Journal of Machine Learning Research 18 (2018) 1-33 Submitted 8/17; Published 4/18

Convergence of Unregularized Online Learning Algorithms

Yunwen Lei YUNWELEI@CITYU.EDU.HK
Shenzhen Key Laboratory of Computational Intelligence
Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen, 518055, China
and
Department of Mathematics
City University of Hong Kong
Kowloon, Hong Kong, China

Lei Shi LEISHI@FUDAN.EDU.CN
School of Mathematical Sciences
Shanghai Key Laboratory for Contemporary Applied Mathematics
Fudan University
Shanghai, 200433, China

Zheng-Chu Guo GUOZHENGCHU@ZJU.EDU.CN
School of Mathematical Sciences
Zhejiang University
Hangzhou, 310027, China

Editor: Andreas Christmann

Abstract

In this paper we study the convergence of online gradient descent algorithms in reproducing kernel Hilbert spaces (RKHSs) without regularization. We establish a sufficient condition and a necessary condition for the convergence of excess generalization errors in expectation. A sufficient condition for the almost sure convergence is also given. With high probability, we provide explicit convergence rates of the excess generalization errors for both averaged iterates and the last iterate, which in turn also imply convergence rates with probability one. To our best knowledge, this is the first high-probability convergence rate for the last iterate of online gradient descent algorithms in the general convex setting. Without any boundedness assumptions on iterates, our results are derived by a novel use of two measures of the algorithm's one-step progress, respectively by generalization errors and by distances in RKHSs, where the variances of the involved martingales are cancelled out by the descent property of the algorithm.

Keywords: Learning theory, Online learning, Convergence analysis, Reproducing kernel Hilbert space

1. Introduction

Online gradient descent is a scalable method able to tackle large-scale data arriving in a sequential manner (Zhang, 2004; Kivinen et al., 2004; Duchi and Singer, 2009; Dieuleveut and Bach, 2016), which is becoming ubiquitous within the big data era. As a first-order method, it iteratively builds an unbiased estimate of the true gradient upon the arrival of

©2018 Yunwen Lei, Lei Shi and Zheng-Chu Guo.

License: CC-BY 4.0, see `https://creativecommons.org/licenses/by/4.0/`. Attribution requirements are provided at `http://jmlr.org/papers/v18/17-457.html`.

a new example and uses this information to guide the learning process (Zinkevich, 2003; Zhang, 2004). As verified by theoretical and empirical analysis, online gradient descent enjoys comparable performance as compared to its batch counterpart such as gradient descent (Zhang, 2004; Yao, 2010; Shalev-Shwartz et al., 2011), while attaining a great computational speed-up since its gradient calculation involves only a single example. As a comparison, the gradient calculation in gradient descent requires to traverse all training examples. Recently, online gradient descent has received renewed attention due to the wide applications of its stochastic analogue, i.e., stochastic gradient descent, in training deep neural networks (Bottou, 1991; Ngiam et al., 2011; Sutskever et al., 2013).

In this paper, we are interested in the setting that training examples $\{z_t = (x_t, y_t)\}_{t\in\mathbb{N}}$ are sequentially and identically drawn from a probability measure ρ defined in the sample space $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the input space and $\mathcal{Y} \subset \mathbb{R}$ is the output space. We focus on the nonparametric setting, where the learning process is implemented in a reproducing kernel Hilbert space (RKHS) H_K associated with a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is assumed to be continuous, symmetric and positive semi-definite. The space H_K is defined as the completion of the linear span of the set of functions $\{K_x(\cdot) := K(x, \cdot) : x \in \mathcal{X}\}$ with the inner producing satisfying the reproducing property $f(x) = \langle f, K_x\rangle$ for any $x \in \mathcal{X}$ and $f \in H_K$. In this setting, the use of Mercer kernels provides a unifying way to measure similarities between pairs of objects (Cortes and Vapnik, 1995; Müller et al., 2001; Steinwart, 2001; Schölkopf and Smola, 2001), which turns out to be a key to the great success of kernel methods in many practical learning problems. We wish to build a prediction rule $f \in H_K$ after seeing a sequence of training examples, the performance of which at an example (x, y) can be quantitatively measured by a loss function $\phi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ as $\phi(y, f(x))$. With a sequence $\{\eta_t\}_{t\in\mathbb{N}}$ of positive step sizes and $f_1 = 0$, online gradient descent is a realization of learning schemes by keeping a sequence of iterates as follows

$$f_{t+1} = f_t - \eta_t\phi'(y_t, f_t(x_t))K_{x_t}, \quad \forall t \in \mathbb{N}, \tag{1.1}$$

where ϕ' denotes the derivative of ϕ with respect to the second argument. Although our focus is on the nonparametric setting, it should be mentioned that the above algorithm also recovers the parametric case in which the kernel is taken to be the linear kernel with $K_x(x') = \langle x, x'\rangle, \forall x, x' \in \mathcal{X}$, to which our results also apply.

Despite its widespread applications, the theoretical understanding of the online gradient descent algorithms is still not satisfactory in the following three aspects. Firstly, boundedness assumptions on the iterates are often imposed in the literature, which may be violated in practical implementations if the underlying domain is not bounded. Although a projection of iterates onto a bounded domain guarantees the boundedness assumption, the projection operator may be time-consuming and this introduces an additional challenging problem of tuning the size of the domain. Secondly, most of the theoretical results are stated in expectation, while we are sometimes more interested in either almost sure convergence or convergence rates with high probability. Indeed, an algorithm may suffer from a high variability and should be used with caution if neither almost sure convergence nor high-probability bounds hold (Shamir and Zhang, 2013). In particular, an almost sure convergence is still lacking for online gradient descent algorithms applied to general convex problems (Ying and Zhou, 2017). Lastly, most existing convergence rates are stated for some average of iterates. Though taking average of iterates can improve the robustness of the solution (Nemirovski et al., 2009), it can either destroy the sparsity of the solution which is crucial for a proper interpretation of models in many applications, or slow down the training speed in practical implementations (Rakhlin et al., 2012).

In this paper, we aim to take a further step to tackle the above mentioned problems. We establish a general sufficient condition and a necessary condition on the step sizes for the convergence of online gradient descent algorithms in expectation. With Doob's martingale convergence theorem and the Borel-Cantelli lemma, a sufficient condition for the almost sure convergence and explicit convergence rates with probability one are also established. Furthermore, we present high-probability bounds for both averaged iterates and the last iterate of online gradient descent algorithms. To our best knowledge, this is the first high-probability convergence rate for the last iterate of online gradient descent algorithms in the general convex setting. Our analysis does not impose any boundedness assumptions on the iterates. Indeed, we show that, although implemented in an unbounded domain, the iterates produced by (1.1) fall into a bounded domain with high probability (up to logarithmic factors). Our analysis is performed by viewing the one-step progress of online gradient descent algorithms from different yet unified perspectives: one in terms of generalization errors and one in terms of RKHS distances. For both viewpoints, we relate the one-step progress to a martingale difference sequence and a negative term due to the descent nature of the algorithm. Our novelty is to show that the dominant variance term appearing in the application of a Bernstein-type inequality to these martingales can be cancelled out by the negative terms in the one-step progress inequalities. Both viewpoints of the one-step progress are indispensable in our analysis.

The remaining parts of this paper are organized as follows. We present main results in Section 2. Discussions and comparisons with related work are given in Section 3. The proofs of main results are given in Section 4.

2. Main Results

Our convergence rates are stated for generalization errors, which, for a prediction rule $f : \mathcal{X} \to \mathbb{R}$, are defined as the expected error $\mathcal{E}(f) = \int_{\mathcal{Z}} \phi(y, f(x))d\rho$ incurred from using f to perform prediction. Our analysis requires to impose mild assumptions on the loss functions.

Assumption 1 *We assume the loss function $\phi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$ is convex and differentiable with respect to the second argument. Let $\alpha \in (0, 1]$ and $L > 0$ be two constants. We assume that the gradients of ϕ are (α, L)-Hölder continuous in the sense*

$$|\phi'(y, s) - \phi'(y, \tilde{s})| \leq L|s - \tilde{s}|^\alpha, \quad \forall s, \tilde{s} \in \mathbb{R}, \forall y \in \mathcal{Y}. \tag{2.1}$$

We say ϕ is smooth if it satisfies (2.1) with $\alpha = 1$. Loss functions satisfying Assumption 1 are wildly used in machine learning. Smooth loss functions include the least squares loss $\phi(y, a) = \frac{1}{2}(y-a)^2$ and the Huber loss $\phi(y, a) = \frac{1}{2}(y-a)^2$ if $|y-a| \leq 1$ and $|y-a|-\frac{1}{2}$ otherwise for regression, as well as the logistic loss $\phi(y, a) = \log(1 + \exp(-ya))$ and the quadratically smoothed hinge loss $\phi(y, a) = \max\{0, 1 - ya\}^2$ for classification (Zhang, 2004). If $p \in (1, 2]$, both the p-norm hinge loss $\phi(y, a) = \max\{0, 1 - ya\}^p$ for classification and the p-th power absolute distance $\phi(y, a) = |y - a|^p$ for regression satisfy (2.1) with $\alpha = p - 1$ (Chen et al., 2004; Steinwart and Christmann, 2008).

Throughout this paper, we assume that a minimizer $f_H = \arg\min_{f \in H_K} \mathcal{E}(f)$ exists in H_K. We also assume

$$\max\big\{\sup_{y\in\mathcal{Y}} \phi(y,0), \sup_{z\in\mathcal{Z}} \phi(y, f_H(x))\big\} < \infty \text{ and } \kappa := \sup_{x\in\mathcal{X}} \sqrt{K(x,x)} < \infty,$$

which is satisfied if the sample space $\mathcal{Z}$ is bounded. Denote $\|\cdot\|$ as the norm in H_K. We always use the notation $A_t = \mathbb{E}[\mathcal{E}(f_t)] - \mathcal{E}(f_H)$ and $\hat{A}_t = \mathcal{E}(f_t) - \mathcal{E}(f_H), \forall t \in \mathbb{N}$ for brevity, which are referred to as the expected excess generalization errors and excess generalization errors, respectively.

In the following, we present the main results of this paper. We consider three types of convergence: convergence in expectation, almost sure convergence and convergence rates with high probability.

2.1 Convergence in Expectation

The first part of our main results to be proved in Section 4.1 establishes a general sufficient condition (Theorem 1) and a necessary condition (Theorem 2) on the step size sequence $\{\eta_t\}_{t\in\mathbb{N}}$ for the convergence of A_t to zero.

Theorem 1 *Let $\{f_t\}_{t\in\mathbb{N}}$ be the sequence produced by* (1.1) *and suppose Assumption 1 holds with $\alpha \in (0,1]$. If*

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \lim_{t\to\infty} \eta_t^{\alpha} \sum_{k=1}^{t} \eta_k^2 = 0, \tag{2.2}$$

then $\lim_{t\to\infty} \mathbb{E}[\mathcal{E}(f_t)] - \mathcal{E}(f_H) = 0$.

Theorem 2 *Let $\{f_t\}_{t\in\mathbb{N}}$ be the sequence produced by* (1.1)*. Suppose that for any $y \in \mathcal{Y}$, the function $\phi(y,\cdot) : \mathbb{R} \to \mathbb{R}_+$ is convex and its derivative $\phi'(y,\cdot)$ is $(1,L)$-Hölder continuous. Assume that the step size sequence satisfies $\eta_t \le 1/(6L\kappa^2), \forall t \in \mathbb{N}$ and $\mathcal{E}(f_1) \neq \mathcal{E}(f_H)$. If $\lim_{t\to\infty} \mathbb{E}[\mathcal{E}(f_t)] = \mathcal{E}(f_H)$, then $\sum_{t=1}^{\infty} \eta_t = \infty$.*

Remark 3 *We now illustrate the above theorems by considering the polynomially decaying step sizes $\eta_t = \eta_1 t^{-\theta}, t \in \mathbb{N}, \theta \ge 0$. The condition $\sum_{t=1}^{\infty} \eta_t = \infty$ requires $\theta \le 1$, while the condition $\lim_{t\to\infty} \eta_t^{\alpha} \sum_{k=1}^{t} \eta_k^2 = 0$ requires $\theta > \frac{1}{2+\alpha}$. Therefore, Theorem 1 shows that the iteration scheme* (1.1) *with $\eta_t = \eta_1 t^{-\theta}$ and $\theta \in \big(\frac{1}{2+\alpha}, 1\big]$ guarantees the convergence of $\{A_t\}_{t\in\mathbb{N}}$. Theorem 2 shows that the condition $\theta \le 1$ is also necessary for the convergence.*

2.2 Almost Sure Convergence

The second part of our main results focuses on a sufficient condition (Theorem 4) for the almost sure convergence of $\{\hat{A}_t\}_{t\in\mathbb{N}}$ to zero and convergence rates with probability 1 (Theorem 6). The proofs of results in this section can be found in Section 4.2.

Theorem 4 *Let $\{f_t\}_{t\in\mathbb{N}}$ be the sequence given by* (1.1)*. If Assumption 1 holds with $\alpha \in (0,1]$ and the step size sequence satisfies*

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^{1+\alpha} < \infty, \tag{2.3}$$

then $\lim_{t\to\infty}\mathcal{E}(f_t)=\mathcal{E}(f_H)$ *almost surely.*

Remark 5 *According to Theorem 4, we know that* $\{\hat{A}_t\}_{t\in\mathbb{N}}$ *would converge almost surely to* 0 *if we consider either the step sizes* $\eta_t=\eta_1 t^{-\theta}$ *with* $\theta\in(\frac{1}{1+\alpha},1]$ *or the step sizes* $\eta_t=\eta_1(t\log^\beta t)^{-\frac{1}{1+\alpha}}$ *with* $\beta>1$*. Specifically, if the loss function is smooth, then we can choose either* $\eta_t=\eta_1 t^{-\theta}$ *with* $\theta\in(\frac{1}{2},1]$ *or* $\eta_t=\eta_1\big(t\log^\beta t\big)^{-\frac{1}{2}}$ *with* $\beta>1$ *to guarantee the convergence of the algorithm* (1.1) *almost surely in the sense of generalization errors.*

Theorem 6 *Suppose that Assumption 1 holds with* $\alpha\in(0,1]$*. Let* $\{f_t\}_{t\in\mathbb{N}}$ *be the sequence given by* (1.1) *with* $\eta_t=\eta_1 t^{-\theta},\theta\in(\frac{1}{\alpha+1},1)$ *and* $\eta_1\leq\frac{1}{A\kappa^2}$ *(*A *is defined in* (4.27)*). Then for any* $\epsilon>0$*,*

$$\lim_{t\to\infty} t^{\min\{(1-\theta),(\alpha+1)\theta-1\}-\epsilon}\hat{A}_t=0 \text{ almost surely.} \tag{2.4}$$

Specifically, if we choose $\theta=\frac{2}{2+\alpha}$*, then* $\lim_{t\to\infty}t^{\frac{\alpha}{2+\alpha}-\epsilon}\hat{A}_t=0$ *almost surely.*

2.3 Convergence Rates with High Probability

The last part of our main results is on high-probability bounds for the excess generalization errors, the proof of which is given in Section 4.3. With high probability, Theorem 7 establishes the boundedness (up to logarithmic factors) of the weighted summation $\sum_{t=1}^T\eta_t\hat{A}_t$, from which the decay rate of the excess generalization error $\mathcal{E}(\bar{f}_T^\eta)-\mathcal{E}(f_H)$ associated to a weighted average of the iterates $\bar{f}_T^\eta:=\frac{\sum_{t=1}^T\eta_t f_t}{\sum_{t=1}^T\eta_t}$ follows directly.

Theorem 7 *Let* $\{f_t\}_{t\in\mathbb{N}}$ *be the sequence given by* (1.1)*. Suppose that Assumption 1 holds with* $\alpha\in(0,1]$*. Assume the step size sequence satisfies* $\eta_t\leq\frac{1}{A\kappa^2},\eta_{t+1}\leq\eta_t$ *for all* $t\in\mathbb{N}$ *and* $\sum_{t=1}^\infty\eta_t^2<\infty$*. Then, there exists a constant* $\widetilde{C}$ *independent of* T *(explicitly given in the proof) such that for any* $\delta\in(0,1)$ *the following inequality holds with probability at least* $1-\delta$

$$\sum_{t=1}^T\eta_t\big[\mathcal{E}(f_t)-\mathcal{E}(f_H)\big]\leq\widetilde{C}\log^{\frac{3}{2}}\frac{2T}{\delta}\quad\text{and}\quad\mathcal{E}(\bar{f}_T^\eta)-\mathcal{E}(f_H)\leq\frac{\widetilde{C}\log^{\frac{3}{2}}\frac{2T}{\delta}}{\sum_{t=1}^T\eta_t}. \tag{2.5}$$

Remark 8 *For the step size sequence* $\eta_t=\eta_1 t^{-\theta},\theta>\frac{1}{2}$*, Theorem 7 implies that* $\mathcal{E}(\bar{f}_T^\eta)-\mathcal{E}(f_H)=O\big(T^{\theta-1}\log^{\frac{3}{2}}\frac{T}{\delta}\big)$ *with probability at least* $1-\delta$*. If we consider* $\eta_t=\eta_1\big(t\log^\beta(et)\big)^{-\frac{1}{2}}$ *with* $\beta>1$*, then with probability* $1-\delta$ *we have* $\mathcal{E}(\bar{f}_T^\eta)-\mathcal{E}(f_H)=O\big(T^{-\frac{1}{2}}\log^{\frac{3+\beta}{2}}\frac{T}{\delta}\big)$*.*

A key feature of Theorem 7 distinguishing it from the existing results is that it avoids boundedness assumptions on the iterates, which are always imposed in the literature (Nemirovski et al., 2009; Duchi et al., 2010). Indeed, an essential ingredient in proving Theorem 7 is to show that $\{f_t\}_{t\in\mathbb{N}}$ produced by (1.1) would fall into a bounded ball of H_K (up to logarithmic factors) with high probability, as shown in the following proposition.

Proposition 9 *Suppose assumptions in Theorem 7 hold. Then, there exists a constant* $\bar{C}\geq1$ *independent of* T *(explicitly given in the proof) such that for any* $\delta\in(0,1)$ *the following inequality holds with probability at least* $1-\delta$

$$\max_{1\leq t\leq T}\|f_t-f_H\|^2\leq\bar{C}\log\frac{T}{\delta}.$$

A key ingredient to prove Proposition 9 is to establish the following one-step progress inequality in terms of the RKHS distances (see (4.37))

$$\|f_{t+1}-f_H\|^2 \leq \|f_t-f_H\|^2 + C\eta_t^2 + 2\eta_t\big(\mathcal{E}(f_H)-\mathcal{E}(f_t)\big)+\xi_t,$$

where C is a constant and $\{\xi_t\}_{t\in\mathbb{N}}$ is a Martingale difference sequence. Our novelty in applying a Bernstein-type inequality to control the martingale $\sum_{t=1}^T \xi_t$ is to show that the associated variances can be cancelled out by the negative term $2\sum_{t=1}^T \eta_t\big(\mathcal{E}(f_H)-\mathcal{E}(f_t)\big)$ (see (4.38) and the last inequality of Proposition 23). Although Theorem 7 only considers the behavior of the weighted average $\bar{f}_T^{\eta}$ of iterates, it is possible to establish similar convergence rates for the uniform average of iterates $\bar{f}_T := \frac{1}{T}\sum_{t=1}^T f_t$ (Proposition 24).

Theorem 10 establishes a general high-probability bound for the excess generalization error of the last iterate in terms of the step size sequence.

Theorem 10 *Suppose that the assumptions in Theorem 7 hold. Then, there exists a constant $\widetilde{C}'$ independent of T (explicitly given in the proof) such that for any $\delta\in(0,1)$ the following inequality holds with probability at least $1-\delta$*

$$\mathcal{E}(f_{T+1})-\mathcal{E}(f_H) \leq \widetilde{C}' \max\Big\{\big[\sum_{t=\lfloor\frac{T}{2}\rfloor}^{T}\eta_t\big]^{-1}, \eta_{\lfloor\frac{T}{2}\rfloor}, \sum_{t=\lfloor\frac{T}{2}\rfloor}^{T}\eta_t^{1+\alpha}\Big\}\log^2\frac{3T}{\delta}, \tag{2.6}$$

where $\lfloor\frac{T}{2}\rfloor$ denotes the largest integer not greater than $\frac{T}{2}$.

To establish high-probability error bounds for the last iterate of online gradient descent algorithm is an interesting problem which is not well studied, to our best knowledge, in the general convex setting. The key ingredient in our analysis is the following one-step progress inequality in terms of generalization errors (see (4.47))

$$\hat{A}_{t+1} \leq \hat{A}_t - \eta_t\|\nabla\mathcal{E}(f_t)\|^2 + \bar{\xi}_t + C\eta_t^{1+\alpha},$$

where C is a constant and $\{\bar{\xi}_t\}$ is a martingale difference sequence. A key observation of our analysis is that the variance of the martingale $\sum_{t=1}^T\bar{\xi}_t$ can be cancelled out by the negative term $-\sum_{t=1}^T\eta_t\|\nabla\mathcal{E}(f_t)\|^2$ in the above one-step progress inequality (see (4.48) and (4.52)), paving the way for the application of a Bernstein-type inequality for martingales.

We can derive explicit convergence rates in Corollary 11 by considering polynomially decaying step sizes in Theorem 10.

Corollary 11 *Let $\{f_t\}_{t\in\mathbb{N}}$ be the sequence given by (1.1) with $\eta_t=\eta_1 t^{-\theta}, \theta\in(\frac{1}{2},1)$ and $\eta_1\leq\frac{1}{A\kappa^2}$ If Assumption 1 holds and $\delta\in(0,1)$, then the following inequality holds with probability $1-\delta$*

$$\mathcal{E}(f_{T+1})-\mathcal{E}(f_H) = O\Big(T^{\max\big\{\theta-1, 1-(1+\alpha)\theta\big\}}\log^2\frac{T}{\delta}\Big).$$

If we choose $\theta=\frac{2}{2+\alpha}$, then with probability at least $1-\delta$ we derive $\mathcal{E}(f_{T+1})-\mathcal{E}(f_H) = O\Big(T^{-\frac{\alpha}{2+\alpha}}\log^2\frac{T}{\delta}\Big)$.

Remark 12 *It should be mentioned that, unlike Theorem 7, the convergence rates in Corollary 11 depend on the smoothness parameter α and are not able to attain the minimax optimal convergence rate $O(T^{-\frac{1}{2}})$ (Agarwal et al., 2009). Indeed, for smooth loss functions, Corollary 11 establishes the convergence rate $O\big(T^{-\frac{1}{3}}\log^2\frac{T}{\delta}\big)$ with high probability, which matches the bounds in-expectation $A_T = O(T^{-\frac{1}{3}})$ up to logarithmic factors established in Moulines and Bach (2011); Ying and Zhou (2017). It remains a challenging problem to further improve the high-probability bounds for $\hat{A}_T$.*

3. Discussions

In this section, we discuss related work on convergence of online/stochastic gradient descent algorithms from three viewpoints: convergence in expectation, almost sure convergence and convergence rates with high probability.

3.1 Related Work on Convergence in Expectation

Most studies of online gradient descent algorithms focus on convergence in expectation (Zhang, 2004; Ying and Zhou, 2006; Duchi and Singer, 2009; Shamir and Zhang, 2013; Lin et al., 2016; Hardt et al., 2016; Ying and Zhou, 2017). Convergence rates $O(T^{-\frac{1}{2}})$ were established for some averaged iterates produced by (1.1) in a parametric setting with the linear kernel $K_x = x$ (Zhang, 2004). These results were extended to online gradient descent algorithms in RKHSs with the specific least squares loss function (Ying and Pontil, 2008; Dieuleveut and Bach, 2016; Guo and Shi, 2017), and online mirror descent algorithms performing updates in Banach spaces (Duchi et al., 2010). Under boundedness assumptions on the iterates and (sub)gradients, the convergence rate $O(T^{-\frac{1}{2}}\log T)$ was established for the expected excess generalization error of the last iterate (Shamir and Zhang, 2013). Recently, a general condition on the step sizes as (2.3) was established for the convergence of the algorithm (1.1), in the sense $\lim_{t\to\infty} A_t = 0$, with loss functions satisfying Assumption 1 (Ying and Zhou, 2017). This sufficient condition is stricter than our condition (2.2). To see this clearly, we consider the polynomially decaying step sizes $\eta_t = \eta_1 t^{-\theta}$, for which the condition (2.3) requires $\theta \in (\frac{1}{1+\alpha}, 1]$ while our condition (2.2) requires $\theta \in (\frac{1}{2+\alpha}, 1]$. Furthermore, our discussion also implies a necessary condition for the convergence in expectation.

Implemented in either a parametric or a nonparametric setting, regularized online learning algorithms have also received considerable attention (Kivinen et al., 2004; Smale and Yao, 2006; Ying and Zhou, 2006; Smale and Zhou, 2009), which differ from (1.1) by introducing a regularization term to avoid overfitting. This algorithm updates iterates as follows

$$f_{t+1} = (1-\lambda\eta_t)f_t - \eta_t\phi'(y_t, f_t(x_t))K_{x_t}, \tag{3.1}$$

where $\lambda > 0$ is a regularization parameter and the term $\lambda f_t + \phi'(y_t, f_t(x_t))K_{x_t}$ is used as an unbiased estimator of the gradient for the regularized generalization error $\mathcal{E}^\lambda(f) := \mathcal{E}(f) + \frac{\lambda}{2}\|f\|^2$ at $f = f_t$. Convergence rates in expectation can be stated for either the excess regularized generalization error $\mathcal{E}^\lambda(f_T) - \mathcal{E}^\lambda(f_\lambda)$ (Shamir and Zhang, 2013) or the RKHS distance $\|f_T - f_\lambda\|$ (Smale and Yao, 2006; Ying and Zhou, 2006; Yao, 2010), where $f_\lambda = \arg\min_{f\in H_K}\mathcal{E}^\lambda(f)$ is the minimizer of the regularized generalization error. When the loss function is smooth, a sufficient and necessary condition as

$$\lim_{t\to\infty}\eta_t = 0 \quad \text{and} \quad \sum_{t=1}^{\infty}\eta_t = \infty \tag{3.2}$$

was recently established for the convergence of $\{\mathbb{E}[\|f_t - f_\lambda\|^2]\}_{t\in\mathbb{N}}$ to zero in the parametric case (Lei and Zhou, 2017). A disadvantage of the regularization scheme (3.1) is that it requires to tune two sequences of hyper-parameters: a regularization parameter and the step sizes. As a comparison, an implicit regularization can be attained in the unregularized scheme (1.1) by tuning only the step sizes.

3.2 Related Work on Almost Sure Convergence

Existing almost sure convergence of online learning algorithm is mainly stated for the RKHS distances, which requires to impose some type of strong convexity assumption on the objective function $\mathcal{E}(f)$. In the parametric setting with the learning scheme (1.1), a sufficient condition as

$$\sum_{t=1}^{\infty}\eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty}\eta_t^2 < \infty$$

was established for the almost sure convergence of $\|f_t - f_H\|^2$ if the objective function attains a unique minimizer and satisfies (Bottou, 1998)

$$\inf_{\|f-f_H\|^2>\epsilon}\langle f - f_H, \nabla\mathcal{E}(f)\rangle > 0, \quad \forall\epsilon > 0,$$
$$\mathbb{E}_Z\big[\|\phi'(Y, f(X))K_X\|^2\big] \le \widetilde{A} + \widetilde{B}\|f - f_H\|^2, \quad \forall f \in H_K,$$

where $\widetilde{A}$ and $\widetilde{B}$ are two constants. This result was extended to the online mirror descent setting under some convexity assumption on the objective function measured by Bregman distances induced by the associated mirror map (Lei and Zhou, 2017). For polynomially decaying step sizes $\eta_t = \eta_1 t^{-\theta}$ with $\theta \in (0, 1)$, almost sure convergence of $\|f_t - f_\lambda\|$ was shown for regularized online learning algorithms (3.1) specified to the least squares loss function (Yao, 2010). The analysis in Yao (2010) roots its foundation on the martingale decompositions of the reminders $f_t - f_\lambda$, which only holds in the least squares regularization setting. Almost sure convergence was recently studied for the randomized Kaczmarz algorithm (Lin and Zhou, 2015), which is an instantiation of (1.1) with $\phi(y, a) = \frac{1}{2}(y - a)^2$ and $K_x = x$. The analysis there heavily depends on a restricted strong convexity of the objective function in a linear subspace where the learning takes place, which can not apply to general loss functions. As compared to the above mentioned results, our almost sure convergence is stated for the excess generalization errors with general loss functions and requires no assumptions on the strong convexity of the objective function $\mathcal{E}(f)$.

3.3 Related Work on Convergence Rates with High Probability

In this section, we survey related work on convergence rates with high probability. We divide our discussions into two parts according to the convexity of the objective function.

As far as we know, all existing high-probability convergence rates of online gradient descent algorithms with general convex functions focus on some average of iterates (here we are not interested in probabilistic bounds with a polynomial dependence on $1/\delta$). The following online projected gradient descent algorithm with $K_x = x$ was studied in Nemirovski et al. (2009); Duchi et al. (2010)

$$f_{t+1} = \mathrm{Proj}_{\widetilde{H}}\Big[f_t - \eta_t \phi'(y_t, f_t(x_t))K_{x_t}\Big], \tag{3.3}$$

where $\widetilde{H}$ is a compact subset of H_K and $\mathrm{Proj}_{\widetilde{H}}(f) = \arg\min_{\tilde{f}\in\widetilde{H}} \|f - \tilde{f}\|$ is the projection of f onto $\widetilde{H}$. Under the boundedness assumption

$$\mathbb{E}\Big[\exp\big[\|\phi'(y, f(x))K_x\|^2/G^2\big]\Big] \le \exp(1) \quad \forall f \in \widetilde{H},$$

it was shown that the weighted average $\bar{f}^{\eta}_T = \frac{\sum_{t=1}^T \eta_t f_t}{\sum_{t=1}^T \eta_t}$ of iterates produced by (3.3) with a constant step size satisfies the following inequality with probability $1-\delta$

$$\mathcal{E}(\bar{f}^{\eta}_T) - \mathcal{E}(f_H) = O\big(GDT^{-\frac{1}{2}}\log\delta^{-1}\big),$$

where $D = \sup_{f,\tilde{f}\in\widetilde{H}} \|f - \tilde{f}\|$ is the diameter of the subspace $\widetilde{H}$. Under a stronger assumption $\|\phi'(y, f(x))K_x\| \le G$ for all $(x, y) \in \mathcal{Z}, f \in \widetilde{H}$, the uniform average $\bar{f}_T = \frac{1}{T}\sum_{t=1}^T f_t$ of iterates produced by (3.3) with step sizes $\eta_t = \eta_1 t^{-\frac{1}{2}}$ was shown to enjoy the bound $\mathcal{E}(\bar{f}_T) - \mathcal{E}(f_H) = O(DGT^{-\frac{1}{2}}\log^{\frac{1}{2}}\frac{1}{\delta})$ with probability at least $1-\delta$. In comparison with these results, the convergence rates in Theorem 7 are derived without the projection step and any boundedness assumption on the gradients. Indeed, most of the efforts in proving Theorem 7 is to show $\|f_t - f_H\|^2 = O(\log\frac{T}{\delta})$ with probability at least $1-\delta$. It is implied that the possibly computationally expensive projection step can be removed without harming the behavior of the online gradient descent algorithms. Furthermore, Theorem 10 gives, to our best knowledge, the first high-probability bounds for the last iterate of online gradient descent algorithms in the general convex setting. A framework to transfer regret bounds of online learning algorithms to high-probability bounds for the uniform average of iterates was established by Cesa-Bianchi et al. (2004).

Now we review some high-probability studies for online gradient descent algorithms in the strongly convex setting, for which some results for the last iterate can be found in the literature. For the online regularized algorithm (3.1) with the least squares loss function and $\eta_t = \eta_1 t^{-\theta}, \theta \in [0, 1)$, the following inequality was derived with probability at least $1-\delta$ (Yao, 2010)

$$\|f_T - f_\lambda\|^2 = O\Big(\lambda^{-2+\frac{1}{1-\theta}}T^{-\theta}\log\frac{1}{\delta}\Big).$$

The analysis in Yao (2010) is based on an integral operator approach, which can not be extended to general loss functions. Under almost sure boundedness assumption $\|(\phi'(y_t, f_t(x_t)) + \lambda)K_{x_t}\| \le G$ for all $t \in \mathbb{N}$, the following improved bound for the last iterate of (3.1) with general loss functions and step sizes $\eta_t = \eta_1(t\lambda)^{-1}$ was established with probability at least $1-\delta$ (Rakhlin et al., 2012)

$$\|f_T - f_\lambda\|^2 = O\Big(G^2\lambda^{-2}T^{-1}\log\frac{\log T}{\delta}\Big). \tag{3.4}$$

Although this bound enjoys a tight dependence on T, its dependence on the regularization parameter λ is suboptimal. To make a clear comparison between this result and ours, we consider here the specific least squares loss function and assume that the regression function $f_\rho(x) := \mathbb{E}[Y|X=x]$ belongs to H_K. In this case, Lemma 13 translates (3.4) to the following high-probability bounds on excess generalization errors

$$\mathcal{E}(f_T) + \frac{\lambda}{2}\|f_T\|^2 = \mathcal{E}(f_\lambda) + \frac{\lambda}{2}\|f_\lambda\|^2 + O\Big(G^2\lambda^{-2}T^{-1}\log\frac{\log T}{\delta}\Big). \tag{3.5}$$

The assumption $f_\rho \in H_K$ implies $D(\lambda) := \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \frac{\lambda}{2}\|f_\lambda\|^2 = O(\lambda)$ (Cucker and Zhou, 2007) and therefore (3.5) reads as

$$\begin{aligned}\mathcal{E}(f_T) - \mathcal{E}(f_\rho) &= \Big(\mathcal{E}(f_T) - \mathcal{E}(f_\lambda) - \frac{\lambda}{2}\|f_\lambda\|^2\Big) + \Big(\mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho) + \frac{\lambda}{2}\|f_\lambda\|^2\Big)\\ &= O\Big(G^2\lambda^{-2}T^{-1}\log\frac{\log T}{\delta}\Big) + O(\lambda).\end{aligned}$$

If we choose $\lambda = c\big(G^2T^{-1}\log\frac{\log T}{\delta}\big)^{\frac{1}{3}}$ for a constant $c>0$, then the above inequality translates to $\mathcal{E}(f_T) - \mathcal{E}(f_\rho) = O\Big(\big(G^2T^{-1}\log\frac{\log T}{\delta}\big)^{\frac{1}{3}}\Big)$, which matches our convergence rates up to logarithmic factors. Note that the regularization parameter λ needs to be tuned according to T to balance the bias and variance in (3.5), which may not be accessible in practical implementations. To deal with this issue, a class of fully online regularized algorithms was proposed and investigated by allowing the regularization parameters to vary along the learning process (Ye and Zhou, 2007; Tarres and Yao, 2014). As a comparison, without a regularization parameter to tune, the unregularized online learning algorithm (1.1) achieves a bias-variance balance by tuning only the step sizes. Furthermore, the convergence rates (3.4) require to impose the non-intuitive boundedness assumptions on the gradients encountered during the iterations, which may be violated in practical implementations. This boundedness assumption is removed in our analysis.

4. Proofs

In this section, we present the proofs for the results given in Section 2. Our discussions require to use a property established in the following lemma on functions with (α, L)-Hölder continuous gradients. This lemma is motivated by similar results in the literature (see, e.g., Ying and Zhou, 2017) and we present the proof in Section A for completeness.

Lemma 13 *Let H be a Hilbert space associated with the inner product $\langle\cdot,\cdot\rangle$. Let $\mathcal{G}: H \to \mathbb{R}$ be a convex and differentiable functional satisfying*

$$\|\nabla\mathcal{G}(f) - \nabla\mathcal{G}(\tilde{f})\| \le L\|f-\tilde{f}\|^\alpha, \quad \forall f, \tilde{f} \in H,$$

where $L>0, \alpha\in(0,1], \nabla$ is the gradient operator and $\|\cdot\|$ is the norm induced by the inner product. Then, the following inequality holds for any $f, \tilde{f} \in H$

$$\frac{\alpha\|\nabla\mathcal{G}(f) - \nabla\mathcal{G}(\tilde{f})\|^{\frac{1+\alpha}{\alpha}}}{(1+\alpha)L^{\frac{1}{\alpha}}} \le \mathcal{G}(f) - \big[\mathcal{G}(\tilde{f}) + \langle f-\tilde{f}, \nabla\mathcal{G}(\tilde{f})\rangle\big] \le \frac{L\|f-\tilde{f}\|^{1+\alpha}}{1+\alpha}. \tag{4.1}$$

With Lemma 13, we can derive the following lemma on gradients of loss functions at iterates of the algorithm (1.1). Its power consists in bounding the gradients for the possibly unbounded iterates $\{f_t\}_{t\in\mathbb{N}}$ by the gradients for f_H and the excess generalization errors, the first of which can be considered as a constant while the second of which are exactly the terms we are interested in. For a random variable z, we use $\mathbb{E}_z[\cdot]$ to denote the conditional expectation with respect to z.

Lemma 14 *Suppose Assumption 1 holds and $\beta \in (0,1]$. Then,*

$$\mathbb{E}_{z_t}\big[|\phi'(y_t,f_t(x_t))|^{1+\beta}\big] \leq 2^\beta L^{\frac{1}{\alpha}}(1+\beta)\big[\mathcal{E}(f_t)-\mathcal{E}(f_H)\big]+ \\ \frac{2^\beta(1-\alpha\beta)}{1+\alpha}+2^\beta\mathbb{E}_{z_t}\big[|\phi'(y_t,f_H(x_t))|^{1+\beta}\big], \quad \forall t\in\mathbb{N}. \tag{4.2}$$

Proof With the elementary inequality $|u+v|^{1+\beta}\leq 2^\beta[|u|^{1+\beta}+|v|^{1+\beta}]$ and the Young's inequality

$$uv \leq p^{-1}|u|^p+q^{-1}|v|^q, \quad \forall u,v\in\mathbb{R}, p^{-1}+q^{-1}=1, p\geq 0, \tag{4.3}$$

the term $|\phi'(y_t,f_t(x_t))|^{1+\beta}$ can be controlled by

$$\begin{aligned}|\phi'(y_t,f_t(x_t))|^{1+\beta} &\leq \Big[|\phi'(y_t,f_t(x_t))-\phi'(y_t,f_H(x_t))|+|\phi'(y_t,f_H(x_t))|\Big]^{1+\beta}\\ &\leq 2^\beta|\phi'(y_t,f_t(x_t))-\phi'(y_t,f_H(x_t))|^{1+\beta}+2^\beta|\phi'(y_t,f_H(x_t))|^{1+\beta}\\ &\leq \frac{2^\beta\alpha(1+\beta)}{1+\alpha}|\phi'(y_t,f_t(x_t))-\phi'(y_t,f_H(x_t))|^{\frac{1+\alpha}{\alpha}}+\frac{2^\beta(1-\alpha\beta)}{1+\alpha}+2^\beta|\phi'(y_t,f_H(x_t))|^{1+\beta}.\end{aligned} \tag{4.4}$$

It follows from the first inequality of (4.1) that

$$\frac{\alpha}{1+\alpha}|\phi'(y_t,f_t(x_t))-\phi'(y_t,f_H(x_t))|^{\frac{1+\alpha}{\alpha}} \leq \\ L^{\frac{1}{\alpha}}\Big[\phi(y_t,f_t(x_t))-\phi(y_t,f_H(x_t))-\phi'(y_t,f_H(x_t))(f_t(x_t)-f_H(x_t))\Big].$$

Plugging the above inequality into (4.4) and taking expectations with respect to z_t (note f_t is independent of z_t), we get

$$\begin{aligned}\mathbb{E}_{z_t}\big[|\phi'(y_t,f_t(x_t))|^{1+\beta}\big] &\leq 2^\beta L^{\frac{1}{\alpha}}(1+\beta)\Big[\mathcal{E}(f_t)-\mathcal{E}(f_H)-\big\langle f_t-f_H,\mathbb{E}_{z_t}\big[\phi'(y_t,f_H(x_t))K_{x_t}\big]\big\rangle\Big]\\ &\quad+\frac{2^\beta(1-\alpha\beta)}{1+\alpha}+2^\beta\mathbb{E}_{z_t}\big[|\phi'(y_t,f_H(x_t))|^{1+\beta}\big]\\ &=2^\beta L^{\frac{1}{\alpha}}(1+\beta)\big[\mathcal{E}(f_t)-\mathcal{E}(f_H)\big]+\frac{2^\beta(1-\alpha\beta)}{1+\alpha}+2^\beta\mathbb{E}_{z_t}\big[|\phi'(y_t,f_H(x_t))|^{1+\beta}\big].\end{aligned}$$

Here the last identity holds since

$$\nabla\mathcal{E}(f_H)=\mathbb{E}_z\big[\phi'(y,f_H(x))K_x\big]=0.$$

The proof is complete. ■

4.1 Proofs for Convergence in Expectation

Before proving Theorem 1 and Theorem 2 on convergence in expectation, we first present some preparatory results. Our first preliminary result is a weak result on convergence in expectation under a weak condition on the step size sequence (4.5). Eq. (4.6) implies the existence of a sub-index sequence $\{i_t\}_{t\in\mathbb{N}}$ satisfying $\lim_{t\to\infty} A_{i_t}=0$, while (4.7) shows the convergence of a weighted average of the expected excess generalization errors. This result is derived based on a one-step progress inequality in terms of distances in RKHSs (see (4.10)).

Proposition 15 *Let $\{f_t\}_{t\in\mathbb{N}}$ be the sequence given by* (1.1) *and suppose Assumption 1 holds. If*

$$\lim_{t\to\infty}\eta_t=0 \quad \textit{and} \quad \sum_{t=1}^{\infty}\eta_t=\infty, \tag{4.5}$$

then

$$\liminf_{t\to\infty}\mathbb{E}[\mathcal{E}(f_t)-\mathcal{E}(f_H)]=0 \tag{4.6}$$

and

$$\lim_{T\to\infty}\Big[\sum_{t=1}^{T}\eta_t\Big]^{-1}\sum_{t=1}^{T}\eta_t\mathbb{E}[\mathcal{E}(f_t)-\mathcal{E}(f_H)]=0. \tag{4.7}$$

Lemma 16 *Let $\{\eta_t\}_{t\in\mathbb{N}}$ be a sequence of positive numbers. If $\lim_{t\to\infty}\eta_t=0$ and $\sum_{t=1}^{\infty}\eta_t=\infty$, then $\lim_{t\to\infty}\big[\sum_{k=1}^{t}\eta_k\big]^{-1}\sum_{k=1}^{t}\eta_k^2=0$.*

Proof of Proposition 15 According to the iteration strategy (1.1), we derive

$$\begin{aligned}\|f_{t+1}-f_H\|^2 &= \|f_t-\eta_t\phi'(y_t,f_t(x_t))K_{x_t}-f_H\|^2\\ &\le \|f_t-f_H\|^2+\eta_t^2|\phi'(y_t,f_t(x_t))|^2\kappa^2-2\eta_t\langle f_t-f_H,\phi'(y_t,f_t(x_t))K_{x_t}\rangle \qquad (4.8)\\ &\le \|f_t-f_H\|^2+\eta_t^2|\phi'(y_t,f_t(x_t))|^2\kappa^2+2\eta_t\big[\phi(y_t,f_H(x_t))-\phi(y_t,f_t(x_t))\big]. \qquad (4.9)\end{aligned}$$

Note that f_t is independent of z_t. Taking expectations with respect to z_t on both sides and using (4.2) with $\beta=1$, we derive

$$\begin{aligned}&\mathbb{E}_{z_t}\big[\|f_{t+1}-f_H\|^2\big]\le\|f_t-f_H\|^2+\eta_t^2\kappa^2\mathbb{E}_{z_t}\big[|\phi'(y_t,f_t(x_t))|^2\big]+2\eta_t[\mathcal{E}(f_H)-\mathcal{E}(f_t)]\\ &\le \|f_t-f_H\|^2+4\eta_t^2\kappa^2L^{\frac{1}{\alpha}}[\mathcal{E}(f_t)-\mathcal{E}(f_H)]+\frac{2(1-\alpha)\eta_t^2\kappa^2}{1+\alpha}\\ &\qquad\qquad +2\eta_t^2\kappa^2\mathbb{E}_{z_t}\big[|\phi'(y_t,f_H(x_t))|^2\big]+2\eta_t[\mathcal{E}(f_H)-\mathcal{E}(f_t)]\\ &=\|f_t-f_H\|^2+2\eta_t\big(1-2\eta_t\kappa^2L^{\frac{1}{\alpha}}\big)[\mathcal{E}(f_H)-\mathcal{E}(f_t)]+2\eta_t^2\kappa^2\Big(\mathbb{E}_{z_t}\big[|\phi'(y_t,f_H(x_t))|^2\big]+\frac{1-\alpha}{1+\alpha}\Big).\end{aligned}$$

Since $\lim_{t\to\infty}\eta_t=0$, we can find an integer $t_1\in\mathbb{N}$ such that $\eta_t\le\frac{1}{4\kappa^2L^{\frac{1}{\alpha}}},\forall t\ge t_1$. This together with $\mathcal{E}(f_H)\le\mathcal{E}(f_t)$ implies

$$\eta_t[\mathcal{E}(f_t)-\mathcal{E}(f_H)]\le\|f_t-f_H\|^2-\mathbb{E}_{z_t}\big[\|f_{t+1}-f_H\|^2\big]+\gamma\eta_t^2,\quad \forall t\ge t_1, \tag{4.10}$$

where we introduce $\gamma = 2\kappa^2\Big(\mathbb{E}_{z_t}\big[|\phi'(y_t, f_H(x_t))|^2\big] + \frac{1-\alpha}{1+\alpha}\Big)$. Taking expectations followed with a summation from $t = t_1$ to $t = T$ gives

$$\sum_{t=t_1}^{T} \eta_t A_t \le \mathbb{E}[\|f_{t_1} - f_H\|^2] + \gamma \sum_{t=t_1}^{T} \eta_t^2.$$

It then follows that

$$\begin{aligned}\lim_{T\to\infty} \big[\sum_{t=1}^{T}\eta_t\big]^{-1}\sum_{t=1}^{T}\eta_t A_t &= \lim_{T\to\infty} \big[\sum_{t=1}^{T}\eta_t\big]^{-1}\sum_{t=1}^{t_1-1}\eta_t A_t + \lim_{T\to\infty} \big[\sum_{t=1}^{T}\eta_t\big]^{-1}\sum_{t=t_1}^{T}\eta_t A_t\\ &\le \lim_{T\to\infty} \big[\sum_{t=1}^{T}\eta_t\big]^{-1}\Big[\mathbb{E}[\|f_{t_1} - f_H\|^2] + \gamma\sum_{t=t_1}^{T}\eta_t^2\Big] = 0,\end{aligned}$$

where we have used $\lim_{t\to\infty}\big[\sum_{k=1}^{t}\eta_k\big]^{-1}\sum_{k=1}^{t}\eta_k^2 = 0$ established in Lemma 16. This establishes (4.7).

We now prove (4.6) by contradiction strategy. Suppose to the contrary that $\liminf\limits_{t\to\infty} A_t = \tilde{a} > 0$. Then, there exists $\tilde{t} \in \mathbb{N}$ such that $A_t \ge 2^{-1}\tilde{a}, \forall t \ge \tilde{t}$, from which we derive from (4.7) that

$$0 = \lim_{T\to\infty}\frac{\sum_{t=1}^{T}\eta_t A_t}{\sum_{t=1}^{T}\eta_t} \ge \frac{\tilde{a}}{2}\lim_{T\to\infty}\frac{\sum_{t=\tilde{t}+1}^{T}\eta_t}{\sum_{t=1}^{T}\eta_t} = \frac{\tilde{a}}{2} - \frac{\tilde{a}}{2}\lim_{T\to\infty}\frac{\sum_{t=1}^{\tilde{t}}\eta_t}{\sum_{t=1}^{T}\eta_t} = \frac{\tilde{a}}{2}.$$

This leads to a contradiction. Therefore, $\liminf_{t\to\infty} A_t = 0$ and the proof is complete. ■

As our second preliminary result, Lemma 17 establishes an upper bound on $\mathbb{E}[\|f_t - f_H\|_2^2]$ in terms of the step size sequence, as well as a lower bound on $\mathbb{E}[\|\nabla\mathcal{E}(f_t)\|^2]$ in terms of the step size sequence and the expected excess generalization errors.

Lemma 17 *Let $\{f_t\}_{t\in\mathbb{N}}$ be the sequence given by* (1.1)*. If Assumption 1 holds and* $\lim_{t\to\infty}\eta_t = 0$*, then there exist constants $\widehat{C}, \gamma > 0$ independent of t such that the following inequalities hold for any $t \in \mathbb{N}$*

$$\mathbb{E}[\|f_t - f_H\|^2] \le \widehat{C} + \gamma\sum_{k=1}^{t}\eta_k^2 \tag{4.11}$$

and

$$\mathbb{E}[\|\nabla\mathcal{E}(f_t)\|^2] \ge \frac{\big(\mathbb{E}[\mathcal{E}(f_t) - \mathcal{E}(f_H)]\big)^2}{\widehat{C} + \gamma\sum_{k=1}^{t}\eta_k^2}. \tag{4.12}$$

Proof Since $\mathcal{E}(f_t) \ge \mathcal{E}(f_H)$ for all $t \in \mathbb{N}$, (4.10) implies

$$\mathbb{E}[\|f_{t+1} - f_H\|^2] \le \mathbb{E}[\|f_t - f_H\|^2] + \eta_t^2\gamma, \quad \forall t \ge t_1,$$

where γ and t_1 are defined in the proof of Proposition 15. Taking a summation of the above inequality from $t = t_1$ to $t = T$ shows

$$\mathbb{E}[\|f_{T+1} - f_H\|^2] \leq \mathbb{E}[\|f_{t_1} - f_H\|^2] + \gamma \sum_{t=t_1}^{T} \eta_t^2 \leq \widehat{C} + \gamma \sum_{t=1}^{T} \eta_t^2,$$

where we introduce $\widehat{C} = \mathbb{E}[\|f_{t_1} - f_H\|^2]$. This establishes (4.11).

We now turn to (4.12). According to the convexity of $\mathcal{E}$ and Schwartz inequality, we get

$$\begin{aligned}\mathbb{E}[\mathcal{E}(f_t)] - \mathcal{E}(f_H) &\leq \mathbb{E}\big[\big\langle \nabla\mathcal{E}(f_t), f_t - f_H\big\rangle\big] \leq \mathbb{E}[\|\nabla\mathcal{E}(f_t)\|\|f_t - f_H\|] \\ &\leq \big(\mathbb{E}\big[\|\nabla\mathcal{E}(f_t)\|^2\big]\big)^{\frac{1}{2}}\big(\mathbb{E}\big[\|f_t - f_H\|^2\big]\big)^{\frac{1}{2}}.\end{aligned}$$

The above inequality together with (4.11) gives

$$\mathbb{E}[\big\|\nabla\mathcal{E}(f_t)\big\|^2] \geq \frac{\big(\mathbb{E}\big[\mathcal{E}(f_t) - \mathcal{E}(f_H)\big]\big)^2}{\mathbb{E}[\|f_t - f_H\|^2]} \geq \frac{\big(\mathbb{E}\big[\mathcal{E}(f_t) - \mathcal{E}(f_H)\big]\big)^2}{\widehat{C} + \gamma\sum_{k=1}^{t}\eta_k^2}.$$

This establishes (4.12) and completes the proof. ■

Remark 18 *Eq.* (4.11) *was derived in Ying and Zhou (2017) under an additional assumption $\sum_{k=1}^{\infty}\eta_k^{1+\alpha} < \infty$, which is removed in Lemma 17. For step sizes of the form $\eta_t = \eta_1 t^{-\theta}$ with $\frac{1}{2} < \theta \leq 1$, it was shown $\mathbb{E}[\|f_t - f_H\|^2] = O\big(t^{1-\theta}\big(\mathcal{E}(f_H) - \inf_f \mathcal{E}(f)\big)\big)$ (Lin and Zhou, 2018). As compared with the results in Ying and Zhou (2017); Lin and Zhou (2018), our discussion implies boundedness of $\mathbb{E}[\|f_t\|^2]$ under a milder condition $\sum_{k=1}^{\infty}\eta_k^2 < \infty$.*

We are now in a position to prove Theorem 1 for the convergence in expectation. Let $\epsilon > 0$ be an arbitrary small number. Our idea is to use Proposition 15, based on one-step progress in terms of the distances in RKHSs, to show that $\{A_t\}_{t\in\mathbb{N}}$ can be smaller than ϵ infinitely often. Once $A_{\tilde{t}} \leq \epsilon$ for a sufficiently large $\tilde{t}$, we can use the assumption $\lim_{t\to\infty}\eta_t^{\alpha}\sum_{k=1}^{t}\eta_k^2 = 0$ and the one-step progress inequality (4.15) in terms of generalization errors to show $A_t \leq \epsilon$ for any $t \geq \tilde{t}$.

Proof of Theorem 1 Since $\phi'(y,\cdot)$ is (α, L)-Hölder continuous, we can apply the second inequality of (4.1) to show that

$$\phi(y, f_{t+1}(x)) \leq \phi(y, f_t(x)) + (f_{t+1}(x) - f_t(x))\phi'(y, f_t(x)) + \frac{L}{1+\alpha}|f_{t+1}(x) - f_t(x)|^{1+\alpha}.$$

According to the reproducing property $f(x) = \langle f, K_x\rangle, \forall f \in H$ and the iteration scheme (1.1), we know

$$\begin{aligned}\phi(y, f_{t+1}(x)) &\leq \phi(y, f_t(x)) + \langle f_{t+1} - f_t, \phi'(y, f_t(x))K_x\rangle + \frac{L}{1+\alpha}|\langle f_{t+1} - f_t, K_x\rangle|^{1+\alpha} \\ &\leq \phi(y, f_t(x)) - \eta_t\langle\phi'(y_t, f_t(x_t))K_{x_t}, \phi'(y, f_t(x))K_x\rangle + \frac{L\kappa^{1+\alpha}}{1+\alpha}\|f_{t+1} - f_t\|^{1+\alpha} \\ &\leq \phi(y, f_t(x)) - \eta_t\langle\phi'(y_t, f_t(x_t))K_{x_t}, \phi'(y, f_t(x))K_x\rangle + \frac{L\kappa^{2(1+\alpha)}\eta_t^{1+\alpha}}{1+\alpha}|\phi'(y_t, f_t(x_t))|^{1+\alpha}.\end{aligned} \tag{4.13}$$

Putting (4.2) with $\beta = \alpha$ back into (4.13) followed with a conditional expectation with respect to z_t and z yields

$$\mathbb{E}_{z_t}[\mathcal{E}(f_{t+1})] = \mathbb{E}_{z_t,z}[\phi(y, f_{t+1}(x))] \leq \mathbb{E}_z[\phi(y, f_t(x))] - \eta_t \langle \mathbb{E}_{z_t}[\phi'(y_t, f_t(x_t))K_{x_t}], \mathbb{E}_z[\phi'(y, f_t(x))K_x]\rangle + \frac{L\kappa^{2(1+\alpha)}\eta_t^{1+\alpha}}{1+\alpha}\Big[2^\alpha L^{\frac{1}{\alpha}}(1+\alpha)(\mathcal{E}(f_t) - \mathcal{E}(f_H)) + 2^\alpha(1-\alpha) + 2^\alpha \mathbb{E}_z[|\phi'(y, f_H(x))|^{1+\alpha}]\Big]$$
$$\leq \mathcal{E}(f_t) - \eta_t\|\nabla\mathcal{E}(f_t)\|^2 + \frac{L\kappa^{2(1+\alpha)}2^\alpha\eta_t^{1+\alpha}\Big[L^{\frac{1}{\alpha}}(1+\alpha)(\mathcal{E}(f_t) - \mathcal{E}(f_H)) + (1-\alpha) + \mathbb{E}_z[|\phi'(y, f_H(x))|^{1+\alpha}]\Big]}{1+\alpha}.$$

Subtracting $\mathcal{E}(f_H)$ from both sides of the above inequality gives

$$\mathbb{E}_{z_t}[\mathcal{E}(f_{t+1})] - \mathcal{E}(f_H) \leq \Big[1 + L^{1+\frac{1}{\alpha}}\kappa^{2(1+\alpha)}2^\alpha\eta_t^{1+\alpha}\Big](\mathcal{E}(f_t) - \mathcal{E}(f_H)) - \eta_t\|\nabla\mathcal{E}(f_t)\|^2 + \frac{L\kappa^{2(1+\alpha)}2^\alpha\eta_t^{1+\alpha}}{1+\alpha}\Big[(1-\alpha) + \mathbb{E}_z[|\phi'(y, f_H(x))|^{1+\alpha}]\Big]. \quad (4.14)$$

Taking expectations over both sides, the above inequality can be written as

$$A_{t+1} \leq (1 + a\eta_t^{1+\alpha})A_t + b\eta_t^{1+\alpha} - \eta_t\mathbb{E}[\|\nabla\mathcal{E}(f_t)\|^2], \quad (4.15)$$

where we introduce the notations

$$a = L^{1+\frac{1}{\alpha}}\kappa^{2(1+\alpha)}2^\alpha \quad \text{and} \quad b = \frac{L\kappa^{2(1+\alpha)}2^\alpha}{1+\alpha}\Big[(1-\alpha) + \mathbb{E}_z[|\phi'(y, f_H(x))|^{1+\alpha}]\Big]. \quad (4.16)$$

Plugging (4.12) into the above inequality gives

$$A_{t+1} \leq (1 + a\eta_t^{1+\alpha})A_t + b\eta_t^{1+\alpha} - \frac{\eta_t A_t^2}{\widehat{C} + \gamma\sum_{k=1}^t \eta_k^2}, \quad (4.17)$$

where $\widehat{C}$ and γ are defined in the proof of Lemma 17. The assumption $\lim_{t\to\infty}\eta_t^\alpha\sum_{k=1}^t\eta_k^2 = 0$ implies $\lim_{t\to\infty}\eta_t = 0$ and therefore the assumptions of Proposition 15 hold. Let $\epsilon \in (0,1)$ be an arbitrary number. According to $\liminf_{t\to\infty} A_t = 0$ established in Proposition 15, we can find a $\tilde{t} \in \mathbb{N}$ ($\tilde{t}$ can be sufficiently large) such that $A_{\tilde{t}} \leq \epsilon$ and

$$\eta_t^\alpha\big(\widehat{C} + \gamma\sum_{k=1}^t \eta_k^2\big) \leq \frac{\epsilon^2}{4(a+b)}, \quad \eta_t^{1+\alpha} \leq \frac{\epsilon}{2(a+b)} \quad \forall t \geq \tilde{t}. \quad (4.18)$$

We now prove by induction that $A_t \leq \epsilon$ for all $t \geq \tilde{t}$. It suffices to show that $A_{t+1} \leq \epsilon$ under the assumption $A_t \leq \epsilon$ and $t \geq \tilde{t}$. Since $A_t \leq 1$, we derive from (4.17) that

$$A_{t+1} \leq A_t + (a+b)\eta_t^{1+\alpha} - \frac{\eta_t A_t^2}{\widehat{C} + \gamma\sum_{k=1}^t \eta_k^2}.$$

We now consider two cases. If $A_t^2 \geq (a+b)\eta_t^\alpha\big(\widehat{C} + \gamma\sum_{k=1}^t \eta_k^2\big)$, then we know $A_{t+1} \leq A_t \leq \epsilon$. Otherwise, we derive from (4.18) that

$$A_{t+1} \leq A_t + (a+b)\eta_t^{1+\alpha} \leq \sqrt{(a+b)\eta_t^\alpha\big(\widehat{C} + \gamma\sum_{k=1}^t \eta_k^2\big)} + (a+b)\eta_t^{1+\alpha} \leq \epsilon.$$

Putting the above two cases together we derive $A_{t+1} \leq \epsilon$. That is, $A_t \leq \epsilon$ for all $t \geq \tilde{t}$. Since $\epsilon \in (0,1)$ is arbitrarily chosen, we get $\lim_{t\to\infty} A_t = 0$. ■

The necessary condition in Theorem 2 is established by applying the co-coercivity given in Lemma 13 to bound $\mathcal{E}(f_{t+1})$ in terms of $\mathcal{E}(f_t)$ from below.

Proof of Theorem 2 Since $\phi'(y,\cdot)$ is $(1,L)$-Hölder continuous for any $y \in \mathcal{Y}$, we have

$$\begin{aligned}\|\nabla\mathcal{E}(f) - \nabla\mathcal{E}(\tilde{f})\| = \big\|\mathbb{E}[\phi'(y,f(x))K_x - \phi'(y,\tilde{f}(x))K_x]\big\| &\leq \mathbb{E}\big[|\phi'(y,f(x)) - \phi'(y,\tilde{f}(x))|\|K_x\|\big] \\ &\leq L\mathbb{E}\big[|\langle f-\tilde{f},K_x\rangle|\|K_x\|\big] \leq L\kappa^2\|f-\tilde{f}\|. \qquad (4.19)\end{aligned}$$

That is, $\nabla\mathcal{E}$ is $(1,L\kappa^2)$-Hölder continuous. Lemma 13 with $\alpha = 1$ and $\nabla\mathcal{E}(f_H) = 0$ then yield the following inequality

$$\mathcal{E}(f_t) \geq \mathcal{E}(f_H) + \langle f_t - f_H, \nabla\mathcal{E}(f_H)\rangle + \frac{\|\nabla\mathcal{E}(f_t) - \nabla\mathcal{E}(f_H)\|^2}{2L\kappa^2} = \mathcal{E}(f_H) + \frac{\|\nabla\mathcal{E}(f_t)\|^2}{2L\kappa^2}. \quad (4.20)$$

It follows from the convexity of $\mathcal{E}$ and (1.1) that

$$\mathcal{E}(f_{t+1}) \geq \mathcal{E}(f_t) + \langle\nabla\mathcal{E}(f_t), f_{t+1} - f_t\rangle = \mathcal{E}(f_t) - \eta_t\big\langle\nabla\mathcal{E}(f_t), \phi'(y_t,f_t(x_t))K_{x_t}\big\rangle.$$

Taking expectations over both sides and using (4.20), we derive the following inequality for all $t \in \mathbb{N}$

$$\mathbb{E}[\mathcal{E}(f_{t+1})] \geq \mathbb{E}[\mathcal{E}(f_t)] - \eta_t\mathbb{E}[\|\nabla\mathcal{E}(f_t)\|^2] \geq \mathbb{E}[\mathcal{E}(f_t)] - 2L\kappa^2\eta_t\mathbb{E}[\mathcal{E}(f_t) - \mathcal{E}(f_H)].$$

Hence,

$$A_{t+1} \geq \big(1 - 2L\kappa^2\eta_t\big)A_t, \quad \forall t \in \mathbb{N}.$$

The assumption $\eta_t \leq 1/(6L\kappa^2)$ and the elementary inequality $1 - \eta \geq \exp(-2\eta), \forall\eta \in (0,1/3)$ (Lin and Zhou, 2015) then show

$$A_{t+1} \geq \exp\big(-4L\kappa^2\eta_t\big)A_t \geq \prod_{k=1}^{t}\exp\big(-4L\kappa^2\eta_k\big)A_1 = \exp\Big(-4L\kappa^2\sum_{k=1}^{t}\eta_k\Big)A_1,$$

which, together with the condition $\lim_{t\to\infty} A_t = 0$ and $A_1 \neq 0$, then establishes the necessary condition $\sum_{t=1}^{\infty}\eta_t = \infty$. ■

4.2 Proofs for Almost Sure Convergence

We use the following Doob's martingale convergence theorem (see, e.g., Doob, 1994, page 195) to prove Theorem 4 on almost sure convergence. Specifically, we will use the one-step progress inequality in terms of generalization errors to construct a supermartingale, whose almost sure convergence would imply the almost sure convergence of $\{\hat{A}_t\}_{t\in\mathbb{N}}$.

Lemma 19 *Let $\{\tilde{X}_t\}_{t\in\mathbb{N}}$ be a sequence of non-negative random variables with $\mathbb{E}[\tilde{X}_1] < \infty$ and let $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be a nested sequence of sets of random variables with $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all $t \in \mathbb{N}$. If $\mathbb{E}[\tilde{X}_{t+1}|\mathcal{F}_t] \leq \tilde{X}_t$ for every $t \in \mathbb{N}$, then $\tilde{X}_t$ converges to a nonnegative random variable $\tilde{X}$ almost surely. Furthermore, $\tilde{X} < \infty$ almost surely.*

Proof of Theorem 4 Eq. (4.14) gives

$$\mathbb{E}_{z_t}[\hat{A}_{t+1}] \le (1+a\eta_t^{1+\alpha})\hat{A}_t + b\eta_t^{1+\alpha}, \quad \forall t\in\mathbb{N}, \tag{4.21}$$

with a and b are defined in the proof of Theorem 1. Denote $c=\prod_{k=1}^\infty(1+a\eta_k^{1+\alpha})$, which, according to the elementary inequality $1+\tau\le\exp(\tau),\tau\ge0$ and (2.3), satisfies

$$c\le\prod_{k=1}^\infty\exp(a\eta_k^{1+\alpha})=\exp\big(a\sum_{k=1}^\infty\eta_k^{1+\alpha}\big)<\infty.$$

Multiplying both sides of (4.21) by $\prod_{k=t+1}^\infty(1+a\eta_k^{1+\alpha})$, we derive

$$\begin{aligned}\prod_{k=t+1}^\infty(1+a\eta_k^{1+\alpha})\mathbb{E}_{z_t}\big[\hat{A}_{t+1}\big] &\le \prod_{k=t}^\infty(1+a\eta_k^{1+\alpha})\hat{A}_t + b\eta_t^{1+\alpha}\prod_{k=t+1}^\infty(1+a\eta_k^{1+\alpha})\\ &\le \prod_{k=t}^\infty(1+a\eta_k^{1+\alpha})\hat{A}_t+bc\eta_t^{1+\alpha}.\end{aligned}\tag{4.22}$$

Introduce the stochastic process

$$\hat{X}_t=\prod_{k=t}^\infty(1+a\eta_k^{1+\alpha})\hat{A}_t+bc\sum_{k=t}^\infty\eta_k^{1+\alpha},\quad t\in\mathbb{N}\tag{4.23}$$

According to (2.3), we know $\mathbb{E}[\hat{X}_1]<\infty$. Eq. (4.22) implies $\mathbb{E}_{z_t}[\hat{X}_{t+1}]\le\hat{X}_t$ for all $t\in\mathbb{N}$, that is, $\{\hat{X}_t\}_{t\in\mathbb{N}}$ is a supermartingale taking non-negative values. Lemma 19 then implies that $\lim_{t\to\infty}\hat{X}_t=\hat{X}$ for a non-negative random variable $\hat{X}$ almost surely. Let $\Omega=\{\omega=\{z_t\}_{t\in\mathbb{N}}\}$ be the set for which $\{\hat{X}_t(\omega)\}_t$ converges to $\hat{X}(\omega)$ as $t\to\infty$ and $\hat{X}(\omega)<\infty$. Then, $\Pr\{\Omega\}=1$, where $\Pr\{\Omega\}$ denotes the probability with which the event Ω happens. Let $\omega\in\Omega$ and $\epsilon>0$. Since $\sum_{t=1}^\infty\eta_t^{1+\alpha}<\infty$, we can find $\tilde{t}\in\mathbb{N}$ such that

$$\sum_{t=\tilde{t}}^\infty\eta_t^{1+\alpha}<\frac{\epsilon}{3bc},\quad\prod_{k=\tilde{t}}^\infty(1+a\eta_k^{1+\alpha})<1+\frac{\epsilon}{3\hat{X}(\omega)+\epsilon}\ \text{ and }\ |\hat{X}_t(\omega)-\hat{X}(\omega)|<\frac{\epsilon}{3},\quad\forall t\ge\tilde{t}.$$

It then follows from (4.23) that

$$\begin{aligned}\hat{A}_t(\omega)\le\hat{X}_t(\omega)&\le\Big(1+\frac{\epsilon}{3\hat{X}(\omega)+\epsilon}\Big)\hat{A}_t(\omega)+\frac{\epsilon}{3}\le\hat{A}_t(\omega)+\frac{\epsilon\hat{X}_t(\epsilon)}{3\hat{X}(\omega)+\epsilon}+\frac{\epsilon}{3}\\ &\le\hat{A}_t(\omega)+\frac{\epsilon\big(\hat{X}(\omega)+\frac{\epsilon}{3}\big)}{3\hat{X}(\epsilon)+\epsilon}+\frac{\epsilon}{3}\le\hat{A}_t(\omega)+\frac{2\epsilon}{3},\quad\forall t\ge\tilde{t},\end{aligned}$$

from which we derive

$$\hat{X}(\omega)-\epsilon\le\hat{X}_t(\omega)-\frac{2\epsilon}{3}\le\hat{A}_t(\omega)\le\hat{X}(\omega)+\epsilon,\quad\forall t\ge\tilde{t}.$$

That is, $\lim_{t\to\infty}\hat{A}_t(\omega)=\hat{X}(\omega)$ for any $\omega\in\Omega$, i.e., $\lim_{t\to\infty}\hat{A}_t=\hat{X}$ almost surely. Since $\sum_{t=1}^\infty\eta_t^{1+\alpha}<\infty$, we know $\sum_{t=1}^\infty\eta_t^2<\infty$ and $\lim_{t\to\infty}\eta_t=0$. This further implies

$$\lim_{t\to\infty}\eta_t^\alpha\sum_{k=1}^t\eta_k^2=0$$

and therefore the assumptions in Theorem 1 hold. Theorem 1 shows that $\lim_{t\to\infty}\mathbb{E}[\hat{A}_t]=0$. By Fatou's lemma, we get

$$0\leq\mathbb{E}[\hat{X}]=\mathbb{E}\big[\lim_{t\to\infty}\hat{A}_t\big]\leq\liminf_{t\to\infty}\mathbb{E}[\hat{A}_t]=0,$$

which implies that $\mathbb{E}[\hat{X}]=0$ and therefore $\hat{X}=0$ almost surely since $\hat{X}$ is non-negative. Combining the above deductions together, we know that $\lim_{t\to\infty}\hat{A}_t=0$ almost surely. ■

Our proof of Theorem 6 is based on the following lemma which can be found in Lin and Zhou (2015) as an easy consequence of the Borel-Cantelli Lemma.

Lemma 20 *Let $\{\xi_t\}_{t\in\mathbb{N}}$ be a sequence of non-negative random variables and $\{\epsilon_t\}_{t\in\mathbb{N}}$ be a sequence of positive numbers satisfying $\lim_{t\to\infty}\epsilon_t=0$. If $\sum_{t=1}^{\infty}\Pr\{\xi_t>\epsilon_t\}<\infty$, then ξ_t converges to 0 almost surely.*

Proof of Theorem 6 Introduce $\delta_t=t^{-2}$ for all $t\in\mathbb{N}$. According to Corollary 11, there exists a constant $\widetilde{C}_1$ such that

$$\Pr\Big\{t^{\min\{1-\theta,(\alpha+1)\theta-1\}-\epsilon}\hat{A}_t\geq\widetilde{C}_1t^{-\epsilon}\log^2\frac{t}{\delta_t}\Big\}\leq\delta_t.$$

Since $\sum_{t=1}^{\infty}\delta_t<\infty$ and $\lim_{t\to\infty}t^{-\epsilon}\log^2\frac{t}{\delta_t}=0$, we can apply Lemma 20 here to show (2.4). The proof is complete. ■

4.3 Proofs for Convergence Rates with High Probability

Our discussion on high-probability convergence rates roots its foundation on the following concentration inequalities of martingales. Part (a) is the Azuma-Hoeffding inequality for martingales with bounded differences (Hoeffding, 1963), while Part (b) is a Bernstein-type inequality which exploits information on variances to derive improved concentration inequalities for martingales (Zhang, 2005). A remarkable property of this Bernstein-type inequality is that it involves a conditional variance which itself is a random variable.

Lemma 21 *Let $z_1,\ldots,z_n$ be a sequence of random variables such that z_k may depend on the previous random variables $z_1,\ldots,z_{k-1}$ for all $k=1,\ldots,n$. Consider a sequence of functionals $\xi_k(z_1,\ldots,z_k),k=1,\ldots,n$.*

(a) Assume that $|\xi_k-\mathbb{E}_{z_k}[\xi_k]|\leq b_k$ for each k. Let $\delta\in(0,1)$. With probability at least $1-\delta$ we have

$$\sum_{k=1}^{n}\xi_k-\sum_{k=1}^{n}\mathbb{E}_{z_k}[\xi_k]\leq\Big(2\sum_{k=1}^{n}b_k^2\log\frac{1}{\delta}\Big)^{\frac{1}{2}}.\tag{4.24}$$

(b) Assume that $\xi_k-\mathbb{E}_{z_k}[\xi_k]\leq b$ for each k. Let $\rho>0$ and $\delta\in(0,1)$. With probability at least $1-\delta$ we have

$$\sum_{k=1}^{n}\xi_k-\sum_{k=1}^{n}\mathbb{E}_{z_k}[\xi_k]\leq\frac{(e^\rho-\rho-1)\sigma_n^2}{\rho b}+\frac{b\log\frac{1}{\delta}}{\rho},\tag{4.25}$$

where $\sigma_n^2=\sum_{k=1}^{n}\mathbb{E}_{z_k}(\xi_k-\mathbb{E}_{z_k}\xi_k)^2$ is the conditional variance.

Since $\phi'(y,\cdot)$ is (α,L)-Hölder continuous, convex and non-negative, Proposition 1 in Ying and Zhou (2017) shows that $\phi(y,\cdot)$ satisfies the following self-bounding property

$$|\phi'(y,s)|^{\frac{1+\alpha}{\alpha}} \leq \frac{(1+\alpha)^{1+\frac{1}{\alpha}}}{\alpha} L^{\frac{1}{\alpha}} \phi(y,s), \quad \forall y \in \mathcal{Y}, s \in \mathbb{R}.$$

The Young's inequality (4.3) then implies

$$\begin{aligned} |\phi'(y,s)|^2 &\leq \alpha^{-\frac{2\alpha}{1+\alpha}} (1+\alpha)^2 L^{\frac{2}{1+\alpha}} \phi(y,s)^{\frac{2\alpha}{1+\alpha}} \\ &\leq \alpha^{-\frac{2\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}} (1+\alpha) \Big(2\alpha\phi(y,s) + 1 - \alpha\Big) = A\phi(y,s) + B, \end{aligned} \tag{4.26}$$

where

$$A = 2\alpha^{\frac{1-\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}} (1+\alpha) \quad \text{and} \quad B = \alpha^{-\frac{2\alpha}{1+\alpha}} L^{\frac{2}{1+\alpha}} (1-\alpha^2). \tag{4.27}$$

Below we will use Part (b) of Lemma 21 to show almost boundedness of $\{f_t\}_{t\in\mathbb{N}}$ with high probability (Proposition 9). To this aim, we first establish a crude bound on the iterates $\{f_t\}_{t\in\mathbb{N}}$ in terms of the step size sequence.

Lemma 22 *Let $\{f_t\}_{t\in\mathbb{N}}$ be the sequence given by (1.1). Assume $\eta_t \leq \frac{1}{A\kappa^2}$ for all $t \in \mathbb{N}$. Then, the following inequalities hold for all $t \in \mathbb{N}$*

$$\|f_{t+1} - f_H\|^2 \leq C_1 \sum_{k=0}^{t} \eta_k \quad \textit{and} \quad \|f_{t+1}\|^2 \leq C_1 \sum_{k=1}^{t} \eta_k, \tag{4.28}$$

where we introduce for brevity $\eta_0 = 1$ and

$$C_1 = \|f_H\|_2^2 + A^{-1}B + 2\max\big\{\sup_{y\in\mathcal{Y}} \phi(y,0), \sup_{z\in\mathcal{Z}} \phi(y, f_H(x))\big\}. \tag{4.29}$$

Furthermore, if $\eta_t \leq \frac{1}{A\kappa^2}$ and $\eta_{t+1} \leq \eta_t$ for all $t \in \mathbb{N}$, we have

$$\sum_{k=1}^{t} \eta_k^2 \phi(y_k, f_k(x_k)) \leq \eta_1 \|f_H\|^2 + C_2 \sum_{k=1}^{t} \eta_k^2, \tag{4.30}$$

where we introduce

$$C_2 = 2\sup_{z\in\mathcal{Z}} \phi(y, f_H(x)) + \eta_1 \kappa^2 B. \tag{4.31}$$

Proof Plugging (4.26) into (4.9) gives

$$\begin{aligned} \|f_{t+1} - f_H\|^2 &\leq \|f_t - f_H\|^2 + \eta_t^2\kappa^2[A\phi(y_t, f_t(x_t)) + B] + 2\eta_t[\phi(y_t, f_H(x_t)) - \phi(y_t, f_t(x_t))] \\ &= \|f_t - f_H\|^2 + 2\eta_t\phi(y_t, f_H(x_t)) + \eta_t^2\kappa^2 B + \eta_t(A\eta_t\kappa^2 - 2)\phi(y_t, f_t(x_t)) \qquad (4.32) \\ &\leq \|f_t - f_H\|^2 + 2\eta_t\phi(y_t, f_H(x_t)) + \eta_t^2\kappa^2 B \leq \|f_t - f_H\|^2 + \eta_t\big(2\phi(y_t, f_H(x_t)) + A^{-1}B\big), \end{aligned}$$

where the last two inequalities follow from the assumption $\eta_t \leq \frac{1}{A\kappa^2}$. According to the definitions of C_1 in (4.29) and η_0, it then follows that

$$\|f_{t+1} - f_H\|^2 = \|f_H\|^2 + \sum_{k=1}^{t} \big[\|f_{k+1} - f_H\|^2 - \|f_k - f_H\|^2\big] \leq C_1 \sum_{k=0}^{t} \eta_k.$$

This establishes the first inequality in (4.28). We now prove the second inequality in (4.28). Notice that (4.9) also holds if we replace f_H with 0. This, together with (4.26) and $\eta_t \leq \frac{1}{A\kappa^2}$, gives

$$\begin{aligned}\|f_{t+1}\|^2 &\leq \|f_t\|^2 + \eta_t^2\kappa^2[A\phi(y_t, f_t(x_t)) + B] + 2\eta_t[\phi(y_t, 0) - \phi(y_t, f_t(x_t))] \\ &= \|f_t\|^2 + 2\eta_t\phi(y_t, 0) + \eta_t^2\kappa^2 B + \eta_t(A\eta_t\kappa^2 - 2)\phi(y_t, f_t(x_t)) \\ &\leq \|f_t\|^2 + 2\eta_t\phi(y_t, 0) + \eta_t A^{-1}B.\end{aligned}$$

It is now clear

$$\|f_{t+1}\|^2 = \sum_{k=1}^{t}\left[\|f_{k+1}\|^2 - \|f_k\|^2\right] \leq C_1\sum_{k=1}^{t}\eta_k.$$

We now show (4.30). Applying $\eta_t \leq \frac{1}{A\kappa^2}$ in (4.32) gives

$$\eta_t\phi(y_t, f_t(x_t)) \leq \|f_t - f_H\|^2 - \|f_{t+1} - f_H\|^2 + 2\eta_t\phi(y_t, f_H(x_t)) + \eta_t^2\kappa^2 B. \tag{4.33}$$

Multiplying both sides of the above inequality by η_t and using $\eta_{t+1} \leq \eta_t$, we derive

$$\begin{aligned}\eta_t^2\phi(y_t, f_t(x_t)) &\leq \eta_t\left[\|f_t - f_H\|^2 - \|f_{t+1} - f_H\|^2\right] + 2\eta_t^2\phi(y_t, f_H(x_t)) + \eta_t^3\kappa^2 B \\ &\leq \eta_t\|f_t - f_H\|^2 - \eta_{t+1}\|f_{t+1} - f_H\|^2 + 2\eta_t^2\phi(y_t, f_H(x_t)) + \eta_t^3\kappa^2 B.\end{aligned}$$

Taking a summation of the above inequality gives (4.30). The proof is complete. ■

Based on the above lemma, Proposition 23 gives a high-probability bound on $\|f_{t+1} - f_H\|^2$ in terms of $\sum_{k=1}^t \eta_k^2\|f_k - f_H\|^2$. Proposition 23 is proved based on a one-step progress inequality (4.37) in terms of the RKHS distances, where the involved martingale is controlled by a Bernstein-type inequality with the dominant variance term cancelled out by the negative term $-2\sum_{k=1}^t \eta_k A_k$ existing in the one-step progress inequality.

Proposition 23 *Suppose assumptions in Theorem 7 hold. Let $\delta \in (0,1)$ and C_η, C_3, C_4 be constants defined by*

$$C_\eta = \sup_{k\in\mathbb{N}}\eta_k\sum_{j=0}^{k}\eta_j < \infty, \tag{4.34}$$

$$C_3 = \sup_{z_k\in\mathcal{Z}}\left\|\phi'(y_k, f_H(x_k))K_{x_k} - \mathbb{E}_z[\phi'(y, f_H(x))K_x]\right\|,\ C_4 = \frac{2(1-\alpha)\kappa^2}{1+\alpha} + 2\kappa^2\mathbb{E}_z\left[|\phi'(y, f_H(x))|^2\right]. \tag{4.35}$$

Then, there exists a constant ρ_1 (explicitly given in the proof and independent of t as well as the step size sequence) such that the following inequality holds with probability at least $1-\delta$

$$\begin{aligned}\|f_{t+1} - f_H\|^2 \leq (\eta_1\kappa^2 A + 1)\|f_H\|^2 + (AC_2 + B)\kappa^2\sum_{k=1}^{t}\eta_k^2 + \frac{C_4\sum_{k=1}^t\left[\eta_k^2\|f_H - f_k\|^2\right]}{2C_1C_\eta\kappa^2 L^{\frac{1}{\alpha}}} \\ + \frac{\left(2C_3C_1^{\frac{1}{2}}C_\eta + 4L(C_1^{\frac{1}{2}}\kappa)^{\alpha+1}C_\eta\right)\log\frac{1}{\delta}}{\rho_1}.\end{aligned} \tag{4.36}$$

Proof The assumption $\sum_{t=1}^{\infty}\eta_t^2 < \infty$ implies that C_η in (4.34) is well defined since $\eta_k\sum_{j=1}^k\eta_j \leq \sum_{j=1}^k\eta_j^2 < \infty$. According to (4.8) and (4.26), we derive

$$\|f_{k+1}-f_H\|^2 \leq \|f_k-f_H\|^2+\eta_k^2\kappa^2\big(A\phi(y_k,f_k(x_k))+B\big)+2\eta_k\langle f_H-f_k,\mathbb{E}_{z_k}[\phi'(y_k,f_k(x_k))K_{x_k}]\big\rangle \\ + 2\eta_k\big\langle f_H-f_k,\phi'(y_k,f_k(x_k))K_{x_k}-\mathbb{E}_{z_k}[\phi'(y_k,f_k(x_k))K_{x_k}]\big\rangle. \quad (4.37)$$

Using the convexity of ϕ followed with a summation from $k=1$ to t gives

$$\begin{aligned}\|f_{t+1}-f_H\|^2 &\leq \|f_H\|^2+\kappa^2\sum_{k=1}^t\eta_k^2\big(A\phi(y_k,f_k(x_k))+B\big)+2\sum_{k=1}^t\eta_k\big[\mathcal{E}(f_H)-\mathcal{E}(f_k)\big] \\ &\quad + 2\sum_{k=1}^t\eta_k\big\langle f_H-f_k,\phi'(y_k,f_k(x_k))K_{x_k}-\mathbb{E}_{z_k}\big[\phi'(y_k,f_k(x_k))K_{x_k}\big]\big\rangle \\ &\leq (\eta_1A\kappa^2+1)\|f_H\|^2+(AC_2+B)\kappa^2\sum_{k=1}^t\eta_k^2+2\sum_{k=1}^t\eta_k\big[\mathcal{E}(f_H)-\mathcal{E}(f_k)\big] \\ &\quad + 2\sum_{k=1}^t\eta_k\big\langle f_H-f_k,\phi'(y_k,f_k(x_k))K_{x_k}-\mathbb{E}_{z_k}\big[\phi'(y_k,f_k(x_k))K_{x_k}\big]\big\rangle, \quad (4.38)\end{aligned}$$

where the last inequality is due to (4.30). We now estimate the last term of the above inequality with Lemma 21. To this aim, we need to control both the magnitudes and variances for the martingale difference sequences.

Introduce a sequence of functionals $\xi_k, k\in\mathbb{N}$ as follows

$$\xi_k = \eta_k\big\langle f_H-f_k,\phi'(y_k,f_k(x_k))K_{x_k}-\mathbb{E}_{z_k}\big[\phi'(y_k,f_k(x_k))K_{x_k}\big]\big\rangle.$$

It is clear

$$\begin{aligned}&\big\|\phi'(y_k,f_k(x_k))K_{x_k}-\mathbb{E}_{z_k}[\phi'(y_k,f_k(x_k))K_{x_k}]\big\| \leq \big\|\phi'(y_k,f_k(x_k))K_{x_k}-\phi'(y_k,f_H(x_k))K_{x_k}\big\| \\ &\quad + \big\|\phi'(y_k,f_H(x_k))K_{x_k}-\mathbb{E}_z[\phi'(y,f_H(x))K_x]\big\|+\mathbb{E}_z[\|(\phi'(y,f_H(x))-\phi'(y,f_k(x)))K_x\|] \\ &\leq \sup_{z_k\in\mathcal{Z}}\big\|\phi'(y_k,f_H(x_k))K_{x_k}-\mathbb{E}_z[\phi'(y,f_H(x))K_x]\big\|+2L\kappa\sup_{x\in\mathcal{X}}|f_k(x)-f_H(x)|^\alpha,\end{aligned}$$

where we have used the Jensen's inequality in the first step. But

$$|f_k(x)-f_H(x)| = |\langle f_k-f_H,K_x\rangle| \leq \|f_k-f_H\|\kappa.$$

Combining the above two inequalities and using the definition of C_3 in (4.35) give

$$\big\|\phi'(y_k,f_k(x_k))K_{x_k}-\mathbb{E}_{z_k}[\phi'(y_k,f_k(x_k))K_{x_k}]\big\| \leq C_3+2L\|f_k-f_H\|^\alpha\kappa^{\alpha+1}. \quad (4.39)$$

It then follows from (4.28) and $\mathbb{E}_{z_k}[\xi_k]=0$ that (note $\eta_0=1$)

$$\begin{aligned}\xi_k-\mathbb{E}_{z_k}[\xi_k] = \xi_k &\leq \eta_k\|f_H-f_k\|\big\|\phi'(y_k,f_k(x_k))K_{x_k}-\mathbb{E}_{z_k}\big[\phi'(y_k,f_k(x_k))K_{x_k}\big]\big\| \\ &\leq \eta_kC_3\|f_H-f_k\|+2L\eta_k\kappa^{\alpha+1}\|f_H-f_k\|^{1+\alpha} \quad (4.40) \\ &\leq \eta_kC_3C_1^{\frac{1}{2}}\big(\sum_{j=0}^{k-1}\eta_j\big)^{\frac{1}{2}}+2L(C_1^{\frac{1}{2}}\kappa)^{\alpha+1}\eta_k\big(\sum_{j=0}^{k-1}\eta_j\big)^{\frac{1+\alpha}{2}} \\ &\leq C_3C_1^{\frac{1}{2}}C_\eta+2L(C_1^{\frac{1}{2}}\kappa)^{\alpha+1}C_\eta.\end{aligned}$$

Here we have used the definition of C_η given in (4.34). Furthermore, according to Lemma 14 with $\beta = 1$ and the definition of C_4 in (4.35), the conditional variances can be controlled by (note $\mathbb{E}[(\xi - \mathbb{E}[\xi])^2] < \mathbb{E}[\xi^2]$ for a real-valued random variable ξ)

$$\begin{aligned}\sum_{k=1}^{t} \mathbb{E}_{z_k}(\xi_k - \mathbb{E}_{z_k}[\xi_k])^2 &\leq \sum_{k=1}^{t} \eta_k^2 \mathbb{E}_{z_k}\left[\langle f_H - f_k, \phi'(y_k, f_k(x_k)) K_{x_k}\rangle^2\right] \\ &\leq \sum_{k=1}^{t} \eta_k^2 \|f_H - f_k\|^2 \kappa^2 \mathbb{E}_{z_k}[|\phi'(y_k, f_k(x_k))|^2] \\ &\leq \sum_{k=1}^{t} \eta_k^2 \|f_H - f_k\|^2 \Big(4\kappa^2 L^{\frac{1}{\alpha}}[\mathcal{E}(f_k) - \mathcal{E}(f_H)] + C_4\Big).\end{aligned}$$

According to (4.28) and the definition of C_η in (4.34), we can further get

$$\begin{aligned}\sum_{k=1}^{t} \mathbb{E}_{z_k}(\xi_k - \mathbb{E}_{z_k}[\xi_k])^2 &\leq 4L^{\frac{1}{\alpha}} C_1 \kappa^2 \sum_{k=1}^{t} \Big[\eta_k^2 \big(\sum_{j=0}^{k-1} \eta_j\big)\big(\mathcal{E}(f_k) - \mathcal{E}(f_H)\big)\Big] + C_4 \sum_{k=1}^{t} \eta_k^2 \|f_H - f_k\|^2 \\ &\leq 4L^{\frac{1}{\alpha}} C_1 C_\eta \kappa^2 \sum_{k=1}^{t} \eta_k \big(\mathcal{E}(f_k) - \mathcal{E}(f_H)\big) + C_4 \sum_{k=1}^{t} \eta_k^2 \|f_H - f_k\|^2.\end{aligned}$$

Let ρ_1 be the largest positive constant such that (such ρ_1 exists since $\lim_{\rho\to 0} \frac{e^\rho - \rho - 1}{\rho} = 0$)

$$\frac{(e^{\rho_1} - \rho_1 - 1) L^{\frac{1}{\alpha}} C_1^{\frac{1}{2}} \kappa^2}{C_3 + 2LC_1^{\frac{\alpha}{2}} \kappa^{\alpha+1}} \leq \frac{\rho_1}{4}.$$

Since C_1 and C_3 do not depend on the step size sequence, ρ_1 is also a constant independent of the step size sequence. Plugging the above estimates on the magnitudes and variances of ξ_k into Part (b) of Lemma 21, we derive the following inequality with probability at least $1 - \delta$

$$\begin{aligned}\sum_{k=1}^{t} \xi_k \leq{}& \frac{(e^{\rho_1} - \rho_1 - 1)}{\rho_1\big(C_3 C_1^{\frac{1}{2}} C_\eta + 2L(C_1^{\frac{1}{2}}\kappa)^{\alpha+1} C_\eta\big)} \Big[4L^{\frac{1}{\alpha}} C_1 C_\eta \kappa^2 \sum_{k=1}^{t} \eta_k \big(\mathcal{E}(f_k) - \mathcal{E}(f_H)\big) \\ &+ C_4 \sum_{k=1}^{t} \eta_k^2 \|f_H - f_k\|^2\Big] + \frac{\big(C_3 C_1^{\frac{1}{2}} C_\eta + 2L(C_1^{\frac{1}{2}}\kappa)^{\alpha+1} C_\eta\big) \log \frac{1}{\delta}}{\rho_1} \\ \leq{}& \sum_{k=1}^{t} \eta_k \big(\mathcal{E}(f_k) - \mathcal{E}(f_H)\big) + \frac{C_4 \sum_{k=1}^{t} \left[\eta_k^2 \|f_H - f_k\|^2\right]}{4C_1 C_\eta \kappa^2 L^{\frac{1}{\alpha}}} + \frac{\big(C_3 C_1^{\frac{1}{2}} C_\eta + 2L(C_1^{\frac{1}{2}}\kappa)^{\alpha+1} C_\eta\big) \log \frac{1}{\delta}}{\rho_1}.\end{aligned}$$

Plugging this inequality into (4.38) gives the stated inequality with probability at least $1-\delta$.
■

According to Proposition 9 and the assumption $\sum_{k=1}^{\infty} \eta_k^2 < \infty$, one can show essentially that $\max_{1\leq t\leq T} \|f_t - f_H\|^2 \leq \frac{1}{2} \max_{1\leq t\leq T} \|f_t - f_H\|^2 + c \log T$ for a constant $c > 0$, from which one can establish the boundedness of the iterates with high probability (up to logarithmic factors).

Proof of Proposition 9 We define the subset $\Omega \subset \mathcal{Z}^T$ by

$$\Omega = \Big\{(z_1,\ldots,z_T) : \|f_{t+1}-f_H\|^2 \leq C_5 + \frac{C_4\sum_{k=1}^{t}\big[\eta_k^2\|f_H-f_k\|^2\big]}{2C_1C_\eta\kappa^2L^{\frac{1}{\alpha}}} + C_6\log\frac{T}{\delta} \text{ for all } t=1,\ldots,T\Big\},$$

where we introduce

$$C_5 = (\eta_1\kappa^2A+1)\|f_H\|^2 + (AC_2+B)\kappa^2\sum_{k=1}^{\infty}\eta_k^2, \quad C_6 = \frac{2C_3C_1^{\frac{1}{2}}C_\eta + 4L(C_1^{\frac{1}{2}}\kappa)^{\alpha+1}C_\eta}{\rho_1}. \tag{4.41}$$

Applying Proposition 23 together with union bounds on probabilities of events, we have $\Pr\{\Omega\} \geq 1-\delta$. Since $\sum_{t=1}^{\infty}\eta_t^2 < \infty$, there exists a $t_2 \in \mathbb{N}$ such that

$$C_4\sum_{k=t_2}^{\infty}\eta_k^2 \leq C_1C_\eta\kappa^2L^{\frac{1}{\alpha}}.$$

Under the event Ω, we know

$$\begin{aligned}\|f_{t+1}-f_H\|^2 &\leq C_5 + \frac{C_4\sum_{k=1}^{t_2}\big[\eta_k^2\|f_H-f_k\|^2\big]}{2C_1C_\eta\kappa^2L^{\frac{1}{\alpha}}} + \frac{C_4\sum_{k=t_2+1}^{t}\big[\eta_k^2\|f_H-f_k\|^2\big]}{2C_1C_\eta\kappa^2L^{\frac{1}{\alpha}}} + C_6\log\frac{T}{\delta}\\ &\leq C_5 + C_7 + \frac{1}{2}\max_{t_2<k\leq t}\|f_k-f_H\|^2 + C_6\log\frac{T}{\delta}\\ &\leq C_5 + C_7 + \frac{1}{2}\max_{1\leq k\leq T}\|f_k-f_H\|^2 + C_6\log\frac{T}{\delta}, \qquad \forall t=1,\ldots,T.\end{aligned}$$

where we have used the inequality

$$\frac{C_4\sum_{k=1}^{t_2}\big[\eta_k^2\|f_H-f_k\|^2\big]}{2C_1C_\eta\kappa^2L^{\frac{1}{\alpha}}} \leq \frac{C_4C_1\sum_{k=1}^{t_2}\big[\eta_k^2\sum_{j=0}^{k-1}\eta_j\big]}{2C_1C_\eta\kappa^2L^{\frac{1}{\alpha}}} := C_7.$$

Under the event Ω, it is now clear that

$$\max_{1\leq t\leq T}\|f_t-f_H\|^2 \leq C_5 + C_7 + \frac{1}{2}\max_{1\leq k\leq T}\|f_k-f_H\|^2 + C_6\log\frac{T}{\delta}.$$

Solving the above linear inequality yields the stated inequality with $\bar{C} = \max\{2(C_5+C_6+C_7), 1\}$ with probability at least $1-\delta$. ■

We are now in a position to prove Theorem 7 on general high-probability convergence rates for a weighted average of iterates. The underlying idea is to construct a modified martingale difference sequence by imposing a constraint on the iterates, which is then estimated by applying the Azuma-Hoeffding inequality on martingales. Furthermore, according to Proposition 9, this modified martingale difference sequence would be identical to the original martingale difference sequence with high probability. Let $\mathbb{I}_{\mathcal{A}}$ denote the indicator function of an event $\mathcal{A}$.

Proof of Theorem 7 We now introduce the following sequence of functionals $\xi_k', k = 1, \ldots, T$ by

$$\xi_k' = \eta_k \langle f_H - f_k, \phi'(y_k, f_k(x_k)) K_{x_k} - \mathbb{E}_{z_k}\big[\phi'(y_k, f_k(x_k)) K_{x_k}\big] \rangle \mathbb{I}_{\{\|f_k - f_H\|^2 \le \bar{C} \log \frac{2T}{\delta}\}},$$

where $\bar{C}$ is defined in Proposition 9. Analogous to (4.40), we have

$$\begin{aligned} |\xi_k'| &\le \Big[\eta_k C_3 \|f_H - f_k\| + 2L\eta_k \kappa^{\alpha+1} \|f_H - f_k\|^{1+\alpha}\Big] \mathbb{I}_{\{\|f_k - f_H\|^2 \le \bar{C} \log \frac{2T}{\delta}\}} \\ &\le \big(C_3 + 2L\kappa^{\alpha+1}\big)\eta_k \max\Big(\|f_H - f_k\|^2, 1\Big) \mathbb{I}_{\{\|f_k - f_H\|^2 \le \bar{C} \log \frac{2T}{\delta}\}} \\ &\le \big(C_3 + 2L\kappa^{\alpha+1}\big)\eta_k \bar{C} \log \frac{2T}{\delta} := b_k. \end{aligned} \tag{4.42}$$

It is clear that $\mathbb{E}_{z_k}[\xi_k'] = 0$ and ξ_k' only depends on $z_1, \ldots, z_k$. According to Part (a) of Lemma 21, there exists a subset $\Omega' = \{(z_1, \ldots, z_T) : z_1, \ldots, z_T \in \mathcal{Z}\} \subset \mathcal{Z}^T$ with probability measure $\Pr\{\Omega'\} \ge 1 - \frac{\delta}{2}$ such that for any $(z_1, \ldots, z_T) \in \Omega'$ the following inequality holds

$$\sum_{k=1}^T \xi_k' \le \Big(2\sum_{k=1}^T b_k^2 \log \frac{2}{\delta}\Big)^{\frac{1}{2}} \le \big(C_3 + 2L\kappa^{\alpha+1}\big)\bar{C} \log \frac{2T}{\delta} \Big(2 \log \frac{2}{\delta} \sum_{k=1}^T \eta_k^2\Big)^{\frac{1}{2}}.$$

According to Proposition 9, there exists a subset $\Omega = \{(z_1, \ldots, z_T) : z_1, \ldots, z_T \in \mathcal{Z}\} \subset \mathcal{Z}^T$ with probability measure $\Pr\{\Omega\} \ge 1 - \frac{\delta}{2}$ such that for any $(z_1, \ldots, z_T) \in \Omega$ the following inequality holds

$$\max_{1 \le k \le T} \|f_k - f_H\|^2 \le \bar{C} \log \frac{2T}{\delta}.$$

Let $\{\xi_k\}_k$ be the martingale difference sequence defined in the proof of Proposition 23. For any $(z_1, \ldots, z_T) \in \Omega \cap \Omega'$, we then have

$$\sum_{k=1}^T \xi_k = \sum_{k=1}^T \xi_k' \le \big(C_3 + 2L\kappa^{\alpha+1}\big)\bar{C} \log \frac{2T}{\delta} \Big(2 \log \frac{2}{\delta} \sum_{k=1}^T \eta_k^2\Big)^{\frac{1}{2}}.$$

Under this intersection of these two events, it follows from (4.38) and the definition of C_5 given in (4.41) that

$$\begin{aligned} 2\sum_{k=1}^T \eta_k \big[\mathcal{E}(f_k) - \mathcal{E}(f_H)\big] &\le (\eta_1 A\kappa^2 + 1)\|f_H\|^2 + (AC_2 + B)\kappa^2 \sum_{k=1}^T \eta_k^2 + 2\sum_{k=1}^T \xi_k \\ &\le C_5 + 2\big(C_3 + 2L\kappa^{\alpha+1}\big)\bar{C} \log \frac{2T}{\delta} \Big(2 \log \frac{2}{\delta} \sum_{k=1}^T \eta_k^2\Big)^{\frac{1}{2}}. \end{aligned}$$

But $\Pr\{\Omega \cap \Omega'\} \ge 1 - \delta$. Therefore, the first inequality of (2.5) holds with probability at least $1 - \delta$ and

$$\widetilde{C} = \frac{C_5}{2} + \big(C_3 + 2L\kappa^{\alpha+1}\big)\bar{C}\Big(2\sum_{k=1}^\infty \eta_k^2\Big)^{\frac{1}{2}}.$$

The second inequality of (2.5) follows from the convexity of $\mathcal{E}(\cdot)$. The proof is complete. ■

Other than the high-probability bounds for the weighted average of iterates $\bar{f}_T^{\eta}$, we can also derive similar results for the uniform average of iterates $\bar{f}_T$. If we choose the step sizes $\eta_t = \eta_1 (t \log^\beta t)^{-\frac{1}{2}}$ with $\beta > 1$, then Proposition 24 implies $\mathcal{E}(\bar{f}_T) - \mathcal{E}(f_H) = O(T^{-\frac{1}{2}} \log^{\frac{3}{2}} \frac{T}{\delta})$ with probability at least $1-\delta$. We present the proof in the appendix due to its similarity to the proof of Theorem 7.

Proposition 24 *Suppose assumptions in Theorem 7 hold. Then, for any $\delta \in (0,1)$, with probability at least $1 - \frac{\delta}{2}$ we have*

$$\sum_{t=1}^T [\mathcal{E}(f_t) - \mathcal{E}(f_H)] = O\Big(\big(T^{\frac{1}{2}} + \sum_{t=1}^T \eta_t\big) \log^{\frac{3}{2}} \frac{2T}{\delta}\Big) \text{ and } \mathcal{E}(\bar{f}_T) - \mathcal{E}(f_H) = O\Big(\big(T^{-\frac{1}{2}} + T^{-1} \sum_{t=1}^T \eta_t\big) \log^{\frac{3}{2}} \frac{2T}{\delta}\Big).$$

Theorem 10 is a specific case of Proposition 25 with $\widetilde{T} = \lfloor \frac{T}{2} \rfloor$. The step-stone in proving this proposition is the inequality (4.48) following from the one-step progress (4.47) in terms of generalization errors. The first term on the right hand side of (4.48) can be tackled by Theorem 7 on a weighted summation of $\hat{A}_t$ deduced from the one-step analysis in terms of RKHS distances. The variance of the martingales $\sum_{t=\tilde{t}}^T \bar{\xi}_t$ can be controlled by $\sum_{t=\tilde{t}}^T \eta_t \|\nabla \mathcal{E}(f_t)\|^2$, which is then cancelled out by the third term $-\sum_{t=\tilde{t}}^T \eta_t \|\nabla \mathcal{E}(f_t)\|^2$. A notable fact is that the martingale difference $\bar{\xi}_t - \mathbb{E}_{z_t}[\bar{\xi}_t]$ is bounded by $O(\eta_{\widetilde{T}})$ for all $t \geq \widetilde{T}$ with high probability, which would be small if $\widetilde{T}$ is large. We can balance the three terms on the right hand side of (4.43) by choosing an appropriate $\widetilde{T}$.

Proposition 25 *Suppose that the assumptions in Theorem 7 hold. Let $\widetilde{T} \in \mathbb{N}$ satisfy $1 \leq \widetilde{T} \leq T$. Then, there exists a constant $\widetilde{C}'$ independent of T and $\widetilde{T}$ (explicitly given in the proof) such that for any $\delta \in (0,1)$ the following inequality holds with probability at least $1-\delta$*

$$\mathcal{E}(f_{T+1}) - \mathcal{E}(f_H) \leq \widetilde{C}' \max\Big\{ \big[\sum_{t=\widetilde{T}}^T \eta_t\big]^{-1}, \eta_{\widetilde{T}}, \sum_{t=\widetilde{T}}^T \eta_t^{1+\alpha} \Big\} \log^2 \frac{3T}{\delta}. \tag{4.43}$$

Proof Recall that $\hat{A}_t = \mathcal{E}(f_t) - \mathcal{E}(f_H)$. According to the proof of Theorem 7, there exists a subset $\Omega = \{(z_1, \ldots, z_T) : z_1, \ldots, z_T \in \mathcal{Z}\} \subset \mathcal{Z}^T$ with $\Pr\{\Omega\} \geq 1 - \frac{2\delta}{3}$ such that for any $(z_1, \ldots, z_T) \in \Omega$, we have

$$\sum_{t=1}^T \eta_t \hat{A}_t \leq \widetilde{C} \log^{\frac{3}{2}} \frac{3T}{\delta} \quad \text{and} \quad \max_{1 \leq t \leq T} \|f_t - f_H\|^2 \leq \bar{C} \log \frac{3T}{\delta}, \tag{4.44}$$

where $\widetilde{C}$ and $\bar{C}$ are constants independent of T and δ. Under the event of Ω, we have $\sum_{t=\widetilde{T}}^T \eta_t \hat{A}_t \leq \widetilde{C} \log^{\frac{3}{2}} \frac{3T}{\delta}$. Therefore, there exists a $\tilde{t} \in \mathbb{N}$ satisfying $\widetilde{T} \leq \tilde{t} \leq T$ and

$$\hat{A}_{\tilde{t}} \leq \big[\sum_{t=\widetilde{T}}^T \eta_t\big]^{-1} \widetilde{C} \log^{\frac{3}{2}} \frac{3T}{\delta}. \tag{4.45}$$

Taking expectations only with respect to z over both sides of (4.13) gives

$$\hat{A}_{t+1} \leq \hat{A}_t - \eta_t \big\langle \phi'(y_t, f_t(x_t))K_{x_t}, \nabla\mathcal{E}(f_t)\big\rangle + \frac{L\kappa^{2(1+\alpha)}\eta_t^{1+\alpha}}{1+\alpha}|\phi'(y_t, f_t(x_t))|^{1+\alpha}. \tag{4.46}$$

According to (4.4), the term $|\phi'(y_t, f_t(x_t))|^{1+\alpha}$ can be controlled by

$$\begin{aligned}|\phi'(y_t, f_t(x_t))|^{1+\alpha} &\leq 2^\alpha|\phi'(y_t, f_t(x_t)) - \phi'(y_t, f_H(x_t))|^{1+\alpha} + 2^\alpha|\phi'(y_t, f_H(x_t))|^{1+\alpha}\\ &\leq 2^\alpha L^{1+\alpha}|\langle f_t - f_H, K_{x_t}\rangle|^{\alpha(1+\alpha)} + 2^\alpha|\phi'(y_t, f_H(x_t))|^{1+\alpha}\\ &\leq 2^\alpha L^{1+\alpha}\kappa^{\alpha(1+\alpha)}\|f_t - f_H\|^{\alpha(1+\alpha)} + 2^\alpha|\phi'(y_t, f_H(x_t))|^{1+\alpha}.\end{aligned}$$

Plugging the above bound into (4.46) gives

$$\begin{aligned}\hat{A}_{t+1} \leq \hat{A}_t - \eta_t\|\nabla\mathcal{E}(f_t)\|^2 &+ \eta_t\big\langle\nabla\mathcal{E}(f_t) - \phi'(y_t, f_t(x_t))K_{x_t}, \nabla\mathcal{E}(f_t)\big\rangle\\ &+ \Big(\tilde{a}\|f_t - f_H\|^{\alpha(1+\alpha)} + \tilde{b}\Big)\eta_t^{1+\alpha},\end{aligned} \tag{4.47}$$

where we introduce

$$\tilde{a} = 2^\alpha L^{2+\alpha}\kappa^{(2+\alpha)(1+\alpha)}(1+\alpha)^{-1} \quad \text{and} \quad \tilde{b} = 2^\alpha L\kappa^{2(1+\alpha)}(1+\alpha)^{-1}\sup_{z\in\mathcal{Z}}|\phi'(y, f_H(x))|^{1+\alpha}.$$

Taking a summation from $t = \tilde{t}$ to T yields

$$\hat{A}_{T+1} \leq \hat{A}_{\tilde{t}} + \sum_{t=\tilde{t}}^{T}\Big(\tilde{a}\|f_t - f_H\|^{\alpha(1+\alpha)} + \tilde{b}\Big)\eta_t^{1+\alpha} - \sum_{t=\tilde{t}}^{T}\eta_t\|\nabla\mathcal{E}(f_t)\|^2 + \sum_{t=\tilde{t}}^{T}\bar{\xi}_t, \tag{4.48}$$

where we introduce the following two sequences of functionals

$$\begin{aligned}\bar{\xi}_t &= \eta_t\big\langle\nabla\mathcal{E}(f_t) - \phi'(y_t, f_t(x_t))K_{x_t}, \nabla\mathcal{E}(f_t)\big\rangle,\\ \bar{\xi}'_t &= \eta_t\big\langle\nabla\mathcal{E}(f_t) - \phi'(y_t, f_t(x_t))K_{x_t}, \nabla\mathcal{E}(f_t)\big\rangle\mathbb{I}_{\{\|f_t - f_H\|^2 \leq \bar{C}\log\frac{3T}{\delta}\}}.\end{aligned}$$

Under the event Ω, it is clear $\bar{\xi}_t = \bar{\xi}'_t$. In the following, we will use Part (b) of Lemma 21 to estimate $\sum_{t=\tilde{t}}^{T}\bar{\xi}'_t$. It is clear that $\mathbb{E}_{z_t}[\bar{\xi}'_t] = 0$ for all $t \in \mathbb{N}$. Let $\bar{t}$ be any integer in $[\widetilde{T}, T]$. It follows from Lemma 14 with $\beta = 1$ and the definition of C_4 given in (4.35) that (note $\mathbb{E}[(\xi - \mathbb{E}[\xi])^2 < \mathbb{E}[\xi^2]$ for a real-valued random variable ξ)

$$\begin{aligned}\sum_{t=\bar{t}}^{T}\mathbb{E}_{z_t}\big(\bar{\xi}'_t - \mathbb{E}_{z_t}[\bar{\xi}'_t]\big)^2 &\leq \sum_{t=\bar{t}}^{T}\eta_t^2\mathbb{E}_{z_t}\big[\big\langle\phi'(y_t, f_t(x_t))K_{x_t}, \nabla\mathcal{E}(f_t)\big\rangle^2\big]\mathbb{I}_{\{\|f_t - f_H\|^2 \leq \bar{C}\log\frac{3T}{\delta}\}}\\ &\leq \sum_{t=\bar{t}}^{T}\eta_t^2\kappa^2\|\nabla\mathcal{E}(f_t)\|^2\mathbb{E}_{z_t}\big[|\phi'(y_t, f_t(x_t))|^2\big]\mathbb{I}_{\{\|f_t - f_H\|^2 \leq \bar{C}\log\frac{3T}{\delta}\}}\\ &\leq \sum_{t=\bar{t}}^{T}\eta_t^2\|\nabla\mathcal{E}(f_t)\|^2\big(4\kappa^2L^{\frac{1}{\alpha}}\big[\mathcal{E}(f_t) - \mathcal{E}(f_H)\big] + C_4\big)\mathbb{I}_{\{\|f_t - f_H\|^2 \leq \bar{C}\log\frac{3T}{\delta}\}}.\end{aligned} \tag{4.49}$$

Analyzing analogously to (4.19), one can show that $\nabla\mathcal{E}$ is $(\alpha, L\kappa^{1+\alpha})$-Hölder continuous. Then, Lemma 13 together with $\nabla\mathcal{E}(f_H)=0$ shows that

$$\hat{A}_t = \mathcal{E}(f_t) - \mathcal{E}(f_H) \le \frac{L\kappa^{1+\alpha}\|f_t - f_H\|^{1+\alpha}}{1+\alpha}. \tag{4.50}$$

Plugging the above inequality into (4.49) shows

$$\sum_{t=\bar{t}}^{T} \mathbb{E}_{z_t}\big(\bar{\xi}'_t - \mathbb{E}_{z_t}[\bar{\xi}'_t]\big)^2 \le C_8 \log\frac{3T}{\delta} \sum_{t=\bar{t}}^{T} \eta_t^2 \|\nabla\mathcal{E}(f_t)\|^2 \le \eta_{\widetilde{T}} C_8 \log\frac{3T}{\delta} \sum_{t=\bar{t}}^{T} \eta_t \|\nabla\mathcal{E}(f_t)\|^2, \tag{4.51}$$

where we have used $\bar{t} \ge \widetilde{T}$ and introduced

$$C_8 = \frac{4\kappa^{3+\alpha} L^{1+\frac{1}{\alpha}} \bar{C}}{1+\alpha} + C_4.$$

According to (4.39), there holds

$$\begin{aligned}
\bar{\xi}'_t - \mathbb{E}_{z_t}[\bar{\xi}'_t] &\le \eta_t \big|\big\langle \phi'(y_t, f_t(x_t))K_{x_t} - \nabla\mathcal{E}(f_t), \nabla\mathcal{E}(f_t)\big\rangle\big| \mathbb{I}_{\{\|f_t - f_H\|^2 \le \bar{C}\log\frac{3T}{\delta}\}} \\
&\le \eta_t \|\nabla\mathcal{E}(f_t)\| \big\|\phi'(y_t, f_t(x_t))K_{x_t} - \mathbb{E}_{z_t}[\phi'(y_t, f_t(x_t))K_{x_t}]\big\| \mathbb{I}_{\{\|f_t - f_H\|^2 \le \bar{C}\log\frac{3T}{\delta}\}} \\
&\le \big(C_3 + 2L\|f_t - f_H\|^\alpha \kappa^{\alpha+1}\big)\eta_t \|\nabla\mathcal{E}(f_t)\| \mathbb{I}_{\{\|f_t - f_H\|^2 \le \bar{C}\log\frac{3T}{\delta}\}}, \quad \forall t \ge \bar{t}.
\end{aligned}$$

Due to the $(\alpha, L\kappa^{1+\alpha})$-Hölder continuity of $\nabla\mathcal{E}$

$$\|\nabla\mathcal{E}(f_t)\| = \|\nabla\mathcal{E}(f_t) - \nabla\mathcal{E}(f_H)\| \le L\kappa^{1+\alpha}\|f_t - f_H\|^\alpha,$$

we further get

$$\begin{aligned}
\bar{\xi}'_t - \mathbb{E}_{z_t}[\bar{\xi}'_t] &\le \eta_t \big(C_3 + 2L\kappa^{\alpha+1}\big) \max\big(\|f_t - f_H\|^\alpha, 1\big) L\kappa^{1+\alpha}\|f_t - f_H\|^\alpha \mathbb{I}_{\{\|f_t - f_H\|^2 \le \bar{C}\log\frac{3T}{\delta}\}} \\
&\le \eta_{\widetilde{T}}\big(C_3 + 2L\kappa^{\alpha+1}\big) L\kappa^{1+\alpha}\bar{C}\log\frac{3T}{\delta} := \eta_{\widetilde{T}} C_9 \log\frac{3T}{\delta}, \quad \forall t \ge \bar{t}.
\end{aligned}$$

We can find a $\rho_2 > 0$ independent of T such that $(e^{\rho_2} - \rho_2 - 1)C_8 \le \rho_2 C_9$. Applying Part (b) of Lemma 21 with the above bounds on variances and magnitudes of $\bar{\xi}'_k$ followed with union bounds on probabilities, we can find a subset $\Omega' = \{(z_1, \ldots, z_T) : z_1, \ldots, z_T \in \mathcal{Z}\} \subset \mathcal{Z}^T$ with $\Pr\{\Omega'\} \ge 1 - \frac{\delta}{3}$ such that for any $(z_1, \ldots, z_T) \in \Omega'$ there holds (note $\mathbb{E}_{z_t}[\bar{\xi}'_t] = 0$)

$$\begin{aligned}
\sum_{t=\bar{t}}^{T} \bar{\xi}'_t &\le \frac{\eta_{\widetilde{T}}(e^{\rho_2} - \rho_2 - 1)C_8 \log\frac{3T}{\delta} \sum_{t=\bar{t}}^{T} \eta_t \|\nabla\mathcal{E}(f_t)\|^2}{\eta_{\widetilde{T}} \rho_2 C_9 \log\frac{3T}{\delta}} + \frac{\eta_{\widetilde{T}} C_9 \log^2\frac{3T}{\delta}}{\rho_2} \\
&\le \sum_{t=\bar{t}}^{T} \eta_t \|\nabla\mathcal{E}(f_t)\|^2 + \frac{\eta_{\widetilde{T}} C_9 \log^2\frac{3T}{\delta}}{\rho_2}, \quad \forall \bar{t} \in [\widetilde{T}, T].
\end{aligned} \tag{4.52}$$

Under the event $\Omega \cap \Omega'$, we can plug the above inequality with $\bar{t} = \tilde{t}, \bar{\xi}'_t = \bar{\xi}_t$ and $\|f_t - f_H\|^2 \le \bar{C}\log\frac{3T}{\delta}, \forall t = 1, \ldots, T$ into (4.48) to derive

$$\begin{aligned}
\hat{A}_{T+1} &\le \hat{A}_{\tilde{t}} + \Big(\tilde{a}\bar{C}\log\frac{3T}{\delta} + \tilde{b}\Big)\sum_{t=\tilde{t}}^{T} \eta_t^{1+\alpha} + \frac{\eta_{\widetilde{T}} C_9 \log^2\frac{3T}{\delta}}{\rho_2} \\
&\le \big[\sum_{t=\widetilde{T}}^{T} \eta_t\big]^{-1} \widetilde{C} \log^{\frac{3}{2}}\frac{3T}{\delta} + \Big(\tilde{a}\bar{C} + \tilde{b}\Big)\log\frac{3T}{\delta}\sum_{t=\tilde{t}}^{T} \eta_t^{1+\alpha} + \frac{\eta_{\widetilde{T}} C_9 \log^2\frac{3T}{\delta}}{\rho_2},
\end{aligned}$$

where the last inequality is due to (4.45). This establishes the stated inequality with probability $1-\delta$ and

$$\widetilde{C}' = \widetilde{C} + \tilde{a}\bar{C} + \tilde{b} + C_9\rho_2^{-1}.$$

It is clear that $\widetilde{C}'$ is independent of T and $\widetilde{T}$. The proof is complete. ■

Proof of Corollary 11 The polynomially decaying step sizes $\eta_t = \eta_1 t^{-\theta}(\theta > \frac{1}{2})$ satisfies the monotonicity and $\sum_{t=1}^{\infty}\eta_t^2 < \infty$ Furthermore, we have

$$\big[\sum_{t=\lfloor\frac{T}{2}\rfloor}^{T}\eta_t\big]^{-1} \leq \frac{2}{T\eta_T} = O(T^{\theta-1}) \quad \text{and} \quad \sum_{t=\lfloor\frac{T}{2}\rfloor}^{T}\eta_t^{1+\alpha} \leq \frac{(T+1)\eta_{\lfloor\frac{T}{2}\rfloor}^{1+\alpha}}{2} = O(T^{1-(1+\alpha)\theta}).$$

The proof is complete if we plug the above estimates into Theorem 10. ■

Acknowledgments

We thank the anonymous referees for their constructive suggestions. The work described in this paper is supported partially by the National Natural Science Foundation of China (Grants No. 11401524, 11531013, 11571078, 11631015, 11771012). Yunwen Lei is also supported by the Science and Technology Innovation Committee Foundation of Shenzhen (Grant No. ZDSYS201703031748284). Lei Shi is also supported by the Joint Research Fund by National Natural Science Foundation of China and Research Grants Council of Hong Kong (Project No. 11461161006 and RGC Project No. N_CityU120/14), Program of Shanghai Subject Chief Scientist (Project No.18XD1400700), and Zhuo Xue program of Fudan University. The corresponding author is Zheng-Chu Guo.

Appendix A. Some Additional Proofs

Proof of Lemma 16 Let $\epsilon > 0$ be an arbitrary number. Since $\lim_{t\to\infty}\eta_t = 0$ we can find a $t_3 \in \mathbb{N}$ such that $\eta_t \leq \frac{\epsilon}{2}$ for all $t \geq t_3$. Since $\sum_{t=1}^{\infty}\eta_t = \infty$, we can also find a $t_4 > t_3$ such that $\sum_{k=1}^{t_3}\eta_k^2 \leq \frac{\epsilon}{2}\sum_{k=1}^{t_4}\eta_k$. Then, for any $t \geq t_4$, it holds

$$\begin{aligned}\big[\sum_{k=1}^{t}\eta_k\big]^{-1}\sum_{k=1}^{t}\eta_k^2 &= \big[\sum_{k=1}^{t}\eta_k\big]^{-1}\sum_{k=1}^{t_3}\eta_k^2 + \big[\sum_{k=1}^{t}\eta_k\big]^{-1}\sum_{k=t_3+1}^{t}\eta_k^2 \\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2}\big[\sum_{k=1}^{t}\eta_k\big]^{-1}\sum_{k=t_3+1}^{t}\eta_k \leq \epsilon.\end{aligned}$$

Since $\epsilon > 0$ is arbitrarily chosen, the proof is complete. ■

Proof of Lemma 13 Fix $f, \tilde{f} \in H$. Define a function $g : \mathbb{R} \to \mathbb{R}$ by $g(t) = \mathcal{G}(\tilde{f} + t(f - \tilde{f}))$. It is clear that $g'(t) = \langle f - \tilde{f}, \nabla \mathcal{G}(\tilde{f} + t(f - \tilde{f}))\rangle$ and

$$\begin{aligned}|g'(t) - g'(\tilde{t})| &= \Big\langle f - \tilde{f}, \nabla \mathcal{G}(\tilde{f} + t(f - \tilde{f})) - \nabla \mathcal{G}(\tilde{f} + \tilde{t}(f - \tilde{f}))\Big\rangle \\ &\le \|f - \tilde{f}\| \big\|\nabla \mathcal{G}\big(\tilde{f} + t(f - \tilde{f})\big) - \nabla \mathcal{G}\big(\tilde{f} + \tilde{t}(f - \tilde{f})\big)\big\| \\ &\le L\|f - \tilde{f}\|^{1+\alpha}|t - \tilde{t}|^{\alpha}.\end{aligned}$$

It then follows that

$$\begin{aligned}g(1) - g(0) - g'(0) &= \int_0^1 [g'(t) - g'(0)]dt \le \int_0^1 |g'(t) - g'(0)|dt \\ &\le L\|f - \tilde{f}\|^{1+\alpha} \int_0^1 t^{\alpha} dt = \frac{L\|f - \tilde{f}\|^{1+\alpha}}{1+\alpha},\end{aligned}$$

which amounts to the second inequality in (4.1)

$$\mathcal{G}(f) \le \mathcal{G}(\tilde{f}) + \langle f - \tilde{f}, \nabla \mathcal{G}(\tilde{f})\rangle + \frac{L\|f - \tilde{f}\|^{1+\alpha}}{1+\alpha}. \tag{A.1}$$

We now turn to the first inequality in (4.1). Fix f and $\tilde{f} \in H$. Define a functional $\mathcal{L} : H \to \mathbb{R}$ by $\mathcal{L}(\bar{f}) = \mathcal{G}(\bar{f}) - \langle \bar{f}, \nabla \mathcal{G}(f)\rangle$. It is clear that $\mathcal{L}$ is a convex function and $\nabla \mathcal{L}(f) = \nabla \mathcal{G}(f) - \nabla \mathcal{G}(f) = 0$. According to the first-order optimality condition, we know $\mathcal{L}$ attains its minimum at f and

$$\begin{aligned}\mathcal{L}(f) &= \min_{\bar{f} \in H} \mathcal{L}(\bar{f}) = \min_{\bar{f} \in H} \big[\mathcal{G}(\bar{f}) - \langle \bar{f}, \nabla \mathcal{G}(f)\rangle\big] \\ &\le \min_{\bar{f} \in H} \Big[\mathcal{G}(\tilde{f}) + \langle \bar{f} - \tilde{f}, \nabla \mathcal{G}(\tilde{f})\rangle + \frac{L\|\tilde{f} - \bar{f}\|^{1+\alpha}}{1+\alpha} - \langle \bar{f}, \nabla \mathcal{G}(f)\rangle\Big] \\ &= \mathcal{L}(\tilde{f}) + \min_{\bar{f} \in H} \Big[\langle \tilde{f} - \bar{f}, \nabla \mathcal{G}(f) - \nabla \mathcal{G}(\tilde{f})\rangle + \frac{L\|\tilde{f} - \bar{f}\|^{1+\alpha}}{1+\alpha}\Big] \\ &= \mathcal{L}(\tilde{f}) + \min_{\bar{f} \in H} \Big[\langle \bar{f}, \nabla \mathcal{G}(f) - \nabla \mathcal{G}(\tilde{f})\rangle + \frac{L\|\bar{f}\|^{1+\alpha}}{1+\alpha}\Big],\end{aligned}$$

where the inequality follows from (A.1). Taking $\bar{f} = L^{-\frac{1}{\alpha}}\|\nabla \mathcal{G}(\tilde{f}) - \nabla \mathcal{G}(f)\|^{\frac{1-\alpha}{\alpha}}\big(\nabla \mathcal{G}(\tilde{f}) - \nabla \mathcal{G}(f)\big)$ in the above inequality, we derive

$$\begin{aligned}\mathcal{L}(f) &\le \mathcal{L}(\tilde{f}) - L^{-\frac{1}{\alpha}}\|\nabla \mathcal{G}(\tilde{f}) - \nabla \mathcal{G}(f)\|^{\frac{1+\alpha}{\alpha}} + \frac{L^{-\frac{1}{\alpha}}\|\nabla \mathcal{G}(\tilde{f}) - \nabla \mathcal{G}(f)\|^{\frac{1+\alpha}{\alpha}}}{1+\alpha} \\ &= \mathcal{L}(\tilde{f}) - \frac{\alpha L^{-\frac{1}{\alpha}}}{1+\alpha}\|\nabla \mathcal{G}(f) - \nabla \mathcal{G}(\tilde{f})\|^{\frac{1+\alpha}{\alpha}}.\end{aligned}$$

This establishes the first inequality in (4.1). The proof is complete. ■

Proof of Proposition 24 Consider the following sequence of functionals $\tilde{\xi}_k, k = 1, \ldots, T$ by

$$\tilde{\xi}_k = \big\langle f_H - f_k, \phi'(y_k, f_k(x_k))K_{x_k} - \mathbb{E}_{z_k}\big[\phi'(y_k, f_k(x_k))K_{x_k}\big]\big\rangle,$$

where $\bar{C}$ is defined in Proposition 9. Eq. (4.42) implies that

$$|\tilde{\xi}_k|\mathbb{I}_{\{\|f_k-f_H\|^2\le\bar{C}\log\frac{2T}{\delta}\}}\le\big(C_3+2L\kappa^{\alpha+1}\big)\bar{C}\log\frac{2T}{\delta}.$$

By Part (a) of Lemma 21, there exists a subset $\Omega'=\{(z_1,\dots,z_T):z_1,\dots,z_T\in\mathcal{Z}\}\subset\mathcal{Z}^T$ with probability measure $\Pr\{\Omega'\}\ge1-\frac{\delta}{2}$ such that for any $(z_1,\dots,z_T)\in\Omega'$ the following inequality holds

$$\sum_{k=1}^T\tilde{\xi}_k\mathbb{I}_{\{\|f_k-f_H\|^2\le\bar{C}\log\frac{2T}{\delta}\}}\le\big(C_3+2L\kappa^{\alpha+1}\big)\bar{C}\log\frac{2T}{\delta}\Big(2T\log\frac{2}{\delta}\Big)^{\frac{1}{2}}.$$

According to Proposition 9, there exists a subset $\Omega=\{(z_1,\dots,z_T):z_1,\dots,z_T\in\mathcal{Z}\}\subset\mathcal{Z}^T$ with probability measure $\Pr\{\Omega\}\ge1-\frac{\delta}{2}$ such that for any $(z_1,\dots,z_T)\in\Omega$ there holds the inequality $\max_{1\le k\le T}\|f_k-f_H\|^2\le\bar{C}\log\frac{2T}{\delta}$. Under the event $\Omega\cap\Omega'$, we then have

$$\sum_{k=1}^T\tilde{\xi}_k\le\big(C_3+2L\kappa^{\alpha+1}\big)\bar{C}\log\frac{2T}{\delta}\Big(2T\log\frac{2}{\delta}\Big)^{\frac{1}{2}}.\tag{A.2}$$

Furthermore, it follows from (4.37) that

$$2[\mathcal{E}(f_k)-\mathcal{E}(f_H)]\le\eta_k^{-1}\big[\|f_k-f_H\|^2-\|f_{k+1}-f_H\|^2\big]+\eta_k\kappa^2\big(A\phi(y_k,f_k(x_k))+B\big)+2\tilde{\xi}_k.$$

Taking a summation of the above inequality from $k=1$ to T yields the following inequality under the event $\Omega\cap\Omega'$

$$2\sum_{k=1}^T[\mathcal{E}(f_k)-\mathcal{E}(f_H)]\le\sum_{k=1}^{T-1}\big(\eta_{k+1}^{-1}-\eta_k^{-1}\big)\|f_{k+1}-f_H\|^2+\eta_1^{-1}\|f_1-f_H\|^2$$
$$+\kappa^2\sum_{k=1}^T\eta_k\big(A\phi(y_k,f_k(x_k))+B\big)+2\sum_{k=1}^T\tilde{\xi}_k.\tag{A.3}$$

It follows from (4.33) that

$$\sum_{k=1}^T\eta_k\phi(y_k,f_k(x_k))\le\|f_H\|^2+2\sum_{k=1}^T\eta_k\phi(y_k,f_H(x_k))+\kappa^2B\sum_{k=1}^T\eta_k^2.$$

Plugging the above bound into (A.3) and using the monotonicity of η_k together with (A.2), we derive the following inequality with probability at least $1-\delta$

$$2\sum_{k=1}^T[\mathcal{E}(f_k)-\mathcal{E}(f_H)]\le(A\kappa^2+\eta_1^{-1})\|f_H\|^2+\kappa^2\sum_{k=1}^T\Big(2A\eta_k\sup_z\phi(y,f_H(x))+B\eta_k+AB\kappa^2\eta_k^2\Big)$$
$$+\big(C_3+2L\kappa^{\alpha+1}\big)\bar{C}(8T)^{\frac{1}{2}}\log^{\frac{3}{2}}\frac{2T}{\delta}.$$

The proof is complete. ■

References

Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

Léon Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8), 1991.

Léon Bottou. Online learning and stochastic approximations. *On-line Learning in Neural Networks*, 17(9):142, 1998.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

Di-Rong Chen, Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Joseph L Doob. *Measure Theory, Graduate Texts in Mathematics*. Springer, 1994.

John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *Conference on Learning Theory*, pages 14–26, 2010.

Zheng-Chu Guo and Lei Shi. Fast and strong convergence of online learning algorithms. *submitted*, 2017.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jyrki Kivinen, Alexander J Smola, and Robert C Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2004.

Yunwen Lei and Ding-Xuan Zhou. Convergence of online mirror descent algorithms. *submitted*, 2017.

Junhong Lin and Ding-Xuan Zhou. Learning theory of randomized Kaczmarz algorithm. *Journal of Machine Learning Research*, 16:3341–3365, 2015.

Junhong Lin and Ding-Xuan Zhou. Online learning algorithms can converge comparably fast as batch learning. *IEEE Transactions on Neural Networks and Learning Systems*, in press, 2018. doi: 10.1109/TNNLS.2017.2677970.

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

Klaus-Robert Müller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, Quoc V Le, and Andrew Y Ng. On optimization methods for deep learning. In *International Conference on Machine Learning*, pages 265–272, 2011.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pages 449–456, 2012.

Bernhard Schölkopf and Alexander J Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2001.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical Programming*, 127(1):3–30, 2011.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

Steve Smale and Yuan Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6(2):145–170, 2006.

Steve Smale and Ding-Xuan Zhou. Online learning with markov sampling. *Analysis and Applications*, 7(01):87–113, 2009.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93, 2001.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, 2013.

Pierre Tarres and Yuan Yao. Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, 2014.

Yuan Yao. On complexity issues of online learning algorithms. *IEEE Transactions on Information Theory*, 56(12):6470–6481, 2010.

Gui-Bo Ye and Ding-Xuan Zhou. Fully online classification by regularization. *Applied and Computational Harmonic Analysis*, 23(2):198–214, 2007.

Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

Yiming Ying and Ding-Xuan Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, 2006.

Yiming Ying and Ding-Xuan Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 2(42):224–244, 2017.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, pages 919–926, 2004.

Tong Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Conference on Learning Theory*, pages 173–187, 2005.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.