Supplementary Material for Implicit Regularization of Random Feature Models

We organize the Supplementary Material (Supp. Mat.) as follows:

- In Section A, we present the details for the numerical results presented in the main text (and in the Supp. Mat.).
- In Section B, we present additional experiments and some discussions.
- In Section C, we present the proofs of the mathematical results presented in the main text.

A. Experimental Details

The experimental setting consists of N training and N_{tst} test datapoints $\{(x_i, y_i)\}_{i=1}^{N+N_{\text{tst}}} \in \mathbb{R}^d \times \mathbb{R}$. We sample P Gaussian features $f^{(1)}, \dots, f^{(P)}$ of $N + N_{\text{tst}}$ dimension with zero mean and covariance matrix entries thereof $C_{i,j} = K(x_i, x_j)$ where $K(x, x') = \exp(-\|x - x'\|^2/\ell)$ is a Radial Basis Function (RBF) Kernel with lengthscale ℓ. The extended data matrix $\bar{F} = \frac{1}{\sqrt{P}}[f^{(1)}, \dots, f^{(P)}]$ of size $(N + N_{\text{tst}}) \times P$ is decomposed into two matrices: the (training) data matrix $F = \bar{F}_{[:N,:]}$ of size $N \times P$, and a test data matrix $F_{\text{tst}} = \bar{F}_{[N:,:]}$ of size $N_{\text{tst}} \times P$ so that $\bar{F} = [F; F_{\text{tst}}]$. For a given ridge λ, we compute the optimal solution using the data matrix F, i.e. $\hat{\theta} = F^T \left(FF^T + \lambda \mathrm{I}_N\right)^{-1} y$ and obtain the predictions on the test datapoints $\hat{y}_{\text{tst}} = F_{\text{tst}} F^T \left(FF^T + \lambda \mathrm{I}_N\right)^{-1} y$.

Using the procedure above, we performed the following experiments:

A.1. Experiments with Sinusoidal data

We consider a dataset of $N = 4$ training datapoints $(x_i, \sin(x_i)) \in [0, 2\pi) \times [-1, 1]$ and $N_{\text{tst}} = 100$ equally spaced test data points in the interval $[0, 2\pi)$. In this experiment, the lengthscale of the RBF Kernel is $\ell = 2$. We compute the average and standard deviation the λ-RF predictor using 500 samplings of $\bar{F}$ (see Figure 1 in the main text and Figure 1 in the Supp. Mat.).

A.2. MNIST experiments

We sample $N = 100$ and $N_{\text{tst}} = 100$ images of digits 7 and 9 from the MNIST dataset (image size $d = 24 \times 24$, edge pixels cropped, all pixels rescaled down to $[0, 1]$ and recentered around the mean value) and label each of them with $+1$ and -1 labels, respectively. In this experiment, the lengthscale of the RBF Kernel is $\ell = d\ell_0$ where $\ell_0 = 0.2$. We approximate the expected λ-RF predictor on the test datapoints using the average of $\hat{y}_{\text{tst}}$ over 50 instances of $\bar{F}$ and compute the MSE (see Figures 2, 3 in the main text; in the ridgeless case –$\lambda = 10^{-4}$ in our experiments– when P is close to N, the average is over 500 instances). In Figure 4 of the main text, using $N_{\text{tst}} = 100$ test points, we compare two predictors trained over $N = 100$ and $N = 1000$ training datapoints.

A.3. Random Fourier Features

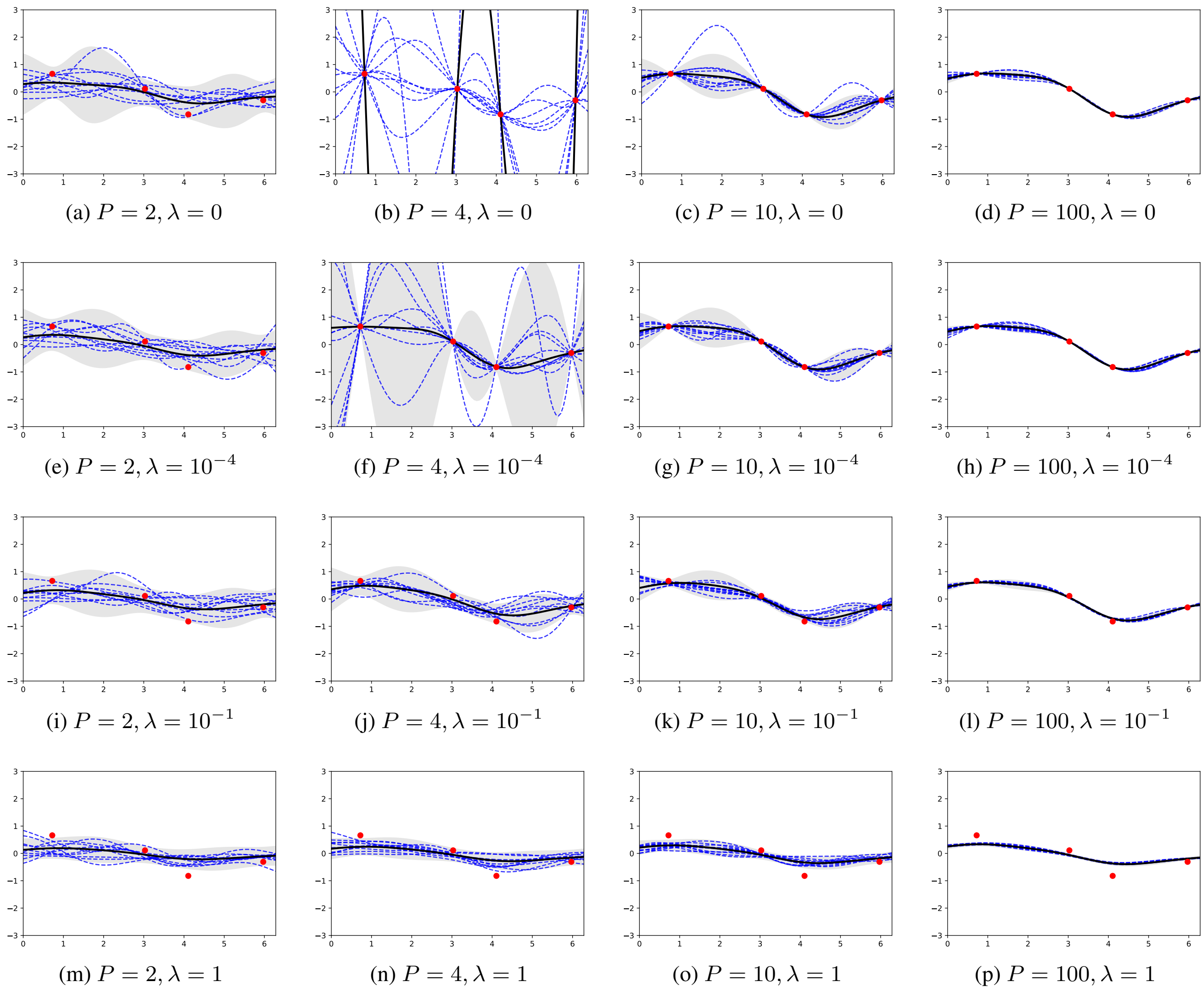
We sample random Fourier Features corresponding to the RBF Kernel with lengthscale $\ell = d\ell_0$ where $\ell_0 = 0.2$ (same as above) and consider the same dataset as in the MNIST experiment. The extended data matrix $\bar{F}$ for Fourier features is obtained as follows: we sample d-dimensional i.i.d. centered Gaussians $w^{(1)}, \dots, w^{(P)}$ with standard deviation $\sqrt{2/\ell}$, sample $b^{(1)}, \dots, b^{(P)}$ uniformly in $[0, 2\pi)$, and define $\bar{F}_{i,j} = \sqrt{\frac{2}{P}} \cos(x_i^T w^{(j)} + b^{(j)})$. We approximate the expected Fourier Features predictor on the test datapoints using the average of $\hat{y}_{\text{tst}}$ over 50 instances of $\bar{F}$ (see Figure 5).

B. Additional Experiments

We present the following complementary simulations:

- In Section B.1, we present the distribution of the λ-RF predictor for the selected P and λ.
- In Section B.2, we present the evolution of $\tilde{\lambda}$ and its derivative $\partial_\lambda \tilde{\lambda}$ for different eigenvalue spectra.
- In Section B.3, we show the evolution of the eigenvalue spectrum of $\mathbb{E}[A_\lambda]$.
- In Section B.4, we present numerical experiments on MNIST using random Fourier features.

B.1. Distribution of the RF predictor



(a) $P = 2, \lambda = 0$ (b) $P = 4, \lambda = 0$ (c) $P = 10, \lambda = 0$ (d) $P = 100, \lambda = 0$

(e) $P = 2, \lambda = 10^{-4}$ (f) $P = 4, \lambda = 10^{-4}$ (g) $P = 10, \lambda = 10^{-4}$ (h) $P = 100, \lambda = 10^{-4}$

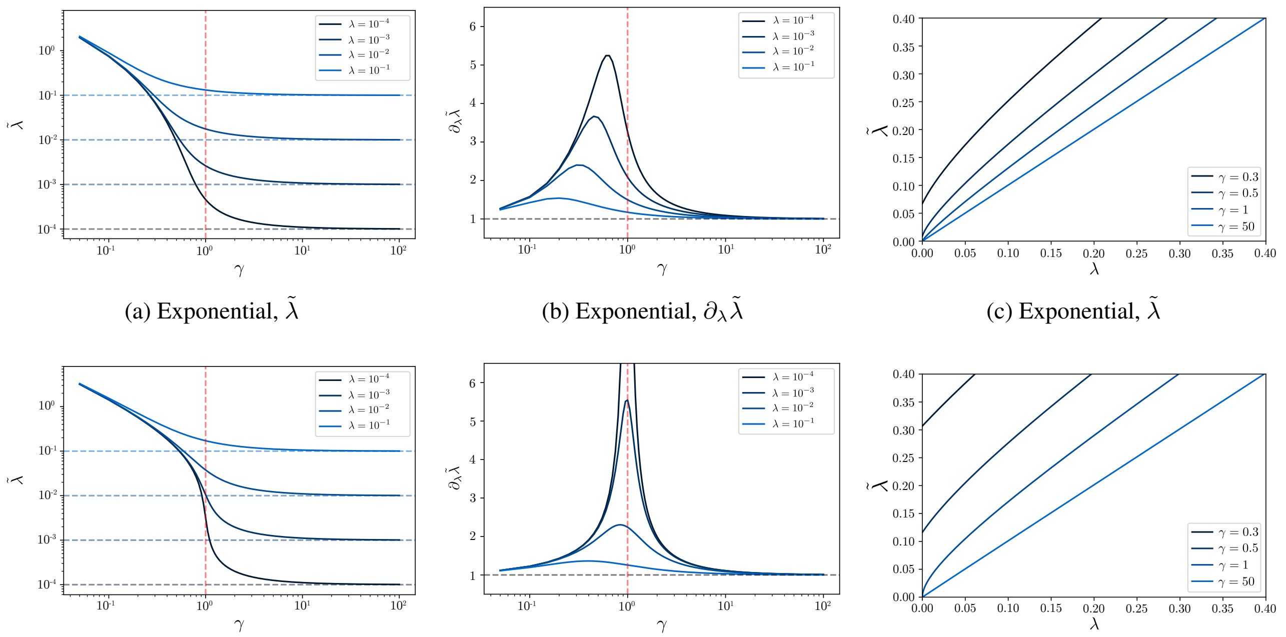
(i) $P = 2, \lambda = 10^{-1}$ (j) $P = 4, \lambda = 10^{-1}$ (k) $P = 10, \lambda = 10^{-1}$ (l) $P = 100, \lambda = 10^{-1}$

(m) $P = 2, \lambda = 1$ (n) $P = 4, \lambda = 1$ (o) $P = 10, \lambda = 1$ (p) $P = 100, \lambda = 1$

Figure 1. Distribution of the RF predictor. Red dots represent a sinusoidal dataset $y_i = \sin(x_i)$ for $N = 4$ points x_i in $[0, 2\pi)$. For $P \in \{2, 4, 10, 100\}$ and $\lambda \in \{0, 10^{-4}, 10^{-1}, 1\}$, we sample ten RF predictors (blue dashed lines) and compute empirically the average RF predictor (black lines) with ± 2 standard deviations intervals (shaded regions).

B.2. Evolution of the Effective Ridge $\tilde{\lambda}$

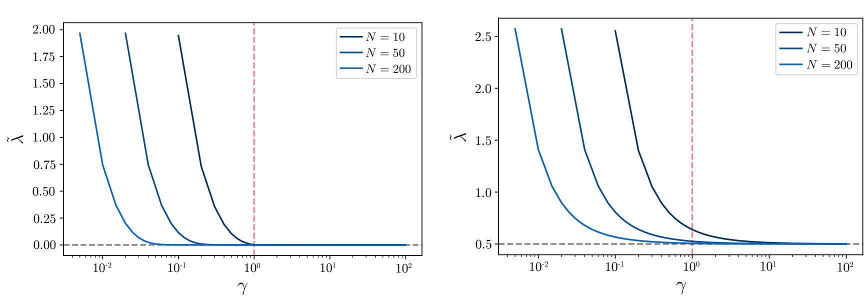
In Figure 2, we show how the effective ridge $\tilde{\lambda}$ and its derivative $\partial_\lambda \tilde{\lambda}$ evolve for the selected eigenvalue spectra with various decays (exponential or polynomial) as a function of γ and λ. In Figure 3, we compare the evolution of $\tilde{\lambda}$ for various N.



(a) Exponential, $\tilde{\lambda}$ (b) Exponential, $\partial_\lambda \tilde{\lambda}$ (c) Exponential, $\tilde{\lambda}$

(d) Polynomial, $\tilde{\lambda}$ (e) Polynomial, $\partial_\lambda \tilde{\lambda}$ (f) Polynomial, $\tilde{\lambda}$

Figure 2. Evolution of the effective ridge $\tilde{\lambda}$ and its derivative $\partial_\lambda \tilde{\lambda}$ for various levels of ridge λ (or γ) and for $N = 20$. We consider two different decays for $d_1, \ldots, d_N$: (i) exponential decay in i (i.e. $d_i = e^{-\frac{(i-1)}{2}}$, top plots) and (ii) polynomial decay in i (i.e. $d_i = \frac{1}{i}$, bottom plots).



(a) $\lambda = 10^{-4}$ (b) $\lambda = 0.5$

*Figure 3. Evolution of effective ridge $\tilde{\lambda}$ as a function of γ for two ridges **(a)** $\lambda = 10^{-4}$ and **(b)** $\lambda = 0.5$ and for various N.* We consider an exponential decay for $d_1, \ldots, d_N$, i.e. $d_i = e^{-\frac{(i-1)}{2}}$.

B.3. Eigenvalues of A_λ

The (random) prediction $\hat{y}$ on the training data is given by $\hat{y} = A_\lambda y$ where $A_\lambda = F(F^T F + \lambda I)^{-1} F^T$. The average λ-RF predictor is $\mathbb{E}[\hat{f}_\lambda^{(RF)}(x)] = K(x, X)K(X, X)^{-1}\mathbb{E}[A_\lambda]y$. We denote by $\tilde{d}_1, \ldots \tilde{d}_N$ the eigenvalues of $\mathbb{E}[A_\lambda]$. By Proposition C.7, the $\tilde{d}_i$'s converge to the eigenvalues $\frac{d_1}{d_1+\tilde{\lambda}}, \ldots, \frac{d_N}{d_N+\tilde{\lambda}}$ of $K(K + \tilde{\lambda} I_N)^{-1}$ as P goes to infinity. We illustrate the evolution of $\tilde{d}_i$ and their convergence to $\frac{d_i}{d_i+\tilde{\lambda}}$ for two different eigenvalue spectrums $d_1, \ldots d_N$.

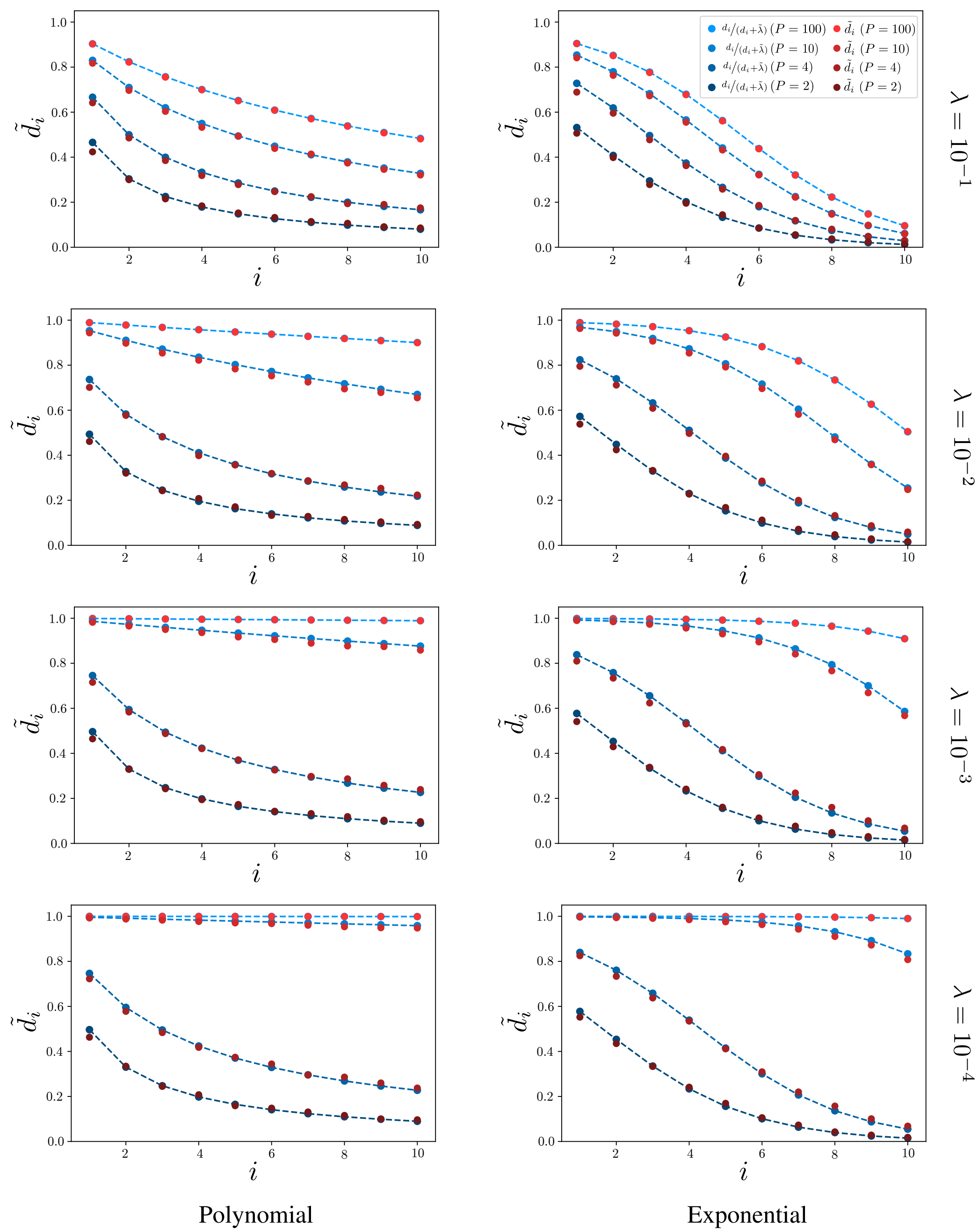


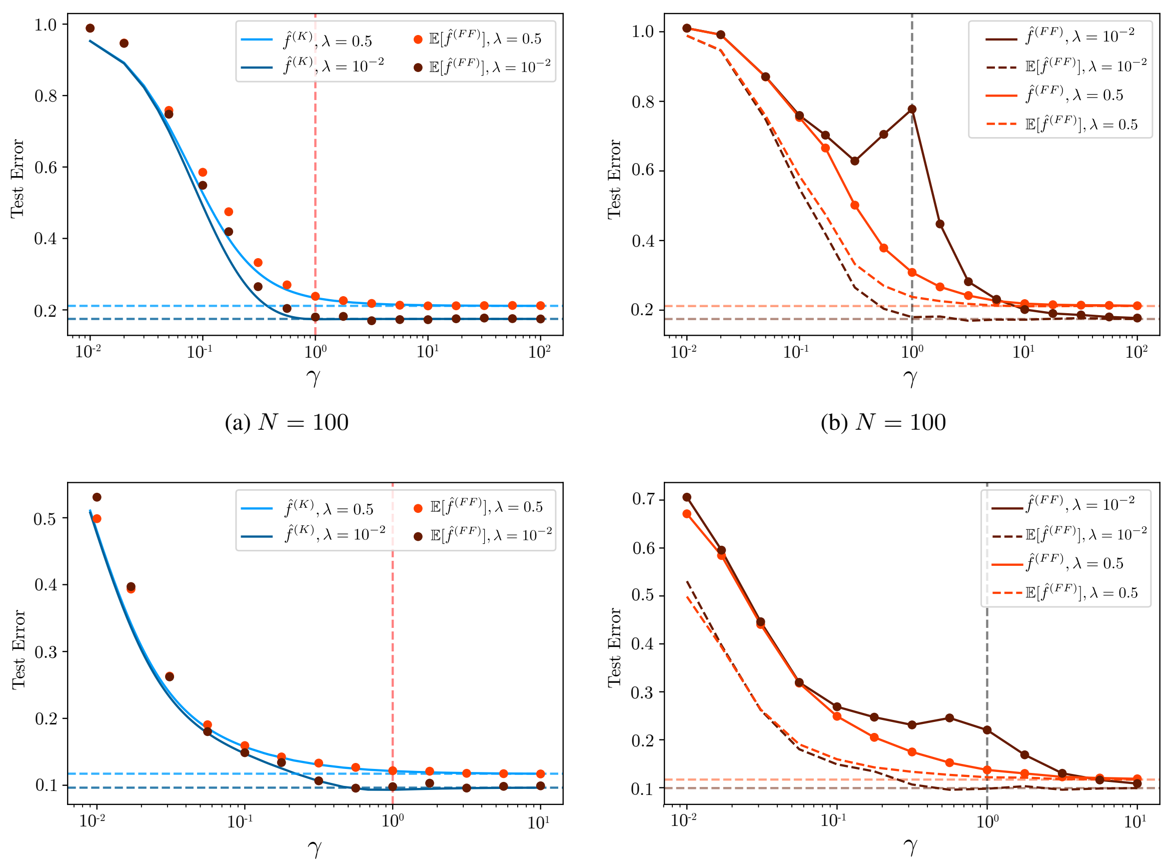
Figure 4. Eigenvalues $\tilde{d}_1, \ldots \tilde{d}_N$ (red dots) vs. eigenvalues $\frac{d_1}{d_1+\tilde{\lambda}}, \ldots, \frac{d_N}{d_N+\tilde{\lambda}}$ (blue dots) for $N = 10$. We consider various values of P and two different decays for $d_1, \ldots, d_N$: (i) exponential decay in i, i.e. $d_i = e^{-\frac{(i-1)}{2}}$ (right plots) and (ii) polynomial decay in i, i.e. $d_i = \frac{1}{i}$ (left plots).

B.4. Average Fourier Features Predictor

The Fourier Features predictor λ-FF is $\hat{f}^{(FF)}(x) = \frac{1}{\sqrt{P}} \sum_{j=1}^{P} \hat{\theta}_j \phi^{(j)}(x)$ where $\phi^{(j)}(x) = \cos(x^T w^{(j)} + b^{(j)})$ and $\hat{\theta} = F^T \left(FF^T + \lambda \mathrm{I}_N\right)^{-1} y$ with the data matrix F as described in Section A.3.

We investigate how close the average λ-FF predictor is to the $\tilde{\lambda}$-KRR predictor and we observe the following:

1. The difference of the test errors of the two predictors decreases as γ increases.
2. In the overparameterized regime, i.e. $P \geq N$, the test error of the $\tilde{\lambda}$-KRR predictor matches with the test error of the λ-FF predictor.
3. For $N = 1000$, strong agreement between the two test errors is observed already for $\gamma > 0.1$. We also observe that Gaussian features achieve lower (or equal) test error than the Fourier features for all γ in our experiments.



(a) $N = 100$ (b) $N = 100$

(c) $N = 1000$ (d) $N = 1000$

Figure 5. Comparision of the test errors of the average λ-FF predictor and the $\tilde{\lambda}$-KRR predictor. In **(a)** and **(c)**, the test errors of the average λ-FF predictor and of the $\tilde{\lambda}$-KRR predictor are reported for various ridge for $N = 100$ and $N = 1000$ MNIST data points (top and bottom rows). In **(b)** and **(d)**, the average test error of the λ-FF predictor and the test error of its average are reported.

C. Proofs

C.1. Gaussian Random Features

Proposition C.1. *Let $\hat{f}^{(RF)}_\lambda$ be the λ-RF predictor and let $\hat{y} = F\hat{\theta}$ be the prediction vector on training data, i.e. $\hat{y}_i = \hat{f}^{(RF)}_\lambda(x_i)$. The process $\hat{f}^{(RF)}_\lambda$ is a mixture of Gaussians: conditioned on F, we have that $\hat{f}^{(RF)}_\lambda$ is a Gaussian process. The mean and covariance of $\hat{f}^{(RF)}_\lambda$ conditioned on F are given by*

$$\mathbb{E}[\hat{f}^{(RF)}_\lambda(x)|F] = K(x,X)K(X,X)^{-1}\hat{y}, \tag{1}$$

$$\mathrm{Cov}[\hat{f}^{(RF)}_\lambda(x), \hat{f}^{(RF)}_\lambda(x')|F] = \frac{\|\hat{\theta}\|^2}{P}\tilde{K}(x,x') \tag{2}$$

where $\tilde{K}(x,x') = K(x,x') - K(x,X)K(X,X)^{-1}K(X,x')$ denotes the posterior covariance kernel.

Proof. Let $F = (\frac{1}{\sqrt{P}} f^{(j)}(x_i))_{i,j}$ be the $N \times P$ matrix of values of the random features on the training set. By definition, $\hat{f}^{(RF)}_\lambda = \frac{1}{\sqrt{P}}\sum_{p=1}^P \hat{\theta}_p f^{(p)}$. Conditioned on the matrix F, the optimal parameters $(\hat{\theta}_p)_p$ are not random and $(f^{(p)})_p$ is still Gaussian, hence, conditioned on the matrix F, the process $\hat{f}^{(RF)}_\lambda$ is a mixture of Gaussians. Moreover, conditioned on the matrix F, for any p, p', $f^{(p)}$ and $f^{(p')}$ remain independent, hence

$$\begin{aligned}
\mathbb{E}\left[\hat{f}^{(RF)}_\lambda(x) \mid F\right] &= \frac{1}{\sqrt{P}}\sum_{p=1}^P \hat{\theta}_p \mathbb{E}\left[f^{(p)}(x) \mid f^{(p)}_N\right] \\
\mathrm{Cov}\left[\hat{f}^{(RF)}_\lambda(x), \hat{f}^{(RF)}_\lambda(x') \mid F\right] &= \frac{1}{P}\sum_{p=1}^P \hat{\theta}_p^2 \mathrm{Cov}\left[f^{(p)}(x), f^{(p)}(x') \mid f^{(p)}_N\right].
\end{aligned}$$

where we have set $f^{(p)}_N = (f^{(p)}(x_i))_i \in \mathbb{R}^N$. The value of $\mathbb{E}\left[f^{(p)}(x) \mid f^{(p)}_N\right]$ and $\mathrm{Cov}\left[f^{(p)}(x), f^{(p)}(x') \mid f^{(p)}_N\right]$ are obtained from classical results on Gaussian conditional distributions (Eaton, 2007):

$$\begin{aligned}
\mathbb{E}\left[f^{(p)}(x) \mid f^{(p)}_N\right] &= K(x,X)K(X,X)^{-1} f^{(p)}_N, \\
\mathrm{Cov}\left[f^{(p)}(x), f^{(p)}(x') \mid f^{(p)}_N\right] &= \tilde{K}(x,x'),
\end{aligned}$$

where $\tilde{K}(x,x') = K(x,x') - K(x,X)K(X,X)^{-1}K(X,x')$. Thus, conditioned on F, the predictor $\hat{f}^{(RF)}_\lambda$ has expectation:

$$\mathbb{E}\left[\hat{f}^{(RF)}_\lambda(x) \mid F\right] = K(x,X)K(X,X)^{-1}\frac{1}{\sqrt{P}}\sum_{p=1}^P \hat{\theta}_p f^{(p)}_N = K(x,X)K(X,X)^{-1}\hat{y}$$

and covariance:

$$\mathrm{Cov}\left[\hat{f}^{(RF)}_\lambda(x), \hat{f}^{(RF)}_\lambda(x') \mid F\right] = \frac{1}{P}\sum_{p=1}^P \hat{\theta}_p^2 \tilde{K}(x,x') = \frac{\|\hat{\theta}\|^2}{P}\tilde{K}(x,x').$$

□

C.2. Generalized Wishart Matrix

Setup. In this section, we consider a fixed deterministic matrix K of size $N \times N$ which is diagonal positive semi-definite, with eigenvalues $d_1, \ldots, d_N$. We also consider a $P \times N$ random matrix W with i.i.d. standard Gaussian entries.

The key object of study is the $P \times P$ generalized Wishart random matrix $F^T F = \frac{1}{P} W K W^T$ and in particular its Stieltjes transform defined on $z \in \mathbb{C} \setminus \mathbb{R}^+$, where $\mathbb{R}^+ = [0, +\infty[$:

$$m_P(z) = \frac{1}{P}\mathrm{Tr}\left[\left(F^T F - z\mathrm{I}_P\right)^{-1}\right] = \frac{1}{P}\mathrm{Tr}\left[\left(\frac{1}{P} W K W^T - z\mathrm{I}_P\right)^{-1}\right],$$

where K is a fixed positive semi-definite matrix.

Since $F^T F$ has positive real eigenvalues $\lambda_1, \ldots, \lambda_P \in \mathbb{R}_+$, and

$$m_P(z) = \frac{1}{P} \sum_{p=1}^{P} \frac{1}{\lambda_p - z},$$

we have that for any $z \in \mathbb{C} \setminus \mathbb{R}^+$,

$$|m_P(z)| \leq \frac{1}{d(z, \mathbb{R}_+)},$$

where $d(z, \mathbb{R}_+) = \inf \{|z - y|, y \in \mathbb{R}^+\}$ is the distance of z to the positive real line. More precisely, $m_P(z)$ lies in the convex hull $\Omega_z = \mathrm{Conv}\left(\left\{\frac{1}{d-z} : d \in \mathbb{R}_+\right\}\right)$. As a consequence, the argument $\arg(m_P(z)) \in (-\pi, \pi)$ lies between 0 and $\arg\left(-\frac{1}{z}\right)$, i.e. $m_P(z)$ lies in the cone spanned by 1 and $-\frac{1}{z}$.

Our first lemma implies that the Stieljes transform concentrates around its mean as N and P go to infinity with $\gamma = \frac{P}{N}$ fixed.

Lemma C.2. *For any integer $m \in \mathbb{N}$ and any $z \in \mathbb{C} \setminus \mathbb{R}^+$, we have*

$$\mathbb{E}\left[|m_P(z) - \mathbb{E}[m_P(z)]|^m\right] \leq \mathbf{c} P^{-\frac{m}{2}},$$

where $\mathbf{c}$ depends on z, γ, and m only.

Proof. The proof follows Step 1 of (Bai & Wang, 2008). Let $w_1, ..., w_N$ be the columns of W from left to right. Let us introduce the $P \times P$ matrices $B(z) = \frac{1}{P} W K W^T - z\mathrm{I}_P$ and $B_{(i)}(z) = \frac{1}{P} W_{(i)} K_{(i)} W_{(i)}^T - z\mathrm{I}_P$ where $W_{(i)}$ is the $P \times (N-1)$ submatrix of W obtained by removing its i-th column w_i, and $K_{(i)}$ is the $(N-1) \times (N-1)$ submatrix of K obtained by removing both its i-th column and i-th row. Since the eigenvalues of WKW^T and $W_{(i)} K_{(i)} W_{(i)}^T$ are all real and positive, $B(z)$ and $B_{(i)}(z)$ are invertible matrices for $z \notin \mathbb{R}^+$.

Noticing that

$$B(z) = \frac{1}{P} W K W^T - z\mathrm{I}_P = \frac{1}{P} W_{(i)} K_{(i)} W_{(i)}^T - z\mathrm{I}_P + \frac{d_i}{P} w_i w_i^T$$

is a rank one perturbation of the matrix $B_{(i)}(z)$, by the Sherman–Morrison's formula, the inverse of $B(z)$ is given by:

$$B(z)^{-1} = \left(B_{(i)}(z)\right)^{-1} - \frac{d_i}{P} \frac{1}{1 + \frac{d_i}{P} w_i^T \left(B_{(i)}(z)\right)^{-1} w_i} \left(B_{(i)}(z)\right)^{-1} w_i w_i^T \left(B_{(i)}(z)\right)^{-1}.$$

We denote $\mathbb{E}_i$ the conditional expectation given $w_{i+1}, ..., w_N$. We have $\mathbb{E}_0[m_P(z)] = m_P(z)$ and $\mathbb{E}_N[m_P(z)] = \mathbb{E}[m_P(z)]$. As a consequence, we get:

$$\begin{aligned}
m_P(z) - \mathbb{E}[m_P(z)] &= \sum_{i=1}^{N} \left(\mathbb{E}_{i-1}[m_P(z)] - \mathbb{E}_i[m_P(z)]\right) \\
&= \frac{1}{P} \sum_{i=1}^{N} \left(\mathbb{E}_{i-1} - \mathbb{E}_i\right) \left[\mathrm{Tr}\left(B(z)^{-1}\right)\right] \\
&= \frac{1}{P} \sum_{i=1}^{N} \left(\mathbb{E}_{i-1} - \mathbb{E}_i\right) \left[\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right].
\end{aligned}$$

The last equality comes from the fact that $\mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)$ does not depend on w_i, hence

$$\mathbb{E}_{i-1}\left[\mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right] = \mathbb{E}_i\left[\mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right].$$

Let $g_i : \mathbb{C} \setminus \mathbb{R}^+ \to \mathbb{C}$ be the holomorphic function given by $g_i(z) := \frac{1}{P} w_i^T \left(B_{(i)}(z)\right)^{-1} w_i$. Its derivative is given by $g_i'(z) = \frac{1}{P} w_i^T \left(B_{(i)}(z)\right)^{-2} w_i$. Hence

$$\begin{aligned}\operatorname{Tr}\left(B(z)^{-1}\right) - \operatorname{Tr}\left(B_{(i)}(z)^{-1}\right) &= -\frac{\frac{d_i}{P} \operatorname{Tr}\left(\left(B_{(i)}(z)\right)^{-1} w_i w_i^T \left(B_{(i)}(z)\right)^{-1}\right)}{1 + d_i g_i(z)} \\ &= -\frac{d_i g_i'(z)}{1 + d_i g_i(z)},\end{aligned}$$

where we used the cyclic property of the trace. We can now bound this difference:

$$\begin{aligned}\left|\operatorname{Tr}\left(B(z)^{-1}\right) - \operatorname{Tr}\left(B_{(i)}(z)^{-1}\right)\right| &= \left|\frac{d_i g_i'(z)}{1 + d_i g_i(z)}\right| \\ &\leq \left|\frac{w_i^T \left(B_{(i)}(z)\right)^{-2} w_i}{w_i^T \left(B_{(i)}(z)\right)^{-1} w_i}\right| \\ &\leq \max_w \left|\frac{w^T \left(B_{(i)}(z)\right)^{-2} w}{w^T \left(B_{(i)}(z)\right)^{-1} w}\right| \\ &\leq \| \left(B_{(i)}(z)\right)^{-1} \|_{op} = \max_j |\frac{1}{\nu_j - z}| \leq \frac{1}{d(z, \mathbb{R}^+)},\end{aligned}$$

where ν_j are the eigenvalues of $\frac{1}{P} W_{(i)} K_{(i)} W_{(i)}^T$.

The sequence

$$\left((\mathbb{E}_{N-i} - \mathbb{E}_{N-i+1}) \left[\operatorname{Tr}\left(B(z)^{-1}\right) - \operatorname{Tr}\left(B_{(N-i+1)}(z)^{-1}\right)\right]\right)_{i=1,\ldots,N}$$

is a martingale difference sequence. Hence, by Burkholder's inequality, there exists a positive constant K_m such that

$$\begin{aligned}\mathbb{E}\left[\left|m_P(z) - \mathbb{E}\left[m_P(z)\right]\right|^m\right] &\leq K_m \frac{1}{P^m} \mathbb{E}\left[\left(\sum_{i=1}^N \left|\left[\mathbb{E}_{i-1} - \mathbb{E}_i\right] \left(\operatorname{Tr}\left(B(z)^{-1}\right) - \operatorname{Tr}\left(B_{(i)}(z)^{-1}\right)\right)\right|^2\right)^{\frac{m}{2}}\right] \\ &\leq K_m \frac{1}{P^m} \left(N \left(\frac{2}{d(z, \mathbb{R}_+)}\right)^2\right)^{\frac{m}{2}} \\ &\leq K_m \gamma^{-\frac{m}{2}} \left(\frac{2}{d(z, \mathbb{R}_+)}\right)^m P^{-\frac{m}{2}},\end{aligned}$$

hence the desired result with $\mathbf{c} = K_m \gamma^{-\frac{m}{2}} \left(\frac{2}{d(z, \mathbb{R}_+)}\right)^m$. □

The following lemma, which is reminiscent of Lemma 4.5 in (Au et al., 2018), is a consequence of Wick's formula for Gaussian random variables and is key to prove Lemma C.4.

Lemma C.3. *If $A^{(1)}, \ldots, A^{(k)}$ are k square random matrices of size P independent from a standard Gaussian vector w of size P,*

$$\mathbb{E}\left[w^T A^{(1)} w w^T A^{(2)} w \ldots w^T A^{(k)} w\right] = \sum_{p \in \boldsymbol{P}_2(2k)} \sum_{\substack{i_1, \ldots, i_{2k} \in \{1, \ldots, P\} \\ p \leq \operatorname{Ker}(i_1, \ldots, i_{2k})}} \mathbb{E}\left[A^{(1)}_{i_1 i_2} \ldots A^{(k)}_{i_{2k-1} i_{2k}}\right], \tag{3}$$

where $\boldsymbol{P}_2(2k)$ is the set of pair partitions of $\{1, \ldots, 2k\}$, $\leq$ is the coarser (i.e. $p \leq q$ if q is coarser than p), and for any $i_1, \ldots, i_{2k}$ in $\{1, \ldots, P\}$, $\operatorname{Ker}(i_1, \ldots, i_{2k})$ is the partition of $\{1, \ldots, 2k\}$ such that two elements u and v in $\{1, ..., 2k\}$ are in the same block (i.e. pair) of $\operatorname{Ker}(i_1, \ldots, i_{2k})$ if and only if $i_u = i_v$.

Furthermore,

$$\mathbb{E}\left[\left(w^T A^{(1)} w - \operatorname{Tr}\left(A^{(1)}\right)\right)\left(w^T A^{(2)} w - \operatorname{Tr}\left(A^{(2)}\right)\right)\dots\left(w^T A^{(k)} w - \operatorname{Tr}\left(A^{(k)}\right)\right)\right] = \sum_{p \in :\boldsymbol{P}_2(2k):} \sum_{\substack{i_1,\dots,i_{2k}\in\{1,\dots,P\} \\ p \leq \operatorname{Ker}(i_1,\dots,i_{2k})}} \mathbb{E}\left[A^{(1)}_{i_1 i_2}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right], \quad (4)$$

where $:\boldsymbol{P}_2(2k):$ *is the subset of partitions* p *in* $\boldsymbol{P}_2(2k)$ *for which* $\{2j-1, 2j\}$ *is not a block of* p *for any* $j \in \{1,\dots,k\}$.

Proof. Expanding the left-hand side of Equation (3), we obtain:

$$\mathbb{E}\left[\sum_{i_1,\dots,i_{2k}\in\{1,\dots,P\}} w_{i_1} A^{(1)}_{i_1 i_2} w_{i_2} w_{i_3} A^{(2)}_{i_3 i_4} w_{i_4} \dots w_{i_{2k-1}} A^{(k)}_{i_{2k-1} i_{2k}} w_{i_{2k}}\right].$$

Using Wick's formula, we get:

$$\sum_{i_1,\dots,i_{2k}\in\{1,\dots,P\}} \sum_{\substack{p\in \boldsymbol{P}_2(2k), \\ p\leq \operatorname{Ker}(i_1,\dots,i_{2k})}} \mathbb{E}\left[A^{(1)}_{i_1 i_2} A^{(2)}_{i_3 i_4}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right],$$

hence, interchanging the order of summation, we recover the left-hand side of Equation (3):

$$\sum_{p\in \boldsymbol{P}_2(2k)} \sum_{\substack{i_1,\dots,i_{2k}\in\{1,\dots,P\} \\ p\leq \operatorname{Ker}(i_1,\dots,i_{2k})}} \mathbb{E}\left[A^{(1)}_{i_1 i_2}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right].$$

We now prove Equation (4). Expanding the product, the left-hand side is equal to:

$$\sum_{I\subset\{1,\dots,k\}} (-1)^{k-\#I}\mathbb{E}\left[\prod_{i\in I} w^T A^{(i)} w \prod_{i\notin I} \operatorname{Tr}(A^{(i)})\right].$$

Expanding the product and the trace, and using Wick's equation, we obtain: a

$$\sum_{I\subset\{1,\dots,k\}} (-1)^{k-\#I} \sum_{i_1,\dots,i_{2k}\in\{1,\dots,P\}} \sum_{\substack{p\in \boldsymbol{P}_2(2k), p\leq p_I \\ p\leq \operatorname{Ker}(i_1,\dots,i_{2k})}} \mathbb{E}\left[A^{(1)}_{i_1 i_2}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right].$$

where p_I is the partition composed of blocks of size 2 given by $\{2l, 2l+1\}$ with $l \notin I$ and the rest of the indices contained in a single block. Interchanging the order of summation, we get:

$$\sum_{i_1,\dots,i_{2k}\in\{1,\dots,P\}} \sum_{\substack{p\in \boldsymbol{P}_2(2k), \\ p\leq \operatorname{Ker}(i_1,\dots,i_{2k})}} \mathbb{E}\left[A^{(1)}_{i_1 i_2}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right]\left[\sum_{\substack{I\subset\{1,\dots,k\}, \\ p\leq p_I}} (-1)^{k-\#I}\right].$$

Since $\left[\sum_{I\subset\{1,\dots,k\} p\leq p_I} (-1)^{\#I}\right] = \delta_{\{I\subset[k], p\leq p_I\}=\{\{1,\dots,k\}\}}$ and $\{I\subset[k], p \leq p_I\} = \{\{1,\dots,k\}\}$ if and only if $p \in :\boldsymbol{P}_2(2k):$, interchanging a last time the order of summation, we recover the left-hand side of Equation (4):

$$\sum_{p\in :\boldsymbol{P}_2(2k):} \sum_{\substack{i_1,\dots,i_{2k}\in\{1,\dots,P\} \\ p\leq \operatorname{Ker}(i_1,\dots,i_{2k})}} \mathbb{E}\left[A^{(1)}_{i_1 i_2}\dots A^{(k)}_{i_{2k-1}i_{2k}}\right].$$

$\square$

For any $z \in \mathbb{C} \setminus \mathbb{R}^+$, we define the holomorphic function $g_i : \mathbb{C} \setminus \mathbb{R}^+ \to \mathbb{C}$ by

$$g_i(z) = \frac{1}{P} w_i^T \left(\frac{1}{P} W_{(i)} K_{(i)} W_{(i)}^T - z\, I_P \right)^{-1} w_i,$$

where $W_{(i)}$ is the $P \times (N-1)$ submatrix of W obtained by removing its i-th column w_i, and $K_{(i)}$ is the $(N-1) \times (N-1)$ submatrix of K obtained by removing both its i-th column and i-th row. In the following lemma, we bound the distance of $g_i(z)$ to its mean. Then we prove that $\mathbb{E}[g_i(z)]$ is close to the expected Stieljes transform of K.

Lemma C.4. *The random function $g_i(z)$ satisfies:*

$$\begin{aligned} \left|\mathbb{E}\left[g_i(z)\right] - \mathbb{E}\left[m_P(z)\right]\right| &\leq \frac{\mathbf{c_0}}{P}, \\ \mathrm{Var}\left(g_i(z)\right) &\leq \frac{\mathbf{c_1}}{P}, \\ \mathbb{E}\left[\left(g_i(z) - \mathbb{E}\left[g_i(z)\right]\right)^4\right] &\leq \frac{\mathbf{c_2}}{P^2}, \\ \mathbb{E}\left[\left(g_i(z) - \mathbb{E}\left[g_i(z)\right]\right)^8\right] &\leq \frac{\mathbf{c_3}}{P^4}, \end{aligned}$$

where $\mathbf{c_0}$, $\mathbf{c_1}$, $\mathbf{c_2}$, and $\mathbf{c_3}$ depend on γ and z only.

Proof. The random variable w_i is independent from $B_{(i)}(z) = \frac{1}{P} W_{(i)} K_{(i)} W_{(i)}^T - z \mathrm{I}_P$ since the i-th column of W does not appear in the definition of $B_{(i)}(z)$. Using Lemma C.3, since there exists a unique pair partition $p \in \boldsymbol{P}_2(2)$, namely $\{\{1,2\}\}$, the expectation of $g_i(z)$ is given by

$$\mathbb{E}\left[g_i(z)\right] = \frac{1}{P} \mathbb{E}\left[\mathrm{Tr}\left[B_{(i)}(z)^{-1}\right]\right].$$

Recall that $\mathbb{E}\left[m_P(z)\right] = \frac{1}{P}\mathbb{E}\left[\mathrm{Tr}\left[B(z)^{-1}\right]\right]$ and $\left|\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right| \leq \frac{1}{d(z,\mathbb{R}_+)}$ (from the proof of Lemma C.2). Hence

$$\left|\mathbb{E}\left[g_i(z)\right] - \mathbb{E}\left[m_P(z)\right]\right| \leq \frac{1}{P}\mathbb{E}\left[\left|\mathrm{Tr}\left(B(z)^{-1}\right) - \mathrm{Tr}\left(B_{(i)}(z)^{-1}\right)\right|\right] \leq \frac{1}{P}\frac{1}{d(z,\mathbb{R}_+)}.$$

which proves the first assertion with $\mathbf{c_0} = \frac{1}{d(z,\mathbb{R}_+)}$.

Now, let us consider the variance of $g_i(z)$. Using our previous computation of $\mathbb{E}\left[g_i(z)\right]$, we have

$$\mathrm{Var}(g_i(z)) \quad = \quad \mathbb{E}\left[w_i^T \frac{\left(B_{(i)}(z)\right)^{-1}}{P} w_i w_i^T \frac{\left(B_{(i)}(z)\right)^{-1}}{P} w_i\right] - \mathbb{E}\left[\frac{1}{P}\mathrm{Tr}\left[B_{(i)}(z)^{-1}\right]\right]^2.$$

The first term can be computed using the first assertion of Lemma C.3: there are 2 matrices involved, thus we have to sum over 3 pair partitions. A simplification arises since $\frac{\left(B_{(i)}(z)\right)^{-1}}{P}$ is symmetric: the partition $\{\{1,2\},\{3,4\}\}$ yields $\mathbb{E}\left[\left(\mathrm{Tr}\left[\frac{\left(B_{(i)}(z)\right)^{-1}}{P}\right]\right)^2\right]$ whereas both $\{\{1,3\},\{2,4\}\}$ and $\{\{1,4\},\{2,4\}\}$ yield $\mathbb{E}\left(\mathrm{Tr}\left[\frac{\left(B_{(i)}(z)\right)^{-2}}{P^2}\right]\right)$.

Thus, the variance of $g_i(z)$ is given by:

$$\mathrm{Var}(g_i(z)) = 2\mathbb{E}\left(\mathrm{Tr}\left[\frac{\left(B_{(i)}(z)\right)^{-2}}{P^2}\right]\right) + \mathbb{E}\left[\left(\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-1}\right]\right)^2\right] - \mathbb{E}\left[\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-1}\right]\right]^2$$

hence is given by a sum of two terms:

$$\mathrm{Var}(g_i(z)) = \frac{2}{P}\mathbb{E}\left(\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-2}\right]\right) + \mathrm{Var}\left(\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-1}\right]\right).$$

Using the same arguments as those explained for the bound on the Stieltjes transform, the first term is bounded by $\frac{2}{Pd(z,\mathbb{R}_+)^2}$. In order to bound the second term, we apply Lemma C.2 for $W_{(i)}$ and $K_{(i)}$ in place of W and K. The second term is bounded by $\frac{\mathbf{c}}{P}$, hence the bound $\mathrm{Var}\left(g_i(z)\right) \leq \frac{\mathbf{c_1}}{P}$.

Finally, we prove the bound on the fourth moment of $g_i(z) - \mathbb{E}\left[g_i(z)\right]$. We denote $m_{(i)}(z) = \frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-1}\right]$. Recall that $\mathbb{E}\left[g_i(z)\right] = \mathbb{E}\left[m_{(i)}(z)\right]$. Using the convexity of $t \mapsto t^4$, we have

$$\begin{aligned}\mathbb{E}\left[\left(g_i(z) - \mathbb{E}[g_i(z)]\right)^4\right] &= \mathbb{E}\left[\left(g_i(z) - m_{(i)}(z) + m_{(i)}(z) - \mathbb{E}\left[m_{(i)}(z)\right]\right)^4\right] \\ &\leq 8\mathbb{E}\left[\left(g_i(z) - m_{(i)}(z)\right)^4\right] + 8\mathbb{E}\left[\left(m_{(i)}(z) - \mathbb{E}\left[m_{(i)}(z)\right]\right)^4\right].\end{aligned}$$

We bound the second term using the concentration of the Stieljes transform (Lemma C.2): it is bounded by $\frac{8\mathbf{c}}{P^2}$. The first term is bounded using the second assertion of Lemma C.3. Using the symmetry of $B_{(i)}(z)$, the partitions in : $\boldsymbol{P}_2(4)$: yield two different terms, namely:

1. $\frac{1}{P^2}\mathbb{E}\left[\left(\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-2}\right]\right)^2\right]$, for example if $p = \{\{1,3\},\{2,4\},\{5,7\},\{6,8\}\}$
2. $\frac{1}{P^3}\mathbb{E}\left[\frac{1}{P}\mathrm{Tr}\left[\left(B_{(i)}(z)\right)^{-4}\right]\right]$, for example if $p = \{\{2,3\},\{4,5\},\{6,7\},\{8,1\}\}$.

We bound the two terms using the same arguments as those explained for the bound on the Stieljes transform at the beginning of the section. The first term is bounded by $\frac{d(z,\mathbb{R}^+)^{-4}}{P^2}$ and the second term by $\frac{d(z,\mathbb{R}^+)^{-4}}{P^3}$ hence the bound $\mathbb{E}\left[\left(g_i(z) - \mathbb{E}\left[g_i(z)\right]\right)^4\right] \leq \frac{\mathbf{c_2}}{P^2}$.

The bound $\mathbb{E}[(g_i(z) - \mathbb{E}\left[g_i(z)\right])^8] \leq \frac{\mathbf{c_3}}{P^4}$ is obtained in a similar way, using the second assertion of Lemma C.3 and simple bounds on the Stieljes transform. $\square$

In the next proposition we show that the Stieltjes transform $m_P(z)$ is close in expectation to the solution of a fixed point equation.

Proposition C.5. *For any $z \in \mathbb{H}_{<0} = \{z : \mathrm{Re}(z) < 0\}$,*

$$|\mathbb{E}\left[m_P(z)\right] - \tilde{m}(z)| \leq \frac{\mathbf{e}}{P},$$

where $\mathbf{e}$ depends on z, γ, and $\frac{1}{N}\mathrm{Tr}(K)$ only and where $\tilde{m}(z)$ is the unique solution in the cone $\mathcal{C}_z := \{u - \frac{1}{z}v : u, v \in \mathbb{R}_+\}$ spanned by 1 and $-\frac{1}{z}$ of the equation

$$\gamma = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i\tilde{m}(z)}{1 + d_i\tilde{m}(z)} - \gamma z\tilde{m}(z).$$

Proof. We use the same notation as in the previous proofs, namely $B(z) = \frac{1}{P}WKW^T - z\mathrm{I}_P$, $B_{(i)}(z) = \frac{1}{P}W_{(i)}K_{(i)}W_{(i)}^T - z\mathrm{I}_P$ and $g_i(z) = \frac{1}{P}w_i^T\left(B_{(i)}(z)\right)^{-1}w_i$. Let $\nu_j \geq 0,\ j = 1,\ldots,P$ be the spectrum of the positive semi-definite matrix $\frac{1}{P}W_{(i)}K_{(i)}W_{(i)}^T$. After diagonalization, we have

$$B_{(i)}(z)^{-1} = O^T\mathrm{diag}(\frac{1}{\nu_1 - z},\ldots,\frac{1}{\nu_P - z})O,$$

with O an orthogonal matrix. Then

$$g_i(z) = \frac{1}{P}\mathrm{Tr}\left(\left(B_{(i)}(z)\right)^{-1}w_iw_i^T\right) = \frac{1}{P}\sum_{j=1}^{P}\frac{((Ow_i)_{jj})^2}{\nu_j - z}. \tag{5}$$

Since $z \in \mathbb{H}_{<0}$, we conclude that $\Re[g_i(z)] \geq 0$ for all $i = 1,\ldots,P$.

In order to prove the proposition, the key remark is that, since $\mathrm{Tr}\left((\frac{1}{P}WKW^T - z\mathrm{I}_P)(B(z))^{-1}\right) = P$, the Stieltjes transform $m_P(z)$ satisfies the following equation:

$$P = \mathrm{Tr}\left(\frac{1}{P}KW^TB(z)^{-1}W\right) - zPm_P(z).$$

From the proof of Lemma C.2, recall that $B^{-1}(z) = B_{(i)}^{-1}(z) - \frac{d_i}{P} \frac{1}{1+\frac{d_i}{P} w_i^T B_{(i)}^{-1}(z) w_i} B_{(i)}^{-1}(z) w_i w_i^T B_{(i)}^{-1}(z)$, hence:

$$\begin{aligned} \frac{1}{P} w_i^T B^{-1}(z) w_i &= g_i(z) - \frac{d_i g_i(z)^2}{1 + d_i g_i(z)} \\ &= \frac{g_i(z)}{1 + d_i g_i(z)}. \end{aligned} \tag{6}$$

Expanding the trace,

$$\operatorname{Tr}\left(\frac{1}{P} K W^T B(z)^{-1} W\right) = \sum_{i=1}^{N} d_i \frac{1}{P} w_i^T B^{-1}(z) w_i = \sum_{i=1}^{N} \frac{d_i g_i(z)}{1 + d_i g_i(z)}.$$

Thus, the Stieljes transform $m_P(z)$ satisfies the following equation $P = \sum_{i=1}^{N} \frac{d_i g_i(z)}{1+d_i g_i(z)} - zPm_P(z)$, or equivalently

$$\gamma = \frac{1}{N} \sum_{i=1}^{N} \frac{d_i g_i(z)}{1 + d_i g_i(z)} - z\gamma m_P(z).$$

Recall that $\gamma > 0$ and $\mathrm{Re}(z) < 0$. The Stieljes transform $m_P(z)$ can be written as a function of $g_i(z)$ for $i = 1, \ldots, n$: $m_P(z) = f(g_1(z), ..., g_N(z))$ where

$$f(g_1, \ldots, g_N) = \frac{1}{\gamma z N} \sum_{i=1}^{N} \frac{d_i g_i}{1 + d_i g_i} - \frac{1}{z} = -\frac{1}{z}\left(1 - \frac{1}{\gamma} + \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + d_i g_i}\right).$$

From Lemma C.6, the map $f(m) = f(m, ..., m)$ has a unique non-degenerate fixed point $\tilde{m}(z)$ in the cone $\mathcal{C}_z$. We will show that $\mathbb{E}[m_P(z)]$ is close to $\tilde{m}(z)$ using the following two steps: we show a non-tight bound $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}'}{\sqrt{P}}$ and use it to obtain the tighter bound $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}}{P}$.

Let us prove the $\frac{\mathbf{e}'}{\sqrt{P}}$ bound. From Lemma C.6, the distance between $m_P(z)$ and the fixed point $\tilde{m}(z)$ of f is bounded by the distance between $f(m_P(z), \ldots, m_P(z))$ and $m_P(z)$. Using the fact that $m_P(z) = f(g_1(z), ..., g_N(z))$, we obtain

$$|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \mathbb{E}\left[|m_P(z) - \tilde{m}(z)|\right] \leq \mathbb{E}\left[|f(m_P(z), \ldots, m_P(z)) - f(g_1(z), ..., g_N(z))|\right].$$

Recall that for any $z \in \mathbb{H}_{<0}$, $\Re(g_i(z)) \geq 0$: we need to study the function f on $\mathbb{H}_{\geq 0}^N$ where $\mathbb{H}_{\geq 0} = \{z \in \mathbb{C} | \Re(z) \geq 0\}$. On $\mathbb{H}_{\geq 0}^N$, the function f is Lipschitz:

$$|\partial_{g_i} f(g_1, .., g_N)| = \left|\frac{1}{\gamma z N} \frac{d_i}{(1 + d_i g_i)^2}\right| \leq \frac{d_i}{\gamma |z| N}.$$

Thus,

$$\mathbb{E}\left[|f(m_P(z), ..., m_P(z)) - f(g_1(z), ..., g_N(z))|\right] \leq \sum_{i=1}^{N} \frac{d_i}{\gamma |z| N} \mathbb{E}\left[|m_P(z) - g_i(z)|\right].$$

Since

$$\mathbb{E}\left[|m_P(z) - g_i(z)|\right] \leq \mathbb{E}\left[|m_P(z) - \mathbb{E}[m_P(z)]|\right] + |\mathbb{E}[m_P(z)] - \mathbb{E}[g_i(z)]| + \mathbb{E}\left[|g_i(z) - \mathbb{E}[g_i(z)]|\right],$$

using Lemmas C.2 and C.4, we get that $\mathbb{E}\left[|m_P(z) - g_i(z)|\right] \leq \frac{\mathbf{d}}{\sqrt{P}}$, where $\mathbf{d}$ depends on γ and z only. This implies that

$$\mathbb{E}\left[|f(m_P(z), ..., m_P(z)) - f(g_1(z), ..., g_N(z))|\right] \leq \frac{1}{\sqrt{P}} \frac{\mathbf{d}}{N} \operatorname{Tr}(K),$$

which allows to conclude that $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}'}{\sqrt{P}}$ where $\mathbf{e}'$ depends on γ, z and $\frac{1}{N}\operatorname{Tr}(K)$ only.

We strengthen this inequality and show the $\frac{\mathbf{e}}{P}$ bound. Using again Lemma C.6, we bound the distance between $\mathbb{E}[m_P(z)]$ and the fixed point $\tilde{m}(z)$ by

$$|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq |\mathbb{E}[f(g_1(z), \ldots, g_N(z))] - f(\mathbb{E}[m_P(z)], \ldots, \mathbb{E}[m_P(z)])|$$

and study the r.h.s. using a Taylor approximation of f near $\mathbb{E}\left[m_P(z)\right]$. For $i = 1, \ldots, N$ and $m_0 \in \mathbb{H}_{\geq 0}$, let $\mathrm{T}_{m_0} h_i$ be the first order Taylor approximation of the map $h_i : m \mapsto \frac{1}{1+d_i m}$ at a point m_0. The error of the first order Taylor approximation is given by

$$h_i(m) - \mathrm{T}_{m_0} h_i(m) = \frac{1}{1+d_i m} - \left(\frac{1}{1+d_i m_0} - \frac{d_i(m-m_0)}{\left(1+d_i m_0\right)^2}\right) = \frac{d_i^2\left(m_0-m\right)^2}{\left(1+d_i m\right)\left(1+d_i m_0\right)^2},$$

which, for $m \in \mathbb{H}_{\geq 0}$ can be upper bounded by a quadratic term:

$$|h_i(m) - \mathrm{T}_{m_0} h_i(m)| = \left|\frac{d_i^2}{\left(1+d_i m\right)\left(1+d_i m_0\right)^2}\right| |m_0 - m|^2 \leq \frac{1}{|m_0|^2} |m_0 - m|^2 . \tag{7}$$

The first order Taylor approximation $\mathrm{T}f$ of f at the N-tuple $\left(\mathbb{E}\left[m_P(z)\right], ..., \mathbb{E}\left[m_P(z)\right]\right)$ is

$$\mathrm{T}f(g_1, .., g_N) = -\frac{1}{z}\left(1 - \frac{1}{\gamma} + \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N} \mathrm{T}_{\mathbb{E}[m_P(z)]} h_i(g_i)\right).$$

Using this Taylor approximation, $\mathbb{E}[f(g_1(z), \ldots, g_N(z))] - f(\mathbb{E}[m_P(z)], \ldots, \mathbb{E}[m_P(z)])$ is equal to:

$$\mathbb{E}\left[\mathrm{T}f(g_1(z), .., g_N(z))\right] - f(\mathbb{E}[m_P(z)], \ldots, \mathbb{E}[m_P(z)]) + \mathbb{E}\left[f(g_1(z), ..., g_N(z)) - \mathrm{T}f(g_1(z), .., g_N(z))\right].$$

Using Lemma C.4, we get

$$\begin{aligned}
|\mathbb{E}\left[f(g_1(z), ..., g_N(z)) - \mathrm{T}f(g_1(z), .., g_N(z))\right]| &\leq \frac{1}{|z|\,\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathbb{E}[m_P(z)]|^2}\mathbb{E}\left[\left|g_i(z) - \mathbb{E}\left[m_P(z)\right]\right|^2\right] \\
&\leq \frac{1}{P}\frac{\alpha}{|\mathbb{E}[m_P(z)]|^2}
\end{aligned}$$

and

$$\begin{aligned}
|\mathbb{E}\left[\mathrm{T}f(g_1(z), .., g_N(z))\right] - f(\mathbb{E}\left[m_P(z)\right], ..., \mathbb{E}\left[m_P(z)\right])| &\leq \frac{1}{|z|\,\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{d_i\left|\mathbb{E}\left[g_i\right] - \mathbb{E}\left[m_P(z)\right]\right|}{\left|1 + d_i\mathbb{E}\left[m_P(z)\right]\right|^2} \\
&\leq \frac{\beta\left(\frac{1}{N}\mathrm{Tr}K\right)}{P}
\end{aligned}$$

where α and β depends on z and γ only. From the bounds $|\mathbb{E}[m_P(z)] - \tilde{m}(z)| \leq \frac{\mathbf{e}'}{\sqrt{P}}$ and $|\tilde{m}(z)| \geq (|z| + \frac{1}{N\gamma}\mathrm{Tr}(K))^{-1}$ (Lemma C.6), the bound $\frac{1}{P}\frac{\alpha}{|\mathbb{E}[m_P(z)]|^2}$ yields a $\frac{\tilde{\alpha}}{P}$ bound. This implies that $|\mathbb{E}[m_P(z)] - f(\mathbb{E}[m_P(z)], \ldots, \mathbb{E}[m_P(z)])| \leq \frac{\mathbf{e}}{P}$, hence the desired inequality $|\mathbb{E}\left[m_P(z)\right] - \tilde{m}(z)| \leq \frac{\mathbf{e}}{P}$. □

For the proof of Proposition C.5, we have used the fact that the map f_z introduced therein has a unique non-degenerate fixed point in the cone $\mathcal{C}_z := \{u - \frac{1}{z}v : u, v \in \mathbb{R}_+\}$. We now proceed with proving this statement.

Lemma C.6. *Let $d_1, \ldots, d_n \geq 0$ and let $\gamma \geq 0$. For any fixed $z \in \mathbb{H}_{<0}$, let $f_z : \mathbb{H}_{\geq 0} \to \mathbb{C}$ be the function $t \mapsto f_z(t) = -\frac{1}{z}\left(1 - \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{d_i t}{1+d_i t}\right)$. Let $\mathcal{C}_z := \{u - \frac{1}{z}v : u, v \in \mathbb{R}_+\}$ be the convex region spanned by the half-lines $\mathbb{R}_+$ and $-\frac{1}{z}\mathbb{R}_+$. Then for every $z \in \mathbb{H}_{<0}$ there exists a unique fixed point $\tilde{t}(z) \in \mathcal{C}_z$ such that $\tilde{t}(z) = f_z(\tilde{t}(z))$. The map $\tilde{t} : z \mapsto \tilde{t}(z)$ is holomorphic in $\mathbb{H}_{<0}$ and*

$$|\tilde{t}(z)| \geq \left(|z| + \frac{\sum_i d_i}{\gamma N}\right)^{-1}.$$

Furthermore for every $z \in \mathbb{H}_{<0}$ and any $t \in \mathbb{H}_{\geq 0}$, one has

$$|t - \tilde{t}(z)| \leq |t - f_z(t)|.$$

Proof. By means of Schwarz reflection principle, we can assume that $\Im(z) \geq 0$. Let $z \in \mathbb{H}_{<0}$ and let $\Pi_z := \{-\frac{w}{z} : \Im(w) \leq 0\}$ and let $\mathcal{C}_z$ be the wedged region $\mathcal{C}_z := \Pi_z \cap \{w \in \mathbb{C} : \Im(w) \geq 0\}$. To show the existence of a fixed point in $\mathcal{C}_z$ we show that 0 is in the image of the function $\psi : t \mapsto f_z(t) - t$. Note that since $d_i \geq 0$, the eventual poles of f_z are all strictly negative real numbers, hence $\psi : \mathcal{C}_z \to \mathbb{C}$ is an holomorphic function.

To prove that $0 \in \psi(\mathcal{C}_z)$ we proceed with a geometrical reasoning: the image $\psi(\mathcal{C}_z)$ is (one of) the region of the plane confined by $\psi\left(\partial\mathcal{C}_z\right)$, so we only need to "draw" $\psi\left(\partial\mathcal{C}_z\right)$ and show that 0 belongs to the "good" connected component confined by it.

The boundary of C_z is made up of two half-lines $\mathbb{R}_+$ and $-\frac{1}{z}\mathbb{R}_+$. Under the map f_z, 0 is mapped to $-\frac{1}{z}$ and ∞ is mapped to $-\frac{1-\frac{1}{\gamma}}{z}$, the two half-lines are hence mapped to paths from $-\frac{1}{z}$ to $-\frac{1-\frac{1}{\gamma}}{z}$. Now under ψ the half-lines will be mapped to paths going $-\frac{1}{z}$ to ∞ because by our assumption $-\frac{1}{z}$ lies in the upper right quadrant, we will show that the image of $\mathbb{R}_+$ under ϕ goes 'above' the origin while the image of $-\frac{1}{z}\mathbb{R}_+$ goes 'under' the origin:

- $\mathbb{R}_+$ is mapped under f_z to the segment $-\frac{1}{z}[1, 1 - \frac{1}{\gamma}]$, as a result, its map under ψ lies in the Minkowski sum $-\frac{1}{z}[1, 1 - \frac{1}{\gamma}] + (-\mathbb{R}_+)$ which is contained in $\overline{\mathbb{C} \setminus \Pi_z}$.

- For any $t \in -\frac{1}{z}\mathbb{R}_+$ we have for all d_i

$$\Im\left(\frac{d_i t}{1 + d_i t}\right) = \Im\left(1 - \frac{1}{1 + d_i t}\right) = \Im\left(\frac{1}{1 + d_i t}\right) \leq 0,$$

  since $\Im(t) \geq 0$. As a result the image of $-\frac{1}{z}\mathbb{R}_+$ under f_z lies in Π_z and its image under ψ lies in the Minkovski sum $\Pi_z + (-\frac{1}{z}\mathbb{R}_+) = \Pi_z$.

Thus we can conclude that $0 \in \psi\left(\mathcal{C}_z\right)$, which shows that there exists at least a fixed point $\tilde{m}$ in $\mathcal{C}_z$.

We observe that, for every $t \in \mathcal{C}_z$, the derivative of f has negative real part:

$$\begin{aligned}
\mathrm{Re}\left(f_z'(t)\right) &= \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N} \mathrm{Re}\left(\frac{d_i}{z\left(1 + d_i t\right)^2}\right) \\
&= \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^{N} \frac{d_i\left[\Re(z) + 2d_i\Re(z)\Re(t) - 2d_i\Im(z)\Im(t) + d_i^2\Re(zt^2)\right]}{|z|^2\left|1 + d_i t\right|^4} \leq 0,
\end{aligned}$$

where we concluded the last inequality by using that $\Re(z) \leq 0$, $\Re(t) \geq 0$, $\Im(z)\Im(t) \geq 0$ and $\Re(zt^2) \leq 0$. Thus, since for no point $t \in \mathcal{C}_z$ has $f_z'(t) = 1$, any fixed point of f_z is a simple fixed point.

We now proceed to show the uniqueness of the fixed point in the region $\mathcal{C}_z$. Suppose there are two fixed points t_1 and t_2, then

$$\begin{aligned}
t_1 - t_2 &= f_z(t_1) - f_z(t_2) \\
&= (t_1 - t_2)\frac{1}{z}\frac{1}{\gamma N}\sum_{i=1}^{N} \frac{d_i}{(1 + d_i t_1)(1 + d_i t_2)}.
\end{aligned}$$

Again, since $\Re(z) \leq 0$, $\Re(t_1), \Re(t_2) \geq 0$, $\Im(z)\Im(t_1), \Im(z)\Im(t_2), \geq 0$ and $\Re(zt_1t_2) \leq 0$, the factor $\frac{1}{z}\frac{1}{N}\sum_{i=1}^{N} \frac{d_i}{(1+d_it_1)(1+d_it_2)}$ has negative real part, and thus the identity is possible only if $t_1 = t_2$. Let's then $\tilde{t}(z)$ be the only fixed point in $\mathcal{C}_z$.

We proceed now to show that $|t - f_z(t)| \geq |t - \tilde{t}(z)|$, i.e. if t and its image are close, then t is not too far from being a fixed point, and so it is close to $\tilde{t}(z)$.

For any $t \in \mathcal{C}_z$, we have

$$\begin{aligned}
|t - f_z(t)| &= |t - \tilde{t}(z) + f_z(\tilde{t}(z)) - \tilde{f}_z(t)| \\
&= \left| (t - \tilde{t}(z)) - \big(t - \tilde{t}(z)\big) \left(\frac{1}{z} \frac{1}{\gamma N} \sum_{i=1}^{N} \frac{d_i}{(1 + d_i t)(1 + d_i \tilde{t}(z))} \right) \right| \\
&= \left| t - \tilde{t}(z) \right| \left| 1 - \frac{1}{z} \frac{1}{\gamma N} \sum_{i=1}^{N} \frac{d_i}{(1 + d_i t)(1 + d_i \tilde{t}(z))} \right| \\
&\geq \left| t - \tilde{t}(z) \right|
\end{aligned}$$

where we have used again that $\frac{1}{z} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{(1+d_i t)(1+d_i \tilde{t}(z))}$ has negative real part.

We provide a lower bound on the norm of the fixed point:

$$\left| \tilde{t}(z) \right| = \frac{1}{|z|} \left| 1 - \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i \tilde{t}(z)}{1 + d_i \tilde{t}(z)} \right| \geq \frac{1}{|z|} \left(1 - \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \left| \frac{d_i \tilde{t}(z)}{1 + d_i \tilde{t}(z)} \right| \right) \geq \frac{1}{|z|} \left(1 - \frac{\left| \tilde{t}(z) \right|}{\gamma N} \sum_{i=1}^{N} d_i \right).$$

hence

$$|\tilde{t}(z)| \geq \left(|z| + \frac{\sum_i d_i}{\gamma N} \right)^{-1}.$$

Finally, note that z can be expressed from the fixed point $\tilde{m}$, hence defining an inverse for the map $\tilde{t}$:

$$\tilde{t}^{-1}(\tilde{m}) = z = -\frac{1}{\tilde{m}} \left(1 - \frac{1}{\gamma} \frac{1}{N} \sum_{i=1}^{N} \frac{d_i \tilde{m}}{1 + d_i \tilde{m}} \right)$$

because the inverse is holomorphic, so is $\tilde{t}$. □

C.3. Ridge

Using Proposition C.1, in order to have a better description of the distribution of the predictor $\hat{f}^{(RF)}_{\lambda,\gamma}$, it remains to study the distributions of both the final labels $\hat{y}$ on the training set and the parameter norm $\|\hat{\theta}\|^2$. In Section C.3.1, we first study the expectation of the final labels $\hat{y}$: this allows us to study the loss of the average predictor $\mathbb{E}\left[\hat{f}^{(RF)}_{\lambda,\gamma}\right]$. Then in Section C.3.3, a study of the variance of the predictor allows us to study the average loss of the RF predictor.

C.3.1. Expectation of the predictor

The optimal parameters $\hat{\theta}$ which minimize the regularized MSE loss is given by $\hat{\theta} = F^T(FF^T + \lambda \mathrm{I}_N)^{-1} y$, or equivalently by $\hat{\theta} = (F^T F + \lambda)^{-1} F^T y$. Thus, the final labels take the form $\hat{y} = A(-\lambda) y$ where $A(z)$ is the random matrix defined as

$$\begin{aligned}
A(z) &:= F \left(F^T F - z \mathrm{I}_P \right)^{-1} F^T \\
&= \frac{1}{P} K^{\frac{1}{2}} W^T \left(\frac{1}{P} W K W^T - z \mathrm{I}_P \right)^{-1} W K^{\frac{1}{2}}.
\end{aligned}$$

Note that the matrix A_λ defined in the proof sketch of Theorem 4.1 in the main text is given by $A_\lambda = A(-\lambda)$.

Proposition C.7. *For any $\gamma > 0$, any $z \in \mathbb{H}_{<0}$, and any symmetric positive definite matrix K,*

$$\|\mathbb{E}\left[A(z)\right] - K(K + \tilde{\lambda}(-z) I_N)^{-1}\|_{op} \leq \frac{c}{P}, \tag{8}$$

where $\tilde{\lambda}(z) := \frac{1}{\tilde{m}(-z)}$ and $c > 0$ depends on z, γ and $\frac{1}{N} Tr(K)$ only.

Proof. Since the distribution of W is invariant under orthogonal transformations, by applying a change of basis, in order to prove Inequality (8), we may assume that K is diagonal with diagonal entries $d_1, \dots, d_N$. Denoting $w_1, \dots, w_N$ the columns of W, for any $i, j = 1, \dots, N$,

$$(A(z))_{ij} = \frac{1}{P}\sqrt{d_i d_j} w_i^T \left(\frac{1}{P}WKW^T - zI_P\right)^{-1} w_j,$$

where $WKW^T = \sum_{i=1}^N d_i w_i w_i^T$. Replacing w_i by $-w_i$ does not change the law W hence does not change the law of $(A(z))_{ij}$. Since WKW^T is invariant under this change of sign, we get that for $i \neq j$, $\mathbb{E}\left[(A(z))_{ij}\right] = -\mathbb{E}\left[(A(z))_{ij}\right]$, hence the off-diagonal terms of $\mathbb{E}\left[A(z)\right]$ vanish.

Consider a diagonal term $(A(z))_{ii}$. From Equation (6), we get

$$(A(z))_{ii} = \frac{d_i}{P} w_i^T B^{-1}(z) w_i = \frac{d_i g_i(z)}{1 + d_i g_i(z)}. \tag{9}$$

By Lemma C.4, g_i lies close to $m_P(z)$ which itself is approximatively equal to $\tilde{m}(z)$ by Proposition C.5. Therefore, we expect $\mathbb{E}\left[(A(z))_{ii}\right] = \mathbb{E}\left[\frac{d_i g_i}{1+d_i g_i}\right]$ to be at short distance from $\frac{d_i \tilde{m}(z)}{1+d_i \tilde{m}(z)}$.

In order to make rigorous this heuristic and to prove that $\mathbb{E}\left[(A(z))_{ii}\right]$ is within $\mathcal{O}(\frac{1}{P})$ distance to $\frac{d_i \tilde{m}(z)}{1+d_i \tilde{m}(z)}$, we consider the first order Taylor approximation $\mathrm{T}_{\tilde{m}(z)} h_i$ of the map $h_i : g \mapsto \frac{1}{1+d_i g}$ (as in the proof Proposition C.5 but this time centered at $\tilde{m}(z)$). Using the fact that $\frac{d_i t}{1+d_i t} = 1 - \frac{1}{1+d_i t} = 1 - h_i(t)$, and inserting the Taylor approximation, $\mathbb{E}\left[(A(z))_{ii}\right] - \frac{d_i \tilde{m}(z)}{1+d_i \tilde{m}(z)}$ is equal to:

$$h_i(\tilde{m}(z)) - h_i(g_i(z)) = \frac{1}{1 + d_i \tilde{m}(z)} - \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z))\right] + \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z)) - h(g_i(z))\right].$$

Thus,

$$\left|\mathbb{E}\left[(A(z))_{ii}\right] - \frac{d_i \tilde{m}(z)}{1 + d_i \tilde{m}(z)}\right| \leq \left|\frac{1}{1 + d_i \tilde{m}(z)} - \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z))\right]\right| + \left|\mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z)) - h(g_i(z))\right]\right|.$$

Using Lemma C.4 and Proposition C.5, the first term $\left|\frac{1}{1+d_i \tilde{m}(z)} - \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z))\right]\right| = \frac{d_i \left|\mathbb{E}[g_i(z)] - \tilde{m}(z)\right|}{|1+d_i \tilde{m}(z)|^2}$ can be bounded by $\frac{\delta}{P} \frac{d_i}{|1+d_i \tilde{m}(z)|^2}$ where δ depends on z, γ and $\frac{1}{N}\mathrm{Tr}(K)$ only. Since $\mathrm{Re}\left[\tilde{m}(z)\right] \geq 0$ thus $|1 + d_i \tilde{m}(z)| \geq \max(1, |d_i \tilde{m}(z)|)$, and $|\tilde{m}(z)| \geq \frac{1}{|z| + \frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K}$ (Lemma C.6), the denominator can be lower bounded:

$$|1 + d_i \tilde{m}(z)|^2 \geq |d_i \tilde{m}(z)| \geq \frac{d_i}{|z| + \frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K},$$

yielding the upper bound:

$$\left|\frac{1}{1 + d_i \tilde{m}(z)} - \mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z))\right]\right| \leq \frac{1}{P}\delta\left[|z| + \frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K\right].$$

For the second term, using the same arguments as for the proof of Proposition C.5, we have:

$$\left|\mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z)) - h(g_i(z))\right]\right| \leq \frac{\mathbb{E}\left[|\tilde{m}(z) - g_i(z)|^2\right]}{|\tilde{m}(z)|^2}.$$

Recall that $|\tilde{m}(z)| \geq \frac{1}{|z| + \frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K}$ and that, by Lemma C.4 and Proposition C.2, $\mathbb{E}\left[|\tilde{m}(z) - g_i(z)|^2\right] \leq \frac{\tilde{\delta}}{P}$ where $\tilde{\delta}$ depends on z, γ and $\frac{1}{N}\mathrm{Tr}(K)$ only. This implies that

$$\left|\mathbb{E}\left[\mathrm{T}_{\tilde{m}(z)} h(g_i(z)) - h(g_i(z))\right]\right| \leq \frac{\tilde{\delta}}{P}\left[|z| + \frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K\right]^2.$$

As a consequence, there exists a constant c which depends on z, γ and $\frac{1}{N}\mathrm{Tr}(K)$ only such that:

$$\left|\mathbb{E}\left[(A(z))_{ii}\right] - \frac{d_i\tilde{m}(z)}{1+d_i\tilde{m}(z)}\right| \leq \frac{c}{P}.$$

Using the effective ridge $\tilde{\lambda}(z) := \frac{1}{\tilde{m}(-z)}$, the term $\frac{d_i\tilde{m}(z)}{1+d_i\tilde{m}(z)} = \frac{d_i}{d_i+\tilde{\lambda}(-z)}$ is equal to $(K(K+\tilde{\lambda}I_N)^{-1})_{ii}$ since, in the basis considered, $K(K+\tilde{\lambda}I_N)^{-1}$ is a diagonal matrix. Hence, we obtain:

$$\left\|\mathbb{E}\left[A(z)\right] - K(K+\tilde{\lambda}I_N)^{-1}\right\|_{op} \leq \frac{c}{P}$$

which allows us to conclude. □

Using the above proposition, we can bound the distance between the expected λ-RF predictor and the $\tilde{\lambda}$-RF predictor.

Theorem C.8. *For $N, P > 0$ and $\lambda > 0$, we have*

$$\left|\mathbb{E}[\hat{f}^{(RF)}_{\lambda,\gamma}(x)] - \hat{f}^{(K)}_{\tilde{\lambda}}(x)\right| \leq \frac{c\sqrt{K(x,x)}\|y\|_{K^{-1}}}{P} \tag{10}$$

where the effective ridge $\tilde{\lambda}(\lambda,\gamma) > \lambda$ is the unique positive number satisfying

$$\tilde{\lambda} = \lambda + \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^{N}\frac{d_i}{\tilde{\lambda}+d_i}, \tag{11}$$

and where $c > 0$ depends on λ, γ, and $\frac{1}{N}\mathrm{Tr}K(X,X)$ only.

Proof. Recall that $\tilde{m}(-\lambda)$ is the unique non negative real such that $\gamma = \frac{1}{N}\sum_{i=1}^{N}\frac{d_i\tilde{m}(-\lambda)}{1+d_i\tilde{m}(-\lambda)} + \gamma\lambda\tilde{m}(-\lambda)$. Dividing this equality by $\gamma\tilde{m}(-\lambda)$ yields Equation (11). From now on, let $\tilde{\lambda} = \tilde{\lambda}(\lambda,\gamma)$.

We now bound the l.h.s. of Equation (10). By Proposition C.1, since $\hat{y} = A(-\lambda)y$, the average λ-RF predictor is $\mathbb{E}\left[f^{(RF)}_{\lambda,\gamma}(x)\right] = K(x,X)K^{-1}\mathbb{E}\left[A(-\lambda)\right]y$. The $\tilde{\lambda}$-KRR predictor is $f^{(K)}_{\tilde{\lambda}}(x) = K(x,X)\left(K+\tilde{\lambda}I_N\right)^{-1}y$. Thus:

$$\left|\mathbb{E}[f^{(RF)}_{\lambda,\gamma}(x)] - f^{(K)}_{\tilde{\lambda}}(x)\right| = \left|K(x,X)K^{-1}\left[\mathbb{E}\left[A(-\lambda)\right] - K\left(K+\tilde{\lambda}I_N\right)^{-1}\right]y\right|.$$

The r.h.s. can be expressed as the absolute value of the scalar product $|\langle w,v\rangle_{K^{-1}}| = \left|v^TK^{-1}w\right|$ where $v = K(x,X)$ and $w = [\mathbb{E}\left[A(-\lambda)\right] - K(K+\tilde{\lambda}I_N)^{-1}]y$. By Cauchy-Schwarz inequality, $|\langle v,w\rangle_{K^{-1}}| \leq \|v\|_{K^{-1}}\|w\|_{K^{-1}}$.

For a general vector v, the K^{-1}-norm $\|v\|_{K^{-1}}$ is equal to the norm mininum Hilbert norm (for the RKHS associated to the kernel K) interpolating function:

$$\|v\|_{K^{-1}} = \min_{f\in\mathcal{H}, f(x_i)=v_i}\|f\|_{\mathcal{H}}.$$

Indeed the minimal interpolating function is the kernel regression given by $f^{(K)}(\cdot) = K(\cdot,X)K(X,X)^{-1}v$ which has norm (writing $\beta = K^{-1}v$):

$$\left\|f^{(K)}\right\|_{\mathcal{H}} = \left\|\sum_{i=1}^{N}\beta_iK(\cdot,x_i)\right\|_{\mathcal{H}} = \sqrt{\sum_{i,j=1}^{N}\beta_i\beta_jK(x_i,x_j)} = \sqrt{v^TK^{-1}KK^{-1}v} = \|v\|_{K^{-1}}.$$

We can now bound the two norms $\|v\|_{K^{-1}}$ and $\|w\|_{K^{-1}}$. For $v = K(x,X)$, we have

$$\|v\|_{K^{-1}} = \min_{f\in\mathcal{H}, f(x_i)=v_i}\|f\|_{\mathcal{H}} \leq \|K(x,\cdot)\|_{\mathcal{H}} = K(x,x)^{\frac{1}{2}}. \tag{12}$$

since $K(x,\cdot)$ is an interpolating function for v.

It remains to bound $\|w\|_{K^{-1}}$. Recall that $K = UDU^T$ with D diagonal, and that, from the previous proposition, $\mathbb{E}\left[A(-\lambda)\right] = UD_AU^T$ where $D_A = \mathrm{diag}\left(\frac{d_1 g_1(-\lambda)}{1+d_1 g_1(-\lambda)}, \dots, \frac{d_N g_N(-\lambda)}{1+d_N g_N(-\lambda)}\right)$. The norm $\|w\|_{K^{-1}}$ is equal to

$$\sqrt{\tilde{y}^T \left[D_A - D\left(D + \tilde{\lambda}(\lambda) I_N\right)^{-1}\right]^T D^{-1} \left[D_A - D\left(D + \tilde{\lambda}(\lambda) I_N\right)^{-1}\right] \tilde{y}},$$

where $\tilde{y} = U^T y$. Expanding the product, $\|w\|_{K^{-1}} = \sqrt{\sum_{i=1}^N \frac{\tilde{y}_i^2}{d_i}\left((D_A)_{ii} - \frac{d_i}{\tilde{\lambda}(\lambda)+d_i}\right)^2}$, hence by Proposition C.7, $\|w\|_{K^{-1}} \leq \frac{c}{P}\sqrt{\sum_{i=1}^N \frac{\tilde{y}^2}{d_i}}$. The result follows from noticing that $\sum_{i=1}^N \frac{\tilde{y}^2}{d_i} = \tilde{y}^T D^{-1} \tilde{y} = \|y\|_{K^{-1}}^2$:

$$\left|\mathbb{E}[f_{\lambda,\gamma}^{(RF)}(x)] - f_{\tilde{\lambda}}^{(K)}(x)\right| \leq \|v\|_{K^{-1}} \|w\|_{K^{-1}} \leq \frac{cK(x,x)^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}.$$

which allows us to conclude. □

Corollary C.9. *If $\mathbb{E}_{\mathcal{D}}[K(x,x)] < \infty$, we have that the difference of errors $\delta_E = \left|L(\mathbb{E}[\hat{f}_{\lambda,\gamma}^{(RF)}]) - L(\hat{f}_{\tilde{\lambda}}^{(K)})\right|$ is bounded from above by*

$$\delta_E \leq \frac{C\|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}_{\tilde{\lambda}}^{(K)}\right)} + \frac{C\|y\|_{K^{-1}}}{P}\right),$$

where C is given by $c\sqrt{\mathbb{E}_{\mathcal{D}}[K(x,x)]}$, with c the constant appearing in (10) above.

Proof. For any function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\|f\| = \left(\mathbb{E}_{\mathcal{D}}\left[f(x)^2\right]\right)^{\frac{1}{2}}$ its $L^2(\mathcal{D})$-norm. Integrating $\left|\mathbb{E}[f_{\lambda,\gamma}^{(RF)}(x)] - f_{\tilde{\lambda}}^{(K)}(x)\right|^2 \leq \frac{c^2 K(x,x)\|y\|_{K^{-1}}^2}{P^2}$ over $x \sim \mathcal{D}$, we get the following bound:

$$\|\mathbb{E}[f_{\lambda,\gamma}^{(RF)}] - f_{\tilde{\lambda}}^{(K)}\| \leq \frac{c\left[\mathbb{E}_{\mathcal{D}}\left[K(x,x)\right]\right]^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}.$$

Hence, if f^* is the true function, by the triangular inequality,

$$\left|\|\mathbb{E}[f_{\lambda,\gamma}^{(RF)}] - f^*\| - \|f_{\tilde{\lambda}}^{(K)} - f^*\|\right| \leq \frac{c\left[\mathbb{E}_{\mathcal{D}}\left[K(x,x)\right]\right]^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}.$$

Notice that $L(\mathbb{E}[\hat{f}_{\gamma,\lambda}^{(RF)}]) = \|\mathbb{E}[f_{\lambda,\gamma}^{(RF)}] - f^*\|^2$ and $L(\hat{f}_{\tilde{\lambda}}^{(K)}) = \|f_{\tilde{\lambda}}^{(K)} - f^*\|^2$. Since $\left|a^2 - b^2\right| \leq |a-b|\left(|a-b| + 2|b|\right)$, we obtain

$$\left|L\left(\mathbb{E}[\hat{f}_{\gamma,\lambda}^{(RF)}]\right) - L\left(\hat{f}_{\tilde{\lambda}}^{(K)}\right)\right| \leq \frac{c\left[\mathbb{E}_{\mathcal{D}}\left[K(x,x)\right]\right]^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}_{\tilde{\lambda}}^{(K)}\right)} + \frac{c\left[\mathbb{E}_{\mathcal{D}}\left[K(x,x)\right]\right]^{\frac{1}{2}} \|y\|_{K^{-1}}}{P}\right),$$

which allows us to conclude. □

C.3.2. Properties of the effective ridge

Thanks to the implicit definition of the effective ridge $\tilde{\lambda}$, we obtain the following:

Proposition C.10. *The effective ridge $\tilde{\lambda}$ satisfies the following properties:*

1. *for any $\gamma > 0$, we have $\lambda < \tilde{\lambda}(\lambda,\gamma) \leq \lambda + \frac{1}{\gamma}T$;*
2. *the function $\gamma \mapsto \tilde{\lambda}(\lambda,\gamma)$ is decreasing;*
3. *for $\gamma > 1$, we have $\tilde{\lambda} \leq \frac{\gamma}{\gamma-1}\lambda$;*
4. *for $\gamma < 1$, we have $\tilde{\lambda} \geq \frac{1-\sqrt{\gamma}}{\sqrt{\gamma}} \min_i d_i$.*

Proof. **(1)** The upper bound in the first statement follows directly from Lemma C.6 where it was shown that $\tilde{m}(-\lambda) \geq \frac{1}{\lambda+\frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}K}$ and from the fact that $\tilde{\lambda}(\lambda,\gamma) = \frac{1}{\tilde{m}(-\lambda)}$. For the lower bound, remark that Equation (11) can be written as:

$$\tilde{\lambda}(\lambda,\gamma) = \lambda + \frac{1}{\gamma}\frac{1}{N}\mathrm{Tr}[\tilde{\lambda}(\lambda,\gamma)K(\tilde{\lambda}(\lambda,\gamma)I_N + K)^{-1}].$$

Since $\tilde{\lambda}(\lambda,\gamma) \geq 0$ and K is a positive symmetric matrix, $\mathrm{Tr}[K[\tilde{\lambda}(\lambda,\gamma)I_N + K]^{-1}] \geq 0$: this yields $\tilde{\lambda}(\lambda,\gamma) \geq \lambda$.

(2) We show that $\gamma \mapsto \tilde{\lambda}(\lambda,\gamma)$ is decreasing by computing the derivative of the effective ridge with respect to γ. Differentiating both sides of Equation (11), $\partial_\gamma \tilde{\lambda} = \partial_\gamma \left[\lambda + \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i}\right]$. The r.h.s. is equal to:

$$\frac{\partial_\gamma \tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i} - \frac{\tilde{\lambda}}{\gamma^2}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i} - \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i \partial_\gamma \tilde{\lambda}}{(\tilde{\lambda}+d_i)^2}.$$

Using Equation (11), $\frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i} = \frac{\tilde{\lambda}-\lambda}{\tilde{\lambda}}$ and thus:

$$\partial_\gamma \tilde{\lambda}\left[\frac{\lambda}{\tilde{\lambda}} + \frac{\tilde{\lambda}}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\left(\tilde{\lambda}+d_i\right)^2}\right] = -\frac{\tilde{\lambda}-\lambda}{\gamma}.$$

Since $\tilde{\lambda} \geq \lambda \geq 0$, the derivative of the effective ridge with respect to γ is negative: the function $\gamma \mapsto \tilde{\lambda}(\lambda,\gamma)$ is decreasing.

(3) Using the bound $\frac{d_i}{\tilde{\lambda}+d_i} \leq 1$ in Equation (11), we obtain $\tilde{\lambda} \leq \lambda + \frac{\tilde{\lambda}}{\gamma}$ which, when $\gamma \geq 1$, implies that $\tilde{\lambda} \leq \lambda\frac{\gamma}{\gamma-1}$.

(4) Recall that $\lambda > 0$ and that the effective ridge $\tilde{\lambda}$ is the unique fixpoint of the map $f(t) = \lambda + \frac{t}{\gamma}\frac{1}{N}\sum_i \frac{d_i}{t+d_i}$ in $\mathbb{R}_+$. The map is concave and, at $t=0$, we have $f(t) = \lambda > 0 = t$: this implies that $f'(\tilde{\lambda}) < 1$ otherwise by concavity, for any $t \leq \tilde{\lambda}$ one would have $f(t) \leq t$. The derivative of f is $f'(t) = \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i^2}{(t+d_i)^2}$, thus $\frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i^2}{\left(\tilde{\lambda}+d_i\right)^2} < 1$. Using the fact that d_0 is the smallest eigenvalue of $K(X,X)$, i.e. $d_i \geq d_0$, we get $1 > \frac{1}{\gamma}\frac{d_0^2}{\left(\tilde{\lambda}+d_0\right)^2}$ hence $\tilde{\lambda} \geq d_0\frac{1-\sqrt{\gamma}}{\sqrt{\gamma}}$. □

Similarily, we gather a number of properties of the derivative $\partial_\lambda \tilde{\lambda}(\lambda,\gamma)$.

Proposition C.11. *For $\gamma > 1$, as $\lambda \to 0$, the derivative $\partial_\lambda \tilde{\lambda}$ converges to $\frac{\gamma}{\gamma-1}$. As $\lambda\gamma \to \infty$, we have $\partial_\lambda \tilde{\lambda}(\lambda,\gamma) \to 1$.*

Proof. Differentiating both sides of Equation (11),

$$\partial_\lambda \tilde{\lambda} = 1 + \partial_\lambda \tilde{\lambda}\frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i} - \tilde{\lambda}\partial_\lambda \tilde{\lambda}\frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{(\tilde{\lambda}+d_i)^2}.$$

Hence the derivative $\partial_\lambda \tilde{\lambda}$ satisfies the following equality

$$\partial_\lambda \tilde{\lambda}\left(1 - \frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\tilde{\lambda}+d_i} + \tilde{\lambda}\frac{1}{\gamma}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{(\tilde{\lambda}+d_i)^2}\right) = 1. \tag{13}$$

(1) Assuming $\gamma > 1$, from the point 3. of Proposition C.10, we already know that $\tilde{\lambda}(\lambda,\gamma) \leq \lambda\frac{\gamma}{\gamma-1}$ hence $\tilde{\lambda}(0,\gamma) = 0$. Actually, using similar arguments as in the proof of point 3., this holds also for $\gamma = 1$. Using the fact that $\tilde{\lambda}(0,\gamma) = 0$, we get $\partial_\lambda \tilde{\lambda}(0,\gamma) = 1 + \frac{\partial_\lambda \tilde{\lambda}(0,\gamma)}{\gamma}$, hence $\partial_\lambda \tilde{\lambda}(0,\gamma) = \frac{\gamma}{\gamma-1}$.

(2) From the first point of Proposition C.10, $\tilde{\lambda} \sim \lambda$ as $\lambda\gamma \to \infty$. Since Equation (13) can be expressed as:

$$\partial_\lambda \tilde{\lambda}\left(1 - \frac{1}{\gamma\lambda}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{\frac{\tilde{\lambda}}{\lambda}+d_i} + \frac{1}{\gamma\lambda}\frac{\tilde{\lambda}}{\lambda}\frac{1}{N}\sum_{i=1}^N \frac{d_i}{(\frac{\tilde{\lambda}}{\lambda}+d_i)^2}\right) = 1,$$

we obtain that $\partial_\lambda \tilde{\lambda} \to 1$ as $\lambda \to \infty$. □

C.3.3. Variance of the predictor

By the bias-variance decomposition, in order to bound the difference between $\mathbb{E}[L(\hat{f}^{(RF)}_{\gamma,\lambda})]$ and $L(\hat{f}^{(K)}_{\tilde{\lambda}})$, we have to bound $\mathbb{E}_{\mathcal{D}}[\mathrm{Var}(f(x))]$. The law of total variance yields $\mathrm{Var}(\hat{f}(x)) = \mathrm{Var}(\mathbb{E}[\hat{f}(x)|F]) + \mathbb{E}[\mathrm{Var}[\hat{f}(x)|F]]$. By Proposition C.1, we have $\mathbb{E}[\hat{f}(x)|F] = K(x,X)K(X,X)^{-1}\hat{y}$ and $\mathrm{Var}[\hat{f}(x)|F] = \frac{1}{P}\|\hat{\theta}\|^2\tilde{K}(x,x)$. Hence, it remains to study $\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right)$ and $\mathbb{E}[\|\hat{\theta}\|^2]$. Recall that we denote $T = \frac{1}{N}\mathrm{Tr}K(X,X)$.

This section is dedicated to the proof of the variance bound of Theorem 5.1 of the paper:

Theorem 5.1 *There are constants $c_1, c_2 > 0$ depending on λ, γ, T only such that*

$$\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) \leq \frac{c_1 K(x,x)\|y\|^2_{K^{-1}}}{P}$$
$$\left|\mathbb{E}\|[\hat{\theta}\|^2] - \partial_\lambda\tilde{\lambda}y^T M_{\tilde{\lambda}}y\right| \leq \frac{c_2\|y\|^2_{K^{-1}}}{P},$$

where $\partial_\lambda\tilde{\lambda}$ is the derivative of $\tilde{\lambda}$ with respect to λ and for $M_{\tilde{\lambda}} = K(X,X)(K(X,X)+\tilde{\lambda}I_N)^{-2}$. As a result

$$\mathrm{Var}\left(\hat{f}^{(RF)}_{\lambda}(x)\right) \leq \frac{c_3 K(x,x)\|y\|^2_{K^{-1}}}{P},$$

where $c_3 > 0$ depends on λ, γ, T.

• **Bound on** $\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right)$**.** We first study the covariance of the entries of the matrix

$$A_\lambda = \frac{1}{P}K^{\frac{1}{2}}W^T\left(\frac{1}{P}WKW^T + \lambda \mathrm{I}_P\right)^{-1}WK^{\frac{1}{2}},$$

where $K = \mathrm{diag}(d_1, \dots, d_N)$ is a positive definite diagonal matrix and W is a $P \times N$ matrix with i.i.d. Gaussian entries. In the next proposition we show a $\frac{c_1}{P}$ bound for the covariance of the entries of A_λ, then we exploit this result in order to prove the bound on the variance of $K(x,X)K(X,X)^{-1}\hat{y}$.

Proposition C.12. *There exists a constant $c_1' > 0$ depending on λ, γ, and $\frac{1}{N}\mathrm{Tr}(K)$ only, such that the following bounds hold:*

$$|\mathrm{Cov}\left((A_\lambda)_{ii}, (A_\lambda)_{jj}\right)| \leq \frac{c_1'}{P}$$
$$\mathrm{Var}\left((A_\lambda)_{ij}\right) \leq \min\left\{\frac{d_i}{d_j}, \frac{d_j}{d_i}\right\}\frac{c_1'}{P}.$$

For all other cases (i.e. if i,j, k and l take more than two different values), $\mathrm{Cov}\left((A_\lambda)_{ij}, (A_\lambda)_{kl}\right) = 0$.

Proof. We want to study the covariances $\mathrm{Cov}\left((A_\lambda)_{ij}, (A_\lambda)_{kl}\right)$ for any i, j, k, l. Using the same symmetry argument as in the proof of Proposition C.7, $\mathbb{E}\left[(A_\lambda)_{ij}(A_\lambda)_{kl}\right] = 0$ whenever each value in $\{i, j, k, l\}$ does not appear an even number of times in (i, j, k, l). Using the fact that A_λ is symmetric, it remains to study $\mathrm{Cov}\left((A_\lambda)_{ii}, (A_\lambda)_{jj}\right)$, $\mathrm{Var}\left((A_\lambda)_{ii}\right)$ and $\mathrm{Var}\left[(A_\lambda)_{ij}\right]$ for all $i \neq j$. By the Cauchy-Schwarz inequality, any bound on $\mathrm{Var}\left((A_\lambda)_{ii}\right)$ will imply a similar bound on $\mathrm{Cov}\left((A_\lambda)_{ii}, (A_\lambda)_{jj}\right)$. Besides, as we have seen in the proof of Proposition C.7, $\mathbb{E}\left[(A_\lambda)_{ij}\right] = 0$ for any $i \neq j$. Thus, we only have to study $\mathrm{Var}\left((A_\lambda)_{ii}\right)$ and $\mathbb{E}\left[(A_\lambda)^2_{ij}\right]$.

• Bound on $\mathrm{Var}\left((A_\lambda)_{ii}\right)$: From Equation (9),

$$\mathrm{Var}\left((A_\lambda)_{ii}\right) = \mathrm{Var}\left(\frac{d_i g_i}{1+d_i g_i}\right) = \mathrm{Var}\left(1 - \frac{1}{1+d_i g_i}\right) = \mathrm{Var}\left(\frac{1}{1+d_i g_i}\right) \leq \mathbb{E}\left[\left(\frac{1}{1+d_i g_i} - \frac{1}{1+d_i\tilde{m}}\right)^2\right],$$

where $g_i := g_i(-\lambda)$. Again, we use the first order Taylor approximation $\mathrm{T}h$ of $h : x \to \frac{1}{1+d_i x}$ centered at $\tilde{m} := \tilde{m}(-\lambda)$, as well as the bound (7), to obtain

$$\mathbb{E}\left[\left(\frac{1}{1+d_i g_i}-\frac{1}{1+d_i\tilde{m}}\right)^2\right]=\mathbb{E}\left[\left(-\frac{d_i}{(1+d_i\tilde{m})^2}(g_i-\tilde{m})+h(g_i)-\mathrm{T}h(g_i)\right)^2\right]$$
$$\leq\frac{2d_i^2}{(1+d_i\tilde{m})^4}\mathbb{E}\left[(g_i-\tilde{m})^2\right]+2\mathbb{E}\left[(h(g_i)-\mathrm{T}h(g_i))^2\right]$$
$$\leq\frac{2}{6\tilde{m}^2}\mathbb{E}\left[(g_i-\tilde{m})^2\right]+\frac{2}{\tilde{m}^4}\mathbb{E}\left[(g_i-\tilde{m})^4\right].$$

Using Lemma C.4, we get $\mathrm{Var}\left((A_\lambda)_{ii}\right)\leq\frac{c_1'}{P}$, where $c_1'>0$ depends on λ, γ, and $\frac{1}{N}\mathrm{Tr}(K)$ only.

• Bound on $\mathbb{E}\left((A_\lambda)_{ij}\right)$ for $i\neq j$: Following the same arguments as for Equation (9), $(A_\lambda)_{ij}$ is equal to

$$(A_\lambda)_{ij}=\frac{\sqrt{d_id_j}}{P}\left[w_i^TB_{(i)}^{-1}w_j-\frac{d_ig_i}{1+d_ig_i}w_i^TB_{(i)}^{-1}w_j\right]=\frac{\sqrt{d_id_j}}{1+d_ig_i}\frac{1}{P}w_i^TB_{(i)}^{-1}w_j,$$

where we set $B_{(i)}:=B_i(-\lambda)$. Since w_i and $B_{(i)}$ are independent, $\mathbb{E}\left[\left(w_i^TB_{(i)}^{-1}w_j\right)^2\right]=\mathbb{E}\left[w_j^TB_{(i)}^{-2}w_j\right]$, and thus, by the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[(A_\lambda)_{ij}^2\right]\leq\frac{1}{P^2}\sqrt{\mathbb{E}\left[\frac{d_i^2d_j^2}{(1+d_ig_i)^4}\right]}\sqrt{\mathbb{E}\left[\left(w_j^TB_{(i)}^{-2}w_j\right)^2\right]}. \tag{14}$$

Recall that $\tilde{m}:=\tilde{m}(-\lambda)$. Using the fact that $\frac{1}{1+d_ig_i}=\frac{1}{1+d_i\tilde{m}}+\frac{1}{1+d_ig_i}-\frac{1}{1+d_i\tilde{m}}$ and inserting the first Taylor approximation $\mathrm{T}h$ of $h:x\to\frac{1}{1+d_ix}$ centered at $\tilde{m}$, we get:

$$\mathbb{E}\left[\left(\frac{1}{1+d_ig_i}\right)^4\right]=\mathbb{E}\left[\left(\frac{1}{1+d_i\tilde{m}}-\frac{d_i}{(1+d_i\tilde{m})^2}(g_i-\tilde{m})+h(g_i)-\mathrm{T}h(g_i)\right)^4\right].$$

Using a convexity argument, the bound (7), and the lower bound on $\tilde{m}$ given by Lemma C.6, there exists three constants $\tilde{c}_1$, $\tilde{c}_2$, $\tilde{c}_3$, which depend on λ, γ and $\frac{1}{N}\mathrm{Tr}(K)$ only, such that $\mathbb{E}\left[\left(\frac{1}{1+d_ig_i}\right)^4\right]$ is bounded by

$$\frac{\tilde{c}_1}{(1+d_i\tilde{m})^4}+\frac{\tilde{c}_2d_i^4}{(1+d_i\tilde{m})^8}\mathbb{E}\left[(g_i-\tilde{m})^4\right]+\tilde{c}_3\mathbb{E}\left[(g_i-\tilde{m})^8\right].$$

Thanks to Lemma C.4 and Proposition C.5, this last expression can be bounded by an expression of the form $\frac{\tilde{e}_1}{d_i^4}+\frac{\tilde{e}_2}{P^2d_i^4}+\frac{\tilde{e}_3}{P^4}$. Note that $\frac{\tilde{e}_2}{P^2d_i^4}\leq\frac{\tilde{e}_2}{d_i^4}$ and $\frac{\tilde{e}_3}{P^4}\leq\frac{\tilde{e}_3}{\gamma^4}\frac{(\frac{1}{N}\mathrm{Tr}(K))^4}{d_i^4}$. Hence, we obtain the bound:

$$\mathbb{E}\left[\left(\frac{1}{1+d_ig_i}\right)^4\right]\leq\frac{\tilde{c}}{d_i^4},$$

where $\tilde{c}=\tilde{e}_1+\tilde{e}_2+\frac{\tilde{e}_3(\frac{1}{N}\mathrm{Tr}(K))^4)}{\gamma^4}$ depends on λ, γ and and $\frac{1}{N}\mathrm{Tr}(K)$ only.

Let us now consider the second term in the r.h.s. of (14) . Using the fact that $\|B_{(i)}\|_{op}\geq\frac{1}{\lambda}$, we get

$$\sqrt{\mathbb{E}\left[\left(w_j^TB_{(i)}^{-2}w_j\right)^2\right]}\leq\sqrt{\frac{1}{\lambda^4}\mathbb{E}\left[\left(w_j^Tw_j\right)^2\right]}=\sqrt{\frac{1}{\lambda^4}N(N+2)}\leq\frac{N+1}{\lambda^2},$$

where we have used the fact that the second moment of a $\chi^2(N)$ distribution is $N(N+2)$. Together, we obtain

$$\begin{aligned}\mathbb{E}\left[(A)_{ij}^2\right] &\leq \frac{1}{P^2}\sqrt{\mathbb{E}\left[\frac{d_i^2 d_j^2}{(1+d_i g_i)^4}\right]}\sqrt{\mathbb{E}\left[\left(w_j^T B_{(i)}^{-2} w_j\right)^2\right]} \\ &\leq \frac{\tilde{c} d_i d_j}{d_i^2}\frac{N+1}{P^2\lambda^2} \\ &\leq \frac{\tilde{c} d_j}{P d_i \lambda^2 \gamma}\frac{N+1}{N} \leq \frac{c_1'}{P}\frac{d_i}{d_j},\end{aligned}$$

for $c_1' = 2\frac{\tilde{c}}{\lambda^2\gamma}$. Since the matrix A_λ is symmetric, we finally conclude that

$$\mathbb{E}\left[(A_\lambda)_{ij}^2\right] \leq \frac{c_1'}{P}\min\left\{\frac{d_i}{d_j}, \frac{d_j}{d_i}\right\}.$$

Note that c_1' is a constant related to the bounds constructed in Lemma C.2 and Proposition C.5 and as such it depends on $\frac{1}{N}\mathrm{Tr}(K)$, γ and λ only. □

Proposition C.13. *There exists a constant $c_1 > 0$ (depending on λ, γ, T only) such that the variance of the estimator is bounded by*

$$\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) \leq \frac{c_1\|y\|_{K^{-1}}^2 K(x,x)}{P}.$$

Proof. As in the proof of Theorem C.8, with the right change of basis, we may assume the Gram matrix $K(X,X)$ to be diagonal.

We first express the covariances of $\hat{y} = A(-\lambda)y$. Using Proposition Proposition C.12, for $i \neq j$ we have

$$\mathrm{Cov}\left(\hat{y}_i, \hat{y}_j\right) = \sum_{k,l=1}^{N}\mathrm{Cov}\left((A_\lambda)_{ik}, (A_\lambda)_{lj}\right) y_k y_l = \mathrm{Cov}\left((A_\lambda)_{ii}, (A_\lambda)_{jj}\right) y_i y_j + \mathbb{E}\left[(A_\lambda)_{ij}^2\right] y_j y_i,$$

whereas for $i = j$ we have

$$\mathrm{Cov}\left(\hat{y}_i, \hat{y}_i\right) = \sum_{k=1}^{N}\mathrm{Cov}\left((A_\lambda)_{ik}, (A_\lambda)_{ki}\right) y_k^2 = \mathrm{Var}\left((A_\lambda)_{ii}\right) y_i^2 + \sum_{k\neq i}\mathbb{E}\left[(A_\lambda)_{ik}^2\right] y_k^2.$$

We decompose $K^{-\frac{1}{2}}\mathrm{Cov}(\hat{y},\hat{y})K^{-\frac{1}{2}}$ into two terms: let C be the matrix of entries

$$C_{ij} = \frac{\mathrm{Cov}((A_\lambda)_{ii}, (A_\lambda)_{jj}) + \delta_{i\neq j}\mathbb{E}\left[(A_\lambda)_{ij}^2\right]}{\sqrt{d_i d_j}} y_i y_j,$$

and let D the diagonal matrix with entries

$$D_{ii} = \frac{\sum_{k\neq i}\mathbb{E}\left[(A_\lambda)_{ik}^2\right] y_k^2}{d_i}.$$

We have the decomposition $K^{-\frac{1}{2}}\mathrm{Cov}(\hat{y},\hat{y})K^{-\frac{1}{2}} = C + D$.

Proposition C.12 asserts that $\mathrm{Cov}((A_\lambda)_{ii}, (A_\lambda)_{jj} \leq \frac{c_1'}{P}$ and $\mathbb{E}\left[(A_\lambda)_{ij}^2\right] \leq \frac{c_1'}{P}$, and thus the operator norm of C is bounded by

$$\begin{aligned}\|C\|_{op} &\leq \|C\|_F \\ &= \sqrt{\sum_{i,j}\frac{\left(\mathrm{Cov}((A_\lambda)_{ii}, (A_\lambda)_{jj}) + \delta_{i\neq j}\mathbb{E}\left[(A_\lambda)_{ij}^2\right]\right)^2}{d_i d_j} y_i^2 y_j^2} \\ &\leq \frac{2c_1'}{P}\sqrt{\sum_{ij}\frac{1}{d_i d_j} y_i^2 y_j^2} \;=\; \frac{2c_1'\|y\|_{K^{-1}}^2}{P}\end{aligned}$$

For the matrix D, we use the bound $\mathbb{E}\left[(A_\lambda)_{ik}^2\right] \leq \frac{c_1'}{P}\frac{d_i}{d_k}$ to obtain

$$D_{ii} = \frac{\sum_{k\neq i}\mathbb{E}\left[(A_\lambda)_{ik}^2\right]y_k^2}{d_i} \leq \frac{c_1'}{P}\sum_{k\neq i}\frac{y_k^2}{d_k} \leq \frac{c_1'\|y\|_{K^{-1}}^2}{P},$$

which implies that $\|D\|_{op} \leq \frac{c_1'\|y\|_{K-1}^2}{P}$. As a result

$$\begin{aligned}\mathrm{Var}\left(K(x,X)K^{-1}\hat{y}\right) &= K(x,X)K^{-1}\mathrm{Cov}(\hat{y},\hat{y})K^{-1}K(X,x)\\ &\leq K(x,X)K^{-\frac{1}{2}}\|C+D\|_{op}K^{-\frac{1}{2}}K(X,x)\\ &\leq \frac{3c_1'\|y\|_{K^{-1}}^2}{P}\|K(x,X)\|_{K^{-1}}^2\\ &\leq \frac{3c_1'K(x,x)\|y\|_{K^{-1}}^2}{P},\end{aligned}$$

where we used Inequality (12). This yields the result with $c_1 = 3c_1'$. □

• **Bound on $\mathbb{E}_\pi\left[\|\hat{\theta}\|^2\right]$.** To understand the variance of the λ-RF estimator $\hat{f}_\lambda^{(RF)}$, we need to describe the distribution of the squared norm of the parameters:

Proposition C.14. *For $\gamma, \lambda > 0$ there exists a constant $c_2 > 0$ depending on λ, γ, T only such that*

$$\left|\mathbb{E}[\|\hat{\theta}\|^2] - \partial_\lambda\tilde{\lambda}y^T K(X,X)\left(K(X,X)+\tilde{\lambda}I_N\right)^{-2}y\right| \leq \frac{c_2\|y\|_{K^{-1}}^2}{P}. \tag{15}$$

Proof. As in the proof of Theorem C.8, with the right change of basis, we may assume the Gram matrix $K(X,X)$ to be diagonal. Recall that $\hat{\theta} = \frac{1}{\sqrt{P}}\left(\frac{1}{P}WK(X,X)W^T + \lambda I_N\right)^{-1}WK(X,X)^{\frac{1}{2}}y$, thus we have:

$$\|\hat{\theta}\|^2 \;=\; \frac{1}{P}y^T K(X,X)^{\frac{1}{2}}W^T(\frac{1}{P}WK(X,X)W^T + \lambda I_P)^{-2}WK(X,X)^{\frac{1}{2}}y \;=\; y^T A'(-\lambda)y, \tag{16}$$

where $A'(-\lambda)$ is the derivative of

$$A(z) = \frac{1}{P}K(X,X)^{\frac{1}{2}}W^T\left(\frac{1}{P}WK(X,X)W^T - z\mathrm{I}_P\right)^{-1}WK(X,X)^{\frac{1}{2}}$$

with respect to z evaluated at $-\lambda$. Let

$$\tilde{A}(z) = K(X,X)(K(X,X)+\tilde{\lambda}(-z)\mathrm{I}_N)^{-1}.$$

Remark that the derivative of $\tilde{A}(z)$ is given by $\tilde{A}'(z) = \tilde{\lambda}'(-z)K(X,X)(K(X,X)+\tilde{\lambda}(-z)I_N)^{-2}$. Thus, from Equation (16), the l.h.s. of (15) is equal to:

$$\left|y^T\left(\mathbb{E}[A'(-\lambda)] - \tilde{A}'(-\lambda)\right)y\right|. \tag{17}$$

Using a classical complex analysis argument, we will show that $\mathbb{E}[A'(-\lambda)]$ is close to $\tilde{A}'(-\lambda)$ by proving a bound of the difference between $\mathbb{E}[A(z)]$ and $\tilde{A}(z)$ for any $z \in \mathbb{H}_{<0}$.

Note that the proof of Proposition C.7 provides a bound on the diagonal entries of $\mathbb{E}[A(z)]$, namely that for any $z \in \mathbb{H}_{<0}$,

$$\left|\mathbb{E}[(A(z))_{ii}] - (\tilde{A}(z))_{ii}\right| \leq \frac{c}{P},$$

where $\hat{c}$ depends on z, γ and T only. Actually, in order to prove (15), we will derive the following slightly different bound: for any $z \in \mathbb{H}_{<0}$,

$$\left|\mathbb{E}[(A(z))_{ii}] - (\tilde{A}(z))_{ii}\right| \leq \frac{\hat{c}}{d_iP}, \tag{18}$$

where $\hat{c}$ depends on z, γ and T only. Let $g_i := g_i(z)$ and $\tilde{m} := \tilde{m}(z)$. Recall that for $h_i : x \mapsto \frac{d_i x}{1+d_i x}$, one has $(A(z))_{ii} = h_i(g_i)$, $(\tilde{A}(z))_{ii} = h_i(\tilde{m})$ and

$$\mathrm{T}_{\tilde{m}} h_i(g_i) = \frac{d_i \tilde{m}}{1+d_i\tilde{m}} - \frac{d_i\,(g_i - \tilde{m})}{(1+d_i\tilde{m})^2},$$

$$h_i(g_i) - \mathrm{T}_{\tilde{m}} h_i(g_i) = \frac{d_i^2\,(g_i-\tilde{m})^2}{(1+d_i g_i)\,(1+d_i\tilde{m})^2},$$

where $\mathrm{T}_{\tilde{m}} h_i$ is the first order Taylor approximation of h_i centered at $\tilde{m}$. Using this first order Taylor approximation, we can bound the difference $|\mathbb{E}[h_i(g_i)] - h_i(\tilde{m})|$:

$$\begin{aligned}|\mathbb{E}[h_i(g_i)] - h_i(\tilde{m})| &\leq \frac{d_i\,|\mathbb{E}[g_i]-\tilde{m}|}{(1+d_i\tilde{m})^2} + \frac{d_i^2}{(1+d_i\tilde{m})^2}\mathbb{E}\left[\frac{|g_i-\tilde{m}|^2}{1+d_i g_i}\right] \\ &\leq \frac{\mathbf{a}}{d_i P} + \mathbf{a}\sqrt{\mathbb{E}\left[\frac{1}{(1+d_i g_i)^2}\right]\mathbb{E}\left[|g_i-\tilde{m}|^4\right]},\end{aligned}$$

where $\mathbf{a}$ depends on z, γ and T. We need to bound $\mathbb{E}\left[\frac{1}{(1+d_i g_i)^2}\right]$. Recall that in the proof of Proposition C.12, we bounded $\mathbb{E}\left[\frac{1}{(1+d_i g_i)^4}\right]$. Using similar arguments, one shows that

$$\mathbb{E}\left[\frac{1}{(1+d_i g_i)^2}\right] \leq \frac{\hat{e}^2}{d_i^2},$$

where $\hat{e}$ depends on z, γ and $\frac{1}{N}\mathrm{Tr}(K(X,X))$ only. The term $\mathbb{E}\left[|g_i-\tilde{m}|^4\right]$ is bounded using Lemmas C.4, C.2 and Proposition C.5. This allows us to conclude that:

$$|\mathbb{E}[h_i(g_i)] - h_i(\tilde{m})| \leq \frac{\hat{c}}{d_i P},$$

where $\hat{c}$ depends on z, γ and $\frac{1}{N}\mathrm{Tr}(K(X,X))$ only, hence we obtain the Inequality (18).

We can now prove Inequality 15. We bound the difference of the derivatives of the diagonal terms of $A(z)$ and $\tilde{A}(z)$ by means of Cauchy formula. Consider a simple closed path $\phi : [0,1] \to \mathbb{H}_{<0}$ which surrounds z. Since

$$\mathbb{E}[(A'(z))_{ii}] - (\tilde{A}'(z))_{ii} = \frac{1}{2\pi i}\oint_{\phi} \frac{\mathbb{E}[(A(z))_{ii}] - (\tilde{A}(z))_{ii}}{(w-z)^2}dw,$$

using the bound (18), we have:

$$\left|\mathbb{E}[(A'(z))_{ii}] - (\tilde{A}'(z))_{ii}\right| \leq \frac{\hat{c}}{d_i P}\frac{1}{2\pi}\oint_{\phi}\frac{1}{|w-z|^2}dw \leq \frac{c_2}{d_i P},$$

where c_2 depends on z, γ, and T only. This allows one to bound the operator norm of $K(X,X)(\mathbb{E}[A'(z)] - \tilde{A}'(z))$:

$$\|K(X,X)(\mathbb{E}[A'(z)] - \tilde{A}'(z))\|_{op} \leq \frac{c_2}{P}.$$

Using this bound and (17), we have

$$\left|\mathbb{E}[\|\hat{\theta}\|^2] - \partial_\lambda\tilde{\lambda}\; y^T K(X,X)\left(K(X,X)+\tilde{\lambda}I_N\right)^{-2} y\right| = \left|y^T\left(\mathbb{E}[A'(-\lambda)] - \tilde{A}'(-\lambda)\right)y\right| \leq \frac{c_2\|y\|^2_{K^{-1}}}{P},$$

which allows us to conclude. □

• **Bound on** $\mathrm{Var}\left(\hat{f}^{(RF)}_\lambda(x)\right)$**.** We have shown all the bounds needed in order to prove the following proposition.

Proposition C.15. *For any $x \in \mathbb{R}^d$, we have*

$$\mathrm{Var}\left(\hat{f}_\lambda^{(RF)}(x)\right) \leq \frac{c_3 K(x,x)\|y\|_{K^{-1}}^2}{P},$$

where $c_3 > 0$ depends on λ, γ, T.

Proof. Recall that for any $x \in \mathbb{R}^d$,

$$\begin{aligned}\mathrm{Var}(\hat{f}_\lambda^{(RF)}(x)) &= \mathrm{Var}\left(\mathbb{E}\left[\hat{f}_\lambda^{(RF)}(x) \mid F\right]\right) + \mathbb{E}\left[\mathrm{Var}\left[\hat{f}_\lambda^{(RF)}(x) \mid F\right]\right] \\ &= \mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) + \frac{1}{P}\mathbb{E}\left[\|\hat{\theta}\|^2\right]\left[K(x,x) - K(x,X)K(X,X)^{-1}K(X,x)\right].\end{aligned}$$

From Proposition C.13,

$$\mathrm{Var}\left(K(x,X)K(X,X)^{-1}\hat{y}\right) \leq \frac{c_1 K(x,x)\|y\|_{K^{-1}}^2}{P},$$

and from Proposition C.14, we have:

$$\mathbb{E}\left[\|\hat{\theta}\|^2\right] \leq \partial_\lambda\tilde{\lambda}\, y^T K\left(K + \tilde{\lambda} I_N\right)^{-2} y + \frac{c_2\|y\|_{K^{-1}}^2}{P} \leq \partial_\lambda\tilde{\lambda}\, \|y\|_{K^{-1}}^2 + \frac{c_2\|y\|_{K^{-1}}^2}{P} \leq \alpha\|y\|_{K^{-1}}^2,$$

where $\alpha = \partial_\lambda\tilde{\lambda} + c_2$. Using the fact that $\tilde{K}(x,x) \leq K(x,x)$, we get

$$\begin{aligned}\mathbb{E}\left[\mathrm{Var}\left[\hat{f}(x) \mid F\right]\right] &= \frac{1}{P}\mathbb{E}\left[\|\hat{\theta}\|^2\right]\left[K(x,x) - K(x,X)K(X,X)^{-1}K(X,x)\right] \\ &\leq \frac{\alpha\|y\|_{K^{-1}}^2 K(x,x)}{P}.\end{aligned}$$

This yields

$$\mathrm{Var}\left(\hat{f}_\lambda^{(RF)}(x)\right) \leq \frac{c_3\|y\|_{K^{-1}}^2 K(x,x)}{P},$$

where $c_3 = \alpha + c_1$. □

C.3.4. Average loss of λ-RF predictor and loss of $\tilde{\lambda}$-KRR:

Putting the pieces together, we obtain the following bound on the difference $\Delta_E = |\mathbb{E}[L(\hat{f}_{\lambda,\gamma}^{(RF)})] - L(\hat{f}_{\tilde{\lambda}}^{(K)})|$ between the expected RF loss and the KRR loss:

Corollary C.16. *If $\mathbb{E}_\mathcal{D}[K(x,x)] < \infty$, we have*

$$\Delta_E \leq \frac{C_1\|y\|_{K^{-1}}}{P}\left(2\sqrt{L(\hat{f}_{\tilde{\lambda}}^{(K)})} + C_2\|y\|_{K^{-1}}\right),$$

where C_1 and C_2 depend on λ, γ, T and $\mathbb{E}_\mathcal{D}[K(x,x)]$ only.

Proof. Using the bias/variance decomposition, Corollary C.9, and the bound on the variance of the predictor, we obtain

$$\begin{aligned}\left|\mathbb{E}\left[L\left(\hat{f}_{\gamma,\lambda}^{(RF)}\right)\right] - L\left(\hat{f}_{\tilde{\lambda}}^{(K)}\right)\right| &\leq \left|L\left(\mathbb{E}\left[\hat{f}_{\gamma,\lambda}^{(RF)}\right]\right) - L\left(\hat{f}_{\tilde{\lambda}}^{(K)}\right)\right| + \mathbb{E}_\mathcal{D}\left[\mathrm{Var}\left(\hat{f}(x)\right)\right] \\ &\leq \frac{C\|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}_{\tilde{\lambda}}^{(K)}\right)} + \frac{C\|y\|_{K^{-1}}}{P}\right) + \frac{c_3\|y\|_{K^{-1}}^2\mathbb{E}_\mathcal{D}\left[K(x,x)\right]}{P} \\ &\leq \frac{C_1\|y\|_{K^{-1}}}{P}\left(2\sqrt{L\left(\hat{f}_{\tilde{\lambda}}^{(K)}\right)} + C_2\|y\|_{K^{-1}}\right),\end{aligned}$$

where C_1 and C_2 depends on λ, γ, T and $\mathbb{E}_\mathcal{D}\left[K(x,x)\right]$ only. □

C.3.5. DOUBLE DESCENT CURVE

Recall that for any $\tilde{\lambda}$, we denote $M_{\tilde{\lambda}} = K(X,X)(K(X,X) + \tilde{\lambda} I_N)^{-2}$. A direct consequence of Proposition C.14 is the following lower bound on the variance of the predictor.

Corollary C.17. *There exists $c_4 > 0$ depending on λ, γ, T only such that $\mathrm{Var}\left(\hat{f}_{\lambda}^{(RF)}(x)\right)$ is bounded from below by*

$$\partial_{\lambda}\tilde{\lambda} \frac{y^T M_{\tilde{\lambda}} y}{P} \tilde{K}(x,x) - \frac{c_4 K(x,x) \|y\|_{K^{-1}}^2}{P^2}.$$

Proof. By the law of total cumulance,

$$\mathrm{Var}\left(\hat{f}_{\lambda}^{(RF)}(x)\right) \geq \mathbb{E}\left[\mathrm{Var}\left[\hat{f}_{\lambda}^{(RF)}(x) \mid F\right]\right] \geq \frac{1}{P}\mathbb{E}\left[\|\hat{\theta}\|^2\right] \tilde{K}(x,x).$$

From Proposition C.14, $\mathbb{E}[\|\hat{\theta}\|^2] \geq \partial_{\lambda}\tilde{\lambda}\, y^T M_{\tilde{\lambda}} y - \frac{c_2 \|y\|_{K^{-1}}^2}{P}$, hence

$$\mathrm{Var}\left(\hat{f}_{\lambda}^{(RF)}(x)\right) \geq \partial_{\lambda}\tilde{\lambda} \frac{y^T M_{\tilde{\lambda}} y}{P} \tilde{K}(x,x) - \frac{c_4 \tilde{K}(x,x) \|y\|_{K^{-1}}^2}{P^2}.$$

The result follows from the fact that $\tilde{K}(x,x) \leq K(x,x)$. □

References

Au, B., Cébron, G., Dahlqvist, A., Gabriel, F., and Male, C. Large permutation invariant random matrices are asymptotically free over the diagonal, 2018. To appear in Annals of Probability.

Bai, Z. and Wang, Z. Large sample covariance matrices without independence structures in columns. *Statistica Sinicia*, 18: 425–442, 2008.

Eaton, M. Multivariate statistics: A vector space approach. *Journal of the American Statistical Association*, 80, 01 2007. doi: 10.2307/20461449.