Nearly-Optimal Private LASSO[*]

Kunal Talwar
Google Research
kunal@google.com

Abhradeep Thakurta
(Previously) Yahoo! Labs
guhathakurta.abhradeep@gmail.com

Li Zhang
Google Research
liqzhang@google.com

Abstract

We present a nearly optimal differentially private version of the well known LASSO estimator. Our algorithm provides privacy protection with respect to each training example. The excess risk of our algorithm, compared to the non-private version, is $\widetilde{O}(1/n^{2/3})$, assuming all the input data has bounded ℓ_∞ norm. This is the first differentially private algorithm that achieves such a bound without the polynomial dependence on p under no additional assumptions on the design matrix. In addition, we show that this error bound is nearly optimal amongst all differentially private algorithms.

1 Introduction

A common task in supervised learning is to select the model that best fits the data. This is frequently achieved by selecting a *loss function* that associates a real-valued loss with each datapoint d and model θ and then selecting from a class of admissible models, the model θ that minimizes the average loss over all data points in the training set. This procedure is commonly referred to as *Empirical Risk Minimization*(ERM).

The availability of large datasets containing sensitive information from individuals has motivated the study of learning algorithms that guarantee the privacy of individuals contributing to the database. A rigorous and by-now standard privacy guarantee is via the notion of differential privacy. In this work, we study the design of differentially private algorithms for Empirical Risk Minimization, continuing a long line of work. (See [2] for a survey.)

In particular, we study adding privacy protection to the classical LASSO estimator, which has been widely used and analyzed. We first present a differentially private optimization algorithm for the LASSO estimator. The algorithm is the combination of the classical Frank-Wolfe algorithm [15] and the exponential mechanism for guaranteeing the privacy [21]. We then show that our algorithm achieves nearly optimal risk among all the differentially private algorithms. This lower bound proof relies on recently developed techniques with roots in Cryptography [4, 14],

Consider the training dataset D consisting of n pairs of data $d_i = (x_i, y_i)$ where $x_i \in \mathbb{R}^p$, usually called the feature vector, and $y_i \in \mathbb{R}$, the prediction. The LASSO estimator, or the sparse linear regression, solves for $\theta^* = \mathsf{argmin}_\theta\, \mathcal{L}(\theta; d_i) = \frac{1}{n}\sum_i |x_i \cdot \theta - y_i|^2$ subject to $\|\theta\|_1 \le c$. To simplify presentation, we assume $c = 1$, but our results directly extend to general c. The ℓ_1 constraint tends to induce sparse θ^* so is widely used in the high dimensional setting when $p \gg n$. Here, we will study approximating the LASSO estimation with minimum possible error while protecting the privacy of each individual d_i. Below we define the setting more formally.

[*] Part of this work was done at Microsoft Research Silicon Valley Campus.

近似最优私有LASSO[*]

库纳尔·塔尔瓦谷歌研
究 kunal@google.com

阿布哈德·塔库特拉(此前) 雅虎实验室
guhathakurta.abhradeep@gmail.com

张丽谷歌研究
liqzhang@google.com

摘要

我们提出了一种近似最优的差分隐私版本来改进众所周知的LASSO估计器。我们的算法针对每个训练样本提供隐私保护。与非私有版本相比，我们的算法的超额风险为 $\widetilde{O}(1/n^{2/3})$，假设所有输入数据具有有界的 ℓ_∞ 范数。这是第一个在无设计矩阵额外假设的情况下，不依赖于 p 的多项式依赖而实现此类界限的差分隐私算法。此外，我们证明了该误差界在所有差分隐私算法中近似最优。

1 引言

在监督学习中，一个常见的任务是选择最适合数据的模型。这通常通过选择一个损失函数 来实现，该函数将一个实值损失与每个数据点 d 和模型 θ 相关联，然后从可接受模型类中选择平均损失在训练集中所有数据点上最小的模型 θ 。该过程通常被称为经验风险最小化(ERM)。

大型数据集的可用性，其中包含个人的敏感信息，推动了保证数据库中贡献者隐私的学习算法的研究。一种严格且现已标准化的隐私保证是通过差分隐私的概念实现的。在这项工作中，我们研究了用于经验风险最小化的差分隐私算法的设计，延续了长期的研究工作。(参见 [2] 的综述。)

特别是，我们研究了为经典LASSO估计器添加隐私保护，该估计器已被广泛使用和分析。我们首先提出了一种用于LASSO估计器的差分隐私优化算法。该算法是经典Frank-Wolfe算法 [15] 和用于保证隐私 [21]的指数机制的组合。然后，我们表明我们的算法在所有差分隐私算法中实现了接近最优的风险。该下界证明依赖于最近开发的技术，这些技术源于密码学 [4, 14],

考虑训练数据集 D，它由 n 对数据 $d_i = (x_i, y_i)$ 组成，其中 $x_i \in \mathbb{R}^p$ 通常称为特征向量，而 $y_i \in \mathbb{R}$ 是预测。LASSO 估计器，或稀疏线性回归，求解 $\theta^* =$ 最小化 $_\theta\, \mathcal{L}(\theta; d_i) = \frac{1}{n}\sum_i |x_i \cdot \theta - y_i|^2$，约束条件为 $\|\theta\|_1 \le c.$。为简化表述，我们假设 $c = 1$，但我们的结果可直接推广到一般的 c。ℓ_1 约束倾向于诱导稀疏 θ^*，因此在高维设置中当 $p \gg n$ 时被广泛使用。在这里，我们将研究在保护每个个体的隐私 d_i 的同时，以最小可能的误差逼近 LASSO 估计。

[*] 部分工作在微软研究院硅谷校区完成。

Problem definition: Given a data set $D = \{d_1, \cdots, d_n\}$ of n samples from a domain $\mathcal{D}$, a constraint set $\mathcal{C} \subseteq \mathbb{R}^p$, and a loss function $\mathcal{L} : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$, for any model θ, define its excess empirical risk as

$$R(\theta; D) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; d_i) - \min_{\theta \in \mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; d_i). \tag{1}$$

For LASSO, the constraint set is the ℓ_1 ball, and the loss is the quadratic loss function. We define the *risk* of a mechanism $\mathcal{A}$ on a data set D as $R(\mathcal{A}; D) = \mathbb{E}[R(\mathcal{A}(D); D)]$, where the expectation is over the internal randomness of $\mathcal{A}$, and the risk $R(\mathcal{A}) = \max_{D \in \mathcal{D}^n} R(\mathcal{A}; D)$ is the maximum risk over all the possible data sets. Our objective is then to design a mechanism $\mathcal{A}$ which preserves (ϵ, δ)-differential privacy (Definition 1.3) and achieves as low risk as possible. We call the minimum achievable risk as *privacy risk*, defined as $\min_{\mathcal{A}} R(\mathcal{A})$, where the min is over all (ϵ, δ)-differentially private mechanisms $\mathcal{A}$.

There has been much work on studying the privacy risk for the LASSO estimator. However, all the previous results either need to make strong assumption about the input data or have polynomial dependence on the dimension p. First [20] and then [24] studied the LASSO estimator with differential privacy guarantee. They showed that one can avoid the polynomial dependence on p in the excess empirical risk if the data matrix X satisfy the restricted strong convexity and mutual incoherence properities. While such assumptions seem necessary to prove that LASSO recovers the exact support in the worst case, they are often violated in practice, where LASSO still leads to useful models. It is therefore desirable to design and analyze private versions of LASSO in the absence of such assumptions. In this work, we do so by analyzing the loss achieved by the private optimizer, compared to the true optimizer.

We make primarily two contributions in this paper. First we present an algorithm that achieves the privacy risk of $\widetilde{O}(1/n^{2/3})$ for the LASSO problem[1]. Compared to the previous work, we only assume that the input data has bounded ℓ_∞ norm. In addition, the above risk bound only has logarithmic dependence on p, which fits particularly well for LASSO as we usually assume $n \ll p$ when applying LASSO. This bound is achieved by a private version of the Frank-Wolfe algorithm. Assuming that each data point d_i satisfies that $\|d_i\|_\infty \leq 1$, we have

Theorem 1.1. *There exists an (ϵ, δ)-differentially private algorithm $\mathcal{A}$ for LASSO such that*

$$R(\mathcal{A}) = O\left(\frac{\log(np)\sqrt{\log(1/\delta)}}{(n\epsilon)^{2/3}}\right).$$

Our second contribution is to show that, surprisingly, this simple algorithm gives a nearly tight bound. We show that this rather unusual $n^{-2/3}$ dependence is not an artifact of the algorithm or the analysis, but is in fact the right dependence for the LASSO problem: no differentially private algorithm can do better! We prove a lower bound by employing fingerprinting codes based techniques developed in [4, 14].

Theorem 1.2. *For the sparse linear regression problem where $\|x_i\|_\infty \leq 1$, for $\epsilon = 0.1$ and $\delta = o(1/n^2)$, any (ϵ, δ)-differentially private algorithm $\mathcal{A}$ must have*

$$R(\mathcal{A}) = \Omega(1/(n \log n)^{2/3}).$$

Our improved privacy risk crucially depends on the fact that the constraint set is a polytope with few (polynomial in dimensions) vertices. This allows us to use a private version of the Frank-Wolfe algorithm, where at each step, we use the exponential mechanism to select one of the vertices of the polytope. We also present a variant of Frank-Wolfe that uses objective perturbation instead of the exponential mechanism. We show that (Theorem 2.6) we can obtain a risk bound dependent on the *Gaussian width* of the constraint set, which often results in tighter bounds compared to bounds based, e.g., on diameter. While more general, this variant adds much more noise than the Frank-Wolfe based algorithm, as it is effectively publishing the whole gradient at each step. When $\mathcal{C}$ is not a polytope with a small number of vertices, one can still use the exponential mechanism as long as one has a small list of candidate points which contains an approximate optimizer for every direction. For many simple cases, for example the ℓ_q ball with $1 < q < 2$, the bounds attained in this way have

[1] Throughout the paper, we use $\widetilde{O}$ to hide logarithmic factors.

问题定义：给定一个包含 $D = \{d_1, \cdots, d_n\}$ 个来自域 $\mathcal{D}$的 n 个样本的数据集、一个约束集 $\mathcal{C} \subseteq \mathbb{R}^p$和一个损失函数 $\mathcal{L} : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$，对于任何模型 θ，定义其超额经验风险为

$$R(\theta; D) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; d_i) - \min_{\theta \in \mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; d_i). \tag{1}$$

对于LASSO，约束集是 ℓ_1 球，损失是二次损失函数。我们定义机制 $\mathcal{A}$ 在数据集 D 上的风险为 $R(\mathcal{A}; D) = \mathbb{E}[R(\mathcal{A}(D); D)]$，其中期望值是关于 $\mathcal{A}$的内部随机性，风险 $R(\mathcal{A}) = \max_{D \in \mathcal{D}^n} R(\mathcal{A}; D)$ 是所有可能数据集上的最大风险。我们的目标是为机制 $\mathcal{A}$ 设计一个机制，它保留(ϵ, δ)-差分隐私（定义 1.3），并尽可能降低风险。我们将最小可达到的风险称为隐私风险，定义为$\min_{\mathcal{A}} R(\mathcal{A})$，其中min是所有$(\epsilon, \delta)$-差分隐私机制 $\mathcal{A}$上的最小值。

在研究LASSO估计器的隐私风险方面已经做了大量工作。然而，所有先前结果要么需要对输入数据做出强假设，要么在维度 p上具有多项式依赖性。首先 [20] 然后 [24] 研究了具有差分隐私保证的LASSO估计器。他们表明，如果数据矩阵 X 满足限制强凸性和相互不相关性特性，则可以避免超额经验风险在 p 上的多项式依赖性。虽然这些假设似乎在证明LASSO在最坏情况下恢复精确支撑时是必要的，但它们在实践中往往被违反，此时LASSO仍然会导致有用的模型。因此，在没有这些假设的情况下设计和分析LASSO的隐私版本是理想的。在这项工作中，我们通过分析私有优化器与真实优化器相比所实现的损失来完成这一目标。

本文主要做出两项贡献。首先我们提出了一种算法，该算法在LASSO问题上实现了 $\widetilde{O}(1/n^{2/3})$ 的隐私风险1。与先前的工作相比，我们仅假设输入数据具有有界的 ℓ_∞ 范数。此外，上述风险界限仅对 p具有对数依赖性，这特别适合LASSO，因为我们通常在应用LASSO时假设 $n \ll p$。该界限是通过Frank-Wolfe算法的隐私版本实现的。假设每个数据点 d_i 满足 $\|d_i\|_\infty \leq 1$，我们有

Theorem 1.1. *There exists an (ϵ, δ)-differentially private algorithm $\mathcal{A}$ for LASSO such that*

$$R(\mathcal{A}) = O\left(\frac{\log(np)\sqrt{\log(1/\delta)}}{(n\epsilon)^{2/3}}\right).$$

我们的第二个贡献是表明，出乎意料的是，这个简单算法给出了一个近似紧的界限。我们证明了这个相当不寻常的 $n^{-2/3}$ 依赖性不是算法或分析的产物，而是LASSO问题的正确依赖性：没有差分隐私算法能做得更好！我们通过采用在 [4, 14]中开发基于指纹码的技术来证明下界。

定理 **1 2** $\|x_i\|_\infty \leq 1$ $\epsilon = 0.1$ $\delta =$ .. 对于稀疏线性回归问题，其中，对于和 (),
o $1/n^2$ ϵ δ $\mathcal{A}$
任何 (,)-差分隐私算法必须满足

$$R(\mathcal{A}) = \Omega(1/(n \log n)^{2/3}).$$

我们改进的隐私风险关键地取决于约束集是一个具有少量（维度的多项式）顶点的多面体这一事实。这使我们能够使用Frank-Wolfe算法的隐私版本，其中每一步，我们使用指数机制来选择多面体的一个顶点。我们还提出了一种使用目标扰动的Frank-Wolfe算法变体，而不是指数机制。我们证明，(定理 2.6)我们可以获得一个依赖于约束集的 高斯宽度 的风险界限，这通常与基于直径的界限相比更紧。虽然更通用，但这种变体比基于Frank-Wolfe的算法增加了更多的噪声，因为它实际上在每一步都发布整个梯度。当 $\mathcal{C}$ 不是一个具有少量顶点的多面体时，只要有一个包含每个方向的大致优化器的候选点的小列表，仍然可以使用指数机制。对于许多简单情况，例如具有 $1 < q < 2$的 ℓ_q 球，通过这种方式达到的界限有

1全文中，我们使用 $\widetilde{O}$ 来隐藏对数因子。

an additional polynomial dependence on the dimension p, instead of the logarithmic dependence in the above result. For example, when $q = 1$, the upper bound from this variant has an extra factor of $p^{1/3}$. Whereas such a dependence is provably needed for $q = 2$, the upper bound jump rather abruptly from the logarithmic dependence for $q = 1$ to a polynomial dependence on p for $q > 1$. We leave open the question of resolving this discontinuity and interpolating more smoothly between the ℓ_1 case and the ℓ_2 case.

Our results enlarge the set of problems for which privacy comes "for free". Given n samples from a distribution, suppose that θ^* is the empirical risk minimizer and θ^{priv} is the differentially private approximate minimizer. Then the non-private ERM algorithm outputs θ^* and incurs expected (on the distribution) loss equal to the $\mathsf{loss}(\theta^*, \text{training-set}) +$ generalization-error, where the *generalization error* term depends on the loss function, $\mathcal{C}$ and on the number of samples n. The differentially private algorithm incurs an additional loss of the privacy risk. If the privacy risk is asymptotically no larger than the generalization error, we can think of privacy as coming for free, since under the assumption of n being large enough to make the generalization error small, we are also making n large enough to make the privacy risk small. In the case when $\mathcal{C}$ is the ℓ_1-ball, and the loss function is the squared loss with $\|x\|_\infty \leq 1$ and $|y| \leq 1$, the best known generalization error bounds dominate the privacy risk when $n = \omega(\log^3 p)$ [1, Theorem 18].

1.1 Related work

There have been much work on private LASSO or more generally private ERM algorithms. The error bounds mainly depend on the shape of the constraint set and the Lipschitz condition of the loss function. Here we will summarize these related results. Related to our results, we distinguish two settings: i) the constraint set is bounded in the ℓ_1-norm and the the loss function is 1-Lipschitz in the ℓ_1-norm. (call it the (ℓ_1/ℓ_∞)-setting). This is directly related to our bounds on LASSO; and ii) the constraint set has bounded ℓ_2 norm and the loss function is 1-Lipschitz in the ℓ_2 norm (the (ℓ_2/ℓ_2)-setting), which is related to our bounds using Gaussian width.

The (ℓ_1/ℓ_∞)***-setting:*** The results in this setting include [20, 24, 19, 25]. The first two works make certain assumptions about the instance (*restricted strong convexity* (RSC) and *mutual incoherence*). Under these assumptions, they obtain privacy risk guarantees that depend logarithmically in the dimensions p, and thus allowing the guarantees to be meaningful even when $p \gg n$. In fact their bound of $O(\text{polylog}\, p/n)$ can be better than our *tight* bound of $O(\text{polylog}\, p/n^{2/3})$. However, these assumptions on the data are strong and may not hold in practice. Our guarantees do not require any such data dependent assumptions. The result of [19] captures the scenario when the constraint set $\mathcal{C}$ is the probability simplex and the loss function is a generalized linear model, but provides a *worse* bound of $O(\text{polylog}\, p/n^{1/3})$. For the special case of *linear loss functions*, which are interesting primarily in the online prediction setting, the techniques of [19, 25] provide a bound of $O(\text{polylog}\, p/n)$.

The (ℓ_2/ℓ_2)***-setting:*** In all the works on private convex optimization that we are aware of, either the excess risk guarantees depend polynomially on the dimensionality of the problem (p), or assumes special structure to the loss (e.g., generalized linear model [19] or linear losses [25]). Similar dependence is also present in the online version of the problem [18, 26]. [2] recently show that in the private ERM setting, in general this polynomial dependence on p is unavoidable. In our work we show that one can replace this dependence on p with the Gaussian width of the constraint set $\mathcal{C}$, which can be much smaller.

Effect of Gaussian width in risk minimization: Our result on general $\mathcal{C}$ has an dependence on the Gaussian width of $\mathcal{C}$. This geometric concept has previously appeared in other contexts. For example, [1] bounds the the excess generalization error by the Gaussian width of the constraint set $\mathcal{C}$. Recently [5] show that the Gaussian width of a constraint set $\mathcal{C}$ is very closely related to the number of generic linear measurements one needs to perform to recover an underlying model $\theta^* \in \mathcal{C}$. The notion of Gaussian width has also been used by [22, 11] in the context of differentially private query release mechanisms but in the very different context of answering multiple linear queries over a database.

在维度 p 上存在额外的多项式依赖，而不是上述结果中的对数依赖。例如，当 $q = 1$ 时，此变体的上界有一个额外的因子 $p^{1/3}$。而此类依赖被证明对于 $q = 2$ 是必需的，但上界从 $q = 1$ 的对数依赖突然跳跃到 $q > 1$ 的 p 多项式依赖。我们留下解决此不连续性的问题，并在 ℓ_1 情况和 ℓ_2 情况之间更平滑地插值。

我们的结果扩大了隐私“免费”出现的问题集合。给定来自分布的 n 个样本，假设 θ^* 是经验风险最小化器， θ^{priv} 是差分隐私近似最小化器。那么非隐私ERM算法输出 θ^* 并产生等于损失(θ^*,训练集) $+$ 泛化误差的预期（在分布上）损失，其中 泛化误差 项取决于损失函数、$\mathcal{C}$ 和样本数量 n。差分隐私算法产生额外的隐私风险损失。如果隐私风险渐近不大于泛化误差，我们可以认为隐私是免费的，因为假设 n 足够大以使泛化误差很小，我们也使 n 足够大以使隐私风险很小。在 $\mathcal{C}$ 是 ℓ_1-球且损失函数是具有 $\|x\|_\infty \leq 1$ 和 $|y| \leq 1$ 的平方损失的情况下，当 $n = \omega(\log^3 p)$ [1，定理 18] 时，已知的最佳泛化误差界限支配隐私风险。

1.1 相关工作

在隐私LASSO或更一般地隐私ERM算法方面已有大量研究。误差界限主要取决于约束集的形状和损失函数的Lipschitz条件。在这里我们将总结这些相关结果。与我们的结果相关，我们区分两种设置：i) 约束集在 ℓ_1-范数下有界，损失函数在 ℓ_1-范数下是1-Lipschitz的（称为(ℓ_1/ℓ_∞)-设置）。这与我们对LASSO的界限直接相关；以及ii) 约束集有有界的 ℓ_2 范数，损失函数在 ℓ_2 -范数下是1-Lipschitz的（(ℓ_2/ℓ_2)-设置），这与我们使用高斯宽度的界限相关。

The (ℓ_1/ℓ_∞)***-setting:*** 该设置的结果包括 [20, 24, 19, 25]。前两项工作对实例做出了一些假设（限制强凸性 (RSC)和相互不相关性）。在这些假设下，它们获得了依赖于维度 p的对数隐私风险保证，因此即使当 $p \gg n$时也能保证其意义。实际上它们的 O(多项式对数 p/n)界限可以比我们的紧 界限 $O(\text{polylog}\, p/n^{2/3})$更好。然而，这些对数据的假设很强，在实践中可能不成立。我们的保证不需要任何这样的数据相关假设。 [19]的结果捕捉了约束集 $\mathcal{C}$ 是概率单纯形且损失函数是广义线性模型的情况，但提供了一个更差 的界限 $O(\text{polylog}\, p/n^{1/3})$。对于线性损失函数的特殊情况，它们主要在在线预测设置中感兴趣， [19, 25]的技术提供了一个界限O(多项式对数 p/n)。

The (ℓ_2/ℓ_2)***-setting:*** 在所有我们了解的关于私有凸优化的工作中，要么超额风险保证依赖于问题的维度(p)，要么假设损失的特定结构（例如，广义线性模型 [19] 或线性损失 [25]）。问题的在线版本 [18, 26]中也存在类似的依赖性。 [2] 最近表明，在私有ERM设置中，通常这种对 p 的多项式依赖是无法避免的。在我们的工作中，我们表明可以用约束集 $\mathcal{C}$的高斯宽度来代替对 p 的依赖，这可以小得多。

高斯宽度在风险最小化中的影响： 我们对一般 $\mathcal{C}$ 的结果依赖于约束集 $\mathcal{C}$的高斯宽度。这个几何概念之前在其他上下文中出现过。例如， [1] 通过约束集 $\mathcal{C}$的高斯宽度来限制超额泛化误差。最近 [5] 表明，约束集 $\mathcal{C}$ 的高斯宽度与为了恢复底层模型 $\theta^* \in \mathcal{C}$所需执行的通用线性测量的数量非常密切相关。高斯宽度的概念也被 [22, 11] 在差分隐私查询释放机制的上下文中使用，但在回答数据库上多个线性查询的非常不同的上下文中。

1.2 Background

Differential Privacy: The notion of differential privacy (Definition 1.3) is by now a defacto standard for statistical data privacy [10, 12]. One of the reasons why differential privacy has become so popular is because it provides meaningful guarantees even in the presence of arbitrary auxiliary information. At a semantic level, the privacy guarantee ensures that *an adversary learns almost the same thing about an individual independent of his presence or absence in the data set.* The parameters (ϵ, δ) quantify the amount of information leakage. For reasons beyond the scope of this work, $\epsilon \approx 0.1$ and $\delta = 1/n^{\omega(1)}$ are a good choice of parameters. Here n refers to the number of samples in the data set.

Definition 1.3. *A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if, for all neighboring data sets $\mathcal{D}$ and $\mathcal{D}'$ (i.e., they differ in one record, or equivalently, $d_H(D, D') = 1$) and for all events S in the output space of $\mathcal{A}$, we have*

$$\Pr(\mathcal{A}(\mathcal{D}) \in S) \leq e^{\epsilon} \Pr(\mathcal{A}(\mathcal{D}') \in S) + \delta\,.$$

Here $d_H(D, D')$ refers to the Hamming distance.

ℓ_q-norm, $q \geq 1$: For $q \geq 1$, the ℓ_q-norm for any vector $v \in \mathbb{R}^p$ is defined as $\left(\sum_{i=1}^{p} v(i)^q\right)^{1/q}$, where $v(i)$ is the i-th coordinate of the vector v.

L-Lipschitz continuity w.r.t. norm $\|\cdot\|$: A function $\Psi : \mathcal{C} \to \mathbb{R}$ is L-Lispchitz within a set $\mathcal{C}$ w.r.t. a norm $\|\cdot\|$ if the following holds.

$$\forall \theta_1, \theta_2 \in \mathcal{C}, |\Psi(\theta_1) - \Psi(\theta_2)| \leq L \cdot \|\theta_1 - \theta_2\|.$$

Gaussian width of a set $\mathcal{C}$: Let $b \sim \mathcal{N}(0, \mathbb{I}_p)$ be a Gaussian random vector in $\mathbb{R}^p$. The Gaussian width of a set $\mathcal{C}$ is defined as $G_{\mathcal{C}} \overset{def}{=} \mathbb{E}_b \left[\sup_{w \in \mathcal{C}} |\langle b, w\rangle|\right]$.

2 Private Convex Optimization by Frank-Wolfe algorithm

In this section we analyze a differentially private variant of the classical Frank-Wolfe algorithm [15]. We show that for the setting where the constraint set $\mathcal{C}$ is a polytope with k vertices, and the loss function $\mathcal{L}(\theta; d)$ is Lipschitz w.r.t. the ℓ_1-norm, one can obtain an excess privacy risk of roughly $O(\log k/n^{2/3})$. This in particular captures the high-dimensional linear regression setting. One such example is the classical LASSO algorithm[27], which computes $\text{argmin}_{\theta:\|\theta\|_1 \leq 1} \frac{1}{n}\|X\theta - y\|_2^2$. In the usual case of $|x_{ij}|, |y_j| = O(1)$, $\mathcal{L}(\theta) = \frac{1}{n}\|X\theta - y\|_2^2$ is $O(1)$-Lipschitz with respect to ℓ_1-norm, we show that one can achieve the nearly optimal privacy risk of $\widetilde{O}(1/n^{2/3})$.

The Frank-Wolfe algorithm [15] can be regarded as a "greedy" algorithm which moves towards the optimum solution in the first order approximation (see Algorithm 1 for the description). How fast Frank-Wolfe algorithm converges depends on $\mathcal{L}$'s "curvature", defined as follows according to [8, 17]. We remark that a β-smooth function on $\mathcal{C}$ has curvature constant bounded by $\beta\|C\|^2$.

Definition 2.1 (Curvature constant)**.** *For $\mathcal{L} : \mathcal{C} \to \mathbb{R}$, define $\Gamma_{\mathcal{L}}$ as below.*

$$\Gamma_{\mathcal{L}} := \sup_{\theta_1, \theta_2, \in \mathcal{C}, \gamma \in (0,1], \theta_3 = \theta_1 + \gamma(\theta_2 - \theta_1)} \frac{2}{\gamma^2} \left(\mathcal{L}(\theta_3) - \mathcal{L}(\theta_1) - \langle \theta_3 - \theta_1, \bigtriangledown\mathcal{L}(\theta_1)\rangle\right).$$

Remark 1. A useful bound can be derived for a quadratic loss $\mathcal{L}(\theta) = \theta A^T A \theta + \langle b, \theta\rangle$. In this case, by [8], $\Gamma_{\mathcal{L}} \leq \max_{a,b \in A \cdot \mathcal{C}} \|a - b\|_2^2$. When $\mathcal{C}$ is centrally symmetric, we have the bound $\Gamma_{\mathcal{L}} \leq 4 \max_{\theta \in \mathcal{C}} \|A\theta\|_2^2$. For LASSO, $A = \frac{1}{\sqrt{n}} X$.

Define $\theta^* = \underset{\theta \in \mathcal{C}}{\text{argmin}}\, \mathcal{L}(\theta)$. The following theorem bounds the convergence rate of Frank-Wolfe algorithm.

1.2 背景

差分隐私： 差分隐私（定义 1.3）的概念目前已成为统计数据隐私的准标准 [10, 12]。差分隐私之所以如此流行，是因为它即使在存在任意辅助信息的情况下也能提供有意义的保证。在语义层面，隐私保证确保攻击者几乎以与个体是否存在于数据集中无关的方式学习到关于该个体的信息。 参数（ϵ, δ）量化了信息泄露的量。由于超出本工作范围的原因，$\epsilon \approx 0.1$ 和 $\delta = 1/n^{\omega(1)}$ 是参数的良好选择。这里的 n 指的是数据集中的样本数量。

定义 1.3。 随机算法 $\mathcal{A}$ 是 (ϵ, δ)-差分隐私的，如果对于所有相邻的数据集 $\mathcal{D}$和 $\mathcal{D}'$（即它们在一个记录中不同，或者等价地， $d_H(D, D') = 1)$ 并且对于所有事件 S在 $\mathcal{A}$ 的输出空间中，我们有

$$\Pr(\mathcal{A}(\mathcal{D}) \in S) \leq e^{\epsilon} \Pr(\mathcal{A}(\mathcal{D}') \in S) + \delta\,.$$

这里 $d_H(D, D')$ 指的是汉明距离。

ℓ_q**-范数**， $q \geq 1$： 对于 $q \geq 1$，任何向量 $v \in \mathbb{R}^p$ 的 ℓ_q-范数定义为 $\left(\sum_{i=1}^{p} v(i)^q\right)^{1/q}$， 其中 $v(i)$

是向量 v 的第 i-个坐标。

L-关于范数的 ***Lipschitz*** 连续性 $\|\cdot\|$： 一个函数 $\Psi : \mathcal{C} \to \mathbb{R}$ 在集合 $\mathcal{C}$ 上关于范数 $\|\cdot\|$ 是 L-Lispchitz，如果满足以下条件。

$$\forall \theta_1, \theta_2 \in \mathcal{C}, |\Psi(\theta_1) - \Psi(\theta_2)| \leq L \cdot \|\theta_1 - \theta_2\|.$$

集合的高斯宽度 $\mathcal{C}$： 设 $b \sim \mathcal{N}(0, \mathbb{I}_p)$ 是 $\mathbb{R}^p$ 中的一个高斯随机向量。集合 $\mathcal{C}$ 的高斯宽度定义为 $G_{\mathcal{C}} \overset{def}{=} \mathbb{E}_b \sup_{w \in \mathcal{C}} \left|\Big[b, w\rangle\right|$ 。$\Big]$

2 基于 Frank-Wolfe 算法的隐私凸优化

在本节中，我们分析经典 Frank-Wolfe 算法的一个差分隐私变体 [15]。我们证明，在约束集 $\mathcal{C}$ 是一个具有 k 个顶点的多面体，并且损失函数 $\mathcal{L}(\theta; d)$ 关于 ℓ_1- 范数 Lipschitz 的情况下，可以得到大约 $O(\log k/n^{2/3})$ 的隐私风险过剩。这特别适用于高维线性回归场景。一个这样的例子是计算 $\text{argmin}_{\theta:\|\theta\|_1 \leq 1} \frac{1}{n}\|X\theta - y\|_2^2$ 的经典 LASSO 算法[27],。在通常情况下 $|x_{ij}|$ $|y_j| = O(1)$ $\mathcal{L}(\theta) = \frac{1}{n}\|X\theta - y\|_2^2$ is $O(1)$-,，关于 ℓ_1-范数 Lipschitz，我们证明可以实现接近最优的隐私风险 $\widetilde{O}(1/n^{2/3})$。

Frank-Wolfe 算法 [15] 可以被视为一种“贪婪”算法，它在第一阶近似中朝向最优解移动（有关描述，请参见算法1）。Frank-Wolfe 算法的收敛速度取决于 $\mathcal{L}$ 的“曲率”，根据 [8, 17] 定义如下。我们指出，在 $\mathcal{C}$ 上的 a β-平滑函数的曲率常数被 $\beta\|C\|^2$ 所界定。

定义 2.1(曲率常数)**.** 对于 $\mathcal{L}$：$\mathcal{C} \to \mathbb{R}$，定义 $\Gamma_{\mathcal{L}}$ 如下。

$$\Gamma_{\mathcal{L}} := \sup_{\theta_1, \theta_2, \in \mathcal{C}, \gamma \in (0,1], \theta_3 = \theta_1 + \gamma(\theta_2 - \theta_1)} \frac{2}{\gamma^2} \left(\mathcal{L}(\theta_3) - \mathcal{L}(\theta_1) - \langle \theta_3 - \theta_1, \bigtriangledown\mathcal{L}(\theta_1)\rangle\right).$$

注 1. 对于二次损失 $\mathcal{L}(\theta) = \theta A^T A \theta + \langle b, \theta\rangle$ 可以推导出一个有用的界。在这种情况下，通过 [8], $\Gamma_{\mathcal{L}} \leq \max_{a\ b \in A \cdot \mathcal{C}} \|a - b\|_{2,}^2$。当 $\mathcal{C}$ 中心对称时，我们有界 $\Gamma_{\mathcal{L}} \leq 4 \max_{\theta \in \mathcal{C}} \|A\theta\|_2^2$。对于 LASSO, $A = \frac{1}{\sqrt{n}} X$。

定义 $\theta^* = \underset{\theta \in \mathcal{C}}{\text{argmin}}\, \mathcal{L}(\theta)$。以下定理界定了 Frank-Wolfe 算法的收敛速度。

Algorithm 1 Frank-Wolfe algorithm

```
Input: $\mathcal{C} \subseteq \mathbb{R}^p$, $\mathcal{L} : \mathcal{C} \to \mathbb{R}$, $\mu_t$
1: Choose an arbitrary $\theta_1$ from $\mathcal{C}$
2: for $t = 1$ to $T - 1$ do
3:     Compute $\widetilde{\theta}_t = \mathrm{argmin}_{\theta \in \mathcal{C}} \langle \triangledown \mathcal{L}(\theta_t), (\theta - \theta_t) \rangle$
4:     Set $\theta_{t+1} = \theta_t + \mu_t(\widetilde{\theta}_t - \theta_t)$
5: return $\theta_T$.
```

Theorem 2.2 ([8, 17]). *If we set* $\mu_t = 2/(t+2)$, *then* $\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) = O(\Gamma_{\mathcal{L}}/T)$.

While the Frank-Wolfe algorithm does not necessarily provide faster convergence compared to the gradient-descent based method, it has two major advantages. First, on Line 3, it reduces the problem to solving a minimization of linear function. When $\mathcal{C}$ is defined by small number of vertices, e.g. when $\mathcal{C}$ is an ℓ_1 ball, the minimization can be done by checking $\langle \triangledown \mathcal{L}(\theta_t), x \rangle$ for each vertex x of $\mathcal{C}$. This can be done efficiently. Secondly, each step in Frank-Wolfe takes a convex combination of θ_t and $\widetilde{\theta}_t$, which is on the boundary of $\mathcal{C}$. Hence each intermediate solution is always inside $\mathcal{C}$ (sometimes called *projection free*), and the final outcome θ_T is the convex combination of up to T points on the boundary of $\mathcal{C}$ (or vertices of $\mathcal{C}$ when $\mathcal{C}$ is a polytope). Such outcome might be desired, for example when $\mathcal{C}$ is a polytope, as it corresponds to a sparse solution. Due to these reasons Frank-Wolfe algorithm has found many applications in machine learning [23, 16, 8]. As we shall see below, these properties are also useful for obtaining low risk bounds for their private version.

2.1 Private Frank-Wolfe Algorithm

We now present a private version of the Frank-Wolfe algorithm. The algorithm accesses the private data only through the loss function in step 3 of the algorithm. Thus to achieve privacy, it suffices to replace this step by a private version.

To do so, we apply the exponential mechanism [21] to select an approximate optimizer. In the case when the set $\mathcal{C}$ is a polytope, it suffices to optimize over the vertices of $\mathcal{C}$ due to the following basic fact:

Fact 2.3. *Let* $\mathcal{C} \subseteq \mathbb{R}^p$ *be the convex hull of a compact set* $S \subseteq \mathbb{R}^p$. *For any vector* $v \in \mathbb{R}^p$, $\arg\min\limits_{\theta \in \mathcal{C}} \langle \theta, v \rangle \cap S \neq \emptyset$.

Thus it suffices to run the exponential mechanism to select θ_{t+1} from amongst the vertices of $\mathcal{C}$. This leads to a differentially private algorithm with risk logarithmically dependent on $|S|$. When $|S|$ is polynomial in p, it leads to an error bound with $\log p$ dependence. We can bound the error in terms of the ℓ_1-Lipschitz constant, which can be much smaller than the ℓ_2-Lipschitz constant. In particular, as we show in the next section, the private Frank-Wolfe algorithm is nearly optimal for the important high-dimensional sparse linear regression problem.

Algorithm 2 $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$: Differentially Private Frank-Wolfe Algorithm (Polytope Case)

```
Input: Data set: $\mathcal{D} = \{d_1, \cdots, d_n\}$, loss function: $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; d_i)$ (with $\ell_1$-Lipschitz
    constant $L_1$ for $\mathcal{L}$), privacy parameters: $(\epsilon, \delta)$, convex set: $\mathcal{C} = conv(S)$ with $\|\mathcal{C}\|_1$ denoting
    $\max_{s \in S} \|s\|_1$.
1: Choose an arbitrary $\theta_1$ from $\mathcal{C}$
2: for $t = 1$ to $T - 1$ do
3:     $\forall s \in S, \alpha_s \leftarrow \langle s, \triangledown \mathcal{L}(\theta_t; D) \rangle + \mathsf{Lap}\left(\frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)}}{n\epsilon}\right)$, where $\mathsf{Lap}(\lambda) \sim \frac{1}{2\lambda} e^{-|x|/\lambda}$.
4:     $\widetilde{\theta}_t \leftarrow \arg\min_{s \in S} \alpha_s$.
5:     $\theta_{t+1} \leftarrow (1 - \mu_t)\theta_t + \mu_t \widetilde{\theta}_t$, where $\mu_t = \frac{2}{t+2}$.
6: Output $\theta^{priv} = \theta_T$.
```

Theorem 2.4 (Privacy guarantee). *Algorithm 2 is* (ϵ, δ)*-differentially private.*

Algorithm 1 Frank-Wolfe algorithm

```
Input: $\mathcal{C} \subseteq \mathbb{R}^p$, $\mathcal{L} : \mathcal{C} \to \mathbb{R}$, $\mu_t$
1: Choose an arbitrary $\theta_1$ from $\mathcal{C}$
2: for $t = 1$ to $T - 1$ do
3:     Compute $\widetilde{\theta}_t = \mathrm{argmin}_{\theta \in \mathcal{C}} \langle \triangledown \mathcal{L}(\theta_t), (\theta - \theta_t) \rangle$
4:     Set $\theta_{t+1} = \theta_t + \mu_t(\widetilde{\theta}_t - \theta_t)$
5: return $\theta_T$.
```

Theorem 2.2 ([8, 17]). *If we set* $\mu_t = 2/(t+2)$, *then* $\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) = O(\Gamma_{\mathcal{L}}/T)$.

虽然与梯度下降方法相比，Frank-Wolfe 算法并不一定能提供更快的收敛速度，但它有两个主要优势。首先，在第 3 行，它将问题简化为求解线性函数的最小化。当 $\mathcal{C}$ 由少量顶点定义时，例如当 $\mathcal{C}$ 是一个 ℓ_1 球时，可以通过检查 $\langle \triangledown \mathcal{L}(\theta_t), x \rangle$ 来对 $\mathcal{C}$ 的每个顶点 x 进行最小化。这可以高效地完成。其次，Frank-Wolfe 的每一步都取 θ_t 和 $\widetilde{\theta}_t$ 的凸组合，该组合位于 $\mathcal{C}$ 的边界上。因此，每个中间解始终在 $\mathcal{C}$ 内（有时称为 无投影），最终结果 θ_T 是 $\mathcal{C}$ 边界上最多 T 个点的凸组合（当 $\mathcal{C}$ 是多面体时，即 $\mathcal{C}$ 的顶点）。这样的结果可能是所需的，例如当 $\mathcal{C}$ 是多面体时，因为它对应于稀疏解。由于这些原因，Frank-Wolfe 算法在机器学习 [23, 16, 8] 中找到了许多应用。正如我们将在下文中看到的，这些特性对于为其私有版本获得低风险界限也很有用。

2.1 私有Frank-Wolfe算法

我们现在提出Frank-Wolfe算法的私有版本。该算法仅通过算法第3步中的损失函数访问私有数据。因此，为了实现隐私保护，只需将这一步替换为私有版本。

为此，我们应用指数机制 [21] 来选择近似优化器。在{set $\mathcal{C}$ }为多面体的情况下，由于以下基本事实，只需在 $\mathcal{C}$ 的顶点上优化即可:

事实2.3。 设 $\mathcal{C} \subseteq \mathbb{R}^p$ 为一个紧致集 $S \subseteq \mathbb{R}^p$的凸包。对于任何向量 $v \in \mathbb{R}^p$，$\arg\min\limits_{\theta \in \mathcal{C}} \langle \theta, v \rangle \cap S = \emptyset$。/

因此，只需运行指数机制从 $\mathcal{C}$的顶点中选择 θ_{t+1} 。这导致一个差分隐私算法，其风险对 $|S|$ 的对数依赖。当$|S|$ 对 p是多项式时，它导致一个对数 p 依赖的错误界。我们可以用 ℓ_1-Lipschitz常数来界定误差，这可以远小于 ℓ_2-Lipschitz常数。特别是，正如我们在下一节中所示，私有Frank-Wolfe算法对于重要的高维稀疏线性回归问题是接近最优的。

Algorithm 2 $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$: Differentially Private Frank-Wolfe Algorithm (Polytope Case)

```
Input: Data set: $\mathcal{D} = \{d_1, \cdots, d_n\}$, loss function: $\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; d_i)$ (with $\ell_1$-Lipschitz
    constant $L_1$ for $\mathcal{L}$), privacy parameters: $(\epsilon, \delta)$, convex set: $\mathcal{C} = conv(S)$ with $\|\mathcal{C}\|_1$ denoting
    $\max_{s \in S} \|s\|_1$.
1: Choose an arbitrary $\theta_1$ from $\mathcal{C}$
2: for $t = 1$ to $T - 1$ do
3:     $\forall s \in S, \alpha_s \leftarrow \langle s, \triangledown \mathcal{L}(\theta_t; D) \rangle + \mathsf{Lap}\left(\frac{L_1 \|\mathcal{C}\|_1 \sqrt{8T \log(1/\delta)}}{n\epsilon}\right)$, where $\mathsf{Lap}(\lambda) \sim \frac{1}{2\lambda} e^{-|x|/\lambda}$.
4:     $\widetilde{\theta}_t \leftarrow \arg\min_{s \in S} \alpha_s$.
5:     $\theta_{t+1} \leftarrow (1 - \mu_t)\theta_t + \mu_t \widetilde{\theta}_t$, where $\mu_t = \frac{2}{t+2}$.
6: Output $\theta^{priv} = \theta_T$.
```

Theorem 2.4 (Privacy guarantee). *Algorithm 2 is* (ϵ, δ)*-differentially private.*

Since each data item is assumed to have bounded ℓ_∞ norm, for two neighboring databases D and D' and any $\theta \in \mathcal{C}$, $s \in S$, we have that

$$|\langle s, \triangledown \mathcal{L}(\theta; D)\rangle - \langle s, \triangledown \mathcal{L}(\theta; D)\rangle| = O(L_1\|\mathcal{C}\|_1/n)\,.$$

The proof of privacy then follows from a straight-forward application of the exponential mechanism [21] or its noisy maximum version [3, Theorem 5]) and the strong composition theorem [13]. In Theorem 2.5 we prove the utility guarantee for the private Frank-Wolfe algorithm for the convex polytope case. Define $\Gamma_{\mathcal{L}} = \max_{D\in\mathcal{D}} C_{\mathcal{L}}$ over all the possible data sets in $\mathcal{D}$.

Theorem 2.5 (Utility guarantee)**.** *Let L_1, S and $\|\mathcal{C}\|_1$ be defined as in Algorithms 2 (Algorithm $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$). Let $\Gamma_{\mathcal{L}}$ be an upper bound on the curvature constant (defined in Definition 2.1) for the loss function $\mathcal{L}(\cdot; d)$ that holds for all $d \in \mathcal{D}$. In Algorithm $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$, if we set $T = \frac{{\Gamma_{\mathcal{L}}}^{2/3}(n\epsilon)^{2/3}}{(L_1\|\mathcal{C}\|_1)^{2/3}}$, then*

$$\mathbb{E}\left[\mathcal{L}(\theta^{priv}; D)\right] - \min_{\theta\in\mathcal{C}} \mathcal{L}(\theta; D) = O\left(\frac{{\Gamma_{\mathcal{L}}}^{1/3}\left(L_1\|\mathcal{C}\|_1\right)^{2/3}\log(n|S|)\sqrt{\log(1/\delta)}}{(n\epsilon)^{2/3}}\right).$$

Here the expectation is over the randomness of the algorithm.

The proof of utility uses known bounds on noisy Frank-Wolfe [17], along with error bounds for the exponential mechanism. The details can be found in the full version.

General $\mathcal{C}$ While a variant of this mechanism can be applied to the case when $\mathcal{C}$ is not a polytope, its error would depend on the size of a cover of the boundary of $\mathcal{C}$, which can be exponential in p, leading to an error bound with polynomial dependence on p. In the full version, we analyze another variant of private Frank-Wolfe that uses objective perturbation to ensure privacy. This variant is well-suited for a general convex set $\mathcal{C}$ and the following result, proven in the Appendix, bounds its excess risk in terms of the Gaussian Width of $\mathcal{C}$. For this mechanism, we only need $\mathcal{C}$ to be bounded in ℓ_2 diameter, but our error now depends on the ℓ_2-Lipschitz constant of the loss functions.

Theorem 2.6. *Suppose that each loss function is L_2-Lipschitz with respect to the ℓ_2 norm, and that $\mathcal{C}$ has ℓ_2 diameter at most $\|\mathcal{C}\|_2$. Let $G_{\mathcal{C}}$ the Gaussian width of the convex set $\mathcal{C} \subseteq \mathbb{R}^p$, and let $\Gamma_{\mathcal{L}}$ be the curvature constant (defined in Definition 2.1) for the loss function $\ell(\theta; d)$ for all $\theta \in \mathcal{C}$ and $d \in \mathcal{D}$. Then there is an (ϵ, δ)-differentially private algorithm $\mathcal{A}_{\mathsf{Noise-FW}}$ with excess empirical risk:*

$$\mathbb{E}\left[\mathcal{L}(\theta^{priv}; D)\right] - \min_{\theta\in\mathcal{C}} \mathcal{L}(\theta; D) = O\left(\frac{{\Gamma_{\mathcal{L}}}^{1/3}\left(L_2 G_{\mathcal{C}}\right)^{2/3}\log^2(n/\delta)}{(n\epsilon)^{2/3}}\right).$$

Here the expectation is over the randomness of the algorithm.

2.2 Private LASSO algorithm

We now apply the private Frank-Wolfe algorithm $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$ to the important case of the sparse linear regression (or LASSO) problem.

Problem definition: Given a data set $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ of n-samples from the domain $D = \{(x, y) : x \in \mathbb{R}^p, y \in [-1, 1], \|x\|_\infty \le 1\}$, and the convex set $\mathcal{C} = \ell_1^p$. Define the mean squared loss,

$$\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i\in[n]} (\langle x_i, \theta\rangle - y_i)^2\,. \tag{2}$$

The objective is to compute $\theta^{priv} \in \mathcal{C}$ to minimize $\mathcal{L}(\theta; D)$ while preserving privacy with respect to any change of individual (x_i, y_i) pair. The non-private setting of the above problem is a variant of the least squares problem with ℓ_1 regularization, which was started by the work of LASSO [27, 28] and intensively studied in the past years.

Since the ℓ_1 ball is the convex hull of $2p$ vertices, we can apply the private Frank-Wolfe algorithm $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$. For the above setting, it is easy to check that the ℓ_1-Lipschitz constant is bounded by $O(1)$. Further, by applying the bound on quadratic programming Remark 1, we have that $C_{\mathcal{L}} \le 4\max_{\theta\in\mathcal{C}} \frac{1}{n}\|X\theta\|_2^2 = O(1)$ since $\mathcal{C}$ is the unit ℓ_1 ball, and $|x_{ij}| \le 1$. Hence $\Gamma = O(1)$. Now applying Theorem 2.5, we have

由于每个数据项都被假设为具有有界的 ℓ_∞ 范数，对于两个相邻的数据库 D 和 D' 以及任何 $\theta \in \mathcal{C}$ $s \in S$，我们有

$$|\langle s, \triangledown \mathcal{L}(\theta; D)\rangle - \langle s, \triangledown \mathcal{L}(\theta; D)\rangle| = O(L_1\|\mathcal{C}\|_1/n)\,.$$

隐私证明则来自于指数机制[21] 或其带噪声的最大版本 [3, 定理 5])和强组合定理 [13]的直接应用。在定理2.5中，我们证明了凸多面体情况下的私有Frank-Wolfe算法的有效性保证。定义 $\Gamma_{\mathcal{L}} = \max_{D\in\mathcal{D}} C_{\mathcal{L}}$ 在所有可能的数据集 $\mathcal{D}$上。

定理2.5 (有效性保证)**.** 令 L_1, S 和 $\|\mathcal{C}\|_1$ 定义如算法*2*（算法 $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$）。令 $\Gamma_{\mathcal{L}}$ 是损失函数 $\mathcal{L}(\cdot; d)$ （该损失函数对所有 $d \in \mathcal{D}$都成立）的曲率常数（定义在定义*2.1*中）的上界。在算法 $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$中，如果我们设置$T = \frac{{\Gamma_{\mathcal{L}}}^{2/3}(n\epsilon)^{2/3}}{(L_1\|\mathcal{C}\|_1)^{2/3}}$ ，则

$$\mathbb{E}\left[\mathcal{L}(\theta^{priv}; D)\right] - \min_{\theta\in\mathcal{C}} \mathcal{L}(\theta; D) = O\left(\frac{{\Gamma_{\mathcal{L}}}^{1/3}\left(L_1\|\mathcal{C}\|_1\right)^{2/3}\log(n|S|)\sqrt{\log(1/\delta)}}{(n\epsilon)^{2/3}}\right).$$

这里期望是关于算法的随机性。

效用证明利用了带噪声的Frank-Wolfe [17], 的已知界限以及指数机制的误差界限。详细信息可以在完整版本中找到。

一般的 $\mathcal{C}$ 虽然这个机制的变体可以应用于 $\mathcal{C}$ 不是多面体的情况，但其误差将取决于 $\mathcal{C}$边界的一个覆盖的大小，这可以随着 p呈指数增长，导致误差界限随着 p呈多项式依赖。在完整版本中，我们分析了另一个使用目标扰动来确保隐私的私有Frank-Wolfe变体。这个变体非常适合于一般凸集 $\mathcal{C}$，并且附录中证明的以下结果将其超额风险界限为 $\mathcal{C}$的高斯宽度。对于这个机制，我们只需要 $\mathcal{C}$ 在 ℓ_2 直径内有界，但我们的误差现在取决于损失函数的 ℓ_2-Lipschitz常数。

定理 2.6。 假设每个损失函数关于 ℓ_2 范数是 L_2-*Lipschitz* 的，并且 $\mathcal{C}$ 的直径最多为 $\|\mathcal{C}\|_2$。设 $G_{\mathcal{C}}$ 凸集 $\mathcal{C} \subseteq \mathbb{R}^p$ 的高斯宽度，并设 $\Gamma_{\mathcal{L}}$ 是损失函数 $\ell\,(\theta; d)$ 对所有 $\theta \in \mathcal{C}$ 和 $d \in \mathcal{D}$ 的曲率常数（定义在 *2.1* 节中）(ϵ, δ)-差分隐私算法 $\mathcal{A}_{\mathsf{Noise-FW}}$ 的超额经验风险：

$$\mathbb{E}\left[\mathcal{L}(\theta^{priv}; D)\right] - \min_{\theta\in\mathcal{C}} \mathcal{L}(\theta; D) = O\left(\frac{{\Gamma_{\mathcal{L}}}^{1/3}\left(L_2 G_{\mathcal{C}}\right)^{2/3}\log^2(n/\delta)}{(n\epsilon)^{2/3}}\right).$$

这里的期望值是关于算法的随机性。

2.2 私有LASSO算法

我们现在将私有 Frank-Wolfe 算法 $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$ 应用于稀疏线性回归（或LASSO）问题这一重要情况。

问题定义： 给定一个来自域 $D = \{(x, y) : x \in \mathbb{R}^p, y \in [-1, 1], \|x\|_\infty \le 1\}$ 的包含 n 个样本的数据集 $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$，以及凸集 $\mathcal{C} = \ell_1^p$。定义均方损失，

$$\mathcal{L}(\theta; D) = \frac{1}{n}\sum_{i\in[n]} (\langle x_i, \theta\rangle - y_i)^2\,. \tag{2}$$

目标是计算 $\theta^{priv} \in \mathcal{C}$ 以最小化 $\mathcal{L}(\theta; D)$，同时保持对任何个体 (x_i, y_i) 对变化的隐私。上述问题的非私有设置是最小二乘问题的 LASSO [27, 28] 正则化变体，该问题由 LASSO [27, 28] 的工作开始，并在过去几年中得到了深入研究。

由于 ℓ_1 球是 $2p$ 顶点的凸包，我们可以应用隐私 Frank-Wolfe 算法 $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$。对于上述设置，很容易验证 ℓ_1-Lipschitz 常数被 $O(1)$ 所限制。此外，通过应用二次规划的界限 Remark 1，我们得到 $C_{\mathcal{L}} \le 4\max_{\theta\in\mathcal{C}} \frac{1}{n}\|X\theta\|_2^2 = O(1)$，因为 $\mathcal{C}$ 是单位 ℓ_1 球，并且 $|x_{ij}| \le 1$。因此 $\Gamma = O(1)$。现在应用 Theorem 2.5，我们得到

Corollary 2.7. *Let $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ of n samples from the domain $\mathcal{D} = \{(x, y) : \|x\|_\infty \leq 1, |y| \leq 1\}$, and the convex set $\mathcal{C}$ equal to the ℓ_1-ball. The output θ^{priv} of Algorithm $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$ ensures the following.*

$$\mathbb{E}[\mathcal{L}(\theta^{priv}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)] = O\left(\frac{\log(np/\delta)}{(n\epsilon)^{2/3}}\right).$$

Remark 2. Compared to the previous work [20, 24], the above upper bound makes no assumption of *restricted strong convexity* or *mutual incoherence*, which might be too strong for realistic settings. Also our results significantly improve bounds of [19], from $\tilde{O}(1/n^{1/3})$ to $\tilde{O}(1/n^{2/3})$, which considered the case of the set $\mathcal{C}$ being the probability simplex and the loss being a generalized linear model.

3 Optimality of Private LASSO

In the following, we shall show that to ensure privacy, the error bound in Corollary 2.7 is nearly optimal in terms of the dominant factor of $1/n^{2/3}$.

Theorem 3.1 (Optimality of private Frank-Wolfe)**.** *Let $\mathcal{C}$ be the ℓ_1-ball and $\mathcal{L}$ be the mean squared loss in equation (2). For every sufficiently large n, for every (ϵ, δ)-differentially private algorithm $\mathcal{A}$, with $\epsilon \leq 0.1$ and $\delta = o(1/n^2)$, there exists a data set $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ of n samples from the domain $\mathcal{D} = \{(x, y) : \|x\|_\infty \leq 1, |y| \leq 1\}$ such that*

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D); D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)] = \widetilde{\Omega}\left(\frac{1}{n^{2/3}}\right).$$

We prove the lower bound by following the fingerprinting codes argument of [4] for lowerbounding the error of (ϵ, δ)-differentially private algorithms. Similar to [4] and [14], we start with the following lemma which is implicit in [4].The matrix X in Theorem 3.2 is the padded Tardos code used in [14, Section 5]. For any matrix X, denote by $X_{(i)}$ the matrix obtained by removing the i-th row of X. Call a column of a matrix a *consensus* column if the entries in the column are either all 1 or all -1. The sign of a consensus column is simply the consensus value of the column. Write $w = m/\log m$ and $p = 1000m^2$. The following theorem follows immediately from the proof of Corollary 16 in [14].

Theorem 3.2. *[Corollary 16 from [14], restated] Let m be a sufficiently large positive integer. There exists a matrix $X \in \{-1, 1\}^{(w+1) \times p}$ with the following property. For each $i \in [1, w+1]$, there are at least $0.999p$ consensus columns W_i in each $X_{(i)}$. In addition, for algorithm $\mathcal{A}$ on input matrix $X_{(i)}$ where $i \in [1, w+1]$, if with probability at least 2/3, $\mathcal{A}(X_{(i)})$ produces a p-dimensional sign vector which agrees with at least $\frac{3}{4}p$ columns in W_i, then $\mathcal{A}$ is not (ε, δ) differentially private with respect to single row change (to some other row in X).*

Write $\tau = 0.001$. Let $k = \tau w p$. We first form an $k \times p$ matrix Y where the column vectors of Y are mutually orthogonal $\{1, -1\}$ vectors. This is possible as $k \gg p$. Now we construct $w+1$ databases D_i for $1 \leq i \leq w+1$ as follows. For all the databases, they contain the common set of examples $(z_j, 0)$ (i.e. vector z_j with label 0) for $1 \leq j \leq k$ where $z_j = (Y_{j1}, \ldots, Y_{jp})$ is the j-th row vector of Y. In addition, each D_i contains w examples $(x_j, 1)$ for $x_j = (X_{j1}, \ldots, X_{jk})$ for $j \neq i$. Then $\mathcal{L}(\theta; D_i)$ is defined as follows (for the ease of notation in this proof, we work with the un-normalized loss. This does not affect the generality of the arguments in any way.)

$$\mathcal{L}(\theta; D_i) = \sum_{j \neq i} (x_j \cdot \theta - 1)^2 + \sum_{j=1}^{k} (y_j \cdot \theta)^2 = \sum_{j \neq i} (x_j \cdot \theta - 1)^2 + k\|\theta\|_2^2.$$

The last equality is due to that the columns of Y are mutually orthogonal $\{-1, 1\}$ vectors. For each D_i, consider $\theta^* \in \left\{-\frac{1}{p}, \frac{1}{p}\right\}^p$ such that the sign of the coordinates of θ^* matches the sign for the consensus columns of $X_{(i)}$. Plugging θ^* in $\mathcal{L}(\theta^*; \hat{D})$ we have the following,

$$\begin{aligned} \mathcal{L}(\theta^*; \hat{D}) &\leq \sum_{i=1}^{w} (2\tau)^2 + \frac{k}{p} && \text{[since the number of consensus columns is at least } (1-\tau)p] \\ &= (\tau + 4\tau^2) w. && (3) \end{aligned}$$

推论 2.7。 设 $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ 是从域 $\underset{iv}{\mathcal{D}} = \{(x, y)$: 中抽取的 n 个样本，并且凸集 $\mathcal{C}$ 等于 ℓ_1-球。算法 $\mathcal{A}_{\mathsf{Noise-FW(polytope)}}$ 的输出 θ^{pr} 确保以下内容。

$$\mathbb{E}[\mathcal{L}(\theta^{priv}; D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)] = O\left(\frac{\log(np/\delta)}{(n\epsilon)^{2/3}}\right).$$

备注 2. 与之前的工作 [20, 24]，相比，上述上界没有假设 限制强凸性 或 相互不相关性，这在实际场景中可能过于严格。此外，我们的结果将 [19]，的界限从 $\tilde{O}(1/n^{1/3})$ 显著提高到 $\tilde{O}(1/n^{2/3})$，后者考虑了集合 $\mathcal{C}$ 为概率单纯形且损失为广义线性模型的情况。

3 隐私LASSO 的最优性

在以下内容中，我们将证明为了确保隐私，引理 2.7 中的误差界在 $1/n^{2/3}$ 的主导因子方面几乎是最佳的。

定理 3.1 (私人 Frank-Wolfe 的最优性)**.** 令 $\mathcal{C}$ 为 ℓ_1-球， $\mathcal{L}$ 为方程 *(2)* 中的均方损失。对于每一个足够大的 n，对于每一个 (ϵ, δ)-差分隐私算法$\mathcal{A}$，具有 $\epsilon \leq 0.1$ 和 $\delta = o(1/n^2)$，存在一个数据集 $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$，包含来自域 $\mathcal{D} = \{(x, y) : \|x\|_\infty \leq 1, |y| \leq 1\}$ 的 n 个样本，使得

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D); D) - \min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)] = \widetilde{\Omega}\left(\frac{1}{n^{2/3}}\right).$$

我们通过遵循 [4] 的指纹码论证来证明下界，以限制 (ϵ, δ)-差分隐私算法的误差。类似于 [4] 和 [14], ，我们从以下引理开始，该引理隐含在 [4] 中。定理 3.2 中的矩阵 X 是在 [14，第 5] 节中使用的填充 Tardos 码。对于任何矩阵 X，用 $X_{(i)}$ 表示通过删除 X 的 i-行得到的矩阵。如果矩阵的某一列中的元素要么全为 1，要么全为 -1，则称该列为矩阵的 共识 列。共识列的符号简单地是该列的共识值。写 $w = m/\log\ m$ 和 $p = 1000m^2$。以下定理直接来自引理 16 的证明，该证明在 [14] 中。

定理 3.2。 *[推论 16 来自 [14], 重新表述]* 设 m 是一个足够大的正整数。存在一个矩阵 $X \in \{-1, 1\}^{(w+1) \times p}$ 具有以下性质。对于每个 $i \in [1, w+1]$，每个 $X_{(i)}$ 中至少有 $0.999p$ 个共识列 W_i 。此外，对于算法 $\mathcal{A}$ 在输入矩阵$X_{(i)}$ 上，其中 $i \in [1, w+1]$，如果以至少 2/3 $\mathcal{A}(X_{(i)})$ 的概率产生一个 p-维符号向量，该向量与 W_i 中的至少 $\frac{3}{4}p$ 列一致，则 $\mathcal{A}$ 不是 (ε, δ) 关于单行更改（相对于 X 中的其他行）的差分隐私。

写出 $\tau = 0.001$。设 $k = \tau w p$。我们首先构造一个 $k \times p$ 矩阵 Y，其中 Y 的列向量是相互正交的 $\{1, -1\}$ 向量。这是可能的，因为 $k \gg p$。现在我们构造 $w+1$ 数据库 D_i，如下所示。对于所有数据库，它们包含公共的示例集 $(z_j, 0)$（即标签为 0 的向量 z_j）对于 $1 \leq j \leq k$，其中 $z_j = (Y_{j1}, \ldots, Y_{jp})$ 是 Y 的第 j 行向量。此外，每个 D_i 包含 w 个示例 $(x_j, 1)$ 对于 $x_j = (X_{j1}, \ldots, X_{jk})$，对于 $j = i$。然后 $\mathcal{L}(\theta; D_i)$ 定义如下（为了方便本证明中的符号表示，我们使用未归一化的损失。这不会影响论点的普遍性。）

$$\mathcal{L}(\theta; D_i) = \sum_{j \neq i} (x_j \cdot \theta - 1)^2 + \sum_{j=1}^{k} (y_j \cdot \theta)^2 = \sum_{j \neq i} (x_j \cdot \theta - 1)^2 + k\|\theta\|_2^2.$$

最后一个等式是由于 Y 的列是相互正交的 $\{-1, 1\}$ 向量。对于每个 D_i，考虑 $\theta^* \in \left\{-\frac{1}{p}\ \frac{1}{p}\right\}^p$，

使得 θ^* 的坐标符号与 $X_{(i)}$ 的共识列的符号匹配。将 θ^* 代入 $\mathcal{L}(\theta^*; \hat{D})$ 我们得到以下结果，

$$\begin{aligned} \mathcal{L}(\theta^*; \hat{D}) &\leq \sum_{i=1}^{w} (2\tau)^2 + \frac{k}{p} && \text{[since the number of consensus columns is at least } (1-\tau)p] \\ &= (\tau + 4\tau^2) w. && (3) \end{aligned}$$

We now prove the crucial lemma, which states that if θ is such that $\|\theta\|_1 \leq 1$ and $\mathcal{L}(\theta; D_i)$ is small, then θ has to agree with the sign of most of the consensus columns of $X_{(i)}$.

Lemma 3.3. *Suppose that $\|\theta\|_1 \leq 1$, and $\mathcal{L}(\theta; D_i) < 1.1\tau w$. For $j \in W_i$, denote by s_j the sign of the consensus column j. Then we have*

$$|\{j \in W_i \ : \ \text{sign}(\theta_j) = s_j\}| \geq \frac{3}{4}p\,.$$

Proof. For any $S \subseteq \{1, \ldots, p\}$, denote by $\theta|_S$ the projection of θ to the coordinate subset S. Consider three subsets S_1, S_2, S_3, where

$$\begin{aligned} S_1 &= \{j \in W_i \ : \ \text{sign}(\theta_j) = s_j\}\,, \\ S_2 &= \{j \in W_i \ : \ \text{sign}(\theta_j) \neq s_j\}\,, \\ S_3 &= \{1, \ldots, p\} \setminus W_i\,. \end{aligned}$$

The proof is by contradiction. Assume that $|S_1| < \frac{3}{4}p$.

Further denote $\theta_i = \theta|_{S_i}$ for $i = 1, 2, 3$. Now we will bound $\|\theta_1\|_1$ and $\|\theta_3\|_1$ using the inequality $\|x\|_2 \geq \|x\|_1/\sqrt{d}$ for any d-dimensional vector.

$$\|\theta_3\|_2^2 \geq \|\theta_3\|_1^2/|S_3| \geq \|\theta_3\|_1^2/(\tau p)\,.$$

Hence $k\|\theta_3\|_2^2 \geq w\|\theta_3\|_1^2$. But $k\|\theta_3\|_2^2 \leq k\|\theta\|_2^2 \leq 1.1\tau w$, so that $\|\theta_3\|_1 \leq \sqrt{1.1\tau} \leq 0.04$.

Similarly by the assumption of $|S_1| < \frac{3}{4}p$,

$$\|\theta_1\|_2^2 \geq \|\theta_1\|_1^2/|S_1| \geq 4\|\theta_1\|_1^2/(3p)\,.$$

Again using $k\|\theta\|_2^2 < 1.1\tau w$, we have that $\|\theta_1\|_1 \leq \sqrt{1.1 * 3/4} \leq 0.91$.

Now we have $\langle x_i, \theta\rangle - 1 = \|\theta_1\|_1 - \|\theta_2\|_1 + \beta_i - 1$ where $|\beta_i| \leq \|\theta_3\|_1 \leq 0.04$. By $\|\theta_1\|_1 + \|\theta_2\|_1 + \|\theta_3\|_1 \leq 1$, we have

$$|\langle x_i, \theta\rangle - 1| \geq 1 - \|\theta_1\| - |\beta_i| \geq 1 - 0.91 - 0.04 = 0.05\,.$$

Hence we have that $\mathcal{L}(\theta; D_i) \geq (0.05)^2 w \geq 1.1\tau w$. This leads to a contradiction. Hence we must have $|S_1| \geq \frac{3}{4}p$. □

With Theorem 3.2 and Lemma 3.3, we can now prove Theorem 3.1.

Proof. Suppose that $\mathcal{A}$ is private. And for the datasets we constructed above,

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] \leq cw\,,$$

for sufficiently small constant c. By Markov inequality, we have with probability at least $2/3$, $\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_\theta \mathcal{L}(\theta; D_i) \leq 3cw$. By (3), we have $\min_{\theta} \mathcal{L}(\theta; D_i) \leq (\tau + 4\tau^2)w$. Hence if we choose c a constant small enough, we have with probability $2/3$,

$$\mathcal{L}(\mathcal{A}(D_i); D_i) < (\tau + 4\tau^2 + 3c)w \leq 1.1\tau w\,. \tag{4}$$

By Lemma 3.3, (4) implies that $\mathcal{A}(D_i)$ agrees with at least $\frac{3}{4}p$ consensus columns in $X_{(i)}$. However by Theorem 3.2, this violates the privacy of $\mathcal{A}$. Hence we have that there exists i, such that

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] > cw\,.$$

Recall that $w = m/\log m$ and $n = w + wp = O(m^3/\log m)$. Hence we have that

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] = \Omega(n^{1/3}/\log^{2/3} n)\,.$$

The proof is completed by converting the above bound to the normalized version of $\Omega(1/(n \log n)^{2/3})$. □

我们现在证明这个关键的引理，它声明如果 θ 是这样的，使得 $\|\theta\|_1 \leq 1$ 和 $\mathcal{L}(\theta; D_i)$ 很小，那么 θ 必须与 $X_{(i)}$ 的多数共识列的符号一致。

引理 3.3. 假设 $\|\theta\|_1 \leq 1$，和 $\mathcal{L}(\theta;\ D_i) < 1.1\tau w$。对于 $j \in W_i$，用 s_j 表示共识列 j 的符号。那么我们有

$$|\{j \in W_i \ : \ \text{sign}(\theta_j) = s_j\}| \geq \frac{3}{4}p\,.$$

证明。 对于任何 $S \subseteq \{1, \ldots, p\}$，用 $\theta|_S$ 表示 θ 到坐标子集 S 的投影。考虑三个子集 S_1 S_2 S_3，，，其中

$$\begin{aligned} S_1 &= \{j \in W_i \ : \ \text{sign}(\theta_j) = s_j\}\,, \\ S_2 &= \{j \in W_i \ : \ \text{sign}(\theta_j) \neq s_j\}\,, \\ S_3 &= \{1, \ldots, p\} \setminus W_i\,. \end{aligned}$$

证明采用反证法。假设 $|S_1| < \frac{3}{4}p$。

进一步用 $\theta_i = \theta|_{S_i}$ 表示 $i = 1,2$， 3。现在我们将使用不等式 $\|x\|_2 \geq \|x\|_1/\sqrt{d}$ 限制 $\|\theta_1\|_1$ 和 $\|\theta_3\|_1$，对于任何 d 维向量。

$$\|\theta_3\|_2^2 \geq \|\theta_3\|_1^2/|S_3| \geq \|\theta_3\|_1^2/(\tau p)\,.$$

因此 $k\|\theta_3\|_2^2 \geq w\|\theta_3\|_1^2$。但 $k\|\theta_3\|_2^2 \leq k\|\theta\|_2^2 \leq 1.1\tau w$，所以 $\|\theta_3\|_1 \leq \sqrt{1.1\tau} \leq 0.04$。

同样地，通过假设 $|S_1| < \frac{3}{4}p$,

$$\|\theta_1\|_2^2 \geq \|\theta_1\|_1^2/|S_1| \geq 4\|\theta_1\|_1^2/(3p)\,.$$

再次使用 $k\|\theta\|_2^2 < 1.1\tau w$，我们得到 $\|\theta_1\|_1 \leq \quad \sqrt{}1.1 * 3/4 \leq 0.91$。

现在我们有 $\langle x_i, \theta\rangle - 1 = \|\theta_1\|_1 - \|\theta_2\|_1 + \beta_i - 1$ 其中 $|\beta_i| \leq \|\theta_3\|_1 \leq 0.04$。通过 $\|\theta_1\|_1 + \|\theta_2\|_1 + \|\theta_3\|_1 \leq 1$，我们得到

$$|\langle x_i, \theta\rangle - 1| \geq 1 - \|\theta_1\| - |\beta_i| \geq 1 - 0.91 - 0.04 = 0.05\,.$$

因此我们有 $\mathcal{L}(\theta; D_i) \geq (0.05)^2 w \geq 1.1\tau w$。这导致了矛盾。因此我们必须有 $|S_1| \geq \frac{3}{4}p$. | |

利用定理 3.2 和引理 3.3，我们现在可以证明定理 3.1。

证明。 假设 $\mathcal{A}$ 是私密的。对于我们上面构造的数据集，

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] \leq cw\,,$$

对于足够小的常数 c。根据马尔可夫不等式，我们有至少以 $2/3$ 的概率，$\mathcal{L}(\mathcal{A}(D_i);\ D_i) - \min_\theta \mathcal{L}(\theta;\ D_i) \leq 3cw$。根据 (3)，我们有 $\min_\theta \mathcal{L}(\theta; D_i) \leq (\tau + 4\tau^2)w$。因此如果我们选择 c 一个足够小的常数，我们有至少以 $2/3$ 的概率，

$$\mathcal{L}(\mathcal{A}(D_i); D_i) < (\tau + 4\tau^2 + 3c)w \leq 1.1\tau w\,. \tag{4}$$

根据引理 3.3，(4) 意味着 $\mathcal{A}(D_i)$ 与 $X_{(i)}$ 中的至少 $\frac{3}{4}p$ 个共识列一致。然而根据定理 3.2，这违反了 $\mathcal{A}$ 的隐私。因此我们有，存在 i，使得

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] > cw\,.$$

回想 $w = m/\log\ m$ 和 $n = w + wp = O(m^3/\log m)$。因此我们有

$$\mathbb{E}[\mathcal{L}(\mathcal{A}(D_i); D_i) - \min_{\theta} \mathcal{L}(\theta; D_i)] = \Omega(n^{1/3}/\log^{2/3} n)\,.$$

通过将上述界限转换为 $\Omega(1/(n\log n)^{2/3})$ 的标准化版本来完成证明。

| |

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[2] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization, revisited. In *FOCS*, 2014.

[3] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta. Discovering frequent patterns in sensitive data. In *KDD*, New York, NY, USA, 2010.

[4] M. Bun, J. Ullman, and S. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, 2014.

[5] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[6] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*, 2008.

[7] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *JMLR*, 12:1069–1109, 2011.

[8] K. L. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transations on Algorithms*, 2010.

[9] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, 2013.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[11] C. Dwork, A. Nikolov, and K. Talwar. Efficient algorithms for privately releasing marginals via convex relaxations. *arXiv preprint arXiv:1308.1385*, 2013.

[12] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science. NOW Publishers, 2014.

[13] C. Dwork, G. N. Rothblum, and S. P. Vadhan. Boosting and differential privacy. In *FOCS*, 2010.

[14] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *STOC*, 2014.

[15] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[16] E. Hazan and S. Kale. Projection-free online learning. In *ICML*, 2012.

[17] M. Jaggi. Revisiting {Frank-Wolfe}: Projection-free sparse convex optimization. In *ICML*, 2013.

[18] P. Jain, P. Kothari, and A. Thakurta. Differentially private online learning. In *COLT*, pages 24.1–24.34, 2012.

[19] P. Jain and A. Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning (ICML)*, 2014.

[20] D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *COLT*, pages 25.1–25.40, 2012.

[21] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103. IEEE, 2007.

[22] A. Nikolov, K. Talwar, and L. Zhang. The geometry of differential privacy: The sparse and approximate cases. In *STOC*, 2013.

[23] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 2010.

[24] A. Smith and A. Thakurta. Differentially private feature selection via stability arguments, and the robustness of the Lasso. In *COLT*, 2013.

[25] A. Smith and A. Thakurta. Follow the perturbed leader is differentially private with optimal regret guarantees. *Manuscript in preparation*, 2013.

[26] A. Smith and A. Thakurta. Nearly optimal algorithms for private online learning in full-information and bandit settings. In *NIPS*, 2013.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.

[28] R. Tibshirani et al. The Lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[29] J. Ullman. Private multiplicative weights beyond linear queries. *CoRR*, abs/1407.1571, 2014.

参考文献

[1] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[2] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization, revisited. In *FOCS*, 2014.

[3] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta. Discovering frequent patterns in sensitive data. In *KDD*, New York, NY, USA, 2010.

[4] M. Bun, J. Ullman, and S. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, 2014.

[5] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[6] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*, 2008.

[7] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *JMLR*, 12:1069–1109, 2011.

[8] K. L. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transations on Algorithms*, 2010.

[9] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, 2013.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[11] C. Dwork, A. Nikolov, and K. Talwar. Efficient algorithms for privately releasing marginals via convex relaxations. *arXiv preprint arXiv:1308.1385*, 2013.

[12] C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science. NOW Publishers, 2014.

[13] C. Dwork, G. N. Rothblum, and S. P. Vadhan. Boosting and differential privacy. In *FOCS*, 2010.

[14] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *STOC*, 2014.

[15] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[16] E. Hazan and S. Kale. Projection-free online learning. In *ICML*, 2012.

[17] M. Jaggi. Revisiting {Frank-Wolfe}: Projection-free sparse convex optimization. In *ICML*, 2013.

[18] P. Jain, P. Kothari, and A. Thakurta. Differentially private online learning. In *COLT*, pages 24.1–24.34, 2012.

[19] P. Jain and A. Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning (ICML)*, 2014.

[20] D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *COLT*, pages 25.1–25.40, 2012.

[21] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103. IEEE, 2007.

[22] A. Nikolov, K. Talwar, and L. Zhang. The geometry of differential privacy: The sparse and approximate cases. In *STOC*, 2013.

[23] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 2010.

[24] A. Smith and A. Thakurta. Differentially private feature selection via stability arguments, and the robustness of the Lasso. In *COLT*, 2013.

[25] A. Smith and A. Thakurta. Follow the perturbed leader is differentially private with optimal regret guarantees. *Manuscript in preparation*, 2013.

[26] A. Smith and A. Thakurta. Nearly optimal algorithms for private online learning in full-information and bandit settings. In *NIPS*, 2013.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.

[28] R. Tibshirani et al. The Lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

[29] J. Ullman. Private multiplicative weights beyond linear queries. *CoRR*, abs/1407.1571, 2014.