Appl. Math. J. Chinese Univ.
2026, 41(2): 480-508

Online outcome weighted learning with varying Gaussians and non-identical distributions

YANG Ao-li[1,2]   FAN Jun[3]   XIANG Dao-hong[1,2,*]

Abstract. Outcome weighted learning is one of the most promising data-driven methods developed for estimating optimal individualized treatment rules (ITR) in precision medicine by reweighting outcomes based on observed data. This paper proposes a fully regularized online outcome weighted learning (ROOWL) algorithm to tackle sequential, independent but non-identically distributed data in precision medicine. We consider a time-varying Gaussian kernel to enhance flexibility in dynamic environments, and a time-varying regularization parameter to better adapt to evolving data. Its generalization ability evaluated by the excess value function is studied for commonly used loss functions, including hinge loss, generalized DWD loss, least square loss, and q-norm SVM loss. Fast learning rates are derived under smoothness or geometric noise conditions on the target function for a broad class of loss functions.

§1 Introduction

The growing interest in precision medicine stems from the significant heterogeneity in patient responses to treatments [11,21]. The primary goal of precision medicine is to determine the optimal individualized treatment rule (ITR) [18–20,26,33,35] based on individual patient traits such as demographics, clinical results, and genetic data, to maximize each patient's expected clinical outcome (referred to as value function). Many data-driven approaches for estimating ITR have been developed, with outcome weighted learning (OWL) being a notable approach proposed by [33]. They demonstrated that maximizing the individual clinical outcome is equivalent to minimizing a weighted classification error, using clinical outcomes as the classification weights. In [33], the hinge loss function is introduced as a surrogate for the 0-1 loss function [19]

Received: 2025-08-29.   Revised: 2026-02-12.
MR Subject Classification: 68Q32, 68W40, 68W27.
Keywords: online learning, outcome weighted learning, varying Gaussian kernels, sampling with non-identical distributions, convergence rates.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-026-5531-7.
The work by FAN Jun is partially supported by the Research Grants Council of Hong Kong (HKBU 12302923 and 12303024), Guangdong Basic and Applied Basic Research Fund (2024A1515011878), and Hong Kong Baptist University (RC-FNRA-IG/22-23/SCI/02). The work by XIANG Dao-hong is supported by the National Natural Science Foundation of China (12471487, 62536006).
*Corresponding author.

to make optimization feasible. The error analysis for OWL with general loss functions is provided in [36]. Additionally, [10] examined the batch OWL algorithm within the context of the Gaussian kernel and l_1-norm regularization. The common assumption of the aforementioned studies is that the sample is independent and identically distributed. However, medical data often show distributional differences due to variations in factors such as age, disease type, and severity. Collecting large and homogeneous datasets is both challenging and costly.

In this paper, we incorporate non-identically distributed sample from multiple sources, including genomic data, clinical records, and imaging, into an online version of OWL. This approach aims to enhance the understanding of disease mechanisms and treatment effects while efficiently handle large-scale and stream medical datasets. Online learning [4, 12, 30] is a powerful approach for large-scale and streaming data. A key implementation of this paradigm is online gradient descent (OGD), which efficiently updates model parameters incrementally using single data points. This method provides notable computational savings over batch approaches [5, 6, 9, 17]. Beyond standard supervised learning, OGD has been effectively applied to pairwise learning such as ranking and metric learning in large-scale settings [28, 31]. In information-theoretic learning, OGD supports model development grounded in information principles [25]. Functional data analysis benefits from OGD in capturing complex patterns in signals and biological data [3,8]. In robust learning, OGD enhances resilience to noise and outliers [7]. Additionally, it advances tensor learning and inference by enabling efficient updates on tensor-structured data [2]. Recently, the authors in [27] studied two online OWL algorithms and established the convergence theory under an i.i.d. setting.

Our contributions are threefold. First, we extend the regularized online outcome weighted learning (ROOWL) algorithm [27] via two key modifications: a time-varying Gaussian kernel (with adaptive bandwidth $\sigma_t > 0$ per iteration) to boost flexibility in dynamic environments, and a time-varying regularization parameter $\lambda(t)$ (replacing the fixed λ) to better fit evolving data. These enable online adaptation to new data, suiting large-scale or streaming precision medicine applications. Second, to handle distributional heterogeneity, we relax i.i.d. assumption of [27] by allowing samples from a time-varying distribution $P^{(t)}$, and establish convergence rates for ROOWL across loss functions including hinge loss, generalized DWD loss, least square loss, and q-norm SVM loss. Finally, we improve the convergence theory of existing online algorithms [13,14]: we derive faster rates for ROOWL under non-smooth target functions (e.g., hinge loss, generalized DWD loss) and, for the first time, lay a theoretical foundation for online learning with a varying Gaussian kernel in this setting.

The structure of this paper is organized as follows. Section 2 elaborates on the mathematical framework underlying the fully ROOWL algorithm, which is built upon varying Gaussian kernels. In Section 3, we formulate the key assumptions, present the primary theoretical results, and provide upper bounds for excess value function under a variety of loss functions. A detailed error analysis of the fully ROOWL algorithm in the scenario of non-identically distributed data is conducted in Section 4. Section 5 concludes the paper and outlines potential directions for future research.

§2 Methodology

We are considering a randomized trial for binary treatment. Let the patient's prognostic variables be denoted as $X \in \mathcal{X} \subset \mathbb{R}^n$. We assume that the assignment of treatment, $A \in \mathcal{A} = \{+1, -1\}$, is independent of X. The clinical outcome, $R \in \mathcal{R} \subseteq \mathbb{R}^+$ serves as the corresponding reward, quantifying the treatment's effectiveness, such as clinical response, disease progression, or survival rates. For simplicity, we assume that $0 < R \leq M$, with larger values of R being more desirable. An ITR $\mathcal{D}$ is a decision rule that maps the patient's covariate space $\mathcal{X}$ into the treatment space $\mathcal{A}$. Let P be the distribution of $\mathcal{Z} = (\mathcal{X}, \mathcal{A}, \mathcal{R})$, and accordingly, let $\mathbb{E}$ stand for the expectation with respect to P. The value function [27,33] $\mathcal{V}(\mathcal{D})$ is employed to evaluate the prediction performance of $\mathcal{D}$, which is defined as follows:

$$\mathcal{V}(\mathcal{D}) = \mathbb{E}\left[\frac{I(A = \mathcal{D}(X))}{A\varrho + (1-A)/2} R\right],$$

where $I(\cdot)$ is the indicator function, and $\varrho = P(A = 1) > 0$. Then the optimal ITR is defined as

$$\mathcal{D}^* = \arg\max_{\mathcal{D}} \mathbb{E}\left[\frac{I(A = \mathcal{D}(X))}{A\varrho + (1-A)/2} R\right].$$

As shown in [27,33], we can write $\mathcal{D}^*$ as $\mathcal{D}^* = \text{sign}(f^*(x))$ with

$$f^*(x) = \mathbb{E}[R|X = x, A = 1] - \mathbb{E}[R|X = x, A = -1].$$

The idea behind OWL proposed in [33] is to transform the ITR model into a weighted classification problem, more precisely, the value function $\mathcal{V}(\mathcal{D})$ can be rewritten as $\mathbb{E}[R|A = 1] + \mathbb{E}[R|A = -1] - \mathcal{R}(\mathcal{D})$, where

$$\mathcal{R}(\mathcal{D}) = \mathbb{E}\left[\frac{I(A \neq \mathcal{D}(X))}{A\varrho + (1-A)/2} R\right] \tag{1}$$

is regarded as a weighted misclassification error. Therefore, $\mathcal{D}^* = \arg\max_{\mathcal{D}} \mathcal{V}(\mathcal{D}) = \arg\min_{\mathcal{D}} \mathcal{R}(\mathcal{D})$.

Since ITR $\mathcal{D}$ is usually induced by $\mathcal{D}(x) = \text{sign}(f(x))$ for a real-valued function $f : \mathcal{X} \to \mathbb{R}$, then we define

$$\mathcal{R}(f) = \mathbb{E}\left[\frac{R}{A\varrho + (1-A)/2} I(Af(X) < 0)\right].$$

The original OWL algorithm in [33] uses the convex hinge loss function $\phi_h(af(x)) = (1-af(x))_+$ to substitute the 0-1 loss function $\phi_{0-1}(af(x)) = I(af(x) < 0)$. In this paper, we consider the fully ROOWL algorithm with general convex loss function and varying Gaussian kernels in an independent but non-identically distributed setting.

Definition 1. *A function $\phi : \mathbb{R} \to \mathbb{R}_+$ is called an admissible loss function if it is convex and differentiable at 0, with the condition that $\phi'(0) < 0$.*

Examples of the admissible loss functions include the hinge loss function $\phi_h(u) = \max\{1 - u, 0\} = (1-u)_+$, the least square loss function $\phi_{ls}(u) = (1-u)^2$, the q-norm SVM loss function $\phi_q(u) = (\phi_h(u))^q, 1 < q < \infty$, and the generalized DWD loss function

$$\phi_{dwd}(u) = \begin{cases} 1 - u, & \text{if } u \leqslant \frac{d}{d+1}, \\ \frac{1}{u^d} \frac{d^d}{(d+1)^{d+1}}, & \text{if } u > \frac{d}{d+1}, \end{cases} \tag{2}$$

where $0 < d < \infty$.

Gaussian kernel is widely used in machine learning due to its smoothness and universal approximation capabilities. It is a function on $\mathcal{X} \times \mathcal{X}$ with variance $\sigma > 0$ given by

$$K_\sigma(x,v) = \exp\left\{-\frac{|x-v|^2}{2\sigma^2}\right\}.$$

It defines [1] a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\sigma$ with the inner product $\langle\cdot,\cdot\rangle_\sigma$ given by $\langle K_\sigma(x,\cdot), K_\sigma(v,\cdot)\rangle_\sigma = K_\sigma(x,v)$. Its reproducing property plays a special role in learning theory

$$\langle K_\sigma(x,\cdot), f\rangle_\sigma = f(x), \quad \forall x \in \mathcal{X},\ f \in \mathcal{H}_\sigma. \tag{3}$$

Let $C(\mathcal{X})$ be the space of continuous functions on $\mathcal{X}$ with the norm $\|\cdot\|_{C(\mathcal{X})}$. Then it follows from (3) that

$$\|f\|_{C(\mathcal{X})} \le \|f\|_\sigma, \quad \forall x \in \mathcal{X}, f \in \mathcal{H}_\sigma. \tag{4}$$

A key factor influencing the structure of the RKHS $\mathcal{H}_\sigma$ is the variance σ of the Gaussian kernel. The value of σ significantly affects the expressive power of the kernel, as it controls the smoothness and locality of the resulting functions. Smaller values of σ lead to kernels that are more localized, allowing the RKHS $\mathcal{H}_\sigma$ to capture finer data variations, thus enhancing the approximation capability of the learning algorithm. However, a small σ may cause overfitting, while a large σ may lead to underfitting, highlighting the importance of appropriately tuning the variance parameter in practical applications.

Assume that $\mathbf{z} = \{(x_i, a_i, r_i)\}_{i=1}^T \subseteq \mathcal{Z}^T$ is a sample drawn independently from P. The following batch OWL algorithm generates an ITR $\operatorname{sign}(f_{\mathbf{z},\lambda,\sigma})$ by implementing an off-line regularization framework in $\mathcal{H}_\sigma$ involving the general convex loss ϕ

$$f_{\mathbf{z},\lambda,\sigma} = \underset{f\in\mathcal{H}_\sigma}{\arg\min}\left\{\frac{1}{T}\sum_{t=1}^{T}\frac{r_t}{a_t\varrho + (1-a_t)/2}\phi\left(a_t f\left(x_t\right)\right) + \frac{\lambda}{2}\|f\|_\sigma^2\right\}, \tag{5}$$

where $\lambda > 0$ is the regularization parameter. While the off-line algorithm (5) demonstrates strong performance both in theory and in various applications, it may encounter practical challenges when dealing with a large sample size T or extensively large data. For example, if $\phi(u) = (1-u)_+ = \max\{1-u, 0\}$ or $(1-u)_+^2$, the scheme (5) becomes a quadratic optimization problem, which has a standard complexity about $O\left(T^3\right)$. To handle scenarios with large sample size, online learning algorithms with linear complexity $O(T)$ can be utilized and provide efficient ITR.

In this paper, we study a family of online OWL algorithms associated with an admissible loss function and varying Gaussian kernels.

Definition 2. *The fully regularized online outcome weighted learning (ROOWL) algorithm with varying Gaussian kernels is defined by* $f_1 = 0$ *and*

$$f_{t+1} = f_t - \eta_t\left\{\frac{r_t}{a_t\varrho + (1-a_t)/2}\phi'_-\left(a_t f_t\left(x_t\right)\right)a_t K_{\sigma_t}\left(x_t,\cdot\right) + \lambda_t f_t\right\}, \quad \text{for } t = 1,\cdots,T, \tag{6}$$

where $\eta_t > 0$ *is the step size,* $\lambda_t > 0$ *the regularization parameter and* $\sigma_t > 0$ *the variance of the Gaussian kernels.*

The ITR is given by $\operatorname{sign}(f_{T+1})$. Its performance is measured by the excess value function

$\mathcal{V}(\mathcal{D}^*)-\mathcal{V}(\mathrm{sign}(f_{T+1}))$. In this fully ROOWL algorithm, the regularization parameter λ_t and the variance σ_t of the Gaussian kernels change with learning step t. We are interested in studying the role of the regularization parameter and the variance of Gaussian kernels in the fully ROOWL algorithm. The main purpose of this paper is to estimate the excess value function $\mathcal{V}(\mathcal{D}^*) - \mathcal{V}(\mathrm{sign}(f_{T+1}))$ as $T \to \infty$. Convergence rates will be derived under the choice of the parameters

$$\lambda_t = \lambda_1 t^{-\gamma}, \quad \sigma_t = t^{-\beta}, \quad \eta_t = \eta_1 t^{-\alpha} \tag{7}$$

for $\gamma, \beta, \alpha > 0$ and conditions on the distribution P and the loss function ϕ. For a general Mercer kernel K and a fixed λ, the error analysis for ROOWL was investigated in [27].

§3 Main results

In this section, we start with the introduction of some assumptions that will enable us to present the main results.

3.1 Assumptions and comparison theorems

In contrast to an i.i.d. setting, this paper assumes that the sequence of probability distributions $\{P^{(t)}\}_{t=1,2,\cdots}$ on $\mathcal{Z}$ vary at each step. Specifically, each probability distribution $P^{(t)}$ produces a sample point $z_t = (x_t, a_t, r_t)$ independently. Denote by $C^\zeta(\mathcal{X})$ the Hölder space defined as the set of all continuous functions on $\mathcal{X}$ with a finite norm given by $\|f\|_{C^\zeta(\mathcal{X})} = \|f\|_{C(\mathcal{X})} + |f|_{C^\zeta(\mathcal{X})}$, where $|f|_{C^\zeta(\mathcal{X})} := \sup_{x\neq y} \frac{|f(x)-f(y)|}{(d(x,y))^\zeta}$. To conduct the error analysis, we assume that the sequence of marginal distributions $\{P_{\mathcal{X}}^{(t)}\}_{t=1,2,\cdots}$ on $\mathcal{X}$ converges polynomially in the dual of $C^\zeta(\mathcal{X})$ for some $0 < \zeta \leq 1$ as below. Let $P_{\mathcal{X}}$ be the marginal distribution on $\mathcal{X}$, and let $P_{\mathcal{X}}^{(t)}$ be the empirical distribution of X at time step t.

Assumption 1. *We assume that the sequence $\{P_{\mathcal{X}}^{(t)}\}_{t=1,2,\cdots}$ converges polynomially to a probability distribution $P_{\mathcal{X}}$ in $\left(C^\zeta(\mathcal{X})\right)^*$ for $0 \leq \zeta \leq 1$, that is, there exist $C > 0$ and $b > 0$ such that*

$$\|P_{\mathcal{X}}^{(t)} - P_{\mathcal{X}}\|_{(C^\zeta(\mathcal{X}))^*} \leq C t^{-b}, \quad t \in \mathbb{N}. \tag{8}$$

By the definition of the dual space $\left(C^\zeta(\mathcal{X})\right)^*$, the decay condition (8) can be reformulated as

$$\left| \int_{\mathcal{X}} f(x) dP_{\mathcal{X}}^{(t)} - \int_{\mathcal{X}} f(x) dP_{\mathcal{X}} \right| \leq C t^{-b} \|f\|_{C^\zeta(\mathcal{X})}, \quad \forall f \in C^\zeta(\mathcal{X}), t \in \mathbb{N}. \tag{9}$$

The power index b quantitatively highlights the distinctions between our non-identical framework and the i.i.d. setting, where the latter is characterized by $b = \infty$.

Another assumption for distributions is as follows, which will be used in the error analysis of algorithm (6) under the sample drawn from non-identical distribution setting.

Assumption 2. *Assume that the conditional distributions $\{P_x(a,r) = P(A, R|X = x) : x \in \mathcal{X}\}$ is Lipschitz ζ in $\left(C^\zeta(\mathcal{A} \times \mathcal{R})\right)^*$. Specifically, there exists a constant $C_P \geq 0$ such that*

$$\|P_x(a,r) - P_u(a,r)\|_{(C^\zeta(\mathcal{A}\times\mathcal{R}))^*} \leq C_P |x - u|^\zeta, \quad \forall x, u \in \mathcal{X}. \tag{10}$$

This assumption imposes a mild regularity condition on the joint distribution of actions and rewards: it requires that the conditional distribution $P(A, R \mid X = x)$ varies smoothly with respect to the patient's prognostic variables X. The Lipschitz continuity in the dual of $C^{\varsigma}(\mathcal{A}\times\mathcal{R})$ ensures stability under small changes in X, which is essential for deriving convergence rates in individualized treatment rule learning.

The following assumption regarding the loss function ϕ sets out two conditions: a growth condition for its left derivative ϕ'_-, and that it is locally Lipschitz at the origin. These conditions are essential for analyzing the error of algorithm (6).

Assumption 3. *Assume that the admissible loss function ϕ satisfies the following increment condition: for some $p \geq 0, N_1 > 0$,*

$$N(w) := \sup\left\{\left|\phi'_-(u)\right| : |u| \leq w\right\} \leq N_1 w^p, \forall w \geq 1, \tag{11}$$

and

$$\tilde{N}(w) := \sup\{\|\phi(u)\|_{C^{\varsigma}(\mathcal{A})} : |u| \leq w\} \leq N_1 w^{p+1}, \forall w \geq 1. \tag{12}$$

Additionally, assume that ϕ'_- is locally Lipschitz at the origin with

$$M_0 := \sup\left\{\frac{\left|\phi'_-(u) - \phi'_-(0)\right|}{|u|}, |u| \leq 1\right\} < \infty. \tag{13}$$

Define $M(w) := \sup\left\{\frac{\left|\phi'_-(u) - \phi'(0)\right|}{|u|}, |u| \leq |w|\right\}$, it follows from [14] that

$$M(w) \leq M_0 + 2N(w) \leq (M_0 + 2N_1)\, |w|^p, \quad \forall |w| \geq 1. \tag{14}$$

This assumption can be easily verified for several commonly used loss functions [15], including the hinge loss function with $p = 0$, the least square loss function with $p = 1$, the q-norm SVM loss function with $p = q - 1$, and the generalized DWD loss function with $p = 0$.

Define the weighted generalization error for the admissible loss function ϕ as

$$\mathcal{E}(f) = \mathbb{E}\Big[\frac{R}{A\varrho + (1 - A)/2}\phi(Af(X))\Big] = \int_{\mathcal{Z}} \frac{r}{a\varrho + (1 - a)/2}\phi(af(x))dP, \tag{15}$$

and f_P^{ϕ} is a minimizer of $\mathcal{E}(f)$ over all measurable functions.

The comparison theorems in the framework of ITR show that the excess value function $\mathcal{V}(\mathcal{D}^*) - \mathcal{V}(\mathcal{D})$ can be bounded by the excess weighted generalization error $\mathcal{E}(f) - \mathcal{E}(f_P^{\phi})$. For the hinge loss function $\phi(u) = (1 - u)_+$, Theorem 3.2 in [33] together with Lemma 3.2 in [10] showed that

$$\mathcal{V}(\mathcal{D}^*) - \mathcal{V}(\mathrm{sign}(f)) \leq \mathcal{E}(f) - \mathcal{E}(\mathcal{D}^*). \tag{16}$$

For the generalized DWD loss function, the authors in [27] proved that

$$\mathcal{V}(\mathcal{D}^*) - \mathcal{V}(\mathrm{sign}(f)) \leq \frac{d+1}{d}\left[\mathcal{E}(f) - \mathcal{E}(f_P^{\phi_{dwd}})\right]. \tag{17}$$

For more general loss functions, the following noise condition is required to derive a tight comparison theorem.

Assumption 4. *Recall $f^*(x) = \mathbb{E}[R|X = x, A = 1] - \mathbb{E}[R|X = x, A = -1]$. Assume that the distribution P satisfies the Tsybakov noise condition with index $\tau \in [0, +\infty]$, that is, there exists a constant $C_\tau > 0$ such that for any $t > 0$,*

$$P_{\mathcal{X}}\left(\{x \in \mathcal{X} : |f^*(x)| \leq C_\tau t\}\right) \leq t^{\tau}. \tag{18}$$

Under the Tsybakov noise condition, Theorem 3.3 in [36], along with Proposition 1 in [27] established the comparison theorem for an admissible loss function with $\phi''(0) > 0$, as stated below

$$\mathcal{V}(\mathcal{D}^*) - \mathcal{V}(\mathrm{sign}(f)) \leq C_1^{\phi}\left[\mathcal{E}(f) - \mathcal{E}(f_P^{\phi})\right]^{\frac{\tau+1}{\tau+2}}, \tag{19}$$

where C_1^{ϕ} is a constant.

3.2 Learning rates

Since the RKHS $\mathcal{H}_\sigma$ is used as a hypothesis space for ROOWL, the learning ability of Algorithm (6) depends on the approximation power of the space $\mathcal{H}_\sigma$ with respect to the target function f_P^{ϕ}. We can quantify the approximation power by the following regularization error.

Definition 3. *Let $f_\lambda^\sigma \in \mathcal{H}_\sigma$ with $\lambda > 0$. The regularization error with (ϕ, P) in $\mathcal{H}_\sigma$ is defined by*

$$\mathcal{D}(\sigma, \lambda) = \mathcal{E}(f_\lambda^\sigma) - \mathcal{E}(f_P^{\phi}) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2. \tag{20}$$

There are different choices for the function $f_\lambda^\sigma \in \mathcal{H}_\sigma$. We usually take it as

$$f_\lambda^\sigma := \arg\min_{f\in\mathcal{H}_\sigma}\left\{\mathcal{E}(f) - \mathcal{E}(f_P^{\phi}) + \frac{\lambda}{2}\|f\|_{\mathcal{H}_\sigma}^2\right\}. \tag{21}$$

Thus, the error decomposition for (6) can now be conducted as follows:

$$\begin{aligned}\mathcal{E}(f_{T+1}) - \mathcal{E}(f_P^{\phi}) &= \left\{\mathcal{E}(f_{T+1}) - \mathcal{E}(f_{\lambda_T}^{\sigma_T})\right\} + \left\{\mathcal{E}(f_{\lambda_T}^{\sigma_T}) - \mathcal{E}(f_P^{\phi})\right\} \\ &\leq \left\{\mathcal{E}(f_{T+1}) - \mathcal{E}(f_{\lambda_T}^{\sigma_T})\right\} + \mathcal{D}(\sigma_T, \lambda_T).\end{aligned} \tag{22}$$

In (22), the regularization error term $\mathcal{D}(\sigma_T, \lambda_T)$ is independent of the sample $\mathbf{z} = \{z_t\}_{t=1}^T$, which will be bounded in Lemma 3 and Lemma 5. The term $\mathcal{E}(f_{T+1}) - \mathcal{E}(f_{\lambda_T}^{\sigma_T})$ is referred to as the sample error which can be controlled by $\left\|f_{T+1} - f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}$.

In the existing studies [13, 14], the convergence analysis of online learning algorithms typically relies on the assumption that the target function satisfies specific regularity conditions, including smoothness within Lipschitz or Sobolev spaces. This assumption is particularly common when utilizing loss functions such as the least square loss or the q-norm SVM loss. In our framework, we further extend the scope to consider the cases where the target function lies in a generalized Lipschitz space [24], which is characterized by the following ℓ-th order modulus of continuity.

Definition 4. *For any Lebesgue measurable function $f : \mathbb{R}^n \to \mathbb{R}$ with $f \in L^v(\mathbb{R}^n)$ for some $0 < v \leq \infty$, its ℓ-th modulus of continuity in L^v-norm is defined as*

$$\omega_v^\ell(f, h) := \sup_{|u|\leq h}\left\|\left|\Delta_u^\ell f(\cdot)\right|\right\|_{L^v(\mathbb{R}^n)}, \ell \in \mathbb{N}, h > 0,$$

where $|u|$ denotes the Euclidean norm of $u \in \mathbb{R}^n$ and the ℓ-th divided difference $\Delta_u^\ell f(\cdot)$ is given by

$$\Delta_u^\ell f(x) := \sum_{i=0}^{\ell}\binom{\ell}{i}(-1)^{\ell-i}f(x+iu), \quad \forall x \in \mathbb{R}^n.$$

Definition 5. *For $0 < s < \ell$, $\ell = \lfloor s \rfloor + 1$ and $1 \le v < \infty$, the generalized Lipschitz space* $\mathrm{Lip}^*(s, L^v(\mathbb{R}^n))$ *is defined as*

$$\mathrm{Lip}^*(s, L^v(\mathbb{R}^n)) := \left\{ f \in L^v(\mathbb{R}^n), |f|_{Lip^*(s,L^v(\mathbb{R}^n))} < \infty \right\}$$

with the semi-norm

$$|f|_{Lip^*(s,L^v(\mathbb{R}^n))} := \sup_{h>0} h^{-s} \omega_v^\ell(f, h)$$

and the norm $\|f\|_{Lip^*(s,L^v(\mathbb{R}^n))} = \|f\|_{L^v(\mathbb{R}^n)} + |f|_{Lip^*(s,L^v(\mathbb{R}^n))}$.

As shown in [36], the target functions corresponding to the least squares loss and the q-norm SVM loss are given by

$$f_P^{\phi_{ls}}(x) = \frac{\mathbb{E}[R \mid X = x, A = 1] - \mathbb{E}[R \mid X = x, A = -1]}{\mathbb{E}[R \mid X = x, A = 1] + \mathbb{E}[R \mid X = x, A = -1]}$$

and

$$f_P^{\phi_q}(x) = \frac{(\mathbb{E}[R \mid X = x, A = 1])^{1/(q-1)} - (\mathbb{E}[R \mid X = x, A = -1])^{1/(q-1)}}{(\mathbb{E}[R \mid X = x, A = 1])^{1/(q-1)} + (\mathbb{E}[R \mid X = x, A = -1])^{1/(q-1)}},$$

respectively. When $q = 2$, and $f_P^{\phi_2} = f_P^{\phi_{ls}}$.

Let us demonstrate our main results by stating learning rates for the least squares loss $\phi_{ls}(u) = (1-u)^2$ and the 2-norm SVM loss $\phi_2(u) = (1-u)_+^2$. Despite both functions being smooth and convex, their effectiveness in our analysis stems from a nice structural property: the excess ϕ_2-risk can be tightly bounded by the L^2 distance between the corresponding least squares solutions. Specifically, we have $\mathcal{E}(f) - \mathcal{E}(f_P^{\phi_{ls}}) \le \widetilde{M} \|f - f_P^{\phi_{ls}}\|_{L^2(\mathcal{X})}^2$, which allows us to derive sharper convergence rates. This relationship plays a crucial role in the subsequent theoretical guarantees. To simplify notation, we denote by $\mathbb{E}_{\mathcal{Z}_t}$ the expectation over the sample set $\{z_1, \ldots, z_t\}$.

Theorem 1. *Let $\widetilde{M} = \frac{M}{\min\{\varrho, 1-\varrho\}}$. For $\phi_{ls}(u) = (1-u)^2$ and $\phi_2(u) = (1-u)_+^2$, define $\{f_t\}_{t=1}^{T+1}$ by (6). Let Assumptions 1, 2 and 3 be satisfied. Assume that for $s > 0$, $\tilde{f}_P^{\phi_{ls}} \in \mathrm{Lip}^*\left(s, L^2(\mathbb{R}^n)\right) \cap L^\infty(\mathbb{R}^n)$ and $f_P^{\phi_{ls}} = \tilde{f}_P^{\phi_{ls}} |_{\mathcal{X}}$. Choose the triple $\{\lambda_t, \sigma_t, \eta_t\}$ as (7) with $0 < \eta_1 \le 1/(8\widetilde{M})$, $\lambda_1 = 2\widetilde{M}$ and $\beta = c_1\gamma$ where $c_1 = 1/(2s+n)$. If*

$$0 < \gamma < 1/(4 + 3nc_1/2),$$
$$\gamma(3 + nc_1) < \alpha < 1 - (nc_1/2 + 1)\gamma,$$

and

$$b > (nc_1 + \zeta c_1 + 3)\gamma,$$

then we have

$$\mathbb{E}_{\mathcal{Z}_T}[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_P^{\phi_{ls}})] = O\left(T^{-\min\{\theta_1, 2sc_1\gamma\}}\right),$$

where

$$\theta_1 := \min\{2 - 2\alpha - (nc_1 + 2)\gamma, \alpha - (3 + nc_1)\gamma, b - (nc_1 + \zeta c_1 + 3)\gamma\}. \tag{23}$$

By combining Theorem 1 with (19), we can derive an explicit learning rate for the least squares loss and the 2-norm SVM loss under specific parameter choices.

Corollary 1. *For $\phi_{ls}(u) = (1-u)^2$ and $\phi_2(u) = (1-u)_+^2$, all assumptions of Theorem 2 and Assumption 4 are satisfied. Choosing $\alpha = \frac{8n+16s}{11n+22s}$, $\beta = \frac{2}{11n+22s}$, $\gamma = \frac{2n+4s}{11n+22s}$, and*

$b > \frac{8n+12s+2\zeta}{11n+22s}$, *then we have*

$$\mathbb{E}_{\mathcal{Z}_T}\left[\mathcal{V}\left(\mathcal{D}^*\right) - \mathcal{V}(\mathrm{sign}(f_{T+1}))\right] = O\left(T^{-\frac{\tau+1}{\tau+2}\cdot\min\left\{\frac{4s}{11n+22s}, b-\frac{8n+12s+2\zeta}{11n+22s}\right\}}\right).$$

Remark 1. *For the least squares and 2-norm SVM loss, Corollary 1 gets a better convergence rate of* $O\left(T^{-\frac{4(\tau+1)s}{(\tau+2)(11n+22s)}}\right)$ *compared to the prior rate of* $O\left(T^{-\frac{(\tau+1)s}{(\tau+2)(22n+19s)}}\right)$ *from [14] in an i.i.d. setting as* $b\to\infty$. *This improvement stems from a more precise variance error analysis, crucial for time-varying kernel online learning. Unlike the loose bound in [14], we use Lemma 2, which exploits the Gaussian kernel's convolutional structure (* $\sigma_t = t^{-\beta}$*). This enhances the decay of variancing error, thereby leading to improved convergence.*

For more general loss functions whose target functions satisfy a smoothness condition in a generalized Lipschitz space, we present the following result.

Theorem 2. *Let* $\widetilde{M} = \frac{M}{\min\{\varrho, 1-\varrho\}}$. *For an admissible loss function* ϕ, *define* $\{f_t\}_{t=1}^{T+1}$ *by* (6). *Let Assumptions 1, 2 and 3 be satisfied. Suppose that for* $s>0$, $\tilde{f}_P^\phi \in \mathrm{Lip}^*\left(s, L^v\left(\mathbb{R}^n\right)\right) \cap L^\infty\left(\mathbb{R}^n\right), f_P^\phi = \tilde{f}_P^\phi|_{\mathcal{X}}$ *and* $\frac{dP_{\mathcal{X}}}{dx} \in L^{v'}(\mathcal{X})$ *with* $\frac{1}{v'}+\frac{1}{v}=1$. *Take the triple* $\{\lambda_t, \sigma_t, \eta_t\}$ *as* (7) *with* $0<\eta_1\le 1/(\widetilde{M}\left(M_0+2N_1\right)+\lambda_1)$, $\lambda_1 \ge \widetilde{M}\left|\phi'(0)\right|$ *and* $\beta = c\gamma$ *with* $c = 1/\left(s+n\right)$. *Then if*

$$0<\gamma<1/(\max\{(5p+1)(nc/2+1)+1, (p+1)(2+3nc/2)+p\}), \tag{24}$$

$$\gamma\max\{p(2nc+4)+1, (p+1)(1+nc)+p\} < \alpha < 1-(p+1)(nc/2+1)\gamma, \tag{25}$$

and

$$b > [(3p+1)(nc/2+1)+\zeta c+1]\gamma, \tag{26}$$

we have

$$\mathbb{E}_{\mathcal{Z}_T}[\mathcal{E}\left(f_{T+1}\right) - \mathcal{E}(f_P^\phi)] = O\left(T^{-\min\left\{\frac{\theta}{2}-p\left(\frac{nc}{2}+1\right)\gamma, sc\gamma\right\}}\right),$$

where

$$\begin{aligned}\theta := \min\{&2-2\alpha-(nc+2)\gamma, \alpha-(1+nc)\gamma, \alpha-(1+npc+2p)\gamma,\\ &b-((p+1)(nc/2+1)+\zeta c+1)\gamma\}.\end{aligned} \tag{27}$$

The following corollary comes immediately from Theorem 2 and comparison theorem (19).

Corollary 2. *Let* ϕ *be an admissible loss function such that* $\phi''(0)>0$. *All assumptions of Theorem 2 and Assumption 4 are satisfied. Then we have*

$$\mathbb{E}_{\mathcal{Z}_T}\left[\mathcal{V}\left(\mathcal{D}^*\right) - \mathcal{V}(\mathrm{sign}(f_{T+1}))\right] = O\left(T^{-\frac{\tau+1}{\tau+2}\cdot\min\left\{\frac{\theta}{2}-p\left(\frac{nc}{2}+1\right)\gamma, sc\gamma\right\}}\right),$$

where θ *is defined by* (27).

Corollary 2 provides the convergent rates for convex loss with $\phi''(0)>0$, such as the q-norm SVM loss function $\phi_q(u) = (1-u)_+^q, 1<q<\infty$.

Example 1. *Let* $\phi_q(u) = (1-u)_+^q$ *with* $1<q<\infty$. *All assumptions of Theorem 2 and Assumption 4 are satisfied. If* $q \ge 2-\frac{2n+2s}{3n+2s}$, *then* $\alpha = \frac{(12q-10)n+(8q-2)s}{(15q-10)n+10qs}$, $\beta = \frac{2}{(15q-10)n+10qs}$, $\gamma = \frac{2n+2s}{(15q-10)n+10qs}$ *and* $b > \frac{(9q-4)n+(6q-2)s+2\zeta}{5(3q-2)n+10qs}$. *If* $1<q<2-\frac{2n+2s}{3n+2s}$, *then* $\alpha = \frac{(6q-2)n+(4q+2)s}{(9q-2)n+(6q+4)s}$, $\beta = \frac{2}{(9q-2)n+(6q+4)s}$, $\gamma = \frac{2n+2s}{(9q-2)n+(6q+4)s}$ *and* $b > \frac{(9q-4)n+(6q-2)s+2\zeta}{(9q-2)n+(6q+4)s}$. *In both cases, we have*

$$\mathbb{E}_{\mathcal{Z}_T}\left[\mathcal{V}\left(\mathcal{D}^*\right) - \mathcal{V}(\mathrm{sign}(f_{T+1}))\right] = O\left(T^{-\theta_q}\right),$$

where

$$\theta_q = \begin{cases} \frac{\tau+1}{\tau+2} \cdot \min\left\{\frac{2s}{(15q-10)n+10qs}, \frac{b}{2} - \frac{(9q-4)n+(6q-2)s+2\zeta}{10(3q-2)n+20qs}\right\}, & q \geq 2 - \frac{2n+2s}{3n+2s}, \\ \frac{\tau+1}{\tau+2} \cdot \min\left\{\frac{2s}{(9q-2)n+(6q+4)s}, \frac{b}{2} - \frac{(9q-4)n+(6q-2)s+2\zeta}{(18q-4)n+(12q+8)s}\right\}, & 1 < q < 2 - \frac{2n+2s}{3n+2s}. \end{cases}$$

Remark 2. *The convergence rate θ_q has two piecewise forms based on whether q exceeds the threshold $2 - \frac{2n+2s}{3n+2s}$. For high regularity ($s \to \infty$), nearly stationary data distribution ($b \to \infty$), and negligible noise ($\tau \to \infty$), the convergence rate approaches $O\left(T^{-\frac{1}{5q}}\right)$ if $q \geq 2 - \frac{2n+2s}{3n+2s}$, and $O\left(T^{-\frac{1}{3q+2}}\right)$ if $1 < q < 2 - \frac{2n+2s}{3n+2s}$. The results show how q (related to loss smoothness) slows convergence for larger q in the worst case.*

The result in Example 1 applies to general q-norm SVM loss with $1 < q < \infty$, including $q = 2$. However, under the stronger regularity $\tilde{f}_P^\phi \in \mathrm{Lip}^*(s, L^2(\mathbb{R}^n))$, Corollary 1 gives a sharper convergence rate for 2-norm SVM loss. Therefore, when L^2-based regularity holds, Corollary 1 should be preferred over the general result in Example 1.

To study loss functions whose associated target functions f_P^ϕ may be discontinuous, such as the hinge loss and the generalized DWD loss, we cannot impose the same smoothness conditions as in previous results. For instance, it is shown in [24] that $f_P^{\phi_h} = \mathcal{D}^*$, and the target function induced by the generalized DWD loss is given by

$$f_P^{\phi_{\rm dwd}}(x) = \frac{d}{d+1}\left[\left(\frac{\bar{p}}{\bar{q}}\right)^{1/(d+1)} I\left\{\bar{p} > \bar{q}\right\} - \left(\frac{\bar{q}}{\bar{p}}\right)^{1/(d+1)} I\left\{\bar{p} < \bar{q}\right\}\right],$$

where $\bar{p} = \mathbb{E}[R \mid X = x, A = 1]$ and $\bar{q} = \mathbb{E}[R \mid X = x, A = -1]$. In such cases, the target function f_P^ϕ may exhibit discontinuities, and thus does not belong to any generalized Lipschitz space. To address this, the study in [24] demonstrates that employing varying Gaussian functions can enhance learning performance across various scenarios. This improvement stems from incorporating geometric noise conditions to estimate approximation error, removing the requirement for smoothness conditions in generalized Lipschitz spaces. To do this, define a distance function $\mathcal{N}_\varepsilon^\phi(x) := \inf_{u \in \mathcal{X}}\{|x-u|, \text{ s.t. } |f_P^\phi(x) - f_P^\phi(u)| \geq \varepsilon\}$ for given $x \in \mathcal{X}$ and $\varepsilon > 0$. For example, if there exists $\varepsilon_0 > 0$ such that $|f_P^\phi(x) - f_P^\phi(u)| < \varepsilon_0$ for all $u \in \mathcal{X}$, then $\mathcal{N}_{\varepsilon_0}^\phi(x) = \infty$. Recall that $P_\mathcal{X}$ denotes the marginal distribution on $\mathcal{X}$.

Assumption 5. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a compact set. We assume that the distribution $P_\mathcal{X}$ satisfies the general geometric noise condition, that is, there exist $C_1 > 0$, $\xi_1, \xi_2 > 0$, for small enough $\varepsilon > 0$, such that*

$$\int_\mathcal{X} \exp\left\{-\frac{(\mathcal{N}_\varepsilon^\phi(x))^2}{h^2}\right\} \mathrm{d}P_\mathcal{X} \leq C_1 h^{\xi_1} \varepsilon^{-\xi_2}, \quad \forall 0 < h < 1. \tag{28}$$

This condition characterizes the complexity of the target function f_P^ϕ near its discontinuities. The function $\mathcal{N}_\varepsilon^\phi(x)$ measures the minimum distance required to move from a point x to observe a change of at least ε in the function value. If the function is smooth or nearly constant around x, this distance is large; otherwise, it is small near sharp transitions or discontinuities. The geometric noise condition (28) controls the concentration of the data distribution near such irregular regions. This enables effective approximation error estimates, even when the target function lacks smoothness, thereby improving the performance of kernel-based learning methods. Under this assumption, we obtain the following main result.

Theorem 3. *Let $\widetilde{M} = \frac{M}{\min\{\varrho, 1-\varrho\}}$. For an admissible loss function ϕ, define $\{f_t\}_{t=1}^{T+1}$ by (6). Let Assumptions 1, 2, 3, and 5 be satisfied. Suppose that $\tilde{f}_P^{\phi} \in L^{\infty}(\mathbb{R}^n), f_P^{\phi} = \tilde{f}_P^{\phi}|_{\mathcal{X}}$. Choose the triple $\{\lambda_t, \sigma_t, \eta_t\}$ as (7) with $0 < \eta_1 \leq 1/(\widetilde{M}(M_0 + 2N_1) + \lambda_1)$, $\lambda_1 \geq \widetilde{M}|\phi'(0)|$ and $\beta = c_2\gamma$ with $c_2 = \frac{1+\xi_2}{\xi_1 + n(1+\xi_2)}$. Then if*

$$0 < \gamma < 1/(\max\{(5p+1)(nc_2/2+1)+1, (p+1)(2+3nc_2/2)+p\}),$$

$$\gamma \max\{p(2nc_2+4)+1, (p+1)(1+nc_2)+p\} < \alpha < 1-(p+1)(nc_2/2+1)\gamma,$$

and

$$b > [(3p+1)(nc_2/2+1)+\zeta c_2+1]\gamma,$$

we have

$$\mathbb{E}_{\mathcal{Z}_T}[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_P^{\phi})] = O\left(T^{-\min\left\{\frac{\theta_2}{2} - p\left(\frac{nc}{2}+1\right)\gamma, \frac{\xi_1 c_2 \gamma}{1+\xi_2}\right\}}\right),$$

where

$$\theta_2 := \min\{2-2\alpha-(nc_2+2)\gamma, \alpha-(1+nc_2)\gamma, \alpha-(1+npc_2+2p)\gamma,$$
$$b-((p+1)(nc_2/2+1)+\zeta c_2+1)\gamma\}.$$

Theorem 3 together with (16) yields the following result for the hinge loss function and generalized DWD loss function.

Corollary 3. *For $\phi = \phi_h$ and $\phi = \phi_{dwd}$, suppose all assumptions of Theorem 3 hold. Assume that $\mathcal{N}_{\varepsilon}^{\phi_h}(x) \geq \bar{c} > 0$ almost surely for some constant $\bar{c} > 0$ and any $0 < \varepsilon < 1$. Choosing $\alpha = \frac{4n+6\xi_1}{7n+10\xi_1}$, $\beta = \frac{2}{7n+10\xi_1}$, $\gamma = \frac{2\xi_1+2n}{7n+10\xi_1}$, and $b > \frac{5n+4\xi_1+2\zeta}{7n+10\xi_1}$, then we have*

$$\mathbb{E}_{\mathcal{Z}_T}[\mathcal{V}(\mathcal{D}^*) - \mathcal{V}(\mathrm{sign}(f_{T+1}))] = O\left(T^{-\min\left\{\frac{2\xi_1}{7n+10\xi_1}, \frac{b}{2} - \frac{5n+4\xi_1+2\zeta}{14n+20\xi_1}\right\}}\right).$$

Remark 3. *The convergence rate in Corollary 3 depends on the minimum of two terms: one from the geometric noise exponent ξ_1, and the other from the distributional shift parameter b. A large ξ_1 implies strong geometric separation of the target function $f_P^{\phi_h}$ or $f_P^{\phi_{dwd}}$. If b is sufficiently large, the convergence rate approaches $O\left(T^{-\frac{2\xi_1}{7n+10\xi_1}}\right)$, improving with ξ_1. This highlights the effectiveness of the geometric noise condition for analyzing non-smooth target functions like hinge or generalized DWD loss.*

We summarize our main results and compare them with existing work in Table 1. This paper establishes the first convergence analysis of the ROOWL algorithm with a varying-bandwidth Gaussian kernel under independent but non-identically distributed (i.n.i.d.) data. We derive convergence rates for commonly used loss functions including the hinge loss, generalized DWD loss, least squares loss, and q-norm SVM loss. Except for the least squares loss under i.i.d. settings (see Remark 1), no prior work has analyzed the convergence of these losses in regularized online algorithms with adaptive kernels, whether for classification or OWL. In practice, data often exhibit distributional shifts, so developing theory under such conditions is essential. Our results enrich the theoretical understanding of ROOWL and provide guidance on loss function selection in dynamic, i.n.i.d. environments, especially for ITR learning.

Table 1. Comparison for learning rates.

Loss function	[14]	Our paper
ϕ_{ls}	$T^{-\frac{\tau+1}{\tau+2}\cdot\frac{s}{22n+19s}}$	$T^{-\frac{\tau+1}{\tau+2}\cdot\min\left\{\frac{4s}{11n+22s},b-\frac{8n+12s+2\zeta}{11n+22s}\right\}}$
ϕ_q, $q=2$	No	$T^{-\frac{\tau+1}{\tau+2}\cdot\min\left\{\frac{4s}{11n+22s},b-\frac{8n+12s+2\zeta}{11n+22s}\right\}}$
ϕ_q, $1<q<2-\frac{2n+2s}{3n+2s}$	No	$T^{\frac{\tau+1}{\tau+2}\cdot\min\left\{\frac{2s}{(9q-2)n+(6q+4)s},\frac{b}{2}-\frac{(9q-4)n+(6q-2)s+2\zeta}{(18q-4)n+(12q+8)s}\right\}}$
ϕ_q, $q\geq 2-\frac{2n+2s}{3n+2s}$ $q\neq 2$	No	$T^{\frac{\tau+1}{\tau+2}\cdot\min\left\{\frac{2s}{(15q-10)n+10qs},\frac{b}{2}-\frac{(9q-4)n+(6q-2)s+2\zeta}{10(3q-2)n+20qs}\right\}}$
ϕ_h	No	$T^{-\min\left\{\frac{2\xi_1}{7n+10\xi_1},\frac{b}{2}-\frac{5n+4\xi_1+2\zeta}{14n+20\xi_1}\right\}}$
ϕ_{dwd}	No	$T^{-\min\left\{\frac{2\xi_1}{7n+10\xi_1},\frac{b}{2}-\frac{5n+4\xi_1+2\zeta}{14n+20\xi_1}\right\}}$

§4 Proofs of main results

4.1 Preliminary Lemmas

We demonstrate some useful bounds, including error caused by non-identical distributions, variancing error resulting from changes in the Gaussian kernels σ_t and approximation error, which all can be used in algorithm (6) and play critical roles in proving the main results.

The key analysis for one-step iteration (Proposition 1) is to handle an extra error arising from the non-identical distributions. A similar result for classification and regression problems is presented in [15].

Lemma 1 (**Error by non-identical**). *Let $\widetilde{M}=\frac{M}{\min\{\varrho,1-\varrho\}}$ and Assumption 2 be satisfied. Then for $h,g\in C^{\zeta}(\mathcal{X})$, we have*

$$\left|\int_{\mathcal{Z}}\frac{r}{a\varrho+(1-a)/2}\left\{\phi\left(ah(x)\right)-\phi\left(ag(x)\right)\right\}d[P^{(t)}-P]\right|$$
$$\leq\left\{2MB_{h,g}(\|h\|_{C^{\zeta}(\mathcal{X})}+\|g\|_{C^{\zeta}(\mathcal{X})})+2\widetilde{M}C_P\widetilde{B}_{h,g}\right\}\|P_{\mathcal{X}}^{(t)}-P_{\mathcal{X}}\|_{(C^{\zeta}(\mathcal{X}))^*},$$

where $B_{h,g}$ and $\widetilde{B}_{h,g}$ are constants given by

$$B_{h,g}=\sup\left\{|\phi'_-(af)|:a\in\mathcal{A}=\{-1,1\},|f|\leq\max\left\{\|h\|_{C(\mathcal{X})},\|g\|_{C(\mathcal{X})}\right\}\right\},$$

and

$$\widetilde{B}_{h,g}=\sup\left\{\|\phi(af)\|_{C^{\zeta}(\mathcal{A})}:a\in\mathcal{A}=\{-1,1\},|f|\leq\max\left\{\|h\|_{C(\mathcal{X})},\|g\|_{C(\mathcal{X})}\right\}\right\}.$$

Proof. Let $P_x(a,r)=P(R,A|X=x)$. We first decompose the probability distributions on $\mathcal{Z}$ into marginal and conditional distributions

$$\int_{\mathcal{Z}}\frac{r}{a\varrho+(1-a)/2}\left\{\phi\left(ah(x)\right)-\phi\left(ag(x)\right)\right\}d[P^{(t)}-P]$$
$$=\int_{\mathcal{X}}\int_{\mathcal{A}\times\mathcal{R}}\frac{r}{a\varrho+(1-a)/2}\left[\phi(ah(x))-\phi(ag(x))\right]dP_x(a,r)d[P_{\mathcal{X}}^{(t)}-P_{\mathcal{X}}].$$

It follows from the definition of the norm in $\left(C^{\varsigma}(\mathcal{X})\right)^*$ that

$$\left|\int_{\mathcal{Z}} \frac{r}{a\varrho+(1-a)/2}\left\{\phi\left(ah(x)\right)-\phi\left(ag(x)\right)\right\} d[P^{(t)}-P]\right|$$
$$\leq \|P_{\mathcal{X}}^{(t)}-P_{\mathcal{X}}\|_{(C^{\varsigma}(\mathcal{X}))^*}\|J\|_{C^{\varsigma}(\mathcal{X})},$$

where J is a function on $\mathcal{X}$ given as

$$\begin{aligned} J(x) &= \int_{\mathcal{A}\times\mathcal{R}} \frac{r}{a\varrho+(1-a)/2}\left[\phi(ah(x))-\phi(ag(x))\right] dP_x(a,r) \\ &= [\phi(h(x))-\phi(g(x))]\int_{\mathcal{R}} r dP(r|x,a=1) \\ &\quad + [\phi(-h(x))-\phi(-g(x))]\int_{\mathcal{R}} r dP(r|x,a=-1) \\ &\leq M\left\{[\phi(h(x))-\phi(g(x))]+[\phi(-h(x))-\phi(-g(x))]\right\}, \quad x\in\mathcal{X}. \end{aligned}$$

The notion of $B_{h,g}$ implies that $|\phi(ah(x))-\phi(ag(x))| \leq B_{h,g}|h(x)-g(x)|$ for each $a\in\mathcal{A}$. Therefore, we have

$$\|J\|_{C(\mathcal{X})} \leq 2MB_{h,g}\|h-g\|_{C(\mathcal{X})}.$$

Next, we bound $|J|_{C^{\varsigma}(\mathcal{X})}$. For $x,u\in\mathcal{X}$, split $J(x)-J(u)$ into two parts

$$\begin{aligned} J(x)-J(u) = &\int_{\mathcal{A}\times\mathcal{R}} \frac{r}{a\varrho+(1-a)/2}\{[\phi(ah(x))-\phi(ag(x))] \\ &-[\phi(ah(u))-\phi(ag(u))]\}dP_x(a,r) \\ &+\int_{\mathcal{A}\times\mathcal{R}} \frac{r}{a\varrho+(1-a)/2}[\phi(ah(u))-\phi(ag(u))]d\left[P_x-P_u\right](a,r). \end{aligned} \tag{29}$$

It is observed from the definition of $B_{h,g}$ and $|f|_{C^{\varsigma}(\mathcal{X})} := \sup_{x\neq y}\frac{|f(x)-f(y)|}{(d(x,y))^{\varsigma}}$ that

$$|\phi(ah(x))-\phi(ah(u))| \leq B_{h,g}|h(x)-h(u)| \leq B_{h,g}|h|_{C^{\varsigma}(\mathcal{X})}(d(x,u))^{\varsigma},$$

and

$$|\phi(ag(x))-\phi(ag(u))| \leq B_{h,g}|g(x)-g(u)| \leq B_{h,g}|g|_{C^{\varsigma}(\mathcal{X})}(d(x,u))^{\varsigma}.$$

Therefore, the first term of (29) is bounded by $2MB_{h,g}\Big(|h|_{C^{\varsigma}(\mathcal{X})}+|g|_{C^{\varsigma}(\mathcal{X})}\Big)(d(x,u))^{\varsigma}$. For the second term of (29), from Assumption 2 and the definition of $\widetilde{B}_{h,g}$, it follows that

$$\begin{aligned} &\left|\int_{\mathcal{A}\times\mathcal{R}} \frac{r}{a\varrho+(1-a)/2}[\phi(ah(u))-\phi(ag(u))]d\left[P_x-P_u\right](a,r)\right| \\ &\leq \widetilde{M}\left|\int_{\mathcal{A}\times\mathcal{R}}[\phi(ah(u))-\phi(ag(u))]d\left(P_x(a,r)-P_u(a,r)\right)\right| \\ &\leq \widetilde{M}\left\|\phi(ah(u))-\phi(ag(u))\right\|_{(C^{\varsigma}(\mathcal{A}))}\left\|P_x(a,r)-P_u(a,r)\right\|_{(C^{\varsigma}(\mathcal{A}\times\mathcal{R}))^*} \\ &\leq 2\widetilde{M}\widetilde{B}_{h,g}C_P(d(x,u))^{\varsigma}. \end{aligned}$$

By integrating the two bounds presented above, we can conclude that

$$|J|_{C^{\varsigma}(\mathcal{X})} = \sup_{x\neq u\in\mathcal{X}}\frac{|J(x)-J(u)|}{(d(x,u))^{\varsigma}} \leq 2MB_{h,g}\left\{|h|_{C^{\varsigma}(\mathcal{X})}+|g|_{C^{\varsigma}(\mathcal{X})}\right\}+2\widetilde{M}C_P\widetilde{B}_{h,g}.$$

Then the desired bound follows, thereby proving the lemma. □

To estimate the variancing error and the approximation error, [24] introduced an approximation function $f_{\lambda}^{\sigma_t}\in\mathcal{H}_{\sigma_t}$ and provided a construction under the smoothness assumption of the generalized Lipschitz space $\mathrm{Lip}^*\left(s, L^v\left(\mathbb{R}^n\right)\right)$ and the general geometric noise condition. In the error analysis of algorithm (6), we adopt the same approximation function $f_\lambda^{\sigma_t}$.

Now, we present the variancing error and the approximation error under smooth conditions $\tilde{f}_P^\phi \in \mathrm{Lip}^*\left(s, L^v\left(\mathbb{R}^n\right)\right) \cap L^\infty\left(\mathbb{R}^n\right), f_P^\phi=\tilde{f}_P^\phi \mid_{\mathcal{X}}$ and $\frac{dP_{\mathcal{X}}}{dx} \in L^{v'}(\mathcal{X})$ with $\frac{1}{v'}+\frac{1}{v}=1$ for some $s>0$. Let $\tilde{K}_\sigma(x):=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{|x|^2}{2\sigma^2}\right\}$ and define a function $\tilde{f}_\sigma(x): \mathbb{R}^n \rightarrow \mathbb{R}$ as

$$\begin{aligned} \tilde{f}_\sigma(x) &= \sum_{i=1}^{\ell} \binom{\ell}{i} (-1)^{\ell-i} \left(\frac{1}{\sqrt{2\pi} i \cdot \sigma}\right)^n \int_{\mathbb{R}^n} \exp\left\{-\frac{|x-u|^2}{2(i \cdot \sigma)^2}\right\} \tilde{f}_P^\phi(u) du \\ &= \int_{\mathbb{R}^n} \tilde{K}_\sigma(u) \sum_{i=1}^{\ell} \binom{\ell}{i} (-1)^{\ell-i} \tilde{f}_P^\phi(x+i \cdot u) du. \end{aligned} \tag{30}$$

Therefore, a sequence $\{f_\lambda^\sigma \in \mathcal{H}_\sigma\}_{\sigma>0}$ can be constructed by taking $f_\lambda^\sigma(x):=\tilde{f}_\sigma(x) \mid_{\mathcal{X}}$, which represents the restriction of $\tilde{f}_\sigma(x)$ onto $\mathcal{X}$. Denote by $h_t=\left\|f_\lambda^{\sigma_t}-f_\lambda^{\sigma_{t-1}}\right\|_{\sigma_t}$ the variancing error. We now present the bound proved in [24] for the variancing error.

Lemma 2 (**Variancing error under smooth condition**). *Define* $\tilde{f}_{\sigma_t}$ *by* (30) *and* $f_\lambda^\sigma(x):=\tilde{f}_\sigma(x) \mid_{\mathcal{X}} \in \mathcal{H}_\sigma$. *Then we have*

$$\|f_\lambda^\sigma\|_\sigma \leq \frac{\ell c_\ell}{(\sqrt{2\pi})^{n/2}} \|\tilde{f}_P^\phi\|_{L^2(\mathbb{R}^n)} \sigma^{-\frac{n}{2}}, \tag{31}$$

where $c_\ell:=\max\left\{\binom{\ell}{i}, 1 \leq i \leq \ell\right\}=\binom{\ell}{\lfloor \ell/2 \rfloor}$. *If the sequence* $\{\sigma_t\}$ *is non-increasing, then*

$$h_t \leq C_{s,P}\left|\left(\frac{\sigma_{t-1}}{\sigma_t}\right)^2-1\right| \sigma_t^{-\frac{n}{2}},$$

where $C_{s,P}$ *is a constant independent of* t. *Specially, let* $\sigma_t=t^{-\beta}$ *with* $0 \leq \beta<1$. *Then for any* $t \geq 1$,

$$\|f_\lambda^{\sigma_t}\|_{\sigma_t} \leq \frac{\ell c_\ell}{(\sqrt{2\pi})^{n/2}} \|\tilde{f}_P^\phi\|_{L^2(\mathbb{R}^n)} t^{\frac{n\beta}{2}}, \tag{32}$$

and

$$h_t \leq C_{s,P,\beta} t^{-1+\frac{n\beta}{2}}, \tag{33}$$

where $C_{s,P,\beta}$ *is a constant independent of* t.

Following the construction of $\{f_\lambda^\sigma \in \mathcal{H}_\sigma\}_{\sigma>0}$ above, we will now estimate the approximation error as follows. The idea of the proof is similar to that in [24, 36].

Lemma 3 (**Approximation error under smooth condition**). *Define* $\tilde{f}_{\sigma_t}$ *by* (30) *and* $f_\lambda^\sigma(x):=\tilde{f}_\sigma(x) \mid_{\mathcal{X}} \in \mathcal{H}_\sigma$. *Suppose that for* $s>0$, $\tilde{f}_P^\phi \in \mathrm{Lip}^*\left(s, L^v\left(\mathbb{R}^n\right)\right) \cap L^\infty\left(\mathbb{R}^n\right), f_P^\phi=\tilde{f}_P^\phi \mid_{\mathcal{X}}$ *and* $\frac{dP_{\mathcal{X}}}{dx} \in L^{v'}(\mathcal{X})$ *with* $\frac{1}{v'}+\frac{1}{v}=1$. *Then there is a function* $\{f_\lambda^\sigma \in \mathcal{H}_\sigma: 0<\sigma \leq 1, \lambda \geq 0\}$ *such that*

$$\|f_\lambda^\sigma\|_{L^\infty(\mathcal{X})} \leq \widetilde{B}, \tag{34}$$

$$\mathcal{E}\left(f_\lambda^\sigma\right)-\mathcal{E}(f_P^\phi)+\frac{\lambda}{2}\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 \leq \widetilde{B}(\sigma^s+\lambda \sigma^{-n}), \tag{35}$$

where the constant $\widetilde{B} \geq 1$ *is independent of* σ *and* λ *(given explicitly in the proof).*

In particular, for $\phi_{ls}(u)=(1-u)^2$ *and* $\phi_2(u)=(1-u)_+^2$, *assume that* $\tilde{f}_P^\phi \in \mathrm{Lip}^*\left(s, L^2\left(\mathbb{R}^n\right)\right) \cap$

$L^\infty(\mathbb{R}^n)$ and $f_P^\phi = \tilde{f}_P^\phi\,|_{\mathcal{X}}$ with $s>0$, the following holds true

$$\|f_\lambda^\sigma\|_{L^\infty(\mathcal{X})} \le \widetilde{B}_1, \tag{36}$$

$$\mathcal{E}\left(f_\lambda^\sigma\right) - \mathcal{E}(f_P^\phi) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 \le \widetilde{B}_1\left(\sigma^{2s} + \lambda\sigma^{-n}\right), \tag{37}$$

where the constant $\widetilde{B}_1 \ge 1$ is independent of σ and λ (given explicitly in the proof).

Proof. According to the proof of Theorem 3 in [24], when $\tilde{f}_P^\phi \in L^\infty(\mathbb{R}^n)$, we have

$$|f_\lambda^\sigma(x)| \le \sum_{i=1}^{\ell}\begin{pmatrix}\ell\\ i\end{pmatrix}\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}, \tag{38}$$

and

$$\left\|f_\lambda^\sigma - f_P^\phi\right\|_{L^v(\mathcal{X})} \le A_{1,v}\sigma^s, \tag{39}$$

where $A_{1,v}$ is a constant independent of σ.

Let $g = \frac{dP_{\mathcal{X}}}{dx} \in L^{\nu'}(\mathcal{X})$ and denote $\bar{p} = \mathbb{E}[R \mid X=x, A=1]$ and $\bar{q} = \mathbb{E}[R \mid X=x, A=-1]$. It follows from (15) and the Schwarz inequality that

$$\begin{aligned}
&\mathcal{E}(f_\lambda^\sigma) - \mathcal{E}(f_P^\phi)\\
\le& \int_{\mathcal{X}} \big|(\phi(f_\lambda^\sigma(x)) - \phi(f_P^\phi(x)))\cdot\bar{p}\\
&+ (\phi(-f_\lambda^\sigma(x)) - \phi(-f_P^\phi(x)))\cdot\bar{q}\big| dP_{\mathcal{X}}\\
\le& 2M\int_{\mathcal{X}}\sup\left\{\left|\phi_+'(\xi)\right| : |\xi| \le \max\left\{\sum_{i=1}^{\ell}\begin{pmatrix}\ell\\ i\end{pmatrix}\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}, \|f_P^\phi\|_{L^\infty(\mathcal{X})}\right\}\right\}\\
&|f_\lambda^\sigma(x) - f_P^\phi(x)|g(x)dx\\
\le& 2M\sup\left\{\left|\phi_+'(\xi)\right| : |\xi| \le \sum_{i=1}^{\ell}\begin{pmatrix}\ell\\ i\end{pmatrix}\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}\right\}\|f_\lambda^\sigma - f_P^\phi\|_{L^v(\mathcal{X})}\|g\|_{L^{v'}(\mathcal{X})}\\
\le& 2M\sup\left\{\left|\phi_+'(\xi)\right| : |\xi| \le \sum_{i=1}^{\ell}\begin{pmatrix}\ell\\ i\end{pmatrix}\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}\right\}\|g\|_{L^{v'}(\mathcal{X})}A_{1,v}\sigma^s.
\end{aligned} \tag{40}$$

This together with (31) implies that (34) and (35) hold with

$$\widetilde{B} = \max\left\{\sum_{i=1}^{\ell}\begin{pmatrix}\ell\\ i\end{pmatrix}\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}, 2M\sup\right.$$
$$\left.\left\{\left|\phi_+'(\xi)\right| : |\xi| \le \sum_{i=1}^{\ell}\begin{pmatrix}\ell\\ i\end{pmatrix}\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}\right\}\|g\|_{L^{v'}(\mathcal{X})}A_{1,v}, \frac{1}{2}\left(\frac{\ell c_\ell}{(\sqrt{2\pi})^{n/2}}\|\tilde{f}_P^\phi\|_{L^2(\mathbb{R}^n)}\right)^2, 1\right\}.$$

Now we prove the second part of Lemma 3. If $\phi_{ls}(u) = (1-u)^2$ or $\phi_2(u) = (1-u)_+^2$, we get that

$$\begin{aligned}
\mathcal{E}\left(f_\lambda^\sigma\right) - \mathcal{E}(f_P^{\phi_2}) &\le \mathcal{E}\left(f_\lambda^\sigma\right) - \mathcal{E}(f_P^{\phi_{ls}})\\
&= \mathbb{E}\Big[\frac{R}{A\varrho + (1-A)/2}(Af_P^{\phi_{ls}}(X) - Af_\lambda^\sigma(X))^2\Big]\\
&\le \widetilde{M}\int_{\mathcal{X}}(f_P^{\phi_{ls}}(x) - f_\lambda^\sigma(x))^2 dP_{\mathcal{X}}\\
&= \widetilde{M}\|f_\lambda^\sigma - f_P^{\phi_{ls}}\|_{L^2(\mathcal{X})}^2.
\end{aligned}$$

Therefore, the above bound combined with (38), (39) and (31) yields that (36) and (37) hold with $\widetilde{B}_1 = \max\left\{1, \sum_{i=1}^{\ell} \begin{pmatrix} \ell \\ i \end{pmatrix} \|\tilde{f}_P^{\phi}\|_{L^\infty(\mathbb{R}^n)},\ \widetilde{M}A_{1,2}^2, \frac{1}{2}\left(\frac{\ell c_\ell}{(\sqrt{2\pi})^{n/2}}\|\tilde{f}_P^{\phi}\|_{L^2(\mathbb{R}^n)}\right)^2\right\}$. □

Remark 4. *In the error analysis of algorithm* (6) *under the smooth condition, setting* $\sigma_t = t^{-\beta}$, *the bound* (35) *can be expressed as regularization error* $\mathcal{D}(\sigma_t, \lambda_t) \leq \widetilde{B}(\sigma_1^s t^{-s\beta} + \lambda_1 \sigma_1^{-n} t^{-(\gamma - n\beta)})$. *Therefore, a natural choice for the parameter is* $\beta = \frac{\gamma}{s+n}$ *in Theorem 2. When* $\phi_{ls}(u) = (1-u)^2$ *and* $\phi_2(u) = (1-u)_+^2$, (37) *implies that* $\mathcal{D}(\sigma_t, \lambda_t) \leq \widetilde{B}_1(\sigma_1^{2s} t^{-2s\beta} + \lambda_1 \sigma_1^{-n} t^{-(\gamma - n\beta)})$. *In this case, the parameter choice is* $\beta = \frac{\gamma}{2s+n}$ *in Theorem 1.*

Next, we present the bounds for the variancing error and the approximation error under the noise condition (28). Consider $\mathcal{X}$ as the closed unit ball within $\mathbb{R}^n$, and let $\tilde{\mathcal{X}} = 3\mathcal{X}$. The function f_P^{ϕ} can be extended onto $\mathbb{R}^n$ as follows

$$\tilde{f}_P^{\phi}(x) = \begin{cases} f_P^{\phi}(x), & \text{if } x \in \mathcal{X}, \\ f_P^{\phi}(\frac{x}{|x|}), & \text{if } x \in \tilde{\mathcal{X}}/\mathcal{X}, \\ 0, & \text{if } x \in \mathbb{R}^n/\tilde{\mathcal{X}}. \end{cases} \tag{41}$$

Recall that $\tilde{K}_\sigma(x) := \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{|x|^2}{2\sigma^2}\right\}$. We define

$$\tilde{f}_\sigma(x) = \int_{\mathbb{R}^n} \tilde{K}_\sigma(x-u)\tilde{f}_P^{\phi}(u)\mathrm{d}u,\ x \in \mathbb{R}^n. \tag{42}$$

Consequently, let $\{f_\lambda^\sigma\}_{\sigma>0}$ be a sequence such that $f_\lambda^\sigma \in \mathcal{H}_\sigma$ represents the restriction of $\tilde{f}_\sigma$ onto $\mathcal{X}$.

We now present the error bounds for variance under noise condition, as detailed in Lemma 4 from [24].

Lemma 4 (**Variancing error under noise condition**). *We define* $\tilde{f}_{\sigma_t}$ *by* (42) *and* $f_\lambda^\sigma(x) := \tilde{f}_\sigma(x)\mid_{\mathcal{X}} \in \mathcal{H}_\sigma$. *If* $\sigma_t = t^{-\beta}$ *with* $0 \leq \beta < 1$, *there holds*

$$\left\|f_\lambda^{\sigma_t}\right\|_{\sigma_t} \leq \frac{1}{(\sqrt{2\pi})^{n/2}}\left\|\tilde{f}_P^{\phi}\right\|_{L^2(\mathbb{R}^n)} t^{n\beta/2}$$

and

$$h_t \leq \frac{2c_\beta + c_\beta^2}{\sqrt{2}(\sqrt{2\pi})^{n/2}}\left\|\tilde{f}_P^{\phi}\right\|_{L^2(\mathbb{R}^n)} t^{-1+\frac{n\beta}{2}},$$

where $c_\beta = 2\max\left\{1, 3^{\beta-1}\beta\right\}$.

Lemma 5 (**Approximation error under noise condition**). *Define* $\tilde{f}_{\sigma_t}$ *by* (42) *and* $f_\lambda^\sigma(x) := \tilde{f}_\sigma(x)\mid_{\mathcal{X}} \in \mathcal{H}_\sigma$. *Let Assumption 5 be satisfied. Suppose that* $\tilde{f}_P^{\phi} \in L^\infty(\mathbb{R}^n), f_P^{\phi} = \tilde{f}_P^{\phi}\mid_{\mathcal{X}}$. *Then there is a function* $\{f_\lambda^\sigma \in \mathcal{H}_\sigma : 0 < \sigma \leq 1, \lambda \geq 0\}$ *such that*

$$\|f_\lambda^\sigma\|_{L^\infty(\mathcal{X})} \leq \widetilde{B}_2, \tag{43}$$

$$\mathcal{E}(f_\lambda^\sigma) - \mathcal{E}(f_P^{\phi}) + \frac{\lambda}{2}\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 \leq \widetilde{B}_2\left(\sigma^{\frac{\xi_1}{1+\xi_2}} + \lambda\sigma^{-n}\right), \tag{44}$$

where the constant $\widetilde{B}_2 \geq 1$ *is independent of* σ *and* λ *(given explicitly in the proof).*

Proof. Since $\tilde{f}_P^{\phi} \in L^\infty(\mathbb{R}^n)$, we directly estimate $\|f_\lambda^\sigma\|_{L^\infty(\mathcal{X})}$ from the expression of $\tilde{f}_\sigma$. Noting that $\int_{\mathbb{R}^n}\tilde{K}_\sigma(y)\mathrm{d}y=1$. For each $x\in\mathcal{X}$,

$$\left|\tilde{f}_\sigma(x)\right|=\left|\int_{\mathbb{R}^n}\tilde{f}_P^\phi(u)\tilde{K}_\sigma(x-u)du\right|=\left|\int_{\mathbb{R}^n}\tilde{f}_P^\phi(x-u)\tilde{K}_\sigma(u)du\right|$$
$$\leq\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}\int_{\mathbb{R}^n}\tilde{K}_\sigma(u)du=\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}.$$

So, we can get that

$$\|f_\lambda^\sigma\|_{L^\infty(\mathcal{X})}\leq\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}.$$

According to the conclusion of Theorem 6 in [24], we have

$$\int_{\mathcal{X}}\left|f_\lambda^\sigma(x)-f_P^\phi(x)\right|dP_{\mathcal{X}}\leq\varepsilon+16C_1\left\|f_P^\phi\right\|_{L^\infty(\mathcal{X})}(8n)^{\xi_1/2}\varepsilon^{-\xi_2}\sigma^{\xi_1}.$$

Using the first inequality of (40), then

$$\begin{aligned}&\mathcal{E}(f_\lambda^\sigma)-\mathcal{E}\left(f_P^\phi\right)\\&\leq 2M\int_{\mathcal{X}}\sup\left\{\left|\phi_+'(\xi)\right|:|\xi|\leq\max\left\{\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)},\|f_P^\phi\|_{L^\infty(\mathcal{X})}\right\}\right\}\left|f_\lambda^\sigma(x)-f_P^\phi(x)\right|dP_{\mathcal{X}}\\&\leq 2M\sup\left\{\left|\phi_+'(\xi)\right|:|\xi|\leq\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}\right\}\left(\varepsilon+16C_1\|f_P^\phi\|_{L^\infty(\mathcal{X})}(8n)^{\xi_1/2}\varepsilon^{-\xi_2}\sigma^{\xi_1}\right).\end{aligned}$$

Choose $\varepsilon=\sigma^{\frac{\xi_1}{1+\xi_2}}$, then (43) and (44) are obtained by taking $\widetilde{B}_2=\max\Big\{\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)},$ $2M\sup\left\{\left|\phi_+'(\xi)\right|:|\xi|\leq\|\tilde{f}_P^\phi\|_{L^\infty(\mathbb{R}^n)}\right\}\left(1+16\big(C_1\|f_P^\phi\|_{L^\infty(\mathcal{X})}(8n)^{\xi_1/2}\big)\right),\frac{1}{2}\|\tilde{f}_P^\phi\|^2_{L^2(\mathbb{R}^n)},1\Big\}$. □

Remark 5. *In the error analysis of algorithm* (6) *under the noise condition, by setting* $\sigma_t=t^{-\beta}$, *the bound* (44) *can be formulated as a regularization error* $\mathcal{D}(\sigma_t,\lambda_t)\leq\widetilde{B}_2\Big(\sigma_1^{\frac{\xi_1}{1+\xi_2}}t^{-\frac{\xi_1\beta}{1+\xi_2}}+\lambda_1\sigma_1^{-n}t^{-(\gamma-n\beta)}\Big)$. *Therefore, a natural choice for the parameter is* $\beta=c\gamma$ *with* $c=\frac{1+\xi_2}{\xi_1+n(1+\xi_2)}$ *as specified in Theorem 3.*

4.2 Drift error

The key analysis for full ROOWL is to estimate the error arising from changing the regularization parameter from λ_{t-1} to λ_t.

Definition 6. *The drift error is defined by*

$$d_t=\|f_{\lambda_t}^{\sigma_{t-1}}-f_{\lambda_{t-1}}^{\sigma_{t-1}}\|_{\sigma_{t-1}}.$$

The drift error can be evaluated through the approximation error, a subject explored in regression [13] and classification [14, 29]. The proof for the OWL algorithm is a straightforward adaptation from [14, 29], and for simplicity, we have omitted it.

Lemma 6. *Let* ϕ *be an admissible loss function,* f_λ^σ *defined as* (21) *and* $\mu>\lambda>0$. *We have*

$$\|f_\lambda^\sigma-f_\mu^\sigma\|_\sigma\leq\frac{\mu}{2}\left(\frac{1}{\lambda}-\frac{1}{\mu}\right)\left(\|f_\lambda^\sigma\|_\sigma+\|f_\mu^\sigma\|_\sigma\right)\leq\frac{\mu}{2}\left(\frac{1}{\lambda}-\frac{1}{\mu}\right)\left(\sqrt{\frac{2\mathcal{D}(\sigma,\lambda)}{\lambda}}+\sqrt{\frac{2\mathcal{D}(\sigma,\mu)}{\mu}}\right).$$

Particularly, if we adopt the choice $\lambda_t=\lambda_1t^{-\gamma}$ *for* $t\geq1$ *with some* $0<\gamma\leq1$, *then*

$$d_t\leq2(t-1)^{\frac{\gamma}{2}-1}\sqrt{\mathcal{D}\left(\sigma_{t-1},\lambda_1(t-1)^{-\gamma}\right)/\lambda_1}.\tag{45}$$

4.3 Bounding the learning sequence

The authors in [23] studied how the RKHS norms in $\mathcal{H}_\sigma$ change with variations in the variance σ.

Lemma 7. *If $\mathcal{X} \subset \mathbb{R}^n$ and $0 < \sigma \leq \sigma' < \infty$, there holds*

$$\mathcal{H}_{\sigma'} \subseteq \mathcal{H}_\sigma \quad \text{and} \quad \|f\|_\sigma \leq \left(\frac{\sigma'}{\sigma}\right)^{n/2} \|f\|_{\sigma'} \quad \forall f \in \mathcal{H}_{\sigma'}. \tag{46}$$

We can now establish the following bound for $\{f_t\}$ by induction.

Lemma 8. *Let $\widetilde{M} = \frac{M}{\min\{\varrho, 1-\varrho\}}$. Define $\{f_t\}_{t=1}^{T+1}$ by (6). Let Assumption 3 be satisfied. Assume that the triple $\{\lambda_t, \sigma_t, \eta_t\}$ is given by (7). If*

$$\lambda_1 \geq \widetilde{M}\, |\phi'(0)|\,, \tag{47}$$

and

$$\eta_t(\widetilde{M}M(t^{\frac{n\beta}{2}+\gamma}) + \lambda_t) \leq 1, \quad t = 1, \ldots, T, \tag{48}$$

then we have for $t = 1, \ldots, T+1$

$$\|f_t\|_{\sigma_t} \leq t^{\frac{n\beta}{2}+\gamma}. \tag{49}$$

Proof. We prove it by induction. Note that $f_1 \in \mathcal{H}_{\sigma_1}$ and $\|f_1\|_{\sigma_1} \leq 1$ given that $f_1 = 0$. For $t < T$, we assume that $f_t \in \mathcal{H}_{\sigma_t}$ satisfies (49). Denote $u_t = a_t f(x_t)$, we now examine f_{t+1}, which can be formulated as

$$\begin{aligned} f_{t+1} = (1 - \lambda_t \eta_t)\, f_t - \eta_t \cdot \frac{r_t}{a_t \varrho + (1-a_t)/2} \left[\phi'_-\,(u_t) - \phi'(0)\right] a_t K_{\sigma_t}(x_t, \cdot) \\ - \eta_t \cdot \frac{r_t}{a_t \varrho + (1-a_t)/2} \phi'(0) K_{\sigma_t}(x_t, \cdot)\,. \end{aligned}$$

Express the part of the middle term as

$$\left[\phi'_-\,(u_t) - \phi'(0)\right] a_t K_{\sigma_t}(x_t, \cdot) = \frac{\phi'_-\,(u_t) - \phi'(0)}{u_t} L_t f_t,$$

where $L_t(g) = \langle g, K_{\sigma_t}(x_t, \cdot)\rangle_{\sigma_t} K_{\sigma_t}(x_t, \cdot)$. It is easy to verify that $L_t : \mathcal{H}_{\sigma_t} \to \mathcal{H}_{\sigma_t}$ is a self-adjoint, rank-1, and positive linear operator. We can establish bound on the operator norm as $\|L_t\|_{\mathcal{H}_{\sigma_t} \to \mathcal{H}_{\sigma_t}} \leq 1$ since $\langle L_t g, g\rangle_{\sigma_t} = |\langle g, K_{\sigma_t}(x_t, \cdot)\rangle|^2_{\sigma_t} \leq \|g\|^2_{\sigma_t}$ for any $g \in \mathcal{H}_{\sigma_t}$.

The local Lipschitz condition in Assumption 3 implies that the expression $\left(\phi'_-\,(u_t) - \phi'(0)\right)/u_t$ is bounded by $M(t^{\frac{n\beta}{2}+\gamma})$. Since ϕ'_- is nondecreasing and $0 < \frac{r_t}{a_t\varrho + (1-a_t)/2} < \widetilde{M}$, we conclude that

$$0 \leq \frac{r_t}{a_t \varrho + (1-a_t)/2} (\phi'_-\,(u_t) - \phi'(0))/u_t \leq \widetilde{M}M(t^{\frac{n\beta}{2}+\gamma}).$$

This leads to the linear operator $\tilde{A} := \frac{r_t}{a_t\varrho + (1-a_t)/2}(\phi'_-\,(u_t) - \phi'(0))L_t/u_t$ defined on $\mathcal{H}_{\sigma_t}$ being self-adjoint and positive with its norm bounded by $\widetilde{M}M(t^{\frac{n\beta}{2}+\gamma})$. By applying (48), the linear operator $(1 - \eta_t\lambda_t)\, I - \eta_t \tilde{A}$ is self-adjoint, positive, and bounded by $1 - \eta_t\lambda_t$. Thus

$$\left\| (1 - \eta_t \lambda_t)\, f_t - \eta_t \cdot \frac{r_t}{a_t \varrho + (1-a_t)/2} \left[\phi'_-\,(u_t) - \phi'(0)\right] a_t K_{\sigma_t}(x_t, \cdot) \right\|_{\sigma_t} \leq (1 - \eta_t\lambda_t)\, \|f_t\|_{\sigma_t}\,.$$

It yields from (7), (47) and $\widetilde{M} = \frac{M}{\min\{\varrho, 1-\varrho\}}$ that

$$\|f_{t+1}\|_{\sigma_t} \leq (1 - \eta_t\lambda_t)\, \|f_t\|_{\sigma_t} + \eta_t \widetilde{M}\, |\phi'(0)|$$

$$= t^{\frac{n\beta}{2}+\gamma} - \eta_1\lambda_1 t^{-\alpha+\frac{n\beta}{2}} + \eta_1 t^{-\alpha}\widetilde{M}\,|\phi'(0)|$$
$$\le t^{\frac{n\beta}{2}+\gamma} - (t^{\frac{n\beta}{2}} - 1)\eta_1 t^{-\alpha}\widetilde{M}\,|\phi'(0)|$$
$$\le t^{\frac{n\beta}{2}+\gamma}.$$

Thus, we obtain from (46) that

$$\|f_{t+1}\|_{\sigma_{t+1}} \le \left(\frac{\sigma_t}{\sigma_{t+1}}\right)^{\frac{n}{2}} \|f_{t+1}\|_{\sigma_t} \le \left(\frac{t+1}{t}\right)^{\frac{n\beta}{2}} t^{\frac{n\beta}{2}+\gamma} \le (t+1)^{\frac{n\beta}{2}+\gamma}.$$

Therefore, we get the desired result using induction. □

4.4 One-step iteration

The first term on the right-hand side of (22) is to be evaluated by means of the error $\left\|f_{T+1} - f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}$. This error is iteratively bounded through a one-step analysis shifting from $\|f_t - f_{\lambda_{t-1}}^{\sigma_{t-1}}\|_{\sigma_{t-1}}$ to $\left\|f_{t+1} - f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}$. The subsequent Lemma 9 can be established following a procedure analogous to that in [14].

Lemma 9. *Define* $\{f_t\}_{t=1}^{T+1}$ *by* (6). *Then we have*

$$\mathbb{E}_{z_t}\left\|f_{t+1} - f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}^2 \le (1-\eta_t\lambda_t)\left\|f_t - f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}^2 + 2\eta_t\Delta_t$$
$$+ \eta_t^2\mathbb{E}_{z_t}\left\|\frac{r_t}{a_t\varrho + (1-a_t)/2}\phi'_-\left(a_t f_t\left(x_t\right)\right)a_t K_{\sigma_t}(x_t,\cdot) + \lambda_t f_t\right\|_{\sigma_t}^2,$$

where

$$\Delta_t = \int_{\mathcal{Z}} \frac{r}{a\varrho + (1-a)/2}\left\{\phi\left(a f_{\lambda_t}^{\sigma_t}(x)\right) - \phi\left(a f_t(x)\right)\right\} d\left[P^{(t)} - P\right]. \tag{50}$$

Proof. Denote $G_t^\lambda := \frac{r_t}{a_t\varrho+(1-a_t)/2}\phi'_-\left(a_t f_t\left(x_t\right)\right)a_t K_{\sigma_t}\left(x_t,\cdot\right) + \lambda_t f_t$. (6) can be expressed as $f_{t+1} = f_t - \eta_t G_t^\lambda$. Then

$$\left\|f_{t+1} - f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}^2 = \left\|f_t - f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}^2 + \eta_t^2\left\|G_t^\lambda\right\|_{\sigma_t}^2 + 2\eta_t\left\langle G_t^\lambda, f_{\lambda_t}^{\sigma_t} - f_t\right\rangle_{\sigma_t}. \tag{51}$$

It follows from the reproducing property (3) that

$$\left\langle\frac{r_t}{a_t\varrho + (1-a_t)/2}\phi'_-\left(a_t f_t\left(x_t\right)\right)a_t K_{\sigma_t}\left(x_t,\cdot\right), f_{\lambda_t}^{\sigma_t} - f_t\right\rangle_{\sigma_t} = \frac{r_t}{a_t\varrho + (1-a_t)/2}\phi'_-\left(u_t\right)\left(u_t^{\sigma_t} - u_t\right),$$

where $u_t^{\sigma_t} = a_t f_{\lambda_t}^{\sigma_t}\left(x_t\right)$ and $u_t = a_t f_t\left(x_t\right)$. It can be deduced from the convexity of ϕ that

$$\phi'_-\left(u_t\right)\left(u_t^{\sigma_t} - u_t\right) \le \phi\left(u_t^{\sigma_t}\right) - \phi\left(u_t\right).$$

For the other part of the last term of (51), we derive

$$\left\langle f_t, f_{\lambda_t}^{\sigma_t} - f_t\right\rangle_{\sigma_t} \le \frac{1}{2}\left\|f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}^2 + \frac{1}{2}\|f_t\|_{\sigma_t}^2 - \|f_t\|_{\sigma_t}^2 = \frac{1}{2}\left\|f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}^2 - \frac{1}{2}\|f_t\|_{\sigma_t}^2.$$

Thus, the last term of (51) can be bounded as follows:

$$\left\langle G_t^\lambda, f_{\lambda_t}^{\sigma_t} - f_t\right\rangle_{\sigma_t}$$
$$\le \left[\frac{r_t}{a_t\varrho + (1-a_t)/2}\phi\left(u_t^{\sigma_t}\right) + \frac{\lambda_t}{2}\left\|f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}^2\right] - \left[\frac{r_t}{a_t\varrho + (1-a_t)/2}\phi\left(u_t\right) + \frac{\lambda_t}{2}\|f_t\|_{\sigma_t}^2\right].$$

Since f_t only depends on $\{z_1,\cdots,z_{t-1}\}$ and is independent of z_t, it follows that

$$\mathbb{E}_{z_t}\left\langle G_t^\lambda, f_{\lambda_t}^{\sigma_t} - f_t\right\rangle_{\sigma_t} \le \left[\mathcal{E}\left(f_{\lambda_t}^{\sigma_t}\right) + \frac{\lambda_t}{2}\left\|f_{\lambda_t}^{\sigma_t}\right\|_{\sigma_t}^2\right] - \left[\mathcal{E}\left(f_t\right) + \frac{\lambda_t}{2}\|f_t\|_{\sigma_t}^2\right] + \Delta_t.$$

This in connection with Lemma 2 in [27] implies that

$$\mathbb{E}_{z_t} \left\langle G_t^\lambda, f_{\lambda_t}^{\sigma_t} - f_t \right\rangle_{\sigma_t} \le -\frac{\lambda_t}{2} \left\| f_{\lambda_t}^{\sigma_t} - f_t \right\|_{\sigma_t}^2 + \Delta_t.$$

Therefore, it yields from the above bound and (51) that

$$\mathbb{E}_{z_t} \left\| f_{t+1} - f_{\lambda_t}^{\sigma_t} \right\|_{\sigma_t}^2 \le (1 - \lambda_t \eta_t) \left\| f_t - f_{\lambda_t}^{\sigma_t} \right\|_{\sigma_t}^2 + 2\eta_t \Delta_t + \eta_t^2 \mathbb{E}_{z_t} \left\| G_t^\lambda \right\|_{\sigma_t}^2,$$

which is the result we desired. □

Define $h_t = \left\| f_\lambda^{\sigma_t} - f_\lambda^{\sigma_{t-1}} \right\|_{\sigma_t}$ with $f_\lambda^{\sigma_t} = f_{\lambda_t}^{\sigma_t}$ and let $g_t = \left\| f_{\lambda_t}^{\sigma_{t-1}} - f_{\lambda_{t-1}}^{\sigma_{t-1}} \right\|_{\sigma_t}$. The one-step iteration result for algorithm (6) is given as follows. To achieve optimal error bounds, we will utilize some free parameters $0 < \vartheta_1 < 2$ and $A_1, A_2 > 0$.

Proposition 1. *Define $\{f_t\}$ by (6). Let $\{\lambda_t, \sigma_t, \eta_t\}$ follow (7) with $0 < \beta < \frac{1}{n}, \alpha + \gamma < 1$ and $t \ge \left(\frac{4}{\lambda_1 \eta_1} + 1\right)^{\frac{1}{1-\alpha-\gamma}}$. For $0 < \vartheta_1 < 2$ and $A_1, A_2 > 0$, the bound holds*

$$\begin{aligned} &\mathbb{E}_{z_t} \left\| f_{t+1} - f_{\lambda_t}^{\sigma_t} \right\|_{\sigma_t}^2 \\ &\le \left(1 + 2A_1 g_t^{\vartheta_1} + 2A_2 h_t^{\vartheta_1} - \frac{3}{4}\eta_t \lambda_t\right) \left\| f_t - f_{\lambda_{t-1}}^{\sigma_{t-1}} \right\|_{\sigma_{t-1}}^2 \\ &\quad + g_t^{2-\vartheta_1}/A_1 + h_t^{2-\vartheta_1}/A_2 + 2g_t^2 + 2h_t^2 + 2\eta_t \Delta_t \\ &\quad + \eta_t^2 \mathbb{E}_{z_t} \left[\left\| \frac{r_t}{a_t \varrho + (1-a_t)/2} \phi'_-\left(a_t f_t(x_t)\right) a_t K_{\sigma_t}(x_t, \cdot) + \lambda_t f_t \right\|_{\sigma_t}^2 \right], \end{aligned} \tag{52}$$

where Δ_t is defined as (50).

Proof. We firstly decompose $\left\| f_t - f_{\lambda_t}^{\sigma_t} \right\|_{\sigma_t}$ as follows:

$$\left\| f_t - f_{\lambda_t}^{\sigma_t} \right\|_{\sigma_t} \le \left\| f_t - f_{\lambda_{t-1}}^{\sigma_{t-1}} + f_{\lambda_{t-1}}^{\sigma_{t-1}} - f_{\lambda_t}^{\sigma_{t-1}} + f_{\lambda_t}^{\sigma_{t-1}} - f_{\lambda_t}^{\sigma_t} \right\|_{\sigma_t} \le \left\| f_t - f_{\lambda_{t-1}}^{\sigma_{t-1}} \right\|_{\sigma_t} + g_t + h_t.$$

We can apply the elementary inequality $2xy \le Ax^2 y^{\vartheta_1} + y^{2-\vartheta_1}/A$ with $0 < \vartheta_1 < 2$ and $A > 0$ to $x = \left\| f_t - f_{\lambda_{t-1}}^{\sigma_{t-1}} \right\|_{\sigma_t}, y = g_t$ and $A = A_1$ or $y = h_t$ and $A = A_2$, then we get

$$\left\| f_t - f_{\lambda_t}^{\sigma_t} \right\|_{\sigma_t}^2 \le \left(1 + A_1 g_t^{\vartheta_1} + A_2 h_t^{\vartheta_1}\right) \left\| f_t - f_{\lambda_{t-1}}^{\sigma_{t-1}} \right\|_{\sigma_t}^2 + g_t^{2-\vartheta_1}/A_1 + h_t^{2-\vartheta_1}/A_2 + 2g_t^2 + 2h_t^2. \tag{53}$$

It follows from Lemma 9 and Lemma 7 that

$$\begin{aligned} &\mathbb{E}_{z_t} \left\| f_{t+1} - f_{\lambda_t}^{\sigma_t} \right\|_{\sigma_t}^2 \\ &\le \left\{ \left(\frac{t}{t-1}\right)^{n\beta} \left(1 + A_1 g_t^{\vartheta_1} + A_2 h_t^{\vartheta_1}\right) \left(1 - \eta_t \lambda_t\right) \left\| f_t - f_{\lambda_{t-1}}^{\sigma_{t-1}} \right\|_{\sigma_{t-1}}^2 \right\} \\ &\quad + g_t^{2-\vartheta_1}/A_1 + h_t^{2-\vartheta_1}/A_2 + 2g_t^2 + 2h_t^2 + 2\eta_t \Delta_t + \eta_t^2 \mathbb{E}_{z_t} \left\| G_t^\lambda \right\|_{\sigma_t}^2, \end{aligned} \tag{54}$$

where $G_t^\lambda := \frac{r_t}{a_t \varrho + (1-a_t)/2} \phi'_-\left(a_t f_t(x_t)\right) a_t K_{\sigma_t}(x_t, \cdot) + \lambda_t f_t$. Since $n\beta < 1$, $\left(\frac{t}{t-1}\right)^{n\beta} \le 1 + (t-1)^{-1}$ for $t \ge 2$, the coefficient for the first component of (54) can be determined as

$$\begin{aligned} &\left(\frac{t}{t-1}\right)^{n\beta} \left(1 + A_1 g_t^{\vartheta_1} + A_2 h_t^{\vartheta_1}\right) (1 - \eta_t \lambda_t) \\ &\le \left(1 + 2A_1 g_t^{\vartheta_1} + 2A_2 h_t^{\vartheta_1} + (t-1)^{-1}\right) (1 - \eta_t \lambda_t) \\ &\le 1 + 2A_1 g_t^{\vartheta_1} + 2A_2 h_t^{\vartheta_1} + (t-1)^{-1} - \eta_t \lambda_t. \end{aligned}$$

Note that $t \ge \left(\frac{4}{\lambda_1 \eta_1} + 1\right)^{\frac{1}{1-\alpha-\gamma}}$ implies $(t-1)^{-1} \le \frac{1}{4}\eta_t \lambda_t$. Then we get the desired conclusion (52). □

4.5 Sample error

We are now ready to apply one-step iteration to bound the sample error $\left\|f_{T+1}-f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}$. While some ideas in the subsequent proof come from [13, 14], our results offer a tighter upper bound.

Proposition 2. *Let $\widetilde{M}=\frac{M}{\min\{\varrho,1-\varrho\}}$. For an admissible loss function ϕ, define the sequence $\{f_t\}_{t=1}^{T+1}$ by (6). Suppose Assumptions 1, 2 and 3 hold. Suppose that for some $s>0$, $\tilde{f}_P^{\phi}\in \mathrm{Lip}^*\left(s, L^v\left(\mathbb{R}^n\right)\right)\cap L^{\infty}\left(\mathbb{R}^n\right), f_P^{\phi}=\tilde{f}_P^{\phi}\mid_{\mathcal{X}}$ and $\frac{dP_{\mathcal{X}}}{dx}\in L^{v'}(\mathcal{X})$ with $\frac{1}{v'}+\frac{1}{v}=1$. Take the triple $\{\lambda_t,\sigma_t,\eta_t\}$ as specified in (7), where $0<\eta_1\leq 1/\left(\widetilde{M}\left(M_0+2N_1\right)+\lambda_1\right)$, $\lambda_1\geq \widetilde{M}\left|\phi'(0)\right|$ and $\beta=c\gamma$ with $c=1/(s+n)$. If the following conditions are satisfied*

$$0<\gamma<1/(1+\max\{3nc/2+1,(2+nc)(p+1/2)\}), \tag{55}$$

$$\gamma\max\{nc+1,(2+nc)p+1\}<\alpha<1-(nc/2+1)\gamma, \tag{56}$$

and

$$b>((p+1)(nc/2+1)+\zeta c+1)\gamma, \tag{57}$$

then the inequality

$$\mathbb{E}_{\mathcal{Z}_T}\left\|f_{T+1}-f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}^2\leq C_2T^{-\theta} \tag{58}$$

holds. Here θ is defined by (27) and C_2 is a constant independent of T (explicitly derived in the proof).

Proof. Firstly, we bound Δ_t using Lemma 8. Applying (14) with $w=t^{\frac{n\beta}{2}+\gamma}$, where $p(n\beta/2+\gamma)\leq\alpha$, and the condition $\eta_1\leq 1/(\widetilde{M}\left(M_0+2N_1\right)+\lambda_1)$, we obtain

$$\begin{aligned}\eta_t\left(\widetilde{M}M\left(t^{\frac{n\beta}{2}+\gamma}\right)+\lambda_t\right)&\leq\eta_t\left(\widetilde{M}(M_0+2N_1)t^{p(n\beta/2+\gamma)}+\lambda_1t^{-\gamma}\right)\\&\leq\eta_1\left(\widetilde{M}(M_0+2N_1)+\lambda_1\right)\leq 1.\end{aligned}$$

The above bound in connection with $\lambda_1\geq\widetilde{M}\left|\phi'(0)\right|$ and (4) implies that

$$\|f_t\|_{C(\mathcal{X})}\leq\|f_t\|_{\sigma_t}\leq t^{\frac{n}{2}\beta+\gamma}.$$

According to (34), we get that $\left\|f_{\lambda_t}^{\sigma_t}\right\|_{C(\mathcal{X})}\leq\widetilde{B}$ for a constant $\widetilde{B}\geq 1$. Therefore, plugging the bounds for f_t and $f_{\lambda_t}^{\sigma_t}$ into Lemma 1 , we get

$$B_{f_{\lambda_t}^{\sigma_t},f_t}\leq N(\widetilde{B}t^{\frac{n}{2}\beta+\gamma}),\quad \widetilde{B}_{f_{\lambda_t}^{\sigma_t},f_t}\leq\tilde{N}(\widetilde{B}t^{\frac{n}{2}\beta+\gamma}).$$

Recall the following result stated in [34]

$$\|g\|_{C^{\zeta}(\mathcal{X})}\leq\left(1+\sigma^{-\zeta}\right)\|g\|_{\sigma}. \tag{59}$$

Then we have

$$\|f_t\|_{C^{\zeta}(\mathcal{X})}\leq A_0't^{\gamma+(n/2+\zeta)\beta},\quad\left\|f_{\lambda_t}^{\sigma_t}\right\|_{C^{\zeta}(\mathcal{X})}\leq A_1't^{\zeta\beta}.$$

Here $A_0'=(1+\sigma_1^{-\zeta})$ and $A_1'=\widetilde{B}(1+\sigma_1^{-\zeta})$. Thus, we can bound Δ_t as

$$\begin{aligned}\Delta_t\leq&\left\|P_{\mathcal{X}}^{(t)}-P_{\mathcal{X}}\right\|_{(C^{\zeta}(\mathcal{X}))^*}\Big\{2M\left(\|f_t\|_{C^{\zeta}(\mathcal{X})}+\left\|f_{\lambda_t}^{\sigma_t}\right\|_{C^{\zeta}(\mathcal{X})}\right)N(\widetilde{B}t^{\frac{n}{2}\beta+\gamma})\\&+2\widetilde{M}C_P\tilde{N}(\widetilde{B}t^{\frac{n}{2}\beta+\gamma})\Big\}\\\leq&\, C_0t^{-b+(np/2+n/2+\zeta)\beta+(p+1)\gamma},\end{aligned}$$

where $C_0=C_b[2M\widetilde{B}^p\left(A_0'+A_1'\right)+2C_P\widetilde{M}\widetilde{B}^{p+1}]N_1$.

Next, we come to estimate the last part of (52)
$\eta_t^2\mathbb{E}_{z_t}\Big[\big\|\frac{r_t}{a_t\varrho+(1-a_t)/2}\phi'_-\left(a_tf_t\left(x_t\right)\right)a_tK_{\sigma_t}(x_t,\cdot)+\lambda_tf_t\big\|_{\sigma_t}^2\Big]$.

According to Lemma 8, it is bounded by

$$2\eta_1^2(\widetilde{M}^2N_1^2+\lambda_1^2)t^{-\min\{2\alpha-p(n\beta+2\gamma),2\alpha-n\beta\}}.$$

To estimate the other parts of (52), we further need to determine the bound of the drift error

$$d_t=\big\|f_{\lambda_t}^{\sigma_{t-1}}-f_{\lambda_{t-1}}^{\sigma_{t-1}}\big\|_{\sigma_{t-1}}\leq 2(t-1)^{\frac{\gamma}{2}-1}\sqrt{\mathcal{D}\left(\sigma_{t-1},\lambda_1(t-1)^{-\gamma}\right)/\lambda_1}. \tag{60}$$

Notice that $s\beta=\gamma-n\beta$. It follows from (60) and Lemma 3 that

$$\begin{aligned}g_t=\big\|f_{\lambda_t}^{\sigma_{t-1}}-f_{\lambda_{t-1}}^{\sigma_{t-1}}\big\|_{\sigma_t}&\leq\left(\frac{t}{t-1}\right)^{\frac{n\beta}{2}}\big\|f_{\lambda_t}^{\sigma_{t-1}}-f_{\lambda_{t-1}}^{\sigma_{t-1}}\big\|_{\sigma_{t-1}}\\&\leq\left(\frac{t}{t-1}\right)^{\frac{n\beta}{2}}2(t-1)^{\frac{\gamma}{2}-1}(t-1)^{-\frac{s\beta}{2}}\sqrt{\widetilde{B}(\sigma_1^s+\lambda_1\sigma_1^{-n})/\lambda_1}\leq g_0t^{\frac{n\beta}{2}-1},\end{aligned}$$

where $g_0=4\sqrt{\widetilde{B}(\sigma_1^s+\lambda_1\sigma_1^{-n})/\lambda_1}$. Applying the above bound and (33), we can choose $\vartheta_1=\frac{\alpha+\gamma}{1-\frac{n\beta}{2}}$, $A_1=\frac{\eta_1\lambda_1}{8g_0^{\vartheta_1}}$ and $A_2=\frac{\eta_1\lambda_1}{8C_{s,P,\beta}^{\vartheta_1}}$. Since $\alpha+\gamma<1$ and $n\beta<1$ lead to $\frac{\alpha+\gamma}{1-\frac{n\beta}{2}}\in(0,2)$, then the coefficient of the first term of (52) can be bounded as

$$1+2A_1g_t^{\vartheta_1}+2A_2h_t^{\vartheta_1}-\frac{3}{4}\eta_t\lambda_t\leq 1-\frac{1}{4}\eta_t\lambda_t.$$

Therefore, we obtain from Proposition 1 that for each $t>\left(\frac{4}{\eta_1\lambda_1}+1\right)^{\frac{1}{1-\alpha-\gamma}}$,

$$\mathbb{E}_{\mathcal{Z}_t}\big\|f_{t+1}-f_{\lambda_t}^{\sigma_t}\big\|_{\sigma_t}^2\leq\left(1-\frac{1}{4}\eta_t\lambda_t\right)\mathbb{E}_{\mathcal{Z}_{t-1}}\big\|f_t-f_{\lambda_{t-1}}^{\sigma_{t-1}}\big\|_{\sigma_{t-1}}^2+A_3t^{-\widetilde{\theta}}, \tag{61}$$

where

$$\widetilde{\theta}=\min\{2-n\beta-\alpha-\gamma,2\alpha-p(n\beta+2\gamma),2\alpha-n\beta,b+\alpha-(np/2+n/2+\zeta)\beta-(p+1)\gamma\}$$

and

$$A_3=g_0^{2-\vartheta_1}/A_1+h_0^{2-\vartheta_1}/A_2+2g_0^2+2h_0^2+2\eta_1C_0+\eta_1^2(\widetilde{M}^2N_1^2+2\lambda_1^2).$$

Now we iteratively substitute the above bound into (61) from $t=T_0:=\left\lceil\left(\frac{4}{\lambda_1\eta_1}+1\right)^{\frac{1}{1-\alpha-\gamma}}\right\rceil$ to $t=T$. Therefore

$$\begin{aligned}\mathbb{E}_{\mathcal{Z}_T}\big\|f_{T+1}-f_{\lambda_T}^{\sigma_T}\big\|_{\sigma_t}^2&\leq A_3\sum_{t=T_0}^{T}\prod_{j=t+1}^{T}\left(1-\frac{\eta_1\lambda_1}{4}j^{-\gamma-\alpha}\right)t^{-\tilde{\theta}}\\&+\left\{\prod_{t=T_0}^{T}\left(1-\frac{\eta_1\lambda_1}{4}t^{-\gamma-\alpha}\right)\right\}\mathbb{E}_{\mathcal{Z}_{T_0-1}}\big\|f_{T_0}-f_{\lambda_{T_0-1}}^{\sigma_{T_0-1}}\big\|_{\sigma_{T_0-1}}^2.\end{aligned} \tag{62}$$

Denote $\iota=1-\gamma-\alpha$. Since $1-u\leq e^{-u}$ for $u\geq 0$, one can deduce that

$$\begin{aligned}\prod_{t=T_0}^{T}\left(1-\frac{\eta_1\lambda_1}{4}t^{-\gamma-\alpha}\right)&\leq\exp\left\{-\frac{\eta_1\lambda_1}{4}\sum_{t=T_0}^{T}t^{-\gamma-\alpha}\right\}\leq\exp\left\{-\frac{\eta_1\lambda_1}{4}\int_{T_0}^{T+1}x^{-\gamma-\alpha}dx\right\}\\&\leq\exp\left\{\frac{\lambda_1\eta_1}{4\iota}T_0^{\iota}\right\}\exp\left\{-\frac{\lambda_1\eta_1}{4\iota}(T+1)^{\iota}\right\}.\end{aligned} \tag{63}$$

Applying the elementary inequality $\exp\{-cx\}\leq\left(\frac{v}{ec}\right)^vx^{-v}$ with $c=\frac{\lambda_1\eta_1}{4\iota}$, $v=\frac{2}{\iota}$ and $x=$

$(T+1)^{\iota}$, we get

$$\prod_{t=T_0}^{T}\left(1-\frac{\eta_1\lambda_1}{4}t^{-\gamma-\alpha}\right)\leq \exp\left\{\frac{\lambda_1\eta_1}{4\iota}T_0^{\iota}\right\}(8/e\lambda_1\eta_1)^{\frac{2}{\iota}}T^{-2}.$$

For the remaining part of (62), we use the following elementary inequality [22], which is valid for $0<q_1<1$ $\hat{c}, q_2>0$ and $t\in\mathbb{N}$

$$\sum_{t=1}^{t-1} i^{-q_2}\exp\left\{-\hat{c}\sum_{j=i+1}^{t} j^{-q_1}\right\}\leq\left(2^{q_1+q_2}/\hat{c}+\left[(1+q_2)/e\hat{c}\left(1-2^{q_1-1}\right)\right]^{(1+q_2)/(1-q_1)}\right)t^{q_1-q_2}.$$

Set $q_1=\gamma+\alpha$ and $q_2=\tilde{\theta}$. Then

$$\sum_{t=T_0}^{T}\prod_{j=t+1}^{T}\left(1-\frac{\eta_1\lambda_1}{4}j^{-\gamma-\alpha}\right)t^{-\tilde{\theta}}\leq\sum_{t=T_0}^{T}\exp\left\{-\frac{\eta_1\lambda_1}{4}\sum_{j=t+1}^{T}j^{-\gamma-\alpha}\right\}t^{-\tilde{\theta}}\leq A_4T^{\gamma+\alpha-\tilde{\theta}},$$

where A_4 is the constant defined as

$$A_4=2^{\gamma+\alpha+\tilde{\theta}+2}/\eta_1\lambda_1+1+\left((4+4\widetilde{\theta})/e\eta_1\lambda_1\left(1-2^{-\iota}\right)\right)^{(1+\tilde{\theta})/\iota}.$$

Subsequently, by substituting the above bound and (63) into (62), it can be deduced that (58) holds for $\theta=\widetilde{\theta}-\alpha-\gamma$, and the constant C_2 is determined by

$$C_2=A_3A_4+\exp\left\{\frac{\lambda_1\eta_1}{4\iota}T_0^{\iota}\right\}(8/e\lambda_1\eta_1)^{2/\iota}\max_{1\leq t\leq T_0}\mathbb{E}_{\mathcal{Z}_t}\left(\left\|f_t-f_{\lambda_{t-1}}^{\sigma_{t-1}}\right\|_{\sigma_{t-1}}^2\right).$$

This completes the proof of Proposition 2. □

Proposition 3. *Let* $\widetilde{M}=\frac{M}{\min\{\varrho,1-\varrho\}}$. *For* $\phi_{ls}(u)=(1-u)^2$ *and* $\phi_2(u)=(1-u)_+^2$, *define* $\{f_t\}_{t=1}^{T+1}$ *by* (6). *Assume Assumptions 1, 2 and 3 hold. Suppose that for* $s>0, \tilde{f}_P^{\phi}\in \operatorname{Lip}^*\left(s,L^2\left(\mathbb{R}^n\right)\right)\cap L^{\infty}\left(\mathbb{R}^d\right)$ *and* $f_P^{\phi}=\tilde{f}_P^{\phi}|_{\mathcal{X}}$. *Take the triple* $\{\lambda_t,\sigma_t,\eta_t\}$ *as specified in* (7), *where* $0<\eta_1\leq 1/(8\widetilde{M})$, $\lambda_1=2\widetilde{M}$ *and* $\beta=c_1\gamma$ *with* $c_1=1/(2s+n)$. *If the following conditions are satisfied*

$$0<\gamma<1/(4+3nc_1/2),$$
$$\gamma(3+nc_1)<\alpha<1-(nc_1/2+1)\gamma,$$

and

$$b>(nc_1+\zeta c_1+3)\gamma,$$

then the inequality

$$\mathbb{E}_{\mathcal{Z}_T}\left\|f_{T+1}-f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}^2\leq C_3T^{-\theta_1}$$

holds. Here θ_1 *is defined as* (23) *and* C_3 *is a constant independent of* T.

Proof. We can verify that the least square loss function $\phi_{ls}(u)=(1-u)^2$ and 2-norm SVM loss function $\phi_2(u)=(max\{1-u,0\})^2$ satisfy (11) with $p=1$, $N_1=2$ and (13) with $M_0=2$. So the restriction $\eta_1\leq 1/\left(\widetilde{M}(M_0+2N_1)+\lambda_1\right)$ can also be expressed as $\eta_1\leq 1/(8\widetilde{M})$ when $\lambda_1=2\widetilde{M}$. It comes immediately from Remark 4 that the parameter choice for both the least square loss function and 2-norm SVM loss function is $\beta=\frac{\gamma}{2s+n}:=c_1\gamma$. Since the proof is similar to the one for Theorem 2, we omit it. □

To derive Proposition 4, we replace the smoothness condition in Proposition 2 with the geometric noise condition. Its proof follows the same steps as that of Proposition 2, substituting Lemma 2 with Lemma 4 and Lemma 3 with Lemma 5 (as in Proposition 2s original proof).

Hence the proof is omitted.

Proposition 4. *Let* $\widetilde{M} = \frac{M}{\min\{\varrho, 1-\varrho\}}$. *For an admissible loss function* ϕ, *define* $\{f_t\}_{t=1}^{T+1}$ *by* (6). *Assume Assumptions 1, 2, 3 and 5 hold. Suppose that* $\tilde{f}_P^\phi \in L^\infty(\mathbb{R}^n), f_P^\phi = \tilde{f}_P^\phi|_{\mathcal{X}}$. *Take the triple* $\{\lambda_t, \sigma_t, \eta_t\}$ *as specified in* (7) *with* $0 < \eta_1 \leq 1/(\widetilde{M}(M_0 + 2N_1) + \lambda_1)$, $\lambda_1 \geq \widetilde{M}|\phi'(0)|$ *and* $\beta = c_2\gamma$, *where* $c_2 = \frac{1+\xi_2}{\xi_1 + n(1+\xi_2)}$. *If the following conditions are satisfied*

$$0 < \gamma < 1/(1 + \max\{3nc_2/2 + 1, (2 + nc_2)(p + 1/2)\}), \tag{64}$$

$$\gamma \max\{nc_2 + 1, (2 + nc_2)p + 1\} < \alpha < 1 - (nc_2/2 + 1)\gamma, \tag{65}$$

and

$$b > ((p + 1)(nc_2/2 + 1) + \zeta c_2 + 1)\gamma, \tag{66}$$

the the inequality

$$\mathbb{E}_{\mathcal{Z}_T} \left\|f_{T+1} - f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}^2 \leq C_4 T^{-\theta_2} \tag{67}$$

holds. Here θ_2 *is defined as* (3) *and* C_4 *is a constant independent of* T.

4.6 Convergence rates

Proof of Theorem 1. Based on Corollary 5.3 in [36] and the fact that $\|f\|_{L^2(\mathcal{X})} \leq \|f\|_{C(\mathcal{X})} \leq \|f\|_{\mathcal{H}_\sigma}$, it can be concluded that

$$\begin{aligned}\mathcal{E}(f_{T+1}) - \mathcal{E}(f_P^{\phi_2}) &\leq \mathcal{E}(f_{T+1}) - \mathcal{E}(f_P^{\phi_{ls}}) \\ &\leq 2M \int_{\mathcal{X}} \left(f_P^{\phi_{ls}}(x) - f_{T+1}(x)\right)^2 dP_{\mathcal{X}} \\ &\leq 4M \left\|f_{T+1} - f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}^2 + 4M \left\|f_{\lambda_T}^{\sigma_T} - f_P^{\phi_{ls}}\right\|_{L^2(\mathcal{X})}^2.\end{aligned} \tag{68}$$

It follows from Proposition 3 that

$$\mathbb{E}_{\mathcal{Z}_T} \left\|f_{T+1} - f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}^2 \leq C_3 T^{-\theta_1}. \tag{69}$$

From equation (39) with $f_\lambda^\sigma = f_{\lambda_T}^{\sigma_T}$ and $v = 2$, we know that for $\phi = \phi_{ls}$ and $\phi = \phi_2$,

$$\left\|f_{\lambda_T}^{\sigma_T} - f_P^\phi\right\|_{L^2(\mathcal{X})}^2 \leq A_{1,2}^2 T^{-2sc_1\gamma}. \tag{70}$$

Therefore, putting (69) and (70) into (68), we get

$$\mathbb{E}_{\mathcal{Z}_T}\left[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_P^\phi)\right] \leq \left[4MC_3 + 4MA_{1,2}^2\right] T^{-\min\{\theta_1, 2sc_1\gamma\}}.$$

Then the proof is complete. □

Lemma 10. *Consider a general convex function* ϕ. *Let* $\left(\frac{n\beta}{2} + \gamma\right)p < 1$ *and suppose the assumptions of Lemma 8 are satisfied. Assume that* $s > 0$, $\tilde{f}_P^\phi \in \mathrm{Lip}^*(s, L^v(\mathbb{R}^n)) \cap L^\infty(\mathbb{R}^n), f_P^\phi = \tilde{f}_P^\phi|_{\mathcal{X}}$ *and* $\frac{dP_{\mathcal{X}}}{dx} \in L^{v'}(\mathcal{X})$ *with* $\frac{1}{v'} + \frac{1}{v} = 1$. *Then we have the following inequality:*

$$\left|\mathcal{E}(f_{T+1}) - \mathcal{E}\left(f_{\lambda_T}^{\sigma_T}\right)\right| \leq 4MN_1\widetilde{B}^p T^{p\left(\frac{n\beta}{2}+\gamma\right)} \left\|f_{T+1} - f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}. \tag{71}$$

Additionally, suppose Assumption 5 is satisfied and let $\tilde{f}_P^\phi \in L^\infty(\mathbb{R}^n), f_P^\phi = \tilde{f}_P^\phi|_{\mathcal{X}}$. *Then the inequality*

$$\left|\mathcal{E}(f_{T+1}) - \mathcal{E}\left(f_{\lambda_T}^{\sigma_T}\right)\right| \leq 4MN_1\widetilde{B}_2^p T^{p\left(\frac{n\beta}{2}+\gamma\right)} \left\|f_{T+1} - f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T} \tag{72}$$

holds.

Proof. To simplify the notations, let $\bar{p} = \mathbb{E}[R \mid X = x, A = 1]$ and $\bar{q} = \mathbb{E}[R \mid X = x, A = -1]$. It is easy to get from the convexity of the loss function and the Chauchy-Schwarz inequality that

$$\begin{aligned} &\left|\mathcal{E}\left(f_{T+1}\right) - \mathcal{E}\left(f_{\lambda_T}^{\sigma_T}\right)\right| \\ &\leq \max\left\{\left|\phi'_-\left(af_{T+1}(x)\right)\right|, \left|\phi'_-\left(af_{\lambda_T}^{\sigma_T}(x)\right)\right|\right\} \int_{\mathcal{X}} (\bar{p}+\bar{q})|f_{T+1}(x) - f_{\lambda_T}^{\sigma_T}(x)| dP_{\mathcal{X}} \\ &\leq 2M \cdot \max\left\{\left|\phi'_-\left(af_{T+1}(x)\right)\right|, \left|\phi'_-\left(af_{\lambda_T}^{\sigma_T}(x)\right)\right|\right\} \left\|f_{T+1} - f_{\lambda_T}^{\sigma_T}\right\|_{\sigma_T}. \end{aligned} \tag{73}$$

Applying Lemma 3 and Lemma 8, we get

$$|af_{T+1}(x)| \leq \|f_{T+1}\|_{\sigma_{T+1}} \leq 2^{\frac{n\beta}{2}+\gamma} T^{\frac{n\beta}{2}+\gamma}, \quad \left|af_{\lambda_T}^{\sigma_T}(x)\right| \leq \left\|f_{\lambda_T}^{\sigma_T}\right\|_{L^\infty(\mathcal{X})} \leq \widetilde{B}.$$

This in connection with (11) and the condition $\left(\frac{n\beta}{2}+\gamma\right)p < 1$ leads to

$$\max\left\{\left|\phi'_-\left(af_{T+1}(x)\right)\right|, \left|\phi'_-\left(af_{\lambda_T}^{\sigma_T}(x)\right)\right|\right\} \leq N(2^{\frac{n\beta}{2}+\gamma}\widetilde{B} \cdot T^{\frac{n\beta}{2}+\gamma}) \leq 2N_1\widetilde{B}^p T^{\left(\frac{n\beta}{2}+\gamma\right)p}. \tag{74}$$

Therefore, the bound (71) can be derived by putting (74) into (73). Similarly, the bound (72) can be obtained by replacing Lemma 3 with Lemma 5. □

With the preliminaries established, we can now proceed to the specific proofs for our main results.

Proof of Theorem 2. Since (56) and $\beta = c\gamma$, we have

$$\left(\frac{n\beta}{2}+\gamma\right)p < 1.$$

Therefore, (71) holds. Consequently, it follows from Theorem 2 that

$$\mathbb{E}_{\mathcal{Z}_T}\left[\mathcal{E}\left(f_{T+1}\right) - \mathcal{E}\left(f_{\lambda_T}^{\sigma_T}\right)\right] \leq 4MN_1\widetilde{B}^p\sqrt{C_2}T^{-\frac{\theta}{2}+p\left(\frac{nc}{2}+1\right)\gamma}. \tag{75}$$

We obtain from Definition 3 and (35) with $f_\lambda^{\sigma_t} = f_{\lambda_t}^{\sigma_t}$ that

$$\mathcal{E}(f_{\lambda_T}^{\sigma_T}) - \mathcal{E}(f_P^\phi) \leq \mathcal{D}(\sigma_T, \lambda_T) \leq \widetilde{B}\left(\sigma_1^s + \lambda_1\sigma_1^{-n}\right)T^{-sc\gamma}. \tag{76}$$

Therefore, we get the following bound from (75) and (76)

$$\mathbb{E}_{\mathcal{Z}_T}\left[\mathcal{E}\left(f_{T+1}\right) - \mathcal{E}(f_P^\phi)\right] \leq \left[\widetilde{M}N_1\widetilde{B}^p\sqrt{C_2} + \widetilde{B}\left(\sigma_1^s + \lambda_1\sigma_1^{-n}\right)\right]T^{-\min\left\{\frac{\theta}{2}-p\left(\frac{nc}{2}+1\right)\gamma, sc\gamma\right\}}.$$

Proof of Example 1. We consider the convergence rate for the q-norm SVM loss function $\phi = \phi_q$ under two cases.

Case 1: $q \geq 2 - \frac{2n+2s}{3n+2s}$. In this case, it is easy to verify that the q-norm SVM loss function $\phi = \phi_q$ satisfies condition (11) in Assumption 3 with $p = q-1$, $N_1 = q$, and condition (13) with $M_0 = q(q-1)2^{q-1}$. As a result, when $\lambda_1 = q\widetilde{M}$, the step size η_1 can be chosen as

$$\eta_1 \leq \frac{1}{\widetilde{M}(q(q-1)2^{q-1}+3q)}.$$

We then take the parameters

$$\alpha = \frac{(12q-10)n+(8q-2)s}{(15q-10)n+10qs}, \quad \beta = \frac{2}{(15q-10)n+10qs}, \quad \gamma = \frac{2n+2s}{(15q-10)n+10qs},$$

and assume

$$b > \frac{(9q-4)n+(6q-2)s+2\zeta}{5(3q-2)n+10qs}.$$

Under these choices, the conditions in Theorem 2 are satisfied. In particular, inequality (24) becomes

$$0 < \gamma < \frac{2n+2s}{(15q-10)n+(10q-6)s},$$

and inequality (25) becomes

$$\frac{(12q-10)n+(8q-6)s}{(15q-10)n+10qs} < \alpha < \frac{(12q-10)n+8qs}{(15q-10)n+10qs},$$

both of which are valid. Therefore, we obtain the convergence rate

$$\mathbb{E}_{\mathcal{Z}_T}\left(\mathcal{E}(f_{T+1})-\mathcal{E}(f_P^\phi)\right) = O\left(T^{-\min\left\{\frac{2s}{(15q-10)n+10qs},\, \frac{b}{2}-\frac{(9q-4)n+(6q-2)s+2\zeta}{10(3q-2)n+20qs}\right\}}\right).$$

Case 2: $1 < q < 2 - \frac{2n+2s}{3n+2s}$. In this case, we verify that the q-norm SVM loss $\phi = \phi_q$ satisfies condition (11) with $p = q-1$, $N_1 = q$, and condition (13) with $M_0 = 4$. Accordingly, when $\lambda_1 = q\widetilde{M}$, we choose the learning rate

$$\eta_1 \le \frac{1}{\widetilde{M}(4+3q)}.$$

We set

$$\alpha = \frac{(6q-2)n+(4q+2)s}{(9q-2)n+(6q+4)s}, \quad \beta = \frac{2}{(9q-2)n+(6q+4)s}, \quad \gamma = \frac{2n+2s}{(9q-2)n+(6q+4)s},$$

and assume

$$b > \frac{(9q-4)n+(6q-2)s+2\zeta}{(9q-2)n+(6q+4)s}.$$

Then, Theorem 2 applies. Specifically, inequality (24) becomes

$$0 < \gamma < \frac{2n+2s}{(9q-2)n+(6q-2)s},$$

and inequality (25) becomes

$$\frac{(6q-2)n+(4q-2)s}{(9q-2)n+(6q+4)s} < \alpha < \frac{(6q-2)n+(4q+4)s}{(9q-2)n+(6q+4)s},$$

both of which are valid. Hence, the convergence rate is

$$\mathbb{E}_{\mathcal{Z}_T}\left(\mathcal{E}(f_{T+1})-\mathcal{E}(f_P^\phi)\right) = O\left(T^{-\min\left\{\frac{2s}{(9q-2)n+(6q+4)s},\, \frac{b}{2}-\frac{(9q-4)n+(6q-2)s+2\zeta}{(18q-4)n+(12q+8)s}\right\}}\right).$$

In summary, considering both cases above, the learning rate for the q-norm SVM loss function follows from inequality (19). □

Proof of Theorem 3. Since (65) and $\beta = c_2\gamma$ hold, we have

$$(\frac{n\beta}{2}+\gamma)p < 1.$$

Hence, (72) holds. Consequently, it can be derived from Proposition 4 that

$$\mathbb{E}_{\mathcal{Z}_T}\left[\mathcal{E}\left(f_{T+1}\right)-\mathcal{E}\left(f_{\lambda_T}^{\sigma_T}\right)\right] \le 4MN_1\widetilde{B}_2^p\sqrt{C_4}T^{-\frac{\theta_2}{2}+p\left(\frac{nc_2}{2}+1\right)\gamma}. \tag{77}$$

We obtain from Definition 3 and (44) with $f_\lambda^{\sigma_t} = f_{\lambda_t}^{\sigma_t}$ that

$$\mathcal{E}(f_{\lambda_T}^{\sigma_T})-\mathcal{E}(f_P^\phi) \le \mathcal{D}(\sigma_T,\lambda_T) \le \widetilde{B}_2(\sigma_1^s+\lambda_1\sigma_1^{-n})T^{-\frac{\xi_1c_2\gamma}{1+\xi_2}}. \tag{78}$$

Therefore, we get the following bound from (75) and (78),

$$\mathbb{E}_{\mathcal{Z}_T}\left[\mathcal{E}\left(f_{T+1}\right)-\mathcal{E}(f_P^\phi)\right] \le \left[\widetilde{M}N_1\widetilde{B}_2^p\sqrt{C_4}+\widetilde{B}_2\left(\sigma_1^s+\lambda_1\sigma_1^{-n}\right)\right]T^{-\min\left\{\frac{\theta_2}{2}-p\left(\frac{nc}{2}+1\right)\gamma,\frac{\xi_1c_2\gamma}{1+\xi_2}\right\}}.$$

Then we get our desired result. □

Proof of Corollary 3. The hinge loss function $\phi = \phi_h$ satisfies (11) with $p = 0$, $N_1 = 1$ and (13) with $M_0 = 0$, while the generalized DWD loss function $\phi = \phi_{dwd}$ satisfies (11) with $p = 0$, $N_1 = 1$ and (13) with $M_0 = (d+1)^2/d$. For $\phi = \phi_h$ with $\lambda_1 = \widetilde{M}$, the restriction $0 < \eta_1 \le 1/(\widetilde{M}(M_0+2N_1)+\lambda_1)$ in Proposition 4 simplifies to $\eta_1 \le 1/(3\widetilde{M})$. For $\phi = \phi_{dwd}$ with $\lambda_1 = \widetilde{M}$, it becomes $\eta_1 \le 1/(((d+1)^2/d+3)\widetilde{M})$. Given $\mathcal{N}_\varepsilon^{\phi_h}(x) \ge \bar{c} > 0$, (28) holds for any $\xi_1 \in (0, \infty)$ and $\xi_2 = 0$. With the specified choices in this corollary, the desired learning rates for the hinge loss function follow from Theorem 3 and (16) or (17). $\square$

§5 Discussion

In summary, we propose a regularized online outcome weighted learning (ROOWL) algorithm for estimating ITR in precision medicine. Our method handles sequential, independent but non-identically distributed (i.n.i.d.) data by employing a time-varying Gaussian kernel and a time-varying regularization parameter. Theoretical analysis establishes convergence rates for several common used loss functions including the hinge loss, generalized DWD loss, least squares loss, and q-norm SVM loss, under two settings: smooth target functions in generalized Lipschitz spaces, and non-smooth target functions under a geometric noise condition. Our results demonstrate improved performance over existing approaches, especially in non-i.i.d. and non-smooth settings.

We did not address the following questions which are beyond the scope of this paper. It would be interesting to incorporate sparsity-inducing penalties such as ℓ_1-regularization, which may further improve model interpretability and performance on high-dimensional genomic and imaging data [16]. Furthermore, developing distributed or federated versions [32] of ROOWL would help address privacy and data decentralization challenges in healthcare, enabling secure collaborative learning across multiple institutions without compromising sensitive patient information.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

[1] N Aronszajn. *Theory of reproducing kernels,* Trans Amer Math Soc, 1950, 68(3): 337-404.

[2] J Li, J F Cai, Y Chen, et al. *Online tensor learning: Computational and statistical trade-offs, adaptivity and optimal regret,* 2023, DOI:10.48550/arXiv.2306.03372.

[3] X Chen, B Tang, J Fan, et al. *Online gradient descent algorithms for functional data learning,* J Complexity, 2022, 70: 101635.

[4] A Dieuleveut, F Bach. *Nonparametric stochastic approximation with large step-sizes,* Ann Statist, 2016, 44(4): 1363-1399.

[5] S Ding, B Mirza, Z Lin, et al. *Kernel based online learning for imbalance multiclass classification,* Neurocomputing, 2018, 277: 139-148.

[6] J Fan, Y Lei. *High-probability generalization bounds for pointwise uniformly stable algorithms,* Appl Comput Harmon Anal, 2024, 70: 101632.

[7] Z C Guo, A Christmann, L Shi. *Optimality of Robust Online Learning,* Found Comput Math, 2024, 24: 1455-1483.

[8] X Guo, Z C Guo, L Shi. *Capacity dependent analysis for functional online learning algorithms,* Appl Comput Harmon Anal, 2023, 67: 101567.

[9] Z C Guo, L Shi. *Fast and strong convergence of online learning algorithms,* Adv Comput Math, 2019, 45(5): 2745-2770.

[10] B L Han, D H Xiang. *Error analysis of coefficient based owl with l_1-regularization and gaussians,* Zhejiang Normal University, Master's thesis, 2022.

[11] R Hodson. *Precision medicine,* Nature, 2015, 537(7619): S49.

[12] S C H Hoi, D Sahoo, J Lu, et al. *Online learning: A comprehensive survey,* Neurocomputing, 2021, 459: 249-289.

[13] T Hu. *Online regression with varying Gaussians and non-identical distributions,* Anal Appl, 2011, 9(4): 395-408.

[14] T Hu, D X Zhou. *Online classification with varying Gaussians,* Appl Math Comput, 2009, 124(1): 65-83.

[15] T Hu, D X Zhou. *Online learning with samples drawn from non-identical distributions,* J Mach Learn Res, 2009, 10(12): 2873-2898.

[16] J Langford, L Li, T Zhang. *Sparse online learning via truncated gradient,* J Mach Learn Res, 2009, 10(28): 777-801.

[17] Y Lei, L Shi, Z C Guo. *Convergence of unregularized online learning algorithms,* J Mach Learn Res, 2018, 18(171): 1-33.

[18] W Mo, Z Qi, Y Liu. *Learning optimal distributionally robust individualized treatment rules,* J Amer Statist Assoc, 2021, 116(534): 659-674.

[19] S A Murphy, M Qian. *Performance guarantees for individualized treatment rules,* Ann Statist, 2011, 39 (2): 1180-1210.

[20] J M Robins. *Optimal Structural Nested Models for Optimal Sequential Decisions,* In: D Y Lin, P J Heagerty(eds). *Proceedings of the Second Seattle Symposium in Biostatistics*, Lecture Notes in Statistics, New York: Springer, 2014, 179: 189-326.

[21] N J Schork. *Personalized medicine: Time for one-person trials,* Nature, 2015, 520: 609-611.

[22] S Smale, D X Zhou. *Online learning with Markov sampling,* Anal Appl, 2009, 7(1): 87-113.

[23] I Steinwart, D Hush, C Scovel. *An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels,* IEEE Trans Inf Theory, 2006, 52(10): 4635-4643.

[24] B B Wang, T Hu. *Unregularized Online algorithms with Varying Gaussians,* Constr Approx, 2021, 53(3): 403-440.

[25] C Wang, T Hu. *Online minimum error entropy algorithm with unbounded sampling,* Anal Appl, 2019, 17(2): 293-322.

[26] S H Xie, T Tarpey, E Petkova, et al. *Multiple domain and multiple kernel outcome-weighted learning for estimating individualized treatment regimes,* J Comput Graph Statist, 2022, 31(4): 1375-1383.

[27] A Yang, J Fan, D H Xiang. *Online outcome weighted learning with general loss functions,* J Complexity, 2025, 88: 101931.

[28] Z Yang, Y Lei, P Wang, et al. *Simple stochastic and online gradient descent algorithms for pairwise learning,* NIPS'21, 2021, 34: 20160-20171.

[29] G B Ye, D X Zhou. *Fully online classification by regularization,* Appl Comput Harmon Anal, 2007, 23(2): 198-214.

[30] Y M Ying, M Pontil. *Online gradient descent learning algorithms,* Found Comput Math, 2008, 8(5): 561-596.

[31] Y M Ying, D X Zhou. *Unregularized online learning algorithms with general loss functions,* Appl Comput Harmon Anal, 2017, 42(2): 224-244.

[32] X Zhang, M Hong, S Dhople, et al. *FedPD: A federated learning framework with adaptivity to non-IID data,* IEEE Trans Signal Process, 2021, 69: 6055-6070.

[33] Y Q Zhao, D L Zeng, A J Rush, et al. *Estimating individualized treatment rules using outcome weighted learning,* J Amer Statist Assoc, 2012, 107: 1106-1118.

[34] D X Zhou. *Derivative reproducing properties for kernel methods in learning theory,* J Comput Appl Math, 2008, 220: 456-463.

[35] X Zhou, N Mayer-Hamblett, U Khan, et al. *Residual weighted learning for estimating individualized treatment rules,* J Amer Statist Assoc, 2017, 112(517): 169-187.

[36] X X Zhu, D H Xiang. *Error analysis of OWL Algorithms with varying gaussians and convex loss,* Zhejiang Normal University, Master's thesis, 2022.

[1]School of Mathematical Sciences, Zhejiang Normal University, Jinhua 321004, China.
[2]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Jinhua 321004, China.

Emails: al.yang@zjnu.edu.cn, daohongxiang@zjnu.cn

[3]Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.

Email: junfan@hkbu.edu.hk