

Understanding and Improving Fairness-Accuracy Trade-offs in Multi-Task Learning

Yuyan Wang, yuyanw@google.com

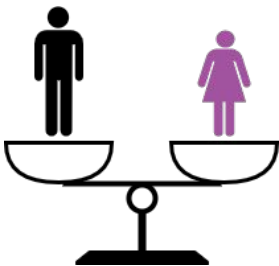
with Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, Ed H. Chi

Faire Tech Talk Series, 07/22/2021

Google Research



Fairness



Objective:

Subgroups are treated equally.

Why: Critical for decision making in employment, education, and criminal justice etc.

Mostly studied in **single-task learning** problems.



Multi-Task Learning

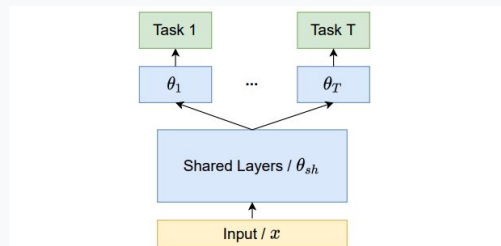


Figure 1: Shared-bottom architecture for a multi-task model.

Objective:

Jointly optimize the performance of multiple tasks.

Why: Transfer learning / regularization / model efficiency/...

Mostly only focused on optimizing **accuracy** across multiple tasks.

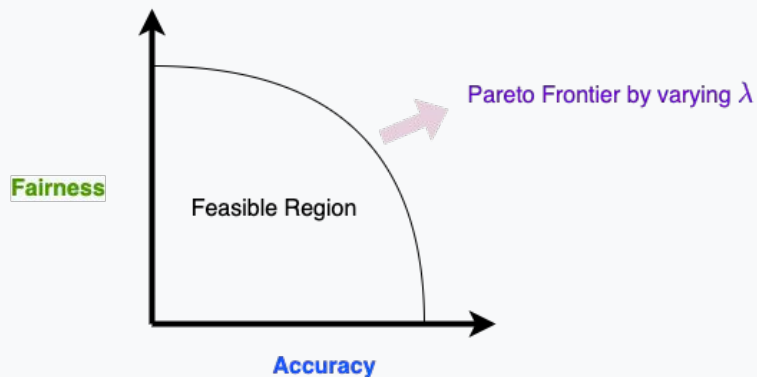
A largely unexplored question:
How does **fairness** play out in **multi-task learning** (MTL) scenarios?

Fairness comes at a cost of accuracy

For a **single task**, there exists a Pareto frontier for fairness-accuracy trade-off.

$$\hat{\mathcal{L}}_{\text{task}}(\theta) = \hat{\mathcal{L}}(\theta) + \lambda \hat{\mathcal{F}}(\theta)$$

↓ accuracy loss ↓ fairness loss



MTL comes with an accuracy trade-off

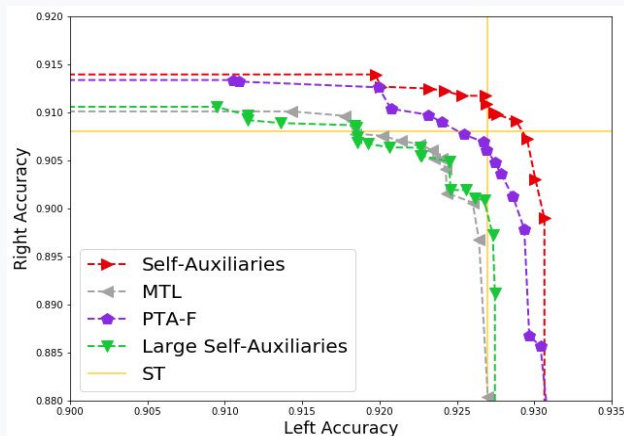
There exists a Pareto frontier for **accuracies** across different tasks.

$$\hat{\mathcal{L}}(\theta) := \sum_{t=1}^T w_t \hat{\mathcal{L}}_t(\theta)$$

weight for task t accuracy loss for task t



Multi-MNIST Dataset [1].



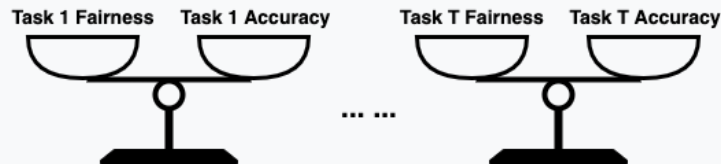
Accuracy trade-off between predicting **left** and **right** digits [2].

Fairness in MTL:

A multi-dimensional Pareto Frontier

Fairness is mostly studied in **single-task** settings.

MTL research mostly focused on optimizing **only the accuracy** across multiple tasks.



With T tasks each with a fairness objective and an accuracy objective, there is a $2T$ -dimensional Pareto frontier.

Research Questions:

1. How does fairness play out in the multi-task scenario?
2. How to characterize the multi-dimensional fairness-accuracy trade-off?
3. Can we improve the Pareto frontier?

Problem Framing

Fairness implications in Multi-Task Learning

Fairness implications in multi-task learning

	T1 Error	T1 FPR Gap	T2 Error	T2 FPR Gap
STL-T1	0.2030	0.2716	-	-
STL-T2	-	-	0.0784	0.0145
MTL	0.2035	0.2846	0.0783	0.0137
Difference	+0.24%	+4.78%	-0.08%	-5.39%

(a) CelebA: MTL hurts Task 1 fairness but improves Task 2 fairness.

	T1 Error	T1 FPR Gap	T2 Error	T2 FPR Gap
STL-T1	0.1659	0.1200	-	-
STL-T2	-	-	0.1313	0.0661
MTL	0.1656	0.1205	0.1299	0.0738
Difference	-0.20%	+0.34%	-1.10%	+11.60%

(b) UCI-Adult: MTL improves Task 2 accuracy but hurts its fairness.

STL-T1: single-task learning for Task 1;

STL-T2: single-task learning for Task 2;

MTL: multi-task learning with equal task weight.

- Datasets: MTL classification problems
 - **CelebA:** Attractive (T1) & Smiling (T2)
 - **UCI-Adult:** Income > \$50k (T1) & Capital Gain > 0 (T2)
- Fairness attribute: **Gender**
- Accuracy metric: prediction error rate
- Fairness metric: Equal opportunity [3] for group fairness
 - False Positive Rate (FPR) gap between male and female

Fairness implications in multi-task learning (cont'd)

	T1 Error	T1 FPR Gap	T2 Error	T2 FPR Gap
STL-T1	0.2030	0.2716	-	-
STL-T2	-	-	0.0784	0.0145
MTL	0.2035	0.2846	0.0783	0.0137
Difference	+0.24%	+4.78%	-0.08%	-5.39%

(a) CelebA: MTL hurts Task 1 fairness but improves Task 2 fairness.

	T1 Error	T1 FPR Gap	T2 Error	T2 FPR Gap
STL-T1	0.1659	0.1200	-	-
STL-T2	-	-	0.1313	0.0661
MTL	0.1656	0.1205	0.1299	0.0738
Difference	-0.20%	+0.34%	-1.10%	+11.60%

(b) UCI-Adult: MTL improves Task 2 accuracy but hurts its fairness.

STL-T1: single-task learning for Task 1;

STL-T2: single-task learning for Task 2;

MTL: multi-task learning with equal task weight.



MTL may have **larger** impacts on fairness goals than on accuracy goals...



.. or **hurt** the fairness of some tasks while benefiting from its accuracy gains

Training multiple tasks together by simply pooling the accuracy objectives may lead to **unwanted fairness consequences**.

New Metrics

Measuring fairness-accuracy trade-offs in MTL

Measuring fairness in multi-task learning

- It's hard to visualize a Pareto frontier with > 3 dimensions
- Even with 2 tasks, we have a 4-dim Pareto frontier for fairness-accuracy trade-off
- Can we efficiently summarize and visualize this **multi-dimensional Pareto frontier**?
 - Moreover, fairness/accuracy metrics could differ largely across different tasks (e.g. some tasks are intrinsically harder to learn / have more bias)

Measuring **relative change** over single-task learning (STL), and **average** across tasks:

Average Relative Fairness Gap (ARFG)

$$ARFG := \frac{1}{T} \sum_{t=1}^T \frac{FPRGap^{(t)}}{\text{single-task FPR gap w/o fairness remediation}^{(t)}},$$

Average Relative Error (ARE)

$$ARE := \frac{1}{T} \sum_{t=1}^T \frac{Err^{(t)}}{\text{single-task error w/o fairness remediation}^{(t)}}.$$

Measuring fairness in multi-task learning (cont'd)

- ARFG and ARE are always positive, and
 - can be either smaller or greater than 1 as MTL could either improve or hurt accuracy / fairness for individual tasks;
 - $ARFG < 1$ ($ARE < 1$) suggests that MTL reduces relative FPR gap (error) on average, and vice versa.
- **ARFG-ARE Pareto frontier:** overall fairness-accuracy trade-off **across all tasks**.

Average Relative Fairness Gap (ARFG)

$$ARFG := \frac{1}{T} \sum_{t=1}^T \frac{FPRGap^{(t)}}{\text{single-task FPR gap w/o fairness remediation}},$$

Average Relative Error (ARE)

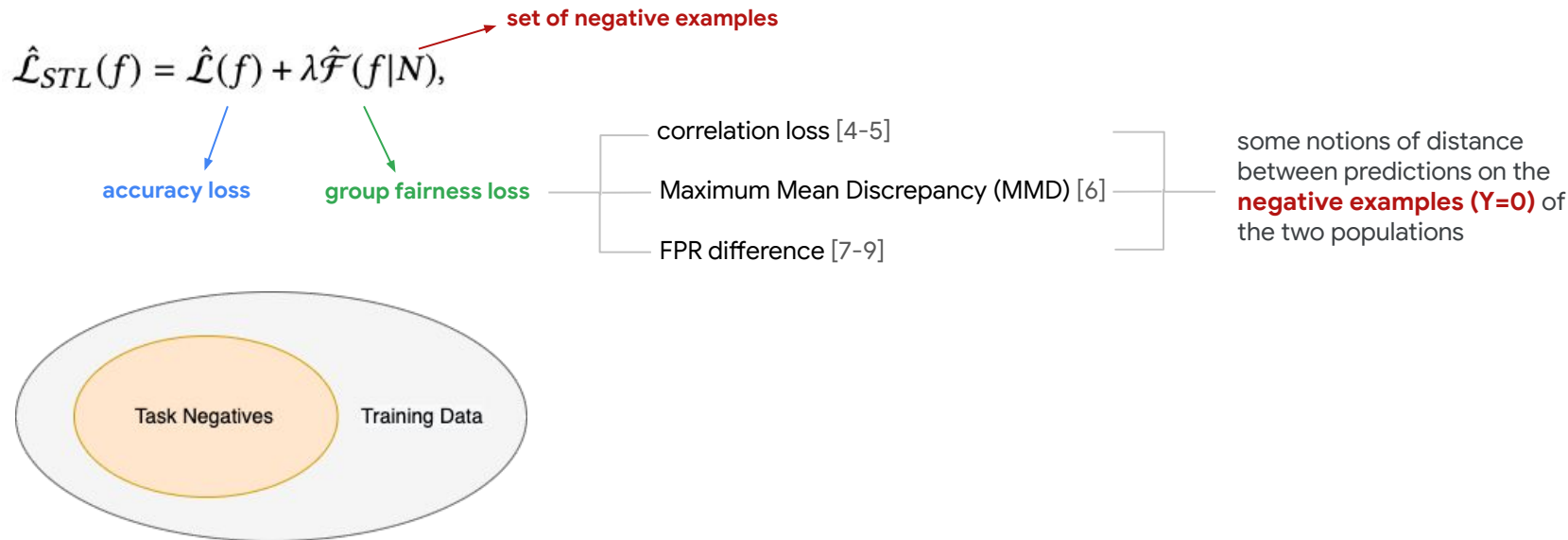
$$ARE := \frac{1}{T} \sum_{t=1}^T \frac{Err^{(t)}}{\text{single-task error w/o fairness remediation}}.$$

New Mitigation

Improving fairness-accuracy
trade-offs in MTL

Fairness loss: single-task learning

Using **FPR gap** for *equal opportunity* as the measure for group fairness...



[4] Beutel et al. Fairness in recommendation ranking through pairwise comparisons. KDD 2019.

[5] Beutel et al. Putting fairness principles into practice: Challenges, metrics, and improvements. AIES 2019.

[6] Prost et al. Toward a better trade-off between performance and fairness with kernel-based distribution matching. NeurIPS 2019 "ML with Guarantees" workshop.

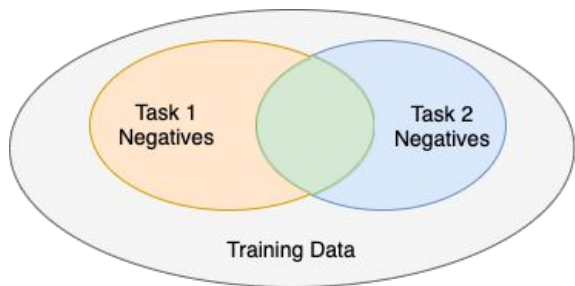
[7] Feldman et al. Certifying and removing disparate impact. KDD 2015.






[8] Menon et al. The cost of fairness in binary classification. FAccT 2018.

[9] Zafar et al. Fairness Constraints: A Flexible Approach for Fair Classification. JMLR 2019.

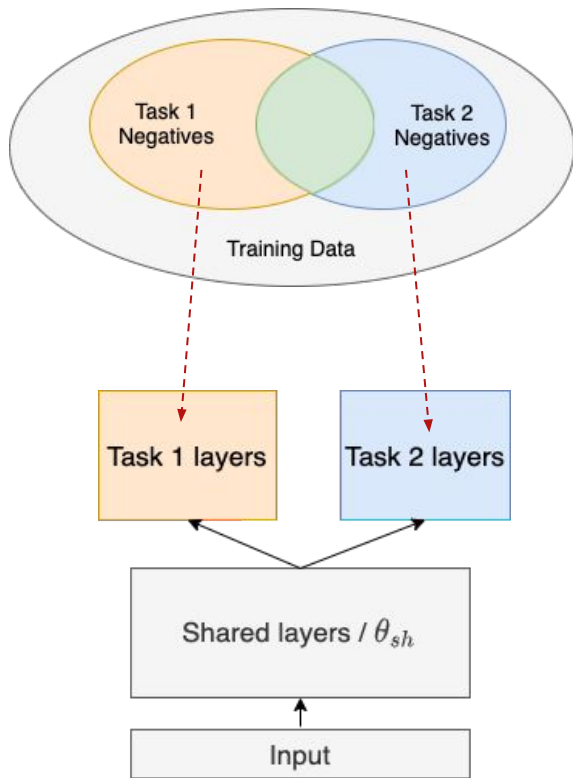
Baseline: Fairness loss generalized to MTL

Multi-task learning



- **Baseline:** Fairness loss computed on  &  &  ;
- However,  is **only** relevant to Task 1 fairness;
- Likewise,  is **only** relevant to Task 2 fairness;
- But Baseline method does not distinguish between them => A **suboptimal** use of model capacity!

Our proposal: A redistribution of fairness losses



MTA-F: Multi-task-aware fairness treatment

Let's address the fairness in a more targeted way:

- Head layers address fairness issues that are **specific** to the task itself;
- Shared layers address fairness issues that are **common** to more than 1 tasks.



A more efficient
allocation of
model capacity

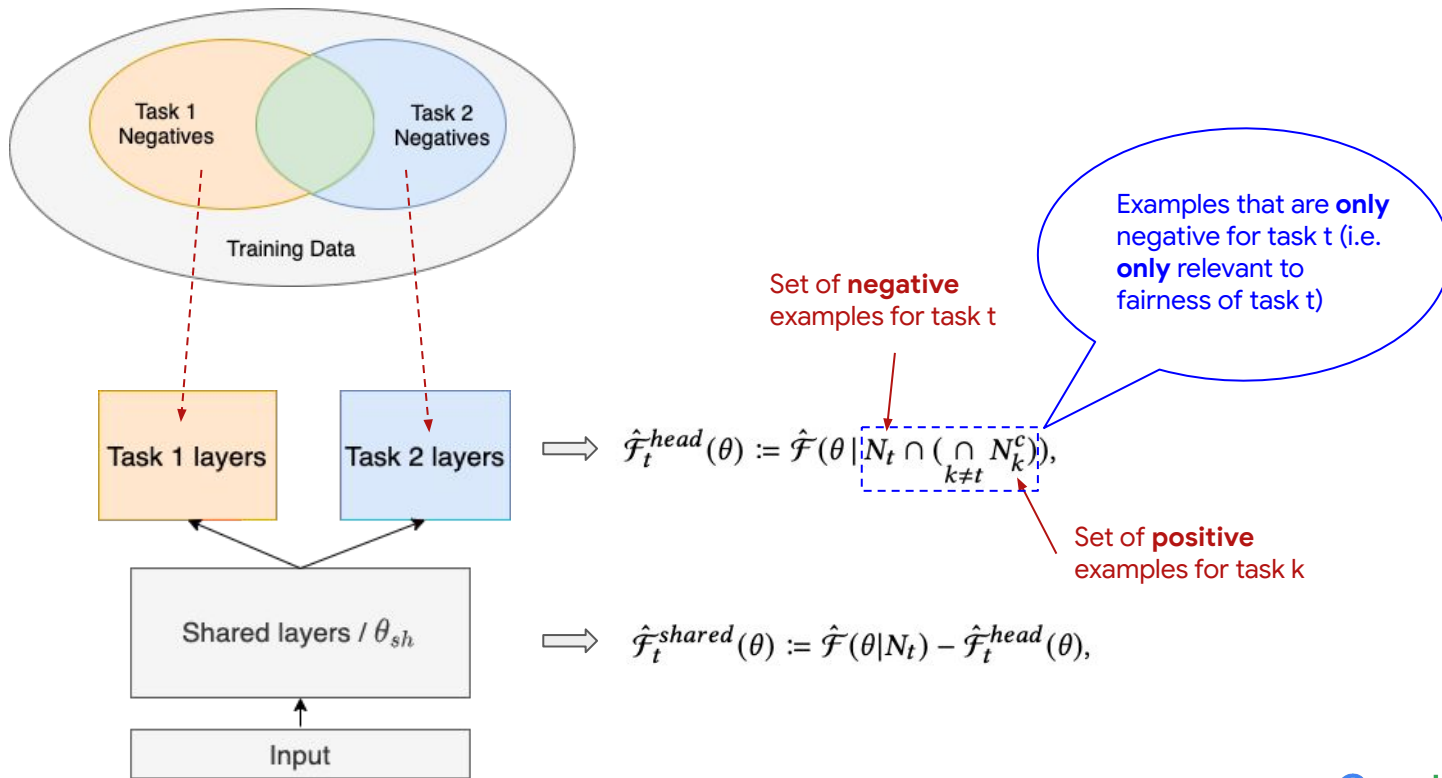


"saves" the model
more capacity for
accuracy objectives

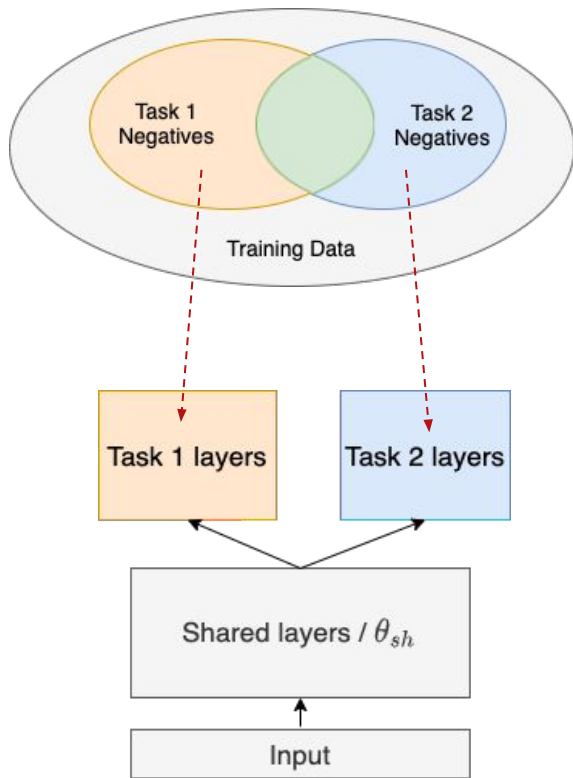


a better overall
fairness-accuracy
trade-off

MTA-F: Multi-task-aware fairness treatment

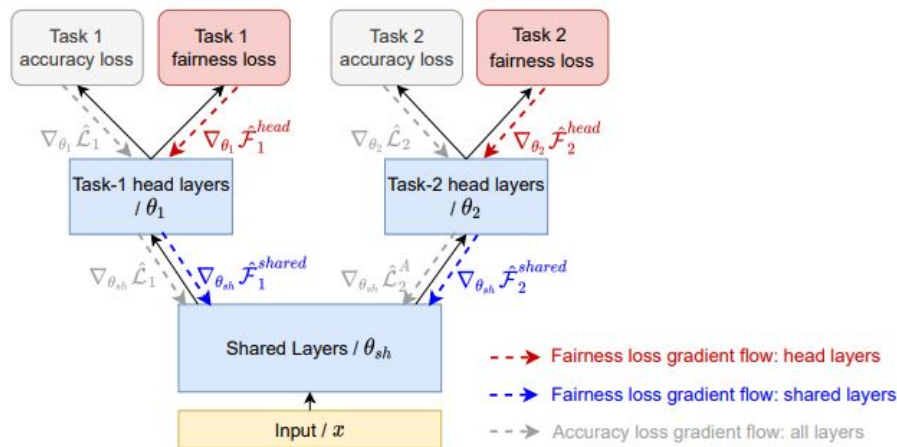


MTA-F: Multi-task-aware fairness treatment



$$\hat{\mathcal{F}}_t^{head}(\theta) := \hat{\mathcal{F}}(\theta | N_t \cap (\bigcap_{k \neq t} N_k^c)),$$

$$\hat{\mathcal{F}}_t^{shared}(\theta) := \hat{\mathcal{F}}(\theta | N_t) - \hat{\mathcal{F}}_t^{head}(\theta),$$



(b) Backpropagation with MTA-F: We backpropagate task-specific fairness losses $\hat{\mathcal{F}}_t^{head}$ to head layers, and the remaining fairness loss $\hat{\mathcal{F}}_t^{shared}$ to shared layers ($t = 1, 2$).

MTA-F: Multi-task-aware fairness treatment

Algorithm 1: MTA-F Update Rule

Input: Mini-batch $\{(x_i, y_i^1, \dots, y_i^T)\}_{i=1}^n$, model parameters

$\theta = (\theta_{sh}, \theta_1, \dots, \theta_T)$, task weights $\{w_t\}_{t=1}^T$, fairness weights $\{\lambda_t\}_{t=1}^T$, head-to-shared ratio $\{r_t\}_{t=1}^T$, and learning rate η

1 **for** $t = 1, \dots, T$ **do**

2 $\hat{\mathcal{L}}_t(\theta_{sh}, \theta_t) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_t(f_t(x_i; \theta_{sh}, \theta_t), y_i^t)$

 ▸ *Compute accuracy losses*

3 $\hat{\mathcal{F}}_t^{head}(\theta_{sh}, \theta_t)$ and $\hat{\mathcal{F}}_t^{shared}(\theta_{sh}, \theta_t)$ as in Eq. (10) / (11)

 ▸ *Compute fairness losses*

4 $\theta_t = \theta_t - \eta(w_t \nabla_{\theta_t} \hat{\mathcal{L}}_t(\theta_{sh}, \theta_t) + \lambda_t r_t \nabla_{\theta_t} \hat{\mathcal{F}}_t^{head}(\theta_{sh}, \theta_t))$

 ▸ *Gradient descent on head parameters*

5 **end**

6 $\theta_{sh} = \theta_{sh} - \eta[\sum_{t=1}^T w_t \nabla_{\theta_{sh}} \hat{\mathcal{L}}_t(\theta_{sh}, \theta_t) + \lambda_t \nabla_{\theta_{sh}} \hat{\mathcal{F}}_t^{shared}(\theta_{sh}, \theta_t)]$

 ▸ *Gradient descent on shared parameters*

Output: Updated model parameters $\theta = (\theta_{sh}, \theta_1, \dots, \theta_T)$

iteratively update T tasks

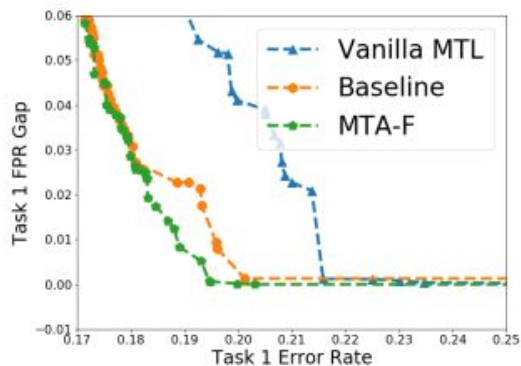
MTA-F differs from baseline method in these places.

Experiments

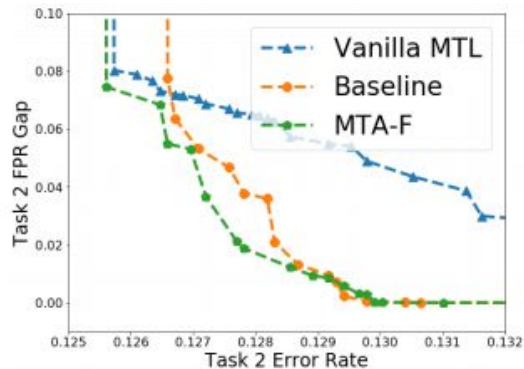
- Datasets:
 - **UCI-Adult**: *Income* > \$50k (T1), *Capital Gain* > 0 (T2)
 - **German Credit Data**: *Good loans* (T1), *Credit* > 2000 (T2)
 - **LSAC Law School**: *Pass bar* (T1), *high GPA* (T2)
- Methods:
 - **Vanilla MTL**: plain MTL **without** fairness mitigation
 - **Baseline**: Per-task fairness treatment
 - **MTA-F**: our proposed method
- Fairness loss: correlation loss / MMD loss / FPR gap loss
- Fairness metric: Equal Opportunity between females and males

Experiments: UCI-Adult

Marginal Pareto frontiers (lower left is better)

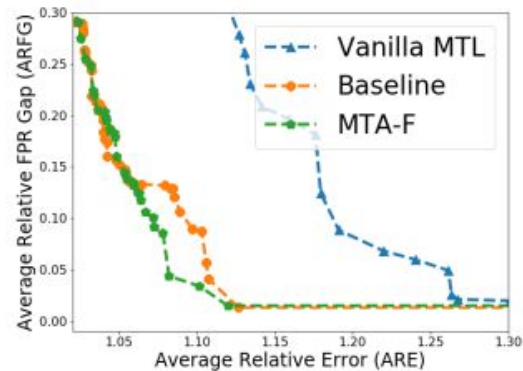


(a) Fairness-accuracy Pareto frontier for Task 1.



(b) Fairness-accuracy Pareto frontier for Task 2.

AFRG-ARE Pareto frontier



(c) ARFG-ARE Pareto frontier.

MTA-F improves **per-task** as well as **overall** fairness-accuracy Pareto frontiers.

Experiments: Numerical results

Overall fairness gap (green arrow) Overall error rate (blue arrow)

Dataset	UCI-Adult		German Credit		LSAC Law School	
Metric	<i>ARFG</i>	<i>ARE</i>	<i>ARFG</i>	<i>ARE</i>	<i>ARFG</i>	<i>ARE</i>
Vanilla MTL	0.3444	1.1040	0.1336	0.8367	0.3497	0.9778
Baseline	0.0871	1.1032	0.0999	0.8356	0.1126	0.9864
MTA-F	0.0437	1.0820	0.0364	0.8264	0.0310	0.9731

Table 3: Average relative fairness gap (*ARFG*) and average relative error (*ARE*) on **UCI-Adult**, **German Credit Data** and **LSAC Law School** datasets, as defined in Section 4. Lower metric values indicate better overall fairness / accuracy across all tasks.

Task 1 (orange arrows) Task 2 (purple arrows)

		T_1 Err	T_1 FPRGap	T_2 Err	T_2 FPRGap
UCI-Adult	Vanilla MTL	0.1911	0.0715	0.1359	0.0091
	Baseline	0.1938	0.0186	0.1336	0.0020
	MTA-F	0.1891	0.0083	0.1319	0.0016
German Credit	Vanilla MTL	0.205	0.0150	0.220	0.0084
	Baseline	0.255	0.0879	0.180	0.0069
	MTA-F	0.200	0.0033	0.220	0.0034
LSAC Law School	Vanilla MTL	0.1555	0.0503	0.1565	0.0004
	Baseline	0.1568	0.0119	0.1580	0.0006
	MTA-F	0.1540	0.0015	0.1565	0.0004

Table 4: Per-task metrics for **UCI-Adult**, **German Credit Data** and **LSAC Law School** datasets.

MTA-F improves **per-task** as well as **overall** fairness-accuracy metrics, across all three datasets and with different fairness losses.

Key Takeaways

- Optimizing only for accuracy trade-off in MTL may lead to **unwanted fairness implications**.
- Quantify the **multi-dimensional trade-off**: we propose Average relative fairness gap (ARFG) and average relative error (ARE).
- **MTA-F**: a **data-dependent multi-task fairness mitigation** approach, which decomposes fairness losses for different model components by exploiting task relatedness and the shared architecture for multi-task models.

Thank You

Yuyan Wang

Google Research, Brain Team

yuyanw@google.com