A COMPARATIVE STUDY OF MODEL ADAPTATION STRATEGIES FOR MULTI-TREATMENT UPLIFT MODELING

Ruyue Zhang[†]*, Xiaopeng Ke*[†]*, Ming Liu*[*]*, Fangzhou Shi, Chang Men, Zhengdan Zhu*

DiDi Chuxing, Beijing, China

ABSTRACT

Uplift modeling has emerged as a crucial technique for individualized treatment effect estimation, particularly in fields such as marketing and healthcare. Modeling uplift effects in multi-treatment scenarios plays a key role in real-world applications. Current techniques for modeling multi-treatment uplift are typically adapted from binary-treatment works. In this paper, we investigate and categorize all current model adaptations into two types: Structure Adaptation and Feature Adaptation. Through our empirical experiments, we find that these two adaptation types cannot maintain effectiveness under various data characteristics (noisy data, mixed with observational data, etc.). To enhance estimation ability and robustness, we propose Orthogonal Function Adaptation (OFA) based on the function approximation theorem. We conduct comprehensive experiments with multiple data characteristics to study the effectiveness and robustness of all model adaptation techniques. Our experimental results demonstrate that our proposed OFA can significantly improve uplift model performance compared to other vanilla adaptation methods and exhibits the highest robustness.

***Index Terms*—** causal inference, multi-treatment, uplift model

1. INTRODUCTION

Recently, Causal Inference has shown its effectiveness in many domains such as online marketing [1, 2, 3], advertising [4], and health care [5]. Among these, in online marketing, causal effect prediction is vital to subsidy distribution, as subsidies are widely adopted to stimulate user engagement and expand platform user bases.

A key component of causal inference modeling is the causal effect prediction problem with multiple discrete treatment candidates. Recent work on causal effect estimation has mainly focused on binary treatment versus control settings, and extensions are usually required for multi-treatment scenarios. Randomized Controlled Trials (RCTs) are widely regarded as the gold standard for causal inference, serving as the foundation for the development of numerous foundational estimation methods. Among them, representative approaches include tree-based methods [6], meta-learners [7, 8], balancing neural networks (BNN) [9], and TARNet [10]. Nonetheless, given the ethical and practical limitations of RCTs, a parallel line of work focuses on causal estimation methods that operate on observational data, such as CFRNet [10], DR-CFR [11], SITE [12], DragonNet [13], and HydraNet [14]. These works mitigate confounding bias through distribution balancing or representation learning. Overall, these methods have demonstrated strong performance in binary treatment settings.

However, how to extend binary treatment models to multi-treatment scenarios remains an open question—unfortunately, current research in this area lacks systematic and reliable investigation. In multi-treatment uplift modeling, binary models are often extended in different ways depending on how the treatment variable is incorporated into the model structure. Particularly for modeling with observational data, the adjustment for distributional differences also needs to be adapted to the multi-treatment setting. Methods such as tree-based models and BNN typically **treat the treatment as an input feature, which is a broadly applicable approach**. In contrast, for binary models with a two-head structure such as CFRNet, **a common practice is to extend the architecture into a multi-head form**. The MEMENTO [15] model exemplifies this approach, with the main difference being its specific method for balancing data distribution. Nonetheless, neither of these approaches consistently delivers stable performance across different datasets, underscoring the limitations of current extension strategies and calling for more efficient and flexible solutions for multi-treatment uplift modeling.

To figure out the robust modeling method for multi-treatment scenarios, we organize all modeling approaches for multi-treatment scenarios and classify them into two groups: **Structural Adaptation** and **Feature Adaptation**. Additionally, we introduce a novel multi-treatment modeling approach, **Orthogonal Function Adaptation**, which is designed as a plug-and-play module that enhances model generalization across diverse datasets. Its architecture consistently improves estimation performance under various data conditions with minimal additional computational overhead. Our experimental evaluation encompasses a diverse collection of datasets, including randomized controlled trials (RCTs), observational studies, synthetically generated data

[†] These authors contributed equally to this work.

[*] This work was conducted during his internship at DiDi Chuxing.

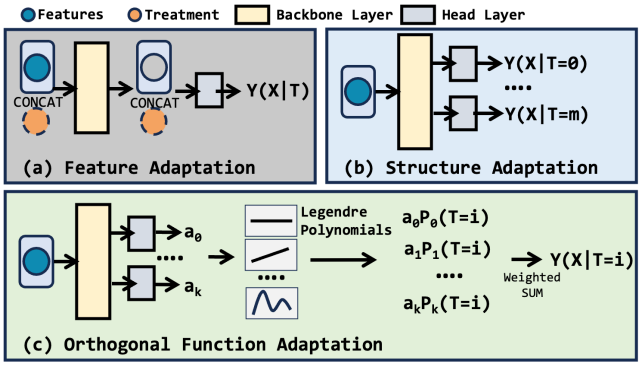

Fig. 1: Overview of different model adaptation categories

with noise, and data satisfying monotonicity constraints. Results demonstrate that the proposed OFA method consistently outperforms existing baselines across the majority of these settings.

Our main contributions are summarized as follows.

- To the best of our knowledge, we are the first to categorize current works on multi-treatment uplift modeling and conduct a comprehensive study on this scenario.
- We propose the Orthogonal Function Adaptation (OFA) technique for multi-treatment modeling based on the Weierstrass approximation theorem. This method is plug-and-play and can maintain computational complexity while increasing the treatment options.
- We conduct comprehensive experiments on synthetic and real-world datasets. The experimental results demonstrate that the proposed OFA outperforms other vanilla adaptations across various data characteristics.

2. METHODOLOGY

2.1. Notation

Let $\mathcal{D} = \{X_i, Y_i, T_i\}_{i=1}^N$ denotes the dataset where $X_i \in \mathbb{R}^d$ represents the d-dimensional input feature vector, Y_i denotes the label, and $T_i \in \mathcal{T}$ denotes the treatment. $\mathcal{T} = \{t_1, t_2, \cdots, t_m\}$ is the set of m possible treatment values. In this paper, we discuss the adaptation for multi-valued treatments under the Neyman-Rubin potential outcome framework [16]. We assume the uplift model has the following form:

$$f(X, T; \theta_f) = g(\varphi_X; T)$$

where φ_X denotes the hidden feature, and $g(\cdot, \cdot)$ denotes the associated effect estimator that operates on the hidden feature.

2.2. Feature Adaptation

Feature Adaptation technique is to concatenate the treatment value with the features as shown in Figure 1(a). The corresponding model can be written as

$$f_{FA}(X, T; \theta_{f_{FA}}) = g(\text{concat}(\varphi_X, T))$$

where $\text{concat}(\cdot, \cdot)$ denotes the concatenation function.

2.3. Structure Adaptation

To handle multi-valued treatments, another common way is to construct branches from the original model as shown in Figure 1(b). This can be represented as the set of effect estimators $\mathcal{G} = \{g_1, g_2, \ldots, g_m\}$. Consequently, the model becomes

$$f_{SA}(X, T; \theta_{f_{SA}}) = g_T(\varphi_X).$$

Note that the hidden feature φ_X could be the input feature $\varphi_X = X$, and the whole model degenerates into several separate submodels.

2.4. Orthogonal Function Adaptation

Through the Weierstrass approximation theorem [17], we can use a set of orthogonal polynomials to approximate every continuous function. Here, we propose the Orthogonal Function Adaptation (OFA) method as shown in Figure 1(c). We modify the model to generate coefficients for each polynomial function. Without loss of generality, we utilize orthogonal polynomials [18] to estimate the target function. The model can be written as

$$f_{OFA}(X, T; \theta_{OFA}) = \sum_{j=0}^{p} a_j(\varphi_X) \cdot P^{(j)}(T)$$

where p is the highest degree of the polynomial, $a_j(\cdot)$ denotes the coefficient estimator, and $P^{(j)}$ is the j-th orthogonal polynomial. In this paper, we select the Legendre orthogonal polynomial [19] to construct the approximation. These polynomials can be written as

$$P^{(0)}(t) = 1, P^{(1)}(t) = t,$$
$$(k+1)P^{(k+1)}(t) = (2k+1)tP^{(k)}(t) - kP^{(k-1)}(t).$$

Using the theorem from [20], we can also prove that the OFA achieves the lowest complexity while the error $\leq \varepsilon$:

$$C_{\text{OFA}} = \mathcal{O}(\varepsilon^{-\frac{m}{r}}) < C_{\text{SA}} = \mathcal{O}(\varepsilon^{-\frac{d}{r}}) < C_{\text{FA}} = \Omega(\varepsilon^{-\frac{m+d}{r}})$$

where C is the model complexity and $m < d$ must hold.

2.5. Loss Design

We select the Binary Cross-Entropy Loss as the basic loss. It could be written as:

$$\mathcal{L}_{\text{BCE}} = \frac{1}{N} \sum_{i=1}^{N} Y_i \log(f(X_i, T_i; \theta)) + (1 - Y_i) \log(1 - f(X_i, T_i; \theta))$$

As for handling the selection bias of the observation data, we apply the balanced representation technique in this paper. We select the Maximum Mean Discrepancy (MMD) [21] and Wasserstein discrepancy (WASS) [22] function to measure the distance between two feature distributions. This discrepancy loss can be written as

$$\mathcal{L}_{\text{disc}} = \text{disc}(\{h(X_i; T_i = t_p)\}_{i=1}^{N}, \{h(X_i; T_i = t_q)\}_{i=1}^{N})$$

where $t_p \neq t_q$ and $1 \leq p, q \leq m$. Thus, our final loss can be written as:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\text{BCE}} + \lambda_2 \cdot \mathcal{L}_{\text{disc}}$$

where $\lambda_1, \lambda_2 \in \mathbb{R}$ are the coefficients.

3. EXPERIMENTS

In this section, we conduct experiments to answer the following research questions.

- **RQ1: How does different adaptations perform on all synthetic and real-world datasets?**
- **RQ2: Is our OFA more robust than other adaptations?**
- **RQ3: Does the OFA maintain the improvement when applied to different models?**

3.1. Datasets and Implementation Details

Synthetic Dataset. We construct synthetic datasets with five treatment candidates. Each sample consists of eight covariates, one treatment variable, and one outcome, with 10,000 samples for training and 10,000 for testing; the test set is always a noise-free RCT. We consider three settings: (1) RCT data where treatments are randomized with small noise, including two variants: RCT-Noise, which injects additional noise into outcomes, and RCT-NM, which employs complex functional forms and removes monotonicity constraints on ITE curves; (2) observational data where treatment assignment depends on covariates, with or without an instrumental variable; and (3) mixed data that combines RCT and observational samples in a 1:1 ratio, also with or without an instrumental variable.

Real-World Dataset. We collect a real-world dataset named RW-RCT from a ride-hailing platform's coupon experiment. Each sample contains 47 covariates and five treatment candidates. The RW-RCT training set contains 975,940 samples; the corresponding test set consists of 811,027 samples.

Hyper-Parameter Setting. Across both synthetic and real-world datasets, the number of parameters in all models is kept comparable, with around 90k parameters for synthetic datasets and about 150k for real-world datasets due to their larger scale. All models underwent multiple rounds of hyperparameter search, and each selected configuration was evaluated through repeated independent runs to ensure that the results reliably reflect the models' performance. For all experiments, we used the Adam optimizer with a fixed learning rate of 1e-4.

Method	RCT	RCT-Noise	RCT-NM
Slearner + FA	<u>0.2854</u>	0.2552	0.2488
BNN + FA	**0.2855**	0.2663	0.2812
TARNet + SA	0.1977	0.2442	0.1948
DR-CFR + SA	0.2023	0.2455	0.2024
TARNet + OFA	0.2547	<u>0.2732</u>	**0.3160**
DR-CFR + OFA	0.2652	**0.2875**	<u>0.2882</u>

Table 1: mQini Score on three synthetic RCT datasets.

3.2. Evaluation Metrics

In this paper, we use the mean of the Qini score [3] to measure the performance of the uplift models. This metric can be written as

$$\begin{aligned}\text{mQini} &= \frac{1}{m}\sum_{i=1}^{m} \text{Qini}(f, t_i, \mathcal{D}) \\ &= \frac{1}{m}\sum_{i=1}^{m} \int \Big[Y_T(k) - \frac{N^T(k)}{N_C(k)} Y_C(k)\Big]_{f,t_i,\mathcal{D}} dk\end{aligned}$$

3.3. Model Performance (RQ1)

On synthetic datasets, OFA consistently shows stable and measurable improvements. Table 1 reports the results on RCT data. Under near-ideal low-noise conditions, FA models such as BNN+FA achieved the highest scores, since their flexibility is maximized in this setting. Nevertheless, even in such minimally confounded cases, OFA still achieved about a 30.0% improvement over SA. When noise was introduced, the advantage of OFA became more evident, yielding an 8.0% improvement over the best non-OFA model, BNN+FA. On RCT-NM data, where monotonicity constraints were removed and response curves exhibited more complex functional forms, OFA maintained its superiority, achieving a 12.4% gain over the best non-OFA model. Furthermore, Tables 2 and 3 present the results on observational and mixed datasets. Regardless of whether instrumental variables were included, OFA-based models consistently ranked at the top. On purely observational data with instrumental variables, the best OFA model improved by 42.1% compared with the best non-OFA model, and by 41.2% without instrumental variables. On mixed datasets, the improvements were 13.7% and 14.1% with and without instrumental variables, respectively.

Table 4 further demonstrates the results on real-world RCT data. TARNet enhanced with OFA achieved the best

Method	OBS w/ IV	OBS w/o IV
BNN + FA + WASS	0.1249	0.1014
BNN + FA + MMD	0.0860	0.0791
CFRNet + SA + WASS	0.1997	0.1843
CFRNet + SA + MMD	0.1795	0.1718
DR-CFR + SA + WASS	0.1747	0.2017
DR-CFR + SA + MMD	0.1721	0.1740
CFRNet + OFA	0.2778	0.2662
CFRNet + OFA + WASS	<u>0.2796</u>	0.2818
CFRNet + OFA + MMD	0.2789	<u>0.2826</u>
DR-CFR + OFA + WASS	0.2652	**0.2848**
DR-CFR + OFA + MMD	**0.2837**	0.2797

Table 2: mQini Score on synthetic observation datasets (with or without instrument variable).

Method	MIX w/ IV	MIX w/o IV
BNN + FA + WASS	0.2345	0.2510
BNN + FA + MMD	0.2460	0.2357
CFRNet + SA + WASS	0.1892	0.2363
CFRNet + SA + MMD	0.2007	0.1971
DR-CFR + SA + WASS	0.1863	0.2019
DR-CFR + SA + MMD	0.2025	0.1714
CFRNet + OFA	0.2522	0.2595
CFRNet + OFA + WASS	0.2684	0.2744
CFRNet + OFA + MMD	0.2674	**0.2864**
DR-CFR + OFA + WASS	<u>0.2730</u>	0.2769
DR-CFR + OFA + MMD	**0.2798**	<u>0.2776</u>

Table 3: mQini Score on synthetic mixed datasets (with or without instrument variable).

performance, with a 6.6% improvement over the best non-OFA model, providing strong evidence of its effectiveness. This result indicates that the robustness demonstrated by OFA on synthetic datasets can directly transfer to real-world applications, leading to superior practical outcomes.

The results demonstrate that although FA can reach peak performance under ideal, low-noise environments, OFA provides substantial, reliable improvements over both SA and FA across all other scenarios, with its advantage particularly pronounced in noisy and complex settings. Furthermore, this robustness makes OFA the best-performing method overall on real-world business RCTs. Across the entire spectrum of datasets, OFA establishes itself as the most reliable and consistently high-performing approach, delivering significant performance gains in the most challenging scenarios.

3.4. Robustness Analysis (RQ2)

As shown in Table 1, under noisy RCT conditions, OFA provides inherent regularization through its orthogonal basis functions, effectively resisting perturbations and yielding an 8.0% improvement over the best non-OFA method. When the response curve follows complex, non-monotonic patterns, OFA exhibits strong adaptability, as evidenced by a 62.2% improvement of TARNet + OFA over its SA counterpart. In observational scenarios (Table 2 to 3), OFA combined with representation learning offers a stable prediction layer that enhances generalization from biased samples, resulting in more reliable bias correction.

Method	RW-RCT
Slearner + FA	0.1082
BNN + FA	0.0880
TARNet + SA	0.0765
DR-CFR + SA	0.0896
TARNet + OFA	**0.1153**
DR-CFR + OFA	<u>0.1100</u>

Table 4: mQini Score on the real-world RCT dataset.

Based on comprehensive experimental results, our proposed OFA method demonstrates significantly superior robustness compared to other adaptation approaches, with quantified advantages across three key dimensions: noise resilience, functional flexibility, and bias correction stability.

3.5. Transferability (RQ3)

As shown in Tables 1 to 4, our proposed OFA consistently improves performance across different model architectures.

When applied to TARNet, OFA enhanced its performance across all challenging scenarios. Most notably, on the RCT-NM dataset, TARNet + OFA achieved a massive 62.2% improvement over the vanilla TARNet + SA. Similarly, when plugged into DR-CFR, the OFA variant also became the top performer on the noisy RCT dataset, showcasing a 17.1% improvement over its standard counterpart.

In conclusion, OFA is not only effective for a single model but acts as a versatile and plug-and-play performance enhancer. Its consistent ability to elevate the performance of diverse underlying architectures confirms its strong generalizability and establishes it as a universally valuable component for multi-treatment causal effect estimation.

4. CONCLUSION

In this paper, we categorize and study different model adaptations for multi-treatment scenarios. We further propose the Orthogonal Function Adaptation (OFA) method, which has superior robustness and generalizability across diverse datasets and baselines. It significantly enhances estimation accuracy under noise, selection bias, and complex functional relationships, while presenting stable performance gains in both synthetic and real-world datasets.

5. REFERENCES

[1] Zexu Sun, Bowei He, Ming Ma, Jiakai Tang, Yuchen Wang, Chen Ma, and Dugang Liu, "Robustness-enhanced uplift modeling with adversarial feature desensitization," in *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2023, pp. 1325–1330.

[2] Dugang Liu, Xing Tang, Han Gao, Fuyuan Lyu, and Xiuqiang He, "Explicit feature interaction-aware uplift network for online marketing," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 4507–4515.

[3] Mouloud Belbahri, Alejandro Murua, Olivier Gandouet, and Vahid Partovi Nia, "Qini-based uplift regression," *The Annals of Applied Statistics*, vol. 15, no. 3, pp. 1247–1272, 2021.

[4] Shogo Kawanaka and Daisuke Moriwaki, "Uplift modeling for location-based online advertising," in *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based recommendations, geosocial networks and geoadvertising*, 2019, pp. 1–4.

[5] Ruoqi Liu, Lingfei Wu, and Ping Zhang, "Kg-treat: pre-training for treatment effect estimation by synergizing patient data with knowledge graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 8805–8814.

[6] Stefan Wager and Susan Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[7] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the national academy of sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[8] Xinkun Nie and Stefan Wager, "Quasi-oracle estimation of heterogeneous treatment effects," *Biometrika*, vol. 108, no. 2, pp. 299–319, 2021.

[9] Fredrik Johansson, Uri Shalit, and David Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*. PMLR, 2016, pp. 3020–3029.

[10] Uri Shalit, Fredrik D Johansson, and David Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International conference on machine learning*. PMLR, 2017, pp. 3076–3085.

[11] Negar Hassanpour and Russell Greiner, "Learning disentangled representations for counterfactual regression," in *International Conference on Learning Representations*, 2019.

[12] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang, "Representation learning for treatment effect estimation from observational data," *Advances in neural information processing systems*, vol. 31, 2018.

[13] Claudia Shi, David Blei, and Victor Veitch, "Adapting neural networks for the estimation of treatment effects," *Advances in neural information processing systems*, vol. 32, 2019.

[14] Borja Velasco-Regulez and Jesus Cerquides, "Hydranet: A neural network for the estimation of multi-valued treatment effects," in *Artificial Intelligence Research and Development*, pp. 16–27. IOS Press, 2023.

[15] Abhirup Mondal, Anirban Majumder, and Vineet Chaoji, "Memento: Neural model for estimating individual treatment effects for multiple treatments," in *Proceedings of the 31st ACM international conference on information & knowledge management*, 2022, pp. 3381–3390.

[16] Donald B Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[17] Marshall H Stone, "The generalized weierstrass approximation theorem," *Mathematics Magazine*, vol. 21, no. 5, pp. 237–254, 1948.

[18] Gabor Szeg, *Orthogonal polynomials*, vol. 23, American Mathematical Soc., 1939.

[19] Eric W Weisstein, "Legendre polynomial," *https://mathworld. wolfram. com/*, 2002.

[20] Tomer Galanti and Lior Wolf, "On the modularity of hypernetworks," *Advances in neural information processing systems*, vol. 33, pp. 10409–10419, 2020.

[21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *The journal of machine learning research*, vol. 13, no. 1, pp. 723–773, 2012.

[22] SS Vallender, "Calculation of the wasserstein distance between probability distributions on the line," *Theory of Probability & Its Applications*, vol. 18, no. 4, pp. 784–786, 1974.