Private-kNN: Practical Differential Privacy for Computer Vision Supplementary Material

Yuqing Zhu[1,2] Xiang Yu[2] Manmohan Chandraker[2,3] Yu-Xiang Wang[1]
[1]University of California, Santa Barbara
[2]NEC Labs America
[3]University of California, San Diego

In this supplementary, we provide the proofs of Theorem 7 and Theorem 8. Moreover, we present a discussion of utility and privacy trade-off in Market1501 dataset. Later, we describe the τ-approximation approach to reduce the global sensitivity in multi-label tasks.

A. Proofs of Theorem 7 and 8 in the paper

Theorem 1 (RDP of "Noisy Screening", Restatement of Theorem 7)**.** *Let $\mathcal{M}_s$ be a randomized algorithm for noisy screening procedure with a predefined Gaussian noise scale σ_1 and the threshold T. Then $\mathcal{M}_s$ obeys RDP with*

$$\epsilon_{\mathcal{M}_s}(\alpha) = \max_{(p,q)\in\mathcal{S}} \frac{1}{\alpha-1}\log\left(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\right).$$

where $\mathcal{S}$ contains the following "pairs":

$$\left(\mathbb{P}[\mathcal{N}(t,\sigma_1^2)\geq T], \mathbb{P}[\mathcal{N}(t+1,\sigma_1^2)]\geq T]\right),$$
$$\left(\mathbb{P}[\mathcal{N}(t,\sigma_1^2)\geq T], \mathbb{P}[\mathcal{N}(t-1,\sigma_1^2)]\geq T]\right)$$

for all integer $\lceil k/c\rceil \leq t \leq k$. The bound can be computed in time time $O(k)$.

Proof. For a given query x, set $n^*(x)$ be the vote count of the plurality and p, q denote the probability of x passes the noisy screening procedure with neighboring private datasets X, X' respectively. The output space of both $\mathcal{M}_s(X)$ and $\mathcal{M}_s(X')$ is $\{\top, \perp\}$, where $\top$ indicates x passes noisy screening process, and vice versa. Then $\mathcal{M}_s(X)$ and $\mathcal{M}_s(X')$ satisfy the Bernoulli distribution with the parameter p, q respectively.

By definition of Renyi Differential privacy and the Renyi Divergence of two Bernoulli distributions:

$$\epsilon_{\mathcal{M}}(\alpha) = \sup_{X,X' \text{are neighbors}} \frac{1}{\alpha-1}\log E_q(\frac{p}{q})^\alpha$$
$$= \sup_{X,X' \text{are neighbors}} \frac{1}{\alpha-1}\log\left(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\right)$$

The key of deriving RDP is to maximize over the two neighboring datasets. We make two observations. First, the notion of datasets X, X' are completely captured by their max votes t, t'. By the fact that the two datasets differ by at most one individual, $|t-t'|\leq 1$. In other word, to enumerate all neighboring datasets, it suffices to consider integer t, t' from $\lceil k/c\rceil$ to k such that $t'\in\{t-1, t+1\}$. Second, p, q can be directly calculated from t and t' respectively: $p = 1-\text{cdf}(\frac{T-t}{\sigma})$ and $q = 1-\text{cdf}(\frac{T-t'}{\sigma^1})$. Where cdf denotes the CDF of a standard normal random variable. Note that p monotonically increases as t increases.

These two observations ensure that we can calculate the RDP $\epsilon_{\mathcal{M}}(\alpha)$ for any fixed α in time $O(k)$. □

Where is t^* in practice? In practice, the worst pair of neighboring datasets occur either around $\max\{\text{votes}\} = T$ or around the boundaries when $\max\{\text{votes}\} = k$ (the largest possible) or $\max\{\text{votes}\} = \lceil k/c\rceil$ (the smallest possible due to pigeon hole principle).

In Figure 1, we plot the data-independent RDP of "Noisy screening" of all possible plurality. The plurality n^* ranges from $\lceil k/c\rceil$ to k and we set $k = 300$, threshold $T = 210, \sigma_1 = 85$. The x-axis is the RDP order α ranges from 1 to 50, the y-axis is the range of possible n^*, and we plot the corresponding RDP $\epsilon(\alpha)$ with the fixed α, n^*. The red curve shows the $\epsilon(\alpha)$ when $\max\{\text{votes}\} = T$, and we plot the red-dash line to view its exact RDP value more clearly. This figure shows that when α is small (below 50), the worst case of data-independent RDP is when $\max\{\text{votes}\} = T$. In Figure 2, we pick 5 curves from Figure 1 to further compare the RDP under the different choices of n^*. It shows that when $\alpha \leq 80$, the maximum data-independent $\epsilon(\alpha)$ is achieved when $n^* \approx T$, and when $\alpha \geq 80$ the $\epsilon(\alpha)$ is maximized when $n^* = k$. So for the upper bound of RDP of noisy screening, we only need to evaluate $\mathcal{M}_s$ for several neighboring datasets. In Figure 3, we plot the privacy cost of answering 8192 queries with 5 different data-independent analysis (from Figure 2 in the noisy screening procedure. The red line shows the privacy cost when $n^* = T$, and it's on the top the five curves which verifies our conjecture: the worst-case appears around $n^* = T$ or $n^* = k$. In the first 10 iterations, $n^* = k$ achieves the maximum. From Lemma 3, we know $\epsilon = \min_\alpha \epsilon(\alpha) + \frac{\log 1/\delta}{\alpha-1}$. When the number of iteration is small, the total privacy cost ϵ is minimized when

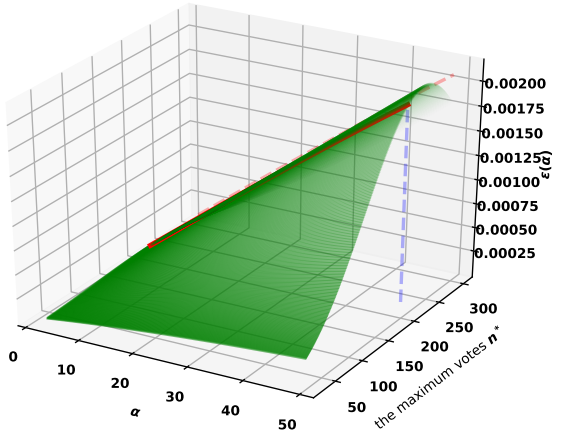

Figure 1. Searching the worst case for data-independent RDP of "Noisy Screening".

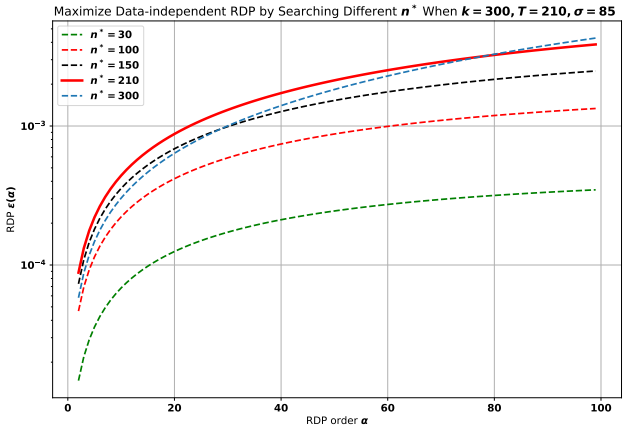

Figure 2. An example for data-independent RDP of " Noisy Screening" with different plurality.

α is large. As the number of iterations keeps increasing, ϵ is minimized when α is small. This phenomenon explains that the maximum data-independent privacy cost could be caused by several choices of n^*, which maximizes $\epsilon(\alpha)$ in the different range of α. However, $\epsilon_{\mathcal{M}}(\alpha)$ is not always maximized when $\max$ votes $= k$or$T$. For a larger α,the max $\epsilon_{\mathcal{M}}(\alpha)$ is attained when $n^* = k$. Check these two cases can give us a fast approximation of $\epsilon_{\mathcal{M}}(\alpha)$.

Theorem 2 (Asymptotic scaling, formal version of Theorem 8)**.** *Assume parameter $\gamma, \sigma_1, \sigma_2, \delta$ are chosen such that $\gamma < 0.1$, $\sigma_1 \geq \sqrt{5}$, $\sigma_2 \geq 2\sqrt{5}$, and moreover $\frac{4\log(1/\delta)\sigma_1^2}{\gamma^2\left(\min\{\sigma_1^2,\sigma_2^2\}\log^2(1/\gamma)-2\right)} \leq m \leq \frac{\sigma_1^2\log(1/\delta)}{3\gamma^2}$, $m_{select} \leq \frac{\sigma_2^2\log(1/\delta)}{6\gamma^2}$. Then, the end-to-end Private-KNN algorithm that processes all m public data points using with noise*

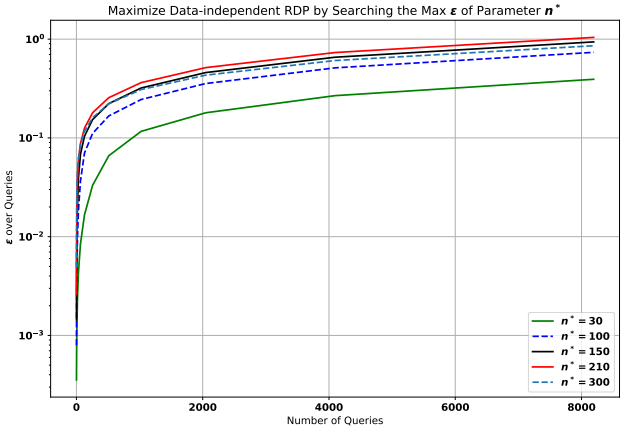

Figure 3. Privacy cost of answering 8192 queries with different data-independent RDP of "Noisy Screening". The sampling ratio $\gamma = 0, 25, \sigma_1 = 85, k = 300$. n^* is the fixed $\max$ votes.

σ_1, σ_2 and sampling ratio γ obeys (ϵ, δ)-DP, with

$$\epsilon = 20\gamma\sqrt{\log(1/\delta)}\left(\frac{\sqrt{m}}{\sigma_1} + \frac{\sqrt{m_{selected}}}{\sigma_2}\right).$$

Proof. The algorithm that process all m data points is an adaptive composition of two steps. In the first step, we release the $\{\top, \perp\}$ with the "noisy screening". In the second step, we release the "noisy max" for those that passes the screening rule. In both steps, the randomized procedure is amplified by Poisson subsampling. As a result, both has an RDP that is upper-bounded by the Poisson subsampled-gaussian mechanism.

The following is an asymptotic scaling of the the subsampled Gaussian mechanism.

Lemma 3 (Theorem 11 of [1])**.** *Let the global ℓ_2 sensitivity be Δ. Assume $\gamma \leq 0.1$, $\sigma/\Delta \geq \sqrt{5}$, then the Poisson-subsampled Gaussian mechanism obeys $(\alpha, \frac{6\gamma^2\Delta^2}{\sigma^2})$-RDP for all $\alpha \leq \frac{\sigma^2\log(1/\gamma)}{2}$.*

The above lemma is implied by the original statement about tCDP [1] for randomly selecting a subset of a fixed size γn, because (1) tCDP is an upper bound of RDP; (2) the exact RDP calculation for the Poisson-subsampled Gaussian mechanism matches the RDP lower bound of the (Random subset) subsampled Gaussian mechanism [6, Proposition 10].

The global sensitivity of the Gaussian mechanism in the "noisy screening" step is 1 because we are releasing only $\max\{$Votes$\}$, while it is 2 in the Gaussian mechanism for releasing the Votes — the histogram. Check that the stated assumptions on $\gamma, \sigma_1, \sigma_2$ satisfy the conditions above.

By the composition rule of Renyi Differential Privacy in Lemma 5 which establish that the end-to-end algorithm

obeys RDP with

$$\epsilon(\alpha) \leq \frac{6\gamma^2 m\alpha}{\sigma_1^2} + \frac{12\gamma^2 m_{\text{select}}\alpha}{\sigma_2^2}.$$

for all α in the range that are permitted by Lemma 3.

Finally, by Lemma 3 in main submission, we can convert RDP to (ϵ, δ)-DP with

$$\epsilon = \alpha\left(\frac{6\gamma^2 m}{\sigma_1^2} + \frac{12\gamma^2 m_{\text{select}}}{\sigma_2^2}\right) + \frac{\log(1/\delta)}{\alpha - 1}.$$

Choose $\alpha = 1 + \frac{\sqrt{\log(1/\delta)}}{\sqrt{\frac{6\gamma^2 m}{\sigma_1^2} + \frac{12\gamma^2 m_{\text{select}}}{\sigma_2^2}}}$ we get that:

$$\epsilon = \frac{6\gamma^2 m}{\sigma_1^2} + \frac{12\gamma^2 m_{\text{select}}}{\sigma_2^2} + 2\gamma\sqrt{\log(\frac{1}{\delta})\left(\frac{6m}{\sigma_1^2} + \frac{12m_{\text{select}}}{\sigma_2^2}\right)}.$$

The proof is complete by checking that under our assumption m, the second term always dominates and the assumption on α in Lemma 3 no matter that m_{select} turns out to be. □

B. The utility and privacy trade-off on Market1501 dataset

Figure 4 shows the utility and privacy trade-off of PATE and ours by varying sampling ratio γ, the noisy scale σ_1 and the number of queries. For GNMAX in PATE, to push the accuracy from 86.80% to 86.90%, we need to increase the privacy budget from 13.41 to 43.14. In the low privacy cost regime, our method achieves accuracy 87.82% with privacy budget 0.2416. In the high privacy cost regime, our algorithm achieves 89.18% with $\epsilon = 1.72$ compared to $\epsilon = 5.298$ and accuracy =86.21% in PATE. Further by checking the same accuracy, i.e., 86.5% for both "GNMAX" and ours with $\gamma = 0.05$, our privacy cost is 0.116 while "GNMAX" is 6.62. Indeed, more than 90% privacy budget is saved from the baseline method.

Privacy and utility trade-off of GNMAX In all experiments of GNMAX, we set the number of teachers with respect to the performance of each teacher. For example, if we set the number of teachers to be 600, then the average non-private accuracy of each teacher is around 76%. Since every partitioned data should not be overlapped with each other regard to the identity, the total identity is 750, and $T = 600$ is the maximum number GNMAX algorithm can afford. If we set a small T for GNMAX, e.x. $T = 100, \sigma_1 = 40$, then the privacy loss of GNMAX achieves $\epsilon = 13.22$ even it only answers 80 queries.

C. Applying Private-kNN to multi-label classification tasks

So far, we have been primarily working with multi-class classification tasks where the global sensitivity of the voting results of the nearest neighbors are naturally bounded. But for multi-label tasks, this is no longer true. Potentially, for a problem with c-labels, any neighbor can potentially vote on all c-labels, which makes the naïve noisy-adding mechanisms inefficient. We propose to fix it by a "clipping" heuristic that limits contribution of every label to at most τ.

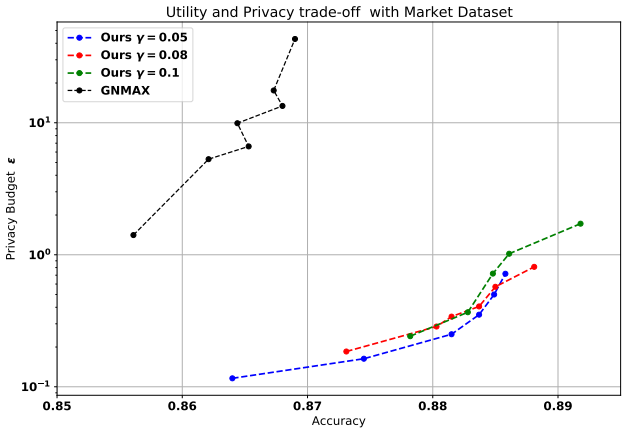

Figure 4. Trade-off between utility and privacy for PATE and ours on Market1501 dataset.

Definition 4. (τ approximation) For a traditional classification task, the global sensitivity of our model in the noisy aggregation process is 2 from Theorem 9. However, consider the more general multi-label task in vision, e.x. Facial attribute classification task, where one face image could have at most 40 attributes and the global sensitivity will increase to 80. To limit the global sensitivity in the multi-label task, we introduce the τ-approximation method, where the basic idea is that each neighbor could vote no more than τ attributes. For simplicity's sake, we only consider binary multi-label tasks here. In a multi-label task, the vote of neighbor j upon query x is $f_j(x) \in \mathcal{N}^c$ now becomes a $c-$way vector. To impose τ approximation on it, we apply

$$\hat{f}_{j,i} = f_{j,i} \cdot \min(\frac{\tau}{|f_j(x)|}, 1), i \in [1, c],$$

with $|f_j(x)|$ the L_1 norm of original neighbor j's voting and $\hat{f}_j$ the neighbor j' prediction upon x with τ approximation.

Theorem 5 below provides a practical privacy bound to guide the analysis for multi-label classification task.

Theorem 5. *Let $\mathcal{M}_\tau$ be a randomized algorithm for a multi-label task with τ-approximation method, the global sensitivity of $f(x)$ here is $2 \cdot \tau$, then we have for integer $\alpha \geq 2$,*

$$D_\alpha(\mathcal{M}_\tau(X)||\mathcal{M}_\tau(X')) = \frac{\alpha \cdot \tau}{\sigma_1^2}$$

Regression problems. Similar clipping tricks can be applied to regression problems so private-kNN applies. We can

also use median, rather than the mean. Careful experimental evaluation on regression problems are left as a future work.

D. Architecture of networks

We plot the network architecture of MNIST in Table 1. The MNIST model contains two convolutional layers with max-pooling and two fully connected layers with ReLUs. For the SVHN task, Table 2 shows that the SVHN model stacks seven convolutional layers with two fully connected layers, which replicates the experimental setup as in [5]. The source code of MNIST and SVHN experiments and a Pytorch implementation of [5] are available on Github.[1]

Table 1. Network architecture of MNIST task

Conv	64 filters of size 5×5
Max pool	2×2
Conv	128 filters of size 5×5
Max pool	2×2
FC	$(384, 192, 10)$

Table 2. Network architecture of SVHN task

Conv	96 filters of size 3×3
Conv	96 filters of size 3×3
Conv	96 filters of size 3×3
Conv	192 filters of size 3×3
Conv	192 filters of size 3×3
Conv	192 filters of size 3×3
Conv	192 filters of size 5×5
FC	$(192, 192, 10)$

E. More discussion about data-dependent noisy-screening

Noisy-Screening vs. Sparse Vector Technique. The noisy screening is closely related to the Sparse Vector Technique [2, 3] (SVT) that screens a sequence of online queries $f_1, f_2, ...$ with global sensitivity 1 and output $\{\top, \perp\}$ with the hope of approximately selecting those queries with value greater than a threshold T and essentially paying only the privacy loss for those that are selected.

The key steps of an SVT include adding Laplace noise to the threshold and also adding Laplace noise to $f_i(x)$ when deciding whether to output $\top$ or $\perp$. When the large majority of the queries have either $\top$ or $\perp$ with sufficiently high margin from the threshold T, then SVT is able to handle an exponentially large set of queries.

"Noisy-screening" is different in two ways. First, it does not aim at "calibrating noise to stability" to achieve a pre-defined privacy budget. Instead the version that we used pays the same amount for every query. Second, we can use Gaussian mechanism on $f_i(x)$ while keeping the threshold T unchanged. This method at a glance does not resemble SVT at all because it does not adapt to the input sequence, and pay only an amount proportional to the $\sqrt{\min\{\#\text{ of }\perp, \#\text{ of }\top\}}$ as in SVT.

That said, the data-dependent RDP of "Noisy-screening" is in fact a lot more closely related to SVT. If a query f_i obeys that either $f_i(x) \gg T$ or $f_i(x) \ll T$, then the data-dependent RDP is going to be exponentially smaller than that is coming from the Gaussian mechanism. Directly composing the data-dependent RDP will lead to qualitatively the same behavior as SVT.

For example, for a sequence of queries where SVT can answer exponentially many without using up a budget of (ϵ, δ), we can answer the same sequence with "noisy-screening" while paying a "data-dependent" privacy loss that is likely to be smaller than (ϵ, δ).

Consider another example, if the sequence of queries are close to $f_i(x) = T$, then the data-dependent calculations for "noisy screening" will arrive at about the same privacy losses as the data-independent counterpart. Similarly, SVT will also stop within just a few rounds because essentially it pays every other iteration on average.

In summary, the data-dependent RDP calculations of "noisy screening" can be thought of as a versatile alternative of SVT, when satisfying a fixed pre-specified privacy budget is not too important and when we do not have to reveal the final privacy loss that is realized (because its value depends on the data). This allows us to use a more concentrated Gaussian noise, and to take advantage of the RDP for a tighter composition.

Both limitations can be resolved by privately releasing the data-dependent RDP using smooth sensitivity [4] as in what was proposed in the appendix of [5]. Details of this procedure and how "noisy screening" compares to SVT in general is left as a future direction of research.

Open problem: Data-dependent RDP of subsampled mechanism. Privacy-amplification by subsampling is not compatible with data-dependent RDP because implicitly, the amplification is coming from the fact that for any subset that is selected, the same RDP bound holds.

A trap is to amplify the data-dependent RDP calculated through the specific sample that is chosen. This is because value probably cannot hold for other subsets.

It remains an open problem how to correctly calculate the data-dependent RDP for a subsampled mechanism. The exact calculation would require enumerating over all subsets and calculating their corresponding data-dependent RDP.

[1] https://github.com/jeremy43/Private_kNN

References

[1] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke. Composable and versatile privacy via truncated cdp. In *STOC-18*, 2018. 2

[2] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390. ACM, 2009. 4

[3] M. Hardt and G. N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 61–70. IEEE, 2010. 4

[4] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *ACM symposium on Theory of computing (STOC-07)*, pages 75–84. ACM, 2007. 4

[5] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018. 4

[6] Y. Zhu and Y.-X. Wang. Poission subsampled rényi differential privacy. In *International Conference on Machine Learning*, pages 7634–7642, 2019. 2