Classical learning guarantees and entanglement manipulation

Evan Peters

Quantum Techniques in Machine Learning 2023
Nov 19, 2023

Overview

1. A guarantee for classical learning
2. A quantum generalization of learning

Overview

1. A guarantee for classical learning
   - *The onset of learning:* When can we learn *something* about nature from some data?
2. A quantum generalization of learning

Overview

1. A guarantee for classical learning
   - *The onset of learning:* When can we learn *something* about nature from some data?
2. A quantum generalization of learning
   - Interact with quantum data while preserving coherence

Overview

1. A guarantee for classical learning
   - *The onset of learning:* When can we learn *something* about nature from some data?
2. A quantum generalization of learning
   - Interact with quantum data while preserving coherence
   - Learning as the classical limit of a quantum task

What is known already

1. Shannon theory for classical ML error *lower bounds*
   - Minimax error: discretization and Fano's inequality (Ibragimov+, 1981)
   - Error upper bounds by *existence proof*

What is known already

1. Shannon theory for classical ML error *lower bounds*
   - Minimax error: discretization and Fano's inequality (Ibragimov+, 1981)
   - Error upper bounds by *existence proof*
2. Quantum information protocols for *singlet fraction*
   - As a teleportation resource (Horodecki+, 1999)
   - Information-theoretic characterization (König+, 2009)

What is known already

1. Shannon theory for classical ML error *lower bounds*
   - Minimax error: discretization and Fano's inequality (Ibragimov+, 1981)
   - Error upper bounds by *existence proof*
2. Quantum information protocols for *singlet fraction*
   - As a teleportation resource (Horodecki+, 1999)
   - Information-theoretic characterization (König+, 2009)
3. Quantum information for quantum ML
   - Classification as state discrimination (Banchi+, 2021)
   - Hypothesis testing as quantum communication (Huang+, 2021)
   - Generalization error from quantum communication (Caro+, 2023)

What is known already

1. Shannon theory for classical ML error *lower bounds*
   - Minimax error: discretization and Fano's inequality (Ibragimov+, 1981)
   - Error upper bounds by *existence proof*
2. Quantum information protocols for *singlet fraction*
   - As a teleportation resource (Horodecki+, 1999)
   - Information-theoretic characterization (König+, 2009)
3. Quantum information for quantum ML
   - Classification as state discrimination (Banchi+, 2021)
   - Hypothesis testing as quantum communication (Huang+, 2021)
   - Generalization error from quantum communication (Caro+, 2023)
   - **Here:** A relationship between classical learning and entanglement

The learning problem

Setting:

- Sample an unknown vector $\alpha \in \mathrm{A} \subset \mathbb{R}^d$. A is compact.
- *Data* $B \in \mathcal{B}$ is generated based on α.

The learning problem

Setting:

- Sample an unknown vector $\alpha \in \mathrm{A} \subset \mathbb{R}^d$. A is compact.
- *Data* $B \in \mathcal{B}$ is generated based on α.

Goal: Output an *estimate* $\hat{\alpha} \in \mathrm{A}$ approximating α to ϵ-accuracy

- Maximize $\ell(d(\alpha, \hat{\alpha}))$ for distance d, score function ℓ

The learning problem

Setting:

- Sample an unknown vector $\alpha \in \mathrm{A} \subset \mathbb{R}^d$. A is compact.
- *Data* $B \in \mathcal{B}$ is generated based on α.

Goal: Output an *estimate* $\hat{\alpha} \in \mathrm{A}$ approximating α to ϵ-accuracy

- Maximize $\ell(d(\alpha, \hat{\alpha}))$ for distance d, score function ℓ

Example: Learning hyperplanes

The learning problem

Setting:

- Sample an unknown vector $\alpha \in \mathrm{A} \subset \mathbb{R}^d$. A is compact.
- *Data* $B \in \mathcal{B}$ is generated based on α.

Goal: Output an *estimate* $\hat{\alpha} \in \mathrm{A}$ approximating α to ϵ-accuracy

- Maximize $\ell(d(\alpha, \hat{\alpha}))$ for distance d, score function ℓ

Example: Learning hyperplanes

- Sample $x_i \in \mathbb{R}^d$ from a distribution, let $y_i = \langle \alpha, x_i \rangle + (\text{noise})$

The learning problem

Setting:

- Sample an unknown vector $\alpha \in \mathrm{A} \subset \mathbb{R}^d$. A is compact.
- *Data* $B \in \mathcal{B}$ is generated based on α.

Goal: Output an *estimate* $\hat{\alpha} \in \mathrm{A}$ approximating α to ϵ-accuracy

- Maximize $\ell(d(\alpha, \hat{\alpha}))$ for distance d, score function ℓ

Example: Learning hyperplanes

- Sample $x_i \in \mathbb{R}^d$ from a distribution, let $y_i = \langle \alpha, x_i \rangle + \text{(noise)}$
- $B = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

The learning problem

Setting:

- Sample an unknown vector $\alpha \in \mathrm{A} \subset \mathbb{R}^d$. A is compact.
- *Data* $B \in \mathcal{B}$ is generated based on α.

Goal: Output an *estimate* $\hat{\alpha} \in \mathrm{A}$ approximating α to ϵ-accuracy

- Maximize $\ell(d(\alpha, \hat{\alpha}))$ for distance d, score function ℓ

Example: Learning hyperplanes

- Sample $x_i \in \mathbb{R}^d$ from a distribution, let $y_i = \langle \alpha, x_i \rangle + \text{(noise)}$
- $B = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Output $\hat{\alpha}$ to minimize $\|\alpha - \hat{\alpha}\|^2$ ($\ell(x) = c - x^2$, $d(x, y) = \|x - y\|$)

The learning problem

Setting:

- Sample an unknown vector $\alpha \in \mathrm{A} \subset \mathbb{R}^d$. A is compact.
- *Data* $B \in \mathcal{B}$ is generated based on α.

Goal: Output an *estimate* $\hat{\alpha} \in \mathrm{A}$ approximating α to ϵ-accuracy

- Maximize $\ell(d(\alpha, \hat{\alpha}))$ for distance d, score function ℓ

Example: Learning hyperplanes

- Sample $x_i \in \mathbb{R}^d$ from a distribution, let $y_i = \langle \alpha, x_i \rangle + \text{(noise)}$
- $B = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Output $\hat{\alpha}$ to minimize $\|\alpha - \hat{\alpha}\|^2$ ($\ell(x) = c - x^2$, $d(x, y) = \|x - y\|$)
- This is at least as hard as linear regression.

The learning guarantee

For a random variable A distributed uniformly on an ϵ-covering of A, best-case score obeys

$$\max_{\alpha \in \mathrm{A}} \max_{\hat{\alpha}} \log \mathbb{E}_{p_{\hat{\alpha}|\alpha}}[\ell(d(\alpha, \hat{\alpha}))] \geq \log \ell(\epsilon) - H(A|B) \quad (1)$$

The learning guarantee

For a random variable A distributed uniformly on an ϵ-covering of A, best-case score obeys

$$\max_{\alpha \in \mathrm{A}} \max_{\hat{\alpha}} \log \mathbb{E}_{p_{\hat{\alpha}|\alpha}}[\ell(d(\alpha, \hat{\alpha}))] \geq \log \ell(\epsilon) - H(A|B) \quad (1)$$

The learning guarantee

For a random variable A distributed uniformly on an ϵ-covering of A, best-case score obeys

$$\max_{\alpha \in \mathrm{A}} \max_{\hat{\alpha}} \log \mathbb{E}_{p_{\hat{\alpha}|\alpha}} [\ell(d(\alpha, \hat{\alpha}))] \geq \log \ell(\epsilon) - H(A|B) \tag{1}$$

The learning guarantee

For a random variable A distributed uniformly on an ϵ-covering of A, best-case score obeys

$$\max_{\alpha \in \mathrm{A}} \max_{\hat{\alpha}} \log \mathbb{E}_{p_{\hat{\alpha}|\alpha}}[\ell(d(\alpha, \hat{\alpha}))] \geq \log \ell(\epsilon) - H(A|B) \tag{1}$$

The learning guarantee

For a random variable A distributed uniformly on an ϵ-covering of A, best-case score obeys

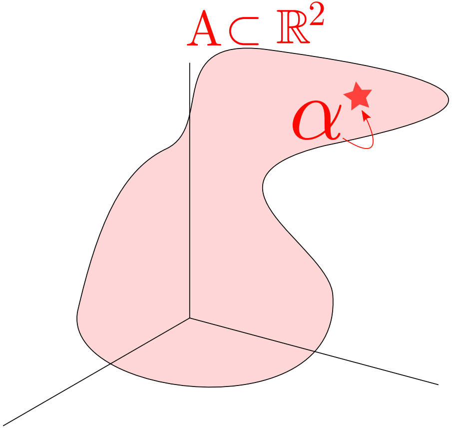
$$\max_{\alpha \in \mathrm{A}} \max_{\hat{\alpha}} \log \mathbb{E}_{p_{\hat{\alpha}|\alpha}} [\ell(d(\alpha, \hat{\alpha}))] \geq \log \ell(\epsilon) - H(A|B) \tag{1}$$

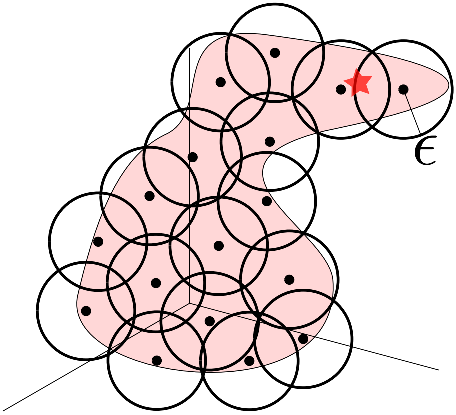
The learning guarantee

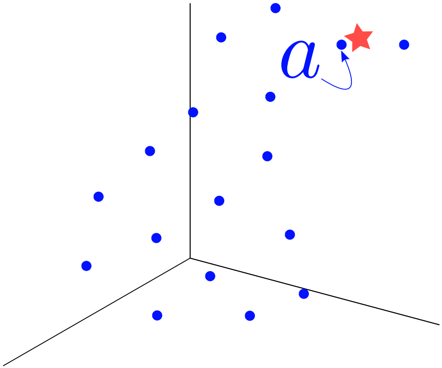
For a random variable A distributed uniformly on an ϵ-covering of A, best-case score obeys

$$\max_{\alpha \in \mathrm{A}} \max_{\hat{\alpha}} \log \mathbb{E}_{p_{\hat{\alpha}|\alpha}}[\ell(d(\alpha, \hat{\alpha}))] \geq \log \ell(\epsilon) - H(A|B) \quad (1)$$

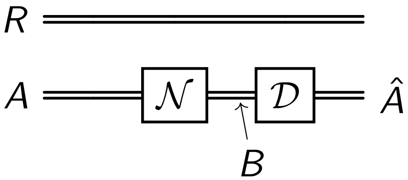
- **LHS**: best score w/r to unknown α and estimate $\hat{\alpha}$
- **RHS**: Score for ϵ-accurate $\hat{\alpha}$, minus uncertainty $H(A|B)$ of discretization given B


$A \subset \mathbb{R}^2$
$\alpha^\star$

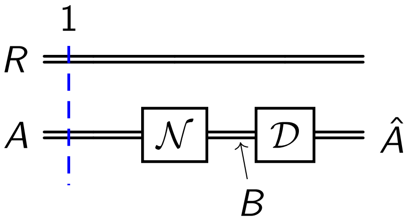

ϵ


a

Classical hypothesis testing

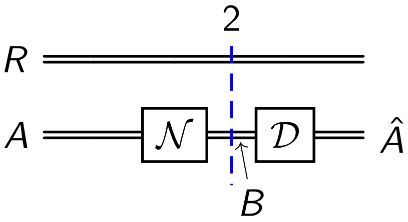


Classical hypothesis testing



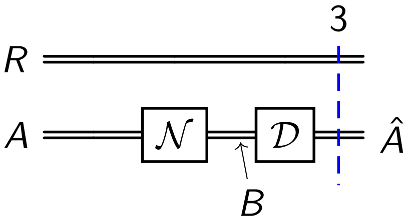
1. Random variables A, R taking values in $\mathcal{A}$ are sampled from p_{AR}

Classical hypothesis testing



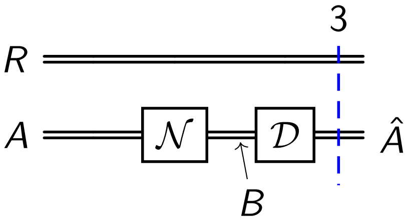
1. Random variables A, R taking values in $\mathcal{A}$ are sampled from p_{AR}
2. Nature uses $\mathcal{N}$ to prepare data B depending on the value of A

Classical hypothesis testing



1. Random variables A, R taking values in $\mathcal{A}$ are sampled from p_{AR}
2. Nature uses $\mathcal{N}$ to prepare data B depending on the value of A
3. Learner predicts A using estimator $\mathcal{D} : B \to \hat{A}$

Classical hypothesis testing

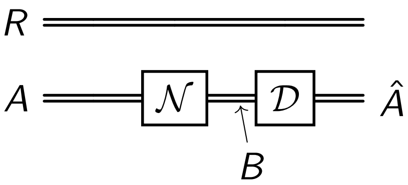


1. Random variables A, R taking values in $\mathcal{A}$ are sampled from p_{AR}
2. Nature uses $\mathcal{N}$ to prepare data B depending on the value of A
3. Learner predicts A using estimator $\mathcal{D} : B \to \hat{A}$

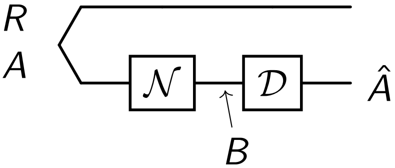
Learner's optimal probability of success obeys:

$$\max_{\mathcal{D}} \Pr(\hat{A} = R) \geq 2^{-H(R|B)} \tag{2}$$

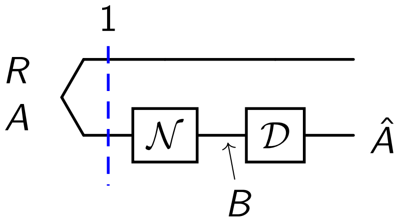
Optimizing singlet fraction



Optimizing singlet fraction

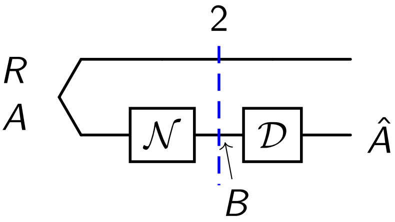


Optimizing singlet fraction



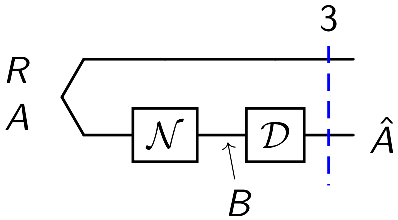
1. RA contains an entangled state $|\psi\rangle_{RA}$

Optimizing singlet fraction



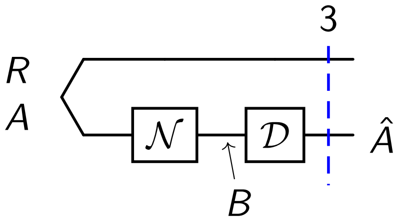
1. RA contains an entangled state $|\psi\rangle_{RA}$
2. Nature evolves A via $\mathcal{N}$ to prepare quantum system B

Optimizing singlet fraction



1. RA contains an entangled state $|\psi\rangle_{RA}$
2. Nature evolves A via $\mathcal{N}$ to prepare quantum system B
3. Learner applies $\mathcal{D}$ to maximize singlet overlap of $R\hat{A}$

Optimizing singlet fraction



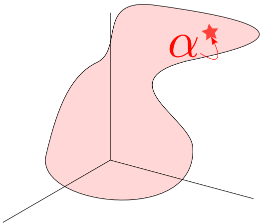
1. RA contains an entangled state $|\psi\rangle_{RA}$
2. Nature evolves A via $\mathcal{N}$ to prepare quantum system B
3. Learner applies $\mathcal{D}$ to maximize singlet overlap of $R\hat{A}$

Learner's *optimal singlet fraction* obeys

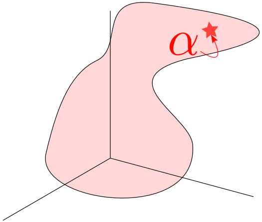
$$\max_{\mathcal{D}} d\langle\Phi|(\mathbb{I}_R \otimes \mathcal{D})\rho_{RB}|\Phi\rangle_{R\hat{A}} \geq 2^{-H(R|B)_\rho} \tag{4}$$

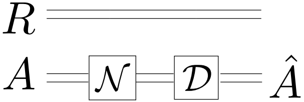
where $|\Phi\rangle_{R\hat{A}}$ is a maximally entangled state on $R\hat{A}$

Classical learning



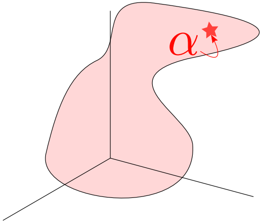
Classical learning



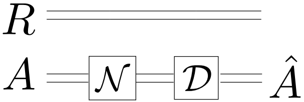


Classical hypothesis testing

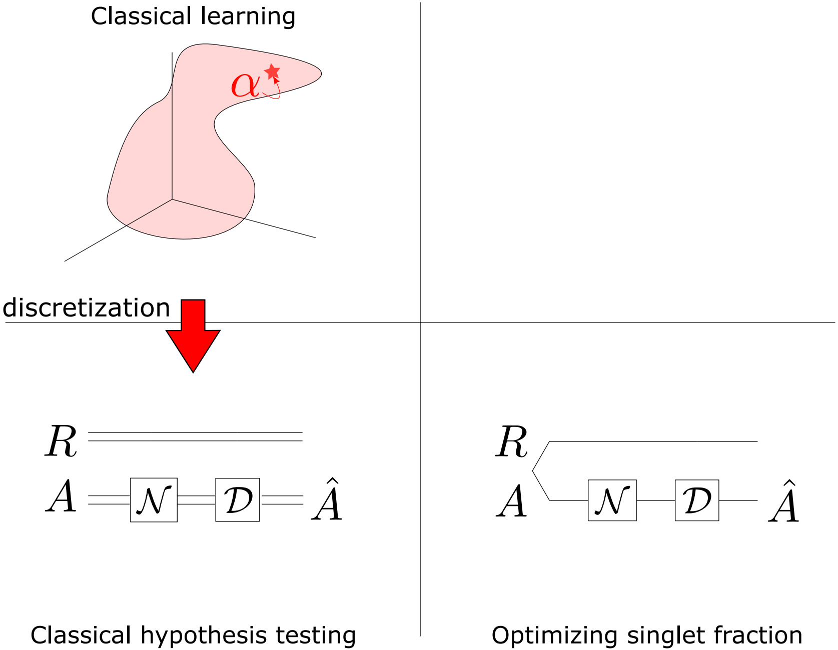
Classical learning

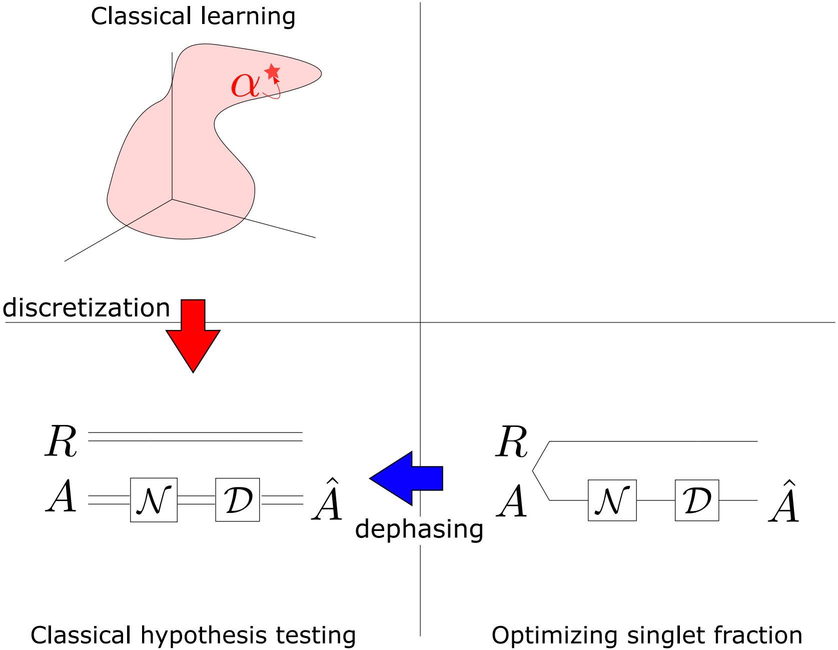


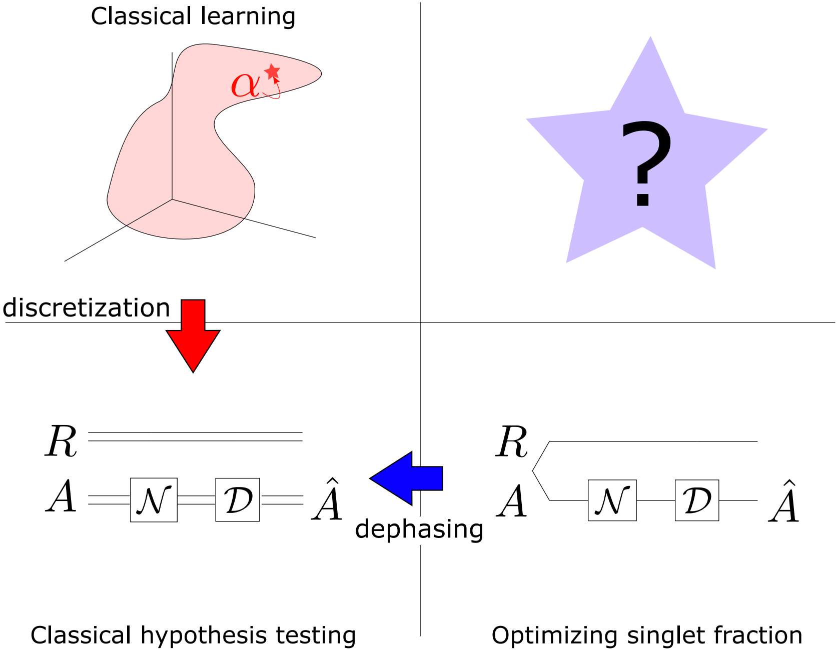
discretization



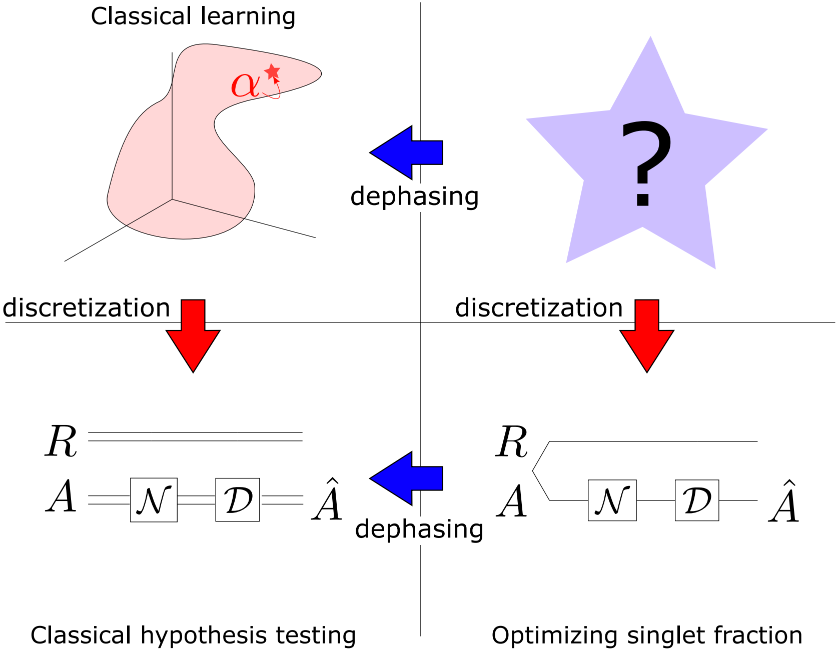
Classical hypothesis testing


Classical learning
α^*
discretization
R
A
$\mathcal{N}$
$\mathcal{D}$
$\hat{A}$
Classical hypothesis testing
R
A
$\mathcal{N}$
$\mathcal{D}$
$\hat{A}$
Optimizing singlet fraction


Classical learning
$\alpha^\star$
discretization
dephasing
R
A
$\mathcal{N}$
$\mathcal{D}$
$\hat{A}$
Classical hypothesis testing
Optimizing singlet fraction



Classical learning
α*
?
discretization
R
A
𝒩
𝒟
Â
dephasing
Classical hypothesis testing
Optimizing singlet fraction


Classical learning
α^*
dephasing
?
discretization
discretization
R
A
$\mathcal{N}$
$\mathcal{D}$
$\hat{A}$
dephasing
R
A
$\mathcal{N}$
$\mathcal{D}$
$\hat{A}$
Classical hypothesis testing
Optimizing singlet fraction

A quantum generalization of learning

Optimizing singlet fraction

Quantum generalization

A quantum generalization of learning

Optimizing singlet fraction

Initial state
$|\psi\rangle_{RA} = \sum_{r,a} \psi_{ra} |r\rangle_R \otimes |a\rangle_A$

Quantum generalization

A quantum generalization of learning

Optimizing singlet fraction

Initial state

$|\psi\rangle_{RA} = \sum_{r,a} \psi_{ra} |r\rangle_R \otimes |a\rangle_A$

Quantum generalization

Initial state

$|\Psi\rangle_{\mathbf{RA}} := \int_{\mathrm{A}} dr \int_{\mathrm{A}} d\alpha \psi(r, \alpha) |r\rangle_{\mathbf{R}} \otimes |\alpha\rangle_{\mathbf{A}}$

A quantum generalization of learning

Optimizing singlet fraction

Initial state

$|\psi\rangle_{RA} = \sum_{r,a} \psi_{ra} |r\rangle_R \otimes |a\rangle_A$

Singlet

$|\Phi\rangle_{R\hat{A}} \propto \sum_r |r\rangle_R |r\rangle_{\hat{A}}$

Quantum generalization

Initial state

$|\Psi\rangle_{\mathbf{RA}} := \int_{\mathrm{A}} dr \int_{\mathrm{A}} d\alpha \psi(r, \alpha) |r\rangle_{\mathbf{R}} \otimes |\alpha\rangle_{\mathbf{A}}$

A quantum generalization of learning

Optimizing singlet fraction

Initial state

$$|\psi\rangle_{RA} = \sum_{r,a} \psi_{ra} |r\rangle_R \otimes |a\rangle_A$$

Singlet

$$|\Phi\rangle_{R\hat{A}} \propto \sum_r |r\rangle_R |r\rangle_{\hat{A}}$$

Quantum generalization

Initial state

$$|\Psi\rangle_{\mathbf{RA}} := \int_{\mathrm{A}} dr \int_{\mathrm{A}} d\alpha \psi(r,\alpha) |r\rangle_{\mathbf{R}} \otimes |\alpha\rangle_{\mathbf{A}}$$

Entangled state

$$|\Phi_\epsilon\rangle_{\mathbf{R\hat{A}}} \propto \int_{\mathrm{A}} dr |r\rangle_{\mathbf{R}} \otimes \int_{\mathbb{B}_\epsilon(r)} d\hat{\alpha} |\hat{\alpha}\rangle_{\mathbf{\hat{A}}}$$

A quantum generalization of learning

Optimizing singlet fraction	**Quantum generalization**
Initial state $\lvert\psi\rangle_{RA} = \sum_{r,a} \psi_{ra} \lvert r\rangle_R \otimes \lvert a\rangle_A$	Initial state $\lvert\Psi\rangle_{\mathbf{RA}} := \int_{\mathbb{A}} dr \int_{\mathbb{A}} d\alpha \psi(r,\alpha) \lvert r\rangle_{\mathbf{R}} \otimes \lvert\alpha\rangle_{\mathbf{A}}$
Singlet $\lvert\Phi\rangle_{R\hat{A}} \propto \sum_r \lvert r\rangle_R \lvert r\rangle_{\hat{A}}$	Entangled state $\lvert\Phi_\epsilon\rangle_{\mathbf{R\hat{A}}} \propto \int_{\mathbb{A}} dr \lvert r\rangle_{\mathbf{R}} \otimes \int_{\mathbb{B}_\epsilon(r)} d\hat{\alpha} \lvert\hat{\alpha}\rangle_{\mathbf{\hat{A}}}$

Is there an analogue for the classical learning guarantee?

A quantum generalization of learning

Optimizing singlet fraction

Initial state
$$|\psi\rangle_{RA} = \sum_{r,a} \psi_{ra} |r\rangle_R \otimes |a\rangle_A$$

Singlet
$$|\Phi\rangle_{R\hat{A}} \propto \sum_r |r\rangle_R |r\rangle_{\hat{A}}$$

Quantum generalization

Initial state
$$|\Psi\rangle_{\mathbf{RA}} := \int_{\mathrm{A}} dr \int_{\mathrm{A}} d\alpha \psi(r,\alpha) |r\rangle_{\mathbf{R}} \otimes |\alpha\rangle_{\mathbf{A}}$$

Entangled state
$$|\Phi_\epsilon\rangle_{\mathbf{R\hat{A}}} \propto \int_{\mathrm{A}} dr |r\rangle_{\mathbf{R}} \otimes \int_{\mathbb{B}_\epsilon(r)} d\hat{\alpha} |\hat{\alpha}\rangle_{\hat{\mathbf{A}}}$$

Is there an analogue for the classical learning guarantee?

$$-H(R|B)_\rho \leq \max_{|\Psi\rangle_{\mathbf{RA}}} \max_{\mathcal{D}} \log \langle \Phi_\epsilon | (\mathbb{I} \otimes \mathcal{D}) \rho_{\mathbf{RB}} | \Phi_\epsilon \rangle_{\mathbf{R\hat{A}}}$$

Summary

- **Classical learning:** Learn $\mathcal{D}$ such that $p_{A\hat{A}}(\alpha, \hat{\alpha})$ is highly correlated

Summary

- **Classical learning:** Learn $\mathcal{D}$ such that $p_{A\hat{A}}(\alpha, \hat{\alpha})$ is highly correlated
- **Generalized task:** Learn $\mathcal{D}$ such that $\rho_{R\hat{A}}$ is highly entangled

Summary

- **Classical learning:** Learn $\mathcal{D}$ such that $p_{A\hat{A}}(\alpha, \hat{\alpha})$ is highly correlated
- **Generalized task:** Learn $\mathcal{D}$ such that $\rho_{R\hat{A}}$ is highly entangled

Applications

- Understanding ability to manipulate *quantum data* coherently
- Continuous-variable resources for teleportation
- Entanglement resources as learning resources

Thank you!

@e6peters $\mathbb{X}$ (twitter) for manuscript eventually

Leonardo Banchi, Jason Pereira, and Stefano Pirandola. Generalization in quantum machine learning: A quantum information standpoint. *PRX Quantum*, 2: 040321, Nov 2021. doi: 10.1103/PRXQuantum.2.040321.

Matthias Caro, Tom Gur, Cambyse Rouzé, Daniel Stilck França, and Sathyawageeswar Subramanian. Information-theoretic generalization bounds for learning from quantum data, 2023. arXiv:2311.05529.

Michał Horodecki, Paweł Horodecki, and Ryszard Horodecki. General teleportation channel, singlet fraction, and quasidistillation. *Phys. Rev. A*, 60:1888–1898, Sep 1999. doi: 10.1103/PhysRevA.60.1888.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Information-theoretic bounds on quantum advantage in machine learning. *Phys. Rev. Lett.*, 126:190505, May 2021. doi: 10.1103/PhysRevLett.126.190505.

I. A. Ibragimov, R. Z. Has'minskii, and Samuel Kotz. *Statistical estimation : asymptotic theory*. Stochastic Modelling and Applied Probability, 16. Springer Science+Business Media, LLC, New York, 1st ed. 1981. edition, 1981. ISBN 1-4899-0027-6.

Robert Koenig, Renato Renner, and Christian Schaffner. The operational meaning of min- and max-entropy. *IEEE Trans. Inform. Theory*, 55(9):4337–4347, 2009. doi: 10.1109/TIT.2009.2025545.