Learning Linear and Nonlinear Low-Rank Structure in Multi-Task Learning - Supplementary Material

Yi Zhang, Yu Zhang, and Wei Wang

I. Proofs

For completeness, we list the lemma and theorems to be proved in the following.

Lemma 1: The right-hand side of Eq. (2) has $2^{p-1}-1$ distinct summands.

Theorem 1: The dual norm of the GTTN defined in Eq. (2) can be computed as

$$\|\|\mathcal{W}\|\|_{*^\star} = \min_{\sum_{\substack{\mathbf{s}\neq\emptyset\\ \mathbf{s}\subset[p]}} \alpha_{\mathbf{s}}\mathcal{Y}^{(\mathbf{s})}=\mathcal{W}} \max_{\substack{\mathbf{s}\neq\emptyset\\ \mathbf{s}\subset[p]}} \|\mathcal{Y}^{(\mathbf{s})}_{\{\mathbf{s}\}}\|_\infty,$$

where $\mathcal{Y}^{(\mathbf{s})}$ is a variable indexed by $\mathbf{s}$ and $\|\cdot\|_\infty$ denotes the spectral norm of a matrix that is equal to the maximum singular value.

Theorem 2: For the solution $\hat{\mathcal{W}}$ of problem (5) and $\delta>0$, with probability at least $1-\delta$, we have

$$L(\hat{\mathcal{W}}) \leq \hat{L}(\hat{\mathcal{W}}) + \frac{2\rho\gamma C}{mn_0}\min_{\substack{\mathbf{s}\neq\emptyset\\ \mathbf{s}\subset[p]}}\left(\frac{\kappa m\sqrt{\ln d_{\mathbf{s}}}}{\alpha_{\mathbf{s}}n_0 d}+\frac{\ln d_{\mathbf{s}}}{\alpha_{\mathbf{s}}n_0}\right)+\sqrt{\frac{2}{m}\ln\frac{1}{\delta}}.$$

In the following sections, we provide proofs for those lemma and theorems.

A. Proof for Lemma 1

Proof: For a valid $\|\mathcal{W}_{\{\mathbf{s}\}}\|_*$, it is required that $\mathbf{s}$ and $\neg\mathbf{s}$ should not be empty, implying that $\mathbf{s}\neq\emptyset$ and $\mathbf{s}\neq[p]$. So the total number of valid summands in the right-hand side of Eq. (2) is 2^p-2. Based on the definition of $\mathcal{W}_{\{\mathbf{s}\}}$, we can see that $\mathcal{W}_{\{\mathbf{s}\}}$ is just the transpose of $\mathcal{W}_{\{\neg\mathbf{s}\}}$, making $\|\mathcal{W}_{\{\mathbf{s}\}}\|_* = \|\mathcal{W}_{\{\neg\mathbf{s}\}}\|_*$. So for $\|\mathcal{W}_{\{\mathbf{s}\}}\|_*$, there will always be an equivalent $\|\mathcal{W}_{\{\neg\mathbf{s}\}}\|_*$, leading to $2^{p-1}-1$ distinct summands in the right-hand side of Eq. (2). ∎

B. Proof for Theorem 1

Proof: We define a linear operator

$$\Phi(\mathcal{W}) = [\mathrm{vec}(\alpha_{\{[1]\}}\mathcal{W}_{\{[1]\}}); \ldots; \alpha_{\{[2:p]\}}\mathrm{vec}(\mathcal{W}_{\{[2:p]\}})],$$

where $\mathrm{vec}(\cdot)$ denotes the columnwise concatenation of a matrix and $[i:j]$ denotes a set of successively integers from i to j. We define the q norm as

$$\|\mathbf{y}\|_q = \sum_i \|\mathcal{Y}^{(\pi(i))}_{\{\pi(i)\}}\|_*,$$

where $\mathcal{Y}^{(\pi(i))}_{\{\pi(i)\}}$ denotes the inverse vectorization of a subvector $\mathbf{y}_{(i-1)*N+1:kN}$ of $\mathbf{y}$ into a $\prod_{j\in\pi(i)}p_j \times \prod_{j\in\neg\pi(i)}p_j$ matrix where $N=\prod_{j=1}^p d_j$ and $\pi(i)$ transforms an index i into a subset of $[p]$. Based on the definition of the dual norm, we have

$$\|\|\mathcal{W}\|\|_{*^\star} = \sup\langle\mathcal{W},\mathcal{X}\rangle \text{ s.t. } \|\|\mathcal{X}\|\|_* \leq 1,$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product between two tensors with equal size. Since this maximization problem satisfies the Slater's condition, the strong duality holds. Thus, due to Fenchel's duality theorem, we have

$$\sup_{\mathcal{X}}(\langle\mathcal{W},\mathcal{X}\rangle - \delta(\|\|\mathcal{X}\|\|_*\leq 1)) = \inf_{\mathbf{y}}(\delta(-\Phi^T(\mathbf{y})+\mathcal{X}) + \|\mathbf{y}\|_{q^\star}),$$

where $\delta(C)$ is an indicator function of condition C and it outputs 0 when C is true and otherwise ∞. Since the dual norm of the trace norm is the spectral norm, then according to Lemma 3 in [1], we reach the conclusion. ∎

C. Proof for Theorem 2

Before presenting the proof for Theorem 2, we first prove the following theorem.

Theorem 3: σ_j^i, a Rademacher variable, is an uniform $\{\pm1\}$-valued random variable, and $\mathcal{M}$ is a $d_1\times\ldots\times d_{p-1}\times d_p$ tensor with $\mathcal{M}_i=\sum_{j=1}^{n_0}\frac{1}{n_0}\sigma_j^i\mathbf{x}_j^i$, where d_p equals m. Then we have

$$\mathbb{E}[\|\|\mathcal{M}\|\|_{*^\star}] \leq \min_{\substack{\mathbf{s}\neq\emptyset\\ \mathbf{s}\subset[p]}}\frac{C}{\alpha_{\mathbf{s}}}\left(\frac{\kappa m}{n_0 d}\sqrt{\ln d_{\mathbf{s}}}+\frac{\ln d_{\mathbf{s}}}{n_0}\right).$$

where $d_{\mathbf{s}}=\prod_{i\in\mathbf{s}}d_i+\prod_{j\in\neg\mathbf{s}}d_j$ and C is an absolute constant.

Proof: According to Theorem 1, we have

$$\|\|\mathcal{M}\|\|_{*^\star} = \min_{\sum_{\substack{\mathbf{s}\neq\emptyset\\ \mathbf{s}\subset[p]}}\alpha_{\mathbf{s}}\mathcal{Y}^{(\mathbf{s})}=\mathcal{M}} \max_{\substack{\mathbf{s}\neq\emptyset\\ \mathbf{s}\subset[p]}}\|\mathcal{Y}^{(\mathbf{s})}_{\{\mathbf{s}\}}\|_\infty$$

Since for each $\mathbf{s}$ we can make $\alpha_{\mathbf{s}}\mathcal{Y}^{\mathbf{s}}$ equal to $\mathcal{M}$, we have

$$\|\|\mathcal{M}\|\|_{*^\star} \leq \frac{1}{\alpha_{\mathbf{s}}}\|\mathcal{M}_{\{\mathbf{s}\}}\|_\infty \ \forall\mathbf{s}\neq\emptyset,\ \mathbf{s}\subset[p],$$

which implies that

$$\|\|\mathcal{M}\|\|_{*^\star} \leq \min_{\mathbf{s}}\frac{1}{\alpha_{\mathbf{s}}}\|\mathcal{M}_{\{\mathbf{s}\}}\|_\infty.$$

So we can get

$$\mathbb{E}[\|\|\mathcal{M}\|\|_{*^\star}] \leq \mathbb{E}\left[\min_{\mathbf{s}}\frac{1}{\alpha_{\mathbf{s}}}\|\mathcal{M}_{\{\mathbf{s}\}}\|_\infty\right] \leq \min_{\mathbf{s}}\mathbb{E}\left[\frac{1}{\alpha_{\mathbf{s}}}\|\mathcal{M}_{\{\mathbf{s}\}}\|_\infty\right].$$

Based on Theorem 6.1 in [2], we can upper-bound each expectation as

$$\mathbb{E}\left[\|\mathcal{M}_{\{\mathbf{s}\}}\|_\infty\right] \leq C(\sigma_{\mathbf{s}}\sqrt{\ln d_{\mathbf{s}}}+\psi_{\mathbf{s}}\ln d_{\mathbf{s}}),$$

Yi Zhang is with Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology. Yu Zhang is with Department of Computer Science and Engineering and Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, and he is also affiliated with Peng Cheng Laboratory. Wei Wang is with National Key Laboratory for Novel Software Technology, Nanjing University.
E-mail: 11930380@mail.sustech.edu.cn, yu.zhang.ust@gmail.com, wangw@nju.edu.cn.
Corresponding author: Yu Zhang.

where $\mathcal{Z}^{i,j}$ is a $d_1 \times \ldots \times d_{p-1} \times d_p$ zero tensor with only the ith slice along the last axis equal to $\frac{1}{n_0}\sigma_j^i \mathbf{x}_j^i$, $\psi_{\mathbf{s}}$ needs to satisfy $\psi_{\mathbf{s}} \geq \|\mathcal{Z}^{i,j}_{\{\mathbf{s}\}}\|_\infty$, and

$$\sigma_{\mathbf{s}}^2 = \max\Big(\big\|\sum_{i=1}^m\sum_{j=1}^{n_0}\mathbb{E}\big[\mathcal{Z}^{i,j}_{\{\mathbf{s}\}}(\mathcal{Z}^{i,j}_{\{\mathbf{s}\}})^T\big]\big\|_\infty, \big\|\sum_{i=1}^m\sum_{j=1}^{n_0}\mathbb{E}\big[(\mathcal{Z}^{i,j}_{\{\mathbf{s}\}})^T\mathcal{Z}^{i,j}_{\{\mathbf{s}\}}\big]\big\|_\infty\Big).$$

As the Frobenius norm of a matrix is larger than its spectral norm, $\|\mathcal{Z}^{i,j}_{\{\mathbf{s}\}}\|_\infty \leq \frac{1}{n_0}$ and we simply set $\psi_{\mathbf{s}} = \frac{1}{n_0}$. For $\sigma_{\mathbf{s}}$, we have

$$\mathbb{E}\Big[\sum_{j=1}^{n_0}\mathcal{Z}^{i,j}_{\{\mathbf{s}\}}(\mathcal{Z}^{i,j}_{\{\mathbf{s}\}})^T\Big] = \frac{1}{n_0}\mathbf{C}_{\mathbf{s}-\{p\}} \preceq \frac{\kappa}{n_0 d}\mathbf{I},$$

implying that

$$\Big\|\sum_{i=1}^m\sum_{j=1}^{n_0}\mathbb{E}\big[\mathcal{Z}^{i,j}_{\{\mathbf{s}\}}(\mathcal{Z}^{i,j}_{\{\mathbf{s}\}})^T\big]\Big\|_\infty \leq \frac{\kappa m}{n_0 d}.$$

Similarly, we have

$$\mathbb{E}\Big[\sum_{j=1}^{n_0}(\mathcal{Z}^{i,j}_{\{\mathbf{s}\}})^T\mathcal{Z}^{i,j}_{\{\mathbf{s}\}}\Big] = \mathrm{diag}\left(\frac{\mathrm{tr}(\mathbf{C}_{\mathbf{s}-\{p\}})}{n_0}\right) \preceq \frac{\kappa}{n_0 d}\mathbf{I},$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix and $\mathrm{diag}(\cdot)$ converts a vector to a diagonal matrix. This inequality implies

$$\Big\|\sum_{i=1}^m\sum_{j=1}^{n_0}\mathbb{E}\big[\mathcal{Z}^{i,j}_{\{\mathbf{s}\}}(\mathcal{Z}^{i,j}_{\{\mathbf{s}\}})^T\big]\Big\|_\infty \leq \frac{\kappa m}{n_0 d}.$$

By combining the above inequalities, we reach the conclusion. ∎

Then we can prove Theorem 2 as follows.

Proof: By following [3], we have

$$\begin{aligned} L(\hat{\mathcal{W}}) &\leq \hat{L}(\hat{\mathcal{W}}) + \sup_{|||\mathcal{W}|||_* \leq \gamma}\Big\{L(\mathcal{W}) - \hat{L}(\mathcal{W})\Big\} \\ &= \hat{L}(\hat{\mathcal{W}}) + \sup_{|||\mathcal{W}|||_* \leq \gamma}\Big\{\mathbb{E}[\hat{L}(\mathcal{W})] - \hat{L}(\mathcal{W})\Big\}. \end{aligned}$$

When each pair of the training data $(\mathbf{x}_j^i, y_j^i)$ changes, the random variable $\sup_{|||\mathcal{W}|||_* \leq \gamma}\Big\{\mathbb{E}[\hat{L}(\mathcal{W})] - \hat{L}(\mathcal{W})\Big\}$ can change by no more than $\frac{2}{mn_0}$ due to the boundedness of the loss function $l(\cdot,\cdot)$. Then by McDiarmid's inequality, we can get

$$\begin{aligned} &P\left(\sup_{\mathcal{W}\in\mathcal{C}}\Big\{\mathbb{E}[\hat{L}(\mathcal{W})] - \hat{L}(\mathcal{W})\Big\} - \mathbb{E}\left[\sup_{\mathcal{W}\in\mathcal{C}}\Big\{\mathbb{E}[\hat{L}(\mathcal{W})] - \hat{L}(\mathcal{W})\Big\}\right] \geq t\right) \\ &\leq \exp\left\{-\frac{t^2 mn_0}{2}\right\}, \end{aligned}$$

where $P(\cdot)$ denotes the probability and $\mathcal{C} = \{\mathcal{W}|||\mathcal{W}|||_* \leq \gamma\}$. This inequality implies that with probability at least $1-\delta$,

$$\begin{aligned} &\sup_{\mathcal{W}\in\mathcal{C}}\Big\{\mathbb{E}[\hat{L}(\mathcal{W})] - \hat{L}(\mathcal{W})\Big\} \\ \leq &\mathbb{E}\left[\sup_{\mathcal{W}\in\mathcal{C}}\Big\{\mathbb{E}[\hat{L}(\mathcal{W})] - \hat{L}(\mathcal{W})\Big\}\right] + \sqrt{\frac{2}{mn_0}\ln\frac{1}{\delta}}. \end{aligned}$$

Based on the the property of the Rademacher complexity, we have

$$\begin{aligned} &\mathbb{E}\left[\sup_{\mathcal{W}\in\mathcal{C}}\Big\{\mathbb{E}[\hat{L}(\mathcal{W})] - \hat{L}(\mathcal{W})\Big\}\right] \\ \leq &2\rho\mathbb{E}\left[\sup_{\mathcal{W}\in\mathcal{C}}\left\{\frac{1}{mn_0}\sum_{i=1}^m\sum_{j=1}^{n_0}\sigma_j^i f_i(\mathbf{x}_j^i)\right\}\right]. \end{aligned}$$

Then based on the definition of $\mathcal{M}$ and the Hölder's inequality, we have

$$\sup_{\mathcal{W}\in\mathcal{C}}\left\{\frac{1}{mn_0}\sum_{i=1}^m\sum_{j=1}^{n_0}\sigma_j^i f_i(\mathbf{x}_j^i)\right\} \leq \frac{\gamma}{m}|||\mathcal{M}|||_{*^*}.$$

By combining the above inequalities, with probability at least $1-\delta$, we have

$$L(\hat{\mathcal{W}}) \leq \hat{L}(\hat{\mathcal{W}}) + \frac{2\rho\gamma}{m}\mathbb{E}[|||\mathcal{M}|||_{*^*}] + \sqrt{\frac{2}{mn_0}\ln\frac{1}{\delta}}.$$

Then by incorporating Theorem 3 into this inequality, we reach the conclusion. ∎

II. Introduction of Three Strategies to Learn $\boldsymbol{\alpha}$

In this section, we introduce the other 3 strategies to learn $\boldsymbol{\alpha}$ as [4] did. First, in the following lemma, we can derive an equivalent problem to problem (3) in the paper by eliminating $\boldsymbol{\alpha}$.

Lemma 2: Problem (3) in the paper is equivalent to

$$\min_{\boldsymbol{\Theta}}\sum_{i=1}^m\frac{1}{n_i}\sum_{j=1}^{n_i}l(f_i(\mathbf{x}_j^i;\boldsymbol{\Theta}), y_j^i) + \lambda\min_{\substack{\mathbf{s}\subset[p]\\ \mathbf{s}\neq\emptyset}}\|\mathcal{W}_{\{\mathbf{s}\}}\|_*, \tag{1}$$

Proof: We first analyze the following problem as

$$\min_{\boldsymbol{\Theta},\boldsymbol{\alpha}\in\mathcal{C}_{\boldsymbol{\alpha}}}\sum_{i=1}^m\frac{1}{n_i}\sum_{j=1}^{n_i}l(f_i(\mathbf{x}_j^i;\boldsymbol{\Theta}), y_j^i) + \lambda|||\mathcal{W}|||_*. \tag{2}$$

Based on Eq. (2) in the paper, we rewrite problem (2) as

$$\min_{\boldsymbol{\Theta},\boldsymbol{\alpha}\in\mathcal{C}_{\boldsymbol{\alpha}}}\sum_{i=1}^m\frac{1}{n_i}\sum_{j=1}^{n_i}l(f_i(\mathbf{x}_j^i;\boldsymbol{\Theta}), y_j^i) + \lambda\sum_{\substack{\mathbf{s}\subset[p]\\ \mathbf{s}\neq\emptyset}}\alpha_{\mathbf{s}}\|\mathcal{W}_{\{\mathbf{s}\}}\|_*,$$

which is equivalent to

$$\min_{\boldsymbol{\Theta}}\sum_{i=1}^m\frac{1}{n_i}\sum_{j=1}^{n_i}l(f_i(\mathbf{x}_j^i;\boldsymbol{\Theta}), y_j^i) + \lambda\min_{\boldsymbol{\alpha}\in\mathcal{C}_{\boldsymbol{\alpha}}}\sum_{\substack{\mathbf{s}\subset[p]\\ \mathbf{s}\neq\emptyset}}\alpha_{\mathbf{s}}\|\mathcal{W}_{\{\mathbf{s}\}}\|_*.$$

We can prove that

$$\min_{\boldsymbol{\alpha}\in\mathcal{C}_{\boldsymbol{\alpha}}}\sum_{\substack{\mathbf{s}\subset[p]\\ \mathbf{s}\neq\emptyset}}\alpha_{\mathbf{s}}\|\mathcal{W}_{\{\mathbf{s}\}}\|_* = \min_{\substack{\mathbf{s}\subset[p]\\ \mathbf{s}\neq\emptyset}}\|\mathcal{W}_{\{\mathbf{s}\}}\|_*.$$

To see that, the optimization problem in the left-hand side of the above equation is a linear programming problem with respect to $\boldsymbol{\alpha}$. It is easy to show that $\sum_{\mathbf{s}\subset[p]}\alpha_{\mathbf{s}}\|\mathcal{W}_{\{\mathbf{s}\}}\|_* \geq \min_{\substack{\mathbf{s}\subset[p]\\ \mathbf{s}\neq\emptyset}}\|\mathcal{W}_{\{\mathbf{s}\}}\|_*$ for $\boldsymbol{\alpha}\in\mathcal{C}_{\boldsymbol{\alpha}}$, where the equality holds when the corresponding coefficient for $\min_{\substack{\mathbf{s}\subset[p]\\ \mathbf{s}\neq\emptyset}}\|\mathcal{W}_{\{\mathbf{s}\}}\|_*$ equals 1 and other coefficients equals 0, which implies that we reach the conclusion. ∎

According to Lemma 2, learning $\boldsymbol{\alpha}$ will tend to choose a tensor flattening with the minimal matrix trace norm. Hence, *the second strategy* is to directly solve problem (1). Though problem (3) in the paper is proved to be equivalent to problem (1), each $\alpha_{\mathbf{s}}$ learned in problem (3) in the paper cannot achieve 0 or 1 exactly due to the softmax reparameterization and

hence numerical solutions of those two problems are slightly different. A benefit of using the first strategy is that the learned $\boldsymbol{\alpha}$ in problem (3) in the paper can visualize the importance of each tensor flattening, which can improve the interpretability of the learning model as we will see in experiments.

Problem (1) is to regularize $\mathcal{W}$ according to its tensor flattening with the minimal matrix trace norm, which can be viewed as an optimistic way to learn $\boldsymbol{\alpha}$. *The third strategy* takes an opposite and pessimistic way to regularize $\mathcal{W}$ according to its tensor flattening with the maximum matrix trace norm. It pays attention to flattenings with the largest trace norm corresponding to those who are the most unlikely to have a low-rank structure. Correspondingly its objective function is formulated as

$$\min_{\boldsymbol{\Theta}} \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} l(f_i(\mathbf{x}_j^i; \boldsymbol{\Theta}), y_j^i) + \lambda \max_{\substack{\mathbf{s} \subset [p] \\ \mathbf{s} \neq \emptyset}} \|\mathcal{W}_{\{\mathbf{s}\}}\|_*. \tag{3}$$

Problem (3) can also be formulated as an optimization problem with $\{\alpha_{\mathbf{s}}\}$ since

$$\max_{\substack{\mathbf{s} \subset [p] \\ \mathbf{s} \neq \emptyset}} \|\mathcal{W}_{\{\mathbf{s}\}}\|_* = \max_{\boldsymbol{\alpha} \in \mathcal{C}_{\boldsymbol{\alpha}}} \sum_{\substack{\mathbf{s} \subset [p] \\ \mathbf{s} \neq \emptyset}} \alpha_{\mathbf{s}} \|\mathcal{W}_{\{\mathbf{s}\}}\|_*.$$

Inspired by meta learning or hyperparameter optimization [5], *the fourth strategy* views $\boldsymbol{\alpha}$ as hyperparameters to be learned. Here the entire training dataset is partitioned into two parts, where the larger part $\mathcal{D}_{tr}$ (i.e., 70% of the entire training dataset) as a training set is used to learn $\boldsymbol{\Theta}$ and the smaller part $\mathcal{D}_{val}$ (i.e., the rest 30%) as a validation set is to learn $\boldsymbol{\alpha}$. Formally, the objective function in this strategy is formulated as

$$\min_{\boldsymbol{\alpha}} L(\mathcal{D}_{val}, \boldsymbol{\Theta}^*) \text{ s.t. } \boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}} h(\mathcal{D}_{tr}, \boldsymbol{\Theta}, \boldsymbol{\alpha}), \tag{4}$$

where $L(\mathcal{D}, \boldsymbol{\Theta}) = \sum_{i=1}^{m} \frac{1}{|\mathcal{D}^i|} \sum_{(\mathbf{x},y) \in \mathcal{D}^i} l(f_i(\mathbf{x}; \boldsymbol{\Theta}), y)$ denotes the loss of different tasks on a dataset $\mathcal{D}$ with $\mathcal{D}^i$ as the subset for the ith task and $|\mathcal{D}^i|$ as its cardinality, and $h(\mathcal{D}_{tr}, \boldsymbol{\Theta}, \boldsymbol{\alpha}) = L(\mathcal{D}_{tr}, \boldsymbol{\Theta}) + \lambda |||\mathcal{W}|||_*$ is just the objective function of problem (3) in the paper on $\mathcal{D}_{tr}$ instead of the entire training dataset. In problem (4), $\boldsymbol{\Theta}^*$ is a function of $\boldsymbol{\alpha}$ and hence the loss on $\mathcal{D}_{val}$ can be used as an objective function in problem (4) to learn $\boldsymbol{\alpha}$. The bilevel optimization method proposed in [5] can be used to solve problem (4).

REFERENCES

[1] R. Tomioka and T. Suzuki, "Convex tensor decomposition via structured schatten norm regularization," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 1331–1339.

[2] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[3] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[4] Y. Zhang, Y. Zhang, and W. Wang, "Multi-task learning via generalized tensor trace norm," in *Proceedings of The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021, pp. 2254–2262.

[5] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 1563–1572.