Loss Balancing for Fair Supervised Learning

Choeun Kim

February 21, 2024

Seoul National University

Introduction

- Focused on EL (Equalized Loss)
- Problem : Imposing EL on the learning process leads to a non-convex optimization problem even if the loss function is convex

- Developed an algorithm with a theoretical performance guarantee for EL fairness.
- also develop a simple algorithm for finding a sub-optimal predictor satisfying EL fairness

Problem Formulation

- Notation

$$(\boldsymbol{X}, A, Y) : \textit{training dataset from two social groups}$$

$$\boldsymbol{X} \in \mathcal{X} : \textit{feature vectore}, \quad A \in \{0, 1\} : \textit{sensitive attribute}$$

$$Y \in \mathcal{Y} \subseteq \mathbb{R} : \textit{label or output}$$

$$\mathcal{F} : \textit{set of predictors } f_w : \mathcal{X} \to \mathbb{R}$$

$$l : \mathcal{Y} \times \mathbb{R} \to \mathbb{R} \textit{ loss function}$$

- Expected loss

$$L(w) := \mathbb{E}\{l(Y, f_w(\mathbb{X}))\} \; w.r.t \; (\boldsymbol{X}, Y)$$

$$L_a(w) := \mathbb{E}\{l(Y, f_w(\mathbb{X}))|A = a\}$$

Problem Formulation

- Assume that $l(y, f_w(\boldsymbol{x}))$ is differentiable and strictly convex in w

Definition

We say f_w satisfies the equalized loss(EL) fairness notion if $L_0(w) = L_1(w)$. Moreover, we say f_w satisfies γ-EL for some $\gamma > 0$ if $-\gamma \leq L_0(w) - L_1(w) \leq \gamma$.

- If $l(Y, f_w(\boldsymbol{X}))$ is convex in w, then both $L_0(w)$ and $L_1(w)$ are also convex in w. However, $L_0(w) - L_1(w)$ is not necessary convex.
- Therefore, the following optimization problem for finding a fair predictor under γ-EL is ***not a convex*** programming,

$$min_w L(w) \ s.t. \ -\gamma \leq L_0(w) - L_1(w) \leq \gamma \tag{1}$$

Optimal Model under γ-EL

- **Assumption 1.** Expected losses $L_0(w), L_1(w)$ *and* $L(w)$ are strictly convex and differentiable in w. Moreover, each of them has a unique minimizer.

$$\boldsymbol{w}_{G_a} = arg \min_{w} L_a(w)$$

  Since it is unconstrained, w_{G_a} can be found efficiently by common convex solvers.

- **Assumption 2.** We assume the following holds,

$$L_0(w_{G_0}) \leq L_1(w_{G_0}) \text{ and } L_1(w_{G_1}) \leq L_0(w_{G_1})$$

Optimal Model under γ-EL

- Under assumptions, the optimal 0-EL fair predictor can be easily found using $\texttt{ELminimizer}(w_{G_0}, w_{G_1}, \epsilon, \gamma)$ *with* $\gamma = 0$.

Algorithm 1 Function `ELminimizer`

Input: $\boldsymbol{w}_{G_0}, \boldsymbol{w}_{G_1}, \epsilon, \gamma$

Parameters: $\lambda^{(0)}_{start} = L_0(\boldsymbol{w}_{G_0}), \lambda^{(0)}_{end} = L_0(\boldsymbol{w}_{G_1}), i = 0$

Define $\tilde{L}_1(\boldsymbol{w}) = L_1(\boldsymbol{w}) + \gamma$

while $\lambda^{(i)}_{end} - \lambda^{(i)}_{start} > \epsilon$ **do**

  $\lambda^{(i)}_{mid} = (\lambda^{(i)}_{end} + \lambda^{(i)}_{start})/2$

  Solve the following convex optimization problem,

$$\boldsymbol{w}_i^* = \arg\min_{\boldsymbol{w}} \tilde{L}_1(\boldsymbol{w}) \text{ s.t. } L_0(\boldsymbol{w}) \le \lambda^{(i)}_{mid} \quad (4)$$

  $\lambda^{(i)} = \tilde{L}_1(\boldsymbol{w}_i^*)$

  if $\lambda^{(i)} \ge \lambda^{(i)}_{mid}$ **then**

    $\lambda^{(i+1)}_{start} = \lambda^{(i)}_{mid}$; $\lambda^{(i+1)}_{end} = \lambda^{(i)}_{end}$;

  else

    $\lambda^{(i+1)}_{end} = \lambda^{(i)}_{mid}$; $\lambda^{(i+1)}_{start} = \lambda^{(i)}_{start}$;

    $i = i + 1$;

  end if

end while

Output: $\boldsymbol{w}_i^*$

Parameter $\epsilon > 0$ specifies the stopping criterion.

Optimal Model under γ-EL

Theorem

Let $\{\lambda_{mid}^{(i)} | i = 0, 1, 2, \ldots\}$ *and* $\{w_i^* | i = 0, 1, 2, \ldots\}$ *be two sequences generated by* `ELminimizer` *when* $\gamma = \epsilon = 0$*, i.e.,* `ELminimizer`$(w_{G_0}, w_{G_1}, 0, 0)$. *Under Assumptions, we have,*

$$\lim_{i \to \infty} w_i^* = w^* \text{ and } \lim_{i \to \infty} \lambda_{mid}^{(i)} = \mathbb{E}\{l(Y, f_{w^*}(X))\}$$

where w^* *is the global optimal solution to (1).*

The theorem implies that when $\gamma = \epsilon = 0$ and i goes to infinity, the solution to convex problem (4) is the same as the global optimal solution under EL constraint.

Optimal Model under γ-EL

Algorithm 2 Solving Optimization (1)

Input: $\boldsymbol{w}_{G_0}, \boldsymbol{w}_{G_1}, \epsilon, \gamma$

$\boldsymbol{w}_{\gamma} = \texttt{ELminimizer}(\boldsymbol{w}_{G_0}, \boldsymbol{w}_{G_1}, \epsilon, \gamma)$

$\boldsymbol{w}_{-\gamma} = \texttt{ELminimizer}(\boldsymbol{w}_{G_0}, \boldsymbol{w}_{G_1}, \epsilon, -\gamma)$

if $L(\boldsymbol{w}_{\gamma}) \leq L(\boldsymbol{w}_{-\gamma})$ **then**

  $\boldsymbol{w}^* = \boldsymbol{w}_{\gamma}$

else

  $\boldsymbol{w}^* = \boldsymbol{w}_{-\gamma}$

end if

Output: w^*

Theorem

Assume that $L_0(w_{G_0}) - L_1(w_{G_0}) < -\gamma$ and $L_0(w_{G_1}) - L_1(w_{G_1}) > \gamma$. If w_O does not satisfy the γ-EL constraint, then, as $\epsilon \to 0$, the output of Algorithm 2 goes to the optimal γ-EL fair solution (i.e., solution to (1)).

Optimal Model under γ-EL

- Complexity Analysis

  If the time complexity of solving (4) is $\mathcal{O}(p(d_w))$, then the overall time complexity of Algorithm 1 is $\mathcal{O}(p(d_w)log(1/\epsilon))$.

- Regularization

  Consider a supervised learning model with regularization.

$$\begin{aligned} \min_w Pr(A=0)L_0(w) + Pr(A=1)L_1(w) + R(w) \\ s.t.,\ |L_0(w) - L_1(w)| < \gamma \end{aligned} \tag{2}$$

  We can re-write (2) as follows,

$$\begin{aligned} \min_w Pr(A=0)(L_0(w) + R(w)) + Pr(A=1)(L_1(w) + R(w)), \\ s.t.,\ |(L_0(w) + R(w)) - (L_1(w) + R(w))| < \gamma \end{aligned}$$

Sub-optimal Model under γ-EL

- `ELminimizer` still requires solving a convex constrained optimization in each iteration.
- In this section, we propose another algorithm that finds a sub-optimal solution to optimization (1) **without solving constrained optimization** in each iteration.
- The algorithm consists of two phases.

  Phase 1. Find two weight vectors by solving two unconstrained convex optimization problems

  Phase 2. Generate a new weight vector satisfying γ-EL using the two weight vectors found in the first phase.

Sub-optimal Model under γ-EL

- **Phase 1. Unconstrained optimization**

$$w_O = arg \min_{w} L(w)$$
$$\hat{a} = arg \max_{a \in \{0,1\}} L_a(w_O)$$
$$w_{G_{\hat{a}}} = arg \min_{w} L_{\hat{a}}(w)$$

  - Since $L(w)$ is strictly convex in w, the above can be solved efficiently.
  - $\hat{a}$ is a disadvantaged under predictor f_{w_O}.

Sub-optimal Model under γ-EL

- **Phase 2. Binary search to find the fair predictor**

$$g(\beta) := L_{\hat{a}}((1-\beta)w_O + \beta w_{G_{\hat{a}}}) - L_{1-\hat{a}}((1-\beta)w_O + \beta w_{G_{\hat{a}}})$$

$$h(\beta) := L((1-\beta)w_O + \beta w_{G_{\hat{a}}})$$

Theorem

Under Assumption 1 and 2,

1. *There exists $\beta_0 \in [0,1]$ such that $g(\beta_0) = 0$*
2. *$h(\beta)$ is strictly increasing in $\beta \in [0,1]$*
3. *$g(\beta)$ is strictly decreasing in $\beta \in [0,1]$*

- If we start from w_O and move toward $w_{G_{\hat{a}}}$ along a straight line, the overall loss increases and the disparity between two groups decreases until we reach $(1-\beta_0)w_O + \beta_0 w_{G_{\hat{a}}}$
- Since $g(\beta)$ is strictly decreasing function, β_0 can be found using binary search.

Sub-optimal Model under γ-EL

Algorithm 3 Sub-optimal solution to optimization (1)

Input: $\boldsymbol{w}_{G_{\hat{a}}}, \boldsymbol{w}_O, \epsilon, \gamma$

Initialization: $g_\gamma(\beta) = g(\beta) - \gamma$, $i = 0$, $\beta_{start}^{(0)} = 0$, $\beta_{end}^{(0)} = 1$

1: **if** $g_\gamma(0) \leq 0$ **then**
2: $\quad \underline{\boldsymbol{w}} = \boldsymbol{w}_O$, and go to line 13;
3: **end if**
4: **while** $\beta_{end}^{(i)} - \beta_{start}^{(i)} > \epsilon$ **do**
5: $\quad \beta_{mid}^{(i)} = (\beta_{start}^{(i)} + \beta_{end}^{(i)})/2$;
6: $\quad$ **if** $g_\gamma(\beta_{mid}^{(i)}) \geq 0$ **then**
7: $\quad\quad \beta_{start}^{(i+1)} = \beta_{mid}^{(i)}$, $\beta_{end}^{(i+1)} = \beta_{end}^{(i)}$;
8: $\quad$ **else**
9: $\quad\quad \beta_{start}^{(i+1)} = \beta_{start}^{(i)}$, $\beta_{end}^{(i+1)} = \beta_{mid}^{(i)}$;
10: $\quad$ **end if**
11: **end while**
12: $\underline{\boldsymbol{w}} = (1 - \beta_{mid}^{(i)})\boldsymbol{w}_O + \beta_{mid}^{(i)}\boldsymbol{w}_{G_{\hat{a}}}$;
13: **Output:** $\underline{\boldsymbol{w}}$

Theorem

Assume that Assumption 1 and 2 hold, and let $g_\gamma(\beta) = g(\beta) - \gamma$. If $g_\gamma(0) \leq 0$, then w_O satisfies the γ-EL fairness; if $g_\gamma(0) > 0$, then $\lim_{i \to \infty} \beta_{mid}^{(i)} = \beta_{mid}^{(\infty)}$ exits, and $(1 - \beta_{mid}^{(\infty)})w_O + \beta_{mid}^{(\infty)} w_{G_{\hat{a}}}$ satisfies the γ-EL fairness constraint.

Sub-optimal Model under γ-EL

- Upper bound of the expected loss of $f_{\underline{w}}$

Theorem

Under Assumption 1 and 2, we have the following :
$L(\underline{w}) \leq \max_{a\in\{0,1\}} L_a(w_O)$. *That is, the expected loss of* $f_{\underline{w}}$ *is not worse than the loss of the disadvantaged group under predictor* f_{w_O}.

- Learning with Finite Samples

$$\hat{L}(w) = \frac{1}{n}\sum_{i=1}^{n} l(Y_i, f_w(X_i)),$$
$$\hat{L}_a(w) = \frac{1}{n_a}\sum_{i:A_i=a} l(Y_i, f_w(X_i))$$

$$\hat{w} = arg\min_{w} \hat{L}(w), \quad s.t. \quad |\hat{L}_0(w) - \hat{L}_1(w)| \leq \hat{\gamma} \tag{3}$$

Solving (3) using γ and empirical loss is equivalent to solving (1) if the number of data points from each group is sufficiently large.

Beyond Linear Models

- To train a deep model under the equalized loss fairness notion, we can take advantage of **Algorithm 2 for fine-tuning under EL** as long as the the objective function is convex with respect to the parameters of the output layer.

Experiments

- Baselines : PM, LinRe, FairBatch
- Overall loss and loss difference between two demographic groups

Table 1: Linear regression model under EL fairness. The loss function in this example is the mean squared error loss.

		$\gamma = 0$	$\gamma = 0.1$
PM	test loss	0.9246 ± 0.0083	0.9332 ± 0.0101
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.1620 ± 0.0802	0.1438 ± 0.0914
LinRe	test loss	0.9086 ± 0.0190	0.8668 ± 0.0164
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.2687 ± 0.0588	0.2587 ± 0.0704
Fair Batch	test loss	0.8119 ± 0.0316	0.8610 ± 0.0884
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.2862 ± 0.1933	0.2708 ± 0.1526
ours Alg 2	test loss	0.9186 ± 0.0179	0.8556 ± 0.0217
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.0699 ± 0.0469	0.1346 ± 0.0749
ours Alg 3	test loss	0.9522 ± 0.0209	0.8977 ± 0.0223
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.0930 ± 0.0475	0.1437 ± 0.0907

Table 2: Logistic Regression model under EL fairness. The loss function in this example is binary cross entropy loss.

		$\gamma = 0$	$\gamma = 0.1$
PM	test loss	0.5594 ± 0.0101	0.5404 ± 0.0046
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.0091 ± 0.0067	0.0892 ± 0.0378
LinRe	test loss	0.3468 ± 0.0013	0.3441 ± 0.0012
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.0815 ± 0.0098	0.1080 ± 0.0098
Fair Batch	test loss	1.5716 ± 0.8071	1.2116 ± 0.8819
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.6191 ± 0.5459	0.3815 ± 0.3470
Ours Alg2	test loss	0.3516 ± 0.0015	0.3435 ± 0.0012
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.0336 ± 0.0075	0.1110 ± 0.0140
Ours Alg3	test loss	0.3521 ± 0.0015	0.3377 ± 0.0015
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.0278 ± 0.0075	0.1068 ± 0.0138

Table 3: Neural Network training under EL fairness. The loss function in this example is the mean squared error loss.

		$\gamma = 0$	$\gamma = 0.1$
PM	test loss	0.9490 ± 0.0584	0.9048 ± 0.0355
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.1464 ± 0.1055	0.1591 ± 0.0847
LinRe	test loss	0.8489 ± 0.0195	0.8235 ± 0.0165
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.6543 ± 0.0322	0.5595 ± 0.0482
Fair Batch	test loss	0.9012 ± 0.1918	0.8638 ± 0.0863
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.2771 ± 0.1252	0.1491 ± 0.0928
ours Alg 2	test loss	0.9117 ± 0.0172	0.8519 ± 0.0195
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.0761 ± 0.0498	0.1454 ± 0.0749
ours Alg 3	test loss	0.9427 ± 0.0190	0.8908 ± 0.0209
	test $\lvert\hat{L}_0 - \hat{L}_1\rvert$	0.0862 ± 0.0555	0.1423 ± 0.0867