Nonparametric Iterative Machine Teaching

Chen Zhang [1] **Xiaofeng Cao** [1] **Weiyang Liu** [2 3] **Ivor W. Tsang** [4] **James T. Kwok** [5]

Abstract

In this paper, we consider the problem of Iterative Machine Teaching (IMT), where the teacher provides examples to the learner iteratively such that the learner can achieve fast convergence to a target model. However, existing IMT algorithms are solely based on parameterized families of target models. They mainly focus on convergence in the parameter space, resulting in difficulty when the target models are defined to be functions without dependency on parameters. To address such a limitation, we study a more general task – Nonparametric Iterative Machine Teaching (NIMT), which aims to teach nonparametric target models to learners in an iterative fashion. Unlike parametric IMT that merely operates in the parameter space, we cast NIMT as a functional optimization problem in the function space. To solve it, we propose both random and greedy functional teaching algorithms. We obtain the iterative teaching dimension (ITD) of the random teaching algorithm under proper assumptions, which serves as a uniform upper bound of ITD in NIMT. Further, the greedy teaching algorithm has a significantly lower ITD, which reaches a tighter upper bound of ITD in NIMT. Finally, we verify the correctness of our theoretical findings with extensive experiments in nonparametric scenarios.

1. Introduction

Machine teaching (MT) (Zhu, 2015; Zhu et al., 2018) is the study of how to design the optimal teaching set, typically with minimal examples, so that learners can quickly learn target models based on these examples. It can be considered an inverse problem of machine learning, where machine learning aims to learn model parameters from a dataset, while MT aims to find a minimal dataset from the target model parameters. MT has proven to be useful in various domains, including robustness (Alfeld et al., 2016; 2017; Ma et al., 2019; Rakhsha et al., 2020), crowd sourcing (Singla et al., 2013; 2014; Zhou et al., 2018; 2020; Collins et al., 2023), and computer vision (Wang et al., 2021; Wang & Vasconcelos, 2021).

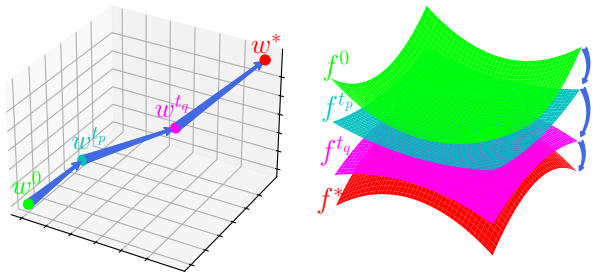
Considering the interaction manner between teachers and learners, MT can be conducted in either batch (Zhu, 2013; 2015; Liu et al., 2016; Mansouri et al., 2019) or iterative (Liu et al., 2017; 2018) fashion. Batch MT only allows the teacher to interact with the learner once. The teacher constructs a teaching dataset and feeds it to the learner in one shot. The learner will learn a target model from this dataset. The minimal number of examples in this teaching set is called *teaching dimension* (Goldman & Kearns, 1995). In contrast, an iterative teacher would feed examples sequentially based on current status of the iterative learner, which further takes the optimization algorithm into consideration. The number of iterations, *i.e.*, the length of this teaching sequence is defined as *iterative teaching dimension* (ITD) (Liu et al., 2017; 2018).

The majority of current research on iterative machine teaching (IMT) (Liu et al., 2017; 2018; Xu et al., 2021; Wang et al., 2021) focuses on the convergence to target models (*i.e.*, functions) f which are usually parameterized by a set of parameters $\boldsymbol{w}$ as it assumes that f can be represented by $\boldsymbol{w}$, *e.g.*, $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ with input $\boldsymbol{x}$. However, there may exist cases where the mapping from input to output cannot be parameterized in terms of $\boldsymbol{w}$, for example, f is defined in a nonparametric fashion (Hollander et al., 2013; Corder & Foreman, 2014; Zhu et al., 2018). Especially in general and more realistic problems (*e.g.*, (Genevay et al., 2016; Blei et al., 2017; Dvurechenskii et al., 2018)), the assumption of parametric learners may not hold. Here comes a natural question: *Can the teacher efficiently guide iterative learners to parameter-free target models?* Our answer is *Yes*. We seek to guide iterative learners to achieve fast convergence

[1]School of Artificial Intelligence, Jilin University, China [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany [3]University of Cambridge, United Kingdom [4]Centre for Frontier AI Research and Institute of High Performance Computing, A*STAR, Singapore [5]Hong Kong University of Science and Technology. Correspondence to: Xiaofeng Cao <xiaofengcao@jlu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Our source code is available at https://github.com/chen2hang/NonparametricTeaching.



(a) Parametric IMT (b) Nonparametric IMT

Figure 1. Comparison between parametric and nonparametric IMT in 3D space. (a): Parameters are precisely vectors represented by a point in 3D space, which would be updated gradually towards w^*. (b): Nonparametric model f can be denoted by a surface in 3D, which would evolve in more complicated fashion.

to a nonparametric target function f^*. Figure 1 provides an intuitive comparison between parametric and nonparametric iterative teaching in a 3-dimensional space.

Shifting our focus to functions, we formulate NIMT as an instance of functional optimization problem (Singer, 1974; Zoppoli et al., 2002; Mroueh et al., 2019; Shen et al., 2020), and then derive two algorithms (one picks examples randomly, and the other picks examples in an greedy fashion). Without loss of generality, we are mainly concerned with the Reproducing Kernel Hilbert Space (RKHS) in this paper. We start with a simple baseline algorithm, called **R**andom **F**unctional **T**eaching (RFT), which essentially adopts uniform sampling and serves as a functional analogue of stochastic gradient descent (Ruder, 2016; Hardt et al., 2016). In the context of IMT, we analyze the functional gradient descent method (Mason et al., 1999a; Shen et al., 2020) in RKHS, and then find that based on the chain rule for functional gradients (Gelfand et al., 2000; Coleman, 2012), the gradient in NIMT can be expressed by the multiplication of a scalar governing the magnitude and the kernel function with the teaching example as its argument. Therefore, steepening gradients is equivalent to maximizing that scalar, which naturally leads to our greedy algorithm – **G**reedy **FT** (GFT). GFT picks examples evaluated at the point where the target and current models reach their maximal difference (Arbel et al., 2019; Cormen et al., 2022). Furthermore, under mild assumptions, we theoretically prove the convergence of both RFT and GFT, and then show that the ITD of GFT is lower than that of RFT. This concludes that GFT yields a tighter upper bound for ITD. Finally, we validate our theoretical findings with a number of experiments in both synthetic and real-world datasets under nonparametric scenarios. To summarize, the contributions of our work are listed as follows.

- To our knowledge, we are the first to comprehensively study Nonparametric Iterative Machine Teaching (NIMT), which focuses on exploring iterative algorithms for teaching parameter-free target models from the optimization perspective. Instead of operating in the finite-dimensional space of parameters, we formulate NIMT as a functional optimization in the space of infinite-dimensional functions, a more general space of models (*i.e.*, RKHS is considered), in Section 4.1. NIMT is a natural generation of IMT (Liu et al., 2017), shifting the parametric paradigm to a nonparametric one.
- We propose two teaching algorithms (RFT and GFT). RFT is based on random sampling with ground truth labels, and the derivation of GFT is based on the maximization of the informative scalar introduced in Proposition 5 in order to steepen gradients. These two teaching algorithms proposed in Section 4.2 fill the gap for teaching nonparametric learners in IMT.
- We theoretically analyze the asymptotic behavior of both RFT and GFT in Section 4.3. We prove that per-iteration reduction of loss $\mathcal{L}$ for RFT and GFT has a negative upper bound expressed by the discrepancy of iterative teaching defined in Definition 10, and we derive that the ITD of GFT is $\mathcal{O}(\psi(\frac{2\mathcal{L}(f^0)}{\tilde{\eta}\epsilon}))$ (detailed notations are introduced in the subsequent sections), which is shown to be lower than the ITD of RFT, $\mathcal{O}(2\mathcal{L}(f^0)/(\tilde{\eta}\epsilon))$.

2. Related Work

Machine teaching. There has been a recent growth of interest in the research of machine teaching (Zhu, 2015; Zhu et al., 2018; Liu et al., 2017; 2018; Wang et al., 2021). Batch machine teaching studies behaviors of version space learners (Chen et al., 2018; Tabibian et al., 2019), linear learners (Liu et al., 2016), reinforcement learners (Kamalaruban et al., 2019; Zhang et al., 2020b) along with forgetful learners (Hunziker et al., 2018; Liu et al., 2018) and multiple learners (Zhu et al., 2017). Further, taking the learner's optimization algorithm into consideration, iterative teaching has been recently studied (Liu et al., 2017; 2018; Peltola et al., 2019; Lessard et al., 2019; Liu et al., 2021; Xu et al., 2021; Qiu et al., 2022). (Liu et al., 2021) considers a label synthesis teacher and (Qiu et al., 2022) proposes a generative teacher. (Xu et al., 2021) improves the scalability and efficiency of the iterative teaching algorithm with locality-sensitive sampling. Different from existing works that focus on parametric learners, we aim to teach a nonparametric learner. In this regime, One of the most related work is (Mansouri et al., 2019) which analyzes sequential teaching from the perspective of hypothesis pruning without specifying a parameter for hypothesis. In contrast, this work systematically investigates nonparametric teaching from the optimization perspective. Besides, (Kumar et al., 2021; Qian et al., 2022) are also highly related, since they study non-gradient-based kernel learners under the batch setting. However, they are not strictly nonparametric teaching since

they assume the hypothesis is determined by parameters, and they cannot produce an iterative algorithm for teaching parameter-free mappings. In contrast, we study a more general task – nonparametric iterative machine teaching, and propose practical iterative functional teaching algorithms.

Functional optimization. Allowing non-parametrically defined mapping from input to output, functional optimization (Singer, 1974; Becke, 1988; Singer, 1974; Friedman, 2001; Zoppoli et al., 2002; Smanski et al., 2014; Zhang et al., 2020a) over more general space of functions, including RKHS, Sobolev space (Adams & Fournier, 2003) and Fréchet space (Narici & Beckenstein, 2010), is a foundational and meaningful task across many domains, such as barycenter problem (Shen et al., 2020; Ye et al., 2017), variational inference (Liu & Wang, 2016; Liu, 2017) and GAN training (Mroueh et al., 2019). (Nitanda & Suzuki, 2018; 2020) make an interesting connection between functional gradient boosting and residual networks (He et al., 2016). We observe that the functional gradient descent algorithm (Mason et al., 1999a;b; Coleman, 2012) for functional optimization in RKHS is well studied because of some regular properties. The iterative interaction (Liu et al., 2017; 2018) between teachers and learners exhibits intriguing similarities to the functional gradient descent algorithm in terms of its gradual improvement. Inspired by such similarities, NIMT starts by analyzing functional gradient descent and then designs algorithms for choosing optimal teaching examples under the iterative teaching framework (Liu et al., 2017; 2018; 2021; Qiu et al., 2022).

3. Notations

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a n dimensional feature space and $\mathcal{Y} \subseteq \mathbb{R}$(Regression) or $\mathcal{Y} = \{-1, 1\}$ (Classification) be a label space. A teaching example refers to a pair of data (*e.g.*, image) and label $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$. A length-$T$ teaching sequence is defined as $\mathcal{D} = \{(\boldsymbol{x}^1, y^1), \dots (\boldsymbol{x}^T, y^T)\} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^T$. The collection of potential teaching sequences is denoted by $\mathbb{D}$ which includes all teaching sequences, *i.e.*, $\mathcal{D} \in \mathbb{D}$ and is also called the knowledge domain of teachers (Liu et al., 2017; 2018).

This paper considers a specific function space – the Reproducing Kernel Hilbert Space, and therefore models are assumed to be mappings in RKHS $f \in \mathcal{H} : \mathcal{X} \mapsto \mathcal{Y}$. This assumption is widely adopted in general functional optimization, *e.g.*, (Liu & Wang, 2016; Mroueh et al., 2019; Arbel et al., 2019; Shen et al., 2020). Operating under RKHS where point evaluation is a continuous linear functional allows us to quantify the iteration quality, which is crucial for convergence analysis. Given a target model[1] $f^* \in \mathcal{H}$, one can uniquely represent a teaching example $(\boldsymbol{x}^\dagger, y^\dagger)$ by its feature $\boldsymbol{x}^\dagger$ for brevity since its label is precisely $y^\dagger = f^*(\boldsymbol{x}^\dagger)$.

Let $K(\boldsymbol{x}, \boldsymbol{x}') : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a positive definite kernel function. Equivalently, $K(\boldsymbol{x}, \boldsymbol{x}') = K_{\boldsymbol{x}}(\boldsymbol{x}') = K_{\boldsymbol{x}'}(\boldsymbol{x})$ and $K_{\boldsymbol{x}}(\cdot)$ can be abbreviated as $K_{\boldsymbol{x}}$. The RKHS $\mathcal{H}$ determined by $K(\boldsymbol{x}, \boldsymbol{x}')$ is the closure of linear span $\{f : f(\cdot) = \sum_{i=1}^r \alpha_i K(\boldsymbol{x}_i, \cdot), \alpha_i \in \mathbb{R}, r \in \mathbb{N}, \boldsymbol{x}_i \in \mathcal{X}\}$ equipped with inner product $\langle f, g\rangle_{\mathcal{H}} = \sum_{ij} \alpha_i \beta_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ when $g = \sum_j \beta_j K_{\boldsymbol{x}_j}$. NIMT reduces to parameterized IMT if we use a linear kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}'\rangle + 1$ (Hofmann et al., 2008). With the Riesz–Fréchet representation theorem (Lax, 2002; Schölkopf et al., 2002), the evaluation functional is defined as follows:

Definition 1. *For a reproducing kernel Hilbert space $\mathcal{H}$ with a positive definite kernel $K_{\boldsymbol{x}} \in \mathcal{H}$, we define the evaluation functional $E_{\boldsymbol{x}}[\cdot] : \mathcal{H} \mapsto \mathbb{R}$ as*

$$E_{\boldsymbol{x}}[f] = \langle f, K_{\boldsymbol{x}}(\cdot)\rangle_{\mathcal{H}} = f(\boldsymbol{x}), f \in \mathcal{H}. \tag{1}$$

Additionally, for a functional $F : \mathcal{H} \mapsto \mathbb{R}$, the Fréchet derivative (Coleman, 2012; Liu, 2017; Shen et al., 2020) of F is given as follows:

Definition 2. *(Fréchet derivative in RKHS) For a functional $F : \mathcal{H} \mapsto \mathbb{R}$, its Fréchet derivative $\nabla_f F[f]$ at $f \in \mathcal{H}$ is defined implicitly as $F[f + \epsilon g] = F[f] + \epsilon\langle \nabla_f F[f], g\rangle_{\mathcal{H}} + \mathcal{O}(\epsilon^2)$ for any $g \in \mathcal{H}$ and $\epsilon \in \mathbb{R}$, which is a function in $\mathcal{H}$.*

4. Nonparametric Iterative Machine Teaching

We start by formulating NIMT as a nested functional minimization (Eq. 2). Then we present a natural baseline called random functional teaching, which samples data randomly (Algorithm 1). After gaining an insight from functional gradient (Proposition 5), we propose the greedy teaching algorithm, called greedy functional teaching, which searches examples with steeper gradients (Algorithm 1). Finally, we analyze the ITD for both RFT and GFT.

4.1. Teaching settings

Different from the parametric cases (Liu et al., 2017; Zhu et al., 2018) reviewed in Appendix A, we define NIMT as a functional minimization over $\mathbb{D}$ in RKHS:

$$\mathcal{D}^* = \underset{\mathcal{D} \in \mathbb{D}}{\arg\min} \quad \mathcal{M}(\hat{f}, f^*) + \lambda \cdot \text{len}(\mathcal{D}) \quad \text{s.t.} \quad \hat{f} = \mathcal{A}(\mathcal{D}), \tag{2}$$

where $\mathcal{M}$ denotes a discrepancy measure, len($\mathcal{D}$), which is regularized by a constant λ, is the length of the teaching sequence $\mathcal{D}$, and $\mathcal{A}$ represents the learning algorithm of

[1] We assume that both f^0 and f^* are from the same RHKS such that f^* is realizable. Generally, f^* can be assigned arbitrarily, but for the convergence to the target model, we consider the projection of f^* into the RHKS constructed by $\mathcal{X}$ with specific kernels.

learners. In fact, $\text{len}(\mathcal{D})$ essentially is the count of iterations, *i.e.*, the ITD defined in (Liu et al., 2017). Specifically, we are concerned with L_2 norm defined in RKHS as the discrepancy measure $\mathcal{M}(\hat{f}, f^*) = \|\hat{f} - f^*\|_{\mathcal{H}}$, and empirical risk minimization as the learning algorithm $\mathcal{A}(\mathcal{D})$ as follows:

$$\hat{f}^* = \arg\min_{f \in \mathcal{H}} \mathbb{E}_{(\boldsymbol{x},y)} \left\{ \mathcal{L}(f(\boldsymbol{x}), y) \right\}, \tag{3}$$

where we have the joint sampling distribution $(\boldsymbol{x}, y) \sim \mathbb{P}(\boldsymbol{x}, y)$ and the convex loss function $\mathcal{L}$. It is optimized by functional gradient descent:

$$f^{t+1} \leftarrow f^t - \eta^t \mathcal{G}(\mathcal{L}; f^t; (\boldsymbol{x}^t, y^t)), \tag{4}$$

where $t = 0, 1, \ldots, \text{len}(\mathcal{D})$ is the iteration index, $\eta^t > 0$ is the learning rate at t-th iteration (a small constant) and $\mathcal{G}$ denotes the gradient functional evaluated at $(\boldsymbol{x}^t, y^t)$.

Compared to the white-box setting where teachers know all information about learners (Liu et al., 2017; 2021; Xu et al., 2021), this paper considers a more practical gray-box teaching setting, where teachers have no access to the learning rate η, specific loss function $\mathcal{L}$ but are able to track f^t. For interaction, we only allow teachers to communicate with learners via teaching examples in $\mathcal{D}$. For teachers with different knowledge domains, we start by deriving the theoretical findings for synthesis-based teachers (Liu et al., 2017), and then extend them to the most practical pool-based teachers discussed in Remark 7. Finally, we study the empirical performance of our method.

4.2. Functional teaching algorithms

Random Functional Teaching. It is straightforward for teachers to pick examples randomly and feed them to learners, which derives a simple teaching baseline called Random Functional Teaching. Given a nonparametric target model f^*, RFT algorithm is to give learners $\mathcal{D} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^{\text{ITD}_{\text{RFT}}}$ where $\boldsymbol{x}^i \in \mathcal{X}$ is picked randomly, $y^i = f^*(\boldsymbol{x}^i)$ and ITD_{RFT} denotes the ITD of RFT. RFT forms a functional counterpart of SGD (Ruder, 2016; Hardt et al., 2016), and RFT provides ground truth $y^i = f^*(\boldsymbol{x}^i)$ as f^* is known. Therefore, it is natural to consider RFT as a very fundamental baseline when comparing against other functional teaching algorithms. Pseudo code is in Algorithm 1.

Greedy Functional Teaching. With Fréchet derivative in RKHS (Definition 2), we introduce Chain Rule for functional gradients (Gelfand et al., 2000) as a Lemma.

Lemma 3. *(Chain rule for functional gradients) For differentiable functions $G : \mathbb{R} \mapsto \mathbb{R}$ that are functions of functionals F, $G(F[f])$, the expression*

$$\nabla_f G(F[f]) = \frac{\partial G(F[f])}{\partial F[f]} \cdot \nabla_f F[f] \tag{5}$$

is usually referred to as the chain rule.

For derivative of evaluation functional (Coleman, 2012), we provide Lemma 4 whose proof is deferred to Appendix B.

Lemma 4. *For an evaluation functional $E_{\boldsymbol{x}}[f] = f(\boldsymbol{x}) : \mathcal{H} \mapsto \mathbb{R}$, its gradient is $\nabla_f E_{\boldsymbol{x}}[f] = K_{\boldsymbol{x}}$.*

f can be viewed as the argument and the loss function $\mathcal{L}$ of interest in NIMT is precisely a functional. Consequently, with Lemma 3 and 4, we gain a critical insight of functional gradients of $\mathcal{L}$ (Mason et al., 1999a; Coleman, 2012).

Proposition 5. *Given a certain example $(\boldsymbol{x}, y)$, the gradient $\mathcal{G}$ of loss function $\mathcal{L}$ w.r.t. the model f can be expressed as a scalar times a unit kernel:*

$$\mathcal{G}(\mathcal{L}; f; (\boldsymbol{x}, y)) = \left.\frac{\partial \mathcal{L}}{\partial f}\right|_{f(\boldsymbol{x}),y} \|K_{\boldsymbol{x}}\|_{\mathcal{H}} \cdot \frac{K_{\boldsymbol{x}}}{\|K_{\boldsymbol{x}}\|_{\mathcal{H}}}. \tag{6}$$

Proposition 5 suggests that the functional gradient is fundamentally determined by an informative real number $\partial \mathcal{L}/\partial f|_{f(\boldsymbol{x}),y} \|K_{\boldsymbol{x}}\|_{\mathcal{H}}$ controlling the magnitude of $\mathcal{G}$ and a unit kernel $K_{\boldsymbol{x}}/\|K_{\boldsymbol{x}}\|_{\mathcal{H}}$ governing the direction (Coleman, 2012). For ease of understanding, such a unit kernel can be viewed as a unit vector in infinite dimensional space (a counterpart of a unit vector in the Euclidean space) since a model can be represented by a infinite series of functions in RKHS, $f = \sum_i^{\infty} \alpha_i K_{\boldsymbol{x}_i}$ (Steinwart & Christmann, 2008).

In IMT (Liu et al., 2017), the target is to achieve fast convergence (maximal reduction of iteration number) by designing the optimal iterative algorithms for example selection. It is natural to consider the properties of the optimal example for reducing ITD at each iteration. This is answered by Theorem 6 proved in Appendix B.

Theorem 6. *Given a nonparametric target model f^*, let $(\boldsymbol{x}^t, y^t)$ be a fed example at t-th iteration and $(\boldsymbol{x}^{t^*}, y^{t^*})$ be the optimal one with the steepest gradient towards f^*:*

$$\left(\boldsymbol{x}^{t^*}, y^{t^*}\right) = \arg\min_{\boldsymbol{x}^t \in \mathcal{X}, y^t \in \mathcal{Y}} \left\| f^t - \eta^t \mathcal{G}(\mathcal{L}; f^t; (\boldsymbol{x}^t, y^t)) - f^* \right\|_{\mathcal{H}}^2. \tag{7}$$

We denote $\mathcal{G}^t := \mathcal{G}(\mathcal{L}; f^t; (\boldsymbol{x}^t, y^t))$ and $\mathcal{G}^{t^*} := \mathcal{G}(\mathcal{L}; f^t; (\boldsymbol{x}^{t^*}, y^{t^*}))$, and then the following holds

$$\langle \mathcal{G}^{t^*} - \mathcal{G}^t, f^t - f^* \rangle_{\mathcal{H}} \geq 0. \tag{8}$$

Eq. 8 indicates a property of $\mathcal{G}^{t^*}$ corresponding to $\boldsymbol{x}^{t^*}$, which is independent of explicit η and specific $\mathcal{L}$ and adapts to gray-box learners. Besides, Theorem 6 intuitively tells that $\mathcal{G}^{t^*} - \mathcal{G}^t$ and $f^t - f^*$ share the same direction. That means if $f^t \geq f^*$, the example with the largest gradient $\mathcal{G}^{t^*} \geq \mathcal{G}^t \geq 0$ would be selected as the optimal example to minimize Eq. 7. For the case of $f^t \leq f^*$, the gradient of the optimal example should be the smallest one, *i.e.*,

$\mathcal{G}^{t^*} \leq \mathcal{G}^t \leq 0$. In a nutshell, the gradient norm at the optimal example should be maximal at every iteration.

Combining Proposition 5 and results in Theorem 6, maximizing gradient norm written in Eq. 6 derives our greedy functional teaching algorithm, namely Greedy-1 Functional Teaching (GFT-1):

Given a nonparametric target model f^*, GFT-1 is to pick the example satisfying

$$\left(\boldsymbol{x}^{t^*} = \arg\max_{\boldsymbol{x}^t \in \mathcal{X}} \left\| \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}^t), y^t = f^*(\boldsymbol{x}^t)} K\left(\boldsymbol{x}^t, \cdot\right) \right\|_{\mathcal{H}}, y^{t^*} = f^*\left(\boldsymbol{x}^{t^*}\right)\right) \quad (9)$$

as the optimal one to learners at t-th iteration, and $t = 0, 1, \ldots, \text{ITD}_{\text{GFT}}$ where ITD_{GFT} is the ITD of GFT.

Practically, we can simplify it as

$$\left(\boldsymbol{x}^{t^*} = \arg\max_{\boldsymbol{x}^t \in \mathcal{X}} \left| \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}^t), y^t = f^*(\boldsymbol{x}^t)} \right|, y^{t^*} = f^*\left(\boldsymbol{x}^{t^*}\right)\right) \quad (10)$$

to save computational cost when choosing normalized kernel functions $\|K_{\boldsymbol{x}}\|_{\mathcal{H}} \approx 1$ or ignoring the trivial influence from $\|K_{\boldsymbol{x}}\|_{\mathcal{H}}$ when the values of $\|K_{\boldsymbol{x}}\|_{\mathcal{H}}$ are the same for all $\boldsymbol{x} \in \mathcal{X}$. Since $\partial\mathcal{L}/\partial f$ has positive correlation with $\|f - f^*\|_{\mathcal{H}}$: $\partial\mathcal{L}/\partial f$ decrease as f gradually approaches f^* (Boyd et al., 2004; Coleman, 2012), it is computationally plausible to maximize $|f(\boldsymbol{x}) - f^*(\boldsymbol{x})|$ rather than $\partial\mathcal{L}/\partial f|_{f^t(\boldsymbol{x}^t), y^t}$ directly, such that GFT-1 also can be implemented under the gray-box setting where $\mathcal{L}$ and η could be unknown. Maximizing $|f(\boldsymbol{x}) - f^*(\boldsymbol{x})|$ is easy to compute, since it avoids calculation of the partial derivative when example selection. Compared to RFT, GFT selects examples with a greedy strategy for fast convergence.

Allowing more examples to be fed, *i.e.*, feeding a pack of teaching examples instead of a single one at each iteration, we present the Greedy-k Functional Teaching (GFT-k) as a heuristic. Given a nonparametric target model f^*, GFT-k is to pick k examples satisfying

$$\left(\boldsymbol{x}_j^{t^*} = \arg\max_{\boldsymbol{x}_i^t \in \mathcal{X} - \{\boldsymbol{x}_i^{t^*}\}_{i=1}^{j-1}} \left\| \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}_i^t), y_i^t} K\left(\boldsymbol{x}_i^t, \cdot\right) \right\|_{\mathcal{H}}, y_j^{t^*} = f^*\left(\boldsymbol{x}_j^{t^*}\right)\right) \quad (11)$$

as the pack of optimal examples to learners at t-th iteration, $t = 0, 1, \ldots, \text{ITD}_{\text{GFT}}$ and $j = 1, \ldots, k$.

The hyper parameter k can take the form of either an integer counting the number of examples, where $k \in \mathbb{N}$, or a decimal representing the ratio of the pack to the whole pool, where $k \in [0, 1]$. The pseudo code for RFT, GFT-1, and GFT-k is given in Algorithm 1 which encapsulates these algorithms.

Remark 7. *For the pool-based teacher who can only provide teaching examples from a pool $\mathcal{P} \subsetneq \mathcal{X}$, RFT and GFT could still work by replacing $\mathcal{X}$ by $\mathcal{P}$. However, f^t might converge to the suboptimal $f^{*\prime}$ when the optimal examples $\boldsymbol{x}^{t^*} \in \mathcal{X} - \mathcal{P}$ and therefore the pool-based teacher cannot provide them to learners.*

Algorithm 1 Random / Greedy Functional Teaching

Input: Target f^*, initial f^0, per-iteration pack size k, small constant $\epsilon > 0$ and maximal iteration number T.

Set $f^t \leftarrow f^0$, $t = 0$.

while $t \leq T$ and $\|f^t - f^*\|_{\mathcal{H}} \geq \epsilon$ **do**

  The teacher selects k teaching examples:

  Initialize the pack of teaching examples $\mathcal{K} = \emptyset$;

  for $j = 1$ **to** k **do**

    (**RFT**) 1. Pick $\boldsymbol{x}_j^{t^*} \in \mathcal{X}$ randomly;

    (**GFT**) 1. Pick $\boldsymbol{x}_j^{t^*}$ with the maximal difference between f^t and f^*:

$$\boldsymbol{x}_j^{t^*} = \arg\max_{\boldsymbol{x}_i^t \in \mathcal{X} - \{\boldsymbol{x}_i^{t^*}\}_{i=1}^{j-1}} \left| f^t(\boldsymbol{x}_i^t) - f^*(\boldsymbol{x}_i^t) \right|;$$

    2. Add $\left(\boldsymbol{x}_j^{t^*}, y_j^{t^*} = f^*\left(\boldsymbol{x}_j^{t^*}\right)\right)$ into $\mathcal{K}$.

  end

  Provide $\mathcal{K}$ to learners.

  The learner updates f^t based on received $\mathcal{K}$:

  $f^t \leftarrow f^t - \eta^t \mathcal{G}(\mathcal{L}; f^t; \mathcal{K})$.

  Set $t \leftarrow t + 1$.

end

4.3. Analysis of Iterative Teaching Dimension

We begin with iterative teaching dimension analysis of RFT under the assumptions (Shen et al., 2020) on $\mathcal{L}$ and the kernel function $K(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{H}$ as below.

Assumption 8. *The loss function $\mathcal{L}(f)$ is $L_{\mathcal{L}}$-Lipschitz smooth,* i.e., *$\forall f, f' \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{X}$*

$$\left|E_{\boldsymbol{x}}\left[\nabla_f \mathcal{L}(f)\right] - E_{\boldsymbol{x}}\left[\nabla_f \mathcal{L}(f')\right]\right| \leq L_{\mathcal{L}} \left|E_{\boldsymbol{x}}\left[f\right] - E_{\boldsymbol{x}}\left[f'\right]\right|,$$

where $L_{\mathcal{L}} \geq 0$ is a constant.

Assumption 9. *The kernel function $K(\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{H}$ is bounded,* i.e., *$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $K(\boldsymbol{x}, \boldsymbol{x}') \leq M_K$, where $M_K \geq 0$ is a constant.*

Recall the definition of the evaluation functional and Fréchet derivative in Definition 1 and 2, respectively, we further introduce a discrepancy (Shen et al., 2020) to quantify the inconsistency between f^t and f^* before theoretical analysis.

Definition 10. *The discrepancy of iterative teaching between f^t and f^* at $\boldsymbol{x}^t$ is defined as*

$$\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) := \left|E_{\boldsymbol{x}^t} \nabla_f \mathcal{L}(f^t, f^*)\right|^2. \quad (12)$$

For succinctness, we rewrite Eq. 12 as $\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) = |E_{\boldsymbol{x}^t} \nabla_f \mathcal{L}(f^t)|^2$ by omitting given f^*. One can observe that $\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t)$ decreases as f^t approaches f^*, thus it can track the convergence state of functional teaching algorithms

and measure the per-iteration improvement about f^t towards f^*. Interestingly, the discrepancy of iterative teaching shares a close connection with the Fisher information (Rissanen, 1996; Schervish, 2012). Note that $|E_{\boldsymbol{x}^t}\nabla_f\mathcal{L}(f^t)|^2$ can be equivalently written as $E_{\boldsymbol{x}}\left(\nabla_f\mathcal{L}(f)\right)^2$. Focus on arithmetic mean rather than point evaluation of $\left(\nabla_f\mathcal{L}(f)\right)^2$, then replacing evaluation functional operator by expectation operator, we have $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}(\boldsymbol{x})}\{\left(\nabla_f\mathcal{L}(\boldsymbol{x};f)\right)^2\}$, which can be viewed as a nonparametric Fisher information for convex loss function. Let f degenerate into the unknown parameter θ and $\mathcal{L}$ be the natural logarithm of the likelihood function $\ell(\boldsymbol{x};\theta)$, we have $\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}(\boldsymbol{x})}\{\left(\nabla_\theta\log\ell(\boldsymbol{x};\theta)\right)^2\}$. Therefore, nonparametric Fisher information for convex loss function can be viewed as a kind of generalized Fisher information, which extends the natural logarithm of likelihood function to a convex loss function and the unknown parameter to a general mapping. More discussion is in Appendix A.

Random Functional Teaching. Recall the teaching settings (Eq. 3, Eq. 4), we analyze per-iteration reduction w.r.t. $\mathcal{L}$.

Lemma 11. *(Sufficient Descent for RFT) Under Assumption 8 and 9, if $\eta^t \leq 1/(2L_{\mathcal{L}}\cdot M_K)$, RFT teachers can reduce the loss $\mathcal{L}$:*

$$\mathcal{L}(f^{t+1})-\mathcal{L}(f^t)\leq -\eta^t/2\cdot\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t). \tag{13}$$

Proof of the Lemma 11 is in Appendix B. Before the convergence of RFT algorithm, the decrease of $\mathcal{L}$ has a negative upper bound expressed by $\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t)$, which is determined by learning rate η^t, loss $\mathcal{L}$, mastery degree f^t and teaching example $\boldsymbol{x}^t$. One can see that these four factors are independent so they affect per-iteration reduction of $\mathcal{L}$ independently. Therefore, even though teachers fail to observe all factors under the gray-box setting, they can also assume that unknown factors are fixed, and optimize example feeding based on tracked f^t to steepen gradients. This is consistent with the motivation of GFT deriving from Proposition 5 and Theorem 6.

Theorem 12. *(Convergence for RFT) Suppose the model of learners is initialized with $f^0\in\mathcal{H}$ and returns $f^t\in\mathcal{H}$ after t iterations, we have the upper bound of minimal $\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t)$:*

$$\min_t\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t)\leq 2\mathcal{L}(f^0)/\left(\tilde{\eta}t\right), \tag{14}$$

where $0<\tilde{\eta}=\min_t\eta^t\leq\frac{1}{2L_{\mathcal{L}}\cdot M_K}$.

The proof of the Theorem 12 is given in Appendix B. It follows from Eq. 14 that the upper bound of minimal $\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t)$ converges at the rate of $\mathcal{O}(1/t)$ and $\min_t\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t)\to 0$ as $t\to\infty$, which means it needs $\mathcal{O}(2\mathcal{L}(f^0)/\left(\tilde{\eta}\epsilon\right))$ iterations for RFT to achieve a stationary point with constant $\epsilon>0$. Therefore, we conclude that ITD of RFT is $\mathcal{O}(2\mathcal{L}(f^0)/\left(\tilde{\eta}\epsilon\right))<\infty$, which suggests feasibility of our extension from the parametric IMT to nonparametric IMT.

Greedy Functional Teaching. Compared to RFT, GFT provably enjoys a faster convergence rate and needs fewer iterations to converge, *i.e.*, lower ITD.

Lemma 13. *(Sufficient Descent for GFT) Under Assumption 8 and 9, if $\eta^t\leq 1/(2L_{\mathcal{L}}\cdot M_K)$, GFT teachers can reduce the loss $\mathcal{L}$ at a faster speed:*

$$\begin{aligned}\mathcal{L}(f^{t+1})-\mathcal{L}(f^t) &\leq -\eta^t/2\cdot\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^{t*})\\ &\leq -\eta^t/2\cdot\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t).\end{aligned} \tag{15}$$

The proof of the Lemma 13 is presented in Appendix B. One can observe that per-round improvement of GFT has a tighter bound than that of RFT. The reason is that with a greedy strategy GFT elaborately selects examples by maximizing norm of difference between current and target models, such that learners improve f^t with a steeper step forward f^* in per iteration. Such tighter bound approves the efficiency of GFT theoretically.

Theorem 14. *(Convergence for GFT) Suppose the model of learners is initialized with $f^0\in\mathcal{H}$ and returns $f^t\in\mathcal{H}$ after t iterations, we have the upper bound of minimal $\mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^{j*})$:*

$$\min_j\mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^{j*})\leq\frac{2}{\tilde{\eta}\psi(t)}\mathcal{L}(f^0), \tag{16}$$

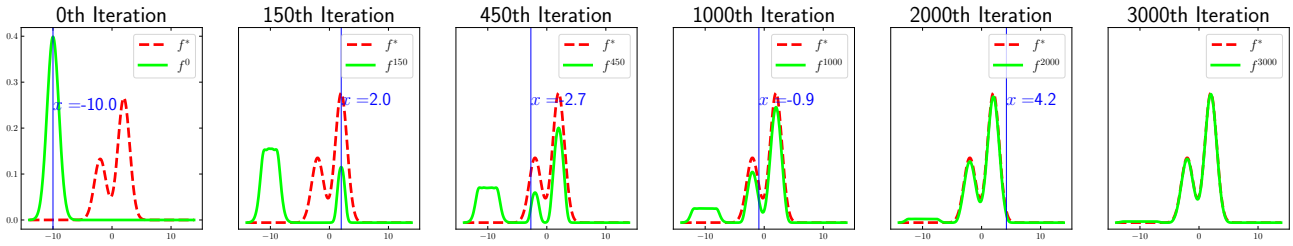
where $0<\tilde{\eta}=\min_t\eta^t\leq\frac{1}{2L_{\mathcal{L}}\cdot M_K}$, $\psi(t)=\sum_{j=0}^{t-1}\gamma^j$ and $\gamma^j=\frac{\mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^j)}{\mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^{j*})}\in(0,1]$ named greedy ratio.

The proof of the Theorem 14 is given in Appendix B. Greedy ratio measures the per-iteration reduction difference between RFT and GFT, and $\psi(t)$ thereby denotes the cumulative difference, *i.e.*superiority of GFT compared to RFT. Intuitively, GFT is strikingly efficient than RFT at beginning and greedy ration is close to 0. As teaching goes on, such divergence vanishes gradually, then greedy ration increasingly close to 1. For $\lim_{t\to\infty}\gamma^t\to 1$, we must have $\lim_{t\to\infty}\psi(t)\to\infty$. Since $\psi(t)\leq t$, one can obtain

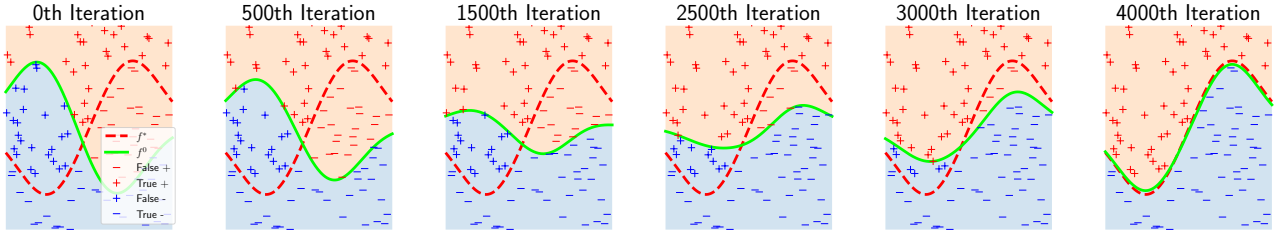
$$\frac{2}{\tilde{\eta}\psi(t)}\mathcal{L}(f^0)\geq\frac{2}{\tilde{\eta}t}\mathcal{L}(f^0), \tag{17}$$

which means $\min_j\mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^{j*})$ has a higher upper bound than $\min_j\mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^j)$. In another word, GFT holds a lower ITD $\mathcal{O}(\psi(\frac{2\mathcal{L}(f^0)}{\tilde{\eta}\epsilon}))$ for convergence.

Remark 15. *Computation complexity. The major computational cost comes from gradient calculation, which could be sped up via parallelization provided in GFT-k. Besides, when the size of an example pool is n, Kernel Operation (KO) and example selection for GFT cost $\mathcal{O}(n^2)$ and $\mathcal{O}(n)$,*



(a) Regression: Teaching a Posterior Probability Density Function.



(b) Classification: Teaching a Nonlinear Decision Boundary.

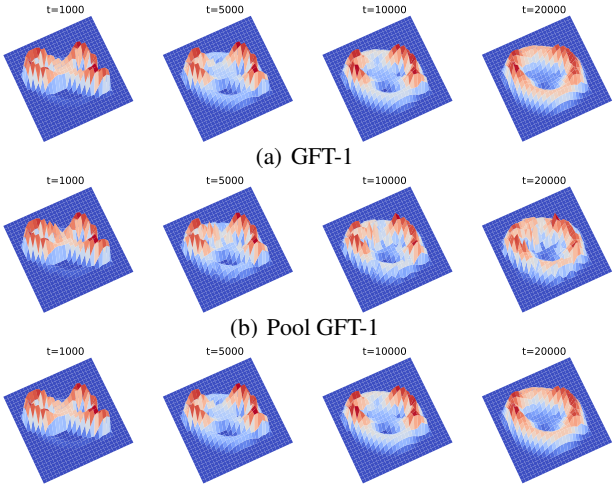
Figure 2. GFT for nonparametric regression and classification teaching problems. (a): The red dashed lines are f^* and the solid lime lines are f^t at different iteration of GFT. Selected examples are pointed out by blue vertical lines. (b): The red dashed lines are f^* when f^0 is represented by the edge between blue and orange regions. x_1 and x_2 are corresponded to x and y axis, respectively. (a)-(b) present the nonparametric teaching ability of helping the learner converge to f^* even from a terrible initial f^0 (without overlap with f^*).

respectively. In large-scale problem, cost of GFT could be saved by implementing it in sub-sampled support of f^* (Politis et al., 1999) and cost of KO could be cut down by a random feature expansion of the kernel (Rahimi & Recht, 2007; Liu, 2017).

5. Experiments and Results

We test our RFT and GFT on both synthetic and real-world data, on which we find these two algorithms present satisfactory capability to tackle nonparametric teaching tasks. Without particular emphasis, experiments are implemented under the synthesis-based teacher setting where the teacher can provide any examples to learners and the knowledge domain is complete. Some detailed settings and extended experiments are given in Appendix C, D.

Synthetic 1D Gaussian Mixture. Consider a nonparametric Bayesian inference problem. The target model is specified as the posterior probability density function (PDF) set to be $f^* = 1/3\mathcal{N}(x; -2, 1) + 2/3\mathcal{N}(x; 2, 1)$, where we denote the PDF of a normal distribution with mean μ and standard deviation σ as $\mathcal{N}(x; \mu, \sigma)$. We assume f^0 for the learner is initialized as $f^0 = \mathcal{N}(x; -10, 1)$. This is a challenging regression teaching problem since f^0 and f^* is far apart (almost without overlap). (a) in Fig. 2 shows that f^0 is guided by GFT to evolve towards f^* directly. It can be found that in spite of obvious difference between f^* and f^0, our GFT can smooth the mode of f^0 where is flatten in f^* and sharpen f^0 towards the mode of f^* via searching x with maximal $|f^*(x) - f^t(x)|$ and feeding it to the learner.



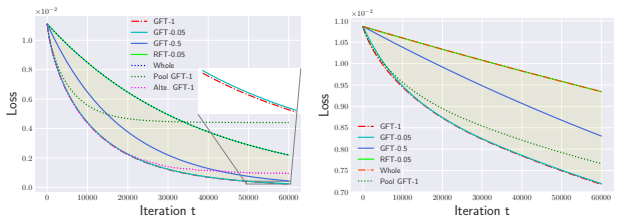
(a) GFT-1

(b) Pool GFT-1

(c) GFT-1 with Alternative

Figure 3. Nonparametric teaching for correcting 8 towards 0. (a): evolution of f^t with GFT-1 algorithm. (b): f^t for GFT-1 under the pool-based teacher. (c): f^t for GFT-1 when occasionally teaching with O. GFT-1 presents satisfied nonparametric teaching capability in these different scenarios.

Synthetic 2D Classification. For a 2D nonparametric classification problem, out of convenience for visualization, the target model is set to be $f^*(x_1, x_2) = x_2 - \exp\left(\frac{x_1-0.5}{0.5}\right)^2 + \exp\left(\frac{x_1+0.5}{0.5}\right)^2$, where x_i represents feature i, $i = 1, 2$. Then, let $f^*(x_1, x_2) = 0$, we can rewrite it as $x_2 = \exp\left(\frac{x_1-0.5}{0.5}\right)^2 - \exp\left(\frac{x_1+0.5}{0.5}\right)^2$ and visualize the decision boundary in a 2D figure. f^0 is set to be $f^0 = x_2 + \exp\left(\frac{x_1-0.3}{0.5}\right)^2 - \exp\left(\frac{x_1+0.6}{0.5}\right)^2$, from which



(a) The Digit Correction (b) The Cheetah Impartation

Figure 4. Comparison of convergence performance for RFT and GFT. The legend of GFT-1 for pool-based teaching is Pool GFT-1 when that for alternative teaching is Alte. GFT-1. The legend, Whole means the teacher provides all pixels to the learner.

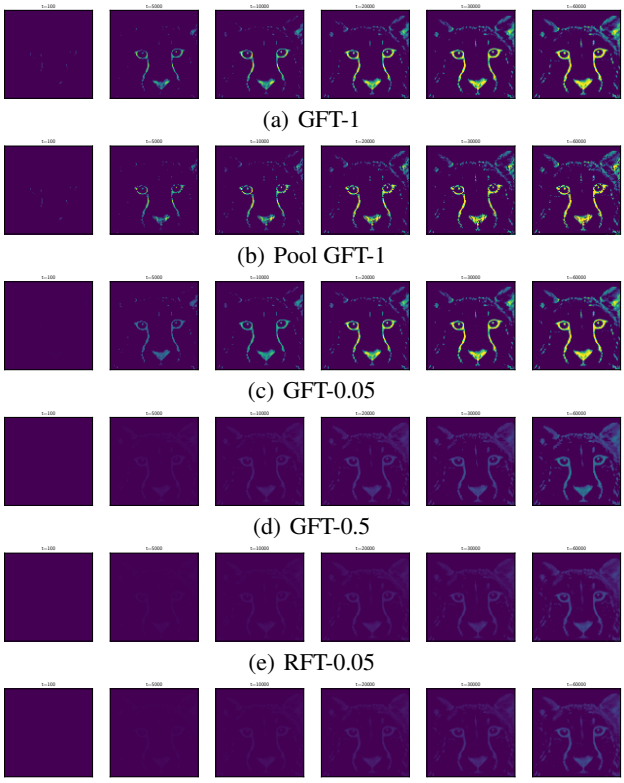
we have $x_2 = \exp\left(\frac{x_1+0.6}{0.5}\right)^2 - \exp\left(\frac{x_1-0.3}{0.5}\right)^2$. (b) in Fig. 2 presents how GFT corrects the inappropriate decision boundary f^0 towards f^*. It can be observed that for a more general function more than PDF, our GFT is also able to amend a bad initialization f^0 towards f^*.

More experiments applying RFT and GFT to teach parameterized target models are given in Appendix D, which shows parametric adaptation of RFT and GFT.

The digit Correction. Consider a digit (MNIST (LeCun, 1998)) teaching instance, one can image a digit figure as a surface in 3D space where z axis is the gray level and x, y axes represent the pixel location. Obviously such complexity surface cannot be identified by a parameter, thus is beyond the capabilities of parametric algorithms (Liu et al., 2017). Initially, the teacher would ask an infant (the learner) *what is digit 0* (f^*)? He would provide a self-convinced but wrong answer as digit 8 (f^0) to the teacher. Based on such a feedback, the teacher would correct f^0 towards f^* via feeding examples (fundamentally is gray value with pixel location). After many rounds of teaching and learning, the learner would evolve its f from incorrect f^0 to ambiguous f^t and final correct f^*, which shares similarity with the process when human beings learn new items (Bengio et al., 2009). We visualize above procedure of our GFT-1 teacher in Fig. 3 (a).

Consider practical pool-based teacher scenario (introduced in Remark 7). We randomly set that 80% pixels are available to the pool-based teacher as $\mathcal{P}$. Fig. 3 (b) shows that our GFT-1 is also effective while f^t cannot converge to f^* due to the limited knowledge domain of the pool-based teacher.

A more interesting case is alternative teaching. Specifically, digit 0 is well-known for the teacher, but lack of Kids Picture Dictionary of 0 at hand he cannot provide wanted teaching examples. Alternatively, notice on similar topological structure between digit 0 and character O (EMNIST from (Cohen et al., 2017)), it is natural to take O as teaching examples. We set the probability of teaching with O as 0.2 in each iteration to test GFT-1. As expected, Fig. 3 (c) shows that GFT-1 also adapts to the alternative teaching with satisfied



(a) GFT-1
(b) Pool GFT-1
(c) GFT-0.05
(d) GFT-0.5
(e) RFT-0.05
(f) Whole

Figure 5. Nonparametric teaching for imparting the cheetah. f of GFT become clear significantly faster than RFT. Part of f are not updated for pool-based teaching, so several dark discontinuity points can be found.

performance. It demonstrates generalizability of GFT-1 as it can be applied in more practical scenarios where only alternative with similar topological structure is accessible. This interesting property may present an intimate connection between our work and transfer learning (Pan & Yang, 2009), domain adaptation (Daume III & Marcu, 2006).

Fig. 4 (a) presents the convergence performance for RFT and GFT under different settings. The yellow region is marked for GFT-k, $k \in (0, 1)$. We see that the loss of GFT declines more dramatically than that of RFT, and it converges to sub-optimal f under the pool-based teacher or alternative teaching scenarios. We leave comparison between RFT and GFT of concrete images like Fig. 3 in Appendix C Fig. 6.

The cheetah impartation. Different from correction tasks where the learner has a preliminary idea of f^*, the impartation problem focus on the learner with no idea about f^*. Concretely, when the teacher asks *what is a cheetah* (Shen et al., 2020), it would be a blank in the learner's mind. As a response, teacher would educate the learner about the cheetah in pixels viewpoint as breaking the whole concept down into smaller points brings better understanding. Fig. 5 compares RFT and GFT under different settings by visualizing f^t therein. We find that GFT is vastly better than RFT that

has roughly the same performance as teaching with whole set (Bottou, 2010). Besides, GFT-1 tends to outperform other GFT algorithms, but fails to teach entirely f^* under the pool-based setting.

We conclude from Fig. 4 (b) that compared with GFT, RFT saves the cost of searching the optimal examples at the expense of slow convergence, and pool-based teaching also suffer from sub-optimization.

6. Concluding Remarks

In this paper, we study a general task, Nonparametric Iterative Machine Teaching (NIMT), which generalizes model space from a finite dimensional one to an infinite dimensional one. We are mainly concerned with the reproduce kernel Hilbert space in this paper. To tackle NIMT, we present a natural baseline algorithm named random functional teaching and propose a greedy one named greedy functional teaching. We theoretically prove that iterative teaching dimension of random functional teaching is $\mathcal{O}(2\mathcal{L}(f^0)/\left(\tilde{\eta}\epsilon\right))$ when greedy functional teaching has a lower iterative teaching dimension $\mathcal{O}(\psi(\frac{2\mathcal{L}(f^0)}{\tilde{\eta}\epsilon}))$ for convergence under mild assumptions. We experimentally demonstrate the efficiency of these two algorithms. Future directions could be more theoretical understanding on NIMT and more efficient functional teaching algorithms with better strategies for potential practical application in deep learning models.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (Grant Number: 62206108), in part by Maritime AI Research Programme (SMI-2022-MTP-06) and AI Singapore OTTC Grant (AISG2-TC-2022-006), and in part by the Research Grants Council of the Hong Kong Special Administrative Region (Grant 16200021).

References

Adams, R. A. and Fournier, J. J. *Sobolev spaces*. Elsevier, 2003.

Alfeld, S., Zhu, X., and Barford, P. Data poisoning attacks against autoregressive models. In *AAAI*, 2016.

Alfeld, S., Zhu, X., and Barford, P. Explicit defense actions against test-set attacks. In *AAAI*, 2017.

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum mean discrepancy gradient flow. In *NeurIPS*, 2019.

Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical review A*, 38(6):3098, 1988.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *ICML*, 2009.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Chen, Y., Singla, A., Mac Aodha, O., Perona, P., and Yue, Y. Understanding the role of adaptivity in machine teaching: The case of version space learners. In *NeurIPS*, 2018.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In *IJCNN*, 2017.

Coleman, R. *Calculus on normed vector spaces*. Springer Science & Business Media, 2012.

Collins, K. M., Bhatt, U., Liu, W., Piratla, V., Sucholutsky, I., Love, B., and Weller, A. Human-in-the-loop mixup. In *UAI*, 2023.

Corder, G. W. and Foreman, D. I. *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons, 2014.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to algorithms*. MIT press, 2022.

Daume III, H. and Marcu, D. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

Dvurechenskii, P., Dvinskikh, D., Gasnikov, A., Uribe, C., and Nedich, A. Decentralize and randomize: Faster algorithm for wasserstein barycenters. In *NeurIPS*, 2018.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Gelfand, I. M., Silverman, R. A., et al. *Calculus of variations*. Courier Corporation, 2000.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *NIPS*, 2016.

Goldman, S. A. and Kearns, M. J. On the complexity of teaching. *Journal of Computer and System Sciences*, 50 (1):20–31, 1995.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Hofmann, T., Schölkopf, B., and Smola, A. J. Kernel methods in machine learning. *The annals of statistics*, 36(3): 1171–1220, 2008.

Hollander, M., Wolfe, D. A., and Chicken, E. *Nonparametric statistical methods*. John Wiley & Sons, 2013.

Hunziker, A., Chen, Y., Mac Aodha, O., Rodriguez, M. G., Krause, A., Perona, P., Yue, Y., and Singla, A. Teaching multiple concepts to a forgetful learner. *arXiv preprint arXiv:1805.08322*, 2018.

Kallenberg, W. C. M., Oosterhoff, J., and Schriever, B. The number of classes in chi-squared goodness-of-fit tests. *Journal of the American Statistical Association*, 80(392): 959–968, 1985.

Kamalaruban, P., Devidze, R., Cevher, V., and Singla, A. Interactive teaching algorithms for inverse reinforcement learning. *arXiv preprint arXiv:1905.11867*, 2019.

Kumar, A., Zhang, H., Singla, A., and Chen, Y. The teaching dimension of kernel perceptron. In *AISTATS*, 2021.

Lax, P. D. *Functional analysis*, volume 55. John Wiley & Sons, 2002.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

Lessard, L., Zhang, X., and Zhu, X. An optimal control approach to sequential machine teaching. In *AISTATS*, 2019.

Liu, J., Zhu, X., and Ohannessian, H. The teaching dimension of linear learners. In *ICML*, 2016.

Liu, Q. Stein variational gradient descent as gradient flow. In *NIPS*, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NIPS*, 2016.

Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., Rehg, J. M., and Song, L. Iterative machine teaching. In *ICML*, 2017.

Liu, W., Dai, B., Li, X., Liu, Z., Rehg, J., and Song, L. Towards black-box iterative machine teaching. In *ICML*, 2018.

Liu, W., Liu, Z., Wang, H., Paull, L., Schölkopf, B., and Weller, A. Iterative teaching by label synthesis. In *NeurIPS*, 2021.

Lutwak, E., Lv, S., Yang, D., and Zhang, G. Extensions of fisher information and stam's inequality. *IEEE transactions on information theory*, 58(3):1319–1327, 2012.

Lv, S. General fisher information matrices of a random vector. *Advances in Applied Mathematics*, 89:18–40, 2017.

Ma, Y., Zhang, X., Sun, W., and Zhu, J. Policy poisoning in batch reinforcement learning and control. In *NeurIPS*, 2019.

Mansouri, F., Chen, Y., Vartanian, A., Zhu, J., and Singla, A. Preference-based batch and sequential teaching: Towards a unified view of models. In *NeurIPS*, 2019.

Mason, L., Baxter, J., Bartlett, P., and Frean, M. Boosting algorithms as gradient descent. In *NIPS*, 1999a.

Mason, L., Baxter, J., Bartlett, P. L., Frean, M., et al. Functional gradient techniques for combining hypotheses. In *NIPS*, 1999b.

Mroueh, Y., Sercu, T., and Raj, A. Sobolev descent. In *AISTATS*, 2019.

Narici, L. and Beckenstein, E. *Topological vector spaces*. Chapman and Hall/CRC, 2010.

Nitanda, A. and Suzuki, T. Functional gradient boosting based on residual network perception. In *ICML*, 2018.

Nitanda, A. and Suzuki, T. Functional gradient boosting for learning residual-like networks with statistical guarantees. In *AISTATS*, 2020.

Noguchi, M. Invariant fisher information. *Differential Geometry and its Applications*, 4(2):179–199, 1994.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2009.

Peltola, T., Çelikok, M. M., Daee, P., and Kaski, S. Machine teaching of active sequential learners. In *NeurIPS*, 2019.

Politis, D. N., Romano, J. P., and Wolf, M. *Subsampling*. Springer Science & Business Media, 1999.

Qian, H., Liu, X.-H., Su, C.-X., Zhou, A., and Yu, Y. The teaching dimension of regularized kernel learners. In *ICML*, 2022.

Qiu, Z., Liu, W., Xiao, T. Z., Liu, Z., Bhatt, U., Luo, Y., Weller, A., and Schölkopf, B. Iterative teaching by data hallucination. In *AISTATS*, 2022.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Rakhsha, A., Radanovic, G., Devidze, R., Zhu, X., and Singla, A. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *ICML*, 2020.

Rissanen, J. J. Fisher information and stochastic complexity. *IEEE transactions on information theory*, 42(1):40–47, 1996.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Schervish, M. J. *Theory of statistics*. Springer Science & Business Media, 2012.

Schölkopf, B., Smola, A. J., Bach, F., et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Shen, Z., Wang, Z., Ribeiro, A., and Hassani, H. Sinkhorn barycenter via functional gradient descent. In *NeurIPS*, 2020.

Singer, I. *The theory of best approximation and functional analysis*. SIAM, 1974.

Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., and Krause, A. On actively teaching the crowd to classify. In *NIPS Workshop on Data Driven Education*, 2013.

Singla, A., Bogunovic, I., Bartók, G., Karbasi, A., and Krause, A. Near-optimally teaching the crowd to classify. In *ICML*, 2014.

Smanski, M. J., Bhatia, S., Zhao, D., Park, Y., BA Woodruff, L., Giannoukos, G., Ciulla, D., Busby, M., Calderon, J., Nicol, R., et al. Functional optimization of gene clusters by combinatorial design and assembly. *Nature biotechnology*, 32(12):1241–1249, 2014.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Tabibian, B., Upadhyay, U., De, A., Zarezade, A., Schölkopf, B., and Gomez-Rodriguez, M. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 116(10): 3988–3993, 2019.

Vajda, I. *Theory of statistical inference and information*. Springer, 1989.

Vajda, I. On convergence of information contained in quantized observations. *IEEE Transactions on Information Theory*, 48(8):2163–2172, 2002.

Wang, P. and Vasconcelos, N. A machine teaching framework for scalable recognition. In *ICCV*, 2021.

Wang, P., Nagrecha, K., and Vasconcelos, N. Gradient-based algorithms for machine teaching. In *CVPR*, 2021.

Xu, Z., Chen, B., Li, C., Liu, W., Song, L., Lin, Y., and Shrivastava, A. Locality sensitive teaching. In *NeurIPS*, 2021.

Ye, J., Wu, P., Wang, J. Z., and Li, J. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9): 2317–2332, 2017.

Zhang, D., Ye, M., Gong, C., Zhu, Z., and Liu, Q. Black-box certification with randomized smoothing: A functional optimization based framework. In *NeurIPS*, 2020a.

Zhang, X., Bharti, S. K., Ma, Y., Singla, A., and Zhu, X. The sample complexity of teaching-by-reinforcement on q-learning. *arXiv preprint arXiv:2006.09324*, 2020b.

Zhou, Y., Nelakurthi, A. R., and He, J. Unlearn what you have learned: Adaptive crowd teaching with exponentially decayed memory learners. In *SIGKDD*, 2018.

Zhou, Y., Nelakurthi, A. R., Maciejewski, R., Fan, W., and He, J. Crowd teaching with imperfect labels. In *WWW*, 2020.

Zhu, X. Machine teaching for bayesian learners in the exponential family. *arXiv preprint arXiv:1306.4947*, 2013.

Zhu, X. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI*, 2015.

Zhu, X., Liu, J., and Lopes, M. No learner left behind: On the complexity of teaching multiple learners simultaneously. In *IJCAI*, 2017.

Zhu, X., Singla, A., Zilles, S., and Rafferty, A. N. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

Zoppoli, R., Sanguineti, M., and Parisini, T. Approximating networks and extended ritz method for the solution of functional optimization problems. *Journal of Optimization Theory and Applications*, 112(2):403–440, 2002.

Appendix

A. Additional Discussions

Broader Impact Machine teaching has been applied in crowd sourcing, computer vision and cyber security – domains with significant societal impacts. This work focuses on theoretical analysis of iterative machine teaching and generalizes parameterized iterative machine teaching to nonparametric scenarios, which is to generalize model space from a finite dimensional one to an infinite dimensional one. This provides possibility of extending parameterized applications to nonparametric cases. Thus, while the contributions of this work are mainly theoretical, there are potential positive impacts in the community of machine teaching and society.

Parametric teaching settings One can rewrite formulations in Section 4.1 into parameterized version via replacing f by w (Liu et al., 2017; 2018; Zhu et al., 2018) as parametric IMT operates in the finite dimensional parameter space. Specifically, the bilevel optimization can be formulated as

$$\mathcal{D}^* = \arg\min_{\mathcal{D}\in\mathbb{D}} \mathcal{M}(\hat{w}, w^*) + \lambda \cdot \text{len}(\mathcal{D}) \quad \text{s.t. } \hat{w} = \mathcal{A}(\mathcal{D}), \tag{18}$$

where notations have same meanings as Eq. 2. Empirical risk minimization $\mathcal{A}(\mathcal{D})$ is as follows

$$\hat{w^*} = \arg\min_{w} \mathbb{E}_{(\boldsymbol{x},y)} \left\{ \mathcal{L}(\langle w, \boldsymbol{x}\rangle, y) \right\}. \tag{19}$$

Besides, parameter w is updated as

$$w^{t+1} \leftarrow w^t - \eta^t \mathcal{G}(\mathcal{L}; w^t; (\boldsymbol{x}^t, y^t)). \tag{20}$$

Nonparametric Fisher information for convex loss function in Section 4.3 Fisher information (Lehmann & Casella, 2006) is a fundamental quantity in statistics and information theory (Vajda, 1989). It measures the information carried by data about an unknown parameter θ. Let

$$\mathcal{I}(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}(\boldsymbol{x})} \left\{ \left(\nabla_\theta \log \ell(\boldsymbol{x}; \theta) \right)^2 \right\} \tag{21}$$

be Fisher information. It can be written (Vajda, 2002) as

$$\mathcal{I}_\phi(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}(\boldsymbol{x})} \left\{ \phi\left(\nabla_\theta \log \ell(\boldsymbol{x}; \theta) \right) \right\}, \tag{22}$$

where $\phi(\cdot) = (\cdot)^2$. There are many works (Noguchi, 1994; Vajda, 2002; Lutwak et al., 2012; Lv, 2017) on generalized Fisher information in terms of explicit form of $\phi(\cdot)$. For example, Kallenberg et al., 1985 considers $\phi(\cdot) = (\cdot)^{4/3}$ and connects it with Pearson goodness of fit test. Besides, let $\phi(\cdot) = -\log(\cdot)$, Eq. 21 is the information divergence (Vajda, 2002).

From another generalized perspective, Eq. 21 can be rewritten in another way as

$$\mathcal{I}(\theta)_{\varphi;\vartheta} = \mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}(\boldsymbol{x})} \left\{ \left(\nabla_\vartheta \varphi(\boldsymbol{x}; \vartheta) \right)^2 \right\}, \tag{23}$$

where $\varphi(\cdot) = \log \ell(\cdot)$ and $\vartheta = \theta$. Meanwhile, concerned with arithmetic mean instead of point evaluation of $\left(\nabla_f \mathcal{L}(f)\right)^2$, $\mathbb{S}_{\mathcal{L}}(f; \boldsymbol{x}) = \left| E_{\boldsymbol{x}} \nabla_f \mathcal{L}(f) \right|^2$ introduced in Definition 10 can be written as

$$\mathbb{E}_{\boldsymbol{x}\sim\mathbb{P}(\boldsymbol{x})} \left\{ \left(\nabla_f \mathcal{L}(\boldsymbol{x}; f) \right)^2 \right\}, \tag{24}$$

which can be viewed as a nonparametric Fisher information for convex loss function. Therefore, extending $\varphi(\cdot)$ from the natural logarithm of the likelihood function, $\log \ell(\cdot)$ to the convex loss function $\mathcal{L}(\cdot)$ and extending ϑ from unknown parameter θ to general mapping f, nonparametric Fisher information for convex loss function can be viewed as a kind of generalized Fisher information.

B. Detailed Proofs

We recommend the literature (Gelfand et al., 2000; Coleman, 2012) for further reading on functional calculus.

Proof of Lemma 4 Let define a function q by adding a small perturbation ϵg ($\epsilon \in \mathbb{R}, g \in \mathcal{H}$) to $f \in \mathcal{H}$, $q = f + \epsilon g$. $q \in \mathcal{H}$ since RKHS is closed under addition and scalar multiplication. Therefore, for a evaluation functional $E_{\boldsymbol{x}}[f] = f(\boldsymbol{x}) : \mathcal{H} \mapsto \mathbb{R}$, we can evaluate q at $\boldsymbol{x}$ as

$$\begin{aligned} E_{\boldsymbol{x}}[q] &= E_{\boldsymbol{x}}[f + \epsilon g] \\ &= E_{\boldsymbol{x}}[f] + \epsilon E_{\boldsymbol{x}}[g] + 0 \\ &= E_{\boldsymbol{x}}[f] + \epsilon \langle K(\boldsymbol{x}, \cdot), g \rangle_{\mathcal{H}} + 0 \end{aligned} \tag{25}$$

Recall implicit definition of Fréchet derivative in RKHS (see Definition 2) $E_{\boldsymbol{x}}[f + \epsilon g] = E_{\boldsymbol{x}}[f] + \epsilon \langle \nabla_f E_{\boldsymbol{x}}[f], g \rangle_{\mathcal{H}} + \mathcal{O}(\epsilon^2)$, it follows from Eq. 25 that we have the gradient of a evaluation functional $\nabla_f E_{\boldsymbol{x}}[f] = K_{\boldsymbol{x}}$.

■

Proof of Theorem 6 Concisely, we omit superscript t for the time being and rewrite Eq. 7 as

$$(\boldsymbol{x}^*, y^*) = \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}} \left\| f - \eta \cdot \mathcal{G} - f^* \right\|_{\mathcal{H}}^2 . \tag{26}$$

Obviously, it is trivial to derive that $\forall (\boldsymbol{x}, y), \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$,

$$\begin{aligned} &\left\| f - \eta \mathcal{G}(\mathcal{L}; f; (\boldsymbol{x}^*, y^*)) - f^* \right\|_{\mathcal{H}}^2 \leq \\ &\quad \left\| f - \eta \mathcal{G}(\mathcal{L}; f; (\boldsymbol{x}, y)) - f^* \right\|_{\mathcal{H}}^2 . \end{aligned} \tag{27}$$

Out of succinctness, we denote $\mathcal{G}^{t^*} \coloneqq \mathcal{G}(\mathcal{L}; f^t; (\boldsymbol{x}^{t^*}, y^{t^*}))$ and $\mathcal{G}^t \coloneqq \mathcal{G}(\mathcal{L}; f^t; (\boldsymbol{x}^t, y^t))$. For l.h.s. of expression 27, we can expand it as

$$\begin{aligned} &\left\| f - \eta \mathcal{G}(\mathcal{L}; f; (\boldsymbol{x}^*, y^*)) - f^* \right\|_{\mathcal{H}}^2 \\ = \;& \left\| f - f^* \right\|_{\mathcal{H}}^2 + \eta^2 \left\| \mathcal{G}^* \right\|_{\mathcal{H}}^2 - \eta \left\langle \mathcal{G}^*, f - f^* \right\rangle_{\mathcal{H}} . \end{aligned} \tag{28}$$

Similarly, we can also expand r.h.s. of expression 27 as

$$\begin{aligned} &\left\| f - \eta \mathcal{G}(\mathcal{L}; f; (\boldsymbol{x}, y)) - f^* \right\|_{\mathcal{H}}^2 \\ = \;& \left\| f - f^* \right\|_{\mathcal{H}}^2 + \eta^2 \left\| \mathcal{G} \right\|_{\mathcal{H}}^2 - \eta \left\langle \mathcal{G}, f - f^* \right\rangle_{\mathcal{H}} . \end{aligned} \tag{29}$$

Combining expansion of expression 27 together, we have

$$\begin{aligned} &\left\| f - f^* \right\|_{\mathcal{H}}^2 + \eta^2 \left\| \mathcal{G}^* \right\|_{\mathcal{H}}^2 - \eta \left\langle \mathcal{G}^*, f - f^* \right\rangle_{\mathcal{H}} \\ \leq \;& \left\| f - f^* \right\|_{\mathcal{H}}^2 + \eta^2 \left\| \mathcal{G} \right\|_{\mathcal{H}}^2 - \eta \left\langle \mathcal{G}, f - f^* \right\rangle_{\mathcal{H}} . \end{aligned} \tag{30}$$

After rearranging, we can obtain

$$\langle \mathcal{G}^* - \mathcal{G}, f - f^* \rangle_{\mathcal{H}} \geq \eta / 2 (\left\| \mathcal{G}^* \right\|_{\mathcal{H}}^2 - \left\| \mathcal{G} \right\|_{\mathcal{H}}^2) \geq 0. \tag{31}$$

■

Proof of Lemma 11 Recall the definition of Fréchet derivative in Definition 2. It follows from the convexity of $\mathcal{L}$ that we have

$$\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \leq \langle f^{t+1} - f^t, \nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \rangle_{\mathcal{H}}. \tag{32}$$

Based on optimization algorithm in Eq. 4, the right term of Eq. 32 can be expressed as

$$\langle f^{t+1} - f^t, \nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \rangle_{\mathcal{H}} = \langle -\eta^t \mathcal{G}^t, \nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \rangle_{\mathcal{H}}.$$

Substituting $\mathcal{G}^t = \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}^t), y^t} \cdot K(\boldsymbol{x}^t, \cdot)$ in and removing constants out of inner product operation, it yields

$$\begin{aligned} & \langle -\eta^t \mathcal{G}^t, \nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \rangle_{\mathcal{H}} \\ = & -\eta^t \left\langle \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}^t), y^t} \cdot K(\boldsymbol{x}^t, \cdot), \nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right\rangle_{\mathcal{H}} \\ = & -\eta^t \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}^t), y^t} \left\langle K(\boldsymbol{x}^t, \cdot), \nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right\rangle_{\mathcal{H}}. \end{aligned} \tag{33}$$

Recall the definition of the evaluation functional in RKHS in Definition 1

$$E_{\boldsymbol{x}}[f] = \langle f, K_{\boldsymbol{x}}(\cdot) \rangle_{\mathcal{H}} \tag{34}$$

and the fact $y = f^*(\boldsymbol{x}) = E_{\boldsymbol{x}}[f^*]$, we can rewrite the last term in Eq. 33 as

$$\begin{aligned} & -\eta^t \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}^t), y^t} \left\langle K(\boldsymbol{x}^t, \cdot), \nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right\rangle_{\mathcal{H}} \\ = & -\eta^t \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}^t), y^t} \times E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right] \\ = & -\eta^t E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^t} \right] \times E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right]. \end{aligned} \tag{35}$$

For succinctness, denote $\xi^t := E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^t} \right]$ and $\xi^{t+1} := E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right]$, then Eq. 12 can be tersely expressed $\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) = \left| E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^t} \right] \right|^2 = (\xi^t)^2$, and we thus can rewrite Eq. 35 as follows:

$$\begin{aligned} & -\eta^t E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^t} \right] \times E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right] \\ = & -\eta^t \xi^t \times \xi^{t+1} \\ = & -\eta^t \xi^t \times \left(\xi^t + \xi^{t+1} - \xi^t \right) \\ = & -\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) - \eta^t \xi^t \times \left(\xi^{t+1} - \xi^t \right) \\ = & -\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) + \eta^t \left(\xi^{t+1} - \xi^t - \xi^{t+1} \right) \times \left(\xi^{t+1} - \xi^t \right) \\ = & -\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) + \eta^t \left(\xi^{t+1} - \xi^t \right)^2 - \eta^t \xi^{t+1} \left(\xi^{t+1} - \xi^t \right) \\ = & -\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) + \eta^t \left(\xi^{t+1} - \xi^t \right)^2 - \eta^t \left(\xi^{t+1} - 1/2\xi^t \right)^2 + 1/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) \\ = & -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) + \eta^t \left(\xi^{t+1} - \xi^t \right)^2 - \eta^t \left(\xi^{t+1} - 1/2\xi^t \right)^2 \\ \leq & -3/4\eta^t \mathbb{S}_{\mathcal{L}}\left(f^t; \boldsymbol{x}^t \right) + \eta^t \left(\xi^{t+1} - \xi^t \right)^2. \end{aligned} \tag{36}$$

Substituting the concrete expression of $\xi^t = E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^t} \right]$ and $\xi^{t+1} = E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right]$ in, it follows from linearity of evaluation functional that

$$\begin{aligned} & -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) + \eta^t \left(\xi^{t+1} - \xi^t \right)^2 \\ = & -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) + \eta^t \left(E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^{t+1}} \right] - E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^t} \right] \right)^2 \\ = & -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) + \eta^t \left(E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^{t+1}} - \nabla_f \mathcal{L}(f)|_{f=f^t} \right] \right)^2. \end{aligned} \tag{37}$$

Under L-Lipschitz smooth Assumption 8 and bounded kernel function Assumption 9, we have

$$\begin{aligned}
& -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) + \eta^t \left(E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^{t+1}} - \nabla_f \mathcal{L}(f)|_{f=f^t} \right] \right)^2 \\
\leq & -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) + \eta^t \left(L_{\mathcal{L}} \cdot E_{\boldsymbol{x}^t}\left[\left| f^{t+1} - f^t \right| \right] \right)^2 \\
= & -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) + \eta^t \left(L_{\mathcal{L}} \eta^t \xi^t E_{\boldsymbol{x}^t}\left[K_{\boldsymbol{x}^t} \right] \right)^2 \\
= & -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) + L_{\mathcal{L}}^2 \left(\eta^t\right)^3 \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) K^2\left(\boldsymbol{x}^t,\boldsymbol{x}^t\right) \\
\leq & -3/4\eta^t \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) + L_{\mathcal{L}}^2 \left(\eta^t\right)^3 \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) (M_K)^2 \\
= & -\eta^t \left(3/4 - L_{\mathcal{L}}^2 (\eta^t)^2 (M_K)^2 \right) \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t).
\end{aligned} \tag{38}$$

Consequently, we obtain

$$\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \leq -\eta^t \left(3/4 - L_{\mathcal{L}}^2 (\eta^t)^2 M_K^2 \right) \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t), \tag{39}$$

and hence $\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \leq -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t)$ if $\eta^t \leq \frac{1}{2L_{\mathcal{L}} \cdot M_K}$

■

Proof of Theorem 12 Recall Lemma 11, when $\eta^t \leq \frac{1}{2L_{\mathcal{L}} \cdot M_K}$,

$$\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \leq -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) \tag{40}$$

Rearranging above, we have:

$$\frac{2\left(\mathcal{L}(f^t) - \mathcal{L}(f^{t+1})\right)}{\eta^t} \geq \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t). \tag{41}$$

Equivalently, $\frac{2\left(\mathcal{L}(f^j) - \mathcal{L}(f^{j+1})\right)}{\eta^j} \geq \mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^j)$. Consequently, plugging $j = 0, 1 \ldots, t-1$ in it and summing them up, we hence have

$$\sum_{j=0}^{t-1} \mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^j) \leq 2 \sum_{j=0}^{t-1} \frac{\mathcal{L}(f^j) - \mathcal{L}(f^{j+1})}{\eta^j} \leq \frac{2}{\tilde{\eta}} \sum_{j=0}^{t-1} \left(\mathcal{L}(f^j) - \mathcal{L}(f^{j+1}) \right), \tag{42}$$

where $\tilde{\eta} = \min_j \eta^j > 0$. Expanding the r.h.s. term in Eq. 42 yields

$$\frac{2}{\tilde{\eta}} \sum_{j=0}^{t-1} \left(\mathcal{L}(f^j) - \mathcal{L}(f^{j+1}) \right) = \frac{2}{\tilde{\eta}} \left(\mathcal{L}(f^0) - \mathcal{L}(f^t) \right) \leq \frac{2}{\tilde{\eta}} \mathcal{L}(f^0). \tag{43}$$

In terms of the l.h.s. term in Eq. 42, we must have

$$\sum_{j=0}^{t-1} \mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^j) \geq t \cdot \min_j \mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^j). \tag{44}$$

Combining expression 43 and 44, we thus have

$$t \cdot \min_t \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) \leq \sum_{j=0}^{t-1} \mathbb{S}_{\mathcal{L}}(f^j;\boldsymbol{x}^j) \leq \frac{2}{\tilde{\eta}} \mathcal{L}(f^0), \tag{45}$$

from which, we can derive

$$\min_t \mathbb{S}_{\mathcal{L}}(f^t;\boldsymbol{x}^t) \leq 2\mathcal{L}(f^0) / \left(\tilde{\eta} t\right). \tag{46}$$

■

It suggests that learners could also conduct its stationary state as: In each round, check if $\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) \leq \epsilon$. If it holds, then they have already reached the ϵ approximating and they can send a terminated signal to teachers; otherwise teachers proceed. The termination occurs within $\mathcal{O}(2\mathcal{L}(f^0)/(\tilde{\eta}\epsilon))$ loops.

Proof of Lemma 13 Recall practical Greedy Functional Teaching in Eq. 10

$$\left(\boldsymbol{x}^{t^*} = \arg\max_{\boldsymbol{x}^t \in \mathcal{X}} \left| E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^t}\right]\right|, y^* = E_{\boldsymbol{x}^{t^*}}\left[f^*\right]\right). \tag{47}$$

Obviously, it is trivial to see that $\forall \boldsymbol{x}^t \in \mathcal{X}$,

$$\left|E_{\boldsymbol{x}^{t^*}}\left[\nabla_f \mathcal{L}(f)|_{f=f^t}\right]\right|^2 \geq \left|E_{\boldsymbol{x}^t}\left[\nabla_f \mathcal{L}(f)|_{f=f^t}\right]\right|^2. \tag{48}$$

Analogous to the Proof of Lemma 11 in B, we can derive

$$\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \quad \leq \quad -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t^*}), \tag{49}$$

if $0 < \eta^t \leq \frac{1}{2L_{\mathcal{L}} \cdot M_K}$. Consequently, we have

$$\mathcal{L}(f^{t+1}) - \mathcal{L}(f^t) \leq -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t^*}) \leq -\eta^t/2 \cdot \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t). \tag{50}$$

■

Proof of Theorem 14 Recall the result of Lemma 11, when $0 < \eta^t \leq \frac{1}{2L_{\mathcal{L}} \cdot M_K}$

$$\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t) \leq \frac{2\left(\mathcal{L}(f^t) - \mathcal{L}(f^{t+1})\right)}{\eta^t}. \tag{51}$$

Before converging to the stationary state, $\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t^*}) > 0$. Therefore, we can express it as

$$\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t^*}) \cdot \frac{\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t)}{\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t^*})} \leq \frac{2\left(\mathcal{L}(f^t) - \mathcal{L}(f^{t+1})\right)}{\eta^t}. \tag{52}$$

For succinctness, denote $\gamma^t := \frac{\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t)}{\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t^*})}$, namely greedy ratio, we have

$$\gamma^t \cdot \mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t^*}) \leq \frac{2\left(\mathcal{L}(f^t) - \mathcal{L}(f^{t+1})\right)}{\eta^t}. \tag{53}$$

Different to expression 44, we have

$$\sum_{j=0}^{t-1} \gamma^j \cdot \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j^*}) \geq \min_j \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j^*}) \cdot \sum_{j=0}^{t-1} \gamma^j. \tag{54}$$

Since $\boldsymbol{x}^{t^*} = \arg\max_{\boldsymbol{x}^t \in \mathcal{X}} \left\| \frac{\partial \mathcal{L}}{\partial f}\Big|_{f^t(\boldsymbol{x}^t), y^*} K(\boldsymbol{x}^t, \cdot)\right\|_{\mathcal{H}}$, we derive $\left|E_{\boldsymbol{x}^t} \nabla_f \mathcal{L}(f^t, f^*)\right|^2 \leq \left|E_{\boldsymbol{x}^{t^*}} \nabla_f \mathcal{L}(f^t, f^*)\right|^2$. Therefore, $\gamma^t = \frac{\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^t)}{\mathbb{S}_{\mathcal{L}}(f^t; \boldsymbol{x}^{t^*})} = \frac{\left|E_{\boldsymbol{x}^t} \nabla_f \mathcal{L}(f^t, f^*)\right|^2}{\left|E_{\boldsymbol{x}^{t^*}} \nabla_f \mathcal{L}(f^t, f^*)\right|^2} \in (0, 1]$ and we have $\min_j \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j^*}) \cdot \sum_{j=0}^{t-1} \gamma^j \leq t \cdot \min_j \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j^*})$. Besides, similar to expression 43, we have

$$\frac{2}{\tilde{\eta}} \sum_{j=0}^{t-1} \left(\mathcal{L}(f^j) - \mathcal{L}(f^{j+1})\right) = \frac{2}{\tilde{\eta}} \left(\mathcal{L}(f^0) - \mathcal{L}(f^t)\right) \leq \frac{2}{\tilde{\eta}} \mathcal{L}(f^0), \tag{55}$$

where $\tilde{\eta} = \min_j \eta^j > 0$. To sum up, we have

$$\min_j \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j^*}) \cdot \sum_{j=0}^{t-1} \gamma^j \leq \frac{2}{\tilde{\eta}} \mathcal{L}(f^0), \tag{56}$$

For succinctness, denote $\psi(t) \coloneqq \sum_{j=0}^{t-1} \gamma^j$. Note that $\lim_{t\to\infty} \gamma^t \to 1 \Rightarrow \lim_{t\to\infty} \psi(t) = \lim_{t\to\infty} \sum_{j=0}^{t-1} \gamma^j \to \infty$. Rearranging, we obtain

$$\min_j \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j^*}) \leq \frac{2}{\tilde{\eta}\psi(t)} \mathcal{L}(f^0). \tag{57}$$

Since $\psi(t) = \sum_{j=0}^{t-1} \gamma^j \leq t \cdot \max_j \gamma^j$ and $\gamma^j \in (0, 1]$, we have $1/\psi(t) \geq 1/(t \cdot \max_j \gamma^j) \geq 1/t$. Therefore, we derive

$$\frac{2}{\tilde{\eta}\psi(t)} \mathcal{L}(f^0) \geq \frac{2}{\tilde{\eta} t \max_j \gamma^j} \mathcal{L}(f^0) \geq \frac{2}{\tilde{\eta} t} \mathcal{L}(f^0), \tag{58}$$

which means $\min_j \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^{j^*})$ has a higher upper bound than $\min_j \mathbb{S}_{\mathcal{L}}(f^j; \boldsymbol{x}^j)$. Note that $\max_j \gamma^j$ is determined by the randomness introduced by sampling and is dependent on t. To be specific, $\max_j \gamma^j$ would be close to 0 at the beginning and $\max_j \gamma^j$ approaches 1 as t increases. It means GFT will drop $\mathcal{L}$ faster than RFT at first, which is also demonstrated in Fig. 4.

We see that t is to measure the iteration number of RFT and $\psi(t)$ is to measure that of GFT. Let set $\psi(t) = \tau \leq t$, then we have $t = \psi^{-1}(\tau)$. For RFT, we can derive

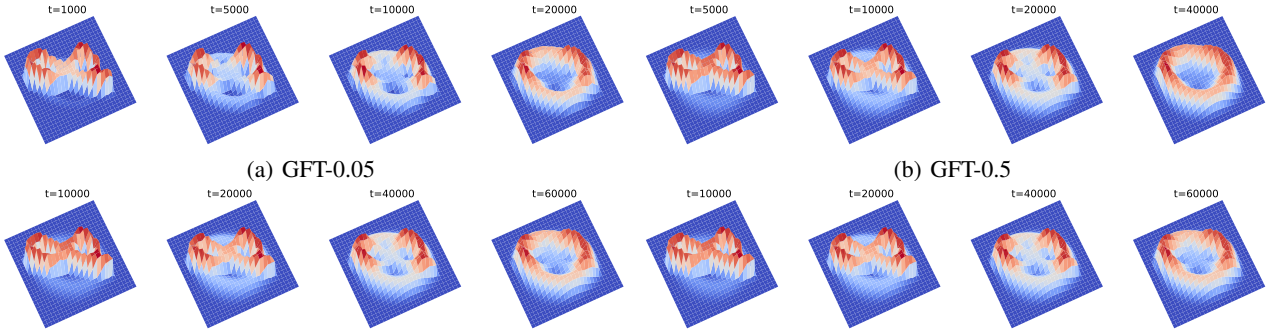
$$t \geq 2\mathcal{L}(f^0)/\left(\tilde{\eta}\epsilon\right). \tag{59}$$

Therefore, plugging $t = \psi^{-1}(\tau)$ into it and $\psi(\cdot)$ is monotonically increasing, we have $\psi^{-1}(\tau) \geq 2\mathcal{L}(f^0)/\left(\tilde{\eta}\epsilon\right)$, that is

$$\tau \geq \psi(2\mathcal{L}(f^0)/\left(\tilde{\eta}\epsilon\right)). \tag{60}$$

τ measures the iteration number of GFT and $\psi(\frac{2\mathcal{L}(f^0)}{\tilde{\eta}\epsilon}) \leq \frac{2\mathcal{L}(f^0)}{\tilde{\eta}\epsilon}$. It means that ITD of GFT is $\mathcal{O}(\psi(\frac{2\mathcal{L}(f^0)}{\tilde{\eta}\epsilon})) \leq \mathcal{O}(2\mathcal{L}(f^0)/\left(\tilde{\eta}\epsilon\right))$. ■

C. Detailed Experiments



(a) GFT-0.05 (b) GFT-0.5

(c) RFT-0.05 (d) Whole

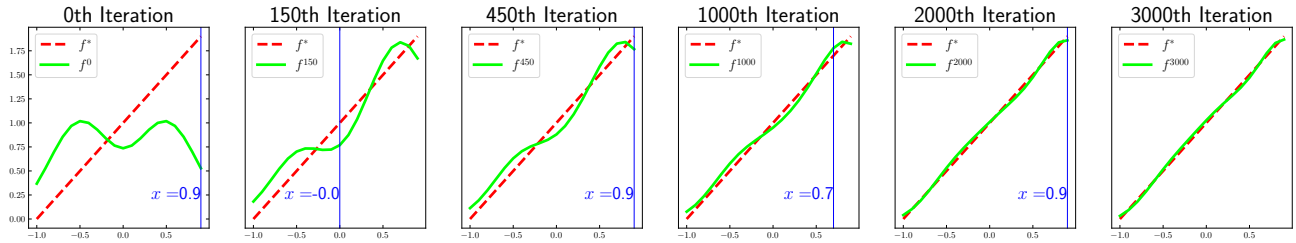
Figure 6. Comparing RFT and GFT when nonparametric teaching for correcting 8 towards 0.

In computer, operations are discrete. Therefore, we use dense pairwise points $\{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}_{i=1}^n$ to represent a function f. For the pool-based teacher (refer to Remark 7), we use sparse pairwise point to denote $\mathcal{P}$. The pool-based teacher knows f^* but cannot provide some teaching examples out of the pool. For all experiments, we set kernel as the popular and general RBF $K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\left\|\frac{\boldsymbol{x}-\boldsymbol{x}'}{2}\right\|_2^2\right)$. We specifically take empirical (average) L_2 norm defined in Hilbert space to measure the difference between f and f^*,

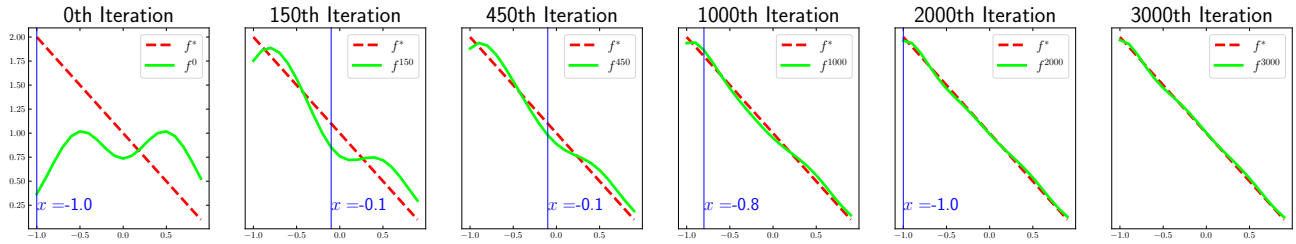
$$\mathcal{M}(f, f^*) = \|f - f^*\|_{\mathcal{H}} = \frac{1}{n}\sqrt{\sum_{i=1}^n \left(f(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)\right)^2}. \tag{61}$$

Our implementation is based on Intel(R) Core(TM) i7-8750H and NVIDIA GTX 1050 Ti with Max-Q Design.

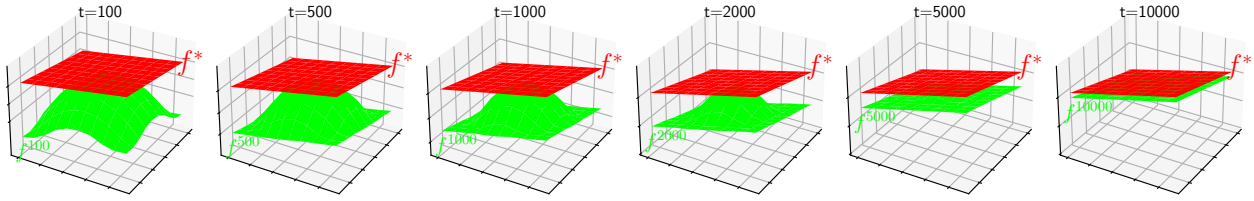
Synthetic 1D Gaussian Mixture. For this regression problem, we assume the loss function of the learner is square loss $\mathcal{L} = (y - f(\boldsymbol{x}))^2$, and we set it unknown for the teacher. We call the dense pairwise points as pixels which are generated by `arange(-14, 14, 0.1)`. The learning rate η^t is fixed as 0.01. Besides, the teacher will stop if $\mathcal{M}(f^t, f^*) < 0.0001$.



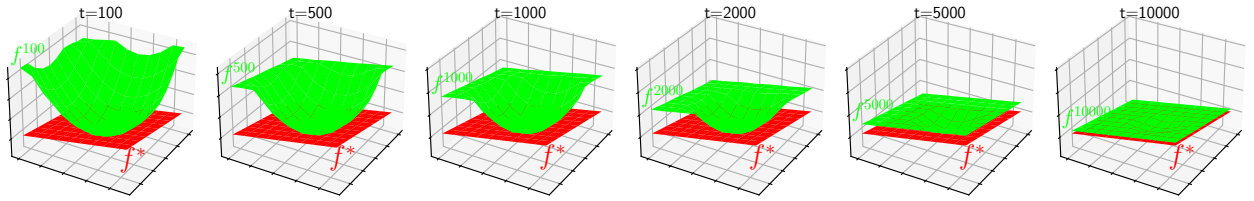
(a) Teaching a Parametrized Target Model $y = x + 1$.



(b) Teaching a Parametrized Target Model $y = -x + 1$.



(c) Teaching a Parametrized Target Model $z = x + y - 8$.



(d) Teaching a Parametrized Target Model $z = x + y - 8$.

Figure 7. GFT for 2D and 3D parameterized target models. (a)-(b): The red dashed lines are f^* and the solid lime lines are f^t at different iteration of GFT. Selected examples are pointed out by blue vertical lines. (c)-(d): The red planes are f^* when f^t is represented by curved lime surfaces. Both 2D and 3D cases show that functional teaching ability of helping the learner converge to f^* even from a bad initial f^0 (without overlap with f^*), which means that the functional teaching algorithm GFT is well-adapted for parameterized target models.

Synthetic 2D Classification. For such classification problem, we assume the loss function of the learner is hinge loss $\mathcal{L} = \max(0, 1 - y \cdot f(\boldsymbol{x}))$ unknown for the teacher. Pixels are generated by `arange(-1, 1, 0.01)`. The learning rate η^t is fixed as 0.001. Besides, the teacher will stop if $\mathcal{M}(f^t, f^*) < 0.001$.

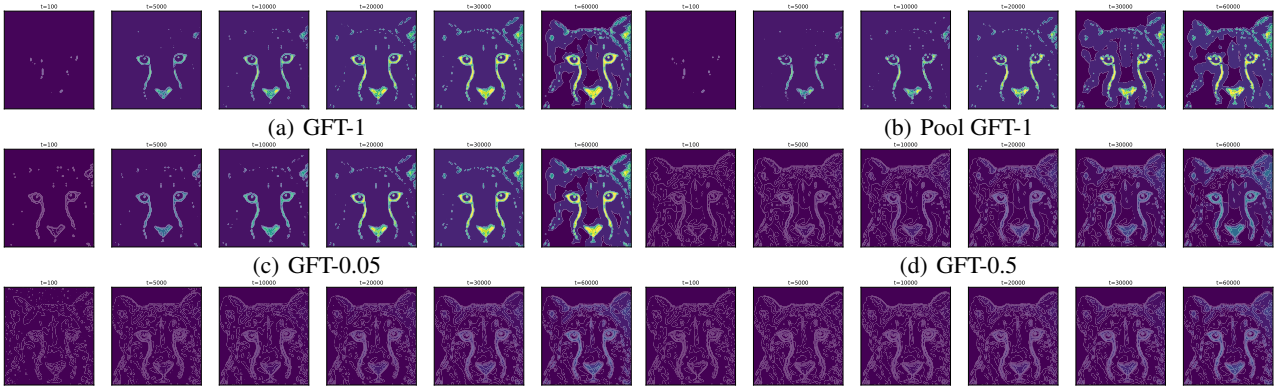
The digit Correction. We tend to recover how an infant (the learner) update its opinion about digit 0 when taught. The learning rate η^t is fixed as 0.01. $\mathcal{L} = (y - f(\boldsymbol{x}))^2$ is unknown for the teacher. We derive the target f^*, optimal 0 via averaging all images of digit 0 in MNIST (LeCun, 1998) (both training and testing sets). We casually pick one digit 8 image as f^*. In Figure 4, the loss is $\mathcal{M}(f^t, f^*)$ rather than that of learners $\mathcal{L}$.

In pool-based teaching, the ratio between the pool and the whole sapce can be adjusted, and we set it as 0.8.

In alternative teaching, the alternative of digit 0, character O is selected from EMNIST (Cohen et al., 2017) via minimizing $\mathcal{M}(f^O, f^*)$, where f^O denotes the character O image. We also scale f^O to match the magnitude of f^* for eliminating influence introduced by the magnitude. the probability of teaching with character O instead of digit 0 can be modified, and we let it as 0.2.

The comparison between RFT and GFT of is presented in Fig. 6. It shows that for GFT, large k (proportion) would delay the convergence by comparing (a) and (b). Besides, GFT is better than RFT when the hyper parameter k is the same via contrasting (a) and (c). Further, (c)-(d) show that RFT has roughly similar performance as teaching with whole set.

The cheetah impartation. The learning rate η^t is fixed as 0.01. $\mathcal{L} = (y - f(\boldsymbol{x}))^2$ is unknown for the teacher. We derive this cheetah figure from Shen et al., 2020 who use pickles to sketch. Differently, we regard a figure as a smooth function and impart it to learners via functional teaching algorithms RFT and GFT. Fig. 8 is the contour version of Fig. 5.



(a) GFT-1 (b) Pool GFT-1

(c) GFT-0.05 (d) GFT-0.5

(e) RFT-0.05 (f) Whole

Figure 8. Nonparametric teaching for imparting the cheetah. f of GFT become clear significantly faster than RFT. Part of f are not updated for pool-based teaching, so several dark discontinuity points can be found. Moreover, (d)-(f) presents a smooth performance, which indicates that pack teaching can effectively smooth the gradient for the convergence to the target model.

D. Experiment Extensions

Teaching parametric target models from nonparametric initialization with GFT. We further test parametric adaptation of GFT. Specifically, we let the target model is identified by the parameter but remain teaching function directly instead of its parameter to see the performance of GFT. Here, we assume the loss function of the learner is square loss $\mathcal{L} = (y - f(\boldsymbol{x}))^2$. In two 2D cases, we set $f^*(x) = x + 1$ and $f^*(x) = -x + 1$, respectively when both $f^0(x) = \exp(-\left(\frac{x-0.5}{0.5}\right)^2) + \exp(-\left(\frac{x+0.5}{0.5}\right)^2)$. The learning rate $\eta = 0.01$ and pixels are generated by `arange(-1, 1, 0.1)`. Besides, the teacher will stop if $\mathcal{M}(f^t, f^*) < 0.0001$. We see that in Fig. 7 (a)-(b) even the target model is a straight line while the initial one is a curve, GFT also can straighten f^t and cover f^* approximately. In two 3D cases, we let both $f^*(x_1, x_2) = x_1 + x_2 - 8$, and let $f^0 = -(x_1 - 5)^2 - (x_2 - 5)^2$ and $f^0 = (x_1 - 5)^2 + (x_2 - 5)^2$, respectively. The learning rate $\eta = 0.01$ and x_1, x_2 pixels are generated by `arange(0, 10, 1)`. We observe that in Fig. 7 (c)-(d) when the target model is identified by the vector $(1, 1, -8)^T$, GFT is also able to teach curved surfaces towards this plane. To summarize, the functional teaching algorithm GFT is well-adapted for parameterized target models and GFT could teach the target function beyond its parameter.

The comparison between nonparametric and parametric teaching under parameterized initialization. We set $\eta^t = 0.01$, $\mathcal{L} = (y - f(\boldsymbol{x}))^2$ and $f^* = \langle w^*, \boldsymbol{x}\rangle = \langle (1, 1)^T, (x, 1)^T\rangle = x + 1$ and $f^0 = \langle w^0, \boldsymbol{x}\rangle = \langle (-0.5, 0.5)^T, (x, 1)^T\rangle = -0.5x + 0.5$. For nonparametric teaching, except for RBF kernel defined before, we introduce a Linear kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}'\rangle + 1$. In each iteration, we let GFT-1 select a teaching example and learners evolve f^t based on RBF and Linear kernels respectively. For parametric teaching, we let learners use parameter gradient descent:

$$w^{t+1} \leftarrow w^t - \eta^t \mathcal{G}(\mathcal{L}; w^t; (\boldsymbol{x}^t, y^t)). \tag{62}$$

For fairness, the provided teaching examples are the same as that of nonparametric teaching derived by GFT-1. We observe that f^t in both nonparametric and parametric teaching converge fast. Interestingly shown in Fig. 9, we find that nonparametric teaching with Linear kernel has same results as parametric teaching in every iteration. This is under expectation since the influence of functional gradient under the Linear kernel in each iteration is just modifying w^t from the parameterized viewpoint. *This means parametric teaching could be viewed as a particular case of nonparametric teaching when kernel is a Linear one $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x}, \boldsymbol{x}'\rangle + c$, where c is a constant.*

The sketch for missing person report. Consider a practical and interesting scenario that associates wish to file a missing person report at a police station without a photograph. The police considered as the learner would randomly provide a initial photograph, then associates (the teacher) can update the initial photograph based on their impressions in mind, which is precisely a teaching process.

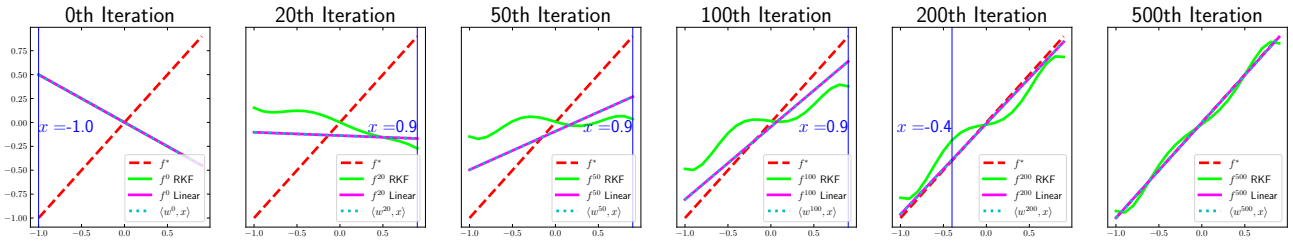


Figure 9. Contrast nonparametric teaching with the RBF and Linar kernels against parametric teaching. For fairness, the fed teaching examples are all from GFT. The red dashed lines are f^*. Nonparametric teaching: the solid lime lines are f^t with RBF kernels and the solid magenta lines are f^t with Linear kernels. Parametric teaching: The dotted lines are $f^t = \langle w^t, \boldsymbol{x} \rangle$. f^t in all settings converge fast. Interestingly, nonparametric teaching with Linear kernel has same performance as parametric teaching in each round. This is reasonable because the contribution of functional gradient under the Linear kernel is just updating w^t from the parameterized viewpoint. It concludes that nonparametric teaching is more general than parametric teaching.

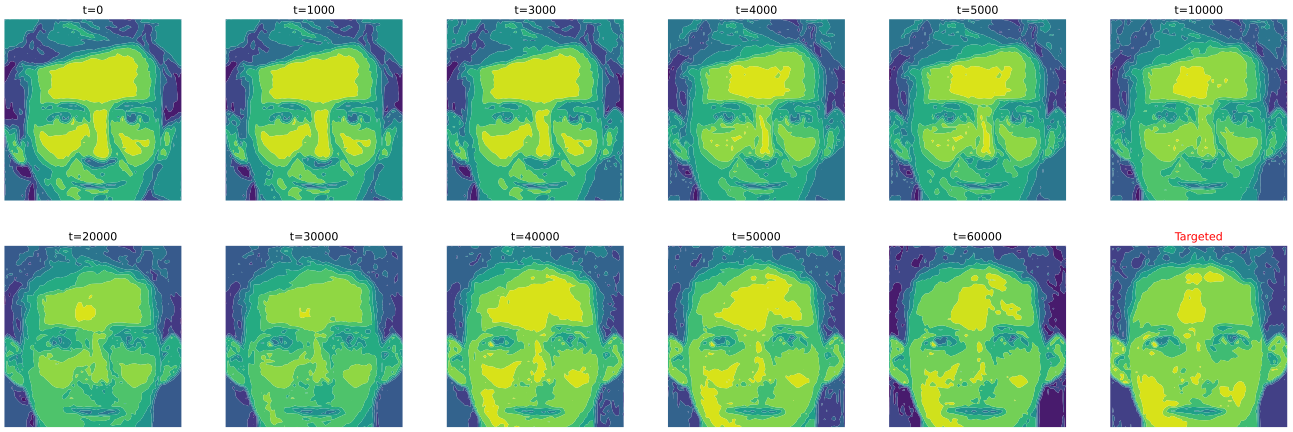


Figure 10. GFT for facial teaching. The left top one is the initial photograph and the right bottom one is the target. Viewing the facial photograph as the function, GFT works well.

GFT is also applicable in above task as a smooth solution. Smooth means f^t is modified gradually instead of replacing. To handle above nonparametric teaching problems, one can view the human face in the photograph as a general function, and GFT would modify the initial one, *i.e.*, random initialization from police, towards the targeted one (the image of the missing person in associates' minds), which is shown in Fig. 10. Specifically, we pick two facial figures form the ORL database (http://www.cam-orl.co.uk), then we set one as initialization and the other as target. The learning rate η^t is fixed as 0.05. $\mathcal{L} = (y - f(\boldsymbol{x}))^2$ is unknown for the teacher. We see that even for the complicated facial figure, our GFT presents expected performance.