Fairness Through Independence via Cramér-von Mises Regularization

Albert Gimó
Criteo AI Lab
a.gimo@criteo.com

Mariia Vladimirova
Criteo AI Lab, Fairplay joint team
m.vladimirova@criteo.com

Olga Petrova
Criteo AI Lab
o.petrova@criteo.com

Federico Pavone*
Theremia
federicopavone@theremia.health

Reda Chhaibi
Université Côte d'Azur, CNRS, LJAD
reda.chhaibi@univ-cotedazur.fr

Abstract

Controlling fairness in machine learning (ML) model outputs is challenging due to complex, unstable and computationally expensive techniques for bias estimation on finite data samples. We propose a simple in-processing method to control group fairness during training by penalizing statistical dependence between model outputs $\hat{Y}$ and a sensitive attribute S. Our approach instantiates the Cramér–von Mises (CvM) dependence coefficient $\xi(S, \hat{Y})$ as a bounded, differentiable regularizer that integrates seamlessly with stochastic optimization. The resulting objective $L + \lambda\, \xi(S, \hat{Y})$ positions models along a fairness–utility Pareto frontier through a single multiplier λ. Our experiments demonstrate the effectiveness of this method for controlling the fairness-utility trade-off in both small fairness-aware and large tabular datasets. In order to control the compromise between fairness metrics and utility metrics, we propose a task-agnostic hyperparameter tuning pipeline and showcase its effectiveness in a large tabular dataset. In practice, we have observed that controlling for CvM leads to lower demographic-parity (DP) scores, providing a tractable and computationally efficient methodology, bridging the gap between policy requirements on DP and a scalable training procedure for ML models.

1 Introduction

ML models often inherit biases from historical data (through selection effects, under-representation, or label bias) yielding unreliable outcomes for protected groups [5]. High-profile failures in hiring and criminal justice illustrate how models can encode disparities even without explicit access to sensitive attributes S [4, 12]. This reality has led to growing legal and regulatory pressure for model deployers to demonstrate and control fairness, as seen in the EU AI Act and New York City's bias-audit duties. This makes the ability to control fairness, i.e achieving a target level with minimal utility loss, a critical objective for modern ML.

Controlling fairness under biased data is challenging for two reasons. First, population-level fairness objectives, such as disparities in error rates, are often hard to estimate from finite, biased samples, which can lead to wrongly estimated risks. Second, many existing fairness-inducing methods rely on computationally intensive techniques like constrained optimization with expensive projections [2] or adversarial training [21]. These approaches do not scale to the massive datasets and models now common in modern training regimes, e.g. in vision, language, and recommendation systems. A

*The work was done during a fellowship at Université Paris Dauphine-PSL.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Reliable ML.

practical solution must therefore (i) account for data bias when estimating fairness-relevant quantities and (ii) integrate with standard stochastic training so it scales with dataset and model size.

Our approach. We introduce a simple in-processing method that augments standard training losses L with a bounded and differentiable regularizer based on the CvM dependence coefficient $\xi(S, \hat{Y})$, which measures how conditioning on sensitive attribute S shifts the distribution of predictions $\hat{Y}$ [10, 13]. To make ξ trainable, we leverage differentiable ranking [7] to backpropagate through the rank-based estimator introduced in Chatterjee [10], enabling end-to-end optimization of a regularized objective $L + \lambda\xi(S, \hat{Y})$. We study how λ positions models along this trade-off and maintain reliable levels of fairness without excessively compromising the model's utility. We provide an analysis on how to perform this adjustment and report results in both standard fairness (*Adults* [1]) and a larger non-fairness specific (*Weather Forecasting* [24]) datasets. Our **contributions** are three-fold:

1. A CvM-based regularizer that promotes independence between $\hat{Y}$ and S on biased datasets, improving the reliability of predictions for protected groups.
2. A differentiable implementation based on soft-ranking with clear stability/complexity properties, exposing a single trade-off parameter λ and a smoothness control ε.
3. A scalable training and tuning protocol on small and large tabular workloads that enables practitioners to adjust the fairness-utility trade-offs when training models.

2 Method

While many independence measures exist (see discussion in Appendix A.2), they often suffer from estimation difficulty, gradient instability, or interpretability issues. We introduce and study a CvM–based regularizer that measures and minimizes statistical dependence between model outputs $\hat{Y}$ and sensitive attributes S, see Appendix B for more details. The CvM dependence coefficient provides a normalized, interpretable scalar in $[0, 1]$ that equals zero iff independence holds and one iff the target is a measurable function of the selected sensitive attribute. It aggregates the variance of conditional expectations of thresholded outcomes, thereby capturing non-linear dependencies without hand-enumerating slices or thresholds. We adopt a finite-sample estimator for the coefficient that is $O(n \log n)$ via sorting and ranking, and we show how to embed it directly in modern optimizers. The CvM coefficient and its estimator are the following:

$$\xi(X, Y) := \frac{\int \mathrm{Var}\left(\mathbb{E}[\mathbb{1}_{\{Y \geq t\}} \mid X]\right)\, dF_Y(t)}{\int \mathrm{Var}\left(\mathbb{1}_{\{Y \geq t\}}\right)\, dF_Y(t)} \quad \text{and} \quad \xi_n(X_n, Y_n) := 1 - \frac{n \sum_{k=1}^{n-1} \left| r_{i_{k+1}} - r_{i_k} \right|}{2 \sum_{k=1}^{n} l_k (n - l_k)} \,. \tag{1}$$

The technical challenge in using this estimation for training deep learning models is that ranking is a discrete operation, making it infeasible for gradient–based optimization. We overcome this by leveraging fast and differentiable soft ranking as projections onto the permutahedron from Blondel et al. [7], yielding order-preserving almost-everywhere differentiable operators with exact Jacobians via isotonic optimization. Plugging these operators into the estimator produces a differentiable CvM penalty term that integrates seamlessly with deep learning frameworks and preserves the statistical relevance of the original coefficient.

Crucially for scalability, the proposed objective is minibatch-friendly. The estimator's behavior is especially well-conditioned when the model outputs are continuous (e.g., regression or probabilities in classification), which we recommend in practice. Under continuity, the sample–based coefficient is stable to small perturbations (as described in Proposition 1), improving optimization under SGD noise and batch shuffling.

Proposition 1 (Robustness to perturbations)**.** *Let (X_n, Y_n) be n i.i.d. samples from $p(X, Y)$. Let Y be continuous and let Z_n^1, Z_n^2 contain n i.i.d. samples from a continuous real-valued noise variable. Define $X^\eta := X + \eta Z^1$ and $Y^\eta := Y + \eta Z^2$. Then, with probability* 1*,*

$$\begin{aligned} \lim_{\eta \to 0} \xi_n(X, Y^\eta) &= \xi_n(X, Y) \,, \\ \lim_{\eta \to 0} \mathbb{E}\big[\xi_n(X^\eta, Y)\big] \;&=\; \lim_{\eta \to 0} \mathbb{E}\big[\xi_n(X^\eta, Y^\eta)\big] = \mathbb{E}\big[\xi_n(X, Y)\big] \,, \end{aligned} \tag{2}$$

where the expectations are with respect to the perturbation noise and any uniformly random tie-breaking mechanism for the right-most term.

The method integrates into existing training procedures as a single regularization term, with two practical hyperparameters: (i) the soft-ranking smoothness ε (controls bias–variance of the gradient signal) and (ii) the multiplier λ (controls the utility–fairness trade-off). In the experiments section, we provide usage guidance based on our practical results: prefer L_2 regularization over L_1 on the CvM penalty to increase robustness; when needed, fine-tune from an unregularized checkpoint, while geometrically increasing λ to trace a stable Pareto frontier; and we provide a 3 stage method for hyperparameter optimization. These practices aim at preserving accuracy while steadily reducing dependence on sensitive attributes.

3 Experiments

We focus on two datasets: (i) `Adult` [1], a canonical fairness benchmark to sanity-check group metrics. We also stress-test by treating education as a sensitive attribute (highly correlated with income) to probe the fairness–utility frontier; (ii) `Weather` (TabReD, Rubachev et al. [24]), a large tabular regression benchmark with deep-learning baselines chosen to test our method's scalability . For more details on the dataset choice and their limitations, we refer to Appendix A.1.

3.1 Adult dataset

Setup. We consider binary income prediction with sensitive attributes comprising (i) a weakly correlated attribute (gender) and (ii) a strongly correlated attribute (education). Utility is measured via accuracy/F1-score; fairness via DP/EO and CvM. A detailed per-attribute analysis, pre-processing choices and discussion are referred to Appendix E.

Results. We (i) demonstrate how adding our CvM-based regularizer translates to decreases in group-based fairness metrics such as DP, and (ii) study the role of the penalty form and fine-tuning.

- *Multiplier control.* Increasing λ monotonically reduces CvM and typically shrinks DP/EO gaps. For weakly correlated attributes, these improvements incur modest utility loss; for highly correlated attributes, utility drops are sharper, consistent with a steeper trade-off frontier, see Figure 1.
- *Penalty form.* Applying an L_2 penalty on the CvM term (Figure 1) yields reduced sensitivity to small coefficient changes and increased robustness when compared to L_1 regularization (Figure 9). We adopt the L_2 formulation henceforth to improve the control over the adjusting of λ. We refer to Appendix E for more details.
- *Fine-tuning.* The presented regularization can also be introduced as a fine-tuning method. Experimentally, introducing λ post-hoc and ramping it geometrically stabilizes training and yields gradual fairness gains with limited utility degradation.

3.2 Weather forecasting dataset

Setup. We study large-scale temperature prediction using the *Weather Forecasting* dataset processed per Rubachev et al. [24]. Utility is tracked by (Neg)MSE; fairness by the CvM coefficient (and, where relevant, group-based summaries). Extended discussion is referred to Appendix F.

Results. We (i) visualize the effect of λ on the CvM–utility frontier, and (ii) study the role of the *smoothness controller* ε in the soft-ranking operator (iii) demonstrate how adding our CvM-based regularizer translates to decreases in group-based fairness metrics such as DP.

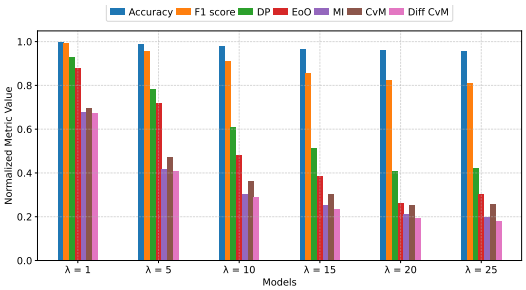


Figure 1: Adult dataset: Plot of utility (accuracy and F1) and fairness (DP, EoO, CvM, MI, Differentiable CvM) metrics. Values of *fair* MLPs ($\lambda \in [1, 5, 10, 15, 20, 25]$, with L_2 penalty) normalized by the values of the *unfair* MLP ($\lambda = 0$).

- λ, ε*-tuning.* CvM decreases predictably as λ increases, exposing a Pareto-like frontier against (Neg)MSE (see Figure 2 and 11). Within practical ranges, ε exhibits negligible impact on both utility and CvM in this setting (see discussion in Appendix F).
- *Impact of λ on DP:* We observe a clear relation between decreasing CvM and lower DP. Since often regulatory policies for fairness in ML focuses on DP, this provides a strategy for determining an appropriate value of λ via the reasoning chain "regulations $\rightarrow$ DP $\rightarrow$ $\xi_n \rightarrow \lambda$" , bridging the gap between regulation and the training of deep models.

3.3 Hyperparameter strategy

Based on our experiments on the weather forecasting dataset, we propose a scalable and dataset-agnostic hyperparameter tuning pipeline consisting of 3 steps:

Step 1 (utility-only baseline). Tune non-fairness hyperparameters with $\lambda = 0$ (architecture, optimizer, regularization,...) to confirm task learnability and provide an initial performance baseline.

Step 2 (fairness-specific tuning). Fix the baseline hyperparameters that maximize utility, then sweep the CvM multiplier λ and the smoothness controller ε via randomized search over wide ranges. This provides an initial reference point that enables to shrink down to the regions of λ and ϵ that are most promising to perform hyperparameter search in step 3.

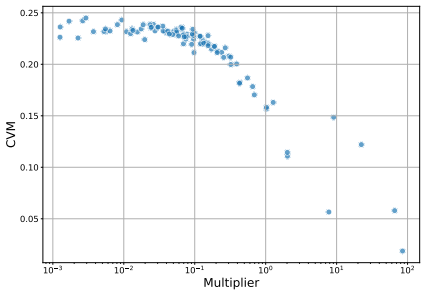


Figure 2: Weather dataset: CvM $\xi(S, \hat{Y})$ vs. regularization multiplier λ. Increasing λ reduces $\xi(S, \hat{Y})$.

Step 3 (penalized-utility selection). Guide the exploration of hyperparameter space based on a fairness-penalized utility $U(l, c, ; \gamma)$. In our experiments we define it as:

$$U(\ell, c; \gamma) = \begin{cases} \ell, & c \leq \gamma, \\ \ell - \alpha(c - \gamma), & c > \gamma, \end{cases}$$

where ℓ is the utility (to maximize), c is the CvM (to minimize), γ is a cutoff, and α is a penalty slope. This methodology applies directly to other large tabular datasets beyond weather, and it concentrates the search on desirable regions of the Pareto frontier. The value of the cutoff γ should be determined to accomplish the desired levels of group fairness as discussed in F.2

3.4 Discussion

With our experiments we (i) demonstrate controllability of the trade-off via the multiplier λ, (ii) assess robustness and scalability of the differentiable CvM term on large tabular data, (iii) provide a minibatch-friendly tuning methodology suitable for modern optimizers, and (iv) show strong correlation between DP and CvM providing a reference to determine λ, bridging a gap between DP-based AI regulations and in-training practices.

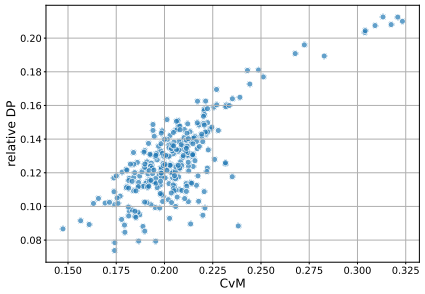


Figure 3: Weather dataset (`sun_elevation` binned): relationship between CvM and the relative DP gap across λ; lower CvM aligns with smaller DP disparities.

4 Conclusion and future work

This work introduces a CvM-based regularizer that makes fairness controllable both during training and as a model fine-tuning method. The CvM term is computed via a finite-sample, rank-based estimator made trainable by replacing non-differentiable ranks with a smooth, order-preserving soft ranking which yields stable gradients that backpropagate efficiently in modern ML training regimes. The approach exposes a single training-time control λ for positioning models along the fairness–utility frontier. We also observe a consistent alignment between CvM reductions and reduction in demographic-parity gaps, providing a direct connection between DP-focused policy and training-time decisions via the CvM regularization. Overall, the method offers a lightweight, scalable mechanism to control fairness within modern ML training regimes offering

a practical path for deploying models that are both accurate and equitable even when the available data is imperfect. Future work will extend experiments in the same datasets and to fairness-specific large-scale datasets as well as develop stronger experimental and theoretical connections between CvM and established fairness metrics such as DP and EO.

5 Acknowledgments

Albert Gimò received support from "La Caixa" Foundation (ID 100010434) fellowship No LCF/BQ/EU24/12060099.

Federico Pavone received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101034255.

References

[1] Machine learning repository. https://archive.ics.uci.edu/. Accessed: 2025-08-22.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *ACM SIGKDD international conference on knowledge discovery & data mining*, 2019.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. 2019.

[6] Sarah Bird, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vladimir Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

[7] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking, 2020. URL `https://arxiv.org/abs/2002.08871`.

[8] Philip Buczak, Andreas Groll, , et al. Cramér-von mises criterion values for different distribution classes with individual additive shifts c in the classification case. *ResearchGate*, 2024.

[9] Flavio Calmon, Dingshi Wei, , et al. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, 2017.

[10] Sourav Chatterjee. A new coefficient of correlation. *The Annals of Statistics*, 2021. arXiv:1909.10140.

[11] Harald Cramér. On the composition of elementary errors. *Skandinavisk Aktuarietidskrift*, 11(1): 13–74, 1928.

[12] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018.

[13] Holger Dette, Karl Siburg, and Pavel Stoimenov. A copula-based nonparametric measure of regression dependence. 02 2013.

[14] Fabrice Gamboa, Thierry Klein, and Agnès Lagnoux. Sensitivity analysis based on cramér von mises distance, 2017. URL `https://arxiv.org/abs/1506.04133`.

[15] Yaroslav Ganin, Evgeniya Ustinova, , et al. Domain-adversarial training of neural networks. In *Journal of Machine Learning Research*, volume 17, pages 1–35, 2016.

[16] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[17] Arthur Gretton, , et al. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1): 723–773, 2012.

[18] Toshihiro Kamishima, Shotaro Akaho, , et al. Fairness-aware learning through regularization. *Proceedings of the 2012 IEEE International Conference on Data Mining*, 2012.

[19] Toshihiro Kamishima, Shotaro Akaho, , et al. Fairness-aware learning: a survey. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 17(1):1–18, 2013.

[20] Alexandru Lopotenco, Ian Tong Pan, Jack Zhang, and Guan Xiong Qiao. Fair representation learning with maximum mean discrepancy distance constraint (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23567–23568, Mar. 2024. doi: 10.1609/aaai.v38i21.30476. URL `https://ojs.aaai.org/index.php/AAAI/article/view/30476`.

[21] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 2018.

[22] Anouar Mary, , et al. A maximal correlation framework for fair machine learning. In *International Conference on Machine Learning*, 2019.

[23] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

[24] Ivan Rubachev, Nikolay Kartashev, Yury Gorishniy, and Artem Babenko. TabReD: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks, 2024.

[25] Le Song, , et al. Feature selection with the hilbert-schmidt independence criterion. In *Proceedings of the 2012 International Conference on Machine Learning*, 2012.

[26] Ryan Turner, David Eriksson, Michael McCourt, Juha Kiili, Eero Laaksonen, Zhen Xu, and Isabelle Guyon. Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In *NeurIPS 2020 Competition and Demonstration Track*, 2021.

[27] Mariia Vladimirova, Federico Pavone, and Eustache Diemert. Fairjob: A real-world dataset for fairness in online systems, 2024. URL `https://arxiv.org/abs/2407.03059`.

[28] Richard von Mises. Theorie der scharen von unabhängigen systemen. *Mathematische Zeitschrift*, 32(1):678–700, 1930.

[29] Jason D Williams, , et al. Evaluating user simulations with the cramér-von mises divergence. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[30] Yuyang Xu, Ziqi Zhang, , et al. In-processing modeling techniques for machine learning fairness: A survey. In *ACM Transactions on Knowledge Discovery from Data*, 2022.

[31] Muhammad Bilal Zafar, , et al. Fairness constraints: a practical approach for fair classification. In *International Conference on Machine Learning*, 2017.

[32] Muhammad Bilal Zafar, , et al. Fairness without spurious correlations. In *International Conference on Machine Learning*, 2019.

[33] Brian Hu Zhang, , et al. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

A Related works

A.1 Datasets

Fairness evaluation and method design have over-relied on tiny, aging benchmarks (e.g., Adult, COMPAS). We explicitly lean into larger tabular settings: alongside standard Adult experiments, we scale to a processed Weather benchmark from TabReD [24] to emulate real training loads and hyperparameter search, precisely because fairness-aware, open, large tabular datasets are scarce. Using this non-fairness-specific but sizable corpus stresses training throughput and stability in ways small datasets cannot.

FairJob [27] targets fairness in online systems and offers a valuable, real-world dataset and highlights the lack of open-source large fairness-aware datasets. Our present study focuses on supervised tabular tasks with standard deep-learning training loops and widely adopted metrics/APIs (Fairlearn [6]). Aligning protocols and baselines across online/interactive settings is non-trivial and would require additional engineering (ranking/replay, exposure bias control) beyond scope for the workshop draft; we therefore leave a Fairjob-style recsys evaluation to future work.

A.2 Methods

Fairness in ML is a rapidly evolving field, with mitigation strategies broadly categorized into pre-processing, in-processing, and post-processing methods. This work focuses on the in-processing paradigm, where the fairness objective is integrated directly into the model training process. Our approach, which uses the CvM statistic for regularization, sits at the intersection of two key research areas: dependence-based fairness methods and the use of integral probability metrics in ML.

A.2.1 In-processing fairness methods

In-processing methods modify the learning algorithm to enforce fairness constraints or penalties during training [23, 30]. One of the most prominent approaches is adversarial debiasing, where a model's representation is trained to be predictive of the target label while simultaneously being unable to predict the sensitive attribute. This is often achieved by training an adversary network to predict the sensitive attribute from the model's latent representation [15, 33]. Another common approach is constrained optimization, which formulates the fairness objective as a constraint on the model's predictions. These methods often use convex optimization techniques to satisfy fairness criteria like demographic parity or equalized odds [31, 32].

Our method differs from these approaches by framing fairness as a direct independence objective between the model's output and the sensitive attribute, and achieving this through a novel regularizer derived from a statistical test, rather than an adversarial game or a hard constraint.

A.2.2 Dependence-based fairness regularization

A large body of work has sought to achieve fairness by minimizing the statistical dependence between the model's predictions $\hat{Y}$ and the sensitive attribute S. This is often achieved by adding a regularization term to the standard loss function with a penalty as a measure of dependence. Classic dependence measures used for this purpose include mutual information (MI), which quantifies the information shared between two variables [18, 19]. Other work has explored maximal correlation [22] and measures of covariance [9].

More recently, research has leveraged kernel-based independence measures, which can capture non-linear dependencies. The Hilbert-Schmidt Independence Criterion (HSIC) and Maximum Mean Discrepancy (MMD) are two prominent examples. HSIC is a powerful non-parametric measure of dependence that has been widely used in ML for feature selection and independent component analysis [16]. In the context of fairness, it can be used to regularize a model to make its representations independent of the sensitive attribute [25]. MMD is an integral probability metric that measures the distance between two probability distributions and has also been applied to fairness, particularly in fair representation learning [17, 20].

Our work contributes to this line of research by proposing a novel dependence regularizer based on the CvM statistic, a classic goodness-of-fit test. While similar in spirit to MMD as an integral

probability metric, the CvM statistic has distinct properties and provides a new perspective on measuring distributional discrepancy for fairness applications.

A.2.3 The CvM statistic in ML

The CvM statistic is a well-established tool in classical statistics used to test the goodness-of-fit of a sample's empirical distribution to a given reference distribution [11, 28]. Its use in ML has been more limited but has appeared in contexts such as evaluating user simulations in dialogue systems [29] or as a general-purpose distance for hyperparameter tuning [8].

To the best of our knowledge, the application of the CvM statistic as a direct regularizer for achieving independence-based fairness is a novel contribution. Unlike HSIC and MMD which are based on kernel inner products, the CvM statistic directly compares the cumulative distribution functions (CDFs) of the model outputs across different sensitive groups. This provides a different theoretical foundation and may offer computational or statistical advantages in certain settings.

B The CvM dependence coefficient

The CvM coefficient has appeared numerous times in the literature, including [13, 14]. In this appendix we provide an explanation of the CvM dependence coefficient for the purposes of motivating its use for dependence measuring and fairness.

B.1 Derivation

Assumptions. We assume Y is continuous and that measurability/integrability conditions hold, so that changes of integration order (Fubini/Tonelli) are valid. All distribution functions are right-continuous and non-decreasing.

The coefficient takes inspiration of the CvM distance. Given CDFs F and G on $\mathbb{R}$, the CvM distance is defined as

$$d^2_{\mathrm{CvM}}(F,G) \;=\; \int_{\mathbb{R}} \big(F(t)-G(t)\big)^2 \, dG(t). \tag{3}$$

Measuring the discrepancy between the conditional and marginal laws of Y leads to the dependence functional

$$\xi(X,Y) \;:=\; \int_{\mathcal{X}} \int_{\mathbb{R}} \big(F_{Y|X}(t \mid x) - F_Y(t)\big)^2 \, dF_Y(t)\, dF_X(x). \tag{4}$$

which equals zero iff $F_{Y|X}(\cdot \mid x) = F_Y(\cdot)$ F_X-a.s., i.e., when X and Y are independent [13].

This coefficient, which we refer to as the CvM coefficient, allows for a variance-based formulation. Let $p_t(X) := \mathbb{E}[\mathbb{1}_{\{Y \geq t\}} \mid X]$. For continuous Y, $p_t(X) = 1 - F_{Y|X}(t \mid X)$ and $\mathbb{E}[p_t(X)] = \mathbb{P}(Y \geq t) = 1 - F_Y(t)$. Expanding the square in (4) and using $\mathbb{E}[F_{Y|X}(t \mid X)] = F_Y(t)$ yields

$$\xi(X,Y) \;=\; \int_{\mathbb{R}} \mathrm{Var}\big(p_t(X)\big)\, dF_Y(t) \;=\; \int_{\mathbb{R}} \mathrm{Var}\Big(\mathbb{E}[\mathbb{1}_{\{Y \geq t\}} \mid X]\Big)\, dF_Y(t). \tag{5}$$

Normalizing by the unconditional variability of the threshold indicators we redefine the coefficeint

$$\xi(X,Y) \;:=\; \frac{\displaystyle\int_{\mathbb{R}} \mathrm{Var}\Big(\mathbb{E}[\mathbb{1}_{\{Y \geq t\}} \mid X]\Big)\, dF_Y(t)}{\displaystyle\int_{\mathbb{R}} \mathrm{Var}\big(\mathbb{1}_{\{Y \geq t\}}\big)\, dF_Y(t)}, \tag{6}$$

which is a form more commonly used in the literature. By the law of total variance applied to $\mathbb{1}_{\{Y \geq t\}}$, the numerator is bounded above by the denominator for every t, hence $0 \leq \xi(X,Y) \leq 1$.

If Y is continuous, then with $u = F_Y(t)$ and $U := F_Y(Y) \sim \mathrm{Unif}(0,1)$,

$$\int_{\mathbb{R}} \mathrm{Var}\big(\mathbb{1}_{\{Y \geq t\}}\big)\, dF_Y(t) = \int_{\mathbb{R}} \big(1 - F_Y(t)\big) F_Y(t)\, dF_Y(t) = \mathbb{E}\big[U(1-U)\big] = \frac{1}{6}. \tag{7}$$

Thus, for continuous Y, the normalizing term in (6) is a constant $1/6$.

B.2 Interpreting the CvM Coefficient

In this section we aim to provide an intuitive explanation of what is being measured by the CvM coefficient. We provide intuitions both from a variance explanation point of view and from a perspective of measuring differences between distributions. We maintain the continuity assumptions from the previous subsection: Y is continuous, measurability/integrability conditions hold, and changes of integration order are valid.

Consider the normalized dependence coefficient defined in (6). Fix a threshold $t \in \mathbb{R}$ and define the Bernoulli variable $\mathbb{1}_{\{Y \geq t\}}$ with mean $p_Y(t) := \mathbb{P}(Y \geq t)$, and its conditional counterpart $p_{Y|x}(t) := \mathbb{P}(Y \geq t \mid X = x)$. When X is informative about Y, the function $x \mapsto p_{Y|x}(t)$ varies across x; when X and Y are nearly independent, this variation is small. The numerator in (6) aggregates this signal as

$$\int \mathrm{Var}\big(\mathbb{E}[\mathbb{1}_{\{Y \geq t\}} \mid X]\big)\, dF_Y(t) \;=\; \int \mathrm{Var}\big(p_{Y|X}(t)\big)\, dF_Y(t),$$

so larger values indicate that X explains more of the thresholded behavior of Y across many t.

Illustrative example

Consider the simple function $Y = \sin(\pi X)$ and an added noise component $\mathcal{N}(0, \sigma^2)$ for different values of σ as shown in 4. Fix $t = 0$ for illustration purposes.

Then $x \mapsto \mathbb{P}(Y \geq 0 \mid X = x)$ is close to $\{0, 1\}$ for most x when σ is small, becomes less extreme but still variable for moderate σ, and is almost constant at $1/2$ for large σ. This differences mean that $\mathrm{Var}(p_{Y|X}(t))$ decreases with σ, lowering the CvM numerator. As discussed, the denominator stays constant at $\frac{1}{6}$ for all values of σ. The heatmap over t in Figure 5 shows the same pattern persists across thresholds; integrating over t therefore preserves this effect, leading to lower noise levels being associated to higher CvM values.

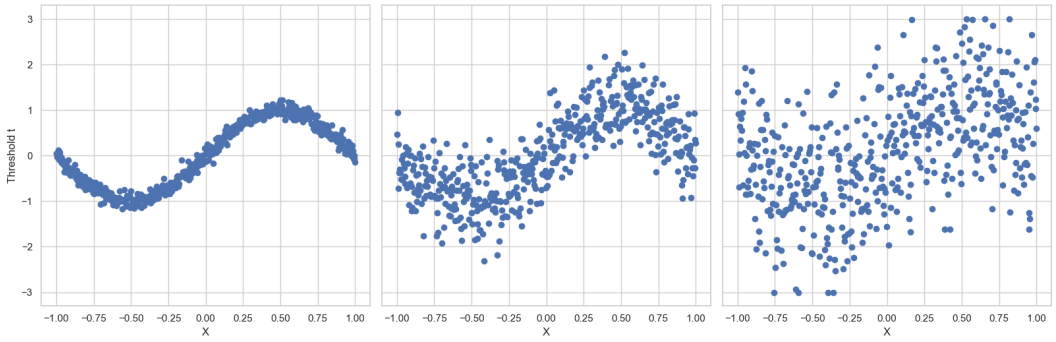


Figure 4: Function $Y = \sin(\pi x) + \mathcal{N}(0, \sigma^2)$ for different σ values.

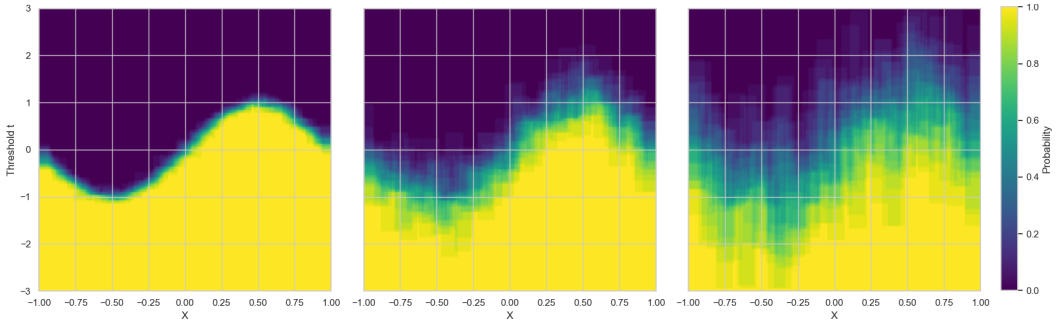


Figure 5: Heatmap encoding the values of $\mathbb{P}(Y \geq t \mid X = x)$ for different cutoffs t (in the Y axes) and different noise levels.

Expanding on this variance-based explanation, by the law of total variance applied to $\mathbb{1}_{\{Y \geq t\}}$, the numerator in (6) measures the component of variance *explained* by X, while the denominator captures

the *total* variance of the threshold indicators. Thus, the coefficient admits the interpretation

$$\xi(X,Y) \;=\; \frac{\int_{\mathbb{R}} \text{variability in } \mathbb{1}_{\{Y\geq t\}} \text{ explained by } X}{\int_{\mathbb{R}} \text{total variability of } \mathbb{1}_{\{Y\geq t\}}}\,.$$

The coefficient allows an alternative interpretation. Going back to (4), one can view the numerator as the average (over x) squared difference between the conditional CDF $F_{Y|X}(\cdot \mid x)$ and the marginal $F_Y(\cdot)$, integrated over the different values of X. This viewpoint emphasizes that $\xi(X,Y)$ is large when conditioning on X substantially deforms the distribution of Y, and small when $F_{Y|X}$ remains close to F_Y for most x.

B.3 Connection to group-fairness metrics

Many group-fairness metrics allow an independence based interpretation. The expectation being that enforcing some type of (conditional) independence between the sensitive attribute and the model's predictions will prevent the sensitive attribute from having a disproportionate effect on the outputs of the model.

Let S be a sensitive attribute and $\hat{Y}$ a (continuous) model output. The independence target of demographic parity (DP), $\hat{Y} \perp S$, is satisfied when

$$\xi(S,\hat{Y}) \;:=\; \frac{\displaystyle\int \mathrm{Var}\Big(\mathbb{E}[\mathbb{1}_{\{\hat{Y}\geq t\}} \mid S]\Big)\, dF_{\hat{Y}}(t)}{\displaystyle\int \mathrm{Var}\big(\mathbb{1}_{\{\hat{Y}\geq t\}}\big)\, dF_{\hat{Y}}(t)} = 0\,. \tag{8}$$

Equalized odds (EO), which requires $\hat{Y} \perp S \mid Y$, can be addressed by applying the same construction within each outcome stratum (i.e., replacing $F_{\hat{Y}}$ with $F_{\hat{Y}|Y=y}$ and averaging over y). We adopt (8) as a differentiable penalty during training; practical estimators are discussed in appendix D.

C Properties of the estimation ξ_n

In this section we explore some properties of the sample-based estimation of the CvM coefficient ξ_n. An important result is discussed in the following subsection and relates to the connection between the theoretical and sample-based coefficients. Continuity properties in terms of robustness to perturbations are also discussed.

C.1 Asymptotic Consistency

The following theorem, presented in Chatterjee [10] provides asymptotic guarantees on the asymptotic accuracy of the estimation:

Proposition 2 (Theorem 1 in Chatterjee [10])**.** *If Y is not almost surely constant, as $n \to \infty$,*

$$\xi_n(X,Y) \to \xi(X,Y) \;:=\; \frac{\displaystyle\int \mathrm{Var}\Big(\mathbb{E}[\mathbb{1}_{\{Y\geq t\}} \mid X]\Big)\, dF_Y(t)}{\displaystyle\int \mathrm{Var}\big(\mathbb{1}_{\{Y\geq t\}}\big)\, dF_Y(t)} \;\in [0,1]\,.$$

A proof is provided in Chatterjee [10].

C.2 Robustness to perturbations (continuity)

We use the estimator presented in Chatterjee [10]:

$$\xi_n \;=\; 1 \;-\; \frac{n\sum_{k=1}^{n-1}\big|r_{k+1}-r_k\big|}{2\sum_{i=1}^{n} l_i\,(n-l_i)}, \qquad r_i := \#\{\,j : x_j \leq x_i\,\}, \quad l_i := \#\{\,j : y_j \geq y_i\,\}, \tag{9}$$

i.e., ξ_n depends only on the relative orderings of $\{x_i\}$ and $\{y_i\}$ (max-ranks), not on their magnitudes.

Proposition 3 (Robustness to perturbations restated)**.** *Let (X_n, Y_n) be n i.i.d. samples from $p(X, Y)$. Let Y be continuous and let Z_n^1, Z_n^2 contain n i.i.d. samples from a continuous real-valued noise variable. Define $X^\eta := X + \eta Z^1$ and $Y^\eta := Y + \eta Z^2$ for some η. Then, with probability 1,*

$$\begin{aligned} \lim_{\eta \to 0} \xi_n(X, Y^\eta) &= \xi_n(X, Y), \\ \lim_{\eta \to 0} \mathbb{E}\big[\xi_n(X^\eta, Y)\big] \;&=\; \lim_{\eta \to 0} \mathbb{E}\big[\xi_n(X^\eta, Y^\eta)\big] = \mathbb{E}\big[\xi_n(X, Y)\big], \end{aligned} \tag{10}$$

where the expectations are with respect to the perturbation noise. If X is also continuous, the expectations can be removed.

Proof. Fix a realized sample $(x_1, \ldots, x_n, y_1, \ldots, y_n)$.

(i) Perturbing Y only. Since Y is continuous, with probability 1 there are no ties among $\{y_i\}$ and the minimum spacing $\Delta_Y := \min_{i \neq j} |y_i - y_j|$ is strictly positive. Because Z^2 takes finite values, there exists some $\eta_0 > 0$ such that for all $\eta \in [0, \eta_0)$ we have $\max_i |\eta Z_i^2| < \Delta_Y/2$, hence the ordering of $\{y_i\}$ is unchanged. The ranks l_i and r_k are also unchanged. By Equation 9, $\xi_n(X, Y^\eta) = \xi_n(X, Y)$ for all sufficiently small η, yielding $\lim_{\eta \to 0} \xi_n(X, Y^\eta) = \xi_n(X, Y)$ almost surely.

(ii) Perturbing X (and optionally Y). If X is continuous, the same spacing argument applies to $\{x_i\}$, so for all sufficiently small η the X-order is unchanged and hence $\xi_n(X^\eta, Y) = \xi_n(X, Y)$ and $\xi_n(X^\eta, Y^\eta) = \xi_n(X, Y)$ almost surely.

If X may have ties, adding arbitrarily small continuous noise acts as a random tie-breaker within each tied block, producing—conditionally on the untied values—the same distribution over strict total orders as uniform random tie-breaking. Taking expectations over the perturbation therefore averages ξ_n over all consistent tie-breakings; this equals the corresponding (noise-free) expectation of $\xi_n(X, Y)$ computed with random tie-breaking. Hence

$$\lim_{\eta \to 0} \mathbb{E}\big[\xi_n(X^\eta, Y)\big] = \mathbb{E}\big[\xi_n(X, Y)\big], \qquad \lim_{\eta \to 0} \mathbb{E}\big[\xi_n(X^\eta, Y^\eta)\big] = \mathbb{E}\big[\xi_n(X, Y)\big].$$

Combining (i) and (ii) proves the claim. □

Implications. For continuous variables, ξ_n is insensitive to infinitesimal perturbations, which supports stable training when used as a regularizer. When X is discrete, randomized (or noise-induced) tie-breaking preserves ξ_n in expectation, providing robustness at the level of average behavior.

D Implementation of the differential CvM coefficient

Let $r_\Psi^\varepsilon(\theta)$ denote the soft (differentiable) ranking operator defined via projections onto the permutahedron, where $\varepsilon > 0$ is the *smoothness controller* that trades faithfulness to hard ranks for smoother Jacobians [7]. Small ε yields near-exact ranks but poorly conditioned/less informative derivatives; large ε produces well-behaved gradients but compresses the dynamic range of the ranks.

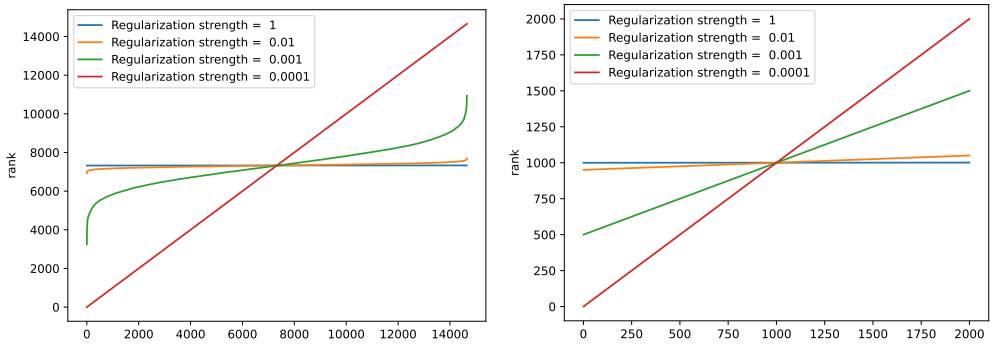
Note: The smoothness operator ε is referred to as *regularization strength* in the paper introducing this soft ranking method. We chose to modify this naming to avoid confusion with the multiplier of the CvM regularization λ.

To showcase the role of this parameter we examine two toy inputs: (i) $n = 15{,}000$ i.i.d. $\mathcal{N}(0, 1)$ samples and (ii) $n = 2{,}000$ equally spaced points on $[0, 1]$ (shuffled). We compare the hard ranks (NumPy) with r_Ψ^ε for several ε and sort both outputs for visualization. Perfect agreement would lie on the $45°$ line. As shown in Figure 6, larger ε preserves order but visibly shrinks the rank spread.

Order-preserving shrinkage and a simple fix

Because r_Ψ^ε is isotonic (order-preserving), the main distortion at larger ε is magnitude shrinkage rather than order mistakes. We therefore post-process the soft ranks with a monotone affine rescaling to match the endpoints of the true rank range. Let $s \in \mathbb{R}^n$ be the soft ranks for a vector, and $t \in \mathbb{R}^n$ its hard ranks. Define

$$s_{\min} = \min_i s_i, \quad s_{\max} = \max_i s_i, \qquad t_{\min} = \min_i t_i, \quad t_{\max} = \max_i t_i,$$



Rankings of random data | Rankings of equally spaced data

Figure 6: Effect of soft-ranking smoothness ε on ranks: larger ε preserves order but shrinks rank spread (Left: 15,000 i.i.d. $\mathcal{N}(0,1)$. Right: 2,000 points in [0,1]).

and apply the mapping

$$\tilde{s}_i \;=\; m(s_i) \;:=\; \frac{s_i - s_{\min}}{s_{\max} - s_{\min}}\,(t_{\max} - t_{\min}) + t_{\min}. \tag{11}$$

This rescaling is strictly increasing, preserves the ordering, and matches boundary values ($\tilde{s}_{\arg\min s} = t_{\min}$, $\tilde{s}_{\arg\max s} = t_{\max}$). Its Jacobian with respect to s is a constant scalar factor $(t_{\max} - t_{\min})/(s_{\max} - s_{\min})$, so gradients remain informative and are merely scaled, which improves numerical conditioning without altering the rank-based structure.

Figure 7 shows that the affine correction restores near-linear alignment to the hard ranks for large-ε soft ranks while retaining smooth derivatives. In all experiments where differentiability is required, we compute r^{ε}_{Ψ} with a moderately large ε and then apply (11) before using the ranks inside the differentiable ξ computation. When training models we suggest adjusting the value of ε to the dataset and tuning it as part of the hyperparameter optimization process.

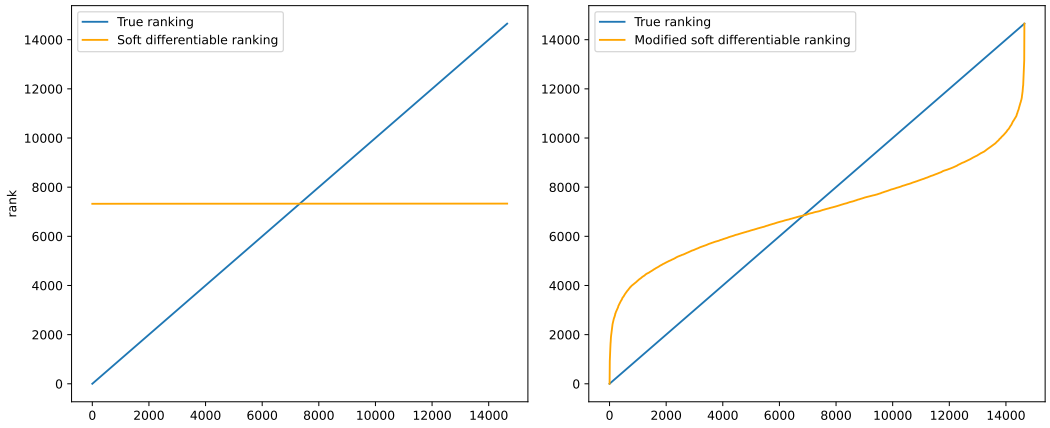


Figure 7: Affine rescaling in Equation 11 restores alignment between soft and hard ranks under large ε (before vs. after).

E Adult dataset: extended results and analysis

Setup. We analyze binary income prediction on *Adult* with two sensitive attributes of different informativeness for the label: *gender* (weakly correlated) and *education* (strongly correlated). Models are 3-layer MLPs with layers of sizes [64, 32, 16]; utility is tracked by accuracy/F1, and fairness by (i) CvM and MI (ii) group gaps for DP/EO (max differences across groups). The displayed results

correspond to the averages and statistics computed on 10 independent runs for each value of the multiplier.

The CvM regularizer's effectiveness varies when switching sensitive attributes, especially when the attribute has imbalanced groups. For instance, when using "education" as the sensitive attribute, groups with low counts (e.g., "Preschool" and "Doctorate") were merged into broader categories, which led to more representative unfairness metrics. The reported results are based on this modified education variable. "Preschool", "1st-4th", "5th-6th", "7th-8th", and "9th" were merged into "Less than HS", while "Doctorate" and "Prof-school" were grouped as "High-income Edu", and the rest as their original categories.

Multiplier effects and correlation regime. Increasing λ lowers CvM and typically shrinks DP/EO gaps, revealing a fairness–utility frontier whose steepness depends on the attribute–label correlation. With *gender*, moderate λ achieves noticeable fairness gains with modest accuracy cost; with *education*, fairness improvements incur sharper utility drops, reflecting a harder trade-off. At $\lambda = 0$, models attain peak utility but exhibit higher dependence; as λ grows, both unfairness and, eventually, performance decrease. Specific values for the metrics and percentual changes can be observed in Tables 1 and 2, respectively. The results corresponding to Table 1 can be visualized with the corresponding error bars in Figure 8.

	Utility (maximize)		**Unfairness (minimize)**				
	Accuracy (↑)	F1 score (↑)	DP (↓)	EoO (↓)	CvM (↓)	MI (↓)	Diff CvM (↓)
$\lambda = 0$	**0.8539**	**0.6665**	0.7868	0.7424	0.2051	0.1468	0.1469
$\lambda = 1$	0.8530	0.6616	0.7304	0.6534	0.1428	0.0997	0.0990
$\lambda = 5$	0.8450	0.6391	0.6153	0.5325	0.0974	0.0614	0.0599
$\lambda = 10$	0.8359	0.6069	0.4797	0.3568	0.0749	0.0444	0.0427
$\lambda = 15$	0.8261	0.5717	0.4059	0.2855	0.0622	0.0373	0.0349
$\lambda = 20$	0.8213	0.5494	**0.3205**	**0.1965**	0.0516	**0.0315**	0.0288
$\lambda = 25$	0.8177	0.5400	0.3308	0.2269	**0.0291**	0.0530	**0.0262**

Table 1: Utility (higher is better) and fairness (lower is better) metrics across regularization strengths λ. Best values are reported in bold.

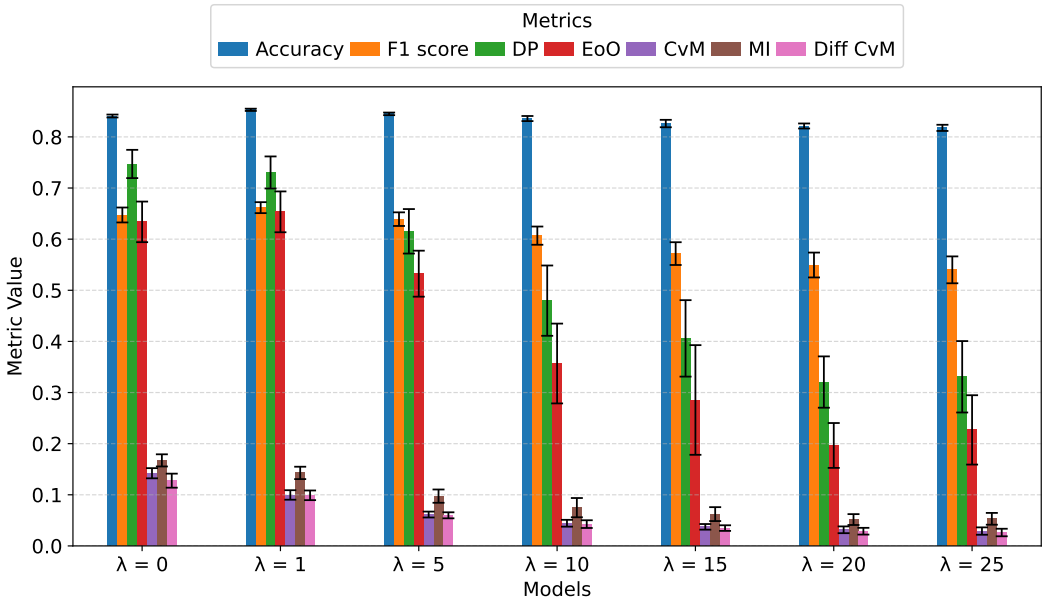


Figure 8: Adult dataset utility (accuracy and F1) and fairness (DP, EoO, CvM, MI, Differentiable CvM) metrics. Values of *unfair* MLP ($\lambda = 0$) and *fair* MLPs ($\lambda \in [1, 5, 10, 15, 20, 25]$, with L_2 penalty). Error bars indicate standard deviation computed over 10 runs.

	Utility (maximize)		Unfairness (minimize)				
	Accuracy (↑)	F1 score (↑)	DP (↓)	EoO (↓)	CvM (↓)	MI (↓)	Diff CvM (↓)
$\lambda = 1$	**-0.10%**	**-0.73%**	-7.17%	-11.99%	-30.36%	-32.09%	-32.63%
$\lambda = 5$	-1.04%	-4.11%	-21.80%	-28.28%	-52.50%	-58.18%	-59.23%
$\lambda = 10$	-2.11%	-8.94%	-39.04%	-51.94%	-63.48%	-69.76%	-70.94%
$\lambda = 15$	-3.25%	-14.22%	-48.41%	-61.55%	-69.67%	-74.59%	-76.25%
$\lambda = 20$	-3.82%	-17.57%	**-59.27%**	**-73.53%**	-74.84%	**-78.55%**	-80.40%
$\lambda = 25$	-4.24%	-18.98%	-57.96%	-69.44%	**-80.18%**	-74.15%	**-82.17%**

Table 2: Relative metrics change with respect to the baseline ($\lambda = 0$) across regularization strengths λ. Best values are reported in bold.

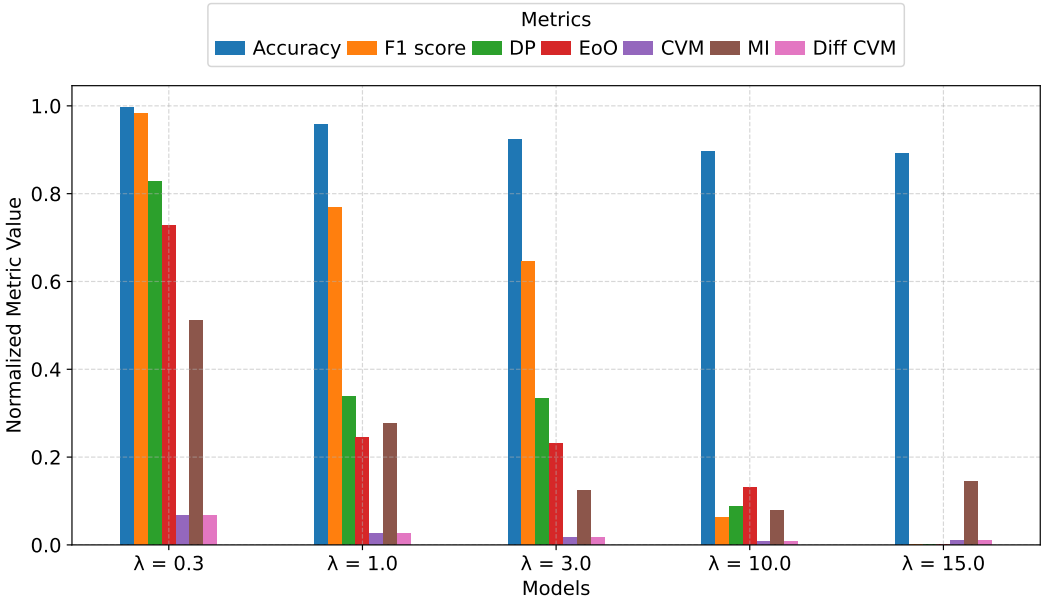


Figure 9: Utility (accuracy, F1 score) and fairness (DP, EO, CvM, MI, Differentiable CvM) metrics comparison of *regularized* models ($\lambda \in [1, 3, 10, 15]$, with L_1 penalty) normalized by values of *unfair* model ($\lambda = 0$). Compared to L_2 regularization, we observe increased instability when training with L_1 loss for the regularizer.

L_1 vs. L_2 on the CvM term. As observed in Figure 9, the increased sensitivity on the λ parameter leads to models failing to learn the task and assigning almost all the labels to the majority class as observed for $\lambda \in \{10, 15\}$ in the plot. Even for small values of the multiplier (compared to those used in Figure 1) the decrease in performance is considerable. As observed in Figure 9, in some cases when the value of the multiplier is set too high (see $\lambda = 15$) the regularizer term can take over and the model fail to learn the task. A practical method to avoid this phenomenon is to use L_2 regularization instead of L_1. Using L_2 regularization makes the magnitude of the corresponding gradient be proportional to the value of the CvM coefficient. More precisely, given a coefficient α for L_1 regularization and β for L_2 regularization then the regularization strength is higher for L_1 if $|\xi_n| < \frac{\alpha}{2\beta}$ and stronger for L_2 if $|\xi_n| > \frac{\alpha}{2\beta}$. Intuitively, L_2 penalizes small ξ more gently when its value is close to 0 (which prevents collapse when λ is large), while increasing pressure on clearly unfair solutions as dependence increases. Empirically, L_2 yields more stable training and preserves utility more consistently than L_1 for comparable fairness gains (Figure 1, Figure 9). We therefore adopt the following L_2 formulation for subsequent runs:

$$\theta^* = \text{argmin}_{\theta \in \Theta} \mathcal{L}(\theta) + \lambda \xi^2(S, Y_\theta). \tag{12}$$

Fine-tuning schedule. We also explore the possibility of using the regularizer for fine-tuning. We start by training the model with no regularization and from that unregularized checkpoint progressively

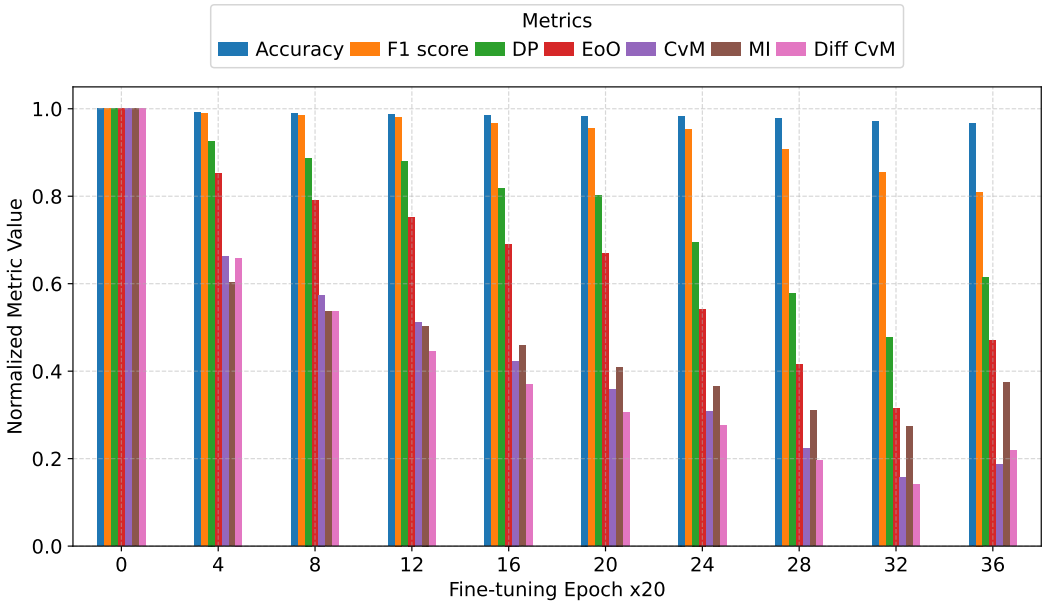


Figure 10: Utility (accuracy, F1 score) and fairness (DP, EO, CvM, MI, Differentiable CvM) metrics comparison normalized by values of *unfair* model ($\lambda = 0$). The unregularized model is trained for 20 epochs and then the value of λ is geometrically increased every 80 epochs. It helps control the stability of the optimization model and not crush into trivial solutions

increase the CvM multiplier to trace a controlled path toward fairness. The process is the following: We train 20 epochs with $\lambda = 0$, then introduce the CvM term and increase λ by $\sqrt{3}$ every 80 epochs. This schedule steadily reduces CvM and DP/EO gaps while maintaining competitive accuracy/F1, avoiding the abrupt utility losses seen when starting with a large λ.

Remark: Finite-sample rank-based estimates can be slightly negative under near-independence. This is expected from sampling variability and is mitigated in practice by using continuous outputs (probabilities) and an L_2 penalty on the CvM term.

Takeaways. On *Adult*, the CvM regularizer enables calibrated movement along the fairness–utility frontier; the effect is gentle for weakly correlated attributes and steeper for strongly correlated ones. L_2 regularization and a staged fine-tuning schedule improve robustness, yielding smoother progress toward lower dependence (CvM) and smaller DP/EO gaps with controlled utility cost.

F Weather forecasting dataset: extended results and analysis

We study large-scale temperature prediction using the *Weather Forecasting* dataset processed per Rubachev et al. [24]. Utility is tracked by (Neg)MSE; fairness by the CvM dependence coefficient and MI. When the sensitive attribute is discrete fairness is also tracked by DP. Because the dataset lacks informative categorical attributes, group structure is obtained by discretizing a continuous variable via binning into subsets of equal size.

F.1 Hyperparameter tuning to control the fairness–performance trade-off

Tuning targets. We tune (i) the CvM multiplier λ and (ii) the derivative smoothness controller ε of the soft-ranking operator. The initial λ sweep spans $[10^{-3}, 10^{2}]$; subsequent sweeps adapt to $[10^{-1}, 10^{3}]$ based on observed frontiers. For ε, we start with an initial wide sweep were no effect is observed. After we reduce the search space to a narrow, practically stable band $[10^{-5}, 10^{-3}]$ to prioritize budget on λ.

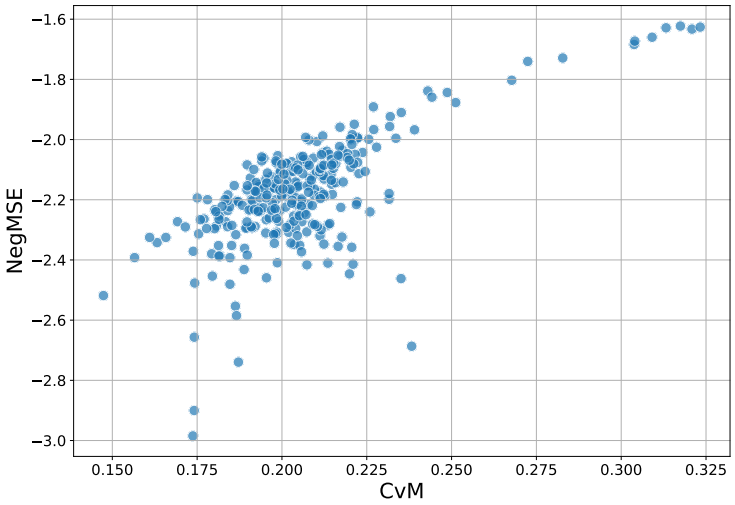


Figure 11: Weather dataset: CvM $\xi(S, \hat{Y})$ vs. NegMSE. This plot enables the visualization of the fairness-utility frontier.

Optimization protocol. We use randomized search with Optuna [3, 26], maximizing a reported objective (see below) with early stopping (patience 16). Budgets: $N{=}300$ trials when using a discretized (binned) sensitive attribute for group metrics, and $N{=}100$ trials when treating the sensitive signal as continuous (CvM-only). We use the default train/val/test splits from the TabReD preprocessing and take `sun_elevation` as the sensitive attribute, which is strongly correlated with the target (≈ 0.47).

F.1.1 Fairness-penalized hyperparameter selection

For generating the results, we follow the same method that we detail in the paper. Namely we conduct the following 3 steps.

Step 1 (utility-only baseline). Tune non-fairness hyperparameters with $\lambda{=}0$ (architecture, optimizer, regularization, early stopping) to establish a performance baseline and confirm task learnability.

Step 2 (fairness-specific tuning). Fix the hyperparameters from the previous step that maximize utility, then sweep the CvM multiplier λ and the smoothness controller ε via randomized search over wide ranges. For this specific case, we identify that ε has a negligible effect on the results so we keep its range limited to $[10^{-6}, 10^{-3}]$ and concentrate most of the exploration power to explore the effect of $\lambda \in [10^{-1}, 10^{3}]$. This initial search over the regularizer-specific hyperparameters provides an initial reference point that enables us to shrink down to the regions of λ and ϵ that are most promising for hyperparameter search in the following step.

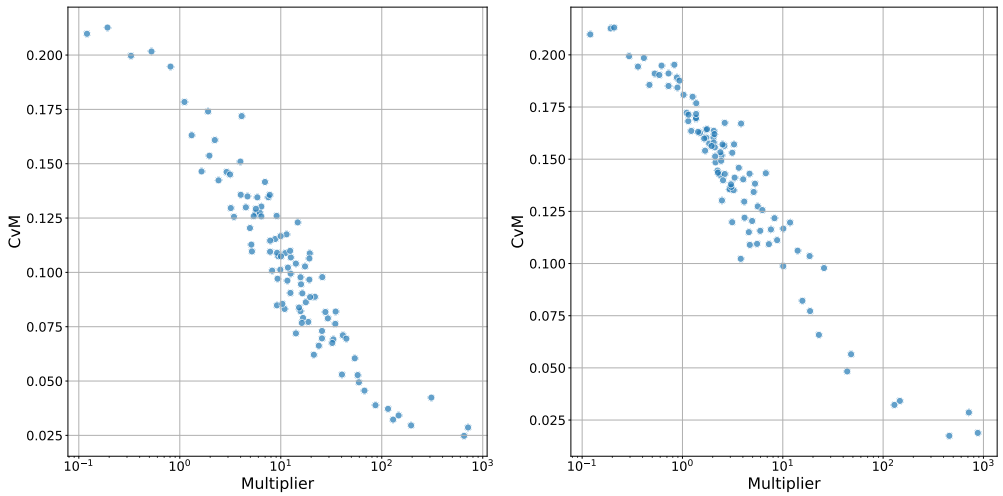
Step 3 (penalized-utility selection). Report to Optuna a fairness-penalized utility score that preserves utility when CvM is below a cutoff and subtracts a linear penalty otherwise:

$$U(\ell, c; \gamma) \;=\; \begin{cases} \ell, & c \le \gamma, \\ \ell - \alpha\,(c - \gamma), & c > \gamma, \end{cases} \tag{13}$$

where ℓ is the utility to maximize (NegMSE), c is the CvM value to minimize, and γ is a user-specified cutoff reflecting the desired group-fairness regime (e.g., via the empirical relation between CvM and DP). We fixed $\alpha{=}10$ and note that other slopes can be explored to adjust selection pressure. This pipeline is dataset-agnostic and directly applicable to other large tabular problems.

F.2 Choosing the cutoff γ

The choice of the cutoff γ will guide the hyperparameter exploration of the hyperparamter optimization module. As seen in Figure 12, the values of the CvM (determined by the choice of γ) will center around the cutoff value.



$\gamma = 0.10$ $\gamma = 0.15$

Figure 12: The cutoff determines what values of the CvM runs will be accumulated around and guides the optimization process.

The choice for the value of the cutoff should be task-specific and can be oriented by the relationship between the CvM and other fairness metrics via the discussed relationship between CvM and metrics such as DP. Plots as the one observed in Figure 3 can be useful to guide this choice.

F.3 Results

We conduct two complementary analyses:

(i) a *continuous* setting, using `sun_elevation` directly for CvM (no groups),

ii) a *discrete* setting, where the same variable is *binned* to form groups for DP evaluation.

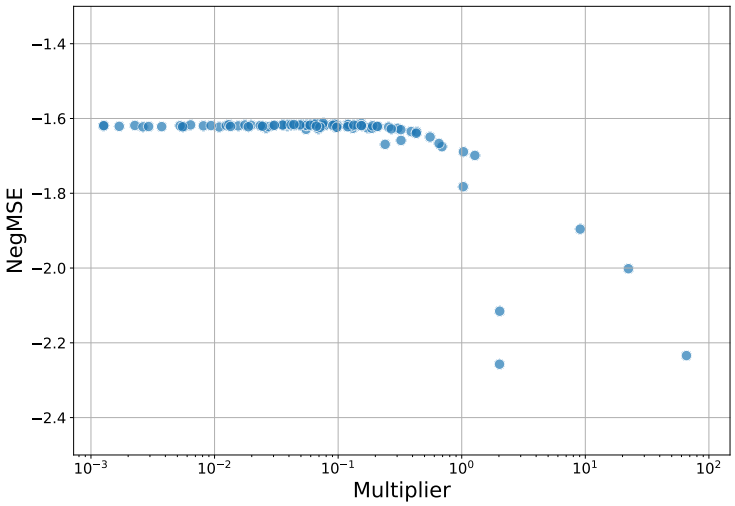


Figure 13: Effect of setting different values of λ on the obtained NegMSE values. An incresase of the multiplier leads to a deterioration of performance.

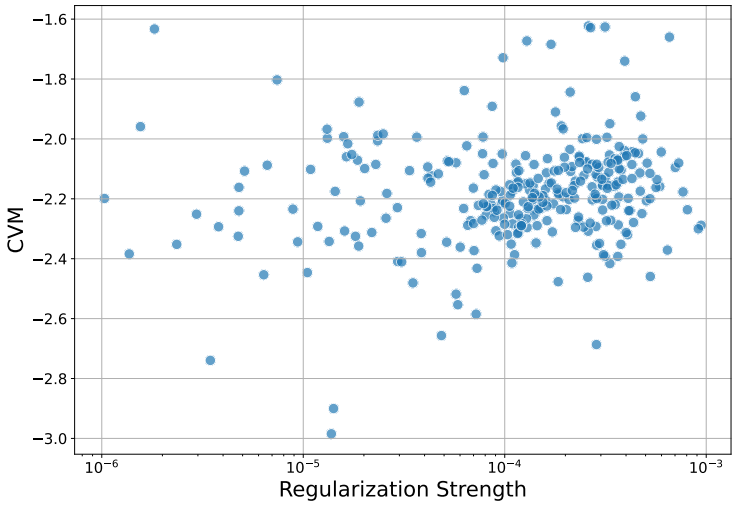


Figure 14: Effect of setting different values of ε on the obtained CvM values. There is a negligible effect of ε on the results.

In both settings, increasing λ monotonically reduces CvM, yielding a Pareto-like frontier against NegMSE, see Figure 11[2]. Moreover, ε exerts limited influence on the CvM–utility trade-off relative to λ, justifying a narrowed search space for ε (Figure 14). In the binned (discrete) analysis, reductions in CvM are accompanied by consistent decreases in DP gaps, enabling a practical mapping from policy targets on DP to choices of λ (Figure 3)[3] with model training.

Visualization. We summarize the trade-off via Pareto-like plots with utility on the horizontal axis and fairness (CvM or DP) on the vertical axis as observed in Figure 11; each point corresponds to a distinct training with a different λ value. To visualize the impact of the regularization parameter λ on the different metrics, Figure 15 depicts the joint evolution of NegMSE, DP, CvM, and MI as λ varies. For this analysis, the training runs were sorted by their corresponding λ values and divided into eight groups containing an equal number of runs, from which the statistics for each bin were computed.

G Additional experimental details

Resources. Experiments were conducted on internal cluster on instances with a RAM of 500Go and 46 CPUs available and 2 GPUs V100.

Practical Guidance

- Use continuous outputs for stability; report DP/EO alongside CvM.
- Prioritize tuning λ; treat ε as a low-priority *derivative smoothness controller* (fix small values unless instability is observed).
- Prefer L_2 on the CvM term; consider ramping λ for fine-tuning when utility is critical.
- For hyper parameter optimization, adopt a fairness-penalized utility to target the desired region of the frontier.

[2] In figures involving NegMSE, two outlier runs were removed for visibility; both corresponded to very large λ yielding low CvM and very poor utility.

[3] Demographic parity can be thought of as a stronger version of the US Equal Employment Opportunity Commission's "four-fifths rule", which requires that the "selection rate for any race, sex, or ethnic group [must be at least] four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate", see the Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. §1607.4(D) (2015).

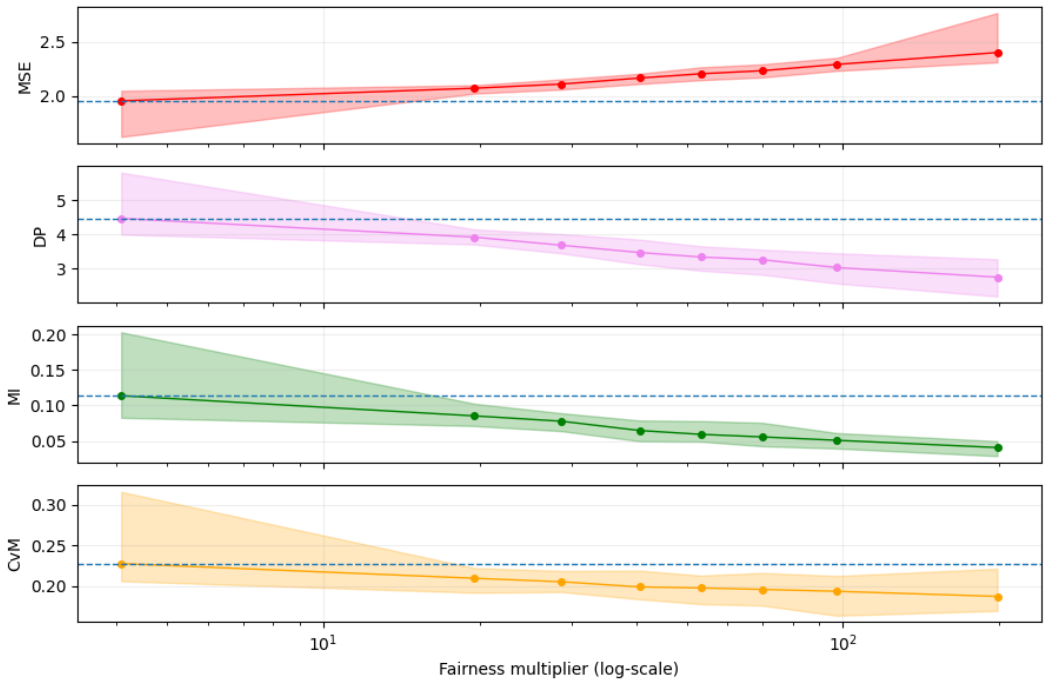


Figure 15: Joint evolution of NegMSE, DP, CvM, and MI as the regularization strength λ varies. The plot summarizes results from 300 training runs, which were sorted by their corresponding λ values and divided into eight groups with an equal number of runs. Dotted lines indicate the metric values obtained from runs with $\lambda = 0$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4 and we state assumptions for theoretical results.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendix B

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Reproducibility results are described in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: the datasets are available, the code is not available (anonymized github is experiencing issues) but will be made available upon acceptance

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In the experiment section in the main text and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: details are explained in the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix G

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We state a connection with AI regulations and improving methods in terms of fairness.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the data, methods, packages we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.