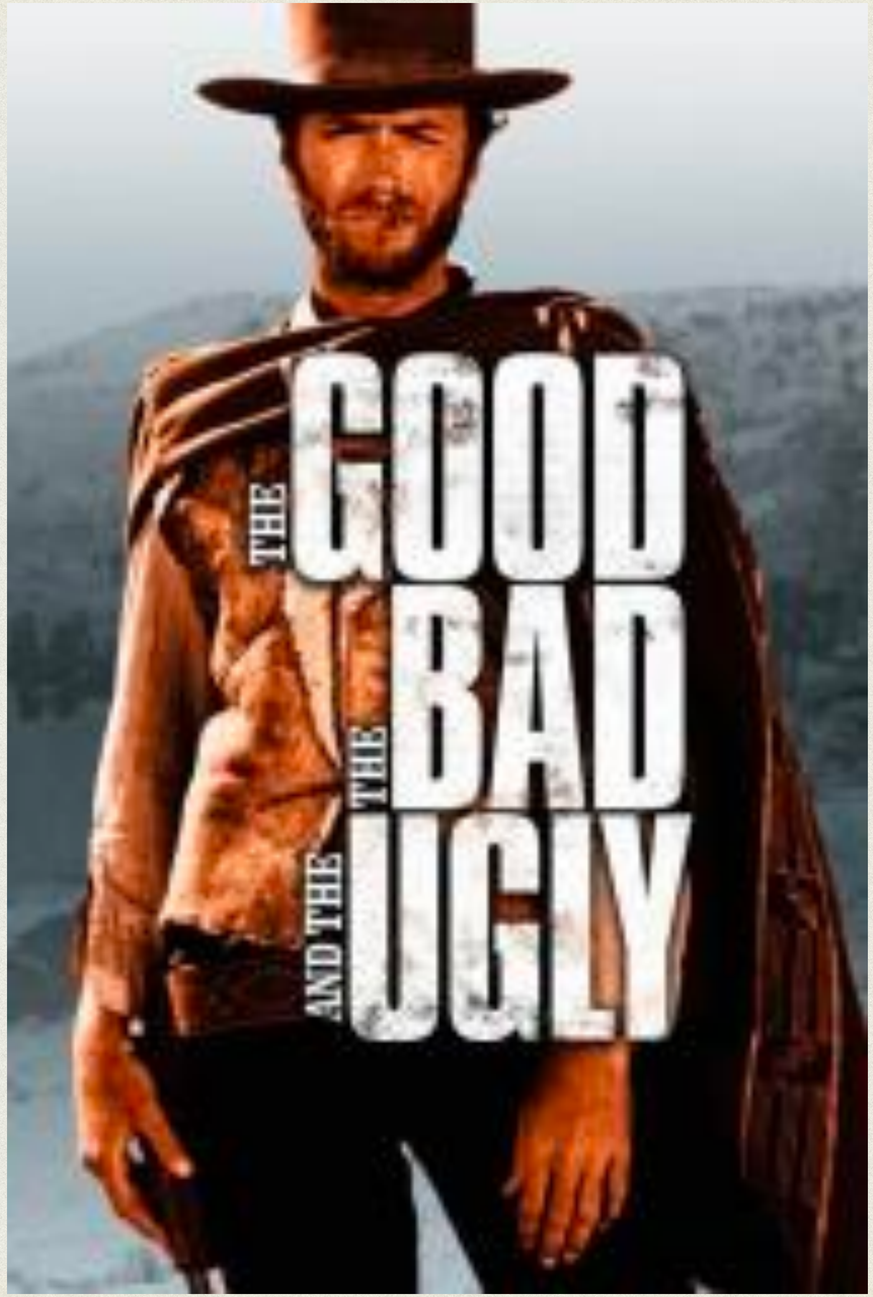
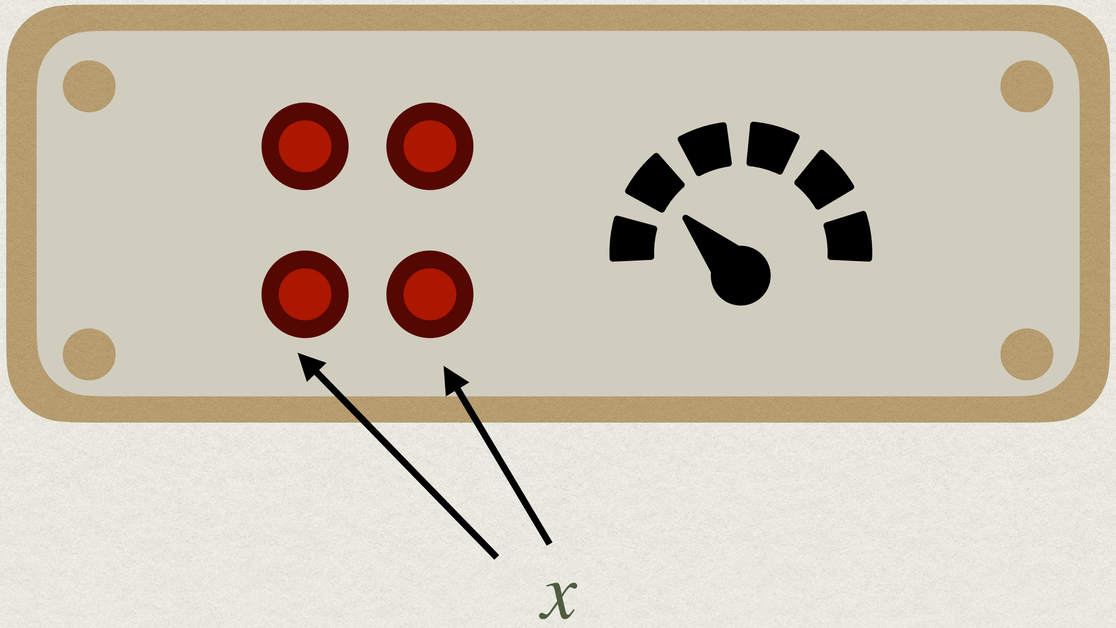
GENERALISATION IN QUANTUM CLASSIFICATION: A QUANTUM-INFO PERSPECTIVE

Leonardo Banchi — University of Florence, Italy

L. Banchi, J. Pereira, S. Pirandola, PRX Quantum 2, 040321 (2021)

«In this world there's two kinds of people, my friend: those with loaded guns and those who dig. You dig.»



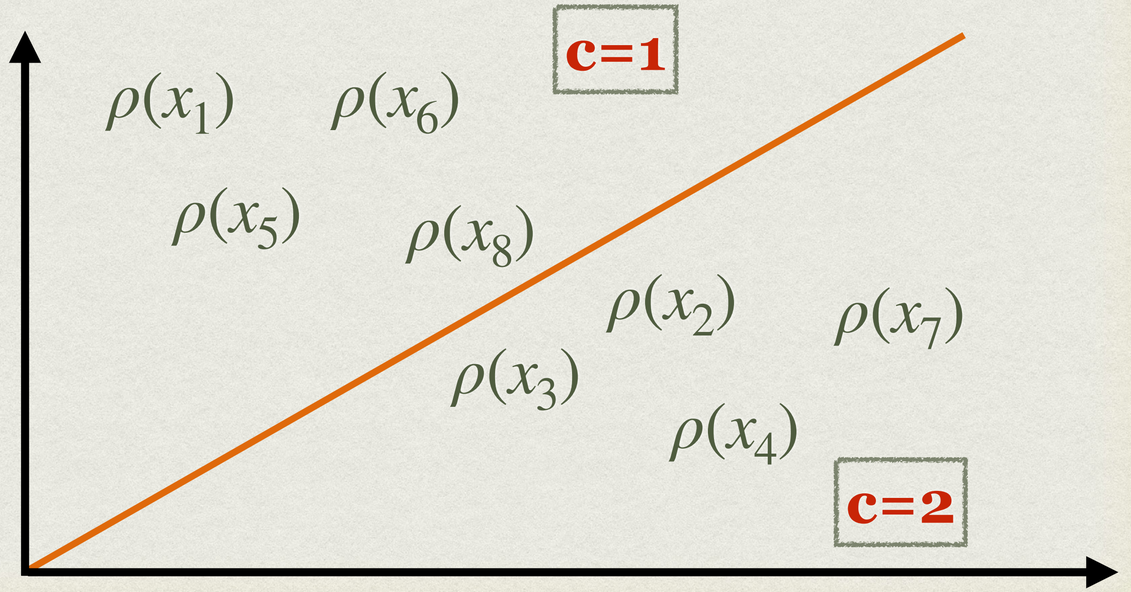


$\rho(x)$

$c = 1 \quad \text{or} \quad c = 2$?

«In this world there's two kinds of quantum states, my friend»

- Entangled / separable
- Ground states of critical/not critical models



THEORY QUESTION: HOW DIFFICULT IS TO OBTAIN A RELIABLE QUANTUM CLASSIFIER FROM DATA?

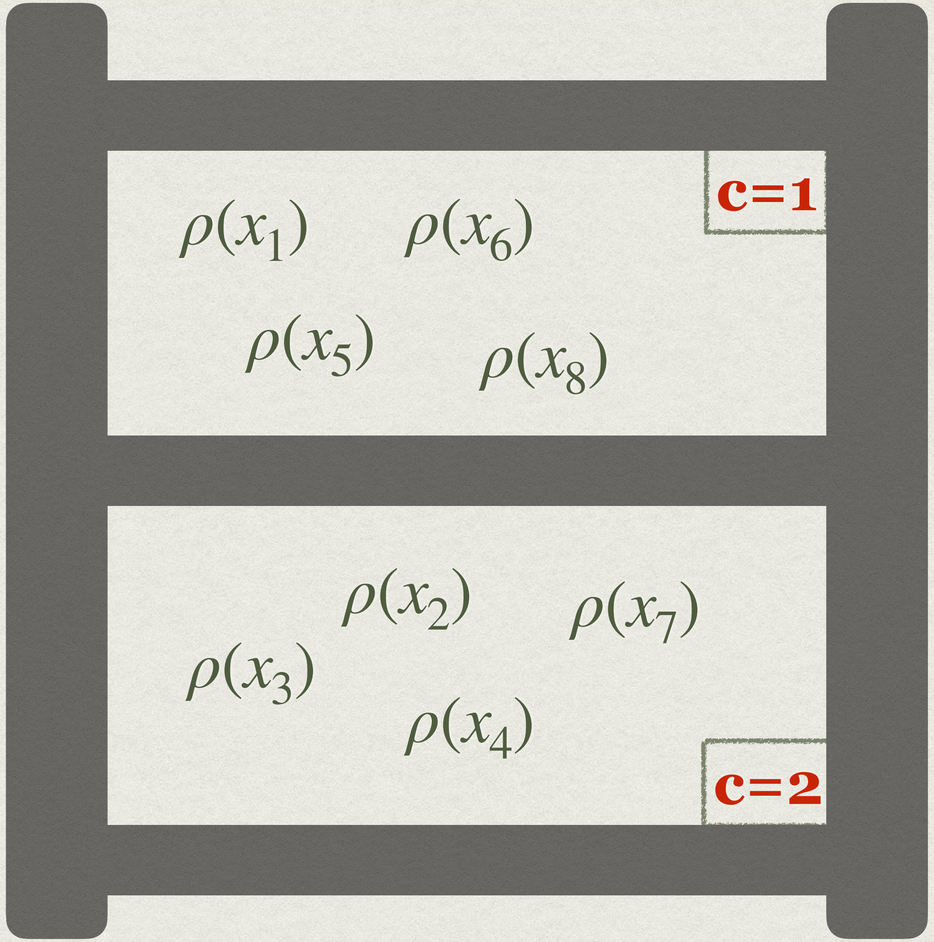
MAIN RESULTS

- In **data driven classification** the big question is generalisation: how does the model perform with **new** samples, not present in the training set?

- **Main theoretical result** (loose):
  generalisation error / out of sample classification error

$$\approx \sqrt{\frac{2^{I_2(X:Q)}}{T}}$$

"Complexity" quantified by $I_2(X{:}Q)$, i.e. number of training samples T to get low error with new data



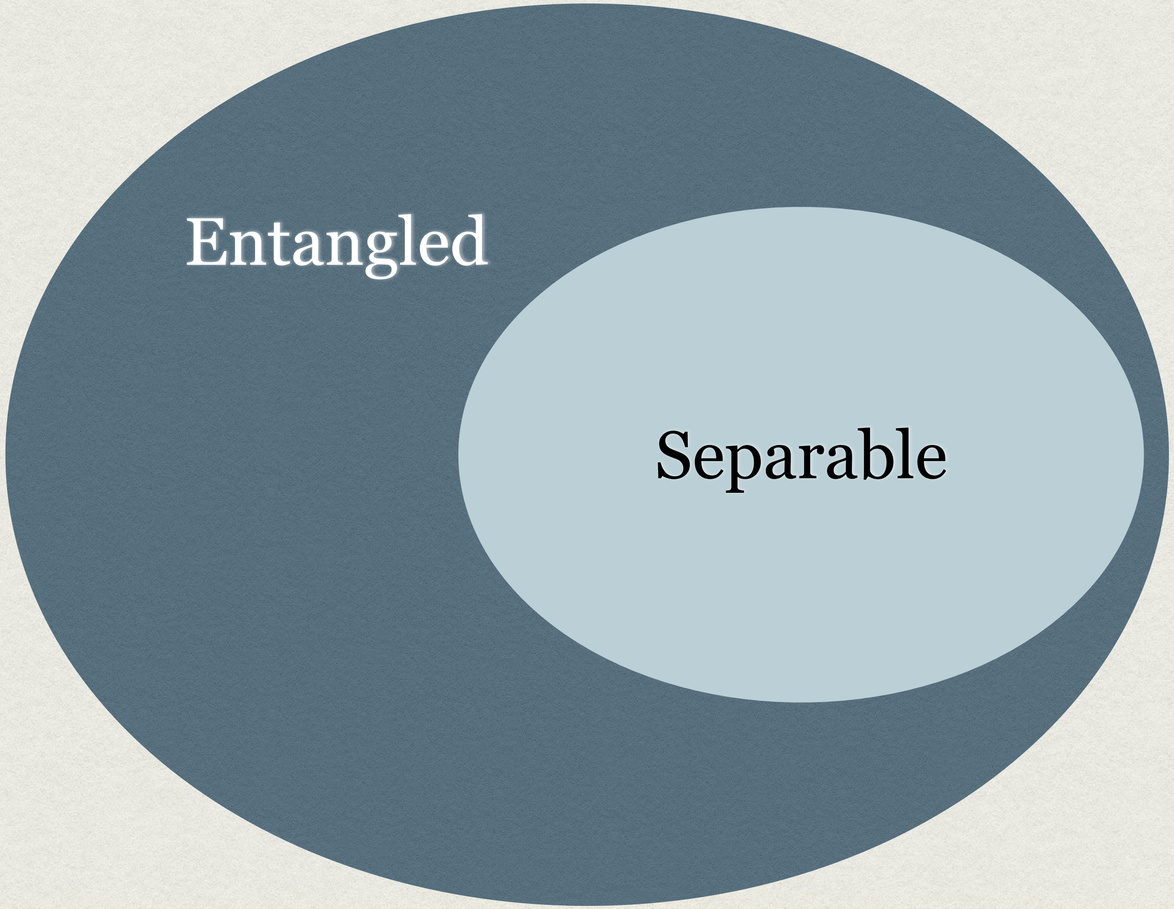
DIFFERENCES WITH QUANTUM STATE DISCRIMINATION

- In quantum state discrimination the states $\rho(x)$ appear with **known** probability distribution $P(x)$ — or $P(x, c)$ for classes c
    - (Sub)-optimal discrimination strategies are known: Helstrom measurements / pretty-good (square root) measurements
- In machine learning $P(x, c)$ is unknown and must be **learnt** from data
    - In state discrimination we don't have to worry about generalisation. Here we do

EXAMPLE APPLICATIONS

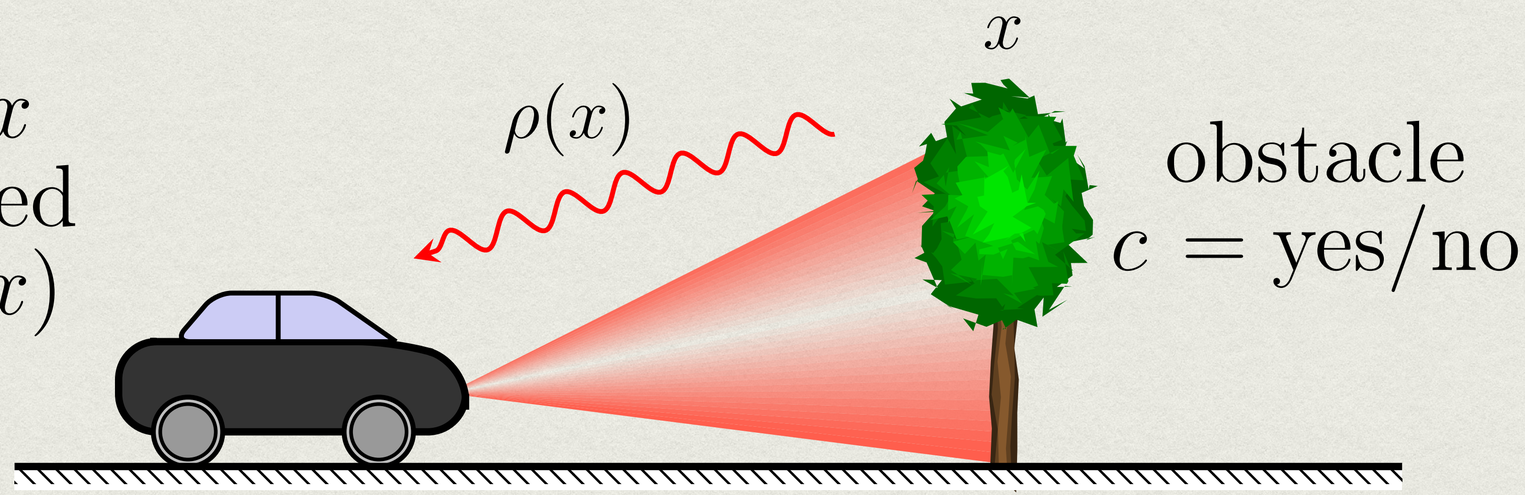
ENTANGLEMENT DETECTION

- Depending on some experimental parameters x a device either outputs entangled or separable states $\rho(x)$
- The training set is made of very well characterised samples
- What about new samples?



QUANTUM CHANNEL DISCRIMINATION

detect objects x
from the scattered
state of light $\rho(x)$

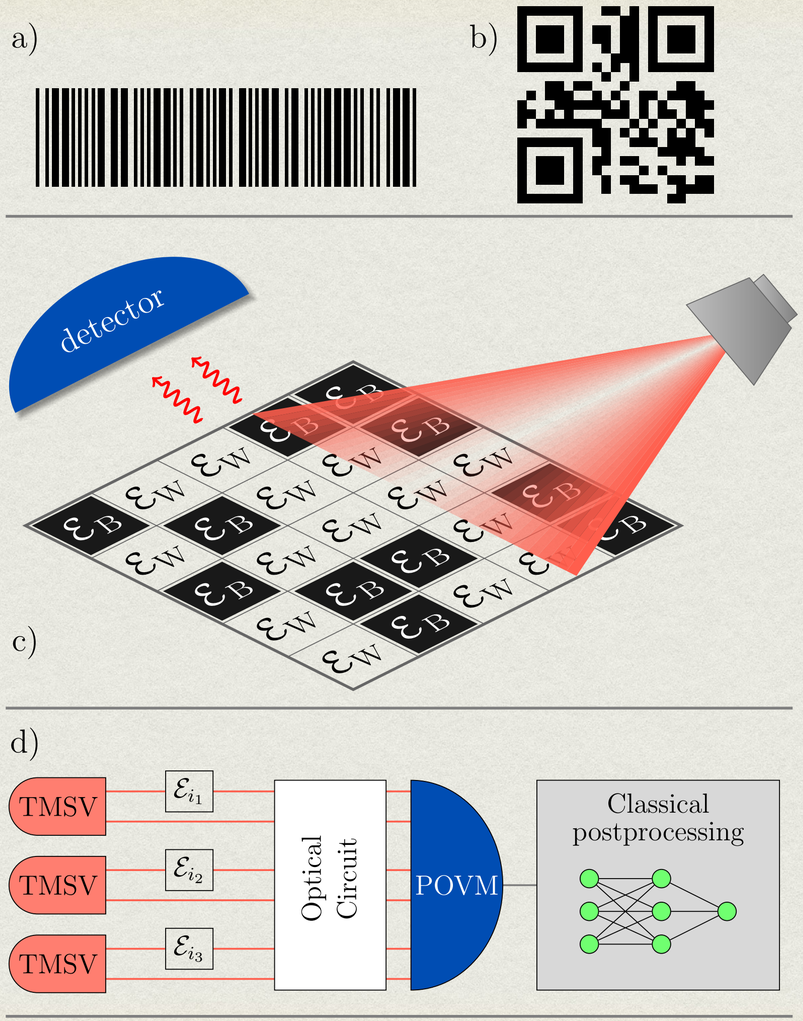


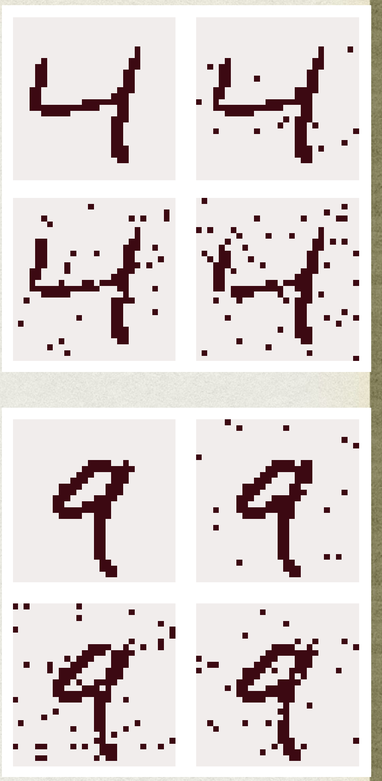
- Images x live in the physical world
- Optimise over the (entangled) input probe state of light and over the detection POVM

QUANTUM BARCODES AND PATTERN RECOGNITION

- Barcode classification must identify each pixel correctly: multiary quantum reading / illumination
- Handwriting classification is easier as errors are tolerated!

$$\text{error} \simeq F(\rho_{\text{black}}, \rho_{\text{white}})^{\text{Hamming}_{4\leftrightarrow 9}}$$





- L. Banchi, Q. Zhuang, S. Pirandola, Phys. Rev. Applied 14, 064026 (2020)
- C Harney, L Banchi, S Pirandola, Phys. Rev. A 103, 052406 (2021)
- JL Pereira, L Banchi, Q Zhuang, S Pirandola, Phys. Rev. A 103, 042614 (2021)

QUANTUM EMBEDDINGS

- Classify classical data (e.g. images)
- Embed images x onto a quantum state $x \mapsto \rho(x)$
- Decide the class from a quantum measurement $\{\Pi_c\}$

- M Schuld, N Killoran, Phys. rev. lett. 122 (4), 040504, (2019)
- V Havlicek, et al, Nature 567 (7747), 209, (2019)
- S Lloyd, et al, arXiv:2001.03622

(a) Data distribution

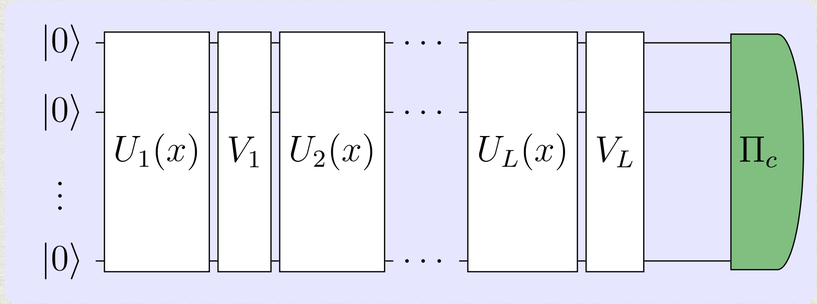
$P(c, x)$

samples $\{c, x\}$

$x =$ 

class/label
$c =$ "cat"

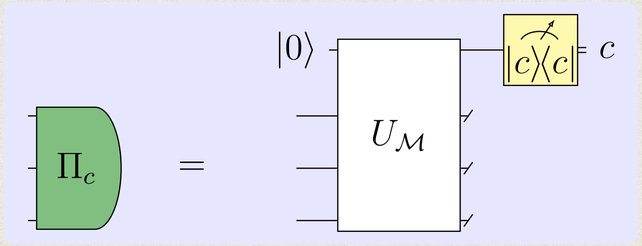
(b) Quantum classifier



predicted class c of x

embedding circuit $x \mapsto \rho(x)$ decision via POVM Π_c

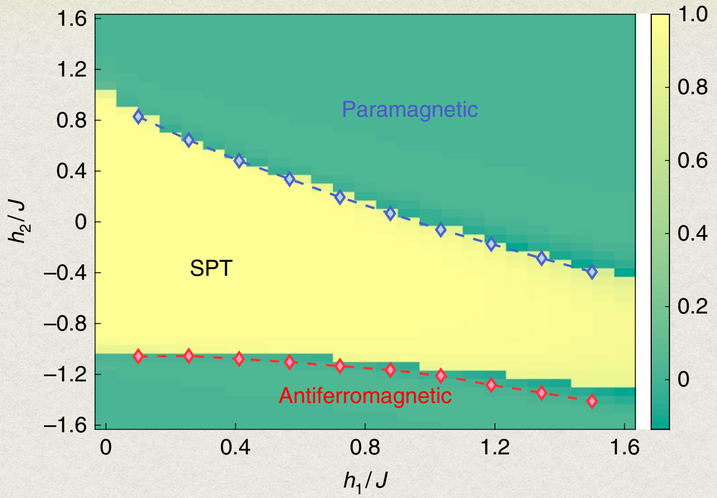
(c) Dilated measurements



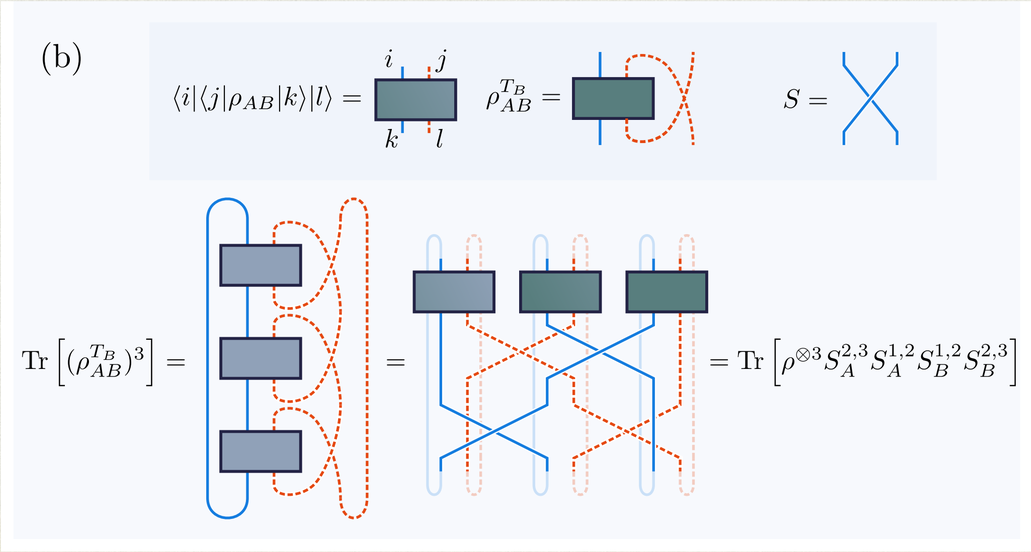
optimization of POVM Π_c
=
optimization of unitary $U_{\mathcal{M}}$

MANY-BODY PHYSICS

- Quantum Phase Recognition
  - I Cong, S Choi, MD Lukin,
    Nature Physics 15, 1273 (2019)



- Many-Body Entanglement Measurement from PPT-moments $\mathrm{Tr}\left[(\rho_{AB}^{T_B})^n\right]$
  - J Gray, L Banchi, A Bayat, S Bose,
    Phys. Rev. Lett. 121, 150503, 2018



RESULTS

PROBLEM DEFINITION

- We have a training set $\mathcal{T}$ made of T correctly classified states: $\mathcal{T} = \{(\rho(x_t), c_t) \text{ for } t = 1, \ldots, T\}$
- We can empirically check generalisation using a testing set $\mathcal{T}'$ with T' correctly classified states
- We consider a fixed **quantum embedding** $x \mapsto \rho(x)$ and optimal discrimination via POVM $\{\Pi_c\}$
- **What training error / testing error can we expect?**

$\rho(x_1)$	$\rho(x_6)$	**class 1**
$\rho(x_5)$	$\rho(x_8)$	
$\rho(x_2)$	$\rho(x_7)$	**class 2**
$\rho(x_3)$	$\rho(x_4)$	

- Empirical loss / training error

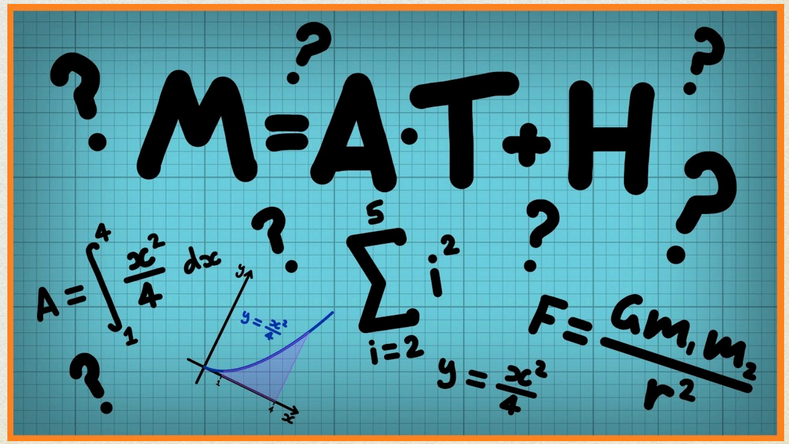
$$R_{\mathcal{T}}(\Pi,\rho)=\frac{1}{T}\sum_{(c_k,\rho(x_k))\in\mathcal{T}}\sum_{c\neq c_k}\mathrm{Tr}\left[\Pi_c\rho(x_k)\right]=1-\frac{1}{T}\sum_{(c_k,\rho(x_k))\in\mathcal{T}}\mathrm{Tr}\left[\Pi_{c_k}\rho(x_k)\right]$$

- Abstract classification error

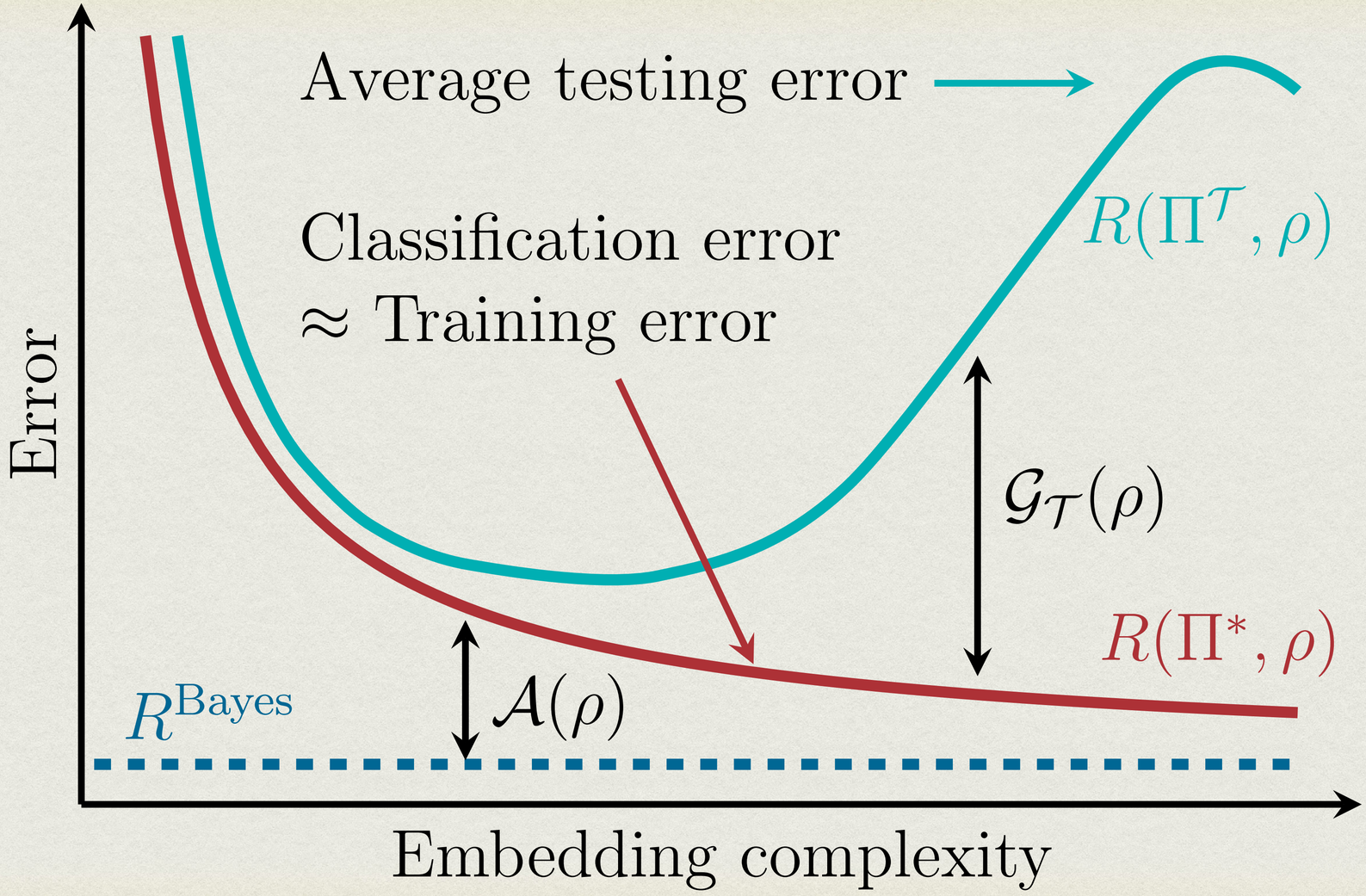
$$R(\Pi,\rho)=\mathbb{E}_{(c,x)\sim P(c,x)}\left(\sum_{c\neq\tilde{c}}\mathrm{Tr}\left[\Pi_{\tilde{c}}\rho(x)\right]\right)=1-\mathbb{E}_{(c,x)\sim P(c,x)}\mathrm{Tr}\left[\Pi_c\rho(x)\right]$$

- Optimal measurement for **empirical risk minimisation** $\Pi^{\mathcal{T}}=\mathrm{argmin}_{\Pi}R_{\mathcal{T}}(\Pi,\rho)$
- Real optimal $\Pi^*=\mathrm{argmin}_{\Pi}R(\Pi,\rho)$
- Testing error

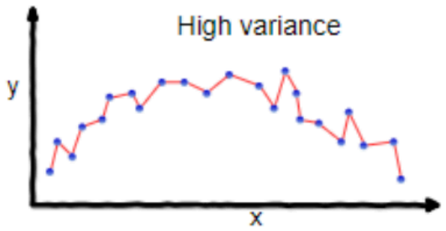
$$R_{\mathcal{T}'}(\Pi_{\mathcal{T}},\rho)$$



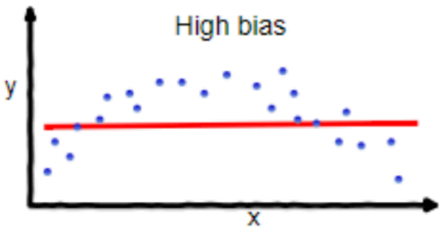
"COMPLEXITY" OF QUANTUM EMBEDDINGS



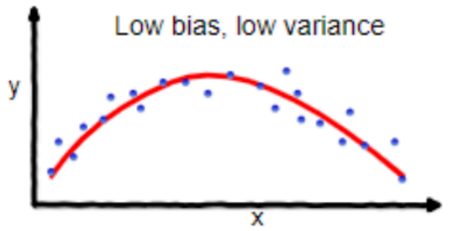
BIAS VARIANCE TRADEOFF VS. DEEP LEARNING



overfitting



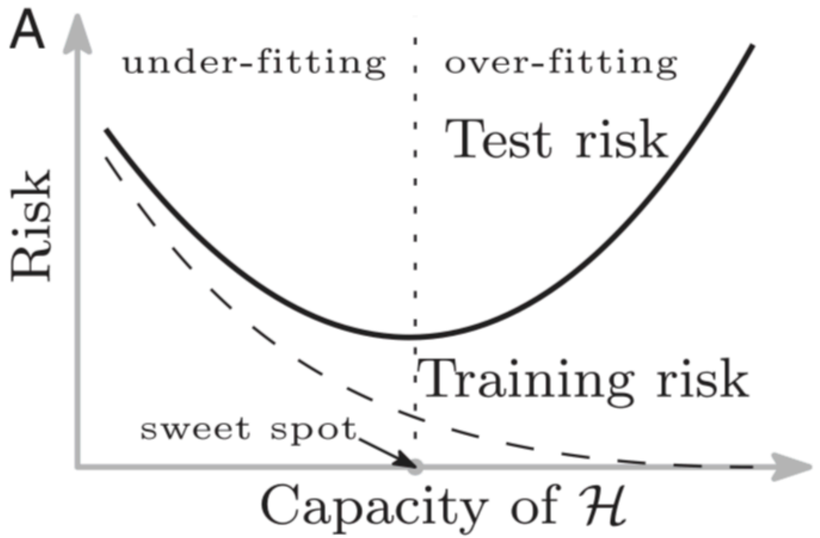
underfitting

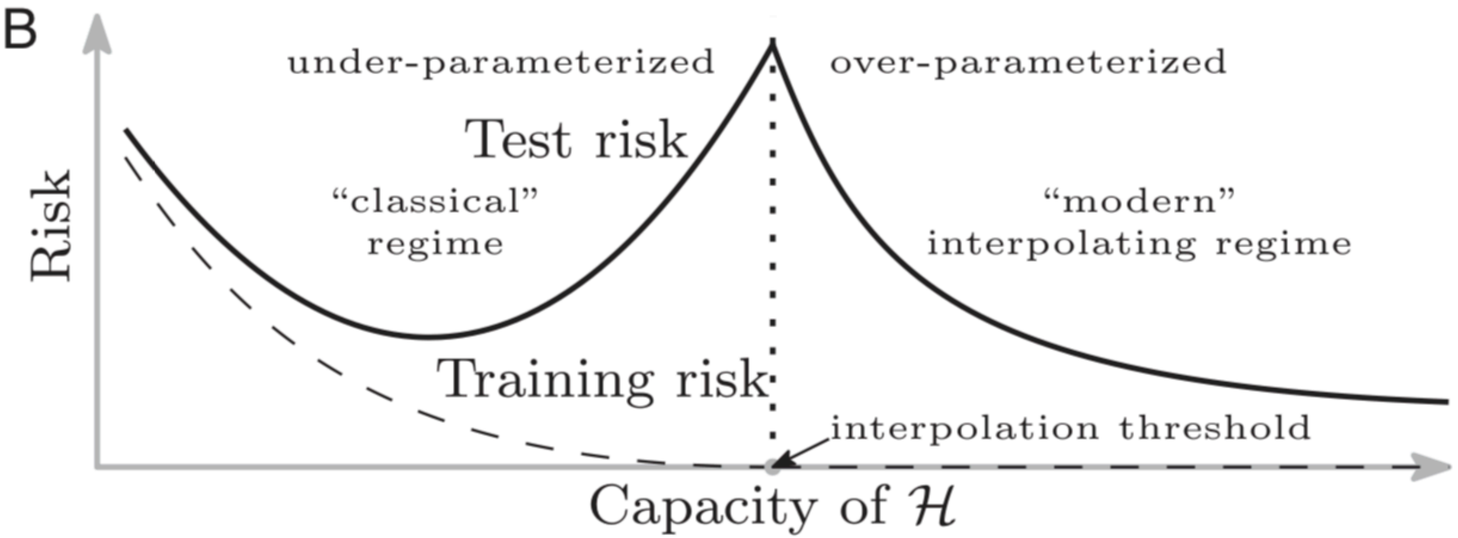


Good balance

PNAS, 2019, 116
(32) 15849-15854

https://medium.com/analytics-vidhya/bias-variance-trade-off-in-machine-learning-ae6fc6e326ba





MAIN RESULT

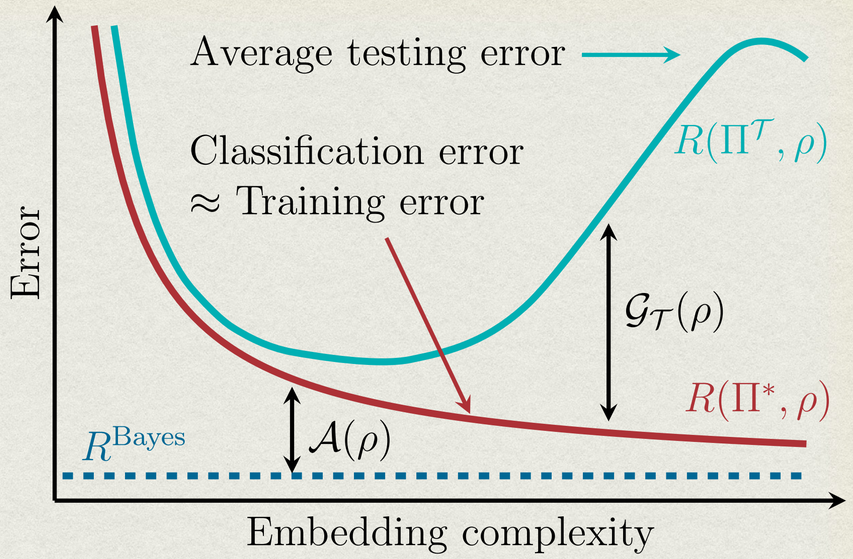
- Bound for the generalisation error

$$G_{\mathcal{T}} \leq 2\sqrt{\frac{\mathscr{B}}{T}} + \sqrt{\frac{2\log(1/\delta)}{T}}$$

- Bound for the approximation error (binary hyp.)

$$\mathscr{A} = \sum_x \frac{\left|P(x|0) - P(x|1)\right|}{2} - \frac{\|\rho_0 - \rho_1\|_1}{2}$$

where $\rho_c = \sum_x P(x|c)\rho(x)$



$$\mathscr{B} = 2^{I_2(X:Q)}$$

$$\mathscr{A} \leq K - \frac{2^{I(C:Q)}}{N_C}$$

Generalisation bound by combining **Rademacher complexity** from statistical learning theory, operator inequalities and the following (new) lemma

Lemma 1. *Let A_i be a set of operators and i a random variable with probability distribution p_i. Then*

$$\mathop{\mathbb{E}}_{i\sim p}\left(\|A_i\|_1\right) \leq \mathrm{Tr}\sqrt{\mathop{\mathbb{E}}_{i\sim p}\left(A_i A_i^\dagger\right)}\,, \qquad \text{(A12)}$$

where $\mathbb{E}_{i\sim p}\, f(i) := \sum_i p_i f(i)$.

This is an operator version of the generalised mean inequality

$$\sqrt[a]{\sum_{i=1}^{n} p_i x_i^a} \leq \sqrt[b]{\sum_{i=1}^{n} p_i x_i^b} \qquad \text{for} \qquad a < b$$

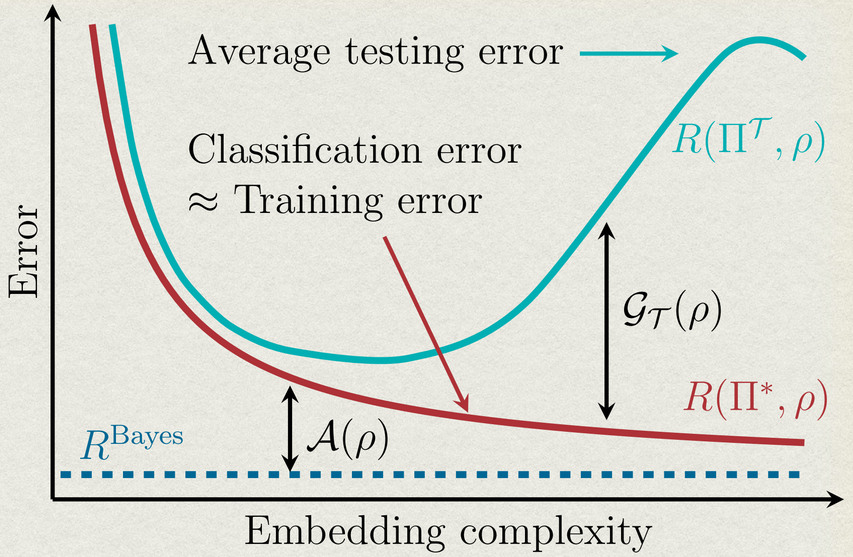
Approximation bound by combining known results from QIP: Helstrom measurements (binary case), Min-Entropy, Entropy inequalities

Konig, Renner, Schaffner, IEEE Trans. Inf. Th. 55, 4337 (2009).

Berta, Seshadreesan, Wilde, J. Math. Phys. 56, 022205 (2015).

Good embeddings should maximise $I(C{:}Q)$ and minimise $I_2(X{:}Q)$

Spoiler: Information Bottleneck!



*Two **extreme cases**:*

- Basis encoding: $x \mapsto \rho(x) = |x\rangle\langle x|$
  minimum $\mathscr{A}(\rho) = 0$, maximum $\mathscr{G}_{\mathscr{T}}(\rho)$

- Constant embedding : $x \mapsto \rho$
  maximum $\mathscr{A}(\rho)$, minimum $\mathscr{G}_{\mathscr{T}}(\rho) = 0$

$$\mathscr{B} = 2^{I_2(X:Q)}$$

$$\mathscr{A} \leq K - \frac{2^{I(C:Q)}}{N_C}$$

RISK/LOSS/ERRORS

- "Single-shot" linear loss, $(c_k, x_k) \in \mathcal{T}$

$$\sum_{c \neq c_k} \mathrm{Tr}\left[\Pi_c \rho(x_k)\right]$$

- Many shots

$$\sum_{c \neq c_k} \mathrm{Tr}\left[\Pi_c \rho(x_k)^{\otimes N}\right]$$

in parallel or with different experiments

For large N we find

$$\mathscr{A}(\rho) \leq K F_{\max}^N$$

where $F_{\max} = \max_{x \neq y} F(\rho(x), \rho(y))$

Conjecture:

$$\mathscr{B} \simeq \begin{cases} \mathcal{O}(\mathrm{poly}(N)) & \\ \mathcal{O}(f^N) & f > 1 \end{cases}$$

depending on the data distribution.

Related result with classical information

$$I(X{:}C_1, \ldots, C_N) = \begin{cases} \mathcal{O}(\log N) & \\ \mathcal{O}(N) & \end{cases}$$

Haussler, Opper, Ann. Stat. 1997

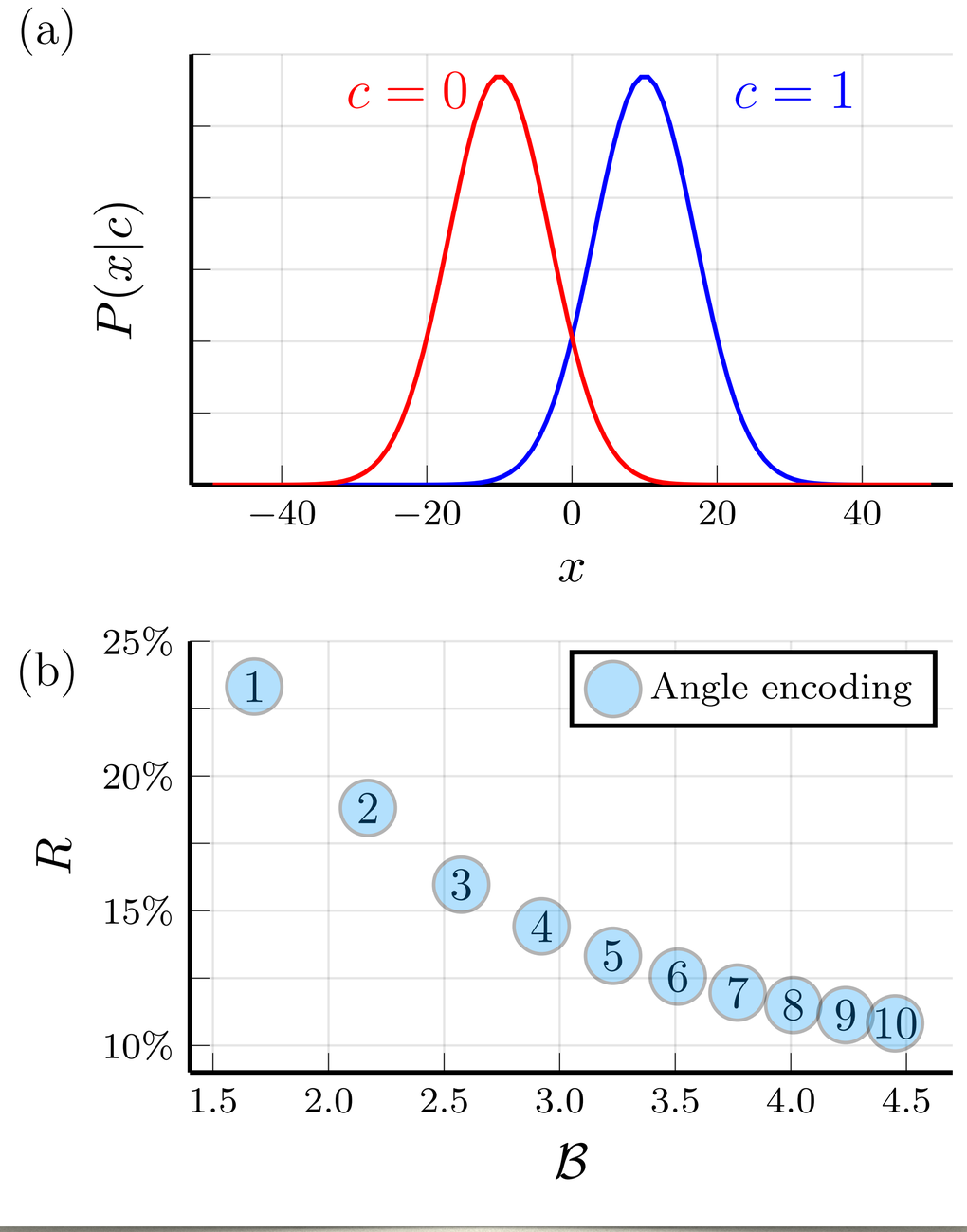
NUMERICAL EXPERIMENTS

Bigger Hilbert spaces have lower approximation error, but larger generalisation error

In the numerical example we consider

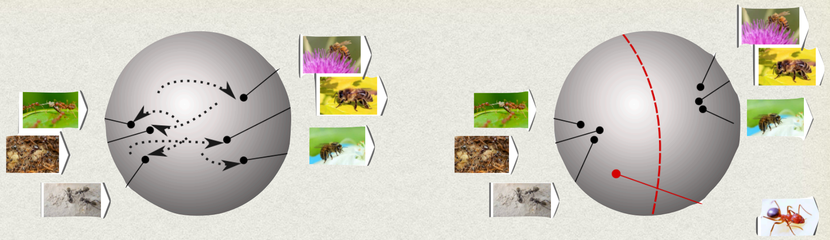
$$x \mapsto |\psi(x)\rangle\langle\psi(x)|^{\otimes N}$$

$$|\psi(x)\rangle = \cos(x)|0\rangle + \sin(x)|1\rangle$$



We proved that low dimensional embeddings / Low entropy datasets generalise well!

If $\rho(x)$ "fully scrambles" x in a d-dimensional subspace of the full Hilbert space, then $\mathcal{B} \approx \mathcal{O}(d)$



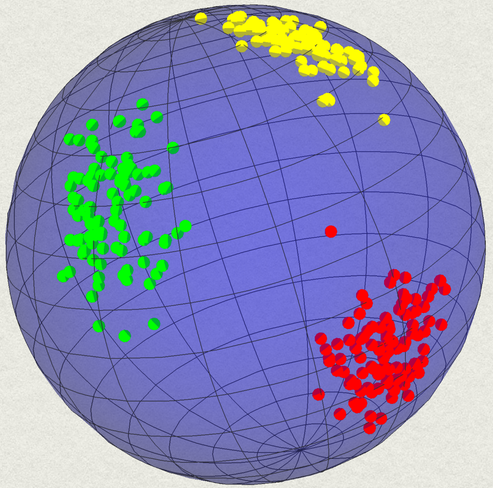
Conjecture from:
S Lloyd, et al, arXiv:2001.03622

Geometric characterisation:

$$\mathcal{B}_c \leq 1 + \sqrt{(r_c^2 - r_c)\left(1 - \mathrm{Tr}\left[\rho_c^2\right]\right)}$$

$$\mathrm{Tr}[\rho_c^2] = \sum_{x,y} P(x\,|\,c)P(y\,|\,c)F(\rho(x),\rho(y))^2$$

$$R \leq F(\rho_0,\rho_1)/2$$



QUANTUM KERNELS

For pure state embeddings $\rho(x) = |\psi(x)\rangle\langle\psi(x)|$ we find

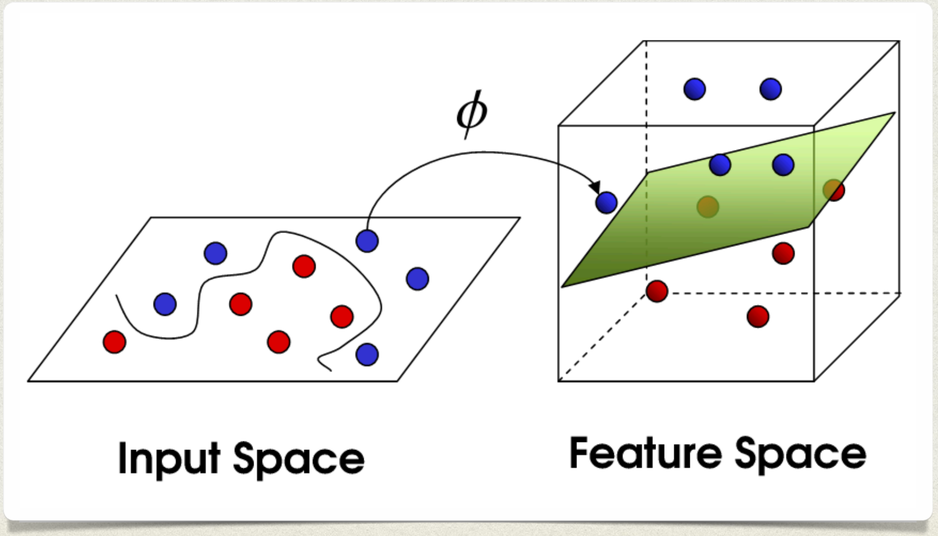
$$\mathscr{B} = \left[\mathrm{Tr}\sqrt{K}\right]^2$$

where $K_{xy} = \sqrt{p(x)p(y)}\,|\langle\psi(x)|\psi(y)\rangle|$ is a (normalised) **kernel** matrix. This makes the calculation $\mathscr{B}$ easier for large-dimensional embeddings.

Quantum kernels are used in

- Quantum support vector machines
- Quantum enhanced-feature space

Take home message: avoid $K \propto$ identity (bad generalisation)



*To favour generalisation the **final** Hilbert space must be small, but the initial one can be big!*

*We may iteratively discard information via **pooling** layers (e.g. QCNN)*

Pooling favours generalisation but harms the accuracy (via data processing)

***Take home message**: if low training error is achievable with pooling layers, then generalisation can only be better!*

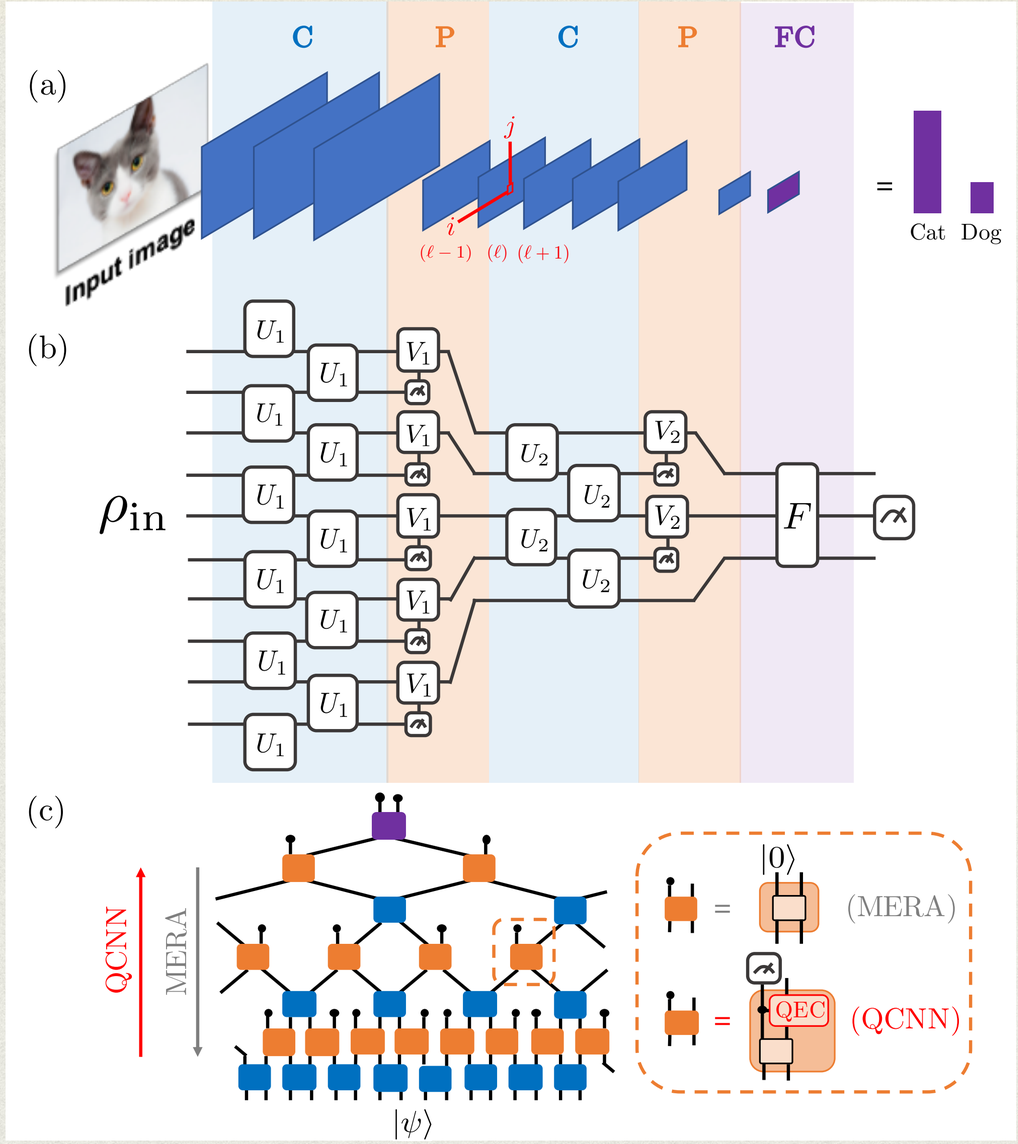


image from:
I Cong, et al. Nature Physics (2019)

INFORMATION BOTTLENECK FOR QUANTUM CLASSIFIERS

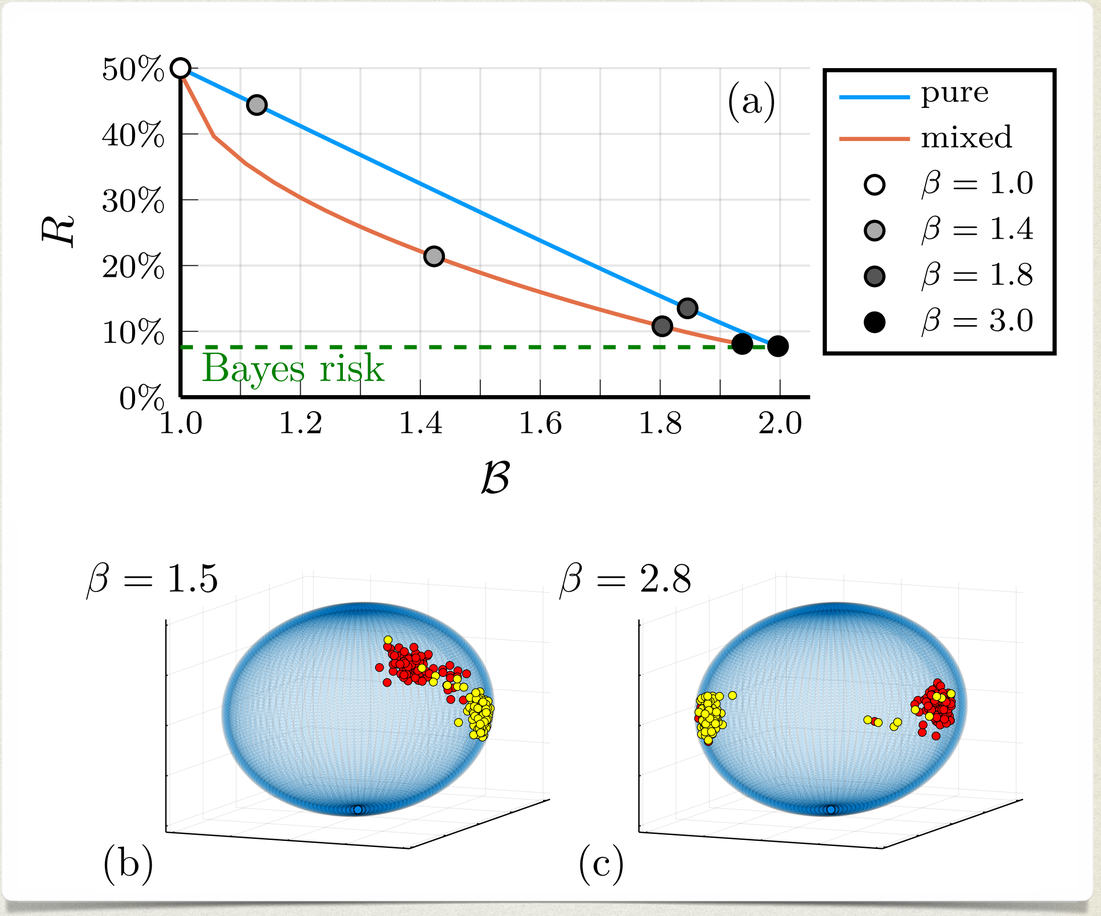
$\rho(x)$ as "bottleneck" that squeezes the relevant information that x provides about c

IB principle (loss independent): minimise

$$\mathscr{L}_{IB} = I(X{:}Q) - \beta I(C{:}Q)$$

Self-consistent solutions (similar for ρ)

$$\tilde{\lambda}_z |\psi(z)\rangle = e^{(1-\beta)\log\bar{\rho}+\beta\sum_c P(c|z)\log\rho_c} |\psi(z)\rangle$$



See also: Salek et al. IEEE Trans. Inf. Th. (2018), Datta et al., IEEE-ISIT, (2019)

APPLICATIONS

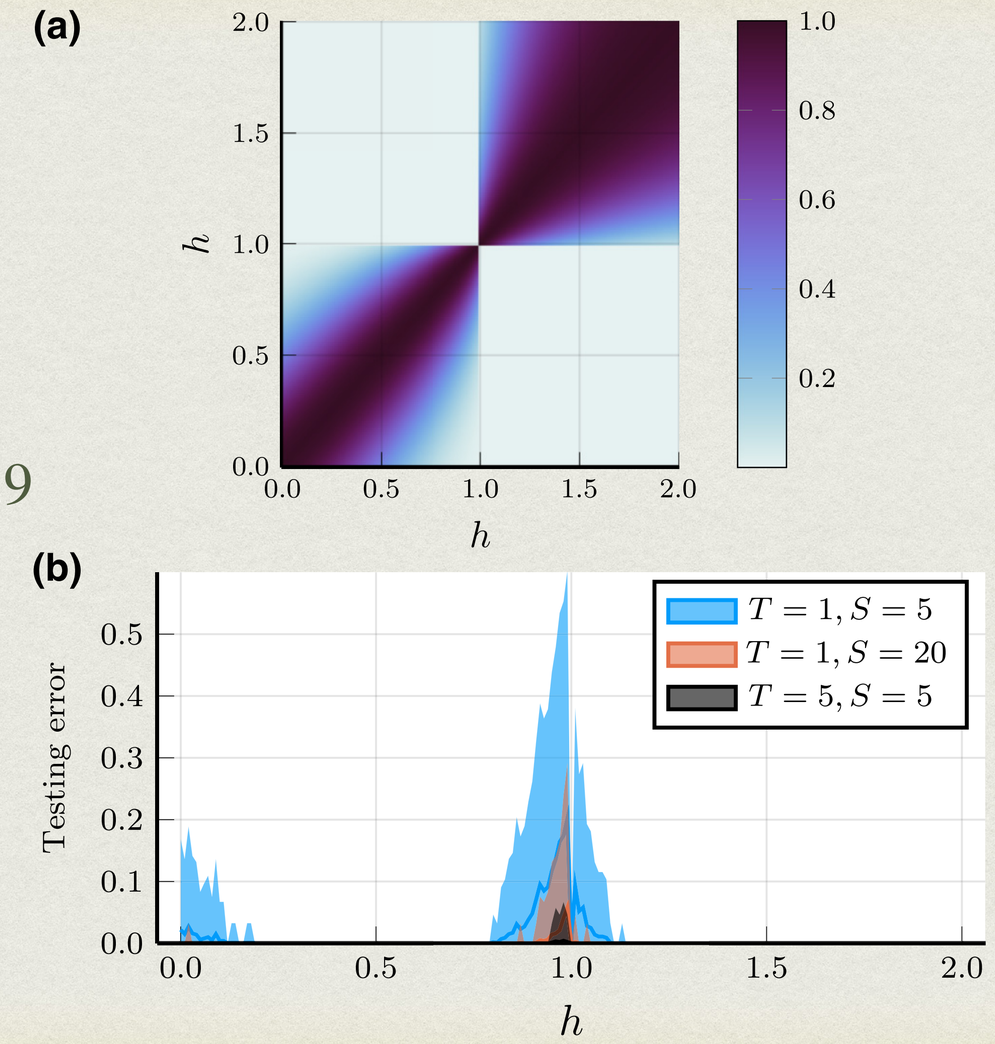
QUANTUM PHASE RECOGNITION

Task: recognize the phases of matter of a quantum many-body system by taking measurements on the quantum system itself

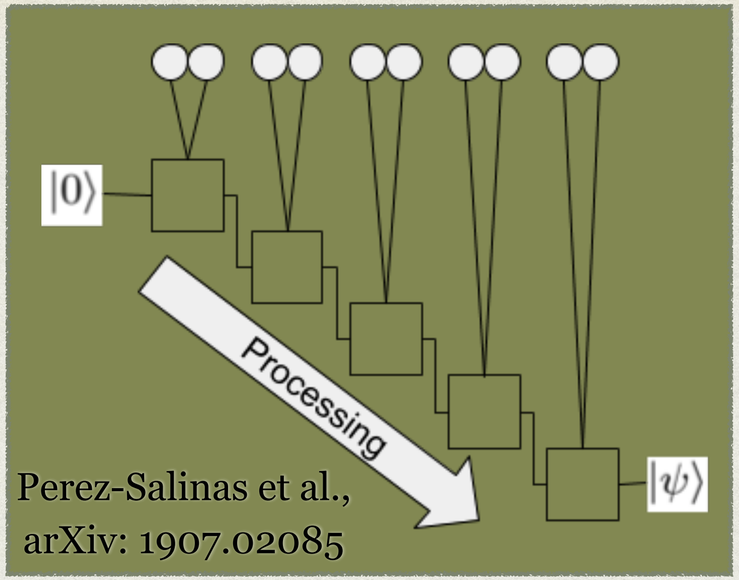
$$H = -\sum_{i=1}^{L} (\sigma_i^x \sigma_{i+1}^x + h\sigma_i^z), \qquad \Longrightarrow \qquad \mathscr{B} \simeq 5.9$$

Ordered ($|h| < 1$) / disordered ($|h| > 1$) phases

T : number of training samples per class
S : number of measurement shots



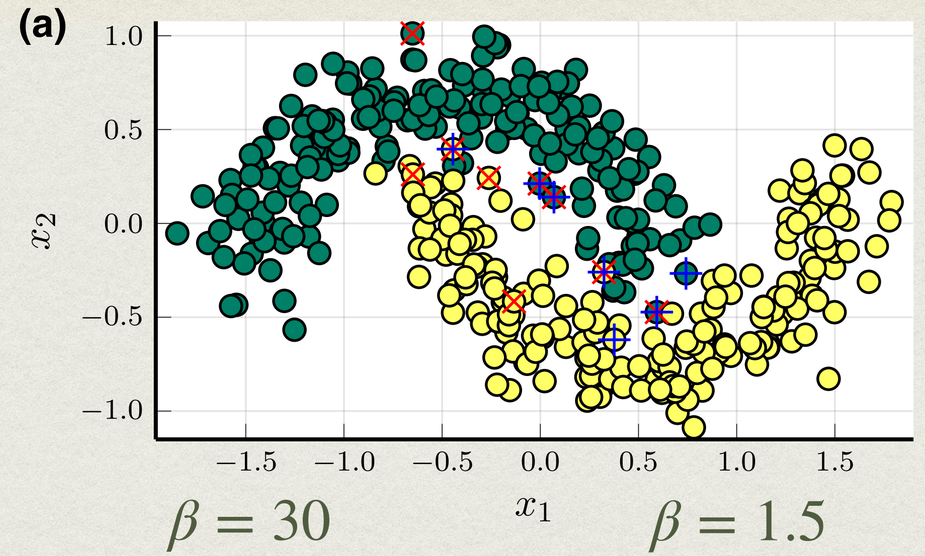
VARIATIONAL QUANTUM INFORMATION BOTTLENECK

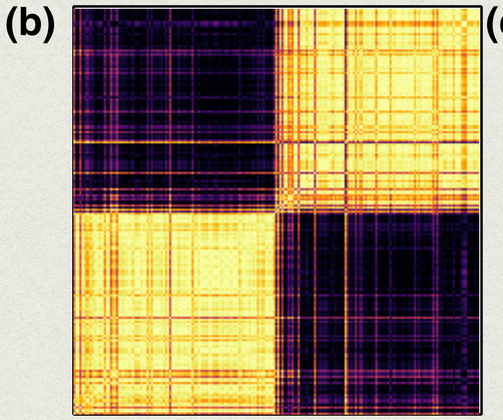


Single-qubit
"Data Re-uploading"
Classifier

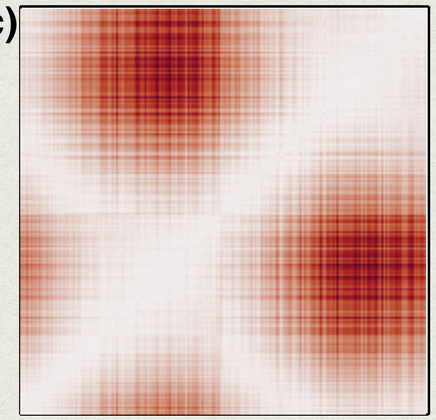
$$|\psi_w(x)\rangle = \prod_{\ell=1}^{L} [R^z(w^{z\ell}\cdot x + w_0^{z\ell}) R^y(w^{y\ell}\cdot x + w_0^{y\ell})]|0\rangle,$$

For two-qubit states the "re-upoading" embedding can be trained with an efficient variational minimisation of the IB Lagrangian



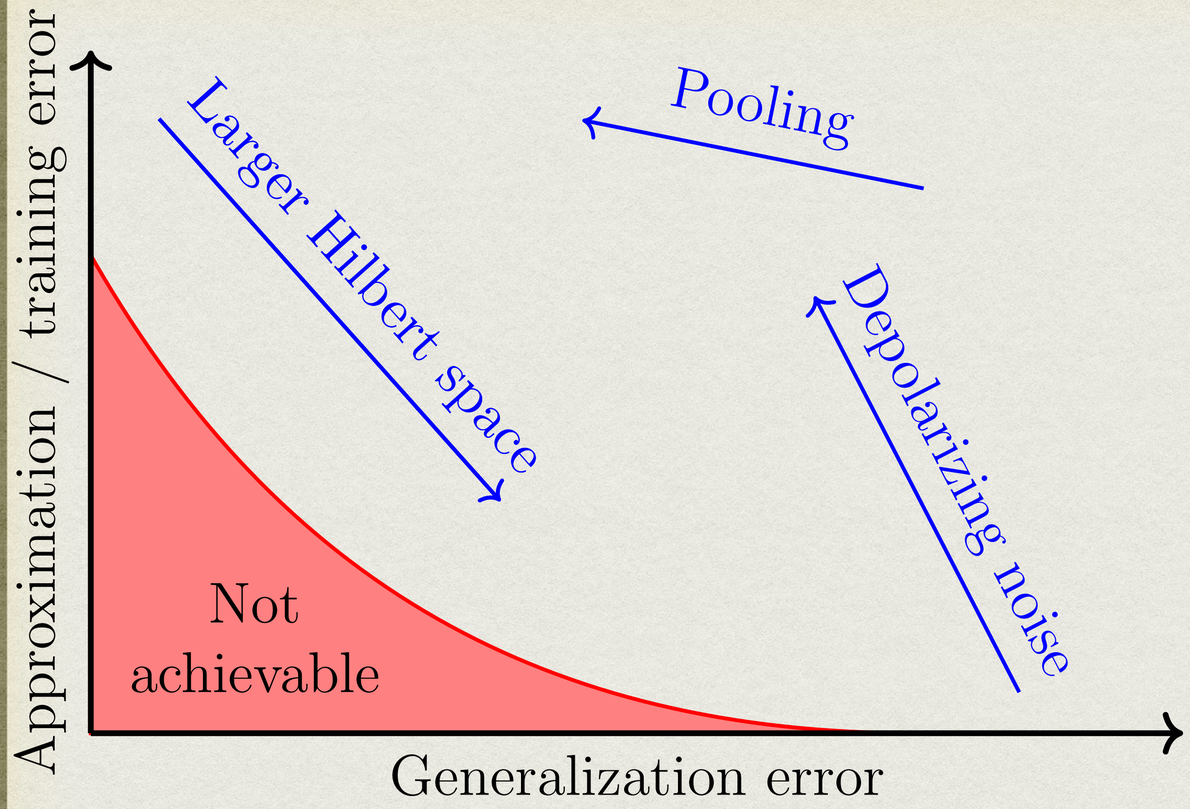


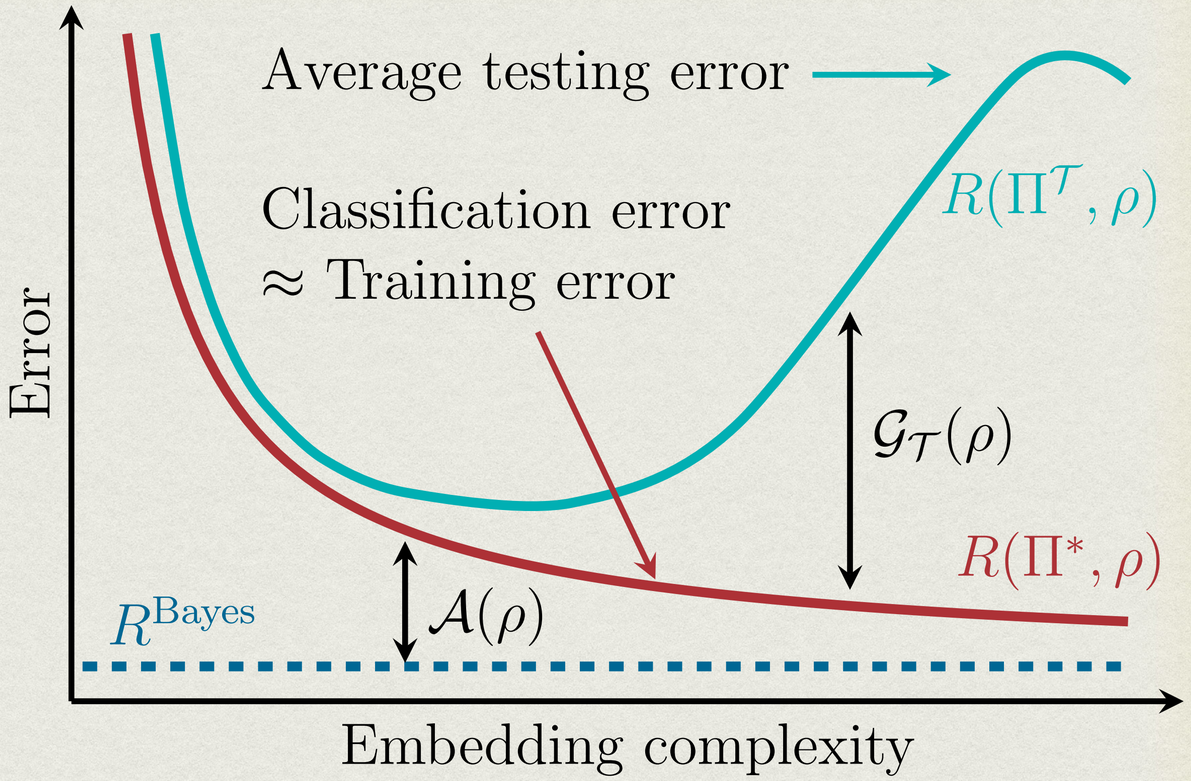
Testing error 4.5%



Testing error 3.5%

SUMMARY





Main open questions: deep learning regime / asymptotics with multiple shots