A Some useful results

Proposition 1 (Spectral functions). Let \( f, g : [0, T] \rightarrow \mathbb{R} \) 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

\[
\begin{align*}
    f(A) & = U f(\Sigma) U^\top, \\
    f(A) + g(A) &= (f + g)(A), \\
    f(A)g(A) &= (fg)(A)
\end{align*}
\]

where \( f(\Sigma) = \text{diag}(f(\sigma_1), \ldots, f(\sigma_n)) \). 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 \( K_{mm} = RR^\top \) and \( A = K_{nm}R \). Then, for any \( \lambda, m > 0 \), \( \alpha_{m, \lambda} \) is characterized by Equation 18.

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

\[
\begin{align*}
    \hat{\alpha}_{m, \lambda} &= K_{mm}^\top K_{mm}(K_{nm} K_{nm}^\top K_{mm} + \lambda nI)^{-1} y \\
    &= RR^\top (RR^\top K_{nm} RR^\top K_{nm}^\top + \lambda nI)^{-1} y \\
    &= RA^\top (AA^\top + \lambda nI)^{-1} y \\
    &= R(A^\top A + \lambda nI)^{-1} A^\top y,
\end{align*}
\]

where the last step is due to Prop. 1.

Proposition 3. Let \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) be a kernel function on \( \mathcal{X} \), \( x_1, \ldots, x_n \) be the given points and \( y = (y_1, \ldots, y_n) \) 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 \( C \in \mathbb{R}^{n \times m} \) independent from \( y \), the following holds

\[
E_y R(f) = \frac{\sigma^2}{n} \text{Tr}(Q^2) + \frac{1}{n} \| P(I-Q)\mu \|_2^2. \tag{Variance \text{Var}(Q)}
\]

\[
+ \frac{1}{n} \| P(I-Q)\mu \|_2^2. \tag{Bias \text{Bias}(Q)}
\]

with \( Q = KC \in \mathbb{R}^{n \times n} \), \( K \) the kernel matrix, \( \mu = E y \in \mathbb{R}^n \) and \( P = K^\top K \) 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 \( i \in \{1, \ldots, n\} \) we have \( f(x_i) = \sum_{j=1}^n \alpha_j k(x_i, x_j) = k_i^\top w \) with \( w = Cy \) and \( k_i = (k(x_i, x_1), \ldots, k(x_i, x_n)) \). Note that \( K = (k_1, \ldots, k_n) \).

Rewriting of \( E, R \) for fixed design. We have

\[
E(w) = \frac{1}{n} \sum_{i=1}^n E(k_i^\top w - y_i) = \frac{1}{n} \sum_{i=1}^n (E(k_i^\top w - \mu_i)^2 \\
- 2(k_i^\top w - \mu_i)(y_i - \mu_i) + (y_i - \mu_i)^2) \\
= \frac{1}{n} \sum_{i=1}^n (k_i^\top w - \mu_i)^2 + \frac{\sigma^2}{n} + \frac{1}{n} \| Kw - \mu \|_2^2,
\]

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 \( \inf_{v \in \mathcal{H}} E(v) = \sigma^2 + \|(I-P)\mu\|_2^2 \), then the excess risk can be rewritten as

\[
R(w) = \frac{1}{n} \| Kw - \mu \|_2^2 - \frac{1}{n} \| (I-P)\mu \|_2^2 \\
= \frac{1}{n} \| PKw - \mu \|_2^2 + \frac{1}{n} \| (I-P)(Kw - \mu) \|_2^2 \\
- \frac{1}{n} \| (I-P)\mu \|_2^2 = \frac{1}{n} \| PKw - \mu \|_2^2.
\]

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

\[
\begin{align*}
    \mathbb{E} R(w) &= \frac{1}{n} \mathbb{E} \| PKC y - P \mu \|_2^2 \\
    &= \frac{1}{n} \mathbb{E} \| P(Q(y - \mu) + P(I-Q)\mu \|_2^2 \\
    &= \frac{1}{n} \mathbb{E} \text{Tr}(Q(y - \mu)(y - \mu)^\top) + \frac{1}{n} \| P(I-Q)\mu \|_2^2 \\
    - \frac{2}{n} \mathbb{E}(y - \mu)^\top QP(I-Q)\mu \\
    &= \frac{1}{n} \text{Tr}(QE(y - \mu)(y - \mu)^\top) + \frac{1}{n} \| P(I-Q)\mu \|_2^2 \\
    &= \frac{\sigma^2}{n} \text{Tr}(Q^2) + \frac{1}{n} \| P(I-Q)\mu \|_2^2.
\end{align*}
\]

Here the third term is due to \( \| a - b \|_2 = \| a \|_2 + \| b \|_2 - 2a^\top b \) and that \( \| a \|_2^2 = \text{Tr}(aa^\top) \), 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 \( Q_{obs} = K^\top K = P \). Now note that \( P^2 = P \) by definition, \( \text{Tr}(P) = d^* \) and that \( P(I-P) = 0 \), therefore

\[
\mathbb{E} R(f_{obs}) = \frac{\sigma^2}{n} \text{Tr}(P^2) + \frac{1}{n} \| P(I-P)\mu \|_2 = \frac{\sigma^2 d^*}{n}.
\]

\]
Proof of Theorem 2. Let $K = UΣU^\top$ be the eigen-decomposition of $K$ with $U$ an orthonormal matrix and $Σ$ a diagonal matrix with $σ_1 ≥ · · · ≥ σ_n ≥ 0$. Let $Q_λ = (K + λnI)^{-1}K$, $β = U^\top P_μ$ with $μ = E_y$ 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 $P(I - Q_λ) = (I - Q_λ)P$, that $I - Q_λ = λn(K + λnI)^{-1}$ and that $σ_i = β_i = 0$ for $i > d^*$, we have

$$
ER(\tilde{f}_λ) = \frac{σ^2}{n} \text{Tr}(\bar{Q}_λ^2) + \frac{1}{n} \|P(I - Q_λ)μ\|^2
$$

$$
= \frac{σ^2}{n} \text{Tr}(\bar{Q}_λ^2) + \frac{1}{n} \|P(I - Q_λ)Pμ\|^2
$$

$$
= \frac{σ^2}{n} \text{Tr}(Σ^2(Σ + λI)^{-2}) + \frac{λ}{n} \|Pμ\|^2
$$

$$
= \frac{1}{n} \sum_{i=1}^{d^*} \frac{σ_i^2 + λ^2n^2β_i^2}{(σ_i + λn)^2} = \frac{1}{n} \sum_{i=1}^{d^*} \frac{σ_i^2 + β_i^2}{(σ_i + λ)^2},
$$

with $σ_i = σ_i/n$ for $1 ≤ i ≤ d^*$. Note that, by defining $τ_i = σ_i^{-1/2}β_i$ for $1 ≤ i ≤ d^*$, we have

$$
\|f_\text{opt}\|_H^2 = \sum_{i,j=1}^{n} \langle α_{i,\text{opt}}, k(x_i, ·), α_{j,\text{opt}}, k(x_j, ·) \rangle_H
$$

$$
= α_{i,\text{opt}}^\top Kα_{\text{opt}} = μ^\top KK^\dagger μ = μ^\top PK^\dagger Pμ
$$

$$
= μ^\top PUSU^\top Pμ = β^\top Σβ = \sum_{i=1}^{d^*} τ_i^2.
$$

Now we study $ER(\tilde{f}_λ^\ast)$. When $λ^\ast = σ^2/T$ with $T = \|f_\text{opt}\|_H^2$. We have

$$
ER(\tilde{f}_λ^\ast) = \frac{σ^2}{n} \sum_{i=1}^{d^*} \frac{σ_i + λ^\ast}{σ_i + λ^\ast n} \bar{σ}_i + \frac{σ^2}{n} \bar{λ}^\ast
$$

$$
= \frac{σ^2}{n} \sum_{i=1}^{d^*} \frac{σ_i}{σ_i + λ^\ast} \left(1 - \frac{1 - σ_i^2}{T}\right)
$$

$$
≤ \frac{σ^2}{n} \sum_{i=1}^{d^*} \frac{σ_i}{σ_i + λ^\ast n} \sum_{i=1}^{d^*} \frac{σ_i}{σ_i + λ^\ast n}
$$

$$
= \frac{σ^2}{n} \text{Tr}(Σ(Σ + λ^\ast nI)^{-1}) = \frac{σ^2}{n} d^\text{eff}(λ^\ast).
$$

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 $C_y$. Now we find the associated $Q$ for NYTRO, that is $Q_{m,γ,t}$. By solving the recursion of Equation (19), we have for any $i ∈ \{1, \ldots, n\}$

$$
\hat{f}_{m,γ,t}(x_i) = k_i^\top C_y, \text{ with } C = \left(\begin{array}{c} C_{m,γ,t} \\ 0_{(n-m)×n} \end{array}\right),
$$

$$
C_{m,γ,t} = γ \sum_{p=0}^{t-1} R(I - γA^\top A)^p A^\top,
$$

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

$$
Q_{m,γ,t} = KC = γ \sum_{p=0}^{t-1} K_{nm}R(I - γA^\top A)^p A^\top
$$

$$
= γ \sum_{p=0}^{t-1} A(I - γA^\top A)^p A^\top.
$$

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

$$
Q_{m,γ,t} = Af(A^\top A)A^\top = f(AA^\top)AA^\top = g(AA^\top),
$$

where $g(σ) = f(σ)σ$. Now note that

$$
g(σ) = γσ \sum_{i=0}^{t-1} (1 - γ/nσ)^p = 1 - (1 - γ/nσ)^t,
$$

therefore we have

$$
Q_{m,γ,t} = g(AA^\top) = I - (I - γ/nAA^\top)^t.
$$

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

$$B(\hat{Q}_{m, \gamma, t}, t) = \frac{1}{n} \| P(I - \hat{Q}_{m, \gamma, t}) \mu \|^2$$

$$= \frac{1}{n} \| P(I - \frac{\gamma}{n} Z) \mu \|^2 = \frac{1}{n} \| (I - \frac{\gamma}{n} Z)^t P \mu \|^2$$

$$= \frac{1}{n} \| (I - \frac{\gamma}{n} Z)^t (Z + \lambda n I)(Z + \lambda n I)^{-1} P \mu \|^2$$

$$\leq \frac{1}{n} q(A, \lambda n) \| (Z + \lambda n I)^{-1} P \mu \|^2$$

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

$$q(A, \gamma t) = \sup_{i \in \{1, \ldots, n\}} (1 - \gamma/n \sigma_i)^t \sigma_i + n \frac{\gamma t}{\sigma_i}$$

$$\leq \sup_{0 \leq \sigma \leq n \gamma} (1 - \gamma/n \sigma)^t \left( \sigma + \frac{n \gamma t}{\sigma} \right) \leq \frac{n^2}{\gamma^2 t^2}.$$

Therefore we have

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

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

$$v(\sigma) = \frac{\sigma}{(t - 1) + \sigma(1 + r(\sigma))} - \lambda n(1 - r(\sigma)).$$

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

$$\lambda n(1 - r(\sigma)) = \lambda n \left( 1 - e^{-t \log \frac{1}{1 - \frac{\sigma}{\gamma}}} \right) \leq \frac{n \gamma}{\gamma t} \left( 1 - \frac{1}{1 - \frac{\sigma}{\gamma}} \right) = \sigma \frac{\gamma}{t - 1} + \sigma$$

$$\leq \frac{\sigma}{t - 1} + \sigma(1 + r(\sigma)),$$

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

$$w(\sigma) = \frac{(2t - 1)(t - 1)}{t} w_1(\sigma) - g(\sigma),$$

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

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

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

$$\text{Tr}(\hat{Q}_{m, \gamma, t}^2) = \text{Tr}(g^2(Z)) = \text{Tr}(U g^2(\Sigma) U^T) = \text{Tr}(g^2(\Sigma))$$

$$= \sum_{i=1}^n q(\sigma_i)^2 \leq c_t \sum_{i=1}^n w_1(\sigma_i)^2 = c_t \text{Tr}(w_1(\Sigma)^2)$$

$$= c_t \text{Tr}(U w_1(\Sigma)^2 U^T) = c_t \text{Tr}(w_1(Z)^2)$$

$$= c_t \text{Tr}(Z^2(Z + \lambda n I)^{-2})$$

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

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

**The expected excess risk for Nyström KRLS**

The Nyström KRLS estimator with linear kernel is a function of the form

$$\tilde{f}(x_i) = k_i^T C y, \quad \text{with } C = \left( \tilde{C}_{m, \lambda} \right)_{0 \leq \lambda \leq n_m \times n},$$

$$\tilde{C}_{m, \lambda} = R(A^T A + \lambda n I)^{-1} A^T,$$

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

$$\hat{Q}_{m, \lambda} = K C = K_{nm} \tilde{C}_{m, \lambda},$$

$$= A(A^T A + \lambda n I)^{-1} A = A A^T (A A^T + \lambda I)^{-1}$$

$$= Z(Z + \lambda n I)^{-1}$$

Thus we have

$$V(\hat{Q}_{m, \lambda}) = \frac{\sigma^2}{n} \text{Tr}(\hat{Q}_{m, \lambda})^2 = \frac{\sigma^2}{n} \text{Tr} \left( Z(Z + \lambda n I)^{-1} \right)^2$$

$$\leq \frac{1}{n} \| P(I - Z(Z + \lambda n I)^{-1}) \mu \|^2$$

$$= \frac{1}{n} \| P(Z + \lambda n I)^{-1} \mu \|^2$$

$$= \frac{\lambda^2 n}{Z(Z + \lambda n I)^{-1} P \mu \|^2},$$

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

$$R(f_{m, \gamma, t}, t) = V(\hat{Q}_{m, \gamma, t}, n) + B(\hat{Q}_{m, \gamma, t})$$

$$\leq c_t V(\hat{Q}_{m, \gamma, t}, n) + B(\hat{Q}_{m, \gamma, t})$$

$$= c_t \left( V(\hat{Q}_{m, \gamma, t}, n) + B(\hat{Q}_{m, \gamma, t}) \right)$$

$$= c_t R(f_{m, \gamma, t}),$$

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

$$\gamma \| Z \|^2 = \gamma \| K_{nm} R R^T K_{nm}^\dagger \| = \gamma \| K_{nm} K_{nm}^\dagger K_{nm}^\top \|$$

$$\leq \gamma \| K \| \leq \gamma n \max_{1 \leq i \leq n} (K)_{ii} \gamma n \max_{1 \leq i \leq n} k(x_i, x_i),$$
where $\|K_{nm}K_{mm}^tK_{nm}\| \leq \|K\|$ can be found in Bach (2013); Alaoui and Mahoney (2014).

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