Information-theoretic aspects of ML for QECC decoding

Evan Peters [1]

[1]Department of Applied Mathematics, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

UNIVERSITY OF WATERLOO

Theory: Decoders and hypothesis testing

Can we use single-parameter "noisiness" of a system to predict the accuracy of decoding?

Decoding an [[n,k]] Quantum Error-Correcting Code (QECC) as *hypothesis testing*:

$$E \to A \to \Sigma \to \hat{A} \tag{1}$$

1. An *error* $E \in \mathcal{P}_n$ occurs
2. E induces *coset label* $A = (L, T)$, combining *logical error* L and *pure error* T .
3. The pure error T gives a *syndrome* Σ
4. We predict $\hat{A}$ for which logical error occured

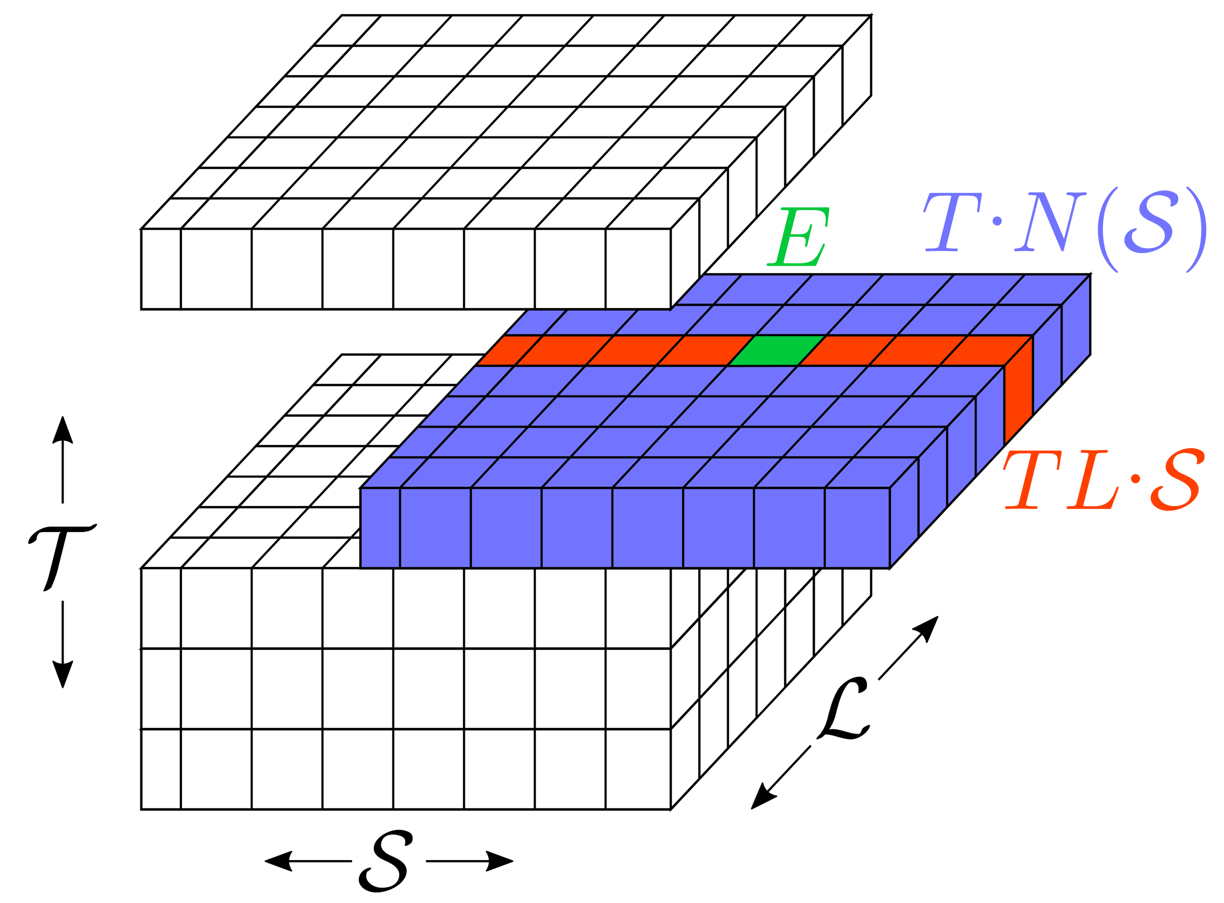

Figure 1. Partition the Pauli group $\mathcal{P}_n$ into a "prism" [1] of $\mathcal{L} \times \mathcal{S} \times \mathcal{T}$ (dimensions $2^{2k} \times 2^{n-k} \times 2^{n-k}$). Logical operators are in $\mathcal{L} := N(\mathcal{S})/\mathcal{S}$, stabilizers are in $\mathcal{S} = \langle g_1, \ldots, g_{n-k} \rangle$, pure errors are in $\mathcal{T} := \{t_1, \ldots, t_{n-k}\}$ with $\{t_i, g_i\} = 0$.

A good decoder has small error probability $p_{err}(A|\Sigma) := \Pr_{p_E}(A \neq \hat{A})$. A ***Maximum Likelihood Decoder (MLD)*** achieves a minimum error p^*_{err}.

Bounding decoder accuracy with entropy

The *Shannon entropy* of a variable $X \sim p_X$ is

$$H(X) := -\sum_x p_X(x) \log p_X(x) := H(p_X) \tag{2}$$

For QECCs, the *conditional* entropy satisfies

$$H(A|\Sigma) = H(A) - H(\Sigma) \tag{3}$$

This gives us upper [2] and lower [3] bounds for the MLD accuracy:

$$H(A|\Sigma) \leq h_2(p_{err}) + 2kp_{err} \tag{4}$$

$$H(A|\Sigma) \geq \Phi_N(p^*_{err}) \tag{5}$$

where Φ_N is some decreasing convex function.

⚠ Computing [upper-bounding] $H(A)$ is worst-case #P-complete [(probably) NP-hard].

Experiment

Consider a noise model $\mathcal{N}^{\otimes n}$, where

$$\mathcal{N}(\rho) = p_I\rho + p_X X\rho X + p_Y Y\rho Y + p_Z Z\rho Z. \tag{6}$$

We can define $H(p) = H(\{p_I, p_X, p_Y, p_Z\})$. Some facts:

$$H(A) \leq \min(nH(p), n+k) \qquad H(\Sigma) \leq (n-k) \tag{7}$$

We generate random [[n, k]] codes and compute (optimal) decoder performance:

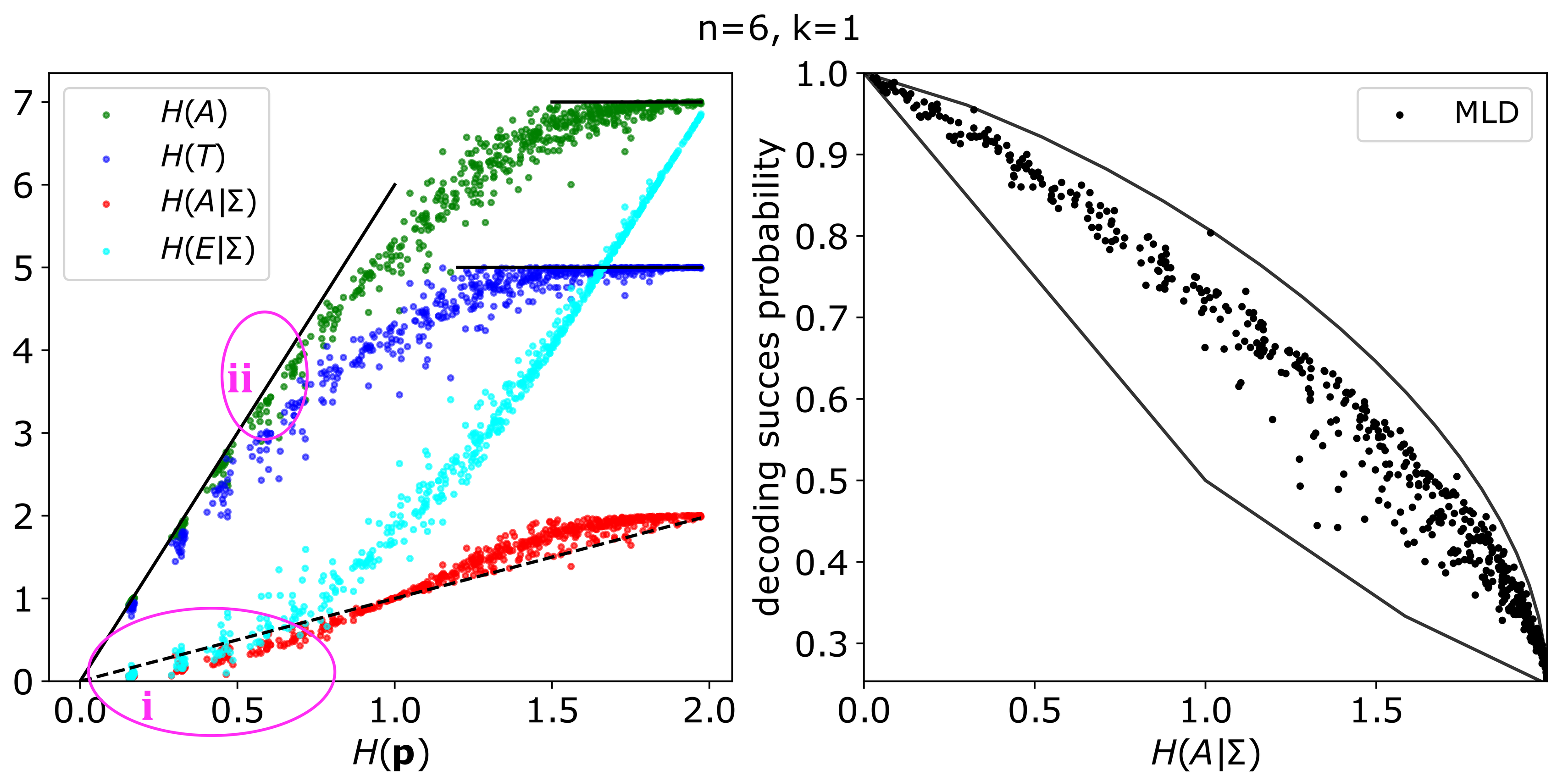

Figure 2. LEFT: (i) Random codes are more nondegenerate for small $H(\mathbf{p})$. (ii) $H(A) < nH(\mathbf{p})$ implies more degeneracy. Dashed line $H(A|\Sigma) = kH(\mathbf{p})$ is analytical solution for a "canonical" stabilizer code. Solid lines indicate bounds from Eq. 7. RIGHT: MLD accuracy for random codes is bounded by Eqs. 4-5

Applications: Shannon/learning theory for MLdec

Decoding an [[n, k]] QECC (no fault tolerance) as a *learning problem (MLdec)*:

1. **Given**: A dataset

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}, \quad x_i = \Sigma(E_i),\ y_i = A(E_i) \tag{8}$$

   and a *loss function* ℓ

2. **Learn**: a decoder function $f : \mathcal{T} \to \mathcal{L}$ (e.g. NN) minimizing empirical risk

$$\hat{f} = \arg\min_f \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell(f(x), y) \tag{9}$$

3. **Hope**: Generalization error is small, e.g. for 0-1 loss

$$\text{loss}(\hat{f}) \geq \min_f \mathop{\mathbb{E}}_{E\sim p_E} \ell(f(x(E)), y(E)) := p^*_{err} \tag{10}$$

Variants

- **Generative models** learn $\hat{f}$ by first approximating $\hat{p}_{LS\Sigma}$, then

$$\hat{f}(x) = \arg\max_\ell \hat{p}_{L|\Sigma}(\ell|x) \tag{11}$$

- **Noisy**: Data (x, y) or underlying circuit may be noisy.
- **Data-driven**: Using empirical data instead of simulating $(\sigma(E_i), A(E_i))$ (unrealistic). e.g. surface code detector data:

$$x = \{\langle g_j \rangle^{(0)}, \ldots, \langle g_j \rangle^{(t)}\}_{\forall j}, \qquad y = |\langle \bar{L} \rangle^{(0)} - \langle \bar{L} \rangle^{(t)}| \tag{12}$$

- **Non-degenerate** Choosing $y_i = E_i$ instead of $y_i = A(E_i)$ (suboptimal, but easier?)

Variant		Refs
simulated data	non-degenerate	[4–10]
	degenerate	[10–15]
data-driven/fault-tolerant		[16–28]

Table 1. Survey of Variants of existing MLdec schemes.

Design considerations for ML decoders

1. Relating the *model architecture* to *out-of-distribution generalization*
   - What models can represent the group structure of Fig. 1?
2. Side-information [24]: x contains upstream raw data (e.g. IQ-plane coords)
3. Lifting ML decoders into fault-tolerant settings [29]
4. Regimes for non-degenerate decoding, where $y_i = E_i$ instead of $A(E_i)$
   - At low $H(\mathbf{p})$, degenerate decoding ≈ non-degenerate decoding
   - Otherwise, non-degenerate decoding is *sub-optimal*
5. Equivariance [12, 28] vs. noise tolerance: How well can group structure be learned with noisy labels?

How much training data do we need?

Naive: $|\mathcal{S}| = 2^{n-k}$ unique data are sufficient for MLD

Shannon theory: If $\mathcal{E}_{typ}$ contains "typical errors" s.t. $\Pr(\mathcal{E}_{typ}) > 1 - \epsilon$, then *asymptotically*,

$$|\mathcal{E}_{typ}| \approx 2^{nH(\mathbf{p})} \tag{13}$$

⚠ In less noisy systems, we don't care about most syndromes.

- Noiseless syndromes: a *look-up table* of $N \approx 2^{H(A)} \leq |\mathcal{E}_{typ}|$ data is sufficient.
- Compare to linear-algebraic sensitivity of Ref. [10] (not noise-specific)

Learning theory: For random stabilizers, suppose labels contain the maximum-likelihood coset:

$$(x_i, y_i) = (\sigma_i, \arg\max_a P_{A|\Sigma}(a|\sigma_i)). \tag{14}$$

Say $f_{\mathcal{D}}$ is trained on a dataset of $|\mathcal{D}| = N$ points. Find h such that

$$\mathop{\mathbb{E}}_{\mathcal{D}} \Pr_{E\sim p_E} (f_{\mathcal{D}}(\sigma(E)) \neq A(E)) < h(N) \tag{15}$$

The form of $h(N)$ will decide how effective we expect MLdec to be.

Acknowledgements

Thanks to Ewan Murphy and Zachary Mann for helpful discussions.

e6peters@uwaterloo.ca