Supplemental Material: Quantum state-agnostic work extraction (almost) without dissipation

Josep Lumbreras,[1,2,*] Ruo Cheng Huang,[1,†] Yanglin Hu (胡杨林),[3,‡] Mile Gu,[3,1,4,§] and Marco Tomamichel[3,4,5,¶]

[1] *Nanyang Quantum Hub, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore*
[2] *Centre for Quantum Technologies, Nanyang Technological University, Singapore*
[3] *Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore*
[4] *MajuLab, CNRS-UNS-NUS-NTU International Joint Research Unit, UMI 3654, 117543, Singapore*
[5] *Department of Electrical and Computer Engineering, National University of Singapore*

(Dated: April 14, 2026)

CONTENTS

* josep.lumbreras@u.nus.edu
† ruocheng001@e.ntu.edu.sg
‡ yanglin.hu@u.nus.edu
§ mgu@quantumcomplexity.org
¶ marco.tomamichel@nus.edu.sg

I. LEARNING PURE QUANTUM STATES (ALMOST) WITHOUT REGRET

In this Section we review the quantum state tomography algorithm presented in [7]. They considered sequential access to an unknown pure quantum state $|\psi\rangle$ and at each time step $k \in \{1, \dots, N\}$, a reward measurement is performed on the direction $|\psi_k\rangle$. Formally the reward measurement is described by a rank-1 two-outcome POVM $\{\psi_k, \psi_k^{\perp}\}$ where $\psi_k = |\psi_k\rangle\langle\psi_k|$ corresponds to $R_k = 1$ and $\psi_k^{\perp} = \mathbb{I} - \psi_k$ corresponds to $R_k = 0$. The observed reward R_k is distributed accordingly to Born's rule, i.e

$$\Pr(R_k = r_k) = \begin{cases} |\langle\psi_k|\psi\rangle|^2, & r_k = 1, \\ 1 - |\langle\psi_k|\psi\rangle|^2, & r_k = 0. \end{cases} \tag{1}$$

The goal of this work was to design an algorithm that uses measurements that minimally disturb the unknown state $|\psi\rangle$. They used as a figure of merit for this task the regret which is defined as

$$\mathrm{Rgrt}(N) = \sum_{k=1}^{N} (1 - |\langle\psi_k|\psi\rangle|^2), \tag{2}$$

and is the cumulative sum of infidelities between the unknown state $|\psi\rangle$ and the selected direction ψ_k. The algorithm is formulated in terms of the Bloch vector which means that at each time step k, the algorithm uses the previous and current information of the reward $\{r_1, \dots, r_k\}$ and corresponding directions $\{|\psi_1\rangle, \dots, |\psi_k\rangle\}$ to output a vector $a_{k+1} \in \mathbb{R}^3$ that is normalized $\|a_{k+1}\| = 1$ and it is linked to the next reward measurement described by the two-outcome POVM $\{\psi_{k+1}, \psi_{k+1}^{\perp}\}$ (or the direction $|\psi_{k+1}\rangle$) as

$$\psi_k = \frac{1}{2}(\mathbb{I} + a_k \cdot \sigma), \tag{3}$$

where $\sigma = (\sigma_x, \sigma_y, \sigma_z)$ are the Pauli matrices (or any other basis of 2×2 Hermitian matrices). The pseudo-code for how the algorithm updates the reward measurements can be found in Algorithm 1 (LinUCB Vanishing Variance Noise) and now we describe all the involved quantities. The algorithm presented in [7] was formulated for a more general case but here we focus for the specific qubit case.

Algorithm 1: LinUCB-VVN

Require: $\lambda_0 \in \mathbb{R}_{>0}$, $t \in \mathbb{N}$
Set initial design matrix $V_0 \leftarrow \lambda_0 \mathbb{I}$.
Set initial estimate of variance $\hat{\sigma}_1^2 \leftarrow 1$.
Set initial Bloch vectors $a_{1,1} = (\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})^{\intercal}$, $a_{1,2} = (-\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})^{\intercal}$, $a_{1,3} = (0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^{\intercal}$, $a_{1,4} = (0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^{\intercal}$.
for $k^* = 1, 2, \dots$ **do**
  Optimistic action selection
  for $i = 1, 2, 3, 4$ **do**
    Perform t independent measurements for each $a_{k^*,i}$
    for $j = 1, ..., t$ **do**
      Measure the unknown $|\psi\rangle$ in the Bloch vector directions given by $a_{k^*,i}$ and receive outcomes $r_{k^*,i,j}$
  Update design matrix
  $V_{k^*} \leftarrow V_{k^*-1} + \frac{1}{\hat{\sigma}_{k^*}^2} \sum_{i=1}^{4} a_{k^*,i} a_{k^*,i}^{\intercal}$
  Update LSE for each subsample
  for $j = 1, 2, ..., t$: **do**
    $\widetilde{\theta}_{k^*,j}^{\mathrm{w}} \leftarrow V_{k^*}^{-1} \sum_{s=1}^{k^*} \frac{1}{\hat{\sigma}_{k^*}^2} \sum_{i=1}^{4} a_{s,i}(2r_{s,i,j} - 1)$
  Update Bloch vectors for estimates
  Compute $\theta_{k^*}^{\mathrm{wMOM}}$ using $\{\widetilde{\theta}_{k^*,j}^{\mathrm{w}}\}_{j=1,\dots t}$ according to Eq. (6)
  Update Bloch vectors for measurements
  Select Bloch vectors $a_{k^*+1,i}$ using $\theta_{k^*}^{\mathrm{wMOM}}$ according to Eq. (9)
  Update estimator of variance for $a_{k^*+1,i}$
  $\hat{\sigma}_{k^*+1}^2 \leftarrow \frac{2\zeta}{\sqrt{\lambda_{\max}(V_{k^*})}}$

The algorithm described in Algorithm 1 takes as input a parameter t that is used to control the success probability and updates the measurements in stages. Each stage k^* contains 4 groups labeled by $i \in \{1,2,3,4\}$, and each group i contains t rounds labeled by $j \in \{1,\dots,t\}$. Each round k can also be labeled by its stage k^*, group i and group index j where $k = 4t(k^*-1) + 4(j-1) + i$. At each time stage k^*, the algorithm takes 4 Bloch vectors $\{a_{k^*,i}\}_{i=1,\dots,4}$ and performs t independent reward measurements defined by $a_{k^*,i}$ in each group i. The updates for $a_{k^*,i}$ can be defined recursively given the previous information of outcomes and measurements. At each stage k^* the algorithm computes the following t weighted least-square estimators of the Bloch vector of the unknown state $|\psi\rangle$ as

$$\widetilde{\theta}^{\mathrm{w}}_{k^*,j} = (V^{\mathrm{w}}_{k^*})^{-1} \sum_{s=1}^{k^*} \frac{1}{\hat{\sigma}^2_s} \sum_{i=1}^{4} a_{s,i}(2r_{s,i,j} - 1), \tag{4}$$

where $r_{s,i,j} \in \{0,1\}$ is the outcome of the reward measurement defined by $a_{s,i}$ in round (s,i,j), $\hat{\sigma}^2_s$ is the estimate of reward variances in stage s and V_{k^*} is the design matrix defined as

$$V^{\mathrm{w}}_{k^*} = V^{\mathrm{w}}_{k^*-1} + \frac{1}{\hat{\sigma}^2_{k^*}} \sum_{i=1}^{4} a_{k^*,i} a^{\intercal}_{k^*,i}, \tag{5}$$

where the initial design matrix is setup to be $V_0 = \lambda_0 \mathbb{I}$ for some parameter $\lambda_0 > 0$. Then the algorithm does an extra step that outputs a normalized version $\theta^{\mathrm{wMoM}}_{k^*}$ of median of means with the k weighted least squares estimator $\widetilde{\theta}^{\mathrm{wMoM}}_{k^*}$ in the following way

$$\theta^{\mathrm{wMoM}}_{k^*} = \frac{\widetilde{\theta}^{\mathrm{wMoM}}_{k^*}}{\|\widetilde{\theta}^{\mathrm{wMoM}}_{k^*}\|_2} \tag{6}$$

where

$$\widetilde{\theta}^{\mathrm{wMoM}}_{k^*} := \tilde{\theta}_{k^*,j^*}, \tag{7}$$

where

$$j^* = \operatorname*{argmin}_{j \in [t]} \operatorname{median}\{\|\tilde{\theta}_{k^*,j} - \tilde{\theta}_{k^*,j'}\|_{V_{k^*}} : j' \in [t]/j\}, \tag{8}$$

and $\|x\|^2_{V_{k^*}} = \langle x, V_{k^*} x\rangle$ is the weighted norm given by V_{k^*} and $\langle \cdot,\cdot\rangle$ is the standard inner product between real vectors. Then the update of the measurements is defined as

$$a_{k^*+1,i} = \frac{\widetilde{a}_{k^*+1,i}}{\|\widetilde{a}_{k^*+1,i}\|_2}, \tag{9}$$

where

$$\tilde{a}_{k^*+1,i} = \theta^{\mathrm{wMoM}}_{k^*} - \frac{(-1)^i}{\sqrt{\lambda_{\min}(V_{k^*})}} v_{k^*,\lceil \frac{i}{2} \rceil} \tag{10}$$

where $\{v_{k^*,0}, v_{k^*,1}\}$ are the two eigenvectors of V_{k^*} with smallest eigenvalues, $\lambda_{\min}(V_{k^*})$ is the smallest eigenvalue of V_{k^*} and the estimate of reward variances $\hat{\sigma}^2_{k^*}$ is set to

$$\hat{\sigma}^2_{k^*} = \frac{2\zeta}{\sqrt{\lambda_{\max}(V_{k^*})}}, \tag{11}$$

where $\lambda_{\max}(V_{k^*})$ is the largest eigenvalue of V_{k^*} and ζ is any constant such that $\zeta \geq 334812\sqrt{2} + 1296\sqrt{6}$. The particular form $\hat{\sigma}^2_{k^*}$ given by Eq. (11) guarantees that $\hat{\sigma}^2_{k^*}$ is a good upper bound for the variance of the outcomes $r_{k^*,i,j}$ that are sampled after performing a measurement on the directions given by $a_{k^*,i}$. With this choice as argued in [7] it allows to obtain rigorous concentration bounds such that $\widetilde{\theta}^{\mathrm{wMoM}}_{k^*}$ is a good estimator for the Bloch vector of the unknown state $|\psi\rangle$.

Now we state the main theorem from [7] that states how the regret scales with the time horizon N and also how the infidelities between the unknown state $|\psi\rangle$ the directions $|\psi_k\rangle$ scale with respect to the unknown state $|\psi\rangle$.

Theorem SI.1 ([7, Theorem 9 and 11]). *Fix $K^* \in \mathbb{N}$, $t = \lceil 24 \ln(K^*/\delta) \rceil$ for some $\delta > 0$ and time horizon $N = 4tK^*$. Then we have that the quantum state tomography Algorithm 1 over an unknown state $|\psi\rangle$ achieves with probability at least $1-\delta$ the regret Eq. (2) scaling*

$$\mathrm{Rgrt}(N) \leq C_1 \ln\left(\frac{N}{\delta}\right) \ln(N), \tag{12}$$

for some universal constant $C_1 > 0$. Also for all $k \in \{1, \ldots, N\}$ the selected 2-outcome POVM's given by the rank-1 projector $\psi_k = |\psi_k\rangle\langle\psi_k|$ achieve infidelity

$$1 - |\langle\psi_k|\psi\rangle|^2 \leq C_2 \frac{\ln\left(\frac{N}{\delta}\right)}{k}, \tag{13}$$

for some universal constant $C_2 > 0$. Moreover setting $\delta = \frac{1}{N^*}$ it holds

$$\mathbb{E}[\mathrm{Rgrt}(N)] = C_3 \ln^2(N), \quad \mathbb{E}[1 - |\langle\psi_k|\psi\rangle|^2] \leq C_4 \frac{\ln(N)}{k}, \tag{14}$$

for some universal constants $C_3, C_4 > 0$ and the expectation is taken over the probability distribution of outcomes and measurements induced by the policy.

II. REGRET PROOF

In this section, we present all the technical tools and proofs needed to establish Theorem SI.1, which was originally introduced in [7] for some of the present authors. For clarity, we will restrict our theorems to $\mathbb{R}^3$ that corresponds to the space of the Bloch sphere (qubits).

A. Median of means concentration bound

We start giving the general concentration bound for the median of means least squares estimator (6). In general, we consider a normalized unknown parameter $\theta \in \mathbb{R}^3$, $\|\theta\|_2 = 1$ such that at each time step k selects a normalized vector $a_k \in \mathbb{R}^3$, $\|a_k\|_2 = 1$ and samples t independent rewards distributed as

$$r_{k,i} = \langle \theta, a_k \rangle + \epsilon_{k,i} \quad \text{for } i \in [t], \tag{15}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product and $\epsilon_{k,i}$ is some noise such that $\mathbb{E}[\epsilon_{k,i}|\mathcal{H}_{k-1}] = 0$ where $\mathcal{H}_{k-1}$ contains the past history of actions and rewards. We refer to t as the number of subsamples per time step. Then at time step k we define t least squares estimators as

$$\widetilde{\theta}_{k,i} = V_k^{-1} \sum_{s=1}^{k} r_{s,i} a_s \quad \text{for } i \in [t], \tag{16}$$

where V_k is the design matrix defined as

$$V_k = \lambda \mathbb{I} + \sum_{s=1}^{k} a_s a_s^\intercal, \tag{17}$$

with $\lambda > 0$ being a parameter that ensures invertibility of V_k. We note that the design matrix is independent of i. Then the median of means for least squares estimator (MOMLSE) is defined as

$$\widetilde{\theta}_{k^*}^{\mathrm{MoM}} := \tilde{\theta}_{k^*, j^*} \quad \text{where } j^* = \operatorname*{argmin}_{j \in [t]} y'_j, \tag{18}$$

where

$$y'_j = \mathrm{median}\{\|\tilde{\theta}_{k,j'} - \tilde{\theta}_{k,i}\|_{V_k} : i \in [t]/j'\} \quad \text{for } j' \in [t]. \tag{19}$$

Using the results in [13] we have that the above estimator has the following concentration property around the true estimator.

Lemma SI.2 (Lemma 2 and 3 in [13]). *Let $\widetilde{\theta}_k^{\text{MoM}}$ be the MOMLSE defined (18) in with t subsamples with $\{r_{s,i}\}_{(s,i)\in[k]\times[t]}$ rewards and corresponding actions $\{a_s\}_{s\in[k]}$. Assume that the noise of all rewards has bounded variance, i.e $\mathbb{E}\left[\epsilon_{s,i}^2|\mathcal{H}_{k-1}\right] \leq 1$ for all $s \in [k]$ and $i \in [t]$. Then we have*

$$\Pr\left(\|\theta - \widetilde{\theta}_k^{MoM}\|_{V_k}^2 \leq 9\left(\sqrt{27} + \lambda\|\theta\|_2\right)^2\right) \geq 1 - \exp\left(\frac{-t}{24}\right). \tag{20}$$

We will use a slight modification of the above result with the weighted least squares estimator like the one used in [8]. The weights will be related to a variance estimator of the noise for action $a \in \mathbb{R}^3$ such that at each time step k can be generally defined as

$$\hat{\sigma}_k^2 : \mathcal{H}_{k-1} \times A \to \mathbb{R}_{>0}, \tag{21}$$

where $\mathcal{H}_{k-1} = \{r_{s,i}\}_{(s,i)\in[k-1]\times[t]} \cup \{a_s\}_{s\in[k-1]}$ contains the past information of rewards and actions played and $\mathcal{A} = \{x \in \mathbb{R}^3 : \|x\|_2 = 1\}$ is the set of normalized vectors. For our purposes we will use only the information of the past actions and in order to simplify notation we will use $\hat{\sigma}_k^2(a)$ to denote an estimator of the variance for the reward associated action $a \in \mathcal{A}$ with the information collected up to time step $k-1$. Then the corresponding weighted versions with t subsamples are defined as

$$\widetilde{\theta}_{k,i}^{\text{w}} = (V_k^{\text{w}})^{-1} \sum_{s=1}^{k} \frac{1}{\hat{\sigma}_s^2(a_s)} r_{s,i} a_s \quad \text{for } i \in [t], \tag{22}$$

with the weighted design matrix

$$V_k^{\text{w}} = \lambda\mathbb{I} + \sum_{s=1}^{k} \frac{1}{\hat{\sigma}_s^2(a_s)} a_s a_s^\intercal. \tag{23}$$

Then the weighted version of the median of means linear estimator is defined analogously to (18) with the corresponding weighted versions (22)(23) and we will denote it as $\widetilde{\theta}_k^{\text{wMOM}}$. In our algorithm analysis we will use the following analogous concentration bound under the condition that the estimators $\hat{\sigma}_k^2$ overestimate the true variance. We use $\mathbb{V}[X]$ to denote the variance of a random variable X.

Corollary SI.3. *Let $\widetilde{\theta}_k^{\text{wMOM}}$ be the weighted version of the MOMLSE with t subsamples, $\{r_{s,i}\}_{(s,i)\in[k]\times[t]}$ rewards with corresponding actions $\{a_s\}_{s\in[k]}$ and variance estimator $\hat{\sigma}_k^2$. Define the following event*

$$G_k := \{(\mathcal{H}_{k-1}, a_k) : \mathbb{V}[\epsilon_{s,i}] \leq \hat{\sigma}^2(a_s)\ \forall s, i \in [k] \times [t]\}. \tag{24}$$

Then we have

$$\Pr\left(\|\theta - \widetilde{\theta}_k^{\text{wMOM}}\|_{V_k}^2 \leq \xi \mid G_k\right) \geq 1 - \exp\left(\frac{-t}{24}\right), \tag{25}$$

where

$$\xi := 9\left(\sqrt{27} + \lambda\|\theta\|_2\right)^2. \tag{26}$$

Proof. The result follows from applying Lemma SI.2 to the sequences of re-normalized rewards $\{\frac{r_{s,i}}{\hat{\sigma}_s(a_s)}\}_{(s,i)\in[k]\times[t]}$ and actions $\{\frac{a_{s,i}}{\hat{\sigma}_s(a_s)}\}_{s\in[k]}$. We only need to check that the sequence $\{\frac{\epsilon_{s,i}}{\hat{\sigma}_s(a_s)}\}_{(s,i)\in[k]\times[t]}$ has finite variance. Conditioning with the event G_k and the fact that by definition $\hat{\sigma}_s^2(a_s)$ only depend on the past $s-1$ action and rewards we have that the re-normalized noise has bounded variance since

$$\mathbb{E}\left[\left(\frac{\epsilon_{s,i}}{\hat{\sigma}_s(a_s)}\right)^2 \middle| \mathcal{H}_{k-1}\right] = \frac{1}{\hat{\sigma}_s^2(a_s)}\mathbb{E}[\epsilon_{s,i}^2|\mathcal{H}_{k-1}] = \frac{\mathbb{V}[\epsilon_{s,i}]}{\hat{\sigma}_s^2(a_s)} \leq 1. \tag{27}$$

□

B. Relation between exploration and exploitation

The other technical ingredient we will employ is a lemma proved in [8] that relates the scaling of the minimum and maximum eigenvalue of the design matrix for our particular choice of measurements (or actions). This lemma ensures that the exploration of the algorithm is tightly controlled by the exploitation. We denote the set of normalized vectors as $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and the set of semidefinite positive matrices $\mathrm{P}_+^d = \{X \in \mathbb{R}^{d\times d} : X \geq 0\}$.

Theorem SI.4 (Theorem 3 in [8]). *Let $\{c_k\}_{k=0}^{\infty} \subset \mathbb{S}^{d-1}$ be a sequence of normalized vectors and $\omega : \mathrm{P}_+^d \to \mathbb{R}_{\geq 0}$ a function such that*

$$\omega(X) \leq C\sqrt{\|X\|_\infty}, \tag{28}$$

for a constant $C > 0$ and any $X \in \mathrm{P}_+^d$. Let $\lambda_0 \geq \max\left\{2, \sqrt{\frac{2}{3(d-1)}}2dC + \frac{2}{3(d-1)}\right\}$, and define a sequence of matrices $\{V_k\}_{k=0}^{\infty} \subset \mathbb{R}^{d\times d}$ as

$$V_0 := \lambda_0 \mathbb{I}_{d\times d}, \quad V_{k+1} := V_k + \omega(V_k)\sum_{i=1}^{d-1} P_{k,i}, \tag{29}$$

where

$$P_{k,i} := a_{k+1,2i-1}(a_{k+1,2i-1})^\intercal + a_{k+1,2i}(a_{k+1,2i})^\intercal, \tag{30}$$

$$a_{k+1,i} := \frac{\tilde{a}_{k+1,i}}{\|\tilde{a}_{k+1,i}\|_2}, \quad \tilde{a}_{k+1,i} := c_k - \frac{(-1)^i}{\sqrt{\lambda_{k,1}}} v_{k,\lceil \frac{i}{2} \rceil}, \tag{31}$$

with $\lambda_{k,i} = \lambda_i(V_k)$ the eigenvalues of V_k with corresponding normalized eigenvectors $v_{k,1}, ..., v_{k,d} \in \mathbb{S}^{d-1}$. Then we have

$$\lambda_{\min}(V_k) \geq \sqrt{\frac{2}{3(d-1)}\lambda_{\max}(V_k)} \quad \textit{for all} \quad k \geq 0. \tag{32}$$

C. Proof of Theorem SI.1

We present the proof of Theorem SI.1 taken from the preprint [7] for the completeness of our work.

Proof. First, we write the regret in terms of Bloch vectors since the measurement updates (9) are given in terms of Bloch vectors. A simple computation gives

$$\mathrm{Rgrt}(N) = \frac{1}{2}\sum_{k=1}^{N} \|\theta - a_k\|_2^2, \tag{33}$$

where θ is the unknown Bloch vector of $|\psi\rangle$ and a_k the Bloch vector of the measurement direction $|\psi_k\rangle$. Thus, we have that to give an upper bound between the distance of the unknown parameter θ and the actions $a_{k,i}$ selected by the algorithm (9).

We denote the step $k^* \in [K^*]$ to run over the batches, the algorithm updates the MoM estimator $\widetilde{\theta}_k^{\mathrm{wMOM}}$. First, we will do the computation assuming that the event

$$E_{k^*} := \{\mathcal{H}_{k^*} : \forall s \in [k^*], \theta \in \mathcal{C}_s\}, \tag{34}$$

holds where $\mathcal{C}_s = \{\theta' \in \mathbb{R}^d : \|\theta' - \widetilde{\theta}_{k^*}^{\mathrm{wMOM}}\|_{V_s^{\mathrm{w}}}^2 \leq \xi\}$. Here, the history $\mathcal{H}_{\tilde{k}}$ is defined with the previous outcomes and actions of our algorithm i.e

$$\mathcal{H}_{k^*} := (r_{s,i,j}, a_{s,i})_{(s,i,j)\in[k^*]\times[2(d-1)]\times[t]} \tag{35}$$

Later, we will quantify the probability that this event always holds. Using the definition of the actions (9), $\theta, \widetilde{\theta}_{k^*}^{\mathrm{wMOM}} \in \mathbb{S}^2$ and the arguments from [8][Appendix C.1, Eq. (165)] we have that through geometrical arguments

$$\|\theta - a_{k^*,i}\|_2^2 \leq \frac{9\xi}{\lambda_{\min}(V_{k^*-1}^{\mathrm{w}})}. \tag{36}$$

Then using that the design matrix $V^{\mathrm{w}}_{k^*}$ (5) is updated as in Theorem SI.4 and the choice of weights (11) we fix

$$\lambda_0 \geq \max\left\{2, 2\sqrt{\frac{1}{3}}\frac{1}{4\sqrt{2}\xi} + \frac{1}{3}\right\} \tag{37}$$

and we have that $\lambda_{\min}(V^{\mathrm{w}}_{k^*}) \geq \sqrt{\frac{1}{3}\lambda_{\max}(V^{\mathrm{w}}_{k^*})}$ applying Theorem SI.4. Inserting this into the above, we have

$$\|\theta - a_{k^*,i}\|_2^2 \leq \frac{12\sqrt{2}\xi}{\sqrt{\lambda_{\max}(V^{\mathrm{w}}_{k^*})}}. \tag{38}$$

Thus, it remains to provide a lower bound on $\lambda_{\max}(V^{\mathrm{w}}_{k^*})$. We note that in [8][Appendix C.1] they also had to provide an upper bound, but this was because the constant ξ they use depends on k. From the definition of $V^{\mathrm{w}}_{k^*}$ (5) we can bound the trace as

$$\mathrm{Tr}(V^{\mathrm{w}}_{k^*}) \geq \sum_{s=2}^{k^*} 4\omega(V^{\mathrm{w}}_{s-1}) = \frac{\sqrt{2}}{6\xi}\sum_{s=1}^{k^*-1}\sqrt{\lambda_{\max}(V^{\mathrm{w}}_{s})}. \tag{39}$$

Then using the bound $\mathrm{Tr}(V^{\mathrm{w}}_{k^*}) \geq \lambda_{\max}(V^{\mathrm{w}}_{k^*})/3$ and some algebra we arrive at

$$\lambda_{\max}(V^{\mathrm{w}}_{k^*}) \geq \frac{1}{1+6\frac{3}{\sqrt{2}}\xi}\sum_{s=1}^{k^*}\sqrt{\lambda_{\max}(V^{\mathrm{w}}_{s})}. \tag{40}$$

Now we have an inequality with the function $\lambda_{\max}(V^{\mathrm{w}}_{s})$ at both sides. In order to solve it we use the technique from [8][Appendix C.1, Eqs. (197)–(208)] which consist on extending $\lambda_{\max}(V^{\mathrm{w}}_{k^*})$ to the continuous with a linear interpolation and then transforming the sum to an integral which leads to a differential inequality. Solving this leads to

$$\lambda_{\max}(V^{\mathrm{w}}_{k^*}) \geq \frac{(k^*)^2}{4(1+6\frac{3}{\sqrt{2}}\xi)^2}. \tag{41}$$

Now we can insert the above into (38) and we have

$$\|\theta - a_{k^*,i}\|_2^2 \leq \frac{24\sqrt{2}\xi(1+6\frac{3}{\sqrt{2}}\xi)}{k^*-1} = \frac{432\xi^2 + 24\sqrt{2}\xi}{k^*-1}. \tag{42}$$

Thus, we can insert the above bound into the regret expression Eq. (33) and we have

$$\begin{aligned}
\mathrm{Rgrt}(N) &= \frac{1}{2}\sum_{k=1}^{N}\|\theta - a_k\|_2^2 = \frac{1}{2}\sum_{k^*=1}^{K^*}\sum_{i=1}^{4}\sum_{j=1}^{t}\|\theta - a_{k^*,i}\|_2^2 && (43)\\
&\leq 8t + \frac{1}{2}\sum_{k^*=2}^{K^*}\sum_{i=1}^{4}\sum_{j=1}^{t}\|\theta - a_{k^*,i}\|_2^2 && (44)\\
&\leq 8t + (864t\xi^2 + 48^{\frac{3}{2}}t\xi)\sum_{k^*=2}^{K^*}\frac{1}{k^*-1} && (45)\\
&\leq 8t + 864t\xi^2\log K^* + 48^{\frac{3}{2}}t\xi\log K^* && (46)\\
&= 8t + 864t\xi^2\log\left(\frac{N}{4t}\right) + 48^{\frac{3}{2}}t\xi\log\left(\frac{N}{4t}\right).
\end{aligned}$$

It remains to quantify the probability that the event E_{k^*} holds. For that, we will use the concentration bounds of the median of means for the least squares estimator stated in Corollary SI.3. First, we note that the relation between the reward and Bloch vector is

$$\langle\theta, a_k\rangle = 2\,\mathbb{E}[r_k] - 1, \tag{47}$$

which justifies the renormalization of the reward we use in (22). Then a simple computation of the variance of our model leads to

$$\mathbb{V}[\epsilon_{k^*,i,j}|\mathcal{H}_{k^*-1}] \leq 1 - \langle \theta, a_{k^*,i}\rangle^2 \leq 2(1 - \langle \theta, a_{k^*,i}\rangle) = \|\theta - a_{k^*,i}\|_2^2, \tag{48}$$

where we used $1 + \langle \theta, a_{k^*,i}\rangle \leq 2$. Thus from our choice of weights (11) and (42) we have that

$$\text{if } \theta \in \mathcal{C}_{s-1} \Rightarrow \mathbb{V}[\epsilon_{k^*,i,j}|\mathcal{H}_{k^*-1}] \leq \hat{\sigma}_s^2(a_{s,i}). \tag{49}$$

Then in order to apply Corollary SI.3 we note that from the choice $\hat{\sigma}_s^2(a_{1,i}) = 1$ the event G_{k^*} at $k^* = 1$ is always satisfied i.e $\Pr(G_1) = 1$. Then applying Bayes theorem, union bound over the events $G_1, E_1, ..., G_{K^*}, E_{K^*}$ and Corollary SI.3 we have

$$\Pr(E_{K^*} \cap G_{K^*}) \geq (1 - \exp(-t/24))^{K^*}. \tag{50}$$

This probability also quantifies the probability that (42) holds since the only assumption we used is $\theta \in \mathcal{C}_{k^*-1}$. Then we can take simply one of the actions $a_{k^*,i}$ as the estimator $\hat{\theta}_k$ and the result follows using the relabeling $k = 4tk^*$ and the inequality $1/(k^* - 1) \leq 2/k^*$ for $k^* \geq 2$. □

D. Extension to Higher-Dimensional Systems

The extension of our algorithm to higher-dimensional quantum systems, such as qudits or systems of interacting qubits (e.g., unknown states of an XY chain), poses mainly technical challenges. A key result underlying our performance guarantees—namely, the eigenvalue relation of the design matrix

$$\lambda_{\min}(V_k) = \Omega\left(\sqrt{\lambda_{\max}(V_k)}\right), \tag{51}$$

proven in [8]—relies on geometric arguments in real vector spaces. For qubits, this result applies directly because the Bloch sphere corresponds to a real 3-dimensional sphere, matching the assumptions of the proof. The geometry of general quantum state spaces, however, is significantly more intricate. In particular, the action space for higher-dimensional quantum states does not form a simple real sphere, and the confidence ellipsoid arguments used in our regret analysis no longer carry over straightforwardly. While we believe that an analogous spectral relation should hold more generally, adapting the proof to arbitrary quantum systems remains an open challenge. We are currently investigating alternative formulations that could lead to generalizations of our method to such systems.

III. ρ^*-WORK EXTRACTION PROTOCOL

Here we discuss more details about the work-extraction protocol that is used, we adapted the protocol formalized in Skrzypczyk's paper [14]. Just as in their formulation, the expected work extracted from a known state ρ will precisely be given by the state's non-equilibrium free energy, which equals the relative entropy between the state and Gibbs' state, γ_β with inverse temperature β, i.e.

$$\beta\,\mathbb{E}(W) = \mathrm{D}(\rho\|\gamma_\beta). \tag{52}$$

We will be applying the protocol to a degenerate Hamiltonian. Discussion on non-degenerate Hamiltonian will be discussed towards the end of the section.

We focus on a specific time step within the N rounds of extraction, doing so simplifies notation by removing the k index. In a specific round, the agent is given a partially unknown qubit system state ψ, and a classical description of the direction $\hat{\psi}$ and an accuracy ϵ. The agent will then choose to optimize the protocol for a state $\rho^* = (1-\epsilon)\hat{\psi} + \epsilon\hat{\psi}^\perp$. Along the unknown state ψ, he also has access to a heat bath at inverse temperature β and a battery state $\varphi(x)$. The Hamiltonian of the system is $H_A = \omega\mathbb{1}$, setting $\hbar = 1$. The heat bath can provide any amount of thermal state with any Hamiltonian at inverse temperature β. We will mainly consider qubit thermal states $\gamma_\beta(\nu) = \frac{1}{Z_R(\nu)} e^{-\beta H_R(\nu)}$ where $Z_R(\nu) = \mathrm{tr}\left(e^{-\beta H_R(\nu)}\right)$ with Hamiltonian $H_R(\nu) = \nu\,|1\rangle\langle 1|$ where $\{|i\rangle\}_{i=0,1}$ are the energy eigenstates. The battery is modeled as a weight at a certain height whose state is described by $\varphi(x) \in L^2(\mathbb{R})$ and Hamiltonian H_B such that $H_B\varphi(x - \mathrm{d}x) = x\varphi(x - \mathrm{d}x)$. We will assume $\varphi(x - \mathrm{d}x)$ to be a battery state whose energy is sharply centered at $\mathrm{d}x$. The energy of the battery can be changed by translating the weight up by a certain height $\mathrm{d}x$, described by the

translation operator $\Gamma^B_{\mathrm{d}x}\varphi(x) = \varphi(x - \mathrm{d}x)$. We aim to extract work from the partially unknown system system state into the battery. We will design the work extraction protocol for the state ρ^*, i.e. the protocol optimally extracts work from ρ^*, see Algorithm 2.

To simplify calculation as well as maintain generality later, we will denote ψ as ρ, $\hat{\psi}$ as $|\phi_0\rangle$ and $\hat{\psi}^\perp$ as $|\phi_1\rangle$, likewise we denote $p_0 = 1 - \epsilon$ and $p_1 = \epsilon$, so $\rho^* = \sum_i p_i |\phi_i\rangle\langle\phi_i|$.

Algorithm 2: ρ^*-ideal work extraction

Require: An unknown system state ψ, a classical description of the state ρ^*, a battery state $\varphi(x)$ with the battery energy $\mu = 0$, a reservoir at inverse temperature β

Set $\{\phi_i\}_i$ and $\{p_i\}_i$ to be the eigenvectors and eigenvalues of ρ^*

Unitary Rotation

Apply unitary $U = \sum_i |i\rangle\langle\phi_i|$ to try to diagonalize the system qubit in computational basis.

for $\ell = 1, 2, \ldots, M$ **do**

Prepare a fresh a reservoir qubit and exchange it with the system

Take a fresh thermal qubit $\gamma_\beta(\nu(\ell,\epsilon)) = \frac{1}{Z_R(\nu(\ell,\epsilon))} e^{-\beta H_R(\nu(\ell,\epsilon))}$ where $\nu(\ell,\epsilon) = \beta^{-1} \ln \frac{p_{0,\ell}}{p_{1,\ell}}$, $p_{i,\ell} = p_i - (-1)^i \ell \delta p$ and $\delta p = \frac{1}{M}(p_0 - \frac{1}{2})$ from the reservoir.

Apply the swap unitary $V_{\rho^*,\ell} = \sum_{ij} |i\rangle\langle j|_A \otimes |j\rangle\langle i|_R \otimes \Gamma_{(i-j)\nu(\ell,\epsilon)}$ on the system, the battery and the reservoir qubit.

Discard the reservoir qubit.

Measure the extracted work

Measure the battery energy, obtain the battery energy μ' and compute the extracted work $\Delta W = \mu' - \mu$.

In the first stage of the protocol, we rotate the unknown qubit via unitary

$$U = \sum_i |i\rangle\langle\phi_i| \ . \tag{53}$$

This operation attempts to diagonalize the system qubit in the computation basis. We then interact the system with the battery. The state of the system together with the battery is

$$\rho_{AB} = \sum_{ij} \langle\phi_i| \rho |\phi_j\rangle |i\rangle\langle j|_A \otimes \varphi(x)_B. \tag{54}$$

In the second stage of the protocol, we perform M repetitions of the following process. In repetition ℓ, we take a fresh thermal qubit $\gamma_\beta(\nu(\ell,\epsilon))$ with Hamiltonian $H_R(\nu(\ell,\epsilon))$ from the reservoir where $\nu(\ell,\epsilon) = \beta^{-1} \ln \frac{p_{0,\ell}}{p_{1,\ell}}$, $p_{i,\ell} = p_i - (-1)^i \ell \delta p$ and $\delta p = \frac{1}{M}(p_0 - \frac{1}{2})$. Note that the reservoir qubit we take depends on which repetition we are in. Then we perform the swap unitary

$$V_{\rho^*,\ell} = \sum_{ij} |i\rangle\langle j|_A \otimes |j\rangle\langle i|_R \otimes \Gamma_{(i-j)\nu(\ell,\epsilon)} \ . \tag{55}$$

This unitary swaps the system and the fresh qubit from the reservoir, extracts work into the battery due to the different energy gap between $\{|i\rangle_A\}_{i=0,1}$ and $\{|i\rangle_R\}_{i=0,1}$ and conserves energy of the system, the qubit from the reservoir and the battery. Finally, the qubit from the reservoir is discarded. At the end of each repetition ℓ, the reduced state is

$$\rho_{AB,\ell} = \mathrm{tr}_R \left(V_{\rho^*,\ell} \left(\rho_{AB,\ell-1} \otimes \gamma_\beta(\nu(\ell,\epsilon)) \right) V^\dagger_{\rho^*,\ell} \right), \tag{56}$$

After the first repetition, we obtain

$$\rho_{AB,1} = \sum_i p_{i,1} |i\rangle\langle i|_A \otimes \rho_{B,i,1}, \tag{57}$$

where

$$\rho_{B,i,1} = \sum_j \langle\phi_j| \rho |\phi_j\rangle \varphi(x - (i-j)\nu(\ell,\epsilon)), \tag{58}$$

and after repetition ℓ where $\ell \geq 2$, we obtain

$$\rho_{AB,\ell} = \sum_i p_{i,\ell} \left|i\right\rangle\!\left\langle i\right|_A \otimes \rho_{B,i,\ell}, \tag{59}$$

where

$$\rho_{B,i,\ell} = \sum_j p_{j,\ell-1} \Gamma_{(i-j)\nu(\ell,\epsilon)} \rho_{B,j,\ell-1} \Gamma^\dagger_{(i-j)\nu(\ell,\epsilon)}. \tag{60}$$

From Eq. (59), we observe that the reduced state of the system changes gradually, which resembles a quasi-static process in thermodynamics. This is the reason why we take the swap unitary in repetition.

A. Work distribution

In this section, we will use the Lagrange mean value theorem and the first mean value theorem for definite integrals [15] as follows:

Theorem SI.5. *Let $f : [a, b] \to \mathbb{R}$ be a continuous on the closed interval $[a, b]$ and differentiable on the open interval (a, b). Then there exists $c \in (a, b)$ such that*

$$f(b) - f(a) = f'(c)(b - a). \tag{61}$$

Theorem SI.6. *Let $f : [a, b] \to \mathbb{R}$ be a continuous function on the closed interval $[a, b]$. Then there exists $c \in (a, b)$ such that*

$$\int_a^b f(x)\mathrm{d}x = f(c)(b - a). \tag{62}$$

We will show the following theorem in the following subsection.

Theorem SI.7. *Let $\{\phi_i\}_{i=0,1}$ and $\{p_i\}_{i=0,1}$ be the eigenvectors and eigenvalues of ρ^*, ΔW be the extracted work (which is a continuous random variable) and M be the number of repetitions as in Algorithm 2. It holds that: if the extraction protocol is operated on a state ρ that is the eigenstate of ρ^*, i.e., $\rho = \phi_i$, then then the expected extracted work $\mathbb{E}[\Delta W]$ converges to a fixed value w_i and the extracted work ΔW converges in probability to its expectation $\mathbb{E}[\Delta W]$. To be precise, it means*

$$\lim_{M\to\infty} \mathbb{E}[\Delta W] = w_i, \tag{63}$$

where

$$w_i \coloneqq \beta^{-1}(\mathrm{D}(\phi_i \| \mathbb{1}/2) + \ln p_i), \tag{64}$$

and for any $\epsilon > 0$

$$\lim_{M\to\infty} \Pr[|\Delta W - \mathbb{E}[\Delta W]| \geq \epsilon] = 0. \tag{65}$$

Proof. We consider the case where $\rho = \phi_i$. We assume that $p_0 > \frac{1}{2}$ as we deal with an estimate for a pure state in Algorithm 2, although similar proof holds for other cases. According to Eq. (57), the state after the first repetition can be viewed as a classical state described as follows: the state after repetition 1 is ϕ_{x_1} where x_1 is a random bit sampled from $\{0, 1\}$ according to the probability distribution $(p_{0,1}, p_{1,1})$; the extracted work after the first repetition conditioned on x_1 is $(x_1 - i)\nu(1, \epsilon)$. According to Eq. (59), the evolution in repetition ℓ where $\ell \geq 2$ can be viewed as a classical process described as follows: the state after repetition ℓ is ϕ_{x_ℓ} where x_ℓ is a random bit sampled from $\{0, 1\}$ according to the probability distribution $(p_{0,\ell}, p_{1,\ell})$; the extracted work in repetition ℓ conditioned on $x_{\ell-1}x_\ell$ and is $(x_\ell - x_{\ell-1})\nu(\ell, \epsilon)$. Suppose that the random bits sampled during the above process is $x_1 \dots x_M$ after M repetitions. The extracted work after repetition M conditioned on $x_1 \dots x_M$ is

$$\Delta W = (x_1 - i)\nu(1, \epsilon) + \sum_{\ell=2}^{M} (x_\ell - x_{\ell-1})\nu(\ell, \epsilon) = -i\nu(1, \epsilon) + \sum_{\ell=1}^{M-1} x_\ell(\nu(\ell, \epsilon) - \nu(\ell + 1, \epsilon)) + x_M\nu(M, \epsilon), \tag{66}$$

recall that

$$\nu(\ell,\epsilon)=\beta^{-1}\ln\frac{p_0-\ell\delta p}{p_1+\ell\delta p}. \tag{67}$$

where $\delta p=\frac{1}{M}(p_0-\frac{1}{2})$. The expected extracted work is

$$\mathbb{E}[\Delta W]=-i\nu(1,\epsilon)+\sum_{\ell=1}^{M-1}\mathbb{E}[x_\ell](\nu(\ell,\epsilon)-\nu(\ell+1,\epsilon))+\mathbb{E}[x_M]\nu(M,\epsilon). \tag{68}$$

Notice that, from the definition of x_ℓ, $\mathbb{E}[x_\ell]=p_{1,\ell}=p_1+\ell\delta p$, we obtain

$$\mathbb{E}[\Delta W]=-i\beta^{-1}\ln\frac{p_0-\delta p}{p_1+\delta p}+\beta^{-1}\sum_{\ell=1}^{M-1}\left(\ln\frac{p_0-\ell\delta p}{p_1+\ell\delta p}-\ln\frac{p_0-(\ell+1)\delta p}{p_1+(\ell+1)\delta p}\right)(p_1+\ell\delta p). \tag{69}$$

We now use the definition of $\nu(\ell,\epsilon)$ in Eq. (67) as well as the Lagrange mean value theorem in Theorem SI.5 to obtain

$$\beta(\nu(\ell,\epsilon)-\nu(\ell+1,\epsilon))=\ln\frac{p_0-\ell\delta p}{p_1+\ell\delta p}-\ln\frac{p_0-(\ell+1)\delta p}{p_1+(\ell+1)\delta p}=\frac{1}{\xi_\ell(1-\xi_\ell)}\delta p, \tag{70}$$

for some $\xi_\ell\in[p_0-(\ell+1)\delta p,p_0-\ell\delta p]$.

Therefore, Eq. (69) can be simplified to

$$\begin{aligned}\mathbb{E}[\Delta W]&=-i\beta^{-1}\left(\ln\frac{p_0}{p_1}-\frac{\delta p}{\xi_0(1-\xi_0)}\right)+\beta^{-1}\sum_{\ell=1}^{M-1}\frac{p_1+\ell\delta p}{\xi_\ell(1-\xi_\ell)}\delta p\\&=-i\beta^{-1}\ln\frac{p_0}{p_1}+\beta^{-1}\sum_{\ell=1}^{M}\frac{p_1+\ell\delta p}{\xi_\ell(1-\xi_\ell)}\delta p\\&\quad+i\beta^{-1}\frac{\delta p}{\xi_0(1-\xi_0)}-\beta^{-1}\frac{\delta p}{2\xi_M(1-\xi_M)}\,.\end{aligned} \tag{71}$$

We will approximate the sum in the second line of Eq. (71) with an integration, where the remainder is bounded due to first mean value theorem for definite integrals as in Theorem SI.6. Namely,

$$\beta^{-1}\sum_{\ell=1}^{M}\frac{p_1+\ell\delta p}{\xi_\ell(1-\xi_\ell)}\delta p=\beta^{-1}\int_{\frac{1}{2}}^{p_0}\frac{\mathrm{d}p}{p}+\beta^{-1}R_1(\delta p)=\beta^{-1}\ln p_0+\beta^{-1}\ln(2)+\beta^{-1}R_1(\delta p), \tag{72}$$

where $R_1(\delta p)$ is the remainder given by

$$R_1(\delta p)=\sum_{\ell=1}^{M}\frac{p_1+\ell\delta p}{\xi_\ell(1-\xi_\ell)}\delta p-\int_{\frac{1}{2}}^{p_0}\frac{\mathrm{d}p}{p}=\sum_{\ell=1}^{M}\left(\frac{p_1+\ell\delta p}{\xi_\ell(1-\xi_\ell)}\delta p-\int_{p_0-\ell\delta p}^{p_0-(\ell-1)\delta p}\frac{\mathrm{d}p}{p}\right) \tag{73}$$

$$=\sum_{\ell=1}^{M}\left(\frac{p_1+\ell\delta p}{\xi_\ell(1-\xi_\ell)}-\frac{1}{\xi'_\ell}\right)\delta p=\sum_{\ell=1}^{M}\frac{\xi_\ell(1-\xi_\ell)-\xi'_\ell(p_1+\ell\delta p)}{\xi'_\ell\xi_\ell(1-\xi_\ell)}\delta p, \tag{74}$$

where from the first line to the second line, we have used the first mean value theorem for definite integrals Theorem SI.6 that

$$\int_{p_0-\ell\delta p}^{p_0-(\ell-1)\delta p}\frac{\mathrm{d}p}{p}=\frac{1}{\xi'_\ell}\delta p, \tag{75}$$

for some $\xi'_\ell\in[p_0-\ell\delta p,p_0-(\ell-1)\delta]$. Therefore, the remainder satisfies

$$|R_1(\delta p)|\le\sum_{\ell=1}^{M}\left|\frac{\xi_\ell(1-\xi_\ell)-\xi'_\ell(p_1+\ell\delta p)}{\xi'_\ell\xi_\ell(1-\xi_\ell)}\right|\delta p\le\sum_{\ell=1}^{M}\left|\frac{p_1+\ell\delta p+\xi_\ell}{\xi'_\ell\xi_\ell(1-\xi_\ell)}\right|(\delta p)^2\le\sum_{\ell=1}^{M}\frac{4}{p_1}(\delta p)^2\le(2p_0-1)\frac{2}{p_1}\delta p. \tag{76}$$

Therefore, the second line in Eq. (71) is finite while the third line is infinitesimal, and we obtain

$$\mathbb{E}[\Delta W] = -i\beta^{-1}\ln\frac{p_0}{p_1} + \beta^{-1}\ln p_0 + \beta^{-1}\ln(2) + O(\delta p) \tag{77}$$

$$= -\beta^{-1}\operatorname{tr}(\phi_i \ln \mathbb{1}/2) - i\beta^{-1}\ln\frac{p_0}{p_1} + \beta^{-1}\ln p_0 + O(\delta p) \tag{78}$$

$$= -\beta^{-1}\operatorname{tr}(\phi_i \ln \mathbb{1}/2) + \beta^{-1}\ln p_i + O(\delta p) \tag{79}$$

$$= \beta^{-1}\left[\mathrm{D}\left(\phi_i \| \mathbb{1}/2\right) + \ln p_i\right] + O(\delta p). \tag{80}$$

Since $\delta p \propto \frac{1}{M}$, we then obtain that

$$\mathbb{E}[\Delta W] = \beta^{-1}\left[\mathrm{D}\left(\phi_i \| \mathbb{1}/2\right) + \ln p_i\right] + O\left(\frac{1}{M}\right). \tag{81}$$

Taking $M \to \infty$, we obtain

$$\lim_{M\to\infty} \mathbb{E}[\Delta W] = \beta^{-1}\left[\mathrm{D}\left(\phi_i \| \mathbb{1}/2\right) + \ln p_i\right]. \tag{82}$$

Now we demonstrate the convergence of ΔW towards its expectation value, recall from Eq. (66), we have that

$$\Delta W = -i\nu(1,\epsilon) + \sum_{\ell=1}^{M-1} x_\ell(\nu(\ell,\epsilon) - \nu(\ell+1,\epsilon)) + x_M \nu(M,\epsilon). \tag{83}$$

By the Lagrange mean value theorem,

$$\nu(\ell,\epsilon) - \nu(\ell+1,\epsilon) = \ln\frac{p_0 - \ell\delta p}{p_1 + \ell\delta p} - \ln\frac{p_0 - (\ell+1)\delta p}{p_1 + (\ell+1)\delta p} = \frac{\delta p}{\xi_\ell(1-\xi_\ell)}, \tag{84}$$

for some $\xi_\ell \in [p_0 - (\ell+1)\delta p, p_0 - \ell\delta p]$ and $\ell = 1, \ldots, (M-1)$ satisfying

$$|\nu(\ell,\epsilon) - \nu(\ell+1,\epsilon)| \leq \frac{2}{p_1}\delta p. \tag{85}$$

We thus obtain that $x_\ell(\nu(\ell,\epsilon) - \nu(\ell+1,\epsilon)) \in [0, \frac{2}{p_1}\delta p]$ for $\ell = 1, \ldots, (M-1)$. Besides, $\nu(M,\epsilon) = 0$. The convergence rate to the expectation, by the Hoeffding inequality, is given by

$$\Pr[|\Delta W - \mathbb{E}[\Delta W]| \geq \zeta] \leq 2e^{-\frac{\zeta^2}{\sum_{\ell=1}^{M-1}\left(\frac{2}{p_1}\delta p\right)^2}} \leq 2e^{-\frac{p_1^2\zeta^2 M}{(2p_0-1)^2}}. \tag{86}$$

Taking $M \to \infty$, we obtain

$$\lim_{M\to\infty} \Pr[|\Delta W - \mathbb{E}[\Delta W]| \geq \zeta] = 0. \tag{87}$$

□

Theorem SI.7 demonstrates that the extracted work is close to either w_0 or w_1, with probability close to $\langle\phi_0|\,\rho\,|\phi_0\rangle$ and $\langle\phi_1|\,\rho\,|\phi_1\rangle$ respectively. Therefore, measuring the extracted work ΔW from the state ρ in Algorithm 2 is effectively measuring the state ρ in the basis $\{\phi_i\}_{i=0,1}$ up to an error probability exponentially vanishing with respect to the number of repetitions M in Algorithm 2. When ρ is a pure state i.e., $\rho = |\psi\rangle\langle\psi|$, the energy measurement of the battery is equivalent to the reward measurement in the quasi-static limit of Algorithm 1, and their correspondence is (without loss of generality assuming $w_0 \geq w_1$)

$$r = \begin{cases} 1, & \Delta W \geq \frac{w_0+w_1}{2}, \\ 0, & \Delta W \leq \frac{w_0+w_1}{2}. \end{cases} \tag{88}$$

The distribution of the reward is

$$\Pr[R = r] = \begin{cases} |\langle\phi_0|\psi\rangle|^2 + \epsilon_{\text{error}}, & r = 1, \\ 1 - |\langle\phi_0|\psi\rangle|^2 - \epsilon_{\text{error}}, & r = 0, \end{cases} \tag{89}$$

where

$$|\epsilon_{\text{error}}| \leq 2e^{-\frac{p_1^2\zeta^2 M}{(2p_0-1)^2}} \tag{90}$$

In the limit of $M \to \infty$, the correspondence reduces to

$$r = \begin{cases} 1, & \Delta W = w_0, \\ 0, & \Delta W = w_1, \end{cases} \tag{91}$$

and the distribution of the reward reduces to

$$\Pr[R = r] = \begin{cases} |\langle\phi_0|\psi\rangle|^2, & r = 1, \\ 1 - |\langle\phi_0|\psi\rangle|^2, & r = 0, \end{cases} \tag{92}$$

The above claims takes $M \to \infty$. In reality, this means M to be sufficiently large. Now we explain how large M should be using Algorithm 1. Without loss of generality, we assume the input state is $\rho = \phi_0$. We are supposed to obtain the correct reward $r = 1$. According to Eq. (88), we obtain a wrong reward $r = 0$ only if

$$|\Delta W - \mathbb{E}[\Delta W]| \geq \frac{1}{2}(w_0 - w_1) - |\,\mathbb{E}[\Delta W] - w_0|. \tag{93}$$

Note that as Algorithm 2 proceeds with increasing k, $\epsilon_k = \Theta(\ln(N)/k)$ and, by definition, $p_1 = \Theta(\ln(N)/k)$ as well. Substituting ϵ_k into Eq. (64), we obtain

$$\mathbb{E}[\Delta W] = w_0 + O\left(\frac{1}{M}\right), \tag{94}$$

$$w_0 = \beta^{-1}\left(\frac{1}{2} + \ln\left(1 - \Theta\left(\frac{\ln(N)}{k}\right)\right)\right), \tag{95}$$

$$w_1 = \beta^{-1}\left(\frac{1}{2} + \ln\left(\Theta\left(\frac{C\ln(N)}{k}\right)\right)\right). \tag{96}$$

Substituting above values into Eq. (93) and noting $\Theta(\ln(N)/k)$ is small for large k, we obtain a wrong reward $r = 0$ only if

$$|\Delta W - \mathbb{E}[\Delta W]| \geq \frac{\beta^{-1}}{2}\ln\Theta\left(\frac{k}{\ln(N)}\right) - O\left(\frac{1}{M}\right) = \Theta(\ln(k)). \tag{97}$$

In the concentration bound in Eq. 86, setting $\zeta = \Theta(\ln(k))$, the error probability is upper bounded by

$$\Pr[\text{Wrong Reward}] \leq 2e^{-\frac{p_1^2\zeta^2 M}{(2p_0-1)^2}} \leq 2e^{-\Theta\left(\frac{\ln(N)^2\ln(k)^2 M}{k^2}\right)}, \tag{98}$$

where we have substituted $p_1 = \Theta(\ln(N)/k)$, $\zeta = \Theta(\ln(k))$ and $2p_0 - 1 = \Theta(1)$. At the same time, Algorithm 1 succeeds in all rounds with probability $1 - \frac{1}{N}$ if the error probability scales as $\Pr[\text{Wrong Reward}] \leq O(\frac{1}{N^2})$. Recalling that $k \in [N]$, this indicates that we have to choose

$$M = \Theta\left(\frac{N^2}{\ln(N)^3}\right). \tag{99}$$

The number of iterations $M = \Theta(N^2/\ln(N)^3)$ in each round leads to the time $T = \Theta(N^3/\ln(N)^3)$ cost by the N-round work extraction algorithm.

B. Extracted work for different inputs

Theorem SI.8. *Let $\{\phi_i\}_{i=0,1}$ and $\{p_i\}_{i=0,1}$ be the eigenvectors and eigenvalues of ρ^* and w_i be the value of work extracted defined in Theorem. SI.7. It holds that,*

1. *When applying the protocol to any state ρ, the probability of measuring $\Delta W = w_i$ is given by*

$$\Pr(\Delta W = w_i) = \langle\phi_i|\,\rho\,|\phi_i\rangle \ . \tag{100}$$

2. *When the extraction protocol is operated on a state ρ where $\rho \neq \rho^*$, the expected work extracted is given by*

$$\mathbb{E}[\Delta W] = \beta^{-1}\left[\mathrm{D}(\rho\|\mathbb{1}/2) - \mathrm{D}(\rho\|\rho^*)\right] \ , \tag{101}$$

where the second term can be defined as the the dissipation due to the agent's imperfect knowledge of ρ.

Proof. We first observe that the off-diagonal term $\langle\phi_i|\,\rho\,|\phi_j\rangle\,|i\rangle\langle j|_A \otimes \varphi(x)_B$ of the join state in Eq. (54) does not affect Eq. (57). Therefore, it is identical for the case where the input is ρ and the case where the input is $\sum_i \langle\phi_i|\,\rho\,|\phi_i\rangle\,\phi_i$. The latter case can be viewed as a probabilistic mixture of cases where the input is ϕ_i with probability $\langle\phi_i|\,\rho\,|\phi_i\rangle$. Therefore, Statement 1 in Theorem SI.8 holds, i.e.,

$$\Pr(\Delta W = w_i) = \langle\phi_i|\,\rho\,|\phi_i\rangle \ . \tag{102}$$

Next, to prove Statement 2, we consider the case where a protocol optimized for $\rho^* = \sum_i p_i\,|\phi_i\rangle\langle\phi_i|$ is applied onto an arbitrary state ρ with possibly $\rho \neq \rho^*$. Using Eq. (64) and (100), the extracted work is given by

$$\begin{aligned}
\mathbb{E}[\Delta W] &= \langle\phi_0|\,\rho\,|\phi_0\rangle\,\beta^{-1}\left[\mathrm{D}\left(\phi_0\|\mathbb{1}/2\right) + \ln p_0\right] + \langle\phi_1|\,\rho\,|\phi_1\rangle\,\beta^{-1}\left[\mathrm{D}\left(\phi_1\|\mathbb{1}/2\right) + \ln p_1\right] && (103)\\
&= \beta^{-1}\left[\langle\phi_0|\,\rho\,|\phi_0\rangle\left(-\operatorname{tr}(\phi_0 \ln \mathbb{1}/2) + \ln p_0\right) - \langle\phi_1|\,\rho\,|\phi_1\rangle\left(-\operatorname{tr}(\phi_1 \ln \mathbb{1}/2) + \ln p_1\right)\right] && (104)\\
&= \beta^{-1}\left[\operatorname{tr}(\mathcal{P}(\rho)\ln\rho^*) - \operatorname{tr}(\mathcal{P}(\rho)\ln\mathbb{1}/2)\right] = \beta^{-1}\left[\operatorname{tr}(\rho\ln\rho^*) - \operatorname{tr}(\rho\ln\mathbb{1}/2)\right], && (105)
\end{aligned}$$

where $\mathcal{P}(\rho) = \sum_i \phi_i\rho\phi_i$ the pinching map. We can also express the expected work in term of relative entropy:

$$\mathbb{E}[\Delta W] = \beta^{-1}\left[\operatorname{tr}(\rho\ln\rho^*) - \operatorname{tr}(\rho\ln\mathbb{1}/2)\right] = \beta^{-1}\left[\mathrm{D}\left(\rho\|\mathbb{1}/2\right) - \mathrm{D}(\rho\|\rho^*)\right] \ . \tag{106}$$

□

Note that in the event $\rho^* = \rho$, i.e., the agent is fully aware of the identity of the quantum state and is able to ensure the work extraction protocol to be entirely quasi-static. Then the extract work is given by

$$\begin{aligned}
\mathbb{E}[\Delta W] &= p_0\beta^{-1}\left[\mathrm{D}\left(\phi_0\|\mathbb{1}/2\right) + \ln p_0\right] + p_1\beta^{-1}\left[\mathrm{D}\left(\phi_1\|\mathbb{1}/2\right) + \ln p_1\right] && (107)\\
&= \beta^{-1}(-p_0\operatorname{tr}(\phi_0\ln\mathbb{1}/2) + p_0\ln p_0 - p_1\operatorname{tr}(\phi_1\ln\mathbb{1}/2) + p_1\ln p_1) && (108)\\
&= \beta^{-1}\left[\operatorname{tr}(\rho\ln\rho) - \operatorname{tr}(\rho\ln\mathbb{1}/2)\right] = \beta^{-1}\mathrm{D}\left(\rho\|\mathbb{1}/2\right). && (109)
\end{aligned}$$

We retrieve the full non-equilibrium free energy. Therefore, the dissipation due to agent's imperfect knowledge of the true state ρ can be quantified as

$$W_{\mathrm{diss}} = \max_{\rho^*}\mathbb{E}[\Delta W] - \mathbb{E}[\Delta W] = \beta^{-1}\mathrm{D}(\rho\|\rho^*). \tag{110}$$

C. Cumulative dissipation

In this section, we consider the setting where we have oracle sequential access to an unknown pure qubit state ψ, and our goal is to extract the maximal amount of work into a battery system. To achieve this, we can use Algorithm 1 with the rewards (91) to learn an approximate direction of the state, and then run Algorithm 2 to extract work based on the approximate input. In general, we can consider mixed-state estimator $\hat{\rho}_k$. Assuming sequential access to the unknown state over N rounds, and using the expected extracted work from Theorem SI.8, we define the dissipation at round $k \in [N]$ with respect to the optimal protocol as

$$W^k_{\mathrm{diss}} := \beta^{-1}\mathrm{D}(\psi\|\hat{\rho}_k), \tag{111}$$

and the cumulative dissipation over all N rounds as

$$W_{\mathrm{diss}}(N) := \beta^{-1}\sum_{k=1}^{N} W^k_{\mathrm{diss}} = \beta^{-1}\sum_{k=1}^{N}\mathrm{D}(\psi\|\hat{\rho}_k). \tag{112}$$

Algorithm 3: Thermal work extraction

Require: sequence $\{\epsilon_k\}_{k=1}^{\infty}$
for $k = 1, 2, \ldots$ **do**
  Receive unknown $|\psi\rangle$ and couple to battery state $\varphi(x - \mu_{k-1})$
  Compute direction $|\psi_k\rangle$ with Algorithm 1 using $\{\psi_s, r_s\}_{s=1}^{k-1}$
  Set $\hat{\rho}_k = \Delta_{2\epsilon_k}(\psi_k)$
  Extract work using Algorithm 2 with input $\hat{\rho}_k$ and get extracted work ΔW_k and energy μ_k
  Set reward $r_k = \{0, 1\}$ according to (91)

To minimize the cumulative dissipation, we use Algorithm 3, which takes as input a sequence of accuracies $\{\epsilon_k\}_{k=1}^{\infty}$. At each round, the estimator uses ψ_k, the direction output by Algorithm 1, and sets $\hat{\rho}_k = \Delta_{2\epsilon_k}(\psi_k)$, where $\Delta_{2\epsilon_k}$ is the completely depolarizing channel. If ϵ_k is a good approximation of the infidelity between the true state ψ and the estimate ψ_k, then the dissipation W_{diss}^k is controlled by ϵ_k. This is formalized in the following theorem.

Theorem SI.9 (Theorem 2.34 in [2])**.** *Let ρ and $\hat{\rho}$ be d-dimensional quantum states achieving infidelity $1 - F(\rho, \hat{\rho}) \leq \epsilon \leq \frac{1}{2}$. Then we have*

$$\mathrm{D}(\rho \| \Delta_{2\epsilon}(\hat{\rho})) \leq 16\epsilon \left(2 + \ln\left(\frac{d}{2\epsilon}\right)\right). \tag{113}$$

Given the above bound we can use the fidelity guarantee of Algorithm 1 in Theorem SI.1 to prove a bound on the cumulative dissipation.

Theorem SI.10. *Given a finite time horizon $N \in \mathbb{N}$ and $\delta \in (0, 1)$ there exists an explicit sequence of accuracies $\{\epsilon_k\}_{k=1}^{\infty}$ such that Algorithm 3 achieves*

$$W_{\text{diss}}(N) = O\left(\beta^{-1} \ln^2(N) \ln\left(\frac{N}{\delta}\right)\right). \tag{114}$$

Proof. We can choose

$$\epsilon_k = \min\left\{C \frac{\ln\left(\frac{N}{\delta}\right)}{k}, \frac{1}{2}\right\}, \tag{115}$$

where C is the constant in Theorem SI.1 for the fidelity bound of the direction ψ_k. Then we can use Theorem SI.9 combined with the fidelity guarantee of Theorem SI.1 to get that with probability at least $1 - \delta$ we have

$$W_{\text{diss}}(N) \leq \beta^{-1} \sum_{k=1}^{N} 16\epsilon_k(2 - \ln \epsilon_k). \tag{116}$$

The result follows by noting that, for a sufficiently large constant k^*, we have $\epsilon_k = C \ln\left(\frac{N}{\delta}\right)/k$ for all $k \geq k^*$. The dissipation incurred during the first k^* rounds contributes a constant term. For the remaining rounds $k \geq k^*$, using $-\ln \epsilon_k \leq \ln N$ for $k \leq N$ and summing the corresponding dissipation terms yields the claimed polylogarithmic scaling. □

D. Non-degenerate Hamiltonian

So far we have discussed the case for when the system Hamiltonian is degenerate, in general though we can consider Hamiltonian's with energy gap of ω, i.e., $H_A = \omega |E_1\rangle\langle E_1|$. In this case, there is a pre-defined energy eigenbasis, hence all the swapping operations will have to be done in such basis. As mentioned before, under the constraint of strict energy conservation, there is no way for one to extract the full non-equilibrium free energy if the initial state is not diagonalized in the energy eigenbasis. However, it is possible if we relax the constraint to energy conservation on average, in which case a unitary rotation, U_{ρ^*} can be carried out on the system qubit before the swap operations were applied. The energy difference inccured during the rotation can be shifted into the battery via the unitary,

$$U_{\rho^*} = \sum_i |E_i\rangle\langle\phi_i| \otimes \Gamma_{e_i}^B \,, \tag{117}$$

where $e_i = \mathrm{tr}(\phi_i H_A) - \omega_i$ is the energy difference. This operation obeys average energy conservation as long as the input is diagonalized along the eigenbasis of $\{\phi_i\}_i$. However, when the input is not diagonalized, which is usually the case where input is unknown, this operation is non-energy conserving even on the average and hence will require additional energy be supplied. The amount of additional work required is at least the energetic difference between the unknown state and the pinched version of it. Although a possible experimental setup was proposed in [14], it is based on the implementation of a time-dependent interaction Hamiltonian. This in turn would require additional resources to keep track of time and change the Hamiltonian smoothly [17]. More recent work has shown that such an operation can only be realized if one has access to unbounded coherence [1, 6], while this can potentially be achieved approximately using lasers, the energy from the laser also has to be accounted for. In order to avoid including unnecessary technicality and shift the focus of the paper, we focus on the degenerate Hamiltonian. Notably, for a non-degenerate Hamiltonian, the parametrization of energy gap for the thermal reservoir will be given by

$$\nu(\ell, \epsilon_k) = \beta^{-1} \ln \left(\frac{(1 - \frac{\ell}{M}(1 + e^{\beta\omega})^{-1}) - \left(1 - \frac{\ell}{M}\right) \epsilon_k}{\frac{\ell}{M}(1 + e^{\beta\omega})^{-1} + \left(1 - \frac{\ell}{M}\right) \epsilon_k} \right) . \tag{118}$$

IV. JAYNES-CUMMINGS WORK EXTRACTION PROTOCOL

Algorithm 4: Jaynes-Cummings work extraction

Require: A sequence of unknown states $|\psi\rangle$
for $k = 1, 2, \ldots$ **do**
  Receive the unknown state $|\psi\rangle$
  Compute direction $|\psi_k\rangle$ using Algorithm 1
  Expose it to a field that induces Hamiltonian $H_A = \omega |\psi_k\rangle\langle\psi_k|$.
  Turn on the interaction between the system and the battery whose interaction Hamiltonian is $H_I = \frac{\Omega}{2}(a \otimes |\psi_k\rangle\langle\psi_k^\perp| + a^\dagger \otimes |\psi_k^\perp\rangle\langle\psi_k|)$ for a time $t_k = \pi\Omega^{-1}(n_k + 1)^{-\frac{1}{2}}$.
  Measure the battery energy to obtain n_{k+1}
  Set reward $r_k = \{0, 1\}$ according to Eq. (129)

We consider the Jaynes-Cummings work extraction protocol in Algorithm 4 which extracts the energy from a field into a battery.

The systems involved in this protocol includes a system in an unknown state $|\psi\rangle$ with a tunable Hamiltonian in the form of $H_A = \nu |\phi\rangle\langle\phi|$ where ν and ϕ can be controlled by the strength and the direction of a field, respectively, as well as a battery system with the Hamiltonian of the battery given by $H_B = \omega a^\dagger a$ where $a = \sum_{n=1}^{\infty} \sqrt{n} |n-1\rangle\langle n|$ and $a^\dagger = \sum_{n=0}^{\infty} \sqrt{n+1} |n+1\rangle\langle n|$. We are also allowed to turn on an interaction $H_I = \frac{\Omega}{2}(a \otimes |\psi_k\rangle\langle\psi_k^\perp| + a^\dagger \otimes |\psi_k^\perp\rangle\langle\psi_k|)$ between the system and the battery.

The protocol works in round $k = 1, 2, \ldots, N$.

In each round k, we possess a battery state $|n_k\rangle$, receive an unknown state $|\psi\rangle$ and compute a direction $|\psi_k\rangle$ from previous records of $\{n_s\}_{s=1}^k$.

We then expose the system to a field that induces Hamiltonian $H_A = \omega |\psi_k\rangle\langle\psi_k|$, which transfer energy from the field to the system.

We turn on the interaction H_I between the system and the battery for time $t_k = \pi\Omega^{-1}(n_k + 1)^{-\frac{1}{2}}$. The time evolution of the system and the battery is under the total Hamiltonian

$$H = \omega(|\psi_k\rangle\langle\psi_k| + a^\dagger a) + \frac{\Omega}{2}(a \otimes |\psi_k\rangle\langle\psi_k^\perp| + a^\dagger \otimes |\psi_k^\perp\rangle\langle\psi_k|). \tag{119}$$

This is exactly given by the famous Jaynes-Cummings model [4], which is nowadays a textbook model [12]. One may verify that the eigenstates of the total Hamiltonian are $|0\rangle |\psi_k^\perp\rangle$ as well as

$$|n, +\rangle = \frac{1}{\sqrt{2}}(|n-1\rangle |\psi_k\rangle + |n\rangle |\psi_k^\perp\rangle), \quad |n, -\rangle = \frac{1}{\sqrt{2}}(|n-1\rangle |\psi_k\rangle - |n\rangle |\psi_k^\perp\rangle), \tag{120}$$

for $n = 1, 2, \ldots$ and the eigenvalues are respectively $E_0 = 0$ and

$$E_{n+} = n\omega + \frac{\Omega}{2}\sqrt{n}, \quad E_{n-} = n\omega - \frac{\Omega}{2}\sqrt{n}. \tag{121}$$

for $n = 1, 2, \ldots$. The state $|n_k\rangle |\psi\rangle$ is decomposed into 4 eigenstates $|n_k, \pm\rangle$ and $|n_k + 1, \pm\rangle$, i.e.

$$|n_k\rangle |\psi\rangle = \frac{1}{\sqrt{2}}\langle\psi_k^\perp|\psi\rangle(|n_k, +\rangle - |n_k, -\rangle) + \frac{1}{\sqrt{2}}\langle\psi_k|\psi\rangle(|n_k + 1, +\rangle + |n_k + 1, -\rangle). \tag{122}$$

These eigenstates gains phase factors during the time evolution. After time $t_k = \pi\Omega^{-1}(n_k + 1)^{-\frac{1}{2}}$, the state evolves to, up to an irrelevant global phase,

$$e^{-iHt_k} |n_k\rangle |\psi\rangle = \frac{1}{\sqrt{2}}\langle\psi_k^\perp|\psi\rangle(e^{i\theta_k} |n_k, +\rangle - e^{-i\theta_k} |n_k, -\rangle) + \frac{1}{\sqrt{2}} e^{i\frac{\omega\pi}{\Omega\sqrt{n_k+1}}} \langle\psi_k|\psi\rangle(i |n_k + 1, +\rangle - i |n_k + 1, -\rangle) \tag{123}$$

$$= \langle\psi_k^\perp|\psi\rangle(i \sin\theta_k |n_k - 1\rangle |\psi_k\rangle + \cos\theta_k |n_k\rangle |\psi_k^\perp\rangle) + \frac{1}{\sqrt{2}} i e^{i\frac{\omega\pi}{\Omega\sqrt{n_k+1}}} \langle\psi_k|\psi\rangle |n_k + 1\rangle |\psi_k^\perp\rangle, \tag{124}$$

where $\theta_k = \frac{\pi}{2}\sqrt{\frac{n_k}{n_k+1}}$.

We finally measure the battery energy. The measurement outcome is n_{k+1}. The probability distribution of n_{k+1} is given by

$$\Pr[n_{k+1}] = \begin{cases} |\langle\psi_k|\psi\rangle|^2, & n_{k+1} = n_k + 1, \\ |\langle\psi_k^\perp|\psi\rangle|^2 \cos^2\theta_k, & n_{k+1} = n_k, \\ |\langle\psi_k^\perp|\psi\rangle|^2 \sin^2\theta_k, & n_{k+1} = n_k - 1. \end{cases} \tag{125}$$

The extracted work is defined as $\Delta W_k = \omega(n_{k+1} - n_k)$. The expected extracted work is given by

$$\mathbb{E}[\Delta W_k] = \omega(|\langle\psi_k|\psi\rangle|^2 - |\langle\psi_k^\perp|\psi\rangle|^2 \sin^2\theta_k) = \omega(|\langle\psi_k|\psi\rangle|^2(1 + \sin^2\theta_k) - \sin^2\theta_k), \tag{126}$$

where we have used $|\langle\psi_k^\perp|\psi\rangle|^2 = 1 - |\langle\psi_k|\psi\rangle|^2$. It is obvious that $\mathbb{E}[\Delta W_k]$ increases as $|\langle\psi_k|\psi\rangle|^2$ increases, therefore, $\mathbb{E}[\Delta W_k]$ is maximized at $|\langle\psi_k|\psi\rangle| = 1$,

$$\max_{|\psi_k\rangle} \mathbb{E}[\Delta W_k] = \omega. \tag{127}$$

The dissipation in this round is thus given by the difference between the maximal and the actual expected extracted work

$$W_{\text{diss}}^{\text{jc},k} = \max_{|\psi_k\rangle} \mathbb{E}[\Delta W_k] - \mathbb{E}[\Delta W_k] = \omega(1 + \sin^2\theta_k^2)(1 - |\langle\psi_k|\psi\rangle|^2) \leq 2\omega(1 - |\langle\psi_k|\psi\rangle|^2). \tag{128}$$

The correspondence between the reward measurement and the battery energy measurement is given by

$$r_k = \begin{cases} 1, & n_{k+1} = n_k + 1, \\ 0, & \text{otherwise.} \end{cases} \tag{129}$$

The probability distribution of the reward is

$$\Pr[R_k = r_k] = \begin{cases} |\langle\psi_k|\psi\rangle|^2, & r_k = 1, \\ |\langle\psi_k^\perp|\psi\rangle|^2, & r_k = 0. \end{cases} \tag{130}$$

The regret in this round is thus

$$\text{Rgrt}^k = 1 - |\langle\psi_k|\psi\rangle|^2. \tag{131}$$

The dissipation and the regret in this round is thus related by

$$W_{\text{diss}}^{\text{jc},k} \leq 2\omega \,\text{Rgrt}^k. \tag{132}$$

The cumulative dissipation over N rounds is thus

$$W_{\text{diss}}^{\text{jc}}(N) = \sum_{k=1}^{N} W_{\text{diss}}^{\text{jc},k} \leq 2\omega \sum_{k=1}^{N} (1 - |\langle\psi_k|\psi\rangle|^2). \tag{133}$$

Which is related to the cumulative dissipation in extracting work from knowledge via

$$W_{\text{diss}}^{\text{jc}}(N) \leq 2\omega \,\text{Rgrt}(N), \tag{134}$$

Theorem SI.11. *There exists an explicit protocol for extracting work from knowledge that adaptively updates the estimate $|\psi_k\rangle$ based on the rewards $\{r_s\}_{s=1}^{k-1}$, achieving, with probability at least $1 - \delta$,*

$$W_{\text{diss}}^{jc}(N) = O\left(\omega \ln(N) \ln\left(\frac{N}{\delta}\right)\right). \tag{135}$$

V. COST OF MEASUREMENT AND ERASURE

The measurement as well as erasure of memory of the agent does not come for free. The cost of measurement along with the cost of memory erasure can be lower bounded by a quantity known as *QC-mutual information*, I_{QC}, which is a measure of how much information a measurement carries about the quantum system [11]. I_{QC} is then upper bounded by the entropy of the classical memory register itself. Hence the total cost of measurement and erasure can be loosely lower bounded by just the entropy of the memory register itself. By Landauer's principle [3, 11, 16], the heat dissipation required to erase the register is lower bounded by the entropy change ΔS of the memory register via

$$\beta Q \geq \Delta S. \tag{136}$$

Furthermore, it is widely accepted that the lower bound can be achieved in the quasi-static limit when the state to be erased is known [5, 9, 10]. In our work extraction model, we repeatedly perform measurements in the energy eigenbasis, which requires memory erasure in order to store new measurement outcomes. With the assumption on the probability distribution of work values scaling linearly with fidelity, when $|\psi\rangle$ is known, the measurement outcome is deterministic and there is no energy dissipation. However, when $|\psi\rangle$ is unknown, the measurement outcome is stochastic and there is energy dissipation. This may also be taken into account in the dissipation, that is,

$$W'_{\text{diss}}(N) = W_{\text{diss}}(N) + \beta^{-1}\sum_{k=1}^{N}\Delta S_k, \tag{137}$$

where ΔS_k is the entropy change of the memory register in round k. Let $\epsilon_k = 1 - |\langle\psi_k|\psi\rangle|^2$ be the infidelity in round k. ΔS_k is closely related to ϵ_k in both models we consider. In the semi-classical battery model, the entropy change is upper bounded by

$$\Delta S_k = -(1-\epsilon_k)\ln(1-\epsilon_k) - \epsilon_k \ln \epsilon_k \leq \epsilon_k - \epsilon_k \ln \epsilon_k, \tag{138}$$

and the dissipation is upper bounded by

$$W^{\text{sc},*}_{\text{diss}}(N) = W^{\text{sc}}_{\text{diss}}(N) + \beta^{-1}\sum_{k=1}^{N}(\epsilon_k - \epsilon_k \ln \epsilon_k). \tag{139}$$

In the Jaynes-Cummings battery model, the entropy change in round k is upper bounded by

$$\Delta S_k = -(1-\epsilon_k)\ln(1-\epsilon_k) - \epsilon_k \ln \epsilon_k - \epsilon_k(\cos^2\theta_k \ln\cos^2\theta_k + \sin^2\theta_k \ln\sin^2\theta_k) \leq 2\epsilon_k - \epsilon_k \ln \epsilon_k, \tag{140}$$

and the dissipation is upper bounded by

$$W^{\text{jc},*}_{\text{diss}}(N) = W^{\text{jc}}_{\text{diss}}(N) + \beta^{-1}\sum_{k=1}^{N}(2\epsilon_k - \epsilon_k \ln \epsilon_k). \tag{141}$$

The quantum state tomography algorithm in [7] ensures a polylogarithmic cumulative infidelity with a high probability. Now we focus on the case where the cumulative infidelity is upper bounded by $\ln \frac{N}{\delta} \ln N$ with high probability $1-\delta$, that is, for some constant C,

$$\sum_{k=1}^{N}\epsilon_k \leq C \ln\frac{N}{\delta}\ln N. \tag{142}$$

By taking the derivative $(-\epsilon_k \ln \epsilon_k)' = -\ln\epsilon_k - 1$, it is obvious that when $0 < \epsilon_k \leq e^{-1}$, $-\epsilon_k \ln\epsilon_k$ increases with ϵ_k while $e^{-1} \leq \epsilon_k \leq 1$, $-\epsilon_k \ln\epsilon_k$ decreases with ϵ_k. Now we consider two cases:

Case 1: $C(\ln(N/\delta)\ln N)/N \leq e^{-1}$**.** This happens when N is large enough. As a result, $-\epsilon_k \ln\epsilon_k$ monotonically increase with respect to ϵ_k when $\epsilon_k \leq C(\ln(N/\delta)\ln N)/N$. On the one hand, for $\epsilon_k \leq C(\ln(N/\delta)\ln N)/N$,

$$-\epsilon_k \ln\epsilon_k \leq \frac{C\ln\frac{N}{\delta}\ln N}{N}\ln\frac{N}{C\ln\frac{N}{\delta}\ln N}, \tag{143}$$

On the other hand, for $\epsilon_k > C(\ln(N/\delta)\ln N)/N$,

$$-\epsilon_k \ln\epsilon_k \leq \epsilon_k \ln\frac{N}{C\ln\frac{N}{\delta}\ln N}, \tag{144}$$

Then,

$$-\sum_{k=1}^{N}\epsilon_k\ln\epsilon_k = -\sum_{\epsilon_k:\epsilon_k\leq\frac{C\ln\frac{N}{\delta}\ln N}{N}}\epsilon_k\ln\epsilon_k - \sum_{\epsilon_k:\epsilon_k>\frac{C\ln\frac{N}{\delta}\ln N}{N}}\epsilon_k\ln\epsilon_k \tag{145}$$

$$\leq \sum_{\epsilon_k:\epsilon_k\leq\frac{C\ln\frac{N}{\delta}\ln N}{N}}\frac{C\ln\frac{N}{\delta}\ln N}{N}\ln\frac{N}{C\ln\frac{N}{\delta}\ln N} + \sum_{\epsilon_k:\epsilon_k>\frac{C\ln\frac{N}{\delta}\ln N}{N}}\epsilon_k\ln\frac{N}{C\ln\frac{N}{\delta}\ln N} \tag{146}$$

$$\leq C\ln\frac{N}{\delta}\ln N\ln\frac{N}{C\ln\frac{N}{\delta}\ln N} + C\ln\frac{N}{\delta}\ln N\ln\frac{N}{C\ln\frac{N}{\delta}\ln N} \leq 2C\ln\frac{N}{\delta}\ln N\ln\frac{N}{C\ln\frac{N}{\delta}\ln N}. \tag{147}$$

Case 2: $C(\ln(N/\delta)\ln N)/N > e^{-1}$**.** This happens when N is not very large. Therefore, $N/e \leq C\ln(N/\delta)\ln N$ and certainly $N \geq 2$. Using the fact that $-\epsilon_k\ln\epsilon_k \leq e^{-1}$, we obtain

$$\sum_{t=1}^{N}\epsilon_k\ln\frac{1}{\epsilon_k} \leq \frac{N}{e} \leq C\ln\frac{N}{\delta}\ln N. \tag{148}$$

Combining both cases, we conclude that

$$-\sum_{t=1}^{N}\epsilon_k\ln\epsilon_k \leq O((\ln N)^3). \tag{149}$$

Substituting into Eq. (137), we obtain the dissipation for for the semi-classical model,

$$W_{\text{diss}}^{\text{sc},*}(N) \leq W_{\text{diss}}^{\text{sc}}(N) + \beta^{-1}O((\ln N)^3) = O((\ln N)^3). \tag{150}$$

and the Jaynes-Cummings battery model,

$$W_{\text{diss}}^{\text{jc},*}(N) \leq W_{\text{diss}}^{\text{jc}}(N) + 2\beta^{-1}O((\ln N)^3) = O((\ln N)^3), \tag{151}$$

Therefore, the regret still scales as $O(\text{polylog}(N))$ even if we take the energy dissipation due to Landauer's principle into account.

As mentioned above, in order to achieve Landauer's bound, the erasure process must necessarily be done quasi-statically. This is not something that a sequential adaptive agent can accomplish since it needs to measure and decide on its action in real time. To circumvent this, an array of empty memory registers can be first prepared. Suppose the agent will be extracting work for N steps, we first prepare a memory register M_0, consisting of N empty registers.

$$M_0 = \{\underbrace{0,0,\ldots,0}_{N}\} \tag{152}$$

At every time step, rather than erasing the old memory that has the previous outcome remembered, the agent simply records the outcome of the measurement on the battery into a new empty register. In order words, after time step t, the memory register takes the form

$$M_t = \{r_1, r_2, \ldots, r_t, 0, \ldots, 0\} \quad \text{for} \quad t \in \{1, 2, \ldots N\}\ . \tag{153}$$

At the end of all the extractions, the memory M_N can then be quasi-statically reset. As previously calculated, as $t \to N$, the distribution of r_t becomes more peaked and the total entropy of the memory registers scale with $O(\ln N)^3$, likewise for the cost of measurement. The resultant dissipation therefore still scales with $O(\text{polylog}(N))$.

[1] J. Åberg. *"Catalytic Coherence"*. Phys. Rev. Lett. **113**: 150402 (2014).

[2] S. T. Flammia and R. O'Donnell. *"Quantum chi-squared tomography and mutual information testing"*. Quantum **8**: 1381 (2024).

[3] J. Goold, M. Paternostro, and K. Modi. *"Nonequilibrium Quantum Landauer Principle"*. Phys. Rev. Lett. **114**: 060602 (2015).

[4] E. Jaynes and F. Cummings. *"Comparison of quantum and semiclassical radiation theories with application to the beam maser"*. Proceedings of the IEEE **51**(1): 89–109 (1963).

[5] Y. Jun, M. c. v. Gavrilov, and J. Bechhoefer. *"High-Precision Test of Landauer's Principle in a Feedback Trap"*. Phys. Rev. Lett. **113**: 190601 (2014).

[6] K. Korzekwa, M. Lostaglio, J. Oppenheim, and D. Jennings. *"The extraction of work from quantum coherence"*. New Journal of Physics **18**(2): 023045 (2016).

[7] J. Lumbreras, M. Terekhov, and M. Tomamichel. *"Learning pure quantum states (almost) without regret"*. arXiv preprint arXiv:2406.18370 , (2024).

[8] J. Lumbreras and M. Tomamichel. *"Linear bandits with polylogarithmic minimax regret"*. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 3644–3682, (2024).

[9] H. J. D. Miller, G. Guarnieri, M. T. Mitchison, and J. Goold. *"Quantum Fluctuations Hinder Finite-Time Information Erasure near the Landauer Limit"*. Phys. Rev. Lett. **125**: 160602 (2020).

[10] P. M. Riechers and M. Gu. *"Initial-state dependence of thermodynamic dissipation for any quantum process"*. Phys. Rev. E **103**: 042145 (2021).

[11] T. Sagawa and M. Ueda. *"Minimal Energy Cost for Thermodynamic Information Processing: Measurement and Information Erasure"*. Phys. Rev. Lett. **102**: 250602 (2009).

[12] M. O. Scully and M. S. Zubairy. *Atom-field interaction - quantum theory*, page 193–219. Cambridge University Press (1997).

[13] H. Shao, X. Yu, I. King, and M. R. Lyu. *"Almost optimal algorithms for linear stochastic bandits with heavy-tailed payoffs"*. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8430–8439, Red Hook, NY, USA(2018).

[14] P. Skrzypczyk, A. J. Short, and S. Popescu. *"Work extraction and thermodynamics for individual quantum systems"*. Nature communications **5**(1): 4185 (2014).

[15] G. Strang. *Calculus*. Wellesley-Cambridge Press (2019).

[16] V. V. Tan and K. Saito. *"Finite-Time Quantum Landauer Principle and Quantum Coherence"*. Phys. Rev. Lett. **128**: 010602 (2022).

[17] M. P. Woods and M. Horodecki. *"Autonomous Quantum Devices: When Are They Realizable without Additional Thermodynamic Costs?"*. Phys. Rev. X **13**: 011016 (2023).