FERERO: A Flexible Framework for Preference-Guided Multi-Objective Learning

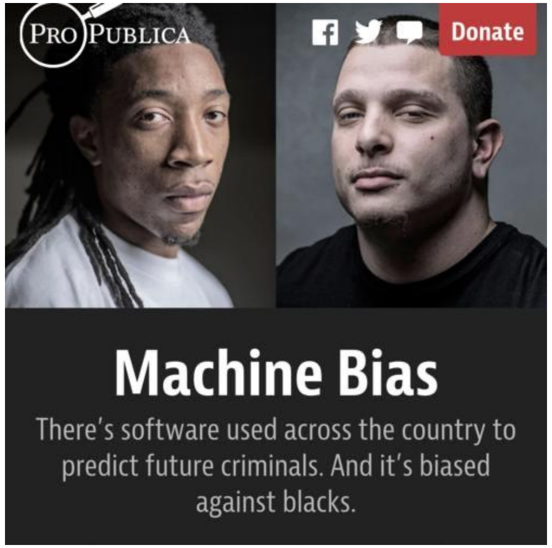
Lisha Chen, AFM Saif, Yanning Shen, Tianyi Chen

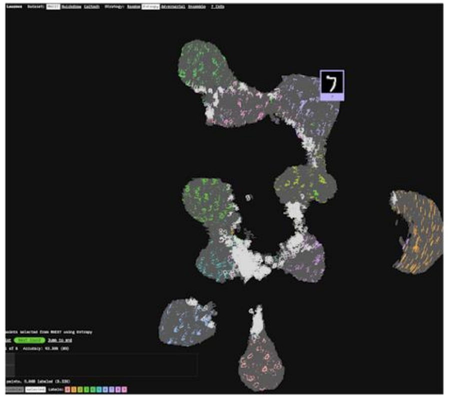
Rensselaer Polytechnic Institute

University of California, Irvine

Multiple metrics arise in machine learning today

Data and model bias





Fast adaptation to new users

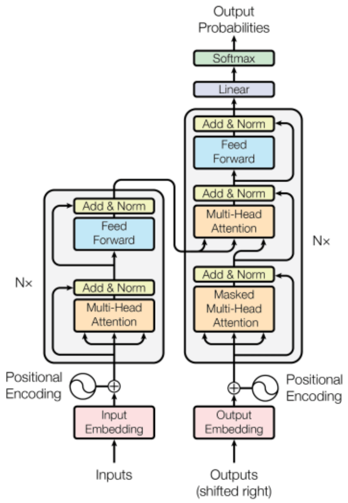
Resource constraints

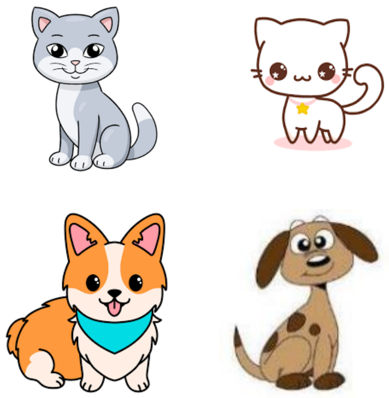
Subject to privacy regulation

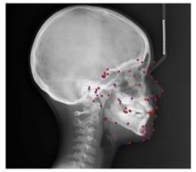
Image from Internet

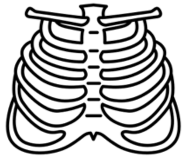
Tasks, data, metrics
all can be modeled as an objective…

$$\min_{\theta} \ \text{loss (model } \theta \text{, training data, metric, tasks)}$$









Unified as multi-objective learning

Formulation for multi-objective learning

$$\min_{\theta} \quad F(\theta) = [f_1(\theta), \dots, f_m(\theta), \dots, f_M(\theta)]$$

A **vector** optimization problem

How to optimize a vector?

$$\begin{bmatrix} 10 \\ 0 \\ 10 \end{bmatrix} \leq_{Pareto} \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix}? \qquad \begin{bmatrix} -5 \\ 2 \\ 2 \end{bmatrix} \leq_{LS} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}?$$

Commonly used dominance notions are not enough

How to optimize a vector?

$$\begin{bmatrix}10\\0\\10\end{bmatrix} \leq_{Pareto} \begin{bmatrix}10\\10\\10\end{bmatrix}? \qquad \begin{bmatrix}-5\\2\\2\end{bmatrix} \leq_{LS} \begin{bmatrix}1\\1\\1\end{bmatrix}?$$

Pareto dominance is not enough: simply minimizing one objective achieves weak Pareto optimality.

Linear scalarization is not enough: objectives can be dominated by the one with the largest scale.

Relative preference with cone-induced partial order

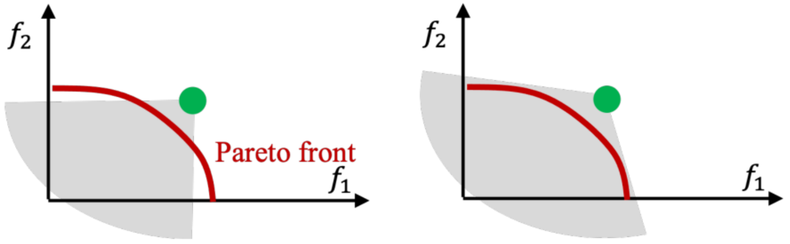
Definition 1 (C_A-dominance). Given $v, w \in R^M, A \in R^{M\times M}$, and $C_A := \{y \in R^M \,|Ay \geq 0\} \neq \emptyset$, we say v strictly dominates w based on C_A if and only if $A(v - w) < 0$.

$A = I_M, \qquad C_A = R_+^M$

reduces to Pareto optimality

Relative preference with cone-induced partial order

Definition 1 (C_A-dominance). Given $v, w \in R^M, A \in R^{M \times M}$, and $C_A := \{y \in R^M \,|Ay \geq 0\} \neq \emptyset$, we say v strictly dominates w based on C_A if and only if $A(v - w) < 0$.



(a) (b)

Solid red curves: Pareto fronts
Green dots: reference points
Gray shaded regions: objectives dominating the reference points, under different C_A in both figures.

Benefits with a general partial order:

1. allows controlled ascent, thus can reach every point on the Pareto front
2. avoid merely minimizing a single objective

Relative preference with cone-induced partial order

Definition 1 (C_A-dominance). Given $v, w \in R^M, A \in R^{M\times M}$, and $C_A := \{y \in R^M \mid Ay \geq 0\} \neq \emptyset$, we say v strictly dominates w based on C_A if and only if $A(v - w) < 0$.

e.g., a user gives at least α relative importance to each objective with $\alpha \in (0, 0.5)$, then this can be achieved by defining a partial order induced by the cone (example in [2]):

$$C = \{F \in R^2 \mid \alpha f_1 + (1-\alpha) f_2 \geq 0, (1-\alpha) f_1 + \alpha f_2 \geq 0\}$$

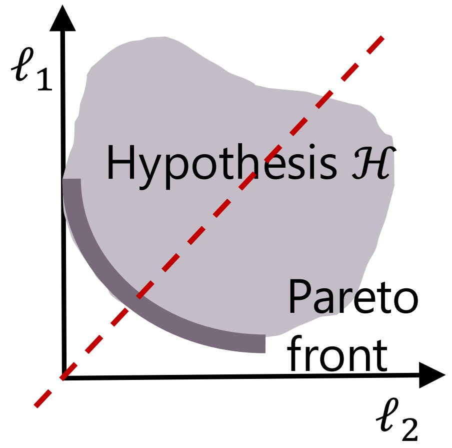
Benefits with a general partial order:

1. allows controlled ascent, thus can reach every point on the Pareto front
2. avoid merely minimizing a single objective

Address the imbalance issue through constraints

In multi-lingual ASR, we want the training losses of all languages to be similar.

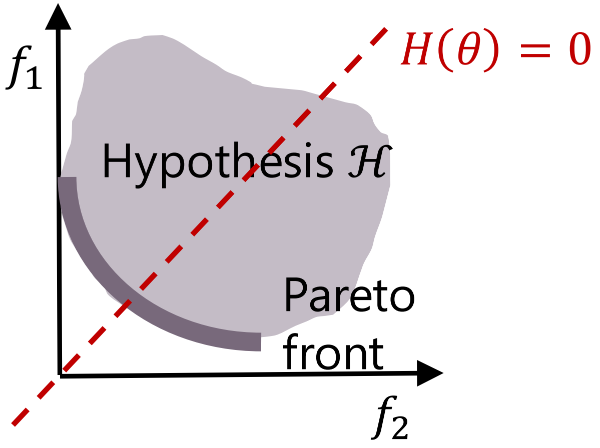
Can we use linear scalarization (LS, a.k.a. static weighting) with carefully tuned weight?



NO! LS cannot achieve certain constraints, even when fine-tuned!

Address the imbalance issue through constraints

Idea: enforce the objectives
to achieve similar values



e.g. use a constraint function

$$H(\theta) = f_1(\theta) - f_2(\theta)$$

$$\min_{C_A} \quad F(\theta) \;\; \text{s.t. } H(\theta) = 0$$

FERERO: a flexible framework to capture preferences

$$\min_{C_A} \quad F(\theta)$$
$$\text{s.t. } H(\theta) = 0, G(\theta) \leq 0$$

Relative preference: captured by the partial order,
determines improving directions

Absolute preference: captured by the constraints

A primal approach to the constrained problem

Main program: $\min_{\theta} \; F(\theta)$ s.t. $H(\theta) = 0, G(\theta) \leq 0$

Define a **subprogram** that finds an update direction d which --
both improves objectives & constraints:

improvement defined by general partial order

$$\min_{c,d} c + \frac{1}{2} \| d \|^2, \qquad \text{s.t.} \;\; A\nabla F(\theta)^\top d \leq c \cdot A\mathbf{1}$$

approximate the Hessian $\nabla^2 F(\theta)$ by identity

$$c_g G(\theta) + \nabla G(\theta)^\top d \leq 0$$
$$c_h H(\theta) + \nabla H(\theta)^\top d = 0$$

Idea similar to SQP:
Use local quadratic approximation to the objectives

A primal approach to the constrained problem

Subprogram:

$$\psi(\theta) := \min_{(d,c)\in\mathbb{R}^q\times\mathbb{R}} c + \frac{1}{2}\|d\|^2 \quad \text{s.t.} \quad A\nabla F(\theta)^\top d \le c \cdot A\mathbf{1}$$
$$\nabla G(\theta)^\top d + c_g G(\theta) \le 0, \nabla H(\theta)^\top d + c_h H(\theta) = 0$$

Dual of the subprogram:

Find dynamic weight λ for the following problem

$$\lambda^*(\theta) \in \arg\min_{\lambda\in\Omega_\lambda} \varphi(\lambda;\theta) := \frac{1}{2}\left\|\nabla F(\theta) A_{ag}^\top \lambda\right\|^2 - c_g \lambda_g^\top G(\theta) - c_h \lambda_h^\top H(\theta)$$

Algorithm update

The optimal direction: $d^*(\theta) = -[\nabla F(\theta)A^\top, \nabla G(\theta), \nabla H(\theta)]\lambda^*(\theta)$

Update $\lambda^*(\theta_t)$: $\lambda^*(\theta_t) = \operatorname{argmin}_{\lambda \in \Omega_\lambda} \varphi(\lambda; \theta_t)$

Update θ_t along $d^*(\theta_t)$: $\theta_{t+1} = \theta_t + \alpha_t d^*(\theta_t)$

Just like MGDA, can be seen as a dynamic weighting method

KKT condition

KKT condition

$$\nabla F(\theta)A^\top \lambda_f + \nabla G(\theta)\lambda_g + \nabla H(\theta)\lambda_h = 0$$ stationarity

$$\lambda_f \in \Delta^M, \lambda_g \in R_+^{M_g}, \lambda_h \in R^{M_h}$$ dual feasibility

$$G(\theta) \leq 0, H(\theta) = 0$$ primal feasibility

$$\lambda_g^*(\theta)^\top [-G(\theta)]_+ = 0$$ complementary slackness

Proper merit functions (KKT score)

$$J_1(\theta) = \underbrace{\| d^*(\theta) \|^2}_{\text{stationarity}} + \underbrace{\lambda_g^*(\theta)^\top [-G(\theta)]_+}_{\text{slackness}} + \underbrace{\| [G(\theta)]_+ \|_1 + \| H(\theta) \|_1}_{\text{feasibility}}$$

$[\cdot]_+$: element-wise ReLU function
$| \cdot |$: element-wise absolute function

it achieves *0* iff the model θ satisfies the first-order KKT optimality condition

Optimization analysis

Theorem (Optimization error guarantee, informal)

Under mild assumptions, with proper choice of step sizes,

for the FERERO meta algorithm, $\frac{1}{T}\sum_{t=0}^{T-1} J_1(\theta_t) = O(T^{-1})$

- The convergence rate matches that of gradient descent for general nonconvex objectives.
- The efficient single-loop and stochastic algorithms developed under this framework also have convergence rate guarantees.

FERERO performance

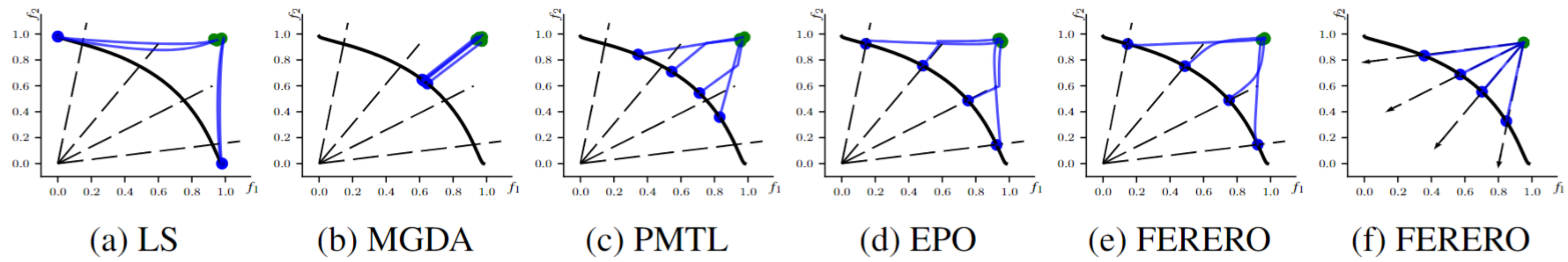
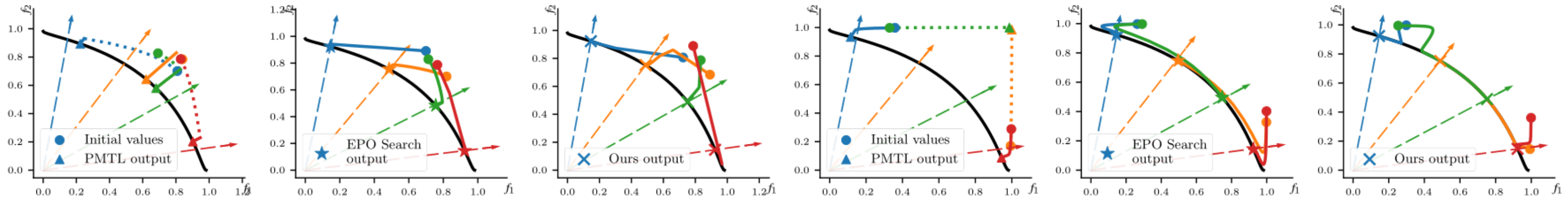


Figure 3: Converging solutions (blue dots) and optimization trajectories (blue lines) on the objective space of different methods on synthetic objectives given in (5.1). Dashed arrows represent pre-specified preference vectors. The green dots represent initial objective values.

- ❑ Linear scalarization (LS) can't converge to certain points on the Pareto front.
- ❑ Multi-gradient descent algorithm (MGDA) does not align perfectly with preference constraints.

FERERO performance



(a) PMTL (b) EPO (c) Ours (d) PMTL (e) EPO (f) FERERO

Figure 4: Outputs (colored markers) and optimization trajectories (colored lines) of different methods when initial objectives are near the Pareto front. Different colors represent different preferences.

- ❑ PMTL does not allow controlled ascent, thus not converging in some problems (d).
- ❑ EPO & FERERO allow controlled ascent and converge in those problems.

Application to multi-lingual ASR

$$\min_{\theta} \quad F(\theta) := \big(\underbrace{f_p(\theta)}_{\text{self-supervised loss}}, \overbrace{f_t^{\mathrm{ch}}(\theta), f_t^{\mathrm{en}}(\theta)}^{\text{supervised loss}}\big)^\top \quad \text{s.t.} \quad f_p(\theta) \le \epsilon_1,\ f_t^{\mathrm{ch}}(\theta) - f_t^{\mathrm{en}}(\theta) = \epsilon_2$$

EPO & PMTL do not capture flexible preferences to solve this problem.

Table 3: WERs (%) on Librispeech and AISHELL v1.

	Method	English	Chinese	Average
	Komatsu et al. [19]	7.11	-	-
	w/o CPC [38]	11.8	10.2	11.0
(MoDo)	Init. (M2ASR) [38]	7.3	6.2	6.7
	LS-FT	6.8	5.9	6.4
	FERERO-FT	**5.4**	**4.9**	**5.1**

Take-home message

We propose a flexible framework capturing absolute & relative preferences for preference-guided MOL.

Algorithms and efficient variants are developed under this framework with convergence rate guarantees.

Thank you!