Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Differentially Private Lasso: An ISTA Framework with Finite-Iteration Guarantees

Jiahui Zhang
Nanyang Technological University
Chi Seng Pun
cspun@ntu.edu.sg

Nanyang Technological University

Research Article

Keywords: High-dimensional Statistics, Sparse Linear Regression, Differential Privacy, Proximal Gradient, Iterative Shrinkage-Thresholding Algorithm (ISTA), Finite-Iteration Guarantees

Posted Date: March 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9153407/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Statistics and Computing on October 14th, 2026. See the published version at https://doi.org/10.1007/s11222-026-11001-9.

Differentially Private Lasso: An ISTA Framework with Finite-Iteration Guarantees

Jiahui Zhang[1] and Chi Seng Pun[1*]

[1]School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore.

*Corresponding author(s). E-mail(s): cspun@ntu.edu.sg;
Contributing authors: jiahui010@e.ntu.edu.sg;

Abstract

Differential privacy (DP) offers a principled way to protect individual records, but in high-dimensional sparse regression, it introduces a delicate accuracy-privacy tradeoff. In this paper, we develop an ISTA-based framework for DP estimation in the high-dimensional sparse linear model, instantiated for the Lasso objective. Our main contribution is a set of *finite-iteration*, high-probability $\boldsymbol{\ell_2}$ guarantees for the returned iterates. Across the considered DP mechanisms, the bounds admit an interpretable form: a nonprivate baseline term, a privacy-induced term determined by the effective noise level of the DP mechanism and its accounting, and an optimization residual that vanishes as the iteration budget increases. To enable stable implementations and principled Gaussian calibration, our algorithms incorporate clipping and an $\boldsymbol{\ell_2}$ projection step. Simulation studies and real-data experiments under matched privacy budgets support the theoretical predictions and demonstrate competitive accuracy in high-dimensional regimes.

Keywords: High-dimensional Statistics; Sparse Linear Regression; Differential Privacy; Proximal Gradient; Iterative Shrinkage-Thresholding Algorithm (ISTA); Finite-Iteration Guarantees

1 Introduction

High-dimensional sparse linear regression is a workhorse in modern statistics and machine learning, providing interpretable linear models and strong predictive performance through sparsity-inducing regularization when the number of variables (p) is much larger than the number of observations (n), i.e., $p \gg n$. In particular, the least

absolute shrinkage and selection operator (Lasso) has been widely adopted for variable selection and prediction, and its statistical guarantees, i.e., prediction error bounds and support recovery, are well understood under restricted eigenvalue or restricted strong convexity conditions (Tibshirani, 1996; Bickel et al., 2009; Wainwright, 2009; Bühlmann and van de Geer, 2011; Negahban et al., 2012; Vershynin, 2018). For its computational aspect, proximal first-order methods such as Iterative Shrinkage-Thresholding Algorithm (ISTA) and Fast ISTA (FISTA) offer scalable solvers with robust convergence behavior for ℓ_1-regularized objectives (Daubechies et al., 2004; Beck and Teboulle, 2009), among other Lasso algorithms such as least angle regression (Efron et al., 2004) and coordinate descent (Wu and Lange, 2008).

Recently, privacy concerns and regulatory requirements increasingly impose non-negotiable constraints on data collection and model training. Differential Privacy (DP) provides a mathematically robust and composable notion of privacy protection (Dwork et al., 2006; Dwork and Roth, 2014). Importantly in realistic systems, the operative privacy model is often dictated by deployment: in the *centralized DP* model, a trusted curator can access raw data and enforce privacy by perturbing training signals or outputs (with explicit composition accounting across iterations); in the more stringent *local DP* (LDP) model, each data holder randomizes their record *before* data transmission, enabling privacy without a trusted server; and in many supervised applications, labels are substantially more sensitive than features, motivating *label privacy* variants that protect only the dependent variable or output (y) while exploiting the structure of the independent variables or inputs (X).

1.1 Deployment Paradigms and Related Works

In practice with DP, the high-dimensional learning is naturally shaped by the deployment paradigms: trusted curator versus client-side randomization; label-only versus record-level protection. To facilitate comparisons and to anchor our contributions within a broader landscape, Table 1 summarizes the related works and typical tasks under three paradigms that are most relevant to our setting. Rather than a comprehensive survey, Table 1 provides a concise taxonomy that motivates why a single algorithmic template (proximal first-order methods with thresholding/projection) behaves differently across regimes, and why one needs theory that separates statistical error, privacy-induced perturbations, and finite-iteration residuals.

Centralized DP methods provide strong utility when a trusted curator is available and the learning pipeline can rely on server-side perturbations. Gradient perturbation such as DP-SGD (Abadi et al., 2016) offers a practical recipe for large models, while private convex empirical risk minimization (ERM) covers linear and logistic regression under convexity assumptions (Abadi et al., 2016; Kifer et al., 2012; Chaudhuri et al., 2011; Bassily et al., 2014). However, for high-dimensional sparse regression, the effect of repeated access under composition must be tracked together with optimization error; our Algorithm 4 and Theorem 15 make this interaction explicit through iteration-dependent privacy terms.

Interactive LDP eliminates the trusted curator via multi-round client-side randomization (Duchi et al., 2018; Kasiviswanathan et al., 2011). While adaptivity can improve statistical efficiency in some regimes, it comes with communication overhead

Table 1: Related works and typical tasks under different DP deployment paradigms.

Paradigm	Reference	Typical Task / Model	Perturbation
Centralized DP	(Abadi et al., 2016)	Deep networks, generalized linear models (GLMs)	Server-side perturbation: gradient queries (iterative)
	(Cai et al., 2020, 2021)	GLMs: algorithms, optimal rates (lower bounds)	Server-side perturbation: output/gradient
	(Sheffet, 2017)	Linear regression and confidence intervals	Server-side perturbation: sufficient statistics/output
Interactive LDP	(Kasiviswanathan et al., 2011)	Locally private learning (interactive protocols)	Client-side local randomizer (multi-round; interactive)
	(Joseph et al., 2019)	Interactivity in LDP: separations and protocols	Client-side local randomizer (multi-round; adaptive)
Non-interactive LDP	(Apple Differential Privacy Team, 2017)	Large-scale client telemetry (iOS/macOS analytics)	Client-side local randomizer (one-shot; non-interactive)
	(Wang et al., 2021)	GLMs with public unlabeled data	Client-side local randomizer (one-shot; non-interactive)

and cumulative noise across rounds, which can be prohibitive in high-dimensional regression unless protocols are carefully designed. Our focus in this paper is on *non-interactive* mechanisms and their post-processing, which aligns more directly with one-shot data collection pipelines.

Non-interactive LDP is attractive from an engineering standpoint: one-shot client-side perturbation followed by arbitrary post-processing that does not consume additional privacy budget (Dwork and Roth, 2014; Duchi et al., 2018). Recent studies demonstrate that under structural assumptions, non-interactive schemes can attain meaningful guarantees for generalized linear and sparse linear models; see Wang et al. (2021); Wang and Xu (2021). In this paper we build on this philosophy and develop one-shot mechanisms tailored to sparse regression, analyzed through proximal first-order optimization with explicit high-probability ℓ_2 guarantees.

A closely related practical consideration is *label privacy*. In supervised learning, the feature matrix X may be derived from public or weakly sensitive sources, whereas labels y (e.g., diagnoses, outcomes, feedback) often constitute the primary privacy bottleneck. Protecting only the labels can therefore preserve much of the design structure while enforcing meaningful privacy on the supervision signal. This idea has been studied in several settings, e.g., label-private learning and teacher-student frameworks; see Papernot et al. (2017, 2018); Ghazi et al. (2021); Esfandiari et al. (2022). In the context of sparse regression, Wang and Xu (2021) provide a non-interactive label-LDP variant based on threshold updates. In contrast, our Algorithm 2 provides a complementary one-shot label mechanism coupled with ISTA post-processing, and Theorem 10 quantifies its high-probability ℓ_2 error.

To highlight the relevance of our work, we briefly indicate where the three regimes arise, emphasizing the deployment constraint rather than industry-specific narratives:

- when data must be privatized at collection and subsequent computation is centralized post-processing, e.g., telemetry and edge reporting, *non-interactive LDP* mechanisms are operationally appealing because post-processing is privacy-free;
- when only outcomes are sensitive while features can be exploited without perturbation, *label-private* mechanisms offer a favorable privacy-utility tradeoff by preserving the geometry of X.
- when a trusted curator can perform iterative training with auditable privacy accounting, *centralized DP* mechanisms are natural but must balance optimization progress against the privacy cost of repeated access.

1.2 Research Gap and Contributions

Despite substantial progress, the existing literature leaves two important gaps unaddressed. First, results on private sparse regression are fragmented across deployment paradigms and privacy models, with most analyses conducted in isolation within a single setting (e.g., centralized DP or non-interactive LDP). This fragmentation makes it difficult to compare algorithms under matched privacy budgets and obscures how privacy accounting interacts with optimization dynamics and sparse recovery geometry across deployments. Second, most theoretical guarantees are asymptotic or optimization-agnostic, whereas practical solvers operate under explicit computational budgets and return finite-iteration iterates subject to stability mechanisms such as clipping or projection. These algorithmic constraints materially affect both privacy calibration and numerical behavior. Consequently, there is a lack of unified theory that cleanly disentangles statistical error, privacy-induced perturbations, and finite-iteration residuals in high-dimensional sparse learning.

We develop a unified proximal framework for private high-dimensional sparse regression that accommodates three privacy regimes within a single analytical template. Specifically, we study (i) one-shot label-side local differential privacy with ISTA post-processing (Algorithm 2), (ii) one-shot local differential privacy applied jointly to (X, y) with ISTA post-processing (Algorithm 3), and (iii) centralized differential privacy with iterative gradient perturbation, thresholding, and ℓ_2 projection (Algorithm 4). Treating these regimes within a common framework allows us to isolate the role of privacy model and access pattern while holding the underlying optimization structure fixed. For each regime, we establish non-asymptotic high-probability ℓ_2 error bounds for parameter estimation (Theorems 10, 12, and 15). The bounds admit an explicit decomposition into three components: a statistical baseline term governed by sparsity and restricted geometry, a privacy-induced perturbation term determined by the effective noise level and privacy accounting of the underlying mechanism, and an optimization residual reflecting finite-iteration effects. In particular, the analysis yields a $k^{-1/2}$ residual for one-shot ISTA-based procedures and geometric contraction for the centralized DP iteration, thereby making precise how computational budget enters the statistical–privacy tradeoff.

A distinguishing feature of our theory is that it is inherently finite-iteration and stability-aware. Rather than analyzing exact minimizers of privatized objectives, we characterize the error of iterates returned by proximal first-order methods under explicit stability constraints, such as clipping and ℓ_2 projection, which are essential for

well-defined sensitivity and valid privacy calibration. This perspective enables a clean separation between statistical error, privacy perturbations, and optimization error, and supports principled comparisons across privacy regimes under matched privacy budgets. We complement the theoretical results with empirical studies that corroborate the predicted regime-dependent tradeoffs.

The remainder of this paper is organized as follows. Section 2 introduces the notation, the sparse linear regression setup, and privacy models with basic optimization facts. Section 3 presents the algorithms and theory, including the one-shot mechanisms, the iterative centralized mechanism, and finite-sample error bounds, together with the main proofs and the auxiliary lemmas needed in the analysis. Section 4 reports simulation tests under controlled high-dimensional settings, and Section 5 provides empirical studies on real datasets. Section 6 concludes with limitations and future work.

2 Preliminaries

In this section we introduce the notations and formalize the privacy framework. We recall the basic definitions of differential privacy (DP) and state several useful preliminary results. Then we set up the sparse linear model, specify the regularity conditions, and collect auxiliary results that support the upper bounds to be derived in the subsequent sections.

2.1 Differential Privacy: Definitions and Preliminary Results

We first start with the definitions of DP and local DP (LDP).

Definition 1 (Differential Privacy (DP)) Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be a randomized mechanism and let $D, D' \in \mathcal{X}^n$ be adjacent datasets. We say $\mathcal{M}$ is (ε, δ)-differentially private (DP) if for all measurable $S \subseteq \mathcal{Y}$,

$$\Pr[\mathcal{M}(D) \in S] \;\le\; e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta.$$

DP enjoys post-processing invariance and composition theorems (Dwork et al., 2006; Dwork and Roth, 2014), which will be detailed.

Definition 2 (Local Differential Privacy (LDP)) A (local) randomized encoder $Q : \mathcal{X} \to \mathcal{Z}$ is (ε, δ)-locally differentially private (LDP) if for all $x, x' \in \mathcal{X}$ and measurable $S \subseteq \mathcal{Z}$,

$$\Pr[Q(x) \in S] \;\le\; e^{\varepsilon} \Pr[Q(x') \in S] + \delta.$$

In the LDP model, each user privatizes her record via Q before transmission; the analyst operates only on privatized reports. Randomized response can be viewed as a canonical LDP mechanism (Warner, 1965). Learning-theoretic foundations are given by Kasiviswanathan et al. (2011); statistical risks under LDP are typically larger than under central DP (Duchi et al., 2018).

Next, we introduce the concept of sensitivity and the related results, especially with the Gaussian mechanism; see Dwork and Roth (2014).

Definition 3 (ℓ_p-sensitivity) For any algorithm f mapping a dataset $X \in \mathcal{X}^n$ to $\mathbb{R}^d$, the ℓ_p-sensitivity of f is

$$\Delta_p(f) \;=\; \sup_{X,\,X' \text{ adjacent } (X \sim X')} \big\| f(X) - f(X') \big\|_p .$$

Proposition 1 (Gaussian mechanism (classical calibration)) *Let $f : \mathcal{X}^n \to \mathbb{R}^m$ and define its ℓ_2-sensitivity*

$$\Delta_2(f) \;=\; \sup_{X \sim X'} \big\| f(X) - f(X') \big\|_2 .$$

The Gaussian mechanism releases $\mathcal{M}(X) = f(X) + Z$ with $Z \sim \mathcal{N}\big(0, \sigma^2 I_m\big)$. A sufficient condition for $\mathcal{M}$ to be (ε, δ)-differentially private is

$$\sigma \;\geq\; \frac{\Delta_2(f)\,\sqrt{2\ln\big(1.25/\delta\big)}}{\varepsilon}, \tag{1}$$

A tighter calibration for the same (ε, δ) is given by the analytical Gaussian mechanism (Balle and Wang, 2018).

Now, we present the key preliminary results about the properties of DP.

Theorem 2 (Post-processing) *Let $\mathcal{M} : \mathcal{X}^n \to \mathcal{Y}$ be an (ε, δ)-differentially private mechanism. For any (possibly randomized) mapping $\mathcal{A} : \mathcal{Y} \to \mathcal{Z}$ that has no access to the dataset other than through $\mathcal{M}(X)$, the composed mechanism $\mathcal{A} \circ \mathcal{M}$ is also (ε, δ)-differentially private. Equivalently, for all adjacent $X \sim X'$ and measurable $S \subseteq \mathcal{Z}$,*

$$\Pr\big[\mathcal{A}(\mathcal{M}(X)) \in S\big] \;\leq\; e^{\varepsilon} \Pr\big[\mathcal{A}(\mathcal{M}(X')) \in S\big] + \delta.$$

This invariance holds even when $\mathcal{A}$ is chosen adaptively using only the output $\mathcal{M}(X)$ and its own independent randomness (Dwork et al., 2006; Dwork and Roth, 2014).

Theorem 3 (Advanced composition (Dwork and Roth, 2014, Thm. 3.20)) *Let $\mathcal{M}_1, \ldots, \mathcal{M}_k$ be (possibly adaptively chosen) mechanisms, each satisfying (ε, δ)-differential privacy on the same database. For any $\delta' \in (0, 1)$, the k-fold composition*

$$\mathcal{M}(X) = \big(\mathcal{M}_1(X), \ldots, \mathcal{M}_k(X)\big)$$

is $(\varepsilon_{\text{tot}}, \delta_{\text{tot}})$-differentially private with

$$\varepsilon_{\text{tot}} \;=\; \sqrt{2k\ln(1/\delta')}\,\varepsilon \;+\; k\,\varepsilon\,(e^{\varepsilon} - 1), \qquad \delta_{\text{tot}} \;=\; k\,\delta + \delta'.$$

In particular, for $\varepsilon \leq 1$, using $e^{\varepsilon} - 1 \leq 1.1\,\varepsilon$ gives $\varepsilon_{\text{tot}} \leq \sqrt{2k\ln(1/\delta')}\,\varepsilon + 1.1\,k\,\varepsilon^2$.

Corollary 4 (Per-step tuning under advanced composition (Dwork and Roth, 2014, Cor. 3.21)) *Fix $k \in \mathbb{N}$ and target (ε', δ') with $0 < \varepsilon' < 1$ and $\delta' > 0$. If each mechanism is (ε, δ)-differentially private with*

$$\varepsilon \;=\; \frac{\varepsilon'}{2\sqrt{2k\ln(1/\delta')}}\,, \tag{2}$$

then the leading term in Theorem 3 satisfies $\sqrt{2k\ln(1/\delta')}\,\varepsilon \leq \varepsilon'/2$, and hence

$$\varepsilon_{\text{tot}} \leq \varepsilon'/2 + 1.1\,k\,\varepsilon^2 \qquad \text{and} \qquad \delta_{\text{tot}} = k\delta + \delta'.$$

In particular, for sufficiently small ε (e.g., when $1.1\,k\,\varepsilon^2 \leq \varepsilon'/2$), the overall privacy loss obeys $\varepsilon_{\text{tot}} \leq \varepsilon'$.

2.2 Sparse Linear Regression and Lasso

We consider the high-dimensional ($p > n$) linear model

$$y \;=\; X\beta^* + w,$$

with design $X \in \mathbb{R}^{n\times p}$, response $y \in \mathbb{R}^n$, sparse target $\beta^* \in \mathbb{R}^p$, and noise $w \in \mathbb{R}^n$. Given $\lambda > 0$, the least absolute shrinkage and selection operator (Lasso) estimator (Tibshirani, 1996) is given by

$$\hat{\beta} \in \arg\min_{\beta\in\mathbb{R}^p} \Big\{ \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \Big\}. \tag{3}$$

For its statistical estimation accuracy, we focus on the ℓ_2 parameter error $\|\hat{\beta}-\beta^*\|_2$. In the high-dimensional regime ($p \gg n$), the least-squares loss is typically not globally strongly convex. To obtain meaningful estimation guarantees and to control optimization on sparse directions, we impose standard *restricted* curvature assumptions on the design (or equivalently on the loss restricted to low-dimensional sets). Classical theory under restricted-eigenvalue/compatibility conditions provides sharp nonasymptotic bounds for prediction and estimation errors and support recovery (Bickel et al., 2009; Wainwright, 2009; Bühlmann and van de Geer, 2011).

Definition 4 (Restricted Eigenvalue (RE) condition) Let $X \in \mathbb{R}^{n\times p}$ and $S \subset [p]$ with $|S| = s$. Define the cone $\mathcal{C}(S;\gamma) = \{v \in \mathbb{R}^p :\ \|v_{S^c}\|_1 \le \gamma\,\|v_S\|_1\}$ for $\gamma \ge 1$. We say X satisfies RE(s,γ) with constant $\phi_0 > 0$ if

$$\frac{1}{n}\|Xv\|_2^2 \;\ge\; \phi_0\,\|v\|_2^2 \quad \text{for all } v \in \mathcal{C}(S;\gamma) \text{ and all } S \subset [p],\ |S| = s.$$

This condition is standard in high-dimensional Lasso analysis (Bickel et al., 2009; Wainwright, 2009).

Definition 5 (Restricted Strong Convexity / Smoothness (RSC/RSS)) For a differentiable loss $L : \mathbb{R}^p \to \mathbb{R}$ with gradient ∇L and a set $\mathcal{A} \subseteq \mathbb{R}^p$, L satisfies RSC on $\mathcal{A}$ with parameters (α,τ) if for all $\beta \in \mathbb{R}^p$ and all $\Delta \in \mathcal{A}$,

$$L(\beta+\Delta) - L(\beta) - \langle \nabla L(\beta), \Delta\rangle \;\ge\; \frac{\alpha}{2}\|\Delta\|_2^2 - \tau\|\Delta\|_1^2.$$

It satisfies RSS with parameter $\ell > 0$ if for all $\beta \in \mathbb{R}^p$ and all $\Delta \in \mathcal{A}$,

$$L(\beta+\Delta) - L(\beta) - \langle \nabla L(\beta), \Delta\rangle \;\le\; \frac{\ell}{2}\|\Delta\|_2^2.$$

These notions are formalized for M-estimators in (Negahban et al., 2012). For least squares $L(\beta) = \frac{1}{2n}\|X\beta - y\|_2^2$, RSC/RSS on $\{\Delta : \|\Delta\|_0 \le s\}$ reduce to

$$\mu_s\|\Delta\|_2^2 \;\le\; \frac{1}{n}\|X\Delta\|_2^2 \;\le\; L_s\|\Delta\|_2^2,$$

used widely in sparse regression (Bickel et al., 2009; Wainwright, 2009).

We will repeatedly need a high-probability, dimension-dependent control that holds uniformly over coordinates. Such bounds are used to justify the tuning choices that appear in our main theorems and to keep track of how the dependence on p enters through logarithmic factors. The following lemma and corollary record a standard maximal inequality and a convenient specialization that we will invoke later.

Lemma 5 (Maximal inequality for sub-Gaussian variables (Vershynin, 2018, Sec. 2.1)) *Let $Z_1, \dots, Z_m$ be random variables (not necessarily independent) such that each Z_i is τ-sub-Gaussian, i.e.,*

$$\mathbb{E}\exp(\lambda Z_i) \;\le\; \exp\Big(\frac{\lambda^2\tau^2}{2}\Big) \quad \textit{for all } \lambda \in \mathbb{R}.$$

Then, for any $t > 0$,

$$\Pr\Big(\max_{1\le i\le m} Z_i \ge t\Big) \le m\,\exp\Big(-\frac{t^2}{2\tau^2}\Big), \qquad \Pr\Big(\max_{1\le i\le m} |Z_i| \ge t\Big) \le 2m\,\exp\Big(-\frac{t^2}{2\tau^2}\Big). \tag{4}$$

Corollary 6 (Coordinate-wise bound for $(1/n)X^\top u$ (Lemma 5 with $m = p$ and $\tau^2 = \sigma^2/n$)) *If $\|X_j\|_2 \le \sqrt{n}$ for all columns X_j and $u \sim \mathcal{N}(0, \sigma^2 I_n)$, then each coordinate $(1/n)X_j^\top u$ is mean-zero Gaussian with variance $\le \sigma^2/n$, and for any $c > 0$,*

$$\Pr\left(\left\|\frac{1}{n}X^\top u\right\|_\infty \le \sigma\sqrt{\frac{2(c+1)\log p}{n}}\right) \;\ge\; 1 - \frac{2}{p^c}.$$

2.2.1 The ISTA framework

We solve the Lasso optimization problem (3) with the Iterative Shrinkage–Thresholding Algorithm (ISTA), a proximal first-order method: each iteration consists of one gradient step for the squared loss followed by coordinate-wise soft-thresholding. For a fixed stepsize $1/L$ with $L \ge \sigma_{\max}(X^\top X/n)$, the ISTA (Daubechies et al., 2004) decreases the objective at rate $\mathcal{O}(1/k)$ while the fast ISTA (FISTA) (Beck and Teboulle, 2009) accelerates to $\mathcal{O}(1/k^2)$. Moreover, ISTA is amenable to privacy mechanisms that perturb the gradient channel while preserving the proximal geometry.

Algorithm 1 ISTA for Lasso (non-private baseline)

Input: Design $X \in \mathbb{R}^{n\times p}$, response $y \in \mathbb{R}^n$, regularization $\lambda > 0$, iterations K, stepsize $1/L$ with $L = \sigma_{\max}(X^\top X/n)$, optional projection radius C.
Output: $\beta^{(K)}$.
Initialize $\beta^{(0)} = 0$
for $k = 0, \ldots, K-1$ **do**
  `// Squared-loss gradient`
  $g \leftarrow \frac{1}{n} X^\top (X\beta^{(k)} - y)$
  `// Gradient step`
  $\beta^{(k+1)}_{\text{temp}} \leftarrow \beta^{(k)} - \frac{1}{L} g$
  `// Soft-thresholding (coordinate-wise shrinkage by` λ/L`)`
  $\beta^{(k+1)} \leftarrow S\left(\beta^{(k+1)}_{\text{temp}}, \frac{\lambda}{L}\right)$
  `// (Optional)` ℓ_2 `projection for robustness`
return $\beta^{(K)}$

Operator: $S(v, \tau) = \text{sign}(v) \odot \max\{|v| - \tau, 0\}$ (coordinate-wise soft-threshold)

3 Algorithms and Theories

This section presents three mechanisms for differentially private high-dimensional sparse linear regression under a common ISTA backbone. These mechanisms differ only in where privacy perturbations are injected: one-shot noise on labels (Algorithm 2), one-shot noise on both features and labels (Algorithm 3), and per-iteration gradient perturbation in the centralized model (Algorithm 4). All schemes use explicit clipping and an ℓ_2 projection step to control sensitivity and stabilize iterates under Gaussian perturbations.

Our main theoretical results give finite-sample, high-probability ℓ_2 error bounds that remain meaningful in high-dimensional regimes. The bounds are stated in comparable forms and clarify how privacy, dimension, sparsity, and the iteration budget jointly affect accuracy (Theorems 10, 12, 15).

3.1 Label Locally Differentially Private ISTA

This subsection studies a one-shot *label-side* locally differentially private mechanism for sparse linear regression. On the client side, each response is clipped to a fixed bound to control sensitivity and then perturbed once by independent Gaussian noise calibrated to the target (ϵ, δ) level. The server receives the privatized labels while keeping the design matrix X unchanged, and runs an ISTA routine to obtain the final estimator. Since privacy is expended only at the release stage, all subsequent iterations are pure post-processing and incur no additional privacy loss. On the no-clipping event, the resulting model is equivalent to linear regression with an inflated

Gaussian noise level, which enables a direct high-probability ℓ_2 analysis via standard high-dimensional arguments after adjusting the effective noise (Theorem 10).

We focus on label-side LDP because it matches common threat models while retaining favorable statistical structure in high dimensions. In many supervised settings, the response y is the dominant sensitive attribute, whereas the feature design is public or substantially less sensitive; protecting labels locally can therefore satisfy the primary privacy constraint without perturbing X. By concentrating randomness on the supervision signal, the method preserves the geometry of the design matrix that drives sparse recovery when $p \gg n$, and avoids both feature corruption and per-iteration noise accumulation. Moreover, the mechanism integrates seamlessly with standard Lasso/ISTA pipelines: it requires only a single client-side perturbation step and dispenses with iterative privacy accounting.

Algorithm 2 Label Locally Differentially Private ISTA (Noise Added Once)

Input: Design matrix $X \in \mathbb{R}^{n\times p}$, private labels $y \in \mathbb{R}^n$ (model $y = X\beta^* + w$, $w \sim \mathcal{N}(0, \sigma_w^2 I_n)$), privacy parameters (ϵ, δ), label clipping bound B, regularization parameter λ, number of iterations K, learning rate $L \geq \sigma_{\max}(X^\top X/n)$, projection radius C.

Output: An estimate of $\beta^{(K)}$.

Client side (one-shot label release):

Clip labels: $y_i^{\text{clip}} = \text{sign}(y_i)\min\{|y_i|, B\}, \quad i = 1, \ldots, n$

Set one-shot noise scale:

$$\sigma_{\text{dp}} = \frac{2B}{\epsilon}\sqrt{2\ln\frac{1.25}{\delta}}.$$

Add Gaussian noise to the clipped labels: $\tilde{y}_i = y_i^{\text{clip}} + z_i, \quad z_i \sim \mathcal{N}(0, \sigma_{\text{dp}}^2), \quad i = 1, \ldots, n$

Server side (post-processing; ISTA on $(X, \tilde{y})$):

Initialize $\beta^{(0)} = 0$

for $k = 0, \ldots, K-1$ **do**

- Compute the gradient: $\nabla L(\beta^{(k)}) = \frac{1}{n}X^\top(X\beta^{(k)} - \tilde{y})$
- Gradient descent step: $\beta_{\text{temp}}^{(k+1)} = \beta^{(k)} - \frac{1}{L}\nabla L(\beta^{(k)})$
- Soft-thresholding: $\beta_{\text{thr}}^{(k+1)} = S\Big(\beta_{\text{temp}}^{(k+1)}, \frac{\lambda}{L}\Big)$
- Projection to ℓ_2-ball: $\beta^{(k+1)} = \Pi_{\mathbb{B}_2(C)}\big(\beta_{\text{thr}}^{(k+1)}\big)$

return $\beta^{(K)}$

3.1.1 Main Upper Bound

We now state the main guarantee for the label-side LDP scheme. Under the standing high-dimensional conditions from Section 2, one-shot privatization of the responses yields a finite-sample, high-probability ℓ_2 error bound for the returned ISTA iterates. The bound separates a baseline term of the usual high-dimensional form, a privacy contribution entering through the effective noise level of the one-shot Gaussian

mechanism, and an optimization residual that vanishes with the iteration budget; see Theorem 10. We first introduce the supporting lemmas needed for the proof and invoke them directly.

The first lemma is a classical Lasso optimality argument: under an appropriate choice of regularization level, the estimation error obeys a cone-type relation and a basic inequality that links prediction and parameter errors. This is the main statistical step that connects noise control to an ℓ_2 bound.

Lemma 7 (Lasso cone and basic inequality) *Let $\widehat{\beta}$ be any minimizer of*

$$F(\beta) = \frac{1}{2n}\|X\beta - \tilde{y}\|_2^2 + \lambda\|\beta\|_1, \qquad \tilde{y} = X\beta^* + u,$$

where u is a noise vector. If $\lambda \geq 2\left\|\frac{1}{n}X^\top u\right\|_\infty$, then with $\Delta = \widehat{\beta} - \beta^*$ we have

$$\|\Delta_{(S^*)^c}\|_1 \leq 3\|\Delta_{S^*}\|_1, \qquad \frac{1}{n}\|X\Delta\|_2^2 \leq 3\lambda\sqrt{s^*}\,\|\Delta\|_2.$$

Proof By optimality,

$$F(\widehat{\beta}) \leq F(\beta^*).$$

Using $\tilde{y} = X\beta^* + u$, we can write

$$F(\beta) = \frac{1}{2n}\|X\beta - (X\beta^* + u)\|_2^2 + \lambda\|\beta\|_1,$$

hence

$$\frac{1}{2n}\|X\widehat{\beta} - (X\beta^* + u)\|_2^2 + \lambda\|\widehat{\beta}\|_1 \;\leq\; \frac{1}{2n}\|X\beta^* - (X\beta^* + u)\|_2^2 + \lambda\|\beta^*\|_1.$$

Since $X\widehat{\beta} - (X\beta^* + u) = X\Delta - u$ and $X\beta^* - (X\beta^* + u) = -u$, expanding the square and canceling $\|u\|_2^2/(2n)$ yields the basic inequality

$$\frac{1}{2n}\|X\Delta\|_2^2 \;\leq\; \frac{1}{n}u^\top X\Delta \;+\; \lambda\big(\|\beta^*\|_1 - \|\widehat{\beta}\|_1\big). \tag{5}$$

By Hölder's inequality and the assumption $\lambda \geq 2\left\|\frac{1}{n}X^\top u\right\|_\infty$,

$$\frac{1}{n}u^\top X\Delta = \left\langle \tfrac{1}{n}X^\top u, \Delta \right\rangle \leq \left\|\tfrac{1}{n}X^\top u\right\|_\infty \|\Delta\|_1 \leq \frac{\lambda}{2}\,\|\Delta\|_1. \tag{6}$$

Let $S^* = \mathrm{supp}(\beta^*)$ with $|S^*| = s^*$. Since $\beta^*_{(S^*)^c} = 0$ and $\widehat{\beta} = \beta^* + \Delta$,

$$\|\widehat{\beta}\|_1 = \|\beta^* + \Delta\|_1 = \|\beta^*_{S^*} + \Delta_{S^*}\|_1 + \|\Delta_{(S^*)^c}\|_1,$$

so that by the reverse triangle inequality $\|a\|_1 - \|a+b\|_1 \leq \|b\|_1$,

$$\|\beta^*\|_1 - \|\widehat{\beta}\|_1 = \|\beta^*_{S^*}\|_1 - \|\beta^*_{S^*} + \Delta_{S^*}\|_1 - \|\Delta_{(S^*)^c}\|_1 \leq \|\Delta_{S^*}\|_1 - \|\Delta_{(S^*)^c}\|_1. \tag{7}$$

Combining (5)–(7) gives

$$\frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{\lambda}{2}\big(\|\Delta_{S^*}\|_1 + \|\Delta_{(S^*)^c}\|_1\big) + \lambda\big(\|\Delta_{S^*}\|_1 - \|\Delta_{(S^*)^c}\|_1\big) = \frac{3\lambda}{2}\,\|\Delta_{S^*}\|_1 - \frac{\lambda}{2}\,\|\Delta_{(S^*)^c}\|_1.$$

Rearranging,

$$\frac{1}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2}\,\|\Delta_{(S^*)^c}\|_1 \leq \frac{3\lambda}{2}\,\|\Delta_{S^*}\|_1. \tag{8}$$

The two stated conclusions follow immediately:

$$\|\Delta_{(S^*)^c}\|_1 \leq 3\,\|\Delta_{S^*}\|_1 \quad \text{(dropping the nonnegative first term in (8))},$$

and, using $\|\Delta_{S^*}\|_1 \leq \sqrt{s^*}\,\|\Delta\|_2$,

$$\frac{1}{n}\|X\Delta\|_2^2 \leq 3\lambda\sqrt{s^*}\,\|\Delta\|_2.$$

This completes the proof. □

We next invoke a classical ISTA convergence guarantee: projected ISTA decreases the objective gap at a rate of order $1/k$ under a standard stepsize condition.

Lemma 8 (ISTA suboptimality gap (Beck and Teboulle, 2009, Thm. 3.1)) *Let $\widehat{\beta}$ minimize $F(\beta) = \frac{1}{2n}\|X\beta - \tilde{y}\|_2^2 + \lambda\|\beta\|_1$ over the convex set $\{\|\beta\|_2 \le C\}$ (with $C > 0$). ISTA with stepsize $1/L$ ($L \ge \sigma_{\max}(X^\top X/n)$), followed by the projection onto $\{\|\beta\|_2 \le C\}$ at each step (as in Algorithm 2), and initialization $\beta^{(0)}$ obeys*

$$F(\beta^{(k)}) - F(\widehat{\beta}) \le \frac{L}{2k}\,\|\beta^{(0)} - \widehat{\beta}\|_2^2.$$

We also use a simple gap-to-distance inequality: for the Lasso objective, a bound on the suboptimality gap implies a corresponding bound on the distance to the exact minimizer under a mild curvature condition along the optimization path.

Lemma 9 *Let $\widehat{\beta}$ be any minimizer of $F(\beta) = \frac{1}{2n}\|X\beta - \tilde{y}\|_2^2 + \lambda\|\beta\|_1$ over the convex set $\{\|\beta\|_2 \le C\}$. Then, for any β with $\|\beta\|_2 \le C$,*

$$F(\beta) - F(\widehat{\beta}) \ge \frac{1}{2n}\,\|X(\beta - \widehat{\beta})\|_2^2.$$

If, in addition, along the optimization path we have the curvature

$$\frac{1}{n}\|X(\beta - \widehat{\beta})\|_2^2 \ge \phi_{\text{opt}}\|\beta - \widehat{\beta}\|_2^2,$$

then

$$\|\beta - \widehat{\beta}\|_2^2 \le \frac{2}{\phi_{\text{opt}}}\,\big(F(\beta) - F(\widehat{\beta})\big).$$

Proof Write $F(\beta) = f(\beta) + g(\beta)$ with $f(\beta) = \frac{1}{2n}\|X\beta - \tilde{y}\|_2^2$ and $g(\beta) = \lambda\|\beta\|_1$. Since $\widehat{\beta}$ minimizes $f+g$ over the convex set $\mathcal{C} := \{\|\beta\|_2 \le C\}$, there exist $s \in \partial g(\widehat{\beta})$ and a normal-cone vector $v \in N_{\mathcal{C}}(\widehat{\beta})$ such that

$$\frac{1}{n}X^\top(X\widehat{\beta} - \tilde{y}) + s + v = 0.$$

For any feasible $\beta \in \mathcal{C}$, denote $\Delta = \beta - \widehat{\beta}$. Using the quadratic identity for f and convexity of g,

$$\begin{aligned} F(\beta) - F(\widehat{\beta}) &= \frac{1}{2n}\|X\Delta\|_2^2 + \Big\langle \frac{1}{n}X^\top(X\widehat{\beta} - \tilde{y}),\, \Delta \Big\rangle \;+\; g(\beta) - g(\widehat{\beta}) \\ &\ge \frac{1}{2n}\|X\Delta\|_2^2 + \Big\langle \frac{1}{n}X^\top(X\widehat{\beta} - \tilde{y}),\, \Delta \Big\rangle \;+\; \langle s,\, \Delta\rangle \\ &= \frac{1}{2n}\|X\Delta\|_2^2 - \langle v,\, \Delta\rangle. \end{aligned}$$

Since $v \in N_{\mathcal{C}}(\widehat{\beta})$ and $\beta \in \mathcal{C}$, we have $\langle v, \beta - \widehat{\beta}\rangle = \langle v, \Delta\rangle \le 0$. Therefore $-\langle v, \Delta\rangle \ge 0$ and

$$F(\beta) - F(\widehat{\beta}) \;\ge\; \frac{1}{2n}\|X(\beta - \widehat{\beta})\|_2^2,$$

proving the first claim. The second claim follows immediately by combining this with the stated curvature condition. □

We are now ready to state the main bound for Algorithm 2. The result shows that one-shot label privatization acts as an additional Gaussian noise source, so the estimator achieves the usual high-dimensional Lasso rate (up to replacing σ_w by an effective noise level that includes the privacy noise), plus an optimization term that decreases along ISTA iterations.

Theorem 10 (High-probability ℓ_2 error for Algorithm 2) *Assume $y = X\beta^* + w$ with $w \sim \mathcal{N}(0, \sigma_w^2 I_n)$, β^* is s^*-sparse, $\|X_j\|_2 \le \sqrt{n}$ for all $j \in [p]$, and the* Restricted Eigenvalue *condition RE$(s^*, 3)$ holds with constant $\phi_0 > 0$, i.e.*

$$\frac{1}{n}\|Xv\|_2^2 \;\ge\; \phi_0\,\|v\|_2^2 \qquad \text{for all } v \in \mathcal{C}(S^*;3) := \{v :\; \|v_{(S^*)^c}\|_1 \le 3\|v_{S^*}\|_1\}.$$

Let Algorithm 2 use stepsize $1/L$ with $L \ge \sigma_{\max}(X^\top X/n)$ and projection radius $C \ge \|\beta^*\|_2$. Let $\widehat{\beta}$ denote the exact Lasso minimizer on the privatized labels over $\{\|\beta\|_2 \le C\}$*, and suppose the ISTA path satisfies*

$$\frac{1}{n}\|X(\beta - \widehat{\beta})\|_2^2 \ge \phi_{\text{opt}}\|\beta - \widehat{\beta}\|_2^2 \qquad \text{(for the iterates } \beta = \beta^{(k)}\text{)}.$$

Fix $c > 0$. Define the no-clipping event $\mathcal{E}_B := \{|y_i| \le B \text{ for all } i \in [n]\}$. Then, with probability at least $1 - \frac{2}{p^c} - \mathbb{P}(\mathcal{E}_B^c)$, for every $k \ge 1$,

$$\|\beta^{(k)} - \beta^*\|_2 \;\le\; \frac{6\sqrt{2}}{\phi_0}\,\sigma_{\text{tot}}\sqrt{\frac{(c+1)\,s^*\,\log p}{n}} + \sqrt{\frac{L}{\phi_{\text{opt}}\,k}}\left[\frac{6\sqrt{2}}{\phi_0}\,\sigma_{\text{tot}}\sqrt{\frac{(c+1)\,s^*\,\log p}{n}} \;+\; C\right], \tag{9}$$

where $\sigma_{\text{tot}}^2 = \sigma_w^2 + \sigma_{\text{dp}}^2$ and $\sigma_{\text{dp}} = \frac{2B}{\epsilon}\sqrt{2\log(1.25/\delta)}$.

Proof Clip the labels so that $|y_i| \le B$ for all i, and release $\tilde{y}_i = \text{clip}(y_i; B) + z_i$ with $z_i \sim \mathcal{N}(0, \sigma_{\text{dp}}^2)$, where

$$\sigma_{\text{dp}} = \frac{2B}{\epsilon}\sqrt{2\log\frac{1.25}{\delta}}.$$

By the Gaussian mechanism (Proposition 1; see also Dwork and Roth (2014)), this ensures (ϵ, δ)-local differential privacy for each released $\tilde{y}_i$. Since the release is one-shot, all subsequent ISTA iterations are post-processing and consume no further privacy budget.

Let $\mathcal{E}_B := \{|y_i| \le B,\ \forall i\}$ denote the no-clipping event. On $\mathcal{E}_B$ we have $\text{clip}(y_i; B) = y_i$, hence $\tilde{y} = y + z = X\beta^* + u$ with

$$u := w + z \sim \mathcal{N}(0, \sigma_{\text{tot}}^2 I_n), \qquad \sigma_{\text{tot}}^2 = \sigma_w^2 + \sigma_{\text{dp}}^2.$$

By Corollary 6 applied to u and column-normalized X,

$$\left\|\frac{1}{n}X^\top u\right\|_\infty \;\le\; \sigma_{\text{tot}}\sqrt{\frac{2(c+1)\log p}{n}} \quad \text{with probability at least } 1 - \frac{2}{p^c}.$$

On this event, with the choice $\lambda = 2\,\sigma_{\text{tot}}\sqrt{\frac{2(c+1)\log p}{n}}$ we have $\lambda \ge 2\|(1/n)X^\top u\|_\infty$.

Since $C \ge \|\beta^*\|_2$, the comparator β^* is feasible for the constrained Lasso problem, so $F(\widehat{\beta}) \le F(\beta^*)$ holds with $\widehat{\beta} = \arg\min_{\|\beta\|_2 \le C} F(\beta)$. Therefore Lemma 7 applies and yields, for $\Delta := \widehat{\beta} - \beta^*$,

$$\|\Delta_{(S^*)^c}\|_1 \le 3\|\Delta_{S^*}\|_1, \qquad \frac{1}{n}\|X\Delta\|_2^2 \le 3\lambda\sqrt{s^*}\,\|\Delta\|_2. \tag{10}$$

$\mathrm{RE}(s^*, 3)$ asserts that for all $v \in \mathcal{C}(S^*; 3)$, $\frac{1}{n}\|Xv\|_2^2 \geq \phi_0 \|v\|_2^2$. Since (10) shows $\Delta \in \mathcal{C}(S^*; 3)$, we combine to get

$$\phi_0 \, \|\Delta\|_2^2 \;\leq\; \frac{1}{n}\|X\Delta\|_2^2 \;\leq\; 3\lambda\sqrt{s^*}\, \|\Delta\|_2,$$

hence

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{3}{\phi_0}\, \lambda\, \sqrt{s^*} = \frac{6\sqrt{2}}{\phi_0}\, \sigma_{\mathrm{tot}} \sqrt{\frac{(c+1)s^* \log p}{n}}.$$

By the ISTA suboptimality bound (Lemma 8) and the gap-to-parameter conversion (Lemma 9) with the stated curvature ϕ_{opt} along the ISTA path,

$$\|\beta^{(k)} - \widehat{\beta}\|_2 \leq \sqrt{\frac{L}{\phi_{\mathrm{opt}}\, k}}\, \|\beta^{(0)} - \widehat{\beta}\|_2.$$

With $\beta^{(0)} = 0$ and $\|\widehat{\beta}\|_2 \leq C$, we have

$$\|\beta^{(0)} - \widehat{\beta}\|_2 \leq \|\beta^{(0)} - \beta^*\|_2 + \|\beta^* - \widehat{\beta}\|_2 \leq C + \|\widehat{\beta} - \beta^*\|_2,$$

where we used $C \geq \|\beta^*\|_2$. Thus

$$\|\beta^{(k)} - \widehat{\beta}\|_2 \leq \sqrt{\frac{L}{\phi_{\mathrm{opt}}\, k}}\, \big(\|\widehat{\beta} - \beta^*\|_2 + C\big).$$

Using the triangle inequality $\|\beta^{(k)} - \beta^*\|_2 \leq \|\widehat{\beta} - \beta^*\|_2 + \|\beta^{(k)} - \widehat{\beta}\|_2$ and substituting the statistical bound for $\|\widehat{\beta} - \beta^*\|_2$ yields exactly (9). Finally, the bound holds on $\mathcal{E}_B$ and the score-bound event; hence the overall probability is at least $1 - \frac{2}{p^c} - \mathbb{P}(\mathcal{E}_B^c)$. □

3.2 Locally Differentially Private ISTA

This subsection studies an end-to-end local privacy mechanism in which both features and responses are privatized *once* on the client side before any training. Each user clips the feature vector to an ℓ_2 radius R and the label to a scalar bound B, and then applies independent Gaussian mechanisms calibrated to the allocated budgets (ϵ_x, δ_x) and (ϵ_y, δ_y). The server receives only the privatized data $(\tilde{X}, \tilde{y})$ and runs an ISTA routine, i.e., a gradient step, soft-thresholding, and an ℓ_2 projection, to obtain the estimator.

Relative to label-only privatization, this scheme provides *record-level* local protection but perturbs the design matrix, which introduces additional stochastic terms that must be controlled in the analysis. Accordingly, our guarantees rely on the imposed clipping/projection constraints and mild curvature conditions stated for the privatized design. Because the perturbation is one-shot and performed locally, all subsequent server-side optimization is post-processing and incurs no further privacy loss.

Algorithm 3 Locally Differentially Private ISTA (Noise Added Once to X and y)

Input: Private records $\{(x_i, y_i)\}_{i=1}^n$ held client-side; privacy budgets (ϵ_x, δ_x) for X and (ϵ_y, δ_y) for y; clipping radii R (features) and B (labels); regularization parameter λ; number of iterations K; learning rate L' (e.g. $L' \geq \sigma_{\max}(\tilde{X}^\top \tilde{X}/n)$ computed on privatized data); projection radius $C \geq \|\beta^*\|_2$.

Output: An estimate of $\beta^{(K)}$.

Local perturbation at each user $i = 1, \ldots, n$:

Clip features and label: $x_i^{\text{clip}} = x_i \cdot \min\Big\{1, \frac{R}{\|x_i\|_2}\Big\}, \quad y_i^{\text{clip}} = \text{sign}(y_i)\ \min\{|y_i|, B\}$

Set LDP noise scales:

$$\sigma_x = \frac{2R}{\epsilon_x}\sqrt{2\ln\frac{1.25}{\delta_x}}, \qquad \sigma_y = \frac{2B}{\epsilon_y}\sqrt{2\ln\frac{1.25}{\delta_y}}$$

Add Gaussian noise locally and send to server: $\tilde{x}_i = x_i^{\text{clip}} + \eta_i,\ \eta_i \sim \mathcal{N}(0, \sigma_x^2 I_p)$; $\tilde{y}_i = y_i^{\text{clip}} + \zeta_i,\ \zeta_i \sim \mathcal{N}(0, \sigma_y^2)$

Server side (using $\tilde{X} = [\tilde{x}_1^\top; \ldots; \tilde{x}_n^\top]$ and $\tilde{y}$):

Initialize $\beta^{(0)} = 0$

for $k = 0, \ldots, K-1$ **do**

- Compute the gradient on privatized data: $\nabla L(\beta^{(k)}) = \frac{1}{n}\tilde{X}^\top\big(\tilde{X}\beta^{(k)} - \tilde{y}\big)$
- Gradient descent step: $\beta_{\text{temp}}^{(k+1)} = \beta^{(k)} - \frac{1}{L'}\nabla L(\beta^{(k)})$
- Soft-thresholding: $\beta_{\text{thr}}^{(k+1)} = S\Big(\beta_{\text{temp}}^{(k+1)}, \frac{\lambda}{L'}\Big)$
- Projection to ℓ_2-ball: $\beta^{(k+1)} = \Pi_{\mathbb{B}_2(C)}\big(\beta_{\text{thr}}^{(k+1)}\big)$

return $\beta^{(K)}$

3.2.1 Main Upper Bound

We now state a finite-sample, high-probability ℓ_2 upper bound for the one-shot LDP scheme on (X, y). Compared with label-only privatization, perturbing the design introduces additional stochastic terms, but under standard high-dimensional conditions on the privatized design the bound retains the familiar $\sqrt{s^* \log p/n}$ scaling, with privacy entering through the effective noise levels of the feature and label mechanisms. As in the label-only case, the ISTA optimization residual decays with the iteration budget and, since optimization is post-processing, does not incur additional privacy loss. The formal statement is given in Theorem 12.

In the one-shot LDP scheme on (X, y), the feature perturbations introduce Gaussian vectors whose squared norms (and related quadratic forms) must be controlled uniformly with high probability. We therefore recall a standard χ^2 tail inequality that allows us to bound these random norms at the scale needed for our subsequent coordinatewise bounds and choice of λ.

Lemma 11 (Laurent–Massart χ^2 tail bound (Laurent and Massart, 2000, Lemma 1)) *Let $U \sim \chi^2_m$. For any $t \geq 0$,*

$$\Pr\big(U - m \geq 2\sqrt{mt} + 2t\big) \leq e^{-t},$$
$$\Pr\big(m - U \geq 2\sqrt{mt}\big) \leq e^{-t}.$$

We now give the principal ℓ_2 performance bound for the setting where each user privatizes *both* the feature vector and the response *once* before any optimization takes place. Relative to label-only privatization, perturbing X affects the effective design seen by the server and introduces additional stochastic terms that must be controlled; the theorem summarizes how these effects enter the estimation error through the regularization level and a single privacy-driven contribution, while the optimization error remains a standard ISTA suboptimality term.

Theorem 12 (High-probability ℓ_2 error for for Algorithm 3) *Assume $y = X\beta^* + w$ with $w \sim \mathcal{N}(0, \sigma_w^2 I_n)$ and β^* is s^*-sparse with support S^*. Assume the columns of X satisfy $\|X_j\|_2 \leq \sqrt{n}$. Each record is locally privatized once as in Algorithm 3:*

$$\tilde{x}_i = x_i^{\mathrm{clip}} + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \sigma_x^2 I_p), \qquad \tilde{y}_i = y_i^{\mathrm{clip}} + \zeta_i, \quad \zeta_i \sim \mathcal{N}(0, \sigma_y^2),$$

independently across i, where x_i^{clip} and y_i^{clip} are the radius-R and bound-B clippings, and

$$\sigma_x \;=\; \frac{2R}{\epsilon_x}\sqrt{2\log\frac{1.25}{\delta_x}}, \qquad \sigma_y \;=\; \frac{2B}{\epsilon_y}\sqrt{2\log\frac{1.25}{\delta_y}}.$$

Let $\tilde{X} = (\tilde{x}_1^\top, \ldots, \tilde{x}_n^\top)^\top$ and $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^\top$. Run Algorithm 3 with stepsize $1/L'$ where $L' \geq \sigma_{\max}(\tilde{X}^\top \tilde{X}/n)$, soft-thresholding parameter λ, and projection radius $C \geq \|\beta^*\|_2$. Let $\widehat{\beta}$ be the exact Lasso minimizer on $(\tilde{X}, \tilde{y})$ under $\|\beta\|_2 \leq C$.

Assume a Restricted Eigenvalue condition holds for $\tilde{X}$ on the cone $\mathcal{C}(S^*; 3)$ with constant $\phi_0' > 0$:

$$\frac{1}{n}\|\tilde{X}v\|_2^2 \;\geq\; \phi_0' \, \|v\|_2^2 \quad \text{for all } v \in \mathcal{C}(S^*; 3),$$

and along the ISTA path we have the curvature

$$\frac{1}{n}\|\tilde{X}(\beta - \widehat{\beta})\|_2^2 \geq \phi_{\mathrm{opt}}'\|\beta - \widehat{\beta}\|_2^2 \qquad \text{(for the iterates } \beta = \beta^{(k)}\text{)}.$$

Fix $c > 0$ and define

$$\alpha_n \;:=\; \sqrt{1 + \frac{2\sqrt{(c+2)\log p}}{\sqrt{n}} + \frac{2(c+2)\log p}{n}}\,.$$

Let $\mathcal{E}_{R,B} := \{\|x_i\|_2 \leq R,\ |y_i| \leq B \text{ for all } i \in [n]\}$ denote the no-clipping event. Choose

$$\lambda \;=\; 2\left[\left(\sqrt{\sigma_w^2 + \sigma_y^2} + \sigma_x C\right)\sqrt{\frac{2(c+2)\log p}{n}} + \sqrt{\sigma_w^2 + \sigma_y^2}\,\sigma_x\,\alpha_n\sqrt{\frac{2(c+2)\log p}{n}} + \sigma_x^2\, C\,\alpha_n^2\right].$$

Then, with probability at least $1 - \frac{2}{p^c} - \mathbb{P}(\mathcal{E}_{R,B}^c)$, for every $k \geq 1$,

$$\|\beta^{(k)} - \beta^*\|_2 \;\leq\; \frac{3}{\phi_0'}\,\lambda\sqrt{s^*} \;+\; \sqrt{\frac{L'}{\phi_{\mathrm{opt}}'\,k}}\left(\frac{3}{\phi_0'}\,\lambda\sqrt{s^*} + C\right). \tag{11}$$

Proof The local clipping guarantees $\|x_i^{\text{clip}}\|_2 \le R$ and $|y_i^{\text{clip}}| \le B$, hence the ℓ_2-sensitivities satisfy $\Delta_{2,x} \le 2R$ and $\Delta_{2,y} \le 2B$. By the Gaussian mechanism (Proposition 1; see also Dwork and Roth (2014)), adding independent noise $\eta_i \sim \mathcal{N}(0, \sigma_x^2 I_p)$ and $\zeta_i \sim \mathcal{N}(0, \sigma_y^2)$ with the stated (σ_x, σ_y) ensures (ϵ_x, δ_x)-LDP for features and (ϵ_y, δ_y)-LDP for labels, respectively. By basic composition, $(\tilde{x}_i, \tilde{y}_i)$ is $(\epsilon_x + \epsilon_y,\ \delta_x + \delta_y)$-LDP. Since the perturbation is one-shot and local, all server-side ISTA iterations are post-processing.

Let $\widehat{\beta}$ be the constrained Lasso minimizer on $(\tilde{X}, \tilde{y})$ under $\|\beta\|_2 \le C$, and set $\Delta = \widehat{\beta} - \beta^*$. To apply Lemma 7, it suffices to control the score term $\left\|\frac{1}{n}\tilde{X}^\top(\tilde{y} - \tilde{X}\beta^*)\right\|_\infty$ and choose λ to dominate it.

Define the feature-noise matrix $Z \in \mathbb{R}^{n\times p}$ by stacking the rows $\eta_i^\top$, and the label-noise vector $\xi \in \mathbb{R}^n$ by stacking ζ_i. On the no-clipping event $\mathcal{E}_{R,B}$ we have $x_i^{\text{clip}} = x_i$ and $y_i^{\text{clip}} = y_i$ for all i, hence

$$\tilde{X} = X + Z, \qquad \tilde{y} = y + \xi = X\beta^* + (w + \xi).$$

Therefore

$$\tilde{y} - \tilde{X}\beta^* = (w + \xi) - Z\beta^*.$$

For each coordinate $j \in [p]$,

$$\frac{1}{n}\tilde{X}_j^\top(\tilde{y} - \tilde{X}\beta^*) = \underbrace{\tfrac{1}{n}X_j^\top(w+\xi)}_{(A)} + \underbrace{\tfrac{1}{n}Z_j^\top(w+\xi)}_{(B)} - \underbrace{\tfrac{1}{n}X_j^\top Z\beta^*}_{(C)} - \underbrace{\tfrac{1}{n}Z_j^\top Z\beta^*}_{(D)}.$$

Control of (A). By Corollary 6 applied to $u = w+\xi \sim \mathcal{N}(0, (\sigma_w^2+\sigma_y^2)I_n)$ and $\|X_j\|_2 \le \sqrt{n}$,

$$\left\|\tfrac{1}{n}X^\top(w+\xi)\right\|_\infty \le \sqrt{\sigma_w^2 + \sigma_y^2}\sqrt{\tfrac{2(c+2)\log p}{n}}$$

with probability at least $1 - 2/p^{c+2}$.

Control of (C). Conditional on β^*, each $X_j^\top Z\beta^*/n$ is Gaussian with variance $\sigma_x^2\|\beta^*\|_2^2\|X_j\|_2^2/n^2 \le \sigma_x^2\|\beta^*\|_2^2/n$. Applying Corollary 6 gives

$$\left\|\tfrac{1}{n}X^\top Z\beta^*\right\|_\infty \le \sigma_x\|\beta^*\|_2\sqrt{\tfrac{2(c+2)\log p}{n}}$$

with probability at least $1 - 2/p^{c+2}$.

Control of (B). Conditional on Z, each coordinate of $\frac{1}{n}Z^\top(w+\xi)$ is Gaussian with conditional variance $(\sigma_w^2 + \sigma_y^2)\|Z_j\|_2^2/n^2$. Hence, conditional on Z,

$$\left\|\tfrac{1}{n}Z^\top(w+\xi)\right\|_\infty \le \sqrt{\sigma_w^2 + \sigma_y^2}\left(\max_{j\in[p]}\tfrac{\|Z_j\|_2}{n}\right)\sqrt{2(c+2)\log p}$$

with conditional probability at least $1 - 2/p^{c+2}$. It remains to bound $\max_j \|Z_j\|_2$. Since $\|Z_j\|_2^2/\sigma_x^2 \sim \chi_n^2$, Lemma 11 with $t = (c+2)\log p$ yields

$$\Pr\left(\max_{j\in[p]}\frac{\|Z_j\|_2}{\sigma_x\sqrt{n}} \le \alpha_n\right) \ \ge\ 1 - \frac{1}{p^{c+1}}.$$

Combining the two displays gives

$$\left\|\tfrac{1}{n}Z^\top(w+\xi)\right\|_\infty \le \sqrt{\sigma_w^2 + \sigma_y^2}\,\sigma_x\,\alpha_n\,\sqrt{\tfrac{2(c+2)\log p}{n}}$$

with probability at least $1 - 2/p^{c+2} - 1/p^{c+1}$.

Control of (D). By Cauchy–Schwarz, $\left|\frac{1}{n}Z_j^\top Z\beta^*\right| \le \frac{\|Z_j\|_2}{n}\|Z\beta^*\|_2$. Moreover, $\|Z\beta^*\|_2^2/(\sigma_x^2\|\beta^*\|_2^2) \sim \chi_n^2$. By Lemma 11 with $t = (c+2)\log p$,

$$\Pr\left(\frac{\|Z\beta^*\|_2}{\sigma_x\|\beta^*\|_2\sqrt{n}} \le \alpha_n\right) \ \ge\ 1 - \frac{1}{p^{c+2}}.$$

On the intersection of this event and $\max_j \|Z_j\|_2 \le \sigma_x\sqrt{n}\,\alpha_n$, we obtain

$$\left\|\tfrac{1}{n} Z^\top Z\beta^*\right\|_\infty \le \sigma_x^2\,\|\beta^*\|_2\,\alpha_n^2.$$

Putting (A)–(D) together and using $\|\beta^*\|_2 \le C$, we conclude that on $\mathcal{E}_{R,B}$, with probability at least $1-2/p^c$ (after a union bound over the displayed events),

$$\left\|\tfrac{1}{n}\tilde{X}^\top(\tilde{y}-\tilde{X}\beta^*)\right\|_\infty \le \left(\sqrt{\sigma_w^2+\sigma_y^2}+\sigma_x C\right)\sqrt{\tfrac{2(c+2)\log p}{n}}+\sqrt{\sigma_w^2+\sigma_y^2}\,\sigma_x\,\alpha_n\sqrt{\tfrac{2(c+2)\log p}{n}}+\sigma_x^2\,C\,\alpha_n^2.$$

With λ chosen as in the theorem statement, this implies $\lambda \ge 2\left\|\tfrac{1}{n}\tilde{X}^\top(\tilde{y}-\tilde{X}\beta^*)\right\|_\infty$.

Since $C \ge \|\beta^*\|_2$, the comparator β^* is feasible for the constrained Lasso objective, so $F(\widehat{\beta}) \le F(\beta^*)$. Therefore Lemma 7 applied to $(\tilde{X},\tilde{y})$ yields

$$\|\Delta_{(S^*)^c}\|_1 \le 3\|\Delta_{S^*}\|_1, \qquad \frac{1}{n}\|\tilde{X}\Delta\|_2^2 \le 3\lambda\sqrt{s^*}\,\|\Delta\|_2.$$

Combining with $\mathrm{RE}(s^*,3)$ for $\tilde{X}$ gives

$$\|\widehat{\beta}-\beta^*\|_2 \le \frac{3}{\phi_0'}\,\lambda\,\sqrt{s^*}.$$

Finally, Lemma 8 and Lemma 9 applied to the constrained problem on $(\tilde{X},\tilde{y})$, together with the curvature condition along the ISTA path, yield

$$\|\beta^{(k)}-\widehat{\beta}\|_2 \le \sqrt{\frac{L'}{\phi'_{\mathrm{opt}}\,k}}\;\|\beta^{(0)}-\widehat{\beta}\|_2 \le \sqrt{\frac{L'}{\phi'_{\mathrm{opt}}\,k}}\;\left(\|\widehat{\beta}-\beta^*\|_2+C\right),$$

where the last inequality uses $\beta^{(0)}=0$ and $\|\beta^*\|_2 \le C$. The triangle inequality then gives (11). The overall success probability is at least $1-2/p^c-\mathbb{P}(\mathcal{E}_{R,B}^c)$. □

3.3 Differentially Private ISTA

This subsection considers a *central* (server-side) differentially private mechanism in which Gaussian noise is injected into the clipped empirical gradient at every ISTA iteration. In contrast to the one-shot local schemes, privacy is enforced through iterative access to the data: each gradient query is perturbed, and the per-iteration noise level is calibrated to a total budget (ε,δ) via standard privacy accounting. To control sensitivity and ensure stable iterates under repeated perturbations, we impose row clipping $\|x_i\|_2 \le R$ and label clipping $|y_i| \le B$, and we incorporate hard-thresholding and an ℓ_2 projection step to control support size and iterate norms. The resulting privacy–accuracy tradeoff reflects the fact that privacy loss aggregates across iterations and must be accounted for (e.g., via advanced composition).

Algorithm 4 Differentially Private ISTA (Noise Added to Gradient)

Input: Design matrix $X \in \mathbb{R}^{n \times p}$; private labels $y \in \mathbb{R}^n$; row bound R; label bound B; privacy (ε, δ); regularization λ; iterations K; sparsity level s^*; stepsize $L \geq L_s$; projection radius $C \geq \|\beta^*\|_2$.

Output: An estimate $\beta^{(K)}$.

Row/label clipping: $X_i^{\text{clip}} = X_i \cdot \min\Big\{1, \frac{R}{\|X_i\|_2}\Big\}, \quad y_i^{\text{clip}} = \text{sign}(y_i)\ \min\{|y_i|, B\}$

Set per-iteration DP noise scale (for a K-step (ε, δ)-DP run):

$$\sigma_{\text{dp}} \;=\; \frac{8\,R\,(RC+B)}{n\,\varepsilon}\,\sqrt{K\,\log\frac{2}{\delta}\,\log\frac{2.5K}{\delta}}\;.$$

Initialize $\beta^{(0)} = 0$

for $k = 0, \ldots, K-1$ **do**

Gradient of squared loss on clipped data:

$\nabla L(\beta^{(k)}) = \frac{1}{n}(X^{\text{clip}})^\top\big(X^{\text{clip}}\beta^{(k)} - y^{\text{clip}}\big)$

Add Gaussian noise:

$\xi^{(k)} \sim \mathcal{N}(0, \sigma_{\text{dp}}^2 I_p), \quad \nabla L_{\text{DP}}(\beta^{(k)}) = \nabla L(\beta^{(k)}) + \xi^{(k)}$

Gradient step:

$\beta_{\text{temp}}^{(k+1)} = \beta^{(k)} - \frac{1}{L}\nabla L_{\text{DP}}(\beta^{(k)})$

Soft-thresholding:

$\beta_{\text{thr}}^{(k+1)} = S\Big(\beta_{\text{temp}}^{(k+1)}, \frac{\lambda}{L}\Big)$

Hard-thresholding (keep largest s^* magnitudes):

$\widehat{\beta}^{(k+1)} = \text{HardThreshold}\big(\beta_{\text{thr}}^{(k+1)}, s^*\big)$

Projection to ℓ_2-ball:

$\beta^{(k+1)} = \Pi_{\mathbb{B}_2(C)}\big(\widehat{\beta}^{(k+1)}\big)$

return $\beta^{(K)}$

3.3.1 Main Upper Bound

We now state a finite-sample, high-probability ℓ_2 upper bound for Algorithm 4. Under the standing sparsity and curvature conditions, the error separates into (i) an optimization term that contracts geometrically with the iteration budget K, (ii) a statistical term driven by the observational noise, and (iii) a privacy-induced term determined by the gradient perturbation level calibrated under the chosen privacy accounting.

We first invoke a basic property of the proximal (soft-thresholding) operator: relative to a sparse target, the shrinkage step deviates from the target by at most the pre-thresholding error plus a sparsity-scaled threshold term.

Lemma 13 (Soft-thresholding bias) *Let $S(u,\tau)$ be coordinate-wise soft-thresholding. If v is s^*-sparse, then*

$$\|S(u,\tau) - v\|_2 \le \|u - v\|_2 + \sqrt{s^*}\,\tau.$$

Proof Let $S^* = \mathrm{supp}(v)$ with $|S^*| \le s^*$. For $i \notin S^*$, $v_i = 0$ and $|S(u_i,\tau) - v_i| = |S(u_i,\tau)| \le |u_i| = |u_i - v_i|$. For $i \in S^*$, $|S(u_i,\tau) - v_i| \le |u_i - v_i| + |S(u_i,\tau) - u_i| \le |u_i - v_i| + \tau$. Thus

$$\|S(u,\tau) - v\|_2^2 \le \|u - v\|_2^2 + 2\tau\|(u-v)_{S^*}\|_1 + s^*\tau^2 \le \big(\|u - v\|_2 + \sqrt{s^*}\tau\big)^2,$$

using $\|(u-v)_{S^*}\|_1 \le \sqrt{s^*}\|u - v\|_2$. Taking square roots yields the claim. □

Next we give a standard approximation guarantee for the hard-thresholding operator: keeping the largest-magnitude coordinates produces a vector that is within a constant factor of the best sparse approximation error (and, in particular, is stable when compared to any sparse reference).

Lemma 14 (Hard-thresholding on union support (Blumensath and Davies, 2009)) *Let $\mathrm{HT}_{s^*}(u)$ keep the s^* largest-magnitude entries of u. For any s^*-sparse v, define $T := \mathrm{supp}(\mathrm{HT}_{s^*}(u)) \cup \mathrm{supp}(v)$, so $|T| \le 2s^*$. Then*

$$\|\mathrm{HT}_{s^*}(u) - v\|_2 \le 2\,\|(u-v)_T\|_2 \le 2\,\|u - v\|_2 \le 2\sqrt{2s^*}\,\|u - v\|_\infty.$$

Proof Let $\hat{u} := \mathrm{HT}_{s^*}(u)$. Since $\hat{u}$ is the best s^*-term approximation to u in ℓ_2, restricted to T we have $\|u_T - \hat{u}_T\|_2 \le \|u_T - v_T\|_2$. Now $\hat{u} - v$ is supported on T, and by triangle inequality,

$$\|\hat{u} - v\|_2 = \|\hat{u}_T - v_T\|_2 \le \|u_T - v_T\|_2 + \|u_T - \hat{u}_T\|_2 \le 2\|u_T - v_T\|_2.$$

The remaining inequalities are immediate. □

We next state an ℓ_2 error bound for Algorithm 4. The bound separates an optimization contraction term, the usual high-dimensional statistical term, and an explicit privacy term determined by the accountant-calibrated noise.

Theorem 15 (High-probability ℓ_2 error for Algorithm 4) *Assume the linear model $y = X\beta^* + w$ with $w \sim \mathcal{N}(0, \sigma_w^2 I_n)$, β^* is s^*-sparse, and $\|X_j\|_2 \le \sqrt{n}$. Assume the RSC/RSS condition holds on every support of size $\le 3s^*$:*

$$\mu_s\|v\|_2^2 \;\le\; \frac{1}{n}\|Xv\|_2^2 \;\le\; L_s\|v\|_2^2, \qquad \text{for all } v \text{ with } |\mathrm{supp}(v)| \le 3s^*.$$

Let the stepsize be $1/L$ with $L \ge L_s$ and define $\kappa := 1 - \mu_s/L \in (0,1)$; assume $2\mu_s > L$. Assume $C \ge \|\beta^*\|_2$. Fix $c > 0$. Then, with probability at least $1 - \dfrac{2}{p^c}$,

$$\begin{aligned}\big\|\beta^{(K)} - \beta^*\big\|_2 \le{}& \big(2 - 2\mu_s/L\big)^K\,\|\beta^{(0)} - \beta^*\|_2 \\ &+ \frac{2\sqrt{2\,s^*}}{2\mu_s - L}\left[\sigma_w\sqrt{\frac{2(c+1)\log p}{n}} + \sigma_{\mathrm{dp}}\sqrt{2\big((c+1)\log p + \log K\big)}\right] \qquad (12)\\ &+ \frac{2\sqrt{s^*}\,\lambda}{2\mu_s - L},\end{aligned}$$

where

$$\sigma_{\mathrm{dp}} = \frac{8\,R\,(RC+B)}{n\,\varepsilon}\,\sqrt{K\,\log\frac{2}{\delta}\;\log\frac{2.5K}{\delta}}\,.$$

Proof Let $e^{(k)} := \beta^{(k)} - \beta^*$ and write (using clipped data in the gradient)

$$\beta^{(k+1)}_{\mathrm{temp}} - \beta^* = \Big(I - \tfrac{1}{L}\cdot\tfrac{1}{n}(X^{\mathrm{clip}})^\top X^{\mathrm{clip}}\Big)e^{(k)} \;-\; \tfrac{1}{L}\,\eta^{(k)}, \qquad \eta^{(k)} := \tfrac{1}{n}(X^{\mathrm{clip}})^\top w + \xi^{(k)}.$$

By the RSC/RSS condition on supports of size $\le 3s^*$ and $L \ge L_s$,

$$\Big\|\Big(I - \tfrac{1}{L}\cdot\tfrac{1}{n}(X^{\mathrm{clip}})^\top X^{\mathrm{clip}}\Big)e^{(k)}\Big\|_2 \le \Big(1 - \frac{\mu_s}{L}\Big)\|e^{(k)}\|_2 =: \kappa\|e^{(k)}\|_2.$$

Let $B_{k+1} := \mathrm{supp}(\widehat{\beta}^{(k+1)}) \cup S^*$, so $|B_{k+1}| \le 2s^*$. Since $\beta^* \in \mathbb{B}_2(C)$, projection is nonexpansive and $\Pi_{\mathbb{B}_2(C)}(\beta^*) = \beta^*$, hence

$$\|e^{(k+1)}\|_2 = \|\beta^{(k+1)} - \beta^*\|_2 \le \|\widehat{\beta}^{(k+1)} - \beta^*\|_2.$$

Apply Lemma 14 with $u = \beta^{(k+1)}_{\mathrm{thr}}$ and $v = \beta^*$, then Lemma 13 on the same union support, to get

$$\|\widehat{\beta}^{(k+1)} - \beta^*\|_2 \le 2\|(\beta^{(k+1)}_{\mathrm{thr}} - \beta^*)_{B_{k+1}}\|_2 \le 2\|(\beta^{(k+1)}_{\mathrm{temp}} - \beta^*)_{B_{k+1}}\|_2 + \frac{2\sqrt{s^*}\lambda}{L}.$$

Next,

$$\|(\beta^{(k+1)}_{\mathrm{temp}} - \beta^*)_{B_{k+1}}\|_2 \le \|\beta^{(k+1)}_{\mathrm{temp}} - \beta^*\|_2 \le \kappa\|e^{(k)}\|_2 + \frac{1}{L}\|\eta^{(k)}_{B_{k+1}}\|_2 \le \kappa\|e^{(k)}\|_2 + \frac{\sqrt{2s^*}}{L}\|\eta^{(k)}\|_\infty.$$

Combining yields the recursion

$$\|e^{(k+1)}\|_2 \le 2\kappa\|e^{(k)}\|_2 + \frac{2\sqrt{2s^*}}{L}\|\eta^{(k)}\|_\infty + \frac{2\sqrt{s^*}\lambda}{L}.$$

Unrolling and using $1 - 2\kappa = (2\mu_s - L)/L$ gives

$$\|e^{(K)}\|_2 \le (2\kappa)^K\|e^{(0)}\|_2 + \frac{1}{1-2\kappa}\left(\frac{2\sqrt{2s^*}}{L}\max_{0\le k<K}\|\eta^{(k)}\|_\infty + \frac{2\sqrt{s^*}\lambda}{L}\right).$$

Using the coordinate-wise Gaussian bound for $(1/n)(X^{\mathrm{clip}})^\top w$ (with $\|X_j\|_2 \le \sqrt{n}$) and a Gaussian maximal inequality over the pK coordinates of $\{\xi^{(k)}\}_{k=0}^{K-1}$, a union bound yields with probability at least $1 - \frac{2}{p^c}$:

$$\max_{0\le k<K}\|\eta^{(k)}\|_\infty \le \sigma_w\sqrt{\frac{2(c+1)\log p}{n}} + \sigma_{\mathrm{dp}}\sqrt{2\big((c+1)\log p + \log K\big)}.$$

It remains to calibrate σ_{dp}. For neighboring datasets differing in one record and any fixed β with $\|\beta\|_2 \le C$, row clipping $\|x_i\|_2 \le R$ and label clipping $|y_i| \le B$ imply the ℓ_2 sensitivity

$$\Delta_2 := \sup_{D\sim D'}\big\|\nabla L(\beta; D) - \nabla L(\beta; D')\big\|_2 \le \frac{2R(RC+B)}{n}.$$

A single iteration is $(\varepsilon_0, \delta_0)$-DP under the Gaussian mechanism provided $\sigma_{\mathrm{dp}} \ge \Delta_2\sqrt{2\log(1.25/\delta_0)}/\varepsilon_0$. Using advanced composition with $\delta' = \delta/2$ and $\delta_0 = \delta/(2K)$, one may take $\varepsilon_0 = \varepsilon/\big(2\sqrt{2K\log(2/\delta)}\big)$ (for $\varepsilon < 1$), which yields

$$\sigma_{\mathrm{dp}} = \frac{8\,R\,(RC+B)}{n\,\varepsilon}\,\sqrt{K\,\log\frac{2}{\delta}\;\log\frac{2.5K}{\delta}}\,.$$

Substituting completes (12). □

4 Simulation Studies

We study the sparse linear model $y = X\beta^* + w$ under three switchable data-generating schemes. We vary the sample size n, set the dimension to $p = 2n$, and fix the sparsity level at $s^* = 20$. For each n, we draw a random support $S^* \subset [p]$ with $|S^*| = s^*$, generate the coefficients on S^* as specified below, normalize them to unit ℓ_2 norm, embed into $\mathbb{R}^p$ to obtain β^*, and then generate $y = X\beta^* + w$. We fix the privacy budget at $(\varepsilon, \delta) = \big(0.5,\, 10/n^{1.1}\big)$.

All iterative methods use stepsize $1/L$ with

$$L = \sigma_{\max}\left(\frac{A^\top A}{n}\right), \qquad A \in \{X, \widetilde{X}\},$$

where A denotes the design matrix available to the algorithm (e.g., X for label-private settings and $\widetilde{X}$ for feature-private settings). The regularization parameter λ is tuned on the same grid across methods that employ an ℓ_1 penalty; the number of iterations K is fixed and shared across iterative methods. Performance is evaluated by the ℓ_2 error $\|\widehat{\beta} - \beta^*\|_2$. Unless stated otherwise, we use a fixed seed for reproducibility (the same seed list is used across all methods).

1. **Setting 1 (Gaussian features, Gaussian noise).**

$$X_{ij} \sim \mathcal{N}(0,1) \text{ i.i.d.}, \qquad w_i \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

   Coefficients: on S^*, draw $\{\beta_j^*\}_{j\in S^*} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then normalize to $\|\beta^*\|_2 = 1$. *This is a canonical light-tailed setting used throughout the high-dimensional Lasso literature.*

2. **Setting 2 (scaled-uniform features, Gaussian noise).**

$$X_{ij} \sim \text{Unif}\big(-1/\sqrt{p},\, 1/\sqrt{p}\big) \text{ i.i.d.}, \qquad w_i \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

   Coefficients: on S^*, draw $\{\beta_j^*\}_{j\in S^*} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then normalize to $\|\beta^*\|_2 = 1$. *This is a bounded, mean-zero design after standard scaling; it also matches the experimental setting used in the DP-IHT baseline (Cai et al., 2021).*

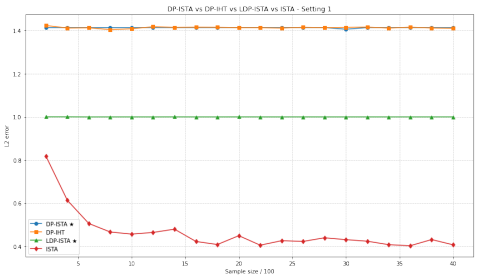
3. **Setting 3 (Rademacher features, bounded-uniform noise).**

$$X_{ij} \in \{-1, +1\} \text{ with prob. } 1/2, \qquad w_i \sim \text{Unif}[-b_0, b_0] \;\; \text{(e.g., } b_0 = 0.05\text{)}.$$

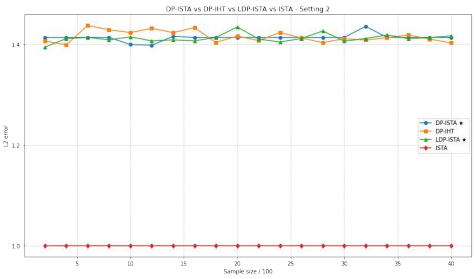
   Coefficients: on S^*, draw $\{\beta_j^*\}_{j\in S^*} \overset{\text{i.i.d.}}{\sim} \text{Unif}[0,1]$, then normalize to $\|\beta^*\|_2 = 1$. *This binarized-design regime is also used in the Label-LDP-IHT baseline (Wang and Xu, 2021).*

4.1 Experiment 1: non-interactive Local DP vs. central DP on (X, y)

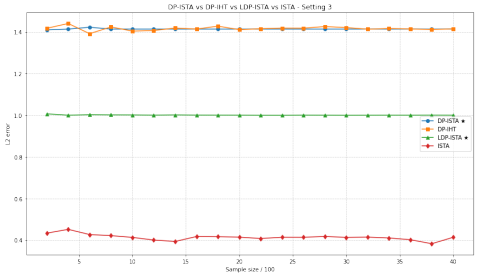
We compare LDP-ISTA (Algorithm 3, non-interactive local DP), DP-ISTA (Algorithm 4, central DP), DP-IHT (comparison, central DP), the differentially private iterative hard-thresholding method of Cai et al. (2021), and ISTA (non-private), the non-private counterpart under the three experimental settings.



(a) Experimental Setting 1



(b) Experimental Setting 2



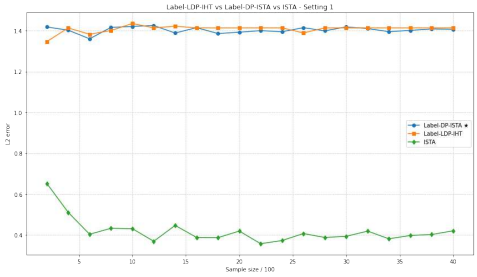
(c) Experimental Setting 3

Fig. 1: Experiment 1 results under three experimental settings (Algorithm 3, Algorithm 4). The horizontal axis is $n/100$.

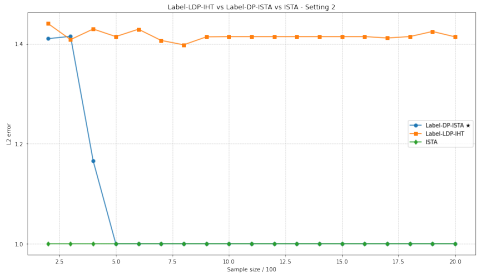
Figure 1 summarizes the results. In Settings 1 and 3, LDP-ISTA attains the smallest ℓ_2 error among the private methods and stays closer to the non-private ISTA baseline. In Setting 2, all private methods cluster in a narrow band, indicating similar performance under the bounded scaled-uniform design (which is also the environment chosen in the DP-IHT baseline). Across multiple regimes, DP-ISTA and DP-IHT exhibit nearly identical behavior, which is consistent with their similar Gaussian noise-injection structure within iterative updates. Overall, LDP-ISTA is competitive with, and in several regimes improves upon, representative central-DP baselines under a common privacy budget.

4.2 Experiment 2: Label-private regression under LDP

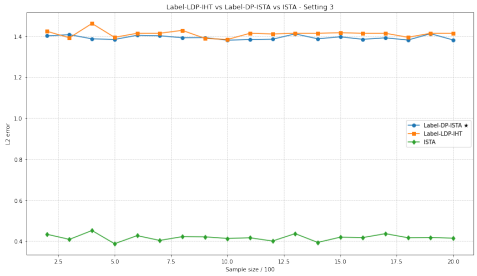
We compare Label-LDP-ISTA (Algorithm 2), Label-LDP-IHT, the label-private IHT method of Wang and Xu (2021), and ISTA (non-private), the non-private counterpart under the three experimental settings.



(a) Experimental Setting 1



(b) Experimental Setting 2



(c) Experimental Setting 3

Fig. 2: Experiment 2 results under three experimental settings (Algorithm 2). The horizontal axis is $n/100$.

Figure 2 reports the results. Across the three regimes, Label-LDP-ISTA is never worse than the comparator and is sometimes better. In Settings 1 and 3, the two private curves nearly overlap, with Label-LDP-ISTA occasionally slightly lower. In Setting 2, Label-LDP-ISTA decreases markedly as the sample size grows and approaches the non-private baseline, whereas Label-LDP-IHT remains at a higher error level, yielding a visible gap. Taken together, Label-LDP-ISTA at least matches the baseline in the baseline's preferred regime (Setting 3) and demonstrates improved convergence behavior in other light-tailed/bounded regimes.

5 Empirical Studies

In this section, we demonstrate the performance of the differentially private algorithms on real data sets.

5.1 Arcene (high-dimensional mass-spectrometry)

Arcene is a high-dimensional benchmark from the NIPS 2003 Feature Selection Challenge, where the goal is to discriminate cancer versus healthy subjects using serum mass-spectrometry profiles (Guyon et al., 2004, 2008). We use the Arcene version available via OpenML (features in $\mathbb{R}^{10,000}$ with binary labels)(Vanschoren et al., 2013; OpenML, 2026).

For each training size n, we randomly sample a test set of size $m_{\text{test}} = 80$ and a training set of size n without replacement, and repeat this procedure over 10 trials with a fixed random seed. Features are standardized using training-set statistics and the same preprocessing is applied to all methods. We evaluate prediction accuracy using NMSE on the test set:

$$\text{NMSE}(\widehat{\beta}) \;=\; \frac{\frac{1}{m_{\text{test}}}\|X_{\text{test}}\widehat{\beta} - y_{\text{test}}\|_2^2}{\text{Var}(y_{\text{test}})}.$$

We compare ISTA (non-private), DP-ISTA (Algorithm 4), DP-IHT (Cai et al., 2021), and LDP-ISTA (Algorithm 3). We fix $(\varepsilon, \delta) = (0.5, 10^{-9})$ and keep algorithmic hyperparameters fixed across methods.

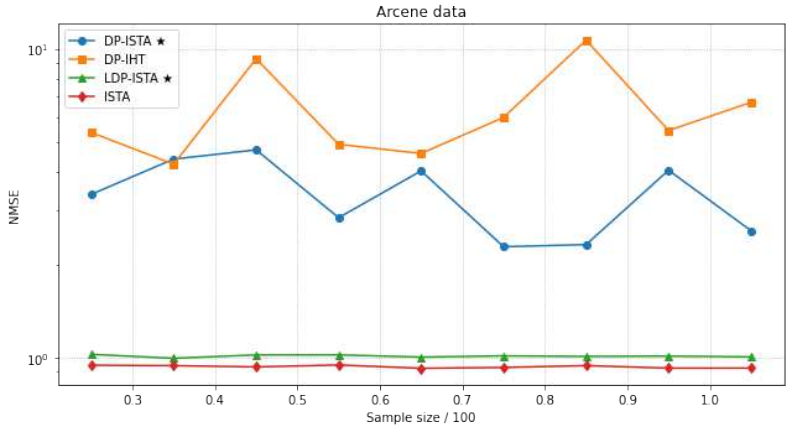


Fig. 3: Arcene (mass-spectrometry). NMSE versus training size n for ISTA, DP-ISTA, DP-IHT, and LDP-ISTA. The horizontal axis is $n/100$.

Figure 3 reports NMSE versus n. Among the private methods, LDP-ISTA achieves the lowest NMSE and remains close to the non-private ISTA curve across the range of n. Under central DP, DP-ISTA yields smaller NMSE than the DP-IHT baseline throughout.

5.2 ACSIncome (label-private regression under LDP)

ACSIncome is a prediction task introduced by Ding et al. (2021a), constructed from the U.S. Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS) (Ding et al., 2021b; U.S. Census Bureau, 2025). The goal is to predict whether an individual's annual income exceeds $50,000.

We use a 2018 ACS snapshot and pool individuals from CA, NY, and TX. Following a standard high-dimensional construction, we discretize numeric covariates and apply feature hashing to obtain a sparse representation with dimension $p \approx 20{,}000$ (fixed hash seed), and map labels $\{0, 1\}$ to $\{-1, +1\}$ for squared-loss training. We form a fixed held-out test set of size 20,000, and for each training size n we subsample (without replacement) from the remaining pool and repeat 5 trials with a fixed random seed. We compare ISTA (non-private), Label-LDP-IHT (Wang and Xu, 2021), and our Label-LDP-ISTA (Algorithm 2). We fix the label-privacy budget $(\varepsilon, \delta) = (0.3, 10^{-9})$ and report NMSE on the test set (defined as above), keeping hyperparameters fixed across methods.

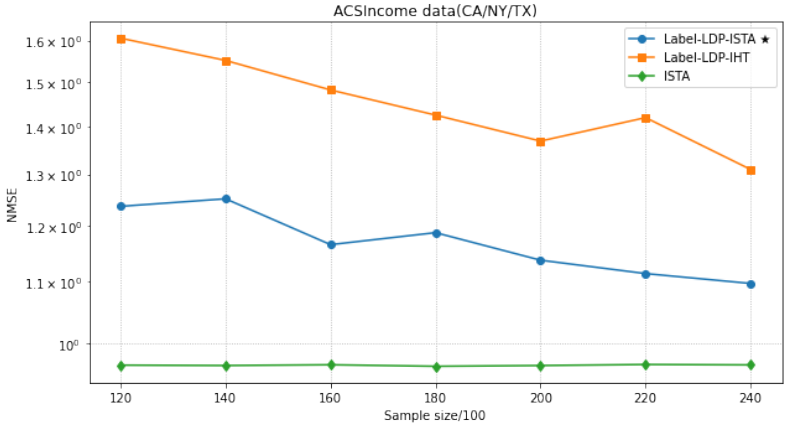


Fig. 4: ACSIncome (CA/NY/TX). NMSE versus training size n for Label-LDP-ISTA, Label-LDP-IHT, and ISTA. The horizontal axis is $n/100$.

Figure 4 shows NMSE versus n. Across all sample sizes considered, Label-LDP-ISTA consistently attains lower NMSE than the Label-LDP-IHT baseline under the same privacy budget.

6 Conclusion

This paper studies high-dimensional sparse linear regression under three common privacy deployments, namely central DP, local DP, and label-side privacy, within a unified proximal first-order (ISTA-type) framework. Across these regimes, we develop concrete privacy mechanisms and stabilized solvers tailored to sparsity. In particular, our non-interactive one-shot perturbation schemes consume privacy only at release time and

therefore avoid privacy-composition-driven iteration-by-iteration noise accumulation, while maintaining algorithmic stability through proximal updates (and ℓ_2 projection where required).

On the theoretical side, we establish explicit finite-sample, high-probability ℓ_2 estimation error bounds under standard RE/RSC/RSS-type conditions. The resulting guarantees recover the familiar high-dimensional scaling $\sqrt{s^* \log p/n}$ in the baseline term, and make transparent how privacy-induced effective noise and finite-iteration optimization error contribute to the total error. This decomposition provides a quantitative characterization of when additional computation improves accuracy and when privacy perturbations become the dominant limitation.

Empirically, we evaluate the proposed methods on synthetic designs and two real-world high-dimensional datasets. Under matched privacy budgets and fixed tuning policies, our methods are competitive with representative DP baselines in the corresponding deployment regimes, and in several settings yield improved accuracy. Moreover, the label-side variant illustrates that when covariates are less sensitive than outcomes, protecting only the supervision signal can substantially reduce the utility loss relative to full record-level privatization.

Taken together, our results clarify practical tradeoffs for sparse regression under privacy constraints. When full record privacy is required, local mechanisms incur the expected dimensional costs, but one-shot perturbation combined with proximal optimization can still produce accurate and stable estimates in high-dimensional regimes. When only labels must be protected, label-side mechanisms can offer markedly better utility than end-to-end privatization while providing formal protection for the supervision signal. We hope the unified analysis and implementations offered here provide a useful foundation for privacy-preserving sparse estimation in modern high-dimensional applications.

6.1 Limitations and future work.

Our guarantees rely on restricted-geometry conditions (RE/RSC/RSS) and boundedness controls (e.g., clipping and projection) that enable principled Gaussian calibration and stability; understanding when these conditions hold and how they can be weakened in specific application domains remains an important direction. In addition, our empirical pipeline fixes preprocessing and hyperparameters for comparability; developing fully end-to-end private model selection and preprocessing with rigorous accounting is complementary to the present focus. Finally, extending the unified proximal analysis to structured sparsity and generalized linear models under comparable privacy models is a natural avenue for future work.

References

Abadi M, Chu A, Goodfellow I, et al (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS 2016). ACM, pp 308–318, https://doi.org/10.1145/2976749.2978318, URL https://doi.org/10.1145/2976749.2978318

Apple Differential Privacy Team (2017) Learning with privacy at scale. Technical report (white paper), URL https://docs-assets.developer.apple.com/ml-research/papers/learning-with-privacy-at-scale.pdf, accessed 2026-02-02

Balle B, Wang YX (2018) Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 80. PMLR, pp 394–403, URL https://proceedings.mlr.press/v80/balle18a.html

Bassily R, Smith A, Thakurta A (2014) Private empirical risk minimization: Efficient algorithms and tight error bounds. In: 2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS). IEEE, pp 464–473, https://doi.org/10.1109/FOCS.2014.56, URL https://doi.org/10.1109/FOCS.2014.56

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences 2(1):183–202. https://doi.org/10.1137/080716542, URL https://epubs.siam.org/doi/10.1137/080716542

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of Lasso and Dantzig selector. The Annals of Statistics 37(4):1705–1732. https://doi.org/10.1214/08-AOS620

Blumensath T, Davies ME (2009) Iterative hard thresholding for compressed sensing. Applied and Computational Harmonic Analysis 27(3):265–274. https://doi.org/10.1016/j.acha.2009.04.002

Bühlmann P, van de Geer S (2011) Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Series in Statistics, Springer, https://doi.org/10.1007/978-3-642-20192-9

Cai TT, Wang Y, Zhang L (2020) The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. arXiv preprint arXiv:201103900 URL https://arxiv.org/abs/2011.03900

Cai TT, Wang Y, Zhang L (2021) The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. The Annals of Statistics 49(5):2825–2850. https://doi.org/10.1214/21-AOS2058

Chaudhuri K, Monteleoni C, Sarwate AD (2011) Differentially private empirical risk minimization. Journal of Machine Learning Research 12:1069–1109. URL https://www.jmlr.org/papers/v12/chaudhuri11a.html

Daubechies I, Defrise M, Mol CD (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics 57(11):1413–1457. https://doi.org/10.1002/cpa.20042

Ding F, Hardt M, Miller J, et al (2021a) Retiring adult: New datasets for fair machine learning. In: Advances in Neural Information Processing Systems (NeurIPS), URL https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html

Ding F, Hardt M, Miller J, et al (2021b) Folktables: Datasets derived from us census. GitHub repository, URL https://github.com/socialfoundations/folktables

Duchi JC, Jordan MI, Wainwright MJ (2018) Minimax optimal procedures for locally private estimation. Journal of the American Statistical Association 113(521):182–201. https://doi.org/10.1080/01621459.2017.1389735, URL https://

doi.org/10.1080/01621459.2017.1389735
Dwork C, Roth A (2014) The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science 9(3-4):211–407. https://doi.org/10.1561/0400000042
Dwork C, McSherry F, Nissim K, et al (2006) Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography (TCC 2006), Lecture Notes in Computer Science, vol 3876. Springer, pp 265–284, https://doi.org/10.1007/11681878_14
Efron B, Hastie T, Johnstone I, et al (2004) Least angle regression. The Annals of Statistics 32(2). https://doi.org/10.1214/009053604000000067
Esfandiari H, Mirrokni V, Syed U, et al (2022) Label differential privacy via clustering. In: Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, pp 7055–7075, URL https://proceedings.mlr.press/v151/esfandiari22a.html
Ghazi B, Golowich N, Kumar R, et al (2021) Deep learning with label differential privacy. In: Advances in Neural Information Processing Systems, URL https://proceedings.neurips.cc/paper/2021/hash/089d3eaf7f140f34f5e32e2dd01161e5-Abstract.html
Guyon I, Gunn SR, Ben-Hur A, et al (2004) Result analysis of the NIPS 2003 feature selection challenge. In: Advances in Neural Information Processing Systems 17, pp 545–552, URL https://papers.nips.cc/paper_files/paper/2004/hash/5e751896e527c862bf67251a474b3819-Abstract.html
Guyon I, Gunn S, Ben-Hur A, et al (2008) Arcene. UCI Machine Learning Repository, https://doi.org/10.24432/C58P55, URL https://archive.ics.uci.edu/dataset/167/arcene, nIPS 2003 feature selection challenge dataset; accessed 2026-02-02
Joseph M, Mao J, Neel S, et al (2019) The role of interactivity in local differential privacy. In: 2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS). IEEE Computer Society, pp 94–105, https://doi.org/10.1109/FOCS.2019.00015
Kasiviswanathan SP, Lee HK, Nissim K, et al (2011) What can we learn privately? SIAM Journal on Computing 40(3):793–826. https://doi.org/10.1137/090756090, URL https://doi.org/10.1137/090756090
Kifer D, Smith A, Thakurta A (2012) Private convex empirical risk minimization and high-dimensional regression. In: Mannor S, Srebro N, Williamson RC (eds) Proceedings of the 25th Annual Conference on Learning Theory, Proceedings of Machine Learning Research, vol 23. PMLR, Edinburgh, Scotland, pp 25.1–25.40, URL https://proceedings.mlr.press/v23/kifer12.html
Laurent B, Massart P (2000) Adaptive estimation of a quadratic functional by model selection. The Annals of Statistics 28(5):1302–1338. https://doi.org/10.1214/aos/1015957395
Negahban SN, Ravikumar P, Wainwright MJ, et al (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statistical Science 27(4):538–557. https://doi.org/10.1214/12-STS400
OpenML (2026) Arcene (openml dataset id 1458). OpenML dataset page, URL https://www.openml.org/d/1458, accessed 2026-02-02

Papernot N, Abadi M, Erlingsson Ú, et al (2017) Semi-supervised knowledge transfer for deep learning from private training data. In: International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=HkwoSDPgg

Papernot N, Song S, Mironov I, et al (2018) Scalable private learning with PATE. In: International Conference on Learning Representations (ICLR), URL https://openreview.net/forum?id=rkZB1XbRZ

Sheffet O (2017) Differentially private ordinary least squares. In: International Conference on Machine Learning (ICML). PMLR, pp 3105–3114, URL https://proceedings.mlr.press/v70/sheffet17a.html

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

U.S. Census Bureau (2025) Public use microdata sample (PUMS). U.S. Census Bureau, URL https://www.census.gov/programs-surveys/acs/microdata.html, page updated 2025-12-15; accessed 2026-02-02

Vanschoren J, van Rijn JN, Bischl B, et al (2013) OpenML: Networked science in machine learning. SIGKDD Explorations 15(2):49–60. https://doi.org/10.1145/2641190.2641198

Vershynin R (2018) High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge University Press, https://doi.org/10.1017/9781108231596

Wainwright MJ (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Transactions on Information Theory 55(5):2183–2202. https://doi.org/10.1109/TIT.2009.2016018

Wang D, Xu J (2021) On sparse linear regression in the local differential privacy model. IEEE Transactions on Information Theory 67(2):1182–1200. https://doi.org/10.1109/TIT.2020.3040406

Wang D, Zhang H, Gaboardi M, et al (2021) Estimating smooth GLM in non-interactive local differential privacy model with public unlabeled data. In: Proceedings of the 32nd International Conference on Algorithmic Learning Theory, pp 1207–1213, URL https://proceedings.mlr.press/v132/wang21f.html, earlier version: arXiv:1910.00482

Warner SL (1965) Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60(309):63–69. https://doi.org/10.1080/01621459.1965.10480775

Wu TT, Lange K (2008) Coordinate descent algorithms for Lasso penalized regression. The Annals of Applied Statistics 2(1). https://doi.org/10.1214/07-aoas147