Appendix A
Proofs

A.1 Weighted Total Least Squares Cost Function Gradient

Denote by Diff the differential operator. It acts on a differentiable function $M_{\text{wtls}} : U \to \mathbb{R}$, where U is an open set in $\mathbb{R}^{\mathtt{m}\times\mathtt{p}}$, and gives as a result another function, the differential of M_{wtls}, $\operatorname{Diff}(M_{\text{wtls}}) : U \times \mathbb{R}^{\mathtt{m}\times\mathtt{p}} \to \mathbb{R}$. $\operatorname{Diff}(M_{\text{wtls}})$ is linear in its second argument, i.e.,

$$\operatorname{Diff}(f) := \mathrm{d}\, M_{\text{wtls}}(X, H) = \operatorname{trace}\big(M'_{\text{wtls}}(X)H^\top\big), \tag{A.1}$$

where $M'_{\text{wtls}} : U \to \mathbb{R}^{\mathtt{m}\times\mathtt{p}}$ is the derivative of M_{wtls}, and has the property

$$M_{\text{wtls}}(X+H) = M_{\text{wtls}}(X) + \mathrm{d}\, M_{\text{wtls}}(X, H) + \mathrm{o}(\|H\|_{\text{F}}), \tag{A.2}$$

for all $X \in U$ and for all $H \in \mathbb{R}^{\mathtt{m}\times\mathtt{p}}$. The notation $\mathrm{o}(\|H\|_{\text{F}})$ has the usual meaning

$$g(H) = \mathrm{o}(\|H\|_{\text{F}}) \quad :\Longleftrightarrow \quad g(H)/\|H\|_{\text{F}} \to 0 \text{ as } \|H\|_{\text{F}} \to 0.$$

We have

$$M_{\text{wtls}}(X) = \sum_{i=1}^{N} e_i^\top(X)\Gamma_i^{-1}(X)e_i(X), \qquad \text{where} \quad \Gamma_i(X) := \begin{bmatrix} X^\top & -I \end{bmatrix} W_i^{-1} \begin{bmatrix} X \\ -I \end{bmatrix}.$$

We find the derivative $M'_{\text{wtls}}(X)$ by first deriving the differential $\operatorname{Diff}(M_{\text{wtls}})$ and then representing it in the form (A.1), from which $M'_{\text{wtls}}(X)$ is extracted. The differential of M_{wtls} is

$$\begin{aligned}
&\mathrm{d}\, M_{\text{wtls}}(X, H) \\
&= \sum_{i=1}^{N} \Big(a_i^\top H \Gamma_i^{-1}(X) e_i(X) + e_i^\top(X)\Gamma_i^{-1}(X) H^\top a_i + e_i^\top(X) \operatorname{Diff}\big(\Gamma_i^{-1}(X)\big) e_i(X) \Big) \\
&= \sum_{i=1}^{N} \bigg(2 \operatorname{trace}\big(a_i e_i^\top(X)\Gamma_i^{-1}(X)H^\top\big) + \operatorname{trace}\Big(\operatorname{Diff}\big(\Gamma_i^{-1}(X)\big) e_i(X) e_i^\top(X) \Big) \bigg).
\end{aligned}$$

Using the rule for differentiation of an inverse matrix valued function, we have

$$\operatorname{Diff}\big(\Gamma_i^{-1}(X)\big) = -\Gamma_i^{-1}(X)\operatorname{Diff}\big(\Gamma_i(X)\big)\Gamma_i^{-1}(X).$$

Using the defining property (A.2), we have

$$\begin{aligned}\operatorname{Diff}\big(\Gamma_i(X)\big) &= \operatorname{Diff}\left(\begin{bmatrix} X^T & -I\end{bmatrix} W_i^{-1}\begin{bmatrix} X \\ -I\end{bmatrix}\right)\\ &= \operatorname{trace}\left(\begin{bmatrix} H^\top & 0\end{bmatrix} W_i^{-1}\begin{bmatrix} X \\ -I\end{bmatrix} + \begin{bmatrix} X & -I\end{bmatrix} W_i\begin{bmatrix} H \\ 0\end{bmatrix}\right)\\ &= 2\ \operatorname{trace}\left(\begin{bmatrix} H^\top & 0\end{bmatrix} W_i^{-1}\begin{bmatrix} X \\ -I\end{bmatrix}\right).\end{aligned}$$

Let $V_i := W_i^{-1}$ and define the partitioning

$$V_i =: \begin{array}{c} \begin{array}{cc} \quad\texttt{m} & \quad\;\texttt{p} \end{array} \\ \begin{bmatrix} V_{a,i} & V_{ab,i} \\ V_{ba,i} & V_{b,i}\end{bmatrix} \end{array} \begin{array}{l} \\ \texttt{m} \\ \texttt{p}\end{array}.$$

Then

$$\operatorname{Diff}\big(\Gamma_i(X)\big) = 2\ \operatorname{trace}\big(H^\top(V_{a,i}X - V_{ab,i})\big).$$

Substituting backwards, we have

$$\begin{aligned}\mathrm{d}\, M_{\mathrm{wtls}}(X,H) &= \sum_{i=1}^{N}\Big(2\ \operatorname{trace}\big(a_i e_i^\top(X)\Gamma_i^{-1}(X)H^\top\big)\\ &\qquad - 2\ \operatorname{trace}\big(\Gamma_i^{-1}(X)H^\top(V_{a,i}X - V_{ab,i})\Gamma_i^{-1}(X)e_i(X)e_i^\top(X)\big)\Big)\\ &= \operatorname{trace}\bigg(\bigg(2\sum_{i=1}^{N}\Big(a_i e_i^\top(X)\Gamma_i^{-1}(X)\\ &\qquad - (V_{a,i}X - V_{ab,i})\Gamma_i^{-1}(X)e_i(X)e_i^\top(X)\Gamma_i^{-1}(X)\Big)\bigg)H^\top\bigg).\end{aligned}$$

Thus

$$M'_{\mathrm{wtls}}(X) = 2\sum_{i=1}^{N}\Big(a_i e_i^\top(X)\Gamma_i^{-1}(X) - (V_{a,i}X - V_{ab,i})\Gamma_i^{-1}(X)e_i(X)e_i^\top(X)\Gamma_i^{-1}(X)\Big).$$

A.2 Structured Total Least Squares Cost Function Gradient

The differential $\operatorname{Diff}(f_0)$ is

$$\operatorname{Diff}(f_0) := \mathrm{d} f_0(X,H) = \operatorname{trace}\big(f_0'(X)H^\top\big) \tag{A.3}$$

and has the property

$$f_0(X+H) = f_0(X) + \mathrm{d} f_0(X,H) + \mathrm{o}(\|H\|_{\mathrm{F}})$$

for all $X \in U$ and for all $H \in \mathbb{R}^{n\times d}$. The function $f_0' : U \to \mathbb{R}^{n\times l}$ is the derivative of f_0. As in Appendix A.1, we compute it by deriving the differential $\mathrm{Diff}(f_0)$ and representing it in the form (A.3), from which $f_0'(X)$ is extracted.

The differential of the cost function $f_0(X) = r^\top(X)\Gamma^{-1}(X)r(X)$ is (using the rule for differentiation of an inverse matrix)

$$\mathrm{d}f_0(X,H) = 2r^\top\Gamma^{-1}\begin{bmatrix} H^\top a_1 \\ \vdots \\ H^\top a_m \end{bmatrix} - r^\top\Gamma^{-1}\big(\mathrm{d}\Gamma(X,H)\big)\Gamma^{-1}r.$$

The differential of the weight matrix

$$\Gamma = V_{\tilde r} = \mathbf{E}\,\tilde r\tilde r^\top = \mathbf{E}\begin{bmatrix} X^\top\tilde a_1 - \tilde b_1 \\ \vdots \\ X^\top\tilde a_m - \tilde b_m \end{bmatrix}\begin{bmatrix}\tilde a_1^\top X - \tilde b_1^\top & \cdots & \tilde a_m^\top X - \tilde b_m^\top\end{bmatrix},$$

where $\tilde A^\top =: \begin{bmatrix}\tilde a_1 & \cdots & a_m\end{bmatrix}$, $\tilde a_i \in \mathbb{R}^n$ and $\tilde B^\top =: \begin{bmatrix}\tilde b_1 & \cdots & b_m\end{bmatrix}$, $\tilde b_i \in \mathbb{R}^d$ is

$$\mathrm{d}\Gamma(X,H) = \mathbf{E}\begin{bmatrix} H^\top\tilde a_1 \\ \vdots \\ H^\top\tilde a_m \end{bmatrix}\tilde r^\top + \mathbf{E}\,\tilde r\begin{bmatrix}\tilde a_1^\top H & \cdots & \tilde a_m^\top H\end{bmatrix}. \tag{A.4}$$

With $M_{ij} \in \mathbb{R}^{d\times d}$ denoting the (i,j)th block of Γ^{-1},

$$\begin{aligned}\mathrm{d}f_0(X,H) &= 2\left(\sum_{i,j=1}^{N} r_i^\top M_{ij}H^\top a_j - \sum_{i,j,k,l=1}^{m} r_l^\top M_{li}H^\top\,\mathbf{E}\,\tilde a_i\tilde c_j^\top X_{\mathrm{ext}}M_{jl}r_l\right)\\ &= 2\ \mathrm{trace}\left(\Big(\sum_{i,j=1}^{m} a_j r_i^\top M_{ij} - \sum_{i,j,k,l=1}^{m}\begin{bmatrix}I & 0\end{bmatrix}V_{\tilde c,ij}X_{\mathrm{ext}}M_{jl}r_l r_l^\top M_{li}\Big)H^\top\right),\end{aligned}$$

so that

$$f_0'(X) = 2\left(\sum_{i,j=1}^{m} a_j r_i^\top M_{ij} - \sum_{i,j=1}^{m}\begin{bmatrix}I & 0\end{bmatrix}V_{\tilde c,ij}X_{\mathrm{ext}}N_{ji}\right),$$

where $N_{ji}(X) := \sum_{l=1}^{m} M_{jl}r_l \cdot \sum_{l=1}^{m} r_l^\top M_{li}$.

A.3 Fundamental Lemma

Of course, $\mathscr{N}_{\mathscr{B}}^{l} \subseteq \ker\big(\mathscr{H}_l^\top(\tilde w)\big)$. Assume by contradiction that $\ker\big(\mathscr{H}_l^\top(\tilde w)\big) \neq \mathscr{N}_{\mathscr{B}}^{l}$. Then there is a lowest degree polynomial $r \in \mathbb{R}^{\mathtt{w}}[z]$, $r(z) =: r_0 + r_1 z + \cdots + r_{l-1}z^{l-1}$, that annihilates $\mathscr{H}_l^\top(\tilde w)$, i.e.,

$$\mathrm{col}^\top(r_0, r_1, \ldots, r_l)\mathscr{H}_l(\tilde w) = 0,$$

but is not an element of $\mathscr{N}_{\mathscr{B}}^{l}$.

Consider $\mathscr{H}_{l+\mathtt{n}}(\tilde{w})$. Then

$$\ker\left(\mathscr{H}_{l+\mathtt{n}}^{\top}(\tilde{w})\right) = \text{image}\Big(r^{(1)}(z), zr^{(1)}(z), \dots, z^{l+\mathtt{n}-\mu_1}r^{(1)}(z)\ ;\ \dots\ ;\\ r^{(\mathtt{p})}(z), zr^{(\mathtt{p})}(z), \dots, z^{l+\mathtt{n}-\mu_{\mathtt{p}}-1}r^{(\mathtt{p})}(z)\ ;\ r(z), zr(z), \dots, z^{\mathtt{n}}r(z)\Big).$$

Note that $r(z), zr(z), \dots, z^{\mathtt{n}}r(z)$ are additional elements due to the extra annihilator r. If all these polynomial vectors were linearly independent on $\mathbb{R}$, then the dimension of $\ker\left(\mathscr{H}_{l+\mathtt{n}}(\tilde{w})\right)$ would be (at least) $\mathtt{p}(l+\mathtt{n})+1$. But the persistency of excitation assumption implies that the number of linearly independent rows of $\mathscr{H}_{l+\mathtt{n}}(\tilde{w})$ is at least $\mathtt{m}(l+\mathtt{n})$, so that

$$\dim\Big(\ker\left(\mathscr{H}_{l+\mathtt{n}}(\tilde{w})\right)\Big) \leq \mathtt{p}(l+\mathtt{n}).$$

Therefore, not all of these elements are linearly independent. By Lemma 7.5 and the assumption that R is row proper, the generators $r^{(1)}, \dots, r^{(\mathtt{p})}$ and all their shifts are linearly independent. It follows that there is $1 \leq k \leq \mathtt{n}$, such that

$$z^k r(z) \in \text{image}\Big(r^{(1)}(z), zr^{(1)}(z), \dots, z^{l+\mathtt{n}-\mu_1}r^{(1)}(z)\ ;\ \dots\ ;\\ r^{(\mathtt{p})}(z), zr^{(\mathtt{p})}(z), \dots, z^{l+\mathtt{n}-\mu_{\mathtt{p}}-1}r^{(\mathtt{p})}(z)\ ;\ r(z), zr(z), \dots, z^{k-1}r(z)\Big).$$

Therefore, there are $g \in \mathbb{R}[z]$ of degree $k \geq 1$ and $f \in \mathbb{R}^{1\times\mathtt{p}}[z]$, such that

$$g(z)r(z) = f(z)R(z).$$

Let λ be a root of $g(z)$. Then $f(\lambda)R(\lambda) = 0$, but by the controllability assumption, $\operatorname{rank}\left(R(\lambda)\right) = \mathtt{p}$ for all $\lambda \in \mathbb{C}$ and, consequently, $f(\lambda) = 0$. Therefore, with

$$g(z) = (z-\lambda)g'(z) \qquad \text{and} \qquad f(z) = (z-\lambda)f'(z),$$

we obtain

$$g'(z)r(z) = f'(z)R(z).$$

Proceeding with this degree lowering procedure yields $r(z) = f(z)R(z)$ and contradicts the assumption that r was an additional annihilator of $\mathscr{H}_l(w)$. Therefore, $\mathscr{H}_l(w)$ had the correct left kernel and therefore $\mathscr{N}_{\mathscr{B}}^{l} = \ker\left(\mathscr{H}_l^{\top}(\tilde{w})\right)$.

A.4 Recursive Errors-in-Variables Smoothing

By the dynamic programming principle (10.9),

$$V_t\big(\hat{x}(t)\big) = \min_{\hat{u}(t)}\left(\begin{bmatrix}\hat{u}(t)\\1\end{bmatrix}^{\top}\begin{bmatrix}V_{\tilde{u}}^{-1} & -V_{\tilde{u}}^{-1}u_{\mathrm{d}}(t)\\ * & u_{\mathrm{d}}(t)^{\top}V_{\tilde{u}}^{-1}u_{\mathrm{d}}(t)\end{bmatrix}\begin{bmatrix}\hat{u}(t)\\1\end{bmatrix}\right.\\ \left.+\begin{bmatrix}\hat{x}(t)\\1\end{bmatrix}^{\top}\begin{bmatrix}C^{\top}V_{\tilde{y}}^{-1}C & -C^{\top}V_{\tilde{y}}^{-1}y_{\mathrm{d}}(t)\\ * & y_{\mathrm{d}}(t)^{\top}V_{\tilde{y}}^{-1}y_{\mathrm{d}}(t)\end{bmatrix}\begin{bmatrix}\hat{x}(t)\\1\end{bmatrix} + V_{t+1}\big(A\hat{x}(t)+B\hat{u}(t)\big)\right), \quad \text{(A.5)}$$

where the $*$'s indicate the symmetric blocks in the matrices. Using induction, we prove that the value function V_t is quadratic for all t. At the final moment of time T, $V_T \equiv 0$ and thus it is trivially quadratic. Assume that V_{t+1} is quadratic for $t \in \{0, 1, \dots, T\}$. Then there are $P_{t+1} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$, $s_{t+1} \in \mathbb{R}^{\mathrm{n}\times 1}$, and $v_{t+1} \in \mathbb{R}^{1\times 1}$, such that

$$V_{t+1}\big(\hat{x}(t)\big) = \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix}^\top \begin{bmatrix} P_{t+1} & s_{t+1} \\ s_{t+1}^\top & v_{t+1} \end{bmatrix} \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix}, \quad \text{for all } \hat{x}(t). \tag{A.6}$$

From (A.5) and (A.6), we have

$$\begin{aligned} V_t\big(\hat{x}(t)\big) = \min_{\hat{u}(t)} \Bigg(& \begin{bmatrix} \hat{u}(t) \\ 1 \end{bmatrix}^\top \begin{bmatrix} V_{\tilde{u}}^{-1} & -V_{\tilde{u}}^{-1} u_{\mathrm{d}}(t) \\ * & u_{\mathrm{d}}(t)^\top V_{\tilde{u}}^{-1} u_{\mathrm{d}}(t) \end{bmatrix} \begin{bmatrix} \hat{u}(t) \\ 1 \end{bmatrix} \\ & + \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix}^\top \begin{bmatrix} C^\top V_{\tilde{y}}^{-1} C & -C^\top V_{\tilde{y}}^{-1} y_{\mathrm{d}}(t) \\ * & y_{\mathrm{d}}(t)^\top V_{\tilde{y}}^{-1} y_{\mathrm{d}}(t) \end{bmatrix} \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix} \\ & + \begin{bmatrix} A\hat{x}(t) + B\hat{u}(t) \\ 1 \end{bmatrix}^\top \begin{bmatrix} P_{t+1} & s_{t+1} \\ s_{t+1}^\top & v_{t+1} \end{bmatrix} \begin{bmatrix} A\hat{x}(t) + B\hat{u}(t) \\ 1 \end{bmatrix} \Bigg). \end{aligned} \tag{A.7}$$

The function to be minimized in (A.7) is a convex quadratic function of $\hat{u}(t)$,

$$\begin{bmatrix} \hat{u}(t) \\ 1 \end{bmatrix}^\top \begin{bmatrix} B^\top P_{t+1} B + V_{\tilde{u}}^{-1} & B^\top P_{t+1} A\hat{x}(t) + B^\top s_{t+1} - V_{\tilde{u}}^{-1} u_{\mathrm{d}}(t) \\ * & \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix}^\top M(t) \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix} \end{bmatrix} \begin{bmatrix} \hat{u}(t) \\ 1 \end{bmatrix},$$

where

$$M(t) := \begin{bmatrix} A^\top P_{t+1} A + C^\top V_{\tilde{y}}^{-1} C & A^\top s_{t+1} - C^\top V_{\tilde{y}}^{-1} y_{\mathrm{d}}(t) \\ * & v_{t+1} + y_{\mathrm{d}}(t)^\top V_{\tilde{y}}^{-1} y_{\mathrm{d}}(t) + u_{\mathrm{d}}(t)^\top V_{\tilde{u}}^{-1} u_{\mathrm{d}}(t) \end{bmatrix}$$

so the minimizing $\hat{u}(t)$ is

$$\hat{u}(t) = -\big(B^\top P_{t+1} B + V_{\tilde{u}}^{-1}\big)^{-1} \big(B^\top P_{t+1} A\hat{x}(t) + B^\top s_{t+1} - V_{\tilde{u}}^{-1} u_{\mathrm{d}}(t)\big). \tag{A.8}$$

Substituting (A.8) back into (A.5), we have

$$\begin{aligned} V_t(\hat{x}(t)) = & \begin{bmatrix} \hat{u}(t) \\ 1 \end{bmatrix} \begin{bmatrix} B^\top P_{t+1} B + V_{\tilde{u}}^{-1} & B^\top P_{t+1} A\hat{x}(t) + B^\top s_{t+1} - V_{\tilde{u}}^{-1} u_{\mathrm{d}}(t) \\ * & u_{\mathrm{d}}(t)^\top V_{\tilde{u}}^{-1} u_{\mathrm{d}}(t) \end{bmatrix} \begin{bmatrix} \hat{u}(t) \\ 1 \end{bmatrix} \\ & + \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix}^\top \begin{bmatrix} A^\top P_{t+1} A + C^\top V_{\tilde{y}}^{-1} C & A^\top s_{t+1} - C^\top V_{\tilde{y}}^{-1} y_{\mathrm{d}}(t) \\ * & v_{t+1} + y_{\mathrm{d}}(t)^\top V_{\tilde{y}}^{-1} y_{\mathrm{d}}(t) \end{bmatrix} \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix}, \end{aligned}$$

which is a quadratic function of $\hat{x}(t)$,

$$V_t(\hat{x}(t)) = \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix}^\top \begin{bmatrix} P_t & s_t \\ s_t^\top & v_t \end{bmatrix} \begin{bmatrix} \hat{x}(t) \\ 1 \end{bmatrix}, \quad \text{for all } \hat{x}(t),$$

with P_t and s_t given in (10.11) and (10.12), respectively. By induction, V_t is quadratic for $t = 0, 1, \dots, T$.