Appendix

A Notation and Useful Propositions

Let V be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_V$ and the induced norm $\| \cdot \|_V$. For $A : V \to V$, we denote an operator norm of A as $\|A\|_{\mathrm{op}}$, that is,

$$\|A\|_{\mathrm{op}} \stackrel{\mathrm{def}}{=} \sup_{v \in V} \frac{\|Av\|_V}{\|v\|_V}.$$

For $a, b \in V$, we define an outer product $a \otimes_V b : V \to V$ as follows:

$$(a \otimes_V b)v \stackrel{\mathrm{def}}{=} \langle b, v \rangle_V a, \quad \forall v \in V.$$

Let W be a closed subspace of V, then a projection onto W is well defined and we denote its operator by $\mathcal{P}_W$. Then we have

$$v = \mathcal{P}_W v + \mathcal{P}_{W^\perp} v, \quad \forall v \in V.$$

Furthermore, we define a partial order $\preceq$ between linear, positive semi-definite and self-adjoint operators $A, B : V \to V$ as follows:

$$A \preceq B \quad \stackrel{\mathrm{def}}{\iff} \quad \langle Av, v \rangle_V \leq \langle Bv, v \rangle_V, \ \forall v \in V.$$

The following inequality shows that the difference between the square root of two self-adjoint positive semi-definite operators is bounded by the square root of the difference of them.

Proposition 1. *Let V be a separable Hilbert space. For any compact, positive semi-definite, self-adjoint operators $S, \widetilde{S} : V \to V$, the following inequality holds:*

$$\|S^{1/2} - \widetilde{S}^{1/2}\|_{\mathrm{op}} \leq \|S - \widetilde{S}\|_{\mathrm{op}}^{1/2} \tag{9}$$

Proof. Since $S^{1/2} - \widetilde{S}^{1/2}$ is also a compact and self-adjoint operator, it allows eigendecomposition of itself. Then let $\lambda_{\max}$ be the eigenvalue with largest absolute value and v be the corresponding normalized eigenfunction of $S^{1/2} - \widetilde{S}^{1/2}$, i.e.,

$$(S^{1/2} - \widetilde{S}^{1/2})v = \lambda_{\max} v.$$

Since (9) obviously holds if $S = \widetilde{S}$, we can assume that $\lambda_{\max} > 0$ without loss of generality. Because $S^{1/2}$ is positive semi-definite, we have

$$\begin{aligned}
\langle v, Sv \rangle_V &= \|S^{1/2} v\|_V^2 \\
&= \|\widetilde{S}^{1/2} v + \lambda_{\max} v\|_V^2 \\
&= \langle v, \widetilde{S} v \rangle_V + \lambda_{\max}^2 + 2\lambda_{\max} \langle v, S^{1/2} v \rangle_V \\
&\geq \langle v, \widetilde{S} v \rangle_V + \lambda_{\max}^2.
\end{aligned}$$

Thus we have

$$\begin{aligned}
\|S - \widetilde{S}\|_{\mathrm{op}} &\geq \langle v, (S - \widetilde{S}) v \rangle_V \\
&\geq \lambda_{\max}^2 = \|S^{1/2} - \widetilde{S}^{1/2}\|_{\mathrm{op}}^2,
\end{aligned}$$

which completes the proof. □

The following inequality is a generalization of the Bernstein inequality to random operators on separable Hilbert space and used in Lemma 1 to derive the concentration of integral operators.

Proposition 2 (Proposition 3 in Rudi and Rosasco (2017)). *Let V be a separable Hilbert space and let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed self-adjoint random operators on V. Assume that $\mathbb{E}X_i = 0$ and there exists $B > 0$ such that $\|X_i\|_{\mathrm{op}} \leq B$ almost surely for any $1 \leq i \leq n$. Let S be the positive operator such that $\mathbb{E}X_i^2 \leq S$. Then for any $\delta \in (0, 1]$, the following inequality holds with probability at least $1 - \delta$:*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} X_i \right\|_{\mathrm{op}} \leq \frac{2B\beta}{3n} + \sqrt{\frac{2\|S\|_{\mathrm{op}}\beta}{n}},$$

where $\beta = \log \frac{2\mathrm{tr}S}{\|S\|_{\mathrm{op}}\delta}$.

B Basic Properties of RKHS

In analyses of kernel methods, it is common to assume $\mathcal{X}$ is compact, $\rho_{\mathcal{X}}$ has the full support and k is continuous because under such assumptions we utilize Mercer's theorem to characterize RKHS Cucker and Smale (2002); Aronszajn (1950). However, such an assumption may not be adopted under the strong low noise condition in which $\rho_{\mathcal{X}}$ may not have full support. In this section, we explain some basic properties of reproducing kernel Hilbert space (RKHS) under more general settings based on Dieuleveut and Bach (2016); Steinwart and Scovel (2012).

First, for given kernel function k and its RKHS $\mathcal{H}$, we define a covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$ as follows:

$$\langle f, \Sigma g \rangle_{\mathcal{H}} = \langle f, g \rangle_{L^2(\rho_{\mathcal{X}})}, \quad \forall f, g \in \mathcal{H}.$$

It is well-defined through Riesz' representation theorem. Using reproducing property, we have

$$\begin{aligned} &\Sigma = \mathbb{E}_{X \sim \rho_{\mathcal{X}}}[k(\cdot, X) \otimes_{\mathcal{H}} k(\cdot, X)], \\ &(\Sigma f)(z) = \mathbb{E}_{X \sim \rho_{\mathcal{X}}}[f(X)k(X, z)], \quad \forall f \in \mathcal{H}. \end{aligned} \tag{10}$$

where expectation is defined via a Bochner integration. From the representation (10), we can extend the covariance operate to $f \in L^2(\rho_{\mathcal{X}})$. We denote this by $T : L^2(\rho_{\mathcal{X}}) \to L^2(\rho_{\mathcal{X}})$ as follows:

$$(Tf)(z) = \mathbb{E}_{X \sim \rho_{\mathcal{X}}}[f(X)k(X, z)], \quad \forall f \in L^2(\rho_{\mathcal{X}}).$$

$\mathrm{Im}(T) \subset L^2(\rho_{\mathcal{X}})$ is verified since $k(\cdot, x)$ is uniformly bounded by Assumption 2. Also, we can write T using feature expansion (4) as

$$T = \mathbb{E}_{\omega \sim \tau}[\varphi(\cdot, \omega) \otimes_{L^2(\rho_{\mathcal{X}})} \varphi(\cdot, \omega)], \tag{11}$$

since

$$\begin{aligned} (Tf)(z) &= \mathbb{E}_{X \sim \rho_{\mathcal{X}}}[f(X)\mathbb{E}_{\omega \sim \tau}[\varphi(X, \omega)\varphi(z, \omega)]] \\ &= \mathbb{E}_{\omega \sim \tau}[\langle f, \varphi(\cdot, \omega) \rangle_{L^2(\rho_{\mathcal{X}})} \varphi(z, \omega)]. \end{aligned}$$

Following Dieuleveut and Bach (2016), here we denote a set of square integral function itself by $\mathcal{L}^2(d\rho_{\mathcal{X}})$, that is, its quotient is $L^2(\rho(\mathcal{X}))$, which is separable Hilbert space. We can also define the extended covariance operator $\mathcal{T} : L^2(\rho_{\mathcal{X}}) \to \mathcal{L}^2(d\rho_{\mathcal{X}})$ as follows:

$$(\mathcal{T}f)(z) = \mathbb{E}_{X \sim \rho_{\mathcal{X}}}[f(X)k(X, z)], \quad \forall f \in L^2(\rho_{\mathcal{X}}).$$

Here we present some properties of these covariance operators $\Sigma, T, \mathcal{T}$ from Dieuleveut and Bach (2016).

Proposition 3.

1. *Σ is self-adjoint, continuous operator and $\mathrm{Ker}(\Sigma) = \{f \in \mathcal{H} \mid \|f\|_{L^2(\rho_{\mathcal{X}})} = 0\}$.*
2. *T is continuous, self-adjoint, positive semi-definite operator.*
3. *$\mathcal{T}^{1/2} : \mathrm{Ker}(T)^{\perp} \to \mathrm{Ker}(\Sigma)^{\perp}$ is well-defined and an isometry. In particular, for any $f \in \mathrm{Ker}(\Sigma)^{\perp} \subset \mathcal{H}$, there exists $g \in \mathrm{Ker}(T)^{\perp} \subset L^2(\rho_{\mathcal{X}})$ such that $\|f\|_{\mathcal{H}} = \|g\|_{L^2(\rho_{\mathcal{X}})}$.*

We denote the extended covariate operator associate with k_M by $T_M : L^2(\rho_{\mathcal{X}}) \to L^2(\rho_{\mathcal{X}})$ and $\mathcal{T}_M : L^2(\rho_{\mathcal{X}}) \to \mathcal{L}^2(d\rho_{\mathcal{X}})$.

As with (11), we have

$$T_M = \frac{1}{M}\sum_{i=1}^{M} \varphi(\cdot,\omega_i) \otimes_{L^2(\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i),$$
$$\mathbb{E}[T_M] = T.$$

The next lemma provides a probabilistic bounds about the difference of the two covariate operators T and T_M.

Lemma 1. *For any $\delta \in [0,1)$, the following inequality holds with probability at least $1-\delta$:*

$$\|T - T_M\|_{\mathrm{op}} \leq R^2\left(\frac{2\beta}{3M} + \sqrt{\frac{2\beta}{M}}\right)$$

where $\beta = \log \frac{2R^2}{\|T\|_{\mathrm{op}}\delta}$.

Proof. Let $X_i = T - \varphi(\cdot,\omega_i) \otimes_{L^2(\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i)$. Then $T - T_M = \frac{1}{M}\sum_{i=1}^{M} X_i$. Also, we have

$$\begin{aligned}
\mathbb{E}X_i &= 0,\\
X_i &\preceq T \preceq R^2 I,\\
X_i &\succeq -\varphi(\cdot,\omega_i) \otimes_{L^2(\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i) \succeq -R^2 I,\\
||X_i||_{\mathrm{op}} &\leq R^2, \text{as a result of two previous inequalities,}\\
\mathbb{E}X_i^2 &= \mathbb{E}\left[\varphi(\cdot,\omega_i) \otimes_{L^2(\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i)\right]^2 - T^2\\
&\preceq \mathbb{E}\left[\varphi(\cdot,\omega_i) \otimes_{L^2(\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i)\right]^2\\
&\preceq \mathbb{E}\left[\langle\varphi(\cdot,\omega_i),\varphi(\cdot,\omega_i)\rangle_{L^2(\rho_{\mathcal{X}})}\varphi(\cdot,\omega_i) \otimes_{L^2(\rho_{\mathcal{X}})} \varphi(\cdot,\omega_i)\right]\\
&\preceq R^2 T,\\
\mathrm{tr}T &= \int_{\mathcal{X}} k(x,x)\mathrm{d}\rho_{\mathcal{X}}(x) \leq R^2.
\end{aligned}$$

Let $B = R^2$ and $S = R^2 T$ in Proposition 2, we have

$$\begin{aligned}
\|T - T_M\|_{\mathrm{op}} &= \left\|\frac{1}{M}\sum_{i=1}^{M} X_i\right\|_{\mathrm{op}}\\
&\leq \frac{2R^2\beta}{3M} + \sqrt{\frac{2R^2\|T\|_{\mathrm{op}}\beta}{M}}\\
&\leq R^2\left(\frac{2\beta}{3M} + \sqrt{\frac{2\beta}{M}}\right),
\end{aligned}$$

which completes the proof. □

Let $\mathcal{H}$ and $\mathcal{H}_M$ be RKHSs associate with kernels k and k_M, respectively. Using Proposition 3 and Lemma 1, we have the following proposition, which is essential in the proof of Theorem 1.

Lemma 2. *For any $\delta \in (0,1]$ and $\xi > 0$, if*

$$M \geq \max\left\{\frac{8}{3}\left(\frac{R}{\xi}\right)^2, 32\left(\frac{R}{\xi}\right)^4\right\}\log\frac{2R^2}{\|T\|_{\mathrm{op}}\delta}$$

holds, the following statement holds with probability at least $1-\delta$:
For any $g \in \mathcal{H}$, there exists $\widetilde{g} \in \mathcal{H}_M$ that satisfies

- $\|g - \widetilde{g}\|_{L^2(\rho_{\mathcal{X}})} \leq \xi \|g\|_{\mathcal{H}}$
- $\|g\|_{\mathcal{H}} \geq \|\widetilde{g}\|_{\mathcal{H}_M}$.

Also, for any $\widetilde{g} \in \mathcal{H}_M$, there exists $g \in \mathcal{H}$ that satisfies

- $\|g - \widetilde{g}\|_{L^2(\rho_{\mathcal{X}})} \leq \xi \|\widetilde{g}\|_{\mathcal{H}_M}$
- $\|g\|_{\mathcal{H}} \leq \|\widetilde{g}\|_{\mathcal{H}_M}$.

Proof. We show the first part of the statement. The latter half can be shown in the same manner. For $g \in \mathcal{H}$, set $\widetilde{g} = \mathcal{T}_M^{1/2} \mathcal{P}_{\mathrm{Ker}(T_M)^\perp} \mathcal{T}^{-1/2} \mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g \in \mathcal{H}_M$. Then we have

$$\begin{aligned}
\|\widetilde{g}\|_{\mathcal{H}_M} &= \|\mathcal{P}_{\mathrm{Ker}(T_M)^\perp} \mathcal{T}^{-1/2} \mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g\|_{L^2(\rho_{\mathcal{X}})} \\
&\leq \|\mathcal{T}^{-1/2} \mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g\|_{L^2(\rho_{\mathcal{X}})} \\
&= \|\mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g\|_{\mathcal{H}} \\
&\leq \|g\|_{\mathcal{H}}.
\end{aligned}$$

Moreover, by Proposition 1 and Lemma 1, with probability at least $1 - \delta$, we have

$$\begin{aligned}
\|g - \widetilde{g}\|_{L^2(\rho_{\mathcal{X}})} &= \|\mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g - \widetilde{g}\|_{L^2(\rho_{\mathcal{X}})} \quad (\because \text{Proposition 3.1}) \\
&= \|\mathcal{T}^{1/2} h - \mathcal{T}_M^{1/2} \mathcal{P}_{\mathrm{Ker}(T_M)^\perp} h\|_{L^2(\rho_{\mathcal{X}})} \\
&= \|T^{1/2} h - T_M^{1/2} h\|_{L^2(\rho_{\mathcal{X}})} \\
&\leq \|T^{1/2} - T_M^{1/2}\|_{\mathrm{op}} \|h\|_{L^2(\rho_{\mathcal{X}})} \\
&\leq \|T - T_M\|_{\mathrm{op}}^{1/2} \|g\|_{\mathcal{H}} \\
&\leq \left(R^2 \left(\frac{2\beta}{3M} + \sqrt{\frac{2\beta}{M}} \right) \right)^{1/2} \|g\|_{\mathcal{H}} \\
&\leq R \left(\left(\frac{2\beta}{3M} \right)^{1/2} + \left(\frac{2\beta}{M} \right)^{1/4} \right) \|g\|_{\mathcal{H}}
\end{aligned}$$

where $h = \mathcal{T}^{-1/2} \mathcal{P}_{\mathrm{Ker}(\Sigma)^\perp} g \in L^2(\rho_{\mathcal{X}})$ and $\beta = \log \frac{2R^2}{\|T\|_{\mathrm{op}} \delta}$.
Solving the equation $\max \left\{ \left(\frac{2\beta}{3M} \right)^{1/2}, \left(\frac{2\beta}{M} \right)^{1/4} \right\} \leq \frac{\xi}{2R}$, we get a desired result. □

C Proof of Theorem 1

In this section, we give the complete statement and proof of Theorem 1.

Theorem 1. *Define $\xi > 0$ such that*

$$\xi = \min \left\{ \left(\frac{\epsilon}{2^{p+1} C(\delta) \|g_*\|_{\mathcal{H}}} \right)^{1/1-p}, \frac{\lambda \epsilon^2}{2^4 \cdot 3 R^2 L \|g_*\|_{\mathcal{H}}}, \left(\frac{\lambda^3 \epsilon^4}{2^7 \cdot 3^2 R^4 L^2 \mathcal{L}(g_*)} \right)^{1/2}, \left(\frac{\lambda^3 \epsilon^4}{2^7 \cdot 3^2 R^4 L^3 \|g_*\|_{\mathcal{H}}} \right)^{1/3} \right\}.$$

Then a number of random features M which satisfies

$$M \geq \max \left\{ \frac{8}{3} \left(\frac{R}{\xi} \right)^2, 32 \left(\frac{R}{\xi} \right)^4 \right\} \log \frac{2R^2}{\|T\|_{\mathrm{op}} \delta}$$

is enough to guarantee, with probability at least $1 - 2\delta$, that

$$\|g_\lambda - g_{M,\lambda}\|_{L^\infty(\rho_{\mathcal{X}})} \leq \epsilon.$$

Proof. By Lemma 2, for given $\xi > 0$, if we have a number of feature M such that

$$M \geq \max\left\{\frac{8}{3}\left(\frac{R}{\xi}\right)^2, 32\left(\frac{R}{\xi}\right)^4\right\}\log\frac{2R^2}{\|T\|_{\mathrm{op}}\delta},$$

we can take $\widetilde{g}_\lambda \in \mathcal{H}_M, \widetilde{g}_{M,\lambda} \in \mathcal{H}$ which satisfy the following conditions:

$$\|g_\lambda\|_{\mathcal{H}} \geq \|\widetilde{g}_\lambda\|_{\mathcal{H}_M} \tag{12}$$
$$\|g_{M,\lambda}\|_{\mathcal{H}_M} \geq \|\widetilde{g}_{M,\lambda}\|_{\mathcal{H}} \tag{13}$$
$$\|\widetilde{g}_{M,\lambda} - g_{M,\lambda}\|_{L^2(\rho_{\mathcal{X}})} \leq \xi\|g_{M,\lambda}\|_{\mathcal{H}_M} \tag{14}$$
$$\|\widetilde{g}_\lambda - g_\lambda\|_{L^2(\rho_{\mathcal{X}})} \leq \xi\|g_\lambda\|_{\mathcal{H}} \tag{15}$$

By λ-strong convexity with respect to RKHS norm, we have

$$\mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2 + \frac{\lambda}{2}\|g_\lambda - \widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 \leq \mathcal{L}(\widetilde{g}_{M,\lambda}) + \frac{\lambda}{2}\|\widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 \tag{16}$$
$$\mathcal{L}(g_{M,\lambda}) + \frac{\lambda}{2}\|g_{M,\lambda}\|_{\mathcal{H}_M}^2 + \frac{\lambda}{2}\|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \leq \mathcal{L}(\widetilde{g}_\lambda) + \frac{\lambda}{2}\|\widetilde{g}_\lambda\|_{\mathcal{H}_M}^2. \tag{17}$$

In addition, by L-Lipschitzness of $\mathcal{L}$ with respect to $L^2(\rho_{\mathcal{X}})$ norm in Assumption 1 and (14)(15), we have

$$\begin{aligned}\mathcal{L}(\widetilde{g}_{M,\lambda}) &\leq \mathcal{L}(g_{M,\lambda}) + L\|\widetilde{g}_{M,\lambda} - g_{M,\lambda}\|_{L^2(\rho_{\mathcal{X}})} \\ &\leq \mathcal{L}(g_{M,\lambda}) + L\xi\|g_{M,\lambda}\|_{\mathcal{H}_M} \qquad (18)\\ \mathcal{L}(\widetilde{g}_\lambda) &\leq \mathcal{L}(g_\lambda) + L\|\widetilde{g}_\lambda - g_\lambda\|_{L^2(\rho_{\mathcal{X}})} \\ &\leq \mathcal{L}(g_\lambda) + L\xi\|g_\lambda\|_{\mathcal{H}} \qquad (19)\end{aligned}$$

By inequalities (16)(17)(18)(19) and (12)(13), we have

$$\begin{aligned}&\mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2 + \frac{\lambda}{2}\left(\|g_\lambda - \widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 + \|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2\right) \\ &\leq \mathcal{L}(\widetilde{g}_{M,\lambda}) + \frac{\lambda}{2}\|\widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 + \frac{\lambda}{2}\|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \\ &\leq \mathcal{L}(g_{M,\lambda}) + L\xi\|g_{M,\lambda}\|_{\mathcal{H}_M} + \frac{\lambda}{2}\|g_{M,\lambda}\|_{\mathcal{H}_M}^2 + \frac{\lambda}{2}\|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \\ &\leq \mathcal{L}(\widetilde{g}_\lambda) + \frac{\lambda}{2}\|\widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 + L\xi\|g_{M,\lambda}\|_{\mathcal{H}_M} \\ &\leq \mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2 + L\xi\left(\|g_\lambda\|_{\mathcal{H}} + \|g_{M,\lambda}\|_{\mathcal{H}_M}\right).\end{aligned}$$

Thus we have

$$\|g_\lambda - \widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 + \|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \leq \frac{2L\xi}{\lambda}\left(\|g_\lambda\|_{\mathcal{H}} + \|g_{M,\lambda}\|_{\mathcal{H}_M}\right). \tag{20}$$

In addition, by (17) and (19), we have

$$\begin{aligned}\frac{\lambda}{2}\|g_{M,\lambda}\|_{\mathcal{H}_M}^2 &\leq \mathcal{L}(\widetilde{g}_\lambda) + \frac{\lambda}{2}\|\widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 \\ &\leq \mathcal{L}(g_\lambda) + L\xi\|g_\lambda\|_{\mathcal{H}} + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2. \qquad (21)\end{aligned}$$

Combining (20) and (21), we obtain

$$\begin{aligned}\|g_\lambda - \widetilde{g}_{M,\lambda}\|_{\mathcal{H}}^2 + \|g_{M,\lambda} - \widetilde{g}_\lambda\|_{\mathcal{H}_M}^2 &\leq \frac{2L\xi}{\lambda}\left(\|g_\lambda\|_{\mathcal{H}} + \left(\frac{2}{\lambda}\mathcal{L}(g_\lambda) + \frac{2L\xi}{\lambda}\|g_\lambda\|_{\mathcal{H}} + \|g_\lambda\|_{\mathcal{H}}^2\right)^{1/2}\right) \\ &\leq \frac{2L\xi}{\lambda}\left(\|g_*\|_{\mathcal{H}} + \left(\frac{2}{\lambda}\mathcal{L}(g_*) + \frac{2L\xi}{\lambda}\|g_*\|_{\mathcal{H}} + \|g_*\|_{\mathcal{H}}^2\right)^{1/2}\right) \\ &\leq \frac{2L\xi}{\lambda}\left(2\|g_*\|_{\mathcal{H}} + \left(\frac{2}{\lambda}\mathcal{L}(g_*)\right)^{1/2} + \left(\frac{2L\xi}{\lambda}\|g_*\|_{\mathcal{H}}\right)^{1/2}\right).\end{aligned}$$

In the second inequality, we used $\|g_*\|_{\mathcal{H}} \geq \|g_\lambda\|_{\mathcal{H}}$ and $\mathcal{L}(g_*) + \frac{\lambda}{2}\|g_*\|_{\mathcal{H}}^2 \geq \mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_{\mathcal{H}}^2$. In the third inequality, we used $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$ for $a, b > 0$. Then by Assumption 2, we obtain

$$\|g_{M,\lambda} - \widetilde{g}_\lambda\|_{L^\infty(\rho_\mathcal{X})} \leq R \max\left\{ \left(\frac{12L\xi}{\lambda}\|g_*\|_{\mathcal{H}}\right)^{1/2}, \left(\frac{72L^2\xi^2}{\lambda^3}\mathcal{L}(g_*)\right)^{1/4}, \left(\frac{72L^3\xi^3}{\lambda^3}\|g_*\|_{\mathcal{H}}\right)^{1/4} \right\}. \tag{22}$$

On the other hand, by Assumption 3, we have

$$\begin{aligned} \|g_\lambda - \widetilde{g}_\lambda\|_{L^\infty(\rho_\mathcal{X})} &\leq C(\delta)\|g_\lambda - \widetilde{g}_\lambda\|_{\mathcal{H}_M^+}^p \|g_\lambda - \widetilde{g}_\lambda\|_{L^2(\rho_\mathcal{X})}^{1-p} \\ &\leq C(\delta)(\|g_\lambda\|_{\mathcal{H}} + \|\widetilde{g}_\lambda\|_{\mathcal{H}_M})^p (\xi\|g_\lambda\|_{\mathcal{H}})^{1-p} \\ &\leq 2^p C(\delta)\xi^{1-p}\|g_*\|_{\mathcal{H}} \end{aligned} \tag{23}$$

with probability at least $1 - \delta$. In the second inequality, we used the fact that

$$\|g\|_{\mathcal{H}_M^+} = \inf\{\|g_1\|_{\mathcal{H}} + \|g_2\|_{\mathcal{H}_M} \mid g = g_1 + g_2, g_1 \in \mathcal{H}, g_2 \in \mathcal{H}_M\}.$$

Combining (22) and (23), we have

$$\begin{aligned} \|g_\lambda - g_{M,\lambda}\|_{L^\infty(\rho_\mathcal{X})} &\leq \|g_\lambda - \widetilde{g}_\lambda\|_{L^\infty(\rho_\mathcal{X})} + \|\widetilde{g}_\lambda - g_{M,\lambda}\|_{L^\infty(\rho_\mathcal{X})} \\ &\leq \max\left\{ 2^{p+1}C(\delta)\|g_*\|_{\mathcal{H}}\xi^{1-p}, R\left(\frac{2^4 \cdot 3L\xi}{\lambda}\|g_*\|_{\mathcal{H}}\right)^{1/2}, \right. \\ &\qquad \left. R\left(\frac{2^7 \cdot 3^2L^2\xi^2}{\lambda^3}\mathcal{L}(g_*)\right)^{1/4}, R\left(\frac{2^7 \cdot 3^2L^3\xi^3}{\lambda^3}\|g_*\|_{\mathcal{H}}\right)^{1/4} \right\}. \end{aligned}$$

As a result, define $\xi > 0$ which satisfies

$$\xi = \min\left\{ \left(\frac{\epsilon}{2^{p+1}C(\delta)\|g_*\|_{\mathcal{H}}}\right)^{1/1-p}, \frac{\lambda\epsilon^2}{2^4 \cdot 3R^2L\|g_*\|_{\mathcal{H}}}, \left(\frac{\lambda^3\epsilon^4}{2^7 \cdot 3^2R^4L^2\mathcal{L}(g_*)}\right)^{1/2}, \left(\frac{\lambda^3\epsilon^4}{2^7 \cdot 3^2R^4L^3\|g_*\|_{\mathcal{H}}}\right)^{1/3} \right\},$$

then we have $\|g_\lambda - g_{M,\lambda}\|_{L^\infty(\rho_\mathcal{X})} \leq \epsilon$ with probability at least $1 - 2\delta$. □

D Proof of Theorem 2

The following theorem shows that if k is a Gaussian kernel and k_M is its random Fourier features approximation, then the norm condition in the assumption is satisfied. The proof is inspired by the analysis of Theorem 4.48 in Steinwart and Christmann (2008).

Theorem 2. *Assume* $\mathrm{supp}(\rho_\mathcal{X}) \subset \mathbb{R}^d$ *is a bounded set and* $\rho_\mathcal{X}$ *has a density with respect to Lebesgue measure which is uniformly bounded away from 0 and* ∞ *on* $\mathrm{supp}(\rho_\mathcal{X})$*. Let* k *be a Gaussian kernel and* $\mathcal{H}$ *be its RKHS, then for any* $m \geq d/2$*, there exists a constant* $C_{m,d} > 0$ *such that*

$$\|f\|_{L^\infty(\rho_\mathcal{X})} \leq C_{m,d}\|f\|_{\mathcal{H}}^{d/2m}\|f\|_{L^2(\rho_\mathcal{X})}^{1-d/2m}$$

for any $f \in \mathcal{H}$*. Also, for any* $M \geq 1$*, let* k_M *be a random Fourier features approximation of* k *with* M *features and* $\mathcal{H}_M^+$ *be a RKHS of* $k + k_M$*. Then with probability at least* $1 - \delta$ *with respect to a sampling of features,*

$$\|f\|_{L^\infty(\rho_\mathcal{X})} \leq C_{m,d}\left(1 + \frac{1}{\delta}\right)^{d/4m}\|f\|_{\mathcal{H}_M^+}^{d/2m}\|f\|_{L^2(\rho_\mathcal{X})}^{1-d/2m}$$

for any $f \in \mathcal{H}_M^+$*.*

Proof. For notational simplicity, we denote $\mathrm{supp}(\rho_\mathcal{X})$ by $\mathcal{X}'$. From the boundedness of $\mathcal{X}'$ and the condition on $\rho_\mathcal{X}$, the following relation holds for any $f \in L^\infty(\rho_\mathcal{X})$:

$$\|f\|_{L^\infty(\rho_\mathcal{X})} = \|f\|_{L^\infty(\mathcal{X}')} \tag{24}$$

$$\|f\|_{L^2(\rho_\mathcal{X})} \geq C_1\|f\|_{L^2(\mathcal{X}')}, \tag{25}$$

where $C_1 > 0$ is a constant. From the discussion after Theorem 2, for any $f \in W^m(\mathcal{X}')$ $(m \geq d/2)$ there exists a constant $C_2 > 0$ such that the following inequality holds:

$$\|f\|_{L^\infty(\mathcal{X}')} \leq C_2 \|f\|^{d/2m}_{W^m(\mathcal{X}')} \|f\|^{1-d/2m}_{L^2(\mathcal{X}')}. \tag{26}$$

Here $W^m(\mathcal{X}')$ is Sobolev space with order m defined as follows:

$$W^m(\mathcal{X}') = \left\{ f \in L^2(\mathcal{X}') \;\middle|\; \partial^{(\alpha)} f \in L^2(\mathcal{X}') \text{ exists for all } \alpha \in \mathbb{N}^d \text{ with } |\alpha| \leq m \right\},$$

where $\partial^{(\alpha)}$ is the α-th weak derivative for a multi-index $\alpha = (\alpha^{(1)}, \ldots, \alpha^{(d)}) \in \mathbb{N}^d$ with $|\alpha| = \sum_{i=1}^d \alpha^{(i)}$. Combining (24), (25) and (26), we have

$$\|f\|_{L^\infty(\rho_\mathcal{X})} \leq C \|f\|^{d/2m}_{W^m(\mathcal{X}')} \|f\|^{1-d/2m}_{L^2(\rho_\mathcal{X})}, \tag{27}$$

where $C > 0$ is a constant. So it suffices to show that $\mathcal{H}$ and $\mathcal{H}^+_M$ can be continuously embedded in $W^m(\mathcal{X}')$. For $\mathcal{H}$, it can be shown in the same manner as Theorem 4.48 in Steinwart and Christmann (2008). For $\mathcal{H}^+_M$, we first define a spectral measure of the kernel function $k + k_M$ as

$$\tau^+(\omega) = \frac{1}{M} \sum_{i=1}^M \delta(\omega - \omega_i) + \tau(\omega),$$

where δ is a Dirac measure on Ω. Then a kernel function $k + k_M$ can be written as

$$(k + k_M)(x, x') = \int_\Omega \varphi(x, \omega) \overline{\varphi(x', \omega)} \mathrm{d}\tau^+(\omega),$$

and from Bach (2017b), for any $f \in \mathcal{H}^+_M$, there exists $g \in L^2(\tau^+)$ such that

$$\begin{aligned} f(x) &= \int_\Omega g(\omega) \varphi(x, \omega) \mathrm{d}\tau^+(\omega), \\ \|f\|_{\mathcal{H}^+_M} &= \|g\|_{L^2(\tau^+)}. \end{aligned}$$

Let us fix a multi-index $\alpha = (\alpha^{(1)}, \ldots, \alpha^{(d)}) \in \mathbb{N}^d$ and $|\alpha| = m$. For $\alpha \in \mathbb{N}^d$, we write $\partial^\alpha = \partial_1^{\alpha^{(1)}} \cdots \partial_d^{\alpha^{(d)}}$. We then have

$$\begin{aligned} \|\partial^\alpha f\|^2_{L^2(\mathcal{X}')} &= \int_{\mathcal{X}'} \left(\partial_x^\alpha \int_\Omega g(\omega) \varphi(x, \omega) \mathrm{d}\tau^+(\omega) \right)^2 \mathrm{d}x \\ &\leq \int_{\mathcal{X}'} \left(\int_\Omega |g(\omega)| \partial_x^\alpha \varphi(x, \omega) d\tau^+(\omega) \right)^2 \mathrm{d}x \\ &\leq \|g\|^2_{L^2(\tau^+)} \int_{\mathcal{X}'} \int_\Omega |\partial_x^\alpha \varphi(x, \omega)|^2 \mathrm{d}\tau^+(\omega) \mathrm{d}x. \end{aligned}$$

Because we consider φ as a random Fourier feature, $\Omega = \mathbb{R}^d$ and

$$\begin{aligned} \varphi(x, \omega) &= C' e^{i\omega^\top x}, \\ \partial_x^\alpha \varphi(x, \omega) &= \omega^\alpha C' e^{i\omega^\top x} \end{aligned}$$

where $C' > 0$ is a normalization constant and $\omega^\alpha = \prod_{i=1}^d {\omega^{(i)}}^{\alpha_i}$ for $\omega = (\omega^{(1)}, \ldots, \omega^{(d)}) \in \mathbb{R}^d$ and $\alpha = (\alpha^{(1)}, \ldots, \alpha^{(d)}) \in \mathbb{N}^d$. So we have

$$\begin{aligned} \|\partial^\alpha f\|^2_{L^2(\mathcal{X}')} &\leq \|g\|^2_{L^2(\tau^+)} \int_{\mathcal{X}'} C'^2 \int_\Omega \omega^{2\alpha} \mathrm{d}\tau^+(\omega) \mathrm{d}x \\ &\leq C'^2 \mathrm{vol}(\mathcal{X}') \|f\|^2_{\mathcal{H}^+_M} \left(\mathbb{E}_{\omega \sim \tau} \left[\omega^{2\alpha} \right] + \frac{1}{M} \sum_{i=1}^M \omega_i^{2\alpha} \right). \end{aligned}$$

We note that because τ is Gaussian, $\mathbb{E}_{\omega\sim\tau}\left[\omega^{2\alpha}\right]$ is finite for any $\alpha \in \mathbb{N}^d$. Because $\omega_i \sim \tau$ and $\omega_i^{2\alpha}$ is non negative, from Markov's inequality we have

$$\frac{1}{M}\sum_{i=1}^{M}\omega_i^{2\alpha} \leq \frac{1}{\delta}\mathbb{E}_{\omega\sim\tau}\left[\omega^{2\alpha}\right]$$

with probability at least $1-\delta$. As a result, we have

$$\|\partial^\alpha f\|^2_{L^2(\mathcal{X}')} \leq \left(1+\frac{1}{\delta}\right) C'^2 \mathrm{vol}(\mathcal{X}')\|f\|^2_{\mathcal{H}^+_M}\mathbb{E}_{\omega\sim\tau}\left[\omega^{2\alpha}\right].$$

So we can compute Sobolev norms of f as follows:

$$\begin{aligned}\|f\|^2_{W^m(\mathcal{X}')} &= \sum_{|\alpha|\leq m}\|\partial^\alpha f\|^2_{L^2(\mathcal{X}')} \\ &\leq \left(1+\frac{1}{\delta}\right) C'^2\mathrm{vol}(\mathcal{X}')\|f\|^2_{\mathcal{H}^+_M}\sum_{|\alpha|\leq m}\mathbb{E}_{\omega\sim\tau}\left[\omega^{2\alpha}\right]. \end{aligned} \tag{28}$$

Substitute (28) to (27) and define $C_{m,d} = C\left(C'^2\mathrm{vol}(\mathcal{X}')\sum_{|\alpha|\leq m}\mathbb{E}_{\omega\sim\tau}\left[\omega^{2\alpha}\right]\right)^{d/4m}$, we get a desired result. □

Remark. We note that the assumption that k is Gaussian is only used to derive $\mathbb{E}_{\omega\sim\tau}\left[\omega^{2\alpha}\right]$ is finite for all $\alpha \in \mathbb{N}^d$. This means that if $\psi(x-y) = k(x-y)$ belongs to Schwartz class (a space of rapidly decreasing function) Yoshida (1995), its Fourier transform τ also belongs to this class, thus the above finite moment property is satisfied.

E Proof of Theorem 3

In this section, we provide the complete statement and the proof of Theorem 3. First, we provide some useful propositions which are appeared in Nitanda and Suzuki (2019).

The first proposition suggests that there exists a sufficiently small $\lambda > 0$ such that g_λ is also the Bayes classifier.

Proposition 4 (Proposition A in Nitanda and Suzuki (2019))**.** *Suppose Assumption 3, 5, 6, 7 hold. Then, there exists $\lambda > 0$ such that $\|g_\lambda - g_*\|_{L^\infty(\rho_\mathcal{X})} \leq m(\delta)/2$.*

The second proposition shows that the distance between expected estimator $E[g_{T+1}]$ and the population risk minimizer $g_{M,\lambda}$ converges sub-linearly.

Proposition 5 (Modified version of Proposition C in Nitanda and Suzuki (2019))**.** *Suppose Assumption 2, 4 holds. Consider Algorithm 1 with $\eta_t = \frac{2}{\lambda(\gamma+t)}$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ and assume assume $\|\overline{g}_1\|_\mathcal{H} \leq (2\gamma_1 + 1/\lambda)GR$ and $\eta_1 \leq \min\{1/L, 1/2\lambda\}$. Then, it follows that*

$$\|\mathbb{E}[\overline{g}_{T+1}] - g_\lambda\|^2_\mathcal{H} \leq \frac{2}{\lambda}\left(\frac{18G^2R^2}{\lambda(2\gamma+T)} + \frac{\lambda\gamma(\gamma-1)}{2(2\gamma+T)(T+1)}\|\overline{g}_1 - g_\lambda\|^2_\mathcal{H}\right).$$

The last proposition is about the concentration of the estimator around its mean.

Proposition 6 (Modified version of Proposition 2 and D in Nitanda and Suzuki (2019))**.** *Suppose Assumption 1, 2 and 4 holds. Consider Algorithm 1 with $\eta_t = \frac{2}{\lambda(\gamma+t)}$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ and assume $\|\overline{g}_1\|_\mathcal{H} \leq (2\gamma_1 + 1/\lambda)GR$ and $\eta_1 \leq \min\{1/L, 1/2\lambda\}$. Then, it follows that*

$$\mathbb{P}\left[\left\|\overline{g}_{T+1} - \mathbb{E}[\overline{g}_{T+1}]\right\|_\mathcal{H} \geq \epsilon\right] \leq 2\exp\left(-\frac{\lambda^2(2\gamma+T)}{2^6\cdot 3^2G^2R^2}\epsilon^2\right).$$

Remark. We note that in Nitanda and Suzuki (2019), they assumed only the Lipschitz smoothness of $\mathcal{L}(g)$ with respect to $\|\cdot\|_\mathcal{H}$-norm, but they used the Lipschitz smoothness of $l(g,z)$ with respect to $\|\cdot\|_\mathcal{H}$-norm in the proof of Proposition B. Thus we deal with the Lipschitz smoothness of $l(\cdot, y)$ with respect to the first variable instead (Assumption 4) and correct these proofs.

Using these propositions, our main result about the exponential convergence of the expected classification error is shown as follows.

Theorem 3. *Suppose Assumptions 1-7 holds. There exists a sufficiently small $\lambda > 0$ such that the following statement holds:*
Taking the number of random features M that satisfies

$$M \geq \max\left\{\frac{8}{3}\left(\frac{R}{\xi}\right)^2, 32\left(\frac{R}{\xi}\right)^4\right\}\log\frac{2R^2}{\|T\|_{\mathrm{op}}\delta} \tag{29}$$

where $\xi > 0$ is defined as below:

$$\xi = \min\left\{\left(\frac{m(\delta)}{2^{p+3}C(\delta')\|g_*\|_{\mathcal{H}}}\right)^{1/1-p}, \frac{\lambda m^2(\delta)}{2^8\cdot 3R^2G\|g_*\|_{\mathcal{H}}}, \left(\frac{\lambda^3 m^4(\delta)}{2^{15}\cdot 3^2R^4G^2\mathcal{L}(g_*)}\right)^{1/2}, \left(\frac{\lambda^3 m^4(\delta)}{2^{15}\cdot 3^2R^4G^3\|g_*\|_{\mathcal{H}}}\right)^{1/3}\right\}.$$

Consider Algorithm 1 with $\eta_t = \frac{2}{\lambda(\gamma+t)}$ and $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ where γ is a positive value such that $\|g_1\|_{\mathcal{H}_M} \leq (2\eta_1 + 1/\lambda)GR$ and $\eta_1 \leq \min\{1/L, 1/2\lambda\}$. Then, with probability $1-2\delta'$, for sufficiently large T such that

$$\max\left\{\frac{36G^2R^2}{\lambda^2(2\gamma+T)}, \frac{\gamma(\gamma-1)\|g_1 - g_{M,\lambda}\|^2_{\mathcal{H}_M}}{(2\gamma+T)(T+1)}\right\} \leq \frac{m^2(\delta)}{64R^2},$$

we have the following inequality for any $t > T$:

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{T+1}) - \mathcal{R}(\mathbb{E}[Y|x])\right] \leq 2\exp\left(-\frac{\lambda^2(2\gamma+t)m^2(\delta)}{2^{12}\cdot 9G^2R^4}\right).$$

Proof. Fix $\lambda > 0$ satisfying the condition in Proposition 4. From Theorem 1, if we set a number of features M satisfying (29), we have

$$\begin{aligned}\|g_{M,\lambda} - g_*\|_{L^\infty(\rho_{\mathcal{X}})} &\leq \|g_{M,\lambda} - g_\lambda\|_{L^\infty(\rho_{\mathcal{X}})} + \|g_\lambda - g_*\|_{L^\infty(\rho_{\mathcal{X}})} \\ &\leq \frac{m(\delta)}{4} + \frac{m(\delta)}{2} = \frac{3m(\delta)}{4}.\end{aligned}$$

Then $\mathrm{sgn}(g(X)) = \mathrm{sgn}(g_*(X))$ almost surely for any $g \in \mathcal{H}_M$ satisfying $\|g - g_{M,\lambda}\|_{\mathcal{H}_M} \leq m(\delta)/4R$, since

$$\begin{aligned}\|g - g_*\|_{L^\infty(\rho_{\mathcal{X}})} &\leq \|g - g_{M,\lambda}\|_{L^\infty(\rho_{\mathcal{X}})} + \|g_{M,\lambda} - g_*\|_{L^\infty(\rho_{\mathcal{X}})} \\ &\leq R\|g - g_{M,\lambda}\|_{\mathcal{H}_M} + \|g_{M,\lambda} - g_*\|_{L^\infty(\rho_{\mathcal{X}})} \\ &\leq \frac{m(\delta)}{4} + \frac{3m(\delta)}{4} = m(\delta)\end{aligned}$$

and $|g_*(X)| \geq m(\delta)$ almost surely. In other words, g is also the Bayes classifier of $\mathcal{R}(g)$. Assume

$$\|\mathbb{E}[\overline{g}_{T+1}] - g_{M,\lambda}\|_{\mathcal{H}_M} \leq \frac{m(\delta)}{8R}. \tag{30}$$

Then, substituting $\epsilon = m(\delta)/8R$ in Proposition 6, we have

$$\left\|\overline{g}_{T+1} - g_{M,\lambda}\right\|_{\mathcal{H}_M} \leq \left\|\overline{g}_{T+1} - \mathbb{E}[\overline{g}_{T+1}]\right\|_{\mathcal{H}_M} + \|\mathbb{E}[\overline{g}_{T+1}] - g_{M,\lambda}\|_{\mathcal{H}_M} \leq \frac{m(\delta)}{4R}$$

with probability at least $1 - 2\exp\left(-\frac{\lambda^2(2\gamma+T)m^2(\delta)}{2^{12}\cdot 3^2G^2R^4}\right)$. In other words, $\overline{g}_{T+1}$ is also the Bayes classifier with same probability. By definition of the expected classification error, we have

$$\mathbb{E}[\mathcal{R}(\overline{g}_{T+1})] - \mathcal{R}(\mathbb{E}[Y|x]) \leq 1 - 2\exp\left(-\frac{\lambda^2(2\gamma+T)m^2(\delta)}{2^{12}\cdot 3^2G^2R^4}\right).$$

Finally, to satisfy (30), the required number of iteration T is obtained by Proposition 5, which completes the proof. □

F Proof of Corollary 1

Although g_λ converges to g_* as $\lambda \to 0$ as shown in Proposition 4, specifying its convergence rate is difficult in general. To derive its rate, first we need the *local strong convexity*, which is a strong convexity on a arbitrary compact set.

Assumption 8. *$\phi : \mathbb{R} \to \mathbb{R}$ is $\mu(U)$-strongly convex on a bounded set $[-U, U] \subset \mathbb{R}$, i.e.,*

$$\phi(\zeta_1) - \phi(\zeta_2) - \phi'(\zeta_2)(\zeta_1 - \zeta_2) \geq \frac{\mu(U)}{2}(\zeta_1 - \zeta_2)^2.$$

holds for any $\zeta_1, \zeta_2 \in [-U, U]$.

Lemma 3. *Assume $\mathrm{supp}(\rho_\mathcal{X}) \subset \mathbb{R}^d$ is a bounded set and $\rho_\mathcal{X}$ has a density with respect to Lebesgue measure, which is uniformly bounded away from 0 and ∞ on $\mathrm{supp}(\rho_\mathcal{X})$. Let k be a Gaussian kernel and l satisfies Assumption 8. Then for arbitrary small $\epsilon > 0$, there exists a constant $C > 0$ such that*

$$\|g_\lambda - g_*\|_{L^\infty(\rho_\mathcal{X})} \leq C\|g_*\|_\mathcal{H} \left(\frac{\lambda}{\mu(R\|g_*\|_\mathcal{H})}\right)^{\frac{1}{2}-\epsilon}.$$

Proof. By definition of g_λ, we have

$$\mathcal{L}(g_*) + \frac{\lambda}{2}\|g_*\|_\mathcal{H}^2 \geq \mathcal{L}(g_\lambda) + \frac{\lambda}{2}\|g_\lambda\|_\mathcal{H}^2, \tag{31}$$

$$\|g_*\|_\mathcal{H} \geq \|g_\lambda\|_\mathcal{H}. \tag{32}$$

In addition, it holds that

$$\begin{aligned} g_*(x) &\leq R\|g_*\|_\mathcal{H}, \\ g_\lambda(x) &\leq R\|g_\lambda\|_\mathcal{H} \leq R\|g_*\|_\mathcal{H} \end{aligned} \tag{33}$$

for all $x \in \mathcal{X}$. Furthermore, since g_* attains infimum of $\mathcal{L}$ among all measurable functions, we have

$$\int_\mathcal{Y} \partial_\zeta l(g_*(\cdot), y)\mathrm{d}\rho(y|\cdot) \equiv 0, \tag{34}$$

where ∂_ζ denotes a partial derivative of l with respect to the first variable.
Then we obtain

$$\begin{aligned} \|g_\lambda - g_*\|_{L^2(\rho_\mathcal{X})}^2 &= \int_\mathcal{X} |g_\lambda(x) - g_*(x)|^2 \,\mathrm{d}\rho_\mathcal{X}(x) \\ &\leq \int_{\mathcal{X}\times\mathcal{Y}} \frac{2}{\mu(R\|g_*\|_\mathcal{H})} \{l(g_\lambda(x), y) - l(g_*(x), y) \\ &\qquad - \partial_\zeta l(g_*(x), y)(g_\lambda(x) - g_*(x))\}\mathrm{d}\rho(x, y) \quad (\because (33) \text{ and Assumption 8}) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \frac{2}{\mu(R\|g_*\|_\mathcal{H})} \{l(g_\lambda(x), y) - l(g_*(x), y)\} \,\mathrm{d}\rho(x, y) \quad (\because (34)) \\ &= \frac{2}{\mu(R\|g_*\|_\mathcal{H})} \left(\mathcal{L}(g_\lambda) - \mathcal{L}(g_*)\right) \\ &\leq \frac{\lambda}{\mu(R\|g_*\|_\mathcal{H})} \left(\|g_*\|_\mathcal{H}^2 - \|g_\lambda\|_\mathcal{H}^2\right) \quad (\because (31)) \\ &\leq \frac{\lambda}{\mu(R\|g_*\|_\mathcal{H})} \|g_*\|_\mathcal{H}^2 \quad (\because (32)). \end{aligned}$$

Finally, applying the first part of Theorem 2 with $p = d/2m$, we obtain

$$\begin{aligned} \|g_\lambda - g_*\|_{L^\infty(\rho_\mathcal{X})} &\leq C_p \|g_\lambda - g_*\|_\mathcal{H}^p \|g_\lambda - g_*\|_{L^2(\rho_\mathcal{X})}^{1-p} \\ &\leq 2^p C_p \left(\frac{\lambda}{\mu(R\|g_*\|_\mathcal{H})}\right)^{\frac{1-p}{2}} \|g_*\|_\mathcal{H} \end{aligned}$$

for any $0 < p < 1$ and get a desired result. □

Corollary 1. *Assume* $\mathrm{supp}(\rho_{\mathcal{X}}) \subset \mathbb{R}^d$ *is a bounded set and* $\rho_{\mathcal{X}}$ *has a density with respect to Lebesgue measure, which is uniformly bounded away from 0 and* ∞ *on* $\mathrm{supp}(\rho_{\mathcal{X}})$*. Let* k *be a Gaussian kernel and* l *be logistic loss. Under Assumption 5-7, the following statement holds:*
Taking a regularization parameter λ *and a number of random features* M *that satisfies*

$$\lambda \lesssim \log^3 \frac{1+2\delta}{1-2\delta} \cdot \frac{1}{(2 + e^{R\|g_*\|_{\mathcal{H}}} + e^{-R\|g_*\|_{\mathcal{H}}})\|g_*\|_{\mathcal{H}}^3},$$
$$M \gtrsim \left(\frac{\left(1 + \frac{1}{\delta'}\right) \|g_*\|_{\mathcal{H}}^4}{\lambda^3 \log^4 \frac{1+2\delta}{1-2\delta}} \right)^2 \log \frac{1}{\delta'}.$$

Consider Algorithm 1 with $\eta_t = \frac{2}{\lambda(\gamma+t)}$ *and* $\alpha_t = \frac{2(\gamma+t-1)}{(2\gamma+T)(T+1)}$ *where* γ *is a positive value such that* $\|g_1\|_{\mathcal{H}_M} \leq (2\eta_1 + 1/\lambda)$ *and* $\eta_1 \leq \min\{4, 1/2\lambda\}$*. Then, with probability* $1 - 2\delta'$*, for a sufficiently large* T *such that*

$$\max\left\{ \frac{36}{\lambda^2(2\gamma+T)}, \frac{\gamma(\gamma-1)\|g_1 - g_{M,\lambda}\|_{\mathcal{H}_M}^2}{(2\gamma+T)(T+1)} \right\} \leq \frac{\log^2 \frac{1+2\delta}{1-2\delta}}{64},$$

we have the following inequality for any $t \geq T$*:*

$$\mathbb{E}\left[\mathcal{R}(\overline{g}_{t+1}) - \mathcal{R}(\mathbb{E}[Y|x])\right] \leq 2\exp\left(-\frac{\lambda^2(2\gamma+t)}{2^{12} \cdot 9} \log^2 \frac{1+2\delta}{1-2\delta} \right).$$

Proof. When l is logistic loss, we have $\phi(v) = \log(1 + \exp(-v))$ and $\phi''(v) = \frac{1}{2+e^v+e^{-v}}$. Thus it follows that Assumption 8 is satisfied with $\mu(U) = \frac{2}{1+e^{-U}+e^{U}}$. To satisfy the condition

$$\|g_\lambda - g_*\|_{L^\infty(\rho_{\mathcal{X}})} \leq m(\delta)/2,$$

required λ is easily derived from Lemma 3 with, for example, $\epsilon = 1/6$. In addition, since $\phi''(v) \leq 1/4$ and $\phi'(v) \leq 1$ for any $v \in \mathbb{R}$, Assumption 4 and Assumption 1 are satisfied with $L = 1/4$ and $G = 1$, respectively. Substituting them and $m(\delta) = \log((1+2\delta)/(1-2\delta))$, $R = 1$ in Theorem 3, we get a desired result. □