A Auxiliary Results

Proposition A.1 (Theorem 2.3. Bellec [2018] Restated)**.** Let $\mathcal{C}$ be a closed convex subset of $\mathbb{R}^n$. Suppose one has the model $\boldsymbol{O} = \boldsymbol{\theta} + \mathbf{e}$, and the estimate based on $\operatorname{argmin}_{\mathbf{v}\in\mathcal{C}} \|\boldsymbol{O} - \mathbf{v}\|_2^2$ is denoted by $\widehat{\boldsymbol{\theta}}$. If for some $\mathbf{u} \in \mathcal{C}$ there exists $t_*(\mathbf{u})$ so that the error vector $\mathbf{e}$ satisfies

$$\sup_{\mathbf{v}\in\mathcal{C}, \|\mathbf{v}-\mathbf{u}\|_2 \le t_*(\mathbf{u})} \mathbf{e}^\top(\mathbf{v} - \mathbf{u}) \le \frac{t_*^2(\mathbf{u})}{2} + Ct_*(\mathbf{u})\sqrt{2x},$$

with probability at least $1 - \exp(-x)$ then

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \le \|\mathbf{u} - \boldsymbol{\theta}\|_2^2 + 2t_*^2(\mathbf{u}) + 4C^2 x,$$

with probability at least $1 - \exp(-x)$.

Lemma A.2 (Fano's inequality)**.** Let (Θ, d) be a metric space, and $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a collection of probability distributions. Then

$$\inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathbb{P}_\theta(d^2(\widehat{\theta}, \theta) \ge \varepsilon^2/4) \ge 1 - \frac{\sup_{\theta,\theta'\in T} D_{\mathrm{KL}}(\mathbb{P}_\theta || \mathbb{P}_{\theta'}) + \log 2}{\log \mathcal{N}(\varepsilon, T, d)},$$

where $T \subset \Theta$ is a totally bounded set, and $\mathcal{N}(\varepsilon, T, d)$ is the packing number of T with respect to d.

B Proofs

Proof of Lemma 2.1. First note that the solution to (2.2) will always satisfy $\widehat{p}_0 \ge 0$ and $\widehat{p}_n \le 1$ since all $0 \le O_i \le 1$.

The proof proceeds to show that if for two probabilities $p_i \le p_{i+1}$ we have $O_i \ge O_{i+1}$ it follows that their final estimates are equal, i.e., $\widehat{p}_i = \widehat{p}_{i+1}$. Therefore one clumps O_i and O_{i+1} and treats the two probabilities as the same one, and solves a similar problem with less parameters. Since this observation gives raise to the pool adjacent violators algorithm (PAVA) [Mair et al., 2009] for optimizing the loss function (2.1), which is the same algorithm for optimizing (2.2) the two solutions must coincide.

Suppose that indeed we have $O_i \ge O_{i+1}$ and $\widehat{p}_i < \widehat{p}_{i+1}$. We will arrive at a contradiction by showing that one can add and subtract a small c and increase the loss function. Consider the function

$$\begin{aligned} c \mapsto\ & O_i \log(\widehat{p}_i + c) + (1 - O_i)\log(1 - \widehat{p}_i - c) \\ & + O_{i+1}\log(\widehat{p}_{i+1} - c) + (1 - O_{i+1})\log(1 - \widehat{p}_{i+1} + c). \end{aligned}$$

Taking the derivative with respect to c yields

$$\frac{O_i}{\widehat{p}_i + c} - \frac{1 - O_i}{1 - \widehat{p}_i - c} - \frac{O_{i+1}}{\widehat{p}_{i+1} - c} + \frac{1 - O_{i+1}}{1 - \widehat{p}_{i+1} + c} \ge 0,$$

if c is small enough so that $\widehat{p}_i + c \leq \widehat{p}_{i+1} - c$. Hence the function is increasing in c which is a contradiction. This proves that $\widehat{p}_i = \widehat{p}_{i+1}$. We note that the proof extends to any $0 \leq O_i \leq 1$, and even if one has weights, i.e., if one optimizes:

$$\text{argmax} \sum_{i \in [n]} w_i O_i \log p_i + w_i (1 - O_i) \log(1 - p_i);$$

(in this case instead of c one needs to consider c/w_i and c/w_{i+1} respectively, and the regression will also have weights).

The fact that (2.3) holds[2] is well known [see Chapter 1 of Robertson et al.]. $\square$

Proof of Theorem 2.3. We will use Proposition A.1. We need to control the tails of the process:

$$Z = \sup_{\mathbf{v} \in \mathcal{S}_n^{\uparrow}, \|\mathbf{v} - \mathbf{u}\|_2 \leq t} \sum_{i \in [n]} (O_i - p_i)(v_i - u_i).$$

We will first argue that Z is close to its expected value, using Theorem 6.7 [Boucheron et al., 2013], and in the second step we will control the expected value of Z. To this end define

$$Z_j = \inf_{o_j \in \{0,1\}} \sup_{\mathbf{v} \in \mathcal{S}_n^{\uparrow}, \|\mathbf{v} - \mathbf{u}\|_2 \leq t} \sum_{i \neq j} (O_i - p_i)(v_i - u_i) + (o_j - p_j)(v_j - u_j),$$

and note that

$$(Z - Z_i)^2 \leq (v_i^* - u_i)^2,$$

where $\mathbf{v}^*$ denotes the value where the $\sup_{\mathbf{v} \in \mathcal{S}_n^{\uparrow}, \|\mathbf{v} - \mathbf{u}\|_2 \leq t} \sum_{i \in [n]} (O_i - p_i)(v_i - u_i)$ is attained (the sup is attained since the set $\mathcal{S}_n^{\uparrow} \cap \{\mathbf{v} : \|\mathbf{v} - \mathbf{u}\|_2 \leq t\}$ is compact). It therefore follows that

$$\sum_{i \in [n]} (Z - Z_i)^2 \leq \sum_{i \in [n]} (v_i^* - u_i)^2 \leq t^2.$$

By Theorem 6.7 [Boucheron et al., 2013] we have

$$\mathbb{P}(Z \geq \mathbb{E}Z + y) \leq e^{-y^2/2t^2},$$

and hence setting $y = \sqrt{2x}t$ we obtain that with probability at least $1 - e^{-x}$ we have

$$Z \leq \mathbb{E}Z + \sqrt{2x}t.$$

Next, using symmetrization as in the proof of Theorem 2.2 we obtain

$$\begin{aligned} \mathbb{E}Z &\leq 2\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{\mathbf{v} \in \mathcal{S}_n^{\uparrow}, \|\mathbf{v} - \mathbf{u}\|_2 \leq t} \sum_{i \in [n]} \varepsilon_i (v_i - u_i) \\ &\leq \sqrt{2\pi}\mathbb{E}_{\boldsymbol{\xi}} \sup_{\mathbf{v} \in \mathcal{S}_n^{\uparrow}, \|\mathbf{v} - \mathbf{u}\|_2 \leq t} \sum_{i \in [n]} \xi_i (v_i - u_i), \quad \text{(B.1)} \end{aligned}$$

where $\boldsymbol{\xi}$ is a standard Gaussian random vector. In Chatterjee et al. [2014] it is proved that the above quantity is $\leq t^2/16$ for values of $t \geq c(1 + V(\mathbf{u}))^{1/3} n^{1/6}$. This completes the proof by Proposition A.1. $\square$

Proof of Theorem 2.4. Our proof will follow the proof of Theorem 2.2 of Guntuboyina and Sen [2017] where modifications are needed since the errors are not i.i.d. Gaussian as required in the original statement. First note that the second inequality follows from the first by a simple application of Jensen's inequality, hence we focus on showing the first inequality. We note that by Lemma 2.1, and any integer k

$$\begin{aligned} \widehat{p}_j &= \min_{v \geq j} \max_{u \leq j} \overline{O}_{uv} \leq \max_{u \leq j} (\overline{p}_{u,j+k} + \overline{O}_{u,j+k} - \overline{p}_{u,j+k}) \\ &\leq \overline{p}_{j,j+k} + \max_{u \leq j} (\overline{O}_{u,j+k} - \overline{p}_{u,j+k}), \end{aligned}$$

where the last inequality follows by the monotonicity of $\mathbf{p}$. Hence

$$\widehat{p}_j - p_j \leq (\overline{p}_{j,j+k} - p_j) + \max_{u \leq j} (\overline{O}_{u,j+k} - \overline{p}_{u,j+k}),$$

which implies

$$\begin{aligned} \mathbb{E}(\widehat{p}_j - p_j)_+^p \leq \mathbb{E}((&\overline{p}_{j,j+k} - p_j) \\ &+ \max_{u \leq j} (\overline{O}_{u,j+k} - \overline{p}_{u,j+k}))_+^p, \end{aligned}$$

Now let $N_1, N_2, \ldots, N_m$ denote the indices of the m different equal probabilities. Take $j \in N_k$, and let there be l_k numbers to the left of j and r_k numbers to the right of j in N_k (i.e. $\max_{i \in N_k} i = j + r_k, \min_{i \in N_k} i = j - l_k$). Note that since all probabilities on N_k are the same we have $\overline{p}_{j,j+r_k} = p_j$ and therefore

$$\mathbb{E}(\widehat{p}_j - p_j)_+^p \leq \mathbb{E}(\max_{u \leq j} (\overline{O}_{u,j+r_k} - \overline{p}_{u,j+r_k}))_+^p.$$

Here it is necessary for the proof to depart substantially from the original argument as the sequence $(\overline{O}_{u,j+r_k} - \overline{p}_{u,j+r_k})$ does not have the required i.i.d. structure. We start by symmetrizing the function similarly to the proof of Theorem 2.2. Let $\widetilde{O}_i$ be i.i.d. copies of O_i. Note that since $(\cdot)_+^p$ is convex we have

$$\begin{aligned} &\mathbb{E}(\max_{u \leq j} (\overline{O}_{u,j+r_k} - \overline{p}_{u,j+r_k}))_+^p \\ &\qquad = \mathbb{E}_{\mathbf{O}} \left(\max_{u \leq j} \frac{\sum_{i=u}^{j+r_k} (O_i - \mathbb{E}\widetilde{O}_i)}{j + r_k - u + 1} \right)_+^p \\ &\qquad \leq \mathbb{E}_{\mathbf{O}, \widetilde{\mathbf{O}}} \left(\max_{u \leq j} \frac{\sum_{i=u}^{j+r_k} (O_i - \widetilde{O}_i)}{j + r_k - u + 1} \right)_+^p, \end{aligned}$$

where the last expectation is taken with respect to both O_i and $\widetilde{O}_i$. We can introduce random sign ε_i since the

[2](2.3) holds for unweighted regression only

distributions $O_i - \widetilde{O}_i$ are symmetric.

$$\begin{aligned}
&\mathbb{E}(\max_{u \le j}(\overline{O}_{u,j+r_k} - \overline{p}_{u,j+r_k}))_+^p \\
&\le \mathbb{E}_{\mathbf{O},\widetilde{\mathbf{O}}}\left(\max_{u \le j} \frac{\sum_{i=u}^{j+r_k} \varepsilon_i (O_i - \widetilde{O}_i)}{j + r_k - u + 1}\right)_+^p \\
&= \mathbb{E}_{\mathbf{O},\widetilde{\mathbf{O}},\boldsymbol{\varepsilon}}\left(\max_{u \le j} \frac{\sum_{i=u}^{j+r_k} \varepsilon_i (O_i - \widetilde{O}_i)}{j + r_k - u + 1}\right)_+^p, \qquad \text{(B.2)}
\end{aligned}$$

where in the last equality the expectation is taken with respect to the ε_i as well. Using the convexity of $(\cdot)_+^p$, the properties of max and sign symmetry we obtain

$$\begin{aligned}
&\mathbb{E}(\max_{u \le j}(\overline{O}_{u,j+r_k} - \overline{p}_{u,j+r_k}))_+^p \\
&\le \mathbb{E}_{\mathbf{O},\boldsymbol{\varepsilon}}\left(2\max_{u \le j} \frac{\sum_{i=u}^{j+r_k} \varepsilon_i O_i}{j + r_k - u + 1}\right)_+^p \\
&\le \mathbb{E}_{\boldsymbol{\varepsilon}}\left(2\max_{u \le j} \frac{\sum_{i=u}^{j+r_k} \varepsilon_i}{j + r_k - u + 1}\right)_+^p.
\end{aligned}$$

where in the last inequality we used the contraction principle (Theorem 11.6 Boucheron et al. [2013]). Importantly, note that the sequence of random variables (indexed by u) $\frac{\sum_{i=u}^{j+r_k}\varepsilon_i}{j+r_k-u+1}$ forms a martingale.

Consider first the case $1 < p < 2$. By Doob's L^p maximal inequality for submartingales [Mörters and Peres, 2010] (which holds for $p > 1$) and Khintchine's inequality we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}}\left(2\max_{u \le j} \frac{\sum_{i=u}^{j+r_k} \varepsilon_i}{j + r_k - u + 1}\right)_+^p \\
&\le 2^p(\frac{p}{p-1})^p \mathbb{E}_{\boldsymbol{\varepsilon}}\left(\frac{\sum_{i=j}^{j+r_k} \varepsilon_i}{r_k + 1}\right)_+^p \\
&\le 2^p(\frac{p}{p-1})^p B_p^p \left(\frac{1}{r_k+1}\right)^{p/2},
\end{aligned}$$

where B_p is the upper constant from Khintchine's inequality. Therefore we conclude that:

$$\mathbb{E}(\widehat{p}_j - p_j)_+^p \le 2^p(\frac{p}{p-1})^p B_p^p \frac{1}{(r_k+1)^{p/2}}.$$

Using similar arguments one can also argue that

$$\mathbb{E}(\widehat{p}_j - p_j)_-^p \le 2^p(\frac{p}{p-1})^p B_p^p \frac{1}{(l_k+1)^{p/2}}.$$

Combining the two inequalities above and summing over all j we have

$$\begin{aligned}
\mathbb{E}\sum_{j \in [n]} (\widehat{p}_j - p_j)^p &\le 2^{p+1}(\frac{p}{p-1})^p B_p^p \sum_{k=1}^m \sum_{j \in [|N_k|]} \left(\frac{1}{j}\right)^{p/2} \\
&\le C_p \sum_{k=1}^m \frac{2}{2-p} |N_k|^{1-p/2},
\end{aligned}$$

which is what we wanted to show for the case $1 < p < 2$ (the last bound follows by simple integration).

When $p = 1$, Doob's maximal L^p inequality does not hold, and we need to slightly change the argument. We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}} 2\left(\max_{u \le j} \frac{\sum_{i=u}^{j+r_k} \varepsilon_i}{j + r_k - u + 1}\right)_+ \\
&\le \tau + \int_\tau^\infty \mathbb{P}_{\boldsymbol{\varepsilon}}\left(2\max_{u \le j}\left(\frac{\sum_{i=u}^{j+r_k} \varepsilon_i}{j + r_k - u + 1}\right)_+ \ge t\right) dt,
\end{aligned}$$

where we set $\tau = \sqrt{\frac{1}{r_k+1}}$. Since $\left(\frac{\sum_{i=u}^{j+r_k}\varepsilon_i}{j+r_k-u+1}\right)_+$ is a submartingale (as a convex function of a martingale) Doob's weak maximal inequality [Mörters and Peres, 2010] gives

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\varepsilon}}\left(2\max_{u \le j}\left(\frac{\sum_{i=j}^{j+r_k} \varepsilon_i}{j + r_k - u + 1}\right)_+ \ge t\right) \\
&\le \frac{\mathbb{E}_{\boldsymbol{\varepsilon}}\left(2\left(\frac{\sum_{i=j}^{j+r_k}\varepsilon_i}{r_k+1}\right)_+ \mathbb{1}\left(2\max_{u\le j}\left(\frac{\sum_{i=u}^{j+r_k}\varepsilon_i}{j+r_k-u+1}\right)_+ \ge t\right)\right)}{t}.
\end{aligned}$$

We square the preceding display and apply Cauchy-Schwartz, followed by an application of Khintchine's inequality to obtain

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\varepsilon}}\left(2\max_{u \le j}\left(\frac{\sum_{i=j}^{j+r_k} \varepsilon_i}{j + r_k - u + 1}\right)_+ \ge t\right) \\
&\le \frac{4\mathbb{E}_{\boldsymbol{\varepsilon}}\left(\frac{\sum_{i=j}^{j+r_k}\varepsilon_i}{r_k+1}\right)^2}{t^2} \le \frac{4B_2^2 \frac{1}{r_k+1}}{t^2} = \frac{4B_2^2\tau^2}{t^2}.
\end{aligned}$$

Changing variables yields:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}}\left(2\max_{u \le j} \frac{\sum_{i=u}^{j+r_k} \varepsilon_i}{j + r_k - u + 1}\right)_+ \\
&\le \tau + 4B_2^2\tau \int_1^\infty \frac{1}{t^2} dt = (1 + 4B_2^2)\tau.
\end{aligned}$$

The rest of the proof goes through. □

Lemma B.1. The KL divergence between $P = Ber(p)$ and $Q = Ber(p + c)$ (for some $0 < |c| < \min(p, 1-p)$) is bounded as

$$\left| D_{\mathrm{KL}}(P||Q) - \frac{c^2}{p(1-p)} \right| \le \frac{|c|^3}{p(1-p)(1-p-|c|)(p-|c|)}.$$

Proof of Lemma B.1. The proof is a simple calculation which we include for completeness.

$$\begin{aligned}
&D_{\mathrm{KL}}(P||Q)\\
&= -p\log\left(1+\frac{c}{p}\right) - (1-p)\log\left(1-\frac{c}{1-p}\right)\\
&= p\sum_{i=1}^{\infty}(-1)^i i^{-1}\left(\frac{c}{p}\right)^i + (1-p)\sum_{i=1}^{\infty} i^{-1}\left(\frac{c}{1-p}\right)^i\\
&= \frac{c^2}{p(1-p)} + c\sum_{i\geq 3}\left[(-1)^i i^{-1}\left(\frac{c}{p}\right)^{i-1} + i^{-1}\left(\frac{c}{1-p}\right)^{i-1}\right].
\end{aligned}$$

It immediately follows that

$$\begin{aligned}
&\left|D_{\mathrm{KL}}(P||Q) - \frac{c^2}{p(1-p)}\right|\\
&\quad\leq |c|\sum_{i\geq 3}\left[\left(\frac{|c|}{p}\right)^{i-1} + \left(\frac{|c|}{1-p}\right)^{i-1}\right]\\
&\quad= \frac{|c|^3}{p^2(1-\frac{|c|}{p})} + \frac{|c|^3}{(1-p)^2(1-\frac{|c|}{1-p})}\\
&\quad\leq \frac{|c|^3}{p(1-p)(1-p-|c|)(p-|c|)}.
\end{aligned}$$

□

Proof of Proposition 2.5. For simplicity suppose that $n/m = k \in \mathbb{N}$. Fix a small $0 < \delta < 1$ and take the following base probability vector:

$$p_i = \delta + \alpha\lfloor (i-1)/k\rfloor,$$

for $i \in [n]$, where $\alpha = \frac{1-5/2\delta}{m-1}$. In this way $p_1 = \delta$ and $p_n = 1 - 3/2\delta$. Using the Varshamov-Gilbert Lemma [Tsybakov, 2009] construct a sequence on the cube $\{0,1\}^m$: $\mathcal{W} = \{\mathbf{w}_i\}_{i\in 0,1,\ldots,M}$ such that $d_{\mathrm{H}}(\mathbf{w}_i, \mathbf{w}_j) \geq \frac{m}{8}$ and $\log M \geq \frac{m}{8}$, where d_{H} denotes the Hamming distance. We perturb the probability vector $\mathbf{p}$ by adding $c\mathbf{w}$ for $\mathbf{w} \in \mathcal{W}$ to the corresponding coordinates:

$$p_i^{\mathbf{w}} = p_i + c w_{\lfloor (i-1)/k\rfloor + 1},$$

where $c < \alpha \wedge \frac{\delta}{2}$ and therefore keeps the relationship $p_i^{\mathbf{w}} \leq p_j^{\mathbf{w}}$ for $i \leq j$. For any two $\mathbf{w}$ and $\mathbf{w}'$ and $1 < p \leq 2$ we have the following bound

$$\frac{1}{n}\|\mathbf{p}^{\mathbf{w}} - \mathbf{p}^{\mathbf{w}'}\|_p^p \geq d_{\mathrm{H}}(\mathbf{w}, \mathbf{w}')c^p\frac{k}{n} \geq \frac{c^p}{8}.$$

Next, using Lemma B.1 the maximum KL divergence between vector valued Bernoulli random variables with probabilities equal to $\mathbf{p}^{\mathbf{w}}$ and $\mathbf{p}^{\mathbf{w}'}$ is bounded as

$$\begin{aligned}
&D_{\mathrm{KL}}(Ber(\mathbf{p}^{\mathbf{w}})||Ber(\mathbf{p}^{\mathbf{w}'}))\\
&\quad\leq d_{\mathrm{H}}(\mathbf{w}, \mathbf{w}')k\frac{c^2}{\delta(1-\delta)}\left(1 + \frac{2c}{(1-2\delta)\delta}\right)\\
&\quad\leq d_{\mathrm{H}}(\mathbf{w}, \mathbf{w}')k\frac{2c^2}{\delta(1-2\delta)} \leq \frac{2nc^2}{\delta(1-2\delta)}.
\end{aligned}$$

Using Fano's inequality (Lemma A.2) in conjunction with Markov's inequality we obtain the lower bound

$$\inf_{\widehat{\mathbf{p}}}\sup_{\mathbf{p}}\mathbb{E}\frac{1}{n}\|\widehat{\mathbf{p}} - \mathbf{p}\|_p^p \geq \frac{c^p}{16}\left(1 - \frac{16nc^2}{\delta(1-2\delta)m}\right),$$

which is what we wanted to show after selecting $c = \sqrt{\frac{\delta(1-2\delta)m}{32n}}$. □

Proof of Proposition 2.6. As in the proof of Theorem 2.2 we need to project onto the tangent cone $\mathcal{T}_{S_n^{\uparrow}}(\mathbf{u})$. The proof relies on symmetrization and the contraction principle. Decompose the i^{th} of the n binomials to sums $k\overline{O}_i = \sum_{j\in[k]} O_{ij}$ where $O_{ij} \sim Ber(p_i)$. We need to control the quantity

$$\mathbb{E}_{\mathbf{O}}\Big[\sup_{\mathbf{t}\in\mathcal{S}_l^{\uparrow}, \|\mathbf{t}\|_2\leq 1}\sum_{i\in[l]}\frac{\sum_{j\in[k]}(O_{ij} - p_i)}{k}t_i\Big]^2,$$

where using a slight abuse of notation we refer to $S^{\uparrow}_{|N_l|}$ with $S_l^{\uparrow}$ for brevity. Just as in Theorem 2.2, using symmetrization we obtain the following bound:

$$4\mathbb{E}_{\mathbf{O}}\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}_l^{\uparrow}, \|\mathbf{t}\|_2\leq 1}\sum_{i\in[l]}\frac{\sum_{j\in[k]}\varepsilon_{ij}O_{ij}}{k}t_i\Big]^2.$$

Using the contraction principle (Theorem 11.6 Boucheron et al. [2013]) we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{O}}\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}_l^{\uparrow}, \|\mathbf{t}\|_2\leq 1}\sum_{i\in[l]}\frac{\sum_{j\in[k]}\varepsilon_{ij}O_{ij}}{k}t_i\Big]^2\\
&\qquad\leq 4\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}_l^{\uparrow}, \|\mathbf{t}\|_2\leq 1}\sum_{i\in[l]}\frac{\sum_{j\in[k]}\varepsilon_{ij}}{\sqrt{k}}t_i\Big]^2\frac{1}{k}
\end{aligned}$$

Just as in the proof of Theorem 2.2 we can now substitute the Rademacher random variables with Gaussians:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}_l^{\uparrow}, \|\mathbf{t}\|_2\leq 1}\sum_{i\in[l]}\frac{\sum_{j\in[k]}\varepsilon_{ij}}{\sqrt{k}}t_i\Big]^2\\
&\qquad\leq \frac{\pi}{2}\mathbb{E}_{\boldsymbol{\varepsilon}}\Big[\sup_{\mathbf{t}\in\mathcal{S}_l^{\uparrow}, \|\mathbf{t}\|_2\leq 1}\sum_{i\in[l]}\frac{\sum_{j\in[k]}\xi_{ij}}{\sqrt{k}}t_i\Big]^2 = \frac{\pi}{2}\sum_{i\in[l]}\frac{1}{i},
\end{aligned}$$

where the last equality is well known [see Amelunxen et al., 2014, e.g.]. This completes the proof after applying Lemma 2.10. □

Proof of Proposition 2.8. The proof follows closely that of Theorem 2.3 hence we only sketch it. We need to control the tails of the process:

$$Z = \sup_{\mathbf{v}\in\mathcal{S}_n^{\uparrow}, \|\mathbf{v}-\mathbf{u}\|_2\leq t}\sum_{i\in[n]}\frac{\sum_j O_{ij} - p_i}{k}(v_i - u_i).$$

We will first argue that Z is close to its expected value, using Theorem 6.7 Boucheron et al. [2013], and in the second step we will control the expected value of Z. To this end define

$$Z_{rs} = \inf_{o_{rs}\in\{0,1\}} \sup_{\substack{\mathbf{v}\in\mathcal{S}_n^{\uparrow},\\ \|\mathbf{v}-\mathbf{u}\|_2\le t}} \sum_{i,j:(i,j)\neq(r,s)} \frac{\sum_j O_{ij}-p_i}{k}(v_i-u_i) + \frac{(o_{rs}-p_s)(v_s-u_s)}{k},$$

and note that

$$(Z-Z_{rs})^2 \le \frac{(v_s^*-u_s)^2}{k^2},$$

where $\mathbf{v}^*$ denotes the value where the $\sup_{\mathbf{v}\in\mathcal{S}_n^{\uparrow},\|\mathbf{v}-\mathbf{u}\|_2\le t}\sum_{i\in[n]}\frac{\sum_j O_{ij}-p_i}{k}(v_i-u_i)$. is attained. It therefore follows that

$$\sum_{r,s}(Z-Z_{rs})^2 \le \sum_{r,s}\frac{(v_i^*-u_i)^2}{k^2} \le \frac{t^2}{k}.$$

By Theorem 6.7 Boucheron et al. [2013] it follows that

$$\mathbb{P}(Z\ge \mathbb{E}Z+y)\le e^{-ky^2/2t^2},$$

and hence setting $y=\sqrt{2x}t$ we obtain that with probability at least $1-e^{-x}$ we have

$$Z\le \mathbb{E}Z+\sqrt{2x/k}t.$$

Next, using symmetrization and changing to Gaussian variables

$$\mathbb{E}Z \le \frac{\sqrt{2\pi}}{\sqrt{k}}\mathbb{E}_{\boldsymbol{\xi}} \sup_{\mathbf{v}\in\mathcal{S}_n^{\uparrow},\|\mathbf{v}-\mathbf{u}\|_2\le t}\sum_{i\in[n]}\xi_i(v_i-u_i), \tag{B.3}$$

where $\boldsymbol{\xi}$ is a standard Gaussian random vector. In Chatterjee et al. [2014] it is proved that the above quantity is $\le t^2/16$ for values of $t\ge c\frac{1}{\sqrt{k}}(1+V(\mathbf{u})\sqrt{k})^{1/3}n^{1/6}$. This completes the proof by Proposition A.1. □

Proof of Lemma 3.1. We have the following identity

$$\operatorname*{argmin}_{\boldsymbol{\beta}}\mathbb{E}(Y-\boldsymbol{X}^\top\boldsymbol{\beta})^2 = \operatorname*{argmax}_{\boldsymbol{\beta}} 2\mathbb{E}Y\boldsymbol{X}^\top\boldsymbol{\beta}-\|\boldsymbol{\beta}\|_2^2.$$

Recall that $\|\boldsymbol{\beta}^*\|_2=1$ and represent $\boldsymbol{\beta}=c\boldsymbol{\beta}^*+\boldsymbol{\beta}^\perp$ where $\boldsymbol{\beta}^{*\top}\boldsymbol{\beta}^\perp=0$. By the properties of the normal distribution we have

$$\mathbb{E}Y\boldsymbol{X}^\top\boldsymbol{\beta}^\perp = \mathbb{E}Y\mathbb{E}\boldsymbol{X}^\top\boldsymbol{\beta}^\perp=0.$$

Therefore by the Pythagorean theorem

$$\operatorname*{argmax}_{\boldsymbol{\beta}} 2c\mathbb{E}Y\boldsymbol{X}^\top\boldsymbol{\beta}^*-(c^2+\|\boldsymbol{\beta}^\perp\|_2^2) \le \operatorname*{argmax}_{\boldsymbol{\beta}} 2c\mathbb{E}Y\boldsymbol{X}^\top\boldsymbol{\beta}^*-c^2.$$

The above parabola is maximized at $c=c_0=\mathbb{E}Y\boldsymbol{X}^\top\boldsymbol{\beta}^*$, and therefore the population minimizer of the least squares satisfies (3.3). Using Chebyshev's association inequality [Boucheron et al., 2013] it is not hard to see that when f is strictly monotone increasing and Y is given by (3.1) we have

$$\begin{aligned}c_0 &= \mathbb{E}Y\boldsymbol{X}^\top\boldsymbol{\beta}^* = \mathbb{E}f(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\boldsymbol{X}^\top\boldsymbol{\beta}^*\\ &> \mathbb{E}f(\boldsymbol{X}^\top\boldsymbol{\beta}^*)\mathbb{E}\boldsymbol{X}^\top\boldsymbol{\beta}^*=0,\end{aligned}$$

and therefore $\operatorname{argmin}_{\boldsymbol{\beta}}\mathbb{E}(Y-\boldsymbol{X}^\top\boldsymbol{\beta})^2=c_0\boldsymbol{\beta}^*$ is proportional to $\boldsymbol{\beta}^*$ with $c_0>0$. □

Lemma B.2. Suppose that $n_{p,s}=o(1)$, $n_{p,s}\gtrsim\frac{\mathbb{E}(Y-c_0\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2}{\lambda^2 s}$ for a sufficiently large constant and $\mathbb{E}Y^4<\infty$. Then the solution $\widetilde{\boldsymbol{\beta}}$ coincides with the solution:

$$\widetilde{\boldsymbol{\beta}}_S = \operatorname*{argmin}_{\boldsymbol{\beta}_S\in\mathbb{R}^s}\frac{1}{2n}\|\boldsymbol{Y}-\mathbf{X}_S\boldsymbol{\beta}_S\|_2^2+\lambda\|\boldsymbol{\beta}_S\|_1,$$

where $S=\operatorname{supp}(\boldsymbol{\beta}^*)$ (i.e., the set of non-zero coefficients of $\boldsymbol{\beta}^*$) with high probability (i.e. at least .99). Moreover we have

$$\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2\lesssim\sqrt{s}\lambda+n_{p,s}^{-\frac{1}{2}},$$

with overwhelming probability.

Proof of Lemma B.2. Theorem 2.3.4 i. of Neykov et al. [2016] shows that under $n_{p,s}\gtrsim\frac{\mathbb{E}(Y-c_0\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2}{\lambda^2 s}$ our first claim follows. We therefore focus on showing our second claim below. Define

$$\mathbf{w}:=\boldsymbol{Y}-c_0\mathbf{X}\boldsymbol{\beta}^*=\boldsymbol{Y}-c_0\mathbf{X}_S\boldsymbol{\beta}_S^*.$$

Furthermore define the quantities:

$$\theta^2:=\operatorname{Var}\{(Y-c_0\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2\},\quad \gamma^2:=\operatorname{Var}(Y\boldsymbol{X}^\top\boldsymbol{\beta}^*),$$
$$\xi^2:=\mathbb{E}\{(Y-c_0\boldsymbol{X}^\top\boldsymbol{\beta}^*)^2\}.$$

Notice that the above quantities are well defined since $\mathbb{E}Y^4<\infty$ by assumption. We will now show that the vector $\widetilde{\boldsymbol{\beta}}_S$ is close to $\boldsymbol{\beta}_S^*$ in Euclidean distance. We start by using the inequality:

$$\begin{aligned}&\frac{1}{2n}\|\boldsymbol{Y}-\mathbf{X}_S\widetilde{\boldsymbol{\beta}}_S\|_2^2+\lambda\|\widetilde{\boldsymbol{\beta}}_S\|_1\\ &\quad\le\frac{1}{2n}\|\boldsymbol{Y}-c_0\mathbf{X}_S\boldsymbol{\beta}_S^*\|_2^2+\lambda\|c_0\boldsymbol{\beta}_S^*\|_1.\end{aligned}$$

Expanding the norms leads to

$$\begin{aligned}&\frac{1}{2n}\|\mathbf{X}_S(c_0\boldsymbol{\beta}_S^*-\widetilde{\boldsymbol{\beta}}_S)\|_2^2+\lambda\|\widetilde{\boldsymbol{\beta}}_S\|_1\\ &\quad\le\frac{1}{n}\mathbf{w}^\top\mathbf{X}_S(\widetilde{\boldsymbol{\beta}}_S-c_0\boldsymbol{\beta}_S^*)+\lambda\|c_0\boldsymbol{\beta}_S^*\|_1\\ &\quad\le\frac{1}{n}\|\mathbf{w}^\top\mathbf{X}_S\|_\infty\|c_0\boldsymbol{\beta}_S^*-\widetilde{\boldsymbol{\beta}}_S\|_1+\lambda\|c_0\boldsymbol{\beta}_S^*\|_1\end{aligned} \tag{B.4}$$

The vector $\mathbf{w}^\top \mathbf{X}_S$ is mean 0. We will now control $n^{-1}\|\mathbf{w}^\top \mathbf{X}_S\|_\infty$. We have

$$n^{-1}\|\mathbf{w}^\top \mathbf{X}_S\|_\infty \le n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}} \mathbf{X}_S^\top \mathbf{w}\|_\infty + n^{-1}\|\boldsymbol{\beta}_S^* \boldsymbol{\beta}_S^{*\top} \mathbf{X}_S^\top \mathbf{w}\|_\infty, \qquad \text{(B.5)}$$

where $\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}} = \mathbf{I}_s - \boldsymbol{\beta}_S^* \boldsymbol{\beta}_S^{*\top}$. Note that $\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}} \mathbf{X}_S$ and $\mathbf{w}$ are independent. It is simple to check that conditionally on $\mathbf{w}$ the vector $n^{-1}\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}} \mathbf{X}_S^\top \mathbf{w} \sim \mathcal{N}(0, \mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}} n^{-2}\|\mathbf{w}\|_2^2)$. We now argue that the term $n^{-1}\|\mathbf{w}\|_2^2 \le 2\xi^2$ with probability at least $1 - \frac{\theta^2}{n\xi^2}$. Since $\mathbf{w} = \boldsymbol{Y} - c_0 \mathbf{X}_S \boldsymbol{\beta}_S^*$ is a vector with non-zero mean. However, by Chebyshev's inequality we have:

$$\mathbb{P}\left(\left|\frac{\|\mathbf{w}\|_2^2}{n} - \xi^2\right| \ge t\right) \le \frac{\theta^2}{nt^2}.$$

Then setting $t = \xi^2$ brings the above probability to 0 at a rate $\frac{\theta^2}{n\xi^4}$. Next, conditioning on this event it follows that the diagonal entries of the covariance matrix $n^{-2}\|\mathbf{w}\|_2^2 \mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}}$ are less than $n^{-2}\|\mathbf{w}\|_2^2 \le \frac{2\xi^2}{n}$. Hence by a standard Gaussian tail bound, on the event $n^{-1}\|\mathbf{w}\|_2^2 \le 2\xi^2$ we have that

$$\mathbb{P}(n^{-1}\|\mathbf{P}_{\boldsymbol{\beta}_S^{*\perp}} \mathbf{X}_S^\top \mathbf{w}\|_\infty \ge t) \le 2se^{-\bar{c}nt^2/\xi^2},$$

for some universal constant $\bar{c}$. Therefore setting $t \ge \sqrt{\frac{2\xi^2 \log p}{\bar{c}n}}$ bounds the above probability by $\frac{2s}{p^2} \le 2p^{-1}$. We now move to the second term of (B.5). Since $\|\boldsymbol{\beta}_S^*\|_\infty \le \|\boldsymbol{\beta}_S^*\|_2 \le 1$ we have

$$n^{-1}\|\boldsymbol{\beta}_S^* \boldsymbol{\beta}_S^{*\top} \mathbf{X}_S^\top \mathbf{w}\|_\infty \le n^{-1}\|\boldsymbol{\beta}_S^{*\top} \mathbf{X}_S^\top \mathbf{w}\|_\infty.$$

Next we have the elementary inequality

$$\begin{aligned}\mathbb{P}(n^{-1}|\boldsymbol{\beta}_S^{*\top} \mathbf{X}_S^\top \boldsymbol{Y} - c_0\|\mathbf{X}_S \boldsymbol{\beta}_S^*\|_2^2| \ge t) &\le \mathbb{P}(|n^{-1}\boldsymbol{\beta}_S^{*\top} \mathbf{X}_S^\top \boldsymbol{Y} - c_0| \ge t/2) \\ &+ \mathbb{P}(|n^{-1}\|\mathbf{X}_S \boldsymbol{\beta}_S^*\|_2^2 - 1| \ge t/(2c_0)),\end{aligned}$$

By Chebyshev's inequality

$$\mathbb{P}(|n^{-1}\boldsymbol{\beta}_S^{*\top} \mathbf{X}_S^\top \boldsymbol{Y} - c_0| \ge t/2) \le \frac{4\gamma^2}{nt^2}, \qquad \text{(B.6)}$$

Setting $t = 2\gamma\sqrt{\frac{\log p}{n}}$ bounds the above probability by $(\log p)^{-1}$. By Lemma 1 of Laurent and Massart [2000]

$$\begin{aligned}\mathbb{P}(|n^{-1}\|\mathbf{X}_S \boldsymbol{\beta}_S^*\|_2^2 - 1| \ge t/(2|c_0|)) \\ \le 2\exp(-n\frac{t}{8|c_0|} \wedge \frac{t^2}{64c_0^2}),\end{aligned}$$

Setting $t = 8|c_0|\sqrt{\frac{\log p}{n}}$ bounds the above probability by $2p^{-1}$. We conclude that with probability at least $1 - 2p^{-1} - (\log p)^{-1} - \frac{\theta^2}{n\xi^4}$

$$n^{-1}\|\mathbf{w}^\top \mathbf{X}_S\|_\infty \le \overline{C}\sqrt{\frac{\log p}{n}}, \qquad \text{(B.7)}$$

where $\overline{C}(\bar{c}_0, c_0, \gamma, \xi) = 8|c_0| + 2\gamma + \bar{c}_0\xi$ and $\bar{c}_0 = \sqrt{2/\bar{c}}$ is a universal constant.

Going back to (B.4) we have established that with high probability

$$\begin{aligned}&\frac{1}{2n}\|\mathbf{X}_S(c_0\boldsymbol{\beta}_S^* - \widetilde{\boldsymbol{\beta}}_S)\|_2^2 \\ &\le \overline{C}\sqrt{\frac{\log p}{n}}\|c_0\boldsymbol{\beta}_S^* - \widetilde{\boldsymbol{\beta}}_S\|_1 + \lambda(\|c_0\boldsymbol{\beta}_S^*\|_1 - \|\widetilde{\boldsymbol{\beta}}_S\|_1) \\ &\le (\overline{C}n_{p,s}^{-\frac{1}{2}} + \sqrt{s}\lambda)\|c_0\boldsymbol{\beta}_S^* - \widetilde{\boldsymbol{\beta}}_S\|_2,\end{aligned}$$

where the inequality $\|\mathbf{v}\|_1 \le \sqrt{s}\|\mathbf{v}\|_2$ for $\mathbf{v} \in \mathbb{R}^s$. Corollary 5.35 of Vershynin [2012] guarantees that

$$\frac{\lambda_{\min}(\mathbf{X}_S^\top \mathbf{X}_S)}{n} \ge \frac{(\sqrt{n} - 2\sqrt{s})^2}{n},$$

with probability at least $1 - 2e^{-s/2}$. Hence, when the above two events happen (with probability at least $1 - 2p^{-1} - (\log p)^{-1} - 2e^{-s/2} - \frac{\theta^2}{n\xi^4}$) we have

$$\|c_0\boldsymbol{\beta}_S^* - \widetilde{\boldsymbol{\beta}}_S\|_2 \le (\overline{C}n_{p,s}^{-\frac{1}{2}} + \sqrt{s}\lambda)\frac{n}{(\sqrt{n} - 2\sqrt{s})^2}. \qquad \text{(B.8)}$$

Denote the RHS of (B.8) with R for brevity. We have $c_0 - R \le \|\widetilde{\boldsymbol{\beta}}\|_2 \le c_0 + R$.

$$\begin{aligned}\left\|\boldsymbol{\beta}_S^* - \frac{\widetilde{\boldsymbol{\beta}}_S}{\|\widetilde{\boldsymbol{\beta}}_S\|_2}\right\|_2 &\le \left\|\frac{c_0\boldsymbol{\beta}_S^* - \widetilde{\boldsymbol{\beta}}_S}{\|\widetilde{\boldsymbol{\beta}}_S\|_2}\right\|_2 + \frac{|c_0 - \|\widetilde{\boldsymbol{\beta}}_S\|_2|}{\|\widetilde{\boldsymbol{\beta}}_S\|_2} \\ &\le 2\frac{R}{c_0 - R}.\end{aligned}$$

□

Proof of Theorem 3.2. Using Theorem 2.3 with a vector $\mathbf{u}$ with components $u_i = f(\boldsymbol{X}_{\pi_i}^\top \widehat{\boldsymbol{\beta}})$, we have with probability at least $1 - \exp(-x)$:

$$\begin{aligned}&\frac{1}{n}\sum_{i=n+1}^{2n}(f(\boldsymbol{X}_i^\top \boldsymbol{\beta}^*) - \widehat{f}(\boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}))^2 \\ &\le \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{X}_{\pi_i}^\top \boldsymbol{\beta}^*) - f(\boldsymbol{X}_{\pi_i}^\top \widehat{\boldsymbol{\beta}}))^2 + \frac{C2^{2/3}}{n^{2/3}} + \frac{4x}{n} \\ &\le L^2(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) + \frac{C2^{2/3}}{n^{2/3}} + \frac{4x}{n},\end{aligned}$$

where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=n+1}^{2n}\boldsymbol{X}_i\boldsymbol{X}_i^\top$. By Lemma B.2 we know that the vector $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ is s-sparse, and therefore by Corollary 5.35 of Vershynin [2012] we have

$$\begin{aligned}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) &\le (1 + \sqrt{s/n} + \sqrt{x/n})^2\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \\ &\lesssim (\sqrt{s}\lambda + n_{p,s}^{-\frac{1}{2}})^2,\end{aligned}$$

with probability at least $1 - \exp(-x)$, where in the last inequality we used Lemma B.2 once again. This completes the proof. □