Appendix

A. Proof of Theorem 3: Convergence of SALSA

Our analysis here is a brute force generalisation of the analysis in Zhang et al. (2013). We handle the additive case using ideas from Aronszajn (1950). As such we will try and stick to the same notation. Some intermediate technical results can be obtained directly from Zhang et al. (2013) but we repeat them (or provide an outline) here for the sake of completeness.

In addition to the definitions presented in the main text, we will also need the following quantities,

$$\beta_t^{(j)} = \sum_{\ell=t+1}^{\infty} \mu_\ell^{(j)}, \qquad \Psi^{(j)} = \sum_{\ell=1}^{\infty} \mu_\ell^{(j)}, \qquad b(n,t,q) = \max\left(\sqrt{\max(q, \log t)}\,,\, \frac{\max(q, \log t)}{n^{1/2-1/q}}\right).$$

Here $\Psi^{(j)}$ is the trace of $k^{(j)}$. $\beta_t^{(j)}$ depends on some $t \in \mathbb{N}$ which we will pick later. Also define $\beta_t = \sum_j \beta_t^{(j)}$ and $\Psi = \sum_j \Psi^{(j)}$.

Note that the excess risk can be decomposed into bias and variance terms, $\mathcal{R}(\hat{f}) - \mathcal{R}(f_*) = \mathbb{E}[\|\hat{f} - f_*\|_2^2] = \|f_* - \mathbb{E}\hat{f}\|_2^2 + \mathbb{E}[\|\hat{f} - \mathbb{E}\hat{f}\|_2^2]$. In Sections A.2 and A.3 respectively, we will prove the following bounds which will yield in Theorem 3:

$$\|f_* - \mathbb{E}\hat{f}\|_2^2 \le M_d\left(8\lambda\|\mathbf{f}_*\|_{\mathcal{F}}^2 + \frac{8{M_d}^{3/2}\rho^4\|\mathbf{f}_*\|_{\mathcal{F}}^2}{\lambda}\Psi\beta_t + \|\mathbf{f}_*\|_{\mathcal{F}}^2 \sum_{j=1}^{M_d} \mu_{t+1}^{(j)} + \left(\frac{CM_d b(n,t,q)\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q \|f_*\|_2^2\right), \quad (9)$$

$$\mathbb{E}[\|\hat{f} - \mathbb{E}\hat{f}\|_2^2] \le M_d\left(12\lambda\|\mathbf{f}_*\|_{\mathcal{F}}^2 + \frac{12\sigma^2\gamma_k(\lambda)}{n} + \left(\frac{2\sigma^2}{\lambda} + 4\|\mathbf{f}_*\|_{\mathcal{F}}^2\right)\left(\sum_{j=1}^{M_d}\mu_{t+1}^{(j)} + \frac{12M_d\rho^4}{\lambda}\Psi\beta_t + \left(\frac{CM_d b(n,t,q)\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q\right)\|f_*\|_2^2\right). \quad (10)$$

Accordingly, this gives the following expression for $\chi(k)$,

$$\chi(k) = \inf_t \left[\frac{8{M_d}^{3/2}\rho^4\|\mathbf{f}_*\|_{\mathcal{F}}^2}{\lambda}\Psi\beta_t + \left(\frac{2\sigma^2}{\lambda} + 4\|\mathbf{f}_*\|_{\mathcal{F}}^2 + 1\right)\left(\frac{CM_d b(n,t,q)\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q \|f_*\|_2^2\right) + \left(\frac{2\sigma^2}{\lambda} + 4\|\mathbf{f}_*\|_{\mathcal{F}}^2\right)\left(\sum_{j=1}^{M_d}\mu_{t+1}^{(j)} + \frac{12M_d\rho^4}{\lambda}\Psi\beta_t\right) + \|\mathbf{f}_*\|_{\mathcal{F}}^2 \sum_{j=1}^{M_d}\mu_{t+1}^{(j)}\right]. \quad (11)$$

Note that the second term in $\chi(k)$ is usually low order for large enough q due to the $n^{-q/2}$ term. Therefore if in our setting $\beta_t^{(j)}$ and $\mu_{t+1}^{(j)}$ are small enough, $\chi(k)$ is low order. We show this for the two kernel choices of Theorem 4 in Appendix B.

First, we review some well known results on RKHS's which we will use in our analysis. Let κ be a PSD kernel and $\mathcal{H}_\kappa$ be its RKHS. Then κ acts as the representer of evaluation – i.e. for any $f \in \mathcal{H}_\kappa$, $\langle f, \kappa(\cdot, x)\rangle_{\mathcal{H}_\kappa} = f(x)$. Denote the RKHS norm $\|f\|_{\mathcal{H}_\kappa} = \sqrt{\langle f, f\rangle_{\mathcal{H}_\kappa}}$ and the L^2 norm $\|f\|_2 = \sqrt{\int f^2}$.

Let the kernel κ have an eigenexpansion $\kappa(x,x') = \sum_{\ell=1}^{\infty}\mu_\ell\phi_\ell(x)\phi_\ell(x')$. Denote the basis coefficients of f in $\{\phi_\ell\}$ via $\{\theta_\ell\}$. That is, $\theta_\ell = \int f\cdot\phi_\ell\,\mathrm{d}\mathbb{P}$ and $f = \sum_{\ell=1}^{\infty}\theta_\ell\phi_\ell$. The following results are well known (Schölkopf & Smola, 2001; Steinwart & Christmann, 2008),

$$\langle\phi_\ell, \phi_\ell\rangle = 1/\mu_\ell, \qquad \|f\|_2^2 = \sum_{\ell=1}^{\infty}\theta_\ell^2, \qquad \|f\|_{\mathcal{H}_\kappa}^2 = \sum_{\ell=1}^{\infty}\frac{\theta_\ell^2}{\mu_\ell}.$$

Before we proceed, we make the following remark on the minimiser of (3).

Remark 6. *The solution of* (3) *takes the form* $\hat{f}(\cdot) = \sum_{i=1}^{n}\alpha_i k(\cdot, X_i)$ *where* k *is the sum kernel* (4).

Proof. The key observation is that we only need to consider n (and not nM_d) parameters even though we are optimising over M_d RKHSs. The reasoning uses a powerful result from Aronszajn (1950). Consider the class of functions $\mathcal{H}' = \{f = \sum_j f^{(j)}; f^{(j)} \in \mathcal{H}_{k^{(j)}}\}$. In (3) we are minimising over $\mathcal{H}'$. Any $f \in \mathcal{H}'$ need *not* have a unique additive decomposition. Consider $\mathcal{H} \subset \mathcal{H}'$ which only contains the minimisers in the expression below.

$$\|f\|_{\mathcal{H}}^2 = \inf_{g^{(j)} \in \mathcal{H}_{k^{(j)}}; f = \sum g^{(j)}} \sum_{j=1}^{M} \|g^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2$$

Aronszajn (1950) showed that $\mathcal{H}$ is an RKHS with the sum kernel $k = \sum_j k^{(j)}$ and its RKHS norm is $\|\cdot\|_{\mathcal{H}}$. Clearly, the minimiser of (3) lies in $\mathcal{H}$. For any $g' \in \mathcal{H}'$, we can pick a corresponding $g \in \mathcal{H}$ with the same sum of squared errors (as $g = g'$) but lower complexity penalty (as g minimises the sum of norms for any $g' = g$). Therefore, we may optimise (3) just over $\mathcal{H}$ and not $\mathcal{H}'$. An application of Mercer's theorem concludes the proof. □

A.1. Set up

We first define the following function class of the product of all RKHS's, $\mathcal{F} = \mathcal{H}_{k^{(1)}} \times \mathcal{H}_{k^{(2)}} \times \cdots \times \mathcal{H}_{k^{(M_d)}} = \left\{\mathbf{f} = (f^{(1)}, \ldots, f^{(M_d)}) \middle| f^{(j)} \in \mathcal{H}_{k^{(j)}}\ \forall j\right\}$ and equip it with the inner product $\langle \mathbf{f}_1, \mathbf{f}_2 \rangle = \langle f_1^{(1)}, f_2^{(1)} \rangle_{\mathcal{H}_{k^{(1)}}} + \cdots + \langle f_1^{(M_d)}, f_2^{(M_d)} \rangle_{\mathcal{H}_{k^{(M_d)}}}$. Here, $f_1^{(j)}$ are the elements of $\mathbf{f}_1$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_{k^{(j)}}}$ is the RKHS inner product of $\mathcal{H}_{k^{(j)}}$. Therefore the norm is $\|\mathbf{f}\|_{\mathcal{F}}^2 = \sum_{j=1}^{M_d} \|f^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2$. Denote $\xi_x^{(j)} = k^{(j)}(x, \cdot)$ and $\xi_x(\cdot) = \mathcal{K}(\cdot, x)$. Observe that for an additive function $f = \sum_j f^{(j)}(x)$,

$$f(x) = \sum_j f^{(j)}(x) = \sum_j \langle f^{(j)}, \xi_x^{(j)} \rangle_{\mathcal{H}_{k^{(j)}}} = \langle \mathbf{f}, \xi_x \rangle.$$

Recall that the solution to (3) is denoted by $\hat{f}$ and the individual functions of the solution are given by $\hat{f}^{(j)}$. We will also use $\mathbf{f}_*$ and $\hat{\mathbf{f}}$ to denote the representations of f_* and $\hat{f}$ in $\mathcal{F}$, i.e., $\mathbf{f}_* = (f_*^{(1)}, \ldots, f_*^{(M_d)})$ and $\hat{\mathbf{f}} = (\hat{f}^{(1)}, \ldots, \hat{f}^{(M_d)})$. Note that $\|\mathbf{f}_*\|_{\mathcal{F}}^2$ is precisely the bound used in Theorem 3. We will also denote $\Delta^{(j)} = \hat{f}^{(j)} - f_*^{(j)} \in \mathcal{H}_{k^{(j)}}$, $\mathbf{\Delta} = (\Delta^{(1)}, \ldots, \Delta^{(M_d)}) \in \mathcal{F}$, and $\Delta = \sum_j \Delta^{(j)} = \hat{f} - f_*$.

For brevity, from now on we will write $k^{(j)}(x, x')$ instead of $k^{(j)}(x^{(j)}, x^{(j)'})$. Further, since d is fixed in this analysis we will write M for M_d.

A.2. Bias (Proof of Bound (9))

Note that we need to bound $\|\mathbb{E}[\Delta]\|_2$ which by Jensen's inequality is less than $\mathbb{E}[\|\mathbb{E}[\Delta|X_1^n]\|_2]$. Since, $\|\mathbb{E}[\Delta|X_1^n]\|_2^2 \leq M \sum_{j=1}^M \|\mathbb{E}[\Delta^{(j)}|X_1^n]\|_2^2$, we will focus on bounding $\sum_{j=1}^M \|\mathbb{E}[\Delta^{(j)}|X_1^n]\|_2^2$.

We can write the optimisation objective (3) as follows,

$$\hat{\mathbf{f}} = \operatorname*{argmin}_{\mathbf{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(\langle \mathbf{f}, \xi_{X_i} \rangle - Y_i\right)^2 + \lambda \|\mathbf{f}\|_{\mathcal{F}}^2 \tag{12}$$

Since this is Fréchet differentiable in $\mathcal{F}$ in the metric induced by the inner product defined above, the first order optimality conditions for $\hat{f}^{(j)}$ give us,

$$\frac{1}{n} \sum_{i=1}^{n} \left(\langle \xi_{X_i}, \hat{\mathbf{f}} - \mathbf{f}_* \rangle - \epsilon_i\right) \xi_{X_i}^{(j)} + 2\lambda \hat{f}^{(j)} = \mathbf{0}.$$

Here, we have taken $Y_i = f_*(X_i) + \epsilon_i$ where $\mathbb{E}[\epsilon_i | X_i] = 0$. Doing this for all $\hat{f}^{(j)}$ we have,

$$\frac{1}{n} \sum_{i=1}^{n} \xi_{X_i} \left(\langle \xi_{X_i}, \mathbf{\Delta} \rangle - \epsilon_i\right) + \lambda \hat{\mathbf{f}} = \mathbf{0} \tag{13}$$

Taking expectations conditioned on X_1^n and rearranging we get,

$$(\widehat{\Sigma} + \lambda I) \mathbb{E}[\mathbf{\Delta} | X_1^n] = -\lambda \mathbf{f}_*, \tag{14}$$

where $\widehat{\Sigma} = \frac{1}{n}\sum_i \xi_{X_i} \otimes \xi_{X_i}$ is the empirical covariance. Since $\widehat{\Sigma} \succeq \mathbf{0}$,

$$\forall j', \quad \|\mathbb{E}[\Delta^{(j')}|X_1^n]\|^2_{\mathcal{H}_{k^{(j')}}} \leq \sum_{j=1}^{M} \|\mathbb{E}[\Delta^{(j)}|X_1^n]\|^2_{\mathcal{H}_{k^{(j)}}} = \|\mathbb{E}[\boldsymbol{\Delta}|X_1^n]\|^2_{\mathcal{F}} \leq \|\mathbf{f}_*\|^2_{\mathcal{F}} \tag{15}$$

Let $\mathbb{E}[\Delta^{(j)}|X_1^n] = \sum_{\ell=1}^{\infty} \delta_\ell^{(j)} \phi_\ell^{(j)}$ where $\phi_\ell^{(j)}$ are the eigenfunctions in the expansion of $k^{(j)}$. Denote $\delta_\downarrow^{(j)} = (\delta_1^{(j)}, \ldots, \delta_t^{(j)})$ and $\delta_\uparrow^{(j)} = (\delta_{t+1}^{(j)}, \delta_{t+2}^{(j)}, \ldots)$. We will set t later. Since $\|\mathbb{E}[\Delta^{(j)}|X_1^n]\|_2^2 = \|\delta_\downarrow^{(j)}\|_2^2 + \|\delta_\uparrow^{(j)}\|_2^2$ we will bound the two terms. The latter term is straightforward,

$$\|\delta_\uparrow^{(j)}\|_2^2 \leq \mu_{t+1}^{(j)} \sum_{\ell=t+1}^{\infty} \frac{{\delta_\ell^{(j)}}^2}{\mu_\ell^{(j)}} \leq \mu_{t+1}^{(j)} \|\mathbb{E}[\Delta^{(j)}|X_1^n]\|^2_{\mathcal{H}_{k^{(j)}}} \leq \mu_{t+1}^{(j)} \|\mathbf{f}_*\|^2_{\mathcal{F}} \tag{16}$$

To control $\|\delta_\downarrow^{(j)}\|$, let $f_*^{(j)} = \sum_\ell \theta_\ell^{(j)} \phi_\ell^{(j)}$. Also, define the following: $\theta_\downarrow^{(j)} = (\theta_1^{(j)}, \ldots, \theta_t^{(j)})$, $\Phi^{(j)} \in \mathbb{R}^{n \times t}$, $\Phi_{i\ell}^{(j)} = \phi_\ell^{(j)}(X_i)$, $\Phi_\ell^{(j)} = (\phi_\ell^{(j)}(X_1), \ldots, \phi_\ell^{(j)}(X_n)) \in \mathbb{R}^n$, $\mathcal{M}^{(j)} = \mathrm{diag}(\mu_1^{(j)}, \ldots, \mu_t^{(j)}) \in \mathbb{R}_+^{t \times t}$ and $v^{(j)} \in \mathbb{R}^n$ where $v_i^{(j)} = \sum_{\ell > t} \delta_\ell^{(j)} \phi_\ell^{(j)}(X_i) = \mathbb{E}[\Delta_\uparrow^{(j)}(X_i)|X_1^n]$.

Further define, $\Phi = [\Phi^{(1)} \ldots \Phi^{(M)}] \in \mathbb{R}^{n \times tM}$, $\mathcal{M} = \mathrm{diag}(\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(M)}) \in \mathbb{R}^{tM \times tM}$, $v_i = \sum_j v^{(j)}$, $\delta_\downarrow = [\delta_\downarrow^{(1)}; \ldots; \delta_\downarrow^{(M)}] \in \mathbb{R}^{tM}$ and $\theta_\downarrow = [\theta_\downarrow^{(1)}; \ldots; \theta_\downarrow^{(M)}] \in \mathbb{R}^{tM}$.

Now compute the $\mathcal{F}$-inner product between $(\mathbf{0}, \ldots, \phi_\ell^{(j)}, \ldots, \mathbf{0})$ with equation (14) to obtain,

$$\frac{1}{n}\sum_{i=1}^{n} \langle \phi_\ell^{(j)}, \xi_{X_i}^{(j)} \rangle_{\mathcal{H}_{k^{(j)}}} \langle \xi_{X_i}, \mathbb{E}[\boldsymbol{\Delta}|X_1^n] \rangle + \lambda \langle \phi_\ell^{(j)}, \mathbb{E}[\Delta^{(j)}|X_1^n] \rangle_{\mathcal{H}_{k^{(j)}}} = -\lambda \langle \phi_\ell^{(j)}, f_*^{(j)} \rangle_{\mathcal{H}_{k^{(j)}}}$$

$$\frac{1}{n}\sum_{i=1}^{n} \phi_\ell^{(j)}(X_i) \sum_{j=1}^{M} \left(\sum_{\ell' \leq t} \phi_{\ell'}^{(j)}(X_i) \delta_{\ell'}^{(j)} + \sum_{\ell' > t} \phi_{\ell'}^{(j)}(X_i) \delta_{\ell'}^{(j)} \right) + \lambda \frac{\delta_\ell^{(j)}}{\mu_\ell^{(j)}} = -\lambda \frac{\theta_\ell^{(j)}}{\mu_\ell^{(j)}}$$

After repeating this for all j and for all $\ell = 1, \ldots, t$, and arranging the terms appropriately this reduces to

$$\left(\frac{1}{n} \Phi^\top \Phi + \lambda \mathcal{M}^{-1} \right) \delta_\downarrow = -\lambda \mathcal{M}^{-1} \theta_\downarrow - \frac{1}{n} \Phi^\top v$$

By writing $Q = (I + \lambda \mathcal{M}^{-1})^{1/2}$, we can rewrite the above expression as

$$\left(I + Q^{-1} \left(\frac{1}{n} \Phi^\top \Phi - I \right) Q^{-1} \right) Q \delta_\downarrow = -\lambda Q^{-1} \mathcal{M}^{-1} \theta_\downarrow - \frac{1}{n} Q^{-1} \Phi^\top v.$$

We will need the following technical lemmas. The proofs are given at the end of this section. These results correspond to Lemma 5 in Zhang et al. (2013).

Lemma 7. $\|\lambda Q^{-1} \mathcal{M}^{-1} \theta_\downarrow\|_2^2 \leq \lambda \|\mathbf{f}_*\|^2_{\mathcal{F}}$.

Lemma 8. $\mathbb{E}\left[\|\frac{1}{n} Q^{-1} \Phi^\top v\|_2^2 \right] \leq \frac{1}{\lambda} M^{3/2} \rho^4 \|\mathbf{f}_*\|^2_{\mathcal{F}} \Psi \beta_t$.

Lemma 9. *Define the event* $\mathcal{E} = \{\|Q^{-1}(\frac{1}{n}\Phi^\top \Phi - I) Q^{-1}\|_{op} \leq 1/2\}$. *Then, there exists a constant* C *s.t.*

$$\mathbb{P}(\mathcal{E}^c) \leq \left(\max\left(\sqrt{\max(q, \log t)}\,,\ \frac{\max(q, \log t)}{n^{1/2 - 1/q}} \right) \times \frac{MC\rho^2 \gamma_k(\lambda)}{\sqrt{n}} \right)^q.$$

When $\mathcal{E}$ holds, by Lemma 9 and noting that $Q \succeq I$,

$$\|\delta_\downarrow\|_2^2 \le \|Q\delta_\downarrow\|_2^2 = \left\| \left(I + Q^{-1}\left(\frac{1}{n}\Phi^\top\Phi - I\right)Q^{-1}\right)^{-1} \left(-\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow - \frac{1}{n}Q^{-1}\Phi^\top v\right)\right\|^2$$
$$\le 4\|\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow + \frac{1}{n}Q^{-1}\Phi^\top v\|_2^2. \le 8\|\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow\|_2^2 + 8\|\frac{1}{n}Q^{-1}\Phi^\top v\|_2^2$$

Now using Lemmas 7 and 8,

$$\mathbb{E}[\|\delta_\downarrow\|_2^2|\mathcal{E}] \le 8\left(\lambda\|\mathbf{f}_*\|_{\mathcal{F}}^2 + \frac{M^{3/2}\rho^4\|\mathbf{f}_*\|_{\mathcal{F}}^2\Psi\beta_t}{\lambda}\right)$$

Since $\mathbb{E}[\|\delta_\downarrow\|_2^2] = \mathbb{P}(\mathcal{E})\mathbb{E}[\|\delta_\downarrow\|_2^2|\mathcal{E}] + \mathbb{P}(\mathcal{E}^c)\mathbb{E}[\|\delta_\downarrow\|_2^2|\mathcal{E}^c]$ and by using the fact that $\|\delta_\downarrow\|^2 \le \|\mathbb{E}[\Delta|X_1^n]\|_2^2 \le \|f_*\|_2^2$, we have

$$\mathbb{E}[\|\delta_\downarrow\|_2^2] \le 8\lambda\|\mathbf{f}_*\|_{\mathcal{F}}^2 + \frac{8M\rho^4\|\mathbf{f}_*\|_{\mathcal{F}}^2\Psi\beta_t}{\lambda} +$$
$$\left(\max\left(\sqrt{\max(q,\log t)}\,,\ \frac{\max(q,\log t)}{n^{1/2-1/q}}\right) \times \frac{MC\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q \|f_*\|_2^2$$

Finally using (16) and by noting that

$$\|\mathbb{E}[\Delta|X_1^n]\|_2^2 \ \le M\sum_{j=1}^M \|\mathbb{E}[\Delta^{(j)}|X_1^n]\|_2^2 \ = M\big(\|\delta_\downarrow\|_2^2 + \sum_j \|\delta_\uparrow^{(j)}\|_2^2\big) \ \le M\big(\|\delta_\downarrow\|_2^2 + \|\mathbf{f}_*\|_{\mathcal{F}}^2\sum_j \mu_{t+1}^{(j)}\big)$$

and then taking expectation over X_1^n, we obtain the bound for the bias in (9).

Proofs of Technical Lemmas

A.2.1. Proof of Lemma 7

Lemma 7 is straightforward.

$$\|Q^{-1}\mathcal{M}^{-1}\theta_\downarrow\|_2^2 = \sum_{j=1}^M \|{Q^{(j)}}^{-1}{\mathcal{M}^{(j)}}^{-1}\theta_\downarrow^{(j)}\|_2^2 = \sum_{j=1}^M {\theta_\downarrow^{(j)}}^\top({\mathcal{M}^{(j)}}^2 + \lambda\mathcal{M}^{(j)})^{-1}\theta_\downarrow^{(j)}$$
$$\le \sum_{j=1}^M {\theta_\downarrow^{(j)}}^\top(\lambda\mathcal{M}^{(j)})^{-1}\theta_\downarrow^{(j)} = \frac{1}{\lambda}\sum_{j=1}^M\sum_{\ell=1}^t \frac{{\theta_\ell^{(j)}}^2}{\mu_\ell^{(j)}} \le \frac{1}{\lambda}\|\mathbf{f}_*\|_{\mathcal{F}}^2$$

A.2.2. Proof of Lemma 8

We first decompose the LHS as follows,

$$\left\|\frac{1}{n}Q^{-1}\Phi^\top v\right\|_2^2 = \left\|(M+\lambda I)^{-1/2}\left(\frac{1}{n}M^{1/2}\Phi^\top v\right)\right\|_2^2 \le \frac{1}{\lambda}\left\|\frac{1}{n}M^{1/2}\Phi^\top v\right\|_2^2 \tag{17}$$

The last step follows by noting that $\|(M+\lambda I)^{-1/2}\|_{op}^2 = \max_{j,\ell} 1/(\mu_\ell^{(j)}+\lambda) \le 1/\lambda$. Further,

$$\mathbb{E}\left[\|M^{1/2}\Phi^\top v\|_2^2\right] = \sum_{j=1}^M\sum_{\ell=1}^t \mu_\ell^{(j)}\mathbb{E}[({\Phi_\ell^{(j)}}^\top v)^2] \le \sum_{j=1}^M\sum_{\ell=1}^t \mu_\ell^{(j)}\mathbb{E}[\|\Phi_\ell^{(j)}\|_2^2\|v\|_2^2] \tag{18}$$

Note that the term inside the summation in the RHS can be bounded by, $\sqrt{\mathbb{E}[\|\Phi_\ell^{(j)}\|_2^4]\mathbb{E}[\|v\|_2^4]}$. We bound the first expectation via,

$$\mathbb{E}\left[\|\Phi_\ell^{(j)}\|^4\right] = \mathbb{E}\left[\left(\sum_{i=1}^n \phi_\ell^{(j)}(X_i)^2\right)^2\right] \le \mathbb{E}\left[n\sum_{i=1}^n \phi_\ell^{(j)}(X_i)^4\right] \le n^2\rho^4$$

where the last step follows from Assumption 2. For the second expectation we first bound $\|v\|^4$,

$$\|v\|_2^4 = \left(\sum_{i=1}^n \left(\sum_{j=1}^M v_i^{(j)}\right)^2\right)^2 \leq \left(M\sum_{i=1}^n\sum_{j=1}^M {v_i^{(j)}}^2\right)^2 \leq M^3 n \sum_{i=1}^n\sum_{j=1}^M {v_i^{(j)}}^4$$

Now by the Cauchy Schwarz inequality,

$${v_i^{(j)}}^2 = \left(\sum_{\ell>t} \delta_\ell^{(j)}\phi_\ell^{(j)}(X_i)\right)^2 \leq \left(\sum_{\ell>t}\frac{{\delta_\ell^{(j)}}^2}{\mu_\ell^{(j)}}\right)\left(\sum_{\ell>t}\mu_\ell^{(j)}\phi_\ell^{(j)}(X_i)^2\right).$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\|v\|^4\right] &\leq M^3 n\sum_{i=1}^n\sum_{j=1}^M \mathbb{E}\left[\|\mathbb{E}[\Delta^{(j)}|X_1^n]\|_{\mathcal{H}_{k^{(j)}}}^4\left(\sum_{\ell>t}\mu_\ell^{(j)}\phi_\ell^{(j)}(X_i)^2\right)^2\right] \\
&\leq M^3 n\|\mathbf{f}_*\|_{\mathcal{F}}^4 \sum_{j=1}^M\sum_{i=1}^n\sum_{\ell,\ell'>t}\mathbb{E}[\mu_\ell^{(j)}\mu_{\ell'}^{(j)}\phi_\ell^{(j)}(X_i)^2\phi_{\ell'}^{(j)}(X_i)^2] \\
&\leq M^3 n\rho^4\|\mathbf{f}_*\|_{\mathcal{F}}^4 \sum_{j=1}^M\sum_{i=1}^n\left(\sum_{\ell>t}\mu_\ell^{(j)}\right)^2 \leq M^3 n^2\rho^4\|\mathbf{f}_*\|_{\mathcal{F}}^4\sum_{j=1}^M {\beta_t^{(j)}}^2
\end{aligned}$$

Here, in the first step we have used the definition of $\|\mathbb{E}[\Delta^{(j)}|X_1^n]\|_{\mathcal{H}_{k^{(j)}}}$, in the second step, equation (15), in the third step assumption 2 and Cauchy Schwarz, and in the last step, the definition of β_t. Plugging this back into (18), we get

$$\mathbb{E}\left[\|M^{1/2}\Phi^\top v\|^2\right] \leq M^{3/2}n^2\rho^4\|\mathbf{f}_*\|_{\mathcal{F}}^2\sqrt{\sum_{j=1}^M {\beta_t^{(j)}}^2}\sum_{j=1}^M\sum_{\ell=1}^t \mu_\ell^{(j)} \leq M^{3/2}n^2\rho^4\|\mathbf{f}_*\|_{\mathcal{F}}^2\Psi\beta_t$$

This bound, along with equation (17) gives us the desired result.

A.2.3. Proof of Lemma 9

Define $\pi_i^{(j)} = \{\phi_\ell^{(j)}(x_i)\}_{\ell=1}^t \in \mathbb{R}^t$, $\pi_i = [\pi_i^{(1)};\ldots;\pi_i^{(M)}] \in \mathbb{R}^{tM}$ and the matrices $A_i = Q^{-1}(\pi_i\pi_i^\top - I)Q^{-1} \in \mathbb{R}^{tm\times tM}$. Note that $A_i = A_i^\top$ and

$$\mathbb{E}[A_i] = Q^{-1}(\mathbb{E}[\pi_i\pi_i^\top] - I)Q^{-1} = \mathbf{0}.$$

Then, if $\epsilon_i, i = 1,\ldots,n$ are i.i.d Rademacher random variables, by a symmetrization argument we have,

$$\mathbb{E}\left[\left\|Q^{-1}\left(\frac{1}{n}\Phi^\top\Phi - I\right)Q^{-1}\right\|_{op}^k\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n A_i\right\|_{op}^k\right] \leq 2^k\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \epsilon_i A_i\right\|_{op}^k\right] \tag{19}$$

The above term can be bounded by the following expression.

$$\begin{aligned}
&2^q\left(\sqrt{e\max(q,\log(t))}\frac{\rho^2\sqrt{M}}{\sqrt{n}}\sqrt{\sum_{\ell=1}^M\gamma^{(j)}(\lambda)^2 + 4e\max(q,\log(t))\rho^2\left(\frac{M}{n}\right)^{1-1/q}\left(\sum_{\ell=1}^M\gamma^{(j)}(\lambda)^q\right)^{1/q}}\right)^q \\
&\leq \left(\frac{C}{2}\right)^q\max\left(\sqrt{M(\max(q,\log t))}, \frac{M^{1-1/q}\max(q,\log t)}{n^{1/2-1/q}}\right)^q\left(\frac{\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q
\end{aligned}$$

The proof mimics Lemma 6 in (Zhang et al., 2013) by performing essentially the same steps over $\mathcal{F}$ instead of the usual Hilbert space. In many of the steps, M terms appear (instead of the one term for KRR) which is accounted for via Jensen's inequality.

Finally, by Markov's inequality,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}^c) &\leq 2^k\mathbb{E}\left[\left\|Q^{-1}\left(\frac{1}{n}\Phi^\top\Phi - I\right)Q^{-1}\right\|_{op}^q\right] \\
&\leq C^q\max\left(\sqrt{M(\max(q,\log t))}, \frac{M^{1-1/q}\max(q,\log t)}{n^{1/2-1/q}}\right)^q\left(\frac{\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q
\end{aligned}$$

A.3. Variance (Proof of Bound (10))

Once again, we follow Zhang et al. (2013). The tricks we use to generalise it to the additive case (i.e. over $\mathcal{F}$) are the same as that for the bias. Note that since $\mathbb{E}[\|\hat{f}-\mathbb{E}\hat{f}\|_2^2] \leq \mathbb{E}[\|\hat{f}-g\|_2^2]$ for all g, it is sufficient to bound $\mathbb{E}[\|\hat{f}-f_*\|_2^2] = \mathbb{E}[\|\Delta\|_2^2]$.

First note that,

$$\lambda\mathbb{E}[\|\hat{\mathbf{f}}\|_{\mathcal{F}}^2|X_1^n] \;\leq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\left(\hat{f}(X_i)-Y_i\right)^2+\lambda\|\hat{\mathbf{f}}\|_{\mathcal{F}}^2\middle|X_1^n\right] \;\leq \frac{1}{n}\sum_{i=1}^n\mathbb{E}[\epsilon_i^2|X_1^n]+\lambda\|\mathbf{f}_*\|_{\mathcal{F}}^2 \leq \sigma^2+\lambda\|\mathbf{f}_*\|_{\mathcal{F}}^2$$

The second step follows by the fact that $\hat{\mathbf{f}}$ is the minimiser of (12). Then, for all j,

$$\mathbb{E}[\|\Delta^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2|X_1^n] \leq \mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathcal{F}}^2|X_1^n] \leq 2\|\mathbf{f}_*\|_{\mathcal{F}}^2+2\mathbb{E}[\|\hat{\mathbf{f}}\|_2^2|X_1^n] \leq \frac{2\sigma^2}{\lambda}+4\|\mathbf{f}_*\|_{\mathcal{F}}^2 \tag{20}$$

Let $\Delta^{(j)} = \sum_{\ell=1}^\infty \delta_\ell^{(j)}\phi_\ell^{(j)}$. Note that the definition of $\delta_\ell^{(j)}$ is different here. Define $\delta_\downarrow^{(j)}, \delta_\uparrow^{(j)}, \Delta_\downarrow^{(j)}, \Delta_\uparrow^{(j)}, \delta_\downarrow$ analogous to the definitions in Section A.2. Then similar to before we have,

$$\mathbb{E}[\|\delta_\uparrow^{(j)}\|_2^2] \leq \mu_{t+1}^{(j)}\mathbb{E}[\|\Delta_\uparrow^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2] \leq \mu_{t+1}^{(j)}\left(\frac{2\sigma^2}{\lambda}+4\|\mathbf{f}_*\|_{\mathcal{F}}^2\right)$$

We may use this to obtain a bound on $\mathbb{E}[\|\Delta_\uparrow\|^2]$. To obtain a bound on $\mathbb{E}[\|\Delta_\downarrow\|^2]$, take the $\mathcal{F}$ inner product of $(\mathbf{0},\ldots,\phi_\ell^{(j)},\ldots,\mathbf{0})$ with the first order optimality condition (13) and following essentially the same procedure to the bias we get,

$$\left(\frac{1}{n}\Phi^\top\Phi+\lambda\mathcal{M}^{-1}\right)\delta_\downarrow = -\lambda\mathcal{M}^{-1}\theta_\downarrow-\frac{1}{n}\Phi^\top v+\frac{1}{n}\Phi^\top\epsilon$$

where $\Phi, \mathcal{M}, \theta_\downarrow$ are the same as in the bias calculation. $v^{(j)} \in \mathbb{R}^n$ where $v_i^{(j)} = \sum_{\ell>t}\delta_\ell^{(j)}\phi_\ell^{(j)}(X_i) = \mathbb{E}[\Delta_\uparrow^{(j)}(X_i)|X_1^n]$ (recall that $\delta_\ell^{(j)}$ is different to the definition in the bias) and $\epsilon \in \mathbb{R}^n$, $\epsilon_i = Y_i - f_*(X_i)$ is the vector of errors. Then we write,

$$\left(I+Q^{-1}\left(\frac{1}{n}\Phi^\top\Phi-I\right)Q^{-1}\right)Q\delta_\downarrow = -\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow-\frac{1}{n}Q^{-1}\Phi^\top v+\frac{1}{n}Q^{-1}\Phi^\top\epsilon$$

where $Q = (I+\lambda\mathcal{M}^{-1})^{1/2}$ as before. Following a similar argument to the bias, when the event $\mathcal{E}$ holds,

$$\begin{aligned}\|\delta_\downarrow\|_2^2 \leq \|Q\delta_\downarrow\|_2^2 &\leq 4\|\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow+\frac{1}{n}Q^{-1}\Phi^\top v+\frac{1}{n}Q^{-1}\Phi^\top\epsilon\|_2^2\\ &\leq 12\|\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow\|^2+12\|\frac{1}{n}Q^{-1}\Phi^\top v\|^2+12\|\frac{1}{n}Q^{-1}\Phi^\top\epsilon\|_2^2\end{aligned} \tag{21}$$

By Lemma 7, the first term can be bounded via $12\lambda\|\mathbf{f}_*\|_{\mathcal{F}}^2$. For the second and third terms we use the following two lemmas, the proofs of which are given at the end of this subsection.

Lemma 10. $\mathbb{E}\left[\|\frac{1}{n}Q^{-1}\Phi^\top v\|_2^2\right] \leq \frac{1}{\lambda}M\rho^4\Psi\beta_t(2\sigma^2/\lambda+4\|\mathbf{f}_*\|_{\mathcal{F}}^2)$.

Lemma 11. $\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^\top\epsilon\right\|_2^2\right] \leq \frac{\sigma^2}{n}\gamma_k(\lambda)$

Note that $\mathbb{E}[\|\delta_\downarrow\|_2^2] \leq \mathbb{P}(\mathcal{E})\mathbb{E}[\|\delta_\downarrow\|_2^2|\mathcal{E}]+\mathbb{E}[\mathbb{1}(\mathcal{E}^c)\|\delta_\downarrow\|_2^2]$. The bound on the first term comes via equation (21) and Lemmas 7, 10 and 11. The second term can be bound via,

$$\begin{aligned}\mathbb{E}[\mathbb{1}(\mathcal{E}^c)\|\delta_\downarrow\|_2^2] &\leq \mathbb{E}[\mathbb{1}(\mathcal{E}^c)\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathcal{F}}^2|X_1^n]\\ &\leq \left(\max\left(\sqrt{\max(q,\log t)}\,,\,\frac{\max(q,\log t)}{n^{1/2-1/q}}\right)\times\frac{MC\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q\left(\frac{2\sigma^2}{\lambda}+4\|\mathbf{f}_*\|_{\mathcal{F}}^2\right)\end{aligned} \tag{22}$$

Here, we have used equation (20) and Lemma 9. Finally, note that

$$\begin{aligned}\mathbb{E}[\|\Delta\|_2^2] &\le M\sum_j \mathbb{E}[\|\Delta^{(j)}\|_2^2] = M\big(\mathbb{E}\|\delta_\downarrow\|_2^2 + \sum_j \mathbb{E}\|\delta_\uparrow^{(j)}\|_2^2\big)\\ &\le M\left(\mathbb{E}\|\delta_\downarrow\|_2^2 + \Big(\frac{2\sigma^2}{\lambda} + 4\|\mathbf{f}_*\|_{\mathcal{F}}^2\Big)\sum_j \mu_{t+1}^{(j)}\right)\end{aligned} \tag{23}$$

When we combine (21), (22) and (23) we get the bound in equation (10).

Proofs of Technical Lemmas

A.3.1. Proof of Lemma 10

Note that following an argument similar to equation (25) in Lemma 8, it is sufficient to bound $\mathbb{E}\|M^{1/2}\Phi^\top v\|_2^2$. We expand this as,

$$\mathbb{E}\left[\|M^{1/2}\Phi^\top v\|_2^2\right] = \sum_{j=1}^M\sum_{\ell=1}^t \mu_\ell^{(j)}\mathbb{E}[({\Phi_\ell^{(j)}}^\top v)^2] \le \sum_{j=1}^M\sum_{\ell=1}^t \mu_\ell^{(j)}\mathbb{E}[\|\Phi_\ell^{(j)}\|^2\|v\|^2]$$

To bound this term, first note that

$$\|v\|^2 = \sum_{i=1}^n\left(\sum_{j=1}^M v_i^{(j)}\right)^2 \le M\sum_{i=1}^n\sum_{j=1}^M {v_i^{(j)}}^2 \le M\sum_{i=1}^n\sum_{j=1}^M\left(\sum_{\ell>t}\frac{{\delta_\ell^{(j)}}^2}{\mu_\ell^{(j)}}\right)\left(\sum_{\ell>t}\mu_\ell^{(j)}\phi_\ell^{(j)}(X_i)^2\right)$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\|M^{1/2}\Phi^\top v\|^2\right] &\le \sum_{j=1}^M\sum_{\ell=1}^t \mu_\ell^{(j)} M\sum_{i=1}^n\sum_{j'=1}^M \mathbb{E}\left[\mathbb{E}[\|\Delta^{(j')}\|_{\mathcal{H}_{k^{(j')}}}^2|X_1^n]\|\Phi_\ell^{(j)}\|^2\sum_{\ell'>t}\mu_{\ell'}^{(j')}\phi_{\ell'}^{(j')}(X_i)^2\right] \\ &\le M\left(\frac{2\sigma^2}{\lambda} + 4\|\mathbf{f}_*\|_{\mathcal{F}}^2\right)\sum_{j=1}^M\sum_{\ell=1}^t\mu_\ell^{(j)}\sum_{i=1}^n\sum_{j'=1}^M\sum_{\ell'>t}\mu_{\ell'}^{(j')}\mathbb{E}\left[\|\Phi_\ell^{(j)}\|^2\phi_{\ell'}^{(j')}(X_i)^2\right]\end{aligned} \tag{24}$$

For all i, the inner expectation can be bounded using assumption 2 and Jensen's inequality via,

$$\begin{aligned}\mathbb{E}\left[\|\Phi_\ell^{(j)}\|^2\phi_{\ell'}^{(j')}(X_i)^2\right] &\le \sqrt{\mathbb{E}\left[\|\Phi_\ell^{(j)}\|^4\right]\mathbb{E}\left[\phi_{\ell'}^{(j')}(X_i)^4\right]} \le \rho^2\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^n\phi_\ell^{(j)}(X_i)^2\right)^2\right]} \\ &\le \rho^2\sqrt{\mathbb{E}\left[n\sum_{i=1}^n\phi_\ell^{(j)}(X_i)^4\right]} \le \rho^2\sqrt{n^2\rho^4} = n\rho^4.\end{aligned}$$

This yields,

$$\mathbb{E}\left[\|M^{1/2}\Phi^\top v\|^2\right] \le Mn^2\rho^4\left(\frac{2\sigma^2}{\lambda} + 4\|\mathbf{f}_*\|_{\mathcal{F}}^2\right)\underbrace{\sum_{j=1}^M\sum_{\ell=1}^t\mu_\ell^{(j)}}_{\le\Psi}\underbrace{\sum_{j'=1}^M\sum_{\ell'>t}\mu_{\ell'}^{(j')}}_{=\beta_t}$$

Finally, we have

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^\top v\right\|_2^2\right] \le \mathbb{E}\left[\frac{1}{\lambda}\left\|\frac{1}{n}M^{1/2}\Phi^\top v\right\|_2^2\right] \le \frac{1}{\lambda}M\rho^4\Psi\beta_t\left(\frac{2\sigma^2}{\lambda} + 4\|\mathbf{f}_*\|_{\mathcal{F}}^2\right) \tag{25}$$

A.3.2. PROOF OF LEMMA 11

We expand the LHS as follows to obtain the result.

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^\top\epsilon\right\|^2\right] = \frac{1}{n^2}\sum_{j=1}^{M}\sum_{\ell=1}^{t}\sum_{i=1}^{n}\frac{1}{1+\lambda/\mu_\ell^{(j)}}\mathbb{E}[\phi_\ell^{(j)}(X_i)^2\epsilon_i^2] \leq \frac{\sigma^2}{n}\sum_{j=1}^{M}\gamma^{(j)}(\lambda) = \frac{\sigma^2}{n}\gamma_k(\lambda)$$

The first step is just an expansion of the matrix. In the second step we have used $\mathbb{E}[\phi_\ell^{(j)}(X_i)^2\epsilon_i^2] = \mathbb{E}[\phi_\ell^{(j)}(X_i)^2\mathbb{E}[\epsilon_i^2|X_i]] \leq \sigma^2$ since $\mathbb{E}[\phi_\ell^{(j)}(X)^2] = 1$. In the last two steps we have used the definitions of $\gamma^{(j)}(\lambda)$ and $\gamma_k(\lambda)$.

B. Proof of Theorem 4: Rate of Convergence in Different RKHSs

Our strategy will be to choose λ so as to balance the dependence on n in the first two terms in the RHS of the bound in Theorem 3.

Proof of Theorem 4-1. **Polynomial Decay**:
The quantity $\gamma_k(\lambda)$ can be bounded via $M_d\sum_{\ell=1}^{\infty}1/(1+\lambda/\tilde{\mu}_\ell)$. If we set $\lambda = n^{\frac{-2s}{2s+d}}$, then

$$\begin{aligned}\frac{\gamma_k(\lambda)}{M_d} = \sum_{\ell=1}^{\infty}\frac{1}{1+n^{\frac{-2s}{2s+d}}/\tilde{\mu}_\ell} &\leq n^{\frac{d}{2s+d}} + \sum_{\ell>n^{\frac{d}{2s+d}}}\frac{1}{1+n^{\frac{2s}{2s+d}}\ell^{\frac{2s}{d}}} \\ &\leq n^{\frac{d}{2s+d}} + n^{-\frac{2s}{2s+d}}\sum_{\ell>n^{\frac{d}{2s+d}}}\frac{1}{n^{\frac{-2s}{2s+d}}+\ell^{\frac{2s}{d}}} \\ &\leq n^{\frac{d}{2s+d}} + n^{\frac{-2s}{2s+d}}\left(n^{\frac{d}{2s+d}} + \int_{n^{\frac{d}{2s+d}}}^{\infty}u^{-2s/d}\mathrm{d}u\right) \in \mathcal{O}(n^{\frac{d}{2s+d}}).\end{aligned}$$

Therefore, $\gamma_k(\lambda)/n \in \mathcal{O}(M_d n^{\frac{-2s}{2s+d}})$ giving the correct dependence on n as required. To show that $\chi(k)$ is negligible, set $t = n^{\frac{3d}{2s-d}}$. Ignoring the poly(D) terms, both $\tilde{\mu}_{t+1}, \beta_t \in \mathcal{O}(n^{\frac{-6s}{2s-d}})$ and $\chi(k)$ is low order. Therefore, by Thereom 3 the excess risk is in $\mathcal{O}(M_d^2 n^{\frac{-2s}{2s+d}})$. □

Proof of Theorem 4-2. **Exponential Decay**:
By setting $\lambda = 1/n$ and following a similar argument to above we have,

$$\begin{aligned}\frac{\gamma_k(\lambda)}{M_d} \leq \sqrt{\frac{\log n}{\alpha}} + \frac{1}{\lambda}\sum_{\ell>\sqrt{\log n/\alpha}}\tilde{\mu}_\ell &\leq \sqrt{\frac{\log n}{\alpha}} + n\tilde{\pi}^d\sum_{\ell>\sqrt{\log n/\alpha}}\exp(-\alpha\ell^2) \\ &\leq \sqrt{\frac{\log n}{\alpha}} + n\tilde{\pi}^d\left(\frac{1}{n} + \int_{\sqrt{\log n/\alpha}}^{\infty}\exp(-\alpha\ell^2)\right) = \sqrt{\frac{\log n}{\alpha}} + \tilde{\pi}^d\left(1 + \frac{\sqrt{\pi}}{2}(1-\Phi(\sqrt{\log n})\right),\end{aligned}$$

where Φ is the Gaussian cdf. In the first step we have bounded the first $\sqrt{\frac{\log n}{\alpha}}$ terms by 1 and then bounded the second term by a constant. Note that the last term is $o(1)$. Therefore ignoring $\log n$ terms, $\gamma_k(\lambda) \in \mathcal{O}(M_d\tilde{\pi}^d)$ which gives excess risk $\mathcal{O}(M_d^2\tilde{\pi}^d/n)$. $\chi(k)$ can be shown to be low order by choosing $t = n^2$ which results in $\tilde{\mu}_{t+1}, \beta_t \in \mathcal{O}(n^{-4})$. □

C. Proof of Theorem 5: Analysis in the Agnostic Setting

As before, we generalise the analysis by Zhang et al. (2013) to the tuple RKHS $\mathcal{F}$. We begin by making the following crucial observation about the population minimiser (7) $f_\lambda = \sum_{j=1}^{M} f_\lambda^{(j)}$,

$$f_\lambda = \operatorname*{argmin}_{g\in\mathcal{H}_{d,\lambda}}\|g - f_*\|_2^2. \tag{26}$$

To prove this, consider any $g = \sum_{j=1}^{M} g^{(j)} \in \mathcal{H}_{d,\lambda}$. Using the fact that $\mathcal{R}(g) = \mathcal{R}(f_*) + \|g - f_*\|_2^2$ for any g and that $\|g\|_{\mathcal{F}} \leq R_{d,\lambda}$ we obtain the above result as follows.

$$\begin{aligned}\mathbb{E}\big[(f_*(X) - Y)^2\big] + \|f_\lambda - f_*\|_2^2 + \lambda R_{d,\lambda}^2 &= \mathbb{E}[(f_\lambda(X) - Y)^2] + \lambda R_{d,\lambda}^2 \\ \leq \mathbb{E}[(g(X) - Y)^2] + \lambda \sum_{j=1}^{M} \|g^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2 &\leq \mathbb{E}\big[(f_*(X) - Y)^2\big] + \|g - f_*\|_2^2 + \lambda R_{d,\lambda}^2.\end{aligned}$$

By using the above, we get for all $\eta > 0$,

$$\begin{aligned}\mathbb{E}\big[\|\hat{f} - f_*\|_2^2\big] &\leq (1+\eta)\mathbb{E}\big[\|f_\lambda - f_*\|_2^2\big] + (1 + 1/\eta)\mathbb{E}\big[\|\hat{f} - f_\lambda\|_2^2\big] \\ &= (1+\eta) \underbrace{\inf_{g \in \mathcal{H}_{d,\lambda}} \|g - f_*\|_2^2}_{\mathbf{AE}} + (1 + 1/\eta) \underbrace{\mathbb{E}\big[\|\hat{f} - f_\lambda\|_2^2\big]}_{\mathbf{EE}}\end{aligned}$$

For the first step, by the AM-GM inequality we have $2\int(\hat{f} - f_\lambda)(f_\lambda - f_*) \leq 1/\eta \int(\hat{f} - f_\lambda)^2 + \eta \int (f_\lambda - f_*)^2$. In the second step we have used (26). The term $\mathbf{AE}$ is exactly as in Theorem 5 so we just need to bound $\mathbf{EE}$.

As before, we consider the RKHS $\mathcal{F}$. Denote the representation of f_λ in $\mathcal{F}$ by $\mathbf{f}_\lambda = (f_\lambda^{(1)}, \ldots, f_\lambda^{(M)})$. Note that $R_{d,\lambda} = \|f_\lambda\|_{\mathcal{F}}$. Analogous to the analysis in Appendix A we define $\Delta^{(j)} = \hat{f}^{(j)} - f_\lambda^{(j)}$, $\Delta = \sum_j \Delta^{(j)} = \hat{f} - f_\lambda$ and $\mathbf{\Delta} = (\Delta^{(1)}, \ldots, \Delta^{(M)})$. Note that $\mathbf{EE} = \mathbb{E}[\|\Delta\|_2^2]$.

Let $\Delta^{(j)} = \sum_{\ell=1}^{\infty} \delta_\ell^{(j)} \phi_\ell^{(j)}$ be the expansion of $\Delta^{(j)}$ in $L_2(\mathbb{P}_X)$. For $t \in \mathbb{N}$, which we will select later, define $\Delta_\downarrow^{(j)} = \sum_{\ell=1}^{t} \delta_\ell^{(j)} \phi_\ell^{(j)}$, $\Delta_\uparrow^{(j)} = \sum_{\ell > t} \delta_\ell^{(j)} \phi_\ell^{(j)}$, $\delta_\downarrow^{(j)} = (\delta^{(1)}, \ldots, \delta^{(t)}) \in \mathbb{R}^t$ and $\delta_\uparrow^{(j)} = (\delta_\ell^{(j)})_{\ell > t}$. Let $\Delta_\downarrow = \sum_j \Delta_\downarrow^{(j)}$ and $\Delta_\uparrow = \sum_j \Delta_\uparrow^{(j)}$. Continuing the analogy, let $f_\lambda^{(j)} = \sum_{j=1}^{M} \theta_\ell^{(j)} \phi_\ell^{(j)}$ be the expansion of $f_\lambda^{(j)}$. Let $\theta_\downarrow^{(j)} = (\theta_1^{(j)}, \ldots, \theta_t^{(j)}) \in \mathbb{R}^t$ and $\theta_\downarrow = [\theta_\downarrow^{(1)}; \ldots; \theta_\downarrow^{(M)}] \in \mathbb{R}^{tM}$. Let $v \in \mathbb{R}^n$ such that $v_i^{(j)} = \sum_{\ell > t} \delta_\ell^{(j)} \phi_\ell^{(j)}(X_i)$ and $v_i = \sum_j v_i^{(j)}$. Let $\epsilon \in \mathbb{R}^n$, $\epsilon_i = Y_i - f_\lambda(X_i)$. Also define the following quantities:

$$\varsigma_\lambda^2(x) = \mathbb{E}[(Y - f_\lambda(X))^2 | X = x], \qquad B_\lambda^4 = 32\|\mathbf{f}_\lambda\|_{\mathcal{F}}^4 + 8\mathbb{E}[\varsigma_\lambda^4(X)]/\lambda^2.$$

We begin with the following lemmas.

Lemma 12. $\mathbb{E}[\varsigma_\lambda^4(X)] \leq 8\Psi^2\|\mathbf{f}_\lambda\|_{\mathcal{F}}^4\rho^4 + 8\nu^4$.

Lemma 13. $\mathbb{E}\left[\left(\mathbb{E}[\|\mathbf{\Delta}\|_{\mathcal{F}}^2 | X_1^n]\right)^2\right] \leq B_\lambda^4$.

We first bound $\mathbb{E}[\|\Delta_\uparrow^{(j)}\|_2^2] = \sum_{\ell > t} \mathbb{E}{\delta_\ell^{(j)}}^2$ using Lemma 13 and Jensen's inequality.

$$\mathbb{E}\left[\|\delta_\uparrow^{(j)}\|_2^2\right] = \sum_{\ell > t} \mathbb{E}[{\delta_\ell^{(j)}}^2] \leq \mu_{t+1}^{(j)} \mathbb{E}\left[\sum_{\ell > t} \frac{{\delta_\ell^{(j)}}^2}{\mu_\ell^{(j)}}\right] \leq \mu_{t+1}^{(j)} \mathbb{E}\left[\|\Delta^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2\right] \leq \mu_{t+1}^{(j)} \mathbb{E}\left[\|\mathbf{\Delta}\|_{\mathcal{F}}^2\right] \leq \mu_{t+1}^{(j)} B_\lambda^2 \tag{27}$$

Next we proceed to bound $\mathbb{E}[\|\Delta_\downarrow\|_2^2]$. For this we will use $\Phi^{(j)}$,$\Phi_\ell^{(j)}$,$\mathcal{M}^{(j)}$,$\mathcal{M}$, Q from Appendix A. The first order optimality condition can be written as,

$$\frac{1}{n}\sum_{i=1}^{n} \xi_{X_i}\left(\langle \xi_{X_i}, \mathbf{\Delta}\rangle - \epsilon_i\right) + \lambda \hat{\mathbf{f}} = \mathbf{0}.$$

This has the same form as (13) but the definitions of $\mathbf{\Delta}$ and ϵ_i have changed. Now, just as in the variance calculation, when we take the $\mathcal{F}$-inner product of the above with $(\mathbf{0}, \ldots, \phi_\ell^{(j)}, \ldots, \mathbf{0})$ and repeat for all j we get,

$$\left(I + Q^{-1}\left(\frac{1}{n}\Phi^\top\Phi - I\right)Q^{-1}\right) Q\delta_\downarrow = -\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow - \frac{1}{n}Q^{-1}\Phi^\top v + \frac{1}{n}Q^{-1}\Phi^\top\epsilon$$

Since Φ, $\mathcal{M}$, Q are the same as before we may reuse Lemma 9. Then, as $Q \succeq I$ when the event $\mathcal{E}$ holds,

$$\begin{aligned}\|\delta_\downarrow\|_2^2 \le \|Q\delta_\downarrow\|_2^2 &\le 4\|\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow + \frac{1}{n}Q^{-1}\Phi^\top v + \frac{1}{n}Q^{-1}\Phi^\top \epsilon\|_2^2 \\ &\le 8\|\frac{1}{n}Q^{-1}\Phi^\top v\|^2 + 8\|\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow - \frac{1}{n}Q^{-1}\Phi^\top\epsilon\|_2^2 \end{aligned} \tag{28}$$

We now bound the two terms in the RHS in expectation via the following lemmas.

Lemma 14. $\mathbb{E}[\|\frac{1}{n}Q^{-1}\Phi^\top v\|^2] \le \frac{1}{\lambda}MB_\lambda^2\rho^4\Psi\beta_t$

Lemma 15. $\mathbb{E}[\|\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow - \frac{1}{n}Q^{-1}\Phi^\top\epsilon\|_2^2] \le \frac{1}{n}\rho^2\gamma_k(\lambda)\sqrt{\mathbb{E}[\varsigma_\lambda^4(X)]}$

Now by Lemma 13 we have, $\mathbb{E}[\|\delta_\downarrow\|_2^2] = \mathbb{P}(\mathcal{E})\mathbb{E}[\|\delta_\downarrow\|_2^2|\mathcal{E}] + \mathbb{E}[\mathbb{1}(\mathcal{E}^c)\|\delta_\downarrow\|_2^2] \le \mathbb{E}[\|\delta_\downarrow\|_2^2|\mathcal{E}] + B_\lambda^2\mathbb{P}(\mathcal{E}^c)$. $\mathbb{E}[\|\delta_\downarrow\|_2^2|\mathcal{E}]$ can be bounded using Lemmas 14 and 15 while $\mathbb{P}(\mathcal{E}^c)$ can be bounded using Lemma 9. Combining these results along with (27) we have the following bound for $\mathbf{EE} = \mathbb{E}[\|\Delta\|_2^2]$,

$$\begin{aligned}\mathbb{E}[\|\Delta\|_2^2] &\le \mathbb{E}\left[\left\|\sum_{j=1}^M \Delta^{(j)}\right\|_2^2\right] \le M\sum_{j=1}^M \mathbb{E}\left[\|\Delta^{(j)}\|_2^2\right] = M\left(\mathbb{E}[\|\delta_\downarrow\|_2^2] + \sum_{j=1}^M \mathbb{E}[\|\delta_\downarrow^{(j)}\|_2^2]\right) \\ &\le \frac{8}{n}M\rho^2\gamma_k(\lambda)\sqrt{\mathbb{E}[\varsigma_\lambda^4(X)]} + \frac{8}{\lambda}M^2B_\lambda^2\rho^4\Psi\beta_t + B_\lambda^2M\left(\frac{CM_db(n,t,q)\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q + B_\lambda^2M\sum_j \mu_{t+1}^{(j)}\end{aligned}$$

Now we choose t large enough so that the following are satisfied,

$$\beta_t \le \frac{\lambda}{M^2nB_\lambda^4}\,, \qquad \sum_{j=1}^M \mu_{t+1}^{(j)} \le \frac{1}{MnB_\lambda^4}\,, \qquad \left(\frac{CM_db(n,t,q)\rho^2\gamma_k(\lambda)}{\sqrt{n}}\right)^q \le \frac{1}{MnB_\lambda^4}.$$

Then the last three terms are $\mathcal{O}(1/nB_\lambda^2)$ and the first term dominates. Using Lemma 12 and recalling that $R_{d,\lambda}^2 = \sum_j {R_\lambda^{(j)}}^2 = \|\mathbf{f}_\lambda\|_{\mathcal{F}}^2$ we get $\mathbf{EE} \in \mathcal{O}\left(n^{-1}M\gamma_k(\lambda)R_{d,\lambda}^2\right)$ as given in the theorem.

Proofs of Technical Lemmas

C.1. Proof of Lemma 13

Since $\hat{f}$ is the minimiser of the empirical objective,

$$\begin{aligned}\mathbb{E}\left[\lambda\|\hat{\mathbf{f}}\|_{\mathcal{F}}^2|X_1^n\right] &\le \mathbb{E}\left[\lambda\sum_{j=1}^M\|\hat{f}^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2 + \frac{1}{n}\sum_{i=1}^n\left(\sum_{j=1}^M\hat{f}^{(j)}(X_i^{(j)}) - Y_i\right)^2\,\middle|\,X_1^n\right] \\ &\le \mathbb{E}\left[\lambda\sum_{j=1}^M\|f_\lambda^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2 + \frac{1}{n}\sum_{i=1}^n\left(\sum_{j=1}^M f_\lambda^{(j)}(X_i^{(j)}) - Y_i\right)^2\,\middle|\,X_1^n\right] \le \lambda\|\mathbf{f}_\lambda\|_{\mathcal{F}}^2 + \frac{1}{n}\sum_{i=1}^n\varsigma_\lambda^2(X_i)\end{aligned}$$

Noting that $\boldsymbol{\Delta} = \hat{\mathbf{f}} - \mathbf{f}_\lambda$ and using the above bound and Jensen's inequality yields,

$$\mathbb{E}[\|\boldsymbol{\Delta}\|_{\mathcal{F}}^2|X_1^n] \le 2\|\mathbf{f}_\lambda\|_{\mathcal{F}}^2 + 2\mathbb{E}[\|\hat{\mathbf{f}}\|_{\mathcal{F}}^2|X_1^n] \le 4\|\mathbf{f}_\lambda\|_{\mathcal{F}}^2 + \frac{2}{n\lambda}\sum_{i=1}^n\varsigma_\lambda^2(X_i)$$

Applying Jensen's inequality once again yields,

$$\mathbb{E}[(\mathbb{E}[\boldsymbol{\Delta}\|_{\mathcal{F}}^2|X_1^n])^2] \le \mathbb{E}\left[\frac{8}{n^2\lambda^2}\left(\sum_{i=1}^n\varsigma_\lambda^2\right)^2 + 32\|\mathbf{f}_\lambda\|_{\mathcal{F}}^4\right] \le \frac{8}{n\lambda^2}\sum_{i=1}^n\mathbb{E}[\varsigma_\lambda^4] + 32\|\mathbf{f}_\lambda\|_{\mathcal{F}}^4 = B_\lambda^4$$

C.2. Proof of Lemma 12

First, using Jensen's inequality twice we have

$$\mathbb{E}[\varsigma_\lambda^4(X)] \;= \mathbb{E}\left[\mathbb{E}[(Y - f_\lambda(X))^2|X]^2\right] \;\le \mathbb{E}\left[(Y - f_\lambda(X))^4\right] \;\le 8\mathbb{E}[f_\lambda^4(X)] + 8\mathbb{E}[Y^4] \tag{29}$$

Consider any $f_\lambda^{(j)}$,

$$f_\lambda^{(j)}(x) = \sum_{\ell=1}^{\infty} \theta_\ell^{(j)} \phi_\ell^{(j)}(x) \overset{(a)}{\le} \left(\sum_{\ell=1}^{\infty} \mu_\ell^{(j)^{1/3}} \theta_\ell^{(j)^{2/3}}\right)^{3/4} \left(\sum_{\ell=1}^{\infty} \frac{\theta_\ell^{(j)^2} \phi_\ell^{(j)}(x)^4}{\mu_\ell^{(j)}}\right)^{1/4}$$
$$\overset{(b)}{\le} \left(\sum_{j=1}^{M} \mu_\ell^{(j)}\right)^{1/2} \left(\sum_{j=1}^{M} \frac{\theta_\ell^{(j)^2}}{\mu_\ell^{(j)}}\right)^{1/4} \left(\sum_{\ell=1}^{\infty} \frac{\theta_\ell^{(j)^2} \phi_\ell^{(j)}(x)^4}{\mu_\ell^{(j)}}\right)^{1/4} = \Psi^{(j)^{1/2}} \|f_\lambda^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^{1/2} \left(\sum_{\ell=1}^{\infty} \frac{\theta_\ell^{(j)^2} \phi_\ell^{(j)}(x)^4}{\mu_\ell^{(j)}}\right)^{1/4}$$

In (a), we used Hölder's inequality on $\mu_\ell^{(j)^{1/4}} \theta_\ell^{(j)^{1/2}}$ and $\theta_\ell^{(j)^{1/2}} \phi_\ell^{(j)}(x)/\mu_\ell^{(j)^{1/4}}$ with conjugates $4/3$ and 4 respectively. In (b) we used Hölder's inequality once again on $\mu_\ell^{(j)^{2/3}}$ and $(\theta_\ell^{(j)^2}/\mu_\ell^{(j)})^{1/3}$ with conjugates $3/2$ and 3. Now we expand f_λ in terms of the $f_\lambda^{(j)}$'s as follows,

$$f_\lambda(x) \le \sum_{j=1}^{M} \Psi^{(j)^{1/2}} \|f_\lambda^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^{1/2} \left(\sum_{\ell=1}^{\infty} \frac{\theta_\ell^{(j)^2} \phi_\ell^{(j)}(x)^4}{\mu_\ell^{(j)}}\right)^{1/4} \le \left(\sum_{j=1}^{M} \Psi^{(j)}\right)^{1/2} \left(\sum_{j=1}^{M} \|f_\lambda^{(j)}\|_{\mathcal{H}_{k^{(j)}}} \left(\sum_{\ell=1}^{\infty} \frac{\theta_\ell^{(j)^2} \phi_\ell^{(j)}(x)^4}{\mu_\ell^{(j)}}\right)^{1/2}\right)^{1/2}$$

where we have applied Cauchy-Schwarz in the last step. Using Cauchy-Schwarz once again,

$$f_\lambda^2(X) \le \Psi \left(\sum_{j=1}^{M} \|f_\lambda^{(j)}\|_{\mathcal{H}_{k^{(j)}}}^2\right)^{1/2} \left(\sum_{j=1}^{M} \sum_{\ell=1}^{\infty} \frac{\theta_\ell^{(j)^2} \phi_\ell^{(j)}(X)^4}{\mu_\ell^{(j)}}\right)^{1/2}$$

Using Cauchy-Schwarz for one last time, we obtain

$$\mathbb{E}[f_\lambda^4(x)] \;\le \Psi^2 \|f_\lambda\|_{\mathcal{F}}^2 \sum_{j=1}^{M} \sum_{\ell=1}^{\infty} \frac{\theta_\ell^{(j)^2} \mathbb{E}[\phi_\ell^{(j)}(x)]^4}{\mu_\ell^{(j)}} \;\le \Psi^2 \|f_\lambda\|_{\mathcal{F}}^4 \rho^2$$

where we have used Assumption 2 in the last step. When we combine this with (29) and use the fact that $\mathbb{E}[Y^4] \le \nu^4$ we get the statement of the lemma.

C.3. Proof of Lemma 14

The first part of the proof will mimic that of Lemma 10. By repeating the arguments for (24), we get

$$\mathbb{E}\left[\|\mathcal{M}^{1/2}\Phi^\top v\|^2\right] \le \sum_{j=1}^{M}\sum_{\ell=1}^{t} \mu_\ell^{(j)} M \sum_{i=1}^{n}\sum_{j'=1}^{M} \mathbb{E}\left[\mathbb{E}[\|\Delta^{(j')}\|_{\mathcal{H}_{k^{(j')}}}^2 |X_1^n] \|\Phi_\ell^{(j)}\|^2 \sum_{\ell'>t} \mu_{\ell'}^{(j')} \phi_{\ell'}^{(j')}(X_i)^2\right]$$
$$\le M \sum_{j=1}^{M}\sum_{\ell=1}^{t}\sum_{i=1}^{n}\sum_{j'=1}^{M}\sum_{\ell'>t} \mu_\ell^{(j)} \mu_{\ell'}^{(j')} \mathbb{E}\left[\mathbb{E}[\|\Delta^{(j')}\|_{\mathcal{H}_{k^{(j')}}}^2 |X_1^n] \|\Phi_\ell^{(j)}\|^2 \phi_{\ell'}^{(j')}(X_i)^2\right]$$

Using Cauchy-Schwarz the inner expectation can be bounded via $\sqrt{\mathbb{E}\left[\left(\mathbb{E}[\|\Delta^{(j')}\|_{\mathcal{H}_{k^{(j')}}}^2]\right)^2\right] \mathbb{E}\left[\|\Phi_\ell^{(j)}\|^4 \phi_{\ell'}^{(j')}(X_i)^4\right]}$. Lemma 13 bounds the first expectation by B_λ^4. To bound the second expectation we use Assumption 2.

$$\mathbb{E}\left[\|\Phi_\ell^{(j)}\|^4 \phi_{\ell'}^{(j')}(X_k)^4\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} \phi_\ell^{(j)}(X_i)^2\right)^2 \phi_\ell^{(j)}(X_k)^4\right] = \mathbb{E}\left[\sum_{i,i'} \phi_\ell^{(j)}(X_i)^2 \phi_\ell^{(j)}(X_{i'})^2 \phi_\ell^{(j)}(X_k)^4\right] \le n^2 \rho^8$$

Finally once again reusing some calculations from Lemma 10,

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^\top v\right\|_2^2\right] \leq \mathbb{E}\left[\frac{1}{\lambda}\left\|\frac{1}{n}M^{1/2}\Phi^\top v\right\|_2^2\right] \leq \frac{M}{n^2\lambda}\underbrace{\left(\sum_{i=1}^n n\rho^4\right)}_{n^2\rho^4}\underbrace{\left(\sum_{j=1}^M\sum_{\ell=1}^t \mu_\ell^{(j)}\right)}_{\Psi}\underbrace{\left(\sum_{j'=1}^M\sum_{\ell'>t}^M \mu_{\ell'}^{(j')}\right)}_{\beta_t}$$

C.4. Proof of Lemma 15

First note that we can write the LHS of the lemma as,

$$\mathbb{E}\left[\left\|\lambda Q^{-1}\mathcal{M}^{-1}\theta_\downarrow - \frac{1}{n}Q^{-1}\Phi^\top\epsilon\right\|^2\right] = \sum_{j=1}^M\sum_{\ell=1}^t \frac{1}{1+\lambda/\mu_\ell^{(j)}}\mathbb{E}\left[\left(\frac{\lambda\theta_\ell^{(j)}}{\mu_\ell^{(j)}} - \frac{1}{n}\sum_{i=1}^n \phi_\ell^{(j)}(X_i^{(j)})\epsilon_i\right)^2\right]$$

To bound the inner expectation we use the optimality conditions of the population minimiser (7). We have,

$$2\mathbb{E}\left[\Big(\sum_{j=1}^M f_\lambda^{(j)}(X_i^{(j)}) - Y\Big)\xi_{X_i}^{(j)}\right] + 2\lambda f_\lambda^{(j)} = \mathbf{0} \implies \mathbb{E}\left[\xi_{X_i}^{(j)}\epsilon_i\right] = \lambda f_\lambda^{(j)} \implies \mathbb{E}\left[\phi_\ell^{(j)}(X_i^{(j)})\epsilon_i\right] = \lambda\frac{\theta_\ell^{(j)}}{\mu_\ell^{(j)}}. \quad (30)$$

In the last step we have taken the $\mathcal{F}$-inner product with $(\mathbf{0},\ldots,\phi_\ell^{(j)},\ldots,\mathbf{0})$. Therefore the term inside the expectation is the variance of $n^{-1}\sum_i \phi_\ell^{(j)}(X_i^{(j)})\epsilon_i$ and can be bounded via,

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n \phi_\ell^{(j)}(X^{(j)})\epsilon_i\right] \leq \frac{1}{n}\mathbb{E}\left[\phi_\ell^{(j)}(X^{(j)})^2\epsilon_i^2\right] \leq \frac{1}{n}\sqrt{\mathbb{E}\left[\phi_\ell^{(j)}(X^{(j)})^4\right]\mathbb{E}\left[\epsilon_i^4\right]} \leq \frac{1}{n}\rho^2\sqrt{\mathbb{E}[\varsigma_\lambda^4(X)]}$$

Hence the LHS can be bounded via,

$$\frac{1}{n}\rho^2\sqrt{\mathbb{E}[\varsigma_\lambda^4(X)]}\sum_{j=1}^M\sum_{\ell=1}^t \frac{1}{1+\lambda/\mu_\ell^{(j)}} = \frac{1}{n}\rho^2\gamma_k(\lambda)\sqrt{\mathbb{E}[\varsigma_\lambda^4(X)]}$$

D. Some Details on Experimental Setup

The function f_d used in Figure 1(a) is the log of three Gaussian bumps,

$$f_d(x) = \log\left(\alpha_1\frac{1}{h_d^d}\exp\left(\frac{\|x-v_1\|^2}{2h_d^2}\right) + \alpha_1\frac{1}{h_d^d}\exp\left(\frac{\|x-v_2\|^2}{2h_d^2}\right) + (1-\alpha_1-\alpha_2)\frac{1}{h_d^d}\exp\left(\frac{\|x-v_3\|^2}{2h_d^2}\right)\right) \quad (31)$$

where $h_d = 0.01\sqrt{d}$, $\alpha_1,\alpha_2 \in [0,1]$ and $v_i \in \mathbb{R}^d$ are constant vectors. For figures 1(b)-1(f) we used f_D where D is given in the figures. In all experiments, we used a test set of 2000 points and plot the mean squared test error.

For the real datasets, we normalised the training data so that the X, y values have zero mean and unit variance along each dimensions. We split the given dataset roughly equally to form a training set and testing set. We tuned hyper-parameters via 5-fold cross validation on the training set and report the mean squared error on the test set. For some datasets the test prediction error is larger than 1. Such datasets turned out to be quite noisy. In fact, when we used a constant predictor at 0 (i.e. the mean of the training instances) the mean squared error on the test set was typically much larger than 1.

Below, we list details on the dataset: the source, the used predictor and features.

1. **Housing:** (UCI), Predictor: CRIM
   Features: All other attributes except CHAS which is a binary feature.
2. **Galaxy:** (SDSS data on Luminous Red Galaxies from Tegmark et al (2006)), Predictor: Baryonic Density
   Features: All other attributes.

3. **fMRI:** (From (Just et al., 2010)), Predictor: Noun representation
   Features: Voxel Intensities. Since the actual dimensionality was very large, we use a random projection to bring it down to 100 dimensions.
4. **Insulin:** (From (Tu, 2012)), Predictor: Insulin levels.
   Features: SNP features
5. **Skillcraft:** (UCI), Predictor: TotalMapExplored
   Features: All other attributes. The usual predictor for this dataset is LeagueIndex but its an ordinal attribute and not suitable for real valued prediction.
6. **School:** (From Bristol Multilevel Modelling), Predictor: Given output
   Features: Given features. We don't know much about its attributes. We used the given features and labels.
7. **CCPP*:** (UCI), Predictor: Hourly energy output EP
   Features: The other 4 features and 55 random features for the other 55 dimensions.
8. **Blog:** (UCI Blog Feedback Dataset), Predictor: Number of comments in 24 hrs
   Features: The dataset had 280 features. The first 50 features were not used since they were just summary statistics. Our features included features 51-62 given in the UCI website and the word counts of 38 of the most frequently occurring words.
9. **Bleeding:** (From (Guillame-Bert et al., 2014)), Predictor: Given output
   Features: Given features reduced to 100 dimensions via a random projection. We got this dataset from a private source and don't know much about its attributes. We used the given features and labels.
10. **Speech:** (Parkinson Speech dataset from UCI), Predictor: Median Pitch
    Features: All other attributes except the mean pitch, standard deviation, minimum pitch and maximum pitches which are not actual features but statistics of the pitch.
11. **Music:** (UCI), Predictor: Year of production
    Features: All other attributes: 12 timbre average and 78 timbre covariance
12. **Telemonit:** (Parkinson's Telemonitoring dataset from UCI), Predictor: total-UPDRS
    Features: All other features except subject-id and motor-UPDRS (since it was too correlated with total-UPDRS). We only consider the female subjects in the dataset.
13. **Propulsion:** (Naval Propulsion Plant dataset from UCI), Predictor: Lever Position
    Features: All other attributes. We picked a random attribute as the predictor since no clear predictor was specifiied.
14. **Airfoil*:** (Airfoil Self-Noise dataset from UCI), Predictor: Sound Pressure Level
    Features: The other 5 features and 35 random features.
15. **Forestfires:** (UCI), Predictor: DC
    Features: All other attributes. We picked a random attribute as the predictor since no clear predictor was specifiied.
16. **Brain:** (From Wehbe et al. (2014)), Predictor: Story feature at a given time step
    Features: Other attributes

Some experimental details: GP is the Bayesian interpretation of KRR. However, the results are different in Table 1. We believe this is due to differences in hyper-parameter tuning. For GP, the GPML package (Rasmussen & Williams, 2006) optimises the GP marginal likelihood via L-BFGS. In contrast, our KRR implementation minimises the least squares cross validation error via grid search. Some Add-GP results are missing since it was very slow compared to other methods. On the Blog dataset, SALSA took less than 35s to train and all other methods were completed in under 22 minutes. In contrast Add-GP was not done training even after several hours. Even on the relatively small speech dataset Add-GP took about 80 minutes. Among the others, BF, MARS, and SpAM were the more expensive methods requiring several minutes on datasets with large D and n whereas other methods took under 2-3 minutes. We also experimented with locally cubic and quartic interpolation but exclude them from the table since LL, LQ generally performed better. Appendix D has more details on the synthetic functions and test sets.