8. Supplement to the paper: Learning Heteroscedastic Models by Convex Programming under Group Sparsity

8.1. Proof of Theorem 3.2

The fact that the feasible set is not empty follows from the fact that it contains the minimizers of (6). This immediately follows from the first-order conditions and their relaxations. Indeed, for a minimizer $(\boldsymbol{\phi}^\circ, \boldsymbol{\alpha}^\circ)$ of (6), the first-order conditions take the following form: there exists $\boldsymbol{\nu}^\circ \in \mathbb{R}^T_+$ such that for all $k \in [K]$ and $\ell \in [q]$,

$$\frac{\partial}{\partial \phi_{G_k}}\mathrm{PL}(\boldsymbol{\phi}^\circ, \boldsymbol{\alpha}^\circ) = -\mathbf{X}_{:,G_k}^\top\big(\mathrm{diag}(\boldsymbol{Y})\mathbf{R}\boldsymbol{\alpha}^\circ - \mathbf{X}\boldsymbol{\phi}^\circ\big) + \lambda_k \mathbf{X}_{:,G_k}^\top \frac{\mathbf{X}_{:,G_k}\boldsymbol{\phi}^\circ_{G_k}}{\big|\mathbf{X}_{:,G_k}\boldsymbol{\phi}^\circ_{G_k}\big|_2} = 0, \tag{18}$$

$$\frac{\partial}{\partial \alpha^\circ_\ell}\mathrm{PL}(\boldsymbol{\phi}^\circ, \boldsymbol{\alpha}^\circ) = -\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^\circ} + \sum\nolimits_{t\in\mathcal{T}} \big(y_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha}^\circ - \boldsymbol{X}_{t,:}\boldsymbol{\phi}^\circ\big) y_t r_{t\ell} - (\boldsymbol{\nu}^\circ)^\top \boldsymbol{R}_{:,\ell} = 0, \tag{19}$$

and $\nu^\circ_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha}^\circ = 0$ for every t. It should be emphasized that relation (18) holds true only in the case where the solution satisfies $\min_k |\boldsymbol{X}_{:,G_k}\boldsymbol{\phi}^\circ_{:,G_k}|_2 \neq 0$, otherwise one has to replace it by the condition stating that the null vector belongs to the subdifferential. Since this does not alter the proof, we prefer to proceed as if everything was differentiable.

On the one hand, $(\boldsymbol{\phi}^\circ, \boldsymbol{\alpha}^\circ)$ satisfies (18) if and only if $\mathbf{\Pi}_{G_k}(\mathrm{diag}(\boldsymbol{Y})\mathbf{R}\boldsymbol{\alpha}^\circ - \mathbf{X}\boldsymbol{\phi}^\circ) = \lambda_k \mathbf{X}_{:,G_k}\boldsymbol{\phi}^\circ_{G_k}/|\mathbf{X}_{:,G_k}\boldsymbol{\phi}^\circ_{G_k}|_2$ with $\mathbf{\Pi}_{G_k} = \mathbf{X}_{:,G_k}(\mathbf{X}_{:,G_k}^\top\mathbf{X}_{:,G_k})^+\mathbf{X}_{:,G_k}^\top$ being the orthogonal projector onto the range of $\mathbf{X}_{:,G_k}$ in $\mathbb{R}^T$. Taking the norm of both sides in the last equation, we get $\big|\mathbf{\Pi}_{G_k}(\mathrm{diag}(\boldsymbol{Y})\mathbf{R}\boldsymbol{\alpha}^\circ - \mathbf{X}\boldsymbol{\phi}^\circ)\big|_2 \leq \lambda_k$. This tells us that $(\boldsymbol{\phi}^\circ, \boldsymbol{\alpha}^\circ)$ satisfy (7). On the other hand, since the minimum of (6) is finite, one easily checks that $\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^\circ \neq 0$ and, therefore, $\boldsymbol{\nu}^\circ = 0$. Replacing in (19) $\boldsymbol{\nu}^\circ$ by zero and setting $v^\circ_t = 1/\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^\circ$ we get that $(\boldsymbol{\phi}^\circ, \boldsymbol{\alpha}^\circ, \boldsymbol{v}^\circ)$ satisfies (8), (9). This proves that the set of feasible solutions of the optimization problem defined in the ScHeDs is not empty.

Let us show that one can compute the ScHeDs $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}})$ by solving an SOCP. More precisely, we show that if $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^K \times \mathbb{R}^T$ is a solution to the following problem of second-order cone programming:

$$\min \quad \sum\nolimits_{k=1}^{K} \lambda_k u_k \tag{20}$$

$$\text{subject to (7) and}$$

$$\big|\mathbf{X}_{:,G_k}\boldsymbol{\phi}_{G_k}\big|_2 \leq u_k, \qquad \forall k \in [K], \tag{21}$$

$$\mathbf{R}^\top \boldsymbol{v} \leq \mathbf{R}^\top \mathrm{diag}(\boldsymbol{Y})(\mathrm{diag}(\boldsymbol{Y})\mathbf{R}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\phi}); \tag{22}$$

$$\big|\big[v_t; \boldsymbol{R}_{t,:}\boldsymbol{\alpha}; \sqrt{2}\,\big]\big|_2 \leq v_t + \boldsymbol{R}_{t,:}\boldsymbol{\alpha}; \qquad \forall t \in \mathcal{T}, \tag{23}$$

then $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{v}})$ is a solution to the optimization problem stated in Definition 3.1. This claim readily follows from the fact that the constraint $\big|\big[v_t; \boldsymbol{R}_{t,:}\boldsymbol{\alpha}; \sqrt{2}\,\big]\big|_2 \leq v_t + \boldsymbol{R}_{t,:}\boldsymbol{\alpha}$ can be equivalently written as $v_t(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}) \geq 1$ and $v_t + \boldsymbol{R}_{t,:}\boldsymbol{\alpha} \geq 0$ for every t. This yields $v_t \geq 0$ and $\boldsymbol{R}_{t,:}\boldsymbol{\alpha} \geq 1/v_t$ for every t. Furthermore, it is clear that if $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}})$ is a solution to the aforementioned optimization problem, then all the inequalities in (21) are indeed equalities. This completes the proof.

8.2. Proof of Theorem 5.1

To prove Theorem 5.1, we first introduce a feasible pair $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}})$, in the sense formulated in Lemma 8.1.

Lemma 8.1. *Consider the model* (10). *Let* $z = 1 + 2C_4\sqrt{\frac{2\log(2q/\varepsilon)}{T}}$ *with some* $\varepsilon > 0$ *and assume that* $z \leq 2$. *Then with probability at least* $1 - 2\varepsilon$, *the triplet* $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{v}}) = \big(z\boldsymbol{\phi}^*, z\boldsymbol{\alpha}^*, (\frac{1}{z\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*})_{t=1,\ldots,T}\big)$ *satisfies constraints* (7), (8) *and* (9). *Moreover, the group-sparsity pattern* $\big\{k : \big|\widetilde{\boldsymbol{\phi}}_{G_k}\big|_1 \neq 0\big\}$ *of* $\widetilde{\boldsymbol{\phi}}$ *coincides with that of* $\boldsymbol{\phi}^*$, *that is with* $\mathcal{K}^*$.

The proof of this lemma can be found in Section 8.3.

Set $\boldsymbol{\Delta} = \widehat{\boldsymbol{\phi}} - \widetilde{\boldsymbol{\phi}}$. On an event of probability at least $1 - 2\varepsilon$, $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}})$ is a feasible solution of the optimization problem of the ScHeDs whereas $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}})$ is an optimal solution, therefore

$$\begin{aligned}\sum_{k=1}^{K} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\Delta}_{G_k}\big|_2 &\le \sum_{k=1}^{K} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\Delta}_{G_k}\big|_2 + \sum_{k=1}^{K} \lambda_k \big|\mathbf{X}_{:,G_k} \widetilde{\boldsymbol{\phi}}_{G_k}\big|_2 - \sum_{k=1}^{K} \lambda_k \big|\mathbf{X}_{:,G_k} \widehat{\boldsymbol{\phi}}_{G_k}\big|_2 \\ &= \sum_{k\in\mathcal{K}^*} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\Delta}_{G_k}\big|_2 + \sum_{k\in\mathcal{K}^*} \lambda_k \big(\big|\mathbf{X}_{:,G_k} \widetilde{\boldsymbol{\phi}}_{G_k}\big|_2 - \big|\mathbf{X}_{:,G_k} \widehat{\boldsymbol{\phi}}_{G_k}\big|_2\big) \\ &\le 2 \sum_{k\in\mathcal{K}^*} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\Delta}_{G_k}\big|_2. \end{aligned} \tag{24}$$

This readily implies that

$$\sum_{k\in\mathcal{K}^{*c}} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\Delta}_{G_k}\big|_2 \le \sum_{k\in\mathcal{K}^*} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\Delta}_{G_k}\big|_2.$$

Applying GRE(κ, s) assumption and the Cauchy-Schwarz inequality, we get

$$\sum_{k=1}^{K} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\Delta}_{G_k}\big|_2 \le 2\Big(\sum_{k\in\mathcal{K}^*} \lambda_k^2\Big)^{1/2} \Big(\sum_{k\in\mathcal{K}^*} \big|\mathbf{X}_{:,G_k} \boldsymbol{\Delta}_{G_k}\big|_2^2\Big)^{1/2} \le \frac{2}{\kappa}\big(\textstyle\sum_{k\in\mathcal{K}^*} \lambda_k^2\big)^{1/2} \big|\mathbf{X}\boldsymbol{\Delta}\big|_2. \tag{25}$$

It is clear that

$$\begin{aligned}\big|\mathbf{X}\boldsymbol{\Delta}\big|_2^2 &= \boldsymbol{\Delta}^\top \mathbf{X}^\top (\mathbf{X}\widehat{\boldsymbol{\phi}} - \mathbf{X}\widetilde{\boldsymbol{\phi}}) \\ &= \boldsymbol{\Delta}^\top \mathbf{X}^\top (\mathbf{X}\widehat{\boldsymbol{\phi}} - \operatorname{diag}(\mathbf{R}\widehat{\boldsymbol{\alpha}})\boldsymbol{Y}) + \boldsymbol{\Delta}^\top \mathbf{X}^\top (\operatorname{diag}(\mathbf{R}\widetilde{\boldsymbol{\alpha}})\boldsymbol{Y} - \mathbf{X}\widetilde{\boldsymbol{\phi}}) + \boldsymbol{\Delta}^\top \mathbf{X}^\top \operatorname{diag}(\boldsymbol{Y})\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}).\end{aligned}$$

In addition, using the relation $\mathbf{X}\boldsymbol{\Delta} = \sum_{k=1}^{K} \mathbf{X}_{:,G_k}\boldsymbol{\Delta}_{G_k} = \sum_{k=1}^{K} \boldsymbol{\Pi}_{G_k}\mathbf{X}_{:,G_k}\boldsymbol{\Delta}_{G_k}$ and the fact that both $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}})$ and $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}})$ satisfy constraint (7), we have

$$\begin{aligned}\big|\mathbf{X}\boldsymbol{\Delta}\big|_2^2 &\le \sum_{k=1}^{K} \boldsymbol{\Delta}_{G_k}^\top \mathbf{X}_{:,G_k}^\top \boldsymbol{\Pi}_{G_k} (\mathbf{X}\widehat{\boldsymbol{\phi}} - \operatorname{diag}(\mathbf{R}\widehat{\boldsymbol{\alpha}})\boldsymbol{Y}) + \sum_{k=1}^{K} \boldsymbol{\Delta}_{G_k}^\top \mathbf{X}_{:,G_k}^\top \boldsymbol{\Pi}_{G_k} (\operatorname{diag}(\mathbf{R}\widetilde{\boldsymbol{\alpha}})\boldsymbol{Y} - \mathbf{X}\widetilde{\boldsymbol{\phi}}) \\ &\qquad + \boldsymbol{\Delta}^\top \mathbf{X}^\top \operatorname{diag}(\boldsymbol{Y})\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}) \\ &\le 2\sum_{k=1}^{K} \lambda_k \big|\mathbf{X}_{:,G_k}\boldsymbol{\Delta}_{G_k}\big|_2 + \big|\mathbf{X}\boldsymbol{\Delta}\big|_2 . \big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}})\big|_2. \end{aligned} \tag{26}$$

Therefore, from (25), $|\mathbf{X}\boldsymbol{\Delta}|_2 \le \frac{4}{\kappa}\big(\sum_{k\in\mathcal{K}^*} \lambda_k^2\big)^{1/2} + |\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}})|_2$ and we easily get

$$|\mathbf{X}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)|_2 \le |\mathbf{X}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)|_2 + |\mathbf{X}\boldsymbol{\Delta}|_2 \le (z-1)|\mathbf{X}\boldsymbol{\phi}^*|_2 + \frac{4}{\kappa}\Big(\textstyle\sum_{k\in\mathcal{K}^*} \lambda_k^2\Big)^{1/2} + |\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}})|_2.$$

where we have used the following notation: for any vector $\boldsymbol{v}$, we denote by $\mathbf{D}_{\boldsymbol{v}}$ the diagonal matrix diag$(\boldsymbol{v})$.

To complete the proof, it suffices to replace z and λ_k by their expressions and to use the inequality

$$\begin{aligned}|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}})|_2 &\le |\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)|_2 + (z-1)|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\boldsymbol{\alpha}^*|_2 \\ &\le |\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)|_2 + (z-1)|\mathbf{X}\boldsymbol{\phi}^* + \boldsymbol{\xi}|_2 \\ &\le |\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)|_2 + (z-1)\big(|\mathbf{X}\boldsymbol{\phi}^*|_2 + |\boldsymbol{\xi}|_2\big).\end{aligned}$$

8.3. Proof of Lemma 8.1

For all $\varepsilon \in (0,1)$, consider the random event $\mathcal{B}_\varepsilon = \bigcap_{\ell=1}^{q}\Big(\mathcal{B}_{\varepsilon,\ell}^2 \cap \mathcal{B}_{\varepsilon,\ell}^1\Big)$, where

$$\mathcal{B}_{\varepsilon,\ell}^2 = \Big\{\textstyle\sum_{t\in\mathcal{T}} \dfrac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*\xi_t \ge -\sqrt{2C_2 T\log(2q/\varepsilon)}\Big\},$$

$$\mathcal{B}^1_{\varepsilon,\ell} = \left\{\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}(\xi_t^2-1) \geq -2\sqrt{C_1 T\log(2q/\varepsilon)}\right\}.$$

Using standard tail estimates for the Gaussian and the χ^2 distributions, in conjunction with the union bound, one easily checks that $P(\mathcal{B}_\varepsilon)\geq 1-\varepsilon$. In what follows, we show that on the event $\mathcal{B}_\varepsilon$, $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{v}})$ satisfies constraints (7)-(9).

Constraints (9) are satisfied (with equality) by definition of $\widetilde{\boldsymbol{v}}$. To check that (8) is satisfied as well, we should verify that for all $\ell = 1,\ldots,q$,

$$\frac{1}{z^2}\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*} \leq \sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*\xi_t + \sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}\xi_t^2.$$

On the event $\mathcal{B}_\varepsilon$, the right-hand side of the last inequality can be lower bounded as follows:

$$\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*\xi_t + \sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}\xi_t^2 \geq -(\sqrt{C_2}+\sqrt{2C_1})\sqrt{2T\log(2q/\varepsilon)} + \sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}.$$

Thus, on $\mathcal{B}_\varepsilon$ if for all $\ell = 1,\ldots,q$

$$\frac{z^2-1}{z^2}\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*} \geq (\sqrt{C_2}+\sqrt{2C_1})\sqrt{2T\log(2q/\varepsilon)} \tag{27}$$

then constraint (9) is fulfilled by $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{v}})$. Inequality (27) is valid since for any $z\geq 1$

$$\frac{z^2-1}{z^2}\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*} = \frac{z-1}{z}\left(1+\frac{1}{z}\right)\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*} \geq \frac{z-1}{z}TC_3$$

and $\frac{z-1}{z}TC_3 \geq (\sqrt{C_2}+\sqrt{2C_1})\sqrt{2T\log(2q/\varepsilon)}$ when $z = 1+2C_4\sqrt{\frac{2\log(2q/\varepsilon)}{T}} \leq 2$.

On the other hand, since $z\leq 2$, a sufficient condition implying that the pair $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}})$ satisfies (7) is

$$2\big|\boldsymbol{\Pi}_{G_k}\boldsymbol{\xi}\big|_2 \leq \lambda_k, \qquad \forall k\in\{1,\ldots,K\}. \tag{28}$$

Recall that r_k denotes the rank of $\boldsymbol{\Pi}_{G_k}$. Let $\mathcal{R}_\varepsilon$ be the random event of probability at least $1-\varepsilon$ defined as follows

$$\mathcal{R}_\varepsilon = \bigcap_{k=1}^K \mathcal{R}_{\varepsilon,k} = \bigcap_{k=1}^K \left\{\big|\boldsymbol{\Pi}_{G_k}\boldsymbol{\xi}\big|_2^2 \leq r_k + 2\sqrt{r_k\log(K/\varepsilon)} + 2\log(K/\varepsilon)\right\}.$$

To prove that $P(\mathcal{R}_\varepsilon)\geq 1-\varepsilon$, we use the fact that $\Big|\boldsymbol{\Pi}_{G_k}\boldsymbol{\xi}\big|_2^2$ is drawn from the $\chi^2_{r_k}$ distribution. Using well-known tail bounds for the χ^2 distribution, we get $P(\mathcal{R}^c_{\varepsilon,k}) \leq \frac{\varepsilon}{K}$. Then, we conclude by the union bound.

Since we chose

$$2(r_k + 2\sqrt{r_k\log(K/\varepsilon)} + 2\log(K/\varepsilon))^{1/2} = \lambda_k,$$

on the event $\mathcal{R}_\varepsilon$ inequality (28) is satisfied by $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}})$.

Finally, the triplet $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{v}})$ fulfills constraints (7)-(9) on the event $\mathcal{B}_\varepsilon\cap\mathcal{R}_\varepsilon$, which is of a probability at least $1-2\varepsilon$.

8.4. Proof of Theorem 5.2

We start by noting that, the ScHeDs $(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\alpha}})$ satisfies $\forall\ell\in\{1,\ldots,q\}$, the relation

$$\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}} = \sum\nolimits_{t\in\mathcal{T}} \big(y_t\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}} - \boldsymbol{X}_{t,:}\widehat{\boldsymbol{\phi}}\big)y_t r_{t\ell}. \tag{29}$$

First, for the ScHeDs, all the inequalities in (9) are equalities. Indeed, v_t's are only involved in (8) and (9) and if we decrease one v_t to achieve an equality in (9), the left-hand side of (8) will decrease as well and the constraint will stay inviolated. Thus, setting $\widehat{v}_t = 1/\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}$, we get from (8)

$$\sum\nolimits_{t\in\mathcal{T}} \frac{r_{t\ell}}{\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}} \le \sum\nolimits_{t\in\mathcal{T}} \big(y_t\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}} - \boldsymbol{X}_{t,:}\widehat{\boldsymbol{\phi}}\big) y_t r_{t\ell}, \qquad \forall \ell\in\{1,\ldots,q\}. \tag{30}$$

To be convinced that Eq. (29) is true, let us consider for simplicity the one dimensional case $q = 1$. If inequality (30) was strict, for some $w \in (0,1)$, the pair $(w\widehat{\boldsymbol{\phi}}, w\widehat{\boldsymbol{\alpha}})$ would also satisfy all the constraints of the ScHeDs and the corresponding penalty term would be smaller than that of $(\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{\alpha}})$. This is impossible since $\widehat{\boldsymbol{\phi}}$ is an optimal solution. Thus we get

$$\sum\nolimits_{t\in\mathcal{T}} \boldsymbol{R}_{t,:}^\top(\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}})^{-1} = \sum\nolimits_{t\in\mathcal{T}} \boldsymbol{R}_{t,:}^\top y_t\big(y_t\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}} - \boldsymbol{X}_{t,:}\widehat{\boldsymbol{\phi}}\big) = \mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}\big(\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\widehat{\boldsymbol{\alpha}} - \mathbf{X}\widehat{\boldsymbol{\phi}}\big). \tag{31}$$

Using the identity $(\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}})^{-1} = (\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^{-1} + (\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^{-1}\boldsymbol{R}_{t,:}(\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}})$, we get

$$\begin{aligned}\Big[\sum_{t\in\mathcal{T}} \frac{1}{(\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}})(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)}\boldsymbol{R}_{t,:}^\top\boldsymbol{R}_{t,:}\Big](\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}}) &= -\sum_{t\in\mathcal{T}} \frac{1}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}\boldsymbol{R}_{t,:}^\top + \mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}\big(\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\widehat{\boldsymbol{\alpha}} - \mathbf{X}\widehat{\boldsymbol{\phi}}\big)\\ &= -\mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{1}_T + \mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}^2\mathbf{R}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) - \mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}\mathbf{X}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)\\ &\quad + \mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}\big(\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\boldsymbol{\alpha}^* - \mathbf{X}\boldsymbol{\phi}^*\big).\end{aligned} \tag{32}$$

In view of the identities $\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\boldsymbol{\alpha}^* - \mathbf{X}\boldsymbol{\phi}^* = \boldsymbol{\xi}$ and $\mathbf{D}_{\boldsymbol{Y}} = \mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*} + \mathbf{D}_{\boldsymbol{\xi}})$, Eq. (32) yields[4]

$$\mathbf{R}^\top\Big[\mathbf{D}_{\boldsymbol{Y}}^2 + \mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{R}\widehat{\boldsymbol{\alpha}}}^{-1}\Big]\mathbf{R}(\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}}) = \mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2 - \mathbf{1}_T) - \mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}\mathbf{X}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*) + \mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}. \tag{33}$$

As a consequence, denoting by $\mathbf{M}$ the Moore-Penrose pseudo-inverse of the matrix $\mathbf{R}^\top\big[\mathbf{D}_{\boldsymbol{Y}}^2 + \mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{R}\widehat{\boldsymbol{\alpha}}}^{-1}\big]\mathbf{R}$,

$$\mathbf{R}(\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}}) = \mathbf{R}\mathbf{M}\mathbf{R}^\top\Big(\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2 - \mathbf{1}_T) - \mathbf{D}_{\boldsymbol{Y}}\mathbf{X}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*) + \mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\Big). \tag{34}$$

Multiplying both sides by $\mathbf{D}_{\boldsymbol{Y}}$ and taking the Euclidean norm, we get

$$\big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}})\big|_2 \le \Big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}\mathbf{R}^\top\Big(\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2 - \mathbf{1}_T) + \mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\Big)\Big|_2 + \Big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}\mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}\mathbf{X}(\boldsymbol{\phi}^* - \widehat{\boldsymbol{\phi}})\Big|_2. \tag{35}$$

At this stage of the proof, the conceptual part is finished and we enter into the technical part. At a heuristic level, the first norm in the right-hand side of (35) is bounded in probability while the second norm is bounded from above by $(1-c)\big|\mathbf{X}(\boldsymbol{\phi}^* - \widehat{\boldsymbol{\phi}})\big|_2$ for some constant $c \in (0,1)$. Let us first state these results formally, by postponing their proof to the next subsection, and to finalize the proof of the theorem.

Lemma 8.2. *Let q and T be two integers such that $1 \le q \le T$ and let $\varepsilon \in (0, 1/3)$ be some constant. Assume that for some constant $\widehat{D}_1 \ge 1$ the inequality $\max_{t\in\mathcal{T}} \frac{\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*} \le \widehat{D}_1$ holds true. Then, on an event of probability at least $1 - 3\varepsilon$, the following inequalities are true*[5]*:*

$$|\!|\!|\mathbf{M}^{1/2}\mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}|\!|\!| \le 1, \tag{36}$$

$$\Big|\mathbf{M}^{1/2}\mathbf{R}^\top\big(\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2 - \mathbf{1}_T) + \mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\big)\Big|_2 \le 10\sqrt{q\widehat{D}_1\log(2T/\varepsilon)\log(2q/\varepsilon)}, \tag{37}$$

$$|\!|\!|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}\mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}|\!|\!| \le 1 - \frac{1}{2\widehat{D}_1\big(|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + |\boldsymbol{\xi}|_\infty^2\big) + 1} \le 1 - \frac{1}{\widehat{D}_1\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon)\big)}. \tag{38}$$

In view of these bounds, we get that on an event of probability at least $1 - 3\varepsilon$,

$$\big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}(\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}})\big|_2 \le 10\sqrt{q\widehat{D}_1\log(2T/\varepsilon)\log(2q/\varepsilon)} + \bigg(1 - \frac{1}{\widehat{D}_1\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon)\big)}\bigg)\big|\mathbf{X}(\boldsymbol{\phi}^* - \widehat{\boldsymbol{\phi}})\big|_2. \tag{39}$$

[4] We denote by $\boldsymbol{\xi}^2$ the vector $(\xi_t^2)_{t\in\mathcal{T}}$.

[5] Here and in the sequel, the spectral norm of a matrix $\mathbf{A}$ is denoted by $|\!|\!|\mathbf{A}|\!|\!|$.

Combining this inequality with Theorem 5.1 and using the inequality $2|\mathbf{X}\boldsymbol{\phi}^*|_2 + |\boldsymbol{\xi}|_2 \le \sqrt{T}\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty + |\boldsymbol{\xi}|_\infty\big)$, we get that the following inequalities are satisfied with probability $\ge 1-5\varepsilon$:

$$\begin{aligned}
\big|\mathbf{X}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*)\big|_2 &\le 2C_4\widehat{D}_1\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon)\big)\sqrt{2\log(2q/\varepsilon)}(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty + |\boldsymbol{\xi}|_\infty)\\
&\quad + \frac{8}{\kappa}\Big(2s^* + 3K^*\log(K/\varepsilon)\Big)^{1/2}\widehat{D}_1\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon)\big)\\
&\quad + 10\widehat{D}_1\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon)\big)\sqrt{q\widehat{D}_1\log(2T/\varepsilon)\log(2q/\varepsilon)}\\
&\le 4C_4\widehat{D}_1\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\mathrm{log}(2T/\varepsilon)\big)^{3/2}\sqrt{2\log(2q/\varepsilon)}\\
&\quad + \frac{8\widehat{D}_1}{\kappa}\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon)\big)\big(2s^* + 3K^*\log(K/\varepsilon)\big)^{1/2}\\
&\quad + 10\widehat{D}_1\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon)\big)\sqrt{q\widehat{D}_1\log(2T/\varepsilon)\log(2q/\varepsilon)}. \qquad (40)
\end{aligned}$$

Using the notation $D_{T,\varepsilon} = \widehat{D}_1\big(2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 5\log(2T/\varepsilon)\big)$, we obtain

$$\begin{aligned}
\big|\mathbf{X}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}^*)\big|_2 &\le 4C_4D_{T,\varepsilon}^{3/2}\sqrt{2\log(2q/\varepsilon)} + \frac{8D_{T,\varepsilon}}{\kappa}\big(2s^* + 3K^*\log(K/\varepsilon)\big)^{1/2}\\
&\quad + 10D_{T,\varepsilon}\sqrt{q\widehat{D}_1\log(2T/\varepsilon)\log(2q/\varepsilon)}. \qquad (41)
\end{aligned}$$

To further simplify the last term, we use the inequalities:

$$\begin{aligned}
10D_{T,\varepsilon}\sqrt{q\widehat{D}_1\log(2T/\varepsilon)\log(2q/\varepsilon)} &= D_{T,\varepsilon}\sqrt{10}\sqrt{5\widehat{D}_1\log(2T/\varepsilon)}\sqrt{2q\log(2q/\varepsilon)}\\
&\le 4D_{T,\varepsilon}^{3/2}\sqrt{2q\log(2q/\varepsilon)}.
\end{aligned}$$

Combining this with (41) yields (16).

To prove (17), we use once again (34) to infer that

$$\begin{aligned}
\big|\mathbf{R}(\boldsymbol{\alpha}^*-\widehat{\boldsymbol{\alpha}})\big|_2 &\le \Big|\mathbf{RMR}^\top\Big(\mathbf{D}^{-1}_{\mathbf{R}\boldsymbol{\alpha}^*}(\boldsymbol{\xi}^2-\mathbf{1}_T) + \mathbf{D}^{-1}_{\mathbf{R}\boldsymbol{\alpha}^*}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\Big)\Big|_2 + \Big|\mathbf{RMR}^\top\mathbf{D}_{\boldsymbol{Y}}\mathbf{X}(\boldsymbol{\phi}^*-\widehat{\boldsymbol{\phi}})\Big|_2\\
&\le |\!|\!|\mathbf{RM}^{1/2}|\!|\!|\Big(\Big|\mathbf{M}^{1/2}\mathbf{R}^\top\Big(\mathbf{D}^{-1}_{\mathbf{R}\boldsymbol{\alpha}^*}(\boldsymbol{\xi}^2-\mathbf{1}_T) + \mathbf{D}^{-1}_{\mathbf{R}\boldsymbol{\alpha}^*}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\Big)\Big|_2 + |\!|\!|\mathbf{M}^{1/2}\mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}|\!|\!|\Big|\mathbf{X}(\boldsymbol{\phi}^*-\widehat{\boldsymbol{\phi}})\Big|_2\Big).
\end{aligned}$$

In view of Lemma 8.2, this leads to

$$\big|\mathbf{R}(\boldsymbol{\alpha}^*-\widehat{\boldsymbol{\alpha}})\big|_2 \le |\!|\!|\mathbf{RM}^{1/2}|\!|\!|\Big(10\sqrt{q\widehat{D}_1\log(2T/\varepsilon)\log(2q/\varepsilon)} + \Big|\mathbf{X}(\boldsymbol{\phi}^*-\widehat{\boldsymbol{\phi}})\Big|_2\Big), \qquad (42)$$

with probability at least $1-5\varepsilon$. Using the bound in (16), we get

$$\big|\mathbf{R}(\boldsymbol{\alpha}^*-\widehat{\boldsymbol{\alpha}})\big|_2 \le |\!|\!|\mathbf{RM}^{1/2}|\!|\!|\Big(4(C_4+2)D_{T,\varepsilon}^{3/2}\sqrt{2q\log(2q/\varepsilon)} + \frac{8D_{T,\varepsilon}}{\kappa}\sqrt{2s^* + 3K^*\log(K/\varepsilon)}\Big). \qquad (43)$$

In view of the inequality[6]

$$\begin{aligned}
(\mathbf{RM}^{1/2})(\mathbf{RM}^{1/2})^\top &= \mathbf{R}\big[\mathbf{R}^\top(\mathbf{D}^2_{\boldsymbol{Y}} + \mathbf{D}^{-1}_{\mathbf{R}\boldsymbol{\alpha}^*}\mathbf{D}^{-1}_{\mathbf{R}\widehat{\boldsymbol{\alpha}}})\mathbf{R}\big]^+\mathbf{R}^\top\\
&\preceq \mathbf{R}\big[\mathbf{R}^\top(\mathbf{D}^{-1}_{\mathbf{R}\boldsymbol{\alpha}^*}\mathbf{D}^{-1}_{\mathbf{R}\widehat{\boldsymbol{\alpha}}})\mathbf{R}\big]^+\mathbf{R}^\top\\
&\preceq (\max_{t\in\mathcal{T}}[\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*\cdot\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}])\mathbf{R}\big[\mathbf{R}^\top\mathbf{R}\big]^+\mathbf{R}^\top
\end{aligned}$$

we get

$$\begin{aligned}
|\!|\!|\mathbf{RM}^{1/2}|\!|\!|^2 &= |\!|\!|(\mathbf{RM}^{1/2})(\mathbf{RM}^{1/2})^\top|\!|\!|\\
&\le \widehat{D}_1\big|\mathbf{R}\boldsymbol{\alpha}^*\big|_\infty^2\cdot|\!|\!|\mathbf{R}\big[\mathbf{R}^\top\mathbf{R}\big]^+\mathbf{R}^\top|\!|\!|\\
&\le \widehat{D}_1\big|\mathbf{R}\boldsymbol{\alpha}^*\big|_\infty^2,
\end{aligned}$$

where the last inequality follows from the fact that $\mathbf{R}\big[\mathbf{R}^\top\mathbf{R}\big]^+\mathbf{R}^\top$ is an orthogonal projector.

[6] We use the notation $\mathbf{A}\succeq\mathbf{B}$ and $\mathbf{B}\preceq\mathbf{A}$ for indicating that the matrix $\mathbf{A}-\mathbf{B}$ is positive semi-definite. For any matrix $\mathbf{A}$, we denote by $\mathbf{A}^+$ its Moore-Penrose pseudoinverse.

8.5. Proof of Lemma 8.2

We start by presenting a proof of (37). We have

$$\begin{aligned}
&\Big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}\mathbf{R}^\top\big(\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2-\mathbf{1}_T)+\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\big)\Big|_2\\
&\qquad\qquad\le \|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}^{1/2}\|\cdot\big|\mathbf{M}^{1/2}\mathbf{R}^\top\big(\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2-\mathbf{1}_T)+\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\big)\big|_2\\
&\qquad\qquad\le \|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}^{1/2}\|\cdot\big(\big|\mathbf{M}^{1/2}\mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2-\mathbf{1}_T)\big|_2+\big|\mathbf{M}^{1/2}\mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\big|_2\big).
\end{aligned}\tag{44}$$

We remark that

$$\mathbf{M}^+=\mathbf{R}^\top\big[\mathbf{D}_{\boldsymbol{Y}}^2+\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{R}\widehat{\boldsymbol{\alpha}}}^{-1}\big]\mathbf{R}\succeq\mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}^2\mathbf{R}\quad\Longrightarrow\quad\|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}^{1/2}\|\le 1.$$

and that

$$\mathbf{M}^+\succeq\bigg(\min_t\frac{y_t^2+(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*\cdot\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}})^{-1}}{(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*/\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2}\bigg)\mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-2}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}^2\mathbf{R},$$

which implies that

$$\begin{aligned}
\big|\mathbf{M}^{1/2}\mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\big|_2^2&=\boldsymbol{\xi}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\mathbf{R}\mathbf{M}\mathbf{R}^\top\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\boldsymbol{\xi}\\
&\le\bigg(\max_{t\in\mathcal{T}}\frac{(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*)^2}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2y_t^2+(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*/\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}})}\bigg)\boldsymbol{\xi}^\top\mathbf{\Pi}_1\boldsymbol{\xi},
\end{aligned}\tag{45}$$

where $\mathbf{\Pi}_1=\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\mathbf{R}\big(\mathbf{R}^\top\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}^2\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-2}\mathbf{R}\big)^+\mathbf{R}^\top\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}$ is the orthogonal projection on the linear subspace of $\mathbb{R}^T$ spanned by the columns of the matrix $\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\mathbf{R}$. By the Cochran theorem, the random variable $\eta_1=\boldsymbol{\xi}^\top\mathbf{\Pi}_1\boldsymbol{\xi}$ is distributed according to the χ_q^2 distribution.

Using similar arguments based on matrix inequalities, one checks that

$$\begin{aligned}
\big|\mathbf{M}^{1/2}\mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2-\mathbf{1}_T)\big|_2^2&\le\bigg(\max_{t\in\mathcal{T}}\frac{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^{-2}}{y_t^2+(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*\cdot\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}})^{-1}}\bigg)(\boldsymbol{\xi}^2-1)^\top\mathbf{\Pi}_2(\boldsymbol{\xi}^2-1)\\
&\le\bigg(\max_{t\in\mathcal{T}}\frac{\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*}\bigg)\underbrace{(\boldsymbol{\xi}^2-1)^\top\mathbf{\Pi}_2(\boldsymbol{\xi}^2-1)}_{=:\eta_2},
\end{aligned}\tag{46}$$

where $\mathbf{\Pi}_2=\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{R}\big(\mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-2}\mathbf{R}\big)^+\mathbf{R}^\top\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}$ is the orthogonal projection on the linear subspace of $\mathbb{R}^T$ spanned by the columns of the matrix $\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{R}$.

To further simplify (45), one can remark that under the condition $\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}}\le\widehat{D}_1\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*$, it holds

$$\frac{(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*)^2}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2y_t^2+(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*/\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}})}\le\frac{(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*)^2}{(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*+\xi_t)^2+\widehat{D}_1^{-1}}\le 1+\widehat{D}_1\xi_t^2.\tag{47}$$

These bounds, combined with (44), yield

$$\Big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}\mathbf{R}^\top\big(\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2-\mathbf{1}_T)+\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\big)\Big|_2\le\sqrt{(1+\widehat{D}_1|\boldsymbol{\xi}|_\infty^2)\eta_1}+\sqrt{\widehat{D}_1\eta_2}.\tag{48}$$

One can also notice that $\mathbf{\Pi}_2$ is a projector on a subspace of dimension at most equal to q, therefore one can write $\mathbf{\Pi}_2=\sum_{\ell=1}^q\boldsymbol{v}_\ell\boldsymbol{v}_\ell^\top$ for some unit vectors $\boldsymbol{v}_\ell\in\mathbb{R}^T$. This implies that

$$\eta_2=\sum_{\ell=1}^q|\boldsymbol{v}_\ell^\top(\boldsymbol{\xi}^2-\mathbf{1}_T)|^2\le q\max_{\ell=1,\ldots,q}\Big|\sum_{t\in\mathcal{T}}v_{\ell,t}(\xi_t^2-1)\Big|^2.$$

Hence, large deviations of η_1 and η_2 can be controlled using standard tail bounds; see, for instance, Laurent and Massart (2000, Lemma 1). This implies that with probability at least $1-2\varepsilon$,

$$\Big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\mathbf{M}\mathbf{R}^\top\big(\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\boldsymbol{\xi}^2-\mathbf{1}_T)+\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\big)\Big|_2\le\sqrt{1+\widehat{D}_1|\boldsymbol{\xi}|_\infty^2}(\sqrt{q}+\sqrt{2\log(q/\varepsilon)})+\sqrt{q\widehat{D}_1}\,4\log(2q/\varepsilon).$$

To conclude, it suffices to remark that $\mathbf{P}(|\boldsymbol{\xi}|_\infty \le \sqrt{2\log(2T/\varepsilon)}) \ge 1-\varepsilon$. This implies that

$$\begin{aligned}
\Big|\mathbf{D}_{\boldsymbol{Y}}\mathbf{RMR}^\top\big(\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}(\mathbf{D}_{\boldsymbol{\xi}}^2-\mathbf{I}_T)\mathbf{1}_T+\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1}\mathbf{D}_{\mathbf{X}\boldsymbol{\phi}^*}\boldsymbol{\xi}\big)\Big|_2&\\
&\le 2\sqrt{\widehat{D}_1\log(2T/\varepsilon)}(\sqrt{q}+\sqrt{2\log(q/\varepsilon)})+\sqrt{q\widehat{D}_1}\,4\log(2q/\varepsilon)\\
&\le 4\sqrt{2q\widehat{D}_1\log(2T/\varepsilon)\log(q/\varepsilon)}+4\sqrt{q\widehat{D}_1}\;\log(2q/\varepsilon)\\
&\le 10\sqrt{q\widehat{D}_1\log(2T/\varepsilon)\log(2q/\varepsilon)}.
\end{aligned}$$

This completes the proof of the first claim of the lemma.

Let us now switch to a proof of (38). It is clear that

$$\begin{aligned}
\|\mathbf{D}_{\boldsymbol{Y}}\mathbf{RMR}^\top\mathbf{D}_{\boldsymbol{Y}}\| &= \|\mathbf{M}^{1/2}\mathbf{R}^\top\mathbf{D}_{\boldsymbol{Y}}\|^2\\
&\le \|\mathbf{M}^{1/2}\mathbf{R}^\top(\mathbf{D}_{\boldsymbol{Y}}^2+\mathbf{D}_{\mathbf{R}\widehat{\boldsymbol{\alpha}}}^{-1}\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1})^{1/2}\|^2\|(\mathbf{D}_{\boldsymbol{Y}}^2+\mathbf{D}_{\mathbf{R}\widehat{\boldsymbol{\alpha}}}^{-1}\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1})^{-1/2}\mathbf{D}_{\boldsymbol{Y}}\|^2\\
&\le \|(\mathbf{D}_{\boldsymbol{Y}}^2+\mathbf{D}_{\mathbf{R}\widehat{\boldsymbol{\alpha}}}^{-1}\mathbf{D}_{\mathbf{R}\boldsymbol{\alpha}^*}^{-1})^{-1/2}\mathbf{D}_{\boldsymbol{Y}}\|^2\\
&= \max_{t\in\mathcal{T}}\frac{y_t^2}{y_t^2+(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*\cdot\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}})^{-1}}.
\end{aligned}\tag{49}$$

Using the fact that $\boldsymbol{R}_{t,:}\widehat{\boldsymbol{\alpha}} \le \widehat{D}_1\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*$ for every t, we obtain

$$\begin{aligned}
\|\mathbf{D}_{\boldsymbol{Y}}\mathbf{RMR}^\top\mathbf{D}_{\boldsymbol{Y}}\| &= \max_{t\in\mathcal{T}}\frac{y_t^2(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2}{y_t^2(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2+\widehat{D}_1^{-1}}\\
&= 1-\min_{t\in\mathcal{T}}\frac{1}{\widehat{D}_1y_t^2(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2+1}\\
&= 1-\min_{t\in\mathcal{T}}\frac{1}{\widehat{D}_1(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*+\xi_t)^2+1}.
\end{aligned}\tag{50}$$

To complete the proof of the lemma, it suffices to remark that $(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*+\boldsymbol{\xi}_t)^2 \le 2(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*)^2+2\xi_t^2 \le 2|\mathbf{X}\boldsymbol{\phi}^*|_\infty^2 + 2|\boldsymbol{\xi}|_\infty^2$ and to apply the well-known bound on the tails of the Gaussian distribution.