Extension Multiclasse et Limites Empiriques des Modèles avec *Treatment Disparity*

Discussion sur les résultats de Lipton et al. (2019) [1]

Fairness and Pricacy in Machine Learning

Dépôt github : https://github.com/refouch/multitype-disparate-learning-processes/tree/main

Rémi Fouchérand
ENSAE 3A

22 janvier 2026

1. Introduction

L'objectif de ce rapport sera de proposer une présentation, une discussion et une extension des résultats obtenus par Lipton et al. dans un article publié en 2017[1]. L'article en question présente une analyse à la fois juridique et technique des notions de parité de traitement (*treatment disparity*) et de parité de résultats (*outcome disparity*) lors de l'entraînement d'un classifieur binaire. Leur résultat principal énonce que, pour une tâche de classification sous contrainte de *fairness*, un classifieur respectant une parité de traitement offre un compromis sous-optimal entre la précision (*accuracy*) et la parité de résultats, là où, au contraire, un classifieur apprenant une règle de décision différente en fonction du groupe protégé donne un résultat mathématiquement optimal.

La première partie du rapport sera donc dédiée à une présentation plus en détail des résultats principaux de l'article et des limites identifiées dans son approche. Dans une deuxième partie sera proposée une extension théorique et empirique des résultats du papier au cas d'un attribut protégé multiclasse, ainsi qu'une réflexion sur la capacité de généralisation du classifieur optimal proposé. Enfin, nous terminerons par une discussion plus générale sur les implications concrètes et juridiques de telles méthodes de classification.

2. Résumé de l'article

2.1 Thèse principale et conclusions

Avec l'usage de plus en plus fréquent d'outils d'apprentissage automatique dans les processus de décision (pour aiguiller le pôle d'admission d'une université par exemple), la question se pose naturellement de la justesse de ces outils, notamment au regard de biais existants dans les données historiques sur lesquels ils sont entraînés. La littérature classique s'inspire alors des notions juridiques de discrimination directe et indirecte (en anglais : *Disparate treatment* et *Disparate impact*) pour créer des contraintes techniques pour approximer ces deux types de discrimination. La première contrainte veut s'assurer d'une forme « d'égalité de résultats » en imposant que la proportion de décisions positives soit égale pour chaque groupe protégé considéré (*impact parity*). Il s'agit de façon classique du critère de parité démographique abordé en cours. La seconde, plus subtile, rend le classifieur aveugle à l'attribut protégé lors de la phase de prédiction afin de s'assurer d'une égalité de traitement entre les groupes. Conceptuellement, ces deux objectifs sont parfois opposés. Une manière simple d'obtenir une parfaite égalité de résultats reste de discriminer positivement en fonction de l'attribut protégé pour s'assurer d'une parité parfaite. En pratique, cette solution n'est pas toujours acceptable ; les débats récurrents sur la parité homme-femme en entreprise en sont un exemple criant.

Afin de réconcilier ces deux critères, la littérature, suivant les travaux de Kamishima et al. (2011) [2] puis Zafar et al. (2017) [3], propose un type de classifieurs nommés *Disparate Learning Processes* (DLP). L'idée centrale est alors d'imposer des contraintes de parité lors de l'entraînement afin de satisfaire une plus grande parité de résultats, mais en interdisant l'accès aux caractéristiques protégées lors de l'inférence, atteignant ainsi une égalité de traitement. On introduit ainsi un biais dans les probabilités prédites par le modèle (corrigeant idéalement le biais des données historiques), tout en gardant une même règle de décision pour tous les groupes protégés.

La contribution principale du papier de Lipton et al. tient alors dans l'analyse de ces DLPs. Ils montrent en effet que ce genre de méthode produit un arbitrage sous-optimal entre critères de *fairness* et précision du modèle. Au contraire, ils établissent que l'inégalité de traitement (*treatment disparity*), bien que souvent illégale, produit toujours un arbitrage optimal entre ces deux critères et en donnent une formalisation et une implémentation explicite. Leurs conclusions peuvent alors être résumées en trois affirmations principales :

(i) L'inégalité de traitement en fonction de l'attribut protégé est la seule stratégie optimale pour maximiser le critère de précision sous les deux contraintes de justice considérées (parité démographique, p%-rule).

(ii) Dans le cas limite où les features X encodent de manière redondante l'attribut protégé Z, un DLP et un traitement inégalitaire sont équivalents.

(iii) Dans le cas général où les features X ne permettent de prédire Z qu'imparfaitement, un DLP produit une solution sous-optimale, et risque en plus d'introduire des éléments discriminatoires au sein de chaque groupe (par exemple : une personne noire imparfaitement prédite comme blanche sur la base des features X pourra se voir refuser une admission par un classifieur « fair » alors qu'elle aurait été admise par un classifieur sans contraintes).

Ces résultats sont démontrés théoriquement par les auteurs, mais aussi empiriquement sur de nombreux jeux de données réels ou simulés. Cette analyse empirique permet d'une part de confirmer leurs résultats et de montrer que, dans la grande majorité des situations, un classifieur discriminant sera plus efficace pour établir une parité de résultats qu'un classifieur aveugle aux attributs protégés. Cela leur permet aussi d'illustrer concrètement des cas de discrimination intra-groupe générés par un DLP. Ce sont là les deux aspects principaux de l'analyse que nous garderons en tête lors de l'expression des résultats.

2.2 Formalisme et résultats théoriques

Sans pour autant répéter l'entièreté du raisonnement des auteurs, nous nous attarderons ici sur le formalisme utilisé et présenterons les résultats mathématiques principaux. Cela permettra par la suite de montrer comment ces résultats peuvent être étendus au cas d'un attribut protégé multiclasse.

Cadre formel. On se place dans un cadre de classification supervisée binaire. Chaque individu est décrit par un triplet (X, Y, Z), où $X \in \mathcal{X}$ désigne un vecteur de caractéristiques observables, $Y \in \{0, 1\}$ la variable cible, et $Z \in \{a, b\}$ un attribut protégé binaire distinguant un groupe avantagé a et un groupe désavantagé b. Le classifieur est supposé probabiliste et fournit une estimation

$$\hat{p}(x) \approx \mathbb{P}(Y = 1 \mid X = x),$$

à partir de laquelle une décision binaire est obtenue par seuillage : $\hat{Y}(x) = \delta[\hat{p}(x) \geq t]$. La décision bayésienne optimale sans contraintes étant bien sûr donnée par $t = 0.5$.

Dans ce cadre, l'objectif standard consiste à maximiser une métrique de performance, typiquement l'*accuracy*, tout en satisfaisant des contraintes de *fairness* portant sur la dépendance entre la décision et l'attribut protégé Z.

Mesures de disparité d'impact. Les auteurs se concentrent sur des notions de parité d'impact fondées sur les taux de décision positive au sein de chaque groupe. En notant

$$q_z = \mathbb{P}(\hat{Y} = 1 \mid Z = z),$$

deux critères usuels sont considérés : (i) le *Calders–Verwer gap*, défini par $q_a - q_b$, et (ii) la règle des $p\%$, qui impose $q_b/q_a \geq p/100$. Ces contraintes visent à limiter les écarts de décisions positives entre groupes, sans imposer d'égalité stricte.

On fera remarquer par ailleurs que le CV-gap n'est autre que la différence de parité démographique pour un attribut binaire définie dans le cours pour un classifieur $f : (X, Z) \to \{0, 1\}$:

$$DDP = \mathbb{P}(f(X, Z) = 1 \mid Z = a) - \mathbb{P}(f(X, Z) = 1 \mid Z = b) = q_a - q_b.$$

À des fins de généralisation et par souci de cohérence avec le reste de la littérature étudiée, c'est par ce nom que nous nous référerons à ce critère plutôt que celui de *CV-gap* utilisé par les auteurs.

Disparate Learning Processes. Les *Disparate Learning Processes* (DLPs) correspondent à une classe de méthodes dans lesquelles l'attribut protégé Z est utilisé lors de l'apprentissage, mais n'est pas accessible au classifieur final au moment de la prédiction. Formellement, un DLP définit une application

$$\mathcal{A} : (X^n, Y^n, Z^n) \longrightarrow (\mathcal{X} \to \{0, 1\}),$$

de sorte que le modèle produit satisfait la parité de traitement (*treatment parity*) au sens où la décision finale ne dépend pas explicitement de Z.

Résultats théoriques principaux. La contribution théorique centrale de Lipton et al. est de montrer que, sous des hypothèses très générales, les DLPs sont fondamentalement sous-optimaux lorsqu'on cherche à maximiser la performance sous contrainte de parité d'impact. Plus précisément, les auteurs établissent que :

— le classifieur optimal, au sens de l'exactitude, sous contrainte de parité démographique ou de règle des $p\%$, est obtenu par un *seuillage dépendant du groupe*, c'est-à-dire

$$\hat{Y}^*(x, z) = \delta\big[\mathbb{P}(Y = 1 \mid X = x, Z = z) \geq t_z\big],$$

où les seuils t_z diffèrent selon le groupe. δ est une fonction indicatrice.

— toute méthode qui n'exploite pas explicitement l'information Z au moment de la décision, et qui satisfait les mêmes contraintes de parité d'impact, ne peut atteindre une exactitude supérieure à celle de ce classifieur à seuils spécifiques. Autrement dit, la règle de décision $\hat{Y}^*$ par seuils

différenciés domine faiblement la règle de décision d'un DLP :

$$\mathbb{E}\big[\hat{Y}^*(X,Z)(p_{Y|X,Z}-0.5)\big]-\mathbb{E}\big[\hat{Y}_{DLP}(X,Z)(p_{Y|X,Z}-0.5)\big]\geq 0.$$

On pourra consulter le Théorème 4 du papier pour la preuve formelle de cette inégalité.

— Il s'ensuit que lorsqu'un DLP et un classifieur à seuils par groupe réalisent les mêmes taux (q_a, q_b), le premier est nécessairement moins performant, sauf dans le cas dégénéré où les deux décisions coïncident presque sûrement.

Ces résultats impliquent que l'introduction volontaire d'une disparité de traitement contrôlée constitue une stratégie optimale pour arbitrer entre performance prédictive et parité d'impact. À l'inverse, les DLPs déplacent cet arbitrage vers l'espace des représentations ou des scores, au prix d'une perte de performance et d'effets indésirables sur l'équité individuelle. En effet, c'est le deuxième résultat important du papier : l'apparition de discriminations intra-groupes si Z est imparfaitement encodé dans les données.

— De manière presque tautologique, il s'ensuit qu'un DLP accomplit *de facto* une inégalité de traitement (*treatment disparity*) dans le cas où X encode parfaitement Z. Supposons en effet qu'il existe une fonction g connue telle que $z = g(x)$. Alors un classifieur $f(x, z)$ des données peut être réécrit comme $f'(x) = f(x, g(x))$. S'il satisfait *techniquement* la parité de traitement, il n'en reste pas moins équivalent à un classifieur ayant connaissance de l'attribut protégé.

— Dans le cas contraire, Z ne pouvant être estimé qu'imparfaitement à partir de X, il s'ensuit qu'un DLP, en tentant d'imposer une contrainte globale de parité d'impact, introduit nécessairement des distorsions dans l'espace des scores, ce qui conduit à des violations de l'ordre rationnel au sein des groupes, et potentiellement à des formes de discrimination intra-groupe fondées sur des attributs corrélés mais non pertinents. Un membre du groupe désavantagé « ressemblant » à un individu avantagé à cause d'attributs corrélés non pertinents pourra alors être assigné à un label négatif. Par ce fait, les DLP violent la condition *Do No Harm*, problème évité par une différence de traitement entre les groupes.

3. Contribution personnelle

Afin de poursuivre l'analyse proposée par les auteurs, j'ai décidé d'aborder la question laissée en suspens de l'extension de ces résultats au cas d'un attribut sensible multiclasse. Après une brève justification théorique, l'essentiel du travail présenté sera alors une adaptation de l'implémentation des auteurs sur des données réelles en considérant cette fois le cas de l'ethnie plutôt que celui du sexe. Cela nous permet de montrer que les conclusions de l'article demeurent correctes, voire s'aggravent dans certains cas en considérant plusieurs groupes au lieu de deux.

3.1 Extension théorique au cas multiclasse

Nous montrons maintenant que les résultats théoriques établis par Lipton et al. s'étendent naturellement au cas où l'attribut protégé Z prend plus de deux valeurs. C'est quelque chose que les

auteurs mentionnent en passant, mais ne prennent pas le temps d'expliciter. Sans pour autant prétendre proposer ici une preuve exhaustive, nous présentons quelques éléments permettant de s'en convaincre, en s'appuyant sur les résultats de l'article. L'argument principal repose sur le fait que les contraintes de parité d'impact considérées ne dépendent que des taux de sélection q_z par groupe, et non de la structure fine des décisions à l'intérieur de chaque groupe ; les résultats tiennent donc quel que soit le nombre de groupes considérés.

Cadre multiclasse. On suppose désormais que l'attribut protégé prend ses valeurs dans un ensemble fini

$$Z \in \mathcal{Z} = \{1, \dots, K\}, \qquad K \geq 2.$$

Chaque individu est décrit par un triplet (X, Y, Z) avec $Y \in \{0, 1\}$. Une règle de décision d associe à chaque individu une décision $d(X, Z) \in \{0, 1\}$.

Pour chaque groupe $z \in \mathcal{Z}$, on définit le taux de sélection

$$q_z := \mathbb{P}(d(X, Z) = 1 \mid Z = z).$$

Contraintes de parité multiclasse. Les résultats de Lipton et al. tiennent uniquement sous les deux contraintes de parité démographique et de p%-rule. Leur extension au cas d'un attribut protégé multiclasse n'est pas forcément triviale et peut être effectuée de différentes façons. La littérature récente sur les questions intersectionnelles a proposé plusieurs solutions, dont une assez conservatrice qui consiste à comparer les deux groupes les plus extrêmes. C'est une extension assez répandue et souvent utilisée dans la littérature récente [4][5]. C'est par ailleurs la manière dont ces métriques sont généralisées dans la librairie Fairlearn [6]. On propose donc les contraintes suivantes :

— La différence de parité démographique (DDP), définie par

$$\max_{z \in \mathcal{Z}} q_z - \min_{z \in \mathcal{Z}} q_z \leq \gamma;$$

— une généralisation de la règle des p%, imposant

$$\frac{\min_{z \in \mathcal{Z}} q_z}{\max_{z \in \mathcal{Z}} q_z} \geq \frac{p}{100}.$$

On remarque que, dans les deux cas, la contrainte ne dépend que du vecteur des proportions $(q_1, \dots, q_K)$.

Réduction à un problème d'utilité immédiate. Comme dans le cas binaire, la maximisation de l'exactitude est équivalente à la maximisation d'une utilité immédiate. Plus précisément, par le Lemme 1 de Lipton et al., maximiser l'accuracy revient à maximiser

$$u(d) = \mathbb{E}\big[(p_{Y|X,Z}(X, Z) - 0.5)\, d(X, Z)\big].$$

Par linéarité de l'espérance, cette quantité se décompose par groupe :

$$u(d) = \sum_{z \in \mathcal{Z}} \mathbb{P}(Z = z)\, \mathbb{E}\big[(p_{Y|X,Z}(X, z) - 0.5)\, d(X, z) \mid Z = z\big].$$

Optimalité des règles à seuil par groupe. Fixons un groupe z et un taux de sélection q_z. Le problème consistant à maximiser

$$\mathbb{E}\big[(p_{Y|X,Z}(X, z) - 0.5)\, d(X, z) \mid Z = z\big]$$

sous la contrainte $\mathbb{P}(d(X, z) = 1 \mid Z = z) = q_z$ correspond exactement au cadre étudié par Corbett-Davies et al. (2017) [7], puis repris ensuite par Lipton et al. Sous l'hypothèse standard que la variable aléatoire $p_{Y|X,Z}(X, z)$ admet une densité strictement positive sur $[0, 1]$, le Théorème 3.2 de Corbett-Davies et al. établit que la règle optimale est unique (presque sûrement) et consiste à sélectionner les q_z individus ayant les plus fortes probabilités conditionnelles. Cette règle s'écrit sous la forme d'un seuillage :

$$d^*(x, z) = \delta\big[p_{Y|X,Z}(x, z) \geq t_z\big],$$

où le seuil t_z est déterminé par la valeur de q_z.

Effet des contraintes DDP et p%. Les contraintes DDP et $p\%$ n'imposent aucune restriction supplémentaire sur la structure de la règle de décision à l'intérieur des groupes, mais seulement sur les valeurs admissibles du vecteur $(q_1, \ldots, q_K)$. C'est ce qui rend les résultats de l'article généralisables au cas multiclasse. Comme décrit dans l'article, le seuil sera néanmoins différent en fonction de la contrainte sur laquelle on optimise (DDP ou p%).

Conclusion. Cette analyse montre que les résultats de Lipton et al. ne dépendent pas du caractère binaire de l'attribut protégé. Dès lors que la contrainte de *fairness* considérée ne dépend que des taux de sélection par groupe, l'introduction explicite d'un traitement différencié par groupe constitue la stratégie optimale pour maximiser la performance sous contrainte de parité d'impact. Les limitations théoriques et pratiques des DLPs mises en évidence dans le cas binaire persistent donc pleinement dans le cadre multiclasse.

3.2 Implémentation et méthodologie

Afin de vérifier cette intuition théorique, nous tentons ici de reproduire les résultats du papier pour un cas pratique sur des données réelles. Cela nous permettra, en plus de les confirmer, de montrer quelques nuances dans le comportement des DLPs lorsque l'attribut protégé n'est plus binaire.

3.2.1 Jeu de données

On se concentrera ici sur le jeu de données *UCI Adult*, très utilisé en machine learning et connu pour être un *benchmark* fiable, dont l'objectif principal est de prédire si un individu gagne plus de 50 000$ par an. Pour notre analyse, il présente le double avantage de recenser des données réelles dans lesquelles l'ethnie de chaque individu est documentée, ce qui nous permet d'appliquer directement nos résultats en considérant la variable *race* comme protégée.

Le jeu de données est donc repris tel quel et fait l'objet d'un *preprocessing* standard pour une tâche de classification binaire : nous avons normalisé les données numériques et transformé les données catégorielles par encodage one-hot. Nous avons aussi jugé utile de retirer certains attributs inutiles ou nocifs à l'analyse, en particulier l'attribut *country of origin*, afin d'éviter le phénomène décrit dans l'article où Z se retrouve parfaitement encodé par X.

3.2.2 Méthodes comparées

Rappelons ici les résultats de l'article que nous cherchons à reproduire. Idéalement, il nous faut retrouver les deux effets suivants :

(i) La faible domination des méthodes avec seuils différenciés sur les DLPs dans l'arbitrage *accuracy–fairness* selon les deux métriques considérées (DDP et p%).

(ii) L'apparition de phénomènes dégénérés et de discriminations intra-classe dans les prédictions d'un DLP. On cherche aussi à montrer que la méthode par seuillage respecte au contraire la condition *Do No Harm*, c'est-à-dire qu'aucun individu du groupe désavantagé ne soit prédit dans la classe négative alors qu'il était dans la classe positive sans contraintes.

Pour cela, on entraîne trois classifieurs différents :

— Une régression logistique simple, sans contraintes de *fairness* ni pénalisation, qui sert de point de référence à l'analyse.

— Un DLP sous contrainte de parité démographique. Les auteurs utilisent pour cela l'implémentation de Zafar et al. (2017) [3]. Par souci de simplicité, on utilise de notre côté l'implémentation présente dans la librairie *fairlearn*, qui propose sensiblement le même algorithme d'optimisation par descente de gradient exponentielle et produit un résultat similaire.

— Un classifieur au traitement disparate qui établit un seuil de décision propre à chaque groupe, dont l'implémentation est détaillée dans la section suivante.

Il nous suffit ensuite de comparer la performance de ces trois modèles pour se convaincre du point (i). Nous entraînons par ailleurs une autre régression logistique afin d'estimer les probabilités $\mathbb{P}(Z = \text{White})$. On mesure ainsi à quel point un individu « ressemble à une personne blanche » sur la seule base d'une approximation par les features X disponibles. Cela nous permet ensuite de mieux analyser et visualiser les effets de discriminations intra-classe pour se convaincre du point (ii).

3.2.3 Algorithme

Afin d'obtenir le classifieur par seuils optimaux décrit par les auteurs, plusieurs méthodes sont possibles. La plus évidente serait d'implémenter complètement le problème d'optimisation de la fonction d'utilité sous contraintes présentée par les auteurs. Ces derniers présentent cependant une méthode moins coûteuse et plus simple sous la forme d'un algorithme glouton. C'est donc cette méthode que nous reprendrons, même si le cadre multiclasse nous oblige à modifier assez substantiellement l'algorithme proposé.

L'algorithme des auteurs est assez simple et fonctionne sur une logique **d'inversion des prédictions**. En partant d'un classifieur non contraint, on transforme les 1 en 0 dans le vecteur des prédictions pour le groupe avantagé a, en commençant par les individus avec les probabilités les plus basses jusqu'à atteindre la proportion q_a désirée, et inversement pour le groupe désavantagé b. On voit par ailleurs comment cette méthode préserve l'ordre intra-groupe et évite d'introduire une discrimination nouvelle.

Formellement, les auteurs n'utilisent pas juste les probabilités du modèle, mais un score c_i pour déterminer les prédictions à inverser. Assez simplement, il s'agit du ratio de la perte d'accuracy sur le gain de fairness généré par le flip. Cela permet de déterminer les flips les plus avantageux de manière gloutonne pour atteindre une solution optimale. Formellement, on peut noter :

$$\Delta\text{acc}_i = \begin{cases} 1 - 2\hat{p}_i & \text{si } \hat{y}_i = 1 \to 0, \\ 2\hat{p}_i - 1 & \text{si } \hat{y}_i = 0 \to 1. \end{cases}$$

Cette quantité est positive dès lors qu'on s'éloigne de la règle bayésienne optimale.

$$\Delta\text{fair}_i = \begin{cases} \dfrac{1}{n_b} & \text{si } \hat{y}_i = 0 \to 1,\ z_i = b, \\ \dfrac{1}{n_a} & \text{si } \hat{y}_i = 1 \to 0,\ z_i = a. \end{cases}$$

Avec n_a, n_b le nombre total de prédictions dans a et b respectivement. Il s'agit de manière intuitive du gain associé à une simple mise à jour de la proportion de positifs dans le groupe. À noter ici qu'il s'agit du gain associé à la mesure de parité démographique, que nous utilisons principalement dans notre analyse. Les auteurs utilisent de leur côté la p%-rule et obtiennent un calcul du score légèrement différent.

Le score final devient alors :

$$c_i = \frac{\Delta\text{fair}_i}{\Delta\text{acc}_i} = \begin{cases} \dfrac{1}{n_b(2\hat{p}_i - 1)} & \text{si } \hat{y}_i = 0 \to 1,\ z_i = b, \\ \dfrac{1}{n_a(1 - 2\hat{p}_i)} & \text{si } \hat{y}_i = 1 \to 0,\ z_i = a. \end{cases}$$

Algorithm 1 Optimisation des seuils (cas binaire)

Estimer tous les $\hat{p}_i$ puis $\hat{y}_i = \delta(\hat{p}_i \geq 0.5)$ à l'aide d'une régression logistique non contrainte
Calculer le score c_i pour chaque individu
while $q_a - q_b > \gamma$ **do**
    Inverser la prédiction $\hat{y}_i$ avec le plus grand score c_i
end while

L'algorithme présentépar les auteurs pour le cas binaire est par conséquent très simple. Il consiste à inverser successivement les prédictions ayant le meileur score dans chaque groupe. Cela nous donne donc une correction empirique directement sur le vecteur des prédictions considérés. Les seuils théoriques peuvent ensuite être retrouvés très simplement en regardant pour chaque groupe la probabilité associée au dernier individu à avoir été inversé par l'algorithme. Cela peut-être utile pour faire de futures prédictions sur de nouvelles données.

La généralisation au cas multiclasse n'est pas triviale cependant, au vu des métriques que nous avons définies. Nous n'avons implémenté l'algorithme que pour le cas de la parité démographique, c'est à dire sous la contrainte : $\max_z q_z - min_z q_z \leq \gamma$. En effet, comme on ne considère que les groupes ou les proportions sont les plus extrêmes sans se soucier de ceux du milieu, il pourra arriver que le groupe avec la plus forte/faible proportion de prédiction posititve change au cours des itérations de l'algorithme. Cela a en particulier un impact sur les scores calculés. En effet, là où dans le cas binaire on peut se permettre de calculer le score d'inversion une seule fois pour chaque individu, l'extension à plusieurs classes nous oblige à le calculer dynamiquement. Une manière simple d'expliquer cette généralisation consiste simplement à admettre le fait que notre nouvelle définition de la DDP impose que les groupes a et b que l'on compare changent en cours de route. On peut donc dans cette optique réutiliser les formules écrites plus haut, simplement en gardant à l'esprit qu'elles s'appliquent à chaque fois à des groupes différents. D'un point de vue algorithmique cela revient à ne considérer comme "candidats" à l'inversion des prédictions que les membres des groupes dont la proportion de positifs est la plus forte/faible en ignorant tous les autres.

Algorithm 2 Optimisation des seuils (cas multiclasse)

Estimer tous les $\hat{p}_i$ puis $\hat{y}_i = \delta(\hat{p}_i \geq 0.5)$
while $\max_z q_z - \min_z q_z > \gamma$ **do**
    Vider la liste des candidats
    for all $z \in \arg\max_z q_z$ **do**
        for all $i : Z_i = z$ **do**
            if $\hat{y}_i = 1$ **then**
                Calculer c_i
                Ajouter i aux candidats
            end if
        end for
    end for
    for all $z \in \arg\min_z q_z$ **do**
        for all $i : Z_i = z$ **do**
            if $\hat{y}_i = 0$ **then**
                Calculer c_i
                Ajouter i aux candidats
            end if
        end for
    end for
    Choisir i dans le candidat avec le plus grand c_i
    Inverser $\hat{y}_i$
    Mettre à jour la proportion q_z du groupe z concerné
end while

Cela nous permet aussi de gérer le cas où deux groupes présenteraient des proportions q_z identiques : il suffit de considérer leurs membres comme également candidats à l'inversion des prédictions ($1 \to 0$ si il s'agit du groupe dominant, et $0 \to 1$ si il s'agit du groupe dominé). L'algorithme obtenu présente donc une complexité accrue puisqu'on réévalue les groupes dominants/dominés à chaque fois. : il s'agit là de la généralisation naïve de l'algorithme des auteurs et nous ne prétendons pas présenter ici la solution algorithmique optimale au problème. L'implémentation technique plus détaillé de l'algorithme pourra être retrouvée sur le dépôt github du projet.

3.3 Résultats expérimentaux et discussion

Les résultats obtenus par notre expérience sont, sans suprise, très similaires à ceux obtenus par les auteurs dans le cas binaires. Nous remarquerons cependant quelques cas limites qui aparaissent et qu'il conviendra de commenter. Enfin, nous terminerons par quelques considérations méthodologiques qui nous permettrons de nuancer quelque peu l'efficacité des méthodes par seuillage.

3.3.1 Discrimination intra-groupe et condition Do No Harm

La premier résultat que l'on peut vérifier est la présence de discrimination intra-groupes par distorision des probabilités de prédiction elles-mêmes par le DLP. Ces graphiques permettent de visualiser quels individus ont vu leur label changer après application de la règle de décision du DLP par rapport à un classifieur non-contraint.

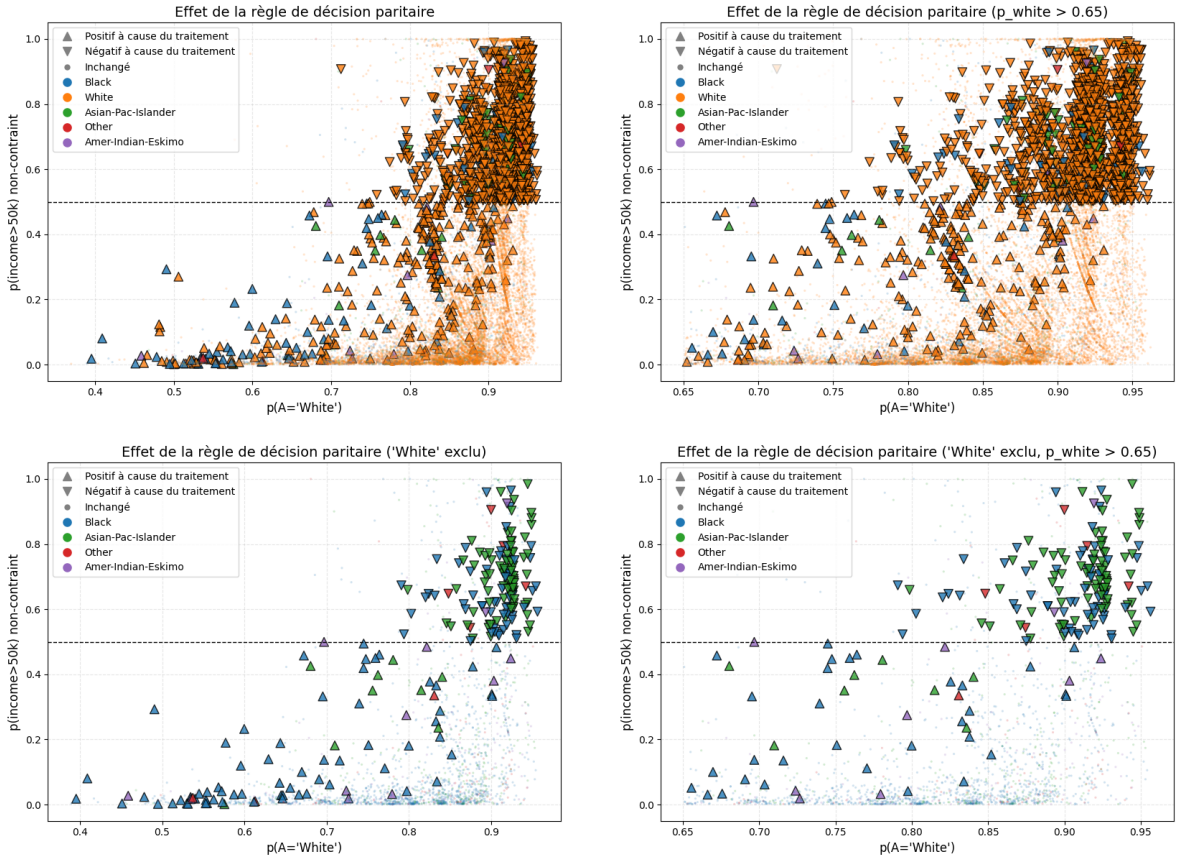

FIGURE 1 – Effet de la règle de décision du DLP comparée au cas standard

On remarque tout de suite que beaucoup de personnes blanches qui auraient lagement été prédites comme gagnant plus de 50k$ se retrouvent finalement dans la classe négative, là où la plupart des bénéficiares de la nouvelle règle de décision se trouvent être afro-américains. On remarque cependant que beaucoup de personnes blanches ne "ressemblant" pas à des blancs d'après les données X se retrouvent avantagées par le DLP. A l'inverse, bon nombre de personnes noires avec une probabilité bien supérieure à 0.5 ne sont pas admises par le classfieur corrigé car elles ont "trop l'air d'être blanches". On retrouve ainsi la discrimination intra-groupe décrite par les auteurs et la violation de la condition "Do No Harm". Certains individus noirs étant désavantagés par un classifieur censé être plus égalitaire simplement car le modèle les considère incorrectement comme faisant partie d'une classe plus privilégiée. D'un point de vue purement philosophique et social c'est une aberration : les programmes de diversité devraient dans l'idéal favoriser les membres les plus talentueux d'une minorité, et certainement pas les rabaisser.

Dans le papier, les auteurs montrent qu'au contraire une règle de décision différenciée pour chaque groupe ne viole pas l'ordre interne du groupe, et respecte donc la condition "Do No Harm". On retrouve ce résultat en pratique avec plus de deux groupes, mais dans une version plus faible : seul le groupe *le plus désavantagé au départ* ne voit aucun de ses membres dans la classe positives être inversés dans la classe négative. On remarque en effet que certains membres d'autres groupes dominés (ici Autre et Amer-Indian-Eskimo) se voient quand même désavantagés. Cela est dû à la nature même de l'algorithme qui recalcule les scores à chaque itération et permet donc ce genre de comportement. On conclut donc qu'il s'ensuit un résultat plus faible que celui annoncé par les auteurs dans le cas binaire : si les membre du groupe le plus défavorisé ne sont jamais désavantagés, cela n'est pas forcément le cas pour les autres groupes défavorisés. À titre de comparaison, la proportion initiale de prédictions positive dans le groupe des afro-américains et dans celui des amérindiens-eskimos est très proche (8% et 9 % respectivement), et pourtant, certains amérindiens se retrouvent tout de même désavantagés

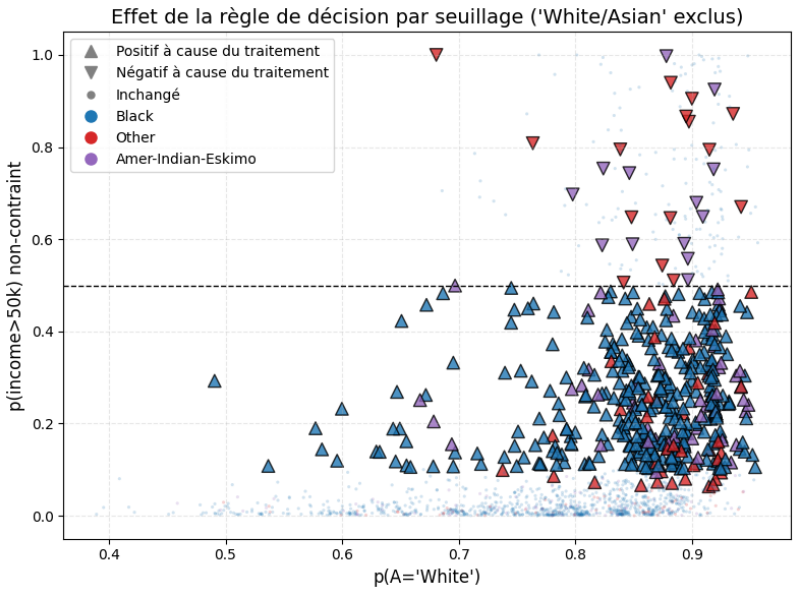

FIGURE 2 – Effet de la règle de décision par seuils différenciés comparé au cas standard

3.3.2 Optimalité de la décision par seuils comparée aux DLPs

Conernant la comparaison de la performance des trois modèles considérés, on retrouve sensiblement les mêmes résultats que les auteurs prouvant ainsi la généralisation naturelle des résultats au cas multiclasse. Le classifieur sans contrainte donne une bonne *accuracy* globale mais au prix d'une piètre performance sur les mesures de parité. Le DLP parvient à optimiser presque parfaitement la parité démographique (qui est la contrainte sur lequel on l'a entraîné) mais au prix d'une perte significative d'*accuracy*. On remarque bien enfin que le traitement différencié par seuils donne le compromis le plus efficace entre *accuracy* et métriques de fairness

TABLE 1 – Comparaison empirique des méthodes en termes de performance et de fairness

Méthode	**Accuracy**	**DDP**	**P%-rule**
Régression logistique (non contrainte)	0.85	0.18	32%
DLP	0.77	0.07	66%
Traitement différencié par seuil (optimal)	0.80	0.00	100%

Cependant, il est important de noter que ces résultats sont obtenus en optimisant l'agortihme des seuils différenciés sur le jeu de test en lui-même. En effet, puisque l'algorithme nécéssite de prendre en entrée un vecteur de prédiction déjà effectuée, il ne peut apprendre les seuils que sur un jeu de données destiné à tester ces mêmes prédictions. Cela pose un énorme problème méthodologique qui n'est pas clairement posé par les auteurs puisque l'algorithme surapprend de manière évidente les contraintes du jeu de test, ce qui explique les métriques parfaites obtenues jusqu'à présent.

Puisque les auteurs ne font pas explicitement mention de ce problème méthodologique, on ne peut pas affirmer si ils l'ont pris en compte dans leur implémentation. On serait amené à penser que non au vu de leurs mesures de parité quasi-parfaites sur l'ensemble de leurs tests.

Nous avons donc voulu mettre cette méthode par seuils différenciés à l'épreuve de la généralisation. Nous proposons un protocole très simmple qui consiste à diviser notre jeu d'entraînement initial en un jeu d'entraînement plus petit doublé d'un jeu de validation ; ce qui est une méthode très fréquente notamment lors de l'entraînement de réseaux de neurones. Nous avons aussi fait en sorte de des proportions pour chaque groupes égales dans les jeux de test et de validation. Ainsi, la régression logistique et le DLP sont donc entraînés normalement sur le nouveau jeu d'entraînement réduit, et testés sur le jeu de test. Pour la méthode par seuillage, on obtient les prédictions initiales à l'aide du jeu d'entraînement, puis on ajuste les seuils à l'aide du jeu de validation. Enfin, les métriques sont calculées sur le jeu de test final, que le classifieur n'a encore jamais vu.

On remarque alors que, dans le contexte très restreint de notre seule expérience empirique, la méthode des seuils semblent tout de même produire un compromis plus avantageux qu'un DLP entre *fairness* et *accuracy* avec des résultats qui semblent plus réalistes. On remarque d'ailleurs que la méthode par seuils obtient une moins bonne DDP qu'un DLP mais une meilleure règle du p%. C'est un résultat assez étonnant étant donné que les seuils sont optimisés sur la parité démographique, mais qui pourrait s'expliquer par des différences de structures entre le jeu de test et de validation. Toujours est-il

que notre seule expérience est bien évidemment insuffisante pour tirer des conclusions fiables quant à la capacité de généralisation de la méthode des seuils différenciés. Il faudrait pousser un peu plus loin l'analyse théorique et empirique pour pouvoir se faire une idée de la pertinence finale de cette méthode. C'est là peut-être la question la plus intéréssante à aborder dans le futur.

TABLE 2 – Comparaison des méthodes sur le jeu de validation et le jeu de test.

Méthode	**Accuracy**	**DDP**	**P%-rule**
Méthodes sans ajustement de seuil			
Régression logistique (non contrainte)	0.85	0.18	0.34
DLP	0.78	0.06	0.63
Traitement différencié par seuil			
Seuils optimisés sur validation (évaluation validation)	0.83	0.00	1.00
Seuils optimisés sur validation (évaluation test)	0.83	0.10	0.72

4. Conclusion

Ce travail avait pour objectif d'analyser de manière critique les *Disparate Learning Processes* (DLPs) proposés dans la littérature pour concilier performance prédictive et contraintes de *fairness*, à partir de l'article fondateur de Lipton et al., puis d'en proposer une extension théorique et empirique au cas plus réaliste d'attributs protégés non binaires.

Dans un premier temps, nous avons montré que les résultats centraux de l'article — à savoir l'optimalité des règles de décision à seuils dépendants du groupe sous contraintes de parité d'impact, et la sous-optimalité structurelle des DLPs — reposent sur des arguments généraux qui ne dépendent pas fondamentalement du caractère binaire de l'attribut sensible. Dès lors que les contraintes de *fairness* considérées ne portent que sur les taux de sélection par groupe, l'introduction explicite d'une disparité de traitement contrôlée demeure la stratégie optimale pour maximiser l'exactitude. Cette observation renforce la portée théorique des résultats de Lipton et al. et confirme que les limites des DLPs sont structurelles plutôt que contingentes à un cadre simplifié.

La première contribution personnelle de ce travail consiste précisément à expliciter cette extension au cas multiclasse, tant sur le plan théorique qu'empirique. Nous avons montré que, si la structure des règles optimales reste inchangée, certaines propriétés mises en avant par les auteurs dans le cas binaire s'affaiblissent lorsque le nombre de groupes augmente. En particulier, la condition dite de *Do No Harm*, selon laquelle aucun membre du groupe le plus désavantagé ne doit être pénalisé par l'introduction de contraintes de *fairness*, n'est plus garantie que pour le groupe le plus extrême. Les groupes intermédiaires peuvent, eux, subir des inversions défavorables de décision, malgré une position initialement défavorisée. Ce résultat met en évidence une limite importante de la généralisation naïve des arguments avancés dans le cas binaire, et invite à une lecture plus prudente des garanties éthiques associées aux méthodes par seuillage lorsque plusieurs groupes sont en concurrence. Cela rejoins par ailleurs la conclusion des auteurs à ce sujet qui invitent avant tout à une utilisation responsable et éclairée de ces méthodes.

La seconde contribution concerne la question de la généralisation des règles de décision optimales par seuils différenciés. Si ces méthodes apparaissent théoriquement dominantes et empiriquement supérieures aux DLPs lorsqu'elles sont ajustées directement sur les données d'évaluation, leur mise en œuvre pratique soulève un problème méthodologique fondamental : l'estimation des seuils repose sur un accès explicite aux distributions observées, ce qui expose la méthode à un surapprentissage des contraintes de *fairness*. Notre expérience montre que, lorsqu'un protocole de validation plus réaliste est mis en place, les performances en termes de parité et d'exactitude se dégradent, bien que le compromis reste globalement plus favorable que celui obtenu par un DLP. Il convient cependant de ne pas tirer de conclusions hâtives à partir de nos résultats, une seule expérience empirique ne suffisant évidemment pas à tirer des conclusions fiables sur la capacité de généralisation des classifieur avec *treatment disparity*. Une analyse théorique et empirique approfondie pourraient constituter un futur travail intéressant.

Ces deux résultats suggèrent que, si le traitement différencié par groupe constitue une solution mathématiquement élégante et conceptuellement transparente au problème de la parité d'impact, son adoption pratique nécessite une réflexion approfondie sur les modalités d'estimation, de validation et d'interprétation des seuils. En définitive, l'extension multiclasse proposée ici renforce le constat critique formulé par Lipton et al. dans leur discussion finale : les contraintes de *fairness* ne peuvent être traitées comme de simples régularisations techniques, indépendantes du contexte social et des choix normatifs sous-jacents.

Références

[1] Zachary C. Lipton, Alexandra Chouldechova et Julian McAuley. *Does mitigating ML's impact disparity require treatment disparity ?* arXiv :1711.07076 [stat]. Jan. 2019. doi : 10.48550/arXiv.1711.07076. url : http://arxiv.org/abs/1711.07076 (visité le 17/01/2026).

[2] Toshihiro Kamishima, Shotaro Akaho et Jun Sakuma. "Fairness-aware Learning through Regularization Approach". en. In : *2011 IEEE 11th International Conference on Data Mining Workshops*. Vancouver, BC, Canada : IEEE, déc. 2011, p. 643-650. isbn : 978-1-4673-0005-6 978-0-7695-4409-0. doi : 10.1109/ICDMW.2011.83. url : http://ieeexplore.ieee.org/document/6137441/ (visité le 17/01/2026).

[3] Muhammad Bilal Zafar et al. *Fairness Constraints : Mechanisms for Fair Classification*. en. arXiv :1507.05259 [stat]. Mars 2017. doi : 10.48550/arXiv.1507.05259. url : http://arxiv.org/abs/1507.05259 (visité le 17/01/2026).

[4] Zhenpeng Chen et al. *Fairness Improvement with Multiple Protected Attributes : How Far Are We ?* arXiv :2308.01923 [cs]. Avr. 2024. doi : 10.48550/arXiv.2308.01923. url : http://arxiv.org/abs/2308.01923 (visité le 17/01/2026).

[5] Giulio Morina et al. *Auditing and Achieving Intersectional Fairness in Classification Problems*. arXiv :1911.01468 [cs]. Juin 2020. doi : 10.48550/arXiv.1911.01468. url : http://arxiv.org/abs/1911.01468 (visité le 17/01/2026).

[6] Hilde Weerts et al. "Fairlearn : Assessing and Improving Fairness of AI Systems". en. In : ().

[7] Sam Corbett-Davies et al. *Algorithmic decision making and the cost of fairness*. en. arXiv :1701.08230 [cs]. Juin 2017. doi : 10.1145/3097983.309809. url : http://arxiv.org/abs/1701.08230 (visité le 18/01/2026).