PRINCETON UNIVERSITY PHYSICS COMPETITION

2016 Online Competition

Guidelines

Student teams will have a total of **one week** to complete the exam from start to finish. There are two separate sections for this examination: one section on Laser and Plasma Physics, and one section on Entropy and Statistical Mechanics. Teams may complete both sections or choose to complete only one, as is specified in the grading explanation below. We recommend that teams set aside approximately 20+ hours to allow enough time for successful completion. Please refer to the submission explanation below for details on both formatting and the submission process.

Grading

Students are encouraged to work on as much of both sections of the exam as possible. However, teams **may choose to submit solutions for only one of the two sections** if they desire. The two sections will be graded separately, and may not necessarily be worth the same amount of points. The award structure will be as follows:

1. Awards will be given to the four teams with the **highest score in each section** (an award for first place, second place, third place, and fourth place). One team can win an award for both sections, such as second place in Laser and Plasma Physics and fourth place in Entropy and Statistical Mechanics. Therefore teams are encouraged to attempt solutions for both sections of the competition.

2. We will additionally award one **overall award to the highest scoring team** on the entire competition. A team which wins this overall award can still receive one of the top four awards for each individual section. The team that wins this overall award will most likely have completed both sections of the exam. It will be at the judges' discretion to choose the overall award for the best submission.

3. **Special awards** will also be given for honorable mentions, the most elegant solution, and the most creative solution.

Collaboration Policy and Resources

Students participating in the competition may only correspond with other members of their team. No other human correspondence is allowed, including: mentors, teachers, professors, and other students. In general, participating students are barred from posting content or asking questions related to the exam on the internet (except where specified below). Students are, however, allowed to use the following resources:

- **Online**: Teams may use any information they find useful on the Internet. However, under no circumstances may they actively post content or ask questions about the exam.

- **Piazza page**: Teams are encouraged to create an account on Piazza and register in the class at the following URL: http://piazza.com/princeton_university_physics_competition/fall2016/pupc2016. The access code is: **pupc2016**. This resource can be used by teams to ask questions about the content of the exam. Please do not post any of your solutions, partial or complete, when asking questions on Piazza.

- **Published Materials**: Teams may take advantage of any published material, both printed or online.
- **Computational**: Teams may use any computational resources they might find helpful, such as Wolfram Alpha/Mathematica, Matlab, Excel, or lower level programming languages (C++, Java, Python, etc). For some sections, the use of computational resources is highly advised.

Citations

All student submissions with outside material must include numbered citations. We do not prefer any style of citation in particular. Students may find the following guide useful in learning when to cite sourced material: http://www.princeton.edu/pr/pub/integrity/pages/cite/.

Submission

All submissions, regardless of formatting, should include a **cover page listing the title of their work, the date, and signatures of all team participants.** The work must be submitted as **one single PDF document with the ".pdf." extension**. All other formatting decisions are delegated to the teams themselves. No one style is favored over another. That being said, we recommend that teams use a typesetting language (e.g., LaTeX) or a word-processing program (e.g. Microsoft Word, Pages). Handwritten solutions are allowed. **Note: we reserve the right to refuse grading of any portion of a team's submission in the case that the writing or solution is illegible.**

Teams must submit their Online Part solutions by e-mailing **pupc@princeton.edu** by 11:59 am (noon) Eastern Time (UTC-5) on Saturday, November 19, 2016. Teams will not be able to submit their solutions to the Online Part at any later time. Any team member may send the submission. The title of the submission e-mail should be formatted as "SUBMISSION - Team Name". Note: all teams may make multiple submissions; however, we will only grade the most recent submission submitted before the deadline. **Teams will receive confirmation once their submission has been received within at most two days.** In the case of extraordinary circumstances, please contact us as soon as possible.

Sponsors

Department of Astrophysical Sciences

3Red

Department of Physics

Jane Street

SIG SUSQUEHANNA

D E Shaw & Co

Collaborators

expii

APS physics

Entropy and Statistical Mechanics Section

The subject matter of this document is statistical mechanics, or the study of how macroscopic results manifest from microscopic interactions in systems with many interacting parts.

Our goal is to provide a unified introduction to statistical mechanics using the concept of entropy. A precise definition of entropy pervades statistical mechanics and other scientific subjects and is useful in its own right. While many students may have heard the word entropy before, entropy is rarely explained in its full detail or with rigorous mathematics, leaving students confused about many of its implications. Moreover, when students learn about thermodynamic laws, laws that describe the macroscopic results of statistical mechanics, like "change in internal energy = heat flow in + work done on a system", the concepts of internal energy, heat, and work are all left at the mercy of a student's vague, intuitive understanding.

In this document, we will see that formulating statistical mechanics with a focus on entropy can provide a more unified and symmetric understanding of many of the laws of thermodynamics. Indeed some laws of thermodynamics which appear confusing and potentially unrelated at first glance can in fact all be seen to follow from the same treatment of entropy in statistical mechanics.

Contents

Learning goals of this topic: This topic is meant as an exercise in learning more so than an exercise in solving problems or external research. The goal of this topic is for the reader, who has not necessarily seen any statistical mechanics in their education so far, to walk away with a set of examples and ideas that will be helpful long into the future.

Topic format: This document consists of long sections of explanatory material with helpful exercises and questions interspersed. Much of the grading will be based on sections that ask you to explain or interpret results in your own words. We are looking to see how well you understand the subject, and are not overly concerned with minor errors in completing exercises.

Expected amount of work: Do not expect to understand the concepts in this document after only one read through. These topics take time to absorb. While it may feel like you are not getting much accomplished as you try to understand the reading, we expect that it may be necessary to read some passages four times in a row before understanding it completely. Because there are not too many questions in this document, you should have time to complete the readings.

Many pages of this document have only one or two places where the author asks for input and response from the readers. Some sections contain no questions. We would encourage you *not* to skip reading these sections completely, as all sections of this document will be beneficial to understand.

Additional reading materials and resources

While this document is meant to generally stand alone as an explanation of entropy, free energy, and related concepts, you may find a few outside sets of information useful:

- Required reading:
  - **Equilibrium Information from Nonequilibrium Measurements in an Experimental Test of Jarzynski's Equality**, J. Liphardt *et al.*, *Science* **296**, 5574 (2002): this paper experimentally verifies the results of the above paper by C. Jarzynski. It is explored in detail in Section 4.1.2.
- Additional potential resources:
  - **Thermal Physics** by Charles Kittel and Herbert Kroemer (1980, W. H. Freeman and Company): an extensive introduction to concepts discussed within this document. While this reference may be useful for clarifications of certain concepts, it is not at all essential.
  - **Nonequilibrium Equality for Free Energy Differences**, C. Jarzynski, *Phys. Rev. Lett.* **78**, 2690 (1997): this reference is discussed in Section 4.1 of this document as an application of the theoretical results derived herein.
  - **Feynman Lectures on Computation** by Richard Feynman, edited by Tony Hey and Robin W. Allen (1996, Perseus Publishing): this reference is an interesting book which contains many useful and intuitive explanations of statistical applications in computing. Of particular interest for this document is the chapter on reversible computation. This reading is relevant to Section 4.2, all though it is by no means necessary to complete the questions in that section.

1 Entropy as information

At this point in your life, you may have heard the word entropy, but chances are, it was given a vague, non-committal definition. The goal of this section is to introduce a more explicit concept of entropy from an abstract standpoint before considering its experimental and observational signatures.

1.1 Quantifying the amount of information in the answer to a question

Entropy, while useful in physics, also has applications in computer science and information theory. This section will explore the concept of information entropy as an abstract object.[1]

We will first consider entropy not as related to the concept of heat in objects, but as a purely axiomatic quantification of what we mean by the information we receive when we hear the answer to a question. For a concrete example, imagine that someone flips a coin and doesn't reveal which side landed upright and we ask "what was the outcome of the coin flip? Heads or tails?" When they now tell us the answer, how much "information" do we gain by learning what the outcome was? In other words, we are faced with determining how much information is received when we hear the answer to a question that has a probability distribution of outcomes. This question was asked by Claude Shannon in 1948.

After pondering this question for a long time, you might come up with a few criteria that any reasonable measure must obey, such as:

1. If the question has two answers that are not dependent on each other, then the measure of information contained in answering both questions should be the same as the sum of the information gained in learning the answer to each one individually.

   For example, if we have two independent coin flip experiments, the information gained in hearing the outcome of one coin flip should be the same as the information gained in hearing the outcome of the other, so that the total information gained is the sum of the individual amounts of information gained.

   How does this concept generalize to questions which have interdependent answers? Two such questions might be "am I wearing gloves?" and "am I wearing a sweater?". The exact statement of this criteria is more complicated and we will not ask you to consider it here.

2. The measure of the information learned should not depend on the language used; it should depend only on the probability distribution of possible answers.

   For example, it should not depend on the fact that we are flipping a coin and asking whether it is heads or tails instead of flipping a dinner plate and asking if it lands upright or not. The measure of the information gained should be the same. For this reason, the measure should depend on the probability distribution of answers to the question, not the content of the answers.

3. The information gained should be largest when all possible answers are equally likely.

   If we have no idea which answer it will be, we will gain the most information possible when we hear an answer. For example, if a fair dice with 6 sides is rolled, it has an equal chance of landing on each side, and you gain the most information possible by learning which side the dice landed on. On the other hand, if the dice was unfair and landed on a specific side every time, then you learn no new information by learning the outcome of the dice role; you already knew what it would be before you rolled the dice.

While there are other possible additional criteria to use when defining the information gained by hearing the answer to a question, these are sufficient for our purposes in this document.

Now, for a definition: given a probability distribution of possible outcomes (such as answers to a question) with associated probabilities p_i for each outcome, the logical quantification of the average information gained if one outcome is obtained (if we learn the answer to a question) is the **Information Entropy** I,

$$\boxed{I = \sum_{i=1}^{N} p_i \log(1/p_i) = -\sum_{i=1}^{N} p_i \log(p_i).} \tag{1}$$

[1] The author would like to acknowledge that some of the writing in this section is based on information from a class in statistical mechanics taught by Prof. William Bialek at Princeton University.

Questions Explore this definition of information entropy for a few example probability distributions over a finite set of outcomes (such as 2 or 4) and summarize your findings. Note that $\log(x)$ refers to the natural logarithm of x (base e, not base 10 or base 2), and will refer to the natural logarithm throughout this document. Is this quantity always positive (despite the minus sign in the definition)? For a fixed number of possible outcomes, N, but varying probabilities p_i, what is the range of possible values that I can take (the maximum minus the minimum)? What is the minimum amount of information you can learn from the answer to a question, in words, and why? A simple example to begin with is the probability distribution of a coin toss, where $p_{\text{heads}} = 1/2$ and $p_{\text{tails}} = 1/2$. What happens if you measure the logarithm with a different base (such as base 10 instead of the natural logarithm)? This is related to the concept of measuring information with different scales, just as we can measure a mass as 1 kilogram or as 1000 grams. Now, go back and consider the definition of information entropy with respect to the criteria listed above. Does this definition of entropy satisfy criteria 1, 2, and 3? (Ignore the case of interdependent answers in criterion 1.)

2 Where does entropy show up in Statistical Mechanics?

We will now consider how entropy relates to physics. This section contains no questions for the reader.

Consider a glass of water. At any given moment, if you were to look at a snapshot of all the atoms of water in the glass, there would be an unimaginably large number of arrangements these atoms could be in. Statistical mechanics on some level requires us to admit that we are not all knowing beings, and cannot always tell if the system is in one configuration or another. In effect, we assume that we could never know the location of all of these molecules at one time, and as such we can only ask other types of questions. This is where entropy shows up. Now, we ask "what is the configuration of atoms in the glass of water?" The set of possible outcomes is the set of all possible positions of all the atoms. This is a large set, and we would not hope to actually be able to calculate the information entropy of the system by finding each p_i and computing the information entropy I. On the other hand, the abstract language of mathematics lets us engage with this problem regardless of our gaps in knowledge. Even if we do not know all of the p_i, they still *exist*, and so the entropy of this question *still exists* and is not meaningless to talk about.

Let us consider a simpler picture. Various symbols will be introduced to define quantities. Do not be intimidated, but rather read slowly to absorb the information. We define Ω (capital omega) as the set of all possible outcomes ω_i (lowercase omega) for the configuration of atoms in the glass, indexed by a number i. That is, the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_9, \omega_{10}\}$ would describe a situation with 10 possible outcomes, and so with 10 elements.[2] These 10 outcomes are the possible configurations of atoms in the glass. It may be hard to imagine some set of atoms having only a finite number of possible configurations, rather than an infinite number, but for now assume that it is true. If we as physicists have not discovered energy or any other variable that might make one outcome more likely than any other (people usually say lower energy states are more likely, and we will see why), we could expect that each outcome ω_i is equally likely. That is, $p_i = 1/10$ for all i. This seems rather simple, and it is reasonable to question why it would be true that each $p_i = 1/10$. For now, assume it is true. We could immediately ask what is the information entropy in the question "what is the state of the water glass?" In this case, $I = \log(10)$.

However, we are not always interested in what the state of the system, is, but rather some other property, such as how high the water in the glass is. If each outcome ω_i has a different value for the height of the water, the situation is more complicated, and the question "what is the height of the water in the glass" would have a different information entropy than the question "what is the state of the system?".

As a concrete example, let us say that 4 of the ω_i have a height 5 cm, 3 have a height 3 cm and 3 have a height 2 cm. Now, the question, "what is the height of the water'?' has a less well defined answer. We can compute the expected height we would see on average (given the probability of each ω_i). This is $h = \sum_i p_i h_i = (4 \times 5 + 3 \times 3 + 3 \times 2)/10$ cm $= 3.5$ cm. We can also compute the information entropy of the question. The information entropy, I, is given by

$$I = \frac{2}{5}\log\left(\frac{5}{2}\right) + \frac{3}{10}\log\left(\frac{10}{3}\right) + \frac{1}{5}\log\left(\frac{5}{1}\right) \approx 1.0495.$$

Note that this information entropy is different from the value $\log(10) \approx 2.3026$ for the entropy of the state of the system. Therefore, you must be careful about what questions you are asking and answering when computing information entropy.[3] Both the average height and the information entropy tell us something different about the answer to the question "what is the height of the water?"

Much of statistical mechanics follows this same vein of reasoning: we assume some set of states of the system are equally likely, but then we ask something about a property that varies across the states (like energy, or the height of water). What ends up mattering for this secondary question is how many states there are with a certain property. Ultimately, entropy is intimately related to how we understand and think about answers to these questions.

[2] In general, Ω could have N elements, where N could be as large as $10^{23}!$, a very large number.

[3] It is also useful to note that the information entropy is not telling us something about how precisely we know the height of the water. The information entropy I would have the same value if the three possible heights were 1 cm, 1.0001 cm, and 58476 cm, for example.

2.1 An example: rigid chain and the "force" that entropy causes

We will now consider a more concrete example system than a glass of water molecules and how entropy informs our understanding of this system.[4]

2.1.1 A description of the physical model

Consider a chain that is made up of N rigid straight sections connected to each other at bendable hinges. Assume that one end of the chain is fixed at position $x = 0$, while the other is free to move around.

We will make one other crucial assumption which may not make intuitive sense. For the purpose of this example, assume that the states of the chain that we observe are *always* with the entire chain confined to one dimension, the x axis. This means that the bends at each hinge are always $0°$ or $180°$ (0 or π radians), and the chain can double back on itself any number of times. We will assume that there is *some way* for the system to go from a bend being $0°$ to $180°$, but we do not care what it is at this point. Examples of the four possible configurations of a rigid chain with two segments are shown in Fig. 1.

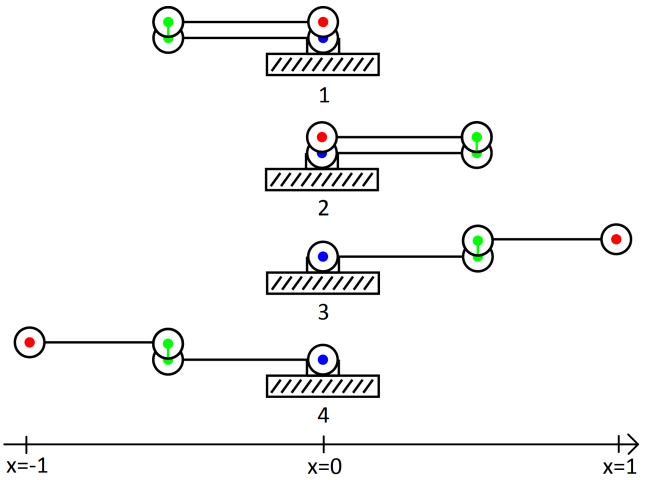

Figure 1: Diagram of the four possible configurations of a rigid chain with two segments. The chain is confined to lie in one dimension and the end of the chain designated by the blue dot is held fixed at $x = 0$.

We will further assume that each segment of the chain has length $1/2$ (in some units). We will also assume that N is an even number so that the location of the free end of the chain is at a position x that is an integer (because each segment of the chain has length $1/2$). This choice is not important, but will make calculations easier. For this reason, a chain with two segments can have its end position lie at $x = -1$, $x = 0$, or $x = 1$, as is shown in Fig. 1.

2.1.2 Finding the probability of certain positions

Like before, we assume that *all configurations of the chain are equally likely*, meaning the chain has no preference for whether a hinge is straight or bent at $180°$.

Questions Now, what is the probability distribution of the location of the free end of the chain if all configurations of the chain are equally likely? There are multiple ways of finding this answer, but for now, try to follow this "chain" of reasoning:

1. List all possible locations of the free end of the chain.
2. List the number of possible configurations of a single link in the chain. Now, what is the number of possible configurations for the entire chain all at once? (We are asking how many outcomes for the

[4]The author would like to acknowledge that this example is based on information from a class in statistical mechanics taught by Prof. William Bialek at Princeton University.

chain configuration are possible, not how many locations. This is the analogous question to the above example of a cup filled with water. We are not asking how many possible heights of the water in the glass there are, but rather how many possible states there are.).

3. For a given $-N/2 \leq n \leq N/2$, how many ways can the free end lie at position $x = n$? (Hint: consider the case of a small chain first and see if you can work your way up to a general expression involving factorials, where m factorial is $m! = (m)(m-1)(m-2)\cdots(2)(1)$.)

4. If all of these individual configurations of the chain are equally likely (meaning any given configuration of hinges being bent or not bent), then what is the the probability $p(x = n)$ that the chain position x is given by a specific value of n for $|n| \leq N/2$?

   You should now have some grasp of what a likely position of the end of the chain is.

Note that the most likely locations of the end of the chain are the locations for which the most number of states of the system have that location. This idea in fact relates back to our idea of entropy; if you were told that the end of the chain was actually at a certain value of $x = n$, but that all states which produced this end of chain location were equally likely, you could compute the information entropy of the question "what is the state of the system given that $x = n$?". For clarity, we can define this information entropy as $\sigma_{\text{Sys}}(x = n)$. In some sense, you have already done the necessary work to compute $\sigma_{\text{Sys}}(x = n))$ in the above steps. Write down a formula for $\sigma_{\text{Sys}}(x = n)$ explicitly.

Next, write down a formula relating $p(x = n)$ to $\sigma_{\text{Sys}}(x = n)$ for the same n. Note that larger values of $\sigma_{\text{Sys}}(x = n)$ relate directly to larger values of $p(x = n)$.

2.1.3 Stirling's approximation: a way to simplify the formula

Now we shall engage in a classic past-time of physicists: approximation. We will use the fact that $n!$ is approximately given by

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \tag{2}$$

when n is large. This equation is known as Stirling's approximation. Elementary explanations of why this is true can be found online if you are intrigued, as well as qualifications of its validity.

Questions Now, derive an approximate expression for $p(x = n)$ that is simpler than the one you found before by using Stirling's approximation. You may also find it useful to know that

$$e^x = \lim_{m\to\infty}\left(1 + \frac{x}{m}\right)^m.$$

We shall refer to this approximate probability formula that you find as $p_s(x = n)$, and it is **not** the exact value of the probability that $x = n$, although it is extremely close in most cases and makes some of the resulting physics clearer. You should obtain, with sufficient approximation, a proportionality of

$$p_s(x = n) \propto \frac{1}{\sqrt{N}}\exp\left(-\frac{4x^2}{N}\right).$$

(The symbol $\propto$ means "is proportional to" up to a constant factor. The notation "$\exp(x)$" means e^x. If you do not get precisely the same answer, this is alright, just show your work. The factor of $-x^2$ in the exponential is what is important.) Stirling's approximation is important because it allows us to make sense of the behavior of the system when N is large and when the exact formula is not particularly enlightening.

In particular, with our result for $p_s(x = n)$, we can find an expression for $\sigma_{\text{Sys}}(x = n)$ which is much easier to work with. Use our formula relating $p(x = n)$ to $\sigma_{\text{Sys}}(x = n)$ for the same n that we derived above to find a simpler expression for $\sigma_{\text{Sys}}(x = n)$.

2.1.4 An aside on energy vs. entropy

One of the important results of thermodynamics is that, for an abstract system exchanging energy with an energy reservoir, the probability of observing any state ω_i of the system is proportional to $\exp(-E(\omega_i)/\tau)$. The reservoir is simply a collection of objects with a large energy that can exchange that energy in some way with the abstract system. The quantity $E(\omega_i)$ is the energy of state ω_i and τ is a property of the reservoir called temperature. We will discuss this in more detail in Section 2.2.

Now, consider a 1 dimensional spring with spring constant k and equilibrium extension 0 that exchanges energy with a reservoir somehow. The probability of observing an extension to a length x would be proportional to $\exp(-(1/2)kx^2/\tau)$.

In some sense, we can make an analogy between the rigid chain and this 1 dimensional spring. In the case of the rigid chain, the negative of the entropy, or $-\sigma_{\text{Sys}}(x = n)$, plays the role of an "energy", and higher values of $-\sigma_{\text{Sys}}(x = n)$ lead to an exponentially suppressed probability of observing that value of x. Taking the analogy further, $-\sigma_{\text{Sys}}(x = n)$ changes approximately as $(1/2)kx^2$, which is similar to the potential energy of an extended spring. In the end, the system behaves nearly identically to a 1 dimensional spring exchanging energy with a bath. However, the potential energy of the analogous "spring" arises only from the *entropy* of the system, not from any real spring forces.

Summary The above discussion was meant to provide an intuition that entropy well and truly matters for a system such as the rigid chain. Now, it remains to be seen whether my interpretation of entropy acting like a spring restoring force has any validity (you should of course be skeptical of this rather contrived example).

2.2 An abstract derivation of entropy in statistical mechanics

Now that we have considered a specific example with an intuitive exhibition of how entropy appears in a mechanical system, we will consider a more general application of entropy in statistical systems.

2.2.1 Definitions

Consider an abstract system and reservoir (think of a small object like a glass of water and the room around it). The system S and reservoir R each have possible outcomes ω_i^S and ω_j^R where i and j range over some set of integers. However, we can consider their outcomes jointly, as possible outcomes $\omega_{i,j}^{S,R}$ defined to be $\omega_{i,j}^{S,R} = \{\omega_i^S, \omega_j^R\}$, meaning that S is in state i and R is in state j at the same time. Now, let us consider another quality of the system. We will call it energy, but what it really *is* doesn't matter to us yet. Let us assume that the total energy $E_{\text{tot}} = E_{\text{Res}} + E_{\text{Sys}}$ is conserved and therefore doesn't change in time (again, this is just an assumption). We may not know the exact value, but it is conserved nonetheless. Let us now make the assumption that all possible outcomes $\omega_{i,j}^{S,R}$ are equally likely, as long as they have the correct total energy. Thus we have the two assumptions:[5]

1. *All possible states for the combined system and reservoir $\omega_{i,j}^{S,R}$ must have the correct total energy.*

   This assumption should be somewhat familiar to physics students, and follows from the fact that energy is conserved globally, and that we assume the system and reservoir together are isolated from anything else in the universe.

2. *All states for the combined system and reservoir that have the correct total energy are equally likely to be observed.*

   Why are all states that have the correct total energy equally likely? To intuitively describe why, we need to change how we think about this physical situation. In a classical system (in the case where we ignore quantum mechanics), the system really *is* in one single state $\omega_{i,j}^{S,R}$ at a given instant in time. However, because of natural interactions, this state will change over time. In some cases, we can

[5] There are additional assumptions we are making, of course, but they are a little more subtle and less relevant to the point here. On some level, stating that we can distinguish the energy of the system from the energy of the reservoir, and the states of the system from the states of the reservoir, requires us to assume that there are not strong interaction energies between the system and reservoir. This is not a relevant detail at our level of analysis here, however.

assume that *the system and reservoir will visit every possible state $\omega_{k,l}^{S,R}$ over time, and that the system and reservoir will spend roughly the same amount of time in each state as the state cycles through all available states.* As long as the changes between states occur very quickly compared to the timescale between when we observe the system, then effectively we are just as likely to observe any one of the states as any other. This is roughly the physical picture of why each state of the combined system and reservoir with the correct total energy is equally likely to be observed.[6] This is basically a statement that we are as maximally confused as possible about *everything* about the system, except that total energy is conserved.

These are not the only assumptions we will make in the course of our analysis, but they are the important ones for now. It is important to keep track of additional assumptions we will make. We will now endeavor to see what these assumptions imply about the likelihood of various states of the system.

2.2.2 A derivation

Having established the necessary notation, we will consider how likely it is for a certain state ω_k^S to be observed which has energy E_k. Because all states $\omega_{i,j}^{S,R}$ are equally likely, the probability of observing ω_k^S is the number of states of the reservoir that can exist in a combination with ω_k^S, divided by the total number of available states of the combined system and reservoir. This is just the usual formula for the probability for event A to occur (a state ω_k^S of the system to be observed); it is the number of ways event A can occur (the number of states of the combined system and reservoir which have the system in state ω_k^S) divided by the total number of possible events (the number of all possible states of the combined system and reservoir). States of the reservoir, ω_j^R, that can pair with ω_k^S must have $E_j = E_{\text{tot}} - E_k$ because of the requirement that the energy of the combined system is E_{tot}. The number of states of the reservoir with energy $E_{\text{tot}} - E_k$ will vary as a function of E_k. Indeed, we can define $\exp(\sigma_{\text{Res}}(E))$ as the number of states of the reservoir which have energy E. In this way, the function $\sigma_{\text{Res}}(E)$ represents the logarithm of a specific number of states of the reservoir. This should remind you of the entropy of a question with many, equally-likely answers.

Explicitly, the probability of observing a state ω_k^S of the system with energy E_k is given by

$$p(\omega_k^S) = \frac{\exp(\sigma_{\text{Res}}(E_{\text{tot}} - E_k))}{\zeta_1}, \tag{3}$$

where ζ_1 (the Greek letter "zeta") is just a normalization constant that depends on the total number of outcomes possible for the combined system and reservoir, and ensures the total probability is 1. We can now define something called **temperature** τ (we will discuss its physical meaning later), to be given by[7]

$$\boxed{\frac{1}{\tau} \equiv \beta \equiv \left.\frac{\partial \sigma_{\text{Res}}(E)}{\partial E}\right|_{E=E_{\text{tot}}}}. \tag{4}$$

If we can assume that $\sigma_{\text{Res}}(E_{\text{tot}} - E_k)$ is only slightly altered from $\sigma_{\text{Res}}(E_{\text{tot}})$, then to a good approximation $\sigma_{\text{Res}}(E_{\text{tot}} - E_k) = \sigma_{\text{Res}}(E_{\text{tot}}) - \beta E_k$, with β being the constant inverse temperature defined above, and

$$\boxed{p(\omega_k^S) = \frac{e^{\sigma_{\text{Res}}(E_{\text{tot}})}}{\zeta_1} \exp(-\beta E_k) = \frac{1}{\zeta_2} \exp\left(-\frac{E_k}{\tau}\right),} \tag{5}$$

where we have used the fact that $\exp(\sigma_{\text{Res}}(E_{\text{tot}}))$ is a constant and we have defined ζ_2 as another normalization constant. The above formula is extremely important and its derivation should be well understood because similar ideas and derivations will be used throughout this document.

One objection you may have to the above argument is whether we can truly consider temperature (as defined here) to be fixed in some physical systems independent of E_k, or in other words you may wonder what the degree of error is in the statement that $\sigma_{\text{Res}}(E_{\text{tot}} - E_k) = \sigma_{\text{Res}}(E_{\text{tot}}) - \beta E_k$. In general, what we

[6] This physical pictures is further explained in Section 4.0.1.

[7] Note that $\equiv$ is shorthand for "is defined to be", as opposed to simply "equals".

are really doing here, if you have a background in Calculus, is Taylor expanding $\sigma_{\text{Res}}(E_{\text{tot}} - E_k)$ close to $\sigma_{\text{Res}}(E_{\text{tot}})$. First we write

$$\sigma_{\text{Res}}(E_{\text{tot}} - E_k) = \sigma_{\text{Res}}\left(E_{\text{tot}}\left(1 - \frac{E_k}{E_{\text{tot}}}\right)\right).$$

Next, to be concise, define $\epsilon_k \equiv E_k/E_{\text{tot}}$. Then by Taylor expanding we have

$$\sigma_{\text{Res}}(E_{\text{tot}} - E_k) = \sigma_{\text{Res}}(E_{\text{tot}}) + (-\epsilon_k)E_{\text{tot}} \left.\frac{\partial \sigma_{\text{Res}}(E)}{\partial E}\right|_{E=E_{\text{tot}}} + (-\epsilon_k)^2 \frac{1}{2!}(E_{\text{tot}})^2 \left.\frac{\partial^2 \sigma_{\text{Res}}(E)}{\partial^2 E}\right|_{E=E_{\text{tot}}} + \cdots.$$

In the event that E_k is "small enough" so that $\epsilon_k = E_k/E_{\text{tot}} \ll 1$, then only keeping terms in the Taylor expansion to first order in ϵ_k (the first two terms on the right hand side of the "=" sign) accurately describes the value $\sigma_{\text{Res}}(E_{\text{tot}} - E_k)$. For a fixed E_{tot}, this just requires that E_k is made small enough.

Questions If we now consider the probability of observing a certain energy rather than a certain state for the system, what is the new formula for $p(E_{\text{Sys}} = E_k)$? If there are $\exp(\sigma_{\text{Sys}}(E))$ number of states of the system at energy E, prove that

$$\boxed{p(E_{\text{Sys}} = E_k) = \frac{1}{\zeta_3}\exp(-\beta(E_k - \tau\sigma_{\text{Sys}}(E_k)) = \frac{1}{\zeta_3}\exp(-\beta\mathcal{F}(E_k))} \tag{6}$$

for some constant ζ_3. Here we have defined a function called the **Free Energy** $\mathcal{F}(E_k) = E_k - \tau\sigma_{Sys}(E_k)$ as a function of E_k. This formula for $p(E_{\text{Sys}} = E_k)$ should be reminiscent of the bendable chain example derived earlier.

Note that the most likely energy of the system to be observed is the one which results in the lowest value of $\mathcal{F}(E_k)$, and this energy is not necessarily the lowest possible value of energy. Can you describe in words why this is true? What two effects are "competing" to determine the most likely state?

2.2.3 Clarifying types of entropy

As a clarification, I want to mention that we need to keep straight a few of the different entropies that we are considering. As was discussed above, the entropy of a system is equivalent to the information entropy of the question "what is the state?" Clearly, this has different meanings if we are talking about the state of the combined system and reservoir, or if we are talking about the state of the system alone, or if we are talking about the state of the system when we are at a fixed energy. Below is a list of the various entropies we have been considering and their distinctions.

1. Entropy of the state of the combined system and reservoir:

   This is the entropy of the question "what is the state of the combined system and reservoir?" While this entropy is generally not considered in this document (it is not particularly useful), you should know what it is. It depends only on the size of the set of all possible outcomes of the combined system and reservoir, or $\{\omega_{i,j}^{S,R}\}$.

2. Entropy of the state of the system:

   This is the entropy of the question "what is the state of the system?" Full stop. We assume we know nothing about the energy of the system. We don't care about the state of the reservoir. The entropy is given by the formula in eq. (1) as

   $$\boxed{\sigma_{\text{Sys}} \equiv -\sum_k p(\omega_k^S)\log(p(\omega_k^S)).} \tag{7}$$

   This value, σ_{Sys}, is distinct from $\sigma_{\text{Sys}}(E)$, although their notation is similar.

3. Entropy of the state of the system at a fixed energy:

   This is the entropy of the question "what is the state of the system *if we know that the energy has a specific value*?" Generally, this will be denoted by $\sigma_{\text{Sys}}(E)$ rather than simply σ_{Sys}, and it is the

appearance of (E) that should alert you to whether we are considering the entropy of the state of the system or of the state of the system at fixed energy. Is the entropy of the system at fixed energy generally less than or greater than σ_{Sys} (this question is just for you to think about, not respond to)?

4. Entropy of the energy of the system:

   While we will not discuss this entropy much in this document, it is also possible to consider the entropy of the question "what is the energy of the system?" We could write down the formula for this entropy using the probability of each value of energy in eq. (6). Note that this is different than σ_{Sys}. The main idea to understand from this example is that the question "what is the energy of the system?" is a distinct question from asking about the state of the system and so has a distinct value of information entropy.

2.2.4 A pause for reflection

Question We now ask that you pause, reflect, and try to list all of the assumptions that went into the derivation of the above result; in other words, effectively summarize your understanding of when the model is valid or not valid. Try to come up with assumptions that were not even explicitly stated. The more you understand about how a result is limited, the more you also understand how to apply it.

2.3 Why is our definition of temperature reasonable?

You probably did not expect temperature to be defined as the abstract quantity in eq. (4). Why would this definition reflect the notion of temperature that you probably already have in your head? One conventional notion of temperature is the amount of vibration of atoms in a material. Another way we think of temperature is as defining a scale of energy. If one object is hotter then another object and we put the two objects in contact, we expect energy to flow from the hot object to the cool object. Is this what happens with our new definition of temperature in eq. (4) as well?

To better understand temperature as defined here, we will now explore a physical example of a reservoir and system that share energy and think about how the temperature of this reservoir is determined.

A physical example: magnetic spins Consider a set of N independent magnetic spins that can point either up or down. You can think of a magnetic spin as a very small bar magnet that points in a certain direction (in the physical world, this is not what spin is, but for our purposes it works). Applying a magnetic field applies a torque to the spin and forces it to align with the magnetic field. If the spin aligns with the field, it is at a lower energy than if it anti-aligns with the field. In this physical example, assume there is an external magnetic field pointing downward, so that the energy of each spin is $\eta > 0$ when the spin points upward, and $-\eta$ when the spin points downward.

From here on out, it is not critical to use the physical interpretation of this example to derive results. However, the physical interpretation still applies.

Questions What is the shape of $\sigma_{\text{Sys}}(E)$? For example, a basic correct description is "this function starts at 0 for $E = -N\eta$, increases to a maximum near $E = 0$ and goes back to 0 at $E = +N\eta$. Moreover it is symmetric for $E \to -E$." Do not bother with finding an exact expression of this function in this section of the exam, but explain in your own words why the description of the shape above is correct.

In particular, first realize that possible values of E range from $-N\eta$ to $+N\eta$. To identify the shape, you can use much of the same mathematical work that you did in the previous section to analyze this example. Just notice the parallels between the number of ways a chain can be at position $x = n$ and the number of ways the total energy of the spins can be at $E = N\eta$ for an integer $|n| \leq N$.

A large reservoir of spins coupled to a smaller system Once you have found how $\sigma_{\text{Sys}}(E)$ behaves, you can instead consider a system of spins with $N \gg 1$ as a reservoir of energy for another smaller system. In this case, we can write the entropy of our reservoir of spins as $\sigma_{\text{Res}}(E)$ with the same functional form as $\sigma_{\text{Sys}}(E)$ which you found above. This is just a way of rewriting a label, and nothing deeper.

Assume that a large bath of spins is hooked up to some smaller subsystem in a way where they can exchange energy, but so that the total energy is fixed. For example, the smaller subsystem could be another set of spins which are close enough to feel the magnetic field of the large reservoir of spins and exchange energy with them.

Questions Using the language of the previous section, what is the temperature of the reservoir of spins as a function of the energy of the reservoir? Again, focus only on the shape of the function based on the shape of $\sigma_{\text{Res}}(E)$ which you found above. Describe the shape in a similar way. When is the temperature decreasing, and when is it increasing as a function of E? When is it maximum and when is it minimum? You will note that the temperature of the reservoir is sometimes negative for some values of reservoir energy. We will later see the meaning of this negative temperature.

Two equal sized systems of spins coupled to each other Now consider two large reservoirs of spins of approximately equal size that are coupled to each other so that they can exchange energy. Again, the total energy shared by the two reservoirs is fixed. If reservoir A and reservoir B have total energy $E_A + E_B = E_{\text{tot}}$, then the total entropy is $\sigma_A(E_A) + \sigma_B(E_B)$ for a fixed distribution of energy between the two systems. Recalling previous discussion, we can ask what distribution of E_{tot} between reservoirs A and B will be the most probable? The answer is, the one that results in the greatest total entropy of the combined two reservoirs.

Questions Show that the most probable distribution of energy between reservoirs A and B is the distribution which causes $\tau_A = \tau_B$, or equal temperatures for reservoirs A and B as defined in eq. (4).

A few claims about the time evolution of two systems that are brought into contact At this point we will appeal to some intuitive reasoning that has not been adequately explained yet. We claim that if you start out with some energy distribution E_A and E_B for the two reservoirs when they are isolated from each other and then you bring the two reservoirs into contact so that they can exchange energy, they will relax to the distribution with $\tau_A = \tau_B$ which still respects the correct total energy $E_A + E_B = E_{\text{tot}}$. When I say "relax to", I mean that the final probability distribution for the division of energy between the two reservoirs will be such that it is highly unlikely to observe anything other than this final energy division that produces $\tau_A = \tau_B$. The reason why this is true is that we are fixing the initial state, but once the reservoirs are brought into contact and can exchange energy, all states with the correct total energy are equally likely if we wait long enough; among all of these possible states, the energy distribution that is by far most likely to be observed is the one with $\tau_A = \tau_B$.

Questions With this assumption about "relaxation" in mind, what happens if we somehow start the system out with $E_A < E_B < 0$? You may use the fact that the final distribution of energy in the system will produce $\tau_A = \tau_B$. For this and subsequent analysis, you should only need to know the shape of $\sigma(E)$ (and therefore τ as a function of energy) for both reservoirs. Consider the initial and final temperature of each reservoir, defined as in eq. (4). Which way does energy flow? Is it from the higher temperature to lower temperature system as we expect?

Now, consider if we begin the system with $E_A = -E_B$. Which way does energy flow in this case? From higher to lower temperature? Is the temperature of one reservoir negative? Is it infinite at some point in time during relaxation? Consider what happens if we instead look at $\beta = 1/\tau$, the inverse of temperature. How does energy flow relative to the initial values of β_A and β_B. Is the energy flow always in the same direction (as in, always from high β to low β, or the reverse)? If so, perhaps the more physical quantity is $\beta = 1/\tau$, not τ...

You should now have some intuition for what the abstract definition of temperature τ as defined in eq. (4) actually means, and how it functions in practice in the physical world. Moreover, we have seen that this quantity τ behaves the way we expect of temperature from our everyday experiences in terms of energy flow, at least in some cases.

2.4 Extensions of the conceptual model

While we only considered systems with a variable called "energy" in our derivation above, whatever that might be, we can all clearly think of other variables that might describe a system, such as volume, or the number of particles contained in the system. These are clearly important for the interface between gas molecules in the air and liquid molecules in water, or for the compression of a gas in the piston of a car. If you have taken any courses on thermodynamics, chances are that you have encountered these variables before, perhaps in terms of the ideal gas law, $\rho V = N\tau$. The beauty of the abstract derivation above is that it easily generalizes to include these other descriptions of a system that we might consider.

As a concrete example, consider a system and reservoir with some volume. Like energy, we can assume that the system and reservoir occupy a total fixed volume, with the volume of the system much smaller than that of the reservoir. Perhaps the system and reservoir occupy adjacent volumes, and push up against each other through a thin, movable wall. We can again consider entropy, although this time as $\sigma_{\text{Res}}(E, V)$, a function of energy and volume. Carrying out the same analysis as above, we would arrive at a probability distribution as a function of E and V that behaves like

$$p(E_{\text{Sys}} = E, V_{\text{Sys}} = V) = \frac{1}{\zeta} \exp\left(-\frac{1}{\tau}\left(E - \tau\sigma_{Sys}(E, V)\right)\right) \exp\left(-V \left.\frac{\partial \sigma_{\text{Res}}(E, V)}{\partial V}\right|_{E=E_{\text{tot}}, V=V_{\text{tot}}}\right).$$

If we define the **pressure** ρ to be

$$\rho \equiv \tau \left.\frac{\partial \sigma_{\text{Res}}(E, V)}{\partial V}\right|_{E=E_{\text{tot}}, V=V_{\text{tot}}},$$

then

$$p(E_{\text{Sys}} = E, V_{\text{Sys}} = V) = \frac{1}{\zeta} \exp\left(-\frac{1}{\tau}\left(E + \rho V - \tau\sigma_{Sys}(E, V)\right)\right).$$

In this case, the effective "free energy function", if you want to call it that, is $E + \rho V - \tau\sigma_{Sys}(E, V)$ rather than the free energy $\mathcal{F}(E)$ defined before. This function is often called Gibbs free energy, or $\mathcal{G}(E, V)$. The point is that, even though our conceptual model got more complicated because we now considered there to be some kind of physical volume to our system and reservoir, we could still use the same method of derivation as before to say something meaningful about the probability distribution of the system. It is important to note that we need to be careful when applying the formula above. It should only be valid when the pressure ρ and temperature τ are sufficiently independent of the system's details and therefore approximately constant. This requires that the energy and volume of the system are very small compared to the energy and volume of the reservoir, for the same reasons that were explained in our earlier introduction of τ.

3 A more precise analysis of free energy and entropy

The natural question to ask at this point in our analysis is what can we do with this all of this theory? What can we calculate that actually has relevance for the real world? In particular, we may want to ask what energy is extractable from a system, on average. In many ways, the answer to these questions is curiously tied up with our understanding of free energy and entropy.

Let us return to considering the free energy functions $\mathcal{F}$ and $\mathcal{G}$. So far we have been a little unclear about what these functions are. We have treated $\mathcal{F}(E)$ and $\mathcal{G}(E, V)$ as functions. To be more precise, these are possible values of combinations of quantities that we observe, and each value of $\mathcal{F}(E)$, for example, has a probability of being observed given by eq. (6). Rather than talk about the probability of each value of $\mathcal{F}(E)$ that can be observed, it is often more useful to talk about something like the average value of $\mathcal{F}(E)$ that will be observed. Similarly, it is often more useful to talk about the average energy that you will observe rather than to talk about how likely each possible energy is to be observed. We can then ask questions about how the average energy changes and how the free energy changes when we alter the system, and this will lead us to understand the work that is extractable from the system.

Motivation for further study of free energy and entropy In this short section, we will give an example of why a more precise understanding of free energy and entropy is needed; this example is meant to be confusing, and to make you realize that some concepts you might have thought you understood were actually very poorly explained.

Consider one of the first concepts that is usually taught in thermodynamics: that the energy of the system obeys $\delta E = \tau\delta\sigma - \rho\delta V$. (Here the notation δx just means a small change in a quantity x.) Without further explanation of what these terms stand for, this statement is virtually useless. Which entropy is being considered here, for example, of all of the types of entropy listed in Section 2.2.3? (You do not need to actually answer this question here. It is rhetorical.) It is further argued that therefore $\delta\mathcal{F} = \sigma\delta\tau - \rho\delta V$ so that $\mathcal{F}$ is independent of entropy for a fixed temperature. It is usually then argued that changes in $\mathcal{F}$ represent changes in the energy available for extraction from the system.

At this point, we note a few potentially confusing issues. First and foremost, $\mathcal{F}(E)$ in the statements above, as well as E, are variables that take on a set of values with certain probability. What do we therefore mean by changes in E, denoted by δE? Moreover, we have yet to make explicit what work done on a system might consist of, and have yet to define energy available for extraction. Clarifying these concepts will be the goal of the following analysis.

Outline of the subsequent analysis The following sections will be an endeavor to make the following statements more precise:

1. What is a more general way to define free energy?
2. How does free energy relate to average values of energy in the system and to the entropy of the system (which will be precisely defined), and how are changes in these values related?
3. How can we define energy that is available for extraction from a system and how can we define work done on a system? In particular, how do we measure changes in these quantities?

First, we will consider in many ways the most simple system possible: a system coupled to an energy reservoir without any concept of volume or other variables. Using this example, we will determine if we can satisfactorily answer the above questions. Second, we will consider in full generality a system coupled to a bath with any number of variables defining the system, such as energy and temperature, volume and pressure, number of particles and chemical potential, and so on. The end result of this section will be a precise understanding of the uses of free energy and entropy when analyzing a statistical mechanical system.

3.1 A system coupled to an energy reservoir

Here we will define a more general notion of free energy for a system coupled to an energy reservoir only. We will then derive how changes in entropy, energy, and free energy are related for this system, as well as what we mean by extractable work. This abstract system will serve as a model for more general systems.

3.1.1 A new definition of free energy

In order to define a more general concept of free energy from an alternative perspective, we will first have to define a few related pieces of machinery in statistical mechanics. One such ubiquitous piece of machinery is the partition function $\mathcal{Z}$, defined as

$$\boxed{\mathcal{Z} \equiv \sum_{\omega_k} \exp(-\beta E_k).} \tag{8}$$

The partition function $\mathcal{Z}$ is the normalization factor which we divide by to get $p(\omega_k) = \exp(-\beta E_k)/\mathcal{Z}$. In fact, $\mathcal{Z}$ is the explicit formula for the quantity ζ_1 that we defined as the normalization constant in eq. (3). We have removed the superscript designation S from ω_k^S because, from now on, ω_k will generally refer to the state of the system and so the specification S is implied. All we have really done is to explicitly write out a formula for the constant of proportionality in eq. (5). Here we should note that $\mathcal{Z}$ can be changed by changing any one of the E_k that exist, or by changing $\beta = 1/\tau$. In a sense, therefore, $\mathcal{Z}$ can be viewed as $\mathcal{Z}(\tau, \{E_i\})$, a function of τ and of the set of all E_i, denoted by $\{E_i\}$.

Now, we will define a new free energy function $\mathcal{F}$ as

$$\boxed{\mathcal{F} \equiv -\tau \log(\mathcal{Z}).} \tag{9}$$

As was stated above, because $\mathcal{Z}$ can be viewed as a function of τ and $\{E_i\}$, likewise so can $\mathcal{F}$. It turns out that this definition of $\mathcal{F}$ has relevance to our earlier concept of free energy. It is your job to find out why:

Questions

1. Using the explicit expression for $p(\omega_k) = \exp(-\beta E_k)/\mathcal{Z}$ mentioned above, show that

$$\sigma_{\text{Sys}} = -\beta^2 \frac{\partial}{\partial \beta}\mathcal{F} = -\frac{\partial}{\partial \tau}\mathcal{F} = \frac{\partial}{\partial \tau}\tau \log(\mathcal{Z})$$

   are all equivalent descriptions of the entropy of the system (this is the entropy of the question "what is the state of the system" if the energy takes on a statistical distribution of values).

2. If we denote the average observed value of energy E as $\langle E \rangle$ (More generally, for any quantity x that can be observed, we denote the average value that you expect to observe by $\langle x \rangle$), show that

$$\langle E \rangle = -\frac{\partial}{\partial \beta} \log(\mathcal{Z}).$$

3. Show that

$$\boxed{\mathcal{F} = \langle E \rangle - \tau \sigma_{\text{Sys}}.} \tag{10}$$

   Two suggested ways to do this are to use the derivative relations above, or to explicitly work with the definitions in terms of $p(\omega_k)$ to get the desired result.

With the result that $\mathcal{F} = \langle E \rangle - \tau \sigma_{\text{Sys}}$, it is clear that our new definition of free energy bears some resemblance to the old definition. Where before $\mathcal{F}(E) = E - \tau \sigma_{\text{Sys}}(E)$ did not tell us directly about the average energy, our new function $\mathcal{F}$ does. Our old definition also depended on the entropy of the state of the system *at fixed energy*, while the new $\mathcal{F}$ is related to the entropy of the state of the system as a whole. It turns out that this new form of $\mathcal{F}$ is more useful, as we will see. From now on, you should assume that *all further references to $\mathcal{F}$ refer to this new definition of free energy in eq. (9) and eq. (10).*

Relations between changes in average energy and entropy We will now try to determine how changes in average values of observables, such as $\langle E \rangle$, are related to changes in σ_{Sys}. Before we begin, we must clarify what we mean by changes. The values $\langle E \rangle$, σ_{Sys}, and $\mathcal{F}$ can all be thought of as functions of τ and $\{E_i\}$, in the sense that they have explicit formulas in terms of these variables. We can therefore explicitly compute their partial derivatives with respect to changes in τ or changes in some E_i. From these partial derivatives, we can find laws between how $\langle E \rangle$ changes and how σ_{Sys} changes that are always true

regardless of the mechanism which produces that change (whether it is changing τ or changing some E_i). With this method, we will be able to show that

$$\boxed{\delta\langle E\rangle = \tau\delta\sigma_{\text{Sys}} + \delta\mathcal{W}_{\text{Ext. on Sys}}} \tag{11}$$

where $\mathcal{W}_{\text{Ext. on Sys}}$ refers to the external work done on the system. Again, the notation δx means a small change in quantity x, and the statement above means that no matter how these small changes are produced, the relation in eq. (11) holds true. This equation is a more precise statement in terms of well defined variables than the vague assertion that we previously made called the first law of thermodynamics that $\delta E = \tau\delta\sigma - p\delta V$ (note that $-p\delta V$ is analogous to $\mathcal{W}_{\text{Ext. on Sys}}$). We will use this relation to define the energy extractable from a system and to analyze applications of statistical mechanics in the real world.

Questions We will now proceed with the derivation of eq. (11). We need to consider all possible ways of changing $\langle E\rangle$ and σ_{Sys} and show that, in all cases, eq. (11) holds. The most general variation in all of our functions is caused by changes in τ and in various E_i. While it is possible to explicitly compute the partial derivatives of $\langle E\rangle$ and σ_{Sys} with respect to all of these variables and use these partial derivatives to prove eq. (11), instead we can take the following shortcut: partial differentiate $\mathcal{F}$ with respect to τ and show that the result implies that

$$\frac{\partial\langle E\rangle}{\partial\tau} = \tau\frac{\partial\sigma_{\text{Sys}}}{\partial\tau}.$$

You will find it helpful to recall that we know an expression for σ_{Sys} in terms of $\partial\mathcal{F}/\partial\tau$. Now partial differentiate $\mathcal{F}$ with respect to E_i for some integer i using the explicit formula $\mathcal{F} = -\tau\log(\mathcal{Z})$, all while holding τ fixed. When you have found your result, equate it to the formula for $\partial\mathcal{F}/\partial E_i$ in terms of $\langle E\rangle$ and σ_{Sys} to obtain

$$\frac{\partial\langle E\rangle}{\partial E_i} = \tau\frac{\partial\sigma_{\text{Sys}}}{\partial E_i} + p(\omega_i).$$

From the two main equations derived above, we can conclude that

$$\delta\langle E\rangle = \tau\delta\sigma_{\text{Sys}} + \sum_{\omega_i} p(\omega_i)\delta E_i. \tag{12}$$

Please explain in your own words why we can conclude this (this is essentially a result from calculus concerning infinitesimal quantities).

Defining external work Yet, we are not done; in order for eq. (12) to match eq. (11) above, we must identify $\sum_{\omega_i} p(\omega_i)\delta E_i$ as an infinitesimal amount of work that has been done on the system, or $\delta\mathcal{W}_{\text{Ext. on Sys}}$. To see why this is a reasonable definition, think about how we might produce a change in some value of E_i. In particular, energy levels change because *you do something to the system.* There might be an external parameter that an observer can control, like a lever, to alter energy levels. Call the value of this parameter x. In this case, as you change x, $\delta E_j = (\partial E_j/\partial x)\delta x$ for all j, and

$$\sum_{\omega_i} p(\omega_i)\delta E_i = \sum_{\omega_i} p(\omega_i)\frac{\partial E_i}{\partial x}\delta x = \left\langle\frac{\partial E}{\partial x}\right\rangle\delta x = \left\langle\frac{\partial E}{\partial x}\delta x\right\rangle. \tag{13}$$

Note that *it is not always true that* $\langle\partial E/\partial x\rangle$ *is the same as* $\partial\langle E\rangle/\partial x$. Instead, $\langle\partial E/\partial x\rangle$ stands for precisely what is shown in the above equation. This is a crucial and subtle point. These quantities are not always equal because $p(\omega_i)$ *actually depends on* E_i and therefore on x. For now, just keep this distinction in mind.

Why is it reasonable to interpret this as the work done on a system? Well, if the system were exclusively in state E_k, then $(\partial E_k/\partial x)\delta x$ would be precisely the infinitesimal amount of work that was done on the system. We are not in any one state of the system, however, so the external work done on the system is only something we can talk about as a statistical average. This is why we must average the quantity inside the $\langle\cdot\rangle$ brackets to get $\delta\mathcal{W}_{\text{Ext. on Sys}}$.[8]

[8]The more accurate picture is that the system switches between all ω_i rapidly, and spends an amount of time in each state ω_i that is proportional to $p(\omega_i)$. Therefore, as the external parameter x is changed slowly, the work that is done on the system is the sum of the work done during each interval that the system spends in each state. This sum is precisely given by eq. (13). This interpretation of what physically happens in the system is explained further in Section 4.0.1 on non-equilibrium and equilibrium changes to a system.

Thus we have found the average external work done on the system if we change a single parameter x by an amount δx. However, the total external work done on the system will come from multiple external parameters $\{x_k\}$ changing, where k ranges over some indices. In this case, the most general expression for the total external work done on the system by all of the infinitesimal changes $\{\delta x_k\}$ is

$$\boxed{\delta \mathcal{W}_{\text{Ext. on Sys}} \equiv \sum_k \left\langle \frac{\partial E}{\partial x_k} \right\rangle \delta x_k = \sum_k \sum_{\omega_i} p(\omega_i) \frac{\partial E_i}{\partial x_k} \delta x_k} \tag{14}$$

3.1.2 Defining energy that is extractable from a system in terms of free energy

We have thus shown that eq. (11) is true (when we define external work correctly), and we have precisely defined all variables that appear in the equation. The next step is to try to construct a useful definition of the energy that can be extracted from a system and used. Because $\mathcal{F} = \langle E \rangle - \tau \sigma_{\text{Sys}}$,

$$\delta \mathcal{F} = \delta \langle E \rangle - \sigma_{\text{Sys}} \delta \tau - \tau \delta \sigma_{\text{Sys}}$$

by the simple product rule of differentiation. If we then substitute our derived expression for $\delta \langle E \rangle$, we obtain

$$\boxed{\delta \mathcal{F} = -\sigma_{\text{Sys}} \delta \tau + \delta \mathcal{W}_{\text{Ext. on Sys}}.} \tag{15}$$

This equation implies that changes in σ_{Sys} alone for a fixed temperature do not alter $\mathcal{F}$, while changes in τ and external work do modify $\mathcal{F}$. How does this help us define work extractable or energy extractable from the system? Well, if we hold temperature τ constant, changes in $\mathcal{F}$ are precisely equivalent to work done on the system, which is the negative of work done by the system (this is Newton's third law in action). Therefore, if an external actor brings $\mathcal{F}$ *down* by some amount $\Delta \mathcal{F}$, then the system has done precisely $\Delta \mathcal{F}$ worth of work on the external actor.

Another argument that might convince you that free energy represents the extractable energy is the following: consider some external work applied to the system at constant temperature τ to change $\mathcal{F}$ from $\langle E \rangle_1 - \tau \sigma_{\text{Sys},1}$ to $\langle E \rangle_2 - \tau \sigma_{\text{Sys},2}$. If you were to now somehow use external parameters x_k to alter the system's entropy $\sigma_{\text{Sys},2}$ back to its original value $\sigma_{\text{Sys},1}$ *without doing any external work on the system*, then $\mathcal{F}$ would be unchanged from its final value, $\mathcal{F}_2$ (this follows from eq. (15)). However, the system would have a new value of average energy, $\langle E \rangle_2^*$, instead of $\langle E \rangle_2$. Therefore $\mathcal{F}_2 - \mathcal{F}_1 = \langle E \rangle_2^* - \langle E \rangle_1$ is an *effective change in the average internal energy of the system if entropy is unchanged.* It might make intuitive sense to interpret this difference in average internal energy as a difference in extractable work because it is the increase in average energy *between comparable systems with the same entropy*; it might only make sense to compare the work extractable from a system if they have the same entropy.

In any case, we can now see that the average change in energy extractable from a system at constant temperature is $\Delta \mathcal{F}$.

3.2 A system coupled to an arbitrary reservoir

This section is meant to extend the analysis of the energy extractable from a system and free energy for systems coupled to arbitrary baths which depend on other variables, such as volume or particle number. This section does not contain questions for the reader directly and can be skipped or skimmed. Readers may prefer to refer to the summary at the end of this section and only reference the derivations within if necessary. However, a more unified understanding of the framework behind free energy can be beneficial in terms of a deeper understanding of related concepts.

Motivation for devising a generalized framework Most fields of physics such as kinematics are taught by introducing a few specific examples and slowly building a physical model or explanation that addresses these examples. For example, in kinematics, the problem of describing a rotating rod leads to the equation $F = ma$ and subsequently to a theory of angular momentum. In this process, it is important to not lose sight of the goal of a generalized understanding; if a kinematics class only taught you how to solve the two examples of a thrown ball and a rotating rod it would clearly miss the point that these examples tell us much

more about the physical world through generalization. Thus, in our analysis of statistical mechanics, we do not seek to only answer the question "how can we solve problems X, Y, and Z?" Instead, we seek to answer "what does knowing how to solve problems X, Y, and Z tell us about an entire class of problems describing the physical world?" An explicit generalized framework for statistical mechanical systems will help us more clearly answer this question about a class of problems.

3.2.1 A more general model of a system

What is the most general system coupled to a reservoir that we can still apply a variant of the above reasoning to? In general, we can divide the external parameters that describe the system into two categories. The fundamental distinction between these categories is that sometimes variables describing the system, such as energy E or volume V are fixed, and sometimes they are statistically distributed. You must examine a system and determine what variable is in each category.

If we denote the set of statistically distributed variables by $\{\lambda_k\}$ and the set of fixed external parameters by $\{\nu_j\}$, then the combined set of λs and νs describes all of the properties of the system. The properties of the system which can be directly altered by an observer, however, are only the νs. In contrast, the λs are statistically distributed and cannot be altered directly. However, the observer can also do something else to affect the system indirectly: alter the bath or reservoir. A few parameters describing the bath are the quantities α_k corresponding to each k and λ_k. We define

$$\alpha_k \equiv \frac{\partial \sigma_{\text{Res}}(\{\lambda_j\})}{\partial \lambda_k}$$

where the notation $\sigma_{\text{Res}}(\{\lambda_j\})$ means that the reservoir jointly depends on the values of all λ_j. For example, if $\lambda_0 = E$, then $\alpha_0 = 1/\tau$, the inverse temperature. Alternatively, if $\lambda_1 = V$, then $\alpha_1 = \rho/\tau$, the pressure divided by the temperature (the factor of temperature is not conceptually important right now). The set $\{\alpha_k\}$ and the set $\{\nu_j\}$ comprise the set of external parameters that can be used to alter the system, both directly and indirectly.

How do we go about determining the state of the system? Well, by definition, we simply set the values of $\{\alpha_k\}$ and $\{\nu_j\}$. However, the variables $\{\lambda_k\}$ are statistically distributed, and we must figure out what this distribution looks like. For these variables $\{\lambda_k\}$ such as $\lambda_0 = E$, you are left with only one assumption to fall back on: that the total combined value of each λ_k for the system and the reservoir is constant. One further assumption that we need to make is that the value of α_k (defined as a derivative of entropy) is a property of the reservoir that is more or less constant independent of the value of λ_k in the system. Once again, this more or less requires that the reservoir has very large values of each λ_k. If these properties are satisfied, we can apply the analysis of the previous sections.

First, we must establish one key piece of notation. The system can have various states ω_s. Each state ω_s will have a certain value of λ_k. For each λ_k, we denote this specific value by $\lambda_k(\omega_s)$. Thus if a particular state ω_2 had $\lambda_0 = 5.1$ and $\lambda_1 = 2.7$, then we would write $\lambda_0(\omega_2) = 5.1$ and $\lambda_1(\omega_2) = 2.7$. Note that generally each λ_k can have different values independent of the other λ_j, so that the set of all ω_s will usually consist of all possible combinations of explicit values of each λ_k.

With this notation, we can proceed with a similar analysis to previous sections. We consider the number of possible combined states of the system and reservoir depending on each value of $\lambda_k(\omega_s)$ and differentiate the entropy of the bath with respect to each λ_k. With this process, we can show that the probability

$$p(\omega_s) = \frac{1}{Z} \exp\left(-\sum_k \alpha_k \lambda_k(\omega_s)\right). \tag{16}$$

Again, we define a partition function Z as

$$Z = \sum_{\omega_s} \exp\left(-\sum_k \alpha_k \lambda_k(\omega_s)\right). \tag{17}$$

This formula immediately lets us see that

$$\langle \lambda_k \rangle = -\frac{\partial}{\partial \alpha_k} \log(Z). \tag{18}$$

Here we pause and note that, by assumption, the only variables that change in this formula are $\{\alpha_k\}$ and $\{\nu_j\}$ (changes in $\{\nu_j\}$ act implicitly but not explicitly because each ν_j affects all of the $\lambda_k(\omega_s)$ individually). Therefore we can treat Z and any function derived from it as a function of $\{\alpha_k\}$ and $\{\nu_j\}$.

3.2.2 A "generalized free energy function" and its uses

As has been demonstrated above, we can more easily derive general formulas if we treat all of the variables λ_k including energy on equal footing. For this reason, when we define something similar to a "free energy" in its definition and in its uses, we will not privilege it by giving it the units of temperature. To that end, we define a "generalized free energy function"[9] G by

$$G \equiv -\log(Z).$$

We can then use the formula for entropy in terms of $p(\omega_s)$ to derive that

$$\sigma_{\text{Sys}} = -\sum_{\omega_s} p(\omega_s)\log(p(\omega_s)) = -\sum_{\omega_s} p(\omega_s)\left(-\sum_k \alpha_k\lambda_k(\omega_s) - \log(Z)\right) = \sum_k \alpha_k\langle\lambda_k\rangle + \log(Z). \quad (19)$$

This directly implies that

$$\boxed{G = \sum_k \alpha_k\langle\lambda_k\rangle - \sigma_{\text{Sys}}.} \quad (20)$$

From the above equation, the parallels between G and free energy as defined before are notable (just let $\alpha_0 = \beta = 1/\tau$).

Changing various α_k: As before, we are still interested in how $\langle\lambda_k\rangle$ change relative to each other and to entropy. Again, we can compute $\partial G/\partial\alpha_k$ using the definition of G in terms of the partition function in eq. (17). Eq. (18) directly implies that this is $\langle\lambda_k\rangle$. We can set this value equal to $\partial G/\partial\alpha_k$ in terms of the $\{\langle\lambda_k\rangle\}$ and σ_{Sys} from eq. (20). The end result, with some cancellation and redistributing, is

$$\frac{\partial\sigma_{\text{Sys}}}{\partial\alpha_k} = \sum_m \alpha_m \frac{\partial\langle\lambda_m\rangle}{\partial\alpha_k}$$

or

$$\delta\sigma_{\text{Sys}} = \sum_m \alpha_m \delta\langle\lambda_m\rangle$$

as long as only the various α_k are varied, not the ν_j.

Changing various ν_j: Now we consider the only possible other change of the system, which is to change some ν_j. By definition, changing ν_j acts on the system by producing changes in each $\lambda_k(\omega_s)$, meaning in the value of λ_k for each state. These changes are given by $\delta\lambda_k(\omega_s) = (\partial\lambda_k(\omega_s)/\partial\nu_j)\delta\nu_j$. Although we did not write it explicitly, $\lambda_k(\omega_s)$ is really a function $\lambda_k(\omega_s, \{\nu_j\})$ of the state and the external variables ν_j. Once again, we consider changes in G. In particular, we consider changes when we alter the value $\lambda_k(\omega_s)$ for a specific value of k and specific value of s. We find $\partial G/\partial\lambda_k(\omega_s)$ using both the explicit formula and eq. (20), and we obtain

$$\frac{\partial\sigma_{\text{Sys}}}{\partial\lambda_k(\omega_s)} = \sum_m \alpha_m \frac{\partial\langle\lambda_m\rangle}{\partial\lambda_k(\omega_s)} + \frac{\partial}{\partial\lambda_k(\omega_s)}\log(Z),$$

as well as

$$\frac{\partial}{\partial\lambda_k(\omega_s)}\log(Z) = -\alpha_k p(\omega_s).$$

[9] This terminology is by no means standard, and that is why it is in quotations.

This follows from differentiating Z explicitly. Putting this all together, as we change a specific ν_j, everything changes by the following formula

$$\sum_k \sum_{\omega_s} \frac{\partial \sigma_{\text{Sys}}}{\partial \lambda_k(\omega_s)} \frac{\partial \lambda_k(\omega_s)}{\partial \nu_j} \delta\nu_j = \sum_m \alpha_m \sum_k \sum_{\omega_s} \frac{\partial \langle \lambda_m \rangle}{\partial \lambda_k(\omega_s)} \frac{\partial \lambda_k(\omega_s)}{\partial \nu_j} \delta\nu_j - \sum_k \alpha_k \sum_{\omega_s} p(\omega_s) \frac{\partial \lambda_k(\omega_s)}{\partial \nu_j} \delta\nu_j.$$

If we jointly consider all ν_j changing at the same time, as well as any changes in α_k, we can write the full expression for changes as

$$\boxed{\delta\sigma_{\text{Sys}} = \sum_m \alpha_m \delta\langle \lambda_m \rangle - \sum_j \sum_k \alpha_k \left\langle \frac{\partial \lambda_k}{\partial \nu_j} \right\rangle \delta\nu_j.} \tag{21}$$

The immediate consequence of the above formula for G is that

$$\boxed{\delta G = \sum_m \langle \lambda_m \rangle \delta\alpha_m + \sum_j \sum_k \alpha_k \left\langle \frac{\partial \lambda_k}{\partial \nu_j} \right\rangle \delta\nu_j.} \tag{22}$$

These formulas generally hold for the functions as defined as long as they obey the stated probability distribution in eq. (16), regardless of their physical interpretation. However, our next step will be to produce a physical interpretation.

3.2.3 A physical interpretation

We want to consider specifically changes in $\langle E \rangle$. For this purpose, suppose that $\lambda_0 = E$ and $\alpha_0 = \beta = 1/\tau$, and relabel sums over λ_k to not include $k = 0$. In this case, if we define $\mathcal{G} \equiv \tau G$, then the three resulting important equations that can be quickly derived are

$$\boxed{\mathcal{G} = \langle E \rangle + \sum_k \tau\alpha_k \langle \lambda_k \rangle - \tau\sigma_{\text{Sys}}.} \tag{23}$$

$$\boxed{\tau\delta\sigma_{\text{Sys}} = \delta\langle E \rangle + \sum_m \tau\alpha_m \delta\langle \lambda_m \rangle - \sum_j \left(\left\langle \frac{\partial E}{\partial \nu_j} \right\rangle + \sum_k \tau\alpha_k \left\langle \frac{\partial \lambda_k}{\partial \nu_j} \right\rangle \right) \delta\nu_j} \tag{24}$$

$$\boxed{\delta\mathcal{G} = -\sigma_{\text{Sys}}\delta\tau + \sum_m \langle \lambda_m \rangle \delta(\tau\alpha_m) + \sum_j \left(\left\langle \frac{\partial E}{\partial \nu_j} \right\rangle + \sum_k \tau\alpha_k \left\langle \frac{\partial \lambda_k}{\partial \nu_j} \right\rangle \right) \delta\nu_j} \tag{25}$$

The final term in the above equation is precisely the generalized definition of external work done on a system:

$$\boxed{\delta\mathcal{W}_{\text{Ext. on Sys}} \equiv \sum_j \left(\left\langle \frac{\partial E}{\partial \nu_j} \right\rangle + \sum_k \tau\alpha_k \left\langle \frac{\partial \lambda_k}{\partial \nu_j} \right\rangle \right) \delta\nu_j.} \tag{26}$$

A valid definition of external work Why is this a valid definition of external work? The first term with $\partial E/\partial \nu_j$ clearly resembles the external work mentioned before, and so is intuitively valid. As an example of why the second term can be interpreted as work, note that if λ_j is volume V, then $\tau\alpha_j$ is pressure ρ, and the final term becomes effectively the average $\rho\delta V$ that occurs as external parameters are altered, precisely what we would interpret as external work done on a system.

Why did we not use $\langle E \rangle$ to define external work? You might still be troubled thatwe didn't use $\langle E \rangle$ instead of $\mathcal{G}$ to measure external work done on the system. Why is this valid? Well, from eq. (24), $\langle E \rangle$ changes with the entropy of the system and with $\tau\alpha_m\delta\langle \lambda_m \rangle$, as well as with external work. Therefore, if we did some external work on the system, it could cause a change in the entropy of the system or $\tau\alpha_m\delta\langle \lambda_m \rangle$ and not change $\langle E \rangle$ at all. Therefore, $\langle E \rangle$ would not measure the work done on the system.

Here is an explicit, simplified example. Consider a system with two possible states, ω_a and ω_b. Assume each state only has one descriptive variable, the energy. Moreover, assume the initial energy is $E = 0$ for both states ($E_a = E_b = 0$). In this case, $\mathcal{G}_{\text{initial}} = -\tau \log(2)$ because $\sigma_{\text{Sys}} = \log(2)$. Moreover, $\langle E \rangle_{\text{initial}} = 0$, clearly. Next, change some external parameter ν so that the energy E_b of state ω_b slowly goes to infinity. Clearly, you are doing work on the system to raise the energy of this state, and in fact we have an explicit formula for this work. In the end, however, $\langle E \rangle_{\text{final}} = 0$ still because the exponential factor in the probability of being in state ω_b dominates the effect of E_b being large. Thus, $\Delta\langle E \rangle$ does not reflect the fact that we did external work on the system. However, $\mathcal{G}_{\text{final}} = 0$ because the entropy of the system goes to 0. Therefore, $\Delta\mathcal{G} = \tau \log(2) > 0$ reflects the work that we had to do on the system to affect this change in state ω_b, while $\Delta\langle E \rangle = 0$ does not.

Summary In one summary sentence: external work done on the system can be stored as $\langle E \rangle$, σ_{Sys}, or a whole host of other variables, and so we must consider all of these variables if we are trying to keep track of the total work done on the system. The generalized free energy $\mathcal{G}$ is the function that does this accounting for us.

3.3 Summary of generalized analysis of free energy and work done on a system

We have shown that, when there is an arbitrary set of variables λ_k that describe the system and are not controlled but instead have a statistical distribution, and an arbitrary set of variables ν_j that describe the system which are fixed and can be used to control the system, then

1. We can always define a generalized free energy function $\mathcal{G}$ as $\mathcal{G} = -\tau \log(Z)$, where Z is the full partition function of the system.[10]

2. This generalized free energy function $\mathcal{G}$ has the property that $\Delta\mathcal{G}$ measures the external work done on the system by changing any of the ν_j variables, if we also assume

   (a) that the system's parameters are changed slowly enough to remain at the equilibrium probability distribution,

   (b) and that each $\partial\sigma_{\text{Res}}/\partial\lambda_k$ remains constant for all k (effectively constant temperature, pressure, etc.).

   Therefore, in general $\mathcal{G}$ is a valid measure of the energy that can be extracted from the system using variables ν_j if we have no control over any λ_k. The precise definition of the most general form of external work is given in eq. (26).

3. Moreover, we have shown that $\mathcal{G}$ is intimately connected to the full entropy of the system (meaning the entropy of the full probability distribution of states for the arbitrary system), and contains a factor $\langle E \rangle - \tau\sigma_{\text{Sys}}$ as well as additional terms related to other λ_k. Work done on the system can be interpreted as being stored in the entropy of the distribution in some cases.

Overall, we have learned what entropy and free energy are, both in the context of information theory and in the context of statistical mechanics, and we are ready to examine exciting new applications of these theories.

[10] Generally textbooks define various types of free energy such as the Hemholtz free energy and Gibbs free energy that are valid in certain physical situations, but if you think about a concept of generalized free energy as defined here, then it is clear that Hemholtz and Gibbs free energy are simply individual manifestations of the exact same object.

4 Applications: non-equilibrium changes in biological and computational systems

The remainder of this document is designed to introduce you to interesting applications of entropy and free energy in a way that will test your understanding of the theoretical concepts described above. However, first we must discuss the concept of non-equilibrium changes, which are essentially changes that occur fast enough so that the system is not described by the equilibrium probability distribution. Generally, these types of non-equilibrium changes are common in real world systems.

Next, we will discuss two applications of the concepts of free energy, entropy, and work done on a system: non-equilibrium biological systems and the process of classical computation. Both of these situations benefit from the fact that we can view small biological systems or computational systems which are immersed in a bath of constant temperature and pressure as precise manifestations of the abstract systems in our previous analysis.

4.0.1 An aside on equilibrium distributions and non-equilibrium distributions

Probability distributions of the form $p(\omega_k) = \exp(-\beta E_k)/\mathcal{Z}$ which were derived above by considering changes in the entropy of the reservoir are called the "**equilibrium probability distribution**", and the system generally remains in this type of distribution as long as you change external parameters slowly enough. A change which maintains the equilibrium probability distribution is called a **reversible** change. A change that does not maintain this probability distribution is called **irreversible**. One important fact is that the definition of external work done on a system that we found above is only valid if the system remains in the equilibrium distribution.

To understand non-equilibrium changes, we must first understand what it physically means to be in an equilibrium of states. More generally, what is the physical meaning of having a probability distribution of states, rather than a distinct state? In reality, when we talk about a system having a probability distribution, the system actually *is* in one of the possible states that are available, and it quickly switches between that state and nearby states over time, doing so in a way that it spends time in each state that is proportional to the probability that we say it has of being in that state. Thus, when we say that we have a probability $p(\omega_i)$ of observing state ω_i, what we really mean is that the system spends about that proportion of its time in state ω_i, so that if we were to look at the system at any random given time, we would have a probability $p(\omega_i)$ of observing state ω_i.

If you change the external parameters of the system too quickly then the state of the system is now described by some **non-equilibrium probability distribution** (any probability distribution that does not have the form $p(\omega_k) = \exp(-\beta E_k)/\mathcal{Z}$ or the generalized analogous formula that we derived above). How does this happen? Well, the state of the system is initially some ω_i, or near some ω_i. When you start changing external parameters of the system quickly, the actual state of the system doesn't have time to change. Therefore, the final state of the system is preferentially more likely to be near the same ω_i when you finish altering the system (although ω_i might now have a different value of E or V). Lastly, this final distribution might not be an equilibrium distribution.

Why is this final state non-equilibrium? Consider a more concrete example. Imagine you had two states available, ω_0 of energy 0 and ω_1 which has very high energy. In equilibrium, the initial state is far more likely going to be ω_0. In fact, we can essentially assume that the initial state of the system is ω_0. However, if you very quickly change the parameters of the system then you are still in state ω_0 with very high probability, even if you cause ω_0 to have huge energy, and ω_1 to have 0 energy. Thus the actual probability distribution which has $p(\omega_0) \approx 1$ and $p(\omega_1) \approx 0$ is *not* the same as what the new equilibrium probability distribution should be based on the new energies of states ω_0 and ω_1.

What are the implications for work done on a system if you drive a non-equilibrium change or an equilibrium change? If the probability distribution is not the equilibrium distribution, then (on average) the external work done on the system as you change an external parameter will **not** be described by eq. (14). Rather, during a non-equilibrium change, the average work done on the system when changing parameter x

by δx is given by a slightly different formula:

$$\delta \mathcal{W}_{\text{non-equil.}} = \sum_i p_{\text{non-equil.}}(\omega_i) \frac{\partial E_i}{\partial x} \delta x.$$

Because the values of $p_{\text{non-equil.}}(\omega_i)$ are no longer taken from the equilibrium distribution, this value of external work may be different from the equilibrium external work in eq. (13). Therefore, our analysis above using changes in free energy $\mathcal{G}$ to measure the work done on a system will not be quite right. However, can this analysis still tell us something about the system? Surprisingly the answer is yes.

It is worth noting, however, that even during a non-equilibrium change of external parameters, the free energy is still well defined as a function of τ and $\{E_i\}$, given by eq. (9). We can therefore still talk about changes in $\mathcal{F}$ or $\mathcal{G}$ during non-equilibrium alterations of external parameters, and we will use this fact in Section 4.1.

4.1 Jarzynski's equality: a non-equilibrium work relation

In the paper "Nonequilibrium Equality for Free Energy Differences", C. Jarzynski, *Phys. Rev. Lett.* **78**, 2690 (1997), the author derives a relation between changes in free energy (an equilibrium concept) and the statistical distribution of work done on a system as you vary external parameters in a way that drives the system out of equilibrium.

4.1.1 Statement of Jarzynski's equality

Consider an equilibrium system. Assume you vary some external parameter λ which changes the associated energies of various possible states of the system. You start with the system in an equilibrium distribution and you vary λ quickly from a value λ_i to a value λ_f, thus potentially driving the system out of equilibrium. If you repeat this change from λ_i to λ_f many times, the work done on the system in the process will not be the same for every experiment. Why is this? When you change λ quickly, you are basically sampling the initial distribution of states for the system, picking a specific state to start with, and then changing that specific state depending on how it is altered by λ. This argument was given in Section 4.0.1. If you happen to start the system in state ω_i, the parameters of the actual state ω_i of the system such as E_i and V_i will take a certain path to their final value and you will perform some specific amount of work on the system. However, if you start in a different ω_j, the work applied over time may be something entirely different (for example, there may be some possible starting states that are not altered at all by λ). This is why there is a statistical distribution for the amount of work that you will do on the system as you change λ quickly. What can we say about this distribution of work, in general?

The result derived by C. Jarzynski, called Jarzynski's equality, which tells us about the statistical distribution of the observed, actual values of external work done on the system, or $\mathcal{W}_{\text{Actual, ext. on Sys}}$, is

$$\boxed{\langle \exp(-\beta \mathcal{W}_{\text{Actual, ext. on Sys}}) \rangle_{\lambda_i \to \lambda_f} = \exp(-\beta \Delta \mathcal{G}).} \tag{27}$$

The averaging symbol $\langle \cdot \rangle_{\lambda_i \to \lambda_f}$ means averaging $\cdot$ over many repeated experiments starting from an equilibrium distribution, where you start with the same initial probability distribution and then change λ in the same way over time. The function $\mathcal{G}$ referenced here is defined to be $-\tau \log(\mathcal{Z})$ for the partition function $\mathcal{Z}$. In different physical situations $\mathcal{Z}$ will depend on different variables such as temperature, pressure, and the distribution of energies for each state, and will most generally be given by eq. (17). The exact interpretation will depend on the context, although the remarkable result is that Jarzynski's equality holds regardless. Although this paper's derivation is interesting from a theoretical standpoint, we do not have the time to analyze it in detail here. It is provided to you for reference should you choose to look at it. Instead we will take eq. (27) as true and consider its implications.

Questions First, look up the mathematical concept of Jensen's inequality, which will allow us to relate $\mathcal{W}_{\text{Actual, ext. on Sys}}$ directly to $\Delta \mathcal{F}$. Explain why Jensen's equality implies that

$$\boxed{\langle \mathcal{W}_{\text{Actual, ext. on Sys}} \rangle_{\lambda_i \to \lambda_f} \geq \Delta \mathcal{F}.} \tag{28}$$

Next, remember that $\mathcal{F}$ measures the work that you can extract back out of the system through a reversible process, and so we can say that $\Delta\mathcal{F} = \mathcal{W}_{\text{Reversible, ext. on Sys}}$, or the reversible external work done on a system. Thus we can define something called dissipated work,

$$\mathcal{W}_{\text{Dis}} = \mathcal{W}_{\text{Actual, ext. on Sys}} - \mathcal{W}_{\text{Reversible, ext. on Sys}}.$$

With this definition, eq. (28) implies that

$$\langle \mathcal{W}_{\text{Dis}} \rangle_{\lambda_i \to \lambda_f} \geq 0 \quad \text{and} \quad \langle \exp(-\beta \mathcal{W}_{\text{Dis}}) \rangle_{\lambda_i \to \lambda_f} = 1 \tag{29}$$

Try to explain why the result quoted above makes intuitive sense based on your real world experience, both for the case of positive external work and negative external work. Your explanation does not need to be mathematically rigorous; we just want to see how you interpret the above result in terms of the real world.

4.1.2 Experimental test of Jarzynski's equality

Equation (28) roughly states that energy is dissipated on average during non-equilibrium changes. Not surprisingly, this fact was known empirically for a long time before the more precise eq. (27) was stated by Jarzynski. Naturally, when eq. (27) was theoretically proposed, it became necessary to test its implications experimentally to ensure that its predictions did not conflict with experiment. We will consider one paper which explores Jarzynski's equality experimentally: "Equilibrium Information from Nonequilibrium Measurements in an Experimental Test of Jarzynski's Equality", J. Liphardt *et al.*, *Science* **296**, 5574 (2002).

Questions Here you are asked to read the above paper by J. Liphardt *et al.* and answer a few key questions about its results. We suggest that you skim the paper once, focusing on general concepts and on understanding the data in the various figures, but *moving on* if a concept takes too much of your time to understand, or if it just seems downright unrelated to anything we've talked about in this document (we won't ask you questions about every detail of the paper). After this initial reading, you should try to answer specific questions posed below by rereading specific sections of the paper in depth.

Please answer the following questions which roughly follow the chronological order of the paper. The questions generally do not focus on minute details that require an understanding of biology, and neither should your answers. Focus on interpreting results in terms of concepts introduced in this document.

1. Describe in your own words the experimental setup. What are the physical parts involved?

2. What can you consider to be the reservoir for this system?

3. How do the experimenters measure work done on the system?

4. How do the experimenters measure the actual change in free energy, $\Delta\mathcal{G}$, between the initial and final states? (Note that this paper uses the Gibbs Free energy in Jarzynski's equality, which is valid because their physical system is coupled to a bath of constant temperature and pressure.) As you have probably realized, computing $\mathcal{G} = -\tau \log(\mathcal{Z})$ from the probability distribution would be nearly impossible in all but the simplest situations. Hint: read through page 2 of the paper if you cannot figure this out and look for the symbol $\Delta\mathcal{G}$.

5. How do the experimenters test Jarzynski's equality? Remember that the experimenters have one objective measure of $\Delta\mathcal{G}$ using the method you found above. Jarzynski's equality also provides an estimate of $\Delta\mathcal{G}$ by averaging something (what is it?) over multiple experiments. What do the experimenters do with these two, potentially different values for $\Delta\mathcal{G}$ that they obtain?

6. What is different about the experiments in Fig. 2(a) that generate the red curves from the experiments that generate the blue curves? Explain why in Fig. 2(a) the two blue curves lie almost on top of each other everywhere while the two red curves do not.

7. Consider Fig. 3(a). What is being plotted? You should refer to the caption. Why does the estimate of $\Delta\mathcal{G}$ from Jarzynski's equality underestimate the actual value of $\Delta\mathcal{G}$? You may find it helpful to reread the bottom two paragraphs of the left column of the page containing Fig. 3.

8. Consider Fig. 3(b). What is the main difference between this data and the data from Fig. 3(a)? Focus on the dashed lines and solid lines. What does it mean that the solid lines are higher than 0 on the graph? What does it mean that the dashed lines are approximately 0 for any extension length? This result is, more or less, the most important result of the paper.

9. Consider Fig. 3(c), Fig. 3(d), and Fig. 3(e). What is plotted is the probability of obtaining certain values of dissipated work, or actual work observed on the system minus the change in free energy. Fig. 3(c) shows the probability distribution after you have extended from 0 nm to 5 nm, while (d) shows the probability distribution after extension from 0 nm to 15 nm, and (e) after extension from 0 nm to 25 nm. The blue, green, and red data sets represent probability distributions for different switching rates (speeds at which the system is altered). Blue corresponds to the blue lines in Fig. 3(a), and red and green correspond to the red and green data in Fig. 3(b).

   Is the switching rate for the blue data set faster or slower than the switching rate for the red data set? Can you explain why it makes sense that the average dissipated work in the blue data set is centered around 0?

   Focus on the red data set in Fig. 3(e). This is a probability distribution with a relatively large spread and an average that is positive. From Jarzynski's equality, we showed that eq. (29), or

$$\langle \exp(-\beta \mathcal{W}_{\text{Dis}}) \rangle_{\lambda_i \to \lambda_f} = 1$$

   should hold true. Why is it that a distribution like the red distribution with a positive average but large spread can still satisfy this equation? To answer this, you might want to consider which data points are weighted more in computing this average than others. You do not have to be mathematically rigorous.

After answering these questions, you hopefully understand what experiment was performed in this paper and how the data can be interpreted to be a valid test of Jarzynski's equality. The end result of this paper is that Jarzynski's equality is confirmed within experimental error. We can therefore accurately predict changes in the equilibrium free energy of a system (which is also the work we can possibly reliably extract from the system) by performing multiple, fast, non-equilibrium experiments and averaging these results according to Jarzynski's equality.

Thus, in a broader sense, we have learned that highly abstract mathematical frameworks can actually be useful for predicting concrete results in physical systems in the world around us. Moreover, these results hold for systems under very few idealizing assumptions. Indeed, the system does not need to be infinitely small, the bath does not need to be infinitely large, and we don't have to change the parameters of the system infinitely slowly.

4.2 Dissipation in computational systems

Perhaps surprisingly, the theoretical results derived earlier in this paper also allow us to explore the thermodynamic limits of computation. In effect, we want to analyze how much energy we dissipate as we perform a computation on a computer, if any. In order to analyze this situation, we must first define a theoretical model of a computing system, meaning define what its logical components are, and we must specify how the computer fits into the framework of a system and reservoir that is used throughout this paper. We can then consider how its free energy changes as a computation is performed, and consider how much energy is dissipated.

4.2.1 Reversible or irreversible computation?

As has been discussed above, when a system's state changes reversibly, the least amount of energy is dissipated (on average) that is possible. However, you often want a system to compute something quickly, in which case you want to drive the system to change quickly, and this can push the system out of equilibrium, leading to more dissipated energy on average.

While the question "how much energy am I willing to dissipate just to have a faster computation?" is an important question, we will not concern ourselves with it here. Rather, we will try to determine limits

on the energy dissipated during a computation which is performed slowly enough so that the system always remains close to the equilibrium probability distribution.

4.2.2 The AND gate as a logical operation

Rather than analyze an entire computer as a system all at once, we will analyze its smallest components. One of the simplest digital computing functions is to take two bits, which are units of information that can be either 0 (false) or 1 (true), and perform the AND operation. This operation outputs 1 *only* if both inputs are 1, and outputs 0 otherwise. Thus, $(0,1) \to (0)$, $(1,0) \to (0)$, $(0,0) \to (0)$, and $(1,1) \to (1)$.

It turns out that most modern computers can be built entirely from AND gates and a few other fundamental building blocks, such as OR gates and NOT gates (which do exactly what they sound like they might do. Feel free to look them up). However, because the AND gate is widely used, we will focus on the AND gate in the discussion below, and leave it to the reader's imagination to generalize these concepts further.

4.2.3 A physical model

We can imagine an AND gate as a little machine that takes two logical inputs and produces one logical output. There are many different ways to implement this machine in the physical world, just as there are many ways to represent 1 and 0, and so it becomes difficult to say something generally meaningful about the energy costs of an AND gate. However, we can say something meaningful if we translate the concept of the AND gate into the language of a system coupled to a reservoir with a statistical distribution of energies.

Our conceptual model is this: imagine a system with two "ports", port A and port B. These ports can each take on the value 0 or 1. To compute the logical operation of an AND gate on two bits, you load the input bits into port A and port B. Next, the computation is applied to the two bits, and then the answer is displayed on port A, leaving the other port B to randomly be either 0 or 1.

How do we compare this to a system and reservoir model? We consider the two bits to be the system which is coupled to an energy reservoir, a computer agent, and a human agent. This system has 4 possible states, corresponding to $(0,0)$, $(1,1)$, $(1,0)$, and $(0,1)$. There are two external actors which interact with the system: the human agent that sets the initial values of the gates, and the computational agent which changes the initial state of the system to the final state of the system according to the logical rules of the AND gate. The initializer agent sets the initial values, steps back, and lets the computer interact with the system to put it into its final state, which can then be observed. While there may be other ways of conceiving of a computer, the physical results below are best understood with this model.

4.2.4 Operation of the AND gate from the perspective of statistical mechanics

This section goes through the sequence of steps that a computation consists of, and determines the value of various statistical mechanical quantities during each step. Ultimately, this allows us to derive the amount of energy dissipated when performing a single computation.

Questions Assume that the system has exactly 4 possible states for port A and B: (0,0), (0,1), (1,0), and (1,1). We will call these states $\omega_1, \omega_2, \omega_3$, and ω_4 respectively. We imagine that all of these states have some energy E_i, and that the states are populated as usual in an equilibrium probability distribution, according to the factor $\exp(-\beta E_i)/\mathcal{Z}$. Imagine that all of these states start out with the same energy $E_i = \epsilon$. This results in an initial value of the free energy which we will call $\mathcal{F}_{\text{Before init.}} = \langle E \rangle - \tau\sigma_{\text{Sys}}$. What is this value?

Next, as the external agent who initializes the ports, we choose to load a certain starting state into the computer. For example, to load $\omega_2 = (0,1)$, we raise all E_i with $i \neq 2$ to $E_i \sim \infty$. If we do this, what is the resulting value of $\langle E \rangle$? Hint: the limit as $x \to \infty$ of xe^{-x} is 0. What is the resulting value of σ_{Sys}? Therefore, what is the resulting value of $\mathcal{F}$? Call this value $\mathcal{F}_{\text{Before comp.}}$. How much work did we have to do to cause this change? To be clear, we as the external agent that initializes the ports are doing this work.

At this point, we step back from the system and we let the computer perform a computation on the system. This operation is a deterministic map from initial to final configurations of the system according to the logical operation of an AND gate. For example, because we are considering an AND gate, if we were in $\omega_2 = (0,1)$ initially, the final state has port A = 0 and port B undetermined. Therefore, the final state is either $\omega_1 = (0,0)$ or $\omega_2(0,1)$, with equal probability. The computer is considered as an external agent

distinct from the system, so from the perspective of the system, the computer does work on the system and changes E_i in some way, just as was done by the human agent during the initialization process. In this interpretation, the computer must have brought E_1 down from $E_1 \sim \infty$ to $E_1 = \epsilon$, so that ω_1 and ω_2 were equally likely, while ω_0 and ω_3 are impossible. When changing this energy, the computer did work on the system. Was this positive or negative work? You can find out by computing the change in free energy. Call the new free energy after this change $\mathcal{F}_{\text{After comp.}}$.

Once we have computed these various free energy changes, we need only keep track of where work is done and where energy is dissipated during the process of computation. The next section will discuss this process in more detail.

4.2.5 Where is energy dissipated?

During the procedure of computation, we do work on the system to start it in one of the computational basis states. We step back, and then the computer steps in and exchanges energy with the system. The system is then in its final state. The work we do to initialize the system is $\mathcal{F}_{\text{Before comp.}} - \mathcal{F}_{\text{Before init.}} = \tau \log(4) = 2\tau \log(2)$. The work that the computer does when it exchanges energy with the system is $\mathcal{F}_{\text{After comp.}} - \mathcal{F}_{\text{Before comp.}} = -\tau \log(2)$. Note that this work done by the computer on the system is negative so that the computer extracts energy from the system. Lastly, if we wanted to reinitialize the system into a different ω_i after finishing a computation, it is not hard to show that we would have to do an additional amount of work given by $\mathcal{F}_{\text{Before comp.}} - \mathcal{F}_{\text{After comp.}} = \tau \log(2)$. From this we see that, in the course of one computation where the system goes from $\mathcal{F}_{\text{After comp.}}$ to $\mathcal{F}_{\text{Before comp.}}$ to $\mathcal{F}_{\text{After comp.}}$, we put in $\tau \log(2)$ worth of energy to the system, and then the computer takes out $\tau \log(2)$ worth of energy. It is not clear exactly where this energy goes after the computer extracts it, but the point is that we as the initializers no longer control the energy, and we no longer can get it back out of the system directly.[11] This loss of energy means that computations with an AND gate costs us as the human initializer $\tau \log(2)$ worth of energy on average per computational cycle.

Questions It is common to look at the above energy cost of computation and call it "irreversible computation" because of the energy cost associated. Why is this potentially a misnomer? Is the energy really dissipated, or is it just stored in a different system (like the computer)? Remember that we are still assuming that we are operating slowly enough so that the system remains in equilibrium with the reservoir at all times.

It is important to realize that the computer itself is a system coupled to a reservoir. In real physical computers, the energy $\tau \log(2)$ is extracted during a computation and resides in the core of the computer. The physical assumption is then that the energy extracted into the computer is eventually dissipated as heat loss, and is therefore well and truly irrecoverable.

4.2.6 Generalizing the computational model

It may have occurred to you that having only 4 states (not 4 logic states but 4 states total) available for the system isn't realistic.

Question Consider if, instead of 4 states, there were 40 states of the system possible, all with the same initial energy. Further assume that 10 of these states correspond to each of the 4 combinations of port A and port B. In this case, we can still perform computation by changing the energy of each set of 10 states at once. Argue why this does not change the essential conclusion that the energy cost of one sequence of computation is $\tau \log(2)$. You will have to consider the same set of steps that we used above to originally derive the $\tau \log(2)$ loss of energy per computation.

Additional information for your own interest This analysis can be extended further using the concept of relative entropy to show that $\delta\sigma_{\text{Sys}} = \delta\sigma_{\text{Logic gates}}$ (meaning that the entropy of the whole system changes

[11] Perhaps there is a conceivable way of getting the energy back out of the computer, but we will not consider that here.

with the entropy of the logical output of the system) as long as a few small constraints on the system's energies are made.[12]

4.2.7 Computing without the cost

In the case described above, although we described the process of computation through changing the energies of the various states, E_i, note that the average energy of the system, $\langle E \rangle$, never changed. Therefore, the change in free energy is $-\tau\sigma_{\text{Sys}}$ which is quite literally the energy available to be extracted from the system. Once again, the entropy of the system seems to be physically relevant.

This observation suggest that, if we can design a computer that does not change the entropy of the system as it changes the input state of the system to the output state, we can perform computations without the associated energy cost. If the free energy does not change, then the average work we, or the computer, performed on the system was 0.

Questions Consider the case where the mapping of the computer between initial and final logic gates is 1 to 1. That is, consider a logical operation that is not the AND gate. This means that for every input logic state there is exactly one possible output logic state, and the reverse. Two examples of such gates for two bit operations are the NOT gate and the identity gate. You can also look up a gate called to Toffoli gate for another example, although this gate uses more computational bits.

If the computer is programmed to carry out this type of gate operation, what will be the change in the entropy of the system from before the computation to after the computation? Assume that we keep average energy fixed still, just as before. Does repeatedly performing this operation cost any energy on the part of the observer? Does the computer extract any energy as it performs its computation? Explain why. These types of logic gates are often referred to as "reversible" gates because the computation itself, if performed in near equilibrium, does not extract any energy on average from the system.

Perform a little research into modern computational systems to answer the question "how close to the thermodynamic limits of computation are modern computers?" In particular, try to find information about modern physical implementations of the AND gate, and determine the energy dissipated in one logical operation. Is it close to $\tau \log(2)$ or is it much larger? You may have to assume a certain temperature of the system and use the fact that $\tau = k_B T$, where k_B is Boltzman's constant and T is the temperature in Kelvin (really temperature should be measured in units of energy, but historical reasons ensure that it is not). With these results, can you argue whether classical computing systems are limited by the theoretical thermodynamic limits of computation, or rather by inherent inefficiencies in the physical implementations of computing architectures.

[12] One possible set of constraints is derived by decomposing the entropy of the states into the entropy of the logical output ports plus the relative entropy of the states given the value of the output ports. The assumptions that must then be made are assumptions to ensure that the the relative entropy term is constant, so that changes in the full entropy of the system are precisely changes in the entropy of the logical output ports.