HyperFair: A Soft Approach to Integrating Fairness Criteria

CHARLES DICKENS, University of California, Santa Cruz, United States of America
RISHIKA SINGH, University of California, Santa Cruz, United States of America
LISE GETOOR, University of California, Santa Cruz, United States of America

Recommender systems are being employed across an increasingly diverse set of domains that can potentially make a significant social and individual impact. For this reason, considering fairness is a critical step in the design and evaluation of such systems. In this paper, we introduce HyperFair, a general framework for enforcing soft fairness constraints in a hybrid recommender system. HyperFair models integrate variations of fairness metrics as a regularization of a joint inference objective function. We implement our approach using probabilistic soft logic and show that it is particularly well-suited for this task as it is expressive and structural constraints can be added to the system in a concise and interpretable manner. We propose two ways to employ the methods we introduce: first as an extension of a probabilistic soft logic recommender system template; second as a fair retrofitting technique that can be used to improve the fairness of predictions from a black-box model. We empirically validate our approach by implementing multiple HyperFair hybrid recommenders and compare them to a state-of-the-art fair recommender. We also run experiments showing the effectiveness of our methods for the task of retrofitting a black-box model and the trade-off between the amount of fairness enforced and the prediction performance.

1 Introduction

As the ubiquity of recommender systems continues to grow, concerns of bias and fairness are becoming increasingly urgent to address. An algorithm oblivious to any form of fairness has the potential to propagate, or even amplify, discrimination [1, 28]. In doing so, certain groups can be severely impacted by the recommendations provided. For instance, one study showed that an algorithm for targeted advertising of jobs in the STEM fields was delivering more advertisements to men than women with similar professional backgrounds [18]. The need to integrate fairness and ensure that different groups of users are experiencing the same level of utility from recommender systems has been acknowledged by the artificial intelligence community [6, 23].

We introduce techniques for integrating fairness metrics as regularizations of a joint inference objective function of a probabilistic graphical model. Our approach naturally leads to novel collections of rules that can be added to a probabilistic soft logic (PSL) [2] model. Furthermore, the weights of the rules can be translated as regularization parameters which can be tuned by the modeler or via weight learning [2]. This motivates a general framework for introducing multiple soft fairness constraints in a hybrid recommender system which we refer to as HyperFair.

HyperFair builds upon the HyPER recommender system introduced by Kouki et al. [15] by adding the ability to enforce multiple soft fairness constraints to the model predictions. This framework is general enough to capture previous work by Farnadi et al. [8] who proposed PSL modelling techniques for addressing disparities stemming from imbalanced training data and observation bias. We develop a generic technique, provide principled derivations of the soft constraints, and show how a set of fairness metrics can be precisely targeted.

Our key contributions are as follows: 1) we introduce the HyperFair framework for enforcing soft fairness constraints in a hybrid recommender system; 2) we show that non-parity and value unfairness can be written as linear combinations of hinge-loss potentials and can thus be integrated into the PSL inference objective via template rules; 3) we perform an empirical analysis using the MovieLens dataset and show both how our fairness rules can be used internally or as a method for retrofitting the output of a black-box algorithm to increase the fairness of predictions; and 4) we show our method improves fairness over baseline models and outperforms a state-of-the-art fair recommender [27] in terms of RMSE and value unfairness.

2 Background

We begin by briefly reviewing related work upon which our approach builds.

2.1 Fairness in Recommender Systems

Methods for addressing fairness can occur at three stages of a recommender pipeline: pre-process, in-process, and post-process [23]. Pre-processing techniques transform the data so that discrimination characteristics are removed prior to model training [5, 17]. In-processing techniques attempt to remove discrimination during the model training process by incorporating changes into the objective function. Post-processing techniques treat the learned model as a black box and modify the output to remove discrimination [10, 11, 20, 24, 25]. Post-processing techniques are particularly attractive to industry since their treatment of the predictor as a black-box makes for a manageable integration into an existing pipeline [9, 10]. Recent work has shown the effectiveness of both adversarial learning [4, 21, 22] and regularization [3, 13, 27]. Our methods can be used as either an in-processing or post-processing method, and builds upon the line of research that addresses fairness via regularization.

2.2 Hybrid Recommenders using Probabilistic Soft Logic

Probabilistic soft logic (PSL) is a probabilistic programming language that has been shown to be effective for hybrid recommender systems [15]. PSL's advantages include the ability to easily write interpretable, extendable, and explainable hybrid systems [16]. PSL models specify probabilistic dependencies using logical and arithmetic rules; the rules, combined with data, are translated into a conditional random field referred to as a *hinge-loss Markov random field (HL-MRF)* [2]. Given a set of evidence $\mathbf{x}$ and continuous unobserved variables $\mathbf{y}$, the inference objective is given by:

$$\min_{\mathbf{y} \in [0,1]^n} \quad \sum_{i}^{k} w_i \phi_i(\mathbf{y}, \mathbf{x}) \tag{1}$$

where k is the number of unique hinge-loss potential functions, $\phi_i(\cdot)$, and w_i are the corresponding scalar weights. In addition to expressivity, an important advantage of HL-MRFs is their scalability; inference is convex, and a variety of specialized optimizers have been proposed [26] (more details about PSL in Section A).

Following Kouki et al. [15], the following collection of rules expresses a simple, intuitive hybrid recommender model:

Demographic and Content Similarity: Demographic-based approaches are built upon on the observation that users with similar demographic properties will tend to make similar ratings. Likewise, content-based approaches are built upon on the observation that items with similar content will be rated similarly by users. This is different from the collaborative filtering approach as the rating patterns of users and items is strictly not considered in the similarity calculation.

$$\textsc{Rating}(\mathtt{U1}, \mathtt{I}) \land \textsc{SimUserDemo}(\mathtt{U1}, \mathtt{U2}) \rightarrow \textsc{Rating}(\mathtt{U2}, \mathtt{I})$$

$$\textsc{Rating}(\mathtt{U}, \mathtt{I1}) \land \textsc{SimItemContent}(\mathtt{I1}, \mathtt{I2}) \rightarrow \textsc{Rating}(\mathtt{U}, \mathtt{I2})$$

The predicate $\textsc{Rating}(\mathtt{U}, \mathtt{I})$ represents the normalized value of the rating that user $\mathtt{U}$ provided for item $\mathtt{I}$.
$\textsc{SimUserDem}(\mathtt{U1}, \mathtt{U2})$ and $\textsc{SimItemContent}(\mathtt{I1}, \mathtt{I2})$ represent the similarity of users $\mathtt{U1}$ and $\mathtt{U2}$ and items $\mathtt{I1}$ and $\mathtt{I2}$.

Neighborhood-based Collaborative Filtering: Neighborhood-based collaborative filtering methods capture the notion that users that have rated items similarly in the past will continue to rate new items similarly. An analogous and transposed notion applies to items, i.e., items that have been rated similarly by many of the same users will continue to be rated similarly. Similarity in this context is based solely on rating patterns and can be measured using various metrics.

$$\textsc{Rating}(\mathtt{U1}, \mathtt{I}) \land \textsc{SimUsers}(\mathtt{U1}, \mathtt{U2}) \rightarrow \textsc{Rating}(\mathtt{U2}, \mathtt{I})$$

$$\textsc{Rating}(\mathtt{U}, \mathtt{I1}) \land \textsc{SimItems}(\mathtt{I1}, \mathtt{I2}) \rightarrow \textsc{Rating}(\mathtt{U}, \mathtt{I2})$$

SimUsers(U1, U2) and SimItems(I1, I2) represent the similarity of users U1 and U2 and items I1 and I2, respectively.

Local Predictor Prior: One of the advantages of the HyPER system is its ability to combine multiple recommendation algorithms into a single model in a principled fashion. Recommender predictions are incorporated as non-uniform priors in the PSL model using the pattern shown below.

$$\textsc{LocalPredictor}(\mathtt{U}, \mathtt{I}) = \textsc{Rating}(\mathtt{U}, \mathtt{I})$$

The predicate LocalPredictor(U, I) represents the prediction made by the external recommendation algorithm.

Mean-Centering Priors: Based on the above rules, ratings are propagated across similar users, and if two users have different average ratings, then these ratings may actually bias one another too much. To counter this effect, the following rules bias ratings towards the average user and item rating calculated from the observed ratings:

$$\textsc{AverageUserRating}(\mathtt{U}) = \textsc{Rating}(\mathtt{U}, \mathtt{I})$$

$$\textsc{AverageItemRating}(\mathtt{I}) = \textsc{Rating}(\mathtt{U}, \mathtt{I})$$

The predicates AverageItemRating(I) and AverageUserRating(U) represent the average normalized value of the ratings associated with user U and item I, respectively.

2.3 FairPSL

Farnadi et al. [8] introduced two collections of rules that can be added to a PSL recommender system to address disparities derived from training data and observations. The authors use the same metrics we consider as proxies to measure this notion of disparity. Our work is different in that we give the modeler a finer level of control. Rather than attempting to capture multiple notions of fairness at once, we propose techniques that integrate specific fairness metrics into the PSL inference objective as regularizers. In this way, the modeler is able to tune the degree of the specific metric to the domain they are working in simply by adjusting the weight of the additional rules.

3 HyperFair

In this section, we introduce HyperFair, a framework for enforcing multiple soft fairness constraints in a hybrid recommender system. HyperFair is a natural development to HyPER[15] that incorporates fairness metrics, U, via regularization of the HL-MRF MAP inference objective (1):

$$\min_{\mathbf{y} \in [0,1]^n} \quad w_f U + \sum_{i}^{k} w_i \phi_i(\mathbf{y}, \mathbf{x}) \tag{2}$$

where $w_f \in \mathcal{R}^+$ is the scalar regularization parameter. A fairness metric, U, in a HyperFair model is expressed as a linear combination of hinge loss potential functions and can then be written as a PSL rule.

A particularly active and productive area of research is defining fairness metrics, and there are many metrics that could be targeted by the proposed HyperFair framework. Following the line of work in [27] and [8], we focus on the unfairness metrics of non-parity and value unfairness defined in the following sections. These metrics were introduced in [27] specifically for addressing disparity stemming from biased training data in collaborative-filtering based recommender systems.

For the remainder of this paper, we will let g represent the protected group of users, i.e., g is a subset of all the users present in the data that have been identified as possessing a protected attribute. Then, $\neg g$ represents the remaining subset of users that do not possess the protected attribute. We let $\mathbf{R}$ be the set of ratings in the dataset, m the number of predictions made by the model, n the number of unique items in the dataset, and $v_{i,j}$ and $r_{i,j}$ the predicted and true rating user i gives item j, respectively. For the fairness metric definitions, we use $E_g[v]$ and $E_{\neg g}[v]$ to represent the average predicted ratings for g and $\neg g$, respectively, $E_g[v]_j$ and $E_{\neg g}[v]_j$ represent the average predicted ratings for item j for g and $\neg g$, respectively, and $E_g[r]_j$ and $E_{\neg g}[r]_j$ represent the average true ratings for item j for g and $\neg g$, respectively.

3.1 Non-parity Unfairness

Non-parity unfairness, U_{par}, aims to minimize the disparity in the overall average predicted ratings of the protected and unprotected groups.

$$U_{par}(v) = |(E_g[v] - E_{\neg g}[v])|$$

In this section, we motivate a collection of rules that can be added to the PSL recommender system as an approach to minimize this metric. We start by introducing two new free variables to the inference problem (1), y_{n+1} and y_{n+2}, and the following hard constraints without breaking the convexity of PSL inference:

$$c_1(\mathbf{y}, \mathbf{x}) := y_{n+1} - \frac{1}{|\{(i,j) : ((i,j) \in \mathbf{R}) \wedge g_i\}|} \sum_{(i,j):((i,j)\in\mathbf{R})\wedge g_i} v_{i,j} = 0$$

$$c_2(\mathbf{y}, \mathbf{x}) := y_{n+2} - \frac{1}{|\{(i,j) : ((i,j) \in \mathbf{R}) \wedge \neg g_i\}|} \sum_{(i,j):((i,j)\in\mathbf{R})\wedge \neg g_i} v_{i,j} = 0$$

With the two additional hard constraints, the solution of the new optimization problem is a state such that $y^*_{n+1} = E_g[v]$ and $y^*_{n+2} = E_{\neg g}[v]$. The two hard constraints can be added to the PSL model by introducing the following pair of rules:

$$\textsc{Rating}(\text{+U}, \texttt{+I})/m_g = \textsc{ProtectedAvgRating}(\texttt{c})\,.\,\{\text{U} : \textsc{Protected}(\text{U})\}\ \{\texttt{I} : \textsc{ProtectedItem}(\texttt{I})\}$$

$$\textsc{Rating}(\text{+U}, \texttt{+I})/m_{\neg g} = \textsc{UnProtectedAvgRating}(\texttt{c})\,.\,\{\text{U} : \textsc{UnProtected}(\text{U})\}\ \{\texttt{I} : \textsc{UnProtectedItem}(\texttt{I})\}$$

where m_g and $m_{\neg g}$ are the total number of ratings for the protected and unprotected group, respectively, and are added as a preprocessing step. The predicates $\textsc{ProtectedAvgRating}(\texttt{c})$ and $\textsc{UnProtectedAvgRating}(\texttt{c})$ hold the values of y_{n+1} and y_{n+2} , respectively. We can now define the two following hinge-loss potentials:

$$\phi_{k+1}(\mathbf{y}, \mathbf{x}) = \max\Big\{1 - y_{n+1} - (1 - y_{n+2}), 0\Big\} \qquad \phi_{k+2}(\mathbf{y}, \mathbf{x}) = \max\Big\{1 - y_{n+2} - (1 - y_{n+1}), 0\Big\}$$

Observe that $U_{par} = \phi_{k+1}(y^*_{n+1}, \mathbf{x}) + \phi_{k+2}(y^*_{n+2}, \mathbf{x})$. This transformation allows us to push the regularizer in (2) into the summation to create a valid PSL objective function. Formally, if we let $w_{k+1} = w_{k+2} = w_f$, then:

$$\underset{\mathbf{y}\in[0,1]^n}{\arg\min} \quad w_f U_{par} + \sum_i^k w_i \phi_i(\mathbf{y}, \mathbf{x}) \quad \equiv \quad \underset{\mathbf{y}\in[0,1]^{n+2}}{\arg\min} \quad \sum_i^{k+2} w_i \phi_i(\mathbf{y}', \mathbf{x}) \qquad \text{s.t.} \quad c_1(\mathbf{y}, \mathbf{x}) = 0,\ c_2(\mathbf{y}, \mathbf{x}) = 0 \tag{3}$$

Furthermore, we now have that the right hand side of (3) is a valid HL-MRF that can be instantiated using PSL.

The following rule can be added to a PSL model to obtain the two desired ground potentials ϕ_{k+1} and ϕ_{k+2}.

$$\textsc{ProtectedAvgRating}(\mathtt{c}) = \textsc{UnProtectedAvgRating}(\mathtt{c})$$

Altogether, this method for addressing non-parity unfairness results in a total of 4 additional ground potentials and 3 additional rules in the PSL template and achieves precisely the desired semantics. Furthermore, the regularization term w_f is directly translated as a weight in PSL that can be tuned by the modeler or via weight learning.

3.2 Value Unfairness

Next, we motivate our approach in addressing value unfairness. Value unfairness aims to minimize the expected inconsistency in the signed estimation error between the protected and unprotected user groups.

$$U_{val}(y) = \frac{1}{n}\sum_{j=1}^{n} |(E_g[v]_j - E_g[r]_j) - (E_{\neg g}[v]_j - E_{\neg g}[r]_j)|$$

A key difference between this metric and non-parity is that the truth values of the predictions are included in the definition of the metric and thus cannot be directly targeted during inference. In PSL the truth values of the target predicates are withheld until evaluation. Therefore, the approach we take is to estimate properties of the rating distribution prior to running the model to approximate the desired inference objective function (2).

We start the derivation of the fairness rules we will be adding to the PSL model by augmenting the inference optimization problem of (1) with two hard constraints for every item in the dataset, that is for all $j \in I = \{j : (i,j) \in \mathbf{R}\}$. Note that these constraints do not break the convexity properties of the original optimization problem.

$$c_{1,j}(y_{n+j}, \mathbf{x}) := y_{n+j} - \frac{1}{|\{i : ((i,j) \in \mathbf{R}) \wedge g_i\}|} \sum_{i:((i,j)\in\mathbf{R})\wedge g_i} v_{i,j} = 0$$
$$c_{2,j}(y_{n+j+|I|}, \mathbf{x}) := y_{n+j+|I|} - \frac{1}{|\{i : ((i,j) \in \mathbf{R}) \wedge \neg g_i\}|} \sum_{i:((i,j)\in\mathbf{R})\wedge \neg g_i} v_{i,j} = 0$$

With these hard constraints, the setting of the free variables $y_{n+1}, \cdots, y_{n+|I|}, y_{n+|I|} \cdots y_{n+2|I|}$ in the optimal solution will be such that $(y^*_{n+1} = E_g[v]_1), \cdots, (y^*_{n+|I|} = E_g[v]_{|I|}), (y^*_{n+1+|I|} = E_{\neg g}[v]_1) \cdots (y^*_{n+2|I|} = E_{\neg g}[v]_{|I|})$. These hard constraints are added to the PSL inference objective with the following rule:

$$\textsc{Rating}(+\mathtt{U}, \mathtt{I})/@Max[1, |\mathtt{U}|] = \textsc{PredGroupAvgItemRating}(\mathtt{G}, \mathtt{I}) \,.\, \{\mathtt{U} : \textsc{target}(\mathtt{U}, \mathtt{I}) \wedge \textsc{Group}(\mathtt{G}, \mathtt{U})\}$$

$\textsc{PredGroupAvgItemRating}(\mathtt{G}, \mathtt{I})$ represents the average of the predicted ratings that users in group G gave item I. The term G in this rule represents either the unprotected or protected group.

At the time of inference, we cannot calculate the average true values of the ratings for either the protected or unprotected group, $E_g[r]_j$ or $E_{\neg g}[r]_j$, since the true rating value information is withheld until evaluation. Instead, the group average item rating is estimated using the observed ratings, $\hat{E}_g[r]_j = \frac{1}{|\{i:((i,j)\in\mathbf{R}_{obs})\wedge g_i\}|} \sum_{i:((i,j)\in\mathbf{R}_{obs})\wedge g_i} v_{i,j}$ and similarly, $\hat{E}_{\neg g}[r]_j = \frac{1}{|\{i:((i,j)\in\mathbf{R}_{obs})\wedge \neg g_i\}|} \sum_{i:((i,j)\in\mathbf{R}_{obs})\wedge \neg g_i} v_{i,j}$, where $\mathbf{R}_{obs}$ is the set of observed ratings. The observed group average item rating is calculated in a preprocessing step and is added to the model as an observed predicate, $\textsc{ObsGroupAvgItemRating}(\mathtt{G}, \mathtt{I})$.

We can now define the following set of hinge-loss potentials:

$$\phi_{k+j}(\mathbf{y}, \mathbf{x}) = \max\left\{(y_{n+j} - \hat{E}_g[r]_j) - (y_{n+j+|I|} - \hat{E}_{\neg g}[r]_j), 0\right\}$$
$$\phi_{k+j+|I|}(\mathbf{y}, \mathbf{x}) = \max\left\{(y_{n+j+|I|} - \hat{E}_{\neg g}[r]_j) - (y_{n+j} + \hat{E}_g[r]_j), 0\right\}$$

Then, in the optimal state: $U_{val} \approx n \sum_{j=1}^{2|I|} \phi_{k+j}(\mathbf{y}^*, \mathbf{x}) =: \hat{U}_{val}$. This transformation allows us to push an approximation of the regularizer in (2) into the summation of the HL-MRF MAP inference objective function. Formally, if we let $w_{k+j} = w' = \frac{1}{n} w_{Val}$ for all $j \geq 0$, then:

$$\underset{\mathbf{y} \in [0,1]^n}{\arg\min} \quad w_f \hat{U}_{val} + \sum_{i}^{k} w_i \phi_i(\mathbf{y}, \mathbf{x}) \quad \equiv \quad \underset{\mathbf{y}' \in [0,1]^{n+2|I|}}{\arg\min} \quad \sum_{i}^{k+2|I|} w_i \phi_i(\mathbf{y}', \mathbf{x}) \qquad \text{s.t.} \quad c_{1,j}(\mathbf{y}', \mathbf{x}) = 0,\ c_{2,j}(\mathbf{y}', \mathbf{x}) = 0,\ \forall j \in I \tag{4}$$

Further, we have (4) is a valid HL-MRF that can be instantiated using PSL. Specifically, the following rule in PSL results in the desired potentials $\phi_{k+1}, \cdots, \phi_{k+2|I|}$:

$$\begin{aligned} &\textsc{PredGroupAvgItemRating}(\mathtt{G1}, \mathtt{I}) - \textsc{ObsGroupAvgItemRating}(\mathtt{G1}, \mathtt{I}) \\ = &\textsc{PredGroupAvgItemRating}(\mathtt{G2}, \mathtt{I}) - \textsc{ObsGroupAvgItemRating}(\mathtt{G2}, \mathtt{I}) \end{aligned}$$

This approach approximates the targeted fairness metric using statistics from the set of observations. The approximation is then transformed into a summation of hinge-loss potential functions that could be pushed into a PSL inference objective function. Furthermore, the weight of the arithmetic rule in this intervention can be interpreted as a scaled version of the regularization parameter of the fairness metric in (2).

4 Empirical Evaluation

We evaluate our proposed PSL fairness interventions on the Movielens 1M dataset [12]. In addition to the ratings, this dataset includes auxiliary information such as movie metadata (e.g., genres, title, and release date) and user demographic information (e.g., gender, age, and occupation). In table 2 of [27], the authors summarize some of the gender-based statistics of the MovieLens 1M dataset which underscores the disparity in the ratings. Therefore, following the same preprocessing steps taken in previous related work [8, 27], the protected and unprotected groups are chosen to be female and male, respectively, and movies are filtered by genre, considering only those with *action*, *romance*, *crime*, *musical*, and *sci-fi* tags, and finally users with fewer than 50 ratings are removed. The remaining dataset contains approximately 450K timestamped ratings made by 3K users across 1K movies.

4.1 In-Process Fairness Intervention

The baseline hybrid recommender system (Section 2.2) uses cosine similarity for the similarity predicates and three local predictors, non-negative matrix factorization (NMF) [19], biased singular value decomposition (SVD) [14], and a content-based multinomial Naive Bayes multi-class classifier with Laplace smoothing that is trained using the demographic and content information of the user and item, respectively. Five versions of the PSL recommender system, extending the baseline model, are implemented:

- **PSL Base:** demographic and content similarity, collaborative filtering, local predictor, and mean-centering rules
- **HyperFair(NP):** The **PSL Base** model with the non-parity fairness rules.
- **HyperFair(Val):** The **PSL Base** model with the value fairness rules.
- **HyperFair(NP + Val):** The **PSL Base** model with both the non-parity and value fairness rules.
- **Fair PSL:** The **PSL Base** model with rules introduced by Farnadi et al. [8].

We also implemented three of the fair matrix factorization methods introduced by Yao and Huang [27] using the hyperparameters and training methods chosen by the authors, i.e., we use a L2 regularization term $\lambda = 10^{-3}$ and learning rate of 0.1 for 500 iterations of Adam optimization using the full gradient. The first is the baseline matrix

Table 1. Prediction performance(RMSE) and unfairness(Non-Parity and Value) of recommender systems on MovieLens1m

Model	**RMSE (SD)**	**Non-Parity (SD)**	**Value (SD)**
MF	0.945(1.0*e*-3)	0.0371(1.6*e*-3)	0.349(6.0*e*-3)
MF NP	0.945(1.0*e*-3)	**0.0106(2.0e-3)**	0.351(6.3*e*-3)
MF Val	0.950(6.4*e*-4)	0.0446(3.0*e*-3)	0.343(3.4*e*-3)
Fair PSL	**0.932(9.7e-4)**	0.0274(9.6*e*-4)	**0.332(5.5e-3)**
PSL Base	**0.931(1.2e-3)**	0.0270(1.4*e*-3)	**0.330(4.4e-3)**
HyperFair(NP)	0.945(1.1*e*-2)	0.0215(5.0*e*-3)	**0.338(9.9e-3)**
HyperFair(Val)	**0.932(1.1e-3)**	0.0267(8.6*e*-4)	**0.333(6.9e-3)**
HyperFair(NP + Val)	**0.932(1.1e-3)**	0.0274(1.4*e*-3)	**0.331(4.5e-3)**

factorization based approach which we refer to as **MF**. The second and third methods are where the matrix factorization objective function is augmented with the smoothed non-parity and value unfairness metrics, which we refer to as **MF NP** and **MF Val**, respectively. For both the HyperFair models we introduced in this work and the matrix factorization methods, we set the fairness regularization parameter to 1.0.

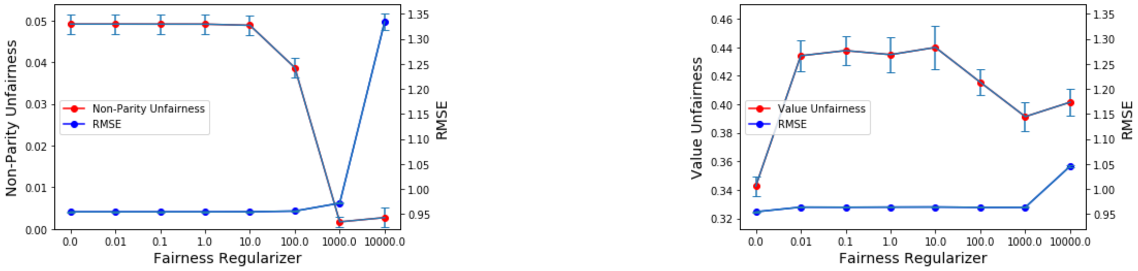
We measure the prediction performance of each of the models using the RMSE of the rating predictions. The unfairness of the predictions are measured using the metrics defined in Section 3. The prediction performance and fairness metrics are measured across 5 folds of the MovieLens dataset and we report the mean and standard deviation for each model. We bold the best value, and values not statistically different from the best with $p < 0.05$ for a paired sample t-test.

We can see from Table 1 that the HyperFair models either improved on the targeted fairness metric over **PSL Base** or achieved results that were not significantly different from the best PSL model. Notably, **HyperFair(NP)** achieved significantly better non-parity unfairness over **PSL Base** and the anticipated performance decrease in the RMSE of the rating predictions was minimal. In fact, the **HyperFair(NP)** still achieved the same level of prediction accuracy as the highest performing matrix factorization method. Another interesting takeaway from Table 1 is that when attempting to optimize for both non-parity and value unfairness simultaneously in **HyperFair(NP + Val)**, the effectiveness of the non-parity rule decreases. This effect can also be observed in the **Fair PSL** of [7], where both fairness notions are attempting to be addressed with a single set of rules. A potential explanation of this could be that the value unfairness and non-parity unfairness metrics are opposing one another, that is to say that a set of ratings that performs well on value unfairness in this dataset may actually perform poorly on non-parity unfairness, and vice-versa. Fully understanding this behavior and controlling the tradeoff between the metrics is a direction for future work.

When comparing the PSL models to the matrix factorization models from [27], we see that PSL consistently achieves better RMSE and value unfairness, while matrix factorization achieves better non-parity unfairness values. It is important to note that Table 1 only reflects metrics values for the regularization parameter $w_f = 1.0$. In the next set of experiments, we show how tuning the non-parity unfairness regularization parameter can effectively yield predictions that fall within a desired non-parity unfairness threshold.

4.2 Post-Process Fairness Intervention

Another way we employ our proposed methods is as an interpretable fair retrofitting procedure for predictions from an arbitrary black-box model. To show the effectiveness of our methods for this task, we create a simple PSL model that contains only the NMF local predictor rule and the fairness rule. We refer to these models as **NMF + NP PSL** and **NMF + Val PSL** for the models with non-parity and value unfairness rules, respectively. The weights of the fairness rules in



(a) (b)

Fig. 1. (a) Non-Parity unfairness and RMSE performance of **NMF** + **NP PSL** vs the value of the fairness regularization parameter. (b) Value unfairness and RMSE performance of **NMF** + **Val PSL** vs the value of the fairness regularization parameter.

the templates are varied to show the tradeoff between prediction performance and the fairness metric. For both models, we run experiments for all $w_f \in \{0.0, 0.01, 0.1, \cdots, 10000.0\}$.

Fig. 1 shows both the RMSE and the fairness of the ratings predicted by the NMF model. We see that **NMF + NP PSL** begins to improve the predictions' fairness without significantly decreasing the performance when the regularization parameter is set to 100.0. When the parameter exceeds 10.0, there is a significant decrease in the non-parity unfairness, reaching nearly 0.0, while the increase in RMSE is still not drastic.

The **Val + NP PSL** model initially does not improve the value unfairness of the NMF rating predictions. We suspect this behavior suggests that the quality of the estimator for the group average item rating needs improvement and initially biases the predictions in a detrimental way and is a direction for future investigation. There is a region where the value unfairness begins a downward trend with respect to the fairness regularizer, as is desired. This is an encouraging result, showing that the weight of the fair rule generally has the desired relationship with value unfairness.

5 Conclusion and Future Work

We introduced the HyperFair framework for enforcing multiple soft fairness constraints in a hybrid recommender system. The effectiveness of HyperFair methods are tested in a movie recommendation setting and it is shown they can improve the fairness of predictions and still achieve state-of-the-art performance. One direction for future work would be to integrate fairness into the weight learning objective of PSL. The experimental results of Section 4.2 showcased that the fairness regularizer is an important hyperparameter to tune, and one way to automate this procedure is via weight learning [2]. Another direction for future work is to integrate more fairness metrics and explore what new modeling patterns can be developed for ranking based recommender systems.

6 Acknowledgements

This work was partially supported by the National Science Foundation grants CCF-1740850 and IIS-1703331, AFRL and the Defense Advanced Research Projects Agency. We also thank the reviewers for their constructive feedback that helped shape the paper.

References

[1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *PACMHCI*, 3(199), 2019.

[2] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *JMLR*, 18, 2017.

[3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *KDD*, 2019.

[4] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *FAT ML*, 2017.

[5] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *NIPS*. 2017.

[6] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *FAT**, 2019.

[7] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness in relational domains. In *AIES*, 2018.

[8] Golnoosh Farnadi, Pigi Kouki, Spencer K. Thompson, Sriram Srinivasan, and Lise Getoor. A fairness-aware hybrid recommender system. In *RecSys*, 2018.

[9] Jean Garcia-Gathright, Aaron Springer, and Henriette Cramer. Assessing and addressing algorithmic bias - but before we get there. In *AAAI Spring Symposium Series*, 2018.

[10] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *SIGKDD*, 2019.

[11] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

[12] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *TIIS*, 5(4), 2015.

[13] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECMLPKDD*, 2012.

[14] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

[15] Pigi Kouki, Shobeir Fakhraei, James Foulds, Magdalini Eirinaki, and Lise Getoor. Hyper: A flexible and extensible probabilistic framework for hybrid recommender systems. In *RecSys*, 2015.

[16] Pigi Kouki, James Schaffer, Jay Pujara, John ODonovan, and Lise Getoor. User preferences for hybrid explanations. In *RecSys*, 2017.

[17] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *ICDE*, 2019.

[18] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7), 2016.

[19] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, 2001.

[20] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. Personalized fairness-aware re-ranking for microlending. In *RecSys*, 2019.

[21] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In *ICLR*, 2016.

[22] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2019.

[23] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[24] Dimitris Paraschakis and Bengt Nilsson. Matchmaking under fairness constraints: a speed dating case study. In *ECIR*, 2020.

[25] Ashudeep Singh and Thorsten Joachims. Fairness of exposure in rankings. *KDD*, 2018.

[26] Sriram Srinivasan, Eriq Augustine, and Lise Getoor. Tandem inference: An out-of-core streaming algorithm for very large-scale relational inference. In *AAAI*, 2020.

[27] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *NIPS*, 2017.

[28] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

Appendix A Appendix

A.1 Probabilistic Soft Logic

Probabilistic soft logic (PSL) is a statistical relational learning (SRL) framework that uses arithmetic and first order like logical syntax to define a specific type of probabilistic graphical model called a hinge-loss Markov random field (HL-MRF) [2]. To do this, PSL derives potential functions from the provided rules which take the form of hinges. Data is used to instantiate several potential functions in a process called grounding. The resulting potential functions are then used to define the HL-MRF. The formal definition of a HL-MRF is as follows:

Definition 1. *Hinge-loss Markov random field. Let* $\mathbf{y} = (y_1, \cdots, y_n)$ *be a vector of* n *variables and* $\mathbf{x} = (x_1, \cdots, x_{n'})$ *a vector of* n' *variables with joint domain* $\mathbf{D} = [0,1]^{n+n'}$*. Let* $\phi = (\phi_1, \cdots, \phi_m)$ *be a vector of* m *continuous potentials of the form* $\phi_i(\mathbf{y}, \mathbf{x}) = (\max\{\ell_i(\mathbf{y}, \mathbf{x}), 0\})^{p_i}$*, where* ℓ_i *is a linear function of* $\mathbf{y}$ *and* $\mathbf{x}$ *and* $p_i \in \{1, 2\}$*. Let* $\mathbf{c} = (c_1, \cdots, c_r)$ *be a vector of* r *linear constraint functions associated with index sets denoting equality constraints* $\mathcal{E}$ *and inequality constraints*

$\mathcal{I}$, which define the feasible set .

$$\tilde{\mathbf{D}} = \left\{ (\mathbf{y}, \mathbf{x}) \in \mathbf{D} \;\middle|\; \begin{array}{ll} c_k(\mathbf{y}, \mathbf{x}) = 0, & \forall k \in \mathcal{E} \\ c_k(\mathbf{y}, \mathbf{x}) \leq 0, & \forall k \in \mathcal{I} \end{array} \right\}$$

*Then, for $(\mathbf{y}, \mathbf{x}) \in \tilde{\mathbf{D}}$, given a vector of m nonnegative parameters, i.e., weights, $\mathbf{w} = (w_1, \cdots, w_m)$, a **hinge-loss Markov random field** $\mathcal{P}$ over random variables $\mathbf{y}$ and conditioned on $\mathbf{x}$ is a probability density defined as:*

$$P(\mathbf{y}|\mathbf{x}) = \begin{cases} \frac{1}{Z(\mathbf{w},\mathbf{x})} \exp(-\sum_{j=1}^{m} w_j \phi_j(\mathbf{y}, \mathbf{x})) & (\mathbf{y}, \mathbf{x}) \in \tilde{\mathbf{D}} \\ 0 & o.w. \end{cases} \tag{5}$$

where $Z(\mathbf{w}, \mathbf{x}) = \int_{\mathbf{y}|(\mathbf{y},\mathbf{x} \in \tilde{\mathbf{D}})} \exp(-f_{\mathbf{w}}(\mathbf{y}, \mathbf{x}))d\mathbf{y}$ is the partition function for the conditional distribution.

Rules in a PSL model capture interactions between variables in the domain and can be in the form of a first order logical implication or a linear arithmetic relation. Each rule is made up of predicates with varying numbers of arguments. Substitution of the predicate arguments with constants present in the data generate ground atoms that can take on a continuous value in the range $[0, 1]$. A logical rule must have a conjunctive clause in the body and a disjunctive clause in the head, while an arithmetic rule must be an inequality or equality relating two linear combinations of predicates.

A logical rule is translated as a continuous relaxation of Boolean connectives using *Lukasiewicz* logic. Specifically, $P \land Q$ results in the potential $\max(0.0, P{+}Q - 1.0)$, $P \lor Q$ results in the potential $\min(1.0, P{+}Q)$, and $\neg Q$ results in the potential $1.0 - Q$. Each grounding of an arithmetic rule is manipulated to $\ell(\mathbf{y}, \mathbf{x}) \leq 0$ and the resulting potential takes the form $\max\{\ell(\mathbf{y}, \mathbf{x}), 0\}$.

We now illustrate the process of instantiating a HL-MRF using PSL with an example in the context of recommender systems.

$\textsc{SimUser}(\texttt{U1}, \texttt{U2}) \land \textsc{Rating}(\texttt{U1}, \texttt{M}) \longrightarrow \textsc{Rating}(\texttt{U2}, \texttt{M})$	(1)
$\neg\textsc{SimUser}(\texttt{U1}, \texttt{U2}) \lor \neg\textsc{Rating}(\texttt{U1}, \texttt{M}) \lor \textsc{Rating}(\texttt{U2}, \texttt{M})$	(2)
$min\{1.0, (1.0 - \textsc{SimUser}(\texttt{Alice}, \text{Bob}){+}(1.0 - \textsc{Rating}(\texttt{Alice}, \texttt{Alien})){+}\textsc{Rating}(\text{Bob}, \texttt{Alien})\}$	(3)

Example 1. *Consider a movie recommendation setting where the task is to predict the ratings users will provide movies that they have not yet rated. We can encode the idea that similar users are likely to enjoy the same movies using the logical statement (1).*

Here, SimUser is an observed predicate that represents the similarity between two users, and Rating is the predicate that we are trying to predict, i.e., the rating a user will give a movie. PSL first converts the rule to its disjunctive normal form, (2). Then, all possible substitutions of constants for the variable arguments in the predicates of the rule are made to make ground atoms. Then, all possible combinations of atoms that can form a ground rule are made.

Finally, by utilizing the Lukasiewicz relaxation, the following hinge-loss function is created. For instance, let us assume we have data with users $\texttt{U} = \{Alice, Bob\}$ and movies $\texttt{M} = \{Alien\}$. This will result in the hinge-loss function (3).

We refer the reader to [2] for a more detailed description of PSL.