HOW TO MAKE SEMI-PRIVATE LEARNING EFFECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

In Semi-Private (SP) learning, the learner has access to both public and private data, and the differential privacy requirement is imposed solely on the private data. We propose a computationally efficient algorithm that, under mild assumptions on the data, provably achieves significantly lower sample complexity and can be efficiently run on realistic datasets. To achieve this, we leverage the features extracted by pre-trained networks. To validate its empirical effectiveness, we propose a particularly challenging set of experiments under tight privacy constraints ($\epsilon = 0.1$) and with a focus on low-data regimes. In all the settings, our algorithm exhibits significantly improved performance over the available baseline.

1 INTRODUCTION

Recent works [31, 38, 11] have shown that attackers can maliciously query ML models in order to reveal the private information contained in them. The de-facto standard remedy to this problem is to enforce $(\epsilon, \delta)-$Differential Privacy (DP) [16] guarantees on the ML algorithms. Intuitively, satisfying these guarantees with low ϵ and δ makes it very unlikely for attackers to reliably determine whether a sample belongs or not to the training set of the attacked model. However, it is well known that satisfying such guarantees comes at the cost of the model's utility, unless the amount of available private training data is significantly increased [21, 10, 8, 9, 17]. To partially address this problem, it has been shown that unlabelled public data [4] (also known as the *Semi-Private* (SP) setting) can lower the private sample complexity. Although of significant theoretical and practical relevance, it is unclear how to scale their proposed algorithm to operate in real-world applications.

The seminal works of Chaudhuri et al. [12], Bassily et al. [6] have shown that the sample complexity of DP empirical risk minimisation suffers from a dependence on the dimensionality of the problem. Our work follows a line of research [29, 20] that exploits the large margin of the data to reduce this dependence. Assuming the data approximately lies on a low rank subspace, we propose an SP training algorithm to learn linear halfspaces with a reduced sample complexity. Our algorithm estimates the Principal Components of the data using public unlabelled data, projects the private dataset on the top-k Principal Components, and learns a private linear classifier for the projected data.

To empirically validate the effectiveness of the proposed technique, we apply it to the common task of image classification. Similarly to a growing line of works [32, 14, 26, 24], we leverage the features extracted by models pre-trained on an additional large-scale dataset (that is assumed to be public) to implement our algorithm. In Figure 1 we display the empirical effectiveness of our SP technique not only on standard classification benchmarks used in the DP literature, but also on a collection of datasets that we argue better represents the actual challenges of (semi-)private training (in agreement with some of the points raised by the concurrent work of Tramèr et al. [33]).

In our evaluations we particularly focus on private data distributions that increasingly differ from the pre-training ones and on low-data regimes. Besides testing on more relevant benchmarks, remarkably, we find the benefits of our SP algorithm increase the tighter the privacy guarantees (i.e., the lower ϵ). In contrast, we observe reducing the dimensionality of the input without imposing privacy guarantees (i.e., $\epsilon = \infty$) to be harmful.

2 AN EFFICIENT SEMI-PRIVATE LEARNER

In this section, we present our main theoretical result for semi-privately learning linear halfspaces. Before that, we introduce some necessary definitions.

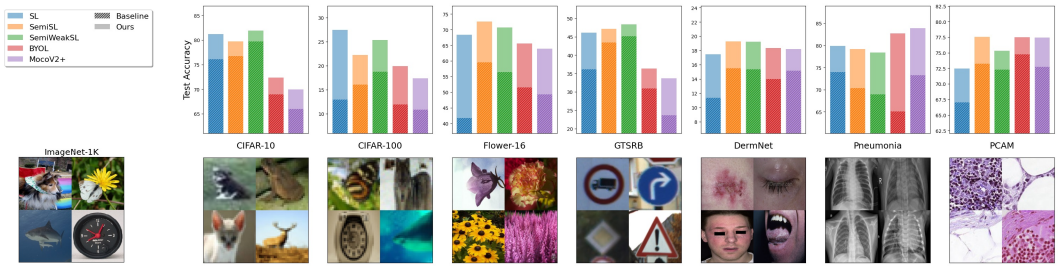


Figure 1: Typical DP fine-tuning benchmarks consider datasets pre-trained on ImageNet-1K (I1K) and fine-tuned on CIFAR-10 and CIFAR-100. We suggest benchmarking also on datasets with less class overlap (e.g., GTSRB, Flower-16) or no overlap (e.g., Dermnet, Pneumonia, PatchCamelyon (PCAM)). Our SP method learning linear classifiers on top of pre-trained features significantly outperforms the DP baseline.

2.1 PRELIMINARIES

Informally, Differential Privacy requires the output distribution of a randomized algorithm to remain similar when one input data point is modified. In the context of this paper, a differentially private learning algorithm generates similar distributions over classifiers when trained on neighbouring datasets. Two datasets are said to be neighbouring when they differ in one entry. Formally,

Definition 1 (Differential Privacy [16]). *A learning algorithm $\mathcal{A}$ is (ϵ, δ)-differential private, if for any two datasets S, S' differing in one entry and for all outputs $\mathcal{Z}$, we have,*

$$\mathbb{P}\left[\mathcal{A}(S) \in \mathcal{Z}\right] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}(S) \in \mathcal{Z}\right] + \delta.$$

For $\epsilon < 1$ and $\delta = o(1/n)$, (ϵ, δ)-differential privacy provides valid protection against potential privacy attacks [11]. Next, we define *semi-private learner*, introduced in Alon et al. [4], where the learning algorithm has access to both a private labelled and a public (labelled or unlabelled) dataset are available. In the theoretical analysis, we assume the more realistic case of having an *unlabelled* public dataset. This specific setting has been referred to as Semi-Supervised Semi-Private learning in Alon et al. [4]. However, for the sake of brevity, we will refer to it as Semi-Private learning (SPL).

Definition 2 ($(\alpha, \beta, \epsilon, \delta)$-semi-private learner on a family of distributions $\mathcal{D}$). *An algorithm $\mathcal{A}$ is said to $(\alpha, \beta, \epsilon, \delta)$-semi-privately learn a hypothesis class $\mathcal{H}$ on a family of distributions $\mathcal{D}$ if for any distribution $D \in \mathcal{D}$, given a labelled dataset S^L of size n^L and an unlabelled dataset S^U of size n^U sampled i.i.d. from D, $\mathcal{A}$ is (ϵ, δ)-DP with respect to S^L and outputs a hypothesis $\hat{h}$ satisfying*

$$\mathbb{P}[\mathbb{P}_{(x,y)\sim D}\left[h(x) \neq y\right] \leq \alpha] \geq 1 - \beta,$$

where the outer probability is over the randomness of S^L, S^U, and $\mathcal{A}$.

Further, the sample complexity n^L and n^U must be polynomial in $\frac{1}{\alpha}, \frac{1}{\beta}$, and the size of the input space. In addition, n^L is also polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. The algorithm is said to be efficient if it also runs in time polynomial in $\frac{1}{\alpha}, \frac{1}{\beta}$, and the size of the input domain.

A key difference between our setting and the one introduced in Alon et al. [4] is that they look at the distribution-independent agnostic learning setting whereas we look at distribution-specific realisable setting. On the other hand, while their algorithm is not computationally efficient, ours is and can be run in time polynomial in the relevant parameters.

2.2 PROBLEM SETTING

Our theoretical results focus on the problem of learning linear halfspaces $\mathcal{H}_L^d$ in d dimensions. Consider the instance space $\mathcal{X}_d = B_2^d = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ as the d-dimensional unit sphere and the binary label space $\mathcal{Y} = \{-1, 1\}$. The hypothesis class of linear halfspaces is defined as $\mathcal{H}_L^d = \{f_w(x) = \text{sign}\left(\langle w, x\rangle\right) | w \in B_2^d\}$. As mentioned above, our results apply in the setting of distribution specific learning. Intuitively, our family of distribution admits a large margin linear

Algorithm 1 $\mathcal{A}_{\epsilon,\delta}(k,\zeta)$

input Labelled dataset S^L, unlabelled dataset S^U, rank parameter k, distributional parameter ζ.

1: Using unlabelled dataset S^U to construct the empirical covariance matrix $\widehat{\Sigma} = \sum_{x \in S^U} xx^T / n^U$.

2: Construct the transformation matrix $\widehat{A}_k$ whose i^{th} column is the ith eigenvector of $\widehat{\Sigma}$.

3: Project S^L with the transformation matrix $\widehat{A}_k$,

$$S^L_{\widehat{A}_k} = \{(\widehat{A}_k^T x, y) : (x,y) \in S^L\}.$$

4: Obtain $\hat{w}_{\widehat{A}_k} = \mathcal{A}_{Noisy-SGD}(S^L_{\widehat{A}_k}, S^L, \ell, (\epsilon,\delta))$ where $\ell : S^L_{\widehat{A}_k} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$ is defined as

$$\ell(w,(x,y)) = \max\left\{1 - \frac{y}{\zeta}\langle w, x\rangle, 0\right\}. \tag{1}$$

output Return $\hat{w} = \widehat{A}_k \hat{w}_{\widehat{A}_k}$.

classifier that contains a large projection on the top principal components of the population covariance matrix. We formalise this as (γ, ξ_k)-Large margin low rank distributions in Definition 3.

Definition 3 ((γ,ξ_k)-Large margin low rank distribution)**.** *A distribution D over $\mathcal{X}_d \times \mathcal{Y}$ is a (γ,ξ_k)-Large margin low rank distribution if there exists $w^* \in B_2^d$ such that*

- $\mathbb{P}_{(x,y)\sim D}\left[\frac{y\langle w^*, x\rangle}{\|w^*\|\|x\|} \geq \gamma\right] = 1$ *(Large-margin),*
- $\left\|A_k A_k^T w^*\right\| \geq 1 - \xi_k$ *(Low-dimension).*

where A_k is a $d \times k$ matrix whose columns are the top k eigenvectors of the covariance matrix of $X \sim D_X$.

Note that, for every distribution that admits a positive margin γ, the low-dimension condition is automatically satisfied for all $k \leq d$ with some $\xi_k \geq 0$. However, the distribution is low rank if it holds for a small k and small ξ_k simultaneously. We use the notation Δ_k to represent the gap between the kth and the $k+1^{\text{th}}$ eigenvalue of the population covariance matrix.

2.3 Theoretical Guarantees

Next, we propose our semi-private learning algorithm in Algorithm 1. The algorithm inputs the privacy parameters ϵ, δ, the labelled dataset S^L, the unlabelled dataset S^U, and an additional parameter ζ that we discuss below as input and outputs a linear separator. Algorithm 1 internally calls $\mathcal{A}_{Noisy-SGD}$ which is the Noisy Stochastic Gradient Descent (SGD) algorithm by [7]. We include this in Appendix B.2.3 for completeness. The parameter ζ depends on the parameters of the large margin low rank distribution (γ, ξ_k). In Appendix C, we also consider a more challenging setting where the unlabelled data and the labelled may come from similar, but not the same distribution. In that case, ζ also depends on the distance between the two distributions.

Theorem 1 shows that if the unlabelled and labelled datasets are sampled from the same large-margin low rank distribution, then Algorithm 1 is both (ϵ,δ)-DP with respect to the labelled dataset and achieves high accuracy on the distribution with relatively small number of labelled data points.

Theorem 1. *For $\gamma_0 \in (0,1), \xi_0 \in [0,1)$, let $\mathcal{D}_{\gamma_0,\xi_0}$ be the family of distributions consisting all (γ,ξ_k)-large margin low rank distributions over $\mathcal{X}_d \times \mathcal{Y}$ with $\gamma \geq \gamma_0$ and $\xi_k \leq \xi_0$. For any $\alpha \in (0,1), \beta \in (0, 1/4)$, $\epsilon \in (0, 1/\sqrt{k})$ and $\delta \in (0,1)$, Algorithm $\mathcal{A}_{\epsilon,\delta}(k,\zeta)$ is an $(\alpha,\beta,\epsilon,\delta)$-semi-private learner of the linear halfspace $\mathcal{H}_L^d$ on $\mathcal{D}_{\gamma_0,\xi_0}$ with sample complexity*

$$n^U = O\left(\frac{\log 2/\beta}{\gamma_0^2 \Delta_k^2}\right), n^L = \widetilde{O}\left(\frac{\sqrt{k}}{\alpha\epsilon\zeta}\right)$$

where $\zeta = \gamma_0(1 - 1\xi_0 - 0.1\gamma_0)$.

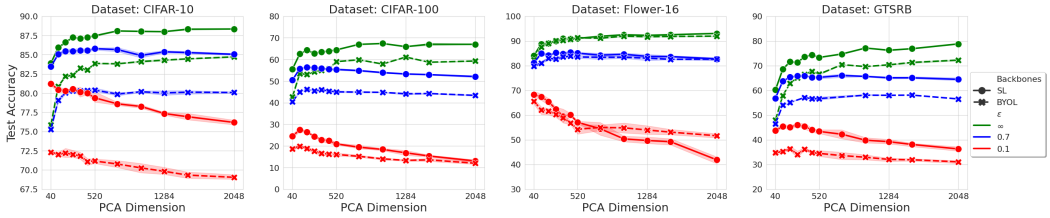


Figure 2: First and second panel: Test Accuracy of DP linear classifier on CIFAR-10 and CIFAR-100. Third and fourth panel: Test accuracy of DP linear classifier on Flower-16 and GTSRB. While for public training ($\epsilon = \infty$) the performance generally increases as dimensions are added, the opposite occurs for DP training. The tighter the privacy constraints, the steeper the decrease. For results on additional feature-extractors refer to Figure 6 in Appendix F.

We compare our result, theoretically, with other relevant works in Appendix D. In order to show the effectiveness of Algorithm 1, we first focus on the task of image classification with CIFAR-10 and CIFAR-100 datasets [23] under differential privacy [14, 32, 24].

3 Results on Standard Image Classification Benchmarks

Common image classification datasets like CIFAR-10 and CIFAR-100 are unlikely to even approximately satisfy the large margin low rank assumptions in Definition 3. However, as Appendix A shows, since features of pre-trained classifiers approximate better those assumptions, in this paper, we use features obtained from a ResNet50 (R50), pre-trained [35] on ImageNet. This aligns with the current trend of leveraging large-scale pre-trained models [14, 32, 26, 24]. In particular, we leverage three different types of pre-training algorithms — supervised training on ImageNet-1K (SL), self-supervised (SSL) training on ImageNet-1K (BYOL [18] and MoCoV2+ [13]), and semi-supervised (SemiSL) and semi-weakly supervised (SemiWeakSL) [37] training (that besides ImageNet-1K leverages a much larger-scale dataset for semi-/semi-weak supervision). For the sake of clarity, in the main paper we only report results for SL and BYOL and relegate the rest to Appendix F.

3.1 Evaluation on CIFAR-10 and CIFAR-100

Differently from past works in the literature [14, 32, 24], we do not focus on values of $\epsilon > 1$. Indeed, while moderately large ϵ allows to measure the progress in the ability of training deep networks with DP at acceptable levels of accuracy, a large ϵ might yield vacuous privacy guarantees [28] and be of little practical relevance. Therefore, we focus on $\epsilon \in \{0.1, 0.7, \infty\}$, where, $\epsilon = \infty$ corresponds to public training of the linear classifier.

In the two left-most panels of Figure 2, we report the test accuracy for CIFAR-10 and CIFAR-100 as a function of the dimensionality of projection. Unless mentioned otherwise, we compute the principal component on a left-out public unlabelled dataset consisting of 10% of the training data. Note that this is precisely the public data allowed by Semi-Private learning in Definition 4.

Figure 2 shows that private-training of the linear classifier benefits from decreasing dimension. To the contrary, public training of the linear classifier either suffers with decreasing dimension or remains stationary. For instance, consider the accuracy on CIFAR-10 for the SL feature extractor at $\epsilon = 0.1$. For $k = 40$, the test accuracy of private training is 81.3%; when no dimensionality reduction is applied it drops to 76.9%. For CIFAR-100, with the SL feature extractor at $\epsilon = 0.7$, the accuracy drops from 56.8% at $k = 200$ to 51.9% for full dimension.

This dichotomy between private and non-private learning in terms of test accuracy as a function of the projection dimension is suggested by Proposition 4 in Appendix D and Theorem 1. Theorem 1 shows that that the private test accuracy increases as the dimension of projection decreases, which is what we observe in Figure 2. For non-private training with moderately large dimension, ($k \geq 520$), we see that the test accuracy remains largely constant which is the setting captured by Proposition 4. The decrease in non-private accuracy for very small values of k is because the approximation error (i.e. how well can the best classifier in k dimensions represent the ground truth), becomes large for very

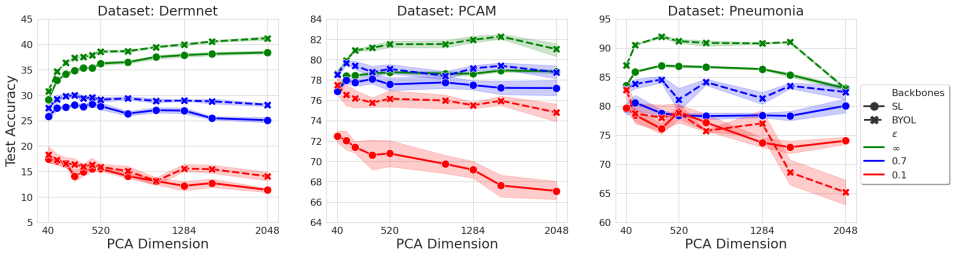


Figure 3: Test Accuracy of DP classification on Dermnet, PCAM and Pneumonia. For results on additional feature-extractors refer to Figure 7 in Appendix F.

small values of k. We will see this difference in behavior between private and non-private learning for decreasing values of k appear in all our experiments and is one of the main contributions of this paper. Although we have shown our algorithm to be effective on the typical CIFAR-10 and CIFAR-100 benchmark [32, 14], as we discuss in Section 4, we find this evaluation setting limiting and misaligned with the goal of private learning.

4 Experimental Results Beyond Standard Benchmarks

4.1 Beyond the Assumptions of Standard Benchmarks

We argue that when the pre-training occurs on ImageNet-1K, as is common practice [32, 14], these evaluation methods imply some unrealistic assumptions (also argued in concurrent work [33]). Therefore, we suggest additional evaluation settings that overcome these assumptions.

Similarity between ImageNet-1k and CIFAR-10, CIFAR-100 Indeed, both datasets contain daily life objects and the label sets of the private CIFAR datasets are partly subsumed by ImageNet (9 out of 10 CIFAR-10 classes and 60 out of 100 CIFAR-100 classes are represented in ImageNet-1K). Such an assumption is unrealistic for several relevant privacy applications (e.g. medical, finance, satellite) and for which a significant distribution shift between the public and private datasets is expected, both in the covariates and the label set. Therefore, we propose to consider additional datasets.

In Figure 1, we report a few samples from the private datasets we consider. For the Flower-16 [2] and GTSRB [19] datasets, the overlap with ImageNet-1K is significantly small. Flower-16 only contains a class that is also present in ImageNet-1K labels, and all the 43 traffic signs of GTSRB are aggregated in a single label in ImageNet-1K. No class present in the Pneumonia [22], Patch Camelyon (PCAM) [34] and DermNet datasets [1] are present in ImageNet-1K.

Private dataset are assumed to be large: Public datasets are usually large scale because they can be scraped from the web [15, 25, 36] and the classes are chosen so that the labelling process does not require specialised domain experts. In contrast, private data is 1) inherently more likely to be collected at smaller scales by private entities, 2) cannot be easily aggregated with other private datasets from other sources due to legal constraints and the competitive advantage it provides, and 3) the labelling process might be extremely expensive (e.g., biochemical or medical data).

To simulate these problem, we consider both naturally occurring small datasets or we decrease their size. For DermNet, we use approximately 12000 training samples (for 23 classes), with significant variability in the amount of samples per class. For Pneumonia, we take 1600 samples of the training set to perform PCA, and therefore we are left with approximately 3400 training samples. To inspect the low-data setting in CIFAR-100 and GTSRB, in Section 4.3, we perform our experiments with various fractions of the training data. Our algorithm performs exceedingly well in these challenging settings as well.

4.2 Performance when Pre-Training and Private Data Significantly Differ

In the two right-most panels of Figure 2, we report the results obtained on Flower-16 and GTSRB. As it can be seen, reducing the dimensionality is still generally beneficial for DP training. For instance, consider the SL feature extractor performance on Flower-16 at $\epsilon = 0.1$. For $k = 40$, the accuracy

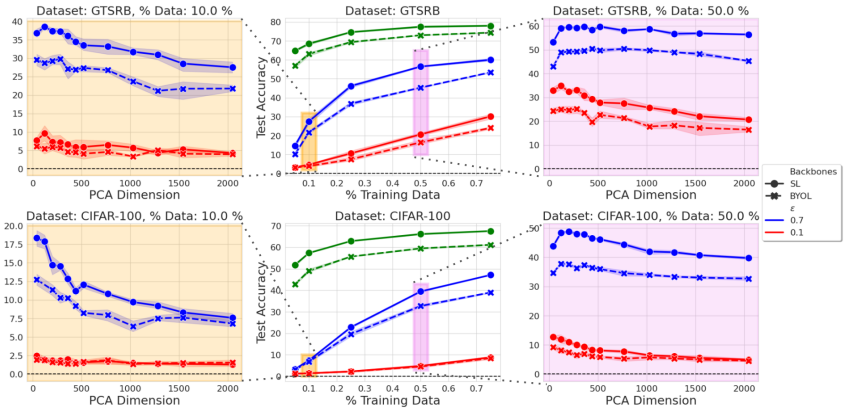


Figure 4: In the first column, test accuracy as the percentage of available training data varies in $\{0.05, 0.1, 0.25, 0.5, 0.75\}$. Second and third columns, results of applying PCA in low-data regime for 50% and 10& of the training data of GTSRB and CIFAR100.

is about 69.3% and drops to 41.2% when the full dimensionality is used. Similarly, on GTSRB at $\epsilon = 0.7$, for the SL feature extractor the accuracy drops from 65.9% at $k = 280$ to 64.2%. The impact of dimensionality reduction becomes more significant the tighter the privacy constraints are. Similar to CIFAR-10 and CIFAR-100, dimensionality reduction can significantly degrade the performance for non-DP training. In Figure 3 we present the results for Pneumonia, PCAM and DermNet (no class overlapping with the pre-training dataset), observing similar trends.

4.3 Performance in Low-Data Regimes

Good Features Are Not Always Enough. In Figure 4 we show the performance of DP and public training by reducing the amount of available training data, in an uniformly random manner, for CIFAR100 and GTSRB. While the authors of Tramer & Boneh [32] indicate that DP training can be facilitated by the availability of good features, we observe that under tight privacy constraints, even in cases of mild or little distribution shift, the performance of DP training is exceptionally low. For instance, when 5% of the training data, the performance of DP training for $\epsilon = 0.1$ is not significantly higher than the random prediction baseline (indicated with a black dashed line) whereas the performance of public training is still considerably high. The large relative drop in accuracy for DP training, compared to non-DP training is especially striking. For $\epsilon = 0.7$, the accuracy of the classifier using SL features on GTSRB drops from 57.3% when 75% of the data is available to 15.3% when 10% of the data is available, which is approximately 0.25 times the original value; however, for $\epsilon = 0.1$ the accuracy drops from 30.2% to 4.3%, which is 0.13 times the original accuracy.

Does PCA still help tackling classification in low data regimes? In Figure 4 we show that our algorithm still exhibits significantly improved performance over using the full-dimensional embeddings. For instance, on 10% of CIFAR-100 for $\epsilon = 0.7$ and the classifier using the SL features, the accuracy can rise from 7.53% to 18.3% by applying PCA with $k = 40$ dimensions.

5 Conclusion

In this paper, we consider the setting of semi-private learning where the learner has access to public unlabelled data in addition to private labelled data. This is a realistic setting in many circumstances e.g. where some people choose to make their data public. Under this setting, we proposed a new algorithm to learn linear halfspaces. Our algorithm uses a mix of PCA on unlabelled data and DP training on private data. Under reasonable theoretical assumptions, we have shown the proposed algorithm is (ϵ, δ)-DP and provably reduces the sample complexity. In practical applications, we performed an extensive set of experiments that show the proposed technique is effective when tight privacy constraints are imposed, even in low-data regimes and with a significant distribution shift between the pre-training and private distribution.

REFERENCES

[1] Dataset for 23 skin lesions. https://www.kaggle.com/datasets/shubhamgoel27/dermnet.

[2] Flowers dataset. https://tinyurl.com/2p8vpsp2.

[3] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery, 2016.

[4] Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. *Advances in neural information processing systems*, 32, 2019.

[5] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. 1999.

[6] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pp. 464–473. IEEE, 2014.

[7] Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. *CoRR*, 2014.

[8] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. 2013.

[9] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*. 2013.

[10] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 128–138, 2005.

[11] Nicholas Carlini, Steve Chien, Milad Nasar, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (SP) '22*, IEEE Symposium on Security and Privacy (SP) '22, 2022.

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

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[14] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[16] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, 2006.

[17] Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. 2014.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[19] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, number 1288, 2013.

[20] Prateek Jain and Abhradeep Guha Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, pp. 476–484. PMLR, 2014.

[21] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 2011.

[22] et al. Kermany DS, Goldbaum M. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell, February 2018.

[23] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[24] Alexey Kurakin, Steve Chien, Shuang Song, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022.

[25] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[26] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

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

[28] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papemoti, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 866–882. IEEE, 2021.

[29] Huy Le Nguyen, Jonathan R. Ullman, and Lydia Zakynthinou. Efficient private algorithms for learning large-margin halfspaces. In *Algorithmic Learning Theory, ALT 2020, 8-11 February 2020, San Diego, CA, USA*, 2020.

[30] Yuge Shi, Imant Daunhawer, Julia E. Vogt, Philip H.S. Torr, and Amartya Sanyal. How robust are pre-trained models to distribution shift? In *International Conference on Learning Representations (ICLR)*, 2023.

[31] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 2017.

[32] Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YTWGvpFOQD-`.

[33] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*, 2022.

[34] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018.

[35] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[36] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE international conference on computer vision*, pp. 1202–1211, 2017.

[37] I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification, 2019.

[38] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pp. 3093–3106, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394505. doi: 10.1145/3548606.3560675. URL https://doi.org/10.1145/3548606.3560675.

[39] Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

[40] Yingxue Zhou, Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private SGD with gradient subspace identification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[41] Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, 2005.

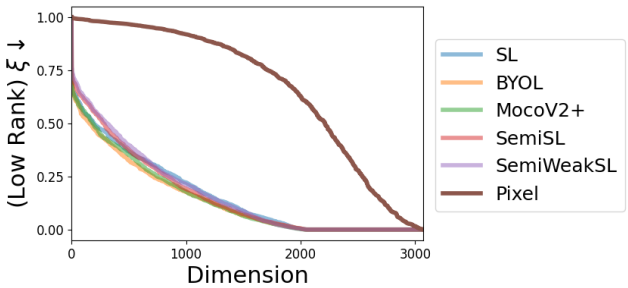


Figure 5: Estimate of ξ for linear classifiers trained on embeddings of two CIFAR-10 classes, extracted from pre-trained ResNet50.

A USING PRE-TRAINED FEATURES TO APPLY OUR ALGORITHM

A.1 PRE-TRAINED FEATURES ARE APPROXIMATELY LARGE-MARGIN AND LOW-RANK

Common image classification datasets like CIFAR-10 and CIFAR-100 are unlikely to even approximately satisfy the large margin low rank assumptions in Definition 3. Indeed, we were not able to fit a linear SVM with 100% training accuracy on the image space of CIFAR-10, even for binary classification, indicating that the pixel-space representations are not linearly separable.

Unlike images in the pixel space, linear SVM can easily achieve 100% training accuracy in the representation space for all pairs of classes from CIFAR-10 and for all of the pre-training strategies we consider. In order to check the low-dimension assumption, we first train a linear SVM w^* on the representation space and then, estimate $\xi_k = 1 - \left\|A_k A_k^T w^*\right\|$ (Definition 3). For a comprehensive comparison, we also compute ξ_k when w^* is trained on the pixel space (as mentioned before this does not yield 100% accuracy). Figure 5 plots ξ_k as a function of k and shows that even upon relaxing the hard margin criterion, images in the pixel space are much worse than the representation space when it comes to satisfying the low-dimension assumption of Definition 3. Therefore, we apply Algorithm 1 on top of the representations of these pre-trained models.

B PROOF OF MAIN RESULT

B.1 PROOF OF THEOREM 1

Theorem 1. *For $\gamma_0 \in (0,1), \xi_0 \in [0,1)$, let $\mathcal{D}_{\gamma_0,\xi_0}$ be the family of distributions consisting all (γ, ξ_k)-large margin low rank distributions over $\mathcal{X}_d \times \mathcal{Y}$ with $\gamma \geq \gamma_0$ and $\xi_k \leq \xi_0$. For any $\alpha \in (0,1), \beta \in (0, 1/4)$, $\epsilon \in (0, 1/\sqrt{k})$ and $\delta \in (0,1)$, Algorithm $\mathcal{A}_{\epsilon,\delta}(k,\zeta)$ is an $(\alpha, \beta, \epsilon, \delta)$-semi-private learner of the linear halfspace $\mathcal{H}_L^d$ on $\mathcal{D}_{\gamma_0,\xi_0}$ with sample complexity*

$$n^U = O\left(\frac{\log 2/\beta}{\gamma_0^2 \Delta_k^2}\right), n^L = \widetilde{O}\left(\frac{\sqrt{k}}{\alpha\epsilon\zeta}\right)$$

where $\zeta = \gamma_0 (1 - 1\xi_0 - 0.1\gamma_0)$.

Proof. **Privacy guarantee** Algorithm $\mathcal{A}_{\epsilon,\delta}(k,\zeta)$ computes the transformation matrix $\widehat{A}_k$ on the public unlabelled dataset. This step is independent of the labelled data S^L and has no impact on the privacy with respect to S^L. Noisy SGD ensures the operations on the labelled dataset S^L to output $\hat{w}_{\widehat{A}_k}$ is (ϵ, δ)-DP with respect to S^L. The final output $\hat{w} = \widehat{A}_k \hat{w}_{\widehat{A}_k}$ is attained by post-processing of $\hat{w}_{\widehat{A}_k}$ and preserves the privacy with respect to S^L by Lemma 1.

Accuracy guarantee For any distribution $D_{\gamma,\xi_k} \in \mathcal{D}_{\gamma_0,\xi_0}$, it is (γ, ξ_k)-large margin low rank distribution for some $\gamma \geq \gamma_0, \xi_k \leq \xi_0$. Let the empirical covariance matrix of D_{γ,ξ_k} calculated with the unlabelled dataset S^U be $\widehat{\Sigma} = \frac{1}{n^U} \sum_{x \in S^U} (x - \bar{x})(x - \bar{x})^T$ and $\widehat{A}_k \in \mathbb{R}^{d \times k}$ be the projection

matrix whose i^{th} column is the ith eigenvector $\widehat{\Sigma}$. Let Σ be the population covariance matrix and similarly, let A_k the matrix of top k eigenvectors of Σ.

Then, the margin is lower bounded by $\frac{y\langle \widehat{A}_k^T z, \widehat{A}_k^T w^* \rangle}{\|\widehat{A}_k^T z\|_2 \|\widehat{A}_k^T w^*\|_2}$ for all $z \in supp(D_{X,(\gamma,\xi_k)})$, where $D_{X,(\gamma,\xi_k)}$ is the marginal distribution of D_{γ,ξ_k}. We first consider the case $y = 1$ and let $z = a_z w^* + b^\perp$ where $b^\perp$ is in the nullspace of w^*. The lower bound for margin can be written as

$$\begin{aligned} \frac{y\left\langle \widehat{A}_k^T z, \widehat{A}_k^T w^* \right\rangle}{\left\|\widehat{A}_k^T z\right\|_2 \left\|\widehat{A}_k^T w^*\right\|_2} &= \frac{\left\langle \widehat{A}_k^T \left(a_z w^* + b^\perp\right), \widehat{A}_k^T w^* \right\rangle}{\left\|\widehat{A}_k^T z\right\|_2 \left\|\widehat{A}_k^T w^*\right\|_2} \\ &\overset{(a)}{=} \frac{a_z \left\|\widehat{A}_k^T w^*\right\|_2^2}{\left\|\widehat{A}_k^T z\right\|_2 \left\|\widehat{A}_k^T w^*\right\|_2} \overset{(b)}{\geq} a_z \left\|\widehat{A}_k^T w^*\right\|_2 \end{aligned} \tag{2}$$

where step (a) is due to $\langle w^*, b^\perp \rangle = 0$ and step (b) follows from $\left\|\widehat{A}_k^T z\right\|_2 \leq \left\|\widehat{A}_k\right\| \|z\|_2 \leq 1$.

It lefts to lower bound a_z and $\left\|\widehat{A}_k w^*\right\|$. First, we can lower bound a_z with the large-margin property in Definition 3 as

$$\begin{aligned} a_z &= \frac{\langle w^*, z \rangle}{\|w^*\|_2 \|z\|_2} \geq \gamma \geq \gamma_0 \qquad \text{for } y = 1 \\ a_z &= \frac{\langle w^*, z \rangle}{\|w^*\|_2 \|z\|_2} \leq -\gamma \leq -\gamma_0 \quad \text{for } y = -1 \end{aligned} \tag{3}$$

Using Lemma 3, we can bound the difference between the population and empirical eigenspace space with probability $1 - \frac{\beta}{2}$,

$$\left\|A_k A_k^T - \widehat{A}_k \widehat{A}_k^T\right\| \leq \frac{4\left(1 + \sqrt{\frac{\log(2/\beta)}{2}}\right)}{\left(\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)\right)\sqrt{n^U}} \leq \frac{\gamma_0}{10}. \tag{4}$$

where the last inequality follows from choosing the size of unlabelled data $n^U \geq \frac{1600\left(1+\sqrt{\frac{\log(2/\beta)}{2}}\right)^2}{\gamma_0^2 (\Delta_{\min}\lambda_k)^2}$.

Then, we can derive a lower bound for bound $\left\|\widehat{A}_k \widehat{A}_k^T w^*\right\|_2$,

$$\begin{aligned} \left\|A_k A_k^T w^*\right\|_2 - \left\|\widehat{A}_k \widehat{A}_k^T w^*\right\|_2 &\leq \left\|\widehat{A}_k \widehat{A}_k^T w^* - A A^T w^*\right\|_2 && \text{by Triangle Inequality} \\ &\leq \left\|\widehat{A}_k \widehat{A}_k^T - A A^T\right\| \|w^*\|_2 && \text{by Cauchy Schwarz Inequality} \\ &= \left\|\widehat{A}_k \widehat{A}_k^T - A A^T\right\| && \|w^*\| = 1 \\ &\leq \frac{\gamma_0}{10} && \text{by Equation (4)} \end{aligned} \tag{5}$$

Rearrange Equation (5) and by the low-rank property in Definition 3, we write the lower bound for $\left\|\widehat{A}_k w^*\right\|_2$,

$$\left\|\widehat{A}_k \widehat{A}_k^T w^*\right\|_2 \geq \left\|A_k A_k^T w^*\right\|_2 - \frac{\gamma_0}{10} = 1 - \xi_k - \frac{\gamma_0}{10} \geq 1 - \xi_0 - \frac{\gamma_0}{10}. \tag{6}$$

Plugging Equation (6) and Equation (3) into Equation (2), we derive a lower bound on the margin if $y = 1$,

$$\frac{y\left\langle \widehat{A}_k^T z, \widehat{A}_k^T w^* \right\rangle}{\left\|\widehat{A}_k^T z\right\|_2 \left\|\widehat{A}_k^T w^*\right\|_2} \geq \gamma_0 \left(1 - \xi_0 - \frac{\gamma_0}{10}\right). \tag{7}$$

Similarly, for $y = -1$,

$$\frac{y\left\langle \widehat{A}_k^T z, \widehat{A}_k^T w^* \right\rangle}{\left\|\widehat{A}_k^T z\right\|_2 \left\|\widehat{A}_k^T w^*\right\|_2} \geq \gamma_0 \left(1 - \xi_0 - \frac{\gamma_0}{10}\right). \tag{8}$$

The margin in the transformed space is at least $\gamma_0\left(1 - \xi_0 - \gamma_0/10\right)$. Thus, the loss function ℓ defined with in Equation (1) is $\frac{1}{\gamma_0(1-\xi_0-\gamma_0/10)}$-Lipschitz. For a hypothesis $w \in B_2^d$, denote the empirical loss of a dataset S by $\hat{L}(w; S) = \frac{1}{|S|}\sum_{(x,y)\in S} \ell(w, (x, y))$ and the loss on the distribution D by $L(w; D) = \mathbb{P}_{(x,y)\sim D}\left[\ell(w, (x, y))\right]$. By the convergence bound on empirical loss function of Noisy SGD (Lemma 4), we have with probability $1 - \frac{\beta}{4}$, Algorithm 1 outputs a hypothesis $\hat{w} \in B_2^d$ such that

$$\hat{L}(\hat{w}; S^L) - \hat{L}(w^*; D) = \hat{L}(\hat{w}; S^L) = \tilde{O}\left(\frac{\sqrt{k}}{n^L \epsilon \gamma_0 \left(1 - \xi_0 - 0.1\gamma_0\right)}\right) \tag{9}$$

where $w^* = \text{argmin}_{w\in B_2^d} \hat{L}(w; S^L)$ and $\hat{L}(w^*; S^L) = 0$ for low rank large margin distributions.

Then, we can bound the empirical 0-1 error by the empirical loss function ℓ. For $n^L = \tilde{O}\left(\frac{\sqrt{k}}{\alpha\beta\gamma_0(1-\xi_0-0.1\gamma_0)}\right)$, with probability $1 - \frac{\beta}{4}$,

$$\frac{1}{n^L}\sum_{(x,y)\in S^L} \mathbb{I}\{y\left\langle x, \hat{w}\right\rangle < 0\} \leq \hat{L}(\hat{w}; S^L) = \tilde{O}\left(\frac{\sqrt{k}}{n^L \epsilon \gamma_0 \left(1 - \xi_0 - 0.1\gamma_0\right)}\right) \leq \frac{\alpha}{2} \tag{10}$$

It remains to bound the generalization error of linear halfspace $\mathcal{H}_L^d$. That is, we still need to show that the empirical error of a linear threshold function is a good approximation of the error on the distribution. To achieve this, we can apply the generalization bound in terms of the growth function (Lemma 6). As the Vapnik-Chervonenkis (VC) dimension of k-dimensional linear halfspace is $k + 1$, its growth function is bounded by $\Pi(2n^L) \leq (2n)^{k+1} + 1$. For $n^L = \tilde{O}\left(\frac{k}{\alpha}\right) \leq \tilde{O}\left(\frac{\sqrt{k}}{\alpha\epsilon\gamma_0(1-\xi_0-\gamma_0/10)}\right)$, with probability $1 - \frac{\beta}{4}$, we have

$$\mathbb{P}_{(x,y)\sim D}\left[y\left\langle \hat{w}, x\right\rangle < 0\right] - \frac{1}{n^L}\sum_{(x,y)\in S^L} \mathbb{I}\{y\left\langle \hat{w}, x\right\rangle \leq 0\} \leq \frac{\alpha}{2}.$$

Thus,

$$\mathbb{P}_{(x,y)\sim D}\left[y\left\langle \hat{w}, x\right\rangle < 0\right] \leq \frac{1}{n^L}\sum_{(x,y)\in S^L} \mathbb{I}\{y\left\langle \hat{w}, x\right\rangle \leq 0\} + \frac{\alpha}{2} = \alpha.$$

This concludes the proof. □

B.2 Useful Lemmas

B.2.1 Differential Privacy

Lemma 1 (Post-processing [16]). *For every (ϵ, δ)-DP algorithm $\mathcal{A} : S \to \mathcal{Y}$ and every (possibly random) function $f : \mathcal{Y} \to \mathcal{Y}'$, $f \circ \mathcal{A}$ is (ϵ, δ)-DP.*

Definition 4 (η-TV tolerant $(\alpha, \beta, \epsilon, \delta)$-semi-private learner on a family of distributions $\mathcal{D}$). *An algorithm $\mathcal{A}$ is an η-TV tolerant $(\alpha, \beta, \epsilon, \delta)$-semi-privately learner for a hypothesis class $\mathcal{H}$ on a family of distributions $\mathcal{D}$ if for any distribution $D^L \in \mathcal{D}$, given a labelled dataset S^L of size n^L sampled i.i.d. from D^L and an unlabelled dataset S^U of size n^U sampled i.i.d. from any distribution D^U with η-bounded TV distance from D_X^L and third moment bounded by η, then $\mathcal{A}$ is (ϵ, δ)-DP with respect to S^L and outputs a hypothesis h satisfying*

$$\mathbb{P}[\mathbb{P}_{(x,y)\sim D}\left[h(x) \neq y\right] \leq \alpha] \geq 1 - \beta,$$

where the outer probability is over the randomness of the samples and the intrinsic randomness of the algorithm. The sample complexity n^L and n^U are polynomial in $\frac{1}{\alpha}$ and $\frac{1}{\beta}$, and n^L is also polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

B.2.2 LEMMAS FOR CONVERGENCE OF EIGENSPACE

Lemma 2 ([41] Theorem 3)**.** *Let $A \in \mathbb{R}^d$ be a symmetric positive definite matrix with nonzero eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_d$. Let $k > 0$ be an integer such that $\lambda_k > 0$. Let $B \in \mathbb{R}^d$ be another symmetric positive definite matrix such that $\|B\| < \frac{1}{4}(\lambda_k - \lambda_{k+1})$ and $A + B$ is still a positive definite matrix. Let $P_k(A), P_k(A+B)$ be the matrices whose columns consists of the first k eigenvectors of $A, A+B$, then*

$$\left\|P_k(A)P_k(A)^T - P_k(A+B)P_k(A+B)^T\right\| \leq \frac{2\|B\|}{\lambda_k - \lambda_{k+1}}.$$

Lemma 3 ([41] Theorem 4)**.** *Let D be a distribution over $\{x \in \mathbb{R}^d | \ \|x\|^2 \leq 1\}$ with covariance matrix Σ and zero mean $\mathbb{E}_{x\sim D}[x] = 0$. For a sample S of size n from D, let $\widehat{\Sigma} = \frac{1}{n}\sum_{x\in S} xx^T$ be the empirical covariance matrix. Let $A_k, \widehat{A}_k$ be the matrices whose columns are the first k eigenvectors of Σ and $\widehat{\Sigma}$ respectively and let $\lambda_1(\Sigma) > \lambda_2(\Sigma) > \ldots > \lambda_d(\Sigma)$ be the ordered eigenvalues of Σ. For any $k > 0, \beta \in (0,1)$ such that $\lambda_k(\Sigma) > 0$ and $n \geq \frac{16\left(1+\sqrt{\beta/2}\right)^2}{(\lambda_k(\Sigma)-\lambda_{k+1}(\Sigma))^2}$, we have that with probability at least $1 - e^{-\beta}$,*

$$\left\|A_kA_k^T - \widehat{A}_k\widehat{A}_k^T\right\| \leq \frac{4\left(1+\sqrt{\frac{\beta}{2}}\right)}{(\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma))\sqrt{n}}. \tag{11}$$

B.2.3 NOISY SGD

Next, we present the Noisy SGD algorithm from [7], which has theoretical guarantees stated in Lemma 4.

Algorithm 2 $\mathcal{A}_{Noisy-SGD}$

input a hypothesis space $\mathcal{W}$, a labelled dataset S^L of size n^L, a loss function ℓ, privacy parameters ϵ, δ and the learning rate functions $\eta : \mathbb{Z} \to \mathbb{R}$.

1: Set noise variance $\sigma^2 \leftarrow \frac{32L^2\left(n^L\right)^2\log(n^L/\delta)\log(1/\delta)}{\epsilon}$.
2: randomize $\hat{w}^0 \in \mathcal{W}$.
3: **for** $t = 1$ to $n^2 - 1$ **do**
4: Uniformly choose $(x, y) \in S^L$.
5: Update $\hat{w}^{t+1} = \Pi_{\mathcal{W}}\left(\hat{w}_t - \eta(t)[n\nabla\ell(\hat{w}^t; (x, y)) + \xi]\right)$ where $\xi \sim N(0, \mathbb{I}_d\sigma^2)$.
6: **end for**

output $\hat{w} = \hat{w}^{n^2}$

Lemma 4 (Empirical accuracy guarantee of Noisy SGD)**.** *[7] Let the loss function ℓ be L-Lipschitz and $\mathcal{C}$ be a d-dimensional convex space with diameter $\|\mathcal{C}\|$. If the learning rate function $\eta(t) = O\left(\frac{L^2n^2\log(n^L/\delta)\log(1/\delta)}{\epsilon^2}\right)$, then $\mathcal{A}_{Noisy-SGD}$ is (ϵ, δ)-DP, and it outputs $\hat{w}$ that satisfies the following excess risk bound*

$$\mathbb{E}\left[\sum_{(x,y)\in S}\ell(\hat{w}, (x, y)) - \sum_{(x,y)\in S}\ell(w^*, (x, y))\right] = O\left(\frac{L\|C\|\log^{3/2}(n/\delta)\sqrt{d\log(1/\delta)}}{\epsilon}\right)$$

for a labelled dataset S of size n^L. Here, w^* is the empirical risk minimizer $w^* = argmin_{w\in\mathcal{C}}\sum_{(x,y)\in S}\ell(w, (x, y))$.

B.2.4 CONVERGENCE OF SECOND MOMENT FOR η-BOUNDED TV DISTRIBUTIONS

Lemma 5. *Let f and g be the Probability Density Functions (PDFs) of two zero-mean distributions F and G over $\mathcal{X}$ with covariance matrices Σ_f and Σ_g respectively. Assume the spectral norm of the third moments of both F and G are bounded by η. If the total variation between the two distributions is bounded by η,i.e.$TV(f, g) = \max_{A\subset\mathcal{X}}|f(A) - g(A)| \leq \eta$, then the discrepancy in the covariance matrices is bounded by 7η, i.e.$\|\Sigma_f - \Sigma_g\| \leq 7\eta$.*

Proof. We first approximate Moment Generating Functions (MGFs) of g and f by their first and second moments. Then, we express the error bound in this approximation by the error bound for Taylor expansion, for any $t \in \mathbb{R}^d$ with $\|t\|_2 > 0$,

$$\begin{aligned}
\left| M_f(t) - 1 + t^T \mathbb{E}_f[X] + \frac{t^T \Sigma_f t}{2} \right| &= \left| M_f(t) - 1 + \frac{t^T \Sigma_f t}{2} \right| \\
&\leq \frac{\mathbb{E}_f\left[e^{t^T x} x x^T x\right] \|t\|_2^3}{3!} \leq \frac{\mathbb{E}_f\left[x x^T x\right] e^{\|t\|_2} \|t\|_2^3}{3!} && \because e^{t^T x} \leq e^{\|t\|_2 \|x\|_2} \leq e^{\|t\|_2} \text{ for } x \in \\
&\leq \eta \|t\|_2^3 && \because e^{\|t\|_2} \leq 3! \text{ for } \|t\|_2 \leq 1
\end{aligned} \tag{12}$$

Similarly,

$$\left| M_g(t) - 1 + t^T \mathbb{E}_g[X] + \frac{t^T \Sigma_g t}{2} \right| \leq \eta \|t\|_2^3 \tag{13}$$

Rewrite Equation (12) and Equation (13), we can bound the terms $\frac{t^T \Sigma_f t}{2}$ and $\frac{t^T \Sigma_g t}{2}$

$$\begin{aligned}
1 - M_f(t) - \eta \|t\|_2^3 \leq \frac{t^T \Sigma_f t}{2} \leq 1 - M_f(t) + \eta \|t\|_2^3 \\
1 - M_g(t) - \eta \|t\|_2^3 \leq \frac{t^T \Sigma_g t}{2} \leq 1 - M_g(t) + \eta \|t\|_2^3
\end{aligned} \tag{14}$$

Next, we show that the discrepancy in covariance matrices of distributions G and F are upper bounded by the difference in their MGFs.

By Equation (14), for all $t \in \mathbb{R}^d$ and $\|t\|_2 \neq 0$,

$$\begin{aligned}
\left| \frac{t^T (\Sigma_f - \Sigma_g) t}{2} \right| &\leq 1 - M_f(t) + \eta \|t\|_2^3 - 1 + M_g(t) + \eta \|t\|_2^3 \\
&= \left| M_g(t) - M_f(t) + 2\eta \|t\|_2^3 \right| \\
&\leq |M_g(t) - M_f(t)| + 2\eta \|t\|_2^3
\end{aligned} \tag{15}$$

Next, we upper bound the difference between the MGFs of distributions G and F by the TV distance between them.

$$\begin{aligned}
|M_f(t) - M_g(t)| &= \left| \int_{x \in B_2^d} e^{t^T x} [f(x) - g(x)] \, dx \right| \\
&\leq \int_{x \in B_2^d} e^{t^T x} |f(x) - g(x)| \, dx \\
&\leq \int_{x \in B_2^d} e^{\|t\|_2 \|x\|_2} |f(x) - g(x)| \, dx \leq \frac{e^{\|t\|_2} \eta}{2}
\end{aligned} \tag{16}$$

where the last inequality follows as $\|x\|_2 = 1$ for $x \in B_2^d$ and $TV(f, g) \leq \eta$.

Combine Equation (15) and Equation (16), we have for all $t \in \mathbb{R}^d$ and $\|t\|_2 \neq 0$,

$$\left| t^T (\Sigma_f - \Sigma_g) t \right| \leq e^{\|t\|_2} \eta_1 + 4\eta \|t\|_2^3 \tag{17}$$

Choose t as a vector in the direction of the first eigenvector of $\Sigma_f - \Sigma_g$. For this t,

$$\left| t^T (\Sigma_f - \Sigma_g) t \right| = \|t\|_2 \|(\Sigma_f - \Sigma_g) t\|_2 = \|\Sigma_f - \Sigma_g\| \|t\|_2^2 \tag{18}$$

where the first equation is because t^T and $(\Sigma_f - \Sigma_g) t$ are linearly dependent. The second equation follows as

$$\begin{aligned}
\|(\Sigma_f - \Sigma_g) t\|_2^2 &= t^T (\Sigma_f - \Sigma_g)^T (\Sigma_f - \Sigma_g) t \\
&= t^T \|\Sigma_f - \Sigma_g\|^2 t = \|\Sigma_f - \Sigma_g\|^2 \|t\|_2^2
\end{aligned} \tag{19}$$

Plugging Equation (18) into Equation (17) and choose the norm of t as the minimizer of $e^{\|t\|_2}\eta_1 + 4\eta \|t\|_2^3$, we get

$$\|\Sigma_f - \Sigma_g\| \leq \min_{0 \leq \|t\|_2 \leq 1} \frac{e^{\|t\|_2}\eta}{\|t\|_2^2} + 4\eta \|t\|_2 \leq \frac{\eta(1 + \|t\|_2 + \|t\|_2^2)}{\|t\|_2^2} + 4\eta \|t\|_2 = 7\eta \tag{20}$$

□

B.2.5 CONCENTRATION BOUNDS

Lemma 6 (Convergence bound on generalization error)**.** *[5] Suppose $\mathcal{H}$ is a hypothesis class with instance space $\mathcal{X}$ and output space $\{-1, 1\}$. Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$ and S be a dataset of size n sampled i.i.d. from D. For $\eta \in (0, 1), \zeta > 0$, we have the following generalization bound*

$$\mathbb{P}_{S \sim D^n}\left[\sup_{h \in \mathcal{H}} L(h; D) - (1 + \zeta)\hat{L}(h; S) \leq \eta\right] \leq 4\Pi_{\mathcal{H}}(2n) \exp\left(-\frac{\eta\zeta n}{4(\zeta + 1)}\right),$$

where L and $\hat{L}$ are the population and the empirical 0-1 error and $\Pi_{\mathcal{H}}$ is the growth function of $\mathcal{H}$.

Lemma 7 (Concentration bound on the norm of Gaussian Random Vectors)**.** *Let $X \sim N(\mu, \Sigma)$, where $v \in B_d^2$. Then, with probability at least $1 - \delta$,*

$$\|X - \mu\|_2 \leq 4 \|\Sigma\|_{op} \sqrt{d} + 2 \|\Sigma\|_{op} \sqrt{\log \frac{1}{\delta}}. \tag{21}$$

C RESULT ON TOLERANCE TO DISTRIBUTION SHIFT

Next, we show a stronger version of Theorem 1 which allows a distribution shift between the labelled and unlabelled data. This is a reasonable relaxation in the real world. If part of a dataset is made public either due to the expiry of statute of limitations or voluntarily by the user, it is reasonable to expect that the public and the private dataset will exhibit a distribution shift. Under this setting, we present similar theoretical guarantees in Theorem 2. We say that two distributions D_X^L and D^U have η-bounded Total Variation (TV) distance if

$$TV(D^U, D_X^L) = \sup_{A \subset \mathcal{X}} |D^U(A) - D_X^L(A)| = \eta.$$

Next, we extend the definition of semi-private learner to this setting. Informally, an algorithm $\mathcal{A}$ is an *η-TV tolerant $(\alpha, \beta, \epsilon, \delta)$-semi-private learner* if it is a $(\alpha, \beta, \epsilon, \delta)$-semi-private learner when the labelled and unlabelled dataset have η-bounded TV.

Theorem 2. *For $\gamma_0 \in (0, 1), \xi_0 \in [0, 1)$, let $\mathcal{D}_{\gamma_0, \xi_0}$ be the family of distributions consisting of all (γ, ξ_k)-large margin low rank distributions over $\mathcal{X}_d \times \mathcal{Y}$ with $\gamma \geq \gamma_0$ and $\xi_k \leq \xi_0$ and small third moment. For any $\alpha \in (0, 1), \beta \in (0, 1/4)$, $\epsilon \in (0, 1/\sqrt{k})$ and $\delta \in (0, 1)$, Algorithm $\mathcal{A}_{\epsilon,\delta}(k, \zeta)$ is an η-TV tolerant $(\alpha, \beta, \epsilon, \delta)$-semi-private learner of the linear halfspace $\mathcal{H}_L^d$ on $\mathcal{D}_{\gamma_0, \xi_0}$ with sample complexity*

$$n^U = O\left(\frac{\log \frac{2}{\beta}}{(\gamma_0 \Delta_k)^2}\right), n^L = \tilde{O}\left(\frac{\sqrt{k}}{\alpha\epsilon\zeta}\right)$$

where $\zeta = \gamma_0(1 - \xi_0 - 0.1\gamma_0 - 7\eta/\Delta_k)$.

Also, the labelled sample complexity bound here is pessimistic and becomes less powerful for large distribuiton shift in terms of TV distance. With better specifications on the family of distribution and the type of shifts, a much tighter bound is expected, which calls for future research.

A formal definition of η-TV tolerant $(\alpha, \beta, \epsilon, \delta)$-learner and detailed versions with proofs of the theorems can be found in Appendix B.

Theorem 3. *For $\gamma_0 \in (0, 1), \xi_0 \in [0, 1)$, let $\mathcal{D}_{\gamma_0, \xi_0}$ be the family of distributions consisting of all (γ, ξ_k)-large margin low rank distributions over $\mathcal{X}_d \times \mathcal{Y}$ with $\gamma \geq \gamma_0$ and $\xi_k \leq \xi_0$ and third moment bounded by η. For any $\alpha \in (0, 1), \beta \in (0, 1/4)$, $\epsilon \in (0, 1/\sqrt{k})$ and $\delta \in (0, 1)$, Algorithm $\mathcal{A}_{\epsilon,\delta}(k, \zeta)$ is*

an η-TV tolerant $(\alpha, \beta, \epsilon, \delta)$-semi-private learner of the linear halfspace $\mathcal{H}_L^d$ on $\mathcal{D}_{\gamma_0,\xi_0}$ with sample complexity

$$n^U = O\left(\frac{\log \frac{2}{\beta}}{(\gamma_0 \Delta_k)^2}\right), n^L = \tilde{O}\left(\frac{\sqrt{k}}{\alpha\epsilon\zeta}\right)$$

where $\Delta_k = \lambda_k\left(\Sigma^L\right) - \lambda_{k+1}\left(\Sigma^L\right)$ and $\zeta = \gamma_0(1 - \xi_0 - 0.1\gamma_0 - {}^{7\eta}/_{\Delta_k})$.

Proof. **Privacy Guarantee** Using a similar argument as in the proof to the privacy guarantee in Theorem 1, we can show that Algorithm $\mathcal{A}_{\epsilon,\delta}(k, \zeta)$ preserves (ϵ, δ)-DP on the labelled dataset S^L.

Accuracy Guarantee For any distribution $D^L_{\gamma,\xi_k} \in \mathcal{D}_{\gamma_0,\xi_0}$, it is (γ, ξ_k)-large margin low rank distribution for some $\gamma \geq \gamma_0$, $\xi_k \leq \xi_0$. For any unlabelled distribution D^U with η-bouneded TV distance from D^L_{γ,ξ_k}, let the empirical covariance matrix of D^U be $\widehat{\Sigma}^U = \frac{1}{n^U}\sum_{x\in S^U} xx^T$ and $\left(\widehat{A}^U_k\right) \in \mathbb{R}^{d\times k}$ be the projection matrix whose i^{th} column is the ith eigenvector $\widehat{\Sigma}^U$. Let Σ^L and Σ^U be the population labelled and unlabelled covariance matrix of D^L and D^U and similarly, let A^L_k and A^U_k be the matrices of top k eigenvectors of Σ^L and Σ^U respectively.

The margin is lower bounded by $\frac{y\left\langle \left(\widehat{A}^U_k\right)^T z, \left(\widehat{A}^U_k\right)^T w^* \right\rangle}{\left\|\left(\widehat{A}^U_k\right)^T z\right\|_2 \left\|\left(\widehat{A}^U_k\right)^T w^*\right\|_2}$ for all $z \in supp(D^L_{X,(\gamma,\xi_0)})$, where $supp(D^L_{X,(\gamma,\xi_0)})$ is the marginal distribution of D_{γ,ξ_k} over $\mathcal{X}$. We first consider the case $y = 1$ and let $z = a_z w^* + b^\perp$ where $b^\perp$ is in the nullspace of w^*. The lower bound for margin can be written as

$$\begin{aligned}\frac{y\left\langle \left(\widehat{A}^U_k\right)^T z, \left(\widehat{A}^U_k\right)^T w^* \right\rangle}{\left\|\left(\widehat{A}^U_k\right)^T z\right\|_2 \left\|\left(\widehat{A}^U_k\right)^T w^*\right\|_2} &= \frac{\left\langle \left(\widehat{A}^U_k\right)^T \left(a_z w^* + b^\perp\right), \left(\widehat{A}^U_k\right)^T w^* \right\rangle}{\left\|\left(\widehat{A}^U_k\right)^T z\right\|_2 \left\|\left(\widehat{A}^U_k\right)^T w^*\right\|_2} \\ &\overset{(a)}{=} \frac{a_z \left\|\left(\widehat{A}^U_k\right)^T w^*\right\|_2}{\left\|\left(\widehat{A}^U_k\right)^T z\right\|_2} \overset{(b)}{\geq} a_z \left\|\left(\widehat{A}^U_k\right)^T w^*\right\|_2\end{aligned} \tag{22}$$

where step (a) is due to $\left\langle w^*, b^\perp \right\rangle = 0$ and step (b) follows from $\left\|\left(\widehat{A}^U_k\right)^T z\right\|_2 \leq \left\|\left(\widehat{A}^U_k\right)\right\| \|z\|_2 \leq 1$.

It lefts to lower bound a_z and $\left\|\left(\widehat{A}^U_k\right)^T w^*\right\|_2$. First, we can lower bound a_z similarly as in proof to Theorem 1. With the large-margin property in Definition 3,

$$\begin{aligned} a_z &= \frac{\langle w^*, z\rangle}{\|w^*\|_2 \|z\|_2} \geq \gamma \geq \gamma_0 \qquad \text{for } y = 1 \\ a_z &= \frac{\langle w^*, z\rangle}{\|w^*\|_2 \|z\|_2} \leq -\gamma \leq -\gamma_0 \quad \text{for } y = -1 \end{aligned} \tag{23}$$

To lower bound $\left\|\left(\widehat{A}^U_k\right)^T w^*\right\|_2$, we first bound $\left\|\widehat{A}^U_k w^*\right\|_2$. The bounded total variation assumption between D^L and D^U implies bounded deviation in the second moment,

By Lemma 2, we can bound the difference in the projection space of A_k^L and A_k^U. With probability $1 - \beta/4$,

$$\begin{aligned}\left\| A_k^L \left(A_k^L\right)^T - A_k^U \left(A_k^U\right)^T \right\| &\leq \frac{2\left\|\Sigma^L - \Sigma^U\right\|}{\lambda_k\left(\Sigma^L\right) - \lambda_{k+1}\left(\Sigma^L\right)} \\ &\leq \frac{2\left\|\Sigma^L - \Sigma^U\right\|_F}{\lambda_k\left(\Sigma^L\right) - \lambda_{k+1}\left(\Sigma^L\right)} \\ &\overset{(a)}{\leq} \frac{7\eta}{\lambda_k\left(\Sigma^L\right) - \lambda_{k+1}\left(\Sigma^L\right)} = \frac{7\eta}{\Delta_k}.\end{aligned} \tag{24}$$

Step (a) follows by the assumption of bounded total variation between the labelled and unlabelled distributions and Lemma 5.

Now, we derive lower bounds on $\left\| A_k^U \left(A_k^U\right)^T w^* \right\|_2$,

$$\begin{aligned}\left\| A_k^L \left(A_k^L\right)^T w^* \right\|_2 - \left\| A_k^U \left(A_k^U\right)^T w^* \right\|_2 &\leq \left\| \left(A_k^U \left(A_k^U\right)^T - A_k^L \left(A_k^L\right)^T \right) w^* \right\|_2 \\ &\leq \left\| A_k^U \left(A_k^U\right)^T - A_k^L \left(A_k^L\right)^T \right\| \left\| w^* \right\|_2 \quad \text{by Cauchy Schwarz Inequality} \\ &= \left\| A_k^U \left(A_k^U\right)^T - A_k^L \left(A_k^L\right)^T \right\| \leq \frac{7\eta}{\Delta_k} \quad \text{By Equation (24) and } \left\| w^* \right\|_2 = 1\end{aligned} \tag{25}$$

Rearrange the terms and by the low-rank property of the labelled distribution D_{γ,ξ_k}, we get the lower bound on $\left\| A_k^U \left(A_k^U\right)^T w^* \right\|_2$,

$$\left\| A_k^U \left(A_k^U\right)^T w^* \right\|_2 \geq \left\| A_k^L \left(A_k^L\right)^T z \right\|_2 - \frac{7\eta}{\Delta_k} \geq 1 - \xi_k + \frac{7\eta}{\Delta_k} \geq 1 - \xi_0 + \frac{7\eta}{\Delta_k}. \tag{26}$$

Next, we lower bound $\left\| \left(\widehat{A}_k^U\right)^T w^* \right\|_2$. Similar to the proof for Theorem 1, by Lemma 3, if the number of unlabelled data is

$$n^U = O\left(\frac{\log \frac{2}{\beta}}{\left(\gamma_0 \Delta_k\right)^2} \right),$$

with probability $1 - \beta/4$, the difference in the eigenspace of A_k^U and $\widehat{A}_k^U$ is bounded,

$$\left\| A_k^U \left(A_k^U\right)^T - \left(\widehat{A}_k^U\right)\left(\widehat{A}_k^U\right)^T \right\| \leq \frac{\gamma_0}{10}. \tag{27}$$

Thus,

$$\left\| \left(\widehat{A}_k^U\right)\left(\widehat{A}_k^U\right)^T w^* \right\|_2 \geq \left\| A_k^U \left(A_k^U\right)^T w^* \right\|_2 - \frac{\gamma_0}{10} \geq 1 - \xi_0 - \frac{7\eta}{\Delta_k} - \frac{\gamma_0}{10}. \tag{28}$$

Plugging Equation (28) and Equation (23) into Equation (22), we derive a lower bound on the margin for $y = 1$,

$$\frac{y \left\langle \left(\widehat{A}_k^U\right)^T z, \left(\widehat{A}_k^U\right)^T w^* \right\rangle}{\left\| \left(\widehat{A}_k^U\right)^T z \right\|_2 \left\| \left(\widehat{A}_k^U\right)^T w^* \right\|_2} \geq \gamma_0 \left(1 - \xi_0 - \frac{7\eta}{\Delta_k} - \frac{\gamma_0}{10} \right). \tag{29}$$

Similarly, for $y = -1$,

$$\frac{y \left\langle \left(\widehat{A}_k^U\right)^T z, \left(\widehat{A}_k^U\right)^T w^* \right\rangle}{\left\| \left(\widehat{A}_k^U\right)_2^T z \right\| \left\| \left(\widehat{A}_k^U\right)_2^T w^* \right\|} \geq \gamma_0 \left(1 - \xi_0 - \frac{7\eta}{\Delta_k} - \frac{\gamma_0}{10} \right). \tag{30}$$

The margin in the transformed space is at least $\zeta = \gamma_0 \left(1 - \xi_0 - 7\eta/\Delta_k - 0.1\gamma_0\right)$. Thus, the loss function ℓ defined with in Equation (1) is $\frac{1}{\zeta}$-Lipschitz. Similar as in the proof for Theorem 1, we

define the empirical loss $\hat{L}(w; S)$ of a dataset S for a hypothesis $w \in B_2^d$ and the loss $L(w; D) = \mathbb{P}_{(x,y)\sim D}\left[\ell(w, (x, y))\right]$ on the distribution D. By the convergence bound on empirical loss function of Noisy SGD (Lemma 4), we have with probability $1 - \frac{\beta}{4}$, Algorithm 1 outputs a hypothesis $\hat{w} \in B_2^d$ such that

$$\hat{L}(\hat{w}; S^L) - \hat{L}(w^*; D) = \hat{L}(\hat{w}; S^L) = \tilde{O}\left(\frac{\sqrt{k}}{n^L \epsilon \zeta}\right) \tag{31}$$

where $w^* = \text{argmin}_{w \in B_2^d} \hat{L}(w; S^L)$ and $\hat{L}(w^*; S^L) = 0$ for low rank large margin distribution.

Then, we can bound the empirical 0-1 error by the empirical loss function ℓ. For $n^L = \tilde{O}\left(\frac{\sqrt{k}}{\alpha\beta\zeta}\right)$, with probability $1 - \frac{\beta}{4}$,

$$\frac{1}{n^L} \sum_{(x,y)\in S^L} \mathbb{I}\{y \langle \hat{w}, x\rangle < 0\} \leq \hat{L}(\hat{w}; S^L) = \tilde{O}\left(\frac{\sqrt{k}}{n^L \epsilon \zeta}\right) \leq \frac{\alpha}{2} \tag{32}$$

It remains to bound the generalization error of linear halfspace with Lemma 6. As the Vapnik-Chervonenkis (VC) dimension of k-dimensional linear halfspace is $k + 1$, its growth function is bounded by $\Pi(2n^L) \leq (2n)^{k+1} + 1$. Thus, for $n^L = \tilde{O}\left(\frac{k}{\alpha}\right) \leq \tilde{O}\left(\frac{\sqrt{k}}{\alpha\epsilon\zeta}\right)$, with probability $1 - \frac{\beta}{4}$, we have

$$\mathbb{P}_{(x,y)\sim D}\left[y \langle \hat{w}, x\rangle < 0\right] - \frac{1}{n^L} \sum_{(x,y)\in S^L} \mathbb{I}\{y \langle \hat{w}, x\rangle < 0\} \leq \frac{\alpha}{2}$$

Thus,

$$\mathbb{P}_{(x,y)\sim D}\left[y \langle \hat{w}, x\rangle < 0\right] \leq \frac{1}{n^L} \sum_{(x,y)\in S^L} \mathbb{I}\{y \langle \hat{w}, w\rangle < 0\} + \frac{\alpha}{2} = \alpha.$$

This concludes the proof. □

D Comparison with existing literature and discussion

We discuss the relevant works in the literature and how they compare with our result in this section.

Generic private algorithms Bassily et al. [7] proposed the Noisy SGD algorithm to privately learn linear thresholds with margin γ. However, their algorithm cannot use unlabelled data and requires $O\left(\sqrt{d}/\alpha\epsilon\gamma\right)$ labelled data. The generic semi-private learner in Alon et al. [4] leverages unlabelled data to reduce the infinite hypothesis class to a finite α-net and applies exponential mechanism [27] to achieve $(\epsilon, 0)$-DP. However, it still requires a dimension-dependent labelled sample complexity $O\left(d/\alpha\epsilon\right)$.

Other dimension reduction based private algorithms Johnson-Lindenstrauss (JL) transformation is another popular technique for dimensionality reduction. Perhaps, most relevant to our work, [29] reduces the dimension of a linear halfspace with margin γ from d to $O\left(1/\gamma\right)$ while preserving the margin in the lower-dimensional space. Private learning in the transformed low-dimensional space requires $O\left(1/\alpha\epsilon\gamma^2\right)$ labelled samples. Our algorithm removes the quadratic dependence on the margin but pays the price of requiring the linear separator to align with the top few principal components of the data. For example, a Gaussian mixture distribution (Definition 5) satisfies low-rank property with parameter $\xi_k = 0$ and $k = 1$. Corollary 5 shows that our algorithm requires labelled sample complexity$O(1/\alpha\epsilon\gamma)$ instead of $O(1/\alpha\epsilon\gamma^2)$ required by Nguyen et al. [29].

Another approach to circumvent the dependency on the dimension is to apply dimension reduction techniques directly to the gradients. For smooth loss functions with ρ-Lipschitz and G-bounded gradients, Zhou et al. [40] showed that applying PCA in the gradient space of DP-SGD [3] achieves dimension-independent labelled sample complexity $O\left(\frac{k\rho G^2}{\alpha\epsilon} + \frac{\rho^2 G^4 \log d}{\alpha}\right)$. However, this algorithm is computationally costly as it applies PCA in every gradient-descent step to a matrix whose size scales with the number of parameters. Low-rank reparametrization in the parameter space [39] is computationally efficient and has gained empirical success in text and vision datasets. However,

their algorithm is unable to use unlabelled public data and their guarantee for linear halfspaces is not immediately clear.

Non-private learning It is interesting to note that our algorithm may not lead to a similar improvement in the non-private case. We show a dimension-independent Rademacher-based labelled sample complexity bound for non-private learning of linear halfspaces. We use a non-private version of Algorithm 1 by replacing Noisy-SGD with Gradient Descent using the same loss function. As before, for any $\gamma_0 \in (0,1), \xi_0 \in (0,1)$, let $\mathcal{D}_{\gamma_0,\xi_0}$ be the family of distributions consisting of all (γ, ξ_k)-large margin low rank distributions with $\gamma \geq \gamma_0$ and $\xi_k \leq \xi_0$.

Proposition 4 (Non-DP learning)**.** *For any $\alpha, \beta \in (0, 1/4)$, and distribution $D \in \mathcal{D}_{\gamma_0,\xi_0}$, given a labelled dataset of size $\tilde{O}\left(1/\zeta\alpha^2\right)$ and unlabelled dataset of size $O\left(\log \frac{2}{\beta}/(\gamma_0\Delta_k)^2\right)$, the non-private version of $\mathcal{A}(k,\zeta)$ produces a linear classifier $\hat{w}$ such that with probability $1-\beta$*

$$\mathbb{P}_D\left[y\left\langle \hat{w}, x\right\rangle < 0\right] < \alpha$$

where $\zeta = \gamma_0(1 - \xi_0 - 0.1\gamma_0)$.

The labelled sample complexity in the above result shows that non-private algorithms do not significantly benefit from decreasing dimensionality[1]. In summary, this section has showed that our computationally efficient algorithm, under certain assumptions, on the data can yield dimension independent private sample complexity. In the result of the paper, we show through a wide variety of experiments that the results transfer to practice in both common benchmarks as well as many newly designed challenging settings.

E AN ILLUSTRATION USING A SIMPLE DISTRIBUTION

As an example, we define (θ, σ^2)-Large Margin Gaussian mixture distributions in Definition 5.

Definition 5. *A distribution D over $\mathcal{X} \times \mathcal{Y}$ is a (θ, σ^2)-Large margin Gaussian mixture distribution if there exists $w^*, \mu \in B_2^d$, such that $\langle \mu, w^* \rangle = 0$, the conditional distribution $X|y$ is distributed according to a normal distribution with mean μy and covariance matrix $\theta w^* \left(w^*\right)^T + \sigma^2 I_d$ and $y \in \{-1, 1\}$ is distributed uniformly.*

For any $\theta, \sigma^2 = O(1/\sqrt{d})$, it is easy to see that this family of distributions satisfies the large margin low rank properties in Definition 3 for $k = 2$ and $\xi = 0$.

For large margin Gaussian mixture distributions, Theorem 1 implies that the dimensionality reduction through PCA on the public dataset leads to a drop in the labelled sample complexity from $O(\sqrt{d})$ to $O(1)$ as shown in Corollary 5.

Corollary 5 (Theoretical guarantees for large margin Gaussian mixture distribution)**.** *For any $\theta, \sigma^2 = O\left(\frac{1}{\sqrt{d}}\right)$, let $\mathcal{D}_{\theta,\sigma^2}$ be the family of all (θ, σ^2)-large margin Gaussian mixture distributions. For any $\alpha \in (0,1), \beta \in (0, 1/4)$, $\epsilon \in (0,1)$ and $\delta \in (0,1)$, Algorithm $\mathcal{A}_{\epsilon,\delta}(k,\zeta)$ is an $(\alpha, \beta, \epsilon, \delta)$-semi-private learner of linear halfspaces $\mathcal{H}_L^d$ with sample complexity*

$$n^U = O\left(\frac{\log 2/\beta}{\gamma^2\theta^2}\right), \;\; n^L = \tilde{O}\left(\frac{L}{\alpha\epsilon\gamma(1-0.1\gamma)}\right) \tag{33}$$

where $\gamma = 1 - C\sqrt{d}\left(\theta + \sigma^2\right)$ and $L = 1 + C\sqrt{d}\left(\theta + \sigma^2\right)$ for some constant C.

Here, in line with the notation of Definition 3, γ intuitively represents the margin in the d-dimensional space and L is the upper bound for the radius of the labelled dataset. For $\theta = \sigma^2 = 1/2C\sqrt{d}$, we get $L = 1.5$ and $\gamma = 0.5$; Corollary 5 implies the labelled sample complexity $\tilde{O}\left(1/\alpha\epsilon\right)$.

Corollary 6 (Theoretical guarantees for large margin Gaussian mixture distribution)**.** *For $\theta, \sigma^2 = \tilde{O}\left(1/\sqrt{d}\right)$, let $\mathcal{D}_{\theta,\sigma^2}$ be the family of all (θ, σ^2)-large margin Gaussian mixture distribution (Definition 5). For any $\alpha \in (0,1), \beta \in (0, 1/4), \epsilon \in (0, 1/\sqrt{L})$, and $\delta \in (0,1)$, Algorithm*

[1]However, this bound uses a standard rademacher complexity result and may be lose. A tighter complexity bound may yield some dependence on the projected dimension

$\mathcal{A}_{\epsilon,\delta}(k, \gamma(1-0.1\gamma))$ is an $(\alpha, \beta, \epsilon, \delta)$-semi-private learner on D_{theta,σ^2} of linear halfspaces $\mathcal{H}_L^d$

$$\begin{aligned} n^U &= O\left(\frac{\log \frac{2}{\beta}}{\gamma^2\theta^2}\right), \\ n^L &= \tilde{O}\left(\frac{L\sqrt{k}}{\alpha\epsilon\gamma(1-0.1\gamma)}\right) \end{aligned} \tag{34}$$

where $\gamma = 1 - \left(4\sqrt{d} + 2\sqrt{\log \frac{2n^L}{\delta}}\right)(\sigma^2+\theta)$, $L = 1 + \left(4\sqrt{d} + 2\sqrt{\log \frac{2n^L}{\delta}}\right)(\sigma^2+\theta)$.

Proof. For $i \in \{1,2\}$, let $f_G(x;\mu,\Sigma_i)$ be the probability density function of a Gaussian random variable with mean μ_i and covariance matrix Σ_i. Then, for any function $D \in \mathcal{D}_{\theta,\sigma^2}$, we can calculate the covariance matrix of its marginal distribution D_X as

$$\begin{aligned} \Sigma_X &= \frac{1}{2}\int_{B_2^d} xx^T f_G(x;\mu_1,\Sigma)dx + \frac{1}{2}\int_{B_2^d} xx^T f_G(x;\mu_2,\Sigma)dx \\ &\overset{(a)}{=} \frac{1}{2}\left(\Sigma + \mu_1\mu_1^T\right) + \frac{1}{2}\left(\Sigma + \mu_2\mu_2^T\right) \\ &= \theta w^* \left(w^*\right)^T + \mu_1\mu_1^T + \sigma^2 I_d \qquad \because \mu_1 = -\mu_2 \end{aligned} \tag{35}$$

where equation (a) follows by the relationship between covaraince matrix and the second moment $\Sigma = \mathbb{E}_X\left[XX^T\right] + \mu\mu^T$.

By the definition of Σ_X, we first show that the first two eigenvectors are μ_1 and w^* with the corresponding eigenvalues $1+\sigma^2$ and $\theta+\sigma^2$ as $\theta = O\left(1/\sqrt{d}\right) \leq 1$. The rest non-spiked eigenvalues of the eigenvalues are σ^2.

$$\begin{aligned} \Sigma_X w^* &= \theta w^* \left(w^*\right)^T w^* + \mu_1\mu_1^T w^* + \sigma^2 w^* \\ &= (\theta+\sigma^2)w^* \qquad \because \left(w^*\right)^T \mu_1 = 0 \\ \Sigma_X \mu_1 &= \theta w^* \left(w^*\right)^T \mu_1 + \mu_1\mu_1^T\mu_1 + \sigma^2\mu_1 \\ &= (\|\mu_1\|^2 + \sigma^2)\mu_1 = (1+\sigma^2)\mu_1 \qquad \because \left(w^*\right)^T \mu_1 = 0 \end{aligned}$$

Thus, with $k=2$, we can show that the low dimension condition in Definition 3. Also, we can show that the low dimensional parameter ξ equals 0,

$$\begin{aligned} \frac{\left\|A_k^T w^*\right\|}{\|w^*\|} &= \frac{1}{\|w^*\|}\begin{bmatrix} \mu_1^T \\ \left(w^*\right)^T \end{bmatrix} w^* \\ &= \frac{\left\|\mu_1^T w^* + \left(w^*\right)^T w^*\right\|}{\|w^*\|} \\ &= 1 = 1 - \xi \end{aligned} \tag{36}$$

Also, we can calculate Δ_k for $k=2$,

$$\Delta_k = \lambda_k\left(\Sigma_X\right) - \lambda_{k+1}\left(\Sigma_X\right) = \theta + \sigma^2 - \sigma^2 = \theta \tag{37}$$

To apply Theorem 1, we show that the labelled dataset lies in a ball with bounded radius with high probability.

Denote the labelled dataset from the subpopulation with $y=1$ by S_1^L and denote the labelled dataset from the subpopulation with $y=-1$ by S_2^L. For any $x \in S_i^L$ for $i=1,2$, by Lemma 7, for some $\frac{\beta}{2n^L} > 0$,

$$\mathbb{P}_{S^L \sim D^{n^L}}\left[\|x-\mu_i\|_2 \leq 4\left(\theta+\sigma^2\right)\sqrt{d} + 2\left(\theta+\sigma^2\right)\sqrt{\log\frac{4n^L}{\beta}}\right] \geq 1 - \frac{\beta}{4n^L}$$

For $i \in \{1, 2\}$, by applying Union bound on all $x \in S_i^L$, we can bound maximum distance of a points $x \in S_i^L$ to the center μ_i,

$$\mathbb{P}_{S^L \sim D^{n^L}} \left[\max_{x \in S_i^L} \|x - \mu_i\|_2 \leq \left(\theta + \sigma^2\right) \left(4\sqrt{d} + 2\sqrt{\log \frac{4n^L}{\beta}} \right) \right] \leq 1 - \frac{\beta}{4}. \tag{38}$$

Note that the distance between the two centers μ_1 and μ_2 is 2. Thus, with probability at least $1 - \frac{\beta}{2}$, all points in the labelled dataset S^L lie in a ball centered at 0 having radius

$$L = 1 + \left(4\sqrt{d} + 2\sqrt{\log \frac{4n^L}{\beta}} \right) \left(\sigma^2 + \theta\right). \tag{39}$$

Also, the margin in the labelled dataset is at least

$$\gamma = 1 - \left(4\sqrt{d} + 2\sqrt{\log \frac{4n^L}{\beta}} \right) \left(\sigma^2 + \theta\right). \tag{40}$$

Following a similar argument as in the proof to Theorem 1, we can show that for $n^U = O\left(\frac{\log 4/\beta}{\gamma^2 \theta^2}\right)$, with probability at least $\frac{\beta}{4}$ over the randomness of the unlabelled dataset, for all points in the labelled dataset, the margin in the 2-dimensional transformed space $\{\widehat{A}_k x | (x, y) \in S^L\}$ is at least $\gamma(1 - 0.1\gamma)$. For a hypothesis $w \in B_2^d$, denote the empirical loss of a dataset S by $\hat{L}(w; S) = \frac{1}{|S|} \sum_{(x,y) \in S} \ell(w, (x, y))$ and the loss on the distribution D by $L(w; D) = \mathbb{P}_{(x,y) \sim D} \left[\ell(w(x, y))\right]$. By the convergence bound on empirical loss fucntion of Noisy SGD (Lemma 4), we have with probability $1 - \frac{\beta}{8}$, Algorithm 1 outputs a hypothesis $\hat{w}$ such that

$$\hat{L}(\hat{w}, S^L) - \hat{L}(w^*; D) = \hat{L}(\hat{w}, S^L) = \tilde{O}\left(\frac{\sqrt{k}}{n^L \epsilon \gamma \left(1 - 0.1\gamma\right)} \right), \tag{41}$$

where $w^* = \text{argmin}_{w \in B_2^d} \hat{L}(w; S^L)$ and $\hat{L}(w^*; S^L) = 0$ for low rank large margin distributions.

As the margin in the labelled dataset bounded below by $\gamma(1 - 0.1\gamma)$, we can upper bound the empirical 0-1 error by the empirical loss function ℓ. For $n^L = \tilde{O}\left(\frac{\sqrt{k}}{\alpha \epsilon \gamma (1 - 0.1\gamma)}\right)$, with probability $1 - \frac{1}{8}$,

$$\frac{1}{n^L} \sum_{(x,y) \in S^L} \mathbb{I}\{y \langle \hat{w}, x \rangle < 0\} \leq \hat{L}(\hat{w}; S^L) = \tilde{O}\left(\frac{\sqrt{k}}{n^L \epsilon \gamma \left(1 - 0.1\gamma\right)} \right) \leq \frac{\alpha}{2} \tag{42}$$

It remains to show that the empirical 0-1 error is a good approximation of the 0-1 error on the distribution. To achieve this, we can apply the generalization bound in terms of the growth function (Lemma 6). As the Vapnik-Chervonenkis (VC) dimension of k-dimensional linear halfspace is $k + 1$, its growth function is bounded by $\Pi(2n^L) \leq (2n)^{k+1} + 1$. Thus, for $n^L = \tilde{O}\left(\frac{Lk}{\alpha}\right) \leq \tilde{O}\left(\frac{\sqrt{k}}{\alpha \epsilon \gamma (1 - 0.1\gamma)}\right)$, with probability $1 - \frac{\beta}{4}$, we have

$$\mathbb{P}_{(x,y) \sim D} \left[y \langle \hat{w}, x \rangle < 0\right] - \frac{1}{n^L} \sum_{(x,y) \in S^L} \mathbb{I}\{y \langle \hat{w}, x \rangle \leq 0\} \leq \frac{\alpha}{2}.$$

The second inequality in the sample complexity bound is due to $\epsilon \in (0, \frac{1}{\sqrt{k}})$.

Thus,

$$\mathbb{P}_{(x,y) \sim D} \left[y \langle \hat{w}, x \rangle < 0\right] \leq \frac{1}{n^L} \sum_{(x,y) \in S^L} \mathbb{I}\{y \langle \hat{w}, x \rangle \leq 0\} + \frac{\alpha}{2} = \alpha.$$

This concludes the proof. □

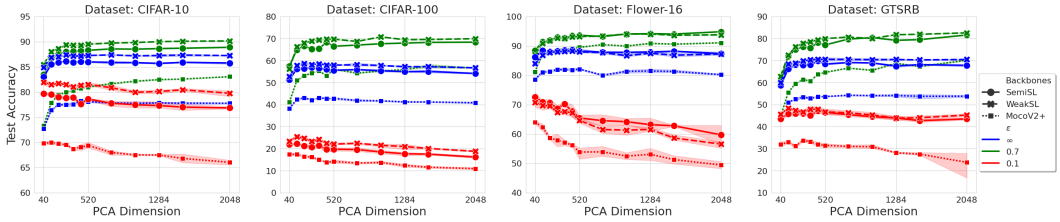


Figure 6: First and second panel: Test Accuracy of DP linear classifier on CIFAR-10 and CIFAR-100. Third and fourth panel: Test accuracy of DP linear classifier on Flowers 16 and GTSRB. While for public training ($\epsilon = \infty$) the performance generally increases as dimensions are added, the opposite occurs for DP training. The tighter the privacy constraints, the steeper the decrease.

F Additional Experimental details and Further Experiments

Details on discrepancy in resolution of images A standard evaluation procedure in DP is to evaluate the accuracy on the CIFAR-10 and CIFAR-100 [23] datasets. When training a linear classifier on pre-trained features, the resolution difference between Imagenet-1K and CIFAR images needs to be taken into account. While the pre-trained models are exposed to ImageNet-1K images at a resolution of 244×244, CIFAR images are at a resolution of 32×32. This discrepancy negatively impacts the performance in practice. To alleviate this issue, we pre-process the CIFAR images with the ImageNet-1K transformation pipeline, which increases the resolution of CIFAR images to that of ImageNet-1K. While the upscaled features still differ from the ones present in the original higher-resolution ImageNet-1K images, this preprocessing yields significantly improved performance. We use this technique throughout this paper whenever a significant difference in resolution exists between the pre-training and private datasets.

The impact of Different Pre-Training Strategies From Figure 2 and 3 it is evident leveraging the features of specific pre-trained models can dominate those produce by other pre-trained models. Therefore, it is important to investigate whether a consistent pattern exists when comparing the DP test accuracy of classifiers trained on different types of pre-trained features. We compute the maximum attainable accuracy with a publicly trained classifier, independently of the pre-trained model used. We measure the minimum drop in performance observed by training a DP classifier on SSL or SL pre-trained features (for SSL we consider both BYOL and MocoV2+). We then plot the fractional reduction (with respect to the best public classifier performance) for both SL and SSL across all the datasets and ϵ values in Figure 8a. We also draw a dashed line to indicate the region where neither methods show an advantage. Points above the line indicate that SL features exhibit an advantage whereas for points below the line, SSL has an advantage. As it can be seen, datasets representing daily life objects and with semantic overlap with ImageNet-1K benefit more from leveraging SL features. However, datasets that do not overlap that much with ImageNet-1K benefit more from leveraging SSL features. This is in line with what has been observed in a different setting by Shi et al. [30]. In Appendix F we also provide similar plots for Semi(Weakly)-Supervised pre-training procedures. Since these techniques leverage even larger amounts of data for pre-training, the obtained features are more convenient to use both with respect to SL and SSL in most cases.

In this section we report the results for MoCov2+ (SSL) and Semi-Supervised and Semi-Weakly Supervised feature extractors (SemiSL, SemiWeakSL). In Figure 6 we report the results on CIFAR-10, CIFAR-100, Flower-16 and GTSRB. In Figure 7 we report the results for PCAM, Pneumonia and DermNet. In Figure 8 we compare the relative reduction in performance when using Semi-(Weakly) supervised pre-training or Self-Supervised training.

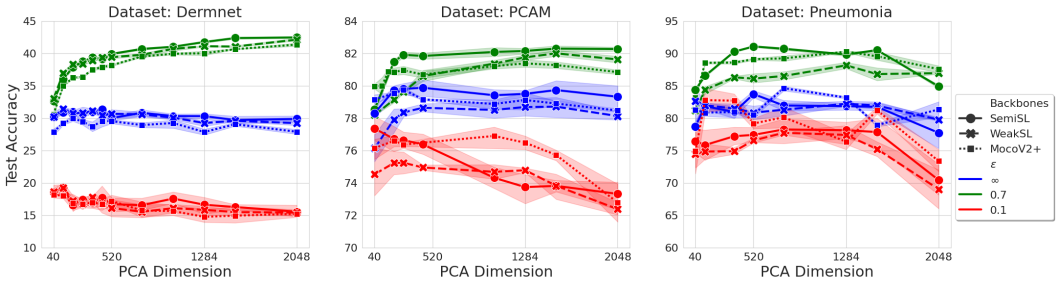
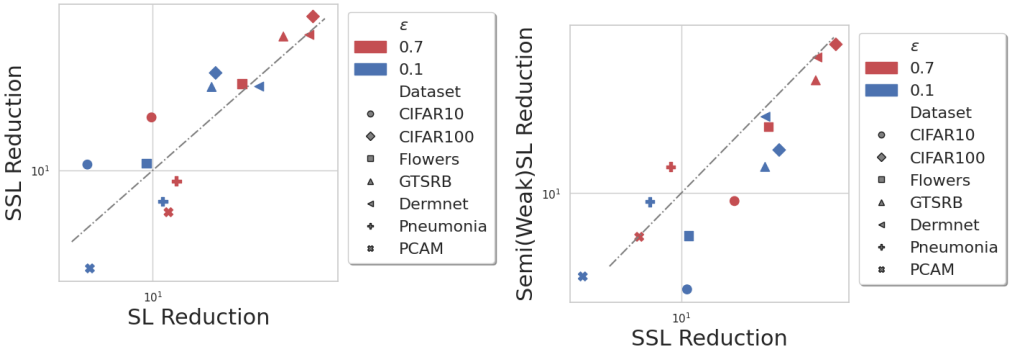


Figure 7: Test Accuracy of DP linear classifier on Dermnet, PCAM and Pneumonia. The proposed procedure performs similarly to what has been observed in Figure 6.



(a)

Figure 8: Comparison of the reduction in test accuracy for different datasets and different ϵ values with respect to the public training using SL, SSL, and semi-weakly SL feature extractors.