Supplementary Material for "Risk consistency of cross-validation with lasso-type procedures"

This supplementary material contains proofs of the theorems and lemmata contained in the manuscript "Risk consistency of cross-validation with lasso-type procedures"

S1 Squared-error loss and quadratic forms

We can rewrite the various formulas for the risk from as quadratic forms. Define the parameter to be $\gamma^\top := (-1, \beta^\top)$, with associated estimator $\widehat{\gamma}_t^\top := (-1, \widehat{\beta}_t^\top)$.

$$R(\beta) = \mathbb{E}_{\mu_n}\left[(\mathcal{Y} - \beta^\top \mathcal{X})^2\right] = \gamma^\top \Sigma_n \gamma \tag{S1.1}$$

where $\Sigma_n := \mathbb{E}_{\mu_n}[\mathcal{Z}\mathcal{Z}^\top]$.

$$\widehat{R}(\beta) = \frac{1}{n}\left|\left|Y - \mathbb{X}\beta\right|\right|_2^2 = \gamma^\top \widehat{\Sigma}_n \gamma,$$

where $\widehat{\Sigma}_n = n^{-1}\sum_{i=1}^n Z_i Z_i^\top$.

$$\widehat{R}_{V_n}(t) = \frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_t^{(v)})^T \widehat{\Sigma}_v \widehat{\gamma}_t^{(v)}, \tag{S1.2}$$

where $\widehat{\Sigma}_v = |v|^{-1}\sum_{r \in v} Z_r Z_r^\top$, $\widehat{\gamma}_t^{(v)} := (-1, \widehat{\beta}_t^{(v)})^\top$, and

$$\widehat{\beta}_t^{(v)} := \operatorname*{argmin}_{\beta \in \mathcal{B}_t} \gamma^\top \widehat{\Sigma}_{(v)} \gamma,$$

with $\widehat{\Sigma}_{(v)} := (n - |v|)^{-1}\sum_{r \notin v} Z_r Z_r^\top$.

S2 Background Results

We use the following results in our proofs. First is a special case of Nemirovski's inequality. See Dümbgen et al. (2010) for more general formulations.

Lemma 1 (Nemirovski's inequality). *Let $(\xi_i)_{i\in v}$ be independent random vectors in $\mathbb{R}^d$, for $d \geq 3$ with $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}\left|\left|\xi_i\right|\right|_2^2 < \infty$. Then, for any validation set v and distribution for the ξ_i's,*

$$\mathbb{E}\left|\left|\sum_{i\in v}\xi_i\right|\right|_\infty^2 \leq (2e\log d - e)\sum_{i\in v}\mathbb{E}\left|\left|\xi_i\right|\right|_\infty^2 \leq 2e\log d\sum_{i\in v}\mathbb{E}\left|\left|\xi_i\right|\right|_\infty^2.$$

Also, we need the following results about the Orlicz norms.

Lemma 2 (van der Vaart and Wellner 1996). *For any ψ_r-Orlicz norm with $1 < r \in \mathbb{N}$ and sequence of $\mathbb{R}$-valued random variables $(\zeta_j)_{j=1,\ldots,m}$*

$$\left|\left|\max_{1\leq j\leq m}\zeta_j\right|\right|_{\psi_r} \leq \Psi\log^{1/r}(m+1)\max_{1\leq j\leq m}\left|\left|\zeta_j\right|\right|_{\psi_r},$$

where Ψ is a constant that depends only on ψ_r.

Lemma 3 (Corollary 5.17 in Vershynin 2012). *Let $\xi_1, \ldots, \xi_n$ be iid centered random variables and let $\left|\left|\xi_i\right|\right|_{\psi_1} \leq \kappa$. Then for every $\delta > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n \xi_i\right| \geq n\delta\right) \leq 2\exp\left(-cn\min\left\{\frac{\delta^2}{\kappa^2}, \frac{\delta}{\kappa}\right\}\right),$$

where $c = 1/8e^2$.

S3 Supporting Lemmas

Several times in our proof of the main results we need to bound a quadratic form given by a symmetric matrix and an estimator indexed by a tuning parameter. To this end, we state the following lemma.

Lemma 4. *Suppose $a \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times p}$. Then*

$$a^\top A a \leq ||a||_1^2 \, ||A||_\infty \, ,$$

where $||A||_\infty := \max_{i,j} |A_{ij}|$ is the entry-wise max norm.

We use Lemma 1 to find the rate of convergence for the sample covariance matrix to the population covariance.

Lemma 5. *Let $v \subseteq \{1, 2, \ldots, n\}$ be an index set and let $|v|$ be its number of elements. If $\mu \in \mathcal{F}_q$, then there exists a constant C, depending only on q, such that*

$$\mathbb{E}_\mu \left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|_\infty \leq C\sqrt{\frac{(\log p)^{1+2/q}}{|v|}},$$

where it is understood that $2/\infty = 0$.

Proof. (Lemma 5) Let $\xi_r \in \mathbb{R}^{(p+1)^2}$ be the vectorized version of the zero-mean matrix $\frac{1}{|v|}(Z_r Z_r^\top - \mathbb{E}\mathcal{Z}\mathcal{Z}^\top)$. Then, by Jensen's inequality

$$\left(\mathbb{E}\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|_\infty\right)^2 \leq \mathbb{E}\left|\left|\widehat{\Sigma}_v - \Sigma_n\right|\right|_\infty^2 = \mathbb{E}\left|\left|\sum_{r\in v} \xi_r\right|\right|_\infty^2 .$$

Using 1 with $d = (p+1)^2$ and writing $||X||^2_{L_2(\mu)} := \mathbb{E}_\mu X^2$ we find

$$\begin{aligned}
\mathbb{E}_\mu \left|\left|\sum_{r\in v} \xi_r\right|\right|^2_\infty &\leq 2e\log\left((p+1)^2\right)\sum_{r\in v}\left|\left|\,||\xi_r||_\infty\,\right|\right|^2_{L_2(\mu)} \\
&\leq 4e\log(p+1)\sum_{r\in v}\left(\left(2(\log 2)^{1/q-1}\right)\left|\left|\,||\xi_r||_\infty\,\right|\right|_{\psi_q}\right)^2 \\
&\lesssim \log(p+1)\sum_{r\in v}\left(\log^{1/q}\left((p+1)^2+1\right)\left|\left|\,||\xi_r||_{\psi_q}\,\right|\right|_\infty\right)^2 \\
&\lesssim \log(p+1)\frac{1}{|v|^2}\sum_{r\in v}\left(\log^{1/q}\left((p+1)^2+1\right)C_q\right)^2 \\
&\leq C'\log(p+1)\frac{1}{|v|^2}\sum_{r\in v}\log^{2/q}\left((p+1)^2+1\right) \\
&\leq C\frac{(\log p)^{1+2/q}}{|v|}.
\end{aligned}$$

Note that ψ_q is the Orlicz norm induced by the measure μ_n and the third inequality follows by 2. □

Corollary 1. *By the definition of* μ_n*,*

$$\mathbb{P}\left(\left|||Y||_2^2 - n\mathbb{E}[Y_1^2]\right| \geq n\delta\right) \leq 2\exp\left(-cn\min\left\{\frac{\delta^2}{C_q'^2}, \frac{\delta}{C_q'}\right\}\right),$$

where $c = 1/8e^2$ *is an absolute constant and* $C_q' = (\log 2)^{1/q-1}C_q$*.*

Proof. This result follows immediately from 3 and the result

$$||\xi||_{\psi_1} \leq (\log 2)^{1/q-1}||\xi||_{\psi_q} \leq (\log 2)^{1/q-1}C_q = C_q'.$$

□

S4 Proof of Main Results

Theorem 4. Let D_n, E_n be any two sets. Then we can make the following decomposition:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}(\widehat{t}, t_n) \geq \delta\right) &= \mathbb{P}\left(\mathcal{E}(\widehat{t}, t_n) \geq \delta \cap D_n \cap E_n\right) + \mathbb{P}\left(\mathcal{E}(\widehat{t}, t_n) \geq \delta \cap D_n^c \cap E_n\right) + \\ &\quad + \mathbb{P}\left(\mathcal{E}(\widehat{t}, t_n) \geq \delta \cap D_n \cap E_n^c\right) + \mathbb{P}\left(\mathcal{E}(\widehat{t}, t_n) \geq \delta \cap D_n^c \cap E_n^c\right) \\ &\leq \mathbb{P}\left(\mathcal{E}(\widehat{t}, t_n) \geq \delta \cap D_n \cap E_n\right) + 2\mathbb{P}\left(D_n^c\right) + \mathbb{P}\left(E_n^c\right). \quad \text{(S4.3)}\end{aligned}$$

Also,

$$\begin{aligned}\mathcal{E}(\widehat{t}, t_n) &= R\left(\widehat{\beta}_{\widehat{t}}\right) - R\left(\beta_{t_n}\right) \\ &= \underbrace{R\left(\widehat{\beta}_{\widehat{t}}\right) - \widehat{R}_{V_n}\left(\widehat{t}\right)}_{(I)} + \underbrace{\widehat{R}_{V_n}\left(\widehat{t}\right) - \widehat{R}_{V_n}\left(t_{\max}\right)}_{(II)} + \\ &\quad + \underbrace{\widehat{R}_{V_n}\left(t_{\max}\right) - \widehat{R}\left(\widehat{\beta}_{t_n}\right)}_{(III)} + \underbrace{\widehat{R}\left(\widehat{\beta}_{t_n}\right) - R\left(\beta_{t_n}\right)}_{(IV)}, \quad \text{(S4.4)}\end{aligned}$$

where we use the notation $\widehat{\beta}_t = \widehat{\beta}(\mathcal{B}_t)$. Now, for any $t \in T_n$, $\widehat{R}_{V_n}\left(\widehat{t}\right) - \widehat{R}_{V_n}(t) \leq 0$, by the definition of $\widehat{t}$, and thus $(II) < 0$.

Let $D_n := \left\{t_{\max} \leq 2nC_q'/a_n\right\}$ and $E_n := \{t_{\max} \geq t_n\}$. On the set E_n, $(III) \leq \widehat{R}_{V_n}(t_{\max}) - \widehat{R}\left(\widehat{\beta}_{t_{\max}}\right) =: \widetilde{(III)}$. Taking the first term from Equation (S4.3) and combining it with the decomposition in Equation (S4.4), we see that

$$\mathbb{P}\left(\mathcal{E}(\widehat{t}, t_n) \geq \delta \cap D_n \cap E_n\right) \leq \mathbb{P}\Bigg((I) \geq \delta/3 \cap D_n\Bigg) + \mathbb{P}\Bigg(\widetilde{(III)} \geq \delta/3 \cap D_n \cap E_n\Bigg) + \mathbb{P}\Bigg((IV) \geq \delta/3\Bigg) \tag{S4.5}$$

We break the remainder of this section into parts based on these terms.

Final predictor and cross-validation risk (I)**:** Using the notation introduced in S1, note that by equations (S1.1) and (S1.2)

$$\begin{aligned}
R\left(\widehat{\beta}_{\widehat{t}}\right) - \widehat{R}_{V_n}\left(\widehat{t}\right) &= \widehat{\gamma}_{\widehat{t}}^\top \Sigma_n \widehat{\gamma}_{\widehat{t}} - \frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \widehat{\Sigma}_v \widehat{\gamma}_{\widehat{t}}^{(v)} \\
&= \left[\widehat{\gamma}_{\widehat{t}}^\top \Sigma_n \widehat{\gamma}_{\widehat{t}} - \widehat{\gamma}_{\widehat{t}}^\top \left(\widehat{\Sigma}_n\right) \widehat{\gamma}_{\widehat{t}}\right] + \left[\widehat{\gamma}_{\widehat{t}}^\top \left(\widehat{\Sigma}_n\right) \widehat{\gamma}_{\widehat{t}} - \frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \widehat{\Sigma}_v \widehat{\gamma}_{\widehat{t}}^{(v)}\right]
\end{aligned}$$

Addressing each of the terms in order,

$$\left[\widehat{\gamma}_{\widehat{t}}^\top \Sigma_n \widehat{\gamma}_{\widehat{t}} - \widehat{\gamma}_{\widehat{t}}^\top \left(\widehat{\Sigma}_n\right) \widehat{\gamma}_{\widehat{t}}\right] = \widehat{\gamma}_{\widehat{t}}^\top \left(\Sigma_n - \widehat{\Sigma}_n\right) \widehat{\gamma}_{\widehat{t}} \leq ||\widehat{\gamma}_{\widehat{t}}||_1^2 \left|\left|\Sigma_n - \widehat{\Sigma}_n\right|\right|_\infty,$$

where the inequality follows by Lemma 4.

Likewise,

$$\begin{aligned}
&\left[\widehat{\gamma}_{\widehat{t}}^\top \left(\widehat{\Sigma}_n\right) \widehat{\gamma}_{\widehat{t}} - \frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \widehat{\Sigma}_v \widehat{\gamma}_{\widehat{t}}^{(v)}\right] \\
&= \left(\widehat{\gamma}_{\widehat{t}}^\top \widehat{\Sigma}_n \widehat{\gamma}_{\widehat{t}} - \frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \widehat{\Sigma}_n \widehat{\gamma}_{\widehat{t}}^{(v)}\right) + \left(\frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \widehat{\Sigma}_n \widehat{\gamma}_{\widehat{t}}^{(v)} - \frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \widehat{\Sigma}_v \widehat{\gamma}_{\widehat{t}}^{(v)}\right) \\
&= \frac{1}{K}\sum_{v \in V_n} \left(\widehat{\gamma}_{\widehat{t}}^\top \widehat{\Sigma}_n \widehat{\gamma}_{\widehat{t}} - (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \widehat{\Sigma}_n \widehat{\gamma}_{\widehat{t}}^{(v)}\right) + \left(\frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \left(\widehat{\Sigma}_n - \widehat{\Sigma}_v\right) \widehat{\gamma}_{\widehat{t}}^{(v)}\right) \\
&\leq \frac{1}{K}\sum_{v \in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \left(\widehat{\Sigma}_n - \widehat{\Sigma}_v\right) \widehat{\gamma}_{\widehat{t}}^{(v)}.
\end{aligned}$$

The last inequality follows as $\widehat{\gamma}_{\widehat{t}}$ is chosen to minimize $\widehat{\gamma}_{\widehat{t}}^\top \widehat{\Sigma}_n \widehat{\gamma}_{\widehat{t}}$, and so for any $v \in V_n$,

$$\widehat{\gamma}_{\widehat{t}}^\top \widehat{\Sigma}_n \widehat{\gamma}_{\widehat{t}} \leq (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \widehat{\Sigma}_n \widehat{\gamma}_{\widehat{t}}^{(v)}. \tag{S4.6}$$

Continuing and using Lemma 4,

$$\begin{aligned}\frac{1}{K}\sum_{v\in V_n} (\widehat{\gamma}_{\widehat{t}}^{(v)})^\top \left(\widehat{\Sigma}_n - \widehat{\Sigma}_v\right) \widehat{\gamma}_{\widehat{t}}^{(v)} &\leq \frac{1}{K}\sum_{v\in V_n} \left|\left|\widehat{\gamma}_{\widehat{t}}^{(v)}\right|\right|_1^2 \left|\left|\widehat{\Sigma}_n - \widehat{\Sigma}_v\right|\right|_\infty \\ &\leq \frac{1}{K}\sum_{v\in V_n} \left|\left|\widehat{\gamma}_{\widehat{t}}^{(v)}\right|\right|_1^2 \left(\left|\left|\widehat{\Sigma}_n - \Sigma_n\right|\right|_\infty + \left|\left|\Sigma_n - \widehat{\Sigma}_v\right|\right|_\infty\right)\end{aligned}$$

Therefore,

$$\begin{aligned}(I) &\leq ||\widehat{\gamma}_{\widehat{t}}||_1^2 \left|\left|\Sigma_n - \widehat{\Sigma}_n\right|\right|_\infty + \frac{1}{K}\sum_{v\in V_n} \left(\widehat{\gamma}_{\widehat{t}}^{(v)}\right)^\top \left(\widehat{\Sigma}_n - \widehat{\Sigma}_v\right) \widehat{\gamma}_{\widehat{t}}^{(v)} \\ &\leq (1+t_{\max})^2 \left(2\left|\left|\Sigma_n - \widehat{\Sigma}_n\right|\right|_\infty + \frac{1}{K}\sum_{v\in V_n} \left|\left|\Sigma_n - \widehat{\Sigma}_v\right|\right|_\infty\right).\end{aligned}$$

By 5 with $V_n = \{\{1, \ldots, n\}\}$ and $c_n = n$,

$$\mathbb{E}\left|\left|\Sigma_n - \widehat{\Sigma}_n\right|\right|_\infty \leq C_1 \sqrt{\frac{(\log p)^{1+2/q}}{n}},$$

while taking $V_n = \{v_1, \ldots, v_K\}$ shows,

$$\frac{1}{K}\sum_{v\in V_n} \mathbb{E}\left|\left|\Sigma_n - \widehat{\Sigma}_v\right|\right|_\infty \leq C_2 \sqrt{\frac{(\log p)^{1+2/q}}{c_n}}.$$

Combining these two bounds together gives

$$\begin{aligned}&\mathbb{P}\left((I) \geq \frac{\delta}{3} \cap D_n\right) \leq \frac{3}{\delta}\mathbb{E}[(I)\,\mathbf{1}_{D_n}] \\ &\leq \frac{3}{\delta}\left(1 + \frac{2n(\log 2)^{1/q-1}C_q)}{a_n}\right)^2 \left(2C_1\sqrt{\frac{(\log p)^{1+2/q}}{n}} + C_2\sqrt{\frac{(\log p)^{1+2/q}}{c_n}}\right).\end{aligned} \tag{S4.7}$$

Cross-validation risk and empirical risk (III): Due to the discussion following Equation (S4.4), it is sufficient to bound $\widetilde{(III)}$ instead. Recall that $\widehat{\Sigma}_{(v)} = (n - c_n)^{-1} \sum_{r \notin v} Z_r Z_r^\top$.

Then,

$$\begin{aligned}
\widehat{R}_{V_n}(t_{\max}) - \widehat{R}\left(\widehat{\beta}_{t_{\max}}\right) &= \frac{1}{K} \sum_{v \in V_n} (\widehat{\gamma}^{(v)}_{t_{\max}})^\top \widehat{\Sigma}_v \widehat{\gamma}^{(v)}_{t_{\max}} - \widehat{\gamma}^\top_{t_{\max}} \widehat{\Sigma}_n \widehat{\gamma}_{t_{\max}} \\
&= \frac{1}{K} \sum_{v \in V_n} \left((\widehat{\gamma}^{(v)}_{t_{\max}})^\top \widehat{\Sigma}_v \widehat{\gamma}^{(v)}_{t_{\max}} - (\widehat{\gamma}^{(v)}_{t_{\max}})^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}^{(v)}_{t_{\max}} \right) + \\
&\quad + \frac{1}{K} \sum_{v \in V_n} \left((\widehat{\gamma}^{(v)}_{t_{\max}})^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}^{(v)}_{t_{\max}} - \widehat{\gamma}^\top_{t_{\max}} \widehat{\Sigma}_n \widehat{\gamma}_{t_{\max}} \right) \\
&\leq \frac{1}{K} \sum_{v \in V_n} \left|\left| \widehat{\gamma}^{(v)}_{t_{\max}} \right|\right|_1^2 \left|\left| \widehat{\Sigma}_v - \widehat{\Sigma}_{(v)} \right|\right|_\infty ,
\end{aligned}$$

where the inequality follows by Lemma 4 and the fact that $\widehat{\gamma}^{(v)}_{t_*}$ is chosen to minimize $(\widehat{\gamma}^{(v)}_{t_*})^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}^{(v)}_{t_*}$, which implies

$$(\widehat{\gamma}^{(v)}_{t_{\max}})^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}^{(v)}_{t_{\max}} \leq \widehat{\gamma}^\top_{t_{\max}} \widehat{\Sigma}_{(v)} \widehat{\gamma}_{t_{\max}}. \tag{S4.8}$$

As before,

$$\frac{1}{K} \sum_{v \in V_n} \left|\left| \widehat{\gamma}^{(v)}_{t_{\max}} \right|\right|_1^2 \left|\left| \widehat{\Sigma}_v - \widehat{\Sigma}_{(v)} \right|\right|_\infty \leq (1+t_{\max})^2 \frac{1}{K} \sum_{v \in V_n} \left(\left|\left| \Sigma_n - \widehat{\Sigma}_v \right|\right|_\infty + \left|\left| \Sigma_n - \widehat{\Sigma}_{(v)} \right|\right|_\infty \right).$$

We can use a straight-forward adaptation of Lemma 5 to show that

$$\frac{1}{K} \sum_{v \in V_n} \mathbb{E} \left|\left| \Sigma_n - \widehat{\Sigma}_{(v)} \right|\right|_\infty \leq C_3 \sqrt{\frac{(\log p)^{1+2/q}}{n - c_n}}.$$

Therefore,

$$\mathbb{P}\left(\widetilde{(III)} \geq \delta/3 \cap D_n \cap E_n \right) \leq \frac{3}{\delta} \mathbb{E}[\widetilde{(III)}\, \mathbf{1}_{D_n}]$$

$$\leq \frac{3}{\delta}\left(1+\frac{2nC'_q}{a_n}\right)^2\left(C_2\sqrt{\frac{(\log p)^{1+2/q}}{c_n}}+C_3\sqrt{\frac{(\log p)^{1+2/q}}{n-c_n}}\right). \tag{S4.9}$$

Empirical risk and expected risk (IV) The proof of these results is given by Greenshtein and Ritov (2004). We include a somewhat different proof for completeness. Observe

$$\begin{aligned} R\left(\widehat{\beta}_{t_n}\right)-R\left(\beta_{t_n}\right) &= R\left(\widehat{\beta}_{t_n}\right)-\widehat{R}\left(\widehat{\beta}_{t_n}\right)+\widehat{R}\left(\widehat{\beta}_{t_n}\right)-R\left(\beta_{t_n}\right) \\ &\leq R\left(\widehat{\beta}_{t_n}\right)-\widehat{R}\left(\widehat{\beta}_{t_n}\right)+\widehat{R}\left(\beta_{t_n}\right)-R\left(\beta_{t_n}\right) \\ &\leq 2\sup_{\beta\in\mathcal{B}_{t_n}}\left|R\left(\beta\right)-\widehat{R}\left(\beta\right)\right|. \end{aligned}$$

Using Lemma 4 (See S1 for notation)

$$\sup_{\beta\in\mathcal{B}_{t_n}}\left|R\left(\beta\right)-\widehat{R}\left(\beta\right)\right| \leq \sup_{\beta\in\mathcal{B}_{t_n}}||\gamma||_1^2\left|\left|\widehat{\Sigma}_n-\Sigma_n\right|\right|_\infty \leq (1+t_n)^2\left|\left|\widehat{\Sigma}_n-\Sigma_n\right|\right|_\infty.$$

Therefore,

$$\mathbb{P}\Big((IV)\geq\delta/3\Big)\leq\frac{3}{\delta}\left(1+t_n\right)^2\left(C_4\sqrt{\frac{(\log p)^{1+2/q}}{n}}\right). \tag{S4.10}$$

The proof follows by combining Equation (S4.3) with Equation (S4.5) and using the bounds from Equations (S4.7), (S4.9), and (S4.10).

Lastly, there are the various constants incurred in the course of this proof. In Lemma 2, the constant Ψ can be chosen arbitrarily small based on inspecting the proof of Lemma 2.2.2 in van der Vaart and Wellner (1996). As this constant premultiplies every term in $\Omega_{n,1}$ and $\Omega_{n,2}$, we can without loss of generality set the constant equal to one. For instance, in Lemma 5, the constant C is upper bounded by $8e^{1/2}(\log 2)^{(2-q)/(2q)}8^{1/q}C_q\Psi$. This constant can be taken arbitrarily small by choosing Ψ small enough. □

Lemma 6. *Define the set* $D_n := \left\{t_{\max} \leq 2nC'_q/a_n\right\}$, *where* a_n *is the normalizing rate defined in Corollary 5 and* $C'_q = C_q(\log 2)^{1/q-1}$. *Then,* $\mathbb{P}(D_n^c) \leq e^{-cn}$.

Proof. By Corollary 1,

$$\mathbb{P}\left(\frac{||Y||_2^2}{a_n} \geq \frac{n(\mathbb{E}[Y_1^2]+\delta)}{a_n}\right) \leq \exp\left(-cn\min\left\{\frac{\delta^2}{C_q'^2}, \frac{\delta}{C'_q}\right\}\right).$$

Furthermore, $\mathbb{E}[Y_1^2] \leq C'_q$, so

$$\mathbb{P}\left(\frac{||Y||_2^2}{a_n} \geq \frac{n(C'_q+\delta)}{a_n}\right) \leq \mathbb{P}\left(\frac{||Y||_2^2}{a_n} \geq \frac{n(\mathbb{E}[Y_1^2]+\delta)}{a_n}\right).$$

Therefore, setting $\delta = C'_q$ yields $\mathbb{P}(||Y||_2^2/a_n \geq 2nh/a_n) \leq e^{-cn}$. □

Lemma 7. *Define the set* $E_n := \{t_{\max} \geq t_n\}$. *If* $a_n t_n = o(n)$, *then, for all* $n > 2a_n t_n/\mathbb{E}[Y_1^2]$, $\mathbb{P}(E_n^c) \leq \exp\{-cn\}$.

Proof. By Corollary 1,

$$\mathbb{P}\left(\frac{||Y||_2^2}{a_n} \leq \frac{n(\mathbb{E}[Y_1^2]-\delta)}{a_n}\right) \leq \exp\left(-cn\min\left\{\frac{\delta^2}{C_q'^2}, \frac{\delta}{C'_q}\right\}\right),$$

for all $\delta > 0$. Setting $\delta = \mathbb{E}[Y_1^2] - a_n t_n/n$ therefore implies

$$\mathbb{P}\left(t_{\max} \leq t_n\right) \leq \exp\left(-cn \min\left\{\frac{(\mathbb{E}[Y_1^2] - a_n t_n/n)^2}{C_q'^2}, \frac{\mathbb{E}[Y_1^2] - a_n t_n/n}{C_q'}\right\}\right)$$
$$\leq \exp\left(-cn \min\left\{\frac{\mathbb{E}[Y_1^2]^2}{4C_q'^2}, \frac{\mathbb{E}[Y_1^2]}{2C_q'}\right\}\right).$$

Since $0 < \mathbb{E}[Y_1^2]/C_q' \leq 1$, the result follows. □

Corollary 6. For the $\sqrt{\text{lasso}}$, the result is nearly immediate as we are considering the same constraint set $||\beta||_1 \leq t$ and the same search space for the tuning parameter $T = [0,\ ||Y||_2^2/a_n]$. However, in Equations (S4.6) and (S4.8), we rely on the empirical minimizer. The analogous results here are $\widehat{\gamma}_{\hat{t}}^\top \widehat{\Sigma}_n \widehat{\gamma}_{\hat{t}} \leq (\widehat{\gamma}_{\hat{t}}^{(v)})^\top \widehat{\Sigma}_n \widehat{\gamma}_{\hat{t}}^{(v)}$ and $(\widehat{\gamma}_{t_{\max}}^{(v)})^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}_{t_{\max}}^{(v)} \leq \widehat{\gamma}_{t_{\max}}^\top \widehat{\Sigma}_{(v)} \widehat{\gamma}_{t_{\max}}$ respectively, but this implies that (S4.6) and (S4.8) hold.

For the group lasso with $\max_g \sqrt{|g|} = O(1)$, we have $t \geq \sum_{g\in G} \sqrt{|g|}\, ||\beta_g||_2 \geq ||\beta||_1$ so that Lemma 7 still applies with $t_{\max}$ as before. We note that in this case, the oracle group linear model is restricted to the ball $\mathcal{B}_{t_n} = \{\beta : \sum_{g\in G} \sqrt{|g|}\, ||\beta_g||_2 \leq t_n\}$ rather than the larger set $\{\beta : ||\beta||_1 \leq t_n\}$. □

References

Dümbgen, L., van de Geer, S. A., Veraar, M. C., and Wellner, J. A. (2010). Nemirovski's inequalities revisited. *American Mathematical Monthly* **117**, 138–160.

Greenshtein, E., and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971–988.

van der Vaart, A. W., and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes* Springer, New York.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. in *Compressed Sensing: Theory and Applications*, eds. Y. C. Eldar and G. Kutyniok pp. 210–268 Cambridge University Press.