Reviews from Prior Submission

Anonymous Author(s)

Abstract

This document provides the reviews of a former version of our paper, which was not accepted for publication. We first provide the meta-review and the detailed reviews. For each reviewer, we provide the original review, the authors' rebuttal and the reviewer response. We finally explain how each specific point is addressed in the current version of the paper. All the *italicized text* is an unedited version of the received reviews, while the text in normal font was written by the authors.

I. META REVIEW

This paper proposes differentially private algorithms for learning rule lists. The main technical step is to estimate "gini impurity" (a measure to evaluate the effectiveness of adding a rule) privately. To do this, the paper uses the smooth sensitivity framework of Nissim, Raskhodnikova, & Smith (2007), which is more effective than the standard global sensitivity approach. The paper presents both theoretical analysis of the privacy and empirical evaluation of the algorithm.

The reviewers appreciated this as an interesting application of differentially private learning. However, they also found aspects of the paper hard to digest. In particular, the experimental evaluation lacks comparisons to other mechanisms, such as the inverse sensitivity mechanism of Asi & Duchi (2020) and variants of the smooth sensitivity approach of Bun & Steinke (2019). The plots (e.g. Fig 3) are also somewhat odd in that they plot ϵ on a log scale (ϵ is already the log of the likelihood ratio, so this is effectively a log log scale) and the lines seem not to reach the non-private baseline or only reach it for extremely large ϵ. It is unclear why the lines have this shape (going up, then flat, then going up again).

Overall, this paper seems solid, but it would benefit from some major revisions prior to presentation at ICML.

II. REVIEWER #1

A. Original Review

1) Summary: This paper focuses on building differentially private (DP) rule lists models. In particular, this work addresses a challenge proposed by Fletcher & Islam (2019) for establishing the smooth sensitivity of the Gini impurity. The Gini impurity is used as a "loss function" in greedy algorithms that construct is used to construct a rule list model. By theoretically characterizing the smooth sensitivity of the Gini impurity, the authors are able to build a DP greedy algorithm.

The experimental results show that the smooth sensitivity greedy algorithm can build rule lists models with comparable accuracy (within fractions of a percent) to the non-private baseline for $\epsilon > 10$, and the smooth sensitivity greedy algorithm consistently outperforms a greedy algorithm based on global sensitivity. These results are especially true for large datasets (> 5k samples). Finally, the authors evaluate how much DP helps against membership inference attacks (MIAs). Due to the difficulty of adapting popular attacks from the literature to rule lists models, the authors propose their own metric (which they name vulnerability) for quantifying the success of an MIA. Their results show that non-DP models consistently exhibit slightly higher vulnerability values than their DP counterparts.

**The minus 1 comes from an edge case regarding the "default" rule.*

2) Strengths:

- *This paper resolves an open problem posed by Fletcher & Islam (2019). This theoretical contribution allows the authors to build the first DP greedy algorithm for constructing rule lists models (also known as decision lists in the literature) with smooth sensitivity with the Gini index. The Gini index is a popular metric: it is scikit-learn's default criterion parameter for the DecisionTreeClassifier class (though the authors leave exploring greedy algorithms for decision trees as future work).*
- *Due to the novelty of this work, there are no existing DP greedy algorithms for rule lists models that use global sensitivity. To establish a global sensitivity baseline, the authors propose two variations of their algorithm (Section 5.2). Finally, experimental results show that the proposed smooth sensitivity greedy algorithm outperforms the proposed global sensitivity algorithm.*

3) Cons:

- *In my opinion, this paper would benefit from a related works section that mentions (1) previous work on DP implementations of interpretable models and (2) cites previous (theoretical) work on DP rules lists models such as Thaler et al. 2014 and Daniely et al. 2019.*
- *Furthermore, the role the Exponential Mechanism plays in the implementation of Algorithm 1 needs clarification. Lines 10-20 in state that the rule $r \in R_{rem}$ with the lowest noisy Gini impurity $\mathcal{G}$ gets selected as the "best rule" (and hence appended to RL). Given the noisy Gini values $\mathcal{G}$, this implies a deterministic process for selecting the best rule. However,*

in Section 5.3, the authors claim "we implemented the Exponential mechanism using the Gini impurity as the utility function for sampling the best rule at each node", which implies a random process (given $\mathcal{G}$) for selecting the best rule. This distinction is not made clear in Algorithm 1. Moreover, the Gini impurity is a "loss function" in the sense that lower is better, whereas the Exponential Mechanism requires a "utility function" (i.e. higher is better). How is the Gini impurity being used as the utility function? This discussion is missing in the current version of the paper.

- *Section 5.3 contains an remark about how these greedy algorithms compare to the DP random forest algorithm of Fletcher & Islam (2017), stating "we incur at $\epsilon = 1$ a significantly lower accuracy loss with respect to the nonprivate model." This paper could benefit from elaborating on this remark. How much is "significantly lower"? How does the proposed greedy algorithm compare to the DP random forest algorithm from Fletcher & Islam (2017) in terms of the privacy accuracy tradeoff?*
- *Finally, the paper would benefit from expanding Table 1 into a plot. Currently, Table 1 only explores DP-GreedyRL and Greedy-RL for $\epsilon = 10$, and shows the values of "vulnerability" and test accuracy across datasets. A plot show casing the tradeoff between vulnerability and accuracy as epsilon is varied would make for a much more compelling plot that clearly shows the robustness to MIAs vs accuracy tradeoff granted by DP.*

I am willing to increase my rating if the concerns outlined in this section are addressed.

REFERENCES

Justin Thaler, Jonathan Ullman, & Salil Vadhan. (2014). Faster Algorithms for Privately Releasing Marginals.

Amit Daniely, & Vitaly Feldman. (2019). Locally Private Learning without Interaction Requires Separation.

4) Questions:

- *The authors state: "Rule Lists can be built either with an exact method such as CORELS (Angelino et al., 2018) or with heuristic approaches (Singh et al., 2021), which we specifically consider in this paper. " Why was the heuristic method considered in this paper? Is it more amendable to differential privacy?*
- *In Section 3.3, the equation describing the smooth sensitivity mechanism that uses Laplace noise is missing an η. I believe it should read $\mathcal{M}^{S^*}_{Laplace}(\mathcal{D}, f, \epsilon) : \mathcal{D} \mapsto f(\mathcal{D}) + \frac{2S^*(\mathcal{D})}{\epsilon}\eta$*
- *In Section 4.2, the authors state "Therefore if the counts are not displayed with the model, then the denominator is only $2K - 1$ for ϵ, which leads to $\epsilon_{node} = \frac{\epsilon}{3K-1}$." The wording here is confusing. Should $2K - 1$ be the denominator or $3K - 1$?*
- *In Section 5.2, it is written that "The second version leverages the global sensitivity of counting queries (equal to 1) rather than using the global sensitivity of the Gini impurity which is very high." This sentence confused me. I thought the Gini impurity global sensitivity was 0.5, which is lower than the global sensitivity of counting queries. Could the authors elaborate on what they meant to say here?*
- *In Table 1, what does the notation $0.507^+ \pm 4e-6$ and $0.507^- \pm 4e-6$ mean?*

5) Limitations: The authors are clear in the limitations of their word. In particular, they clearly show that the extra protection given by DP against for membership inference attacks is small.

6) Soundness: 3: good

7) Presentation: 3: good

8) Contribution: 4: excellent

9) Rating: 5: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.

10) Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

B. Authors' Rebuttal

Thank you very much for all your detailed comments and recommendations on this paper.

Complete related works section. We thank the reviewer for suggesting these two papers. While they both consider decision lists, they focus on a different context than ours. Nonetheless, we will include them in the revised version of the paper and highlight the differences. More precisely, Daniely et al. (2019) discuss the sample complexity of learning decision lists in a non-interactive local differential privacy context while Thaler et al. (2014) focus on the design of efficient privacy-preserving data publishing. In particular, they consider the problem of privately releasing a database whose rows consist of pre-defined decision lists. In this setting, the problem considered is not the privacy-preserving learning of decision lists.

Clarification about the use of the Exponential Mechanism in Algorithm 1. The Exponential Mechanism is another generic DP mechanism that we have implemented based on global sensitivity (blue full line in Figure 3). In contrast, Algorithm 1 details the private algorithm leveraging the smooth sensitivity of the Gini Impurity and does not use the Exponential Mechanism. We apologize for the confusion and in the revised version we will mention it (and explicitly state that we denote it gl-Exponential) in the first paragraph of 5.3. To obtain the utility function for the Exponential Mechanism, we used "u(D,r) = 1-G_D(r)" so that it yields "the higher the better" property while the sensitivity is obviously unchanged.

Comparison with (Fletcher and Islam 2017) in Section 5.3. We compared the difference between the DP model and the baseline model in terms of accuracy at $\epsilon = 1$ on the UCI Adult Income dataset. Unfortunately, the results from Fletcher & Islam (2017) were mostly based on synthetic datasets we could not compare on. For this level of privacy, our smooth sensitivity-based algorithm has less than 0.5% (the accuracy of the proposed DP algorithm is 78.7% and 79.1% for the non private version) in accuracy loss against at least 1.0% for theirs (the accuracy of the DP algorithm is 82% and the non DP version is about 83%). Thus, even though (non-DP) decision trees perform better than (non-DP) rule lists, we estimate our trade-off between privacy and accuracy to be better as our mechanisms incur a lower utility cost, and this trend can only increase with larger datasets.

Presentation of Table 1. One issue that we have encountered with the baseline model is that it is already resilient to membership inference attacks as demonstrated by both the ROC curves and the overall vulnerability metric. We believe that this is an intrinsic advantage of working with small interpretable models for which the capacity of the model to memorizing training data is limited. As such, decreasing ϵ will only worsen the accuracy of our DP model without a noticeable increase in resiliency to membership inference attack.

Integration of differential privacy in CORELS. We actually worked on implementing DP into CORELS but we had to face several issues. First, CORELS relies on optimality-based bounds to efficiently prune out solutions but these bounds impede DP. For instance, for all permutations of a given rule list, CORELS only keeps the rule list with the best accuracy for further exploration of the space, which breaks the DP guarantees as the probability of outputting a sub-optimal rule list becomes 0. A possible approach to address this is to deactivate all bounds but then CORELS perform a full exploration of the latent space, which highly impacts its performance. Second, working on the Gini impurity given its widespread use both (i) allows to build on previous results to build more elaborate mechanisms and (ii) encourages re-use and extension of our work.

In Section 3.3, the equation describing the smooth sensitivity mechanism that uses Laplace noise is missing an η. Thanks for pointing this out, there is indeed the $\eta \sim Lap(1)$ missing. We will correct this in the final version of our paper.

In Section 4.2, the authors state "Therefore if the counts are not displayed with the model, then the denominator is only ... We mean that since the counts are released for each rule with a maximum of K rules per rule lists, they account for a budget of K in the denominator. This corresponds to a privacy budget of $\epsilon_{\text{node}} = \frac{\epsilon}{3K-1}$ with the noisy counts of the leaves made public and $\epsilon_{\text{node}} = \frac{\epsilon}{2K-1}$ without the noisy counts. The remaining $2K-1$ of the budget are split between the Gini impurity Noisy Max Report ($K-1$ in total because no Gini is computed for the default rule) and the label decision for each leaf (K in total).

In Section 5.2, it is written that "The second version leverages the global sensitivity of counting queries (equal to 1) rather than using the global sensitivity of the Gini impurity which is very high." This sentence confused me. I thought the Gini impurity global sensitivity was 0.5, which is lower than the global sensitivity of counting queries. Could the authors elaborate on what they meant to say here? We qualify the global sensitivity of the Gini impurity (0.5) as very high according to the possible range of values it takes (i.e., $[0, 1]$). In contrast, while the sensitivity of counting queries is 1, the counts can go up to several thousands in value, hence the noise distortion is relatively weaker. In practice, this makes noisy counts infinitely more accurate than the noisy Gini impurity but employing them to compute the Gini Impurity is not a noisy max report anymore. To achieve DP, we have to access the noisy counts for all possible rules and then compare the respective Gini computed using the counts, which explains why $|\mathcal{R}|$ gets in the denominator of the privacy budget.

In Table 1, what does the notation $0.507^{+} \pm 4e-6$ and $0.507^{-} \pm 4e-6$ mean? In Table 1, we used notation + to indicate that the non truncated value was above the displayed one, and - to indicate it was below. We used this notation to be fully transparent on our results, and in particular show that even if the truncated values are similar in some experiments, the DP version of our algorithms exhibits a slightly lower vulnerability. We however agree that this notation can be misleading and will clarify this in the final version of our paper.

C. Reviewer Response

I thank the authors for their response. They addressed all of my questions. My score was based mainly on a misunderstanding I had about the role the Exponential Mechanism played in their algorithm. I have increased my score accordingly.

Regardless of the acceptance or rejection of this paper, I encourage authors to further explore their empirical observation that DP rule lists have a "better" privacy/utility tradeoff than (current) DP decision trees. This observation, if true, would make this paper's contribution much stronger.

III. Reviewer #2

A. Original Review

1) Summary: Smooth sensitivity was proposed by Nissim et al. (2007) to construct pure DP or ADP mechanisms that exploit functions whose local sensitivity (max change to function value in any neighborhood of specific dataset) tends to be a lot smaller than the global sensitivity (max change to function value in any neighborhood of any possible dataset) by utilising noise scaled proportional to a "smooth" upper bound of the local sensitivity that considers non-immediate neighbourhoods.

One example of such a function is the widely known Gini impurity for which the manuscript proposes ADP constructions based on smooth sensitivity building on prior results by Fletcher & Islam 2015 for local sensitivity. This is not trivial to achieve and could be useful for differentially private decision tree and rule list learning (in this manuscript only the former is explicitly explored).

2) Strengths:

- *S1. Interesting, non-trivial results on smooth sensitivity for Gini impurity that answers an open question by Fletcher & Islam (2019).*
- *S2. Proposed approach tested against membership inference attacks.*
- *S3. Claims appear to be sufficiently substantiated, manuscript is well-written and easy to follow.*

3) Weaknesses:

- *W1. Discussion of alternative directions as in Asi & Duchi (2020) would be useful. It would seem useful to discuss: Asi, H., & Duchi, J. C. (2020). Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. Advances in neural information processing systems, 33, 14106-14117.*
  For instance, Asi & Duchi (2020) derive specific results that compare their proposed approximate mechanism with the smooth Laplace mechanism based on smooth sensitivity and it would be useful to discuss why the inverse sensitivity approach may not be applicable/challenging for Gini impurity, if that remains and open challenge or is not promising for some reason.
- *W2. Comparison with Fletcher & Islam (2017) could be more explicit.*
  For instance, Fletcher & Islam (2019) argues that smooth sensitivity approaches for Gini impurity could be better for knowledge discovery tasks than the approach in Fletcher & Islam (2017) that is limited to majority class labels. It would be useful to spell these differences out and discuss particular advantages of the proposed approach. Furthermore, it could be useful to be more explicit which particular challenges are left open with regards to DP decision trees.
- *W3. Minor Issues*
  - *While the proofs seem logical and sufficiently comprehensive, they could provide a bit more instruction on how exactly to read the diagrams and in this instance numerical/empirical sanity checks of full/partial results might be possible to make it even easier to quickly verify main results.*
  - *Notation "1-3" in Rule list 1 seems to refer to the set "1,2,3" as in Rivest (1987) rather than "1-3 = -2", which may be confusing to some readers.*

4) Questions:

- *Q1. How do these work's results relate in broad terms to the results for the approximate mechanism by Asi & Duchi (2020)?*
- *Q2. What are the explicit advantages compared to results by Fletcher & Islam (2017) and what would be different when applying the results to decision trees rather than decision rule lists?*
- *Q3. How exactly are the diagrams in the proofs meant to be read and would it potentially be possible to add any sanity checks for the correctness of main results to make them easier to verify?*

5) Limitations: The authors discussed the implications of automated decision making and importance of transparency and explainability of learned models.

6) Soundness: 3: good

7) Presentation: 3: good

8) Contribution: 3: good

9) Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

10) Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

B. Authors' Rebuttal

Thank you for your valuable feedback. We answer the specific points raised in your review hereafter.

Q1. How do these work's results relate in broad terms to the results for the approximate mechanism by Asi & Duchi (2020)? Asi and Duchi's inverse sensitivity mechanism could be a future step in the implementation of our algorithm. In a nutshell, their proposed approach consists in an Exponential Mechanism scaled with the inverse sensitivity (or path length). Rather than measuring the variations of a function f between two adjacent databases, the authors compute the minimum distance from a database x to reach a database x' so that $f(x') = t$ (target value) hence coining the term of inverse sensitivity. While the exact path length introduced in their paper is often intractable, they derived a method using smooth sensitivity to approximate the path length. We observe that their method provides a better alternative to classic noisy mechanisms using smooth sensitivity for pure-DP mechanisms since they do not have to use heavy-tailed distributions such as Cauchy. It would

actually be a great future work to delve into since we have found a closed formula for the smooth sensitivity of the Gini Impurity. We expect to obtain similar results with the smooth Laplace mechanism and better results when it comes to pure DP.

Q2. What are the explicit advantages compared to results by Fletcher & Islam (2017) and what would be different when applying the results to decision trees rather than decision rule lists? In their paper, Fletcher and Islam (2017) propose a method for building trees for random forests. To save on the privacy budget, they propose to only allocate privacy to find the majority labels and use the smooth sensitivity onto the Exponential mechanism to further reduce privacy costs. The splits in the trees (i.e., attribute and value to split on) are (entirely) randomly determined. This works well for random forests for a high number of trees, as it is expected that poor splits will be compensated by the sheer number of "good" trees. Nonetheless, this approach is intrinsically not designed to learn single decision trees or rule lists. Indeed, choosing attributes to split at random while solely using the privacy budget to accurately determine the best label would be too susceptible to heavily skew the split and lead to leaves with very impure Gini score. In our approach, given that we have much more privacy budget to spend on each node (as we learn a single model), we split it between finding the best splitting criterion and ensuring that each leaf has the right label predicted. Thus, there is no doubt that this approach is more efficient for single trees.

Compared to rule Lists, decision trees right split is not necessarily a leaf and can be further divided into subsplits with higher purity. The benefits of the adaptation of our DP algorithm leveraging smooth sensitivity onto decision trees is two-fold. First, we can expect to obtain higher accuracy than with rule lists. Second, decision trees enables to use parallel composition more efficiently to compute the associated DP guarantees. Indeed, all the splitting nodes at a given depth work on disjoint parts of the dataset, so the budget to allocate corresponds only to the budget necessary for one node per depth ! With rule lists, parallel composition was only dividing the budget per depth by a factor of 2. On a decision tree, there are around 2^d nodes at a given depth d hence the obvious benefits.

Unfortunately, we were unable to assess the correctness of the Exponential Mechanism presented by Fletcher and Islam in which they directly replace the global sensitivity by the smooth sensitivity in the Exponential Mechanism.

Q3. How exactly are the diagrams in the proofs meant to be read and would it potentially be possible to add any sanity checks for the correctness of main results to make them easier to verify? The proof is a proof by exhaustion and we would like to apologize for the lack of clarity that was relayed by several reviewers. The variational tables were designed to help at observing the variations of the function graphically according to the different cases (namely the values of β and the position of the roots t_1 and t_2). Here is a brief outline of the method: the smooth sensitivity is computed as $S^*_{\mathcal{G},\beta}(x) = \max_{k\in\mathbb{N}} e^{-k\beta}\mathcal{T}_k(x)$. We first determine the function $\mathcal{T}_k(x)$ in which x is a dataset. We observe that it does not depend on the actual value of the dataset but solely the number of samples it contains. We obtain $\mathcal{T}_k(x) = g\left(\max(1, ||x||_1 - k)\right)$. Given that we managed to obtain a closed form for $\mathcal{T}_k(x)$ in which k directly intervenes, we now consider x to be a parameter and put k as a variable of our function hence the introduction of function $\xi_{x,\beta}(t) = \ e^{-k\beta}\mathcal{T}_k(x)$.

Rather, we study this function on $\mathbb{R}^+$ since it is differentiable. Through cancelling the derivative, we are able to find the minima and the maxima of the function. However, since these are $\mathbb{R}$-valued maxima, we finally truncate them to the closest higher and inferior integers to obtain the smooth sensitivity. Note that since we associate the sign of the derivative to a polynomial, it gives us extra control over the monotony of the function ξ since we know a polynomial takes the sign of the highest degree coefficient outside of the roots (granted that they exist) and the opposite inside the roots. For certain values of β such maxima may not exist because they are computed as the roots of a polynomial whose values depend on β. For this reason, we state that the formula should encompass t only if it is well defined (i.e., the value inside the square root is not negative) and otherwise replace it by 0. Proving that the smooth sensitivity only depends on $||x||$ is a core result of our approach. Thanks to this, at each iteration we only need to query the number of elements left to classify and apply the same smooth sensitivity to all rules (the Gini depends on the split made by the rule, but its smooth sensitivity is indifferent to it!). We hope this will shed some light on the theoretical proof of the smooth sensitivity - which will also be enriched with additional explanations to ease its understanding in the revised version of our paper.

Notations. Regarding the notation clarity issue you point out, it will be modified in the final version of our paper to match that of Rivest (1987).

C. Reviewer Response

Thanks, that answers my questions except the second part of Q3: while empirical testing cannot replace formal proofs it can often be useful to double-check claims for a few numerical examples to spot minor mistakes and make it easier for readers to gain some confidence in the major results without having to verify all proofs, particularly if the proofs are harder to follow.

IV. Reviewer #3

A. Original Review

1) Summary: The paper considers problem of privately estimating the rule lists using gini impurity under the smooth sensitivity framework. The authors propose a private version of greedy algorithm to learn the rule lists utilizing the smooth sensitivity and empirically show improved performance on real world datasets.

2) Strengths: The analysis of smooth sensitivity for Gini impurity is a novel contribution. The paper is well written and the experimental results depict improved performance in terms of accuracy for larger ϵ and bigger datasets.

3) Weakness: The problem setting this work considers seems limited - Gini impurity for rule lists when the rule set is publicly available. It's not clear how straight forward it is to extend this for other interpretable machine learning models or when the rule set needs to be iterated over or found using some heuristic. The experiments also seem limited. Current experiments seem to suggest only dataset size matters for obtaining better accuracy with smooth sensitivity, but it's not clear if additional factors like number of rules, their complexity or letting each mechanism choose appropriate number of rules for given privacy budget affects accuracy.

Minor comment: Fig. 3 (a) is not very readable due to the purple color hiding some curves for $\epsilon < 0.1$ part.

4) Questions:

- *A more fair comparison might be to tune hyper parameters per mechanism (which might need some privacy budget). If any experiments were performed, did the authors observe any difference?*
- *In last paragraph for section 5.2, the authors mention "noisy counts remains nonetheless interesting when mined ruleset is pre-processed beforehand to small cardinality ..." - which seems to be the case for datasets considered but then they comment focusing on the noisy gini version. Could the authors expand on this comment?*

5) Limitations: Discussed in weakness above.

6) Soundness: 3: good

7) Presentation: 3: good

8) Contribution: 2: fair

9) Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

10) Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

B. Authors' Rebuttal

Thank you for your feedback. Hereafter, we provide insights that we hope will answer your key concerns.

On the fact that we assume knowledge of the rule set. The ruleset mining and the (private) learning algorithm are two problems that we consider distinct as is the modus operandi of other papers on this topic. We consider only the algorithmic vision of the problem: for a given set of entries (the ruleset and the training data) then our algorithm is Differentially Private with respect to these input data. The best way to deal with the eventual need for a private ruleset is to allocate two distinct privacy budgets ϵ_{rules} and ϵ_{algo} so that the mining of the ruleset satisfies ϵ_{rules}-DP and the algorithm is ϵ_{algo}-DP. This two-step approach is indeed common practice, even in the non-DP setting (for instance, the popular CORELS algorithm for learning optimal rule lists also takes as input a set of pre-mined rules).

Minor comment: Figure 3 readability. Thank you for pointing out the lack of visibility in Figure 3, we will improve it in our revised version.

On considering additional factors such as the number of rules, their complexity and on performing a per-mechanism hyperparameters tuning. Other factors can indeed influence the privacy-utility trade-off. We believe this is an interesting discussion, and we will include it in the final version of our paper - thank you for your suggestion!

- Effect of the size of the ruleset: Naturally, with rules of higher cardinality (i.e., a higher number of conditions on the attributes), we can expect higher accuracy since the splits would be more refined. Observe that it would not affect the privacy budget of the model since the noisy max report mechanism is independent of the number of elements from which the argmax is searched. However, it yields an exponential increase in time complexity.

- Effect of the number of rules within the rule list and of the minimum support parameter: Optimizing the number of rules within the rule list proves to be interesting. Indeed, the maximum number of rules K heavily influences the privacy budget per node but it is also related to the minimum support condition. There can be no more than $min(K, \lfloor \frac{1}{\lambda} \rfloor)$ rules in the output rule list. Decreasing λ enables more rules but there is a trade-off with the precision of the Laplace noise using smooth sensitivity (the higher λ the less noise added). Overall, our smooth sensitivity method consistently beats the global sensitivity methods for all considered values of K and λ, as was theoretically expected (naturally without considering extreme values such as $K < 3$ or $\lambda \geq 0.25$ since it is simply a matter of model under-fitting). A value of $K = 5$ was on average the best performing for all models. When using $K = 7$, this value was most of the time not reached (i.e., in practice, the number of produced rules was smaller than K) as the minimum support condition was not achieved anymore. Models using global sensitivity were also terminated before reaching this depth since the algorithm stops when the Gini is not improved anymore. In the final version of our paper, we will clearly depict all experiments we performed and the key insights they provided (as aforementioned). In particular, they answer the question on performing a per-mechanism hyper-parameters tuning - since for the different tested combinations of hyper-parameters, the observed trends were consistent with those exposed in the paper.

On the choice of the noisy gini variant rather than the noisy counts. The "Noisy Gini" mechanism leads to better accuracy than "Noisy Counts" at low ϵ. Conversely, the "Noisy Counts" yield better utility for high values of ϵ - but such values do not provide meaningful privacy guarantees. We chose not to consider the "Noisy Counts" mechanism since it is only competitive for (meaningless) high values of ϵ. This holds true even for low-cardinality rule sets, although the trend is slightly mitigated. We will clarify this aspect in the final version of our paper, and apologize for the misleading sentence.

C. Reviewer Response

I thank the authors for their response. I feel the work could be made stronger with more empirical evaluations but with the current state I am still on the borderline. Based on it and the rest of discussions, I maintain my score.

V. Reviewer #4

A. Original Review

1) Summary: This paper proposes a differentially private implementation of a greedy algorithm for learning rule-list models, which classify input according to interpretable rules. The authors establish a smooth sensitivity bound for the Gini impurity index as a building block to their approach, and empirically compare their method to DP rule-list models based on global sensitivity.

2) Strengths And Weaknesses:

- *The goal of building interpretable DP models is certainly a worthy one, and the smooth sensitivity of the Gini impurity seems like a really nice result that could also serve as a building block for other interpretable DP algorithms such as DP decision trees.*
- *However it looks to me like the paper has a few big holes. One of my concerns is soundness: the privacy guarantee for Algorithm 1 is stated at a high level and is really more of a proof sketch. The proof uses only parallel and sequential composition, but due to the interactive setting (e.g., at each iteration the support and the available rules change according to the previous iteration) I am skeptical as to why adaptive composition wouldn't be necessary. The proof for the smooth sensitivity of the Gini impurity — one of the main contributions — is poorly presented and therefore it's difficult to verify whether it's correct.*
- *The experiments also raise some red flags for me. I am surprised that Figure 2 includes ϵ values as high as 10^4, and similarly large ϵ's in Figure 3. It also looks like the baselines based on global sensitivity ("Noisy Gini" and "Noisy Counts") would satisfy pure DP, in which case I think it's misleading to compare them to (ϵ,δ)-DP algorithms as the privacy budgets won't match up. On top of that, I feel like there is a lack of appropriate baselines: the paper emphasizes that this is follow-up work to an open question posed in a survey paper of DP decision trees and DP random forests, so it seems reasonable to compare the proposed DP greedy rule-list algorithm with prior related work.*

3) Questions:

- *Do "Noisy Gini" and "Noisy Counts" shown in Figure 2 satisfy pure DP? If not, what is the value of δ?*
- *How should I interpret the pictures in the proof of the smooth sensitivity of the Gini impurity (Appendix A.1)?*
- *Why wouldn't adaptive composition apply to the proof of the privacy guarantee for Algorithm 1?*
- *Why are the brackets reversed in Appendix A?*

4) Limitations: The authors have adequately addressed the limitations and impact of their work.

5) Soundness: 2: fair

6) Presentation: 2: fair

7) Contribution: 3: good

8) Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

9) Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

B. Authors' Rebuttal

Thank you for you detailed feedback and precise questions. We use this response to clarify some of the concerns discussed in the report. We hope these clarifications can improve your perception and evaluation of this paper.

We answer your two first concerns (regarding the privacy guarantee for Algorithm 1 and the proof of the smooth sensitivity of the Gini Impurity) when answering your detailed questions, after addressing your other comments.

I am surprised that Figure 2 includes ϵ values as high as 10^4, and similarly large ϵ's in Figure 3. As mentioned in the paper, we focus on the range $[0.1, 20]$ for ϵ. As to why values as high as 10^4 appear in the graph (with a log scale on the x-axis), it is simply a means to ensure that the DP algorithms converge towards the baseline algorithm when the privacy becomes very loose (and how fast they do). We do absolutely agree that such high values of privacy do not yield any privacy

guarantees and that is why when we chose between the "Noisy Counts" and "Noisy Gini" methods, we did not account for the fact that the "Noisy Counts" method was more accurate past a high value of ϵ.

It also looks like the baselines based on global sensitivity ("Noisy Gini" and "Noisy Counts") would satisfy pure DP, in which case I think it's misleading to compare them to (ϵ,δ)-DP algorithms as the privacy budgets won't match up. We specifically picked a value of δ equal to $\frac{1}{N^2}$ with N the size of the dataset (in the case of Adult $\delta = 4e-10$). We think the comparison is reasonable for such values of δ. Nevertheless, the Cauchy mechanism using Smooth Sensitivity is guaranteed to yield pure Differential Privacy and when comparing them to the Global Sensitivity mechanisms, the increase in accuracy is very much apparent. In the revised version of the paper, we distinguish in the caption of Figure 3 between the pure DP and the approximate DP mechanisms.

On top of that, I feel like there is a lack of appropriate baselines: the paper emphasizes that this is follow-up work to an open question posed in a survey paper of DP decision trees and DP random forests, so it seems reasonable to compare the proposed DP greedy rule-list algorithm with prior related work. The aim of our paper is to propose a new Differentially Private mechanism for greedy learning algorithms using the Gini Impurity. Indeed, our key contribution is a privacy tool, whose usefulness is illustrated for learning Rule Lists. While we consider rule lists induction as an interesting application for such mechanism (in particular, because they are less studied in the literature than decision trees or random forests), the main purpose of our experiments is to assess how our smooth sensitivity-based approaches compare to traditional ones (based on global sensitivity) in terms of accuracy-privacy tradeoffs. However, more complex models such as Random Forests or larger Decision Trees may reach higher predictive performances. Integrating our proposed framework with such models' learning is a promising direction, as outlined in our answer to Q2 of Reviewer 12jS. As for comparison with other DP learning algorithms producing Rule Lists, we could not find any experimental results.

Do "Noisy Gini" and "Noisy Counts" shown in Figure 2 satisfy pure DP? If not, what is the value of δ ? The "Noisy Gini" and "Noisy Counts" in Figure 2 are two versions of a DP algorithm using a Laplace Mechanism with Global Sensitivity. Both therefore satisfy pure DP.

How should I interpret the pictures in the proof of the smooth sensitivity of the Gini impurity (Appendix A.1)? The proof is a proof by exhaustion and we would like to apologize for the lack of clarity that was relayed by several reviewers. The variational tables were designed to help at observing the variations of the function graphically according to the different cases (namely the values of β and the position of the roots t_1 and t_2). Here is a brief outline of the method: the smooth sensitivity is computed as $S^*_{\mathcal{G},\beta}(x) = \max_{k\in\mathbb{N}}\ e^{-k\beta}\mathcal{T}_k(x)$. We first determine the function $\mathcal{T}_k(x)$ in which x is a dataset. We observe that it does not depend on the actual value of the dataset but solely the number of samples it contains. We obtain $\mathcal{T}_k(x) = g\big[\max(1, ||x||_1 - k)\big]$. Given that we managed to obtain a closed form for $\mathcal{T}_k(x)$ in which k directly intervenes, we now consider x to be a parameter and put k as a variable of our function hence the introduction of a new function: $\xi_{x,\beta}(t) = \ e^{-k\beta}\mathcal{T}_k(x)$. We study this function on $\mathbb{R}^+$ since it is differentiable. Through cancelling the derivative, we are able to find the minima and the maxima of the function. However, since these are $\mathbb{R}$-valued maxima, we finally truncate them to the closest higher and inferior integers to obtain the smooth sensitivity. Note that since we associate the sign of the derivative to a polynomial, it gives us extra control over the monotony of the function ξ since we know a polynomial takes the sign of the highest degree coefficient outside of the roots (granted that they exist) and the opposite inside the roots. For certain values of β such maxima may not exist because they are computed as the roots of a polynomial whose values depend on β. For this reason, we state that the formula should encompass t only if it is well defined (i.e., the value inside the square root is not negative) and otherwise replace it by 0. Proving that the smooth sensitivity only depends on $||x||$ is a core result of our approach. Thanks to this, at each iteration we only need to query the number of elements left to classify and apply the same smooth sensitivity to all rules (the Gini depends on the split made by the rule, but its smooth sensitivity is indifferent to it!). We hope this will shed some light on the theoretical proof of the smooth sensitivity - which will also be enriched with additional explanations to ease its understanding in the revised version of our paper.

Why wouldn't adaptive composition apply to the proof of the privacy guarantee for Algorithm 1? In the previous literature on learning DP tree-based ML models (see the thorough survey on Differentially Private Decision Trees by Fletcher & Islam), the privacy budget is always determined using parallel and sequential composition. To the best of our knowledge, we have not found any article using adaptive composition for this particular application. Although it may be challenging, applying advanced composition mechanisms such as adaptive composition could be an interesting future work to save DP budget and improve the accuracy-privacy trade-off. We mention this as a future work in the revised version of our paper.

Why are the brackets reversed in Appendix A? Our apologies for the confusion, reversed brackets are an equivalent notation for parenthesis when denoting open intervals. In particular : $[a, b[\equiv [a, b)$ and $]a, b] \equiv (a, b]$.

C. Reviewer Response

Thanks for the detailed response to my review. Based on that and the other reviews I will raise my score; I do appreciate the novelty of the paper and the usefulness of its results, but I am still very much on the borderline as I feel that the empirical evaluation and the clarity of the theoretical contribution could be much improved.

VI. Description of the improvements contained in the current version of the paper

Overall, the reviewers acknowledged the significance of our core contribution, namely establishing the smooth sensitivity of the Gini impurity to allow the learning of interpretable models that comply with DP while minimally harming predictive accuracy. As pointed out by the reviewers' responses (after the authors' rebuttals), most concerns were resolved by the additional clarifications and explanations we provided during the rebuttal period and integrated in the revised version of our paper. Besides clarifications, minor changes and detailed explanations, we additionally performed the following updates, as suggested by the reviewers:

- (Meta review, Reviewers #2 and #4) We have modified the proof of Theorem IV.1 (Smooth Sensitivity of the Gini impurity) to enhance its readability, and we have included a detailed proof sketch in the main paper body.
- (Meta review, Reviewer #2) We now discuss as an interesting future work the use of the inverse sensitivity mechanism of Asi and Duchi [1] along with the Gini impurity, leveraging the fact that we now have a closed formula for its smooth sensitivity.
- (Meta review) We have clarified the choice of a log scale on the x-axis of Figure 4. Indeed, a linear scale would completely hide the smallest values of ϵ (in particular, 0.01 and 0.1). On the contrary, our chosen values for ϵ range from very tight privacy budgets to loose budgets, the objective being to evaluate all privacy regimes, from very strong to loose privacy guarantees. Note that for these reasons, log scales are often used in the literature (*e.g.,* in the seminal work of Abadi et al. [2]).
- (Reviewer #1) Although they are not directly related to our work, we briefly discuss the works of Thaler et al. [3] and Daniely and Feldman [4]. More precisely, Daniely and Feldman [4] discuss the sample complexity of learning decision lists in a non-interactive local differential privacy context while Thaler et al. [3] focus on the design of efficient privacy-preserving data publishing. In particular, they consider the problem of privately releasing a database whose rows consist of pre-defined decision lists. Hence in this setting, the problem considered is not the privacy-preserving learning of decision lists.
- (Reviewer #1, Reviewer #4) We have elaborated on the experimental comparison with Fletcher and Islam [5]. Indeed, we compared the difference between the DP model and the baseline model in terms of accuracy at $\epsilon = 1$ on the UCI Adult Income dataset. Unfortunately, the results from Fletcher and Islam [5] were mostly based on synthetic datasets we could not compare on. For this level of privacy, our smooth sensitivity-based algorithm has less than 0.5% (the accuracy of the proposed DP algorithm is 78.7% and 79.1% for the non private version) in accuracy loss against at least 1.0% for theirs (the accuracy of the DP algorithm is 82% and the non DP version is about 83%). Thus, even though (non-DP) decision trees perform better than (non-DP) rule lists, we estimate our trade-off between privacy and accuracy to be better as our mechanisms incur a lower utility cost, and this trend can only increase with larger datasets.
- (Reviewer #1) We have elaborated on the fact that the learnt rule lists, because they exhibit a simple and sparse structure, are already rather resilient to Membership Inference Attacks (MIAs), even without the use of DP. While this can be seen as an intrinsic advantage of using such models, this nevertheless does not provide formal DP guarantees, and the empirical vulnerability measure are moderately but consistently greater for non-DP rule lists than for their DP counterparts, as shown in Table II.
- (Reviewer #1) We have included an additional discussion regarding our choice of a baseline rule lists learning algorithm. More precisely, we chose to focus on greedy learning algorithms using the Gini impurity as information gain, because this framework encompasses a wide range of learning algorithms and models (in particular, beyond rule-based ones, decision trees and random forests). Considering optimal learning algorithms such as CORELS [6, 7] is possible, and is now discussed as an interesting research direction. However, several technical details - mainly related to the fact that CORELS aims at learning optimal models - should be carefully considered as mentioned in our revised paper.
- (Reviewer #1) We have clarified the fact that the global sensitivity of the Gini impurity is considered "high" compared to that of counting queries, related to the range of values these queries can take. More precisely, the Gini impurity has global sensitivity of 0.5 but lies in $[0, 1]$ while counting queries have global sensitivity 1 but take values in $[0, n]$.
- (Reviewer #2, Reviewer #4) We elaborate on the differences of applying our smooth sensitivity based framework for learning decision trees or random forests (as was done by Fletcher and Islam [8] and Fletcher and Islam [5]). Crucially, we explain how decision trees, because they intrinsically have disjoint supports over all their leaves, can inherently benefit from parallel composition while rule lists can not. This leads to tighter privacy bounds, and was theoretically shown by Ferry et al. [9] to leak less information regarding the training data compared to rule lists.
- (Reviewer #3) We briefly discuss the influence of the different hyperparameters of our method, as well as the choice of their value, in a new dedicated subsection (V-D).
- (Reviewer #3) We further elaborate on the choice of the "noisy Gini" variant for our DP baselines considering global sensitivity (compared to the "noisy counts" one), providing new curves and detailed explanations. Indeed, the objective of this discussion was to select the best DP greedy rule lists learning algorithm to compare against our smooth sensitivity

algorithms.

- (Reviewer #4) While - in line with the literature on learning DP tree-based ML models [10] - we consider parallel and sequential composition for the analysis of the privacy budget of our proposed algorithm, we now mention the use of adaptive composition [11] as a promising research direction to possibly tighten the privacy analysis of our approach.
- (Reviewer #4) As stated in the paper, we focus on privacy budgets $\epsilon \in [0.01, 100]$ when comparing the different DP learning methods. However, larger values up to $\epsilon = 10^4$ (which provide no meaningful privacy guarantees) appear in Figures 2 and 3 as a means of verifying the asymptotic behaviour of the two compared variants of global sensitivity-based approaches. This is clarified in the revised version of the paper.
- (Reviewer #4) We have updated the legend of Figure 4 to recall which methods satisfy pure or approximate DP. In particular, our proposed framework based on smooth sensitivity is able to provide either pure or approximate DP, depending on the distribution of the added noise. In particular, the use of Cauchy noise within our smooth sensitivity framework provides pure DP guarantees while Laplace noise provides approximate DP. In either case, our framework provides better utility than the other considered approaches relying on global sensitivity, for similar (pure of approximate) DP budgets.

REFERENCES

[1] H. Asi and J. C. Duchi, "Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms," *Advances in neural information processing systems*, vol. 33, pp. 14 106–14 117, 2020.

[2] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, oct 2016.

[3] J. Thaler, J. Ullman, and S. Vadhan, "Faster algorithms for privately releasing marginals," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2012, pp. 810–821.

[4] A. Daniely and V. Feldman, "Locally private learning without interaction requires separation," *Advances in neural information processing systems*, vol. 32, 2019.

[5] S. Fletcher and M. Z. Islam, "Differentially private random decision forests using smooth sensitivity," *Expert Systems with Applications*, vol. 78, p. 16–31, 2017.

[6] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin, "Learning Certifiably Optimal Rule Lists," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, p. 35–44.

[7] ——, "Learning Certifiably Optimal Rule Lists for Categorical Data," *Journal of Machine Learning Research*, vol. 18, no. 234, pp. 1–78, 2018.

[8] S. Fletcher and M. Islam, "A differentially private decision forest," in *Proceedings of the Thirteenth Australasian Data Mining Conference (AusDM 15)*, 2015, pp. 99–108.

[9] J. Ferry, U. Aïvodji, S. Gambs, M.-J. Huguet, and M. Siala, "Probabilistic Dataset Reconstruction from Interpretable Models," in *2nd IEEE Conference on Secure and Trustworthy Machine Learning*, Toronto, Canada, Apr. 2024.

[10] S. Fletcher and M. Z. Islam, "Decision Tree Classification with Differential Privacy: A Survey," *ACM Comput. Surv.*, vol. 52, no. 4, 2019.

[11] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, p. 211–407, 2014.