A Omitted details from Section 3.1

A.1 Implementation of Figure 1

Figure 1 is generated from a simple three-task linear MTL problem that we constructed, using Equation (4). Specifically, we set $\hat{y}_1 \approx (0.98, 0, 0.2)$, $\hat{y}_2 \approx (-0.49, -0.85, 0.2)$, $\hat{y}_3 \approx (-0.49, 0.85, 0.2)$ (the number of data points is $n = 3$; this is a rotated version of the equiangular tight frame), set $q = 1$ (the width of the network is one, i.e., under-parameterized), and plotted the achievable points of Equation (4) by sweeping P_Z (the set of rank-1 projection matrices). The software we used is Mathematica.

A.2 Over-parametrization reduces the Pareto front to a singleton

For general non-linear multi-task neural networks, we will show that increasing the width reduces the Pareto front to a singleton, where all tasks simultaneously achieve zero training loss. As a consequence, linear scalarization with any convex coefficients will be able to achieve this point.

To illustrate this point, we follow the same setting as in Section 2, except changing the model to be a two-layer ReLU multi-task network with bias terms. Concretely, for task i, the prediction on input x is given by $f_i(x, W, b, a_i) = a_i^\top \max(Wx + b, 0)$. We have the following result.

Theorem A.1. *There exists a two-layer ReLU multi-task network with hidden width $q = nk$ and parameters $(W, b, a_1, \cdots, a_k)$, such that*

$$f_i(x_i^j, W, b, a_i) = y_i^j, \quad \forall j \in [n], \ \forall i \in [k]. \tag{14}$$

This implies that the network achieves zero training loss on each task.

Proof. For every $i \in [k]$, we can apply Theorem 1 in (Zhang et al., 2021) and find $(W_i, b_i, \tilde{a}_i)$ such that

$$\tilde{a}_i^\top \max(W_i x_i^j + b_i, 0) = y_i^j, \quad \forall j \in [n]. \tag{15}$$

Here, $W_i \in \mathbb{R}^{n \times p}$ and $b_i, \tilde{a}_i \in \mathbb{R}^n$. Now consider

$$W = \begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{pmatrix} \in \mathbb{R}^{nk \times p}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^{nk}, \quad a_i = e_i \otimes \tilde{a}_i \in \mathbb{R}^{nk} \quad \forall i \in [n], \tag{16}$$

where e_i denotes the i-th unit vector in $\mathbb{R}^n$ and $\otimes$ stands for the Kronecker product. It is straightforward to see that

$$a_i^\top \max(W_i x_i^j + b_i, 0) = y_i^j, \quad \forall j \in [n], \ \forall i \in [k]. \tag{17}$$

This finishes the proof as desired. $\qquad\square$

A.3 Proof of Theorem 3.1

Proof of Theorem 3.1. Define $v_i = \langle \hat{y}_i, s \rangle$ and $v = (v_1, \cdots, v_k)^\top$. We are interested in the set

$$S := \left\{ v \mid \|s\| = 1, s \in \mathrm{span}(\{\hat{y}_i\}_{i \in [k]}) \right\}. \tag{18}$$

We will show S is equivalent to the following set

$$B := \left\{ v \mid v^\top Q v = 1 \right\}, \quad Q = (\hat{Y}^\top \hat{Y})^{-1}, \tag{19}$$

which is essentially the boundary of an ellipsoid.

Step 1: $S \subset B$.   Note $v = \hat{Y}^\top s$, so it suffices to show

$$s^\top \hat{Y}(\hat{Y}^\top \hat{Y})^{-1} \hat{Y}^\top s = 1. \tag{20}$$

Denote $P_{\hat{Y}} = \hat{Y}(\hat{Y}^\top \hat{Y})^{-1}\hat{Y}^\top$, which is a projection matrix that maps to the column space of $\hat{Y}$. Since $s \in \mathrm{span}(\{\hat{y}_i\}_{i \in [k]})$, we have

$$s^\top P_{\hat{Y}} s = s^\top s = \|s\|^2 = 1. \tag{21}$$

Step 2: $B \subset S$. Since $\hat{Y}^\top$ has full row rank, for every $v \in \mathbb{R}^k$, we can always find $s \in \mathbb{R}^n$ such that $\hat{Y}^\top s = v$. We can further assume $s \in \text{span}(\{\hat{y}_i\}_{i \in [k]})$ by removing the component in the null space of $\hat{Y}^\top$. Assuming $v \in B$, we have

$$s^\top P_{\hat{Y}} s = 1, \quad s \in \text{span}(\{\hat{y}_i\}_{i \in [k]}). \tag{22}$$

Since $P_{\hat{Y}}$ is a projection matrix that maps to the column space of $\hat{Y}$, we have $s^\top s = 1$, which implies $\|s\| = 1$.

Now we have $S = B$. The set

$$S' := \left\{ |v| \mid \|s\| = 1, s \in \text{span}(\{\hat{y}_i\}_{i \in [k]}) \right\} \tag{23}$$

can be obtained by reflecting E to the non-negative orthant. Formally, a reflection is determined by a collection of axes $\{i_1, \cdots, i_l\}$ ($0 \le l \le k$). The image of such reflection in the non-negative orthant is given by

$$B_{i_1 \cdots i_l} = \{v \mid v^\top Q_{i_1 \cdots i_l} v = 1, v \ge 0\}, \tag{24}$$

where $Q_{i_1, \cdots i_l} = D_{i_1 \cdots i_l} Q D_{i_1 \cdots i_l}$. As a consequence, we have

$$S' = \bigcup_{0 \le l \le k} \bigcup_{1 \le i_1 < \cdots < i_l \le k} B_{i_1 \cdots i_l}. \tag{25}$$

Finally, the feasible region can be equivalently characterized as

$$\mathcal{F}_1 = \left\{ v^2 \mid \|s\| = 1, s \in \text{span}(\{\hat{y}_i\}_{i \in [k]}) \right\}, \tag{26}$$

which can be obtained by squaring the coordinates of S'. Therefore, if we denote

$$E_{i_1 \cdots i_l} = \{v \mid \sqrt{v}^\top Q_{i_1 \cdots i_l} \sqrt{v} = 1\}, \tag{27}$$

then (note the square root naturally implies $v \ge 0$)

$$\mathcal{F}_1 = \bigcup_{0 \le l \le k} \bigcup_{1 \le i_1 < \cdots < i_l \le k} E_{i_1 \cdots i_l}. \tag{28}$$

This finishes the proof as desired. $\qquad\square$

A.4 Proof of Theorem 3.2

Proof of Theorem 3.2. Note the orthogonal complement of a 1-dimensional subspace of $\text{span}(\{\hat{y}_i\}_{i \in [k]})$ is a $(k-1)$-dimensional subspace, and that the objective function for each task is the square of the projection of $\hat{y}_i$. Denote $t := (\|\hat{y}_1\|^2, \cdots, \|\hat{y}_k\|^2)^\top$. By the Pythagorean theorem (Bhatia, 2013, Section I.6), if $v \in \mathcal{F}_1$, then we must have $t - v \in \mathcal{F}_{q-1}$, and vice versa. As a consequence, denote

$$I_{i_1 \cdots i_l} = \{v \mid \sqrt{t-v}^\top Q_{i_1 \cdots i_l} \sqrt{t-v} = 1\}, \tag{29}$$

and we have

$$\mathcal{F}_{q-1} = \bigcup_{0 \le l \le k} \bigcup_{1 \le i_1 < \cdots < i_l \le k} I_{i_1 \cdots i_l} \tag{30}$$

by Theorem 3.1. $\qquad\square$

A.5 Why scalarization fails in the presence of gradient disagreement

We explain why linear scalarization is incapable of exploring an intersection point with gradient disagreement.

We first recall the geometric interpretation (see Figure 4.9 in (Boyd et al., 2004)) of linear scalarization: a point P lying at the boundary of the feasible region can be achieved by scalarization, if and only if there exists a hyperplane at P and the feasible region lies above the hyperplane. The normal vector of the hyperplane is proportional to the scalarization coefficients. By definition, if a hyperplane at P lies below the feasible region, its normal vector must be a subgradient of the surface. When the boundary of the feasible region is differentiable and the subdifferential set is non-empty, the normal vector must be the gradient of the surface, and the hyperplane becomes the tangent plane at P.

Now suppose P lies at the intersection of two differentiable surfaces S_1 and S_2, and that P can be achieved by scalarization. Applying the above argument to S_1 and P, we know that the scalarization coefficients are proportional to the gradient w.r.t. S_1. Similarly, applying the above argument to S_2 and P yields that the scalarization coefficients are proportional to the gradient w.r.t. S_2. This will result in a contradiction if the two gradients w.r.t. S_1 and S_2 at P lie in different directions, a phenomenon we refer to as *gradient disagreement*. In this case, scalarization cannot reach P.

B Omitted details from Section 3.2

B.1 Proof of Lemma 3.4

We begin by reviewing some basics of convex analysis. We refer the readers to Rockafellar (1997) for more background knowledge.

Definition B.1 (Convex hull). *Let $X = \{x_i\}_{i \in [m]}$ be a collection of points in $\mathbb{R}^n$. The **convex hull** of X, denoted as $\mathrm{conv}(X)$, is the set of all convex combinations of points in X, i.e.,*

$$\mathrm{conv}(X) := \left\{ \sum_{i \in [m]} \alpha_i x_i \mid \alpha_i \geq 0, \sum_{i \in [m]} \alpha_i = 1 \right\}. \tag{31}$$

Definition B.2 (Cone and convex cone). *A set $K \subset \mathbb{R}^n$ is called a **cone** if $x \in K$ implies $\alpha x \in K$ for all $\alpha > 0$. A **convex cone** is a cone that is closed under addition. The convex cone generated by $X = \{x_i\}_{i \in [m]} \subset \mathbb{R}^n$, denoted as $\mathrm{cone}(X)$, is given by*

$$\mathrm{cone}(X) := \mathrm{conv}(\mathrm{ray}(X)), \tag{32}$$

where $\mathrm{ray}(X) = \{\lambda x_i \mid \lambda \geq 0, i \in [m]\}$.

Definition B.3 (Dual cone). *The **dual cone** of $C \subset \mathbb{R}^n$, denoted as C^*, is given by*

$$C^* := \{y \mid \langle y, x \rangle \geq 0, \ \forall x \in C\} \tag{33}$$

Definition B.4 (Projection onto a convex set). *Let C be a closed convex set in $\mathbb{R}^n$. The **orthogonal projection** of $y \in \mathbb{R}^n$ onto the set C, denoted as $P_C y$, is the solution to the optimization problem*

$$P_C y := \inf_{x \in C} \{\|x - y\|\}. \tag{34}$$

Theorem B.5 (Bourbaki-Cheney-Goldstein inequality (Ciarlet, 2013)). *Let C be a closed convex set in $\mathbb{R}^n$. For any $x \in \mathbb{R}^n$, we have*

$$\langle x - P_C x, y - P_C x \rangle \leq 0, \quad \forall y \in C. \tag{35}$$

Theorem Theorem B.5 is also known as the variational characterization of convex projection. We will use it to prove the following lemma, which serves as a crucial ingredient in the proof of Lemma 3.4.

Lemma B.6. *Let C be a closed convex set in $\mathbb{R}^n$. For any $z \in \mathbb{R}^n$, we have*

$$\langle z', x \rangle \geq \langle z, x \rangle, \quad \forall x \in C, \tag{36}$$

where $z' = 2 P_C z - z$ is the reflection of z w.r.t. C.

Proof. Plugging in $z' = 2 P_C z - z$, it suffices to show

$$\langle P_C z - z, x \rangle \geq 0, \quad \forall x \in C. \tag{37}$$

This is true because

$$\langle P_C z - z, x - P_C z \rangle \geq 0, \quad \forall x \in C \tag{38}$$

by Theorem B.5, and that

$$\langle z - P_C z, P_C z \rangle = 0. \tag{39}$$

$\square$

We are now ready to prove Lemma 3.4.

Proof of Lemma 3.4. For the *if* part, we assume w.l.o.g. that $G = \hat{Y}^\top \hat{Y}$ is doubly non-negative. This implies that the angle between each pair of optimal predictors is non-obtuse. Denote $\mathcal{A} = \text{cone}(\{\hat{y}_i\}_{i \in [k]})$ and $\mathcal{A}^*$ as its dual cone. We have $\mathcal{A} \subset \mathcal{A}^*$.

Our goal is to show that, for every $\|s\| = 1$, we can always find $s' \in \mathcal{A}^*$, such that 1) $\|s'\| = 1$; 2) $\langle s', \hat{y}_i \rangle \geq |\langle s, \hat{y}_i \rangle|, \ \forall i \in [k]$. This implies that when restricting our discussion to the Pareto front, we can safely ignore the absolute value which appears in the proof of Theorem 3.1. As a consequence, the Pareto front belongs to the surface determined by $Q = G^{-1}$.

Consider $s' = 2P_{\mathcal{A}^*}s - s$. It is straightforward to see that $\|s'\| = 1$ since it is the reflection of s w.r.t. $\mathcal{A}^*$. To show 2), we will prove:

$$\langle s', x \rangle \geq |\langle s, x \rangle|, \quad \forall x \in \mathcal{A}. \tag{40}$$

We break the discussion into two cases.

Case 1: $\langle s, x \rangle \geq 0$. Since $x \in \mathcal{A} \subset \mathcal{A}^*$, we have

$$\langle s', x \rangle \geq \langle s, x \rangle = |\langle s, x \rangle| \tag{41}$$

by Lemma B.6.

Case 2: $\langle s, x \rangle < 0$. Consider $u = -s$ and denote $\mathcal{C} := \{a \mid \langle a, P_{\mathcal{A}^*}s \rangle \geq 0\}$. It is straightforward to see that $\mathcal{A} \subset \mathcal{C}$.

We will show $P_{\mathcal{C}}u = P_{\mathcal{A}^*}s - s$. In fact, for any $z \in \mathcal{C}$, we have

$$
\begin{aligned}
\|z - u\|^2 &= \|z - (P_{\mathcal{A}^*}s - s) + P_{\mathcal{A}^*}s\|^2 \\
&= \|z - (P_{\mathcal{A}^*}s - s)\|^2 + \|P_{\mathcal{A}^*}s\|^2 + 2\langle z - (P_{\mathcal{A}^*}s - s), P_{\mathcal{A}^*}s \rangle \\
&\geq \|P_{\mathcal{A}^*}s\|^2 + 2\langle z, P_{\mathcal{A}^*}s \rangle \\
&\geq \|P_{\mathcal{A}^*}s\|^2 \\
&= \|P_{\mathcal{A}^*}s - s - u\|^2.
\end{aligned}
$$

Therefore, we have

$$s' = 2P_{\mathcal{A}^*}s - s = 2P_{\mathcal{C}}u - u. \tag{42}$$

With another application of Lemma B.6, we have

$$\langle s', x \rangle \geq \langle u, x \rangle = \langle -s, x \rangle = |\langle s, x \rangle|. \tag{43}$$

This finishes the proof of the *if* part.

For the *only if* part, we consider the point p_i which achieves maximum value along the i-axis. For a given i, it is straightforward to see that such p_i is unique, and therefore is a PO of the feasible region.

Now, p_i belongs to a surface determined by Q', if and only if there exists a non-negative vector v_i, such that $Q'v_i = e_i$ (the normal vector at p_i aligns with the i-th axis). In other words, $G'e_i \geq 0$ where $G' = (Q')^{-1}$. When C1 does not hold, there does not exist a $G^* \in \mathcal{G}$, such that $G^*e_i \geq 0$ for all $i \in [k]$. This implies that these p_i must belong to different surfaces.

$\square$

B.2 Proof of Lemma 3.5

Proof of Lemma 3.5. For the *if* part, we assume w.l.o.g. that $Q_\varnothing = G_\varnothing^{-1}$ is doubly non-negative. This essentially implies that $E_\varnothing$ is dominated by all other surfaces in $\mathcal{E}$, which further implies that $I_\varnothing$ dominates all other surfaces in $\mathcal{I}$. Therefore, the Pareto front must belong to $I_\varnothing$.

For the *only if* part, we study the Pareto front of $\mathcal{F}_{q-1}$ through the lens of $\mathcal{F}_1$. Specifically, a point z is a PO of $\mathcal{F}_{q-1}$ if and only if $\mathcal{F}_{q-1}$ and the non-negative orthant of z intersects only at z. Denote the corresponding point of z on $\mathcal{F}_1$ as $z' = t - z$. An equivalent characterization is that $\mathcal{F}_1$ and the non-positive orthant of z' intersects only at z'. We can therefore consider a special type of z' whose coordinates are zero except for i, j-th entry. If z' further lies on a surface Q with $q_{ij} > 0$, then the corresponding z will be a PO of $\mathcal{F}_{q-1}$.

When C2 does not hold, there exists a row of $Q_\varnothing$ which contains both positive and negative entries. We assume w.l.o.g. the first row of $Q_\varnothing$ satisfies this condition. By the above observation, we can find two PO of $\mathcal{F}_{q-1}$ that correspond to $Q_\varnothing$ and Q_1, respectively. This finishes the proof as desired. $\square$

B.3 Proof of Theorem 3.8

Proof of Theorem 3.8. We only prove for the case of $q = 1$, and the proof for $q = k - 1$ can be done similarly. Suppose P (whose coordinate is v) lies at the intersection of two surfaces defined by Q and Q'. We denote the intersection of Q and Q' as $\mathcal{S}$, which is a non-linear manifold with dimension $(k - 2)$. Since P is a relative interior point of the PF, there exists some $\varepsilon > 0$, such that any point in $S \cap B_\varepsilon(P)$ is a PO. We will show there exists some $P' \in S \cap B_\varepsilon(P)$ (whose coordinate is v'), such that the gradients at P' w.r.t. the two surfaces disagree, i.e.,

$$Q\sqrt{v'} \neq Q'\sqrt{v'}. \tag{44}$$

To see why this is true, note that Q' can be expressed as $Q' = DQD$, where D is a diagonal matrix whose diagonal entries are either 1 or -1. As a consequence, the set

$$\{v \mid Qv = Q'v\} \tag{45}$$

is a subspace of $\mathbb{R}^k$ whose dimension is at most $(k - 2)$, so it cannot fully contain a non-linear manifold whose dimension is $(k - 2)$. This finishes the proof as desired. $\square$

B.4 Proof of Theorem 3.8 without assumptions ($k = 3$)

Here we provide a proof for Theorem 3.8 under $k = 3$, without relying on Assumptions 3.6 and 3.7.

Proof. We prove the necessity of C1 and C2 separately.

C1 is necessary. Suppose C1 is not true. We assume w.l.o.g. that $\langle \hat{y}_1, \hat{y}_2 \rangle < 0, \langle \hat{y}_1, \hat{y}_3 \rangle > 0, \langle \hat{y}_2, \hat{y}_3 \rangle > 0$. Now we can write

$$Q = \begin{bmatrix} q_{11} & q_{12} & -q_{13} \\ q_{21} & q_{22} & -q_{23} \\ -q_{31} & -q_{32} & q_{33} \end{bmatrix}, \tag{46}$$

where $q_{ij} > 0$ for all pairs of (i, j). We also have

$$q_{11}q_{23} > q_{12}q_{13}, \quad q_{22}q_{13} > q_{12}q_{23}, \quad q_{33}q_{12} > q_{13}q_{23}. \tag{47}$$

The boundary of the feasible region is formed by $E_\varnothing, E_1, E_2$. Our goal is to find a point $I \in E_\varnothing \cap E_1$, such that 1) I is a PO of the feasible region; 2) the gradients at I w.r.t. $E_\varnothing$ and E_1 disagree. We write the coordinate of I as $(x, y = q_{13}^2/s, z = q_{12}^2/s)^\top$, where

$$q_{11}x + \frac{q_{22}q_{13}^2 + q_{33}q_{12}^2 - 2q_{12}q_{23}q_{31}}{s} = 1, \tag{48}$$

and $s < q_{11}q_{23}^2 + q_{22}q_{13}^2 + q_{33}q_{12}^2 - 2q_{12}q_{23}q_{31}$.

It is straightforward to see that 2) can be satisfied as long as $x \neq 0$. Therefore, we will focus on the realization of 1). This further requires two conditions: i) I is a PO w.r.t. $E_\varnothing$ and E_1, meaning that the normal vectors are non-negative; ii) I is a PO w.r.t. E_2, meaning that the non-negative orthant at I does not intersect with E_2.

Some simple calculations yield that i) is equivalent to

$$q_{22}\sqrt{y} \geq q_{23}\sqrt{z} + q_{12}\sqrt{x} \quad \text{and} \quad q_{33}\sqrt{z} \geq q_{23}\sqrt{y} + q_{13}\sqrt{x}. \tag{49}$$

Setting $x = 0$, we have

$$q_{22}\sqrt{y} > q_{23}\sqrt{z} \quad \text{and} \quad q_{33}\sqrt{z} > q_{23}\sqrt{y} \tag{50}$$

by Equation (47). Therefore, it suffices to show the non-negative orthant at $(0, q_{13}^2/s, q_{12}^2/s)^\top$ ($s = q_{22}q_{13}^2 + q_{33}q_{12}^2 - 2q_{12}q_{23}q_{31}$) does not intersect with E_2, and by continuity we can find some $x_0 > 0$ such that i) and ii) hold simultaneously.

To prove the above claim, suppose $a \geq 0, b \geq q_{13}^2/s, c \geq q_{12^2}/s$. We will show $(a, b, c)^\top$ must lie above the surface defined by E_2, i.e.,

$$f(a, b, c) = q_{11}a + q_{22}b + q_{33}c - 2q_{12}\sqrt{ab} - 2q_{13}\sqrt{ac} + 2q_{23}\sqrt{bc} > 1. \tag{51}$$

Let $a^* = \frac{q_{12}\sqrt{b}+q_{13}\sqrt{c}}{q_{11}}$. We have

$$
\begin{aligned}
f(a,b,c) &\geq f(a^*,b,c) \\
&= \left(q_{22} - \frac{q_{12}^2}{q_{11}}\right)b + \left(q_{33} - \frac{q_{13}^2}{q_{11}}\right)c + 2\left(\frac{q_{11}q_{23} - q_{12}q_{13}}{q_{11}}\right)\sqrt{bc} \\
&\geq \frac{q_{11}(q_{22}q_{13}^2 + q_{33}q_{12}^2 + 2q_{12}q_{23}q_{31}) - 4q_{12}^2 q_{13}^2}{sq_{11}} \\
&> \frac{q_{11}(q_{22}q_{13}^2 + q_{33}q_{12}^2 - 2q_{12}q_{23}q_{31})}{sq_{11}} \\
&= 1.
\end{aligned}
$$

As a consequence, by choosing a small x_0, we can guarantee both 1) and 2). I is therefore a PO that cannot be achieved via scalarization.

C2 **is necessary.** Suppose C2 is not true. We can write

$$
Q = \begin{bmatrix} q_{11} & -q_{12} & -q_{13} \\ -q_{21} & q_{22} & -q_{23} \\ -q_{31} & -q_{32} & q_{33} \end{bmatrix}, \tag{52}
$$

where $q_{ij} > 0$ for all pairs of (i,j). The inner boundary of $\mathcal{F}_1$ is formed by E_1, E_2, E_3, and the coordinates of their intersection I is given by $v = (q_{23}^2/r, q_{13^2}/r, q_{12}^2/r)^\top$, where $r = q_{11}q_{23}^2 + q_{22}q_{13}^2 + q_{33}q_{12}^2 + 2q_{12}q_{23}q_{31}$. Our goal is to show the corresponding I' on $\mathcal{F}_2$ (i.e., $t - v$) is a PO which cannot be achieved by scalarization.

To see why this is true, first note that the gradients at I w.r.t. the three surfaces disagree, since $q_{ij} > 0$ for all pairs of (i,j). Now, to demonstrate I' is a PO of $\mathcal{F}_2$, it suffices to show the non-positive orthant of I intersects with $\mathcal{F}_1$ only at I. By symmetry, it suffices to show that I does not dominate any point on E_1.

To prove the above claim, suppose $a \leq q_{23}^2/r, b \leq q_{13}^2/r, c \leq q_{12}^2/r$. We will show, $J = (a,b,c)^\top$ must lie below the surface defined by E_1 unless J equals I. In fact, let

$$
g(a,b,c) = q_{11}a + q_{22}b + q_{33}c + 2q_{12}\sqrt{ab} + 2q_{13}\sqrt{ac} - 2q_{23}\sqrt{bc}. \tag{53}
$$

We have

$$
\begin{aligned}
g(a,b,c) &\leq g(q_{23}^2/r, b, c) \\
&\leq g(q_{23}^2/r, q_{13}^2/r, c) \\
&\leq g(q_{23}^2/r, q_{13}^2/r, q_{12}^2/r) \\
&= 1,
\end{aligned}
$$

where the equalities hold iff $I = J$. This finishes the proof as desired. $\qquad\square$

B.5 Proof of Theorem 3.10

We first present several useful lemmas regarding non-negative matrices.

Lemma B.7. *Let A be a non-negative, irreducible and real symmetric matrix, then there only exists one eigenvector v of A (up to scaling), such that $v > 0$.*

Proof. Since A is non-negative and irreducible, by Perron-Frobenius theorem (Shaked-Monderer and Berman, 2021), the eigenvector associated with the largest eigenvalue $\rho(A)$ (a.k.a. *Perron root*) can be made positive, and is unique up to scaling. We denote it as v.

We will show the eigenvectors associated with other eigenvalues cannot be positive. In fact, assume $u > 0$ is an eigenvector associated with $\lambda < \rho(A)$. Since A is symmetric, we have

$$
\rho(A)v^\top u = v^\top A^\top u = v^\top A u = \lambda v^\top u. \tag{54}
$$

This is a contradiction since $v^\top u > 0$ and $\rho(A) > \lambda$. $\qquad\square$

Lemma B.8. *Let A be a non-negative and real symmetric matrix. Suppose v is an eigenvector of A such that $v > 0$, then v corresponds to the largest eigenvalue of A.*

Proof. Consider the normal form (Varga, 1999) of A:

$$PAP^\top = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1h} \\ 0 & B_{22} & \cdots & B_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_{hh} \end{bmatrix} \triangleq B, \tag{55}$$

where P is a permutation matrix and B_{ii} ($i \in [h]$) are irreducible. It is straightforward to see that B is non-negative. Furthermore, since a permutation matrix is an orthogonal matrix, B is also symmetric (implying that $B_{ij} = 0$ for $i \neq j$) and has the same spectrum as A.

Now suppose $Av = \lambda v$ for some $v > 0$. Since $B = PAP^\top$, we have $B\psi = \lambda\psi$ with $\psi = Pv > 0$. Denote

$$\psi = \begin{pmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_n \end{pmatrix}, \tag{56}$$

where each ψ_i has the same dimension as B_{ii}. Now we have

$$B_{ii}\psi_i = \lambda\psi_i, \quad \forall i \in [h]. \tag{57}$$

Since B_{ii} is non-negative, irreducible and real symmetric, by Lemma B.7, ψ_i corresponds to the largest eigenvalue of B_{ii}. Note B is a block diagonal matrix, so the spectrum of B is essentially the union of the spectrum of B_{ii}. As a consequence,

$$\lambda_{\max}(A) = \lambda_{\max}(B) = \max_{i \in [h]} \lambda_{\max}(B_{ii}) = \lambda. \tag{58}$$

$\square$

Proof of Theorem 3.10. We prove the sufficiency of `C1` and `C2` separately.

`C1` **is sufficient.** We have demonstrated in Lemma 3.4 that the Pareto front belongs to a single surface E, whose gram matrix is doubly non-negative. Let $p = (a_1, \cdots, a_k)^\top$ be a PO of E, and denote $t = \sqrt{p}$. We will show that the tangent plane at this point lies above the feasible region, implying that it can be achieved by scalarization.

Concretely, the tangent plane is given by

$$\sum_{i \in [k]} \frac{\sum_{j \in [k]} q_{ij}\sqrt{a_j}}{\sqrt{a_i}} v_i = 1. \tag{59}$$

To show it lies above the feasible region, it suffices to prove that it lies above every surface in $\mathcal{E}$, i.e.,

$$\sum_{i \in [k]} \frac{\sum_{j \in [k]} q_{ij}\sqrt{a_j}}{\sqrt{a_i}} v_i \leq \sum_{i \in [k]} q_{ii}^{i_1 \cdots i_l} v_i + 2 \sum_{1 \leq i < j \leq k} q_{ij}^{i_1 \cdots i_l} \sqrt{v_i v_j}, \tag{60}$$

where $q_{ij}^{i_1 \cdots i_l}$ represents the (i, j)-th entry of $Q_{i_1 \cdots i_l}$. Note Equation (60) can be treated as a quadratic form w.r.t. $\sqrt{v}$, so it suffices to prove the corresponding matrix is PSD. Writing compactly, it suffices to show

$$Q_{i_1 \cdots i_l} - \mathrm{diag}\{Qt \oslash t\} = D_{i_1 \cdots i_l}(Q - \mathrm{diag}\{Qt \oslash t\})D_{i_1 \cdots i_l} \tag{61}$$

is PSD. As a consequence, we only need to show the positive semi-definiteness of the following matrix

$$T := Q - \mathrm{diag}\{Qt \oslash t\}. \tag{62}$$

Since p is a PO of E, we have $Qt \geq 0$. Let $w = Qt$, we have

$$T = G^{-1} - \mathrm{diag}\{w \oslash Gw\}. \tag{63}$$

Assume w.l.o.g. that $w > 0$ (otherwise we simply remove the corresponding rows and columns in G^{-1}). Denote

$$R = \sqrt{\text{diag}\{w \oslash Gw\}} \quad \text{and} \quad \phi = \sqrt{w \odot Gw} > 0. \tag{64}$$

We have

$$
\begin{aligned}
RGR\phi &= RG\sqrt{\text{diag}\{w \oslash Gw\}}\sqrt{w \odot Gw} \\
&= RGw \\
&= \sqrt{\text{diag}\{w \oslash Gw\}}Gw \\
&= \sqrt{w \odot Gw} \\
&= \phi.
\end{aligned}
$$

By Lemma B.8, we have $\rho(RGR) = 1$. Since RGR is positive definite, we have

$$I \succeq RGR \succ 0. \tag{65}$$

This further implies that $R^{-2} \succeq G$, and

$$T = G^{-1} - R^2 \succeq 0. \tag{66}$$

C2 **is sufficient.** We have demonstrated in Lemma 3.5 that the Pareto front belongs to a single surface $I_\varnothing$. Here we will further show $I_\varnothing$ is concave, and apply the supporting hyperplane theorem (Boyd et al., 2004, Section 2.5.2) to conclude the proof. In fact, $E_\varnothing$ is parametrized by

$$\sum_{i \in [k]} q_{ii} v_i + 2 \sum_{1 \le i < j \le k} q_{ij} \sqrt{v_i v_j} = 1. \tag{67}$$

Since the first term is a summation of linear functions, $f(x,y) = \sqrt{xy}$ is a concave function and that $q_{ij} \ge 0$, we conclude that $E_\varnothing$ is convex, and $I_\varnothing$ is concave. This finishes the proof as desired. □

C   A polynomial algorithm for checking C1

We explain how Algorithm 1 works.

Essentially, the goal is to find a set of flipping signs $\{s_i\}_{i \in [k]} \in \{+, -\}$ that ensure positive correlation between all pairs in $\{s_i \hat{y}_i\}_{i \in [k]}$ (except for those pairs that are not correlated). We state an observation which is critical for our algorithm—given a pair $\hat{y}_i$ and $\hat{y}_j$, whether their signs are the same or opposite can be determined by whether $\hat{y}_i$ and $\hat{y}_j$ is positively or negatively correlated. Thus, given a determined s_i, the signs of its subsequent predictors s_j, $j > i$ can be uniquely determined by evaluating $\langle \hat{y}_i, \hat{y}_j \rangle$. This allows us to determine the signs for the predictors s_i in ascending order, and check for potential conflicts in the mean time.

We state one loop invariant in our algorithm—when entering iteration i in the outer loop, s_i has been determined by its *preceding* predictors and no conflict has occurred, or it has remained undetermined (i.e., None) all the way because it is not correlated with any preceding predictors. At this point, what is left to be done is to examine the *subsequent* predictors. If s_i is undetermined (lines 2-7), we attempt to use its subsequent predictors to determine it (line 5), and break out of the loop once it gets determined. Now that we make sure that s_i is determined after line 9, we can proceed to check the remaining subsequent predictors, to either determine their signs (lines 12-13) or check for potential conflicts (lines 14-15).

We comment that caution needs to be taken for the pairs that are not correlated (i.e., $\langle \hat{y}_i, \hat{y}_j \rangle = 0$). In this case, one cannot determine the sign of s_j by the relationship between these two predictors; basically, the sign has no influence on their correlation. Thus, the predictor $\hat{y}_j$ retains the flexibility to be determined by other predictors other than $\hat{y}_i$. Additionally, it is possible that at the end of line 9 s_j is still undetermined; this only occurs when $\hat{y}_i$ is not correlated with any other optimal predictor. In this case, t will take the value of $k + 1$ and the following loop (lines 10-15) will not be executed.

D   Additional experimental details and results

D.1   Additional experimental details

Dataset. The SARCOS dataset (Vijayakumar and Schaal, 2000) consists of the configurations (position, velocity, acceleration) of the robotic arms, and the problem is to predict the torque of the

Algorithm 1: An $O(k^2)$ algorithm of checking `C1`

Input: Optimal predictors $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k$
Output: `True` if `C1` is met, and `False` otherwise
Initialize: $s_1 = 1$, $s_2 = \cdots = s_k =$ `None`
Define: $\operatorname{sgn}(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x < 0 \\ \texttt{None}, & \text{if } x = 0 \end{cases}$

1 **for** $i = 1$ **to** $k - 1$ **do**
    ▷ Check whether s_i has been determined by its preceding optimal predictors $\{\hat{y}_1, \ldots, \hat{y}_{i-1}\}$; if not, determine it using the subsequent ones.
2   **if** $s_i =$ `None` **then**
3   **for** $j = i + 1$ **to** k **do**
4   **if** $s_j \neq$ `None` ***and*** $\operatorname{sgn}(\langle \hat{y}_i, \hat{y}_j \rangle) \neq$ `None` **then**
5                 $s_i \leftarrow s_j \cdot \operatorname{sgn}(\langle \hat{y}_i, \hat{y}_j \rangle)$
6   **break**
7         $t \leftarrow j + 1$
8   **else**
9         $t \leftarrow i + 1$
    ▷ Now that s_i has been determined, we proceed to check the remaining optimal predictors to either determine the sign for them or check for conflicts.
10   **for** $j = t$ **to** k **do**
11   **if** $\operatorname{sgn}(\langle \hat{y}_i, \hat{y}_j \rangle) \neq$ `None` **then**
12   **if** $s_j =$ `None` **then**
13                 $s_j \leftarrow s_i \cdot \operatorname{sgn}(\langle \hat{y}_i, \hat{y}_j \rangle)$
14   **else if** $s_i \cdot s_j \cdot \operatorname{sgn}(\langle \hat{y}_i, \hat{y}_j \rangle) = -1$ **then**
15   **return** `False`

16 **return** `True`

---

respective arm given its triplet configuration. Since our study concentrates on the training procedure and does not concern generalization, we conduct the experiments on the training set alone, where the size of the training set of the SARCOS dataset (Vijayakumar and Schaal, 2000) is $40,000$. Regression tasks normally benefit from standardization as a pro-processing step (Hastie et al., 2009), so we standardize the dataset to zero mean and unit variance.

Tasks. Following our theoretical analysis in Section 3, we study the scenario where $q = 1$. In this case, it is straightforward to see that the minimal k that may lead to a violation of `C1` is 3. To show the violation of `C1` leading to the failure of scalarization exploring the Pareto Front, we constrain our study to $k \geq 3$; furthermore, for ease of visualization, we focus on $k = 3$. In our experiments, we take the arms 3, 4, and 5 for which do not satisfy `C1`.

Network. We use a two-layer linear network for the reason explained in Section 2. We merge the bias term b into the weight matrix W by adding an additional all-one column to the input X. The input dimension of the network is 22 (seven (position, velocity, acceleration) triplets plus the dummy feature induced by b), the hidden size is $q = 1$, and the output size is 1 (i.e., the predicted torque). The same network architecture is used for both experiments in scalarization and that in SMTOs.

Implementation of linear scalarization. For linear scalarization, we uniformly sample $100,000$ sets of convex coefficients from a three-dimensional simplex (i.e., a tetrahedron). Concretely, we first sample $m_1, m_2 \overset{i.i.d.}{\sim} U(0, 1)$, and then craft $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) = (\min(m_1, m_2), \max(m_1, m_2) - \min(m_1, m_2), 1 - \max(m_1, m_2) + \min(m_1, m_2))$ as the weights. For a given set of convex coefficients, we calculate the optimal value directly based on the analysis in Section 2 instead of performing actual training. More specifically, the optimum of the scalarization objective is given by $\hat{Y}_{q,\Lambda}\Lambda := \sum_{i=1}^{q} \sigma_i u_i v_i^\top$, from which we can compute the corresponding MSE for each task.

SMTOs. We consider two state-of-the-art SMTOs, MGDA (Désidéri, 2012) and MGDA-UB (Sener and Koltun, 2018). MGDA is an optimization algorithm specifically devised for handling multiple objectives concurrently. It utilizes a composite gradient direction to facilitate simultaneous progress across all objectives. MGDA-UB is an efficient variant of MGDA that focuses on maximizing the minimum improvement among all objectives. It is a gradient-based multi-objective optimization algorithm aimed at achieving balanced optimization outcomes across all objectives.

Implementation of SMTOs. Our code is based on the released implementation[5] of MGDA-UB, which also includes the code for MGDA. We apply their code on the SARCOS dataset. For both methods, we use vanilla gradient descent with a learning rate of 0.5 for 100 epochs, following the default choice in the released implementation. We comment that early stopping can also be adopted, i.e., terminate once the minimum norm of the convex hull of gradients is smaller than a threshold, for which we set to be 10^{-3}.

D.2 Additional experiments on random initialization

To eliminate the influence of random initialization, we perform 300 trials for each algorithm using different random seeds. We filter out solutions whose maximum MSE is larger than 1 for clearer presentation, which leads to 198 solutions for MGDA and 240 solutions for MGDA-UB. We plot all these solutions in Figure 4. We see that MGDA and MGDA-UB are consistently capable of finding balanced solutions regardless of the random seed.

(a) MGDA with multiple initializations   (b) MGDA-UB with multiple initializations
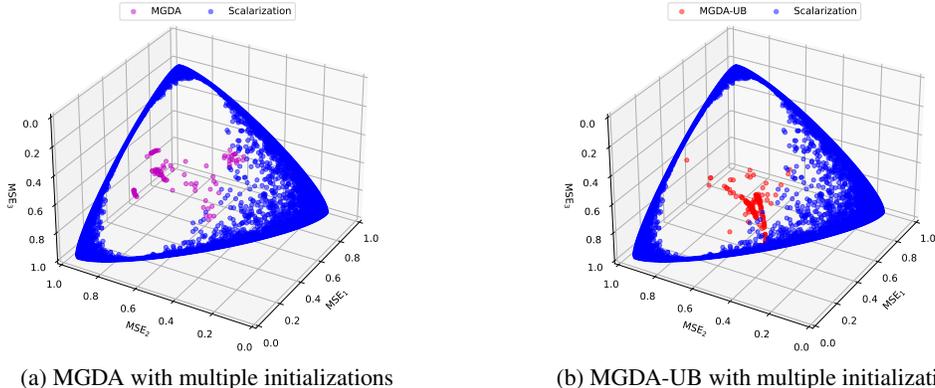
Figure 4: More solutions of MGDA and MGDA-UB obtained by varying the initializations. Both algorithms tend to find solutions located at the interior of the Pareto front.

D.3 Additional experiments on randomization

To verify the effectiveness of randomization as discussed in Section 4.2, we plot the region achievable by randomized linear scalarization. Specifically, according to Equation (13), the region can be expressed as the collection of the convex combination of two points from the feasible region. To this end, we randomly sample 100,000 weight pairs (the same number as in Figure 3). For each weight pair (w_1, w_2), we uniformly draw $t \sim U(0, 1)$ and get two corresponding optimal networks f_1 and f_2 by SVD. For each sample, with probability t, the model uses f_1, otherwise f_2, to calculate the MSE. The final result is demonstrated in Figure 5. It is straightforward to see that randomization allows scalarization to trace out the PF since there is no hole within the blue region, thus validating our theoretical analysis. We additionally comment that randomization convexifies the feasible region, as such, the solutions found by MGDA and MGDA-UB are dominated.

---

[5]https://github.com/isl-org/MultiObjectiveOptimization
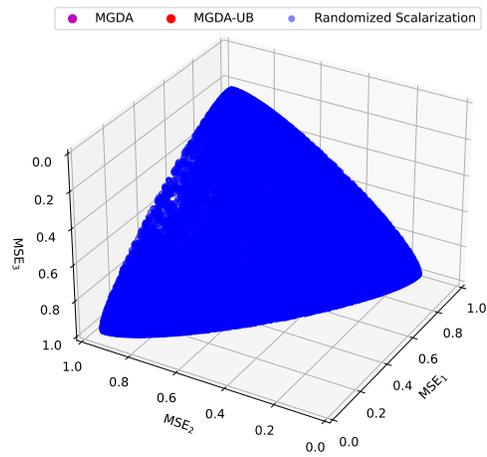
Figure 5: The region achievable by randomized linear scalarization