NISQ: Error Correction, Mitigation, and Noise Simulation

Ningping Cao,† Junan Lin,† David Kiibs,2 Yiu-Tung Poon,3 Bei Zeng,4,1 and Raymond Laflamme1
1Institute for Quantum Computing and Department of Physics and Astronomy, University of Waterloo, Waterloo, ON N2L 3G1, Canada
2Department of Mathematics & Statistics, University of Guelph, Guelph, ON N1G 2W1, Canada
3Department of Mathematics, Iowa State University, Ames, IA, USA 50011
4Department of Physics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
(Dated: November 4, 2021)

Error-correcting codes were invented to correct errors on noisy communication channels. Quantum error correction (QEC), however, may have a wider range of uses, including information transmission, quantum simulation/computation, and fault-tolerance. These invite us to rethink QEC, in particular, about the role that quantum physics plays in terms of encoding and decoding. The fact that many quantum algorithms, especially near-term hybrid quantum-classical algorithms, only use limited types of local measurements on quantum states, leads to various new techniques called Quantum Error Mitigation (QEM). This work addresses the differences and connections between QEC and QEM, by examining different application scenarios. We demonstrate that QEM protocols, which aim to recover the output density matrix, from a quantum circuit do not always preserve important quantum resources, such as entanglement with another party. We then discuss the implications of noise invertibility on the task of error mitigation, and give an explicit construction called quasi-inverse for non-invertible noise, which is trace preserving while the Moore-Penrose pseudoinverse may not be. We also study the consequences of erroneously characterizing the noise channels, and derive conditions when a QEM protocol can reduce the noise.

I. INTRODUCTION

The field of quantum information processing has entered an era featuring noisy, intermediate-scale quantum (NISQ) devices. Despite some recent demonstrations of computational advantages compared to classical computers [1,2], NISQ devices still face significant challenges before eventually becoming practically useful. In particular, noise in NISQ processors can spoil the computation process and possibly lead to incorrect final results.

Conventionally, the main tool for protecting the processor from noise has been quantum error correction (QEC). QEC protocols are designed to allow a user to detect, and eventually correct, errors that happen during a quantum computation. While many approaches for QEC have been developed, few have been tested on real quantum processors due to the significant requirements on the hardware. First, QEC generally encodes quantum information into a much larger Hilbert space, which requires the hardware size to be large as well. Second, quantum operations (gates) on a processor must below a certain threshold value for QEC to successfully reduce the effective error, instead of introducing more errors. Meeting both requirements is generally difficult on most state-of-the-art devices available today.

Recently, the field of quantum error mitigation (QEM) emerged with the goal of decreasing the effective noise level, while circumventing these two obstacles, on near term devices. The general consideration is that, if one has some knowledge about the noise processes happening in a particular hardware, then one should be able to utilize that knowledge to reduce (part of) the effect of that noise. Importantly, it is more desirable to have protocols that does not require (or requires very little) additional hardware overhead in order to improve the computation accuracy. Numerous protocols have been developed during the past few years [3–7] that fall into this category.

The parallel development of both fields naturally leads to the question: under what circumstances should one apply QEC over QEM, and vice versa? The current experimental apparatus favors QEM due to limitations on hardware quality. However, there exists deeper distinctness between the two that restricts the use of QEM under some experimental goals.

In this work, we first examine the relation between QEC and QEM from a high-level perspective. In Section II, we give examples from classical and quantum communication, demonstrating the different usage scenes of QEC and QEM. We argue that the invertibility of noise limits the performance of optimal QEM protocols, and propose a construction called quasi-inverse in case of non-invertible noise. We prove that compared to a conventional choice of pseudoinverse, the Moore-Penrose pseudoinverse, the quasi-inverse has the advantage of being trace preserving, which is advantageous in running computer simulations. In Section III, we study the effects due to imperfect characterizations of noise channels, and give a sufficient condition for when an optimal QEM can improve the expectation value of any observable.

* These authors contribute equally to this work.
† zengb@ust.hk
II. THE CASE OF COMMUNICATION

A. Classical communication

We start by considering a communication task, and for simplicity, we start from the classical setting. Here, a sender Alice would like to transmit a \( k \)-bit string to Bob. An example of the string would look like:

\[
s = 11010101000100110101010011...
\]  

(1)

Alice and Bob share a classical communication channel \( C \) which is subject to noise. She does so by sending the text through \( C \) only once. Suppose that there are 40% of 0’s and 60% of 1’s in the text string. For simplicity, assume that the noise is described by a binary symmetric noise channel with strength \( p \), denoted as \( \text{BSC}_p \) (note that we’ll reserve script letters for quantum channels). This channel preserves the sent bit with probability \( 1-p \), and flips it (symmetrically from 0 to 1 and from 1 to 0) with probability \( p \).

Classical error correcting codes have been developed to fight this noise. An illustrative figure is given in Fig. 1.

The simplest example is the 3-bit repetition code: it is defined by the encoding

\[
0 \rightarrow 000, \ 1 \rightarrow 111,
\]  

(2)
i.e., each bit is repeatedly encoded 3 times. The number of uses of the channel has now increased 3 times to \( 3k \). The decoding is done by performing a majority vote on the received bits, so that at the receiving end, Bob again obtains a bit string of length \( k \). Assuming that \( \text{BSC}_p \) acts independently on each bit, the probability of error is reduced from \( p \) to \( 3p^2(1-p) = O(p^2) \) by this code.

It has long been recognized that the key ingredient that enables classical error correction is trading fidelity with redundancy. This concept has been transferred to quantum error correcting codes (QECCs), where a direct mimicking of classical codes is forbidden due to the no-cloning principle. The complexity of quantum noise has also posed a great challenge to building quantum codes. Nonetheless, many QECCs have been developed that fight against different kinds of noise. Beyond communication, QECCs have also found applications in fields like quantum computation, quantum simulation, and fault-tolerance. We will discuss further about QECCs in Section IIIB.

Next, consider another seemingly “natural” way of reducing the effect of noise. We start by writing down a matrix representation of the noise,

\[
\text{BSC}_p \rightarrow \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} := N_{\text{BSC}_p}
\]  

(3)

where both the input and output basis are ordered as \( \{0,1\} \). The action of \( N_{\text{BSC}_p} \) is through matrix multiplication, so that if Alice sends a bit 0, it becomes at the output end

\[
N_{\text{BSC}_p}(1) = \begin{pmatrix} 1-p \\ p \end{pmatrix},
\]  

(4)

so that Bob gets 0 with probability \( 1-p \) and 1 with probability \( p \).

Note that since classical noise is governed by laws of classical physics, in principle an almighty Bob can learn exactly whether a noise process has occurred or not during the transmission. If this is the case, then the uncertainty in our probabilistic error model disappears, and he can perfectly recover the input message. But in reality, Bob does not have such information, and Eq. (3) represents his complete knowledge about the noise. Suppose that in addition, Bob knows the value of \( p < 1/2 \). What is then the best that Bob can do in attempt to reduce the noise effects? If Bob receives a bit 1, then he only knows that Alice more likely sent a 1 than a 0, so the best deterministic procedure is simply to keep the bit. Applying this argument to all received bits, we see that the best that Bob can do is to simply keep all received bits intact.

There is, however, another possibility for Bob to use his knowledge about \( p \). If \( k \) is sufficiently large, then Bob can fully recover the 40% - 60% distribution of Alice’s input. Specifically, he applies the inverse map of \( N_{\text{BSC}_p} \) on his received distribution, resulting in

\[
N_{\text{BSC}_p}^{-1} N_{\text{BSC}_p} v_A = v_A,
\]  

(5)

where \( v_A = (0.4, 0.6)^T \) is Alice’s input distribution. However, Bob cannot further use this restored distribution to recover Alice’s message. The best he can do is to use the restored distribution to randomly generate a new \( k \)-bit string, during which Alice’s message is completely destroyed. This procedure is analogous to QEM under the classical communication setting, so we will call it classical error mitigation (CEM). In particular, CEM does not increase the channel capacity, defined by

\[
C = \max_{p(x)} I(X : Y)
\]  

(6)

where \( X \) and \( Y \) are random variables denoting the input and output, respectively, \( I(X : Y) \) is the mutual information between \( X \) and \( Y \), and the maximum is over all probability distributions of the input, \( p(x) \). This is a direct consequence of the data processing inequality.
B. Quantum Communication

The problem of preserving information becomes much more interesting and complicated for quantum communication. For example, many different definitions for a “quantum” channel capacity can be proposed based on different considerations [8]. Below, we consider two situations in order to contrast QEC and QEM.

1. When to use both QEC and QEM

First, consider the situation in Fig. 2a where Alice prepares $k$ copies of the input state $\rho_n$, and send it to Bob using a noisy quantum channel $\mathcal{N}$. This involves $k$ uses of the quantum channel. How can Alice an Bob reduce the effect of the noise $\mathcal{N}$?

Earliest development of QECCs have initiated by considering exactly this model [9]. QECCs works by using a certain encoding scheme, which is a completely positive and trace preserving (CPTP) map $\mathcal{E} : C_{2^n} \rightarrow C_{2^n}$ (where $n > k$), and then decode at Bob’s end using another CPTP map $\mathcal{R} : C_{2^n} \rightarrow C_{2^n}$. Here $C_{2^n}$ denotes the complex Euclidean space with dimension $2^n$. This involves $n$ uses of the quantum channel which increases the redundancy. It has been shown that doing so may protect the system against certain types of noise.

Let’s contrast this with QEM. In QEM, we again assume that Bob has some knowledge about the noise $\mathcal{N}$; in order not to constrain Bob’s power, we further assume that Bob knows the exact form of $\mathcal{N}$. Upon receiving $k$ copies of $\rho_{out}$, Bob first makes some measurements to reconstruct $\rho_{out}$ (e.g., by using quantum state tomography), then applies the inverse channel $\mathcal{N}^{-1}$ (assumed exists for now) to recover $\rho_n$.

One should now recognize the similarity between QEM and CEM for reconstructing the input distribution. Indeed, a density matrix is probabilistic description of outcomes of any possible measurement on a quantum system. The only difference is perhaps that quantum noise is more complex in nature. In this scenario, it is sufficient for Bob to reconstruct $\rho_n$ as a mathematical object in order to eliminate the effect of $\mathcal{N}$, since this determines the outcome of any measurement Bob can possibly make on the system. Thus, QEM can be useful in this case.

We can further materialize this view by considering an example using current QEM protocols, such as the quasiprobability decomposition method [3]. The goal here is strictly weaker than the one above; namely, we would like to recover the expectation value $\text{Tr}[O\rho_{in}]$ for some observable $O$. It is weaker because we are now only concerned with some particular measurements corresponding to $A$. The protocol assumes that a set of imperfect gates $\tilde{G}_i$, which forms a spanning set of all 1-qubit gates, is available. In particular, the ideal channel (which is identity in this case) can be decomposed into a quasiprobability distribution (a probability distribution with $+$ or $-$ signs) of all $\tilde{G}_i$’s. Upon receiving the state, Bob performs gates from $\{\tilde{G}_i\}$ according to the underlying probability distribution, measure the observable $O$, and updates the measurement outcome with the $+$ or $-$ sign. This effectively recovers $\text{Tr}[O\rho_{in}]$. We can clearly see in this example that error mitigation is helpful if our goal is to restore the outcome of some measurements, i.e., some classical “shadows” of the quantum system.

2. When to use QEC only

Next, consider the situation where not only its classical image $\rho$, but also the quantum object itself, is of interest. For example, this is the case when entanglement is being used as a resource by two spatially separated users Alice and Bob to achieve some tasks. Consider Fig. 3 where a central source would like to distribute $k$ copies of maximally entangled Bell pairs

$$\Phi^+ = (|00\rangle_{AB} + |11\rangle_{AB})/\sqrt{2} \quad \text{(7)}$$

to Alice and Bob. The quantum channel to Bob is noisy and is described by a channel $\mathcal{N}$. Again, assume for now that $\mathcal{N}$ is invertible.

![FIG. 2:](image1)

![FIG. 3:](image2)
First, the channel to Alice is noiseless, so Alice alone has no knowledge on $\mathcal{N}$. This implies that allowing a 1-way communication channel from Alice and Bob cannot improve the fidelity of Bob’s channel. So any possible operation must be performed on Bob’s side only. A “natural” QEM protocol Bob can apply is then to first measure his qubit using the $k$ copies, obtain $\rho_B$, then apply the inverse channel $\mathcal{N}^{-1}$. This is shown in Fig. 3b. The recovered state then represents all of Bob’s knowledge of his qubit, since from this he can calculate the probability of any measurement outcome that he can make.

The only problem with the above protocol is that it is obviously useless. In particular, all entanglement between Alice and Bob has been destroyed due to the measurement. In fact, analogous to the classical case where Bob recovers Alice’s input distribution and generate a random $k$-bit string, here Bob knows in advance that he will ideally get a maximally mixed state $I/2$; so the above protocol is simply equivalent to Bob generating $(I/2)^k$ locally, and discarding all qubits received from the Source!

The “correct” way of reducing the effect of $\mathcal{N}$ is by using a family of procedures called entanglement purification protocols (EPPs) [10]. This is shown in Fig. 3c. In EPP, Alice and Bob needs to start from $n > k$ copies of the noisy Bell state, and obtain $k$ pairs at the end which are closer to the pure state $\Phi^+$. We omit the details of different possible procedures here, but emphasize that only local operations and classical communications (LOCC) are used in all variants of EPP. Here again, we see that redundancy is necessary in this task, similar to classical EC in preserving classical information.

The above analysis shows that EPP protocols can protect entanglement against noise. But what is perhaps more profound is that a class of EPP protocols called one-way EPP (1-EPP) also permits the creation of a QECC [10]. This is enabled by quantum teleportation: the one-way constraint creates time-separated EPR pairs like $\Phi^+$, allowing a quantum object in an arbitrary state $|\xi\rangle$ to be teleported forward in time. This effectively creates a faithful transportation of quantum information from Alice to Bob, despite the presence of noise from the source of Bell pairs. This type of QECC is particularly relevant in distributed quantum computing, where quantum information needs to be transported among spatially separated locations. In this case the local density matrices, although representing the full knowledge by each individual location, do not cover the whole picture; so using QEM to recover the ideal local density matrices will not be useful.

The above example shows a deep distinction between QEM, which is only capable of restoring the classical image of a quantum system, and QEC, which is capable of restoring the quantum system itself, along with all possible non-classical resources that the quantum system possesses. It is instructive to compare the above with the classical counterpart, namely, that when one needs to preserve classical information (see Section II A). There, we have also argued that EC is helpful for such a task, while EM is not. Indeed, it has long been recognized that entanglement share a similar role to that of classical information [10]. So the task of preserving entanglement may also be viewed as a quantum analog of preserving classical information, which can only be achieved by using QEC. Furthermore, recall in Section III B II we argued that recovering density matrices in QEM is analogous to recovering classical distributions in CEM. These completes our comparisons between EM and EC, which are summarized in Table I.

| | Classical | Quantum |
|---|---|---|
| EC | Classical information | Entanglement |
| EM | Classical distribution | Density matrices |

TABLE I: Comparison between EC and EM under the classical and quantum settings. The table inputs list what the two protocols is capable of restoring, under both settings.

C. Invertible Noise

Finally, before moving on to discuss QEM in quantum computation, in this section we will address how the nature of the noise will determine the “upper limit” of any QEM protocol. Again we start our discussion from classical communication. Consider again the BSC$_p$ noise model. The matrix representation of the noise in Eq. 3 is only a probabilistic description of the underlying physical processes, which can be either nothing (identity) or a bit-flip. The classical bit-flip channel, denoted $F$, acts as follows:

$$F(0) = 1, \quad F(1) = 0.$$  (8)

We see that $F$ is invertible since it is a bijection: i.e., there is a one-to-one correspondence between its inputs and outputs. Moreover, $F^{-1} = F$. If one has the full knowledge of an invertible noise map in a particular transmission (not the probabilistic one as in Eq. (3)), then the noise can in principle be eliminated without redundancy. In the BSC language: if Bob knows that a bit flip happened during a transmission, then he can flip it back to eliminate the noise.

Consider another common classical error model called the binary erasure channel with probability $p$, denoted as BEC$_p$. In this model both 0 and 1 are transmitted with probability $1 - p$, and erased with probability $p$ (Bob knows when a bit is erased). The underlying physical process in this example is an erasure channel, denoted as $E$: it has the effect

$$E(0) = e, \quad E(1) = e.$$  (9)

where $e$ denotes the state of being erased. Since both 0 and 1 correspond to the same output $e$, $E$ is not a bijection and thus non-invertible. In other words, even if
Bob knows that an erasure has occurred, he cannot infer what Alice intended to send. Therefore, such errors cannot be corrected without using degeneracy. Importantly, the 3-bit repetition code introduced in Eq. (2) can protect against the erasure error.

Now we consider the quantum case. A quantum noise process is, on the physical level, described by a CPTP map \( \mathcal{N} \). Just as classical noise, quantum noise can also be either invertible or non-invertible. The invertibility of a quantum noise map can be deduced from its matrix representation, as shown in Theorem 1. There are three distinct possibilities. The first is that \( \mathcal{N} \) is invertible, and \( \mathcal{N}^{-1} \) is CPTP. For an invertible \( \mathcal{N} \), the inverse \( \mathcal{N}^{-1} \) is unique, and is Hermitian preserving (HP) and trace preserving (TP) \(^\text{11}\). If the dimensions of input and output space are the same, the channel has a CPTP inverse iff the channel is an unitary channel \(^\text{12}\) \(^\text{13}\).

It is particularly instructive to re-examine the case in Fig. 6 under this assumption on \( \mathcal{N} \). Since \( \mathcal{N}^{-1} \) is unitary, Bob can in principle implement it on his qubit, which would fully restore the Bell pair \( \left| \Phi^+ \right> \) between him and Alice. This, in fact, also correspond to a QECC, with a trivial encoding map \( \mathcal{E} = I \) and a recovery map \( \mathcal{R} = \mathcal{N}^{-1} \) (see Fig. 2b). Thus, no redundancy is needed in principle to recover this noise. However, note that Bob’s recovery operation is local, which cannot increase the amount of entanglement by any valid measure according to the fundamental postulate of entanglement theory \(^\text{14}\). Since the state after recovery is maximally entangled, the one prior to recovery must be maximally entangled as well, meaning that this (local unitary) noise model does not decrease the entanglement between Alice and Bob. Thus, this noise model is rather trivial from the point of preserving entanglement.

The second possibility is that \( \mathcal{N} \) is invertible, but \( \mathcal{N}^{-1} \) is not CPTP. A condition for when this will happen is later given in Proposition 1. Many experimentally relevant noise models, such as the phase damping channel and the depolarizing channel, fall under this category. Since \( \mathcal{N}^{-1} \) is not a physically realizable operation, it cannot be experimentally implemented on the target system, so our above method to restore quantum information without redundancy fails. Using QEM procedures, one can still recover the classical information in principle, by first extracting the classical information through measurements, and numerically apply the inverse map \( \mathcal{N}^{-1} \). But the process of measurement will inevitably disturb the system being measured, and destroy any entanglement it possibly has with other systems. So in this case, the best possible QEM is capable of restoring the classical information, but not entanglement.

To describe the inverse maps, it is useful to first define a matrix representation for quantum states and maps. In this work we denote the space of linear operators mapping Hilbert space \( H_A \) to \( H_B \) as \( L(H_A, H_B) \), or \( L(H_A) \) in short if \( H_A = H_B \). Let \( T(H_A, H_B) \) be the space of linear maps from \( L(H_A) \) to \( L(H_B) \). Let \( e_i \) be the standard basis of \( H_i \) with a 1 at position \( i \) and 0 elsewhere.

Let \( E_{a,b} \) be the standard basis of \( L(H_A, H_B) \) with a 1 at position \( (a, b) \) and 0 elsewhere.

**Definition 1.** (Vectorization of linear operators.) The vec mapping \( \mathbf{vec} \) : \( L(H_A, H_B) \rightarrow H_B \otimes H_A \) is the unique mapping that satisfies \( \mathbf{vec}(E_{a,b}) = e_b \otimes e_a \).

Next we define representations for quantum maps.

**Definition 2.** (Choi representation.) The Choi representation of a map \( \mathcal{M} \in T(H_A, H_B) \) is defined by \( C(\mathcal{M}) = \sum_{a,b} \mathcal{M}(E_{a,b}) \otimes E_{a,b} \).

**Definition 3.** (Natural representation.) The natural (or equivalently, superoperator) representation of a map \( \mathcal{M} \in T(H_A, H_B) \) is defined by the unique linear operator \( \mathbf{v}(\mathcal{M}) \in L(H_A \otimes H_A, H_B \otimes H_B) \) that satisfies \( \mathbf{v}(\mathcal{M})\mathbf{v}(A) = \mathbf{v}(\mathcal{M}(A)) \) for all \( A \in L(H_A) \).

In the natural representation of quantum channels, the channel \( \mathcal{N} \) acting on a quantum state \( \rho \) can be written as the superoperator \( \mathbf{v}(\mathcal{N}) \) multiply the vector representation \( \mathbf{v}(\rho) \) of the quantum state \( \rho \) \(^\text{15}\). The vector representation \( \mathbf{v}(\rho) \) of \( \rho \) inherits its ordering from the superoperator, hence we abuse the notation \( \mathbf{v}(\cdot) \) for vector representations of quantum states and observables (it often written as the double ket \( \left< \right| \rho \left| \right> \) in other literature).

The following theorem directly comes from representation theory of linear maps.

**Theorem 1.** The quantum channel \( \mathcal{N} \) is invertible iff \( \mathbf{v}(\mathcal{N}) \) is an invertible matrix.

Here is an example that the inverse \( \mathcal{N}^{-1} \) of a CPTP map \( \mathcal{N} \) is non-CP.

**Example 1.** Let the Choi representation of a quantum channel \( \mathcal{N} \) be

\[
C(\mathcal{N}) = \begin{pmatrix}
\frac{3}{4} & 0 & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{pmatrix}.
\]

The superoperator is

\[
\mathbf{v}(\mathcal{N}) = \begin{pmatrix}
\frac{3}{4} & \frac{1}{2} & \frac{1}{2} & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{pmatrix}.
\]

Therefore, the inverse of \( \mathbf{v}(\mathcal{N}) \) is

\[
\mathbf{v}(\mathcal{N}^{-1}) = \begin{pmatrix}
\frac{3}{4} & \frac{1}{2} & \frac{1}{2} & 0 \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{pmatrix}.
\]

Its Choi representation is

\[
C(\mathcal{N}^{-1}) = \begin{pmatrix}
\frac{3}{4} & 0 & \frac{1}{2} & \frac{1}{2} \\
0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{pmatrix}.
\]
The Choi representation $C(N^{-1})$ has negative eigenvalues. Therefore, $N^{-1}$ is a HPTP map, but not CP.

D. Non-invertible Noise

The third possibility is that $N$ is non-invertible. Under this noise, even the classical information cannot be completely recovered without using redundancy in principle. However, the information can be partly restored.

It is known that the superoperator $v(N^{-1})$ of the inverse channel $N^{-1}$ equals to the inverse $v(N)^{-1}$ of the superoperator $v(N)$. However, if the channel $N$ is not invertible, the generalized inverse of $v(N)$ is not unique.

The question of how to construct the inverse-like channel for a non-invertible channel naturally arises.

A commonly used generalized inverse is the Moore-Penrose inverse $\text{MP}$. Later, we will show that the Moore-Penrose inverse of a CPTP map may not be TP anymore. A qubit channel example is shown in Example 2. When the generalized inverse $N^g$ is not TP, the state came out of $N^g$ is not trace 1, which causes failure on metrics like fidelity.

Here we provide a construction of inverse-like channel $N^+$. Let the dimension of input and output space be $d$.

Take the Jordan decomposition of the superoperator of $N$,

$$
v(N) = Q \cdot J \cdot Q^{-1} \tag{10}
$$

where $J$ is the Jordan normal form, $Q$ is a invertible matrix contains the generalized eigenvectors of $v(N)$. If $v(N)$ is diagonalizable, the Jordan normal form $J = \text{diag}(\lambda_1, \cdots, \lambda_d)$ is the diagonal matrix contains eigenvalues $\lambda_i$ of $v(N)$.

We take the inverse-like channel $N^+$ to be

$$
v(N^+) = Q \cdot J^* \cdot Q^{-1} \tag{11}
$$

If $v(N)$ is diagonalizable, $J^*$ is the diagonal matrix that leaves the 0's in $J$ untouched and take the reciprocal of the rest elements in $J$. If $v(N)$ is defective, we can construct each Jordan block in the following way: a $k$ by $k$ Jordan block $J_{\lambda_i}$ of $\lambda_i$ ($\lambda_i \neq 0$) in $J$ is

$$
J_{\lambda_i} = \begin{pmatrix}
\lambda_i & 1 \\
& \lambda_i & \ddots \\
& & \ddots & 1 \\
& & & \lambda_i
\end{pmatrix},
$$

let the corresponding block $J'_{\lambda_i}$ in $J'$ be the inverse of $J_{\lambda_i}$

$$
J'_{\lambda_i} := J_{\lambda_i}^{-1} = \begin{pmatrix}
\frac{1}{\lambda_i} & -\frac{1}{\lambda_i^2} & \cdots & (-1)^{k+1} \frac{1}{\lambda_i^k} \\
\frac{1}{\lambda_i} & \frac{1}{\lambda_i} & \cdots & (-1)^{k} \frac{1}{\lambda_i^{k-1}} \\
& \frac{1}{\lambda_i} & \cdots & \frac{1}{\lambda_i} \\
& & \ddots & \ddots
\end{pmatrix}.
$$

For a $k$ by $k$ Jordan block of diagonal zero ($\lambda_i = 0$), which is the nilpotent matrix $N$, we can set the corresponding block in $J'$ as a zero matrix $0_k$. Since $N$ is not invertible, letting the block be $0_k$ will have the same result as setting it as $N^{k-1}$. There is a certain freedom in the choice of this block.

Note that, for invertible channels, $N^+$ described above provides the inverse $N^{-1}$ of the channel ($N^+ = N^{-1}$). For non-invertible channels, this construction Eq. (11) does not satisfy the condition of generalized inverse ($N \circ N^+ \circ N \neq N$ when the dimension of the nilpotent Jordan block is greater than one), and we will call $N^+$ the quasi-inverse.

The resulting composed map $v(N)v(N^+) = QJ'\cdot Q^{-1}$, where $J'$ is a diagonal matrix, only containing 0's and 1's on its main diagonal. When the noise channel $N$ already contains 0's and 1's in its spectrum, the quasi-inverse is itself, and does not recover more information. In fact, any generalized inverse would not improve the outcome in this case.

The following proposition tells us another condition for a quantum channel to have a non-CP (quasi-) inverse.

**Proposition 1.** If a non-zero eigenvalue $\lambda_i$ of a quantum channel $N$ has modulus less than 1 ($|\lambda| < 1$), then the inverse (or quasi-inverse) channel $N^+$ is not complete positive.

**Proof.** $N$ is a CPTP map, therefore its spectral radius is one [15], i.e. $|J_{ii}| \leq 1$ for any main diagonal element $J_{ii}$ in $J$. Since $N$ has eigenvalues less than 1, there exists $|J_{jj}| < 1$ for some $j \in \{1, \cdots, d^2\}$. As defined above, $|J'_{jj}| > 1$, i.e. the spectral radius of $N^+$ is greater than one. Therefore, $N^+$ is not complete positive. \qed

Note that the spectrum of a quantum channel can be defined independently from its representations. In this section, we mainly work with superoperators (natural representation), but the Proposition 1 still holds in other representations (e.g. the Pauli representation).

Superoperators are powerful when calculating channel compositions and their actions on quantum states. Unlike the Choi representation, the natural representation does not directly show a lot of critical properties of quantum channels, like CP, TP, and HP. However, we found that the eigen structure of the superoperator is essential for its property. Lemma 1 and Lemma 2 provide an insight of why Moore-Penrose inverse is not TP in certain cases. And then, in Theorem 2, we prove that the quasi-inverse for a TP map is also TP.

Denote the trace operation in the vector representation $v(A)$ of $d$ by $d$ matrix $A$ as $\text{sTr} [\cdot]$, where $\text{sTr} [v(A)] = \text{Tr}(A)$.

**Lemma 1.** If a linear map $N : M_d \rightarrow M_d$ is trace preserving, the eigenvectors $v$ and generalized eigenvectors $v^g$ of eigenvalue $\lambda \neq 1$ of the superoperator $v(N)$ is trace zero, i.e. $sTr[v] = sTr[v^g] = 0$. 


Proof. For an eigenvector $v$ of $\mathbf{v}(N)$, we have $\mathbf{v}(N)v = \lambda v$. Since $N$ is trace preserving, $\text{sTr}[v] = \text{sTr}[\lambda v]$. And the eigenvalue $\lambda \neq 1$, we have $\text{sTr}[v] = 0$.

For a $k$ by $k$ Jordan block of eigenvalue $\lambda^g$, where $k > 1$, denote the first generalized eigenvector as $v^{g_1}$, we have

$$[\mathbf{v}(N) - \lambda^g \mathbf{I}]v^{g_1} = v, \quad (12)$$

where $v$ is the eigenvector corresponding to $\lambda^g$. Taking the trace on both sizes, $\text{sTr}([\mathbf{v}(N) - \lambda^g \mathbf{I}]v^{g_1}] = \text{sTr}[v]$, the left hand side is $\text{sTr}[[v^{g_1} - \lambda^g v^{g_1}] = (1 - \lambda^g)\text{sTr}[v^{g_1}]$, and the right hand side is zero from the argument above. Since $\lambda^g \neq 1$, $\text{sTr}[v^{g_1}] = 0$. By deduction, all $v^{g_i}$ are trace zero for $i \in \{1, \ldots, k - 1\}$. \qed

**Lemma 2.** For a trace persevering linear map $N : M_k \rightarrow M_k$, if there is a $k$ by $k$ ($k > 1$) defective Jordan Block of eigenvalue $\lambda = 1$ in $\mathbf{v}(N)$, the eigenvector $v$ and first $k - 2$ generalized eigenvector $v^{g_i}$ has to be trace zero, i.e. $\text{sTr}[v] = \text{sTr}[v^{g_i}] = 0$ for $i \in \{1, \ldots, k - 2\}$.

Proof. Assume that $\text{sTr}[v] \neq 0$. The first generalized eigenvector $v^g$ satisfy that $[\mathbf{v}(N) - \mathbf{I}]v^g = v$. Taking trace on both size, the left hand side equals to, and the right hand side does not equal to zero. It is a contradiction. The same argument holds for the rest of generalized eigenvectors except the last one. \qed

From Lemma 1 and Lemma 2, we know that all eigenvectors $v_\lambda$ for $\lambda \neq 1$ of a TP map has to be traceless. When $\lambda = 1$, if its algebraic multiplicity equals to its geometry multiplicity, $\text{sTr}[\mathbf{v}(N)v_\lambda] = \text{sTr}[v_\lambda]$ (i.e. the trace of $v_\lambda$ will not be changed under the action of $\mathbf{v}(N)$); if the algebraic multiplicity does not equal to the geometry multiplicity, the eigenvectors and generalized eigenvectors is traceless except for the last generalized eigenvector. This tells us that the eigen structure of the superoperator $\mathbf{v}(N)$ is crucial for $N$ to be TP. The way that we construct the quasi-inverse $N^+$ largely preserves the eigen structure, while the Moore-Penrose inverse $N^p$ focuses more on the singular value structure. It hints that $N^+$ should be TP and $N^p$ may not.

**Theorem 2.** The quasi-inverse $N^+$ of a trace preserving map $N$ is also trace preserving.

To prove that $N^+$ is trace preserving, we need to prove

$$\text{sTr}[\mathbf{v}(N^+)v_\lambda] = \text{sTr}[v_\lambda],$$

for every eigenvector and generalized eigenvectors $v_\lambda$ of $\mathbf{v}(N)$ in $Q$. From the construction of $N^+$, we almost get trace preserving for free. The proof can be found in Appendix A. Moreover, it is easy to see from the proof that the composed map $N^+ \circ N$ is also trace preserving.

**Example 2.** Here we give an example where the Moore-Penrose inverse $N^p$ of a CPTP map is not TP, but our constructed quasi-inverse $N^+$ is TP. Consider a noise channel $N$ whose Choi representation is given by

$$C(N) = \frac{1}{20} \begin{pmatrix} 8 & 0 & 1 & 6 \\ 0 & 12 & 2 & -1 \\ 1 & 2 & 8 & 0 \\ 6 & -1 & 0 & 12 \end{pmatrix},$$

then its superoperator is

$$\mathbf{v}(N) = \frac{1}{20} \begin{pmatrix} 8 & 1 & 1 & 8 \\ 0 & 6 & 2 & 0 \\ 0 & 2 & 6 & 0 \\ 12 & -1 & -1 & 12 \end{pmatrix}.$$ 

$$J = \text{diag}(0, 1, \frac{2}{5}, \frac{1}{5}), \text{ and } J' = \text{diag}(0, 1, \frac{5}{2}, 5).$$

The superoperator of quasi-inverse $N^+$ is

$$\mathbf{v}(N^+) = \frac{1}{20} \begin{pmatrix} \frac{2}{5} & \frac{2}{5} & \frac{5}{5} & \frac{2}{5} \\ 0 & \frac{2}{5} & \frac{2}{5} & \frac{2}{5} \\ 0 & -\frac{2}{5} & \frac{2}{5} & \frac{2}{5} \\ \frac{5}{2} & \frac{5}{2} & \frac{5}{2} & \frac{5}{2} \end{pmatrix}.$$ 

The Choi representation of $N^+$ is

$$C(N^+) = \frac{1}{20} \begin{pmatrix} 5 & 0 & 10 & 15 \\ 0 & \frac{2}{5} & -\frac{2}{5} & -\frac{2}{5} \\ \frac{2}{5} & \frac{2}{5} & \frac{2}{5} & \frac{2}{5} \\ 0 & -\frac{2}{5} & \frac{2}{5} & \frac{2}{5} \end{pmatrix}.$$ 

The Choi representation has negative eigenvalues. Therefore, the channel $N^+$ is trace preserving, Hermitian preserving, but not complete positive.

The Moore-Penrose inverse of $\mathbf{v}(N)$ is

$$\mathbf{v}(N^p) = \frac{1}{20} \begin{pmatrix} 115 & 10 & 10 & 505 \\ 294 & 9260 & 147 & 882 \end{pmatrix},$$

and its Choi representation is

$$C(N^p) = \frac{1}{20} \begin{pmatrix} 115 & 50 & 10 & 3245 \\ 115 & 9260 & 147 & 882 \end{pmatrix}.$$ 

which is Hermitian preserving but not trace preserving.

Fig. 4 shows the impact of different recovery channels. For this particular channel $N$, the quasi-inverse perfectly recovers the expectation value for Pauli operator $X$.

The composition of a singular channel and its quasi-inverse, $N^{re} := N^+ \circ N$, has a kernel dimension at least one. The effect of the composed channel $N^{re}$ on transmitted quantum states is analyzed in Appendix B.
First, we will consider different methods of implementing physical gates, but still require that noise not be ideal and can introduce more errors. In practice the gates $U_i$ are implemented imperfectly. Making the standard Markovian assumption on the noise, each imperfect $U_i$ can be decomposed as $N_iU_i$, where each $N_i$ is a CPTP map and can be distinct for different $i$. We thus have

$$\rho_{\text{out}} = N_n \circ U_n \circ \cdots \circ N_1 \circ U_1(\rho_{\text{in}})$$

(13)

where $\rho_{\text{in}}$ is the input quantum state, $\rho_{\text{out}}$ is the quantum state came out of the noisy circuits, $U_i$ are the desired operations, and $N_i$ are the noise corresponding to gate $U_i$.

To perform QEM, one first tries to learn (part or all of) the noise models, then recover the ideal gates through either physical or numerical means. Thus, if we wish to analyze the performance of the best possible QEM strategy, we may assume that all $N_i$’s are known exactly. And in reality, these $N_i$’s are obtained from experiments either during the calibration stage or as part of the QEM process, which necessarily involves inaccuracies when being reconstructed. Denote the noise models we received from experiments $\tilde{N}_i$, which are approximations of $N_i$. Let $\tilde{N}_i^{-1}$ denote the inverse of $\tilde{N}_i$. In this section we will consider channels with the same input and output dimensions.

First, consider the case where $\tilde{N}_i^{-1}$ exists and is CPTP for all $i$. Recall that this is true iff $\tilde{N}_i$ is a unitary channel. Then in principle one can insert an additional gate implementing $\tilde{N}_i^{-1}$ after each $U_i$ to fully invert the noise effect [11]. In reality, the experimentally obtained noise models are $\tilde{N}_i$. Thus, the output from this method will be

$$\rho_{\text{EM}} = \tilde{N}_n^{-1} \circ \tilde{N}_n \circ U_n \circ \cdots \circ \tilde{N}_1^{-1} \circ \tilde{N}_1 \circ U_1(\rho_{\text{in}}).$$

(14)

Naturally, there are two main sources of additional error. First, the experimentally learned noise model $\tilde{N}_i$ does not always equal $N_i$, so $\tilde{N}_i \circ N_i$ does not necessarily equal to the identity. Second, even if $N_i$ can be learned ideally, physically implementing $\tilde{N}_i^{-1}$ will also not be ideal and can introduce more errors.

Next, consider the case where $\tilde{N}_i^{-1}$ exists but is not CPTP. In this case it is impossible to physically restore the ideal output state. However, we can still perform the inverse numerically to recover the density matrix of the output state. The error mitigated output density matrix is given by

$$\rho_{\text{EM}} = \mathcal{E}_{\text{EM}}^{-1}(\rho_{\text{out}}) = \tilde{N}_n^{-1} \circ \tilde{N}_n \circ \cdots \circ \tilde{N}_1^{-1} \circ \tilde{N}_1 \circ U_1(\rho_{\text{in}}).$$

(15)

where

$$\tilde{N}_i^{-1} = U_n \circ \cdots \circ U_{i+1} \circ \tilde{N}_i^{-1} \circ U_{i+1} \circ \cdots \circ U_i.$$

(16)

For example, a depth-3 circuit can be numerically inverted by

$$\rho_{\text{EM}} = \tilde{N}_3^{-1} \circ \tilde{N}_2^{-1} \circ \tilde{N}_3^{-1}(\rho_{\text{out}}) = U_3 \circ U_2 \circ \tilde{N}_1^{-1} \circ U_3 \circ \tilde{N}_2^{-1} \circ U_3 \circ \tilde{N}_3^{-1}(\rho_{\text{out}}).$$

(17)

The numerical inverse method does not involve implementing physical gates, but still require that noise processes are accurately characterized. Many current

FIG. 4: The expectation values of the Pauli $X$ operator, $\text{Tr}(\rho X)$, of 50 randomly generated quantum states. The $x$-axis is a dummy label for tested states. The quasi-inverse channel $N^+$ perfectly recover the expectation values (green triangles) while the Moore-Penrose inverse $N^p$ sometimes worsen the results (blue dots).
numerical QEM protocols can be categorized as trying to obtain the exact noise-inverted output, meaning that their optimal performance is upper bounded by Eq. (17).

For example, the quasi-probability sampling method [8] directly attempts to implement the ideal gates through imperfect ones numerically; and learning-based QEM attempts to implicitly learn the noise models and invert them through regression [19].

In general, we can expand Eq. (15) to get
\[
\rho_{\text{EM}} = \mathcal{E}_{\text{EM}}^{-1}(\rho_{\text{out}}) = [U_n \circ \cdots \circ U_2] \circ \{N_1^{-1} \circ \cdots \circ N_{n-1}^{-1} \circ U_{n-1}^\dagger \circ N_n^{-1} \circ U_n^\dagger \}
\]
\[
= U_{n-1} \circ (U_1^\dagger N_1^{-1})_{1\cdots n}(\rho_{\text{out}})
\]
(18)

where we defined
\[
U_{n-1} := U_n \circ \cdots \circ U_1
\]

and
\[
(U_1^\dagger N_1^{-1})_{1\cdots n} := U_1^\dagger \circ N_1^{-1} \circ \cdots \circ U_n^\dagger \circ N_n^{-1}.
\]
(19)

Here \(U_{n-1}\) is the ideal circuit sequence, and \((U_1^\dagger N_1^{-1})_{1\cdots n}\) is the channel that maps \(\rho_{\text{out}}\) back to \(\rho_{in}\). The composition of such channels first maps the experimental output state \(\rho_{\text{out}}\) back to the input state \(\rho_{in}\), then perform the ideal operations \(U_{n-1}\); this is illustrated by the blue arrows in Fig. 5.

A naive numerical implementation of the channel inverse requires simulating the quantum circuit \(U_{n-1}\), which is naturally expensive. Generally speaking, the computational complexity for computing Eq. (18) is even higher than classically simulating the ideal circuit. This is in addition to the cost of characterizing noise channels. Therefore, directly computing such an inverse channel is not efficient for the purpose of mitigating error, thus not very useful in practice. However, theoretically, comparing the error mitigated results with classically simulated ones can unfold how precise our knowledge about the device noise is. Moreover, the result from this “optimal method” upper bounds the performance of any error mitigation protocol.

Finally, we mention briefly that when only an approximate version of the ideal output is wanted, one may apply an effective channel method where only one effective recover map \(N_{\text{eff}}^{-1}\) is applied (either physically or numerically) at the end, i.e.,
\[
\rho_{\text{EM}} = N_{\text{eff}}^{-1}(\rho_{\text{out}}).
\]
(20)

The effective recover map \(N_{\text{eff}}^{-1}\) is one with tunable parameters. Normally, this method involves first estimating the parameters in \(N_{\text{eff}}\) (the effective noise channel) according to experimental data, then calculating and applying \(N_{\text{eff}}^{-1}\) to new experimental data to mitigate errors. Methods that fall into this category include decoherence compensation in NMR experiments, and depolarizing-model-based EM [23]. Recent work also considered continuous inversion through the Petz recovery map [21].

\section{B. Imperfect Characterization of Noise Channels}

We now study the effects of imperfectly characterized noise channels on the performance of QEM. It is generally acknowledged that characterizing noise models in a quantum system is highly resource demanding [22]. In many current error mitigation protocols, the noise channel is assumed to be particular models [23], such as a depolarizing channel \(D\). It is then natural to ask the question of how incorrectly characterized noise channels \(\{\mathcal{N}_i\}\) would affect the mitigation outcome. As mentioned before, we will assume all \(\{\mathcal{N}_i\}\)’s to be invertible in this subsection.

As shown in Fig. 5 while the ideal circuits and the experimental operations are CPTP maps, the channels \((U_1^\dagger N_1^{-1})_{n\cdots 1}\) and \((U_1^\dagger N_1^{-1})_{n\cdots 1}\) (defined in Eq. (19)) are not necessarily CPTP anymore. The difference \(N_i - \mathcal{N}_i\) between the estimations \(\mathcal{N}_i\) and the actual channels \(N_i\) upper bounds the result of EM, independent of how the inverses are achieved. And it only affects the difference between \((U_1^\dagger N_1^{-1})_{n\cdots 1}\) and \((U_1^\dagger N_1^{-1})_{n\cdots 1}\).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig5.png}
\caption{The schematic diagram of maps. The blue arrows indicate the channel \(\mathcal{E}_{\text{EM-ideal}}^{-1} := U_{n-1} \circ (U_1^\dagger N_1^{-1})_{1\cdots n}\) for ideal error mitigation, and the red arrows indicate the channel \(\mathcal{E}_{\text{EM}}^{-1} = U_{n-1} \circ (U_1^\dagger \mathcal{N}_1^{-1})_{1\cdots n}\) for actual error mitigation. The error between actual noise channels \(\mathcal{N}_i\) and estimations \(\mathcal{N}_i\) cause the difference between \((U_1^\dagger N_1^{-1})_{1\cdots n}\) and \((U_1^\dagger N_1^{-1})_{1\cdots n}\), which leads to a deviation in the mitigated result.}
\end{figure}

From the perspective of output states, the goal of EM is to bring the output states closer to the ideal. In terms of state fidelity, this is to ensure that
\[
F(\rho_{\text{EM}}; \rho_{\text{ideal}}) > F(\rho_{\text{EM}}; \rho_{\text{out}}),
\]
(21)

where \(F(\rho_1, \rho_2) := \text{tr}\left(\sqrt{\sqrt{\rho_1} \rho_2 \sqrt{\rho_1}}\right)\) is the fidelity between \(\rho_1\) and \(\rho_2\).

If the actual noise channels \(\{\mathcal{N}_i\}\) are invertible and the noise characterization is perfect \((\mathcal{N}_i = \mathcal{N}_i)\), theoretically the errors can be perfectly mitigated, with Eq. (21) naturally satisfied. Realistically, \(\mathcal{N}_i \neq \mathcal{N}_i\), which opens the gap between ideal output states \(\rho_{\text{out}}\) and error mit-
igated state $\rho^{\text{EM}}$. We would like to know how much will the imperfections in characterizing $\mathcal{N}$ worsen the fidelity. Let $\Delta N_i := \hat{N}_i - N_i$ and $\Delta N^{-1}_i := N^{-1}_i - \hat{N}^{-1}_i$. Note that $\Delta N_i \Delta N^{-1}_i + \Delta N^{-1}_i \Delta N_i = 0$, therefore $\Delta N_i$ and $\Delta N^{-1}_i$ are related to each other. We mainly use $\Delta N^{-1}_i$ in later discussion.

Fig. 5 shows that the errors $\{\Delta N^{-1}_i\}$ only affect $(\mathcal{U}^{\dagger} \mathcal{N}^{-1})_{1\cdots n}$ and $(\mathcal{U}^{\dagger} \mathcal{N})_{1\cdots n}$ in the EM inverse channels $\mathcal{E}_{\text{EM-ideal}}$ and $\mathcal{E}_{\text{EM}}$ respectively. The difference between $\rho_{\text{EM}}$ and $\rho_{\text{out}}^{\text{ideal}}$ is

$$\rho_{\text{EM}} - \rho_{\text{out}}^{\text{ideal}} = \mathcal{E}_{\text{EM}}(\rho_{\text{out}}) - \mathcal{U}_{1\cdots n}(\rho_{\text{in}}) = \mathcal{U}_{n-1} \circ \left[ (\mathcal{U}^{\dagger} \mathcal{N}^{-1})_{1\cdots n} - (\mathcal{U}^{\dagger} \mathcal{N})_{1\cdots n} \right] (\rho_{\text{out}}).$$

(22)

In the middle bracket in Eq. (22), the errors $\{\Delta N^{-1}_i\}$ scramble in the layers of unitaries $\mathcal{U}^{\dagger}_i$. Let $\Delta N := (\mathcal{U}^{\dagger} \mathcal{N})_{1\cdots n} - (\mathcal{U}^{\dagger} \mathcal{N}^{-1})_{1\cdots n}$, Eq. (22) becomes $\rho_{\text{EM}} - \rho_{\text{out}}^{\text{ideal}} = \mathcal{U}_{1\cdots n} \circ \Delta N(\rho_{\text{out}}^{\text{exp}})$. Take the first order estimation $\Delta N^{(1)}$ of $\Delta N$, where each term in $\Delta N^{(1)}$ only contain one of $\Delta N^{-1}_i$ (see Eq. (C2) in Appendix C for the explicit expression). The first order error between states is $\Delta \rho_{\text{EM}} := \mathcal{U}_{1\cdots n} \circ \Delta N^{(1)}(\rho_{\text{out}}^{\text{exp}})$. We then define $F(\rho_{\text{EM}}, \rho_{\text{EM}}^{\text{ideal}}, \rho_{\text{out}}^{\text{ideal}})$ to be the first order estimation $F^{(1)}(\rho_{\text{EM}}, \rho_{\text{out}}^{\text{ideal}})$ of the state fidelity $F(\rho_{\text{EM}}, \rho_{\text{out}}^{\text{ideal}})$. The following proposition gives a bound on this quantity.

**Proposition 2.** The first order estimation of fidelity between $\rho_{\text{EM}}$ and $\rho_{\text{out}}^{\text{ideal}}$ is

$$\left( 1 - \frac{1}{2} \sqrt{\text{d}C_{\text{exp}}} \| v(\Delta N^{(1)}) \| \right)^2 \leq F^{(1)}(\rho_{\text{EM}}, \rho_{\text{out}}^{\text{ideal}}) \leq \left( 1 - \frac{1}{4} \mathcal{U}_i \| v(\Delta N^{(1)}) v(\rho_{\text{out}}^{\text{exp}}) \| \right)^2,$$

(24)

where $C_{\text{exp}} := \| v(\mathcal{U}_{n-1}) \| \cdot \| v(\rho_{\text{out}}^{\text{exp}}) \|$ is an experiment-related constant, and $\mathcal{U}_i := \inf_{\| x \| = 1} \| v(\mathcal{U}_{n-1} x) \|$ is the lower Lipschitz constant of the ideal operations $\mathcal{U}_{n-1}$. The norm $\cdot \| \cdot $ is 2-norm for vectors and is the induced matrix norm for matrices.

We can see that $F^{(1)}(\rho_{\text{EM}}, \rho_{\text{out}}^{\text{ideal}})$ is bounded by $\Delta N^{(1)}$ and other experimental constants. Therefore, by bounding the errors $\{\Delta N^{-1}_i\}$ in channel estimation, one can constrain the fidelity by using Eq. (24). In fact, this result can be understood easily from the left side of Fig. 5. Details can be found in Appendix C.

If the task realized by the given circuit only concerns the expectation value of a set of observables $\{A_i\}$, then the goal of QEM can be simplified as recovering the ideal expectation value, $\text{Tr}(\rho_{\text{out}}^{\text{ideal}} A_i)$. As shown in Fig. 5, one would wish the error mitigated result to lie within the green area. Since we cannot perfectly characterize the noise models $N_i$, we would like to know the condition which guarantees that $\text{Tr}(\rho_{\text{EM}} A)$ lands in the green zone. We show in Appendix D that the following is a sufficient condition for such a goal.

**Proposition 3.** If the following condition Eq. (25) is satisfied, EM is guaranteed to improve the expectation value of any observable $A$ for any circuit $\mathcal{U}_{n-1}$.

$$\| v(\Delta N) \| \leq l_{\text{ideal-exp}},$$

(25)

where $l_{\text{ideal-exp}} := \inf_{\| x \| = 1} \| v(\mathcal{U}^{\dagger} \mathcal{N}^{(1)}_{1\cdots n} - \mathcal{U}^{\dagger}_{1\cdots n} x) \|$ is the lower Lipschitz constant of $v(\mathcal{U}^{\dagger} \mathcal{N}^{(1)}_{1\cdots n} - \mathcal{U}^{\dagger}_{1\cdots n})$.

In the above result, the channels $(\mathcal{U}^{\dagger} \mathcal{N}^{(1)}_{1\cdots n}$ and $\mathcal{U}^{\dagger}_{1\cdots n})$ maps $\rho_{\text{out}}^{\text{exp}}$ and $\rho_{\text{out}}^{\text{ideal}}$ back to $\rho_{\text{in}}$ respectively. This condition Eq. (25), in general, is asking that the $\Delta N$ (Eq. (23)) to be smaller than $(\mathcal{U}^{\dagger} \mathcal{N}^{(1)}_{1\cdots n} - \mathcal{U}^{\dagger}_{1\cdots n})$. It is straightforward to observe from the brackets in Fig. 5. Since this proposition is for any observables and any circuit, it will also work for quantum state fidelity.

Note that Eq. (25) is a stringent requirement. If $v(\mathcal{U}^{\dagger} \mathcal{N}^{(1)}_{1\cdots n} - \mathcal{U}^{\dagger}_{1\cdots n})$ has a nontrivial null space, then it will focus the noise channel estimation $\hat{N}_i$ to be perfect, i.e. $\hat{N}_i = N_i$ for all $i \in \{1, \cdots , n\}$. We do not introduce extra assumptions on circuits and noises while deriving this sufficient condition. Knowing more information about the circuit and noises can loosen the requirement.

Normally, certain noise models are assumed while identifying device noise. The assumptions made on noise models lead to savings in parameters and resources in characterization. However, the distance between the actually noise $N$ in the system and the model assumed cannot be arbitrarily close. It opens a gap between the ideal outcomes and error mitigated outcomes of the given circuit. In particular, if the error model is overly simplified, it can cause problems on EM performance.

We consider a simple example of a depth-1, single qubit quantum channel, where the actual noise $N$ is a Pauli Channel. Suppose one believes that the noise in the system is mainly depolarizing, and tries to use the depolarizing channel $D$ to approximate the actual noise. Af-
ter fitting the parameters in $D$, the (quasi-) inverse $D^+$ of the estimated $D$ is used to recover information (i.e. $D^+ \circ \mathcal{N}(\rho)$).

The Karus representation of $\mathcal{N}$ and $D$ are $\mathcal{N} : \{\sqrt{p_1}I, \sqrt{p_2}X, \sqrt{p_3}Y, \sqrt{(1 - p_1 - p_2 - p_3)}Z\}$ and $D : \{\sqrt{1 - \frac{3\lambda}{4}}I, \sqrt{\frac{1}{2}X}, \sqrt{\frac{1}{2}Y}, \sqrt{\frac{1}{2}Z}\}$. For a given set of $\{p_1, p_2, p_3\}$, the optimal $\lambda$ to minimize $\|\mathcal{N} - D\|_1$ varies according to different representations and different choices of norm $\| \cdot \|_1$. The symmetry on the parameters in $D$ makes it impossible to perfectly capture the noise $\mathcal{N}$ for $p_i$’s that do not have such a symmetry.

Note that, the two vectors, $\vec{n} := (\sqrt{p_1}, \sqrt{p_2}, \sqrt{p_3}, \sqrt{(1 - p_1 - p_2 - p_3)})$ and $\vec{d} := (\sqrt{1 - \frac{3\lambda}{4}}, \sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}})$, are also representations for $\mathcal{N}$ and $D$ respectively. Since $\vec{n}$ and $\vec{d}$ are normalized, minimizing the distance between $\mathcal{N}$ and $D$ is equivalent to maximizing $\vec{n} \cdot \vec{d}$, i.e.

$$\max_{\lambda \in [0, 1]} \left\{ \sqrt{p_1(1 - \frac{3\lambda}{4})} + \sqrt{p_2 + p_3 + \sqrt{(1 - p_1 - p_2 - p_3)}(\frac{\lambda}{4})} \right\}.$$

When $p_1 = \frac{1}{2}$ and $p_2 = p_3 = 0$, a channel will have a phase flip error with probability $\frac{1}{2}$ and will stay unchanged with probability $\frac{1}{2}$, corresponding to an optimal $\lambda_{\max}$ value of $\frac{1}{3}$. This $\lambda_{\max}$ bounds the distance between $\mathcal{N}$ and $D$ from above for this metric. Assume one fits the parameter $\lambda$ from experiments, and obtains the estimation that $\lambda = \frac{1}{3}$, the channel $\{\sqrt{3}I, \sqrt{\frac{1}{4}X}, \sqrt{\frac{1}{4}Y}, \sqrt{\frac{1}{4}Z\}}$ will be believed to be $\mathcal{N}$. Then $\mathcal{N}^{-1} = D^{-1}$ will be used to perform error mitigation. In Fig. 7 we can see that the actual channel $\mathcal{N}$ in fact preserves the expectation value of $Z$. Because of the incorrect assumptions on noise model, the mitigated results are actually worse (see the blue triangles in Fig. 7). Also note that, since $D^{-1}$ is non-CP, the outputs $D^{-1} \circ \mathcal{N}(\rho)$ are not valid quantum states anymore. In this case the fidelity function is not bounded below 1, thus is no longer a valid metric. We give further details in Appendix B.

While the above is a rather extreme example of channel mismatching, the message in this example is alerting. The gap between our knowledge of the noise and the actual noise in devices should also be considered while mitigating errors. Although we can lower bound the fidelity of the error mitigated state $\rho_{\text{REM}}$ and $\rho_{\text{outideal}}$ from Proposition 2, mitigating errors to improve the results still implies a competition between the experiment accuracy and the noise characterization (Proposition 3 and Fig. 5). In order to improve experimental readout from EM protocols, the increasing accuracy of the experiments demands better knowledge of device noise, which will translate into expensive procedures and sampling costs on noise char-

**IV. CONCLUSION**

In this paper, we discussed the different scenarios for using error mitigation and error correction. While error mitigation has demonstrated its use for most near-term quantum algorithms, it may destroy quantum resources like entanglement under some other scenarios, such as distributed quantum computation. It is thus an interesting open question to further study and classify the use cases of QEM protocols. We also show that the nature of noise processes limits the optimal performance of QEM, and analyzed three distinct cases where the noise is invertible and CPTP, invertible but not CPTP, and non-invertible. The first case is where both classical and quantum information is preserved; the second is where classical information can be perfectly restored but part of the quantum information is lost; and the third is when both classical and quantum information will be lost.

We next focused on the case of non-invertible noise. For non-invertible noise channels, at least part of the information carried in the quantum states $\rho$ will be inevitably erased, and its generalized inverse is not unique. In this case, we constructed an inverse-like channel, called quasi-inverse, to restore the information. We proved that the quasi-inverse is always trace preserving while the Moore-Penrose inverse may not. This has important implications for computing upper bounds on the performance of QEM protocols. Many previous works concern the CPTP inverses of quantum channels, mainly due to the fact that only CPTP maps can be physically implemented. However, any HPTP maps can be written as the linear combination of CPTP maps, thus opening the probability of physically implementing CPTP components by parts and then post-processing results from each branch together. Moreover, a widely considered recovery map, called the Petz recovery channel, is specified

![FIG. 7: Expectation value of $Z$ for 50 randomly generated states. The $x$-axis is the dummy label for these tested states.](image)
to a particular state. An HPTP channel that is opti-
mally recovering over all input states for a non-invertible
channel is rarely considered in previous literature. Our
results in Section II D provide a new point of view of
channel is more expensive than classically simulating the
ideal result. It limits the power of numerically imple-
menting such a inverse. We also need to be cautious
when implementing a non-CP inverse. If the device noise
is perfectly captured, the non-CP inverse would not cause
trouble. However, the gap between $\{N_i\}$ and $\{\tilde{N}_i\}$ may
lead to non-physical numerical outcomes. In this case,
commonly used metrics (such as fidelity) fail.

ACKNOWLEDGMENTS

N.C. thanks Maxwell Fitzsimmons for helpful discus-
sions. This research was undertaken thanks in part to
funding from the Government of Canada through the
Natural Sciences and Engineering Research Council of
Canada (NSERC).

[1] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon,
Joseph C. Bardin, Rami Barends, Rupak Biswas, Sergio
Boixo, Fernando G. S. L. Brandao, David A. Buell, Brian
Burkett, Yu Chen, Zijun Chen, Ben Chiaro, Roberto
Collins, William Courtney, Andrew Dunsworth, Edward
Farhi, Brooks Foxen, Austin Fowler, Craig Gidney,
Marissa Giustina, Rob Graff, Keith Guerin, Steve Habeg-
ger, Matthew P. Harrigan, Michael J. Hartmann, Alan
Ho, Markus Hoffmann, Trent Huang, Travis S. Humble,
Sergei V. Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri,
Kostyantyn Kechedzhi, Julian Kelly, Paul V. Klimov,
Sergey Knysh, Alexander Korotkov, Fedor Kostritsa,
David Landhuis, Mike Lindmark, Erik Lucero, Dmitry
Lyakh, Salvatore Mandrà, Jarrod R. McClean, Matthew
McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen,
Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew
Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Os-
by, Andre Petukhov, John C. Platt, Chris Quintana,
Eleanor G. Rieffel, Pedram Roushan, Nicholas C. Ru-in, Daniel Sank, Kevin J. Satzinger, Vadim Smelyan-
sky, Kevin J. Sung, Matthew D. Trevithick, Amit
Vainsencher, Benjamin Villalonga, Theodore White,
Z. Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven,
and John M. Martinis, “Quantum supremacy using a
programmable superconducting processor,” Nature
574, 505–510 (2019).

[2] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng
Chen, Li-Chao Peng, Yi-Han Luo, Qian Jin, Dian Wu,
Xing Ding, Yi Hu, Peng Hu, Xiao-Yan Yang, Wei-Jun
Zhang, Hao Li, Xuyuan Li, Xiao Jiang, Lin Gan, Guan-
gwen Yang, Lixing You, Zhen Wang, Li Li, Nai-Le Liu,
Chao-Yang Lu, and Jian-Wei Pan, “Quantum computa-
tional advantage using photons,” Science 370, 1460–1463
(2020).

[3] Kristan Temme, Sergey Bravyi, and Jay M Gambetta,
“Error Mitigation for Short-Depth Quantum Circuits,”
Physical Review Letters 119, 180509 (2017).

[4] Suguru Endo, Simon C Benjamin, and Ying Li, “Prac-
tical Quantum Error Mitigation for Near-Future Ap-
lications,” Physical Review X 8, 031027 (2018).

[5] Sam McArdle, Xiao Yuan, and Simon Benjamin, “Error-
Mitigated Digital Quantum Simulation,” Physical Re-
view Letters 122, 180501 (2019).

[6] Filip B. Maciejewski, Zoltán Zimborás, and Michal Os-
maniec, “Mitigation of readout noise in near-term quan-
tum devices by classical post-processing based on detec-
tor tomography,” Quantum 4, 257 (2020).

[7] Bálint Koczor, “Exponential Error Suppression for Near-
Term Quantum Devices,” Physical Review X 11, 031057
(2021).

[8] Laszlo Gyongyosi, Sandor Imre, and Hung Viet Nguyen,
“A Survey on Quantum Channel Capacities,” IEEE
Communications Surveys & Tutorials 20, 1149–1205
(2018).

[9] Peter W Shor, “Scheme for reducing decoherence in
quantum computer memory,” Physical Review A 52
R2493–R2496 (1995).

[10] Charles H Bennett, David P. DiVincenzo, John A Smolin,
and William K Wootters, “Mixed-state entanglement and
quantum error correction,” Physical Review A 54, 3824–
3851 (1996).

[11] Jiaqing Jiang, Kun Wang, and Xin Wang, “Physical
implementability of quantum maps and its application in
error mitigation,” (2020).

[12] John Preskill, “Lecture notes for physics 229: Quan-
tum information and computation,” California Institute
of Technology 16 (1998).

[13] Ashwin Nayak and Pranab Sen, “Invertible quantum
operations and perfect encryption of quantum states,”
arXiv preprint quant-ph/0605041 (2006).

[14] Michal Horodecki, “Entanglement measures,” Quantum
Information and Computation 1, 3–26 (2001).

[15] John Watrous, The theory of quantum information
(Cambridge university press, 2018).

[16] Eliakim H Moore, “On the reciprocal of the general alge-
braic matrix,” Bull. Am. Math. Soc. 26, 394–395 (1920).
Appendix A: The quasi-inverse of a TP map is also TP

Here we provide the proof for Theorem 2.

Proof. Let the Jordan block of eigenvalue λ in J be $J_\lambda$, where $J$ is defined in Eq. (10).

The inverse of a $k$ by $k$ Jordan block $J_\lambda$ (λ ≠ 0) of $v(N)$ is

$$J_\lambda^{-1} = (\lambda I + N)^{-1} = \lambda^{-1}(I - \lambda^{-1}N + \cdots \lambda^{-(k-1)}N^{(k-1)})$$

$$= \lambda^{-1} \left( \sum_{i=0}^{k-1} (-\lambda^{-1}N)^i \right) =: J'_\lambda$$

where $N$ is the $k$ by $k$ nilpotent matrix, $\lambda$ is the eigenvalue.

By the construction (Eq. (11)), we have

$$v(N^+Q = QJ'),$$

(A1)

where $Q$ contains the eigenvectors and generalized eigenvectors of $v(N)$. For the particular block that we concern, the corresponding eigenvector and generalized eigenvector is

$$Q = (\cdots \; v^{g_0} \; v^{g_1} \; \cdots \; v^{g_k-1} \; \cdots)$$

Let $e_i$, $i \in \{0, \cdots, k-1\}$ be the standard basis vectors for this block. Acting $e_j$ on both sides of Eq. (A1), the left hand side is

$$v(N^+Qe_j = v(N^+v^{g_j})$$

and the right hand side is

$$QJ'e_j = Q\lambda^{-1} \left( \sum_{i=0}^{k-1} (-\lambda)^i N^i e_j \right).$$

It is easy to show that $N^i e_j = e_{j-i}$ for $j \geq i$, and $N^i e_j = 0 \cdot e_j$ for $j < i$. Therefore

$$QJ'e_j = Q\lambda^{-1} \left( \sum_{i=0}^{j} (-\lambda)^i e_{j-i} \right) = \lambda^{-1} \left( \sum_{i=0}^{j} (-\lambda)^i v^{g_{j-i}} \right).$$

Thus

$$v(N^+v^{g_j}) = \lambda^{-1} \left( \sum_{i=0}^{j} (-\lambda)^i v^{g_{j-i}} \right).$$

(A2)

For $\lambda$ not equals to 1 and 0, taking trace on both sides of Eq. (A2),

$$sTr \left[ v(N^+v^{g_j}) \right] = sTr \left[ \lambda^{-1} \left( \sum_{i=0}^{j} (-\lambda)^i v^{g_{j-i}} \right) \right].$$
From Lemma 1, we know that \( s\text{Tr} [v^{g_j}] = 0 \), so the right hand side is also 0. That is, \( s\text{Tr} [v(N^+)v^{g_j}] = s\text{Tr} [v^{g_j}] \) holds for every \( j \in \{0, \cdots, k-1\} \).

When \( \lambda = 1 \), according to Lemma 2, we have the same results except for \( j = k - 1 \). Now we check \( j = k - 1 \) case for \( \lambda = 1 \),

\[
v(N^+)v^{g_{k-1}} = \left( \sum_{i=0}^{k-1} (-1)^i v^{g_{k-1-i}} \right).
\]

Taking trace on both sides,

\[
s\text{Tr} [v(N^+)v^{g_{k-1}}] = s\text{Tr} \left[ \left( \sum_{i=0}^{k-1} (-1)^i v^{g_{k-1-i}} \right) \right]
\]

From Lemma 2 we know that all the eigenvector and generalized eigenvectors have trace zero except for the \((k-1)\)th one. We have

\[
s\text{Tr} [v(N^+)v^{g_{k-1}}] = s\text{Tr} [v^{g_{k-1}}]
\]

Finally, when \( \lambda = 0 \), \( J'_\lambda = 0 \), where 0\(_k\) is the \( k \) by \( k \) zero matrix. Thus, \( v(N^+)v^{g_j} = 0 \cdot e_j \). From Lemma 1 the trace of both sides are zero.

Now we have proved that the trace of all columns (eigenvectors and generalized eigenvectors) in \( Q \) are unchanged under the action of \( v(N^+) \).

Since \( Q \) is invertible, any \( v(\rho) \) can be expended by columns \( v^\lambda_j \) in \( Q \). And we have

\[
s\text{Tr} \left[ v(N^+) (\sum_{ij} a_{ij} v^\lambda_j) \right] = s\text{Tr} \left[ \sum_{ij} a_{ij} v(N^+) (v^\lambda_j) \right] = s\text{Tr} \left[ \sum_{ij} a_{ij} v^\lambda_j \right].
\]

Hence, \( N^+ \) is trace preserving.

**Appendix B: Survived States**

![Graphs showing fidelity results](image)

**FIG. 8:** (a) Non-recovered and recovered fidelities for \( \rho_\theta \) where \( \rho_\theta = \frac{1}{2} \left( \begin{array}{cc} 1 & e^{i\theta} \\ e^{-i\theta} & 1 \end{array} \right) \); (b) results for 50 randomly generated states: the green horizontal line is the average fidelity for states that recovered by the quasi-inverse channel \( N^+ \); the blue line and red line are for using Moore-Penrose inverse \( N^P \) and not applying recovery respectively.
For the composition of a non-invertible channel and it quasi-inverse defined in Eq. (11), \( \mathbf{v}(\mathcal{N}^+)\mathbf{v}(\mathcal{N}) \), the kernel dimension is at least one. For such a channel, part of the information inevitably leaks out of the systems.

Consider the case that the dimension of the kernel is one. Denote the vector in the kernel as \( \tilde{k} \). A density matrix \( \rho \) which does not effect by such channel \( \mathbf{v}(\mathcal{N}^+)\mathbf{v}(\mathcal{N}) \) satisfies

\[
[\mathbf{v}(\mathcal{N}^+)\mathbf{v}(\mathcal{N})] \mathbf{v}(\rho) = \mathbf{v}(\rho),
\]

i.e. \( \mathbf{v}(\rho) \cdot \tilde{k} = 0 \).

In Example 2, \( \tilde{k} = [\frac{\sqrt{2}}{2}, 0, 0, -\frac{\sqrt{2}}{2}] \). The quantum states that have a chance to perfectly survive this composed channel are

\[
\rho = \begin{bmatrix}
\frac{1}{2} & b+ic \\
\frac{1}{2} & b-ic
\end{bmatrix}.
\]

It is the disc of \( \text{tr}(\rho Z) = 0 \) in Bloch Sphere. Now we can see the effect of the composed channel in Example 2. Any 1-qubit state \( \rho \) can be decomposed to two parts \( \rho = \lambda_1 \rho_{xy} + \lambda_2 \rho_z \). The channel erases \( \rho_z \) and left \( \rho_{xy} \) untouched.

Let \( F_+(\rho) := F_2(\mathcal{N}^+ \circ \mathcal{N})(\rho), \) \( F_+\mathbf{v}(\rho) := F_2(\mathcal{N}^+ \circ \mathcal{N})(\rho, \mathbf{v}) \) and \( F_{\text{out}}(\rho) := F_2[\mathcal{N}(\rho), \rho] \), where \( F(\rho_1, \rho_2) = \text{tr}(\sqrt{\rho_1 \rho_2 \sqrt{\rho_1}}) \) is the fidelity between \( \rho_1 \) and \( \rho_2 \). In Fig. 3a, we can see that the Moore-Penrose inverse channel \( \mathcal{N}_p^+ \) does not always preserve \( \rho_\theta \), while the quasi-inverse \( \mathcal{N}_p \) perfectly recovered such states. Fig. 8b shows imperfect knowledge about noise channels on fidelity. In this example, the quasi-inverse channel \( \mathcal{N}_p^+ \) does not decrease the fidelity \( (F_+ \geq F_{\text{out}}) \), and the Moore-Penrose inverse \( \mathcal{N}_p^+ \) sometimes make recovered states even further away from the original states (Fig. 1a partly explained the reason). Although, in certain cases, \( \mathcal{N}_p^+ \) has better performance than \( \mathcal{N}_p \), the average fidelity \( F_+ \) is greater than \( F_p \).

**Appendix C: The effect of imperfect knowledge about noise channels on fidelity**

From the main text, we know that

\[
\rho_{\text{EM}} = \mathcal{U}_{n-1} \circ (\mathcal{U}^\dagger \mathcal{N}^{-1})_{1\ldots n}(\rho_{\text{out}}^{\text{exp}}),
\]

\[
\rho_{\text{ideal}}^{\text{out}} = \mathcal{U}_{n-1} \circ (\mathcal{U}^\dagger \mathcal{N}^{-1})_{1\ldots n}(\rho_{\text{out}}^{\text{exp}}),
\]

where \( \mathcal{U}_{n-1} := \mathcal{U}_n \circ \cdots \circ \mathcal{U}_1 \) is the ideal set of circuits. (See Fig. 5)

Imperfect knowledge about \( \mathcal{N}_i \) leads to imperfect inverse \( \mathcal{N}^{-1}_i \). Let \( \mathcal{N}^{-1}_i = \mathcal{N}^{-1}_i + \Delta \mathcal{N}_i^{-1} \), where \( \mathcal{N}^{-1}_i \) is the perfect inverse of \( \mathcal{N}_i \).

Let \( \Delta \rho_{\text{EM}} := \rho_{\text{EM}} - \rho_{\text{ideal}}^{\text{out}} \), then

\[
\Delta \rho_{\text{EM}} = \mathcal{U}_{n-1} \circ \left[(\mathcal{U}^\dagger \mathcal{N}^{-1})_{1\ldots n} - (\mathcal{U}^\dagger \mathcal{N}^{-1})_{1\ldots n}\right](\rho_{\text{out}}^{\text{exp}}) = \mathcal{U}_{n-1} \circ \Delta \mathcal{N}(\rho_{\text{out}}^{\text{exp}}). \tag{C1}
\]

If we only consider the first order approximation \( \Delta \mathcal{N}^{(1)} \) of \( \Delta \mathcal{N} \), the first order correction term \( \Delta \rho_{\text{EM}}^{(1)} \) would be

\[
\Delta \rho_{\text{EM}}^{(1)} = \mathcal{U}_{n-1} \circ \Delta \mathcal{N}^{(1)} \circ \mathcal{U}_{\text{exp}}(\rho_{\text{in}}) \quad \tag{C2}
\]

\[
= \mathcal{U}_{n-1} \circ \left( \sum_{i=1}^{n} \mathcal{U}_i^\dagger \circ \mathcal{N}^{-1}_i \cdots \mathcal{U}_i^\dagger \circ \Delta \mathcal{N}^{-1}_i \cdots \mathcal{U}_i^\dagger \circ \mathcal{N}^{-1}_i \right) \circ \mathcal{U}_{\text{exp}}(\rho_{\text{in}}), \tag{C3}
\]

where \( \mathcal{U}_{\text{exp}} := \mathcal{N}_n \circ \mathcal{U}_n \circ \cdots \circ \mathcal{N}_1 \circ \mathcal{U}_1 \) is the actual experimental operator.

Then \( \rho_{\text{EM}} \approx \rho_{\text{ideal}}^{\text{out}} + \Delta \rho_{\text{EM}}^{(1)} \). The fidelity between \( \rho_{\text{EM}} \) and \( \rho_{\text{ideal}}^{\text{out}} \) is approximately

\[
F(\rho_{\text{EM}}, \rho_{\text{ideal}}^{\text{out}}) \approx F(1)(\rho_{\text{EM}}, \rho_{\text{ideal}}^{\text{out}}) := F(\rho_{\text{ideal}}^{\text{out}} + \Delta \rho_{\text{EM}}^{(1)}, \rho_{\text{ideal}}^{\text{out}}) = \| \rho_{\text{ideal}}^{\text{out}} + \Delta \rho_{\text{EM}}^{(1)} \|_{\text{tr}},
\]

it is the first order approximation of \( F(\rho_{\text{EM}}, \rho_{\text{ideal}}^{\text{out}}) \).

By the Fuchs–van de Graaf inequalities,

\[
[1 - D(\Delta \rho_{\text{EM}}^{(1)})]^2 \leq F(1)(\rho_{\text{EM}}, \rho_{\text{ideal}}^{\text{out}}) \leq 1 - D^2(\Delta \rho_{\text{EM}}^{(1)}), \tag{C4}
\]
where \( D(\cdot) := \frac{1}{2} \| \cdot \|_1 \) is the trace distance, and \( \| \cdot \|_F \) is the trace norm. It is also known that \( \| A \|_F \leq \| A \|_{tr} \leq \sqrt{\| A \| F} \), where \( \| A \|_F \) is the Frobenius norm which equals to \( \| v(A) \|_F \). The norm \( \| \cdot \| \) is the 2-norm.

\[
\| \Delta \rho_{EM}^{(i)} \|_F = \| v(\Delta \rho_{EM}^{(i)}) \| = \| v(U_{n-1}) v(\Delta N^{(i)}) v(\rho_{in}) \| = \| v(U_{n-1}) v(\Delta N^{(i)}) v(\rho_{out}) \|
\]

\[
l_U \cdot \| v(\Delta N^{(i)}) v(\rho_{out}) \| \leq \| \Delta \rho_{EM}^{(i)} \|_F \leq \| v(U_{n-1}) \| \cdot \| v(\Delta N^{(i)}) \| \cdot \| v(\rho_{out}) \|
\]

(\ref{eq:bound})

where \( l_U := \inf_{\| x \|=1} \| v(U_{n-1}) x \| \) is the lower Lipschitz constant of the superoperator of the ideal circuits. Notice that, on the right hand of Eq. \( (C5) \), \( \| v(U_{n-1}) \| \) and \( \| v(\rho_{out}) \| \) are known for a given experiment. Denote \( \| v(U_{n-1}) \| \cdot \| v(\rho_{out}) \| \) as \( C_{exp} \). From Eq. \( (C4) \), we know the fidelity between the mitigated state and the ideal state is bounded

\[
\left( 1 - \frac{1}{2} \sqrt{\| C_{exp} \| v(\Delta N^{(i)}) \|^2} \right)^2 \leq F^{(i)}(\rho_{EM}, \rho_{out}^{ideal}) \leq 1 - \frac{1}{4} \left( l_U \cdot \| v(\Delta N^{(i)}) v(\rho_{out}) \| \right)^2.
\]

In addition, the norm of \( v(\Delta N^{(i)}) \) satisfies that

\[
\| v(\Delta N^{(i)}) \| \leq \prod_{k=1}^{n} \| v(U_i^k) \| \cdot \sum_{i=1}^{n} \| v(\Delta N_i^{-1}) \| \cdot \prod_{j \in \{1, \ldots, n\} \setminus \{i\}} \| v(\tilde{N}_j^{-1}) \|,
\]

(\ref{eq:norm})

where \( U_i^k \) and \( \tilde{N}_j^{-1} \) are known for a given EM tasks. The error on each inverse \( \Delta N_i^{-1} \) can be exposed at the lower bound of the fidelity. And by counting the sampling cost of getting \( \Delta N^{-1} \), one can bound the sampling fidelity from the sampling cost.

**Appendix D: A sufficient condition on improving expectation values**

The goal of error mitigation on the expectation value of an observables \( A \) is

\[
\| \text{Tr}(\rho_{EM} A) - \text{Tr}(\rho_{out}^{ideal} A) \| \leq \| \text{Tr}(\rho_{out} A) - \text{Tr}(\rho_{out}^{exp} A) \|.
\]

(\ref{eq:d1})

The left hand side of Eq. \( (D1) \) is

\[
|\text{Tr}[(\rho_{EM} - \rho_{out}^{ideal}) A]| = |\text{Tr}(\Delta \rho_{EM} A)| = |\text{Tr}(U_{n-1} \circ \Delta N(\rho_{out}^{exp}) \cdot A)| = \left| \left< v(U_{n-1}) v(A^\dagger), v(\Delta N) v(\rho_{out}^{exp}) \right> \right|.
\]

(\ref{eq:d2})

The right hand side of Eq. \( (D1) \) equals to

\[
|\text{Tr}[\rho_{out}^{ideal} - \rho_{out}^{exp} A]| = |\text{Tr}[U_{n-1} \circ (U^1 N^{-1})_{1 \ldots n} - I](\rho_{out}^{exp}) \cdot A)| = |\text{Tr}[U_{n-1} \circ [(U^1 N^{-1})_{1 \ldots n} - U^1_{1 \ldots n}](\rho_{out}^{exp}) \cdot A]| = \left| \left< v(U_{n-1}) v(A^\dagger), v((U^1 N^{-1})_{1 \ldots n} - U^1_{1 \ldots n}) v(\rho_{out}^{exp}) \right> \right|.
\]

(\ref{eq:d3})

It is difficult to draw conclusions directly from Eq. \( (D2) \) and Eq. \( (D3) \) since \( v(\Delta N) \) and \( v(U^1 N^{-1}_{1 \ldots n} - U^1_{1 \ldots n}) \) can be arbitrary. However,

\[
\left| \left< v(U_{n-1}) v(A^\dagger), v(\Delta N) v(\rho_{out}^{exp}) \right> \right| \leq \left| v(U_{n-1}) v(A^\dagger) \right| \| v(\rho_{out}^{exp}) \| \| v(\Delta N) \|,
\]

\[
\left| \left< v(U_{n-1}) v(A^\dagger), v(U^1 N^{-1}_{1 \ldots n} - U^1_{1 \ldots n}) v(\rho_{out}^{exp}) \right> \right| \geq \left| v(U_{n-1}) v(A^\dagger) \right| \| v(\rho_{out}^{exp}) \| \inf_{\| x \|=1} \left| v((U^1 N^{-1}_{1 \ldots n} - U^1_{1 \ldots n}) x) \right|.
\]

Therefore, if

\[
\| v(\Delta N) \| \leq \inf_{\| x \|=1} \left| v((U^1 N^{-1}_{1 \ldots n} - U^1_{1 \ldots n}) x) \right|,
\]

Eq. \( (D1) \) is guaranteed. That means the EM process will improve the expectation value for any observable \( A \) and any desired circuit \( U_{n-1} \) when the above is satisfied. It is a harsh requirement. If \( v((U^1 N^{-1}_{1 \ldots n} - U^1_{1 \ldots n}) \) has a nontrivial null space, than it will focus the noise channel estimation \( \tilde{N}_i \) to be perfect, i.e. \( \tilde{N}_i = N_i \) for \( \forall i \in \{1, \ldots, n\} \).
Appendix E: Examples of oversimplified noise channels

The Karus representation of $\mathcal{N}$ and $\mathcal{D}$ are

$$\mathcal{N} : \left\{ \sqrt{p_1}I, \sqrt{p_2}X, \sqrt{p_3}Y, \sqrt{(1-p_1-p_2-p_3)}Z \right\} ;$$

$$\mathcal{D} : \left\{ \sqrt{1-\frac{3\lambda}{4}}I, \sqrt{\frac{\lambda}{4}}X, \sqrt{\frac{\lambda}{4}}Y, \sqrt{\frac{\lambda}{4}}Z \right\} .$$

For a given set of $\{p_1, p_2, p_3\}$, what is the optimal $\lambda$ to minimize $\|\mathcal{N} - \mathcal{D}\|$ of a chosen norm $\|\cdot\|$? One approach is that we can write down a matrix representation of $\mathcal{N}$ and $\mathcal{D}$, then solve $\lambda$ by minimizing $\|\mathcal{N} - \mathcal{D}\|$ (use a particular norm $\|\cdot\|$). For different representations and different norms, the optimization outcome could be different. The optimal $\lambda$ will bound the distance $\|\mathcal{N} - \mathcal{D}\|$ from below for any possible experimental implementation for this particular norm $\|\cdot\|$.

As mentioned in the main text, the two vectors, $\vec{n} := (\sqrt{p_1}, \sqrt{p_2}, \sqrt{p_3}, \sqrt{(1-p_1-p_2-p_3)})$ and $\vec{d} := \left(\sqrt{1-\frac{3\lambda}{4}}, \sqrt{\frac{\lambda}{4}}, \sqrt{\frac{\lambda}{4}}, \sqrt{\frac{\lambda}{4}}\right)$, are also representations for $\mathcal{N}$ and $\mathcal{D}$ respectively. Since $\vec{n}$ and $\vec{d}$ are normalized, minimizing the distance between $\mathcal{N}$ and $\mathcal{D}$ is equivalent to maximizing $\vec{n} \cdot \vec{d}$, i.e.

$$\max_{\lambda \in [0,1]} \left\{ \sqrt{p_1(1-\frac{3\lambda}{4})} + \sqrt{\frac{p_2 + \sqrt{p_3 + (1-p_1-p_2-p_3)^2}p_1}{4}} \right\} .$$

This can be solved by taking the derivative of the expression, and setting it to be zero. The result is

$$\lambda_{\text{max}} = \frac{\left[ \sqrt{p_2} + \sqrt{p_3} + \sqrt{(1-p_1-p_2-p_3)^2}p_1 \right]}{\frac{9}{4}p_1^2 + \frac{3}{4}p_1(\sqrt{p_2} + \sqrt{p_3} + \sqrt{(1-p_1-p_2-p_3)^2})^2} = 0 \quad \text{or} \quad \lambda = 1, \quad \text{or} \quad \lambda = 0. \quad (E1)$$

The superoperators of $\mathcal{N}$ and $\mathcal{D}$ are

$$\mathbf{v}(\mathcal{N}) = p_1 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + p_2 \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} + p_3 \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} + (1-p_1-p_2-p_3) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{v}(\mathcal{D}) = \left(1-\frac{3\lambda}{4}\right) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \frac{\lambda}{4} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} + \frac{\lambda}{4} \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} + \frac{\lambda}{4} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} .$$

Even with the optimal $\lambda$ in Eq. (E1), when $p_2, p_3$ and $1-p_2-p_3$ are not equal to each other, the distance between $\mathcal{N}$ and $\mathcal{D}$ is not zero.

The following are two examples of different $\{p_i\}$ sets.

1. When $p_1 = p_3 = 0$ and $p_2 = 1$, the optimal $\lambda_{\text{max}}$ is 1. Therefore

$$\mathbf{v}(\mathcal{N}) = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{v}(\mathcal{D}) = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} .$$

In this case, the estimated $\mathcal{D}$ is not a invertible channel while $\mathcal{N}$ is invertible. The inverse $\mathcal{D}$ will definitely worsen the outcomes.
(a) Expectation values of Pauli $Y$ for 50 randomly generated states.

(b) Fidelities of 50 randomly generated states.

**FIG. 9:** The $x$-axis is a dummy label for the tested states. Because the channel $D^{-1} \circ \mathcal{N}$ is not physical (not CP), the output $D^{-1} \circ \mathcal{N}(\rho)$ are not eligible quantum states. In this case the fidelity is no longer a good metric for distinguishing two “states”.

2. When $p_1 = \frac{1}{2}$ and $p_2 = p_3 = 0$, according to Eq. (E1), $\lambda_{\text{max}} = \frac{1}{3}$.

$$v(\mathcal{N}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad v(D) = \frac{1}{6} \begin{pmatrix} 5 & 0 & 0 & 1 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 1 & 0 & 0 & 5 \end{pmatrix}$$

The inverse of $D$ is

$$v(D^{-1}) = \frac{1}{4} \begin{pmatrix} 5 & 0 & 0 & -1 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ -1 & 0 & 0 & 5 \end{pmatrix}.$$  

Therefore,

$$v(\mathcal{N}^{\text{re}}) := v(D^{-1})v(\mathcal{N}) = \frac{1}{4} \begin{pmatrix} 5 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 5 \end{pmatrix}.$$  

This resulting channel $\mathcal{N}^{\text{re}}$ has eigenvalues \{\frac{3}{2}, 1, 0, 0\}, which will worsen the outcome. In Fig. 9 we tested 50 randomly generated quantum state $\rho$ for this example. Fig. 7 shows the expectation value of Pauli $Y$ for these 50 states. The information of the expectation value $\text{Tr}(Y\rho)$ is erased by the noise channel $\mathcal{N}$ cannot be helped by $D^{-1}$. For $\text{Tr}(Z\rho)$ in Fig. 7 the channel $D^{-1}$ has made the outcome worse. Fig. 9b shows the fidelities $F(\mathcal{N}(\rho), \rho)$ and $F(\mathcal{N}^{\text{re}}(\rho), \rho)$. Since $D^{-1}$ is non-CP, the outputs $D^{-1} \circ \mathcal{N}(\rho)$ are not valid quantum states anymore. The fidelity function does not always smaller than 1, thus is no longer a good metric. This explains why the recovery $D^{-1}$ does not improve any expectation value but seems to have higher fidelities.