Appendix of "Nonconvex Optimization for Regression with Fairness Constraints"

A Derivation of the SDP Optimization

For the ease of discussion, we write down Eqn. (3.3) of the main paper in the following:

$$\max_{\xi \geq 0} \ \phi(\xi),$$

where $\phi(\xi)$ is the optimal function defined as

$$\begin{aligned}\phi(\xi) = \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ & \begin{bmatrix}\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix}\mathbf{V}_s & 0 \\ 0 & \mathbf{V}_u\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - 2 \begin{bmatrix}\boldsymbol{q}_s^\top & \boldsymbol{q}_u^\top\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \\ & + \xi \begin{bmatrix}\boldsymbol{\alpha}^\top \boldsymbol{\beta}^\top\end{bmatrix} \begin{bmatrix}(1-\epsilon)\mathbf{V}_s & 0 \\ 0 & -\epsilon\mathbf{V}_u\end{bmatrix} \begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix}.\end{aligned}$$

Let $\mathbf{Q}_{\xi,s}(\xi) := (1 + \xi(1-\epsilon))\mathbf{V}_s$ and $\mathbf{Q}_{\xi,u}(\xi) := (1 - \xi\epsilon)\mathbf{V}_u$. Taking a derivative with respect to each of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ we see that the minimum is achieved when

$$\mathbf{Q}_{\xi,s}(\xi)\boldsymbol{\alpha} = \boldsymbol{q}_s, \ \mathbf{Q}_{\xi,u}(\xi)\boldsymbol{\beta} = \boldsymbol{q}_u.$$

Letting $\mathbf{A} \succeq 0$ denote that $\mathbf{A}$ is positive semidefinite (PSD) and $\dagger$ denotes the pseudo-inverse of a matrix, the optimization is transformed as:

$$\max_{\xi} \ -\boldsymbol{q}_s^\top \mathbf{Q}_{\xi,s}^\dagger(\xi)\boldsymbol{q}_s - \boldsymbol{q}_u^\top \mathbf{Q}_{\xi,u}^\dagger(\xi)\boldsymbol{q}_u \tag{1}$$

$$\text{s.t.} \quad \mathbf{Q}_{\xi,s}(\xi) \succeq 0, \mathbf{Q}_{\xi,u}(\xi) \succeq 0, \xi \geq 0, \tag{2}$$

which is equivalent to

$$\max_{\gamma,\xi} \ \gamma \tag{3}$$

$$\begin{aligned}\text{s.t.} \ & -\gamma - \boldsymbol{q}_s^\top \mathbf{Q}_{\xi,s}^\dagger(\xi)\boldsymbol{q}_s - \boldsymbol{q}_u^\top \mathbf{Q}_{\xi,u}^\dagger(\xi)\boldsymbol{q}_u \geq 0, \\ & \mathbf{Q}_{\xi,s}(\xi) \succeq 0, \mathbf{Q}_{\xi,u}(\xi) \succeq 0, \xi \geq 0.\end{aligned} \tag{4}$$

From Assumption 1 in the main paper, $\mathbf{Q}_{\xi,s}(\lambda)$ and $\mathbf{Q}_{\xi,u}(\lambda)$ are invertible for $\epsilon \in (0,1)$. A standard discussion on the Schur complement of

$$\mathbf{M} = \left[\begin{array}{c|cc} -\gamma & -\boldsymbol{q}_s^\top & -\boldsymbol{q}_u^\top \\ \hline -\boldsymbol{q}_s & \mathbf{Q}_{\xi,s}(\xi) & 0 \\ -\boldsymbol{q}_u & 0 & \mathbf{Q}_{\xi,u}(\xi) \end{array}\right]$$

implies (e.g., Proposition 2.1.(2) in Gallier [Gallier, 2010]) that the constraint in (4) is equivalent to $\mathbf{M} \succeq 0$. Therefore, solving the problem boils down into the equivalent optimization:

$$\max_{\gamma,\xi} \ \gamma \tag{5}$$

$$\begin{aligned}\text{s.t.} \ & \begin{bmatrix}0 & -\boldsymbol{q}_s^\top & -\boldsymbol{q}_u^\top \\ -\boldsymbol{q}_s & \mathbf{V}_s & 0 \\ -\boldsymbol{q}_u & 0 & \mathbf{V}_u\end{bmatrix} - \gamma \begin{bmatrix}1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0\end{bmatrix} \\ & + \xi \begin{bmatrix}0 & 0 & 0 \\ 0 & (1-\epsilon)\mathbf{V}_s & 0 \\ 0 & 0 & -\epsilon\mathbf{V}_u\end{bmatrix} \succeq 0, \ \xi \geq 0,\end{aligned} \tag{6}$$

which is Eqn. (3.4) of the main paper.

Table 1: Comparison of SDP and convex QCQP optimizations. "Runtime (SDP)" and "Runtime (QCQP)" indicate the total running time (measured in seconds) of 100 instances of SDP and QCQP, respectively. The running time is nearly square to the number of features, which is better than the theoretical bound we discussed in Section 3.3 of the main paper. This is not surprising because such a theoretical complexity bound considers the scaling of the hardest instance. We confirmed that in all runs the objective values of the two optimizations were identical within the margin of 0.1%. The simulation here was run on a modern PC server with a Xeon E5-2680 v2 CPU and 198GB memory.

d	Runtime (SDP)	Runtime (QCQP)
10	1.85	1.21
100	12.75	46.85
1000	6149.56	4036.50
3000	103103.18	33859.97

Table 2: Variables selected from the NLSY 79 survey. We used the first two variables as sensitive attributes. The target y is the income of people in 1990 divided by 10,000.

RNUM	Variable Title	Year	Used as
R0000600	AGE OF R	1979	$\boldsymbol{s}$
R0214800	SEX OF R	1979	$\boldsymbol{s}$
H0003400	SF-12 - ASSESSMENT OF R'S GENERAL HEALTH	-	$\boldsymbol{x}$
R0304900	ILLEGAL ACTIVITY 80 INT - TIMES INTENTIONALLY DAMAGED PROPERTY IN PAST YEAR	1980	$\boldsymbol{x}$
R0307100	EVER CHARGED WITH ILLEGAL ACTIVITY? 80 INT (EXC MINOR TRAFFIC OFFENSE)	1980	$\boldsymbol{x}$
R3127300	TYPE OF BUSINESS OR INDUSTRY OF MOST RECENT JOB	1990	$\boldsymbol{x}$
	(80 CENSUS 3 DIGIT) CPS ITEM		
R3146100	ATTENDED VOCATIONAL/TECHNICAL PGM OR ON THE JOB TRAINING SINCE LAST INT?	1990	$\boldsymbol{x}$
R3279401	TOTAL INCOME FROM WAGES AND SALARY IN PAST CALENDAR YEAR (TRUNC) (REVISED)	1990	y
R3403500	NUMBER OF DIFFERENT JOBS EVER REPORTED AS OF INTERVIEW DATE	1990	$\boldsymbol{x}$
R3401501	HIGHEST GRADE COMPLETED AS OF MAY 1 SURVEY YEAR (REVISED)	1990	$\boldsymbol{x}$
R0618300	PROFILES, ARMED FORCES QUALIFICATION TEST (AFQT) PERCENTILE SCORE - REVISED 1989	1989	$\boldsymbol{x}$

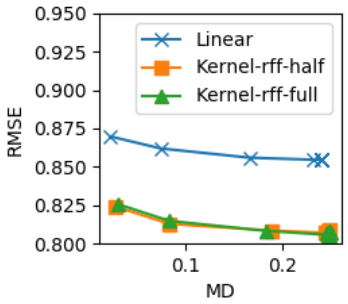
B Comparison of SDP and Convex QCQP Optimizations

This section shows the comparison of the empirical performances of the SDP and the convex QCQP optimizations by using synthetic data. We solved SDPs by using the Mosek optimizer[1] and convex QCQPs by using the Gurobi optimizer[2].

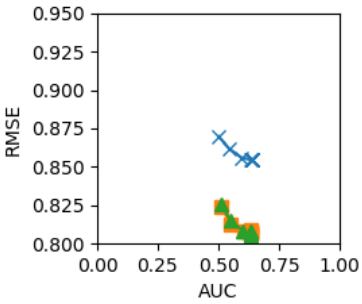
Let $d = d_x + d_x$. We assigned 10% of the features to $\boldsymbol{s}$ (i.e., $d_s = (1/10)d$). The number of datapoints was set to $n = 10d$. Each of features in $\boldsymbol{s}$ and $\boldsymbol{u}$ are drawn from standard normal distribution, and $y = (1, 1, \ldots, 1)s + (1/100, 1/100, \ldots, 1/100)\boldsymbol{u} + \eta$, where η is drawn from the standard normal distribution. We set the strength of the fairness constraint to $\epsilon = 0.1$, which urges the use of both sensitive and non-sensitive features.

Table 1 shows the result of our simulations. While the SQPs ran faster than the convex QCQPs with $d = 100$, the QCQPs ran significantly faster with larger d. As the convex QCQP method showed competitive performance with all size of d, we used the QCQP method for the subsequent simulations in the main paper.

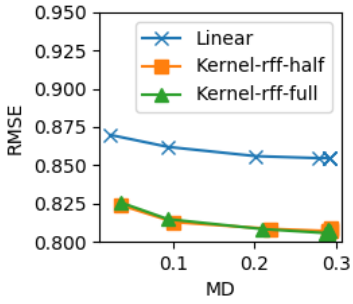
[1] https://www.mosek.com/
[2] http://www.gurobi.com/



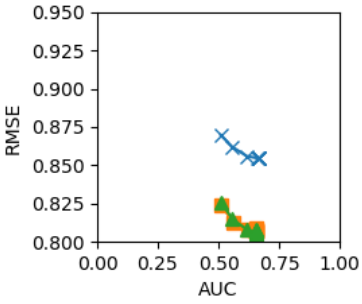
(a) COMPAS (MD Gender-RMSE)



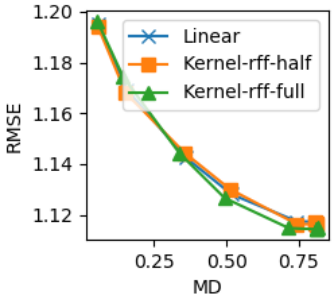
(b) COMPAS (AUC Gender-RMSE)



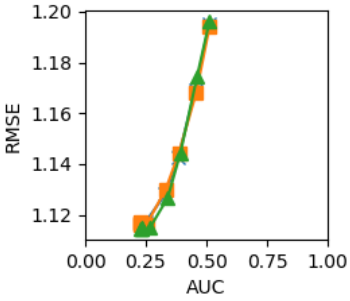
(c) COMPAS (MD Race-RMSE)



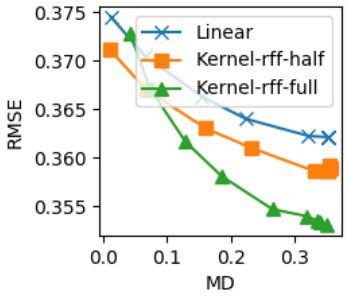
(d) COMPAS (AUC Race-RMSE)



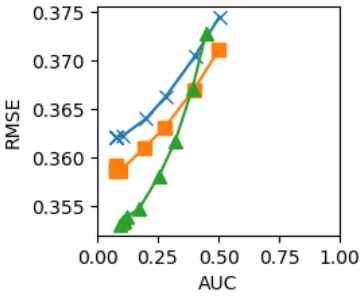
(e) NLSY (MD Gender-RMSE)



(f) NLSY (AUC Gender-RMSE)



(g) LSAC (MD Race-RMSE)

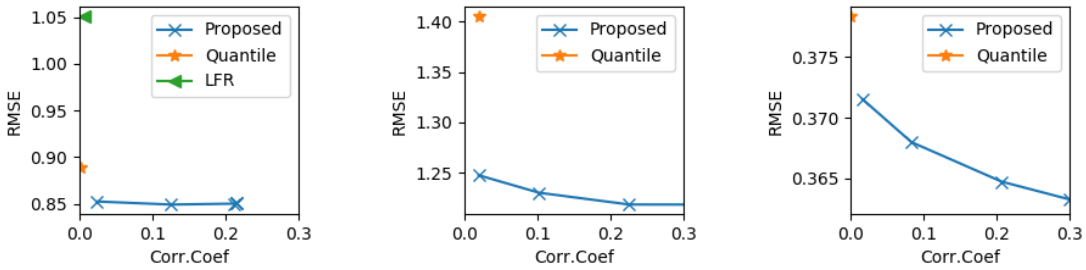


(h) LSAC (AUC Race-RMSE)

Figure 1: RMSE as a function of MD and AUC of the binary sensitive attributes. Only binary sensitive attributes are displayed. Note that the experiment settings were the same as those of Figure 2 in the main paper.

C Details of the Datasets

This section describes the details on the datasets that are not described in the main paper due to the page limitation.



(a) COMPAS (b) NLSY (c) LSAC

Figure 2: RMSE as a function of the correlation coefficient in the three datasets. We chose gender (COMPAS, NLSY), race (LSAC) as the only sensitive attribute. The results are averaged over five random training-test data splittings. "LFR" and "Quantile" are the result of the ridge regressor after the corresponding preprocessing methods [Zemel et al., 2013, Feldman et al., 2015]. Note that the objective of LFR involves the cross-entropy loss and thus LFR does not apply to the NLSY and LSAC datasets where y is numeric. "Proposed" is the (non-kernelized) convex QCQP optimization. The reguralization parameters of all algorithms are the default one ($\lambda = 1.0$).

Data source: We retrieved the COMPAS dataset[3]. Moreover, we retrieved the C&C dataset from the UCI repository[4].

Sensitive attributes: In the COMPAS dataset, we adapted (i) a person's gender and (ii) whether the person is African-American or not as the sensitive attributes. In the LSAC dataset, we adapted (i) whether the person's race is black or not and (ii) age as the sensitive attributes.

C.1 List of Attributes Extracted from the NLSY79 Survey

We build the NLSY dataset by using the NLSY79 investigator tool[5]. The selected variables are shown in Table 2. Note that the categorical features are expanded into dummies, and thus the number of selected variables shown in Table 2 is smaller than $d_s + d_x$.

D Mean Difference and AUC

This section describes the mean difference (MD) and the area under the curve (AUC) that are studied in Calders et al. [Calders et al., 2013] and the empirical results of them. By definition, MD and AUC are only available for a binary sensitive attribute $s^{(l)}$: Let $n_{s=1}$ and $n_{s=0}$ are the number of datapoints where $s^{(l)} = 1$ and $s^{(l)} = 0$, respectively. MD and AUC are defined as

$$\begin{aligned} \mathrm{MD} &= \left| \mathbb{E}_n[\hat{y} | s^{(l)} = 1] - \mathbb{E}_n[\hat{y} | s^{(l)} = 0] \right|, \\ \mathrm{AUC} &= \left| \frac{\sum_{i \in \{1,2,\ldots,n\}: s_i^{(l)}=1} \sum_{j \in \{1,2,\ldots,n\}: s_j^{(l)}=0} \mathbf{1}[\hat{y}_i > \hat{y}_j]}{n_{s=1} \times n_{s=0}} \right|, \end{aligned} \tag{7}$$

where $\mathbb{E}_n$ indicates the sample mean and $\mathbf{1}[x] = 1$ if x is true and 0 otherwise. The larger MD indicates a stronger dependency between $s^{(l)}$ and $\hat{y}$. AUC is 0.5 when $s^{(l)}$ and $\hat{y}$ are independent, and AUC far from 0.5 implies a dependency between $s^{(l)}$ and $\hat{y}$. The MD and AUC of binary attributes on our experiment are shown in Figure 1. In summary, MD and AUC behaved very similarly to the correlation coefficient shown in the main paper.

E Comparison with Preprocessing Methods

To obtain some idea on the accuracy of the preprocessing methods, we compared our optimization (convex QCQP) with existing preprocessing methods (Figure 2). "LFR" is the data-

[3] https://github.com/propublica/compas-analysis
[4] http://archive.ics.uci.edu/ml/datasets/communities+and+crime
[5] The data is publicly available at https://www.nlsinfo.org/investigator/pages/login.jsp.

transformation algorithm proposed in Zemel et al. [Zemel et al., 2013] with their recommended parameters $A_x, A_y, A_z = 0.01, 1, 50$ and $K = 10$. The optimization in LFR is solved by using the l-bfgs global minimizer. "Quantile" is the algorithm proposed in Feldman et al. [Feldman et al., 2015] that merges the two distributions $\boldsymbol{x}|s = 0$ and $\boldsymbol{x}|s = 1$ into a single distribution. One can see that the two preprocessing methods deteriorate the predictive power of features in return for their high level of fairness. Indeed, the LFR did not yield a useful estimator in our environment. Note that the objective of Zemel et al. [Zemel et al., 2013] is nonconvex, and thus the quality of the preprocessing depends on the optimization methods and related parameters.

References

[Calders et al., 2013] Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. (2013). Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 71–80.

[Feldman et al., 2015] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268.

[Gallier, 2010] Gallier, J. (2010). The schur complement and symmetric positive semidefinite (and definite) matrices.

[Zemel et al., 2013] Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 325–333.