## A Some useful results

**Proposition 1** (Spectral functions)**.** *Let $f, g : [0, T] \to \mathbb{R}$ be continuous functions and $A \in \mathbb{R}^{n\times n}$ symmetric with $\|A\| \leq T$, for $T > 0$, $n \geq 1$. Let $A = U\Sigma U^\top$ be its eigenvalue decomposition with $U \in \mathbb{R}^{n\times n}$ an orthonormal matrix, $U^\top U = UU^\top = I$ and $\Sigma$ a diagonal matrix, then*

$$f(A) = Uf(\Sigma)U^\top,$$
$$f(A) + g(A) = (f + g)(A), \quad f(A)g(A) = (fg)(A)$$

*where $f(\Sigma) = diag(f(\sigma_1), \dots, f(\sigma_n))$. Moreover, let $B \in \mathbb{R}^{n\times m}$ with $n, m \geq 1$, then*

$$f(B^\top B)B^\top = B^\top f(BB^\top).$$

**Proposition 2.** *With the notation of Section 2.3 let $R \in \mathbb{R}^{m\times p}$ such that $K^\dagger_{mm} = RR^\top$ and $A = K_{nm}R$. Then, for any $\lambda, m > 0$, $\tilde{\alpha}_{m,\lambda}$ is characterized by Equation 18.*

*Proof.* By Equation 7.7 of Rifkin et al. we have that

$$\begin{aligned}\tilde{\alpha}_{m,\lambda} &= K^\dagger_{mm}K^\top_{nm}(K_{nm}K^\dagger_{mm}K^\top_{nm} + \lambda nI)^{-1}y \\ &= RR^\top K^\top_{nm}(K_{nm}RR^\top K^\top_{nm} + \lambda nI)^{-1}y \\ &= RA^\top(AA^\top + \lambda nI)^{-1}y \\ &= R(A^\top A + \lambda nI)^{-1}A^\top y,\end{aligned}$$

where the last step is due to Prop. 1. □

**Proposition 3.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function on $\mathcal{X}$, $x_1, \dots, x_n$ be the given points and $y = (y_1, \dots, y_n)$ be the labels of the dataset. For any function of the form $f(x) = \sum_{i=1}^n w_i k(x, x_i)$ with $w = Cy$ for any $x \in \mathcal{X}$, with $C \in \mathbb{R}^{n\times n}$ independent from $y$, the following holds*

$$\mathbb{E}_y R(f) = \underbrace{\frac{\sigma^2}{n}\operatorname{Tr}(Q^2)}_{\text{Variance } V(Q)} + \underbrace{\frac{1}{n}\|P(I - Q)\mu\|^2}_{\text{Bias } B(Q)},$$

*with $Q = KC \in \mathbb{R}^{n\times n}$, $K$ the kernel matrix, $\mu = \mathbb{E}y \in \mathbb{R}^n$ and $P = K^\dagger K$ the projection operator on the range of $K$.*

*Proof.* A function $f \in \mathcal{H}$ is of the form $f(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$ for any $x \in \mathcal{X}$. If we compute it on a point of the dataset $x_i$, with $i \in \{1, \dots, n\}$ we have $f(x_i) = \sum_{j=1}^n \alpha_j k(x_i, x_j) = k_i^\top w$ with $w = Cy$ and $k_i = (k(x_i, x_1), \dots, k(x_i, x_n))$. Note that $K = (k_1, \dots, k_n)$.

**Rewriting of E, R for fixed design.** We have

$$\begin{aligned}\mathcal{E}(w) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}(k_i^\top w - y_i) = \frac{1}{n}\sum_{i=1}^n (\mathbb{E}\left(k_i^\top w - \mu_i\right)^2 \\ &\quad - 2\left(k_i^\top w - \mu_i\right)(y_i - \mu_i) + (y_i - \mu_i)^2) \\ &= \frac{1}{n}\sum_{i=1}^n (k_i^\top w - \mu_i)^2 + \frac{\sigma^2}{n} = \frac{\sigma^2}{n} + \frac{1}{n}\|Kw - \mu\|^2,\end{aligned}$$

Now note that $PK = K$ and $(I - P)K = 0$, that $\|q\|^2 = \|Pq\|^2 + \|(I - P)q\|^2$ for any $q \in \mathcal{H}$ and that $\inf_{v\in\mathcal{X}} \mathcal{E}(v) = \sigma^2 + \|(I - P)\mu\|^2$, then the excess risk can be rewritten as

$$\begin{aligned}R(w) &= \frac{1}{n}\|Kw - \mu\|^2 - \frac{1}{n}\|(I - P)\mu\|^2 \\ &= \frac{1}{n}\|P(Kw - \mu)\|^2 + \frac{1}{n}\|(I - P)(Kw - \mu)\|^2 \\ &\quad - \frac{1}{n}\|(I - P)\mu\|^2 = \frac{1}{n}\|P(Kw - \mu)\|^2.\end{aligned}$$

**Expected Excess Risk.** Now we focus on the expectation of $R$ with respect to the dataset for linear functions that depend linearly on the observed labels $y$. Indeed we have

$$\begin{aligned}\mathbb{E}R(w) &= \frac{1}{n}\mathbb{E}\|P(KCy - P\mu)\|^2 \\ &= \frac{1}{n}\mathbb{E}\|PQ(y - \mu) + P(I - Q)\mu\|^2 \\ &= \frac{1}{n}\mathbb{E}\operatorname{Tr}(Q(y - \mu)(y - \mu)^\top Q) + \frac{1}{n}\|P(I - Q)\mu\|^2 \\ &\quad - \frac{2}{n}\mathbb{E}(y - \mu)^\top QP(I - Q)\mu \\ &= \frac{1}{n}\operatorname{Tr}(Q\mathbb{E}(y - \mu)(y - \mu)^\top Q) + \frac{1}{n}\|P(I - Q)\mu\|^2 \\ &= \frac{\sigma^2}{n}\operatorname{Tr}(Q^2) + \frac{1}{n}\|P(I - Q)\mu\|^2.\end{aligned}$$

Here the third step is due to $\|a - b\|^2 = \|a\|^2 + \|b\|^2 - 2a^\top b$ and that $\|a\|^2 = \operatorname{Tr}(aa^\top)$, for any vector $a, b$. The last term in the third step vanishes due to the fact that $y - \mu$ is a zero mean random variable. Moreover, note that $(\mathbb{E}(y - \mu)(y - \mu)^\top)_{ij} = \mathbb{E}(y_i - \mu_i)(y_j - \mu_j) = \sigma^2\delta_{ij}$, therefore $\mathbb{E}(y - \mu)(y - \mu)^\top = \sigma^2 I$. □

## B Proofs

*Proof of Theorem 1.* By applying Prop. 3 to the estimator of Equation 3 we have $Q_{\text{ols}} = K^\dagger K = P$. Now note that $P^2 = P$ by definition, $\operatorname{Tr}(P) = d^*$ and that $P(I - P) = 0$, therefore

$$\mathbb{E}R(f_{\text{ols}}) = \frac{\sigma^2}{n}\operatorname{Tr}(P^2) + \frac{1}{n}\|P(I - P)\mu\| = \frac{\sigma^2 d^*}{n}.$$

□

*Proof of Theorem 2.* Let $K = U\Sigma U^\top$ be the eigen-decomposition of $K$ with $U$ an orthonormal matrix and $\Sigma$ a diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. Let $\bar{Q}_\lambda = (K + \lambda n I)^{-1}K$, $\beta = U^\top P\mu$ with $\mu = \mathbb{E}y$ as in Eq. (5), $P = K^\dagger K$ the projection operator on the range of $K$. By applying Prop. 3 to the estimator of Eq. (3), considering that $P(I - \bar{Q}_\lambda) = (I - \bar{Q}_\lambda)P$, that $I - \bar{Q}_\lambda = \lambda n(K + \lambda nI)^{-1}$ and that $\sigma_i = \beta_i = 0$ for $i > d^*$, we have

$$
\begin{aligned}
\mathbb{E}R(\bar{f}_\lambda) &= \frac{\sigma^2}{n}\operatorname{Tr}(\bar{Q}_\lambda^2) + \frac{1}{n}\|P(I - \bar{Q}_\lambda)\mu\|^2 \\
&= \frac{\sigma^2}{n}\operatorname{Tr}(\bar{Q}_\lambda^2) + \frac{1}{n}\|(I - \bar{Q}_\lambda)P\mu\|^2 \\
&= \frac{\sigma^2}{n}\operatorname{Tr}(\Sigma^2(\Sigma + \lambda I)^{-2}) + \frac{\lambda}{n}\|(\Sigma + \lambda I)^{-1}\beta\|^2 \\
&= \frac{1}{n}\sum_{i=1}^{d^*}\frac{\sigma^2\sigma_i^2 + \lambda^2 n^2\beta_i^2}{(\sigma_i + \lambda n)^2} = \frac{1}{n}\sum_{i=1}^{d^*}\frac{\sigma^2\bar{\sigma}_i^2 + \lambda^2\beta_i^2}{(\bar{\sigma}_i + \lambda)^2},
\end{aligned}
$$

with $\bar{\sigma}_i = \sigma_i/n$ for $1 \leq i \leq d^*$. Note that, by defining $\tau_i = \sigma_i^{-1/2}\beta_i$ for $1 \leq i \leq d^*$, we have

$$
\begin{aligned}
\|f_{\text{opt}}\|_{\mathcal{H}}^2 &= \sum_{i,j=1}^{n}\langle\alpha_{\text{opt},i}k(x_i,\cdot), \alpha_{\text{opt},j}k(x_j,\cdot)\rangle_{\mathcal{H}} \\
&= \alpha_{\text{opt}}^\top K\alpha_{\text{opt}} = \mu^\top K^\dagger K K^\dagger\mu = \mu^\top PK^\dagger P\mu \\
&= \mu^\top PU\Sigma^\dagger U^\top P\mu = \beta^\top\Sigma^\dagger\beta = \sum_{i=1}^{d^*}\tau_i^2.
\end{aligned}
$$

Now we study $\mathbb{E}R(\bar{f}_{\lambda^*})$. When $\lambda^* = \sigma^2/T$ with $T = \|f_{\text{opt}}\|_{\mathcal{H}}^2$. We have

$$
\begin{aligned}
\mathbb{E}R(\bar{f}_{\lambda^*}) &= \frac{\sigma^2}{n}\sum_{i=1}^{d^*}\frac{\bar{\sigma}_i}{\bar{\sigma}_i + \lambda^*}\frac{\bar{\sigma}_i + \sigma^2\frac{\tau_i^2}{T^2}}{\bar{\sigma}_i + \frac{\sigma^2}{T}} \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{d^*}\frac{\bar{\sigma}_i}{\bar{\sigma}_i + \lambda^*}\frac{(\bar{\sigma}_i + \frac{\sigma^2}{T}) - \frac{\sigma^2}{T}(1 - \frac{\tau_i^2}{T})}{\bar{\sigma}_i + \frac{\sigma^2}{T}} \\
&= \frac{\sigma^2}{n}\sum_{i=1}^{d^*}\frac{\bar{\sigma}_i}{\bar{\sigma}_i + \lambda^*}\left(1 - \frac{1 - \tau_i^2/T}{1 + T\bar{\sigma}_i/\sigma^2}\right) \\
&\leq \frac{\sigma^2}{n}\sum_{i=1}^{d^*}\frac{\bar{\sigma}_i}{\bar{\sigma}_i + \lambda^*} = \frac{\sigma^2}{n}\sum_{i=1}^{d^*}\frac{\sigma_i}{\sigma_i + \lambda^* n} \\
&= \frac{\sigma^2}{n}\operatorname{Tr}(\Sigma(\Sigma + \lambda^* nI)^{-1}) = \frac{\sigma^2}{n}d_{\text{eff}}(\lambda^*).
\end{aligned}
$$

□

*Proof of Theorem 3.* It is an application of Theorem 5 when we select the whole training set ($m = n$) for the Nyström approximation. In that case the expected excess risks of Nyström KRLS and NYTRO are just equal to the ones of KRLS and Early Stopping, indeed when $m = n$ we have that $K_{mm} = K_{nm} = K$. If we call $\bar{Q}_\lambda$ and $\tilde{Q}_{n,\lambda}$ the $Q$-matrices for the two algorithms (see Prop. 3) and $R$ such that $RR^\top = K_{mm}^\dagger$, for any $\lambda > 0$ we have

$$
\begin{aligned}
\bar{Q}_\lambda &= (K + \lambda nI)^{-1}K = (KK^\dagger K + \lambda nI)^{-1}KK^\dagger K \\
&= (KRR^\top K + \lambda nI)^{-1}KRR^\top K \\
&= KR(R^\top K^2 R + \lambda nI)^{-1}R^\top K = \tilde{Q}_{n,\lambda}.
\end{aligned}
$$

□

*Proof of Theorem 5.* In the following we assume without loss of generality that the selected points $\tilde{x}_1, \ldots, \tilde{x}_m$ are the first $m$ points in the dataset. In Prop. 3 we have seen that the behavior of an algorithm in a fixed design setting is completely described by a matrix $Q = KC$ when the coefficients of the estimator of the algorithm are of the form $Cy$. Now we find the associated $Q$ for NYTRO, that is $\hat{Q}_{m,\gamma,t}$. By solving the recursion of Equation (19), we have for any $i \in \{1, \ldots, n\}$

$$
\begin{aligned}
\hat{f}_{m,\gamma,t}(x_i) &= k_i^\top Cy, \text{ with } C = \begin{pmatrix} C_{m,\gamma,t} \\ 0_{(n-m)\times n}\end{pmatrix}, \\
C_{m,\gamma,t} &= \gamma\sum_{p=0}^{t-1}R(I - \gamma A^\top A)^p A^\top,
\end{aligned}
$$

with $A = K_{nm}R$ and $k_i = (k(x_i, x_1), \ldots, k(x_i, x_n))$. Therefore, we have

$$
\begin{aligned}
\hat{Q}_{m,\gamma,t} = KC &= \gamma\sum_{p=0}^{t-1}K_{nm}R(I - \gamma A^\top A)^p A^\top \\
&= \gamma\sum_{p=0}^{t-1}A(I - \gamma A^\top A)^p A^\top.
\end{aligned}
$$

**Rewriting of $\hat{Q}_{m,\gamma,t}$.** Now we rewrite $\hat{Q}_{m,\gamma,t}$ in a suitable form to bound the bias and variance error. First of all we apply Prop. 1 to $\hat{Q}_{m,\gamma,t}$. Let $f(\sigma) = \gamma\sum_{i=0}^{t-1}(1 - \gamma/n\sigma)^p$ with $\sigma \in [0, n/\gamma]$, we have that

$$
\hat{Q}_{m,\gamma,t} = Af(A^\top A)A^\top = f(AA^\top)AA^\top = g(AA^\top),
$$

where $g(\sigma) = f(\sigma)\sigma$. Now note that

$$
g(\sigma) = \gamma\sigma\sum_{i=0}^{t-1}(1 - \gamma/n\sigma)^p = 1 - (1 - \gamma/n\sigma)^t,
$$

therefore we have

$$
\hat{Q}_{m,\gamma,t} = g(AA^\top) = I - (I - \gamma/nAA^\top)^t.
$$

**Bound of the bias.** Now we are going to bound the bias for NYTRO. Let $\lambda = 1/(\gamma t)$ and $Z = AA^\top$, then

$$\begin{aligned}B(\hat{Q}_{m,\gamma,t}) &= \frac{1}{n}\|P(I-\hat{Q}_{m,\gamma,t})\mu\|^2 \\ &= \frac{1}{n}\|P(I-\frac{\gamma}{n}Z)^t\mu\|^2 = \frac{1}{n}\|(I-\frac{\gamma}{n}Z)^t P\mu\|^2 \\ &= \frac{1}{n}\|(I-\frac{\gamma}{n}Z)^t(Z+\lambda nI)(Z+\lambda nI)^{-1}P\mu\|^2 \\ &\leq \frac{1}{n}q(A,\lambda n)\|(Z+\lambda nI)^{-1}P\mu\|^2\end{aligned}$$

and $q(A,\lambda n) = \|(I-\gamma/nAA^\top)^t(AA^\top+\lambda nI)\|^2$. Note that the third step is due to the fact that $\operatorname{ran} Z \subseteq \operatorname{ran} K = \operatorname{ran} P$ and $Z$ is symmetric, therefore $Ph(Z) = h(Z)P$ as a consequence of Prop. 1 for any spectral function $h$. Let $\sigma_1, \dots, \sigma_n$ be the singular values of $Z$, we have

$$\begin{aligned}q\left(A, \frac{n}{\gamma t}\right) &= \sup_{i\in\{1,\dots,n\}} (1-\gamma/n\,\sigma_i)^{2t}\left(\sigma_i + \frac{n}{\gamma t}\right)^2 \\ &\leq \sup_{0\leq\sigma\leq n/\gamma}(1-\gamma/n\,\sigma)^{2t}\left(\sigma+\frac{n}{\gamma t}\right)^2 \leq \frac{n^2}{\gamma^2 t^2}.\end{aligned}$$

Therefore we have

$$B(\hat{Q}_{m,\gamma,t}) \leq \lambda^2 n\|(Z+\lambda n)^{-1}P\mu\|^2.$$

**Bound for the Variance.** Let $t \geq 2$, $\lambda = \frac{1}{\gamma t}$, $r(\sigma) = (1-\gamma/n\,\sigma)^t$ and

$$v(\sigma) = \sigma/(t-1) + \sigma(1+r(\sigma)) - \lambda n(1-r(\sigma)).$$

We have $v(\sigma) \geq 0$ for $0 \leq \sigma \leq n/\gamma$. Indeed for $\lambda n < \sigma \leq n/\gamma$ we have $v(\sigma) \geq 0$ since $0 \leq r(\sigma) \leq 1$, while for $0 \leq \sigma \leq \lambda n$ we have

$$\begin{aligned}\lambda n(1-r(\sigma)) &= \lambda n\left(1 - e^{-t\log\frac{1}{1-\frac{\gamma\sigma}{n}}}\right) \leq \frac{n}{\gamma t} t\log\frac{1}{1-\frac{\gamma\sigma}{n}} \\ &\leq \frac{n}{\gamma}\frac{\gamma/n\,\sigma}{1-\gamma/n\,\sigma} \leq \frac{\sigma}{1-\frac{1}{t}} = \frac{\sigma}{t-1}+\sigma \\ &\leq \frac{\sigma}{t-1} + \sigma(1+r(\sigma)),\end{aligned}$$

therefore $v(\sigma) \geq 0$. Now let $0 \leq \sigma \leq n/\gamma$. Since $v(\sigma) \geq 0$, the function $w(\sigma) = v(\sigma)/(\sigma+\lambda n)$ is $w(\sigma) \geq 0$. Now we rewrite $w$ a bit. First of all, note that

$$w(\sigma) = (2t-1)/(t-1)w_1(\sigma) - g(\sigma),$$

with $w_1(\sigma) = \sigma/(\sigma+\lambda n)$. The fact that $w(\sigma) \geq 0$ and that $g(\sigma) \geq 0$ implies that

$$\left(\frac{2t-1}{t-1}\right)^2 w_1(\sigma)^2 \geq g(\sigma)^2. \quad \forall 0 \leq \sigma \leq \frac{n}{\gamma}, t \geq 2$$

Now we focus on $\operatorname{Tr}(\hat{Q}^2_{\gamma t})$. Let $Z = U\Sigma U^\top$ be its eigenvalue decomposition with $U$ an orthonormal matrix and $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ with $\sigma_1 \geq \dots \geq \sigma_n \geq 0$,

$$\begin{aligned}\operatorname{Tr}(\hat{Q}^2_{m,\gamma,t}) &= \operatorname{Tr}(g^2(Z)) = \operatorname{Tr}(Ug^2(\Sigma)U^\top) = \operatorname{Tr}(g^2(\Sigma)) \\ &= \sum_{i=1}^n g(\sigma_i)^2 \leq c_t\sum_{i=1}^n w_1(\sigma_i)^2 = c_t\operatorname{Tr}(w_1(\Sigma)^2) \\ &= c_t\operatorname{Tr}(Uw_1(\Sigma)^2U^\top) = c_t\operatorname{Tr}(w_1(Z)^2) \\ &= c_t\operatorname{Tr}(Z^2(Z+\lambda nI)^{-2})\end{aligned}$$

where we applied many times Prop. 1 and the fact that the trace is invariant to unitary transforms. Thus,

$$V(\hat{Q}_{m,\gamma,t},n) \leq \frac{\sigma^2}{n}\left(\frac{2t-1}{t-1}\right)^2\operatorname{Tr}\left(Z\,(Z+n/(\gamma t)I)^{-1}\right)^2.$$

**The expected excess risk for Nyström KRLS** The Nyström KRLS estimator with linear kernel is a function of the form

$$\begin{aligned}&\tilde{f}(x_i) = k_i^\top Cy, \quad \text{with } C = \begin{pmatrix}\tilde{C}_{m,\lambda} \\ 0_{(n-m)\times n}\end{pmatrix}, \\ &\tilde{C}_{m,\lambda} = R(A^\top A+\lambda nI)^\dagger A^\top,\end{aligned}$$

with $k_i = (k(x_i,x_1),\dots,k(x_i,x_n))$ for any $i \in \{1,\dots,n\}$. Now, by applying Prop. 1 we have

$$\begin{aligned}\tilde{Q}_{m,\lambda} &= KC = K_{nm}\tilde{C}_{m,\lambda} \\ &= A(A^\top A+\lambda nI)^{-1}A = AA^\top(AA^\top+\lambda I)^{-1} \\ &= Z(Z+\lambda nI)^{-1}\end{aligned}$$

Thus we have

$$\begin{aligned}V(\tilde{Q}_{m,\lambda}) &= \frac{\sigma^2}{n}\operatorname{Tr}(\tilde{Q}_{m,\lambda})^2 = \frac{\sigma^2}{n}\operatorname{Tr}\left(Z\,(Z+\lambda nI)^{-1}\right)^2 \\ B(\tilde{Q}_{m,\lambda}) &= \frac{1}{n}\|P(I-Z\,(Z+\lambda nI)^{-1})\mu\|^2 \\ &= \lambda^2 n\|P(Z+\lambda nI)^{-1}\mu\|^2 \\ &= \lambda^2 n\|(Z+\lambda nI)^{-1}P\mu\|^2,\end{aligned}$$

where the last step is due to the same reasoning as in the bound for the bias of NYTRO. Finally, by applying twice Prop. 3 and calling $c_t = \left(\frac{2t-1}{t-1}\right)^2$, we have that

$$\begin{aligned}R(\hat{f}_{m,\gamma,t}) &= V(\hat{Q}_{m,\gamma,t},n) + B(\hat{Q}_{m,\gamma,t}) \\ &\leq c_t V(\tilde{Q}_{m,\frac{1}{\gamma t}},n) + B(\tilde{Q}_{m,\frac{1}{\gamma t}}) \\ &\leq c_t\left(V(\tilde{Q}_{m,\frac{1}{\gamma t}},n) + B(\tilde{Q}_{m,\frac{1}{\gamma t}})\right) \\ &= c_t R(\tilde{f}_{m,\frac{1}{\gamma t}})\end{aligned}$$

for $\|Z\| \leq n/\gamma$ and $t \geq 2$. Now the choice $\gamma = 1/(\max_{1\leq i\leq n} k(x_i,x_i))$ is valid, indeed

$$\begin{aligned}\gamma\|Z\|^2 &= \gamma\|K_{nm}RR^\top K_{nm}^\top\| = \gamma\|K_{nm}K_{mm}^\dagger K_{nm}^\top\| \\ &\leq \gamma\|K\| \leq \gamma n\max_{1\leq i\leq n}(K)_{ii} = \gamma n\max_{1\leq i\leq n}k(x_i,x_i),\end{aligned}$$

where $\|K_{nm}K_{mm}^{\dagger}K_{nm}^{\top}\| \leq \|K\|$ can be found in Bach (2013); Alaoui and Mahoney (2014). $\square$

*Proof of Corollary 1.* Theorem 5 combined with Theorem 1 of Bach (2013). $\square$