Modelling Neuroplasticity: Using Machine Learning to Learn About TMS Data

Santiago López Pereyra, Jennifer Goldschmied, Coauthor 3, Coauthor 4, ...

June 11, 2023

1 Introduction

The emergence of new computational models has provided neuroscientists with novel ways of extracting significant insights from experimental data. This paper takes a computational approach to challenges arising from the analysis of transcranial magnetic stimulation (TMS) data under the double pulse paradigm. We argue that using indicators of neuroplasticity for each specific double stimulation, as opposed to their average, allows for the use of data-driven methods of analysis.

Particularly, we formulate pulse-specific measures of relative amplitude between paired and test pulses. These measures represent the proportionality of specific paired-pulse responses with respect to test-pulse responses. Namely, they serve as proxies for neuroplasticity in the brain by capturing the degree of inhibition or facilitation induced by each double-stimulation.

The benefits of data-driven methods of analysis cannot be overstated. Their potential implication in the production of scientific evidence and the diagnosis of clinical disorders go beyond the scope of this paper and deserve further investigation.

2 Background

Transcranial magnetic stimulation (TMS) is a common, non-invasive experimental technique used to evoke action potentials in cortical regions of the brain. In particular, researchers often target the motor cortex and measure the motor evoked potential (MEP) aroused by the stimulation. When the purpose of the experiment is to inquire on neuroplasticity, such stimulations are performed under the *paired pulse paradigm.*

The *paired pulse paradigm* (or *double pulse paradigm*) consists in eliciting a series of two temporally proximate pulses (in the order of miliseconds). The evoked potentials of the double-stimulation are compared to those of single test stimulations, and their relative amplitude is taken as a proxy of neuroplasticity in the brain. The time separating each of the paired pulses is termed *interstimulus interval* (ISI). It is the general case that low intervals (4 or 5 miliseconds) produce intracortical inhibition, with the evoked potentials of paired stimulations being generally lower than those of single pulses. Greater intervals, on the other hand, tend to produce facilitation.

In the context of this paper, we shall term any coefficient that serves to represent the proportional relationship between the potentials evoked by paired and test pulses a *measure of relative amplitude* (MRA).

The goal of this paper is to provide pulse-specific measures of relative amplitude. This is, coefficients that represent the relative amplitude of each individual paired pulse with respect to the set of test pulses in an experimental session. The benefit of this is that it keeps data resolution at its highest, which allows for the implementation of data-driven artificial intelligence in the analysis of the experimental results. Thus, from a computer science perspective, our objective is constrained to the sphere of feature engineering.

We will show how pulse-specific measures of relative amplitude allow for otherwise unfeasible computational analyses of TMS data, such as the use of machine learning models for the detection of different pulse-response patterns among different groups of clinical subjects. In particular, we will show they allow for a machine learning classifier to correctly determine whether a subject belongs to one of four clinical categories, based only on its evoked potentials and across different inter-stimulus intervals, with an accuracy of up to 90%.

3 Dataset

Our case study uses data from $N = 43$ subjects at UPenn's *Lab for the Study of Slow-wave sleep activity*, with $H = 17$ healthy controls (HC) and $D = 26$ with major depressive disorder (MDD). Subjects received transcranial magnetic stimulation of the motor cortex after both a night of baseline sleep and a night of slow-wave disruption (SWD) sleep. The stimulations were performed under the double pulse paradigm. Per session, subjects received twenty-four (24) test pulses in total and eight (8) paired pulses per ISI. The ISIs were $4, 5, 8, 10, 15$ and 20. The evoked potentials were measured by placing an EMG electrode on the targeted muscle of the thumb. In the slow-wave disruption session, we sent a constant auditory stimulus with sufficient strength to interrupt the normal occurrence of slow-wave sleep, yet not strong enough to wake the subjects.

Please see the table below to see the resulting subject categories:

	Baseline	Slow-wave disruption
Healthy control	HC BL	HC SWD
Major depressive disorder	MDD BL	MDD SWD

Each observation comes from an individual transcranial stimulation. The features of the resulting database are:

1. An ***EMG*** variable with the peak-to-peak EMG of each elicited pulse.
2. A ***Label*** categorical feature, which represents the group of the subject of each observation.
3. An ***ISI*** feature that represents the inter-stimulus interval of each pulse. A value of -1 indicated that the given pulse was a test pulse (no inter-stimulus interval).

The distribution of the evoked potentials is always exponential. However, the $\boldsymbol{\beta}$ parameter of the distribution varies across subject groups and session types, as shown in Figures 1 and 2 and in the table below:

	Mean	Median
HC BL	196.77	149.36
HC SWD	165.98	99.76
MDD BL	187.56	127.69
MDD SWD	247.57	151.07

Observe that each of the means in the above table corresponds to the estimated $\boldsymbol{\beta}$ parameter of the distribution of the paired pulses on each subject group.

4 Limitations of the traditional TMS method of analysis

Measures of relative amplitude in neuroscience are generally computed at the subject and group level by repeatedly taking averages. This is natural, since scientific interest generally resides in evaluating differences across subject groups. However, it also raises problems pertaining to the downscaling of the data resolution and the reliability of the resulting values.

The procedure to establish the average relative amplitude of a group of subjects (for example, healthy controls) is simple. Traditionally, it involves taking the quotient of the average paired-pulse and the average test pulse responses per every subject. Next, those quotients are averaged out across all subjects in the subject group. In short, it is an average of averages.

Although this practice is not arbitrary —since hypotheses generally deal with differences across subject groups— it has three major limitations. Firstly, it implies greatly downscaling the data resolution, making data-driven methods of analysis unfeasible. Secondly, potentials evoked by test and paired pulses present a very high variance (Rossini et. al., 20015; Orth, Snijders and Rothwell, 2003; Wassermann, 2002), which —being that it is a quotient of two means— raises questions about the quality of the representation where relatively few pulses are elicited. Thirdly, our observations suggest potentials evoked by TMS stimulations follow an exponential distribution, which puts in question whether arithmetic means are a reliable measure of central tendency.

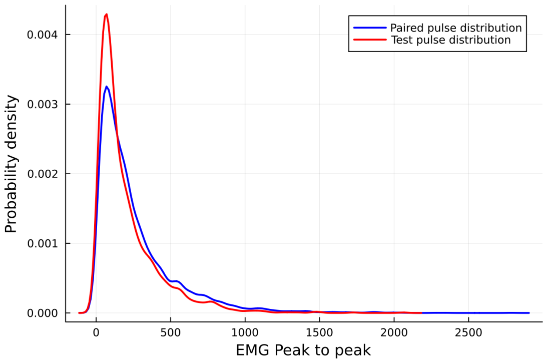


Figure 1: Distribution of paired and test pulses

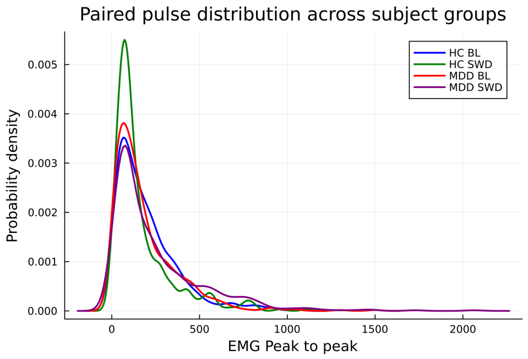


Figure 2: Distribution of paired pulses across subject groups

To make matters worse, when considering the size of the data across multiple experimental subjects, the shrinkage in data resolution is far from negligible. And in times when scientific inquiry can greatly benefit from forms of artificial intelligence that rely on large amounts of data (e.g. machine learning), it is clearly limiting to count solely with low-resolution measures to interpret experimental results.

5 Proposed approach

5.1 Pulse-specific relative amplitude measures

In light of the technical concerns about the traditional approach, this paper proposes two pulse-specific relative amplitude measures. This allows for otherwise unfeasible computational analyses of TMS data, such as the use of machine learning (ML) models for the detection of different pulse-response patterns among different groups of clinical subjects.

Pulse-specific measures of relative amplitude are coefficients that represent, for every experimental subject, the proportionality of each individual paired-pulse response with respect to the set of *all* test-pulse responses. From a neuroscientific perspective, they are a proxy for neuroplasticity at the level of each particular stimulation, measuring the amount facilitation or inhibition in each specific response. From a computational perspective, these measures prescind from the need of averaging paired-pulses out, and thus keep data resolution at its highest.

In order to test these measures, we look at whether they improve the performance of a ML classification model and in what degree.[1] Our case study is a random forest model in Python tasked with classifying every observation with the appropriate subject label. The model we use is an AdaBoost classifier consisting of 20 decision trees. Every tree in the ensemble has a maximum depth of 8. The boosting algorithm is *SAMME.R*. The learning rate is set to $\eta = 0.01$. We use holdout cross-validation to evaluate the performance of the model. The total number of values in our dataset is $n = 9382$.

The goal of the classifier is to infer, based on the properties of each TMS response, the group of the subject upon which the stimulation was elicited, as well as the type of sleep session after which it occurred. Our hypothesis is that pulse-specific measures of relative amplitude will significantly improve the model's performance.

Section **5** provides the mathematical definitions of our measures of relative amplitude. Section **6** discusses the impact of each measure on the performance of the random forest model. Specifically, section **6.1** shows the performance of the model when trained over the raw data (i.e. the features described in section **3.1**). Section **6.2** shows the performance of the model when including pulse-specific measures of relative amplitude. Section **6.3** shows the performance of the model when combining the pulse-specific measures with logarithmic transformations of the raw data and the use of weighted instead of arithmetic averages.

As for the computation of these measures, we implemented the formulas of section **5.2** for every TMS

[1] We use a classification model for practical reasons —namely, because the samples in our data were conveniently labeled by subject group and sleep session.

response using the Julia programming language.

5.2 Definitions: Pulse specific relative amplitudes

For simplicity, we will first deal with the hypothetical situation in which a single ISI is used for paired stimulations in a unique TMS session. Our formulations generalize easily to different inter-stimulus intervals and multiple experimental sessions.

Firstly, we need to establish some convenient notation. Let n and m be the number of paired and test pulses elicited to the subject in the session, respectively. Let $\mathbf{x}$, $\mathbf{t}$ be vectors over the $\mathbb{R}^+$ vector space whose components are understood to be the evoked potentials of paired and test stimulations, respectively.

Definition 1 *Let $x \in \boldsymbol{x}$ be the evoked potential of a single paired stimulation. Then we define two pulse-specific relative amplitude measures, $\rho : \mathbb{R} \to \mathbb{R}, \delta : \mathbb{R} \to \mathbb{R}$, where*

$$\rho(x) := \frac{mx}{\sum_{i=1}^{m} t_j} \tag{1}$$

$$\delta(x) := \frac{x}{m} \sum_{j=1}^{m} \frac{1}{t_j} \tag{2}$$

Theorem 1 $\forall x : x \in \mathbb{R}^+ : \delta(x) \geq \rho(x)$. *(For a proof of this property, consult the appendix.)*

Notice that $\rho(x)$ is the proportion between the potential x, evoked by a paired stimulation, with respect to the average potential of single test stimulations. On the other hand, $\delta(x)$ is the average proportion of x with respect to each single test pulse.

To put it differently, ρ is a representation of the relative importance of each double pulse in relation to the overall distribution of the test pulses in the subject. It measures the deviation of the double pulse x with respect to the average test pulse. δ, in contrast, measures the proportionality of a double pulse with respect to different values in the distribution of the test pulses of the subject.

Notice as well that ρ and δ are defined over the evoked potentials of a specific experimental session. In other words, whenever we speak of the δ and ρ *features* (*id est*, of the data resulting from broadcasting ρ and δ over a collection of evoked potentials), such features are understood to be subject- and session-specific.

Theorem 2 *Let π denote the coefficient traditionally used for estimating the relative amplitude between test and paired pulses. In accordance to what was described in the second paragraph of Section 4,*

$$\pi = \frac{\mu_x}{\mu_t}.$$

where μ_x, μ_t are the mean paired and test pulse responses, respectively. It follows from the definition of π and ρ that

$$\pi = \frac{\rho(x_1) + \ldots + \rho(x_n)}{n}.$$

(For a proof, consult the appendix.)

The second remark is of great importance. It states that the way in which relative amplitude was traditionally computed—the way in which neuroplasticity was traditionally proxied—is different from our proposed ρ function only in terms of scale. More specifically, the traditional coefficient used to proxy neuroplasticity at the subject level is nothing but the mean ρ measure in an experimental session. We believe this bestows ρ with a certain degree of confidence, insofar as not only it is not entirely new, but neuroscientists have implicitly used it all along.

5.3 Generalizing to multiple subjects

Definition 2 *Let $\boldsymbol{x} \in \mathbb{R}^j$ with $j \in \mathbb{N}$. Then $T : \mathbb{R}^j \to \mathbb{R}^j, R : \mathbb{R}^j \to \mathbb{R}^j$ are the linear transformations defined as*

$T(\boldsymbol{x}) := \begin{bmatrix}\rho(x_1) & \ldots & \rho(x_j)\end{bmatrix}^\intercal$ and $R(\boldsymbol{x}) := \begin{bmatrix}\delta(x_1) & \ldots & \delta(x_j)\end{bmatrix}^\intercal$.

It is trivial to show that T and R are linear mappings (consult the Appendix). Now let $k \in \mathbb{N}$ denote the number of experimental subjects in an experiment, to each of whom n, m paired and test pulses were elicited, respectively.

Definition 3 *We define $\boldsymbol{P} \in \mathbb{R}^{k \times n}$ a matrix where the ith row vector $\boldsymbol{x}^{(i)}$ is the vector whose components are the paired-pulse responses of the ith subject.*

Definition 4 *Let $\mu_i^{(t)}$ be average test-pulse response of the ith subject. Let $\mu_i^{(t)^{-1}}$ be the average reciprocal test-pulse response of the ith subject. Then $\boldsymbol{M}^{k \times k}$ is a diagonal matrix with $\boldsymbol{M}_{(ii)} := \frac{1}{\mu_i^{(t)}}$, and $\boldsymbol{M}'^{\,k \times k}$ is a diagonal matrix with $\boldsymbol{M}'_{ii} := \mu_i^{(t^{-1})}$.*

Theorem 3 *The row-vectors of* $\boldsymbol{X}_\rho = \boldsymbol{MP}$ *are the vectors* $T(\boldsymbol{x}^{(1)}), \dots, T(\boldsymbol{x}^{(k)})$*, and the row-vectors of* $\boldsymbol{X}_\delta = \boldsymbol{M'P}$ *are the vectors* $R(\boldsymbol{x}^{(1)}), \dots, R(\boldsymbol{x}^{(k)})$*.*

If experiments were conducted using multiple ISI, then to each ISI r it corresponds a distinct matrix $\mathbf{X}_\delta^{ISI=r}, \mathbf{X}_\rho^{ISI=r}$. The same applies to different subject groups.

We have expanded ρ, δ to higher dimensions for the purpose of providing a theoretical framework of their application to larger data sets and many experimental subjects. However, it is important to note that the matrix representations here presented only make sense under the assumption that, on each subject, none of n paired pulses were anomalous, and hence all can be safely included in the model. Such assumption rarely holds in reality, where it is usual to excluded some pulses due to the presence of artifacts. There are many ways to deal with this situation in the implementation of our model—the safest being the direct application of T and R to the data vectors—. Thus, more than serving the algorithmic implementation of the model, which may prescind from them, these matrices provide insight into the exact nature of the transformations involved in the model.

6 Results

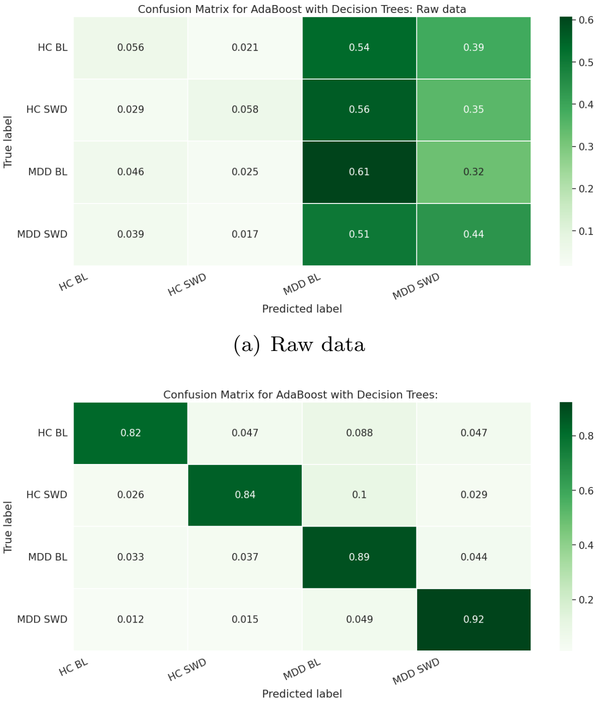
6.1 Random forest without MRA

The first random forest model is our control group. We trained it on the raw data —that is, without the inclusion of pulse-specific MRA. The model's accuracy on the testing set was of ≈ 34.2%. The confusion matrix of the model, **Figure 3.***a*, shows that errors concentrated towards healthy control categories. Additionally, the categorization of diagnosed subjects was poor, although substantially better than that of healthy controls.

Although the model almost never confuses an MDD subject with a healthy control, this does not necessarily imply that the model is capturing key differences between these groups. On the contrary, this is more likely a consequence of the model's bias in favour of the MDD categories. This bias is probably caused by the fact that the dataset has more MDD subjects than healthy controls. In consequence, the model learns the rudimentary strategy of always guessing for one of the MDD categories.

6.2 Random forest with MRA

The following models incorporated the ρ feature, then the δ feature, and then both, respectively. The inclusion of the ρ feature increased accuracy by a factor of ≈ 2.1 to 72.6%. Subsequently, the inclusion of the δ feature had a similar impact, increasing accuracy to 73.4%. Lastly, the inclusion of both the δ and ρ features improved accuracy to ≈ 87.1%. The confusion matrix of the model trained with both ρ and δ can be observed in **Figure 3.***b*.



(a) Raw data

(b) ρ and δ features included.

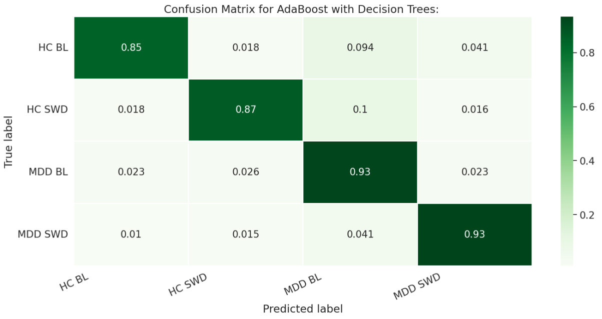
Figure 3: Compared performance: raw data and data with measures of relative amplitude.

6.3 Random forest with additional features

Since the neural response to transcranial stimulations follows an exponential distribution (see Figures 1 and 2), we tested the ρ and δ functions over the logarithmically transformed peak-to-peak EMG. This had a negligible positive impact in the overall performance.

On the contrary, the inclusion of weighted variants of the features defined above greatly improves the performance of the model. The use of weighted instead of arithmetic averages in the formulas for δ and ρ is useful in dealing with the excessive influence of outliers or highly spread out points in the feature. For example, the weight vectors may be computed using the MAD or the inverse-variance of each $t \in \mathbf{t}$.

For example, using inverse-variance weights accuracy increased to $\approx 90.2\%$. The confusion matrix of this last model can be appreciated in **Figure 4**.



(a) $\rho, \delta, \rho(\ln x), \delta(\ln x), \rho_w$ and δ_w included.

Figure 4: Confusion matrix

In summary, our results demonstrate that pulse-specific MRA greatly improve the accuracy of a random forest model trained over TMS data for subject classification. More importantly, it becomes clear that, by using the engineered features δ and ρ, researchers can attain evidence in favour or against hypotheses pertaining to differences among subject groups. Particularly, the model provides otherwise unattainable evidence in favour of sleep-modulated, depression-induced differences in neuroplasticity.

7 Conclusion

This paper shows that pulse-specific relative amplitude measures allow for the use of machine learning models over TMS experimental data. Such models can play an important part in TMS research by introducing new ways of both analyzing and extracting meaningful insights from the experimental results.

The use of machine learning models in neuroscientific observations is not only a promising tool in the production of evidence. There is a diagnostic potential that is yet to be evaluated. By detecting distinct neural patterns between healthy control and diagnosed subjects,[2] these models can potentially serve as a useful tool in the diagnostic process.

Although long and serious scientific effort is still required to appraise the diagnostic potential of these technologies, we believe our results allow for cautious optimism on the matter.

It should however be noted that, although machine learning models reveal distinct response patterns and differences across subjects, they do not disclose the neurobiology modulating those differences nor what those patterns are.

Lastly, the impact of pulse-specific MRA in data-driven models other than random forests is yet to be tested. These empirical trials are not trivial because failure or success of the models is dependent not only on the engineered features, but on the nature of the experimental data itself.

8 References

1. Lefaucheur, JP. “Transcranial Magnetic Stimulation.” Handbook of clinical neurology. U.S. National Library of Medicine. Accessed April 14, 2023. https://pubmed.ncbi.nlm.nih.gov/31277876.

2. Orth, M, AH Snijders, and JC Rothwell. “The Variability of Intracortical Inhibition and Facilitation.” Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology. U.S. National Library of Medicine. Accessed April 14, 2023. https://pubmed.ncbi.nlm.nih.gov/14652096/.

3. Rossini, PM, D Burke, LG Cohen, Z Daskalakis, R Di Iorio, V Di Lazzaro, F Ferreri, et al. “Non-Invasive Electrical and Magnetic Stimulation of the Brain, Spinal Cord, Roots and Peripheral Nerves: Basic Principles and Procedures for Routine Clinical and Research Application. an Updated Report from an I.F.C.N. Committee.” Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology. U.S. National Library of Medicine. Accessed April 14, 2023. https://pubmed.ncbi.nlm.nih.gov/25797650/.

[2] At least under specific experimental conditions.

4. Wassermann, EM. “Variation in the Response to Transcranial Magnetic Brain Stimulation in the General Population.” Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology. U.S. National Library of Medicine. Accessed April 14, 2023. https://pubmed.ncbi.nlm.nih.gov/12088713/.

9 Appendix

Proof 1. In **Remark 1**, we observed the following property:

$$\forall x : x \in \mathbb{R}^+ : \delta(x) \geq \rho(x).$$

Such property can be proven via induction. Firstly, recall that

$$\delta(x) = \frac{x}{m}\sum^{m} \frac{1}{t_j}$$
$$\rho(x) = \frac{xm}{\sum^{m} t_j}$$

Let $S_1^m = \sum^m \frac{1}{t_j}, S_2^m = \sum^m t_j$. We operate under the assumption that $t_i \in \mathbb{R}^+$. It is the case that

$$\frac{x}{m}\sum^{m} \frac{1}{t_j} \geq \frac{xm}{\sum^{m} t_j}$$
$$S_2^m S_1^m \geq m^2$$

This holds for $m = 1$, since $\frac{1}{t_1} + t_1 \geq 1 \iff 1 + t_1^2 \geq t_1$. So we may assume $S_1^k S_2^k \geq k^2$. We now set out to show that

$$S_1^{k+1} S_2^{k+1} \geq (k+1)^2$$

This can be proven as follows.

$$\begin{aligned} S_1^{k+1} S_2^{k+1} &\geq (k+1)^2 \\ (S_1^k + \frac{1}{t_{t+1}})(S_2^k + t_{k+1}) &\geq k^2 + 2k + 1 \\ S_1^k S_2^k + t_{k+1} S_1^k + \frac{1}{t_{k+1}} S_2^k + 1 &\geq k^2 + 2k + 1 \\ S_1^k S_2^k + t_{k+1} S_1^k + \frac{1}{t_{k+1}} S_2^k &\geq k^2 + 2k \end{aligned}$$

We know $S_1^k S_2^k \geq k^2$ and then it suffices to show $t_{k+1} S_1^k + \frac{S_2^k}{t_{k+1}} \geq 2k$. To prove this, simply observe that

$$\begin{aligned} \frac{1}{t_{k+1}} \sum_{j=1}^{m} t_j + t_{k+1} \sum_{j=1}^{m} \frac{1}{t_j} &\geq 2k \\ \left(\frac{t_1}{t_{k+1}} + ... + \frac{t_k}{t_{k+1}}\right) + \left(\frac{t_{k+1}}{t_1} + ... + \frac{t_{k+1}}{t_k}\right) &\geq 2k \\ \Longleftrightarrow \overbrace{a + \frac{1}{a} + b + \frac{1}{b} + ... + n + \frac{1}{n}}^{2k \text{ terms}} &\geq 2k \end{aligned}$$

We have $\min f = 2$ for $f(x) = x + \frac{1}{x}$ for $x \in \mathbb{R}^+$. Then $\min(a + \frac{1}{a} + ... + n + \frac{1}{n}) = 2k$ for $a, ..., n \in \mathbb{R}^+$, which concludes the demonstration.

Proof 2. In **Remark 2**, we observe that ρ is implicitly used in the traditional measures of relative amplitude. Recall that such methods consist in taking the average paired-pulse response and dividing it by the average test-pulse response. It is easily observed that

$$\frac{\mu_x}{\mu_d} = \frac{\frac{1}{n}\sum_{j=1}^{n} x_j}{\frac{1}{m}\sum_{j=1}^{m} t_j} = \frac{\rho(x_{1i}) + \cdots + \rho(x_{ni})}{n}$$

Proof 1 (Theorem 1) *Let $\sigma = \sum_{i=1}^{m} t_j$ and $c \in \mathbb{R}$. It follows directly from the definition of ρ that*

$$\begin{aligned} T(c\boldsymbol{x} + \boldsymbol{y}) &= \begin{bmatrix} \frac{m(cx_1)+y_1}{\sigma} & \ldots & \frac{m(cx_n+y_n)}{\sigma} \end{bmatrix}^{\mathsf{T}} \\ &= \begin{bmatrix} \frac{mcx_1}{\sigma} + \frac{my_1}{\sigma} & \ldots & \frac{mcx_n}{\sigma} + \frac{y_n}{\sigma} \end{bmatrix} \\ &= cT(\boldsymbol{x}) + T(\boldsymbol{y}) \end{aligned}$$

The proof the R is linear is practically identical.