Asymptotic properties of RLS and of the tuning unit

F. A. Ramponi
Rev. 0.3, 2021-06-04

This document contains an asymptotic analysis of the Least Squares estimate. As far as the final exam is concerned, the RLS derivation, the important definitions, and the take-home messages are requested as part of the program, but the tedious computations in the convergence analysis are not.

References:
1) Sergio Bittanti and Marco C. Campi. *Least squares based self-tuning control systems.* In: Identification, Adaptation, Learning. The science of learning models from data (S. Bittanti and G. Picci eds.). Springer-Verlag NATO ASI series - Computer and systems sciences, pages 339-365, 1996.
2) Sergio Bittanti, *Identificazione dei modelli e sistemi adattativi*, Pitagora Editrice, Bologna.
3) Private notes by Simone Garatti.

0.1 ***Non-weighted*** RLS algorithm: existence of the limit

Assumption 0.1.1

The plant is deterministic, and there exists a "true" model in the model class that describes it perfectly. In other words, there exists a parameter $\bar{\theta}$ that explains the measures exactly:

$$y_t = \varphi_t^\top \bar{\theta} \qquad \text{for all } t.$$

<u>*Note*</u>: most of what follows holds under fairly general conditions, *mutatis mutandis*, also if a process noise is present and the regressors are random vectors, i.e. if

$$y_t = \boldsymbol{\varphi}_t^\top \bar{\theta} + \boldsymbol{\varepsilon}_t$$

where $\boldsymbol{\varphi}_t$ are a random vectors and $\boldsymbol{\varepsilon}_t$ is a zero-mean *white noise* independent of $\boldsymbol{\varphi}_t$. But for the purposes of these notes you can safely forget about the noise and assume that the regressors are deterministic: if not for other reasons, at least to fix ideas.

To estimate $\bar{\theta}$ we employ the regularized, *non-weighted* Least Squares estimator regularized with $\lambda > 0$:

$$R_t = \lambda I + \sum_{\tau=0}^{t} \varphi_\tau \varphi_\tau^\top, \qquad S_t = \sum_{\tau=0}^{t} \varphi_\tau y_\tau,$$
$$\hat{\theta}_t = R_t^{-1} S_t \qquad \text{(the unique solution of the normal equations).}$$

For the analysis we will exploit the structure of the *non-weighted* Recursive Least Squares (RLS) algorithm, which is a scheme to update $\hat{\theta}_t \to \hat{\theta}_{t+1}$ without re-doing the entire computation and the matrix inversion. Here follows a fast paced review of the algorithm's derivation.

First, notice that

$$R_{t+1} = \sum_{\tau=0}^{t+1} \varphi_\tau \varphi_\tau^\top = R_t + \varphi_{t+1} \varphi_{t+1}^\top,$$
$$S_{t+1} = \sum_{\tau=0}^{t+1} \varphi_\tau y_\tau = S_t + \varphi_{t+1} y_{t+1}.$$

Second, work out:

$$\begin{aligned}
\hat{\theta}_{t+1} &= R_{t+1}^{-1} S_{t+1} \\
&= R_{t+1}^{-1} \left(S_t + \varphi_{t+1} y_{t+1}\right) \\
&= R_{t+1}^{-1} \left(R_t \hat{\theta}_t + \varphi_{t+1} y_{t+1}\right) \\
&= R_{t+1}^{-1} \left(R_{t+1} \hat{\theta}_t - \varphi_{t+1}\varphi_{t+1}^\top \hat{\theta}_t + \varphi_{t+1} y_{t+1}\right) \\
&= \hat{\theta}_t + \underbrace{R_{t+1}^{-1}\varphi_{t+1}}_{K_{t+1}} \left(y_{t+1} - \varphi_{t+1}^\top \hat{\theta}_t\right). \qquad (1)
\end{aligned}$$

(The vector K_{t+1} is a *gain* multiplying the *residual* $y_{t+1} - \varphi_{t+1}^\top \hat{\theta}_t$.) Third, to update from R_t^{-1} to R_{t+1}^{-1} apply Woodbury's identity (matrix inversion lemma) with $A = R_t$, $B = \varphi_{t+1}$, $C = 1$, and $D = \varphi_{t+1}^\top$:

$$\begin{aligned}
R_{t+1}^{-1} &= \left(R_t + \varphi_{t+1} \cdot 1 \cdot \varphi_{t+1}^\top\right)^{-1} \\
&= R_t^{-1} - R_t^{-1}\varphi_{t+1} \left(1 + \varphi_{t+1}^\top R_t^{-1} \varphi_{t+1}\right)^{-1} \varphi_{t+1}^\top R_t^{-1} \\
&= R_t^{-1} - \frac{R_t^{-1}\varphi_{t+1}\varphi_{t+1}^\top R_t^{-1}}{1 + \varphi_{t+1}^\top R_t^{-1}\varphi_{t+1}}.
\end{aligned}$$

(Note the great advantage: $1 + \varphi_{t+1}^\top R_t^{-1}\varphi_{t+1}$ is a *number*, so in the end to apply RLS we need divisions by numbers instead of inversions of matrices/solutions of linear equations.) It follows:

$$K_{t+1} = R_{t+1}^{-1}\varphi_{t+1} = R_t^{-1}\varphi_{t+1}\left(1 - \frac{\varphi_{t+1}^\top R_t^{-1}\varphi_{t+1}}{1 + \varphi_{t+1}^\top R_t^{-1}\varphi_{t+1}}\right) = \frac{R_t^{-1}\varphi_{t+1}}{1 + \varphi_{t+1}^\top R_t^{-1}\varphi_{t+1}},$$

so that, in particular,

$$R_{t+1}^{-1} = R_t^{-1} - K_{t+1}\varphi_{t+1}^\top R_t^{-1} = \left(I - K_{t+1}\varphi_{t+1}^\top\right) R_t^{-1}.$$

At this stage one realizes that there are many inversions of R around, but none of them is actually performed. Indeed the usual RLS scheme prescribes to deal with the inverse matrix $P_t := R_t^{-1}$ only; substituting it in the previous expressions we get the final algorithm:

Take-home message 0.1.2 (RLS algorithm with regularization)

Initialization:

$$\begin{aligned}
\hat{\theta}_{-1} &= \text{an arbitrary vector in } \mathbb{R}^p; \\
P_{-1} &= (\lambda I)^{-1} = I/\lambda.
\end{aligned}$$

As regressors/measures $(\varphi_0, y_0), (\varphi_1, y_1), \ldots, (\varphi_t, y_t), (\varphi_{t+1}, y_{t+1})$ come one after another, keep computing the gain K and updating the estimate $\hat{\theta}$ and the inverse matrix P:

$$\begin{aligned}
K_{t+1} &= \frac{P_t\varphi_{t+1}}{1 + \varphi_{t+1}^\top P_t \varphi_{t+1}}; \\
\hat{\theta}_{t+1} &= \hat{\theta}_t + K_{t+1}\left(y_{t+1} - \varphi_{t+1}^\top \hat{\theta}_t\right); \\
P_{t+1} &= P_t - \frac{P_t\varphi_{t+1}\varphi_{t+1}^\top P_t}{1 + \varphi_{t+1}^\top P_t \varphi_{t+1}} = \left(I - K_{t+1}\varphi_{t+1}^\top\right) P_t.
\end{aligned}$$

Note: this is just a *fast* procedure to compute the LS estimate at subsequent times. Indeed, $\hat{\theta}_{t+1}$ continues to be the LS estimate at time $t+1$ irrespective of the recursion used to compute it; if one computed it by means of the formula $\hat{\theta}_{t+1} = R_{t+1}^{-1} S_{t+1}$ with the matrix inversion s/he would obtain the same result.

And now some cool matrix stuff...

Tools from linear algebra 0.1.3 (Positive semi-definite matrices)

A symmetric matrix $P = P^\top \in \mathbb{R}^{p\times p}$ is called *positive semi-definite* if $\theta^\top P\theta \geq 0$ for all $\theta \in \mathbb{R}^p$. This is denoted $P \geq 0$. Some facts follow:

- For any matrix $\Phi \in \mathbb{R}^{n\times p}$, the matrix $\Phi^\top\Phi$ is positive semidefinite; indeed $\theta^\top\Phi^\top\Phi\theta = (\Phi\theta)^\top\Phi\theta = \|\Phi\theta\|^2 \geq 0$. For example, if $\varphi_\tau \in \mathbb{R}^{p\times 1}$ is a column vector then $\varphi_\tau\varphi_\tau^\top \geq 0$.
- A sum of positive semi-definite matrices is also positive semi-definite. So, $\sum_{\tau=0}^{t}\varphi_\tau\varphi_\tau^\top \geq 0$.
- The notion of positive semi-definiteness induces a partial ordering between symmetric matrices: we write $P_1 \geq P_2$ (or $P_2 \leq P_1$) whenever $P_1 - P_2 \geq 0$. This is a well-defined order relation between symmetric matrices; however, it is *partial*, i.e. not all pairs P_1, P_2 are comparable.
- If $P_0 \geq P_1 \geq P_2 \geq \ldots \geq P_t \geq \ldots$ is a monotone non-increasing sequence of matrices bounded from below ($P_t \geq \bar{P}$ for all t, for example $P_t \geq 0$), then the sequence has a limit:

  $$\lim_{t\to\infty} P_t = P_\infty.$$

  (This is a generalization of what happens with monotone sequences of numbers.)

Tools from linear algebra 0.1.4 (Positive definite matrices)

A symmetric matrix $P = P^\top \in \mathbb{R}^{p\times p}$ is called *positive definite* if $\theta^\top P\theta > 0$ for all $\theta \neq 0$. This is denoted $P > 0$. Of course if $P > 0$ then also $P \geq 0$; some other facts follow:

- If $P > 0$ then P is nonsingular and hence invertible (with positive eigenvalues). Its inverse is also positive definite: $P^{-1} > 0$.
- If $P_1 \geq P_2$ and $P_2 > 0$, then also $P_1 > 0$. Moreover, $P_1^{-1} \leq P_2^{-1}$.
- If λ is a positive *number*, then $\lambda I > 0$. Indeed $\theta^\top(\lambda I)\theta = \lambda\|\theta\|^2 > 0$ for all $\theta \neq 0$.
  So, $\lambda I + \sum_{\tau=0}^{t}\varphi_\tau\varphi_\tau^\top > 0$.
- Any positive definite matrix $P \in \mathbb{R}^{p\times p}$ induces a well-defined *scalar product* $\langle\cdot,\cdot\rangle$ over $\mathbb{R}^p$ trough the following definition:

  $$\langle\theta_1, \theta_2\rangle := \theta_1^\top P\theta_2.$$

  The converse is also true: every scalar product over $\mathbb{R}^p$ is induced by a positive definite matrix (the standard one, $\langle\theta_1, \theta_2\rangle = \theta_1^\top\theta_2$, is induced by the identity matrix I).

From the properties of positive definite matrices we get the following facts:

- The symmetric matrices $R_t = \lambda I + \sum_{\tau=0}^{t}\varphi_\tau\varphi_\tau^\top$, $t \in \mathbb{N}$, are all positive definite, and form a monotone *non-decreasing* sequence:

  $$\ldots \leq R_{t-1} \leq R_t \leq R_{t+1} \leq \ldots$$

- therefore, their inverses $P_t = R_t^{-1}$, $t \in \mathbb{N}$, are also positive definite, and they form a monotone *non-increasing* sequence

  $$\ldots \geq P_{t-1} \geq P_t \geq P_{t+1} \geq \ldots$$

  bounded from below by $\bar{P} = 0$;
- so, the sequence $\ldots, P_t, P_{t+1}, \ldots$ converges, i.e. $\lim_{t\to\infty} P_t = P_\infty$ exists. (The limit can be a singular matrix, or even the zero matrix.)

Recall an intermediate step in the derivation of the RLS algorithm (equation (1)):

$$\begin{aligned}\hat{\theta}_{t+1} &= \hat{\theta}_t + P_{t+1}\varphi_{t+1}\left(y_{t+1} - \varphi_{t+1}^\top \hat{\theta}_t\right)\\ &= \hat{\theta}_t + P_{t+1}\varphi_{t+1}\left(\varphi_{t+1}^\top \bar{\theta} - \varphi_{t+1}^\top \hat{\theta}_t\right)\\ &= \hat{\theta}_t - P_{t+1}\varphi_{t+1}\varphi_{t+1}^\top\left(\hat{\theta}_t - \bar{\theta}\right).\end{aligned} \tag{2}$$

Now define the estimation error $\tilde{\theta}_t := \hat{\theta}_t - \bar{\theta}$. Substituting it into (2) and subtracting $\bar{\theta}$ from each side we get:

$$\hat{\theta}_{t+1} - \bar{\theta} = \hat{\theta}_t - \bar{\theta} - P_{t+1}\varphi_{t+1}\varphi_{t+1}^\top \tilde{\theta}_t;$$

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - P_{t+1}\varphi_{t+1}\varphi_{t+1}^\top \tilde{\theta}_t; \tag{3}$$

$$R_{t+1}\tilde{\theta}_{t+1} = R_{t+1}\tilde{\theta}_t - \varphi_{t+1}\varphi_{t+1}^\top \tilde{\theta}_t = \left(R_{t+1} - \varphi_{t+1}\varphi_{t+1}^\top\right)\tilde{\theta}_t = R_t\tilde{\theta}_t. \tag{4}$$

Conclusion: repeating the argument (4) at subsequent times one can derive $R_t\tilde{\theta}_t = R_{t+1}\tilde{\theta}_{t+1} = R_{t+2}\tilde{\theta}_{t+2} = R_{t+3}\tilde{\theta}_{t+3} = \ldots$, and hence the vector $R_t\tilde{\theta}_t$ is constant as a function of t. (Now it would be a good time to review the tank example from the beginning of the course, because this is exactly the same line of reasoning.) Consequently:

$$\begin{aligned}\lim_{t\to\infty}\tilde{\theta}_t &= \lim_{t\to\infty}\left(P_t \cdot R_t\tilde{\theta}_t\right)\\ &= \left(\lim_{t\to\infty} P_t\right)\cdot\left(\lim_{t\to\infty} R_t\tilde{\theta}_t\right)\\ &= (\text{a limit that exists})\cdot(\text{the limit of a constant})\,;\end{aligned} \tag{5}$$

both the limits within parentheses exist finite, the second one being the constant in (4), and hence:

Take-home message 0.1.5

The limit of $\tilde{\theta}_t$ exists finite; of course it follows that the limit of $\hat{\theta}_t = \tilde{\theta}_t + \bar{\theta}$ also exists.
We will denote them, respectively,

$$\tilde{\theta}_\infty := \lim_{t\to\infty}\tilde{\theta}_t, \qquad \hat{\theta}_\infty := \lim_{t\to\infty}\hat{\theta}_t.$$

Note: although in the spirit of the Least Squares method one wishes that $\hat{\theta}_\infty = \bar{\theta}$, in general $\hat{\theta}_\infty \neq \bar{\theta}$. The limit of $\hat{\theta}_t$ depends both on the "true" parameter and on the sequence of regressor vectors; a particular sequence of regressors can drive the estimate to any "true" parameter only if R_t "diverges in all directions" and consequently if $P_t = R_t^{-1}$ tends to 0;[a] when this happens we say that the LS estimate is *consistent*. But we will not yet assume that this is the case: indeed the rest of the document contains an analysis of the general case when $\hat{\theta}_t$ is *not* consistent.

[a] See equation (5): if $P_t \to 0$, then $\tilde{\theta}_t \to 0$ and $\hat{\theta}_t \to \bar{\theta}$.

0.2 *Non-weighted* least squares: properties of the limit estimate

The following is a somewhat tedious, but otherwise straightforward computation to make an important point about the limit $\tilde{\theta}_\infty$ (see the take-home message 0.2.1 below). Always keep in mind that the matrices $R_t, P_t \in \mathbb{R}^{p\times p}$ are the inverses of each other, and that the quantity $R_t\tilde{\theta}_t$ is constant with respect to t (therefore, so is its *transpose* $\tilde{\theta}_t^\top R_t$). Here is an intermediate step in the derivation of the RLS algorithm (update of P_t) and the update of $\tilde{\theta}_t$ (equation (3)):

$$P_{t+1} = P_t - \frac{P_t\varphi_{t+1}\varphi_{t+1}^\top P_t}{1 + \varphi_{t+1}^\top P_t\varphi_{t+1}}; \tag{6}$$

$$\tilde{\theta}_{t+1} = \tilde{\theta}_t - P_{t+1}\varphi_{t+1}\varphi_{t+1}^\top\tilde{\theta}_t \qquad (\text{substitute } P_{t+1})$$

$$= \tilde{\theta}_t - P_t\varphi_{t+1}\varphi_{t+1}^\top\tilde{\theta}_t + \frac{P_t\varphi_{t+1}\varphi_{t+1}^\top P_t\varphi_{t+1}\varphi_{t+1}^\top\tilde{\theta}_t}{1 + \varphi_{t+1}^\top P_t\varphi_{t+1}}. \tag{7}$$

It follows:

$$\begin{aligned}
\tilde{\theta}_{t+1}^{\top} R_{t+1} \tilde{\theta}_{t+1} &= \tilde{\theta}_t^{\top} R_t \tilde{\theta}_{t+1} \qquad \text{(because } \tilde{\theta}_t^{\top} R_t \text{ is constant; now substitute (7))} \\
&= \tilde{\theta}_t^{\top} R_t \left(\tilde{\theta}_t - P_t \varphi_{t+1} \varphi_{t+1}^{\top} \tilde{\theta}_t + \frac{P_t \varphi_{t+1} \varphi_{t+1}^{\top} P_t \varphi_{t+1} \varphi_{t+1}^{\top} \tilde{\theta}_t}{1 + \varphi_{t+1}^{\top} P_t \varphi_{t+1}} \right) \\
&= \tilde{\theta}_t^{\top} R_t \tilde{\theta}_t - \tilde{\theta}_t^{\top} R_t P_t \varphi_{t+1} \varphi_{t+1}^{\top} \tilde{\theta}_t + \frac{\tilde{\theta}_t^{\top} R_t P_t \varphi_{t+1} \varphi_{t+1}^{\top} P_t \varphi_{t+1} \varphi_{t+1}^{\top} \tilde{\theta}_t}{1 + \varphi_{t+1}^{\top} P_t \varphi_{t+1}} \\
&= \tilde{\theta}_t^{\top} R_t \tilde{\theta}_t - \tilde{\theta}_t^{\top} \varphi_{t+1} \varphi_{t+1}^{\top} \tilde{\theta}_t + \frac{\tilde{\theta}_t^{\top} \varphi_{t+1} \left(\varphi_{t+1}^{\top} P_t \varphi_{t+1} \right) \varphi_{t+1}^{\top} \tilde{\theta}_t}{1 + \varphi_{t+1}^{\top} P_t \varphi_{t+1}} \qquad \text{(note: } \varphi_{t+1}^{\top} P_t \varphi_{t+1} \text{ is scalar)} \\
&= \tilde{\theta}_t^{\top} R_t \tilde{\theta}_t - \left(\tilde{\theta}_t^{\top} \varphi_{t+1} \right) \left(\varphi_{t+1}^{\top} \tilde{\theta}_t \right) \cdot \left(1 - \frac{\varphi_{t+1}^{\top} P_t \varphi_{t+1}}{1 + \varphi_{t+1}^{\top} P_t \varphi_{t+1}} \right) \\
&= \tilde{\theta}_t^{\top} R_t \tilde{\theta}_t - \frac{\left(\varphi_{t+1}^{\top} \tilde{\theta}_t \right)^2}{1 + \varphi_{t+1}^{\top} P_t \varphi_{t+1}} \\
&\leq \tilde{\theta}_t^{\top} R_t \tilde{\theta}_t, \qquad \text{because the second term in the previous expression is negative or null.}
\end{aligned}$$

Since this inequality is valid for all t, we have

$$\ldots \leq \tilde{\theta}_{t+1}^{\top} R_{t+1} \tilde{\theta}_{t+1} \leq \tilde{\theta}_t^{\top} R_t \tilde{\theta}_t \leq \tilde{\theta}_{t-1}^{\top} R_{t-1} \tilde{\theta}_{t-1} \leq \ldots \leq \tilde{\theta}_1^{\top} R_1 \tilde{\theta}_1 \leq \tilde{\theta}_0^{\top} R_0 \tilde{\theta}_0 := V_0,$$

where the number $V_0 = \tilde{\theta}_0^{\top} R_0 \tilde{\theta}_0 = \tilde{\theta}_0^{\top} \left(\lambda I + \varphi_0 \varphi_0^{\top} \right) \tilde{\theta}_0$ is unknown, because we don't know the initial estimation error $\tilde{\theta}_0$; however, we know that V_0 is a *finite* number. Therefore, the sequence $\tilde{\theta}_t^{\top} R_t \tilde{\theta}_t$ is *bounded* as a function of t:

$$\tilde{\theta}_t^{\top} R_t \tilde{\theta}_t \leq V_0 \qquad \text{for all } t. \tag{8}$$

Now recall: $\ldots, R_t, R_{t+1}, \ldots$ is a non-decreasing sequence *of matrices*. This means that, for all $\theta \in \mathbb{R}^p$, the following is a non-decreasing sequence *of numbers*:

$$\ldots \leq \theta^{\top} R_t \theta \leq \theta^{\top} R_{t+1} \theta \leq \ldots \leq \theta^{\top} R_{t+k} \theta \leq \ldots$$

In particular, this is true for $\theta = \tilde{\theta}_{t+k}$:

$$\begin{aligned}
\tilde{\theta}_{t+k}^{\top} R_t \tilde{\theta}_{t+k} &\leq \tilde{\theta}_{t+k}^{\top} R_{t+k} \tilde{\theta}_{t+k} \qquad \text{for all } t \text{ and } k\text{, because the sequence } \ldots R_t \ldots \text{ is non decreasing} \\
&\leq V_0 \qquad \text{because of (8).}
\end{aligned}$$

Letting k (but not t) tend to infinity, we recover:

$$\tilde{\theta}_{\infty}^{\top} R_t \tilde{\theta}_{\infty} = \lim_{k \to \infty} \tilde{\theta}_{t+k}^{\top} R_t \tilde{\theta}_{t+k} \leq V_0 \qquad \text{for all } t$$

and finally, since the sequence of numbers $\tilde{\theta}_{\infty}^{\top} R_t \tilde{\theta}_{\infty}$, $t \in \mathbb{N}$, is non-decreasing and bounded from above by V_0, it has a finite limit.

Take-home message 0.2.1

The quantity $\tilde{\theta}_{\infty}^{\top} R_t \tilde{\theta}_{\infty} = \tilde{\theta}_{\infty}^{\top} \left(\lambda I + \sum_{\tau=0}^{t} \varphi_\tau \varphi_\tau^{\top} \right) \tilde{\theta}_{\infty}$ has a <u>finite</u> limit as $t \to \infty$.

The above take-home message paves the ground for the following definition:

Definition 0.2.2 (Unexcitation subspace)

The set $\mathcal{U} \subseteq \mathbb{R}^p$ defined as follows,

$$\mathcal{U} = \left\{ \theta \in \mathbb{R}^p \; : \; \lim_{t \to \infty} \theta^{\top} R_t \theta < +\infty \right\},$$

is called the *unexcitation subspace*.

In order to show that the set $\mathcal{U} \subseteq \mathbb{R}^p$ is indeed a subspace of $\mathbb{R}^p$ we need to recall a fundamental fact of Euclidean spaces:

Tools from linear algebra 0.2.3 (Cauchy-Schwarz inequality)

Let $\langle \cdot, \cdot \rangle$ denote *any* scalar product[a] in a vector space $\mathcal{V}$, and define the function $\|\cdot\|$ as $\|x\| := \sqrt{\langle x, x \rangle}$. For all $x, y \in \mathcal{V}$ it holds

$$|\langle x, y \rangle| \leq \|x\| \, \|y\|.$$

For the proof see e.g. Luenberger, *Optimization by vector space methods.* An immediate consequence of this *fundamental* inequality is that $\|\cdot\|$ is a *norm*:

$$\begin{aligned} \|x+y\|^2 &= \langle x+y, x+y \rangle \\ &= \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle \\ &\leq \|x\|^2 + \|y\|^2 + 2\|x\| \, \|y\| \\ &= (\|x\| + \|y\|)^2 . \end{aligned} \tag{9}$$

Taking square roots we get the triangular inequality $\|x+y\| \leq \|x\| + \|y\|$, the key property of a norm.

[a]Examples: in $\mathbb{R}^n$ the canonical scalar product is $\langle x, y \rangle = x^\top y$; but another totally legitimate scalar product is $\langle x, y \rangle = x^\top P y$, where P is an arbitrary symmetric and positive definite matrix. In a space of functions: $\langle x, y \rangle = \int_{-\infty}^{+\infty} x(t) y(t) dt$; etc.

Since R_t is a positive definite symmetric matrix for all t (indeed $R_t \geq \lambda I > 0$), it defines a scalar product $\langle \theta_1, \theta_2 \rangle = \langle \theta_1, \theta_2 \rangle_{R_t} := \theta_1^\top R_t \theta_2$ and a norm $\|\theta\| = \sqrt{\langle \theta, \theta \rangle_{R_t}}$ over $\mathbb{R}^p$. Now let $\theta_1, \theta_2 \in \mathcal{U}$; we have

$$\begin{aligned} (\theta_1 + \theta_2)^\top R_t (\theta_1 + \theta_2) &= \|\theta_1 + \theta_2\|^2 \\ &\leq (\|\theta_1\| + \|\theta_2\|)^2 \qquad \text{(see (9))} \\ &= \left(\theta_1^\top R_t \theta_1 + \theta_2^\top R_t \theta_2\right)^2 . \end{aligned}$$

Since both θ_1 and θ_2 belong to $\mathcal{U}$, taking limits as $t \to \infty$ the last expression converges to a finite quantity; therefore, so does the first expression in the chain of (in)equalities. This proves that if $\theta_1, \theta_2 \in \mathcal{U}$ then $\theta_1 + \theta_2 \in \mathcal{U}$; the proof for the multiplication by a scalar is trivial.

Definition 0.2.4 (Excitation subspace)

The orthogonal complement of $\mathcal{U}$ in $\mathbb{R}^p$ with respect to the standard scalar product $\langle \theta_1, \theta_2 \rangle = \theta_1^\top \theta_2$,

$$\mathcal{E} = \mathcal{U}^\perp,$$

is called the *excitation subspace.* It is indeed a subspace of $\mathbb{R}^p$ due to the properties of orthogonal complements. Note that, by this very definition, for any vector $\theta \in \mathcal{E}$, $\theta \neq 0$, the quantity $\theta^\top R_t \theta$ diverges as $t \to \infty$.

As you know, any vector in $\mathbb{R}^p$ admits a *unique* decomposition as the sum of a vector in a subspace and a vector in the orthogonal complement of that subspace:

$$\theta = \theta^{\mathcal{E}} + \theta^{\mathcal{U}}. \tag{10}$$

(Think at $\theta^{\mathcal{E}}$ and $\theta^{\mathcal{U}}$ as the *orthogonal projections* of θ on $\mathcal{E}$ and $\mathcal{U}$ respectively.)

The entire point of the definition of $\mathcal{E}$ and $\mathcal{U}$ is that if the component $\theta^{\mathcal{E}} \in \mathcal{E}$ is not zero or, stated another way, if θ does not belong to $\mathcal{U}$, then $\theta^\top R_t \theta$ *must diverge* as $t \to \infty$.

To formally prove this, we rely on another immediate consequence of the Cauchy-Schwarz inequality. Fix *any* scalar product $\langle \cdot, \cdot \rangle$ and its induced norm $\| \cdot \|$ in $\mathbb{R}^p$. It holds

$$\begin{aligned} \|\theta_1 + \theta_2\|^2 &= \|\theta_1\|^2 + \|\theta_2\|^2 + 2\langle \theta_1, \theta_2 \rangle \\ &\geq \|\theta_1\|^2 + \|\theta_2\|^2 - 2\|\theta_1\|\|\theta_2\| \qquad \text{(Cauchy-Schwarz inequality)} \\ &= (\|\theta_1\| - \|\theta_2\|)^2 . \end{aligned} \tag{11}$$

Now: consider again the decomposition (10), but apply to it the scalar product and norm defined by

$$\langle \theta_1, \theta_2 \rangle := \theta_1^\top R_t \theta_2, \qquad \|\theta\| := \sqrt{\langle \theta, \theta \rangle} = \sqrt{\theta^\top R_t \theta}. \tag{12}$$

I know, I know: switching back and forth from a scalar product to another may lead to some confusion, but I promise that this actually simplifies computations a lot. Let me stress again that the one in (12) is a totally legitimate scalar product in $\mathbb{R}^p$ because we have assumed the presence of the regularization term $\lambda I > 0$ from the beginning, hence R_t is *always* positive definite. We obtain:

$$\begin{aligned} \theta^\top R_t \theta = \|\theta\|^2 &= \left\|\theta^{\mathcal{E}} + \theta^{\mathcal{U}}\right\|^2 \\ &\geq \left(\|\theta^{\mathcal{E}}\| - \|\theta^{\mathcal{U}}\|\right)^2 \qquad \text{(see equation (11))} \\ &= \left(\underbrace{\sqrt{(\theta^{\mathcal{E}})^\top R_t \theta^{\mathcal{E}}}}_{\text{diverges as } t \to \infty} - \underbrace{\sqrt{(\theta^{\mathcal{U}})^\top R_t \theta^{\mathcal{U}}}}_{\text{converges as } t \to \infty} \right)^2 . \end{aligned}$$

This proves the claim: if the component $\theta^{\mathcal{E}} \in \mathcal{E}$ is not 0, the last expression diverges as $t \to \infty$, and so does $\theta^\top R_t \theta$. (It is now OK to forget about (12) and return to the standard scalar product $\langle \theta_1, \theta_2 \rangle = \theta_1^\top \theta_2$ of Euclidean geometry, i.e. the one used in the definition of $\mathcal{E}$.)

So what? What's the point of this mess?

Read again: for any $\theta \in \mathbb{R}^p$, if $\theta^{\mathcal{E}} \neq 0$ then $\lim_{t\to\infty} \theta^\top R_t \theta = +\infty$.

Now have another look at the take-home message 0.2.1...

Take-home message 0.2.5 (Limit of the projection on $\mathcal{E}$)

The point becomes clear if we apply the above discussion to the vector $\theta = \tilde{\theta}_\infty$, because we know already that $\tilde{\theta}_\infty^\top R_t \tilde{\theta}_\infty$ does, instead, converge as $t \to \infty$. Therefore, the only possibility is that the component of $\tilde{\theta}_\infty$ in the subspace $\mathcal{E}$ is 0:

- if $\tilde{\theta}_t^{\mathcal{E}}$ and $\tilde{\theta}_\infty^{\mathcal{E}}$ are, respectively, the projections of $\tilde{\theta}_t$ and $\tilde{\theta}_\infty$ on the excitation subspace, then

$$\lim_{t\to\infty} \tilde{\theta}_t^{\mathcal{E}} = \tilde{\theta}_\infty^{\mathcal{E}} = 0;$$

- if $\hat{\theta}_t^{\mathcal{E}}$ and $\bar{\theta}^{\mathcal{E}}$ are, respectively, the projections of $\hat{\theta}_t$ and $\bar{\theta}$ on the excitation subspace, then

$$\lim_{t\to\infty} \hat{\theta}_t^{\mathcal{E}} = \bar{\theta}^{\mathcal{E}}.$$

In words: *the component of the LS estimate in the excitation subspace converges, as t tends to infinity, to the component of the "true" parameter on the excitation subspace.*

The intuition behind all this analysis is that, as time goes on, the regressor vectors φ_t keep "adding energy" and "exploring", consistently, along some directions, but possibly not along others: the directions along which the regressor vectors "keep exploring" span the excitation subspace $\mathcal{E}$, and along this subspace the LS estimate is, in the long run, exact.

In general, we cannot say anything about what happens along directions *orthogonal* to $\mathcal{E}$, that is, we cannot say anything about the component of the estimate in the *unexcitation* subspace $\mathcal{U}$, other that it converges to a finite (possibly wrong) vector $\hat{\theta}_\infty^{\mathcal{U}} \neq \bar{\theta}^{\mathcal{U}}$, because in $\mathcal{U}$ the regressors do not "explore" enough. On the other hand *sometimes*, if some stringent condition holds on the actual sequence of regressors, we can prove that the unexcitation subspace is trivial, i.e. $\mathcal{U} = \{0\}$ or, which is the same thing, that the excitation subspace $\mathcal{E}$ is the whole $\mathbb{R}^p$. When this is the case, the LS estimate is, in the long run, always exact, and we call it *consistent*.

Definition 0.2.6 (Consistency)

The LS estimate $\hat{\theta}_t$ is called *consistent* if $\mathcal{U} = \{0\}$ and $\mathcal{E} = \mathbb{R}^p$. When this is the case it holds

$$\lim_{t\to\infty} \hat{\theta}_t = \bar{\theta},$$

whatever the actual "true" parameter $\bar{\theta}$ happens to be.

For example, here is such a consistency condition. If the limit

$$\lim_{t\to\infty} \frac{1}{t+1} \sum_{\tau=0}^{t} \varphi_\tau \varphi_\tau^\top = \Sigma$$

exists and is a positive definite matrix, then $\mathcal{E} = \mathbb{R}^p$ and $\lim_{t\to\infty} \hat{\theta}_t = \bar{\theta}$. Indeed, for all $\theta \in \mathbb{R}^p$, $\theta \neq 0$,

$$\begin{aligned}
\lim_{t\to\infty} \theta^\top R_t \theta &= \lim_{t\to\infty} \theta^\top \left(\lambda I + \sum_{\tau=0}^{t} \varphi_\tau \varphi_\tau^\top \right) \theta \\
&= \lim_{t\to\infty} \lambda \|\theta\|^2 + \theta^\top \left(\sum_{\tau=0}^{t} \varphi_\tau \varphi_\tau^\top \right) \theta \\
&= \underbrace{\lambda \|\theta\|^2}_{\text{constant}} + \lim_{t\to\infty} \underbrace{(t+1)}_{\text{diverges}} \cdot \lim_{t\to\infty} \underbrace{\theta^\top \left(\frac{1}{t+1} \sum_{\tau=0}^{t} \varphi_\tau \varphi_\tau^\top \right) \theta}_{\text{converges to the number } \theta^\top \Sigma \theta > 0} = +\infty,
\end{aligned}$$

therefore all non-zero vectors in $\mathbb{R}^p$ belong to $\mathcal{E}$, which proves the claim.

In *system identification* a hypothesis of this kind, that here is stated as a condition *on the regressors*, depends ultimately *on the information carried by the input signal*; such a condition is typically called *persistent excitation* (of the input), and an experimenter identifying a plant needs it to guarantee that the "true" parameter of the plant is identified correctly in the long run. However, when dealing with self-tuning regulators such condition may be desirable but is *not* needed: we can show that, under fairly general conditions, internal stability of the closed loop and reference tracking for the desired class of reference signals are ensured even if complete identifiability of the plant is not there.

Take-home message 0.2.7 (Recall the tank example)

The tank example from the beginning of the course adopts the RLS algorithm regularized with $\lambda = 1$. To keep things simple, the example is 1-dimensional ($p = 1$), hence the regressor $\varphi_\tau = x(\tau)$ is just a number.

The example is indeed simple because $\mathbb{R}^1$ has only two subspaces: the trivial subspace $\{0\}$, and $\mathbb{R}^1$ itself. Then there are only two possibilities: either $R_t = 1 + \sum_{\tau=0}^{t} x(\tau)^2$ diverges as $t \to \infty$ (the term 1+ is irrelevant here), and in this case $\mathcal{E} = \mathbb{R}^1$ and the estimate is consistent; or R_t converges, and in this case $\mathcal{U} = \mathbb{R}^1$, $\mathcal{E} = \{0\}$ and the limit estimate may be wrong. Nevertheless, the self-tuning regulator achieves internal stability and reference tracking in both cases.