Shakin' All Over: Proving Landauer's Principle without neglect of fluctuations

Wayne C. Myrvold
Department of Philosophy
The University of Western Ontario
wmyrvold@uwo.ca

July 22, 2020

Abstract

Landauer's principle is, roughly, the principle that there is an entropic cost associated with implementation of logically irreversible operations. Though widely accepted in the literature on the thermodynamics of computation, it has been the subject of considerable dispute in the philosophical literature. Both the cogency of proofs of the principle and its relevance, should it be true, have been questioned. In particular, it has been argued that microscale fluctuations entail dissipation that always greatly exceeds the Landauer bound. In this article Landauer's principle is treated within statistical mechanics, and a proof is given that neither relies on neglect of fluctuations nor assumes the availability of thermodynamically reversible processes. In addition, it is argued that microscale fluctuations are no obstacle to approximating thermodynamic reversibility as closely as one would like.

1 Introduction

The statement that has come to be known as *Landauer's Principle* is, roughly, that there is an entropic cost associated with implementation of logically irreversible operations, that is, operations whose input states cannot be recovered from their output states. It is widely

accepted in the literature on the thermodynamics of computation; see Leff and Rex (2003) for a sampling of the relevant literature and an extensive bibliography. Nonetheless, it has been the subject of considerable controversy in the philosophical literature (Earman and Norton, 1999; Norton, 2005; Ladyman et al., 2007, 2008; Norton, 2011; Ladyman and Robertson, 2013; Norton, 2013a,b,c; Ladyman and Robertson, 2014; Ladyman, 2018; Norton, 2018).

Ladyman, Presnell, Short, and Groisman (2007), hereinafter referred to as LPSG, presented a proof of Landauer's principle. The proof, like any proof, rests on assumptions. The operative assumptions of the proof are that a probabilistic version of the second law of thermodynamics holds, and that certain processes can be performed reversibly. These processes include, crucially, expansion of a single-molecule gas. Norton (2011, 2013b,c) has argued that inevitable fluctuations at the molecular level invalidate the assumption of even approximate thermodynamic reversibility of processes at the microscale, and that any process involves dissipation in excess of the bounds required by Landauer's principle, rendering the principle moot. This is regarded by Norton as a 'no-go' result, invalidating the basic framework within which most of the work on thermodynamics of computation has been carried out.

Ladyman and Robertson (2014) addressed the purported no-go result, arguing that the conclusion has not been established. They acknowledged, however, a concern about the assumption, ubiquitous in the literature on thermodynamics of computation, of molecular-scale processes carried out with negligible dissipation.

In this article, the subject of Landauer's principle is addressed from the point of view of statistical mechanics. It is shown that the relevant version of the second law of thermodynamics is provable within statistical mechanics, in two versions, classical and quantum. It is therefore not required as an independent assumption. A derivation within statistical mechanics of the Landauer principle is given, that neither relies on neglect of fluctuations nor assumes the availability of thermodynamically reversible processes.

As Norton has rightly emphasized, a theorem of this sort is moot if the processes involved depart sufficiently far from thermodynamic reversibility. This is explicit in the theorem we prove. Unless there are processes available that approximate reversibility sufficiently closely, the theorem places no bounds on *extra* dissipation associated with logical irreversibility. For that reason, I will argue that, given the notion

of thermodynamic reversibility relevant to the context at hand, fluctuations, even ones that are large on the scale at which the processes are taking place, pose no threat to the assumption that processes can take place that approximate thermodynamic reversibility as closely as we would like.

2 The set-up

As is usual in thermodynamics, the thermodynamic state of a system A is defined with respect to some set of manipulable variables $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, which may represent, for example, the positions of the walls of a container the system is constrained to be in, or the value of applied fields. We thus consider a family of Hamiltonians $\{H_{\boldsymbol{\lambda}}\}$. The variables $\boldsymbol{\lambda}$ are treated as exogenous, meaning that we do not include in our physical description the systems that are the sources of these applied fields, and we do not consider the influence of the system A on those systems. They may also be freely specified, independently of the state of the system. We consider some set $\mathcal{M}$ of manipulations of the system, where each manipulation consists of some specification of $\boldsymbol{\lambda}(t)$ through some interval $t_0 \leq t \leq t_1$. In addition, we may assume that there are available one or more heat reservoirs $\{B_i\}$ at temperatures T_i, with which the system can exchange heat. The system A may be coupled and decoupled from these heat baths during the course of its evolution. That is, the interaction terms in the total Hamiltonian consisting of the system A and the reservoirs $\{B_i\}$ are also treated as manipulable variables.

Since the time of Maxwell (1871, 1878) it has been recognized that the kinetic theory of heat entails that the second law of thermodynamics, as originally formulated, cannot hold strictly. When compressing a gas with a piston, we might find on some occasion that, due to a fluctuation in the force exerted by the gas on the piston, less work is needed to compress the gas than one would expect on average, and so in a given cycle of a heat engine might obtain more work than is allowed by the second law from a given quantity of heat extracted. By the same token we might obtain less work than expected. We do not, however, expect that we will be able to *consistently and reliably* violate the Carnot limit on efficiency of a heat engine. The original version of the second law should be replaced by a probabilistic one. The second law will then be, to employ Szilard's vivid analogy, like

a theorem about the impossibility of a gambling system intended to beat the odds set by a casino.

> Consider somebody playing a thermodynamical gamble with the help of cyclic processes and with the intention of decreasing the entropy of the heat reservoirs. Nature will deal with him like a well established casino, in which it is possible to make an occasional win but for which no system exists ensuring the gambler a profit (Szilard 1972, p. 73, from Szilard 1925, p. 757).

On a macroscopic scale, we expect fluctuations to be negligible, but, as Norton has emphasized, at the microscale on which in-principle limitations on the thermal cost of computation are investigated, they are far from negligible. Accordingly, we will invoke probabilistic considerations, and treat of the evolution of probability distributions over the state of a system subjected to various manipulations. When considering the amount of work needed to perform an operation, or the amount of heat exchanged in the course of the evolution of the system, we will consider *expectation values* of work and heat exchanges, calculated with respect to those probability distributions.

We assume it makes sense to associate a probability distribution with a preparation procedure, and to compute on its basis probabilities for outcomes of subsequent manipulations. We need not enquire into the status of these probabilities, so long as they serve this purpose.

Since the late nineteenth century it has been common to think of probability statements as involving veiled reference to relative frequencies in an actual or hypothetical sequence of events, or in an ensemble of similarly prepared systems. There is no commitment here to such frequentism about probabilities; probability considerations may be applied to single events. There is, however, a link between probabilities and mean outcomes in a long sequence of trials, afforded by the weak law of large numbers. Suppose that we are able to conduct multiple runs of a procedure, in such a way that the probabilities of the outcomes are the same on each trial, and the outcomes of any trial are probabilistically independent of the outcomes of all the others. Then, if we take any outcome variable, and compute its mean value across the results of a long sequence of trials, with high probability this mean value will be close to its expectation value on a single trial. We can make the probability of any degree of approximation as high as we like by increasing the number of trials. Though the expectation values to

be invoked are not *defined* in terms of mean values in a long sequence of trials, they have implications for such mean values. If we could construct a heat engine such that the expectation value of work extracted on each run exceeded the Carnot bound, we could, by running sufficiently many cycles, make the probability of a net violation of the Carnot bound as close to unity as we like.

We will treat of "states" $a = (\rho_a, H_a)$, consisting of a probability distribution over the phase space of the system A (or, in the quantum context, a density operator on the system's Hilbert space), represented by a density ρ_a, sand a Hamiltonian, which, as already noted, may depend on exogenous, manipulable variables. We consider the effects on those states of manipulations in some class $\mathcal{M}$.

As is usual in statistical mechanics, the distributions associated with the heat reservoirs B_i will be canonical distributions, uncorrelated with the system A (see Maroney 2007 for discussion of the justification for this use of canonical distributions). In the classical context, a canonical distribution is a distribution that has density, with respect to Liouville measure,

$$\rho_\beta = Z^{-1} e^{-\beta H}, \tag{1}$$

where β is the inverse temperature $1/kT$, and Z is the normalization constant required to make the integral of this density over all phase space unity. This depends both on the Hamiltonian H and on β, and is called the *partition function*. In the quantum context, a canonical state is represented by density operator

$$\hat{\rho}_\beta = Z^{-1} e^{-\beta \hat{H}}, \tag{2}$$

where, again, Z is the constant required to normalize the state.

As the reservoirs interact with A, correlations will be built up, but we will assume that the reservoirs are big enough and noisy enough that these are, as far as subsequent interactions with A are concerned, effectively effaced, meaning that the effect of the reservoirs on A is *as if* they are uncorrelated. This means, not that the probability distribution over the full state of A and B_i is a product distribution, but that the dynamical variables of B_i and A relevant to interactions their interactions with A are effectively independent.

The manipulations of a system A we will be considering will be ones of the following form.

- At time t_0, the system has some probability distribution ρ_a, and the Hamiltonian of the system A is H_a.

- At time t_0, the heat reservoirs B_i have canonical distributions at temperatures T_i, uncorrelated with A, and are not interacting with A.
- During the time interval $[t_0, t_1]$, the composite system consisting of A and the reservoirs $\{B_i\}$ undergoes Hamiltonian evolution, governed by a time-dependent Hamiltonian $H(t)$, which may include successive couplings between A and the heat reservoirs $\{B_i\}$.
- The internal Hamiltonians of the reservoirs $\{B_i\}$ do not change.
- At time t_1, the Hamiltonian of the system A is H_b, and, as a result of Hamiltonian evolution of the composite system, the marginal probability distribution of A is ρ_b.

This is a manipulation that takes a state $a = (\rho_a, H_a)$ to state $b = (\rho_b, H_b)$.

It should be noted that we are *not* considering manipulations that consist of a measurement performed on the system A followed by a manipulation of the exogenous variables whose choice depends on the outcome of the measurement. Controlled operations are allowed, but the control mechanism must be internalized, that is, included in the system under study. The system A could consist of two parts A_1 and A_2, which interact in such a way that the state of A_1 affects what happens to A_2, which subsequently affects what happens to A_1. But all of this must be encoded in the Hamiltonian $H(t)$, which may be time-varying but which undergoes a preprogrammed evolution that is *not* dependent on the state of the system A. Otherwise, there may be dissipation associated with the operation of the control mechanism that gets left out of the analysis.

We will count energy exchanges with the reservoirs B_i as heat (to be counted as positive if A gains energy from B, negative if A loses energy), and energy changes to A due to changes in the external potentials as work (again, counted as positive if A gains energy, negative if it loses energy).

Dropping the assumption of the availability of reversible processes requires revision of the familiar framework of thermodynamics, as it means dropping the assumption of the availability of an entropy function. In its place we will define quantities $S_{\mathcal{M}}(a \to b)$, defined relative to a class of available manipulations $\mathcal{M}$, to be thought of as analogs, in the current context, of entropy differences between states a and b. These will be representable as differences in the values of some state

function only in the limiting case in which all states can be connected reversibly.

For any manipulation M, that takes a state a to a state b, we can define $\langle Q_i(a \to b)\rangle_M$ as the expectation value of the heat obtained by A from reservoir B_i. We can use these to define,

$$\sigma_M(a \to b) = \sum_i \frac{\langle Q_i(a \to b)\rangle_M}{T_i}. \tag{3}$$

Let $\mathcal{M}(a \to b)$ be the set of manipulations in $\mathcal{M}$ that take a to b, and define, as analogs of entropies (which we will henceforth just call "entropies"),

$$S_{\mathcal{M}}(a \to b) = \text{l.u.b.}\{\sigma_M(a \to b) \,|\, M \in \mathcal{M}(a \to b)\}. \tag{4}$$

Via the obvious extension of this definition we also define quantities such as $S_{\mathcal{M}}(a \to b \to c)$ for processes with any number of intermediate steps. It is assumed that manipulations can be composed, that is, that any manipulation that takes a to b can be followed by one that takes b to c to form a manipulation that takes a to b and then to c. It follows from this composition assumption and the definition of the entropies that

$$S_{\mathcal{M}}(a \to b \to c) = S_{\mathcal{M}}(a \to b) + S_{\mathcal{M}}(b \to c), \tag{5}$$

and similarly for processes consisting of longer chains of intermediates states.

One version of the second law of thermodynamics says that, for any cyclic process, the sum of Q_i/T_i over all heat exchanges cannot be positive. Since we're working in the context of statistical mechanics, and we do not want to ignore fluctuations, the appropriate revision of the second law involves expectation values of heat exchanges. A cyclic process will be one that restores the marginal probability distribution of the system A to the one it started out with. The revised second law that we will prove in the next section states that, for any cyclic process, the sum of $\langle Q_i\rangle/T_i$ over all heat exchanges cannot be positive. In the notation we have introduced, this is:

The Statistical Second Law. For any state a,

$$S_{\mathcal{M}}(a \to a) \leq 0.$$

It follows from this that

$$S_{\mathcal{M}}(a \to b \to a) = S_{\mathcal{M}}(a \to b) + S_{\mathcal{M}}(b \to a) \leq 0, \tag{6}$$

and similarly for processes involving longer chains of intermediate states.

In any process M that takes a state a to a state b, some of the work done, or heat discarded into a reservoir, may be recovered by some process that takes b back to a. If the process can be reversed with the signs of all $\langle Q_i \rangle$ reversed, then full recovery is possible. If full recovery is not possible, and cannot even be approached arbitrarily closely, we will say that the process is *dissipatory*. A manipulation M' that takes b to a and recovers work done and heat discarded would be one such that

$$\sigma_M(a \to b) + \sigma_{M'}(b \to a) = 0. \tag{7}$$

There might be a limit to how closely this can be approached. Define the dissipation associated with the process of M taking a to b as the distance between this limit and perfect recovery.

$$\begin{aligned} \delta_M(a \to b) &= \text{g.l.b.}\{-(\sigma_M(a \to b) + \sigma_{M'}(b \to a)) \mid M' \in \mathcal{M}(b \to a)\} \\ &= -S_{\mathcal{M}}(b \to a) - \sigma_M(a \to b). \end{aligned} \tag{8}$$

It follows from the statistical second law that this is non-negative.

If there is no limit to how much the dissipation associated with processes that connect a to b can be diminished,

$$S_{\mathcal{M}}(a \to b \to a) = 0. \tag{9}$$

When this holds, it is traditional to say that a and b can be connected reversibly, and to imagine a fictitious process that can proceed in either direction, reversing the signs of all heat exchanges. There is no harm in doing so, as long as this is not taken too literally.[1] Following convention, we will say, for any a, b for which (9) is satisfied, that a and b can be connected reversibly. When this locution is used, bear in mind that it is shorthand for (9), and does not presume the existence of an actual reversible process.

From the statistical second law it follows that, if all states can be connected reversibly—that is, if, for all a, b, $S_{\mathcal{M}}(a \to b \to a) = 0$—then there exists a state function $S_{\mathcal{M}}$, defined up to an additive constant, such that

$$S_{\mathcal{M}}(a \to b) = S_{\mathcal{M}}(b) - S_{\mathcal{M}}(a). \tag{10}$$

[1] As Norton (2016) has argued, taking talk of irreversible processes too literally can lead to contradictions.

This is the familiar entropy function. The reason we have been expressing things in an unfamiliar way is that we *don't* want to assume reversibility as a general rule.

Any manipulation that takes a to b must have dissipation of at least $-S_{\mathcal{M}}(a \to b \to a)$. Define the *inefficiency* associated with a manipulation that takes a to b as the amount by which its dissipation exceeds this minimal value.

$$\begin{aligned} \eta_M(a \to b) &= \delta_M(a \to b) - (-S_{\mathcal{M}}(a \to b \to a)) \\ &= S_{\mathcal{M}}(a \to b) - \sigma_M(a \to b). \end{aligned} \tag{11}$$

If a and b can be connected reversibly, the distinction between dissipation and inefficiency vanishes, and the inefficiency is equal to the dissipation.

We are now in a position to state the version of Landauer's principle that we will be proving. Consider a logical operation L that is not logically reversible, meaning that the input is not recoverable from the output. This means that there are two or more inputs $\{\alpha_i\}$ that are mapped by L to the same output β. In a device that implements the logical operation L, the inputs will be represented by statistical mechanical states $\{a_i\}$, and the output by a statistical mechanical state b. Distinct inputs are to be represented by distinguishable states, which, in the classical context, means probability distributions with non-overlapping support, and in the quantum-mechanical context, by orthogonal density operators. An implementation of L is a manipulation M_L that maps each member of the set $\{a_i\}$ into b.

The question to be asked is: can the manipulation M_L do this without incurring any inefficiency? That is, can we have $\eta_{M_L}(a_i \to b)$ equal to zero, for every a_i? Failing that, can we, by appropriate choice of manipulation, make every element of the set $\{\eta_{M_L}(a_i \to b)\}$ arbitrarily small?

In the next section we will prove the following.

> **Landauer bound on dissipations.** If manipulation M takes each of a distinguishable set $\{a_i, i = 1, \ldots, n\}$ of states to the same state b, then
>
> $$\sum_{i=1}^{n} e^{-\delta_M(a_i \to b)/k} \leq 1.$$

This entails that every member of the set $\{\delta_M(a_i \to b)\}$ is greater than zero. It also entails a formulation that is often presented as a

gloss of Landauer's principle, that the mean of the set is not smaller than $k \log n$.[2]

$$\frac{1}{n} \sum_{i=1}^{n} \delta_M(a_i \to b) \geq k \log n. \tag{12}$$

That is, there an *average* dissipation, taken over members of the set $\{a_i\}$, of at least $k \log n$.[3] We might be able to reduce the dissipation associated with any particular member of the set as much as we like, but we cannot simultaneously make all of them arbitrarily small. For the case of $n = 2$, the most commonly discussed case, the constraint is graphed in Figure 1. The shaded region is the set of permitted pairs $(\delta_1, \delta_2) = (\delta_M(a_1 \to b)/k, \delta_M(a_2 \to b)/k)$.

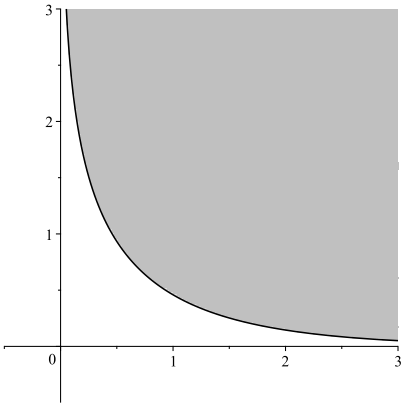

Figure 1: Values of (δ_1, δ_2) permitted by Landauer's principle.

If, as is usually assumed in these discussions, the states $\{a_i\}$ can be connected reversibly to b, then any dissipation is inefficiency, and bounds on dissipations are bounds on inefficiencies. If reversibility is not assumed, there may be unavoidable levels of dissipation associated with some state transitions; if this is the case, not every dissipation represents an inefficiency. We can re-state the Landauer principle in terms of inefficiencies.

> **Landauer bound on inefficiencies.** If manipulation M takes each of a distinguishable set $\{a_i, i = 1, \ldots, n\}$ of

[2]In this article, all logarithms are natural logarithms, that is, logarithms to the base e.

[3]See Appendix for proof that this is entailed by our formulation of the Landauer principle.

states to the same state b, then

$$\sum_{i=1}^{n} e^{-(\eta_i - S_{\mathcal{M}}(a_i \to b \to a_i))/k} \leq 1,$$

where η_i is the inefficiency $\eta_M(a_i \to b)$.

If we have reversibility, then this entails that all of the inefficiencies $\eta_M(a_i \to b)$ must be positive, and that they cannot all be made arbitrarily small in the same process. Far enough from reversibly, it places no constraint on inefficiencies at all. The condition for the Landauer bound to place a constraint on inefficiencies is,

$$\sum_{i=1}^{n} e^{S_{\mathcal{M}}(a_i \to b \to a_i)/k} > 1. \tag{13}$$

A necessary condition for (13) to be satisfied, and thus for the Landauer principle to have teeth, is the condition that, for some a_i,

$$S_{\mathcal{M}}(a_i \to b \to a_i) > -k \log n. \tag{14}$$

If Norton (2011, 2013b,c) is right about the minimum dissipation required for carrying out processes at the molecular level, then (14) is *not* satisfiable; because of fluctuations at the molecular level, any process departs from reversibility by an amount that far exceeds the Landauer bound. In section 4 it will be argued that this is not correct, and the Landauer principle does have teeth.

The Landauer bound we have stated involves a distinguishable set of states. Distinguishability, like reversibility, is something that we should not expect to hold perfectly; in actual implementations we will at best approximate perfect distinguishability. For this reason, the theorem that we will prove in the next section will not require perfect distinguishability, and will entail the version of the Landauer bound we have stated in this section as a special case.

3 Proving the Second Law, and Landauer's Principle

The theorems we will be concerned with come in two versions, classical and quantum, each proven in pretty much the same way. To

avoid saying everything twice, we adopt a systematically ambiguous notation, and state each theorem in such a way that it can be read either as a theorem of classical statistical mechanics, or as a theorem of quantum statistical mechanics.

In what follows, ρ will be used either for a density function, with respect to Liouville measure, on a classical phase space, or, in the quantum context, a density operator on a Hilbert space. $S[\rho]$ is the Gibbs entropy (classical), or the von Neumann entropy (quantum).

$$S[\rho] = -k \langle \log \rho \rangle_\rho. \tag{15}$$

We also define the relative entropy of two distributions.

$$S[\rho \,\|\, \sigma] = -k \left(\langle \log \sigma \rangle_\rho - \langle \log \rho \rangle_\rho \right). \tag{16}$$

$S[\rho \,\|\, \sigma]$ is one way to measure how much the distribution represented by σ departs from that represented by ρ. It is equal to zero for $\sigma = \rho$, and is positive for any other σ.

Suppose $\bar{a}$ is a probabilistic mixture of states $\{a_i\}$.

$$\rho_{\bar{a}} = \sum_{i=1}^{n} p_i \, \rho_{a_i}, \tag{17}$$

where $\{p_i\}$ are positive numbers that add up to one. Then the Gibbs/von Neumann entropy of $\bar{a}$ is related to that of the a_i's via,

$$S[\rho_{\bar{a}}] = \sum_{i=a}^{n} p_i \, S[\rho_{a_i}] + \sum_{i=a}^{n} p_i \, S[\rho_{a_i} \,\|\, \rho_{\bar{a}}]. \tag{18}$$

If the states $\{a_i\}$ are distinguishable, then $S[\rho_{a_i} \,\|\, \rho_{\bar{a}}] = -k \log p_i$, and so

$$S[\rho_{\bar{a}}] = \sum_{i=a}^{n} p_i \, S[\rho_{a_i}] - k \sum_{i=1}^{n} p_i \log p_i. \tag{19}$$

As outlined in the previous section, we are concerned with a system A evolving between times t_0 and t_1 according to a time-varying Hamiltonian, and interacting successively with one or more heat baths $\{B_i\}$, which initially have canonical distributions at temperatures T_i. The Hamiltonians of the heat baths remain fixed throughout the evolution. We define

$$\langle Q_i \rangle = -\Delta \langle H_{B_i} \rangle = -\left(\langle H_{B_i} \rangle_{\rho_{B_i}(t_1)} - \langle H_{B_i} \rangle_{\rho_{B_i}(t_0)} \right). \tag{20}$$

This is the expectation value of the heat energy obtained by A from B_i.

Our first theorem relates the entropies as defined in the previous section to the Gibbs/von Neumann entropies. Though a simple one, it is of fundamental importance in the foundations of statistical mechanics, and deserves to be called the Fundamental Theorem of Statistical Mechanics.[4]

Proposition 1. *If $\mathcal{M}$ is a class of manipulations of the sort outlined in Section 2, then, for any states a, b,*

$$S_{\mathcal{M}}(a \to b) \leq S[\rho_b] - S[\rho_a].$$

The following are immediate corollaries of this.

Corollary 1.1. ***The second law of statistical thermodynamics.*** *For any state a,*

$$S_{\mathcal{M}}(a \to a) \leq 0.$$

Corollary 1.2. *If a and b can be connected reversibly—that is, if*

$$S_{\mathcal{M}}(a \to b \to a) = 0,$$

then

$$S_{\mathcal{M}}(a \to b) = S[\rho_b] - S[\rho_a].$$

Thus, the Gibbs/von Neumann entropy is the state function whose existence is guaranteed by the second law plus reversibility.

Now, to the Landauer principle. If a manipulation M takes each of the states $\{a_i\}$ to the same state b, then it must also take any probabilistic mixture $\bar{a}$ of these states to b. Let $\bar{a}$ be a mixture of the states $\{a_i\}$ with weights $\{p_i\}$. The expectation value of heat exchanged when M is applied to this mixture is a weighted average of exchanges that would occur in the states $\{a_i\}$, and so

$$\sigma_M(\bar{a} \to b) = \sum_{i=1}^{n} p_i \, \sigma_M(a_i \to b). \tag{21}$$

[4]This is not a new theorem. The classical version of it is found in Gibbs (1902, pp. 160–164), and the quantum version, in Tolman (1938, §128–130). Nonetheless, it is not as well-known in the philosophical literature on statistical mechanics and thermodynamics as it should be. Maroney (2009) refers to it as a *generalized Landauer principle*.

We must have, of course,

$$\sigma_M(\bar{a} \to b) \leq S_{\mathcal{M}}(\bar{a} \to b). \tag{22}$$

By the Fundamental Theorem,

$$S_{\mathcal{M}}(\bar{a} \to b) \leq S[\rho_b] - S[\rho_{\bar{a}}]. \tag{23}$$

From (18), the right-hand side of this is

$$S[\rho_b] - S[\rho_{\bar{a}}] = \sum_{i=1}^{n} p_i \left(S[\rho_b] - S[\rho_{a_i}]\right) - \sum_{i=1}^{n} p_i \, S[\rho_{a_i} \,\|\, \rho_{\bar{a}}]. \tag{24}$$

Employing the Fundamental Theorem again,

$$S[b] - S[a_i] \leq -S_{\mathcal{M}}(b \to a_i). \tag{25}$$

Combining (21), (22), (23), (24), and (25) gives us

$$\sum_{i=1}^{n} p_i \, \sigma_M(a_i \to b) \leq \; -\sum_{i=1}^{n} p_i \, S_{\mathcal{M}}(b \to a_i) - \sum_{i=1}^{n} p_i \, S[\rho_{a_i} \,\|\, \rho_{\bar{a}}]. \tag{26}$$

Rearranging, and recalling the definition (8) of the dissipations, we get

$$\sum_{i=1}^{n} p_i \, \delta_M(a_i \to b) \geq \sum_{i=1}^{n} p_i \, S[\rho_{a_i} \,\|\, \rho_{\bar{a}}]. \tag{27}$$

Thus, we have the result,

Proposition 2. *For any manipulation M that takes each of $\{a_i\}$ to b, and any positive numbers $\{p_i\}$ such that*

$$\sum_{i=1}^{n} p_i = 1,$$

we have

$$\sum_{i=1}^{n} p_i \, \delta_M(a_i \to b) \geq \sum_{i=1}^{n} p_i \, S[\rho_{a_i} \,\|\, \rho_{\bar{a}}],$$

where $\bar{a}$ is a mixture of states $\{a_i\}$ with weights $\{p_i\}$.

This is our general version of Landauer's principle. If we apply this to the case in which the states $\{a_i\}$ are distinguishable, we get the following corollary.

Corollary 2.1. *For any manipulation M that takes each of a distinguishable set $\{a_i\}$ to b, and any positive numbers $\{p_i\}$ such that*

$$\sum_{i=1}^{n} p_i = 1,$$

we have

$$\sum_{i=1}^{n} p_i \, \delta_M(a_i \to b) \geq -k \sum_{i=1}^{n} p_i \log p_i.$$

As shown in the Appendix, this is equivalent to the following,

Corollary 2.2. *For any manipulation M that takes each of a distinguishable set $\{a_i\}$ to b,*

$$\sum_{i=1}^{n} e^{-\delta_M(a_i \to b)/k} \leq 1.$$

This is the version stated in the previous section.

4 Approximating reversibility

The second law of statistical thermodynamics entails that, for any a, b,

$$S_{\mathcal{M}}(a \to b \to a) \leq 0. \tag{28}$$

We do not expect there to be any process that takes a to b and then back to a without any dissipation. However, if the array of permitted manipulations is sufficiently rich, there maybe no bound on dissipation short of zero, and we may have $S_{\mathcal{M}}(a \to b \to a) = 0$.

One way to have a process that proceeds with negligibly small dissipation is to keep the system A in contact with a heat reservoir large and noisy enough that the reservoir may be regarded as canonically distributed throughout the process, and to vary the parameters $\boldsymbol{\lambda}$ slowly enough that the time it takes for any appreciable change in these parameters is long compared to the equilibration time-scale of the system A. Then the system A may be treated as if it is in equilibrium with the reservoir at each stage of the process.[5] We can also

[5]This does not, of course, mean that it *is* in equilibrium, only that, for the purposes at hand, differences between quantities calculated on the basis of the equilibrium distribution and quantities calculated on the basis of the actual distribution are small enough that they may be neglected.

consider slowly varying the temperature of the reservoir. For a process like that, at any time t during the process A may be treated as having a canonical distribution for the instantaneous parameter values $(\boldsymbol{\lambda}(t), \beta(t))$.

If ρ_1 is a canonical distribution for parameters $(\boldsymbol{\lambda}, \beta)$, and ρ_2 a canonical distribution for slightly differing parameters $(\boldsymbol{\lambda}+d\boldsymbol{\lambda}, \beta+d\beta)$, then, to first order in the parameter differences,[6]

$$d\langle H\rangle = \langle H_2\rangle_{\rho_2} - \langle H_1\rangle_{\rho_1} = \sum_i \left\langle \frac{\partial H}{\partial \lambda_i} \right\rangle_{\rho_1} d\lambda_i - \beta^{-1}\, d\langle \log \rho\rangle. \qquad (29)$$

The first term on the right-hand side of this equation is the expectation value of the work done in changing the external parameters; the remainder is the expectation value of the heat obtained from the reservoir.

$$\langle đ Q\rangle = -kT\, d\langle \log \rho\rangle, \qquad (30)$$

where $kT = \beta^{-1}$. This means that, for a process in the course of which the system A is in continual contact with a heat reservoir at temperature T and the parameters $\boldsymbol{\lambda}$ are varied slowly from values $\boldsymbol{\lambda}_a$ to $\boldsymbol{\lambda}_b$, the expectation value of total heat absorbed will have the approximate value

$$\langle Q(a \to b)\rangle \approx -kT(\langle \log \rho_b\rangle_{\rho_b} - \langle \log \rho_a\rangle_{\rho_a}) = T\left(S[\rho_b] - S[\rho_a]\right). \qquad (31)$$

As long as there is no in-principle limit to how much time a state-transformation may take, there is no in-principle limit to how closely this approximation can hold, and equality will be approached as the time-scale of the changes in the parameters $\boldsymbol{\lambda}$ is increased, relative to the time-scale of equilibration of the system A.

The result (31) is a result about expectation values. It is *not* assumed that the actual value of heat exchanged will be close to its expectation value, or even that it will *probably* be close to its expectation value. The probability distribution for the heat exchange may have a large variance, and probabilities of large deviations from the expectation value may be far from negligible. That is, the result does *not* depend on disregard of fluctuations. When we say that the system has time to equilibrate, this does not mean that it is ever in a quiescent state, only that its distribution may be treated as canonical at each stage of the process.

[6] The classical version of this eq. (112) on p. 44 of Gibbs (1902), and the quantum, eq. (121.8) on p. 534 of Tolman (1938).

Let a, b be canonical states with parameters $(\boldsymbol{\lambda}_a, \beta_a)$, $(\boldsymbol{\lambda}_b, \beta_b)$. We will say that a class of manipulations $\mathcal{M}$ connects a and b quasi-statically if

1. $\mathcal{M}$ contains manipulations of the following form
   (a) During time interval $[t_0, t_0 + T]$, the parameters undergo smooth evolution $\boldsymbol{\lambda}(t)$, with $\boldsymbol{\lambda}(t_0) = \boldsymbol{\lambda}_a$ and $\boldsymbol{\lambda}(t_0+T) = \boldsymbol{\lambda}_b$.
   (b) At time t the system A is in thermal contact with a heat reservoir at inverse temperate $\beta(t)$, where $\beta(t)$ is a smooth function with $\beta(t_0) = \beta_a$ and $\beta(t_0 + T) = \beta_b$.
2. For any such manipulation, there is one that proceeds twice as slowly. That is, there is a manipulation that takes place in time interval $[t_0, t_0 + 2T]$, with parameter values λ', β' where

$$\boldsymbol{\lambda}'(t_0 + t) = \boldsymbol{\lambda}(t_0 + t/2); \quad \beta'(t_0 + t) = \beta(t_0 + t/2).$$

   for $t \in [0, 2T]$.

Then we have the following result.

Proposition 3. *If a, b are canonical states, and $\mathcal{M}$ is a class of manipulations that connects a to b quasi-statically, then*

$$S_{\mathcal{M}}(a \to b) = S[\rho_b] - S[\rho_a].$$

We have, as a trivial corollary,

Corollary 3.1. *If a, b are canonical states, and $\mathcal{M}$ is a class of manipulations that connects a to b quasi-statically, and also connects b to a quasi-statically, then*

$$S_{\mathcal{M}}(a \to b \to a) = 0.$$

Suppose that we have a system to which can be applied a manipulable external potential $V_{\boldsymbol{\lambda}}$, and which can also be confined, by suitable barriers, to various regions $\{\Gamma_i\}$ of its state space. Let $\{a_i\}$ be a finite set of canonical states, confined to the regions $\{\Gamma_i\}$, with values $\boldsymbol{\lambda}_a$ of the manipulable parameters $\boldsymbol{\lambda}$ on which the external potential depends, and let $\{b_i\}$ be a set of canonical distributions confined to the same regions, with parameter values $\boldsymbol{\lambda}_b$. Then, for any desired degree of approximation to the quasistatic limit, we can find a sufficiently slow variation of the parameters $\boldsymbol{\lambda}$ that yields the desired degree of approximation for *all* of the transitions $a_i \to b_i$. We will say, of such a situation, that $\mathcal{M}$ uniformly quasi-statically connects $\{a_i\}$ to $\{b_i\}$. We have, as another corollary to Proposition (3).

Corollary 3.2. *Let $\{a_i\}$, $\{b_i\}$ be sets of canonical states, such that $\mathcal{M}$ uniformly quasi-statically connects $\{a_i\}$ to $\{b_i\}$ and $\{b_i\}$ to $\{a_i\}$. Let $\{p_i\}$ be a set of non-negative numbers that sum to 1, and let $\bar{a}$ and $\bar{b}$ be probabilistic mixtures of $\{a_i\}$ and $\{b_i\}$ with weights $\{p_i\}$. Then*

$$S_{\mathcal{M}}(\bar{a} \to \bar{b} \to \bar{a}) = 0.$$

5 Example: One-particle gas

The simplest example I can think of for illustrating erasure that is a single particle in a box, with a partition that can be inserted and removed. If this is the only available manipulation, $S_{\mathcal{M}}(a \to b)$ will be zero for all states a, b of the same temperature. To get nontrivial entropies, we need to introduce the possibility of doing work on and obtaining work from the system.

Suppose that the particle can be subjected to an external potential V_λ, that varies in the x-direction only. We take the system to be in thermal equilibrium with a heat bath at temperature T. On a canonical distribution, the distributions of the momentum $\mathbf{p}$ and the coordinates other than x are unchanged when the potential V_λ is varied. We therefore integrate these out, and consider the marginal distribution of the coordinate x.

$$\rho_{\lambda,\beta}(x) = \begin{cases} Z_{\lambda,\beta}^{-1}\, e^{-\beta\, V_\lambda(x)}, & \text{inside the container;} \\ 0. & \text{outside.} \end{cases} \tag{32}$$

Take the x-coordinate within the container to range from $-l$ to l. The partition function is

$$Z_{\lambda,\beta} = \int_{-l}^{l} e^{-\beta\, V_\lambda(x)}\, dx. \tag{33}$$

We need not assume that the potential V_λ is under perfect control. It, too, may fluctuate, with its own probability distribution. Evolution of a probability distribution, via the Liouville equation, of a system subject to a potential V that fluctuates with a probability distribution of its own, independent of the state of the system, is the same as evolution under a steady potential equal to the expectation value $\langle V \rangle$ of the potential. Thus, if the external force fluctuates, the stable distribution is the same as (32), with $V_\lambda(x)$ replaced by its expectation value at the point x. Fluctuations of the external potential, even large ones, do not invalidate our analysis.

Suppose the force on the particle is constant within the box, and may be varied in both strength and direction. The particle could, for example, be a charged particle, and the applied field an electric field. Then the external potential varies linearly with x. Take it to be,

$$V_\lambda(x) = \lambda kT\, x/l. \tag{34}$$

where λ is a dimensionless parameter.

The analog of compressing or expanding the one-particle gas is varying the external potential. As λ is increased from zero, the distribution of the particle becomes more and more concentrated towards the left end of the container. We can make the probability that it is to the left of any chosen location as high as we want by taking λ sufficiently large. Similarly, for negative values of λ, the distribution is concentrated towards the right end of the container.

Relative to a canonical distribution with $\lambda = 0$, a distribution for a large value of λ has a large value of free energy, and so we have to do work on the gas while increasing the potential. The work done may be recovered by decreasing the potential back to zero. If the process is done slowly enough that the particle can be treated as canonically distributed at each state of the process, the expectation value of the work recovered while decreasing the potential is equal to the expectation value work of the work done in increasing it.

Let b be a state in which no partition is present and the applied potential is zero. The probability distribution of the particle is evenly distributed throughout the container. Now insert a partition that divides the container into subvolumes with ratio $p : (1-p)$. Let $a_1(p)$ be a state in which the particle is to the left of the partition, and let $a_2(p)$ be a state in which the particle is to the right of the partition.

The states $a_1(p)$ and $a_2(p)$ are perfectly distinguishable states. There's a complication, however: given our class of manipulations, we have no way to prepare them, starting from state b. If we start from b and increase the potential, we can make the probability that the particle is to the left of where we intend to drop the partition as high as we like, but it can never be equal to 1.

In place of these states $a_1(p)$ and $a_2(p)$, which are perfectly distinguishable but not preparable using the manipulations considered, we consider a pair of states that are *almost* distinguishable, and are preparable. Let ϵ be a small positive number, and let $a_1^\epsilon(p)$ be a state in which V_λ is zero, and a partition is present, dividing the container into subvolumes with ratio $p : (1-p)$, and in which there is a prob-

ability of $1-\epsilon$ that the particle is to the left of the partition, and probability ϵ that it is to the left. Define $a_2^\epsilon(p)$ similarly, with the probabilities reversed.

One manipulation that takes $a_1^\epsilon(p)$ to b is removal of the partition, after which the particle equilibrates. This is inefficient, as we could have performed an expansion of the gas, in the course of which work is obtained and heat enters the gas from the reservoir.

To see how much inefficiency, we consider the following process, which is analogous to a controlled expansion of a gas. We start in state $a_1^\epsilon(p)$.

1. We first slowly increase λ to the point at which, on the canonical distribution for V_λ, the particle has probability $1-\epsilon$ of being to the left of the partition, and probability ϵ of being on the right.
2. We remove the partition, allowing the particle to move freely throughout the container. The probability distribution does not change, as the probability, on the equilibrium distribution, of the particle being on the left of the former location of the partition is the same as it was before the partition was removed.[7]
3. The potential is slowly decreased to zero.

The process can be performed in reverse order to create $a_1^\epsilon(p)$ from b. If we have available to us arbitrarily slow processes,

$$S_{\mathcal{M}}(a_1^\epsilon(p) \to b \to a_1^\epsilon(p)) = S_{\mathcal{M}}(a_2^\epsilon(p) \to b \to a_2^\epsilon(p)) = 0. \tag{35}$$

The expectation value of heat gained in the process of expansion is, in the quasistatic approximation,

$$\begin{aligned}\langle Q(a_1^\epsilon(p) \to b)\rangle &= T(S[b] - S[a_1^\epsilon(p)]) \\ &= -kT[(1-\epsilon)\log p + \epsilon \log(1-p) - v(\epsilon)],\end{aligned} \tag{36}$$

where

$$v(\epsilon) = \epsilon \log \epsilon + (1-\epsilon)\log(1-\epsilon). \tag{37}$$

We can make $\langle Q(a_1^\epsilon(p) \to b)\rangle$ as close to $-kT\log p$ as we like by taking ϵ sufficiently small.

[7] General rule: if we take state space Γ and partition the space into disjoint regions Γ_i, a canonical distribution ρ defined on Γ is a mixture of canonical distributions ρ_i confined to the regions Γ_i, with weights being the probabilities, on ρ, that the system is in Γ_i.

Therefore, erasure by removing the partitions has associated with it inefficencies,

$$\begin{aligned}\eta_1 &= -k[(1-\epsilon)\log p + \epsilon \log(1-p) - v(\epsilon)] \approx -k \log p, \\ \eta_2 &= -k[\epsilon \log p + (1-\epsilon)\log(1-p) - v(\epsilon)] \approx -k \log(1-p).\end{aligned} \tag{38}$$

Suppose that we want an erasure process that takes both $a_1^\epsilon(p)$ and $a_2^\epsilon(p)$ to the state b. One such process goes by removal of the partition. This has the inefficiencies exhibited in (38). But we have only availed ourselves of a fairly limited set of operations. Would it be possible to concoct a different set of operations, which might include the employment of auxiliary systems subject to any sort of Hamiltonian we might dream up, whether or not realization of such Hamiltonians is remotely feasible, and thereby construct an operation that takes both $a_1^\epsilon(p)$ and $a_2^\epsilon(p)$ to b, with lower inefficiency for both input states than the lossy removal-of-partition operation? Alas, the answer is negative. As the reader can verify, as long as $\epsilon < p < 1-\epsilon$, the pair of inefficiencies (38) saturate the Landauer bound exhibited in Proposition 2. This means that no process, no matter how elaborate, will achieve a lower inefficiency for both input states, so long as all exchanges of heat are with canonically distributed reservoirs, there are at the beginning of the process no dynamically relevant correlations between the state of A and either the auxiliary systems or the reservoirs, the evolution of the total system is Hamiltonian, and at the end of the evolution the auxiliary systems are restored to their initial states.

6 The LPSG proof vindicated

The LPSG proof proceeds as follows. Suppose we have a manipulation M_L that takes each of a distinguishable set of states $\{a_i, i = 1, \ldots, n\}$ of a device D to a common destination state b. The proof employs as an auxiliary system a one-molecule gas in a box into which partitions may be inserted and removed, and which can be expanded reversibly. LPSG reason that, on pain of violating the statistical second law of thermodynamics, the manipulation M_L must satisfy the Landauer principle. This involves considering the following cycle of operations (performed with both the device D and the gas G in contact with a heat reservoir at all times). The starting state is one in which device D is in state b, and there are no partitions in the box.

1. $n-1$ partitions are inserted into the box, dividing its volume into n subvolumes, with volumes that are fractions p_i of the total volume. With probability p_i, the gas molecule is in the ith subvolume.
2. A controlled operation is performed on D, using the state of the gas G as control. If the gas molecule is in the ith subvolume, b is taken to a_i.
3. A controlled operation is performed on the gas G, using the state of D as control. The ith subvolume is expanded reversibly, obtaining heat $-kT\log p_i$ from the reservoir. The gas has now been restored to its initial state.
4. The operation M_L is performed, restoring the device D to the state b.

If one works through the expectation values of heat exchanges in the course of this cycle, assuming the statistical second law but not assuming reversibility of the processes $b \to a_i$, then what is obtained is precisely our Corollary 2.1 of section 3. Obviously, if one replaces the assumption that heat $-kT\log p_i$ can be obtained in step 3 with the assumption that there are operations such that the expectation value of heat obtained can come arbitrarily close to $-kT\log p_i$, the result still obtains.

The point of contention is whether expansion of a one-molecule gas can be performed in such a way that expectation value of heat obtained is arbitrarily close to $-kT\log p_i$. Norton, in the works cited, contends that this is false. In my opinion Ladyman and Robertson (2014) are right when they say that he has not established this. However, if one has doubts about this being true for a one-molecule gas expanded by a piston, because of lack of control over a sufficiently sensitive piston, our example from the previous section of a one-molecule gas subjected to an external potential may be substituted.

We replace step 3 with the following process. For simplicity we illustrate it for the case of a single partition; extension to multiple partitions is straightforward. Suppose the particle is found to be to the left of the partition. The initial state is $a_1(p)$.

1. Slowly increase λ to a high positive value λ^*.
2. Remove the partition, and allow the system to equilibrate. Some heat is absorbed from the reservoir, but, for large λ^*, this is small.
3. Slowly decrease λ to zero.

If the particle is found to the right of the partition, one takes λ to a large negative value instead. It is not difficult to calculate the expectation value of heat obtained in such a process in the adiabatic limit. The details of this calculation need not concern us; what matters if that it can be made arbitrarily close to $-kT \log p$ by taking λ^* sufficiently large.[8]

7 Conclusion

Landauer's principle is a theorem of statistical mechanics. The worries raised by Norton about assuming reversibility can be addressed; fluctuations pose no threat to the extent we can approximate reversibility, in the relevant sense. If the system being manipulated is in contact with a heat reservoir at temperature T throughout a cycle of operations, the *expectation value* of heat exchanged over the course of the cycle can be made as small as one likes if one is patient enough. On any given run of the cycle, the actual heat exchanged may differ wildly from this expectation value, but it is the expectation value that is relevant to the statistical version of Landauer's principle.

8 Acknowledgements

I am grateful to a number of people with whom I have discussed these matters over the years. In particular, I thank Owen Maroney for drawing my attention to what I have called the Fundamental Theorem, John Norton for discussions of reversible processes, and Katie Robertson for comments on an earlier draft of this article.

[8]For those who are interested, the result is

$$\langle Q \rangle = -kT \log p - kT \log \left(\frac{1 - e^{-2\lambda^*}}{1 - e^{-2p\lambda^*}} \right).$$

For any p, $0 < p < 1$, for large λ^* we have

$$\langle Q \rangle \approx -kT \log p - kT e^{-2p\lambda^*}.$$

Therefore, $\langle Q \rangle$ approaches $-kT \log p$ exponentially with increase of λ^*.

9 Appendix

9.1 Proof of the Fundamental Theorem

To be proven: If $\mathcal{M}$ is a class of manipulations of the sort outlined in Section 2, then, for any states a, b,

$$S_{\mathcal{M}}(a \to b) \leq S[\rho_b] - S[\rho_a].$$

We use the following lemmas.

Lemma 1. *For any Hamiltonian H, and any $T > 0$, the canonical distribution at temperature T minimizes*

$$\langle H \rangle_\rho - TS[\rho].$$

Lemma 2. Subadditivity. *For a composite system AB,*

$$S[\rho_{AB}] \leq S[\rho_A] + S[\rho_B],$$

with equality if and only if the subsystems are probabilistically independent.

Lemma 3. *$S[\rho]$ is conserved under Hamiltonian evolution.*

We consider some manipulation $M \in \mathcal{M}$ that takes a state a of A at t_0 to a state b at t_1. At time t_0 the composite system consisting of A and $\{B_i\}$ has distribution represented by density $\rho_{tot}(t_0)$. At time t_1 the density is $\rho_{tot}(t_1)$. We will write $S_{tot}(t)$ as an abbreviation for $S[\rho_{tot}(t)]$, and similarly for $S_A(t)$ and $S_{B_i}(t)$.

By Lemma 1 we have, for each reservoir B_i,

$$\langle H_{B_i}(t_0) \rangle - T_i S_{B_i}(t_0) \leq \langle H_{B_i}(t_1) \rangle - T_i S_{B_i}(t_1), \tag{39}$$

or,

$$\Delta \langle H_{B_i} \rangle - T_i \Delta S_{B_i} \geq 0. \tag{40}$$

Since $\langle Q_i \rangle = -\Delta \langle H_{B_i} \rangle$, this gives

$$\frac{\langle Q_i \rangle}{T_i} \leq -\Delta S_{B_i}. \tag{41}$$

Because A is uncorrelated with each B_i at t_0,

$$S_{tot}(t_0) = S_A(t_0) + \sum_{i=1}^{n} S_{B_i}(t_0). \tag{42}$$

Because of subadditivity,

$$S_{tot}(t_1) \leq S_A(t_1) + \sum_{i=1}^{n} S_{B_i}(t_1). \tag{43}$$

Because Hamiltonian evolution conserves S,

$$S_{tot}(t_1) = S_{tot}(t_0). \tag{44}$$

Taken together, (42), (43), and (44) yield,

$$\Delta S_A + \sum_{i=1}^{n} \Delta S_{B_i} \geq 0. \tag{45}$$

This, together with (41), gives us the result,

$$\sigma_M(a \to b) = \sum_{i=1}^{n} \frac{\langle Q_i \rangle}{T_i} \leq \Delta S_A. \tag{46}$$

Since this must hold for every manipulation in the set $\mathcal{M}$, it must hold also for $S_{\mathcal{M}}(a \to b)$, which we defined as the least upper bound of the set of all $\sigma_M(a \to b)$ for $M \in \mathcal{M}$. This gives us the desired result,

$$S_{\mathcal{M}}(a \to b) \leq \Delta S_A. \tag{47}$$

9.2 Proof of equivalence of two formulations.

Lemma 4. *Let $\{x_i, i = 1, \ldots, n\}$ be any sequence of n real numbers. The following are equivalent.*

A) For all non-negative $\{p_i, i = 1, \ldots, n\}$ such that $\sum_i p_i = 1$,

$$\sum_{i=1}^{n} p_i \, x_i \geq -\sum_{i=1}^{n} p_i \log p_i.$$

B)

$$\sum_{i=1}^{n} e^{-x_i} \leq 1.$$

To prove this, we use the following.

Lemma 5. *For any positive numbers* $\{p_i\}$, $\{q_i\}$,

$$\sum_{i=1}^{n} p_i \log q_i - \log\left(\sum_i q_i\right) \leq \sum_{i=1}^{n} p_i \log p_i - \log\left(\sum_i p_i\right).$$

To prove this: given $\{p_i\}$, find $\{q_i\}$ that maximizes the LHS; this maximum value is the RHS. Details omitted. We now proceed to the proof of Lemma 4.

Proof that $(A) \Rightarrow (B)$. Suppose that $\{x_i\}$ are such that (A) holds. Take

$$p_i = \frac{e^{-x_i}}{\sum_{j=1}^{n} e^{-x_j}}. \tag{48}$$

Then $\sum_i p_i = 1$, and

$$\sum_{i=1}^{n} p_i\, x_i = -\sum_{i=1}^{n} p_i \log p_i - \log\left(\sum_{j=1}^{n} e^{-x_j}\right). \tag{49}$$

In order for (A) to be satisfied, we must have

$$\log\left(\sum_{j=1}^{n} e^{-x_j}\right) \leq 0, \tag{50}$$

which is equivalent to

$$\sum_{j=1}^{n} e^{-x_j} \leq 1. \tag{51}$$

Proof that $(B) \Rightarrow (A)$. Suppose that $\{x_i\}$ are such that (B) holds. Let $q_i = e^{-x_i}$. Then

$$\sum_{i=1}^{n} p_i x_i = -\sum_{i=1}^{n} p_i \log q_i. \tag{52}$$

By Lemma 5, for any $\{p_i\}$ such that $\sum_i p_i = 1$,

$$-\sum_{i=1}^{n} p_i \log q_i \geq -\sum_{i=1}^{n} p_i \log p_i - \log\left(\sum_{i=1}^{n} q_i\right), \tag{53}$$

and so

$$\sum_{i=1}^{n} p_i x_i \geq -\sum_{i=1}^{n} p_i \log p_i - \log\left(\sum_{i=1}^{n} q_i\right). \tag{54}$$

Because of (B),

$$\log\left(\sum_{i=1}^{n} q_i\right) = \log\left(\sum_{i=1}^{n} e^{-x_i}\right) \leq 0, \tag{55}$$

and so,

$$\sum_{i=1}^{n} p_i x_i \geq -\sum_{i=1}^{n} p_i \log p_i. \tag{56}$$

References

Earman, J. and J. D. Norton (1999). Exorcist XIV: The wrath of Maxwell's Demon. Part II. From Szilard to Landauer and beyond. *Studies in History and Philosophy of Modern Physics 30*, 1–40.

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. New York: Charles Scribner's Sons.

Ladyman, J. (2018). Intension in the physics of computation: Lessons from the debate about Landauer's principle. In M. E. Cuffaro and S. C. Fletcher (Eds.), *Physical Perspectives on Computation, Computational Perspectives in Physics*, pp. 219–239. Cambridge: Cambridge University Press.

Ladyman, J., S. Presnell, and A. J. Short (2008). The use of the information-theoretic entropy in thermodynamics. *Studies in History and Philosophy of Modern Physics 39*, 315–324.

Ladyman, J., S. Presnell, A. J. Short, and B. Groisman (2007). The connection between logical and thermodynamic irreversibility. *Studies in History and Philosophy of Modern Physics 38*, 58–79.

Ladyman, J. and K. Robertson (2013). Landauer defended: Reply to Norton. *Studies in History and Philosophy of Modern Physics 44*, 263–271.

Ladyman, J. and K. Robertson (2014). Going round in circles: Landauer *vs.* Norton on the thermodynamics of computation. *Entropy 16*, 2278–2290.

Leff, H. S. and A. F. Rex (Eds.) (2003). *Maxwell's Demon 2: Entropy, Classical and Quantum Information, Computing*. Bristol and Philadelphia: Institute of Physics Publishing.

Maroney, O. (2007). The physical basis of the Gibbs-von Neumann entropy. `arXiv:quant-ph/0701127v2`.

Maroney, O. J. E. (2009). Generalizing Landauer's principle. *Physical Review E 79*, 031105.

Maxwell, J. C. (1871). *Theory of Heat*. London: Longmans, Green, and Co.

Maxwell, J. C. (1878). Tait's "Thermodynamics", II. *Nature 17*, 278–280.

Norton, J. D. (2005). Eaters of the lotus: Landauer's principle and the return of Maxwell's demon. *Studies in History and Philosophy of Modern Physics 36*, 375–411.

Norton, J. D. (2011). Waiting for Landauer. *Studies in History and Philosophy of Modern Physics 42*, 184–198.

Norton, J. D. (2013a). Author's reply to Landauer defended. *Studies in History and Philosophy of Modern Physics 44*, 272.

Norton, J. D. (2013b). The end of the thermodynamics of computation: A no-go result. *Philosophy of Science 80*, 1182–1192.

Norton, J. D. (2013c). All shook up: Fluctuations, Maxwell's demon and the thermodynamics of computation. *Entropy 15*, 4432–4483.

Norton, J. D. (2016). The impossible process: Thermodynamic reversibility. *Studies in History and Philosophy of Modern Physics 55*, 43–61.

Norton, J. D. (2018). Maxwell's demon does not compute. In M. E. Cuffaro and S. C. Fletcher (Eds.), *Physical Perspectives on Computation, Computational Perspectives in Physics*, pp. 240–256. Cambridge: Cambridge University Press.

Szilard, L. (1925). Über die Ausdehnung der phänomenologischen Thermodynamik auf die Schwankungserscheinungen. *Zeitschrift für Physik 32*, 753–788. English translation in Szilard (1972).

Szilard, L. (1972). On the extension of phenomenological thermodynamics to fluctuation phenomena. In B. T. Feld, G. W. Szilard, and K. R. Winsor (Eds.), *The Collected Works of Leo Szilard: Scientific Papers*, pp. 70–102. Cambridge, MA: The MIT Press.

Tolman, R. C. (1938). *The Principles of Statistical Mechanics*. Oxford: Clarendon Press.