Differentially Private Fair Support Vector Machines

Lennon Seiders, Meera Kumar

December 5, 2024

Abstract

In high stakes classification tasks such as college admissions, predicting loan defaulting, criminal recidivism, or medical diagnoses, machine learning is often employed [Far+23][BHN23][Fel+15]. However, this does raise concerns around the fairness and privacy-preserving nature of such classification methods, especially when drawing upon crucial protected data or using sensitive attributes of individuals. For this paper, we define fairness as the satisfaction of equalized odds, which serves to balance the cost of misclassification. In the context of high-stakes classification tasks, historically marginalized and disadvantaged groups tend to experience higher error rates, and we view equalized odds as an effective means of countering this. We also assume that satisfying ϵ-differential privacy is sufficient in preserving the privacy of individuals in the provided dataset. This project introduces an approach to ensure a differentially private Support Vector Machine (SVM) by perturbing its support vectors in the post-processing phase. We demonstrate that this method fails to improve the upper bound on error and fairness violations that is proven in Jagielski et al. (2019) [Jag+19] for a generic classifier, and we explore the implications of this in the Discussion and Future Work.

1 Introduction

Consider a scenario in which a bank is evaluating loan applications. The bank employs a machine learning model to predict loan default risk based on applicant data, which includes sensitive attributes like race, gender, or income bracket. In this high-stakes decision-making environment, major concerns arise: privacy, interpretability, and fairness. Private data owners expect their personal information to remain confidential, while regulators and ethical guidelines demand that the bank's model not unfairly disadvantage certain demographic groups.

This scenario illustrates the utility of a differentially private Support Vector Machine (SVM) that satisfies equalized odds. Differential privacy ensures that an individual's inclusion or exclusion in the dataset does not significantly influence the model's output, safeguarding sensitive information. Meanwhile, the equalized odds criterion ensures that the model's true positive and false positive rates are balanced across all demographic groups, mitigating systemic bias.

This paper addresses the intersection of these two critical dimensions by exploring the implementation of a differentially private, fair SVM. We focus on achieving privacy through post-processing perturbations to the model's support vectors while maintaining fairness through equalized odds. By combining these approaches, we aim to create a classifier suitable for sensitive applications where both ethical and regulatory standards must be met.

In the following sections, we provide background and review related work in the domains of fairness, privacy, and SVM optimization. We then describe our approach to introducing differential privacy and fairness constraints to SVMs and analyze theoretical trade-offs, followed by a review of our results and ideas for future work.

2 Previous Work

Our work builds on Jagielski et. al (2019) [Jag+19], where Jagielski et. al define methods to construct a differentially private fair classifier through both in-processing and post-processing, comparing each of these methods and their viability across multiple datasets. This work draws upon equalized odds as defined in Hardt, Price, Srebro (2016) [HPS16] as a notion of fairness, which we use in this paper as well. The post-processing method in Jagielski et. al is introduced as a differentially private fair learning algorithm that is an adaptation of the fair learning algorithm defined in Hardt, Price, Srebro (2016) [HPS16], in which a classifier is derived by selecting the classifier with the lowest error out of all fair classifiers.

Research by Ruan et. al. (2022) [Rua+23] illustrates how in the case of SVMs specifically, given an unfair classifier, introducing differential privacy constraints in post-processing will worsen metrics such as the TPR between groups distinguished by sensitive attribute. The way that these differentially private algorithms are designed is through three different methods: approximate minimal perturbation (adding noise to the loss function of the margin classifier, also called objective perturbation), differentially private stochastic gradient descent (adding noise to gradients), and private convex permutation-based stochastic gradient descent (introducing noise to the final model parameters, also known as output perturbation). However, Ruan et. al. also prove experimentally that in the case of a fair model (ie. a minimal TPR gap between groups in the dataset), introducing differential privacy does not significantly affect model fairness.

In a similar vein, Fish, Kun, and Lelkes (2016) [FKL16] demonstrate that fairness (in this case, in the form of statistical parity) can be achieved for a margin boundary without a significant impact to classification error. This paper had a significant influence on our initial efforts to achieve differential privacy and equalized odds for an SVM, however it became clear that a more complex approach was required to achieve equalized odds in lieu of statistical parity, which only requires the shifting of the decision boundary for the protected group.

Zafar et. al (2019) [Zaf+19] proposes a constraint-based framework to design fair margin-based classifiers, including SVMs. The authors propose a measure of decision boundary unfairness, and by incorporating this measure into the training process, the framework enables a balance between accuracy and fairness. Experiments on synthetic and real-world datasets demonstrate the framework's effectiveness in achieving fair classification outcomes.

Xu et. al. (2020) [XDW20] propose a new post-processing method to introduce differential privacy to a classifier called DPSGD-F, which is another means of achieving gradient perturbation, similar to differentially private stochastic gradient descent. The motivation behind this is that the DPSGD used by Ruan et. al. does not account for the impact on differential privacy that group sample size and group clipping bias have.

Agarwal et. al (2018) [Aga+18] introduces a method to ensure fairness in binary classification by reducing the problem to cost-sensitive classification tasks. It supports fairness definitions like demographic parity and equalized odds, producing a classifier that balances accuracy and fairness. The approach is versatile, works with various algorithms, and shows competitive performance across datasets compared to prior methods.

Cummings et. al. (2019) [Cum+19] studies whether privacy and fairness are simultaneously achievable in different models, using equalized odds. The authors prove algorithmically that it is theoretically impossible for a differentially private model to achieve exact fairness, even in the case of having full distributional access to the dataset being used. Building off of this, the authors prove the existence of a PAC (Probably Approximately Correct) learner — a mathematical model framework used in computational learning theory — that is differentially private and satisfies approximate equality of true positive rates across all values for a protected attribute (equal opportunity, a more relaxed version of the equalized odds constraint).

3 Background

3.1 Differential Privacy

We introduce differential privacy as a robust, mathematical means of ensuring privacy-preserving data analysis. What differential privacy promises is that a data subject will not be affected by allowing their data to be used in a dataset. This means that it will be nearly impossible to identify a subject as a contributor to a dataset that follows the constraint of differential privacy. Additionally, because differentially private algorithms inherently limit each participant's influence on the outcome, they reduce incentives to misreport information [MT07].

Formally, a randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ε, δ)-differentially private if for all $\mathcal{S} \subseteq \mathit{Range}(\mathcal{M})$ and for all databases $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $||x-y||_1 \leq 1$:

$$\mathbb{P}[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\varepsilon)\mathbb{P}[\mathcal{M}(y) \in \mathcal{S}] + \delta$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, we say that $\mathcal{M}$ is ε-differentially private. [DR+14]. Informally, differential privacy protects individuals' private information by ensuring that the output of an analysis on a dataset remains nearly the same whether or not any single individual's data is included.

To implement $\epsilon-$differential privacy in this paper, we will rely on the Laplace mechanism to introduce noise to the data. The Laplace function samples noise on a Laplacian distribution determined by the sensitivity of the target dataset. We utilize the following from the Jagielski [Jag+19] paper:
Definition 1 The ℓ_1-sensitivity of $f : \mathcal{D}^m \to \mathbb{R}^k$ is

$$\Delta f = \max_{\substack{D, D' \in \mathcal{D}^m \\ D \sim D'}} \left\| f(D) - f\left(D'\right) \right\|_1 ,$$

which measures how much the ℓ_1 norm of the function f changes if up to one entry is changed in the database.
Definition 2 (Laplace Mechanism [Dwo+06]). Given a query function $f : \mathcal{D}^m \to \mathbb{R}^k$, a database $D \in \mathcal{D}^m$, and a privacy parameter ϵ, the Laplace mechanism outputs:

$$\tilde{f}_\epsilon(D) = f(D) + (W_1, \ldots, W_k)$$

where W_i 's are i.i.d. random variables drawn from $\mathrm{Lap}(\Delta f / \epsilon)$.

Theorem 1 (Privacy vs. Accuracy of the Laplace Mechanism [Dwo+06]). The Laplace mechanism guarantees ϵ-differential privacy and that with probability at least $1 - \delta$,

$$\left\| \tilde{f}_\epsilon(D) - f(D) \right\|_\infty \leqslant \ln\left(\frac{k}{\delta}\right) \cdot \left(\frac{\Delta f}{\epsilon}\right)$$

3.2 Equalized Odds

In this paper, we define fairness through the constraint of equalized odds. Equalized odds is a statistical fairness constraint requiring the equalization of true positive rate (TPR) and false positive rate (FPR) across all values for a protected attribute. Formally, for an outcome Y, equalized odds is satisfied if protected attribute A and predictor $\hat{Y}$ are independent conditional on Y. [HPS16]

$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 1, A = a\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 0, A = a\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 0, A = b\}$$

The outcomes $y = 0$ and $y = 1$ represent the true negative rate and true positive rate, respectively.

The motivation for equalized odds as a notion of fairness lies in the idea of equal claim to acceptance [HPS16]. This definition requires that all groups corresponding to the values $a, b \in A$ have the same TPR and FPR. Therefore, violating this constraint indicates that different groups experience different costs of misclassification, which would align with the idea that higher error rates are historically correlated with marginalized groups. By requiring this parity in error rates, decision makers are incentivized to improve error rates by building better models and collecting better data, reducing the likelihood of seeing the positive feedback loop that we commonly associate with using machine learning with this kind of decision making [BHN23].

3.3 Support Vector Machines

Support Vector Machines (SVMs) are a class of supervised learning algorithms widely used for classification and regression tasks, valued for their ability to handle high-dimensional data and define clear decision boundaries. SVMs identify a hyperplane that divides data into distinct classes with the maximum possible margin between the nearest points of each class. Because in many cases data is not linearly separable in its original feature space, SVMs utilize a ”kernel trick” to map data to a higher dimension where a dividing hyperplane may exist. SVMs are used for crucial decision-making tasks [Wan+10] because of this interpretability and effectiveness in high-dimensions, and in this paper we prove their compatibility with differentially private data and equalized odds.

We define the optimization problem and variables for a soft-margin SVM as

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i \end{aligned}$$

– $\mathbf{w}$: Weight vector defining the orientation of the decision boundary.

– b: Bias term defining the position of the decision boundary.

- ξ_i: Slack variable for the i-th data point, allowing for margin violation.
- C: Regularization parameter balancing the trade-off between maximizing the margin and minimizing classification error.
- y_i: Label of the i-th data point, typically $+1$ or -1.
- $\mathbf{x}_i$: Feature vector of the i-th data point.
- n: Total number of training data points.

[Cor95]

In this paper, we work with the support vectors, which can be defined as the feature vectors x_i which satisfy the following condition:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) = 1 - \xi_i$$

Support vectors are either on the margin or within it, defining the optimal hyperplane. The position of the hyperplane is directly influenced by these vectors. The support vectors ensure that the margin between the classes is as wide as possible, adhering to the maximum margin principle. We also note that typically, the number of support vectors make up a fraction of the dataset. This means that a support vector machine does not necessarily need all of the training examples to define the optimal decision boundary.

4 Differentially Private Fair SVM Algorithm

In this section we introduce our algorithm for differentially private fair post-processing of an SVM with k support vectors. This algorithm is a variant of the "DP-Postprocessing" algorithm from Jagielski et. al, which is derived from the paper by Hardt et. al.

This algorithm takes as input parameters determining privacy and equalized odds constraints, the joint distribution of sensitive attributes A and binary labels Y, and support vectors x_i^{sv}, which define classifier $\hat{Y}$. The algorithm uses the Laplace mechanism to introduce noise to the support vectors and joint distribution separately, and then combines these to form joint distribution $\tilde{q}_{\hat{y}ay}$. This composite joint distribution is then input into the linear program $\widetilde{\text{LP}}$, which is identical to that in the Jagielski et. al paper, to get the optimal minimizer $\tilde{p}^{\star}$. This minimizer allows us to then make formal guarantees on the satisfaction of the privacy parameter ϵ and fairness violation parameter γ with confidence $1 - \beta$.

```
Algorithm 1 $\epsilon$-differentially private fair post-processing
Input: privacy parameter $\epsilon$, confidence parameter $\beta$, fairness violation $\gamma$, SVM $\widehat{Y}$ with $k$ support vectors $x_i^{sv}$, joint distribution $\hat{q}_{ay} = \widehat{\mathbb{P}}[A = a, Y = y]$.
  → Sample $W_{x^{sv}} \overset{\text{i.i.d.}}{\sim} \text{Lap}(2/k\epsilon)$ for all $x_i^{sv}$.
  → Perturb each support vector $\hat{x}_i^{sv} : \hat{x}_i^{sv} = x_i^{sv} + W_{x^{sv}}$.
  → Sample $W_{ay} \overset{\text{i.i.d.}}{\sim} \text{Lap}(2/m\epsilon)$ for $\hat{\mathbb{P}}[A = a, Y = y]$.
  → Perturb $\hat{q}_{ay} : \tilde{q}_{ay} = \hat{\mathbb{P}}[A = a, Y = y] + W_{ay}$.
  → Calculate $\tilde{q}_{\hat{y}ay} = \tilde{P}[\widehat{Y} = \hat{y}, A = a, Y = y]$.
  → Solve $\widetilde{\text{LP}}$ to get the minimizer $\tilde{p}^{\star}$.
Output: $\tilde{p}^{\star}$, the trained classifier $\widehat{Y}$
```

$\widetilde{\text{LP}}$: **ϵ-Differentially Private Linear Program:**

$$\begin{aligned}
&\arg\min_{p} \quad \widetilde{\text{err}}\left(\widehat{Y}_p\right) \\
&\text{s.t. } \forall a \underset{a\neq 0}{\in} \mathcal{A} \quad \Delta\widetilde{\text{FP}}_a\left(\widehat{Y}_p\right) \leqslant \gamma + \frac{4\ln(4|\mathcal{A}|/\beta)}{\min\{\tilde{q}_{a0}, \tilde{q}_{00}\}\, m\epsilon} \\
&\Delta\widetilde{\text{TP}}_a\left(\widehat{Y}_p\right) \leqslant \gamma + \frac{4\ln(4|\mathcal{A}|/\beta)}{\min\{\tilde{q}_{a1}, \tilde{q}_{01}\}\, m\epsilon} \\
&0 \leqslant p_{\hat{y}a} \leqslant 1 \quad \forall \hat{y}, a
\end{aligned}$$

$$\begin{aligned}
\widetilde{\text{err}}\left(\hat{Y}_p\right) &:= \sum_{\hat{y},a} \left(\tilde{q}_{\hat{y}a0} - \tilde{q}_{\hat{y}a1}\right) \cdot p_{\hat{y}a} + \sum_{\hat{y},a} \tilde{q}_{\hat{y}a1} \\
\Delta\widetilde{\text{FP}}_a\left(\hat{Y}_p\right) &:= \left|\widetilde{\text{FP}}_a(\hat{Y}) \cdot p_{1a} + \left(1 - \widetilde{\text{FP}}_a(\hat{Y})\right) \cdot p_{0a} - \widetilde{\text{FP}}_0(\hat{Y}) \cdot p_{10} - \left(1 - \widetilde{\text{FP}}_0(\hat{Y})\right) \cdot p_{00}\right| \\
\Delta\widetilde{\text{TP}}_a\left(\hat{Y}_p\right) &:= \left|\widetilde{\text{TP}}_a(\hat{Y}) \cdot p_{1a} + \left(1 - \widetilde{\text{TP}}_a(\hat{Y})\right) \cdot p_{0a} - \widetilde{\text{TP}}_0(\hat{Y}) \cdot p_{10} - \left(1 - \widetilde{\text{TP}}_0(\hat{Y})\right) \cdot p_{00}\right|
\end{aligned}$$

5 Proof

It is proven in the Jagielski et. al paper [Jag+19] that the l_∞ norm $|\hat{q} - \tilde{q}\|_\infty \leq \frac{2\ln\left(\frac{4|A|}{\beta}\right)}{m\epsilon}$. This is derived from the theorem we've listed as Theorem 1 in this paper ($\left\|\tilde{f}_\epsilon(D) - f(D)\right\|_\infty \leqslant \ln\left(\frac{k}{\delta}\right) \cdot \left(\frac{\Delta f}{\epsilon}\right)$). For a more detailed explanation of this derivation, see Appendix A. From this upper bound, Jagielski et. al are able to guarantee an upper bound on the error of the perturbed classifier, as well as an upper bound on the sensitivity of perturbed false positive and true positive rates. Being able to bound the FPR and TPR sensitivities is beneficial as it allows us to guarantee preservation of $\gamma-$equalized odds with perturbed data. These upper bounds on error, FPR sensitivity, and TPR sensitivity are listed below:

With probability at least $1-\beta$,

$$\widehat{err}\left(\widehat{Y}_{\widetilde{p}^\star}\right) \leqslant \widehat{err}\left(\widehat{Y}_{\widehat{p}^\star}\right) + \frac{24|\mathcal{A}|\ln(4|\mathcal{A}|/\beta)}{m\epsilon}$$

and for all $a \neq 0$,

$$\begin{aligned}\Delta\widehat{FP}_a\left(\widehat{Y}_{\widetilde{p}^\star}\right) &\leqslant \gamma + \frac{8\ln(4|\mathcal{A}|/\beta)}{\min\left\{\hat{q}_{a0}, \hat{q}_{00}\right\} m\epsilon - 4\ln(4|\mathcal{A}|/\beta)} \\ \Delta\widehat{TP}_a\left(\widehat{Y}_{\widetilde{p}^\star}\right) &\leqslant \gamma + \frac{8\ln(4|\mathcal{A}|/\beta)}{\min\left\{\hat{q}_{a1}, \hat{q}_{01}\right\} m\epsilon - 4\ln(4|\mathcal{A}|/\beta)}\end{aligned}$$

Because we are perturbing the input data $\{A, Y\}$ and classifier output $\{\hat{Y}\}$ separately, we can determine a different upper bound accordingly as follows:

For the input data: Since we are adding noise to the same number of data points m, sensitivity $\Delta f = \frac{2}{m}$. However, k changes to account for the reduced number of outputs we are getting through the joint distribution $\{A, Y\}$ as compared to $\{\hat{Y}, A, Y\}$; $k = 2 \cdot |A|$ (as opposed to the $4 \cdot |A|$ that comes from having the added dimension of $\hat{Y}$ in the joint distribution).

For the output data: We are looking only at adding noise to the support vectors of the final classifier, so sensitivity $\Delta f = \frac{2}{k}$. Also, the k value in Theorem 1 will decrease to $k = 2$ since there are two possible outputs for $\hat{Y}$.

In order to set a clear upper bound, we take the sum of the two l_∞ norms produced by the perturbations:

$$|\hat{q} - \tilde{q}\|_\infty \leq \frac{2\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{2\ln\left(\frac{2}{\beta}\right)}{k\epsilon}$$

Now that we have established this upper bound on the $|\hat{q} - \tilde{q}\|_\infty$ norm produced after perturbing and combining the data, we can use it to modify the guarantees related to accuracy and fairness that Jagielski et. al prove for their algorithm for a generic classifier $\hat{Y}$. See Appendix A for our modified guarantees.

Since the l_∞ norm is directly proportional to each of the three bounds (relating to error, FPR sensitivity, and TPR sensitivity of the perturbed classifier), we can compare Jagielski et. al's l_∞ norm with our own to understand how our upper bounds will relate to theirs. We do this as follows:

Assume that our l_∞ norm is less than or equal to theirs. This would mean that our guarantees will be stronger than theirs (ie. we will have smaller upper bounds on error, as well as TPR and FPR sensitivity). We thus get the inequality:

$$\frac{2\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon}+\frac{2\ln\left(\frac{2}{\beta}\right)}{k\epsilon}\leq\frac{2\ln\left(\frac{4|A|}{\beta}\right)}{m\epsilon}$$

$$\frac{\ln\left(\frac{2|A|}{\beta}\right)}{m}+\frac{\ln\left(\frac{2}{\beta}\right)}{k}\leq\frac{\ln\left(\frac{4|A|}{\beta}\right)}{m}$$

$$\frac{\ln\left(\frac{2}{\beta}\right)}{k}\leq\frac{\ln\left(\frac{4|A|}{\beta}\right)-\ln\left(\frac{2|A|}{\beta}\right)}{m}$$

$$\frac{\ln\left(\frac{2}{\beta}\right)}{k}\leq\frac{\ln\left(\frac{\frac{4|A|}{\beta}}{\frac{2|A|}{\beta}}\right)}{m}$$

$$\frac{\ln\left(\frac{2}{\beta}\right)}{k}\leq\frac{\ln(2)}{m}$$

$$m\cdot\ln\left(\frac{2}{\beta}\right)\leq k\cdot\ln(2)$$

$$\ln\left(\frac{2}{\beta}\right)\leq\frac{k}{m}\ln(2)$$

Exponentiating both sides:

$$\frac{2}{\beta}\leq e^{\frac{k}{m}\ln(2)}=2^{\frac{k}{m}}$$

Solving for β:

$$\beta\geq\frac{2}{2^{\frac{k}{m}}}$$

$$\beta\geq 2^{1-\frac{k}{m}}$$

However, $\beta\in[0,1]$, and since the number of support vectors $k\leq$ the number of data points m, we know that $\frac{k}{m}\in[0,1]$. Thus, $1-\frac{k}{m}\in[0,1]$ as well. This means that $2^{1-\frac{k}{m}}\geq 1$. Thus, we find that if we assume our l_∞ norm to be less than or equal to that of Jagielski et. al, we arrive at a contradiction. So, through proof by contradiction, we have proven that $\frac{2\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon}+\frac{2\ln\left(\frac{2}{\beta}\right)}{k\epsilon}>\frac{2\ln\left(\frac{4|A|}{\beta}\right)}{m\epsilon}$, which means that our SVM-specific guarantees will be worse than those proven by Jagielski et. al in their algorithm for a generic classifier.

6 Discussion and Future Work

The implementation of a differentially private Support Vector Machine (SVM) under the equalized odds constraint highlights the complexities inherent in balancing privacy, fairness, and utility. Our analysis shows that perturbing support vectors in a post-processing algorithm does not outperform the theoretical upper bounds on misclassification error and fairness violations established in Jagielski et al. (2019).

A notable challenge that this paper illustrates is the trade-off between privacy guarantees and fairness outcomes. By introducing differential privacy through post-processing, the addition of noise significantly impacts the alignment of decision boundaries with the fairness criterion of equalized odds. This result underscores the need for more granular methods of incorporating noise, perhaps directly into the training process or via data-dependent perturbation strategies.

Another limitation of our approach is its reliance on the equalized odds fairness definition. While this metric is effective for balancing true positive and false positive rates, it does not address other fairness concerns such as subgroup accuracy or disparate impact, which may be more relevant in certain applications. Future work could explore alternative fairness definitions or hybrid approaches that balance multiple fairness objectives while maintaining privacy guarantees.

Potential avenues for further research:

- Alternative Perturbation Techniques. Investigating methods such as noise addition tailored to specific groups or regions of the decision boundary, which could improve the model's ability to satisfy fairness constraints without significantly compromising utility.

- Empirical Evaluations. Experiments on real-world datasets and sensitive application domains may provide better insight on the empirical error introduce by perturbing the support vectors in post-processing.

These challenges pose relevant avenues for future work as machine learning continues to be deployed in sensitive, high-stakes decision-making environments. Theoretical and empirical advancements in privacy-fairness trade-offs have significant implications for ethical and regulatory practices in AI.

7 Conclusion

In this work, we investigated the implementation of a differentially private Support Vector Machine classifier under the constraint of equalized odds. Our analysis and results show that even when carefully adjusting the decision boundary through perturbation of the support vectors in post-processing, our approach fails to improve the known theoretical bounds on misclassification error and fairness violation. As data-driven methods continue to be deployed in sensitive domains, it remains crucial to understand the trade-offs and theoretical limits involved. Attaining meaningful fairness and privacy is not merely a matter of combining known techniques; it may require very specific algorithmic strategies, richer theoretical frameworks, or more granular control over model training and post-processing stages. Looking forward, developing alternative private learning methods or incorporating alternative fairness notions could lead to notable results. Additionally, investigating more data-dependent perturbation strategies may yield stronger theoretical guarantees.

Appendix A

We prove the guarantees made on error, TPR sensitivity, and FPR sensitivity here for our SVM-specific algorithm.

Claim A.1.(ℓ_1-Sensitivity of $\hat{\boldsymbol{q}}$ to A). As in [Jag+19], Let $\hat{\boldsymbol{q}} = [\hat{q}_{\hat{y}ay}]_{\hat{y},a,y}$ be the empirical distribution of $\{\hat{Y}, A, Y\}$ and let $\Delta\hat{\boldsymbol{q}}$ be the ℓ_1-sensitivity of $\hat{\boldsymbol{q}}$ to A.

$$\Delta\hat{\boldsymbol{q}} = \max_{\substack{A,A' \in \mathcal{A}^m \\ A \sim A'}} \|\hat{\boldsymbol{q}}(A) - \hat{\boldsymbol{q}}(A')\|_1 = \frac{2}{m}$$

This, in conjunction with Theorem 1, gives us our l_∞ norm upper bound:

$$|\hat{q} - \tilde{q}\|_\infty \leq \frac{2\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{2\ln\left(\frac{2}{\beta}\right)}{k\epsilon}$$

Lemma A.2. Suppose $\min_{a,y}\{\hat{q}_{ay}\} > \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{4\ln\left(\frac{2}{\beta}\right)}{k\epsilon}$. We have that with probability $\geqslant 1-\beta$,

1. $\left|\widetilde{err}\left(\widehat{Y}_p\right) - \widehat{err}\left(\widehat{Y}_p\right)\right| \leqslant \frac{12|A|\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{12|A|\ln\left(\frac{2}{\beta}\right)}{k\epsilon} \quad ;\forall p.$
2. $\tilde{q}_{ay} > 0 \quad ;\forall a, y.$
3. $\left|\widetilde{FP}_a(\widehat{Y}) - \widehat{FP}_a(\widehat{Y})\right| \leqslant \frac{2\ln\left(\frac{2|A|}{\beta}\right)}{\tilde{q}_{a0}m\epsilon} + \frac{2\ln\left(\frac{2}{\beta}\right)}{\tilde{q}_{a0}k\epsilon}$, $\left|\widetilde{TP}_a(\widehat{Y}) - \widehat{TP}_a(\widehat{Y})\right| \leqslant \frac{2\ln\left(\frac{2|A|}{\beta}\right)}{\tilde{q}_{a1}m\epsilon} + \frac{2\ln\left(\frac{2}{\beta}\right)}{\tilde{q}_{a1}k\epsilon} \quad ;\forall a.$
4. $\left|\Delta\widetilde{FP}_a\left(\widehat{Y}_p\right) - \Delta\widehat{FP}_a\left(\widehat{Y}_p\right)\right| \leqslant \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{\min\{\tilde{q}_{a0},\tilde{q}_{00}\}m\epsilon} + \frac{4\ln\left(\frac{2}{\beta}\right)}{\min\{\tilde{q}_{a0},\tilde{q}_{00}\}k\epsilon}$, $\left|\Delta\widetilde{TP}_a\left(\widehat{Y}_p\right) - \Delta\widehat{TP}_a\left(\widehat{Y}_p\right)\right| \leqslant \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{\min\{\tilde{q}_{a1},\tilde{q}_{01}\}m\epsilon} + \frac{4\ln\left(\frac{2}{\beta}\right)}{\min\{\tilde{q}_{a1},\tilde{q}_{01}\}k\epsilon};\forall a, p.$
5. $\hat{p}^\star$, the optimal solution of $\widehat{LP}$, is feasible in $\widetilde{LP}$.

To prove this, using our l_∞ norm upper bound, we will do the following:

1. By definition of error as per the linear programs:

$$\forall_p \left|\tilde{\text{err}}\left(\hat{Y}_p\right) - \hat{\text{err}}\left(\hat{Y}_p\right)\right| \leq \Sigma_{\hat{y},a,y}\,|\tilde{q}_{\hat{y}ay} - \hat{q}_{\hat{y}ay}| + \Sigma_{\hat{y},a}\,|\tilde{q}_{\hat{y}a1} - \hat{q}_{\hat{y}a1}|$$

By definition of the $l-\infty$ norm:

$$\Sigma_{\hat{y},a,y}\,|\tilde{q}_{\hat{y}ay} - \hat{q}_{\hat{y}ay}| \leq 4|A| \cdot \|\hat{q} - \tilde{q}\|_\infty$$
$$\Sigma_{\hat{y},a}\,|\tilde{q}_{\hat{y}a1} - \hat{q}_{\hat{y}a1}| \leq 2|A| \cdot \|\hat{q} - \tilde{q}\|_\infty$$

By transitivity, we can say:

$$\left|\text{e}\tilde{\text{r}}\text{r}\left(\hat{Y}_p\right) - \text{e}\hat{\text{r}}\text{r}\left(\hat{Y}_p\right)\right| \leq \frac{12|A|\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{12|A|\ln\left(\frac{2}{\beta}\right)}{k\epsilon}$$

2.

$$\begin{aligned}\forall a, y : |\tilde{q}_{ay} - \hat{q}_{ay}| &= |\tilde{q}_{1ay} + \tilde{q}_{0ay} - \hat{q}_{1ay} - \hat{q}_{0ay}| \\ &\leq |\tilde{q}_{1ay} - \hat{q}_{1ay}| + |\tilde{q}_{0ay} - \hat{q}_{0ay}| \\ &\leq \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{4\ln\left(\frac{2}{\beta}\right)}{k\epsilon}\end{aligned}$$

by the stated assumption that $\hat{q}_{ay} > \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{4\ln\left(\frac{2}{\beta}\right)}{k\epsilon}$, we can conclude that $\tilde{q}_{ay} > 0$.

3. $\forall a :$

$$\begin{aligned}\left|\tilde{FP}_a(\hat{Y}) - \hat{FP}_a(\hat{Y})\right| &= \left|\frac{\tilde{q}_{1a0}}{\tilde{q}_{1a0} + \tilde{q}_{0a0}} - \frac{\hat{q}_{1a0}}{\hat{q}_{1a0} + \hat{q}_{0a0}}\right| \\ &= \left|\frac{\tilde{q}_{1a0}\left(\hat{q}_{1a0} + \hat{q}_{0a0}\right) - \hat{q}_{1a0}\left(\tilde{q}_{1a0} + \tilde{q}_{0a0}\right)}{\left(\tilde{q}_{1a0} + \tilde{q}_{0a0}\right)\left(\hat{q}_{1a0} + \hat{q}_{0a0}\right)}\right| \\ &= \left|\frac{\tilde{q}_{1a0}\hat{q}_{0a0} - \hat{q}_{1a0}\tilde{q}_{0a0}}{\left(\tilde{q}_{1a0} + \tilde{q}_{0a0}\right)\left(\hat{q}_{1a0} + \hat{q}_{0a0}\right)}\right| \\ &= \left|\frac{\hat{q}_{0a0}\left(\tilde{q}_{1a0} - \hat{q}_{1a0}\right) - \hat{q}_{1a0}\left(\tilde{q}_{0a0} - \hat{q}_{0a0}\right)}{\tilde{q}_{a0} \cdot \hat{q}_{a0}}\right| \\ &= \left|\frac{\hat{q}_{oa0}\left(\tilde{q}_{1a0} - \hat{q}_{1a0}\right) + \hat{q}_{1a0}\left(\hat{q}_{0a0} - \tilde{q}_{0a0}\right)}{\tilde{q}_{a0} \cdot \hat{q}_{a0}}\right| \\ &\leq \frac{\hat{q}_{0a0}\|\hat{q} - \tilde{q}\|_\infty + \hat{q}_{1a0}\|\hat{q} - \tilde{q}\|_\infty}{\tilde{q}_{ao} \cdot \hat{q}_{a0}} \\ &= \frac{\left(\hat{q}_{0a0} + \hat{q}_{1a0}\right)\|\hat{q} - \tilde{q}\|_\infty}{\tilde{q}_{a0} \cdot \hat{q}_{a0}} \\ &= \frac{\|\hat{q} - \tilde{q}\|_\infty}{\tilde{q}_{a0}} \\ &\leq \frac{2\ln\left(\frac{2|A|}{\beta}\right)}{\tilde{q}_{a0}m\epsilon} + \frac{2\ln\left(\frac{2}{\beta}\right)}{\tilde{q}_{a0}k\epsilon}\end{aligned}$$

similarly

$$\left|\tilde{TP}_a(\hat{Y}) - \hat{TP}_a(\hat{Y})\right| \leq \frac{2\ln\left(\frac{2|A|}{\beta}\right)}{\tilde{q}_{a1}m\epsilon} + \frac{2\ln\left(\frac{2}{\beta}\right)}{\tilde{q}_{a1}k\epsilon}$$

4. $\forall a_1 p$:

$$\begin{aligned}
&\left|\Delta \tilde{FP}_a\left(\hat{Y}_p\right)-\Delta \hat{FP}_a\left(\hat{Y}_p\right)\right| \\
&\leq\mid \tilde{FP}_a(\hat{Y})_{P_{1a}}+\left(1-\widetilde{FP}_a(\hat{Y})\right)_{p_{0a}}-\widetilde{FP}_0(\hat{Y})_{P_{10}}-\left(1-\widetilde{FP}_0(\hat{Y})\right)p_{00} \\
&-\hat{FP}_a(\hat{Y})p_{1a}+\left(1-\hat{FP}_a(\hat{Y})\right)p_{0a}-\hat{FP}_0(\hat{Y})p_{10}-\left(1-\widehat{FP}_0(\hat{Y})\right)p_{00}| \\
&=\mid \tilde{FP}_a(\hat{Y})_{p_{1a}}-\hat{FP}_a(\hat{Y})_{p_{1a}}-\tilde{FP}_a(\hat{Y})_{p_{0a}}+\hat{FP}_a(\hat{Y})_{p0a} \\
&-\widetilde{FP}_0(\hat{Y})_{P_{10}}+\hat{FP}_0(\hat{Y})_{p_{10}}+\tilde{FP}_0(\hat{Y})p_{00}-\hat{FP}_0(\hat{Y})_{P_{00}} \\
&=\left|(p_{1a}-p_{0a})\left(\widetilde{FP}_a(\hat{Y})-\hat{FP}_a(\hat{Y})\right)+(p_{00}-p_{10})\left(\tilde{FP}_0(\hat{Y})-\hat{FP}_0(\hat{Y})\right)\right| \\
&\leqslant\left|\tilde{FP}_a(\hat{Y})-\hat{FP}_a(\hat{Y})\right|\cdot\left|p_{1a}-p_{0a}\right|+\left|\tilde{FP}_0(\hat{Y})-\hat{FP}_0(\hat{Y})\right|\cdot\left|p_{10}-p_{00}\right|
\end{aligned}$$

we know that $|p_{1a}-p_{0a}|\in[0,1]$ for all a, so:

$$\begin{aligned}
&\left|\Delta \tilde{FP}_a\left(\hat{Y}_p\right)-\Delta \hat{FP}_a\left(\hat{Y}_p\right)\right|\leq\left|\tilde{FP}_a(\hat{Y})-\hat{FP}_a(\hat{Y})\right|+\left|\tilde{FP}_0(\hat{Y})-\hat{FP}_0(\hat{Y})\right| \\
&\quad=\left(\frac{2\ln\left(\frac{2|A|}{B}\right)}{\tilde{q}_{a0}m\epsilon}+\frac{2\ln\left(\frac{2}{\beta}\right)}{\tilde{q}_{a0}k\epsilon}\right)+\left(\frac{2\ln\left(\frac{2|A|}{B}\right)}{\tilde{q}_{00}m\epsilon}+\frac{2\ln\left(\frac{2}{\beta}\right)}{\tilde{q}_{00}k\epsilon}\right) \\
&\quad\leq\frac{4\ln\left(\frac{2|A|}{B}\right)}{\min\left\{\tilde{q}_{a0},\tilde{q}_{00}\right\}m\epsilon}+\frac{4\ln\left(\frac{2}{B}\right)}{\min\left\{\tilde{q}_{a0},\tilde{q}_{00}\right\}k\epsilon}
\end{aligned}$$

similarly:

$$\left|\Delta\widetilde{TP}_a(\hat{r})-\Delta\hat{TP}_a(\hat{y})\right|\leq\frac{4\ln\left(\frac{2|A|}{B}\right)}{\min\left\{\tilde{q}_{a1},\tilde{q}_{01}\right\}m\epsilon}+\frac{4\ln\left(\frac{2}{B}\right)}{\min\left\{\tilde{q}_{a1},\tilde{q}_{01}\right\}k\epsilon}$$

5. The first constraint of $\tilde{LP}$ is proven here:

$$\begin{aligned}
\left|\Delta\widetilde{\mathrm{FP}}_a\left(\widehat{Y}_{\hat{p}^\star}\right)\right|&=\left|\Delta\widetilde{\mathrm{FP}}_a\left(\widehat{Y}_{\hat{p}^\star}\right)-\Delta\widehat{\mathrm{FP}}_a\left(\widehat{Y}_{\hat{p}^\star}\right)+\Delta\widehat{\mathrm{FP}}_a\left(\widehat{Y}_{\hat{p}^\star}\right)\right| \\
&\leqslant\left|\Delta\widehat{\mathrm{FP}}_a\left(\hat{Y}_{\widehat{p}^\star}\right)\right|+\left|\Delta\widetilde{\mathrm{FP}}_a\left(\widehat{Y}_{\widehat{p}^\star}\right)-\Delta\widehat{\mathrm{FP}}_a\left(\widehat{Y}_{\widehat{p}^\star}\right)\right| \\
&\leqslant\gamma+\frac{4\ln(4|\mathcal{A}|/\beta)}{\min\left\{\tilde{q}_{a0},\tilde{q}_{00}\right\}m\epsilon}
\end{aligned}$$

by part 4 of this Lemma and the fact that $\left|\Delta\widehat{\mathrm{FP}}_a\left(\widehat{Y}_{\widehat{p}^*}\right)\right|\leqslant\gamma($ see $\widehat{\mathrm{LP}}$ in Appendix B).

From this, satisfying the second constraint of $\tilde{LP}$ can be similarly shown, and the third is trivial.

Following Lemma A.2, with probability at least $1 - \beta$:

$$\begin{aligned}
\widehat{\mathrm{err}}\left(\widehat{Y}_{\widetilde{p}^\star}\right) &\leqslant \widetilde{\mathrm{err}}\left(\widehat{Y}_{\widetilde{p}^\star}\right) + \frac{12|A|\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{12|A|\ln\left(\frac{2}{\beta}\right)}{k\epsilon} \\
&\leqslant \widetilde{\mathrm{err}}\left(\widehat{Y}_{\widehat{p}^\star}\right) + \frac{12|A|\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{12|A|\ln\left(\frac{2}{\beta}\right)}{k\epsilon} \quad \text{(part 1 of Lemma A.2)} \\
&\leqslant \widehat{\mathrm{err}}\left(\widehat{Y}_{\widehat{p}^\star}\right) + \frac{24|A|\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{24|A|\ln\left(\frac{2}{\beta}\right)}{k\epsilon}
\end{aligned}$$

(part 5 of Lemma A.2)

Also, for all $a \neq 0$,

$$\begin{aligned}
\Delta\widehat{\mathrm{FP}}_a\left(\widehat{Y}_{\widetilde{p}^\star}\right) &\leqslant \Delta\widetilde{\mathrm{FP}}_a\left(\widehat{Y}_{\widetilde{p}^\star}\right) + \frac{4\ln\left(\frac{2|A|}{B}\right)}{\min\left\{\tilde{q}_{a0}, \tilde{q}_{00}\right\} m\epsilon} + \frac{4\ln\left(\frac{2}{B}\right)}{\min\left\{\tilde{q}_{a0}, \tilde{q}_{00}\right\} k\epsilon} \quad \text{(part 4 of Lemma A.2)} \\
&\leqslant \gamma + \frac{8\ln\left(\frac{2|A|}{B}\right)}{\min\left\{\tilde{q}_{a0}, \tilde{q}_{00}\right\} m\epsilon - \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} - \frac{4\ln\left(\frac{2}{\beta}\right)}{k\epsilon}} + \frac{8\ln\left(\frac{2}{B}\right)}{\min\left\{\tilde{q}_{a0}, \tilde{q}_{00}\right\} k\epsilon - \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} - \frac{4\ln\left(\frac{2}{\beta}\right)}{k\epsilon}}
\end{aligned}$$

The last inequality follows from the fact that $|\tilde{q}_{ay} - \hat{q}_{ay}| \leqslant \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} + \frac{4\ln\left(\frac{2}{\beta}\right)}{k\epsilon}$ for all a, y. It follows similarly that,

$$\Delta\widehat{\mathrm{TP}}_a\left(\widehat{Y}_{\widehat{p}^\star}\right) \leqslant \gamma + \frac{8\ln(4|\mathcal{A}|/\beta)}{\min\left\{\hat{q}_{a1}, \hat{q}_{01}\right\} m\epsilon - \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} - \frac{4\ln\left(\frac{2}{\beta}\right)}{k\epsilon}} + \frac{8\ln\left(\frac{2}{B}\right)}{\min\left\{\tilde{q}_{a1}, \tilde{q}_{01}\right\} k\epsilon - \frac{4\ln\left(\frac{2|A|}{\beta}\right)}{m\epsilon} - \frac{4\ln\left(\frac{2}{\beta}\right)}{k\epsilon}}$$

Appendix B

$\widehat{\mathrm{LP}}$: **Empirical Linear Program:**

$$\begin{aligned}
\underset{p}{\arg\min} \quad & \widehat{\mathrm{err}}\left(\widehat{Y}_p\right) \\
\text{s.t. } \forall a \in \dashv\mathcal{A}\mathcal{A} \quad & \Delta\widehat{\mathrm{FP}}_a\left(\widehat{Y}_p\right) \leqslant \gamma \\
& \Delta\widehat{\mathrm{TP}}_a\left(\widehat{Y}_p\right) \leqslant \gamma \\
& 0 \leqslant p_{\hat{y}a} \leqslant 1 \quad \forall \hat{y}, a
\end{aligned}$$

$$\widehat{\mathrm{err}}\left(\widehat{Y}_p\right) = \sum_{\hat{y},a}\left(\hat{q}_{\hat{y}a0} - \hat{q}_{\hat{y}a1}\right) \cdot p_{\hat{y}a} + \sum_{\hat{y},a} \hat{q}_{\hat{y}a1}$$

$$\Delta\widehat{\mathrm{FP}}_a\left(\widehat{Y}_p\right) = \left|\widehat{\mathrm{FP}}_a(\hat{Y}) \cdot p_{1a} + \left(1 - \widehat{\mathrm{FP}}_a(\hat{Y})\right) \cdot p_{0a} - \widehat{\mathrm{FP}}_0(\hat{Y}) \cdot p_{10} - \left(1 - \widehat{\mathrm{FP}}_0(\hat{Y})\right) \cdot p_{00}\right|$$

$$\Delta\widehat{\mathrm{TP}}_a\left(\widehat{Y}_p\right) = \left|\widehat{\mathrm{TP}}_a(\widehat{Y}) \cdot p_{1a} + \left(1 - \widehat{\mathrm{TP}}_1(\widehat{Y})\right) \cdot p_{0a} - \widehat{\mathrm{TP}}_0(\widehat{Y}) \cdot p_{10} - \left(1 - \widehat{\mathrm{TP}}_0(\widehat{Y})\right) \cdot p_{00}\right|$$

References

[Cor95] Corinna Cortes. “Support-Vector Networks”. In: *Machine Learning* (1995).

[Dwo+06] Cynthia Dwork et al. “Calibrating noise to sensitivity in private data analysis”. In: *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer. 2006, pp. 265–284.

[MT07] Frank McSherry and Kunal Talwar. “Mechanism design via differential privacy”. In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07)*. IEEE. 2007, pp. 94–103.

[Wan+10] Ping Wang et al. “Predicting Criminal Recidivism with Support Vector Machine”. In: *2010 International Conference on Management and Service Science*. 2010, pp. 1–9. DOI: 10.1109/ICMSS.2010.5575352.

[DR+14] Cynthia Dwork, Aaron Roth, et al. “The algorithmic foundations of differential privacy”. In: *Foundations and Trends® in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[Fel+15] Michael Feldman et al. “Certifying and removing disparate impact”. In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 259–268.

[FKL16] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. “A confidence-based approach for balancing fairness and accuracy”. In: *Proceedings of the 2016 SIAM international conference on data mining*. SIAM. 2016, pp. 144–152.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro. “Equality of opportunity in supervised learning”. In: *Advances in neural information processing systems* 29 (2016).

[Aga+18] Alekh Agarwal et al. “A reductions approach to fair classification”. In: *International conference on machine learning*. PMLR. 2018, pp. 60–69.

[Cum+19] Rachel Cummings et al. “On the compatibility of privacy and fairness”. In: *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*. 2019, pp. 309–315.

[Jag+19] Matthew Jagielski et al. “Differentially private fair learning”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3000–3008.

[Zaf+19] Muhammad Bilal Zafar et al. “Fairness constraints: A flexible approach for fair classification”. In: *Journal of Machine Learning Research* 20.75 (2019), pp. 1–42.

[XDW20] Depeng Xu, Wei Du, and Xintao Wu. “Removing disparate impact of differentially private stochastic gradient descent on model accuracy”. In: *arXiv preprint arXiv:2003.03699* (2020).

[BHN23] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities.* MIT press, 2023.

[Far+23] Michael Mayowa Farayola et al. "Fairness of AI in Predicting the Risk of Recidivism: Review and Phase Mapping of AI Fairness Techniques". In: *Proceedings of the 18th International Conference on Availability, Reliability and Security.* 2023, pp. 1–10.

[Rua+23] Wenqiang Ruan et al. "Towards Understanding the fairness of differentially private margin classifiers". In: *World Wide Web* 26.3 (2023), pp. 1201–1221.