Mean Estimation Strikes Back! An Efficient Solution to Private Classification

Yuqing Zhu* Huanyu Zhang†

1 Introduction

Differentially private (DP) mean estimation is arguably the most comprehensively studied problems in the DP literature, leading to various efficient estimators that achieve optimal results across diverse applications [Dwork et al., 2006, Kamath et al., 2019, Acharya et al., 2021, Hopkins et al., 2022]. However, given the inherent simplicity of the mean estimation, it becomes nearly impossible to tailor these well-designed algorithms to tasks involving more complex data structures, such as DP machine learning.

Conversely, the application of private deep learning has emerged as a key driver of the recent successes in real-world DP deployment. While training models from scratch [Abadi et al., 2016] has been reported to incur significant performance degradation, the adoption of DP-SGD with fine-tuning has rapidly gained traction [Yu et al., 2021, Li et al., 2021]. Notably, private fine-tuning of only the last layer (also known as linear probing) of a pre-trained ResNet model has demonstrated an accuracy of 80.0% on CIFAR-10 under $(\varepsilon = 2, \delta = 10^{-5})$-DP, as compared to 84.0% without any privacy constraint.

Various theories attempt to explain this phenomenon. For instance, Ganesh et al. [2023] hypothesizes that the model reaches a good "basin" in the loss landscape after pre-training. Motivated by both theoretical insights and promising empirical findings, we are intrigued by the prospect that a basic technique like mean estimation could leverage the potentially enhanced model and data structure, thus achieving satisfactory model accuracy in private classification.

Baselines: There are mainly two fine-tuning schemes for classification:

1. *DP linear probing:* fine-tuning the last layer while keeping the model weights of the other layers frozen, or equivalently, applying DP logistic regression on the last-layer embeddings.
2. *Full fine-tuning:* fine-tuning all layers of the pre-trained model.

Empirical findings [Ke et al., 2024] suggest that linear probing typically performs well in low privacy regimes while full fine-tuning excels in high privacy regimes.

Our contributions can be summarized as follows:

1. In low privacy regimes, DP mean estimation demonstrates remarkable empirical performance, offering a feasible alternative to linear probing. The implementation is quite straightforward, without the complexities associated with iterative processes and hyperparameter tuning. Our findings may also inspire future researches into the loss landscape of fine-tuning.
2. In high privacy regimes, we propose a warm-up phase for full fine-tuning via mean estimation, which accelerates the convergence of DP-SGD. We anticipate this approach will be advantageous in scenarios

*TikTok. yuqingzhu@ucsb.edu. This work was done while the author was an intern at Meta.
†Meta. huanyuzhang@meta.com.

where training resources are constrained.

2 DP Mean Estimation for Classification

2.1 Direct Utilization for Model Inference

Intuitively, if the model has been pre-trained on a similar public dataset, the last-layer feature embeddings of private data might naturally form distinct clusters corresponding to each label class.[1]

Shed by this insight, our proposed training algorithm is outlined in Algorithm 1: we initially encode all private data using the last-layer feature extractor $\phi(\cdot)$ from the pre-trained model. We impose feature clipping within $\phi(\cdot)$, constraining $||\phi(x)||_2 = 1$ for all $x \in \mathcal{X}$. Subsequently, Let K denote the number of label classes. We privately estimate the feature mean $(q_1, ..., q_K) \in \mathcal{R}^{d \times K}$ for each class, utilizing the feature sum and the count perturbed by the Gaussian mechanism.

Algorithm 1 DP mean estimation for classification (training)

1: **Input**: Dataset $S \in (\mathcal{X} \times \mathcal{Y})^n$, noise scale σ and the last-layer feature embedding $\phi(\cdot) : \mathcal{X} \to \mathcal{R}^d$.
2: **for** $j = 1, ..., K$ **do**
3: Release the feature sum per class: $S_j = \left(\sum_{i=1}^{n} \mathbb{I}[y_i = j]\phi(x_i) \right) + \mathcal{N}(0, \sigma^2 \mathbf{1}_d)$
4: Release the number of private data per class: $n_j = \left(\sum_{i=1}^{n} \mathbb{I}[y_i = j] \right) + \mathcal{N}(0, \sigma^2)$
5: Release the class center $q_j = S_j / n_j$
6: **end for**
7: **Return feature means for each class** $q = (q_1, ..., q_K)$

During the inference, we apply the same feature extractor $\phi(\cdot)$ to the testing data and classify it based on the proximity to the nearest class center.

Theorem 2.1. *Algorithm 1 is (ε, δ)-DP if and only if*

$$\Phi\left(\frac{1}{2\sigma} - \frac{\sigma\varepsilon}{2}\right) - e^{\varepsilon}\Phi\left(-\frac{1}{2\sigma} - \frac{\sigma\varepsilon}{2}\right) \leq \frac{\delta}{2},$$

where $\Phi(\cdot)$ is the Gaussian CDF.

The proof directly follows from the analaysis of the Gaussian Mechanism [Balle and Wang, 2018].

Remark 2.2. We present a comparison between our algorithm and linear probing, where the latter can be viewed as a private logistic regression on the last-layer embeddings. While both algorithms theoretically incur the same order of additional loss due to DP [Bassily et al., 2019], we conjecture (also demonstrated by our experiments) that DP mean estimation achieves a superior constant factor. Consequently, as ε tends towards zero, our method provides an enhanced utility.

Remark 2.3. Our algorithm is quite simple in terms of implementation, getting rid of the complexities associated with iterative processes and hyperparameter tuning, which are often pain points for DP-SGD based approaches.

[1]This observation aligns with the recent neural collapse theory. Neural collapse [Papyan et al., 2020] is a phenomenon wherein, during training a deep learning model (without privacy consideration) until convergence, the last-layer features tend to converge to a K-simplex equiangular tight frame.

2.2 DP-SGD Warm-up via Mean Estimation

In this section, we propose utilizing Algorithm 1 to warm up DP-SGD (described in Algorithm 2), with the objective of enhancing both accuracy and computational efficiency.

While fine-tuning all layers in DP-SGD typically yields a state-of-the-art performance under a large ε, DP-SGD suffers from a slow convergence rate, thereby compromising the model's privacy-utility trade-offs, as well as training efficiency. In light of this, our approach aims to initially warm up the pre-trained model via private mean estimation, followed by training a DP-SGD algorithm that focuses solely on learning the residual between ground-truth labels and the pseudo-labels predicted by DP mean estimation.

Algorithm 2 DP-SGD warm-up via mean estimation (training)

1: **Input**: Dataset $S \in (\mathcal{X} \times \mathcal{Y})^n$, the privacy budget (ε, δ), the pre-trained model θ_0 and the last-layer feature embedding $\phi(\cdot) : \mathcal{X} \to \mathcal{R}^d$.
2: Release private class centers $(q_1, ..., q_K)$ = Algorithm 1 $(\varepsilon_1, \delta_1)$.
3: Encode each private data x: $h(x) = \Big(\cos(\phi(x), q_1), ..., \cos(\phi(x), q_K) \Big)$
4: DP-SGD full training on private set S using the loss function in Definition 2.4 under $(\varepsilon_2, \delta_2)$-DP.

As outlined in Algorithm 2, we initially allocate a privacy budget of $(\varepsilon_1, \delta_1)$ to reveal the private class centers q. As a next step, we encode each private individual x using the cosine distance to each center, denoted as $h(x) \in \mathcal{R}^K$. After softmax, the soft label serves as a prior for the subsequent algorithm. Finally, we employ a DP-SGD algorithm on the private dataset leveraging the residual loss introduced below.

Definition 2.4 (Residual loss)**.** Let $g(x) \in \mathcal{R}^K$ denote the logit value (model output before the soft-max layer) for the model θ. Define a probability simplex $\mathrm{softmax}(h) \in \mathcal{R}^K$ of x via the private mean estimation. The residual loss is computed as

$$\mathcal{L}_\theta = \ell_{KL}\bigg(\mathrm{softmax}\Big(g(x) \Big) + \mathrm{softmax}\Big(h(x) \Big); y \bigg)$$

where ℓ_{KL} denote the Kullback-Leibler loss.

In contrast to the conventional cross-entropy loss $\mathcal{L}_\theta = \ell_{KL}\big(\mathrm{softmax}\big(g(x)\big); y\big)$ employed in DP-SGD, the model θ in Algorithm 2 exclusively learns the difference between the prior $\mathrm{softmax}(h(x))$ and the true label y.

For inference, we classify each x as $\arg\max_{j \in [K]} \Big(\mathrm{softmax}(g(x)) + \mathrm{softmax}(h(x)) \Big)_{[j]}$.

Naturally from the composition theorem,

Theorem 2.5. *Algorithm 2 satisfies $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$-DP.*

3 Experiments

In this section, we conduct an empirical comparison between our methods and other private fine-tuning algorithms, aiming to assess both efficiency and privacy-utility trade-offs.

3.1 Low ε regimes

We compare the privacy-utility trade-offs between our Algorithm 1 and linear probing (known for its superior performance compared to full fine-tuning) when ε is small.

We consider two pre-trained models (vision transformer ViT [Dosovitskiy et al., 2020] and the language transformer RoBERTa [Reimers and Gurevych, 2019]. Our evaluation considers two image classification tasks, CIFAR-10 and CIFAR-100, as well as two language classification tasks, AG News [Zhang et al., 2015] and SST2 [Socher et al., 2013].

Table 1: Privacy-utility trade-off for Algorithm 1 and DP linear probing.

Model	Dataset	Method	$\varepsilon = 0.1$	$\varepsilon = 0.5$	$\varepsilon = 2.0$	No Privacy
ViT	CIFAR-10	Linear probing	88.4%	95.5%	**96.3%**	96.6%
		Our algorithm	**95.0%**	**95.6%**	95.6%	95.8%
	CIFAR-100	Linear probing	52.0%	79.1%	**82.1%**	86.0%
		Our algorithm	**58.0%**	**79.5%**	81.7%	82.4%
RoBERTa	SST2	Linear probing	86.6%	88.9%	**91.3%**	91.4%
		Our algorithm	**88.7%**	**89.1%**	89.2%	89.2%
	AG News	Linear probing	**88.1%**	**88.7%**	**89.3%**	91.2%
		Our algorithm	87.5%	87.8%	88.0%	88.3%

When $\varepsilon \leq 0.5$, our algorithm outperforms DP linear probing under some situations. Even when $\varepsilon = 2$, our algorithm functions as a viable alternative, offering simplified implementation compared to its counterpart.

3.2 High ε regimes

In the following experiment, we investigate the impact of accelerating the convergence of DP-SGD by the private mean estimation. Specifically, we keep the number of epochs fixed and ensure the overall privacy budget remains constant at $(\varepsilon = 2.0, \delta = \frac{1}{50000})$-DP for different values of epoch. That being said, our algorithm demonstrates superior performance compared to DP-SGD, when training resources are constrained.

Table 2: **CIFAR-100** Privacy-utility trade-offs at a fixed epoch. We fix epoch in $\{2, 10, 40\}$ and ensure the overall privacy budget is fixed to $(\varepsilon = 2.0, \delta = \frac{1}{50000})$-DP for all the methods (pre-trained model: ViT).

Method	Epoch=2	Epoch=10	Epoch=40	$\varepsilon = \infty$
DP linear probing	79.6%	82.4%	83.6%	86.0%
DP-SGD	79.3%	**84.7%**	85.6%	88.9%
Algorithm 2 (ours)	**81.3%**	84.8%	**86.1%**	89.3%

Table 3: **CIFAR-10** Privacy-utility trade-offs at a fixed epoch.

Method	Epoch=2	Epoch=20
DP-SGD	94.7%	97.0%
Algorithm 2 (ours)	**95.2%**	**97.0%**

References

Martín Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS-16)*, pages 308–318. ACM, 2016.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

Borja Balle and Yu-Xiang Wang. Improving gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. *International Conference in Machine Learning (ICML)*, 2018.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer, 2006.

Arun Ganesh, Mahdi Haghifam, Milad Nasr, Sewoong Oh, Thomas Steinke, Om Thakkar, Abhradeep Guha Thakurta, and Lun Wang. Why is public pretraining necessary for private model training? In *International Conference on Machine Learning*, pages 10611–10627. PMLR, 2023.

Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1406–1417, 2022.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

Shuqi Ke, Charlie Hou, Giulia Fanti, and Sewoong Oh. On the convergence of differentially-private fine-tuning: To linearly probe or to fully fine-tune? *arXiv preprint arXiv:2402.18905*, 2024.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2021.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D13-1170`.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2021.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, 2015.