Optimization under Fairess Constraints

Jing Gong
Department of Mechanical Engineering
Purdue University
gong106@purdue.edu

Abstract

To solve the problem of discrimination or unfairness treatment in decision making process, this project introduce a new method called fairness constrained optimal decisions. We apply our model in different data type domain with different machine learning algorithms. We also explore the varieties of applications in industry with different utility models. We also analyze the performance of our models with different causal effect constraints which shows a trade-off between causal effect and prediction accuracy or maximized utility.

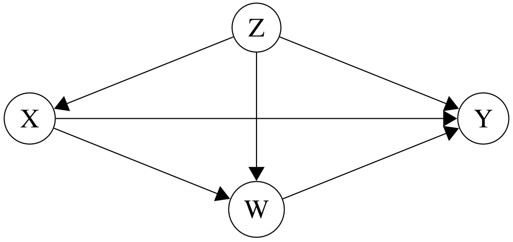
1 Introduction

Algorithms from decision making and machine learning are increasingly playing important roles on people's lives, including policing, pricing, hiring, criminal sentencing and so on. However, discrimination may happen during the decision making process. This happens when sensitive attributes information has disparate impacts to the outcomes. Such sensitive attributes usually include gender, religion, physical ability and so on. These are referred as the protected attributes in Federal Laws and regulations. Take criminal sentencing as an example, blacks were more than twice as likely as whites to be labeled risky even though they didn't continue to commit a crime.

In this project, we focus on the problem of fairness in decision making process and machine learning algorithms, and we apply the causal fairness principles to balance the outcome with respect to sensitive attributes to decision variables.

Our contribution in this project is we apply causal effect measures as constraints in decision making process. And we solve the utility maximization problem in discrete domain and continuous domain where relationship between variables is linear. Another contribution is that we also apply the constraints on classification and regression algorithms. For the classification, we apply the non convex constraints and optimize the model over discrete domain. For the regression, we apply convex optimization over continuous domain.

The domain is restricted to a general causal structural model below:



which X stands for the protected attribute, Y for the decision variable, W the mediator and Z an observed background variable.

2 Optimization over Discrete Domain

Assumption:

1. The distribution is in binary domain.

2. The distribution of the states is known.

3. Specified utility function.

4. Action or decision is going to be taken only when the maximize utility is achieved. Higher value of utility corresponds to greater likelihood of Y=1.

e.g. In a hiring practice, we are going to take an action of whether or not to hire the person. We have a utility function that evaluates the person's ability in work and the utility function is known. And there is a optimal action of hiring or not hiring when the maximized utility value is achieved. Decision Formulation:

State: X, W, Z

Decision variable: Y, if Y=1, take action or make decision. e.g. hiring the person or releasing the criminal.

Utility function: U = f(x, w, z, y). Decision is made when maximized utility is achieved under certain state.

Objective: Maximize the expected utility

$$\underset{Y}{maximize} \quad \mathbb{E}[U]$$
$$= maximize \sum_{y,x,w,z} P(x,w,z,y)U(x,w,z,y)$$
$$= maximize \sum_{x,w,z,y} P(Y=y|x,w,z)P(x,w,z)U(x,w,z,y)$$

Constraints: causal effect with respect to sensitive attribute X and action Y.(DE, IE, SE)

$$DE = \sum_{z,w}(P(Y|x_1,w,z) - P(Y|x_0,w,z))P(w|x_0,z)P(z|x_0)$$
$$IE = \sum_{z,w} P(Y|x_0,w,z)(P(w|x_1,z) - P(w|x_0,z))P(z|x_1)$$
$$SE = \sum_{z,w} P(Y|x_0,w,z)P(w|x_0,z)(P(z|x_1) - P(z|x_0))$$

Formulation of linear programming:

Let $p_{i,jkl}$ denotes $P(Y = i|X = j, W = k, Z = l)$, e.g. $p_{1,000}$ represents $P(Y=1|X=0,W=0,Z=0)$.

The whole problem can be rewritten as:

$$\underset{p_{i,jkl}}{maximize} \quad \textstyle\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1}\sum_{l=0}^{1} p_{i,jkl}P_{x_j w_k z_l}U(x_j,w_k,z_l,y_i) \tag{1}$$

s. t. $\sum_{i=0}^{1} p_{i,jkl} = 1, \forall j,k,l$

$0 \leq p_{i,jkl} \leq 1, \forall i,j,k,l$

$|\sum_{k=0}^{1}\sum_{l=0}^{1}(p_{1,1kl} - p_{1,0kl})P_{w_k|x_0 z_l}P_{z_l|x_0}| \leq c_1$, (constrained on DE)

In the above equation, the big P denotes the conditional distribution regrading X, Z, W which is assumed to be given. The small p denotes the conditional distribution regrading Y which is the objective.

Constraints can also be set in the following ways:

Constrained on DE+IE:

$$|\textstyle\sum_{k=0}^{1}\sum_{l=0}^{1}(p_{1,1kl} - p_{1,0kl})P_{w_k|x_0 z_l}P_{z_l|x_0} + \sum_{k=0}^{1}\sum_{l=0}^{1} p_{1,0kl}(P_{w_k|x_1 z_l} - P_{w_k|x_0 z_l})P_{z_l|x_1}| \leq c_2$$

Constrained on total variation:

$$|\textstyle\sum_{k=0}^{1}\sum_{l=0}^{1}(p_{1,1kl} - p_{1,0kl})P_{w_k|x_0 z_l}P_{z_l|x_0} + \sum_{k=0}^{1}\sum_{l=0}^{1} p_{1,0kl}(P_{w_k|x_1 z_l} - P_{w_k|x_0 z_l})P_{z_l|x_1} +$$

$$\sum_{k=0}^{1}\sum_{l=0}^{1} p_{1,0kl}P_{w_k|x_0 z_l}(P_{z_l|x_1} - P_{z_l|x_0})| \leq c_3$$

The optimization problem (1) is a linear program in variables respect to CPDs conditioning on states: (1)The objective function is a linear function of these parameters. (2)The constraints are linear or convex which specify the optimal solution in a convex hull.

And we can say that the value of Y can be defined when the conditional probability of y given X, Z,W is greater than some threshold.

3 Optimization over Continuous Domain

Assumption:

1. Suppose the causal relationship is linear between variables.

2. Decision variables are continuous random variables, and take any value between the lower bound and upper bound. e.g. $d \in (\underline{d}, \overline{d})$

3. Decision is made when utility maximization is achieved. e.g. the maximized reward or revenue.

4. Utility function is deterministic and known.

5. The utility function is continuous and differentiable.

3.1 Linear Utility

The constraints are linear:

$$\begin{aligned} &\text{DE: } DE_{x_0,x_1}(Y) = \beta_{xy}, \\ &\text{IE: } IE_{x_0,x_1}(Y) = \beta_{wy}\beta_{xw}, \\ &\text{SE: } SE_{x_0,x_1}(Y) = (\beta_{zy} + \beta_{zw}\beta_{wy})\beta_{xz}, \end{aligned} \tag{2}$$

where β represent the corresponding regression coefficient.

Hence, Y can be represented by linear regression: $Y = \beta_{xy}X + \beta_{wy}W + \beta_{zy}Z + \beta_0$.

$\beta = [\beta_{xy}, \beta_{wy}, \beta_{zy}, \beta_0]^T$

Objective function: $\underset{Y}{maximize} \quad \mathbb{E}[U(Y)] = \underset{\beta}{maximize} \quad \mathbb{E}[U(\beta)]$

The whole problem is a linear program:

$$\underset{\beta}{maximize} \quad \mathbb{E}[U(\beta)] \tag{3}$$

$$\text{s. t. } \underline{y} \leq A^T\beta \leq \overline{y}$$

$$|\beta_{xy}| \leq c_1 \qquad \text{(constrained on DE)}$$

Simplex method can be applied to solve this problem easily.

3.2 Nonlinear Utility

3.2.1 Concave Utility

Since the utility function is a differentiable convex function in the feasible region and the constrained region is a convex hull, then the KKT optimality conditions are sufficient for the global optimality of a feasible solution.

3.2.2 Non-concave Utility

Generally, arbitrary nonconcave or nonconvex optimization problems are not easy to solve. However, since we've restricted the domain in a linear causal model, the whole problem becomes into linear constrained nonlinear optimization problem.

Let $|A^T\beta| \leq c$ denotes the linear fairness constraints. It could be any combination of DE, SE and IE. The constraints are all continuous and differentiable over R^n.

Assume the utility function is determinstic. Hence, the general problem is:

$$\underset{\beta}{maximize} \quad f(\beta), s.t.|A^T\beta| \leq c \tag{4}$$

Let g(β) denotes the constrain as: g(β) ≤ 0, g(β) is continuous over β.

Sometimes the global optimal solution can be found if the nonconcave utility function has special structures:

1. Monotonic Utility

Monotonicity optimization problems are a very important part in economic and communication networks.

For linear constraints: g($\beta_1, ...\beta_n$) is monotone in β_i, for $\forall$ i=1, 2, ... n

According to the monotonicity principles, we can have the following result:

Assume β has lower bound and upper bound, $\beta \in (\beta^L, \beta^U)$

Assume f is monotonic utility function with respect to β.

i) Suppose the constraints $g(\beta)$ are monotone increasing in β, then there exists a global optimal solution of problem (4) if at least one variable β_k in f in not monotone increasing.

Proof The proof is very straightforward. Let $\beta^* = (\beta_1^*, \beta_2^*, ..., \beta_n^*)$ denote a global optimum of problem (4). If f is not monotone increasing in β_k, obtaining the value β_k such that function $g(\beta_1^*, \beta_2^*, ..., \beta_n^*)$ becomes equal to zero. And this is always possible since β has lower bound.

ii) More generally, suppose the constraints $g(\beta)$ are monotone increasing in β_i, monotone decreasing in β_j, and independent of β_k. If f has the different monotonicity with respect to β_i and β_j, and is monotone increasing or decreasing in β_k, then the constraints are tight and there exists a global optimal solution of problem (4).

iii)One special case is the problem to maximize an increasing function subject to an increasing constraint with nonnegative coordinates:

$$max\{f(\beta)|x \in G\},$$

where G is a compact normal set $G := \{\beta \in R_+^n | g(\beta) \leq 0\}$

Suppose $g(\beta)$ is increasing, then a normal set G can be approximated by ployblocks.

2. Generic Utility

If the function is generic, there are several methods that can be applied:

1. Linearly constrained Lagrangian Methods: We assume the utility function is differentiable. LCL methods can solve a sequence of subproblems that maximize an augmented Lagrangian function subject to linear inequality constraints. But sometimes it may not converge from arbitrary starting points. And the optimal solution might not be global optimal.

2. Genetic Algorithm: Genetic algorithm can approximate the global optimal point. It depends on the specific problems. Some problems might be effectively solved by genetic algorithm. Sometimes other optimization algorithms may be more efficient that genetic algorithm.

4 Fairness in Machine Learning Algorithm

4.1 Fairness in Classification

Apply the causal fairness measures in logistic regression model:

$$p(y_i = 1|x_i, \theta) = \frac{1}{1+e^{-\theta^T x_i}}$$

Consider the direct causal effect:

$$DE = (\frac{1}{1+e^{-\theta_x-\theta_z-\theta_w}} - \frac{1}{1+e^{-\theta_z-\theta_w}})P_{w_1|x_0z_1}P_{z_1|x_0} + (\frac{1}{1+e^{-\theta_x-\theta_z}} - \frac{1}{1+e^{-\theta_z}})P_{w_0|x_0z_1}P_{z_1|x_0} + \\ (\frac{1}{1+e^{-\theta_x-\theta_w}} - \frac{1}{1+e^{-\theta_w}})P_{w_1|x_0z_0}P_{z_0|x_0} + (\frac{1}{1+e^{-\theta_x}} - \frac{1}{2})P_{w_0|x_0z_0}P_{z_0|x_0}$$

Rewrite the expression in a simpler way:

$$DE = k_1(\frac{1}{1+e^{-\theta_x-\theta_z-\theta_w}} - \frac{1}{1+e^{-\theta_z-\theta_w}}) + k_2(\frac{1}{1+e^{-\theta_x-\theta_z}} - \frac{1}{1+e^{-\theta_z}}) + \\ k_3(\frac{k_3}{1+e^{-\theta_x-\theta_w}} - \frac{1}{1+e^{-\theta_w}}) + k_4(\frac{1}{1+e^{-\theta_x}} - \frac{1}{2})$$

The problem becomes into:

$$\underset{\theta}{minimize} - \sum_{i=1}^{N} \; logp(y_i = 1|x_i, \theta) \tag{5}$$

$$\text{subject to } |k_1(\frac{1}{1+e^{-\theta_x-\theta_z-\theta_w}} - \frac{1}{1+e^{-\theta_z-\theta_w}}) + k_2(\frac{1}{1+e^{-\theta_x-\theta_z}} - \frac{1}{1+e^{-\theta_z}}) + \\ k_3(\frac{k_3}{1+e^{-\theta_x-\theta_w}} - \frac{1}{1+e^{-\theta_w}}) + k_4(\frac{1}{1+e^{-\theta_x}} - \frac{1}{2})| \leq c_1, \text{(constrained on DE)}$$

The constraints are not linear and not convex. The problem is a convex objective function with nonlinear constraints.

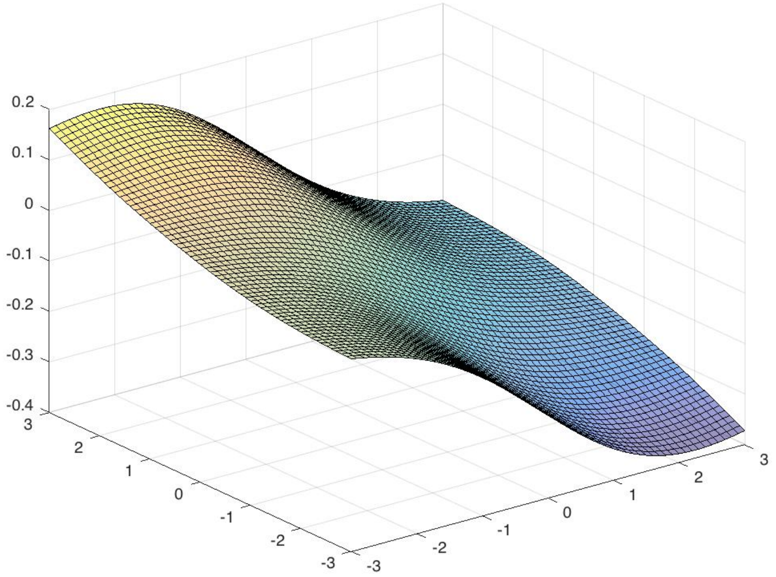


Figure 1: non-linear constraint

From the above figure, we can see that the constraints are partly concave, partly convex and might be linear in some regions.

Since the loss function and constraints are all differentiable, we can apply several methods to locate to optimal solution:

1. Lagrangian Methods: Lagrange multipliers may still be applied as long as the algorithm converge.

2. Piecewise linearity: approximate the constraints to a convex set of piecewise linear expression. Then the problem is very easy to solve.

3. Genetic algorithm: suppose θ is bounded, genetic algorithm can effectively find the global optimal solution.

We apply genetic algorithm to optimize constrained logistic regression model,Data are generated from the causal model with Bernoulli distribution. Below is the optimization results(constrained on

DE):

	DE	Accuracy
Unconstrained	0.1874	0.805
Constrained(c=0.1)	0.010	0.771
Constrained(c=0.05)	0.010	0.764
Constrained(c=0.001)	0.0090	0.455

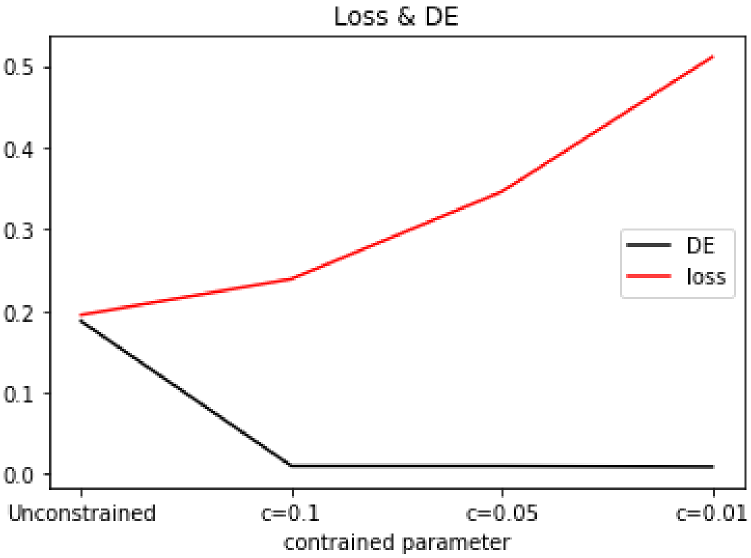


Figure 2: Logistic regression with constrained DE

We can also consider the decision boundary as expression for constraints. For linear classifier, the decision boundary equation can be reduced to $\theta^T x = 0$. In this case, the constraints are all linear. Then it's a convex optimization problem:
The objective is:

$$\underset{\theta}{minimize} - \textstyle\sum_{i=1}^{N} \;\; logp(y_i = 1|x_i, \theta) \tag{6}$$

subject to$|A^T\theta| \leq c$(constrained on total variation)

A: is a matrix with numbers of coefficients of xw, xz or zw.

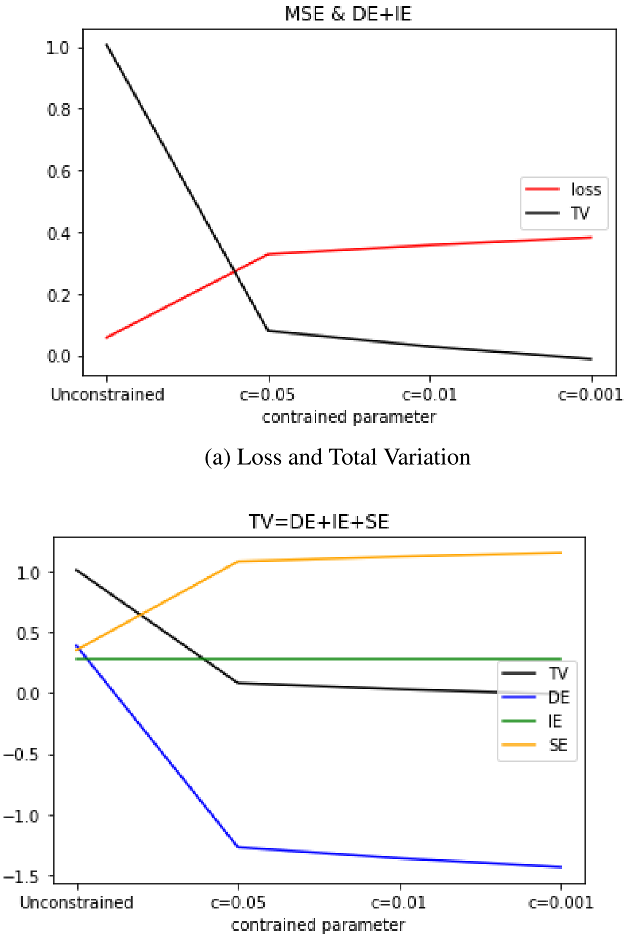
Lagrange multipliers can be employed here:

$$L(\theta, \lambda) = \;\; - \textstyle\sum_{i=1}^{n}((y_i\theta x) - log(1 + e^{\theta x})) + \lambda_1(A^T\theta - c) + \lambda_2(-A^T\theta - c)$$

Experiments: generate random data based on the causal model given before. u_x, u_z, u_w, u_y are Bernoulli distributed with random parameter p and the f_x, f_z, f_w,f_y are defined according to the causal model we just referred at the beginning.

Below is the constrained result for DE, IE, SE and total variation.

	DE	IE	SE	TV	Loss
Unconstrained	0.3849	0.2714	0.3508	1.0071	0.0574
Constrained(c=0.1)	-1.2702	0.2714	1.0780	0.0792	0.3278
Constrained(c=0.05)	-1.3607	0.2714	1.1178	0.0285	0.3574
Constrained(c=0.01)	-1.4330	0.2714	1.1495	-0.0121	0.3823



(a) Loss and Total Variation

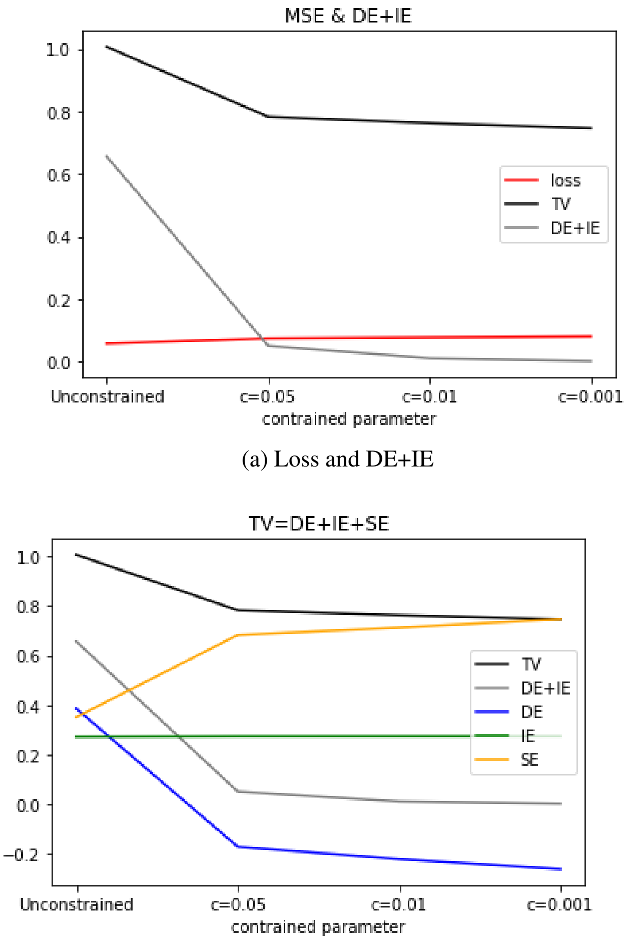
(b) DE, IE, SE, TV

Figure 3: logistic regression constrained on total variation

From the plot we can see that as the total variation decreases, the loss increases. DE and SE increase to the inverse effect.

Constrained on DE+IE:

	DE	IE	SE	TV	DE+IE	Loss
Unconstrained	0.3849	0.2714	0.3508	1.0071	0.6563	0.0574
Constrained(c=0.1)	-0.1731	0.2731	0.6829	0.7829	0.100	0.0735
Constrained(c=0.05)	-0.2232	0.2732	0.7127	0.7627	0.050	0.0768
Constrained(c=0.01)	-0.2634	0.2734	0.7466	0.7466	0.009	0.0797



(a) Loss and DE+IE

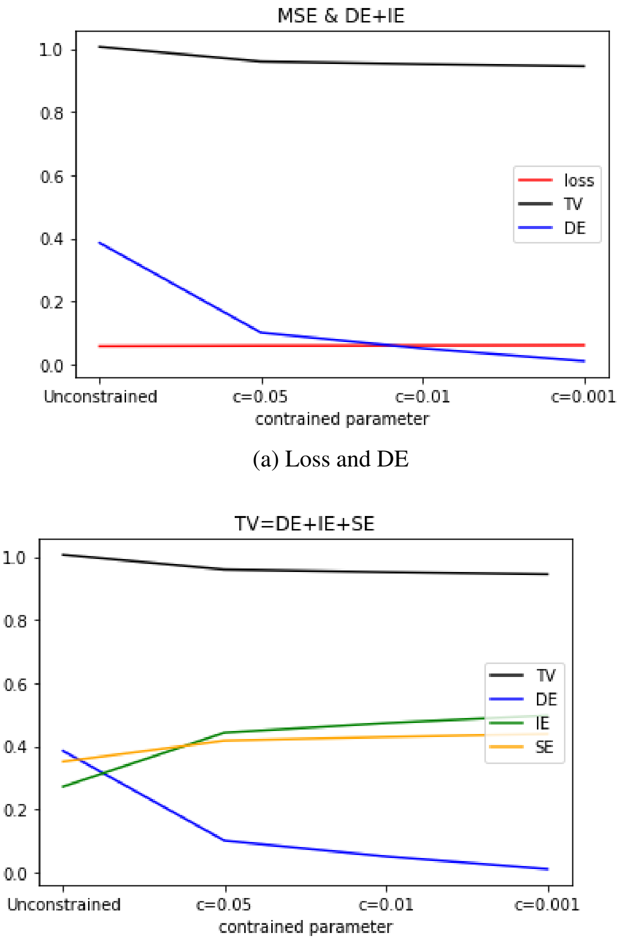
(b) DE, IE, SE, TV

Figure 4: logistic regression constrained on DE+IE

From the above result, we can see that when constrained on DE+IE, the loss has a slightly increase.DE and total variation drops down. SE increases while IE keeps almost the same.

Constrained on DE:

	DE	IE	SE	TV	Loss
Unconstrained	0.3849	0.2714	0.3508	1.0071	0.0574
Constrained(c=0.1)	0.100	0.4429	0.4174	0.9603	0.0588
Constrained(c=0.05)	0.050	0.4730	0.4291	0.9521	0.0595
Constrained(c=0.01)	0.0100	0.4971	0.4384	0.9455	0.0602



(a) Loss and DE

(b) DE, IE, SE, TV

Figure 5: logistic regression constrained on DE

When constrained on DE, there is a trade-off between DE and loss. The total variation has a sightly decrease while IE and SE are all increasing.

4.2 Fairness in Regression

We apply the causal fairness measures in linear regression models:

Causal constraints:

$$\begin{aligned}&\text{DE: } DE_{x_0,x_1}(Y) = \beta_{xy},\\&\text{IE: } IE_{x_0,x_1}(Y) = \beta_{wy}\beta_{xw},\\&\text{SE: } SE_{x_0,x_1}(Y) = (\beta_{zy} + \beta_{zw}\beta_{wy})\beta_{xz},\end{aligned} \tag{2}$$

Since we only care about the coefficients with respect to y, other coefficients can be considered as constants.

which is : $|k_1\beta_{xy} + k_2\beta_{zy} + k_3\beta_{wy}| < c$(constrained on total variation)

The boundary is convex with respect to coefficients β

4. The whole problem can be written as:

$$\underset{\beta}{minimize} \;\; \sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

$$\text{subject to } k^T\beta \leq c, -k^T\beta \leq c$$

The constraints can be written as Lagrange multipliers:

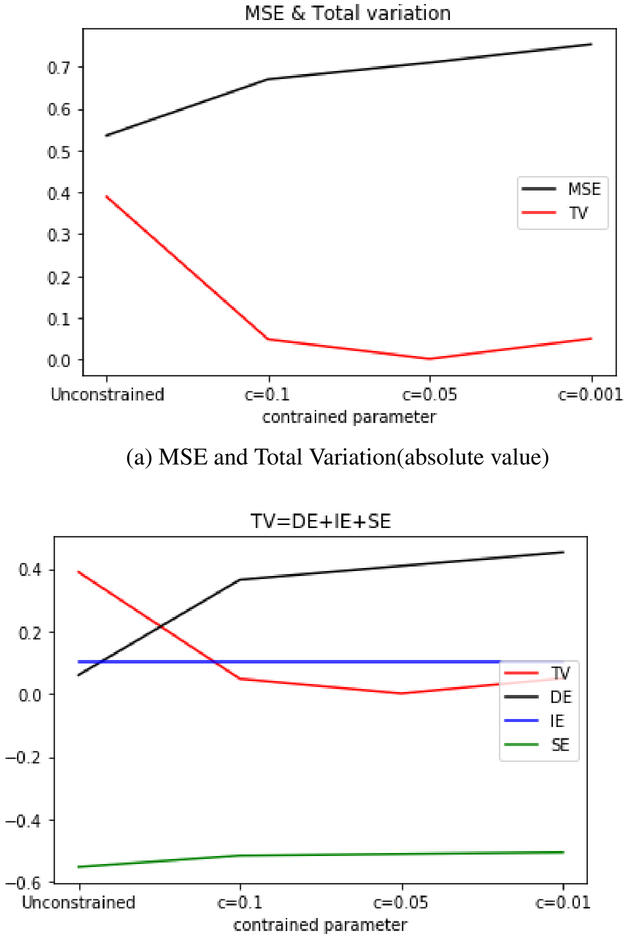
$$\underset{\beta}{minimize} \;\; \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda_1(k^T\beta - c) + \lambda_2(-k^T\beta - c)$$

5.Experiments:

Generate random data from standard normal distribution: u_x, u_y, u_z, u_w are normally distributed. And also randomly select structural coefficients α, θ, γ from standard normal distribution.

Below is the discrimination after optimization compared to the discrimination measures without optimization:

	DE	IE	SE	TV	MSE
Unconstrained	0.0600	0.1035	-0.5528	-0.3893	0.5358
Constrained(c=0.1)	0.3646	0.1035	-0.5164	-0.0483	0.6703
Constrained(c=0.05)	0.4088	0.1035	-0.5111	0.0012	0.7100
Constrained(c=0.001)	0.4522	0.1035	-0.5060	0.0498	0.7538



(a) MSE and Total Variation(absolute value)

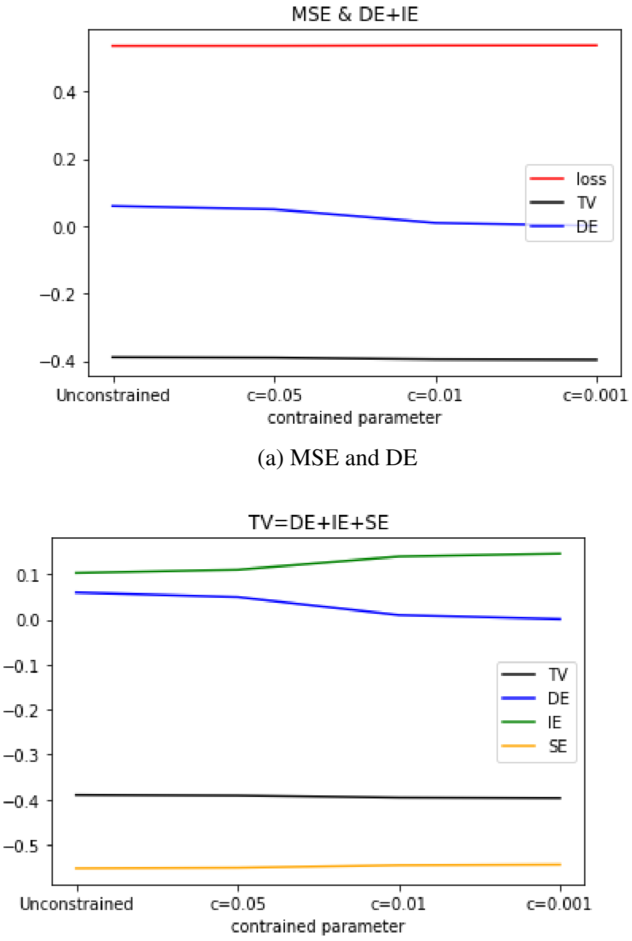
(b) DE, IE, SE, TV

Figure 6: linear regression constrained on total variation

As we can see from the table, fairness constraints control the total variation by balancing the DE and SE while keeping IE almost the same. And the mean squared error is larger than the unconstrained case.

If we constrain on DE or DE+IE, below is the discrimination after optimization:

	DE	IE	SE	TV	MSE
Unconstrained	0.0600	0.1035	-0.5528	-0.3893	0.5358
Constrained(c=0.05)	0.0500	0.1108	-0.5514	-0.3906	0.5360
Constrained(c=0.01)	0.0100	0.1400	-0.5455	-0.3955	0.5371
Constrained(c=0.001)	9.9985e-04	0.1465	-0.5442	-0.3967	0.5374



(a) MSE and DE

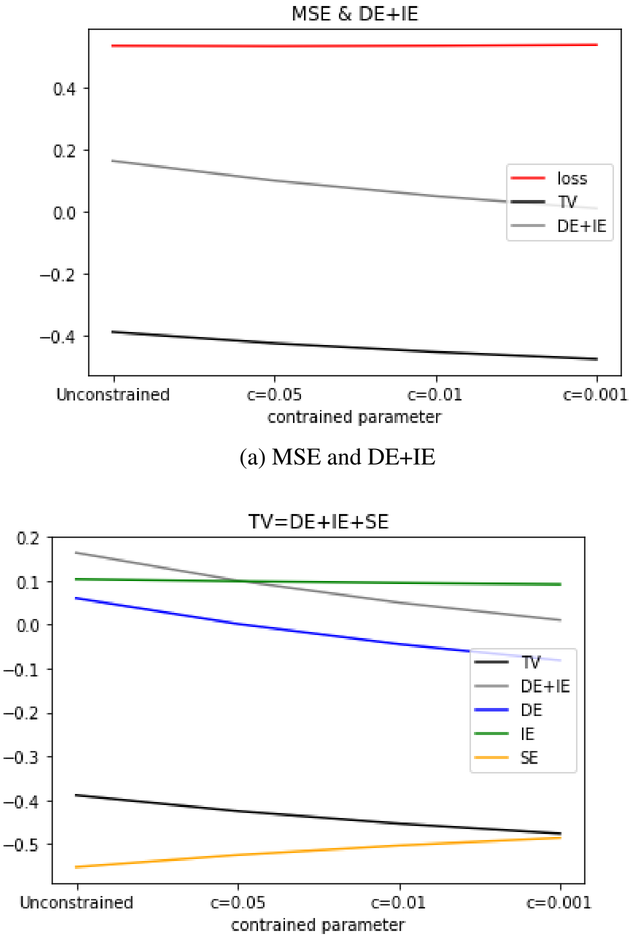
(b) DE, IE, SE, TV

Figure 7: linear regression constrained on DE

When constrained on DE, the mean squared error seems to be the same. This because in this case the unconstrained direct counterfactual effect is small. The change of DE is very small to the whole model.

Constrained on DE+IE:

	DE	IE	DE+IE	SE	TV	MSE
Unconstrained	0.0600	0.1035	0.1635	-0.5528	-0.3893	0.5358
Constrained(c=0.1)	0.0014	0.0986	0.100	-0.5253	-0.4253	0.5345
Constrained(c=0.05)	-0.0447	0.0947	0.050	-0.5037	-0.4537	0.5360
Constrained(c=0.01)	-0.0816	0.0916	0.0100	-0.4864	-0.4764	0.5389



(a) MSE and DE+IE

(b) DE, IE, SE, TV

Figure 8: linear regression constrained on DE+IE

When constrained on DE+IE, DE decreases and then has reverse effect and IE keeps almost the same level.

5 Conclusion

In conclusion, our work firstly apply causal effect measures as constraints in decision making process. Binary domain is to maximize the likelihood with linear programming. In continuous domain, linear and convex utility maximization can be solved. Problems with special structural utilities can also be located to the global optimal solution.

The future goal for this project is to apply the causal constraints to high dimensional data where things may be different. And also in nonlinear models the constraints are not linear with respect to regression coefficients. In addition, in many cases, the utility function is not well-defined or not deterministic.

6 Reference

[1] J. Zhang, E. Bareinboim. Fairness in Decision-Making – The Causal Explanation Formula. AAAI-18. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018, forthcoming. Purdue AI Lab, Technical Report (R-30), Nov, 2017.
[2] S. C. Davies, E. Pierson, A. Feller. Algorithmic Decision Making and the Cost of Fairness, 2017
[3] M. B. Zafar, I. Valera, M. G. Rodriguez, K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification, 2017
[4] M. Hardt, E. Price, N. Srebro. Equality of Opportunity in Supervised Learning, 2016
[5] P. Hansen, B. Jaumard, S. H. Lu. Some Further Results on Monotonicity in Globally Optimal Design, 1989.
[6] Alexander. Rubinov , Hoang. Tuy & Heather. Mays (2001) An alogrithm for monotonic global optimization problems , Optimization, 49:3, 205-221, DOI: 10.1080/02331930108844530