PHYS 414 Problem Set 4:
Fluctuation theorems, Maxwellian demonology

Problem 1

The goal of this problem is to directly verify the integral fluctuation theorem through a computational experiment: simulating an atomic force microscope (AFM) apparatus pulling on a biomolecule. The simulation will be a simplified version of a landmark experiment [Liphardt *et. al.*, Science **296**, 1832 (2002)] from the Carlos Bustamante group at Berkeley. The experiment used optical tweezers (rather than AFM) to pull on single RNA molecules, but the underlying principle is the same. The significance of the 2002 experiment was such that it was cited as part of the justification for awarding half of the 2018 Nobel Prize in Physics to Arthur Ashkin, inventor of optical tweezers.

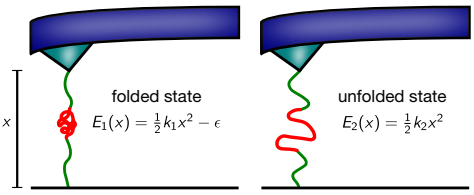

Figure 1: Two possible states of a biomolecular system in an AFM apparatus with a given extension x, and the corresponding state energies $E_n(x)$.

The model consists of a protein molecule (red) in solution attached to the AFM tip and the surface of a stage by polymer handles (green), as shown in Fig. 1. There is nothing special about using biomolecules (like proteins or RNA) in this setup, since the thermodynamic relations hold for any system. However the original experiments used RNA because it was a well characterized molecule with known states. In our case the system (protein + handles) has $N = 2$ states: state 1 is where the protein is folded, and state 2 is where it becomes unfolded.

The entire protein-handle system can be roughly approximated as an elastic spring of total extension x with spring constant k_1 (in the case where the protein is folded) or k_2 (if the protein is unfolded). Since the folded structure is more rigid than the unfolded one, $k_1 > k_2$. There is one additional energy contribution besides the spring potential: folding lowers the protein energy by an amount $\epsilon > 0$, which makes the folded state more favorable at small extensions. The resulting energies of state 1 (folded) and state 2 (unfolded) are:

$$E_1(x) = \frac{1}{2}k_1x^2 - \epsilon, \qquad E_2(x) = \frac{1}{2}k_2x^2. \tag{1}$$

The transition rates between the two states, $W_{12}(x)$ (from 2 to 1) and $W_{21}(x)$ (from 1 to 2), satisfy local detailed balance, $W_{21}(x)/W_{12}(x) = e^{-\beta(E_2(x)-E_1(x))}$, where $\beta = 1/k_BT$. Hence we will write them in the form:

$$W_{12}(x) = \omega e^{\beta E_2(x)}, \qquad W_{21}(x) = \omega e^{\beta E_1(x)}, \tag{2}$$

with a prefactor ω. This form guarantees that they satisfy the local detail balanced relation. Note that both the state energies and rates depend on the parameter x that can be controlled by the experimentalist. In the first part of the problem we will keep x constant, but later change it over time.

The parameters we will use for the simulations are as follows:

$$k_1 = 0.05\, k_B T/\text{nm}^2, \quad k_2 = 0.01\, k_B T/\text{nm}^2, \quad \epsilon = 80\, k_B T, \quad \omega = 5 \times 10^{-10}, \quad \delta t = 1\,\text{s}. \quad (3)$$

Note that all energies will be measured in units of $k_B T$, all extensions in units of nm, spring constants in units of $k_B T/\text{nm}^2$, and the time step δt in units of seconds. If you stay within this unit scheme, you can use the raw numerical values above (like 0.05, 0.01, or 80) in your code without doing any kind of unit conversions.

The experiment is highly sensitive to the value of the extension x. Fig. 2 shows a plot of $E_1(x)$ and $E_2(x)$ given the parameters above, for a range of x between 62.5 and 63.5 nm. At $x = 63.25$ nm the two curves cross. Below that value state 1 has lower energy (and hence the system has a greater likelihood of being folded than unfolded in equilibrium), and above that value state 2 has lower energy (so the unfolded state dominates in equilibrium). If we keep x constant, the system should eventually equilibrate, and we should find in the long run that the probabilities of the two states are described by the Boltzmann distribution, $p_n^s(x) = \exp(-\beta E_n(x))/Z(x)$, for $n = 1, 2$, where $Z(x) = \exp(-\beta E_1(x)) + \exp(-\beta E_2(x))$ is the partition function. Note that since the energies depend on x, the probabilities and partition function also depend on x.

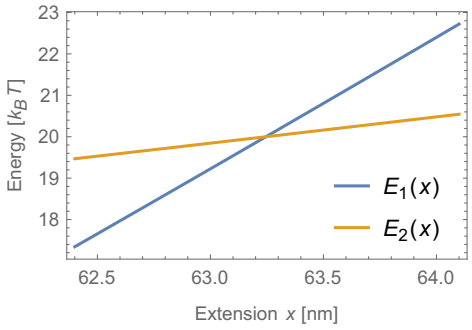

Figure 2: State energies $E_n(x)$ versus extension x.

Part A: Approaching equilibrium

The first step of the computational experiment will be to let the system equilibrate for $\tau = 50$ time steps at a fixed value of extension $x_0 = 62$ nm.

a) Before we do any numerics, let us first determine theoretically what we expect to happen. Since this is a system with fixed external parameters, doing no work, and simply exchanging energy with the environment at temperature T, we have a clear expectation for the behavior as $t \to \infty$. The state probabilities should approach the Boltzmann distribution, and the Helmholtz free energy $F(t) = E(t) - TS(t)$ should approach its equilibrium minimum F^{eq}. Calculate the following quantities in the $t \to \infty$ limit, plugging in the parameters above to get numerical values: i) the state probabilities $p_1(t)$ and $p_2(t)$; ii) the mean energy $E(t)$, in units of $k_B T$; iii) the entropy $S(t)$, in units of k_B; iv) the Helmholtz free energy $F(t)$, in units of $k_B T$.

b) Now let us check numerically if these expectations are fulfilled. We need to develop code that will calculate an ensemble of trajectories for the system. Each trajectory will be of the form $\nu = (n_0, n_1, \ldots, n_\tau)$, where n_i is the state of the system at the ith time step. To get good statistics, we will need a fairly large ensemble, something like 10^4 trajectories or more. The rough idea of how to generate one trajectory is as follows:

1. Our initial distribution $\mathbf{p}(t_0)$ will have states 1 and 2 equally distributed, $p_1(t_0) = 1/2$,

$p_2(t_0) = 1/2$. Hence pick the first state n_0 in the trajectory as either 1 or 2 with equal probability.

2. For each of the τ time steps, carry out the following procedure, which depends on the value of the current state. For example, let us say we are at time step i and the current state is $n_i = 1$. The probability that state $n_{i+1} = 2$ is just $W_{21}(x)$. Hence draw a uniform random real number r between 0 and 1, and if $r < W_{21}(x)$ then set $n_{i+1} = 2$. Otherwise set $n_{i+1} = 1$. The procedure is analogous if $n_i = 2$, except we would use the probability $W_{12}(x)$ to decide whether the next state is 1.

3. After you have completed a full trajectory $\nu = (n_0, \ldots, n_\tau)$, save it as a row in an array, and move on to the next trajectory. Repeat step 1 and 2 above for each trajectory, until you have completed the array. This is your ensemble of trajectories, which you will use for subsequent analysis.

c) Using the ensemble from part b, calculate the state 1 probability $p_1(t_i)$ at each time step i, and plot it. Recall that $p_1(t_i)$ is the fraction of trajectories in the ensemble where $n_i = 1$. Check from the plot that $p_1(t_i)$ approaches the expected limiting value for large times that you calculated from part a.

d) Using the ensemble from part b, and the state probabilities from part c, calculate the mean energy $E(t_i)$, entropy $S(t_i)$ and free energy $F(t_i)$ at each time step, and plot them. Check that they approach the corresponding limits determined in part a.

e) Now let us verify the integral fluctuation theorem for this portion of the experiment. For each trajectory ν in the ensemble from part b, calculate the irreversibility $I(\nu)$. Recall that in the case of no work coupling, this is given by

$$I(\nu) = -k_B \left(\ln p_{n_\tau}(t_\tau) - \ln p_{n_0}(t_0)\right) - \frac{1}{T}\left(E_{n_\tau}(x_0) - E_{n_0}(x_0)\right), \tag{4}$$

and has units of k_B. If you average the irreversibility over all the trajectories, verify that you get the second law relation derived in class

$$\langle I(\nu)\rangle = -\frac{1}{T}(F(t_\tau) - F(t_0)), \tag{5}$$

where you know the free energy difference on the right-hand side from the results of part d. Note that $\langle I(\nu)\rangle$ should be non-negative to satisfy the second law of thermodynamics. Finally calculate $\exp(-I(\nu)/k_B)$ for each trajectory ν, and average them to determine $\langle \exp(-I(\nu)/k_B)\rangle$. Assuming you used a sufficiently large number ($> 10^4$) of trajectories, you should find this number magically equals 1, up to a small statistical error ($\ll 0.01$ discrepancy) due to the fact that the ensemble is finite. If successful, you have just verified the integral fluctuation theorem! If not, you have just violated a fundamental law of nature: probably time to debug the code. Plot a histogram of the $\exp(-I(\nu)/k_B)$ values in the ensemble, and observe that there are just enough "entropy destroying" trajectories (where $I(\nu) < 0$ and and hence $\exp(-I(\nu)/k_B) < 1$) to counterbalance the "entropy producing" trajectories (where $I(\nu) > 0$ and and hence $\exp(-I(\nu)/k_B) > 1$) in order for the whole ensemble to have mean 1.

Part B: Driving the system out of equilibrium

Now let us modify the code from part b to incorporate the full experiment. The basic structure will be the same, but now our trajectories will be longer, with three periods. The first $\tau = 50$ time steps will be exactly as before: letting the system equilibrate at constant $x_0 = 62$ nm. The second period will last $\sigma = 10$ time steps, and during this period will will increase the extension $x(t)$ from x_0 to $x_f = 63.5$ nm at a constant rate. In other words, $x(t_{\tau+j}) = x_0 + j(x_f - x_0)/\sigma$, for $j = 1, \ldots, \sigma$. In this middle period we drive the system out of equilibrium. Remember to change your transition probabilities $W_{12}(x(t_i))$ and $W_{21}(x(t_i))$ as the extension changes. During the final period, which lasts $\tau = 50$ time steps, the extension is kept constant at the value x_f, allowing the system to reach equilibrium again, but now with a different external parameter. Thus each trajectory ν will consist of a sequence of $2\tau + \sigma$ states, $\nu = (n_0, n_1, \ldots, n_{2\tau+\sigma})$.

In addition to calculating each trajectory ν, we also want the code to keep track of the total work $W(\nu)$ associated with each trajectory. Experimentally this is a quantity that can be measured directly, since the AFM measures the force applied to the molecule at each extension. Hence for every run, we get $W(\nu)$ by integrating the force vs. extension (distance) curve. For our simulation, we can use the expression for $W(\nu)$ derived in class. There is no explicit work coupling ω_{nm} in our case, but we do have a contribution to work during the middle period when we vary the extension. If the system is in state n_i at time t_i, and we have changed the extension from $x(t_{i-1})$ to $x(t_i)$, then we do work $E_{n_i}(x(t_i)) - E_{n_i}(x(t_{i-1}))$ on the system. The work done *by* the system is just the negative of this, and the sum of each such contribution over the whole trajectory is just:

$$W(\nu) = -\sum_{i=\tau+1}^{\tau+\sigma} \left[E_{n_i}(x(t_i)) - E_{n_i}(x(t_{i-1}))\right]. \tag{6}$$

Note that the sum only includes the time steps during the middle period of the trajectory, where we actually have work done on the system. For the purposes of the simulation, it is easiest to keep track of each contribution to the sum as we generate the trajectory state by state, and then save the total $W(\nu)$ as an element in an array.

f) As derived in class, one of the consequences of the integral fluctuation theorem (for the special case where you equilibrate, drive, and then equilibrate again) is the Jarzynski equality. This states that

$$\langle e^{\beta W(\nu)} \rangle = e^{-\beta \Delta F^{\text{eq}}}, \tag{7}$$

where $\beta = (k_B T)^{-1}$. Verify the Jarzynski equality (up to small statistical error) by numerically calculating both sides of the above equation using the modified code described above. Note the equilibrium free energy difference ΔF^{eq} on the right-hand side is the difference between the final and initial equilibrium free energies. The final one is just $F(t_{2\tau+\sigma})$, after the system has equilibrated at x_f, and the initial one is $F(t_\tau)$, after the system has equilibrated at x_0. Hence $\Delta F^{\text{eq}} \approx F(t_{2\tau+\sigma}) - F(t_\tau)$. You can calculate these free energies similarly to the way you did it in part d. Alternatively you can calculate ΔF^{eq} using the analytical expressions from part a.

g) Redo part f, but using a different σ, for example $\sigma = 20$. This corresponds to changing the extension at a different pulling velocity. Check that both sides of the Jarzynski equality stay the same as in part f. It is valid no matter how fast or slow we pull. In fact the Jarzynski equality

allows us to pull arbitrarily fast, and use the experimentally measured work distribution to find out information about the equilibrium properties of the system (the equilibrium free energies). The 2002 Liphardt *et. al.* paper was the first experimental verification of this fundamental relation.

Introduction to Problems 2 and 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 ($T_h > T_c$). Because of the temperature difference, the mean speeds of particles in H are faster than those in C. This is because in a chamber at temperature T, the velocities v of the individual particles are distributed according to the Boltzmann distribution $p(v) \propto \exp(-E(v)/(k_B T))$, where $E(v) = (1/2)mv^2$ and m is the mass of the gas particle. Hence higher T favors the chances of finding larger v. 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. Such fast particles in C may be rare, but from the Boltzmann distribution we know that all velocities are possible, just with different probabilities. (ii) Similarly whenever the demon observes a particle in H moving toward the door with a speed slower than than average speed in C, he allows it pass through to C. We assume the door is essentially massless and can be slid frictionlessly, so that opening and closing the door requires no work on the part of the demon. Over time, the particles in C get slower on average, and those in H get faster, which is equivalent to cooling C and heating up H. 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".

In other words we have a demonic refrigerator, moving heat from a cold to a hot reservoir, without seeming to expend any work. The challenge in describing this scenario thermodynamically is to consider all the components, including the demon itself. In the terminology of our course, the demon is a system coupled to two heat baths (or reservoirs) at temperatures T_h and T_c. Over some interval of interest τ, the demon does no work, so $W = 0$. Similarly the demon does not change its internal energy during the process, so $\Delta E = 0$. From lecture, the first and second laws in the case of two temperature reservoirs look like:

$$\begin{aligned} Q_c + Q_h &= \Delta E + W \\ I &= \Delta S - \frac{Q_c}{T_c} - \frac{Q_h}{T_h} \geq 0 \end{aligned} \tag{8}$$

We assume that the time interval τ is short enough (and the reservoirs big enough) that the temperatures T_h and T_c have not changed substantially during the process. Of course we know in the very long run they will gradually shift. I is the mean irreversibility, which we know must always satisfy $I \geq 0$. ΔS is the change in entropy, and Q_i is the heat extracted from the ith reservoir. Since $\Delta E = W = 0$, we know that $Q_c = -Q_h$ from the first law. Since heat is leaving

the cold reservoir, $Q_c > 0$. Thus

$$-\frac{Q_c}{T_c} - \frac{Q_h}{T_h} = -Q_c\left(\frac{1}{T_c} - \frac{1}{T_h}\right) < 0. \tag{9}$$

In order to satisfy the second line in Eq. (8), we thus must have $\Delta S > 0$. The system (demon) entropy must increase. But where is this increase of demonic entropy manifested? One idea is that the entropy increase occurs in the demon's brain, through a recording of a "memory" of each door opening event. 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 in the next two problems.

Problem 2: The demonic refrigerator

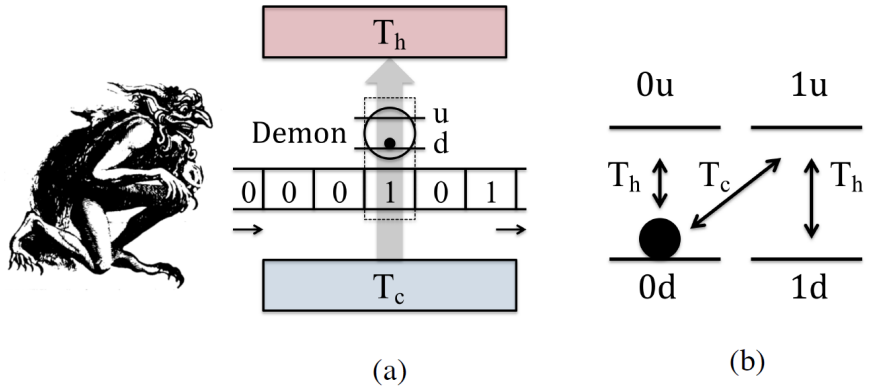

Figure 3: A model for Maxwell's demon, adapted from Ref. [3].

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 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 probabilities x (demon going from d to u) and y (demon going from u to d). These are probabilities of the transition occurring over some microscopic time interval $\delta t \ll \tau$. The probabilities satisfy detailed balance with the hot reservoir,

$$\frac{x}{y} = e^{-\beta_h \epsilon} \tag{10}$$

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 probability w per time step δt. Or, conversely, if they are in states u and 1 they can both flip to give d and 0, with a probability z per time step δt. These probabilities satisfy detailed balance with the cold reservoir,

$$\frac{w}{z} = e^{-\beta_c \epsilon} \tag{11}$$

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 if 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 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:

$$p^{Ds}_i = \sum_{j=0,1} p^s_{ij}, \quad p^{Bs}_j = \sum_{i=u,d} p^s_{ij}. \tag{12}$$

If you do the calculation correctly, you should find that p^s_{ij} factorizes as: $p^s_{ij} = p^{Ds}_i p^{Bs}_j$. *Recommendation:* You know the stationary probability is the right eigenvector of the 4×4 matrix W from part a) with eigenvalue 1. From the elements of this four component eigenvector you can then read off the stationary probabilities, $\mathbf{p}^s = (p^s_{u0}, p^s_{d0}, p^s_{u1}, p^s_{d1})$. 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, $p_{ij}(\tau) \approx p^s_{ij} = p^{Ds}_i p^{Bs}_j$. 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^{Ds}_i during the interaction interval prior to time $t = 0$, we can assume that at $t = 0$ it starts with distribution $p^D_i(0) = p^{Ds}_i$. At $t = 0$ a new bit appears on the tape, with a state 0 that is independent of the demon, $p^B_j(0) = \delta_{j0}$. Because the demon and bit are uncorrelated at $t = 0$, we can write $p_{ij}(0) = p^D_i(0) p^B_j(0)$, 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 $p^D_i(t)$ 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:

$$\begin{aligned} S(t) &= -k_B \sum_{ij=u0,d0,u1,d1} p_{ij}(t) \ln p_{ij}(t), \\ S^D(t) &= -k_B \sum_{i=u,d} p^D_i(t) \ln p^D_i(t), \\ S^B(t) &= -k_B \sum_{j=0,1} p^B_j(t) \ln p^B_j(t). \end{aligned} \tag{13}$$

Here the demon and bit marginal probabilities are $p^D_i(t) = \sum_{j=0,1} p_{ij}(t)$ and $p^B_j(t) = \sum_{i=u,d} p_{ij}(t)$. 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:

$$\Delta S = \Delta S^D + \Delta S^B. \tag{14}$$

Moreover, show that the demon and bit entropy changes are

$$\begin{aligned} \Delta S^D &\equiv S^D(\tau) - S^D(0) = 0 \\ \Delta S^B &\equiv S^B(\tau) - S^B(0) = k_B \left[\ln(1+\alpha) - \frac{\alpha \ln \alpha}{1+\alpha} \right] \end{aligned} \tag{15}$$

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, a concept introduced by Claude Shannon: this is simply defined as thermodynamic entropy divided by $k_B \ln 2$, and measured in units of bits. Thus $k_B \ln 2$ of thermodynamic entropy is equivalent to 1 bit of information entropy. 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) The energy of the demon at time t just depends on the demon state $i = u, d$ rather than the state of the tape, so the mean system energy is

$$E(t) = \sum_{i=u,d} p_i^D(t) E_i. \tag{16}$$

Argue that $\Delta E = E(\tau) - E(0) = 0$.

e) Plug in the results of parts c) and d) into the first and second laws in Eq. (8), keeping in mind $W = 0$. Show that

$$I = \Delta S^B - Q_c \left(\frac{1}{T_c} - \frac{1}{T_h} \right). \tag{17}$$

f) On the right-hand side of Eq. (17), we know ΔS^B from part c), but what is Q_c? Find an expression for Q_c in terms of ϵ and α. *Hint:* we know Q_c is the average heat extracted from the cold reservoir during the interval τ. From the argument in the introduction, we know that if the bit is in state 1 at $t = \tau$, this indicates a net energy ϵ was extracted from the cold reservoir; if it is in state 0 at $t = \tau$, no net energy was extracted. Since we know that probabilities of the bit being in each state at $t = \tau$ from part b), we can calculate the mean amount of energy Q_c extracted.

g) Using the result of part f) and the expression for ΔS^B from part c), write down an explicit expression for I in Eq. (17). After simplifying, you should find $I = k_B \ln(1+\alpha)$. This satisfies $I > 0$, since $0 < \alpha < 1$. *Hint:* use the definition of α to write $1/T_c - 1/T_h$ in terms of α, ϵ, and k_B.

Postscript: It is instructive to compare Eq. (17) to the analogous result we would get from an ordinary refrigerator. For a conventional refrigeration cycle we return the system to its original

state (there is no "memory" tape) and hence $\Delta S = 0$. On the other hand $W < 0$, since we do work on the system (plug the refrigerator into an outlet). Hence Eq. (8) can be rewritten as

$$I = -\frac{W}{T_h} - Q_c \left(\frac{1}{T_c} - \frac{1}{T_h} \right) \geq 0. \tag{18}$$

In our Maxwell demon case, the role of $-W/T_h$ 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 3: Eternal sunshine of the demonic mind

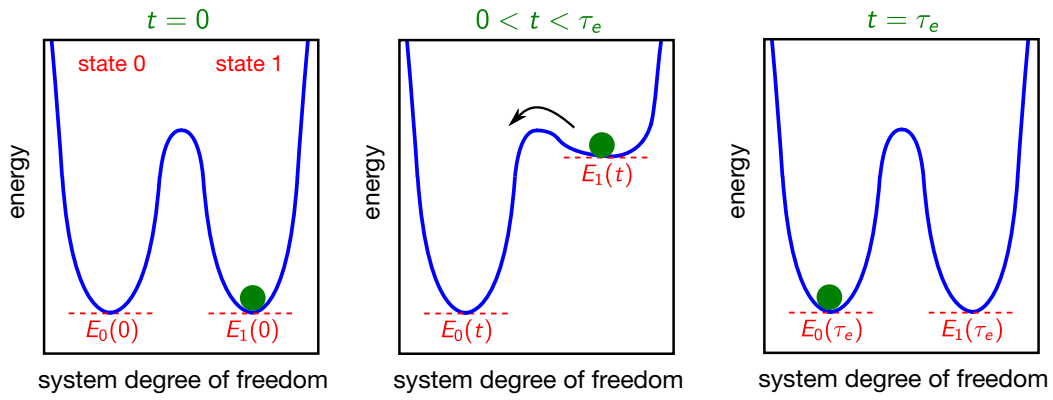

Figure 4: Schematic of erasing a physical bit (returning it to state 0). In this case it starts at $t = 0$ in state 1.

The Maxwell demon from Problem 2 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) What are $\Delta E = E(\tau_e) - E(0)$ and $\Delta S = S(\tau_e) - S(0)$ for the erasure protocol? You can leave the expression for ΔS in terms of $S(0)$.

b) Using the first and second laws for a system connected to a single temperature reservoir T, prove that $Q = W \leq -TS(0)$ for the erasure protocol, where Q is the heat extracted from the environment, and W is the work done by the system. Since $S(0) > 0$, this means heat must be dumped into the environment during erasure ($Q < 0$), and we have to do work on the system to erase the bit ($W < 0$).

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).