}^{mp}$ and $\Sigma^{(1)}\in\mathbb{R}^{mp\times mp}$ be its mean and covariance, respectively. We define the analogous notation for the other two collections. Our calculations will only require us to index into these objects "blockwise," so we abuse notation and define, for $\ell,k\in[m]$,

$$\mu_\ell^{(1)} = \mathbf{E}\left[\theta_\ell^{(1)}\right]\in\mathbb{R}^p \quad\text{and}\quad \Sigma_{\ell,k}^{(1)} = \mathbf{E}\left[\left(\theta_\ell^{(1)}-\mu_\ell^{(1)}\right)\left(\theta_k^{(1)}-\mu_k^{(1)}\right)^\dagger\right]\in\mathbb{R}^{p\times p}.$$

We establish that the vectors from our algorithms are close in total variation to the ideal case by showing that the relevant means and covariances are close. We use the following standard fact.

Claim C.11 (See, e.g., (Diakonikolas et al., 2019))**.** *There exists a constant K such that, for any $\beta\le\frac{1}{2}$, vectors $\mu_1,\mu_2\in\mathbb{R}^d$, and positive definite $\Sigma_1,\Sigma_2\in\mathbb{R}^{d\times d}$, if $\|\mu_1-\mu_2\|_{\Sigma_1}\le\beta$ and $\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}-\mathbb{I}\|_F\le\beta$ then $\mathrm{TV}(\mathcal{N}(\mu_1,\Sigma_1),\mathcal{N}(\mu_2,\Sigma_2))\le K\beta$.*

Ideal Case Each vector is an independent draw from a fixed Gaussian, so for any ℓ and any $k\neq\ell$ we have

$$\mu_\ell^{(1)} = \hat{\theta},\quad \Sigma_{\ell,\ell}^{(1)} = \eta^2\lambda^2\mathbf{A}^{(\infty)},\text{ and }\quad \Sigma_{\ell,k}^{(1)} = 0.$$

Independent Runs Here the vectors are also independent Gaussian random variables, but with different parameters. For any ℓ and any $k\neq\ell$ we have

$$\mu_\ell^{(2)} = \hat{\theta} + (\mathbb{I}-\eta\Sigma)^t(\theta_0-\hat{\theta}),\quad \Sigma_{\ell,\ell}^{(2)} = \eta^2\lambda^2\mathbf{A}^{(t)},\text{ and }\quad \Sigma_{\ell,k}^{(2)} = 0.$$

Checkpoints Here the vectors are no longer independent, so we have additional work. We have

$$\mu_\ell^{(3)} = \hat{\theta} + (\mathbb{I} - \eta\Sigma)^{\ell t}(\theta_0 - \hat{\theta}), \quad \Sigma_{\ell,\ell}^{(3)} = \eta^2\lambda^2\mathbf{A}^{(\ell t)}, \text{ and } \quad \Sigma_{\ell,k}^{(3)} = \mathbf{E}\left[(\mathbf{z}'_{\ell t})(\mathbf{z}'_{kt})^\dagger\right].$$

We now analyze this third term. Recall from Lemma C.4 that these vectors $\mathbf{z}'_{\ell t}$ stand in for sums that depend on all the noise vectors added so far:

$$\mathbf{z}'_t = \eta\sum_{i=1}^{t}(\mathbb{I} - \eta\Sigma)^{i-1}\mathbf{z}^{t-i} = \eta\sum_{i=0}^{t-1}(\mathbb{I} - \eta\Sigma)^{t-i-1}\mathbf{z}^i,$$

where the second equation was re-indexed to simplify the next operation. The random variables $\mathbf{z}_i$ are drawn i.i.d. from $\mathcal{N}(0, \lambda^2\mathbb{I})$. This allows us to understand the covariance in $\Sigma_{\ell,k}^{(3)}$, as some of the noise terms are duplicated and some are not. For any pair of integers $\tau > t$, we have

$$\begin{aligned}
\mathbf{z}'_\tau &= \eta\sum_{i=0}^{\tau-1}(\mathbb{I} - \eta\Sigma)^{\tau-i-1}\mathbf{z}^i \\
&= \eta\sum_{i=0}^{t-1}(\mathbb{I} - \eta\Sigma)^{\tau-i-1}\mathbf{z}^i + \eta\sum_{i=t}^{\tau-1}(\mathbb{I} - \eta\Sigma)^{\tau-i-1}\mathbf{z}^i \\
&= (\mathbb{I} - \eta\Sigma)^{\tau-t}\cdot\eta\sum_{i=0}^{t-1}(\mathbb{I} - \eta\Sigma)^{t-i-1}\mathbf{z}^i + \eta\sum_{i=t}^{\tau-1}(\mathbb{I} - \eta\Sigma)^{\tau-i-1}\mathbf{z}^i \\
&= (\mathbb{I} - \eta\Sigma)^{\tau-t}\mathbf{z}'_t + \eta\sum_{i=t}^{\tau-1}(\mathbb{I} - \eta\Sigma)^{\tau-i-1}\mathbf{z}^i.
\end{aligned}$$

The second sum is independent of $\mathbf{z}'_t$, since it involves only later terms in the algorithm. Applying this with $\tau \leftarrow \ell t$ and $t \leftarrow kt$ (assuming without loss of generality that $\ell > k$), we have

$$\begin{aligned}
\Sigma_{\ell,k}^{(3)} = \mathbf{E}\left[(\mathbf{z}'_{\ell t})(\mathbf{z}'_{kt})^\dagger\right] &= (\mathbb{I} - \eta\Sigma)^{(\ell-k)t}\,\mathbf{E}\left[(\mathbf{z}'_{kt})(\mathbf{z}'_{kt})^\dagger\right] \\
&= \eta^2\lambda^2(\mathbb{I} - \eta\Sigma)^{(\ell-k)t}\mathbf{A}^{(kt)}.
\end{aligned}$$

Setup We derive some facts which will be useful for both of our case-specific analyses. Since the data satisfies the No-Clipping Condition, we have $\frac{1}{2}\mathbb{I} \preceq \Sigma \preceq 2\mathbb{I}$ and $\|\mathbb{I} - \eta\Sigma\| \leq \frac{7}{8}$. These in turn establish operator norm bounds for $\mathbf{D}, \mathbf{A}^{(\infty)}, \mathbb{I} - \mathbf{D}$ (and their inverses) of at most c for some absolute constant c, which yields a Frobenius norm bound of $c\sqrt{m}$. We also have $\|\mathbf{D}^t\|_F \leq \sqrt{m}\cdot\left(\frac{7}{8}\right)^t$.

We pass from the rescaled norms needed in Claim C.11 to Frobenius and ℓ_2 norms:

$$\begin{aligned}
\|(\Sigma^{(1)})^{-1/2}(\Sigma^{(2)})(\Sigma^{(1)})^{-1/2} - \mathbb{I}\|_F &\leq \frac{1}{\lambda_{\min}(\Sigma^{(1)})}\cdot\|\Sigma^{(2)} - \Sigma^{(1)}\|_F, \quad \text{and} \\
\|\mu^{(1)} - \mu^{(2)}\|_{\Sigma^{(1)}} &\leq \frac{1}{\sqrt{\lambda_{\min}(\Sigma^{(1)})}}\cdot\|\mu^{(1)} - \mu^{(2)}\|_2.
\end{aligned}$$

Since the minimum eigenvalue of a block-diagonal matrix is the minimum eigenvalue of the blocks and, in the ideal case, the blocks are diagonal, we have $\lambda_{\min}(\Sigma^{(1)}) = \lambda_{\min}(\eta^2\lambda^2\mathbf{A}^{(\infty)}) \geq \eta^2\lambda^2 c$ for some constant c.

Analyze Independent Runs With the above tools in hand, it suffices to analyze $\|\Sigma^{(2)} - \Sigma^{(1)}\|_F$. Observe that $\Sigma^{(2)}$ is equal to $\Sigma^{(1)}$ plus a block-diagonal matrix where each block is $\eta^2\lambda^2$ times $\mathbf{A}^{(\infty)} - \mathbf{A}^{(t)} = \mathbf{A}^{(\infty)}\mathbf{D}^t$. So we have

$$\|\Sigma^{(2)} - \Sigma^{(1)}\|_F \leq \eta^2\lambda^2 m\cdot\|\mathbf{A}^{(\infty)}\mathbf{D}^t\|_F \leq c\eta^2\lambda^2 m^2\left(\frac{7}{8}\right)^t$$

for some constant c. Cancelling the $\eta^2\lambda^2$ from $\lambda_{\min}$ in Equation (C.2.2), we have $\|(\Sigma^{(1)})^{-1/2}(\Sigma^{(2)})(\Sigma^{(1)})^{-1/2} - \mathbb{I}\|_F \leq \beta$ when $t \geq c\log m/\beta$ for some constant c.

The mean is similar: the m blocks of the vector are identical, so we have

$$\|\mu^{(1)} - \mu^{(2)}\|_2 \le \sqrt{m}\|(\mathbb{I} - \eta\Sigma)^t(\theta_0 - \hat{\theta})\|_2.$$

The No-Clipping Condition implies that $\|\theta_0 - \hat{\theta}\|_2 \le 2c_0^2\sigma\sqrt{p}$ (as in the proof of Theorem 2.7). Thus, combining with Equation (C.2.2), we have $\|\mu^{(1)} - \mu^{(2)}\|_{\Sigma^{(1)}} \le \beta$ when $t \ge c\log\frac{\sigma mp}{\eta\lambda\beta}$ for some constant c.

Analyze Checkpoints We apply similar arguments. The upper bound on $\|\mu^{(1)} - \mu^{(3)}\|_{\Sigma^{(1)}}$ is identical to the previous case except that we apply $\|\mathbf{D}^{\ell t}\|_F \le \|\mathbf{D}^t\|_F$ for all $\ell \in [m]$. The covariance analysis is similar, but $\Sigma^{(1)} - \Sigma^{(3)}$ is no longer block-diagonal (as the checkpoints are not independent). So the main additional work is to control the effect of these off-diagonal blocks.

We expand over the m^2 blocks, moving temporarily to the squared Frobenius norm:

$$\begin{aligned}\|\Sigma^{(3)} - \Sigma^{(1)}\|_F^2 &= \eta^4\lambda^4\left(\sum_{\ell=1}^m \|\mathbf{A}^{(\infty)} - \mathbf{A}^{(\ell t)}\|_F^2 + 2\sum_{\ell>k}\|0 - (\mathbb{I} - \eta\Sigma)^{(\ell-k)t}\mathbf{A}^{(kt)}\|_F^2\right)\\ &\le \eta^4\lambda^4\left(\sum_{\ell=1}^m \|\mathbf{A}^{(\infty)}\mathbf{D}^t\|_F^2 + 2\sum_{\ell>k}\|(\mathbb{I} - \eta\Sigma)^{(\ell-k)t}\|_2^2 \cdot \|\mathbf{A}^{(kt)}\|_F^2\right).\end{aligned}$$

As before, we combine with Equation (C.2.2) to obtain an upper bound on the (unsquared) norm of $\mathrm{poly}\,(m)\cdot\left(\frac{7}{8}\right)^t$, which is less than β when $t \ge c\log m/\beta$ for some constant c.