

Quantum state-agnostic work extraction (almost) without dissipation

Josep Lumbraeras,^{1,*} Ruo Cheng Huang,^{2,†} Yanglin Hu (胡杨林),^{1,‡} Mile Gu,^{1,2,3,§} and Marco Tomamichel^{1,3,4,¶}

¹Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore

²Nanyang Quantum Hub, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore

³MajuLab, CNRS-UNS-NUS-NTU International Joint Research Unit, UMI 3654, 117543, Singapore

⁴Department of Electrical and Computer Engineering, National University of Singapore

(Dated: June 13, 2025)

We investigate work extraction protocols designed to transfer the maximum possible energy to a battery using sequential access to N copies of an unknown pure qubit state. The core challenge is designing interactions to optimally balance two competing goals: charging of the battery optimally using the qubit in hand, and acquiring more information by qubit to improve energy harvesting in subsequent rounds. Here, we leverage exploration-exploitation trade-off in reinforcement learning to develop adaptive strategies achieving energy dissipation that scales only poly-logarithmically in N . This represents an exponential improvement over current protocols based on full state tomography.

Introduction—Given sequential access to finite, identical samples of an unknown quantum system, what is the optimal strategy for extracting work from them and charging a battery? The extraction of useful work from available resources has long been a central problem in classical thermodynamics and continues to attract significant attention in the quantum domain. While numerous protocols have been proposed [2–4, 6, 25], the majority assume that the agent has full knowledge of the quantum state—these are so-called *state-aware* protocols. In practice we may not always know how the state is prepared, leading us to the *state-agnostic* scenario at hand. In this case, how can we design a protocol for work extraction?

A natural strategy might involve first performing quantum state tomography to estimate the unknown state, followed by work extraction based on this estimate [11, 31, 32]. Indeed, knowledge of the quantum state is essential for efficient work extraction; consider, for example, Szilard’s engine, where the information about the system’s configuration—such as the position of a particle—enables work to be extracted [27]. However, with only finitely many samples available, any estimate of the true state has statistical uncertainty [9, 21], resulting in unavoidable heat dissipation during work extraction [22]. Furthermore, quantum systems that are measured during state tomography are no longer available for work extraction and therefore contribute to irreversible heat dissipation. Combining these observations, we will show that any two-step procedure that first uses tomography to inform an efficient work extraction procedure will lead to a cumulative dissipation that scales at least as $\Omega(\sqrt{N})$ in the number N of copies of the unknown state. Such strategies are only minimally adaptive though and this raises a natural question: *Can more adaptive strategies,*

which simultaneously balance learning and work extraction, offer better performance in terms of cumulative dissipation? We answer this in the affirmative, showing an exponential improvement for pure qubit states.

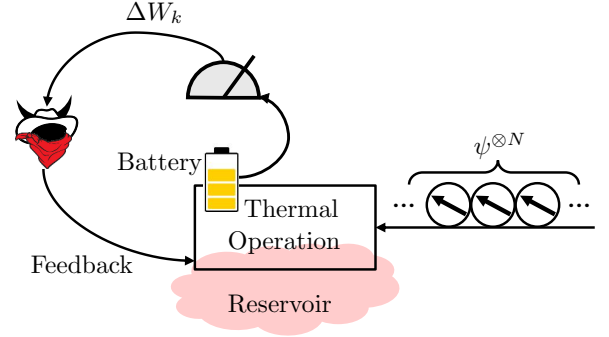


Figure 1. Sketch of the sequential work extraction protocol with a thermal reservoir. At each time step $k \in [N]$, the agent receives a copy of an unknown qubit state ψ and performs a thermal operation involving the reservoir and a battery. A measurement in the battery system is carried out to determine the extracted work ΔW_k , which is then used as feedback to improve the extraction strategy in subsequent rounds.

More specifically, we consider a source that sequentially emits a total of N identical unknown pure qubit states, which are used to charge a battery using an interaction that couples the qubit and battery to a thermal reservoir. For this we adopt a semi-classical thermodynamic model where the battery is represented as a classical weight that stores energy via vertical displacement. We then design a *state-agnostic* protocols that controls the interaction between the qubit and the battery depending on data collected in previous rounds and examine how the cumulative dissipation scales with the number of available samples. We find that adaptive strategies can result in an exponential improvement of the cumulative dissipation: whereas tomography-based algorithms yield a cumulative dissipation that scales as $\Omega(\sqrt{N})$, we show that a scaling of $O(\text{polylog}(N))$ is possible for suitable adaptive work-extraction algorithms for pure states.

* josep.lumbraeras@u.nus.edu

† ruocheng001@e.ntu.edu.sg

‡ yanglin.hu@u.nus.edu

§ mgu@quantumcomplexity.org

¶ marco.tomamichel@nus.edu.sg

Our main technical contribution is an upper bound on the cumulative dissipation in this model. This result builds on a novel connection between adaptive quantum control and classical reinforcement learning [15], which we believe to be of independent interest. More precisely, we show that the dissipation in our framework can be interpreted as the *regret* incurred during the learning of pure quantum states, formulated within a multi-armed bandit setting [17, 18].

Extracting work from knowledge—We consider a commonly studied extraction model, where the goal is to extract energy from a qubit system in contact with a thermal bath [10, 25]. Energy extraction is quantified by its use to raise the potential energy of an external battery — represented by the lifting of an external weight. Here, we assume that the qubit system is initially unknown. The three physical subsystems involved in the process are:

- **Unknown pure state source (System A):** A qubit in a pure state $\psi_A = |\psi\rangle\langle\psi|$ with a degenerate Hamiltonian $H_A = \omega \mathbb{1}/2$. This is the system from which free energy is to be extracted.
- **Battery (System B):** A semi-classical weight described by a continuous-variable state $\varphi(x) \in L^2(\mathbb{R})$. The battery Hamiltonian is defined as $H_B\varphi(x) = x\varphi(x)$, where x represents the height of the weight. The energy of the battery can be changed by translating the weight up by a certain height x_0 , described by the translation operator $\Gamma_{x_0}^B\varphi(x) = \varphi(x - x_0)$.
- **Thermal reservoir (System R):** A heat bath at a fixed inverse temperature β , modeled as a supply of qubit states $\gamma_\beta(\nu) = Z_R(\nu)^{-1}e^{-\beta H_R(\nu)}$, where the Hamiltonian is $H_R(\nu) = \nu|1\rangle\langle 1|$ and $Z_R(\nu) = \text{tr}(e^{-\beta H_R(\nu)})$ is the partition function. Here, $|0\rangle$ and $|1\rangle$ are energy eigenstates, and ν is the energy gap. This energy gap can be tuned.

We consider a degenerate system Hamiltonian, such that every orthogonal basis is an energy eigenbasis. This choice avoids the phenomenon of “work-locking” — the inability of thermal operations to extract work from coherence in the energy eigenbasis [16]. In contrast, systems with non-degenerate Hamiltonian requires additional energy expenditure. Further details are provided in Appendix B.

We consider an agent with oracle access to the unknown system A over a finite number of rounds $N \in \mathbb{N}$, indexed by $k \in [N]$. The goal is to convert the maximum amount of free energy from A and store it in the battery B through interactions with a thermal reservoir R . However, since the state ψ of the system is unknown, the agent cannot extract work optimally from the outset. Instead, it must gradually improve its strategy by learning from each round, where the learning is done by measuring the battery. The work extraction protocol is defined by two key components: a policy that

updates the agent’s guess ψ_k of the unknown state ψ at each round, and a sequence $\{\epsilon_k\}_{k=1}^N$ that determines the accuracy of these guesses. Both quantities can be chosen adaptively and together they determine the dissipation at each round. We first present the general structure of the protocol and later discuss how the dissipation depends on these choices, and how to optimize them for best performance. The protocol proceeds as follows at each round $k \in [N]$:

1. The agent receives a sample of the unknown qubit state ψ .
2. Based on the outcomes from previous rounds, the agent selects a direction ψ_k on the Bloch sphere, sets an accuracy $\epsilon_k \in [0, 1]$ and defines a basis $\{\psi_k, \psi_k^\perp\}$ for system A . This computation is done using the previously selected directions and measured battery energies $\{\psi_s, \mu_s\}_{s=1}^{k-1}$.
3. We first implement a unitary on the system qubit in the form of

$$U_k = |0\rangle\langle\psi_k| + |1\rangle\langle\psi_k^\perp|, \quad (1)$$

satisfying $[H_A, U_k] = 0$ which tries to diagonalize the system qubit in computational basis.

4. The agent then performs a thermal operation by repeatedly appending a reservoir qubit R , applying an energy-conserving unitary on the combined system ABR to transfer energy from R to the battery B and discarding R . Specifically, for each of M repetitions (indexed by $\tau \in [M]$), the agent:
 - Sets the energy gap of the reservoir qubit to $\nu(\tau, \epsilon_k)$ (see Eq. (6)) and gets a fresh qubit state $\gamma_\beta(\nu(\tau, \epsilon_k))$.
 - Applies the following unitary

$$V_{\psi_k, \tau} = \sum_{i,j} |i\rangle\langle j|_A \otimes |j\rangle\langle i|_R \otimes \Gamma_{(i-j)\nu(\tau, \epsilon_k)}^B, \quad (2)$$

satisfying $[H_A + H_B + H_R, V_{\psi_k, \tau}] = 0$ to ABR .

- Discards the reservoir qubit.
5. After completing the M steps, the agent measures the energy of the battery in its eigenbasis and records the energy μ_k .

For the above protocol, the expected work extracted at round $k \in [N]$, as a function of the chosen direction ψ_k and accuracy parameter ϵ_k , is given by

$$\mathbb{E}[\Delta W_k] = \beta^{-1} [\text{D}(\psi \| \mathbb{1}/2) - \text{D}(\psi \| \Delta_{2\epsilon_k}(\psi_k))], \quad (3)$$

where $\Delta_\epsilon(\rho) = (1 - \epsilon)\rho + \epsilon \mathbb{1}/2$ denotes the depolarizing channel and $\text{D}(\rho \| \sigma) = \text{tr}(\rho(\ln \rho - \ln \sigma))$ is the quantum relative entropy. The maximal expected work per round, $\beta^{-1}\text{D}(\psi \| \mathbb{1}/2)$, is achieved when the agent perfectly estimates the state, i.e., when $\psi_k = \psi$ and $\epsilon_k = 0$. Accordingly, we define the dissipation at round k as:

$$W_{\text{diss}}^k := \max_{\psi_k, \epsilon_k} \mathbb{E}[\Delta W_k] - \mathbb{E}[\Delta W_k] = \beta^{-1}\text{D}(\psi \| \Delta_{2\epsilon_k}(\psi_k)). \quad (4)$$

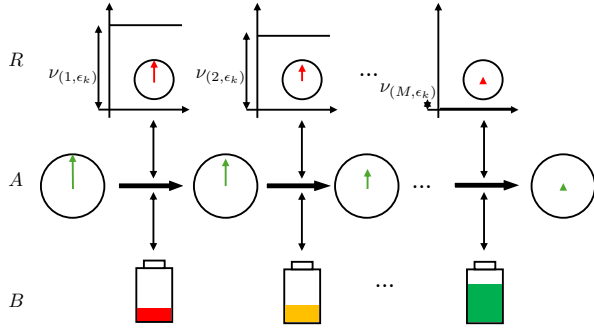


Figure 2. Illustration of the repetitions of the thermal operation in the full system ABR , where arrows represent Bloch vectors of states, showing that the system qubit becomes more and more mixed as the process goes. The energy gaps $\{\nu_{k,i}\}_i$ of successive reservoir Hamiltonian forms a strictly decreasing sequence, making the successive thermal states more mixed. At each step, we take a new qubit from the reservoir and swap the system qubit with the reservoir qubit, the energy from reservoir will flow into the battery. At the end of the process, the qubit in system A is the thermal state.

This expression highlights a key trade-off: to reduce dissipation, the agent must align ψ_k with the true state ψ , but it cannot set ϵ_k too small unless the estimate is sufficiently accurate, as otherwise the divergence becomes large. The parameter ϵ_k thus plays a dual role, quantifying both the uncertainty in the estimate and its thermodynamic penalty. The agent's objective is to extract the maximum amount of work into the battery using the N copies of the unknown system. Equivalently, the agent aims to minimize the cumulative dissipation over N rounds, which is given by

$$W_{\text{diss}}(N) = \beta^{-1} \sum_{k=1}^N W_{\text{diss}}^k = \sum_{k=1}^N D(\psi \| \Delta_{2\epsilon_k}(\psi_k)). \quad (5)$$

Going back to the protocol we note that the unitary $V_{\psi_k, \tau}$ in Eq. (2) swaps the states of the unknown system A and the thermal reservoir R . This operation induces an energy exchange determined by the energy gap $\nu(\tau, \epsilon_k)$ between the two systems. The energy difference is transferred to the battery through a translation operation $\Gamma_{(i-j)\nu(\tau, \epsilon_k)}^B$ to conserve the total energy of the combined system.

To maximize the extractable non-equilibrium free energy from system A , it is essential to minimize dissipation during the energy transfer to the battery. This can be achieved by ensuring that the process is quasi-static — that is, slow and nearly reversible. Under such conditions, the system qubit remains close to the thermal state of the reservoir's Hamiltonian, thereby suppressing heat flow during the interaction. Conversely, if the process is carried out rapidly, i.e., the system state deviates from reservoir's thermal state, the system will appear out of thermal equilibrium with the reservoir, resulting in heat exchange, which contributes to the entropy production

during the protocol. The fundamental trade-off between efficiency and power in such protocols is well established: in general, the reduction in work output scales as $1/\tau_0$ where τ_0 denotes time taken for extraction protocol [30]. This can also be characterized by discrete number of interactions required to implement the transformation [29].

To approximate a quasi-static process, we repeat the thermal interaction M times within each round. In the limit $M \rightarrow \infty$, the process becomes effectively reversible. At each repetition $\tau \in [M]$, we introduce a slight mismatch between the energy levels of system A and the reservoir R by varying the energy gap. We use the following parametrization of the gap

$$\nu(\tau, \epsilon_k) = \beta^{-1} \ln \left(\frac{1 - \frac{\tau}{2M} - (1 - \frac{\tau}{M}) \epsilon_k}{\frac{\tau}{2M} + (1 - \frac{\tau}{M}) \epsilon_k} \right), \quad (6)$$

where $\epsilon_k \in [0, 1]$ controls the mismatch. The choice of ϵ_k will be discussed later for specific protocols. Further details on this gap parametrization can be found in Appendix B. The full interaction of systems A , B , and R through the unitary $V_{\psi_k, \tau}$ is illustrated in Figure 2.

To compute the total extracted work, we evaluate at each repetition $\tau \in [M]$ how much energy is transferred to the battery. This is done by considering the two initial input states that can trigger a swap operation, namely $|0\rangle_A |1\rangle_R$ and $|1\rangle_A |0\rangle_R$. When the unitary Eq. (2) is applied repeatedly, there is a certain probability that the states of the system and reservoir qubits are swapped, thereby enabling energy exchange with the battery. In the quasi-static limit $M \rightarrow \infty$, each of these initial states contributes a fixed amount of energy $\Delta W_k = \mu_{k+1} - \mu_k$ to the battery, which are respectively

$$\begin{aligned} w_{k,0} &:= \beta^{-1} (D(\psi_k \| 1/2) + \ln(1 - \epsilon_k)), \\ w_{k,1} &:= \beta^{-1} (D(\psi_k^\perp \| 1/2) + \ln \epsilon_k). \end{aligned} \quad (7)$$

To build intuition for $w_{k,i}$, consider $w_{k,0}$ as an example. This represents the extracted work given ψ_k . The relative entropy $D(\psi_k \| 1/2)$ reflects the non-equilibrium free energy of ψ_k , which corresponds to the extracted work in the quasi-static limit. The logarithmic term arises because the process is not entirely quasi-static — it is quasi-static for all but the first repetition. During the first repetition, the system qubit $|0\rangle\langle 0|$ is swapped with the reservoir qubit $(1 - \epsilon_k) |0\rangle\langle 0| + \epsilon_k |1\rangle\langle 1|$. This swap generates entropy $\Delta S = -\epsilon_k \ln \epsilon_k - (1 - \epsilon_k) \ln(1 - \epsilon_k)$ and yields work extraction $\Delta U = -\beta^{-1} \epsilon_k \nu(1, \epsilon_k) = -\beta^{-1} \epsilon_k \ln \frac{\epsilon_k}{1 - \epsilon_k}$. Combining both, the resulting change in free energy of the system and the battery is $\Delta U - \beta^{-1} \Delta S = \ln(1 - \epsilon_k)$, which precisely accounts for the logarithmic term. A similar argument holds for $w_{k,1}$.

We now consider the case where the initial state is ψ , the total energy extracted will be distributed as

$$\Pr(\Delta W_k = w_{k,i}) = i + (-1)^i |\langle \psi | \psi_k \rangle|^2, \quad (8)$$

which combined with Eq. (7) gives the expected extracted work stated in Eq. (3). More details of this computation are provided in Appendix B.

Extracting work after learning—An attentive reader may notice that achieving sublinear dissipation in the number of copies N is, in principle, not difficult. The sequence of battery measurement outcomes follows the distribution Eq. (8), which corresponds to a measurement in the direction ψ_k on the unknown state ψ . Hence, these outcomes can be used to perform state tomography of ψ and then select ψ_k and ϵ_k that minimizes $D(\psi \| \Delta_{2\epsilon_k}(\psi_k))$. A natural strategy is to allocate a fraction αN , with $0 \leq \alpha \leq 1$, of the available copies to learning the unknown state and producing an estimate that will fix a direction $\hat{\psi}$ and $\hat{\epsilon}$, then use the remaining $(1 - \alpha)N$ copies for work extraction by setting $\psi_k = \hat{\psi}$ and $\epsilon_k = \hat{\epsilon}$. However, such a two-phase approach is fundamentally limited by the sample complexity needed to achieve accuracy in relative entropy. Standard tomography algorithms are not designed to minimize dissipation during learning and typically result in linear dissipation. The cumulative dissipation for this approach can be lower bounded as

$$W_{\text{diss}}(N) = \Omega(\alpha N + (1 - \alpha)N\epsilon_{\alpha N}), \quad (9)$$

where $\epsilon_{\alpha N}$ is the smallest achievable relative entropy error using αN copies. This error is lower bounded as $\epsilon_{\alpha N} = \Omega(\frac{1}{\alpha N})$ for αN samples [5, 7, 9], which implies $W_{\text{diss}}(N) = \Omega(\alpha N + \frac{1}{\alpha})$. Minimizing this expression over α yields a lower bound of $W_{\text{diss}}(N) = \Omega(\sqrt{N})$. Moreover, a matching upper bound $W_{\text{diss}}(N) = O(\sqrt{N})$ can be achieved using an optimal learning algorithm for relative entropy [7]. This naturally raises the question of whether the dissipation for pure states can be further reduced through more adaptive or integrated strategies, or whether the $\Omega(\sqrt{N})$ scaling already represents the fundamental limit in this setting.

Extracting work while learning—Indeed, using a more adaptive approach, it is possible to surpass the $\sqrt{N}$ barrier. To achieve this, we use the fully adaptive state tomography algorithm studied in [18], which is based on the multi-armed bandit algorithm from [19]. The algorithm assumes sequential access to N identical copies of an unknown pure qubit state $|\psi\rangle$ and performs adaptive single-copy measurements using rank-1, two-outcome POVMs. At each round $k \in [N]$, the algorithm selects a direction ψ_k and performs a projective measurement in the basis $\{\psi_k, \psi_k^\perp\}$. The binary outcome after the measurement or “reward” $r_k \in \{0, 1\}$ is distributed according to Born’s rule

$$\Pr(R_k = r_k) = \begin{cases} |\langle \psi_k | \psi \rangle|^2, & r_k = 1, \\ 1 - |\langle \psi_k | \psi \rangle|^2, & r_k = 0. \end{cases} \quad (10)$$

These outcomes match the energy distribution of the thermal protocol Eq. (8), and can be used within that framework by setting $r_k = i$ if $\Delta W_k = w_{k,1-i}$ for $i \in \{0, 1\}$.

The directions of the measurements are chosen to ensure low infidelity with the unknown state, which can be

converted into a bound on the relative entropy. This can be done by selecting ϵ_k such that $1 - |\langle \psi | \psi_k \rangle|^2 \leq \epsilon_k \leq \frac{1}{2}$ with high probability and then bound $D(\psi \| \Delta_{2\epsilon_k}(\psi_k))$. This can be done applying [7, Proposition 2.34] which leads to

$$W_{\text{diss}}(N) \leq \beta^{-1} \sum_{k=1}^N 16\epsilon_k(2 - \ln \epsilon_k). \quad (11)$$

Now we review the exact updating rule of ψ_k at each time step $k \in \{1, 2, \dots, N\}$. Each ψ_k is adaptively based on the previous rewards $\{r_s\}_{s=1}^{k-1}$ and measured directions $\{\psi_s\}_{s=1}^{k-1}$. The updating rule is formulated in the Bloch vector representation. The algorithm works in $K^* \in \mathbb{N}$ stages where for each stage $k^* \in [K^*]$ performs t measurements on the Bloch vector directions $\{a_{k^*,i}\}_{i=1}^4 \subset \mathbb{R}^3$ which we are going to define. The total number of rounds is then $N = 4tK^*$. The sampled reward are then used to build the following t weighted least-square estimators

$$\tilde{\theta}_{k^*,j}^w = V_{k^*}^{-1} \sum_{l=1}^{k^*} \frac{1}{\hat{\sigma}_l^2} \sum_{i=1}^4 a_{l,i} (2r_{l,i,j} - 1), \quad (12)$$

where $r_{l,i,j} \in \{0, 1\}$ is the reward for the measurement direction along $a_{l,i}$ for $l \in [K^*]$ and $j \in [t]$. Then $\hat{\sigma}_l^2$ is an upper bound on the variance of $r_{l,i,j}$ and V_{k^*} is the design matrix

$$V_{k^*} = \mathbb{1} + \sum_{l=1}^{k^*} \frac{1}{\hat{\sigma}_l^2} \sum_{i=1}^4 a_{l,i} (a_{l,i})^\top. \quad (13)$$

As studied in [18, 19], a weighted least-squares estimator boosts learning along directions with low variance, which are precisely those close to ψ , since $\text{Var}(r_k) \leq 1 - |\langle \psi_k | \psi \rangle|^2$. For the specific update used by the algorithm, it can be shown that $\hat{\sigma}_l^2 \sim 1/\sqrt{\lambda_{\max}(V_{k^*-1})}$, (where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue) provides an upper bound on the outcome variances and ensures well-defined least-squares estimators. However, a single estimator is not sufficient to guarantee proximity to the unknown state. For this reason, the algorithm constructs multiple estimators and combines them into a normalized weighted median-of-means $\theta_{k^*}^{\text{wMoM}}$, which satisfies the required probabilistic guarantees to be close to ψ . Details on the construction of $\theta_{k^*}^{\text{wMoM}}$ are provided in Appendix A. Finally it outputs the measurement directions on the Bloch sphere representation $\{a_{k^*+1,i}\}_{i=1,2,3,4}$ of

$$\tilde{a}_{k^*+1,i} = \theta_{k^*}^{\text{wMoM}} - \frac{(-1)^i}{\sqrt{\lambda_{\min}(V_{k^*})}} v_{k^*, \lceil \frac{i}{2} \rceil}, \quad (14)$$

where $\lambda_{\min}(V_{k^*})$ is the minimum eigenvalue of V_{k^*} and $\{v_{k^*,0}, v_{k^*,1}\}$ are the two eigenvectors of V_{k^*} with smallest eigenvalues. The normalized Bloch vectors $a_{k^*+1,i}$ are used to update the measurement directions ψ_k for round k in stage $k^* + 1$. More details can be found in Appendix A.

It has been proven [18, Theorems 9 and 11] that, with high probability, the algorithm achieves an infidelity scaling of $1 - |\langle \psi_k | \psi \rangle|^2 \leq C \ln(N)/k$ for all $k \in \{1, 2, \dots, N\}$. The same bound also holds in expectation. We can thus use this guarantee to set $\epsilon_k = C \ln(N)/k$ in the thermal protocol, using the measurement directions ψ_k given by Eq. (14), and since ϵ_k upper bounds the infidelity between ψ and ψ_k we can use the dissipation bound Eq. (11) to obtain the following result.

Main Result—There exists an explicit protocol for the semi-classical battery model that adaptively updates the direction ψ_k and the accuracy ϵ_k based on the rewards $\{r_s\}_{s=1}^{k-1}$, achieving, with probability at least $1 - \delta$

$$W_{\text{diss}}(N) = O\left(\beta^{-1} \ln^2(N) \ln\left(\frac{N}{\delta}\right)\right). \quad (15)$$

Notice that in this model, measurements in the energy eigenbasis of the battery were performed. According to Landauer’s principle, such measurements reduce entropy and thus incur an energetic cost associated with information erasure. This form of dissipation — arising from cost of resetting memory—should also be taken into account. In Appendix D we show that the dissipation required by Landauer’s principle also scales as $O(\text{polylog}(N))$ when our adaptive algorithm is employed.

Furthermore, this algorithm is not specific to the work extraction protocol discussed above. We have also applied it to an alternative protocol modeled using the Jaynes-Cummings Hamiltonians, where the battery is represented as a uniform energy ladder. In this setting, the work dissipation similarly exhibits $O(\text{polylog}(N))$ scaling. Interested readers are referred to Appendix C

for further details.

Conclusion and outlook—We are the first to frame sequential work extraction as an instance of the exploration-exploitation trade-off and to quantify cumulative dissipation in the finite-copy regime. Beyond establishing a provable upper bound on the total dissipation, we explicitly construct an adaptive algorithm that achieves it. Our approach not only advances the understanding of work extraction under partial information but also opens a pathway for extending this framework to the extraction of other quantum resources beyond free energy—such as entanglement, coherence, and magic—within their respective resource theories.

Acknowledgments JL and RCH contributed equally to this work. JL thanks Jan Seyfried for discussions about quantum state tomography with quantum relative entropy loss. This work is supported by the National Research Foundation of Singapore through the NRF Investigatorship Program (Award No. NRF-NRFI09-0010), the National Quantum Office, hosted in A*STAR, under its Centre for Quantum Technologies Funding Initiative (S24Q2d0009), grant FQXi-RFP-IPW-1903 (“Are quantum agents more energetically efficient at making predictions?”) from the Foundational Questions Institute (FQXI) and Fetzer Franklin Fund (a donor-advised fund of Silicon Valley Community Foundation), the Singapore Ministry of Education Tier 1 Grant RT4/23, MT, YH and JL are supported by the National Research Foundation, Singapore and A*STAR under its CQT Bridging Grant and its Quantum Engineering Programme (NRF2021-QEP2-02-P05)

-
- [1] J. Åberg. “Catalytic Coherence”. Phys. Rev. Lett. **113**:150402 (2014).
 - [2] J. Åberg. “Truly work-like work extraction via a single-shot analysis”. Nature communications **4**(1): 1925 (2013).
 - [3] A. E. Allahverdyan, R. Balian, and T. M. Nieuwenhuizen. “Maximal work extraction from finite quantum systems”. Europhysics Letters **67**(4): 565 (2004).
 - [4] F. G. Brandao, M. Horodecki, J. Oppenheim, J. M. Renes, and R. W. Spekkens. “Resource theory of quantum states out of thermal equilibrium”. Physical review letters **111**(25): 250404 (2013).
 - [5] S. Chen, B. Huang, J. Li, A. Liu, and M. Sellke. “When Does Adaptivity Help for Quantum State Learning?”. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 391–404, (2023).
 - [6] C. Elouard and A. N. Jordan. “Efficient quantum measurement engines”. Physical review letters **120**(26): 260601 (2018).
 - [7] S. T. Flammia and R. O’Donnell. “Quantum chi-squared tomography and mutual information testing”. Quantum **8**: 1381 (2024).
 - [8] J. Goold, M. Paternostro, and K. Modi. “Nonequilibrium Quantum Landauer Principle”. Phys. Rev. Lett. **114**: 060602 (2015).
 - [9] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu. “Sample-Optimal Tomography of Quantum States”. IEEE Transactions on Information Theory **63**(9): 5628–5641 (2017).
 - [10] R. C. Huang, P. M. Riechers, M. Gu, and V. Narasimhachar. “Engines for predictive work extraction from memoryful quantum stochastic processes”. Quantum **7**: 1203 (2023).
 - [11] D. Šafránek, D. Rosa, and F. C. Binder. “Work Extraction from Unknown Quantum Sources”. Phys. Rev. Lett. **130**: 210401 (2023).
 - [12] 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).
 - [13] Y. Jun, M. c. v. Gavrilo, and J. Bechhoefer. “High-Precision Test of Landauer’s Principle in a Feedback Trap”. Phys. Rev. Lett. **113**: 190601 (2014).
 - [14] K. Korzekwa, M. Lostaglio, J. Oppenheim, and D. Jennings. “The extraction of work from quantum coherence”. New Journal of Physics **18**(2): 023045 (2016).
 - [15] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press (2020).
 - [16] M. Lostaglio, D. Jennings, and T. Rudolph. “Descrip-

- tion of quantum coherence in thermodynamic processes requires constraints beyond free energy”. *Nature communications* **6**(1):6383 (2015).
- [17] J. Lumbrellas, E. Haapasalo, and M. Tomamichel. “Multi-armed quantum bandits: Exploration versus exploitation when learning properties of quantum states”. *Quantum* **6**:749 (2022).
- [18] J. Lumbrellas, M. Terekhov, and M. Tomamichel. “Learning pure quantum states (almost) without regret”. arXiv preprint arXiv:2406.18370, (2024).
- [19] J. Lumbrellas 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).
- [20] 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).
- [21] R. O’Donnell and J. Wright. “Efficient quantum tomography”. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC ’16, pages 899–912, New York, NY, USA (2016).
- [22] P. M. Riechers and M. Gu. “Initial-state dependence of thermodynamic dissipation for any quantum process”. *Phys. Rev. E* **103**:042145 (2021).
- [23] T. Sagawa and M. Ueda. “Minimal Energy Cost for Thermodynamic Information Processing: Measurement and Information Erasure”. *Phys. Rev. Lett.* **102**:250602 (2009).
- [24] M. O. Scully and M. S. Zubairy. *Atom-field interaction - quantum theory*, page 193–219. Cambridge University Press (1997).
- [25] P. Skrzypczyk, A. J. Short, and S. Popescu. “Work extraction and thermodynamics for individual quantum systems”. *Nature communications* **5**(1):4185 (2014).
- [26] G. Strang. *Calculus*. Wellesley-Cambridge Press (2019).
- [27] L. Szilard. “Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen”. *Zeitschrift für Physik* **53**(11):840–856 (1929).
- [28] V. V. Tan and K. Saito. “Finite-Time Quantum Landauer Principle and Quantum Coherence”. *Phys. Rev. Lett.* **128**:010602 (2022).
- [29] P. Taranto, F. Bakhshinezhad, A. Bluhm, R. Silva, N. Friis, M. P. Lock, G. Vitagliano, F. C. Binder, T. Debarba, E. Schwarzhans, et al. “Landauer versus Nernst: What is the true cost of cooling a quantum system?”. *PRX Quantum* **4**(1):010332, (2023).
- [30] T. Van Vu and K. Saito. “Finite-Time Quantum Landauer Principle and Quantum Coherence”. *Phys. Rev. Lett.* **128**:010602 (2022).
- [31] K. Watanabe and R. Takagi. “Black box work extraction and composite hypothesis testing”. *Physical Review Letters* **133**(25):250401 (2024).
- [32] K. Watanabe and R. Takagi. “Universal work extraction in quantum thermodynamics”. arXiv preprint arXiv:2504.12373, (2025).
- [33] M. P. Woods and M. Horodecki. “Autonomous Quantum Devices: When Are They Realizable without Additional Thermodynamic Costs?”. *Phys. Rev. X* **13**:011016 (2023).

Appendix A: Learning pure quantum states (almost) without regret

In this Section we review the quantum state tomography algorithm presented in [18]. 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

$$\Pr(R_k = r_k) = \begin{cases} |\langle\psi_k|\psi\rangle|^2, & r_k = 1, \\ 1 - |\langle\psi_k|\psi\rangle|^2, & r_k = 0. \end{cases} \quad (\text{A1})$$

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

$$\text{Rgrt}(N) = \sum_{k=1}^N (1 - |\langle\psi_k|\psi\rangle|^2), \quad (\text{A2})$$

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

$$\psi_k = \frac{1}{2}(\mathbb{I} + a_k \cdot \sigma), \quad (\text{A3})$$

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 [18] 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 $a_{1,1} = (\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})^\top$, $a_{1,2} = (-\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})^\top$, $a_{1,3} = (0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top$, $a_{1,4} = (0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top$.

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 $j = 1, \dots, t$ **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

$V_{k^*} \leftarrow V_{k^*-1} + \frac{1}{\hat{\sigma}_{k^*}^2} \sum_{i=1}^4 a_{k^*,i} a_{k^*,i}^\top$

Update LSE for each subsample

for $j = 1, 2, \dots, t$ **do**

$\tilde{\theta}_{k^*,j}^w \leftarrow V_{k^*}^{-1} \sum_{s=1}^{k^*} \frac{1}{\hat{\sigma}_{k^*}^2} \sum_{i=1}^4 a_{s,i} (2r_{s,i,j} - 1)$

Update Bloch vectors for estimates

 Compute $\theta_{k^*}^{\text{wMoM}}$ using $\{\tilde{\theta}_{k^*,j}^w\}_{j=1,\dots,t}$ according to Eq. (A6)

Update Bloch vectors for measurements

 Select Bloch vectors $a_{k^*+1,i}$ using $\theta_{k^*}^{\text{wMoM}}$ according to Eq. (A9)

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

$$\tilde{\theta}_{k^*,j}^w = V_{k^*}^{-1} \sum_{l=1}^{k^*} \frac{1}{\hat{\sigma}_l^2} \sum_{i=1}^4 a_{l,i} (2r_{l,i,j} - 1), \quad (\text{A4})$$

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

$$V_{k^*} = V_{k^*-1} + \frac{1}{\hat{\sigma}_{k^*}^2} \sum_{i=1}^4 a_{k^*,i} a_{k^*,i}^\top, \quad (\text{A5})$$

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 $\theta_{k^*}^{\text{wMoM}}$ of median of means with the k weighted least squares estimator $\tilde{\theta}_{k^*}^{\text{wMoM}}$ in the following way

$$\theta_{k^*}^{\text{wMoM}} = \frac{\tilde{\theta}_{k^*}^{\text{wMoM}}}{\|\tilde{\theta}_{k^*}^{\text{wMoM}}\|_2} \quad (\text{A6})$$

where

$$\tilde{\theta}_{k^*}^{\text{wMoM}} := \tilde{\theta}_{k^*,j^*}, \quad (\text{A7})$$

where

$$j^* = \underset{j \in [t]}{\operatorname{argmin}} \operatorname{median}\{\|\tilde{\theta}_{t^*,j} - \tilde{\theta}_{t^*,j'}\|_{V_{k^*}} : j' \in [k]/j\}, \quad (\text{A8})$$

and $\|x\|_{V_{k^*}}^2 = \langle x, V_{k^*} x \rangle$ 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

$$a_{k^*+1,i} = \frac{\tilde{a}_{k^*+1,i}}{\|\tilde{a}_{k^*+1,i}\|_2}, \quad (\text{A9})$$

where

$$\tilde{a}_{k^*+1,i} = \theta_{k^*}^{\text{wMoM}} - \frac{(-1)^i}{\sqrt{\lambda_{\min}(V_{k^*})}} v_{k^*, \lceil \frac{i}{2} \rceil} \quad (\text{A10})$$

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 $\hat{\sigma}_{k^*}^2$ is set to

$$\hat{\sigma}_{k^*}^2 = \frac{2\zeta}{\sqrt{\lambda_{\max}(V_{k^*})}}, \quad (\text{A11})$$

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 $\hat{\sigma}_{k^*}^2$ given by Eq. (A11) guarantees that $\hat{\sigma}_{k^*}^2$ 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 [18] it allows to obtain rigorous concentration bounds such that $\tilde{\theta}_{k^*}^{\text{wMoM}}$ is a good estimator for the Bloch vector of the unknown state $|\psi\rangle$.

Now we state the main theorem from [18] 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 1 ([18, 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. (A2) scaling*

$$\text{Rgrt}(N) \leq C_1 \ln\left(\frac{N}{\delta}\right) \ln(N), \quad (\text{A12})$$

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

$$1 - |\langle\psi_k|\psi\rangle|^2 \leq C_2 \frac{\ln\left(\frac{N}{\delta}\right)}{k}, \quad (\text{A13})$$

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

$$\mathbb{E}[\text{Rgrt}(N)] = C_3 \ln^2(N), \quad \mathbb{E}[1 - |\langle\psi_k|\psi\rangle|^2] \leq C_4 \frac{\ln(N)}{k}, \quad (\text{A14})$$

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.

Appendix B: 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 [25]. 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.

$$\beta \mathbb{E}(W) = D(\rho \| \gamma_\beta). \quad (\text{B1})$$

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 $H_A = \omega \mathbb{1}/2$. 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 $Z_R(\nu) = \text{tr}(e^{-\beta H_R(\nu)})$ 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 $H_B \varphi(x - x_0) = x \varphi(x - x_0)$. We will assume $\varphi(x - x_0)$ to be a battery state whose energy is sharply centered at x_0 . The energy of the battery can be changed by translating the weight up by a certain height x_0 , described by the translation operator $\Gamma_{x_0}^B \varphi(x) = \varphi(x - x_0)$. We aim to extract work from the partially unknown 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 $\tau = 1, 2, \dots, M$ **do**

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

Take a fresh thermal qubit $\gamma_\beta(\nu(\tau, \epsilon)) = \frac{1}{Z_R(\nu(\tau, \epsilon))} e^{-\beta H_R(\nu(\tau, \epsilon))}$ where $\nu(\tau, \epsilon) = \beta^{-1} \ln \frac{p_{0,\tau}}{p_{1,\tau}}$,

$p_{i,\tau} = p_i - (-1)^i \tau \delta p$ and $\delta p = \frac{1}{M}(p_0 - \frac{1}{2})$ from the reservoir.

Apply the swap unitary $V_{\rho^*,\tau} = \sum_{ij} |i\rangle\langle j|_A \otimes |j\rangle\langle i|_R \otimes \Gamma_{(i-j)\nu(\tau, \epsilon)}$ 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

$$U = \sum_i |i\rangle\langle\phi_i|. \quad (\text{B2})$$

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

$$\rho_{AB} = \sum_{ij} \langle\phi_i|\rho|\phi_j\rangle |i\rangle\langle j|_A \otimes \varphi(x)_B. \quad (\text{B3})$$

In the second stage of the protocol, we perform M repetitions of the following process. In repetition τ , we take a fresh thermal qubit $\gamma_\beta(\nu(\tau, \epsilon))$ with Hamiltonian $H_R(\nu(\tau, \epsilon))$ from the reservoir where $\nu(\tau, \epsilon) = \beta^{-1} \ln \frac{p_{0,\tau}}{p_{1,\tau}}$, $p_{i,\tau} = p_i - (-1)^i \tau \delta p$ 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

$$V_{\rho^*,\tau} = \sum_{ij} |i\rangle\langle j|_A \otimes |j\rangle\langle i|_R \otimes \Gamma_{(i-j)\nu(\tau, \epsilon)}. \quad (\text{B4})$$

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 τ , the reduced state is

$$\rho_{AB,\tau} = \text{tr}_R \left(V_{\rho^*,\tau} (\rho_{AB,\tau-1} \otimes \gamma_\beta(\nu(\tau, \epsilon))) V_{\rho^*,\tau}^\dagger \right), \quad (\text{B5})$$

After the first repetition, we obtain

$$\rho_{AB,1} = \sum_i p_{i,1} |i\rangle\langle i|_A \otimes \rho_{B,i,1}, \quad (\text{B6})$$

where

$$\rho_{B,i,1} = \sum_j \langle \phi_j | \rho | \phi_j \rangle \varphi(x - (i-j)\nu(\tau, \epsilon)), \quad (\text{B7})$$

and after repetition τ where $\tau \geq 2$, we obtain

$$\rho_{AB,\tau} = \sum_i p_{i,\tau} |i\rangle\langle i|_A \otimes \rho_{B,i,\tau}, \quad (\text{B8})$$

where

$$\rho_{B,i,\tau} = \sum_j p_{j,\tau-1} \Gamma_{(i-j)\nu(\tau,\epsilon)} \rho_{B,j,\tau-1} \Gamma_{(i-j)\nu(\tau,\epsilon)}^\dagger. \quad (\text{B9})$$

From Eq. (B8), 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.

1. Work distribution

In this section, we will use the Lagrange mean value theorem and the first mean value theorem for definite integrals [26] as follows:

Theorem 2. *Let $f : [a, b] \rightarrow \mathbb{R}$ 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*

$$f(b) - f(a) = f'(c)(b - a). \quad (\text{B10})$$

Theorem 3. *Let $f : [a, b] \rightarrow \mathbb{R}$ be a continuous function on the closed interval $[a, b]$. Then there exists $c \in (a, b)$ such that*

$$\int_a^b f(x) dx = f(c)(b - a). \quad (\text{B11})$$

We will show the following theorem in the following subsection.

Theorem 4. *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*

$$\lim_{M \rightarrow \infty} \mathbb{E}[\Delta W] = w_i, \quad (\text{B12})$$

where

$$w_i := \beta^{-1}(\mathcal{D}(\phi_i \| \mathbb{1}/2) + \ln p_i), \quad (\text{B13})$$

and for any $\epsilon > 0$

$$\lim_{M \rightarrow \infty} \Pr[|\Delta W - \mathbb{E}[\Delta W]| \geq \epsilon] = 0. \quad (\text{B14})$$

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. (B6), 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. (B8), the evolution in repetition τ where $\tau \geq 2$ can be viewed as a classical process described as follows: the state after repetition τ is ϕ_{x_τ} where x_τ is a random bit sampled from $\{0, 1\}$ according to the probability distribution $(p_{0,\tau}, p_{1,\tau})$; the extracted work in repetition τ conditioned on $x_{\tau-1}x_\tau$ and is $(x_\tau - x_{\tau-1})\nu(\tau, \epsilon)$. 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

$$\Delta W = (x_1 - i)\nu(1, \epsilon) + \sum_{\tau=2}^M (x_\tau - x_{\tau-1})\nu(\tau, \epsilon) = -i\nu(1, \epsilon) + \sum_{\tau=1}^{M-1} x_\tau(\nu(\tau, \epsilon) - \nu(\tau+1, \epsilon)) + x_M\nu(M, \epsilon), \quad (\text{B15})$$

recall that

$$\nu(\tau, \epsilon) = \beta^{-1} \ln \frac{p_0 - \tau\delta p}{p_1 + \tau\delta p}. \quad (\text{B16})$$

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

$$\mathbb{E}[\Delta W] = -i\nu(1, \epsilon) + \sum_{\tau=1}^{M-1} \mathbb{E}[x_\tau](\nu(\tau, \epsilon) - \nu(\tau+1, \epsilon)) + \mathbb{E}[x_M]\nu(M, \epsilon). \quad (\text{B17})$$

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

$$\mathbb{E}[\Delta W] = -i\beta^{-1} \ln \frac{p_0 - \delta p}{p_1 + \delta p} + \beta^{-1} \sum_{\tau=1}^{M-1} \left(\ln \frac{p_0 - \tau\delta p}{p_1 + \tau\delta p} - \ln \frac{p_0 - (\tau+1)\delta p}{p_1 + (\tau+1)\delta p} \right) (p_1 + \tau\delta p). \quad (\text{B18})$$

We now use the definition of $\nu(\tau, \epsilon)$ in Eq. (B16) as well as the Lagrange mean value theorem in Theorem 2 to obtain

$$\beta(\nu(\tau, \epsilon) - \nu(\tau+1, \epsilon)) = \ln \frac{p_0 - \tau\delta p}{p_1 + \tau\delta p} - \ln \frac{p_0 - (\tau+1)\delta p}{p_1 + (\tau+1)\delta p} = \frac{1}{\xi_\tau(1 - \xi_\tau)} \delta p, \quad (\text{B19})$$

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

Therefore, Eq. (B18) can be simplified to

$$\begin{aligned} \mathbb{E}[\Delta W] &= -i\beta^{-1} \left(\ln \frac{p_0}{p_1} - \frac{\delta p}{\xi_0(1 - \xi_0)} \right) + \beta^{-1} \sum_{\tau=1}^{M-1} \frac{p_1 + \tau\delta p}{\xi_\tau(1 - \xi_\tau)} \delta p \\ &= -i\beta^{-1} \ln \frac{p_0}{p_1} + \beta^{-1} \sum_{\tau=1}^M \frac{p_1 + \tau\delta p}{\xi_\tau(1 - \xi_\tau)} \delta p \\ &\quad + i\beta^{-1} \frac{\delta p}{\xi_0(1 - \xi_0)} - \beta^{-1} \frac{\delta p}{2\xi_M(1 - \xi_M)}. \end{aligned} \quad (\text{B20})$$

We will approximate the sum in the second line of Eq. (B20) with an integration, where the remainder is bounded due to first mean value theorem for definite integrals as in Theorem 3. Namely,

$$\beta^{-1} \sum_{\tau=1}^M \frac{p_1 + \tau\delta p}{\xi_\tau(1 - \xi_\tau)} \delta p = \beta^{-1} \int_{\frac{1}{2}}^{p_0} \frac{dp}{p} + \beta^{-1} R_1(\delta p) = \beta^{-1} \ln p_0 + \beta^{-1} \ln(2) + \beta^{-1} R_1(\delta p), \quad (\text{B21})$$

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

$$R_1(\delta p) = \sum_{\tau=1}^M \frac{p_1 + \tau\delta p}{\xi_\tau(1 - \xi_\tau)} \delta p - \int_{\frac{1}{2}}^{p_0} \frac{dp}{p} = \sum_{\tau=1}^M \left(\frac{p_1 + \tau\delta p}{\xi_\tau(1 - \xi_\tau)} \delta p - \int_{p_0 - \tau\delta p}^{p_0 - (\tau-1)\delta p} \frac{dp}{p} \right) \quad (\text{B22})$$

$$= \sum_{\tau=1}^M \left(\frac{p_1 + \tau\delta p}{\xi_\tau(1 - \xi_\tau)} - \frac{1}{\xi'_\tau} \right) \delta p = \sum_{\tau=1}^M \frac{\xi_\tau(1 - \xi_\tau) - \xi'_\tau(p_1 + \tau\delta p)}{\xi'_\tau \xi_\tau(1 - \xi_\tau)} \delta p, \quad (\text{B23})$$

where from the first line to the second line, we have used the first mean value theorem for definite integrals Theorem 3 that

$$\int_{p_0 - \tau\delta p}^{p_0 - (\tau-1)\delta p} \frac{dp}{p} = \frac{1}{\xi'_\tau} \delta p, \quad (\text{B24})$$

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

$$|R_1(\delta p)| \leq \sum_{\tau=1}^M \left| \frac{\xi_\tau(1 - \xi_\tau) - \xi'_\tau(p_1 + \tau\delta p)}{\xi'_\tau \xi_\tau(1 - \xi_\tau)} \right| \delta p \leq \sum_{\tau=1}^M \left| \frac{p_1 + \tau\delta p + \xi_\tau}{\xi'_\tau \xi_\tau(1 - \xi_\tau)} \right| (\delta p)^2 \leq \sum_{\tau=1}^M \frac{4}{p_1} (\delta p)^2 \leq (2p_0 - 1) \frac{2}{p_1} \delta p. \quad (\text{B25})$$

Therefore, the second line in Eq. (B20) is finite while the third line is infinitesimal, and we obtain

$$\mathbb{E}[\Delta W] = -i\beta^{-1} \ln \frac{p_0}{p_1} + \beta^{-1} \ln p_0 + \beta^{-1} \ln(2) + O(\delta p) \quad (\text{B26})$$

$$= -\beta^{-1} \text{tr}(\phi_i \ln \mathbb{1}/2) - i\beta^{-1} \ln \frac{p_0}{p_1} + \beta^{-1} \ln p_0 + O(\delta p) \quad (\text{B27})$$

$$= -\beta^{-1} \text{tr}(\phi_i \ln \mathbb{1}/2) + \beta^{-1} \ln p_i + O(\delta p) \quad (\text{B28})$$

$$= \beta^{-1} [\text{D}(\phi_i \| \mathbb{1}/2) + \ln p_i] + O(\delta p). \quad (\text{B29})$$

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

$$\mathbb{E}[\Delta W] = \beta^{-1} [\text{D}(\phi_i \| \mathbb{1}/2) + \ln p_i] + O\left(\frac{1}{M}\right). \quad (\text{B30})$$

Taking $M \rightarrow \infty$, we obtain

$$\lim_{M \rightarrow \infty} \mathbb{E}[\Delta W] = \beta^{-1} [\text{D}(\phi_i \| \mathbb{1}/2) + \ln p_i]. \quad (\text{B31})$$

Now we demonstrate the convergence of ΔW towards its expectation value, recall from Eq. (B15), we have that

$$\Delta W = -i\nu(1, \epsilon) + \sum_{\tau=1}^{M-1} x_\tau(\nu(\tau, \epsilon) - \nu(\tau + 1, \epsilon)) + x_M \nu(M, \epsilon). \quad (\text{B32})$$

By the Lagrange mean value theorem,

$$\nu(\tau, \epsilon) - \nu(\tau + 1, \epsilon) = \ln \frac{p_0 - \tau\delta p}{p_1 + \tau\delta p} - \ln \frac{p_0 - (\tau + 1)\delta p}{p_1 + (\tau + 1)\delta p} = \frac{\delta p}{\xi_\tau(1 - \xi_\tau)}, \quad (\text{B33})$$

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

$$|\nu(\tau, \epsilon) - \nu(\tau + 1, \epsilon)| \leq \frac{2}{p_1} \delta p. \quad (\text{B34})$$

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

$$\Pr[|\Delta W - \mathbb{E}[\Delta W]| \geq \epsilon] \leq 2e^{-\frac{\epsilon^2}{\sum_{\tau=1}^{M-1} \left(\frac{2}{p_1} \delta p\right)^2}} \leq 2e^{-\frac{p_1^2 \epsilon^2 M}{(2p_0 - 1)^2}}. \quad (\text{B35})$$

Taking $M \rightarrow \infty$, we obtain

$$\lim_{M \rightarrow \infty} \Pr[|\Delta W - \mathbb{E}[\Delta W]| \geq \epsilon] = 0. \quad (\text{B36})$$

□

Theorem 4 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 $w_0 \geq w_1$)

$$r = \begin{cases} 1, & \Delta W \geq \frac{w_0 + w_1}{2}, \\ 0, & \Delta W \leq \frac{w_0 + w_1}{2}. \end{cases} \quad (\text{B37})$$

The distribution of the reward is

$$\Pr[R = r] = \begin{cases} |\langle \phi_0 | \psi \rangle|^2 + \epsilon_{\text{error}}, & r = 1, \\ 1 - |\langle \phi_0 | \psi \rangle|^2 - \epsilon_{\text{error}}, & r = 0, \end{cases} \quad (\text{B38})$$

where $|\epsilon_{\text{error}}| \leq O(e^{-CM})$ for some constant C . In the limit of $M \rightarrow \infty$, the correspondence reduces to

$$r = \begin{cases} 1, & \Delta W = w_0, \\ 0, & \Delta W = w_1, \end{cases} \quad (\text{B39})$$

and the distribution of the reward reduces to

$$\Pr[R = r] = \begin{cases} |\langle \phi_0 | \psi \rangle|^2, & r = 1, \\ 1 - |\langle \phi_0 | \psi \rangle|^2, & r = 0, \end{cases} \quad (\text{B40})$$

Note that as Algorithm 2 proceeds with increasing k , the error probability that the algorithm can tolerate in the reward measurement decreases polynomially, which requires the number of repetitions M increases polynomially.

2. Extracted work for different inputs

Theorem 5. 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. 4. It holds that,

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

$$\Pr(\Delta W = w_i) = \langle \phi_i | \rho | \phi_i \rangle. \quad (\text{B41})$$

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

$$\mathbb{E}[\Delta W] = \beta^{-1} [\text{D}(\rho \| \mathbb{1}/2) - \text{D}(\rho \| \rho^*)], \quad (\text{B42})$$

where the second term can be defined as 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 joint state in Eq. (B3) does not affect Eq. (B6). 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 5 holds, i.e.,

$$\Pr(\Delta W = w_i) = \langle \phi_i | \rho | \phi_i \rangle. \quad (\text{B43})$$

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. (B13) and (B41), the extracted work is given by

$$\mathbb{E}[\Delta W] = \langle \phi_0 | \rho | \phi_0 \rangle \beta^{-1} [\text{D}(\phi_0 \| \mathbb{1}/2) + \ln p_0] + \langle \phi_1 | \rho | \phi_1 \rangle \beta^{-1} [\text{D}(\phi_1 \| \mathbb{1}/2) + \ln p_1] \quad (\text{B44})$$

$$= \beta^{-1} [\langle \phi_0 | \rho | \phi_0 \rangle (-\text{tr}(\phi_0 \ln \mathbb{1}/2) + \ln p_0) - \langle \phi_1 | \rho | \phi_1 \rangle (-\text{tr}(\phi_1 \ln \mathbb{1}/2) + \ln p_1)] \quad (\text{B45})$$

$$= \beta^{-1} [\text{tr}(\mathcal{P}(\rho) \ln \rho^*) - \text{tr}(\mathcal{P}(\rho) \ln \mathbb{1}/2)] = \beta^{-1} [\text{tr}(\rho \ln \rho^*) - \text{tr}(\rho \ln \mathbb{1}/2)], \quad (\text{B46})$$

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:

$$\mathbb{E}[\Delta W] = \beta^{-1} [\text{tr}(\rho \ln \rho^*) - \text{tr}(\rho \ln \mathbb{1}/2)] = \beta^{-1} [\text{D}(\rho \| \mathbb{1}/2) - \text{D}(\rho \| \rho^*)]. \quad (\text{B47})$$

□

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

$$\mathbb{E}[\Delta W] = p_0 \beta^{-1} [\text{D}(\phi_0 \| \mathbb{1}/2) + \ln p_0] + p_1 \beta^{-1} [\text{D}(\phi_1 \| \mathbb{1}/2) + \ln p_1] \quad (\text{B48})$$

$$= \beta^{-1} (-p_0 \text{tr}(\phi_0 \ln \mathbb{1}/2) + p_0 \ln p_0 - p_1 \text{tr}(\phi_1 \ln \mathbb{1}/2) + p_1 \ln p_1) \quad (\text{B49})$$

$$= \beta^{-1} [\text{tr}(\rho \ln \rho) - \text{tr}(\rho \ln \mathbb{1}/2)] = \beta^{-1} \text{D}(\rho \| \mathbb{1}/2). \quad (\text{B50})$$

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

$$W_{\text{diss}} = \max_{\rho^*} \mathbb{E}[\Delta W] - \mathbb{E}[\Delta W] = \beta^{-1} \text{D}(\rho \| \rho^*). \quad (\text{B51})$$

3. 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 (B39) 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 5, we define the dissipation at round $k \in [N]$ with respect to the optimal protocol as

$$W_{\text{diss}}^k := \beta^{-1} D(\psi \| \hat{\rho}_k), \quad (\text{B52})$$

and the cumulative dissipation over all N rounds as

$$W_{\text{diss}}(N) := \beta^{-1} \sum_{k=1}^N W_{\text{diss}}^k = \beta^{-1} \sum_{k=1}^N D(\psi \| \hat{\rho}_k). \quad (\text{B53})$$

Algorithm 3: Thermal work extraction

Require: sequence $\{\epsilon_k\}_{k=1}^\infty$

for $k = 1, 2, \dots$ **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 (B39)

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 6 (Theorem 2.34 in [7]). *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*

$$D(\rho \| \Delta_{2\epsilon}(\hat{\rho})) \leq 16\epsilon \left(2 + \ln \left(\frac{d}{2\epsilon} \right) \right). \quad (\text{B54})$$

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

Theorem 7. *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*

$$W_{\text{diss}}(N) = O \left(\beta^{-1} \ln^2(N) \ln \left(\frac{N}{\delta} \right) \right). \quad (\text{B55})$$

Proof. We can choose

$$\epsilon_k = \min \left\{ C \frac{\ln \left(\frac{N}{\delta} \right)}{k}, \frac{1}{2} \right\}, \quad (\text{B56})$$

where C is the constant in Theorem 1 for the fidelity bound of the direction ψ_k . Then we can use Theorem 6 combined with the fidelity guarantee of Theorem 1 to get that with probability at least $1 - \delta$ we have

$$W_{\text{diss}}(N) \leq \beta^{-1} \sum_{k=1}^N 16\epsilon_k (2 - \ln \epsilon_k). \quad (\text{B57})$$

The result follows by noting that, for a sufficiently large constant k^* , we have $\epsilon_k = C \frac{\ln \left(\frac{N}{\delta} \right)}{k}$ 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. $\square$

4. 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 incurred during the rotation can be shifted into the battery via the unitary,

$$U_{\rho^*} = \sum_i |E_i\rangle\langle\phi_i| \otimes \Gamma_{e_i}^B, \quad (\text{B58})$$

where $e_i = \text{tr}(\phi_i H_A) - \omega_i$ 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 [25], 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 [33]. More recent work has shown that such an operation can only be realized if one has access to unbounded coherence [1, 14], 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.

Appendix C: Jaynes-Cummings work extraction protocol

Algorithm 4: Jaynes-Cummings work extraction

Require: A sequence of unknown states $|\psi\rangle$

for $k = 1, 2, \dots$ **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. (C11)

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 $k = 1, 2, \dots, N$.

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

$$H = \omega(|\psi_k\rangle\langle\psi_k| + a^\dagger a) + \frac{\Omega}{2}(a \otimes |\psi_k\rangle\langle\psi_k^\perp| + a^\dagger \otimes |\psi_k^\perp\rangle\langle\psi_k|). \quad (\text{C1})$$

This is exactly given by the famous Jaynes-Cummings model [12], which is nowadays a textbook model [24]. One

may verify that the eigenstates of the total Hamiltonian are $|0\rangle |\psi_k^\perp\rangle$ as well as

$$|n, +\rangle = \frac{1}{\sqrt{2}}(|n-1\rangle |\psi_k\rangle + |n\rangle |\psi_k^\perp\rangle), \quad |n, -\rangle = \frac{1}{\sqrt{2}}(|n-1\rangle |\psi_k\rangle - |n\rangle |\psi_k^\perp\rangle), \quad (\text{C2})$$

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

$$E_{n+} = n\omega + \frac{\Omega}{2}\sqrt{n}, \quad E_{n-} = n\omega - \frac{\Omega}{2}\sqrt{n}. \quad (\text{C3})$$

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

$$|n_k\rangle |\psi\rangle = \frac{1}{\sqrt{2}}\langle\psi_k^\perp|\psi\rangle(|n_k, +\rangle - |n_k, -\rangle) + \frac{1}{\sqrt{2}}\langle\psi_k|\psi\rangle(|n_k + 1, +\rangle + |n_k + 1, -\rangle). \quad (\text{C4})$$

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,

$$e^{-iHt_k} |n_k\rangle |\psi\rangle = \frac{1}{\sqrt{2}}\langle\psi_k^\perp|\psi\rangle(e^{i\theta_k} |n_k, +\rangle - e^{-i\theta_k} |n_k, -\rangle) + \frac{1}{\sqrt{2}}e^{i\frac{\omega\pi}{\Omega\sqrt{n_k+1}}}\langle\psi_k|\psi\rangle(i |n_k + 1, +\rangle - i |n_k + 1, -\rangle) \quad (\text{C5})$$

$$= \langle\psi_k^\perp|\psi\rangle(i \sin \theta_k |n_k - 1\rangle |\psi_k\rangle + \cos \theta_k |n_k\rangle |\psi_k^\perp\rangle) + \frac{1}{\sqrt{2}}ie^{i\frac{\omega\pi}{\Omega\sqrt{n_k+1}}}\langle\psi_k|\psi\rangle |n_k + 1\rangle |\psi_k^\perp\rangle, \quad (\text{C6})$$

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

$$\Pr[n_{k+1}] = \begin{cases} |\langle\psi_k|\psi\rangle|^2, & n_{k+1} = n_k + 1, \\ |\langle\psi_k^\perp|\psi\rangle|^2 \cos^2 \theta_k, & n_{k+1} = n_k, \\ |\langle\psi_k^\perp|\psi\rangle|^2 \sin^2 \theta_k, & n_{k+1} = n_k - 1. \end{cases} \quad (\text{C7})$$

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

$$\mathbb{E}[\Delta W_k] = \omega(|\langle\psi_k|\psi\rangle|^2 - |\langle\psi_k^\perp|\psi\rangle|^2 \sin^2 \theta_k) = \omega(|\langle\psi_k|\psi\rangle|^2(1 + \sin^2 \theta_k) - \sin^2 \theta_k), \quad (\text{C8})$$

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 $|\langle\psi_k|\psi\rangle| = 1$,

$$\max_{|\psi_k\rangle} \mathbb{E}[\Delta W_k] = \omega. \quad (\text{C9})$$

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

$$W_{\text{diss}}^{\text{jc},k} = \max_{|\psi_k\rangle} \mathbb{E}[\Delta W_k] - \mathbb{E}[\Delta W_k] = \omega(1 + \sin^2 \theta_k^2)(1 - |\langle\psi_k|\psi\rangle|^2) \leq 2\omega(1 - |\langle\psi_k|\psi\rangle|^2). \quad (\text{C10})$$

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

$$r_k = \begin{cases} 1, & n_{k+1} = n_k + 1, \\ 0, & \text{otherwise.} \end{cases} \quad (\text{C11})$$

The probability distribution of the reward is

$$\Pr[R_k = r_k] = \begin{cases} |\langle\psi_k|\psi\rangle|^2, & r_k = 1, \\ |\langle\psi_k^\perp|\psi\rangle|^2, & r_k = 0. \end{cases} \quad (\text{C12})$$

The regret in this round is thus

$$\text{Rgrt}^k = 1 - |\langle\psi_k|\psi\rangle|^2. \quad (\text{C13})$$

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

$$W_{\text{diss}}^{\text{jc},k} \leq 2\omega \text{Rgrt}^k. \quad (\text{C14})$$

The cumulative dissipation over N rounds is thus

$$W_{\text{diss}}^{\text{jc}}(N) = \sum_{k=1}^N W_{\text{diss}}^{\text{jc},k} \leq 2\omega \sum_{k=1}^N (1 - |\langle \psi_k | \psi \rangle|^2). \quad (\text{C15})$$

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

$$W_{\text{diss}}^{\text{jc}}(N) \leq 2\omega \text{Rgrt}(N), \quad (\text{C16})$$

Theorem 8. *There exists an explicit protocol for extracting work from knowledge that adaptively updates the estimate $|\hat{\psi}_k\rangle$ and the probe state $|\psi_k\rangle$ based on the rewards $\{r_s\}_{s=1}^{k-1}$, achieving, with probability at least $1 - \delta$,*

$$W_{\text{diss}}^{\text{jc}}(N) = O\left(\omega \ln(N) \ln\left(\frac{N}{\delta}\right)\right). \quad (\text{C17})$$

Appendix D: 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 [23]. 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 [8, 23, 28], the heat dissipation required to erase the register is lower bounded by the entropy change ΔS of the memory register via

$$\beta Q \geq \Delta S. \quad (\text{D1})$$

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 [13, 20, 22]. 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,

$$W'_{\text{diss}}(N) = W_{\text{diss}}(N) + \beta^{-1} \sum_{k=1}^N \Delta S_k, \quad (\text{D2})$$

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

$$\Delta S_k = -(1 - \epsilon_k) \ln(1 - \epsilon_k) - \epsilon_k \ln \epsilon_k \leq \epsilon_k - \epsilon_k \ln \epsilon_k, \quad (\text{D3})$$

and the dissipation is upper bounded by

$$W_{\text{diss}}^{\text{sc},*}(N) = W_{\text{diss}}^{\text{sc}}(N) + \beta^{-1} \sum_{k=1}^N (\epsilon_k - \epsilon_k \ln \epsilon_k). \quad (\text{D4})$$

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

$$\Delta S_k = -(1 - \epsilon_k) \ln(1 - \epsilon_k) - \epsilon_k \ln \epsilon_k - \epsilon_k (\cos^2 \theta_k \ln \cos^2 \theta_k + \sin^2 \theta_k \ln \sin^2 \theta_k) \leq 2\epsilon_k - \epsilon_k \ln \epsilon_k, \quad (\text{D5})$$

and the dissipation is upper bounded by

$$W_{\text{diss}}^{\text{jc},*}(N) = W_{\text{diss}}^{\text{jc}}(N) + \beta^{-1} \sum_{k=1}^N (2\epsilon_k - \epsilon_k \ln \epsilon_k). \quad (\text{D6})$$

The quantum state tomography algorithm in [18] 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 ,

$$\sum_{k=1}^N \epsilon_k \leq C \ln \frac{N}{\delta} \ln N. \quad (\text{D7})$$

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 $\epsilon_k \leq C(\ln(N/\delta) \ln N)/N$,

$$-\epsilon_k \ln \epsilon_k \leq \frac{C \ln \frac{N}{\delta} \ln N}{N} \ln \frac{N}{C \ln \frac{N}{\delta} \ln N}, \quad (\text{D8})$$

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

$$-\epsilon_k \ln \epsilon_k \leq \epsilon_k \ln \frac{N}{C \ln \frac{N}{\delta} \ln N}, \quad (\text{D9})$$

Then,

$$-\sum_{k=1}^N \epsilon_k \ln \epsilon_k = - \sum_{\epsilon_k: \epsilon_k \leq \frac{C \ln \frac{N}{\delta} \ln N}{N}} \epsilon_k \ln \epsilon_k - \sum_{\epsilon_k: \epsilon_k > \frac{C \ln \frac{N}{\delta} \ln N}{N}} \epsilon_k \ln \epsilon_k \quad (\text{D10})$$

$$\leq \sum_{\epsilon_k: \epsilon_k \leq \frac{C \ln \frac{N}{\delta} \ln N}{N}} \frac{C \ln \frac{N}{\delta} \ln N}{N} \ln \frac{N}{C \ln \frac{N}{\delta} \ln N} + \sum_{\epsilon_k: \epsilon_k > \frac{C \ln \frac{N}{\delta} \ln N}{N}} \epsilon_k \ln \frac{N}{C \ln \frac{N}{\delta} \ln N} \quad (\text{D11})$$

$$\leq C \ln \frac{N}{\delta} \ln N \ln \frac{N}{C \ln \frac{N}{\delta} \ln N} + C \ln \frac{N}{\delta} \ln N \ln \frac{N}{C \ln \frac{N}{\delta} \ln N} \leq 2C \ln \frac{N}{\delta} \ln N \ln \frac{N}{C \ln \frac{N}{\delta} \ln N}. \quad (\text{D12})$$

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

$$\sum_{t=1}^N \epsilon_k \ln \frac{1}{\epsilon_k} \leq \frac{N}{e} \leq C \ln \frac{N}{\delta} \ln N. \quad (\text{D13})$$

Combining both cases, we conclude that

$$-\sum_{t=1}^N \epsilon_k \ln \epsilon_k \leq O((\ln N)^3). \quad (\text{D14})$$

Substituting into Eq. (D2), we obtain the dissipation for the semi-classical model,

$$W_{\text{diss}}^{\text{sc},*}(N) \leq W_{\text{diss}}^{\text{sc}}(N) + \beta^{-1} O((\ln N)^3) = O((\ln N)^3). \quad (\text{D15})$$

and the Jaynes-Cummings battery model,

$$W_{\text{diss}}^{\text{jc},*}(N) \leq W_{\text{diss}}^{\text{jc}}(N) + 2\beta^{-1} O((\ln N)^3) = O((\ln N)^3), \quad (\text{D16})$$

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.

$$M_0 = \underbrace{\{0, 0, \dots, 0\}}_N \quad (\text{D17})$$

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 other words, after time step t , the memory register takes the form

$$M_t = \{r_1, r_2, \dots, r_t, 0, \dots, 0\} \quad \text{for } t \in \{1, 2, \dots, N\} . \quad (\text{D18})$$

At the end of all the extractions, the memory M_N can then be quasi-statically reset. As previously calculated, as $t \rightarrow N$, 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 $O(\text{polylog}(N))$.