Supplement

A. Upper bound for the supremum of Gaussian processes

Proof of Lemma 5.3. By the Gaussian concentration theorem (Boucheron et al., 2013, Theorem 5.8), with probability at least $1 - e^{-x}$ we have

$$\sup_{B\in T^*} G_B \le \mathbb{E} \sup_{B\in T^*} G_B + \sigma\sqrt{2x} \sup_{B\in T^*} \|[2I_{n\times n} - (B - \bar{A})](B - \bar{A})\mu\|. \tag{A.1}$$

$$\le C_{16}\gamma_2(T^*, d_G) + \sigma\sqrt{2x} \sup_{B\in T^*} 3\|(B - \bar{A})\mu\| \tag{A.2}$$

where for the second inequality we used Talagrand's majorizing measure theorem (cf., e.g., (Vershynin, 2018, Section 8.6)) and the fact that $B, \bar{A}$ have operator norm at most one, where d_G is the canonical metric of the Gaussian process,

$$d_G(A, B)^2 = \mathbb{E}[(G_A - G_B)^2].$$

If $D = B - A$ is the difference and P commutes with A and B,

$$G_B - G_A = \epsilon^T [2D\mu - \tfrac{1}{2}(A + B - 2\bar{A})D\mu - \tfrac{1}{2}D(A + B - 2P)\mu] + \epsilon^T D \left(\bar{A} - P\right) \mu.$$

By the triangle inequality and using that $A, B, P, \bar{A}$ have operator norm at most one, $d_G(A, B) \le 6\sigma\|D\mu\| + \sigma\|D(\bar{A} - P)\mu\|$. This shows that

$$\gamma_2(T^*, d_G) \le 6\sigma\gamma_2(T^*, d_1) + \sigma\gamma_2(T^*, d_2)$$

where $d_1(A, B) = \|(B - A)\mu\|$ and $d_2(A, B) = \|(A - B)(\bar{A} - P)\mu\|$. By Lemma 5.2, $\gamma_2(T^*, d_1) \le C_{17}\Delta(T^*, d_1)$ and similarly for d_2 (note that d_2 is similar to d_1 with μ replaced by $\mu' = (P - \bar{A})\mu$).

If $\sup_{B\in T^*} d(B, \bar{A}) \le \delta^*$ for the metric d in (5.1), then $\sup_{B\in T^*} \|(B - \bar{A})\mu\| \le \delta^*$ and $\Delta(T^*, d_1) \le 2\delta^*$. Furthermore if P is the convex projection of $\bar{A}$ onto the convex hull of T^* with respect to the Hilbert metric d in (5.1), then

$$\Delta(T^*, d_2) = \sup_{B,B'\in T^*} d_2(B, B') \le 2\|(P - \bar{A})\mu\| \le 2d(P, \bar{A}) \le 2d(B_0, \bar{A}) \le 2\delta^*$$

for any $B_0 \in T^*$ where we used that by definition of the convex projection, $d(P, \bar{A}) \le d(B_0, \bar{A})$. □

B. Upper bound for the supremum of Quadratic processes

The following inequality, known as the Hanson-Wright inequality, will be useful for the next Lemma. If $\varepsilon \sim N(0, \sigma^2 I_{n\times n})$ is standard normal, then

$$\mathbb{P}\left[|\varepsilon^T Q\varepsilon - \sigma^2 \operatorname{trace} Q| > 2\sigma^2(\|Q\|_F\sqrt{x} + \|Q\|_{op}x)\right] \le 2e^{-x}, \tag{B.1}$$

for any square matrix $Q \in \mathbb{R}^{n\times n}$. We refer to (Boucheron et al., 2013, Example 2.12) for a proof for normally distributed ε and (Rudelson & Vershynin, 2013; Hsu et al., 2012; Bellec, 2014; Adamczak, 2015) for proofs of (B.1) in the sub-gaussian case.

Proof of Lemma 5.4. We apply Theorem 2.4 in (Adamczak, 2015) which implies that if $W_B = \varepsilon^T Q_B \varepsilon - \operatorname{trace}[Q_B]$ where $\varepsilon \sim N(0, I_{n\times n})$ and Q_B is a symmetric matrix of size $n \times n$ for every B, then

$$\mathbb{P}\Big(\sup_{B\in T^*} W_B \le \mathbb{E} \sup_{B\in T^*} W_B + C_{18}\sigma\sqrt{x} \sup_{B\in T^*} \mathbb{E}\|Q_B\varepsilon\| + C_{19}x\sigma^2 \sup_{B\in T^*} \|Q_B\|_{op}\Big) \ge 1 - 2e^{-x}.$$

For the third term, $Q_B = 2(B - \bar{A}) - (B - \bar{A})^2/2$ hence $\|Q_B\|_{op} \le 6$ because $B, \bar{A}$ both have operator norm at most one. For the second term, since T^* is a family of ordered linear smoothers, there exists extremal matrices $B_0, B_1 \in T^*$ such that $B_0 \preceq B \preceq B_1$ for all $B \in T^*$; we then have $B - B_0 \preceq B_1 - B_0$ and

$$\|Q_B\varepsilon\| \le 3\|(B - \bar{A})\varepsilon\| \le 3\|(B_1 - B_0)\varepsilon\| + 3\|(B_0 - \bar{A})\varepsilon\| \le 3\|(B_1 - \bar{A})\varepsilon\| + 6\|(B_0 - \bar{A})\varepsilon\|.$$

Hence $\mathbb{E}\|Q_B\varepsilon\| \le \mathbb{E}[\|Q_B\varepsilon\|^2]^{1/2} \le 3\sigma\|B_1 - \bar{A}\|_F + 6\sigma\|B_0 - \bar{A}\|_F \le 9\delta^*$.

We finally apply a generic chaining upper bound to bound $\mathbb{E} \sup_{B\in T^*} W_B$. For any fixed $B_0 \in T^*$ we have $\mathbb{E}[W_{B_0}] = 0$ hence $\mathbb{E} \sup_{B\in T^*} W_B = \mathbb{E} \sup_{B\in T^*}(W_B - W_{B_0})$. For two matrices $A, B \in T^*$ we have $W_B - W_A = \varepsilon^T(Q_B - Q_A)\varepsilon - \operatorname{trace}[Q_B - Q_A]$, and

$$\varepsilon^T(Q_B - Q_A)\epsilon = \varepsilon^T[(B - A)(2I_{n\times n} - \tfrac{1}{2}(A + B - 2\bar{A}))]\varepsilon,$$

hence by the Hanson-Wright inequality (B.1), with probability at least $1 - 2e^{-x}$,

$$|W_B - W_A| \le 2\sigma^2\|(B - A)(2I_{n\times n} - \tfrac{1}{2}(A + B - 2\bar{A}))\|_F(\sqrt{x} + x) \le 8\sigma^2\|A - B\|_F(x + \sqrt{x}).$$

Hence by the generic chaining bound given in Theorem 3.5 in (Dirksen, 2015), we get that

$$\mathbb{E} \sup_{B\in T^*} |W_B - W_{B_0}| \le C_{20}\sigma^2 \left[\gamma_1(T^*, \|\cdot\|_F) + \gamma_2(T^*, \|\cdot\|_F) + \Delta(T^*, \|\cdot\|_F)\right].$$

For each $\alpha = 1, 2$ we have $\gamma_\alpha(T^*, \|\cdot\|_F) \leq C_{21}\Delta(T^*, \|\cdot\|_F)$ by Lemma 5.2. Since $\sigma\|B - \bar{A}\| \leq \delta^*$ for any $B \in T^*$, we obtain $\Delta(T^*, \|\cdot\|_F) \leq 2\delta^*/\sigma$. □

C. Proof of Theorem 3.2

Proof. Consider $\mu \in \mathbf{R}^n$ with norm $\|\mu\|^2 = n(1-c/\sqrt{n})$ for a numerical constant $c > 0$ to be determined. Set $A_1 = 0$ and $A_2 = I_n$, assume $\sigma^2 = 1$ for simplicity. The loss of A_1 is $\|\mu\|^2$ and the loss of A_2 is $\|\varepsilon\|^2$.

A_1 has smaller MSE than A_2 since $\|\mu\|^2 < n$. The regret for selecting based on C_p is thus $I_{\Omega_2}(\|\varepsilon\|^2 - \|\mu\|^2)$ where I_{Ω_2} is the indicator of the event $C_p(A_2) < C_p(A_1)$, this event is

$$\Omega_2 = \left\{C_P(A_2) = 2n < \|y\|^2 = C_P(A_2)\right\}.$$

Consider now for some absolute constants A, B, the events

$$\Omega_A = \{-1 \leq \varepsilon^T\mu/\|\mu\| \leq 0\}$$

and

$$\Omega_B = \{\|(I_n - \|\mu\|^{-2}\mu\mu^T)\varepsilon\|^2 - n \geq 3\sqrt{n}\}.$$

The first event Ω_A involves the standard normal $\varepsilon^\top\mu/\|\mu\|$ and the second event Ω_B involves the random variable $\|(I_n - \|\mu\|^{-2}\mu\mu^T)\varepsilon\|^2$ which has χ^2 distribution with $n-1$ degrees-of-freedom. The two random variables are independent by properties of $\varepsilon \sim N(0, I_n)$ so that Ω_A and Ω_B are independent and $\mathbb{P}(\Omega_A \cap \Omega_B) = \mathbb{P}(\Omega_A)\mathbb{P}(\Omega_B) \geq C_{22} > 0$ for some absolute constant.

Furthermore, on $\Omega_A \cap \Omega_B$ we have

$$\begin{aligned}\|y\|^2 - 2n &= \|\mu\|^2 + \|\varepsilon\|^2 + 2\varepsilon^T\mu - 2n \\ &\geq -c\sqrt{n} + 3\sqrt{n} - 2\|\mu\| \\ &\geq (-c+1)\sqrt{n}\end{aligned}$$

so that $\Omega_A \cap \Omega_B \subset \Omega_2$ if, for instance, we choose $c = 1/2$.

Since $\|y\|^2 = \|\mu\|^2 + 2\varepsilon^T\mu + \|\varepsilon\|^2$, Ω_2 can be rewritten

$$\Omega_2 = \left\{2c\sqrt{n} - 2\varepsilon^T\mu = 2(n - \|\mu\|^2) - 2\varepsilon^T\mu < \|\varepsilon\|^2 - \|\mu\|^2\right\}.$$

Hence the regret is bounded from below on $\Omega_A \cap \Omega_B$ as

$$\begin{aligned}(\|A_{\hat{k}}y - \mu\|^2 - \|A_1 y - \mu\|^2) &= (\|\varepsilon\|^2 - \|\mu\|^2) \\ &\geq (2c\sqrt{n} - 2\varepsilon^T\mu) \\ &\geq 2c\sqrt{n} = \sqrt{n}.\end{aligned}$$

Here, $\sqrt{n} \asymp \|\mu\| = (R^*)^{1/2}$ up to an absolute multiplicative constant, so that the claim is proved. □