Hyperparameter Transfer Using Item Response Theory

Takeaki Sakabe and Yuko Sakurai
Graduate School of Engineering
Nagoya Institute of Technology
Nagoya, Japan

Hisashi Kashima
Graduate School of Informatics
Kyoto University
Kyoto, Japan

Satoshi Oyama
Graduate School of Data Science
Nagoya City University
Nagoya, Japan

***Abstract*—Hyperparameter optimization (HPO) is computationally expensive due to large search spaces and dataset-dependent optimal configurations. To reduce this cost, transfer-based HPO methods have been proposed. However, existing methods rely solely on observed performance metrics, failing to disentangle the latent capability of hyperparameter configurations from dataset characteristics, which can cause biased transfer. In this paper, we propose a hyperparameter transfer method using Item Response Theory (IRT), a testing theory that enables the independent estimation of problem characteristics and examinee capability. By associating hyperparameter configurations with examinees and individual data samples with test items, the proposed method estimates the latent capability of configurations on source datasets. This is then combined with the dataset characteristics estimated from a small number of evaluations on a target dataset, enabling efficient hyperparameter transfer. Experimental results on multiple datasets demonstrate that the proposed method can select effective hyperparameter configurations with fewer evaluation trials compared to an existing method.**

***Index Terms*—Hyperparameter transfer, Item Response Theory, Performance prediction**

I. Introduction

In recent years, as the performance of machine learning models has continued to improve, the importance of HPO for fully exploiting their potential has grown significantly [1]–[3]. Hyperparameters, such as learning rates, optimization algorithms, and network architectures, are values specified outside the training process and are known to have a substantial impact on training dynamics and final model performance. Consequently, selecting appropriate hyperparameter configurations is a crucial step in building practical machine learning systems.

However, optimal hyperparameter configurations strongly depend on the characteristics of the dataset and task; even with the same model architecture, differences in data distribution or difficulty composition can lead to substantially different optimal solutions [4], [5]. Moreover, the search space of HPO is generally large and often exhibits a complex structure that combines continuous and discrete variables. In addition, for many machine learning models, particularly deep learning models, a single evaluation requires substantial computational time. As a result, rerunning HPO from scratch to identify suitable hyperparameter configurations for each new dataset or task is computationally impractical.

To address this computational cost issue, transfer-based HPO methods that leverage past optimization results have been proposed [6], [7]. However, many existing approaches evaluate hyperparameter configurations solely based on observed performance metrics, such as accuracy or loss, obtained on source datasets. Observed performance reflects not only the latent capability of a model configuration but also factors such as the difficulty composition of samples and biases in the evaluation data. Consequently, existing transfer-based HPO methods do not explicitly disentangle the latent capability of hyperparameter configurations from dataset-specific characteristics. This limitation can lead to ineffective hyperparameter transfer when the source and target datasets differ substantially in their properties [8].

In this study, we focus on IRT as a framework to address this issue. IRT is a statistical model that estimates examinee ability and item characteristics separately from observed correctness data and has been widely used in the field of educational measurement [9]. By associating examinees with hyperparameter configurations and items with individual samples in a dataset, we use IRT to disentangle the latent capability of hyperparameter configurations from dataset-specific characteristics based on correctness data, and perform hyperparameter transfer based on these estimates. Figure 1 illustrates an overview of our framework. Specifically, we estimate the latent capabilities of hyperparameter configurations using a source dataset, while dataset characteristics of the target dataset are inferred from a small number of evaluation results. By integrating these estimates, the proposed framework enables efficient performance prediction and hyperparameter transfer without requiring extensive search.

In the evaluation experiments, we compared the proposed method with an existing transfer-based HPO method using multiple image classification datasets. The results demonstrate that our method can effectively identify promising hyperparameter configurations even under a limited number of evaluations. The implementation is publicly available.[1]

This work was partially supported by JSPS KAKENHI (Grant Number JP24K01112) and JST CREST (Grant Numbers JPMJCR21D1 and JPMJCR2564).

[1]https://github.com/T-Sakabe/irt-hyperparameter-transfer

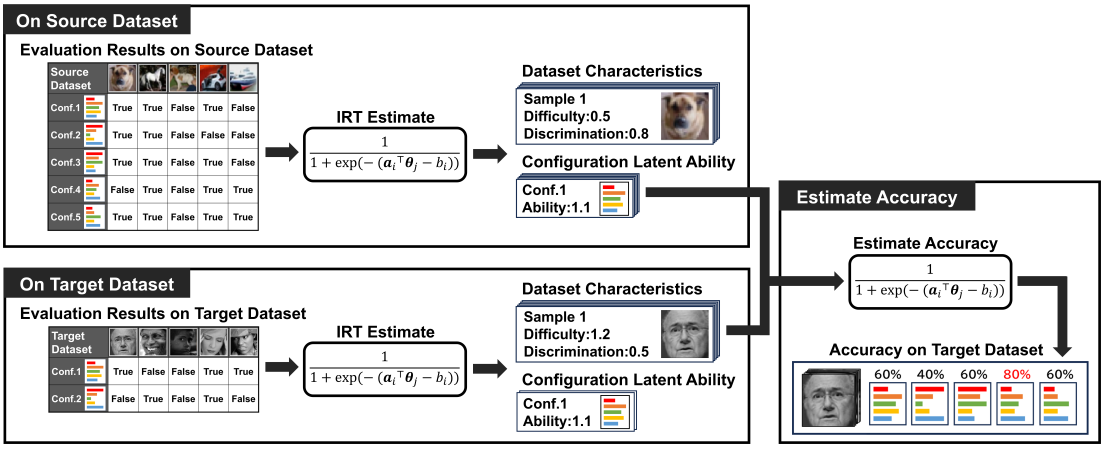

Fig. 1. The figure illustrates how the latent abilities of hyperparameter configurations and dataset-specific characteristics are disentangled using IRT, and how performance prediction is enabled with only a limited number of target evaluations.

II. Related Work

To reduce the computational cost of HPO, numerous transfer-based HPO methods have been proposed, which leverage optimization histories obtained from previously solved tasks to accelerate the search process on a new target task [6], [7], [10]. These methods aim to improve search efficiency on a target dataset by reusing knowledge acquired from source tasks.

Early studies on transfer-based HPO primarily employed warm-start approaches, which directly transferred hyperparameter configurations that achieved high performance on source tasks as initial candidates for optimization on a target task [11], [12]. Subsequently, similarity-based methods were proposed, in which task similarity is defined based on dataset meta-features or performance correlations, and optimization histories from similar tasks are weighted and reused [7], [13]. More recently, meta-learning–based approaches that learn structures shared across multiple tasks have been explored, reporting further improvements in the efficiency of hyperparameter search [6], [14].

In addition to these approaches, recent work has also explored transfer HPO methods that focus on transferring information from the exploration process [15], [16]. One such recent method is MCTS-transfer [17]. It transfers and integrates search trees and exploration statistics constructed by Monte Carlo Tree Search (MCTS) into a target task, enabling efficient exploration of promising regions while maintaining balanced exploration.

Despite their differences in how knowledge is transferred, existing transfer-based HPO methods commonly rely on observed performance values to guide the search process. As a result, dataset-specific characteristics and the intrinsic effectiveness of hyperparameter configurations are not explicitly disentangled. This limitation may lead to degraded transfer performance when the difficulty structures differ across datasets.

III. Preliminaries

In this study, we introduce a new framework for hyperparameter transfer based on IRT, and estimate model performance on a target dataset from a limited number of evaluation trials.

IRT is a test theory that estimates examinee ability and item characteristics separately from observed binary response data [9], [18]. An important property of IRT is that estimated parameters can be represented on a common scale across different tests, which enables comparison across datasets.

A. Multidimensional Two-Parameter Logistic Model

As a representative IRT model, we use the Two-Parameter Logistic Model (2PLM), which characterizes each item by its discrimination and difficulty [19]. In this study, we adopt its multidimensional extension to model multiple latent abilities [20]. Let $\boldsymbol{\theta} \in \mathbb{R}^D$ denote a D-dimensional ability vector of an examinee. The probability that the examinee correctly responds to item i is defined as

$$P_i(\boldsymbol{\theta}) = \frac{1}{1+\exp\bigl(-(\boldsymbol{a}_i^\top \boldsymbol{\theta} - b_i)\bigr)}, \tag{1}$$

where $\boldsymbol{a}_i \in \mathbb{R}^D$ represents the discrimination vector of item i, and $b_i \in \mathbb{R}$ denotes its difficulty. The discrimination parameter controls the sensitivity of the response probability to ability, while the difficulty parameter determines the location of the item on the ability scale.

B. Parameter Estimation via Joint Maximum Likelihood

We estimate the ability and item parameters of the multidimensional 2PLM from the observed response data.

Let I be the number of items and J the number of examinees, and let $y_{ij} \in \{0,1\}$ denote the binary response of examinee j to item i. Under the 2PLM, the response probability is given by $p_{ij} = P(y_{ij} = 1 \mid \boldsymbol{\theta}_j, \boldsymbol{a}_i, b_i)$. The likelihood of the observed response matrix is then written as

$$L = \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{y_{ij}} (1-p_{ij})^{1-y_{ij}}. \tag{2}$$

In this work, both ability parameters $\{\boldsymbol{\theta}_j\}$ and item parameters $\{\boldsymbol{a}_i, b_i\}$ are treated as unknown, and are estimated simultaneously using Joint Maximum Likelihood (JML) estimation [21].

C. Linking by Procrustes Transformation

In IRT, the latent ability space is not uniquely determined, and estimated parameters may differ by sign flips or rotations. As a result, parameters estimated from different tests cannot be directly compared. To address this issue, linking methods are commonly employed. In this study, we adopt an orthogonal Procrustes transformation based on a set of common examinees [22], [23].

Let $\boldsymbol{\theta}_j^{(A)}$ and $\boldsymbol{\theta}_j^{(B)}$ denote the estimated ability vectors of a common examinee j obtained from tests A and B, respectively. The orthogonal transformation matrix $\boldsymbol{R}$ is obtained by solving

$$\boldsymbol{R} = \arg\min_{\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}} \sum_{j \in C} \left\| \boldsymbol{\theta}_j^{(B)} \boldsymbol{R} - \boldsymbol{\theta}_j^{(A)} \right\|^2, \tag{3}$$

where C denotes the set of common examinees. The resulting transformation is applied to both ability and discrimination parameters to align the latent scales. Difficulty parameters are not affected by the rotation and are therefore left unchanged.

IV. Proposed Method

We propose an IRT-based hyperparameter transfer method. The proposed method consists of three stages: (1) estimation of ability parameters on a source dataset, (2) estimation of item parameters on a target dataset, and (3) linking and performance prediction based on estimated parameters. In IRT, we associate examinees with hyperparameter configurations and questions with individual samples in the dataset. The dimensionality of the ability parameter D is specified in advance and shared across all stages. In this study, we focus on the case where the hyperparameter search space is a finite discrete space.

A. Ability Estimation on the Source Dataset

On the source dataset, we first evaluate all hyperparameter configurations and collect their prediction results. Specifically, each configuration is trained and evaluated on the source dataset, and the correctness of its prediction for each sample during evaluation is recorded. Let J denote the number of hyperparameter configurations and I the number of samples in the source dataset. We define

$$y_{ij} \in \{0, 1\} \tag{4}$$

as the binary response indicating whether the hyperparameter configuration j correctly classifies the sample i. The response matrix $\{y_{ij}\}$ is constructed from these binary outcomes and used as input to the IRT model.

In typical machine learning datasets, the number of samples is often very large. Using all samples as items in IRT may cause instability in parameter estimation.

To address this issue, we estimate ability parameters based on a fixed-size subset of samples selected from the response matrix $\{y_{ij}\}$. The subset size is determined based on the number of hyperparameter configurations to ensure stable parameter estimation. To improve estimation stability, the sample subset is selected such that the distribution of sample-wise accuracy

$$p_i = \frac{1}{J} \sum_{j=1}^{J} y_{ij} \tag{5}$$

approximately follows a normal-shaped distribution. If the estimation does not converge, sample selection and parameter estimation are repeated until stable ability parameters $\boldsymbol{\theta}^{(s)}$ are obtained according to the convergence criterion of the parameter estimation procedure.

B. Item Parameter Estimation on the Target Dataset

On the target dataset, evaluating all hyperparameter configurations is computationally prohibitive. Therefore, we estimate item parameters using prediction results obtained from a limited number of hyperparameter configurations.

Let J' denote the number of selected hyperparameter configurations evaluated on the target dataset, and I' the number of samples in the target dataset. We define

$$y'_{ij} \in \{0, 1\} \tag{6}$$

as the binary response indicating whether the hyperparameter configuration j correctly classifies the sample i on the target dataset. The response matrix $\{y'_{ij}\}$ constructed from these binary outcomes is used to estimate item parameters of the target dataset.

When the performance of the selected hyperparameter configurations is heavily biased, this parameter estimation becomes unstable. To mitigate this issue, the hyperparameter configurations evaluated on the target dataset are selected based on their performance on the source dataset. Specifically, configurations are chosen such that their accuracies on the source dataset are approximately evenly distributed.

As in the source dataset, using all samples as items may lead to unstable estimation. Therefore, item parameter estimation is performed on a fixed-size subset of samples randomly selected from the target dataset.

To reduce dependency on a specific subset of samples, sample selection and parameter estimation are repeated multiple times, and only converged estimation results are retained. This procedure yields stable estimates of the target item parameters (the discrimination parameters $\boldsymbol{a}^{(t)}$ and difficulty parameters $b^{(t)}$) while controlling both estimation instability and selection-induced bias.

C. Hyperparameter Transfer via Linking and Performance Prediction

The latent spaces estimated on the source and target datasets are obtained independently and thus are not directly comparable. To link these spaces, we apply Procrustes-based linking using hyperparameter configurations evaluated on both datasets as anchors. Through this linking, the target item parameters are mapped onto the source dataset scale and are denoted as $\boldsymbol{a}^{(t \to s)}$ and $b^{(t \to s)}$.

TABLE I
HYPERPARAMETER VALUES OF THE CNN USED IN THIS STUDY

Hyperparameter	Values
Learning rate	0.01, 0.001
Number of training epochs	5, 10
Optimizer	Adam, SGD
Activation function	ReLU, LeakyReLU, ELU
Number of convolutional blocks	2, 3, 4
Channel width	16, 32, 64
Convolutional kernel size	3, 5
Dropout rate	0.0, 0.2

Using the linked item parameters and the ability parameters $\boldsymbol{\theta}_j^{(s)}$ estimated on the source dataset, we predict the performance of each hyperparameter configuration on the target dataset. Let $\mathcal{I}^{(t)}$ denote the set of samples used for estimating the target item parameters. The predicted accuracy of configuration j on the target dataset, $\hat{P}_j^{(t)}$, is computed as

$$\hat{P}_j^{(t)} = \frac{1}{|\mathcal{I}^{(t)}|} \sum_{i \in \mathcal{I}^{(t)}} \frac{1}{1 + \exp\Big(-\Big((\boldsymbol{a}_i^{(t \to s)})^\top \boldsymbol{\theta}_j^{(s)} - b_i^{(t \to s)}\Big)\Big)}. \tag{7}$$

This predicted accuracy enables comparison of hyperparameter configurations on the target dataset without exhaustively evaluating all configurations.

V. EXPERIMENTS

In this section, we evaluate the proposed IRT-based hyperparameter transfer method from two perspectives: (1) the validity of the estimated item parameters, and (2) the performance in transfer-based HPO.

A. Experimental Setup

We conduct experiments on three image classification datasets: CIFAR-10 [24], FER-2013 [25], and Quick, Draw! [26]. For Quick, Draw!, to reduce computational cost, we randomly select 50 classes and use 1,000 images per class.

The target model is a CNN implemented in PyTorch, and the hyperparameter search space is summarized in Table I. All hyperparameter configurations are evaluated in advance, and binary response data are constructed based on classification results, enabling offline transfer-based HPO experiments.

In this study, we consider both one-to-one and two-to-one transfer settings. In the two-to-one transfer setting, the proposed method integrates binary response data obtained from multiple source datasets after aligning the number of samples. All experiments are repeated five times, and the average results are reported.

The proposed method is evaluated with ability dimensions $D = 1$ and $D = 2$, and hyperparameter configurations are evaluated in descending order of the estimated accuracy on the target dataset. The main parameter settings of the proposed method are summarized in Table II.

As a baseline, we use MCTS-transfer [17]. Evaluation metrics include the progression of the maximum accuracy, its area under the curve (AUC), and the internal computation time of each method.

TABLE II
PARAMETER SETTINGS OF THE PROPOSED METHOD

Parameter	Values
Selected items (source)	500
Selected items (target)	500
Selected hyperparameter configurations (target)	10
Converged item parameter estimates (target)	5
Ability dimension D	1, 2

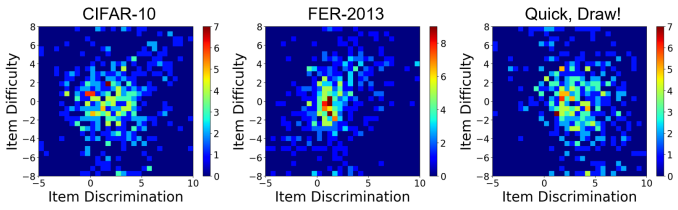

Fig. 2. Distributions of estimated parameters (item discrimination and item difficulty) for CIFAR-10, FER-2013, and Quick, Draw!

B. Analysis of Item Parameters

We first examine the validity of the item parameters estimated by the proposed method. Figure 2 shows the distributions of item parameters estimated by the 1D 2PLM for CIFAR-10, FER-2013, and Quick, Draw!.

While most samples in FER-2013 and Quick, Draw! are distributed in regions with positive discrimination, CIFAR-10 contains a relatively large number of samples with negative discrimination. Samples with negative discrimination correspond to samples for which higher-ability models tend to make incorrect predictions, indicating that sample characteristics differ substantially across datasets. This result suggests that transfer methods based solely on observed accuracy may fail to sufficiently capture dataset-specific characteristics.

Furthermore, Figure 3 illustrates the relationship between sample images in CIFAR-10 and the estimated item parameters. In the 1D 2PLM, samples with negative discrimination often contain unusual backgrounds or compositions, indicating deviations from the features learned by the model. While the first dimension of the 1D and 2D 2PLMs shows similar distributions, the same samples are mapped to different regions across the dimensions of the 2D 2PLM, suggesting that distinct abilities are captured.

These results demonstrate that the proposed method effectively captures dataset-specific sample characteristics through the estimated item parameters.

C. Hyperparameter Transfer Performance

Figure 4 shows the progression of the maximum accuracy on the target dataset under the one-to-one and two-to-one transfer settings, respectively. Table III summarizes the results for both settings, including AUC values and the internal computation times of each method on the source and target datasets.

During the first 10 evaluations, the proposed method tends to show lower performance than MCTS-transfer, as it uniformly evaluates hyperparameter configurations on the target

TABLE III
PERFORMANCE COMPARISON OF TRANSFER-BASED HPO GROUPED BY TARGET DATASET.

Transfer setting	MCTS-transfer			Proposed Method (1D 2PLM)			Proposed Method (2D 2PLM)		
	AUC ↑	Src (s) ↓	Tgt (s) ↓	AUC ↑	Src (s) ↓	Tgt (s) ↓	AUC ↑	Src (s) ↓	Tgt (s) ↓
FER-2013 → CIFAR-10	77.07	4.4	3606.9	77.21	10.6	2.0	77.01	12.3	18.2
Quick, Draw! → CIFAR-10	77.15	4.8	6925.9	77.06	14.4	2.0	76.95	54.4	21.8
FER-2013, Quick, Draw! → CIFAR-10	76.78	7.7	3130.2	77.16	16.1	2.2	77.13	40.8	73.6
CIFAR-10 → FER-2013	59.44	4.4	3727.9	59.35	10.3	1.9	59.50	13.2	21.8
Quick, Draw! → FER-2013	59.64	3.6	4815.6	59.70	15.5	1.9	59.38	55.6	24.8
CIFAR-10, Quick, Draw! → FER-2013	59.60	6.2	4157.7	59.54	15.9	1.9	59.60	36.9	27.9
CIFAR-10 → Quick, Draw!	77.19	4.3	3207.5	77.02	10.5	2.0	76.87	13.2	60.0
FER-2013 → Quick, Draw!	76.86	4.2	3538.7	77.12	10.5	1.9	77.07	12.3	47.2
CIFAR-10, FER-2013 → Quick, Draw!	76.78	8.0	2638.2	77.12	13.6	2.0	76.94	13.1	93.3

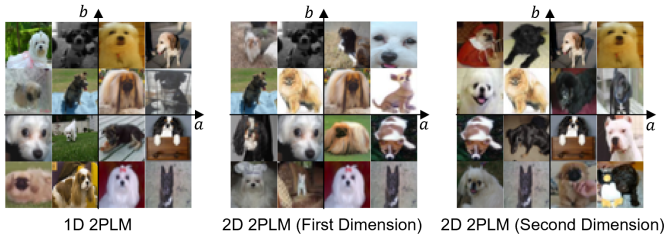

Fig. 3. Relationship between image samples and estimated item parameters (horizontal axis: item discrimination a, vertical axis: item difficulty b)

dataset regardless of their performance. From the 11th evaluation onward, it achieves superior performance by estimating the performance of all hyperparameter configurations based on the IRT framework. This trend is consistently observed in both one-to-one and two-to-one transfer settings.

Furthermore, when comparing the one-to-one and two-to-one transfer results for each method, while MCTS-transfer shows a decrease in AUC when moving from one-to-one to two-to-one transfer, the proposed method achieves AUC improvements in some settings, indicating its more effective integration of multiple source datasets.

From the perspective of computational efficiency, the proposed method significantly reduces the internal computation time on the target dataset. This advantage arises from the difference in the evaluation strategy: while MCTS-transfer sequentially selects the next hyperparameter configuration to explore, the proposed method estimates the accuracy of all hyperparameter configurations at once. This efficiency is particularly important in transfer-based HPO scenarios, where computational resources on the target dataset are often limited.

Regarding the effect of the ability dimension, the 1D 2PLM consistently outperforms the 2D 2PLM in terms of AUC. This result suggests that, for the considered datasets and hyperparameter space, a single dominant latent ability primarily determines model performance, while the additional dimension may introduce noise.

VI. CONCLUSION

In this paper, we proposed an IRT-based hyperparameter transfer method that explicitly separates the latent ability of hyperparameter configurations from dataset-specific characteristics. By disentangling these two factors, the proposed approach enables hyperparameter transfer that is less dependent on the characteristics of the source datasets. Through an analysis of the estimated item parameters, we confirmed that the proposed method appropriately captures sample-level characteristics specific to each dataset. Furthermore, in the evaluation of transfer-based HPO, the proposed method achieved competitive or superior performance compared to an existing approach, while significantly reducing the computational cost on the target dataset.

As future work, we plan to extend the proposed method to hyperparameter spaces that include continuous values. While this study focused on a finite discrete search space, continuous hyperparameters are widely used in practical machine learning problems. Another important direction is the application to the Combined Algorithm Selection and Hyperparameter Optimization (CASH) problem. In settings where differences among algorithms are substantial, the advantages of multidimensional ability models may become more pronounced.

REFERENCES

[1] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated machine learning: Methods, systems, challenges*. Springer International Publishing Cham, 2019, pp. 3–33.

[2] A. Morales-Hernández, I. Van Nieuwenhuyse, and S. Rojas Gonzalez, "A survey on multi-objective hyperparameter optimization algorithms for machine learning," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8043–8093, 2023.

[3] F. Karl, T. Pielok, J. Moosbauer, F. Pfisterer, S. Coors, M. Binder, L. Schneider, J. Thomas, J. Richter, M. Lang *et al.*, "Multi-objective hyperparameter optimization in machine learning—an overview," *ACM Transactions on Evolutionary Learning and Optimization* (ACM-2023), vol. 3, no. 4, pp. 1–50, 2023.

[4] J. N. Van Rijn and F. Hutter, "Hyperparameter importance across datasets," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (ACM-2018), 2018, pp. 2367–2376.

[5] P. Probst, A.-L. Boulesteix, and B. Bischl, "Tunability: Importance of hyperparameters of machine learning algorithms," *Journal of Machine Learning Research* (JMLR-2019), vol. 20, no. 53, pp. 1–32, 2019.

[6] V. Perrone, R. Jenatton, M. W. Seeger, and C. Archambeau, "Scalable hyperparameter transfer learning," *Advances in neural information processing systems* (NeurIPS-2018), vol. 31, 2018.

[7] K. Swersky, J. Snoek, and R. P. Adams, "Multi-task bayesian optimization," *Advances in neural information processing systems* (NeurIPS-2013), vol. 26, 2013.

[8] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering* (IEEE-2009), vol. 22, no. 10, pp. 1345–1359, 2009.

[9] S. E. Embretson and S. P. Reise, *Item response theory for psychologists*. Psychology Press, 2013.

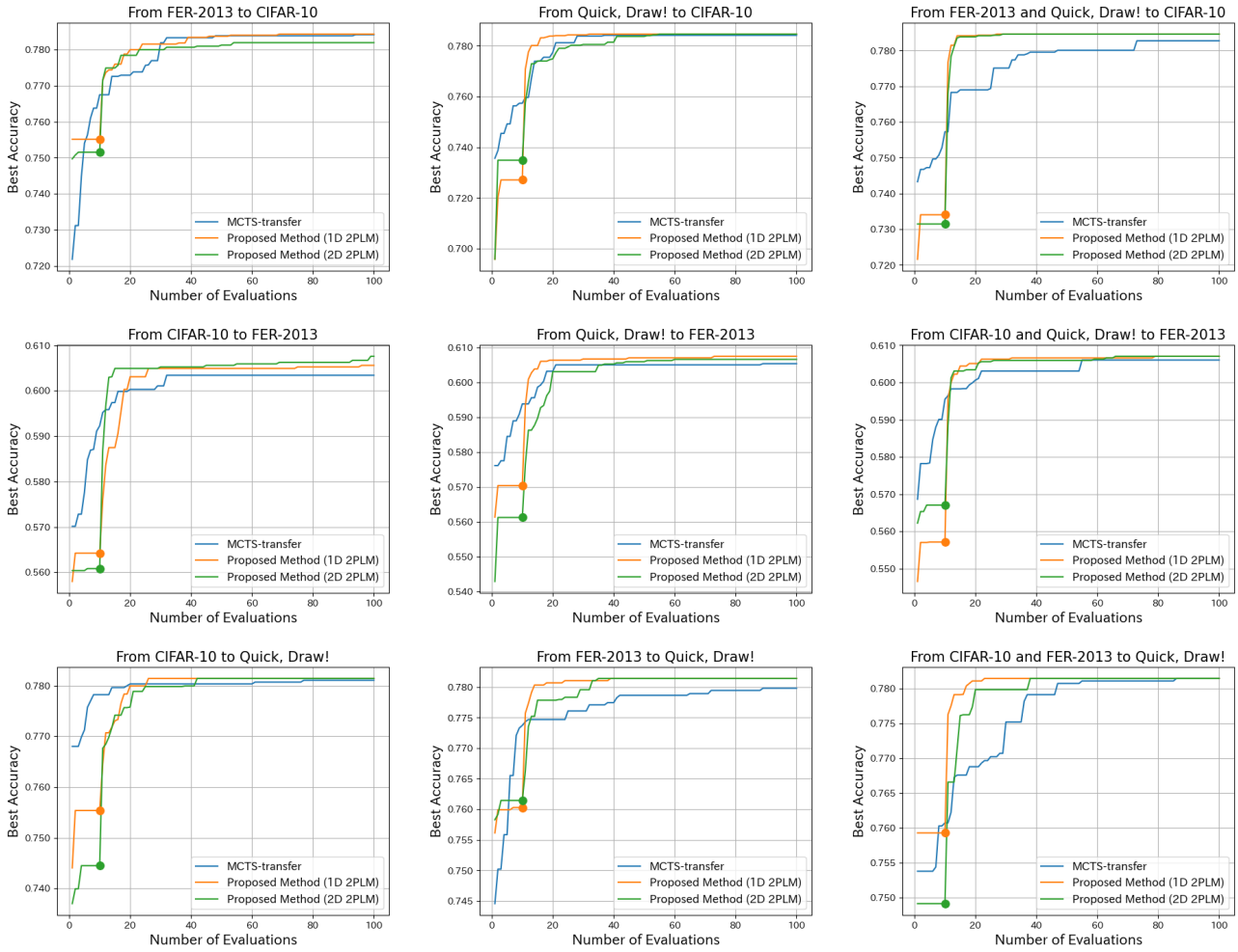

Fig. 4. Comparison of the progression of the maximum accuracy in transfer-based HPO between the baseline method (MCTS-transfer) and the proposed method (1D 2PLM and 2D 2PLM).

[10] M. Feurer, J. T. Springenberg, and F. Hutter, "Using meta-learning to initialize bayesian optimization of hyperparameters." in *MetaSel@ ECAI*, 2014, pp. 3–10.

[11] M. Feurer, J. Springenberg, and F. Hutter, "Initializing bayesian hyperparameter optimization via meta-learning," in *Proceedings of the AAAI conference on artificial intelligence* (AAAI-2015), vol. 29, no. 1, 2015.

[12] M. Wistuba, N. Schilling, and L. Schmidt-Thieme, "Learning hyperparameter optimization initializations," in *2015 IEEE international conference on data science and advanced analytics* (DSAA-2015). IEEE, 2015, pp. 1–10.

[13] D. Salinas, H. Shen, and V. Perrone, "A quantile-based approach for hyperparameter transfer learning," in *International conference on machine learning* (ICML-2020). PMLR, 2020, pp. 8438–8448.

[14] Y. Wei, P. Zhao, and J. Huang, "Meta-learning hyperparameter performance prediction with neural processes," in *International Conference on Machine Learning* (ICML-2021). PMLR, 2021, pp. 11 058–11 067.

[15] B.-J. Hsieh, P.-C. Hsieh, and X. Liu, "Reinforced few-shot acquisition function learning for bayesian optimization," *Advances in Neural Information Processing Systems* (NeurIPS-2021), vol. 34, pp. 7718–7731, 2021.

[16] S. Müller, M. Feurer, N. Hollmann, and F. Hutter, "Pfns4bo: In-context learning for bayesian optimization," in *International Conference on Machine Learning* (ICML-2023). PMLR, 2023, pp. 25 444–25 470.

[17] S. Wang, K. Xue, S. Lei, X. Huang, and C. Qian, "Monte carlo tree search based space transfer for black box optimization," *Advances in Neural Information Processing Systems* (NeurIPS-2024), vol. 37, pp. 49 591–49 624, 2024.

[18] Y. Chen, X. Li, J. Liu, and Z. Ying, "Item response theory—a statistical framework for educational and psychological measurement," *Statistical Science*, vol. 40, no. 2, pp. 167–194, 2025.

[19] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of item response theory*. Sage, 1991, vol. 2.

[20] M. D. Reckase, "Multidimensional item response theory," *Handbook of statistics*, vol. 26, pp. 607–642, 2006.

[21] Y. Chen, X. Li, and S. Zhang, "Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis," *Psychometrika*, vol. 84, no. 1, pp. 124–146, 2019.

[22] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[23] Y. H. Li and R. W. Lissitz, "An evaluation of the accuracy of multidimensional irt linking," *Applied Psychological Measurement*, vol. 24, no. 2, pp. 115–138, 2000.

[24] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[25] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International conference on neural information processing* (ICONIP-2013). Springer, 2013, pp. 117–124.

[26] D. Ha and D. Eck, "A neural representation of sketch drawings," *arXiv preprint arXiv:1704.03477*, 2017.