A Preliminary Extension of LoRT for "No Negative Transfer"

To mitigate the risk of negative transfer in multi-source settings—especially when some source domains are poorly aligned with the target—we present a preliminary extension of the LoRT framework as a supplementary theoretical exploration, rather than a core contribution. Motivated by the empirical observation that naive aggregation of all sources can degrade performance under significant divergence, we propose a weighted regularized least squares formulation that adaptively adjusts the influence of each domain. This design preserves LoRT's structural principles while improving robustness in heterogeneous environments. The accompanying derivation follows the spirit of He et al. (2024), while being carefully tailored to the assumptions and architecture of the LoRT framework. We also find the adaptive weighting idea of Duan & Wang (2023) particularly inspiring. ***Although we provide a preliminary exploration of how LoRT can be extended to mitigate negative transfer, developing a complete and principled solution lies beyond the scope of this submission and is left for future work.***

1. Model and Results

The core idea is to assign a non-negative weight w_k to each domain (with $k = 0$ denoting the target), and jointly estimate both the model parameters and these weights. Each w_k affects both the empirical loss and the regularization terms. In particular, w_k appears in the data-fitting term and in the discrepancy penalty $\|\underline{\mathbf{W}}^{(0)} - \underline{\mathbf{W}}^{(k)}\|_\star$, thus modulating how much each source influences the estimation of $\underline{\mathbf{W}}^{(0)}$.

We fisrt show that the estimation error of $\underline{\mathbf{W}}^{(0)}$ can be decomposed into two interpretable components (Proposition 1): (i) A variance term $\kappa_{\mathbf{w}} \cdot \frac{rd_1d_3}{N}$, reflecting the effective sample complexity; (ii) A bias term $\bar{h}_{\mathbf{w}}\sqrt{d_1/N_T}$, reflecting the source-target dissimilarity. This decomposition highlights a fundamental tradeoff between variance reduction (via data reuse) and bias increase (due to misalignment). Our analysis formalizes this tradeoff and characterizes when and how properly chosen weights can provably mitigate negative transfer.

The key insight is summarized by the following informal result of Theorem 2:

Theorem 1 (Informal)**.** *Suppose the source informativeness levels $\{h_k\}_{k=1}^K$ are known. Under the normalization constraint $\sum_{k=0}^K w_k \cdot N_k/N = 1$ with $w_k \geq 0$, and an appropriate choice of weights $\{w_k\}$ that down-weight less informative sources, the target estimation error satisfies, with high probability,*

$$\|\hat{\underline{\mathbf{W}}}^{(0)} - \underline{\mathbf{W}}_\star^{(0)}\|_{\mathrm{F}}^2 \lesssim \frac{rd_1d_3}{N_T}. \tag{1}$$

In particular, when $K = 1$, we obtain a sharper bound:

$$\|\hat{\underline{\mathbf{W}}}^{(0)} - \underline{\mathbf{W}}_\star^{(0)}\|_{\mathrm{F}}^2 \lesssim \frac{rd_1d_3}{N} + \min\left\{h_1\sqrt{\frac{d_1}{N_T}},\ \frac{rd_1d_3}{N_T}\right\}. \tag{2}$$

This result confirms that, under appropriately selected weights, the estimator is provably immune to negative transfer, even when some source domains are irrelevant or misaligned with the target.

1.1. A Weighted Extension of LoRT

Specifically, to estimate the target parameter $\underline{\mathbf{W}}_\star^{(0)}$, we formulate and solve a joint optimization problem. This involves computing $\hat{\underline{\underline{\mathbf{W}}}} = (\hat{\underline{\mathbf{W}}}^{(0)}, \ldots, \hat{\underline{\mathbf{W}}}^{(K)})$, where we use $\hat{\underline{\mathbf{W}}}^{(0)}$ as our estimate for $\underline{\mathbf{W}}_\star^{(0)}$:

$$\hat{\underline{\underline{\mathbf{W}}}} \in \underset{\underline{\underline{\mathbf{W}}}}{\operatorname{argmin}} \underbrace{\frac{1}{2N}\sum_{k=0}^K w_k \sum_{i=1}^{N_k}(y_i^{(k)} - \langle \underline{\mathbf{W}}^{(k)}, \underline{\mathbf{X}}_i^{(k)}\rangle)^2 + \lambda_0\sqrt{\sum_{k=0}^K \frac{N_k}{N} w_k^2}\|\underline{\mathbf{W}}^{(0)}\|_\star + \lambda_1 \sum_{k=1}^K w_k\|\underline{\mathbf{W}}^{(0)} - \underline{\mathbf{W}}^{(k)}\|_\star}_{\text{(a) Weighted data fitting and regularization}}, \tag{3a}$$

$$\text{s.t.} \quad \underbrace{\frac{1}{N_T}\|\mathfrak{X}^{*(0)}(\boldsymbol{y}^{(0)} - \mathfrak{X}^{(0)}(\underline{\mathbf{W}}^{(0)}))\|_{\text{tsp}} \leq \lambda_T}_{\text{(b) Gradient-based constraint}}. \tag{3b}$$

where $w_k \geq 0$ for $k = 0, 1, \ldots, K$ are the weights for teach tasks, which is required to be normalized $\sum_{k=0}^{K} w_k N_k / N = 1$. We explain each component of (3) below:

(a) The objective jointly optimizes the model parameters and domain weights. The first term represents the weighted empirical risk, where each w_k reflects the relative importance of domain k. The second and third terms regularize the target model via low-rank constraints: the first shrinks $\underline{\mathbf{W}}^{(0)}$ toward a low-rank structure, while the second encourages similarity to source parameters, scaled by their weights.

(b) This constraint is inspired by the generalized tensor Dantzig selector (Wang et al., 2019), ensuring the identifiability of $\underline{\mathbf{W}}^{(0)}$ by bounding the norm of the gradient of the target loss. When there is no noise, the left-hand side vanishes at the true parameter. The slack variable λ_T accounts for possible label noise.

This formulation generalizes the first step of the original LoRT framework in Eq. (6) of the main paper, which solves

$$\min_{\underline{\mathbf{W}}} \frac{1}{2N} \sum_{k=0}^{K} \sum_{i=1}^{N_k} (y_i^{(k)} - \langle \underline{\mathbf{W}}^{(k)}, \underline{\mathbf{X}}_i^{(k)} \rangle)^2 + \lambda_0 \left(\|\underline{\mathbf{W}}^{(0)}\|_\star + \sum\nolimits_{k=1}^{K} a_k \|\underline{\mathbf{W}}^{(0)} - \underline{\mathbf{W}}^{(k)}\|_\star \right). \tag{4}$$

In this earlier formulation, the influence of each source is fixed by pre-specified constants a_k. In contrast, our extended formulation (3) introduces learnable weights w_k to adaptively determine the contribution of each domain. This enhancement allows the model to suppress uninformative or misleading sources and therefore enables robustness to negative transfer.

To understand when this approach leads to provable performance guarantees, we next present a theoretical analysis. Our main result decomposes the estimation error into two interpretable components: a variance-like term reflecting effective sample size, and a bias-like term quantifying the mismatch between sources and the target. We then show how the choice of weights w_k governs the tradeoff between these two terms, and derive both optimal and data-driven strategies for selecting them.

1.2. When Informative Level h_k Are Known

We begin our theoretical analysis by considering an idealized setting in which an upper bound on the rank r of the true parameter, as well as the informativeness of each source domain—quantified by $h_k := |\underline{\mathbf{W}}_\star^{(0)} - \underline{\mathbf{W}}_\star^{(k)}|_\star$—is known. Under this setting, we consider the following parameter space:

$$\mathbb{W}(r, \mathbf{h}) := \left\{ (\underline{\mathbf{W}}^{(0)}, \{\underline{\mathbf{W}}^{(k)}\}_{k=1}^{K}) : r_{\mathrm{t}}(\underline{\mathbf{W}}^{(0)}) \leq r,\ \|\underline{\mathbf{W}}^{(0)} - \underline{\mathbf{W}}^{(k)}\|_\star \leq h_k,\ \forall k \in [K] \right\}. \tag{5}$$

Within this space, we establish an upper bound on the target estimation error when using appropriately chosen weights.

Proposition 1. *Suppose Assumptions 3.3 and 3.4 hold, and the sample sizes satisfy $N_S > N_T$ and $r d_1 d_3 / N_T + K \bar{h}_{\mathbf{w}} \sqrt{d_1 / N_T} = o(1)$. Assume further that the normalized weights $\{w_k\}$ either satisfy $w_k \geq \underline{w}$ for some small constant $\underline{w} > 0$, or are set to zero, and that the regularization parameters are chosen by*

$$\lambda_0 = c_0 \left[\left(\sum_{k=0}^{K} \frac{N_k}{N} w_k^2 \right)^{-\frac{1}{2}} \left(\frac{\bar{h}_{\mathbf{w}}^2 d_1}{(r d_3)^2 N_T} \right)^{\frac{1}{4}} + \left(\frac{d_1}{N} \right)^{\frac{1}{2}} \right], \lambda_1 = c_0 \frac{N_S}{N} \left(\frac{d_1}{N_T} \right)^{\frac{1}{2}}, \text{ and } \lambda_T = c_1 \left(\frac{d_1}{N_T} \right)^{\frac{1}{2}}.$$

Then, the solution $\hat{\underline{\mathbf{W}}}^{(0)}$ to Problem (3) satisfies

$$\|\hat{\underline{\mathbf{W}}}^{(0)} - \underline{\mathbf{W}}_\star^{(0)}\|_{\mathrm{F}}^2 \lesssim \kappa_{\mathbf{w}} \cdot \frac{r d_1 d_3}{N} + c_\epsilon \cdot \bar{h}_{\mathbf{w}} \cdot \sqrt{\frac{d_1}{N_T}}, \quad \textit{w.h.p.}, \tag{6}$$

where

$$\kappa_{\mathbf{w}} := \sum_{k=0}^{K} \frac{N_k w_k^2}{N} \text{ and } \bar{h}_{\mathbf{w}} := \sum_{k=1}^{K} \frac{N_k w_k}{N} h_k,$$

and c_ϵ is a universal constant depending on the noise level.

This result offers two insights:

- The first term, $\kappa_{\mathbf{w}} \cdot rd_1d_3/N$, behaves like a variance term. It reflects the effective sample size after weighting. Notably, $\kappa_{\mathbf{w}} \geq 1$ under the normalization constraint, and achieves its minimum value 1 when all weights are equal. Thus, $\kappa_{\mathbf{w}}$ captures the price of weighting and partial source usage.
- The second term, $c_\epsilon \cdot \bar{h}_{\mathbf{w}} \cdot \sqrt{d_1/N_T}$, represents the bias due to mismatched sources. By down-weighting sources with large h_k, this bias can be reduced.

Together, these two terms illustrate a bias–variance tradeoff: increasing source usage lowers variance but may increase bias. Therefore, optimal weight design must balance this tradeoff.

To explore how this tradeoff depends on sample sizes and informativeness levels, we introduce the sample-adjusted weights:

$$w'_k := w_k N_k/N, \quad \forall k = 0, 1, \ldots, K$$

and define the simplex:

$$\Delta^K := \left\{ \mathbf{w}' \in \mathbb{R}^{K+1} : \sum_{k=0}^{K} w'_k = 1,\ w'_k \geq 0 \right\}.$$

Under this transformation, minimizing the error bound becomes the following constrained optimization:

$$\min_{\mathbf{w}' \in \Delta^K} \left\{ \mathcal{Q}(\mathbf{w}') := \frac{rd_1d_3}{N_T}(w'_0)^2 + \sum_{k=1}^{K} \left(\frac{rd_1d_3}{N_S}(w'_k)^2 + c_\epsilon h_k \sqrt{\frac{d_1}{N_T}} w'_k \right) \right\}, \tag{7}$$

which is a quadratic program over a simplex and can be efficiently solved (Duchi et al., 2008).

The next result gives a closed-form expression for the optimal weights.

Theorem 2. *Let $h_{(j)}$ denote the j-th smallest value among $\{h_k\}_{k=1}^K$, and let $K^\sharp$ denote the number of strictly positive weights in the optimal solution. Then the optimal weights minimizing the bound in Eq. (6) are given by:*

$$w'_0 = \frac{N_T}{N_T + K^\sharp N_S} \left(1 + \sum_{j=1}^{K^\sharp} \frac{h_{(j)}}{2} \cdot \frac{c_\epsilon \sqrt{d_1/N_T}}{rd_1d_3/N_S} \right), \tag{8}$$

$$w'_k = \max \left\{ \frac{N_S}{N_T + K^\sharp N_S} \left(1 + \sum_{j=1}^{K^\sharp} \frac{h_{(j)}}{2} \cdot \frac{c_\epsilon \sqrt{d_1/N_T}}{rd_1d_3/N_S} \right) - \frac{h_k}{2} \cdot \frac{c_\epsilon \sqrt{d_1/N_T}}{rd_1d_3/N_S},\ 0 \right\}, \quad k \in [K]. \tag{9}$$

Moreover, the following properties hold:

(I) (No worse than target-only baseline) *The vector $\mathbf{w}' = (1, 0, \ldots, 0)^\top$, which corresponds to using only the target domain, is always feasible to the optimization problem (7). Hence, the optimal estimator satisfies with high probability:*

$$\|\underline{\hat{\mathbf{W}}}^{(0)} - \underline{\mathbf{W}}_\star^{(0)}\|_{\mathrm{F}}^2 \lesssim \frac{rd_1d_3}{N_T}. \tag{10}$$

(II) (Special case: *$K = 1$) When $K = 1$, the closed-form solution simplifies to:*

$$w'_0 = 1 - w'_1 \ \text{ and } \ w'_1 = \max \left\{ \frac{N_S}{N_T + N_S} - \frac{N_T}{N_T + N_S} \cdot \frac{h_1}{2} \cdot \frac{c_\epsilon \sqrt{d_1/N_T}}{rd_1d_3/N_S},\ 0 \right\} \tag{11}$$

and under the assumptions of Proposition 1, the corresponding error satisfies

$$\|\underline{\hat{\mathbf{W}}}^{(0)} - \underline{\mathbf{W}}_\star^{(0)}\|_{\mathrm{F}}^2 \lesssim \frac{rd_1d_3}{N} + \min \left\{ c_\epsilon h_1 \cdot \sqrt{\frac{d_1}{N_T}},\ \frac{rd_1d_3}{N_T} \right\}. \tag{12}$$

This finding demonstrates the adaptive nature of optimal weights: they are responsive to both individual source characteristics h_k and the comparative informativeness across different sources. Sources that poorly align with the target (indicated by large h_k values) receive automatically reduced weights, thereby safeguarding against negative transfer effects.

1.3. Weight Choice When h_k Are Unknown

The optimal weight design in Theorem 2 assumes prior knowledge of the source-target similarities h_k and the rank r of $\underline{\mathbf{W}}^{(0)}$. In practice, since these quantities are typically unknown, we adopt a data-driven approach inspired by He et al. (2024), estimating the weights $\{w'_k\}$ via the following convex optimization:

$$\min_{\mathbf{w}'} \left\{ \sum_{k=0}^{K} \frac{r_{\mathrm{t}}(\hat{\underline{\mathbf{W}}}_{\text{init}}^{(0)}) \cdot d_1 d_3}{N_k} (w'_k)^2 + \lambda_W \sum_{k=1}^{K} \|\hat{\underline{\mathbf{W}}}_{\text{init}}^{(k)} - \hat{\underline{\mathbf{W}}}_{\text{init}}^{(0)}\|_\star \cdot \sqrt{\frac{d_1}{N_T}} w'_k \right\} \quad \text{s.t.} \quad \sum_{k=0}^{K} w'_k = 1, \quad w'_k \geq 0. \tag{13}$$

Here, we approximate the rank r of $\underline{\mathbf{W}}^{(0)}$ using the initial estimate $\hat{\underline{\mathbf{W}}}_{\text{init}}^{(0)}$, and similarly approximate h_k via the discrepancy $\|\hat{\underline{\mathbf{W}}}_{\text{init}}^{(k)} - \hat{\underline{\mathbf{W}}}_{\text{init}}^{(0)}\|_\star$ between the target and each source. These initial estimates can be obtained using existing methods such as the tensor nuclear norm (TNN) estimator (Wang et al., 2019). The hyperparameter λ_W can be tuned through cross-validation on the target domain.

Problem (13) is a quadratic program with simplex constraints and can be efficiently solved via standard convex optimization techniques (Duchi et al., 2008). Once the weights $\{\hat{w}_k\}$ are obtained, we substitute them into the main estimation objective to compute the final estimator of the target parameter:

$$\begin{aligned} \left(\hat{\underline{\mathbf{W}}}^{(0)}, \{\hat{\underline{\mathbf{\Theta}}}^{(k)}\}_{k=1}^{K}\right) \in \underset{\underline{\mathbf{W}}^{(0)}, \{\underline{\mathbf{\Theta}}^{(k)}\}}{\operatorname{argmin}} \; & \frac{1}{N} \sum_{k=0}^{K} \hat{w}_k \sum_{i=1}^{N_k} \left(y_i^{(k)} - \langle \underline{\mathbf{X}}_i^{(k)}, \underline{\mathbf{W}}^{(0)} + \underline{\mathbf{\Theta}}^{(k)} \rangle \right)^2 \\ & + \lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} \hat{w}_k^2} \|\underline{\mathbf{W}}^{(0)}\|_\star + \lambda_1 \sum_{k=1}^{K} \hat{w}_k \|\underline{\mathbf{\Theta}}^{(k)}\|_\star \\ \text{s.t.} \quad & \frac{1}{N_T} \|\mathfrak{X}^{*(0)}(\boldsymbol{y}^{(0)} - \mathfrak{X}^{(k)}(\underline{\mathbf{W}}^{(0)}))\|_{\text{tsp}} \leq \lambda_T. \end{aligned} \tag{14}$$

In summary, this adaptive scheme provides a preliminary and still immature design for mitigating negative transfer. ***Developing a fully principled and robust algorithm that completely avoids negative transfer is beyond the scope of this ICML submission and is left for future work.***

2. Proofs of Proposition 1 and Theorem 2

2.1. Proof of Proposition 1

Similar to the proof of Theorem 4.2, we begin by introducing some key notations and transformations that will facilitate our analysis.

For each source task $k \in [K]$, let $\underline{\mathbf{\Theta}}_\star^{(k)} := \underline{\mathbf{W}}_\star^{(k)} - \underline{\mathbf{W}}_\star^{(0)}$ denote the difference between the ground truth parameters of the k-th source task and the target task, representing the model shift. We define:

$$\underline{\underline{\mathbf{\Theta}}}_\star = (\underline{\mathbf{\Theta}}_\star^{(0)}, \underline{\mathbf{\Theta}}_\star^{(1)}, \ldots, \underline{\mathbf{\Theta}}_\star^{(K)}) = (\underline{\mathbf{W}}_\star^{(0)}, \underline{\mathbf{W}}_\star^{(1)} - \underline{\mathbf{W}}_\star^{(0)}, \ldots, \underline{\mathbf{W}}_\star^{(K)} - \underline{\mathbf{W}}_\star^{(0)}) \in \mathbb{R}^{(K+1)\times d_1 \times d_2 \times d_3}. \tag{15}$$

Here, $\underline{\mathbf{W}}_\star^{(0)} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is the tensor parameter of the target task model, and $\underline{\mathbf{W}}_\star^{(k)} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is the parameter of the k-th source task model for all $k \in [K]$.

We rewrite the loss function as:

$$\mathcal{L}(\underline{\underline{\mathbf{\Theta}}}) := \frac{1}{2N} \left(w_0 \|\mathbf{y}^{(0)} - \mathfrak{X}^{(0)}(\underline{\mathbf{\Theta}}^{(0)})\|_2^2 + \sum_{k=1}^{K} w_k \|\mathbf{y}^{(k)} - \mathfrak{X}^{(k)}(\underline{\mathbf{\Theta}}^{(k)} + \underline{\mathbf{\Theta}}^{(0)})\|_2^2 \right) \tag{16}$$

where we use the change of variable:

$$\underline{\underline{\mathbf{\Theta}}} = (\underline{\mathbf{\Theta}}^{(0)}, \underline{\mathbf{\Theta}}^{(1)}, \ldots, \underline{\mathbf{\Theta}}^{(K)}) = (\underline{\mathbf{W}}^{(0)}, \underline{\mathbf{W}}^{(1)} - \underline{\mathbf{W}}^{(0)}, \ldots, \underline{\mathbf{W}}^{(K)} - \underline{\mathbf{W}}^{(0)}) \in \mathbb{R}^{(K+1)\times d_1 \times d_2 \times d_3} \tag{17}$$

With this change of variables, solving problem (4) is equivalent to solving:

$$\hat{\underline{\underline{\mathbf{\Theta}}}} = \underset{\underline{\underline{\mathbf{\Theta}}}}{\operatorname{argmin}} \{\mathcal{L}(\underline{\underline{\mathbf{\Theta}}}) + \lambda_0 \mathcal{R}(\underline{\underline{\mathbf{\Theta}}})\}, \text{ s.t. } \|\mathfrak{X}^{*(0)}(\boldsymbol{y}^{(0)} - \mathfrak{X}^{(0)}(\underline{\mathbf{\Theta}}^{(0)}))\|_{\text{tsp}} \leq \lambda_T, \tag{18}$$

where

$$\mathcal{R}(\underline{\underline{\boldsymbol{\Theta}}}) := \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2} \|\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)}\|_\star + \sum_{k=1}^{K} \frac{\lambda_1}{\lambda_0} w_k \|\underline{\underline{\boldsymbol{\Theta}}}_\star^{(k)}\|_\star. \tag{19}$$

We define the estimation error as $\underline{\underline{\boldsymbol{\Delta}}} := \hat{\underline{\underline{\boldsymbol{\Theta}}}} - \underline{\underline{\boldsymbol{\Theta}}}_\star \in \mathbb{R}^{(K+1)\times d_1 \times d_2 \times d_3}$, with the corresponding k-th block $\underline{\boldsymbol{\Delta}}^{(k)} := \hat{\underline{\boldsymbol{\Theta}}}^{(k)} - \underline{\boldsymbol{\Theta}}_\star^{(k)}$, for all $k = 0, 1, \ldots, K$.

Our goal is to bound $\|\underline{\boldsymbol{\Delta}}^{(0)}\|_{\mathrm{F}}^2$, the F-norm error of the estimated target parameter. The proof of Proposition 1 relies on three key technical lemmas, whose proofs are provided later. The first lemma establishes an upper bound for the first-order term of the Taylor series expansion of $\mathcal{L}(\underline{\underline{\boldsymbol{\Theta}}})$.

Lemma 1 (Concentration of Gradient)**.** *Under Assumption 3.3 and 3.4, if $N_S \gtrsim d_1$, then by choosing $\lambda_0 \gtrsim c_0\sqrt{d_1/N}$ and $\lambda_1 \gtrsim c_0\sqrt{N_S d_1}/N$ for some appropriate constant c_0, then for for any $\underline{\underline{\boldsymbol{\Delta}}} = \left(\underline{\boldsymbol{\Delta}}^{(0)}, \underline{\boldsymbol{\Delta}}^{(1)}, \ldots, \underline{\boldsymbol{\Delta}}^{(k)}\right) \in \mathbb{R}^{(K+1)\times d_1 \times d_2 \times d_3}$, it holds w.h.p.that*

$$\left|\left\langle \nabla \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right), \underline{\underline{\boldsymbol{\Delta}}}\right\rangle\right| \leq \sum_{k=1}^{K} \frac{\lambda_1}{2} w_k \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star + \frac{\lambda_0}{2}\sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2}\left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_\star$$

We omit the proof due to its similarity to Lemma D.1.

The next lemma establishes a restricted set of directions in which $\underline{\underline{\boldsymbol{\Delta}}}$ lies. The proof is omitted due to its similarity to Lemma D.2.

Lemma 2. *Under Assumption 3.3 and 3.4, and the conditions of Lemma 1, the estimation error $\underline{\underline{\boldsymbol{\Delta}}}$ satisfies the inequality w.h.p.*

$$2\lambda_0\sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2}\left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_\star + \sum_{k=1}^{K} \lambda_1 w_k \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star \leq 8\lambda_0\sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2}\left\|\mathcal{P}_\star(\underline{\boldsymbol{\Delta}}^{(0)})\right\|_\star + 8\sum_{k=1}^{K} \lambda_1 w_k h_k.$$

The following lemma ensures a property analogous to restricted strong convexity for $\underline{\underline{\boldsymbol{\Delta}}}$ similar to Lemma D.3. The proof is provided in § 3.1.

Lemma 3. *Define $S_N = \{k : w_k = 0\}$ and $\underline{w} = \min_{k \in S_N^c} w_k$. Under Assumption 3.3 and 3.4 and the conditions of Lemma 2, if $N_S > N_T$, the estimation error $\underline{\underline{\boldsymbol{\Delta}}}$ satisfies w.h.p.*

$$\begin{aligned}
&\mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \underline{\underline{\boldsymbol{\Delta}}}\right) - \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right) - \left\langle \nabla \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right), \underline{\underline{\boldsymbol{\Delta}}}\right\rangle \\
&\geq \left(\frac{\alpha_{\min}^2}{\gamma_0}\sum_{k=0}^{K} \frac{N_k w_k}{N} - u_n\right)\left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 + \frac{\alpha_{\min}^2}{\gamma_0}\sum_{k=1}^{K} \frac{N_k w_k}{N}\left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_{\mathrm{F}}^2 - \frac{2\alpha_{\max}}{\gamma_0}\sum_{k=1}^{K} \frac{N_k w_k}{N}\lambda_T \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star - v_n \sum_{k=1}^{K} \lambda_1 w_k h_k
\end{aligned} \tag{20}$$

where

$$\begin{aligned}
u_n &= \frac{256(\alpha_{\max}\tau_0 + \beta_{\max}\gamma_0)}{\gamma_0}\frac{N_S}{N_T}\frac{d_1}{N} \cdot \frac{\lambda_0^2 r d_3 (\sum_{k=0}^{K} \frac{N_k}{N} w_k^2)}{\lambda_1^2 \underline{w} \wedge [(\lambda_0^2 \sum_{k=0}^{K} \frac{N_S}{N} w_k^2)/((N_T/N_S)w_0 + \sum_{k=1}^{K} w_k)]}, \\
v_n &= \frac{256(\alpha_{\max}\tau_0 + \beta_{\max}\gamma_0)}{\gamma_0}\frac{N_S}{N_T}\frac{d_1}{N} \cdot \frac{\sum_{k=1}^{K} \lambda_1 w_k h_k}{\lambda_1^2 \underline{w} \wedge [(\lambda_0^2 \sum_{k=0}^{K} \frac{N_S}{N} w_k^2)/((N_T/N_S)w_0 + \sum_{k=1}^{K} w_k)]}, \\
\alpha_{\min} &= \min_{0\leq k\leq K} \alpha_k, \quad \alpha_{\max} = \max_{0\leq k\leq K} \alpha_k \text{ and } \quad \beta_{\max} = \max_{0\leq k\leq K} \beta_k
\end{aligned}$$

with RSC constants (α_k, β_k) and RSM constants (γ_k, τ_k) defined in Lemma C.6.

Proof of Proposition 1 For convenience define the function $\mathcal{F} : \mathbb{R}^{(K+1)\times d_1 \times d_2 \times d_3} \to \mathbb{R}$, given by

$$\mathcal{F}(\underline{\underline{\boldsymbol{\Delta}}}) = \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \underline{\underline{\boldsymbol{\Delta}}}\right) - \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right) + \lambda_0 \mathcal{R}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \underline{\underline{\boldsymbol{\Delta}}}\right) - \lambda_0 \mathcal{R}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right).$$

By applying Lemma 1, we can establish the following inequality with high probability:

$$
\begin{aligned}
\mathcal{F}(\underline{\underline{\boldsymbol{\Delta}}}) =& \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \underline{\underline{\boldsymbol{\Delta}}}\right) - \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right) + \lambda_0 \mathcal{R}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \underline{\underline{\boldsymbol{\Delta}}}\right) - \lambda_0 \mathcal{R}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right) \\
\overset{(i)}{\geq}& -\sum_{k=0}^{K} \left\|\nabla \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star^{(k)}\right)\right\|_{\text{tsp}} \|\underline{\underline{\boldsymbol{\Delta}}}^{(k)}\|_\star + \texttt{vec}(\underline{\underline{\boldsymbol{\Delta}}})^\top \nabla^2 \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \gamma \underline{\underline{\boldsymbol{\Delta}}}\right) \texttt{vec}(\underline{\underline{\boldsymbol{\Delta}}}) \quad (\gamma \in (0,1)) \\
&+ \sum_{k=1}^{K} \lambda_1 w_k \left(\left\|\underline{\underline{\boldsymbol{\Theta}}}_\star^{(k)} + \underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star - \left\|\underline{\underline{\boldsymbol{\Theta}}}_\star^{(k)}\right\|_\star\right) + \lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2} \left(\left\|\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)} + \underline{\boldsymbol{\Delta}}^{(0)}\right\|_\star - \left\|\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)}\right\|_\star\right) \\
\overset{(ii)}{\geq}& -\sum_{k=1}^{K} \frac{\lambda_1 w_k}{2} \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star - \frac{\lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2}}{2} \left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_\star + \texttt{vec}(\underline{\underline{\boldsymbol{\Delta}}})^\top \nabla^2 \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \gamma \underline{\underline{\boldsymbol{\Delta}}}\right) \texttt{vec}(\underline{\underline{\boldsymbol{\Delta}}}) \\
&+ \sum_{k=1}^{K} \lambda_1 w_k \left(\left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star - 2\left\|\underline{\underline{\boldsymbol{\Theta}}}_\star^{(k)}\right\|_\star\right) \\
&+ \lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2} \left(\left\|\mathcal{P}_\star(\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)})\right\|_\star - \left\|\mathcal{P}_{\star^\perp}(\underline{\boldsymbol{\Delta}}^{(0)})\right\|_\star + \left\|\mathcal{P}_{\star^\perp}(\underline{\boldsymbol{\Delta}}^{(0)})\right\|_\star - \left\|\mathcal{P}_{\star^\perp}(\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)})\right\|_\star - \left\|\mathcal{P}_\star(\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)})\right\|_\star - \left\|\mathcal{P}_{\star^\perp}(\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)})\right\|_\star\right) \\
\overset{(iii)}{\geq}& \texttt{vec}(\underline{\underline{\boldsymbol{\Delta}}})^\top \nabla^2 \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \gamma \underline{\underline{\boldsymbol{\Delta}}}\right) \texttt{vec}(\underline{\underline{\boldsymbol{\Delta}}}) + \frac{\lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2}}{2} \left(\left\|\mathcal{P}_{\star^\perp}(\underline{\boldsymbol{\Delta}}^{(0)})\right\|_\star - 3\left\|\mathcal{P}_\star(\underline{\boldsymbol{\Delta}}^{(0)})\right\|_\star\right) \\
&+ \sum_{k=1}^{K} \frac{\lambda_1 w_k}{2} \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star - 2\sum_{k=1}^{K} \lambda_1 w_k h_k.
\end{aligned}
$$

where (i) follows by the mean value theorem with $\gamma \in (0,1)$ and Holder's inequality; (ii) holds as a result of Lemma 1 and we also use $\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)} = \underline{\mathbf{W}}_\star^{(0)}$, $\|\mathcal{P}_{\star^\perp}(\underline{\underline{\boldsymbol{\Theta}}}_\star^{(0)})\|_\star = 0$, and $\|\underline{\mathbf{W}}_\star^{(0)} + \underline{\boldsymbol{\Delta}}\|_\star = \|\underline{\mathbf{W}}_\star^{(0)} + \mathcal{P}_{\star^\perp}(\underline{\boldsymbol{\Delta}}) + \mathcal{P}_\star(\underline{\boldsymbol{\Delta}})\|_\star \geq \|\underline{\mathbf{W}}_\star^{(0)} + \mathcal{P}_{\star^\perp}(\underline{\boldsymbol{\Delta}})\|_\star - \|\mathcal{P}_\star(\underline{\boldsymbol{\Delta}})\|_\star = \|\underline{\mathbf{W}}_\star^{(0)}\|_\star + \|\mathcal{P}_{\star^\perp}(\underline{\boldsymbol{\Delta}})\|_\star - \|\mathcal{P}_\star(\underline{\boldsymbol{\Delta}})\|_\star$ due to the decomposibility of TNN; (iii) holds because $\|\underline{\underline{\boldsymbol{\Theta}}}_\star^{(k)}\|_\star \leq h_k$ for $1 \leq k \leq K$.

Applying Lemma 3, the quadratic term $\texttt{vec}(\underline{\underline{\boldsymbol{\Delta}}})^\top \nabla^2 \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \gamma \underline{\underline{\boldsymbol{\Delta}}}\right) \texttt{vec}(\underline{\underline{\boldsymbol{\Delta}}})$ can be lower bounded *w.h.p.*as

$$
\begin{aligned}
\mathcal{F}(\underline{\underline{\boldsymbol{\Delta}}}) \geq& \left(\frac{\alpha_{\min}^2}{\gamma_0} \sum_{k=0}^{K} \frac{N_k w_k}{N} - u_n\right) \left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\text{F}}^2 + \frac{\alpha_{\min}^2}{\gamma_0} \sum_{k=1}^{K} \frac{N_S w_k}{N} \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_{\text{F}}^2 - \frac{2\alpha_{\max}}{\gamma_0} \sum_{k=1}^{K} \frac{N_S w_k}{N} \lambda_T \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star - v_n \sum_{k=1}^{K} \lambda_1 w_k h_k \\
&+ \frac{\lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2}}{2} \left(\left\|\mathcal{P}_{\star^\perp}(\underline{\boldsymbol{\Delta}}^{(0)})\right\|_\star - 3\left\|\mathcal{P}_\star(\underline{\boldsymbol{\Delta}}^{(0)})\right\|_\star\right) + \sum_{k=1}^{K} \frac{\lambda_1 w_k}{2} \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star - 2\sum_{k=1}^{K} \lambda_1 w_k h_k \\
\geq& \left(\frac{\alpha_{\min}^2}{\gamma_0} \sum_{k=0}^{K} \frac{N_k w_k}{N} - u_n\right) \left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\text{F}}^2 - \frac{3\sqrt{rd_3}\lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2}}{2} \left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\text{F}} \\
&+ \left(\sum_{k=1}^{K} \frac{\lambda_1 w_k}{2} - \frac{2\alpha_{\max}}{\gamma_0} \sum_{k=1}^{K} \frac{N_S w_k}{N} \lambda_T\right) \left\|\underline{\boldsymbol{\Delta}}^{(k)}\right\|_\star - (2 + v_n) \sum_{k=1}^{K} \lambda_1 w_k h_k. \qquad (21)
\end{aligned}
$$

Recall that we select $\lambda_1 = c_1 \sqrt{d_1 N_S^2 / (N_T N^2)}$ and $\lambda_T = c_2 \sqrt{d_1 / N_T}$, therefore, with a proper choice of the constants c_1 and c_2, we have $\lambda_1 \geq 4 \frac{\alpha_{\max}}{\gamma_0} \frac{N_S}{N} \lambda_T$. In addition, notice that $\hat{\underline{\underline{\boldsymbol{\Theta}}}}$ is the solution to the Problem (18) and $\underline{\underline{\boldsymbol{\Theta}}}_\star$ is feasible, we have $\mathcal{F}(\underline{\boldsymbol{\Delta}}) \leq 0$. These results, combining with Eq. (21), lead to

$$
0 \geq \left(\frac{\alpha_{\min}^2}{\gamma_0} \sum_{k=0}^{K} \frac{N_k w_k}{N} - u_n\right) \left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\text{F}}^2 - \frac{3\sqrt{rd_3}\lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_k}{N} w_k^2}}{2} \left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\text{F}} - (2 + v_n) \sum_{k=1}^{K} \lambda_1 w_k h_k, \qquad (22)
$$

which is an inequality quadratic in $\|\underline{\boldsymbol{\Delta}}^{(0)}\|_{\text{F}}$.

To establish the convergence rate of $\|\underline{\boldsymbol{\Delta}}^{(0)}\|_{\mathrm{F}}$, it remains to find out the order of u_n and v_n. We now discuss by cases based on the order of λ_0. Recall the parameter choices given in the theorem:

$$\lambda_0 = c_0\left[\Big(\sum_{k=0}^{K}\frac{N_k}{N}w_k^2\Big)^{-1/2}\Big(\frac{\bar{h}_{\mathbf{w}}^2 d_1}{(rd_3)^2 N_T}\Big)^{1/4} + \Big(\frac{d_1}{N}\Big)^{1/2}\right], \quad \lambda_1 = c_0\frac{N_S}{N}\Big(\frac{d_1}{N_T}\Big)^{1/2}, \quad \text{and} \quad \lambda_T = c_1\Big(\frac{d_1}{N_T}\Big)^{1/2}.$$

Case 1: If $\frac{rd_1d_3}{N} \lesssim \Big(\sum_{k=0}^{K}\frac{N_k}{N}w_k^2\Big)^{-1}\bar{h}_{\mathbf{w}}\sqrt{\frac{d_1}{N_T}}$, then we have

$$\lambda_0 \asymp \Big(\sum_{k=0}^{K}\frac{N_k}{N}w_k^2\Big)^{-1/2}\left(\frac{\bar{h}_{\mathbf{w}}^2 d_1}{(rd_3)^2 N_T}\right)^{1/4}, \quad \lambda_1 \asymp \frac{N_S}{N}\sqrt{\frac{d_1}{N_T}}, \quad \lambda_T \asymp \sqrt{\frac{d_1}{N_T}}$$

Recalling that $N_k = N_S$ for $k \in [K]$, we consider the following two cases:

- If $\lambda_1^2\underline{w} \lesssim (\lambda_0^2\sum_{k=0}^{K}\frac{N_k}{N}w_k^2)/((N_T/N_S)w_0 + \sum_{k=1}^{K}w_k)$, then we have

  $$u_n \lesssim \frac{1}{K}\frac{d_1}{N_T}\frac{\lambda_0^2 rd_3(\sum_{k=0}^{K}\frac{N_k}{N}w_k^2)}{\lambda_1^2} \lesssim K\bar{h}_{\mathbf{w}}\sqrt{\frac{d_1}{N_T}} = o(1),$$

  where the first inequality is based on the assumption that $\underline{w}$ is bounded away from 0 and $(\beta_{\max}\vee\tau_0)/\gamma_0$ is bounded above, while for the last equality we uses the assumption that $K\bar{h}_{\mathbf{w}}\sqrt{\frac{d_1}{N_T}} = o(1)$. Similarly, we can establish that

  $$v_n \lesssim \frac{1}{K}\frac{d_1}{N_T}\frac{\left(\sum_{k=1}^{K}\lambda_1 w_k h_k\right)}{\lambda_1^2} \lesssim \sqrt{\frac{d_1}{N_T}}\sum_{k=1}^{K}w_k h_k \lesssim K\bar{h}_{\mathbf{w}}\sqrt{\frac{d_1}{N_T}} = o(1).$$

  Therefore, in this case, solving the inequality (22) yields

  $$\left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 \lesssim \frac{1}{\left(\sum_{k=0}^{K}\frac{N_k w_k}{N}\right)^2}\bar{h}_{\mathbf{w}}\sqrt{\frac{d_1}{N_T}} + \frac{1}{\sum_{k=0}^{K}\frac{N_k w_k}{N}}\sum_{k=1}^{K}\frac{N_S w_k}{N}h_k\sqrt{\frac{d_1}{N_T}}. \tag{23}$$

- If $\lambda_1^2\underline{w} \gtrsim (\lambda_0^2\sum_{k=0}^{K}\frac{N_S}{N}w_k^2)/((N_T/N_S)w_0 + \sum_{k=1}^{K}w_k)$, we can similarly establish that

  $$u_n \lesssim \frac{1}{K}\frac{d_1}{N_T}rd_3((N_T/N_S)w_0 + \sum_{k=1}^{K}w_k) \lesssim \frac{rd_1d_3}{N_T} = o(1)$$

  where the last inequality follows from the assumptions that $\frac{rd_1d_3}{N_T} = o(1)$.

  In this case, noting that

  $$\frac{\left(\sum_{k=1}^{K}\lambda_1 w_k h_k\right)}{\left(1/((N_T/N_S)w_0 + \sum_{k=1}^{K}w_k)\right)} \lesssim K\bar{h}_{\mathbf{w}}\sqrt{\frac{d_1}{N_T}},$$

  thereby we have

  $$v_n \lesssim \frac{1}{K}\frac{d_1}{N_T}\frac{\left(\sum_{k=1}^{K}\lambda_1 w_k h_k\right)}{\left(\lambda_0^2\sum_{k=0}^{K}\frac{N_k}{N}w_k^2/((N_T/N_S)w_0 + \sum_{k=1}^{K}w_k)\right)} \lesssim \frac{rd_1d_3}{N_T} = o(1).$$

  Hence, the bound in (23) still holds.

Case 2: If $\frac{rd_1d_3}{N} \gtrsim \left(\sum_{k=0}^{K} \frac{N_k}{N} w_k^2\right)^{-1} \bar{h}_{\mathbf{w}} \sqrt{\frac{d_1}{N_T}}$, then we have

$$\lambda_0 \asymp \sqrt{\frac{d_1}{N}}, \quad \lambda_1 \asymp \frac{N_S}{N} \sqrt{\frac{d_1}{N_T}}, \quad \lambda_T \asymp \sqrt{\frac{d_1}{N_T}}.$$

In this case, we have $\lambda_1^2 \gtrsim \frac{\lambda_0^2}{K} \gtrsim \left(\lambda_0^2 \sum_{k=0}^{K} \frac{N_k}{N} w_k^2 / ((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k)\right)$ as $N_S > N_T$. So following the discussion in the first case, we have

$$u_n \lesssim \frac{1}{K} \frac{d_1}{N_T} r d_3 ((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k) \lesssim \frac{r d_1 d_3}{N_T} = o(1).$$

Further notice that in this case, it holds that

$$\sum_{k=1}^{K} \lambda_1 w_k h_k \lesssim \bar{h}_{\mathbf{w}} \sqrt{\frac{d_1}{N_T}} \lesssim \Big(\sum_{k=0}^{K} \frac{N_k}{N} w_k^2\Big) \frac{r d_1 d_3}{N}.$$

Therefore, we further obtain

$$v_n \lesssim \frac{1}{K} \frac{d_1}{N_T} \frac{\left(\sum_{k=1}^{K} \lambda_1 w_k h_k\right)}{\left(\lambda_0^2 \sum_{k=0}^{K} \frac{N_k}{N} w_k^2 / ((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k)\right)} \lesssim \frac{1}{K} \frac{d_1}{N_T} r d_3 \left((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k\right) \lesssim \frac{r d_1 d_3}{N_T} = o(1).$$

So in this case plugging in the choice of λ_0 and λ_1, the F-norm error bound for $\underline{\boldsymbol{\Delta}}^{(0)}$ becomes

$$\left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 \lesssim \frac{1}{\left(\sum_{k=0}^{K} \frac{N_k w_k}{N}\right)^2} \Big(\sum_{k=0}^{K} \frac{N_k w_k^2}{N}\Big) \frac{r d_1 d_3}{N} + \frac{1}{\sum_{k=0}^{K} \frac{N_k w_k}{N}} \bar{h}_{\mathbf{w}} \sqrt{\frac{d_1}{N_T}} \tag{24}$$

In summary, by combining the results from the two cases discussed above and applying the constraint $\sum_{k=0}^{K} \frac{N_K w_k}{N} = 1$, we have

$$\left\|\underline{\boldsymbol{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 \lesssim \Big(\sum_{k=0}^{K} \frac{N_k w_k^2}{N}\Big) \frac{r d_1 d_3}{N} + \bar{h}_{\mathbf{w}} \sqrt{\frac{d_1}{N_T}},$$

which completes the proof.

2.2. Proof of Theorem 2

The proof follows the arguments in He et al. (2024), which builds on Section 3 of Duchi et al. (2008). According to Proposition 1, the optimal weights can be formally described as the solution to the following problem:

$$\min_{\mathbf{w}'} \left\{ \sum_{k=0}^{K} (w_k')^2 \frac{r d_1 d_3}{N_k} + c_\epsilon \sum_{k=1}^{K} w_k' h_k \sqrt{\frac{d_1}{N_T}} \right\} \quad \text{s.t.} \ \sum_{k=0}^{K} w_k' = 1, \ w_k' \geq 0, \ k = 0, \ldots, K. \tag{25}$$

where recall that we define $w_k' = \frac{N_k}{N} w_k$ with $N_k = N_S$ for $k \in [K]$ and $N_k = N_T$ for $k = 0$.

The corresponding Lagrangian function of Problem (25) is

$$\mathcal{L}(\mathbf{w}', \boldsymbol{\eta}) = \sum_{k=0}^{K} (w_k')^2 \frac{r d_1 d_3}{N_k} + c_\epsilon \sum_{k=1}^{K} w_k' h_k \sqrt{\frac{d_1}{n_T}} + \xi \left(1 - \sum_{k=1}^{K} w_k'\right) - \sum_{k=1}^{K} \eta_k w_k'$$

where $\boldsymbol{\eta} \in \mathbb{R}_+^K$ and $\xi \in \mathbb{R}$ are Lagrange multipliers. According to the KKT conditions, we have

$$\frac{\partial \mathcal{L}(\mathbf{w}', \boldsymbol{\eta})}{\partial w_k'} = \frac{2 r d_1 d_3}{N_S} w_k' + c_\epsilon h_k \sqrt{\frac{d_1}{N_T}} - \xi - \eta_k = 0, \ k \in [K] \tag{26}$$

$$\frac{\partial \mathcal{L}(\mathbf{w}', \boldsymbol{\eta})}{\partial w_0'} = \frac{2 r d_1 d_3}{N_T} w_0' - \xi - \eta_0 = 0 \tag{27}$$

and $\eta_k \cdot w'_k = 0$ for $k = 0, \ldots, K$. Notice for any source task that is assigned a non-zero weight w_k, we must have $\eta_k = 0$. Therefore, the non-zero weights are tied to the single variable ξ. To find ξ, we use the following lemma, with its proof deferred to Section 3.

Lemma 4. *Let $w'_0, \ldots, w'_K$ be the optimal solution to the problem (25), then*

- *For any $j, l \in [K]$ such that $h_l > h_j$, $w'_j = 0$ implies $w'_l = 0$.*
- *For any choice of $\boldsymbol{h} \in \mathbb{R}_+^K$, it always holds that $w'_0 > 0$.*

Lemma 4 establishes a systematic approach for determining non-zero weights based on $\boldsymbol{h}$. Let us define $K^\sharp = \#\{k \in [K] : w_k > 0\}$ as the number of sources with positive weights, and let $h_{(1)} \leqslant h_{(2)} \leqslant \ldots \leqslant h_{(K)}$ represent the elements of $\boldsymbol{h}$ arranged in ascending order. By combining Lemma 4 with Eqs. (26) and (27), we obtain:

$$\sum_{k=1}^{K} \frac{2rd_1d_3}{N_S} w'_k = \sum_{k=1}^{K^\sharp} \left(\xi - c_\epsilon h_{(k)} \sqrt{\frac{d_1}{N_T}} \right), \quad \text{and} \quad \frac{2rd_1d_3}{N_T} w'_0 = \xi.$$

This relationship allows us to express ξ as:

$$\xi = \frac{2rd_1d_3 + N_S \sum_{k=1}^{K^\sharp} c_\epsilon h_{(k)} \sqrt{\frac{d_1}{N_T}}}{N_T + K^\sharp N_S} \tag{28}$$

By substituting this expression for ξ back into Eqs. (26) and (27), we derive the desired result.

To complete the proof, we must establish the bound in (12). We proceed by analyzing two exhaustive cases:

- Case 1: When $rd_1d_3/N_T \leq c_\epsilon/2 \cdot h_1\sqrt{d_1/N_T}$. According to Eq. (11), this condition implies $w_1 = 0$. In this scenario, the upper bound in Eq. (6) simplifies to:
$$\min\left\{ \frac{c_\epsilon}{2} h_1 \sqrt{\frac{d_1}{N_T}}, \frac{rd_1d_3}{N_T} \right\}.$$
- Case 2: When $rd_1d_3/N_T \leq c_\epsilon/2 \cdot h_1\sqrt{d_1/N_T}$. By substituting the expression for w_1 into the upper bound in (6), we obtain:
$$\frac{rd_1d_3}{N} + \frac{N_S}{N} \min\left\{ c_\epsilon h_1 \sqrt{\frac{d_1}{N_T}}, 2\frac{rd_1d_3}{N_T} \right\}.$$

The combination of these two cases yields the comprehensive bound stated in (12).

3. Proof of Lemmas

3.1. Proof of Lemma 3

We begin by expanding the objective function using a second-order Taylor expansion to establish the relationship between the loss function and parameter deviation. Leveraging the Restricted Strong Convexity (RSC) and Restricted Smoothness (RSM) properties in Lemma C.6 of the main paper, we decompose the expansion into several controllable terms. We then analyze these terms systematically to establish the final error bound.

Initially, we apply a second-order Taylor expansion to the loss function $\mathcal{L}$, which yields

$$\begin{aligned}
&\mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \underline{\underline{\boldsymbol{\Delta}}}\right) - \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right) - \left\langle \nabla\mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star\right), \underline{\underline{\boldsymbol{\Delta}}} \right\rangle \\
&= \ \mathtt{vec}(\underline{\underline{\boldsymbol{\Delta}}})^\top \nabla^2\mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \gamma\underline{\underline{\boldsymbol{\Delta}}}\right) \mathtt{vec}(\underline{\underline{\boldsymbol{\Delta}}}) \quad (\gamma \in (0,1)) \\
&= \frac{N_T}{N} w_0 \mathtt{vec}\left(\underline{\boldsymbol{\Delta}}^{(0)}\right)^\top \hat{\boldsymbol{\Sigma}}^{(0)} \mathtt{vec}(\underline{\boldsymbol{\Delta}}^{(0)}) + \sum_{k=1}^{K} \frac{N_S}{N} w_k \mathtt{vec}\left(\underline{\boldsymbol{\Delta}}^{(k)} + \underline{\boldsymbol{\Delta}}^{(0)}\right)^\top \hat{\boldsymbol{\Sigma}}^{(k)} \mathtt{vec}\left(\underline{\boldsymbol{\Delta}}^{(k)} + \underline{\boldsymbol{\Delta}}^{(0)}\right).
\end{aligned}$$

where $\hat{\mathbf{\Sigma}}^{(0)}$ and $\hat{\mathbf{\Sigma}}^{(k)}$ are given as

$$\hat{\mathbf{\Sigma}}^{(0)} := \sum_{k=0}^{K}\sum_{i=1}^{N_k} \mathtt{vec}(\underline{\mathbf{X}}_i^{(k)})\mathtt{vec}(\underline{\mathbf{X}}_i^{(k)})^\top \in \mathbb{R}^{d_1d_2d_3\times d_1d_2d_3}$$
$$\hat{\mathbf{\Sigma}}^{(k)} := \sum_{i=1}^{N_k} \mathtt{vec}(\underline{\mathbf{X}}_i^{(k)})\mathtt{vec}(\underline{\mathbf{X}}_i^{(k)})^\top \in \mathbb{R}^{d_1d_2d_3\times d_1d_2d_3}, \quad \forall k = 1,\dots,K.$$

Applying the RSC and RSM properties established in Lemma C.6, we further derive

$$\begin{aligned}
&\mathcal{L}\left(\underline{\underline{\mathbf{\Theta}}}_\star + \underline{\underline{\mathbf{\Delta}}}\right) - \mathcal{L}\left(\underline{\underline{\mathbf{\Theta}}}_\star\right) - \left\langle \nabla\mathcal{L}\left(\underline{\underline{\mathbf{\Theta}}}_\star\right), \underline{\underline{\mathbf{\Delta}}}\right\rangle \\
\geq& \sum_{k=1}^{K} \frac{N_S\alpha_k w_k}{N}\left\|\underline{\mathbf{\Delta}}^{(k)} + \underline{\mathbf{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 + \frac{N_T\alpha_0 w_0}{N}\left\|\underline{\mathbf{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 - R_1(\underline{\mathbf{\Delta}}) \\
\geq& \sum_{k=1}^{K} \frac{N_S\alpha_k w_k}{N}\cdot\frac{1}{\gamma_0}\mathtt{vec}\left(\underline{\mathbf{\Delta}}^{(k)} + \underline{\mathbf{\Delta}}^{(0)}\right)^\top \hat{\mathbf{\Sigma}}^{(0)}\mathtt{vec}\left(\underline{\mathbf{\Delta}}^{(k)} + \underline{\mathbf{\Delta}}^{(0)}\right) - R_1(\underline{\mathbf{\Delta}}) + \frac{N_T\alpha_0 w_0}{N}\left\|\underline{\mathbf{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 - R_2(\underline{\mathbf{\Delta}}) \\
\geq& \sum_{k=1}^{K} \frac{N_S\alpha_k w_k}{N}\cdot\frac{1}{\gamma_0}\left(\mathtt{vec}\left(\underline{\mathbf{\Delta}}^{(k)}\right)^\top \hat{\mathbf{\Sigma}}^{(0)}\mathtt{vec}(\underline{\mathbf{\Delta}}^{(k)}) + \mathtt{vec}\left(\underline{\mathbf{\Delta}}^{(0)}\right)^\top \hat{\mathbf{\Sigma}}^{(0)}\mathtt{vec}(\underline{\mathbf{\Delta}}^{(0)}) + 2\mathtt{vec}\left(\underline{\mathbf{\Delta}}^{(k)}\right)^\top \hat{\mathbf{\Sigma}}^{(0)}\mathtt{vec}(\underline{\mathbf{\Delta}}^{(0)})\right) \\
&+ \frac{N_T\alpha_0 w_0}{N}\left\|\underline{\mathbf{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 - R_1(\underline{\mathbf{\Delta}}) - R_2(\underline{\mathbf{\Delta}}),
\end{aligned} \tag{29}$$

where $R_1(\underline{\underline{\mathbf{\Delta}}})$ and $R_2(\underline{\underline{\mathbf{\Delta}}})$ are defined as

$$R_1(\underline{\mathbf{\Delta}}) := \frac{\beta_0 w_0 d_1}{N}\left\|\underline{\mathbf{\Delta}}^{(0)}\right\|_\star^2 + \sum_{k=1}^{K}\frac{\beta_k w_k d_1}{N}\left\|\underline{\mathbf{\Delta}}^{(k)} + \underline{\mathbf{\Delta}}^{(0)}\right\|_\star^2, \quad R_2(\underline{\mathbf{\Delta}}) := \sum_{k=1}^{K}\frac{N_S\alpha_k w_k}{N}\frac{\tau_0}{\gamma_0}\frac{d_1}{N_T}\left\|\underline{\mathbf{\Delta}}^{(k)} + \underline{\mathbf{\Delta}}^{(0)}\right\|_\star^2. \tag{30}$$

Moreover, observe that $\hat{\underline{\mathbf{W}}}^{(0)}$ satisfies the constraint specified in Eq. (3). Consequently, by invoking the results from Lemma 1, we obtain

$$\begin{aligned}
\left\|\mathfrak{X}^{*(0)}(\mathfrak{X}^{(0)}(\underline{\mathbf{\Delta}}^{(0)}))\right\|_{\text{tsp}} &= \left\|\frac{1}{N_T}\mathfrak{X}^{*(0)}\mathfrak{X}^{(0)}\left(\hat{\underline{\mathbf{W}}}^{(0)} - \underline{\mathbf{W}}_\star^{(0)}\right)\right\|_{\text{tsp}} \\
&= \left\|\frac{1}{N_T}\mathfrak{X}^{*(0)}\left(\boldsymbol{y}^{(0)} - \mathfrak{X}^{(0)}(\underline{\mathbf{W}}_\star^{(0)})\right) - \frac{1}{N_T}\mathfrak{X}^{*(0)}\left(\boldsymbol{y}^{(0)} - \mathfrak{X}^{(0)}(\hat{\underline{\mathbf{W}}}^{(0)})\right)\right\|_{\text{tsp}} \\
&\leq \left\|\frac{1}{N_T}\mathfrak{X}^{*(0)}\left(\boldsymbol{y}^{(0)} - \mathfrak{X}^{(0)}(\underline{\mathbf{W}}_\star^{(0)})\right)\right\|_{\text{tsp}} + \left\|\frac{1}{N_T}\mathfrak{X}^{*(0)}\left(\boldsymbol{y}^{(0)} - \mathfrak{X}^{(0)}(\hat{\underline{\mathbf{W}}}^{(0)})\right)\right\|_{\text{tsp}} \\
&\leq 2\lambda_T.
\end{aligned}$$

We use Hölder's inequality to get

$$\mathtt{vec}\left(\underline{\mathbf{\Delta}}^{(k)}\right)^\top \hat{\mathbf{\Sigma}}^{(0)}\mathtt{vec}(\underline{\mathbf{\Delta}}^{(0)}) = \langle\underline{\mathbf{\Delta}}^{(k)}, \mathfrak{X}^{*(0)}(\mathfrak{X}^{(0)}(\underline{\mathbf{\Delta}}^{(0)}))\rangle \leq 2\lambda_T\left\|\underline{\mathbf{\Delta}}^{(k)}\right\|_\star. \tag{31}$$

Recall that $N_k = N_S$ for $k = 1,\dots,K$ and $N_k = N_T$ for $k = 0$. Combining Eq. (31) with Eq. (29) and applying the RSC property again, we have

$$\begin{aligned}
&\mathcal{L}\left(\underline{\underline{\mathbf{\Theta}}}_\star + \underline{\underline{\mathbf{\Delta}}}\right) - \mathcal{L}\left(\underline{\underline{\mathbf{\Theta}}}_\star\right) - \left\langle \nabla\mathcal{L}\left(\underline{\underline{\mathbf{\Theta}}}_\star\right), \underline{\underline{\mathbf{\Delta}}}\right\rangle \\
\geq& \sum_{k=1}^{K}\frac{N_S\alpha_k w_k}{N}\cdot\frac{1}{\gamma_0}\left[\alpha_0\left\|\underline{\mathbf{\Delta}}^{(k)}\right\|_{\mathrm{F}}^2 + \alpha_0\left\|\underline{\mathbf{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 - 2\lambda_T\left\|\underline{\mathbf{\Delta}}^{(k)}\right\|_\star\right] + \frac{N_T\alpha_0 w_0}{N}\left\|\underline{\mathbf{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 - R_1(\underline{\mathbf{\Delta}}) - R_2(\underline{\mathbf{\Delta}}) - R_3(\underline{\mathbf{\Delta}}) \\
\geq& \frac{\alpha_{\min}^2}{\gamma_0}\left[\sum_{k=0}^{K}\frac{N_k w_k}{N}\left\|\underline{\mathbf{\Delta}}^{(0)}\right\|_{\mathrm{F}}^2 + \sum_{k=1}^{K}\frac{N_S w_k}{N}\left\|\underline{\mathbf{\Delta}}^{(k)}\right\|_{\mathrm{F}}^2\right] - \frac{2\alpha_{\max}}{\gamma_0}\sum_{k=1}^{K}\frac{N_S w_k}{N}\lambda_T\left\|\underline{\mathbf{\Delta}}^{(k)}\right\|_\star - \sum_{j=1}^{3}R_j(\underline{\mathbf{\Delta}}),
\end{aligned}$$

where $R_3$is defined as

$$R_3(\underline{\underline{\boldsymbol{\Delta}}}) := \sum_{k=1}^{K} \frac{N_S \alpha_k w_k}{N} \frac{\beta_0}{\gamma_0} \frac{d_1}{N_T} \left(\left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star^2 + \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 \right).$$

We further bound the term

$$\begin{aligned}
\sum_{j=1}^{3} R_j(\underline{\underline{\boldsymbol{\Delta}}}) =& \sum_{j=1}^{K} \frac{\beta_k w_k d_1}{N} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} + \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 + \frac{\beta_0 w_0 d_1}{N} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 + \sum_{k=1}^{K} \frac{N_S \alpha_k w_k}{N} \frac{\tau_0}{\gamma_0} \frac{d_1}{N_T} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} + \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 \\
&+ \sum_{k=1}^{K} \frac{N_S \alpha_k w_k}{N} \frac{\beta_0}{\gamma_0} \frac{d_1}{N_T} \left(\left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star^2 + \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 \right) \\
\leq& \left(\frac{\beta_0 w_0 d_1}{N} + 2 \sum_{k=1}^{K} \frac{\beta_k w_k d_1}{N} + 2 \sum_{k=1}^{K} \frac{N_S \alpha_k w_k}{N} \frac{\tau_0}{\gamma_0} \frac{d_1}{N_T} \right) \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 + \sum_{k=1}^{K} \left(\frac{2\beta_k w_k d_1}{N} + \frac{2N_S \alpha_k w_k}{N} \frac{\tau_0}{\gamma_0} \frac{d_1}{N_T} \right) \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star^2 \\
\leq& \left(2 \sum_{k=0}^{K} \frac{\beta_k w_k d_1}{N} + 2 \sum_{k=1}^{K} \frac{N_S \alpha_k w_k}{N} \frac{\tau_0}{\gamma_0} \frac{d_1}{N_T} \right) \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 + \sum_{k=1}^{K} \left(\frac{2\beta_k w_k d_1}{N} + \frac{2N_S \alpha_k w_k}{N} \frac{\tau_0}{\gamma_0} \frac{d_1}{N_T} \right) \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star^2 \\
\leq& \frac{2(\alpha_{\max} \tau_0 + \beta_{\max} \gamma_0)}{\gamma_0} \left[\left(w_0 \frac{d_1}{N} + \sum_{k=1}^{K} w_k \frac{N_S}{N_T} \frac{d_1}{N} \right) \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 + \sum_{k=1}^{K} w_k \frac{N_S}{N_T} \frac{d_1}{N} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star^2 \right].
\end{aligned}$$

As $w_k \geq \underline{w}$, we use Lemma 2 to get

$$\begin{aligned}
&\sum_{j=1}^{3} R_j(\widehat{\underline{\underline{\boldsymbol{\Delta}}}}) \cdot \left(\frac{2(\alpha_{\max} \tau_0 + \beta_{\max} \gamma_0)}{\gamma_0} \right)^{-1} \\
\leq& \sum_{k=1}^{K} w_k \frac{N_S}{N_T} \frac{d_1}{N} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star^2 + \left(w_0 + \sum_{k=1}^{K} w_k \frac{N_S}{N_T} \right) \frac{d_1}{N} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 \\
\leq& \frac{N_S}{N_T} \frac{d_1}{N} \sum_{k=1}^{K} \frac{\lambda_1^2 w_k^2}{\lambda_1^2 w_k} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star^2 + \frac{d_1}{N} \frac{1}{N_T} \left(N_T w_0 + \sum_{k=1}^{K} N_S w_k \right) \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 \\
=& \frac{N_S}{N_T} \frac{d_1}{N} \left(\sum_{k=1}^{K} \frac{\lambda_1^2 w_k^2}{\lambda_1^2 w_k} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star^2 + \frac{\lambda_0^2 \sum_{k=0}^{K} \frac{N_S}{N} w_k^2}{(\lambda_0^2 \sum_{k=0}^{K} \frac{N_S}{N} w_k^2)/((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k)} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star^2 \right) \\
\leq& \frac{N_S}{N_T} \frac{d_1}{N} \cdot \frac{1}{\lambda_1^2 \underline{w} \wedge (\lambda_0^2 \sum_{k=0}^{K} \frac{N_S}{N} w_k^2)/((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k)} \left(\sum_{k=1}^{K} \lambda_1 w_k \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star + \lambda_0 \sqrt{\sum_{k=0}^{K} \frac{N_S}{N} w_k^2} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_\star \right)^2 \\
\leq& \frac{N_S}{N_T} \frac{d_1}{N} \cdot \frac{1}{\lambda_1^2 \underline{w} \wedge (\lambda_0^2 \sum_{k=0}^{K} \frac{N_S}{N} w_k^2)/((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k)} \left(32 \lambda_0^2 r d_3 (\sum_{k=0}^{K} \frac{N_k}{N} w_k^2) \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_{\mathrm{F}}^2 + 32 \left(\sum_{k=1}^{K} \lambda_1 w_k h_k \right)^2 \right).
\end{aligned}$$

Thus, by letting

$$\begin{aligned}
u_n &= \frac{256(\alpha_{\max} \tau_0 + \beta_{\max} \gamma_0)}{\gamma_0} \frac{N_S}{N_T} \frac{d_1}{N} \cdot \frac{\lambda_0^2 r d_3 (\sum_{k=0}^{K} \frac{N_k}{N} w_k^2)}{\lambda_1^2 \underline{w} \wedge [(\lambda_0^2 \sum_{k=0}^{K} \frac{N_S}{N} w_k^2)/((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k)]} \\
v_n &= \frac{256(\alpha_{\max} \tau_0 + \beta_{\max} \gamma_0)}{\gamma_0} \frac{N_S}{N_T} \frac{d_1}{N} \cdot \frac{\sum_{k=1}^{K} \lambda_1 w_k h_k}{\lambda_1^2 \underline{w} \wedge [(\lambda_0^2 \sum_{k=0}^{K} \frac{N_S}{N} w_k^2)/((N_T/N_S) w_0 + \sum_{k=1}^{K} w_k)]},
\end{aligned}$$

we have

$$\begin{aligned}
&\mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star + \underline{\underline{\boldsymbol{\Delta}}} \right) - \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star \right) - \left\langle \nabla \mathcal{L}\left(\underline{\underline{\boldsymbol{\Theta}}}_\star \right), \underline{\underline{\boldsymbol{\Delta}}} \right\rangle \\
\geq& \left(\frac{\alpha_{\min}^2}{\gamma_0} \sum_{k=0}^{K} \frac{N_k w_k}{N} - u_n \right) \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(0)} \right\|_{\mathrm{F}}^2 + \frac{\alpha_{\min}^2}{\gamma_0} \sum_{k=1}^{K} \frac{N_k w_k}{N} \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_{\mathrm{F}}^2 - \frac{2\alpha_{\max}}{\gamma_0} \sum_{k=1}^{K} \frac{N_k w_k}{N} \lambda_T \left\| \underline{\underline{\boldsymbol{\Delta}}}^{(k)} \right\|_\star - v_n \sum_{k=1}^{K} \lambda_1 w_k h_k,
\end{aligned}$$

which finishes the proof.

3.2. Proof of Lemma 4

Our proof is adapted from He et al. (2024) and follows, via contradiction, a strategy similar to that employed in Shalev-Shwartz et al. (2006).

Claim 1: Assume by contradiction that for some indices j and l (with $j \neq l$), we have $w'_j = 0$ while $w'_l > 0$. Note that swapping the roles of w_j and w_l in the optimization problem (25) does not affect the normalization constraint. By the optimality condition, the objective value after swapping must not be lower than before. Formally, this yields

$$
\begin{aligned}
0 &\geq \left(\sum_{k=0}^{K}(w'_k)^2\frac{rd_1d_3}{N_k} + c_\epsilon\sum_{k=1}^{K}w'_k h_k\sqrt{\frac{d_1}{N_T}}\right) - \left(\sum_{k=0}^{K}(w'_k)^2\frac{rd_1d_3}{N_k} + c_\epsilon\sum_{\substack{k=1\\k\neq j,l}}^{K}w'_k h_k\sqrt{\frac{d_1}{N_T}} + c_\epsilon\left(w'_j h_l + w'_l h_j\right)\sqrt{\frac{d_1}{N_T}}\right) \\
&= c_\epsilon\sqrt{\frac{d_1}{N_T}}\left(w'_j(h_j - h_l) + w'_l(h_l - h_j)\right) \\
&= c_\epsilon\sqrt{\frac{d_1}{N_T}}\, w'_l(h_l - h_j).
\end{aligned}
$$

Since $c_\epsilon\sqrt{d_1/N_T} > 0$ and $w'_l > 0$, it follows that $h_l - h_j \leq 0$, i.e., $h_l \leq h_j$. This contradicts the assumption that $h_l > h_j$.

Claim 2: Next, suppose by contradiction that $w'_0 = 0$. Then, according to the KKT condition (see Eq. (27)), we have $\xi = -\eta_0 \leq 0$. However, the normalization constraint in Eq. (25) requires $\sum_{k=0}^{K} w'_k = 1$. which implies that the total weight assigned to the source domains (i.e., for $k \geq 1$) is 1. Following similar arguments as in the proof of Theorem 2, we deduce that there exists a positive integer $K^\sharp$ such that

$$
\frac{N_S}{2rd_1d_3}\sum_{k=1}^{K^\sharp}\left(\xi - c_\epsilon\, h_{(k)}\sqrt{\frac{d_1}{N_T}}\right) = 1.
$$

This implies

$$
\xi = \frac{1}{K^\sharp}\left(\frac{2rd_1d_3}{N_S} + c_\epsilon\sum_{k=1}^{K^\sharp}h_{(k)}\sqrt{\frac{d_1}{N_T}}\right) > 0,
$$

which contradicts the earlier conclusion that $\xi \leq 0$.

Since both claims lead to a contradiction, we conclude that in any optimal solution the weights must satisfy $w'_j > 0$ for sources with sufficiently small h_k and that w'_0 is strictly positive. This completes the proof.

References

Duan, Y. and Wang, K. Adaptive and robust multi-task learning. *The Annals of Statistics*, 51(5):2015–2039, 2023.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *ICML*, pp. 272–279, 2008.

He, Z., Sun, Y., Liu, J., and Li, R. Adatrans: Feature-wise and sample-wise adaptive transfer learning for high-dimensional regression. *arXiv preprint arXiv:2403.13565*, 2024.

Shalev-Shwartz, S., Singer, Y., Bennett, K. P., and Parrado-Hernández, E. Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, 7(7), 2006.

Wang, A., Song, X., et al. Generalized dantzig selector for low-tubal-rank tensor recovery. In *ICASSP*, pp. 3427–3431, 2019.