PHYS 414 Problem Set 3:
Demonic refrigerators and eternal sunshine

In a famous thought experiment discussing the second law of thermodynamics, James Clerk Maxwell imagined an intelligent being (a "demon") standing guard at a door in an insulated wall between two large, enclosed volumes (H and C) filled with gases at different temperatures ($T_h > T_c$). The door remains closed, with two exceptions: (i) Whenever the demon observes a particle in C moving toward the door with a speed faster than the average speed of particles in H, he opens the door to allow the particle to pass into H. Remember that the root-mean-squared speeds of the particles in each chamber are $v_{c,\text{rms}} = (3k_BT_c/m)^{1/2}$ and $v_{h,\text{rms}} = (3k_BT_h/m)^{1/2}$ respectively, where m is the mass of a gas particle. The velocities are Maxwell-Boltzmann distributed, so there will be some small fraction of unusually fast particles in C with speeds larger than $v_{h,\text{rms}}$. (ii) Similarly whenever the demon observes a particle in H moving toward the door with a speed slower than $v_{c,\text{rms}}$, he allows it pass through to C. In Maxwell's words, the net result is "the hot system has got hotter and the cold colder and yet no work has been done, only the intelligence of a very observant and neat-fingered being has been employed".

The problem with analyzing such a demonic refrigerator—and verifying that it still obeys the second law—has always been in describing the thermodynamics of the entire system, including the demon itself. There is no external work done on the total demon/gas system. If we treat the two gas volumes as large thermal reservoirs, the net heat transfer rate $\dot{Q} > 0$ from C to H means that the there is an entropy flow $\dot{Q}/T_h$ into H and an entropy flow $-\dot{Q}/T_c$ out of C. The total entropy flow $\dot{Q}/T_h - \dot{Q}/T_c < 0$. This decrease in gas entropy must be compensated for by a larger increase of entropy in the demon, so that the total entropy increases. (We have implicitly assumed that the total system is isolated from the surrounding universe, so entropy must increase.) But where is this increase of demonic entropy manifested? What if the net result is only a "memory" of each door opening event, imprinted in the demon's mind? What is the relationship between recording information and thermodynamic entropy? If the demon's mind is finite, what are the thermodynamics of eventually erasing that information, to make room for more events?

With the advent of nanotechnology, and experimental analogues to Maxwell's demon [1, 2], these issues have become more than merely philosophical puzzles. Perhaps the most elegant way of understanding this problem is through an exactly solvable model published in 2013 by D. Mandal, H.T. Quan, and C. Jarzynski [3], which we will explore below.

Problem 1: The demonic refrigerator

The model (Fig. 1) consists of two thermal reservoirs, at temperatures T_h and T_c, with $T_h > T_c$. The demon is a simple two-state system, with states denoted by u and d having corresponding energies $E_u > E_d$. In addition, there is a tape consisting of a sequence of bits (0 or 1) which slides frictionlessly past the demon. As will become clear, this will play the role of the demon's "memory". The demon can interact with the two heat reservoirs and the bit on the tape which is nearest to it. The tape moves at constant velocity v, and the bits are spaced at intervals of length l, so it has a finite time $\tau = l/v$ during which it can interact with a given bit, before the next

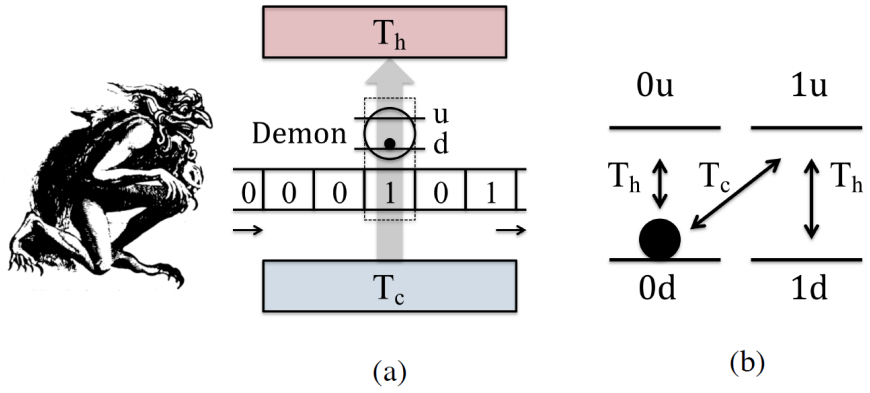

Figure 1: A model for Maxwell's demon, adapted from Ref. [3].

bit comes along. In the simplest version of the model, which is what we will consider here, all the bits on the tape are initially 0 before reaching the demon. The demon can change the state of the nearest bit, according to rules which we will lay out below. Once a bit leaves the demon interaction zone, it is permanently fixed in the state which it attained at the end of the interaction interval.

The demon has two types of transitions, mediated by the two different reservoirs:

i) *Intrinsic* ones that occur regardless of the state of the nearest bit, and leave the bit unchanged. These involve the demon exchanging energy with the hot reservoir. Let us call the intrinsic transition rates x (demon going from d to u) and y (demon going from u to d). The rates satisfy detailed balance with the hot reservoir,

$$\frac{x}{y} = e^{-\beta_h \epsilon} \tag{1}$$

where $\beta_h = (k_B T_h)^{-1}$ and $\epsilon = E_u - E_d > 0$.

ii) *Cooperative* ones that simultaneously change both the states of the demon and the bit. These involve exchanging energy with the cold reservoir. During the interaction interval two such transitions can occur: if the demon is in state d and the bit is 0, they can both flip, yielding states u and 1, with a rate w. Or, conversely, if they are in states u and 1 they can both flip to give d and 0, with a rate z. These transitions satisfy detailed balance with the cold reservoir,

$$\frac{w}{z} = e^{-\beta_c \epsilon} \tag{2}$$

where $\beta_c = (k_B T_c)^{-1}$. We assume that the two states of each bit have the same energy, so $\epsilon = E_u - E_d$ comes just from the demon switching states, as above.

Note that every cooperative transition described by w extracts energy ϵ from the cold reservoir, and every transition described by z deposits energy ϵ into the cold reservoir. Imagine an interaction interval for a given bit which starts at time $t = 0$ and ends at time $t = \tau$. Since the bit is initially in state 0, if it is also 0 at time τ, this means that the number of w transitions was exactly equal to the number of z transitions during that time interval, and the total energy exchanged with the cold reservoir is zero. However is the bit is in state 1 at time τ, the number

of w transitions was one more than the number of z transitions. Hence there is a net energy ϵ extracted from the cold reservoir, and a record of this event has been imprinted permanently in the demon's "memory". Where does this energy go eventually? Well, the demon does not have the capacity to store more than ϵ of energy, but it does exchange energy with the hot reservoir, described by the intrinsic transitions x and y. Since the end result of an interaction with a bit is either extraction of energy from the cold reservoir or no energy taken from the cold reservoir, over the course of many interactions there must be a net flow of energy from the cold to the hot reservoir. Thus the system should behave like a demonic refrigerator. For simplicity, we assume that the reservoirs are arbitrarily large, so this movement of energy does not appreciably change the temperatures T_h and T_c on the time scales of interest. (Though if the demon were allowed to operate indefinitely, T_h would increase and T_c would decrease.)

To make these ideas concrete, we will work out the statistical physics of the system:

a) Initially, let us focus on a single interaction interval between a demon and a certain bit, occurring between times $t = 0$ and τ. Let us call the joint probability of the demon and bit as $p_{ij}(t)$, where $i = u$ or d denotes the state of the demon, and $j = 0$ or 1 denotes the state of the bit. Thus there are four possible states, $ij = u0, d0, u1, d1$. Write down the 4×4 transition matrix W for this system. Note this matrix has elements $W_{ij,i'j'}$. Each off-diagonal element $W_{ij,i'j'}$ where $ij \neq i'j'$ is just the transition rate from state $i'j'$ to ij. The diagonal elements $W_{ij,ij}$ are found by demanding that the columns of $W\delta t$ each sum to 1. This is the same as the W matrix familiar from class, except that the integer state labels n have been replaced by the integer pair labels ij.

b) If the interaction time τ is made very long (longer than the equilibration times of the demon-bit system) the system relaxes to a stationary probability p^s_{ij} by the end of the interval. Using the result of part a, find this probability (make sure you properly normalize it). Also find the marginal stationary probabilities of the demon by itself and the bit by itself, defined as:

$$p_i^{Ds} = \sum_{j=0,1} p^s_{ij}, \quad p_j^{Bs} = \sum_{i=u,d} p^s_{ij}. \tag{3}$$

If you do the calculation correctly, you should find that p^s_{ij} factorizes as: $p^s_{ij} = p_i^{Ds} p_j^{Bs}$. *Recommendation:* you can get cleaner expressions by introducing the constants $\mu \equiv y/x = e^{\beta_h \epsilon}$ and $\alpha \equiv wy/(xz) = e^{(\beta_h - \beta_c)\epsilon}$. Note that $\mu > 1$ and $0 < \alpha < 1$ since $0 < \beta_h < \beta_c$ and $\epsilon > 0$.

From now on we will assume τ is long enough that full relaxation can occur, $p_{ij}(\tau) \approx p^s_{ij} = p_i^{Ds} p_j^{Bs}$. This fully specifies the joint probability at the end of the interaction interval, $t = \tau$. We can also infer the joint probability at the beginning, $t = 0$. Since the demon already achieved the stationary distribution p_i^{Ds} during the interaction interval prior to time $t = 0$, we can assume that at $t = 0$ it starts with distribution $p_i^D(0) = p_i^{Ds}$. At $t = 0$ a new bit appears on the tape, with a state 0 that is independent of the demon, $p_j^B(0) = \delta_{j0}$. Because the demon and bit are uncorrelated at $t = 0$, we can write $p_{ij}(0) = p_i^D(0) p_j^B(0)$, fully specifying the probability at the beginning of the interaction interval. Of course in-between during the relaxation of the system the marginal demon probability $p_i^D(t)$ can deviate from the stationary distribution because of interactions with the bit, but it turns out that we do not need to calculate these deviations: knowing the beginning and end states is sufficient for our purposes. This greatly simplifies the calculation, because solving the full master equation for $p_{ij}(t)$ over time becomes unnecessary.

c) The entropy of the full system $S(t)$, as well as the marginal entropies of the demon ($S^D(t)$) and bit ($S^B(t)$) are defined as:

$$\begin{aligned} S(t) &= -k_B \sum_{ij=u0,d0,u1,d1} p_{ij}(t) \ln p_{ij}(t), \\ S^D(t) &= -k_B \sum_{i=u,d} p_i^D(t) \ln p_i^D(t), \\ S^B(t) &= -k_B \sum_{j=0,1} p_j^B(t) \ln p_j^B(t). \end{aligned} \tag{4}$$

Here the demon and bit marginal probabilities are $p_i^D(t) = \sum_{j=0,1} p_{ij}(t)$ and $p_j^B(t) = \sum_{i=u,d} p_{ij}(t)$. Let $\Delta S = S(\tau) - S(0)$ be the total system entropy change over the interaction interval, and analogously define $\Delta S^D = S^D(\tau) - S^D(0)$ and $\Delta S^B = S^B(\tau) - S^B(0)$. Prove that in our case the entropy changes are additive:

$$\Delta S = \Delta S^D + \Delta S^B. \tag{5}$$

Moreover, show that the demon and bit entropy changes are

$$\begin{aligned} \Delta S^D &\equiv S^D(\tau) - S^D(0) = 0 \\ \Delta S^B &\equiv S^B(\tau) - S^B(0) = k_B \left[\ln(1+\alpha) - \frac{\alpha \ln \alpha}{1+\alpha} \right] \end{aligned} \tag{6}$$

Verify that ΔS^B satisfies that bounds $0 < \Delta S^B < k_B \ln 2$.

We can interpret ΔS^B in terms of information entropy (thermodynamic entropy divided by $k_B \ln 2$, measured in units of bits). Please note the potentially confusing use of bits both to describe the physical objects on the tape, and as a unit of information entropy. If we consider the physical bits on the tape before the demon, each of them has zero information entropy (all the prior bits are in the same state 0). If we look at the physical bits on the tape after the demon, each of them has gained $0 < \Delta S^B/(k_B \ln 2) < 1$ bits of information entropy. This corresponds to the fact that if we had an ensemble of such systems, the pre-demon tape would be identical for each system in the ensemble (all 0's, perfect certainty about the ensemble, zero entropy), while the post-demon tape would be different in each system (some 1's mixed with 0's, less certainty about the ensemble, entropy greater than zero).

d) To complete our description of the system, we need to look at the instantaneous entropy rate

$$\dot{S}(t) = \dot{S}^{\mathrm{i}}(t) + \dot{S}^{\mathrm{e}}(t) \tag{7}$$

during the interaction interval, where the decomposition into entropy production $\dot{S}^{\mathrm{i}} \geq 0$ and entropy flow $\dot{S}^{\mathrm{e}}$ was derived in lecture:

$$\dot{S}^{\mathrm{i}}(t) = \frac{k_B}{2} \sum_{ij,i'j'} J_{ij,i'j'}(t) \ln \frac{W_{ij,i'j'} p_{i'j'}(t)}{W_{i'j',ij} p_{ij}(t)}, \quad \dot{S}^{\mathrm{e}}(t) = -\frac{k_B}{2} \sum_{ij,i'j'} J_{ij,i'j'}(t) \ln \frac{W_{ij,i'j'}}{W_{i'j',ij}}. \tag{8}$$

Here $J_{ij,i'j'}(t)$ is the current from state $i'j'$ to ij, and the sums are double sums over all states (hence the need for a $1/2$ to prevent double-counting). Using the W matrix from part a, show that the entropy flow is given by:

$$\dot{S}^{\mathrm{e}}(t) = \epsilon \left(\frac{J_{u0,d0}(t) + J_{u1,d1}(t)}{T_h} - \frac{J_{d0,u1}(t)}{T_c} \right). \tag{9}$$

e) If we integrate Eq. (7) from $t = 0$ to τ, we get $\Delta S = \Delta S^{\mathrm{i}} + \Delta S^{\mathrm{e}}$. Show that the explicit form of ΔS^{e} is:

$$\Delta S^{\mathrm{e}} = \frac{\alpha\epsilon}{1+\alpha} \left(\frac{1}{T_c} - \frac{1}{T_h} \right). \tag{10}$$

Together with part c, show that this implies:

$$\Delta Q \left(\frac{1}{T_h} - \frac{1}{T_c} \right) + \Delta S^B = \Delta S^i \geq 0 \tag{11}$$

where $\Delta Q \equiv \alpha\epsilon/(1+\alpha)$. The ΔQ expression has a simple interpretation: $\alpha/(1+\alpha) = p_1^{Bs} \approx p_1^B(\tau)$ is the probability that the bit is in state 1 at the end of the interaction interval. From the argument in the introduction, we know that if the bit is in state 1 at $t = \tau$, this indicates a net transfer of energy ϵ from the cold to the hot reservoir; if it is in state 0 at $t = \tau$, no net energy was transferred. Hence $\alpha\epsilon/(1+\alpha) = \Delta Q > 0$ is the average energy moved from the cold to the hot reservoir during one interaction interval. *Hint:* To integrate the currents in the $\dot{S}^{\mathrm{e}}$ expression from part d, use the continuous time master equation which relates the currents to time derivatives of the probabilities dp_{ij}/dt. You can carry out the integrals over dp_{ij}/dt since you know the beginning and ending probabilities at $t = 0$ and τ.

In class we derived an analogous equation for a conventional refrigerator operating in a stationary state between reservoirs T_h and T_c:

$$\dot{Q}_c \left(\frac{1}{T_h} - \frac{1}{T_c} \right) - \frac{\dot{W}}{T_h} = S^{\mathrm{i}} \geq 0 \tag{12}$$

where $\dot{Q}_c > 0$ is the net rate of energy extracted from the cold reservoir, and $-\dot{W} > 0$ is the corresponding amount of external work that needs to be done on the system to make the refrigerator run. In our Maxwell demon case, the role of $-\dot{W}/T_h$ is played by ΔS^B, since there is no external source of work. The high certainty (low information entropy) about the state of the tape entering the demon is effectively like an information reservoir powering the demonic refrigerator, doing the "work" required to move energy from the cold to the hot reservoir. The tape that comes out of the demon is depleted (has greater uncertainty, higher information entropy). Thus the demon literally enacts Sir Francis Bacon's "ipsa scientia potestas est": knowledge itself is power.

Problem 2: Eternal sunshine of the demonic mind

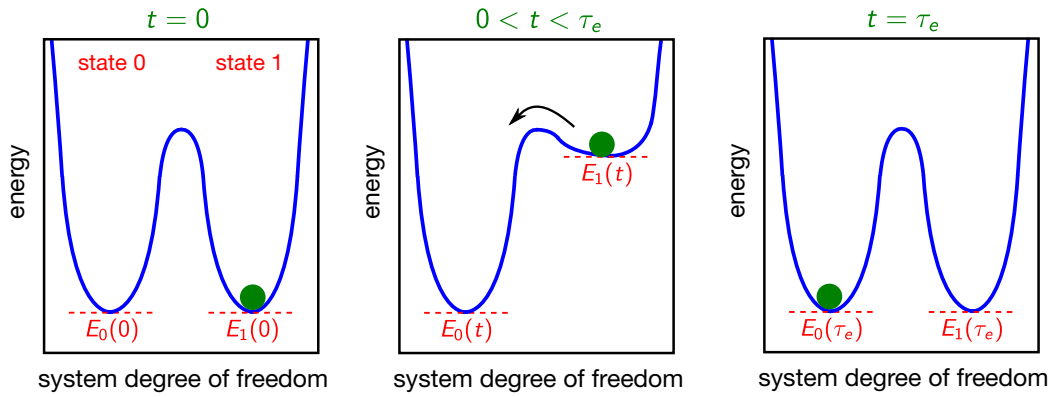

Figure 2: Schematic of erasing a physical bit (returning it to state 0). In this case it starts at $t = 0$ in state 1.

The Maxwell demon from Problem 1 stores entropy in its memory register, but eventually the universe must get its due: there is no infinite memory register, and if you eventually want to loop the tape back into the demon to keep the process going, you need to pass the tape through another device which erases it, restoring all the bits to 0. What are the thermodynamics of erasing a physical bit?

Let us concentrate solely on one bit, which we can imagine has already passed through the demon. Let us call the current time $t = 0$ (the beginning of the erasing procedure). In a hypothetical ensemble of tapes, this physical bit has a probability distribution $p_j(0)$, and corresponding entropy $S(0) = -k_B \sum_j p_j(0) \ln p_j(0)$. For simplicity, we drop the B superscripts to refer to the entropy of the bit, since the only system we are considering here is the bit. $S(0)$ is the entropy stored in the bit by the demon, so its value is equal to ΔS^B from the previous problem. For our purposes, all that really matters is that $0 < S(0) < k_B \ln 2$.

We need a basic physical description of the bit: let us model it as two deep energy wells in the space of some degree of freedom (for example particle spin or position). The wells correspond to states 0 and 1, and the energy barrier between the wells is so large compared to $k_B T$ that spontaneous flipping between the wells is negligible (left panel of Fig. 2). Thus if the system is in a particular state, and we do not disturb it, it should remain there for arbitrarily long times (for a solid state bit in a hard drive, possibly hundreds of years).

To erase the bit (return it to state 0), we can carry out a procedure as follows: the bit is coupled to a single thermal reservoir at temperature T. (Remember that the erasing device is completely different than the demon.) At $t \geq 0$, we perform an erasing protocol on the system, which involves changing the system energy level $E_1(t)$ over time, making it a time-dependent function. There are two possible scenarios: a) the system was in state 1 at $t = 0$, so $p_1(0) = 1$. Eventually, if $E_1(t)$ reaches a level comparable to or higher than the barrier energy, the system will (with extremely high probability) spontaneously switch due to thermal fluctuations into state 0. The high uphill energy slope in the reverse direction prevents it from switching back. In the second

part of our process, we lower $E_1(t)$ back to its original level, which we reach at $t = \tau_e$, the end of the erasure period. Hence $E_1(0) = E_1(\tau_e)$ and $p_0(\tau_e) = 1$. During this process $E_0(t)$ stays constant. b) If the system happened to be in state 0 at $t = 0$, it would do nothing during the same $E_1(t)$ protocol, staying in state 0. The end result is the same: we have a bit in state 0, so $p_0(\tau_e) = 1$.

a) Let us define $\Delta W_{\text{erase}} \equiv -\int_0^{\tau_e} dt\, \dot{W}(t)$ as the net work we do on the system during the erasure process, and $\Delta Q_{\text{erase}} \equiv \int_0^{\tau_e} dt\, \dot{Q}(t)$ as the net heat from the reservoir to the system. Note the minus sign in the definition of ΔW_{erase}, since $\dot{W}$ is by convention the rate at which the system does work on the environment. So $\Delta W_{\text{erase}} > 0$ would correspond to us doing work on the system. Use the first law of thermodynamics, $\dot{Q} = \dot{\bar{E}} + \dot{W}$, to derive a simple relationship between ΔQ_{erase} and ΔW_{erase}.

b) The entropy change of the system during the erasure is $\Delta S = \Delta S^{\text{i}} + \Delta S^{\text{e}}$. Use the fact that $S^{\text{i}} \geq 0$ to prove that $\Delta W_{\text{erase}} \geq TS(0)$ and $\Delta Q_{\text{erase}} \leq -TS(0)$. Thus if we had 1 bit of information entropy, $S(0) = k_B \ln 2$, it would require doing at least $k_B T \ln 2$ of work to erase it, leading to at least $k_B T \ln 2$ of heat dumped into the reservoir (increasing the entropy of the universe).

This fundamental bound on the work required to erase a physical bit was first pointed out by Rolf Landauer in 1961, and since then has been dubbed the *Landauer principle* [4]. By being forced to erase the bit, you contribute to the entropy increase of the universe, so ultimately even our intelligent demon cannot evade the slow creep toward heat death.

References

[1] M. G. Raizen, "Comprehensive Control of Atomic Motion", *Science* **324**, 1403 (2009).

[2] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, "Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality", *Nature Physics* **6**, 988 (2010).

[3] D. Mandal, H. T. Quan, and C. Jarzynski, "Maxwell's Refrigerator: An Exactly Solvable Model", *Phys. Rev. Lett.* **111**, 030602 (2013).

[4] R. Landauer, "Irreversibility and heat generation in the computing process", *IBM J. Res. Devel.* **5**, 183 (1961).