Supplementary note 1: Potential and limitations of quantum extreme learning machines

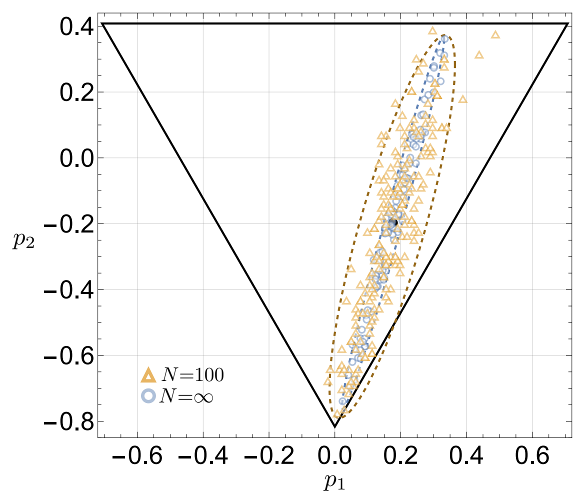
L. Innocenti,[1] S. Lorenzo,[1] I. Palmisano,[2] A. Ferraro,[2,3] M. Paternostro,[2] and G. M. Palma[1,4]

[1] *Università degli Studi di Palermo, Dipartimento di Fisica e Chimica - Emilio Segrè, via Archirafi 36, I-90123 Palermo, Italy*
[2] *Centre for Theoretical Atomic, Molecular, and Optical Physics, School of Mathematics and Physics, Queen's University Belfast, BT7 1NN, United Kingdom*
[3] *Quantum Technology Lab, Dipartimento di Fisica Aldo Pontremoli, Università degli Studi di Milano, I-20133 Milano, Italy*
[4] *NEST, Istituto Nanoscienze-CNR, Piazza S. Silvestro 12, 56127 Pisa, Italy*

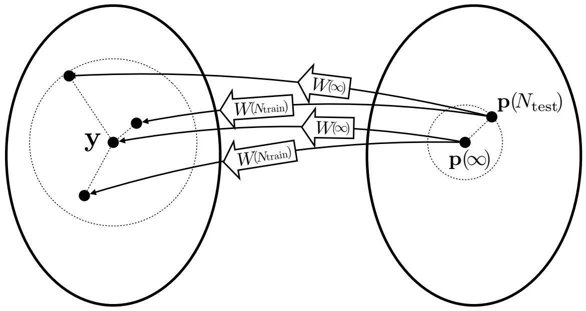
The purpose of this Supplementary Information is to discuss more in detail the phenomenon, briefly mentioned in the main text, of the underestimation of the condition number for higher training statistics.

As in the main text, we denote with P_N the matrix whose columns are the frequencies obtained estimating the outcome probabilities with N samples for different training states. The columns of P_N are thus finite-sample estimates of the columns of P. Even though $\kappa(P_N)$ can be larger for larger N, the corresponding estimation error always decreases with N, because the inaccuracies in the estimated probability vectors also decrease with N counteracting the increased noise sensitivity flagged by $\kappa(P_N)$. A rough intuition for why $\kappa(P_N)$ often increases with N can be obtained as follows: an arbitrary matrix P_N can be pictured as the ellipsoid that it maps the unit (hyper)sphere to. The singular values of P_N are then proportional to the lengths of the principal axes of this ellipsoid. The condition number equals the ratio between larger and smaller (nonzero) principal axes. In our case, the columns of P_N are subject to statistical noise that causes them to fluctuate by a quantity of the order of $1/N$. Consequently, directions corresponding to singular values smaller than the statistical noise, will appear larger, with lengths in the order of $\sim 1/N$. The overall result is that for small N the directions corresponding to the smallest singular values of P_N might appear larger, thus causing $\kappa(P_N)$ to be underestimated. This underestimation will become negligible when the statistical noise has magnitude significantly smaller than the smallest singular value of P_N. This also explains why the underestimation is most prominent in situations where $\kappa(P_N)$ is larger, which is generally due to the smallest singular value of P_N being smaller. This phenomenon is displayed in a simple case in Supplementary Fig. 1.

Another way to understand the potentially odd behaviour of the condition number discussed in the main text, and in particular its increasing with the number of training statistics $N_{\rm train}$, is to observe that when $N_{\rm train}$ is significantly larger than the testing statistics $N_{\rm test}$, one might incur in a phenomenon analogous to overfitting. Indeed, even though using large $N_{\rm train} \gg 1$ results in a weight matrix $W(\infty)$ which sends ideal output probabilities $\mathbf{p}(\infty)$ to the corresponding accurate expectation values, it is possible that this $W(\infty)$ significantly amplifies errors in the probabilities $\mathbf{p}(N_{\rm test})$ estimated with finite statistics, and thus results in an overall larger estimation error, unless $N_{\rm test}$ is also large enough to overcome this effect. For this reason, having $N_{\rm train} \gg N_{\rm test}$ may result in overall decreased performances, because even though $W(\infty)$ sends ideal probabilities $\mathbf{p}(\infty)$ into perfectly estimated expectation values, noisy probabilities $\mathbf{p}(N_{\rm test})$ might be sent to estimated expectation values worse than those that would have been produced with $W(N_{\rm train})$. This phenomenon is shown schematically in Supplementary Fig. 2.



Supplementary Fig. 1. *Underestimation of condition number for limited statistics* — We consider random qubit states measured with a random three-outcome POVM with unit-rank operators. Each state is thus represented as a length-3 probability vector. Exploiting the normalisation, each such vector can be projected onto the two dimensions tangent space to the 3-dimensional simplex, p_1 and p_2 in the figure. The blue dots represent here the probability vector associated to each state. The orange triangles, those obtained sampling from the same probability vectors with finite statistics (in this case $N = 100$ sampled were used). The dashed ellipses are drawn using as principal axes the principal components of the corresponding matrices of probabilities, and have principal lengths corresponding to the associated singular values. The condition number of P_N is then proportional to the ratio between largest and smallest principal axes of its ellipse. As clearly seen from the spread of the orange triangles here, points estimated from finite statistics result in larger smallest singular values, and therefore smaller condition numbers.



Supplementary Fig. 2. Schematic illustration of how statistical estimation errors are amplified for different training and testing statistics $N_{\rm train}$ and $N_{\rm test}$. When $N_{\rm train} \to \infty$, the associated weight matrix $W(\infty)$ sends ideal probabilities $\mathbf{p}(\infty)$ to ideally estimated expectation values $\mathbf{y}$. However, if $N_{\rm test}$ is finite, then $W(\infty)$ operates on estimates probabilities $\mathbf{p}(N_{\rm test})$, and the statistical errors in $\mathbf{p}(N_{\rm test})$ might be significantly amplified by $W(\infty)$. On the other hand, using finite training statistics, $W(N_{\rm train})$ incorrectly estimates $\mathbf{p}(\infty)$, but might amplify the errors in $\mathbf{p}(N_{\rm train})$ less than $W(\infty)$. The overall effect is that using $N_{\rm train} \gg N_{\rm test}$ might negatively impact the estimation MSE.