Chapter 2
Statistical Analysis in Solution Space

> " 'You see,' he exclaimed, 'I consider that a man's brain originally is like a little empty attic, and you have to stock it with such furniture as you choose ... It is a mistake to think that that little room has elastic walls and can distend to any extent. Depend upon it, there comes a time when for every addition of knowledge you forget something that you knew before. It is of the highest importance, therefore, not to have useless facts elbowing out the useful ones.' "
>
> *Sherlock Holmes* [Arthur Conan Doyle, A Study in Scarlet]

In the solution of optimization problems, many factors act in concert to achieve the cumulative effect that we measure using a single cost function. We are dealing with finding a particular microscopic arrangement of many constituent parts – called a *microstate* – in order to attain a desired macroscopic result – called the *macrostate*.

Suppose that you are in a room. This room has many molecules of air that move around in the room. The knowledge of the positions and momenta of all these molecules is the microstate of the room. The macrostate is comprised of a few parameters of interest to you, such as the temperature and pressure of the air. If you were to move a single molecule from one side of the room to the other, would the temperature in the room change perceptibly? No. This observation means that (1) even though a particular microstate leads to a particular macrostate, (2) any one macrostate can potentially be achieved by more than one microstate. The relationship between microstate and macrostate is thus not a one-to-one relationship. By analogy to maps we have one altitude for a specified location but possibly several locations for one specified altitude; as such the location is the microstate and the altitude the macrostate.

The same observation holds true for optimization problems: A particular value for the cost function is usually achieved with many settings of the process parameters. *The optimum state is an exception and is often achieved using only one parameter setting just as the altitude of 8850 meters is achieved only in one location, namely Mount Everest.* In the analysis of the relationship between microstates and macrostates, the analogy to the molecules in the room applies.

P. Bangert (ed.), *Optimization for Industrial Problems*,
DOI 10.1007/978-3-642-24974-7_2, © Springer-Verlag Berlin Heidelberg 2012

As this problem was first investigated by physicists in the context of thermodynamics, the language of the theory uses vocabulary that is reminiscent of thermodynamic processes. This should not be misunderstood as the suggestion that optimization problems are thermodynamic. They are not. The theory that governs thermodynamic processes is, however, so general that it can easily encompass our situation of optimization problems.

The relevant field of physics is called *statistical mechanics*. It derives its name from the fact that the macrostate is essentially a statistical summary of the microstate just as the mean, or average, is a statistical summary of a set of numbers.

In this chapter, we will treat the relationship between microstate and macrostate as developed in statistical mechanics. The vocabulary of thermodynamics will be retained but the ideas will be made sufficiently general that it will become clear how they apply to our situation. For the purposes of this chapter, please suspend any ideas of optimizing. First, we must become clear about how the state of the problem relates to the cost function or, in other words, we must first understand the problem that we are faced with and the answer we desire. Only when this relationship is clear, are we permitted to ask what the state of the problem is that corresponds to a minimum in the cost function.

2.1 Basic Vocabulary of Statistical Mechanics

The energy of a physical system is essentially the same as the cost function in optimization in that nature seeks the configuration of least energy. To understand this from the physical perspective, we quote a description of the concept of *energy* here:

> "Consider a volume of water stationary in a pool at the head of a waterfall. It has what we may call 'privilege of position,' in that once it has dropped over the fall we must do work to return it to its original position. As the water passes over the fall its 'privilege of position' vanishes, but at the same time it acquires *vis viva*, the 'living force' of motion. By passing the water through a turbodynamo, we strip it of its *vis viva* and simultaneously acquire electric power which, vanishing when the dynamo is shorted through a resistance, there gives rise to an evolution of heat. If the water drops directly to the bottom of the fall, without passing through the turbine, *vis viva* disappears without the production of electric power; but at the bottom of the fall the water has a temperature slightly higher than that with which it left the top of the fall – just as though it had received the heat from the above-noted resistor. Now *a priori* there is *no* reason to suppose that 'privilege of motion,' *vis viva*, electric power, and heat – qualitatively apparently utterly different – stand in an relation whatever to each other. Experience, however, teaches us to regard them all as diverse manifestations of a single fundamental potency: energy (Gr. *energos*, active; from *en*, in + *ergon*, work)." [98]

Let us consider a particular instance of an optimization problem. For definiteness, consider a particular instance of the traveling salesman problem. The number of cities and the distances between each city pair is known.

A *microstate* is a complete detailed description of any arrangement of the most basic elements of the problem such that no boundary conditions are violated. Any

microstate is thus a *solution* of the problem instance. In the context of the traveling salesman, any ordering of the cities, without repetition, is a microstate and thus a solution in the sense that all such orderings are legal traveling salesman tours. Remember that we are not optimizing yet, we are just describing the problem. If you had an ordering of the cities in which a particular city featured more than once or a city was missing, then this would violate a boundary condition of the problem and thus not be a microstate or solution. In terms of mathematics, a microstate can be expressed as a vector.

A *macrostate* is a global description of a microstate in terms of all the functions that we will later use to optimize the solution. In most optimization contexts the macrostate is the value of the cost function and thus a single number. For the traveling salesman, the macrostate is the total length of the tour.

A *system* is the instance of the problem viewed as an evolutionary entity that changes in time. Mathematically speaking, a system is a series of microstates ordered in time. In the context of thermodynamics, the microstate of the molecules in a room will change from moment to moment in accordance with the laws of physics. In the context of optimization, the microstate of the traveling salesman problem will change from one step of the optimization procedure to the next. In both cases, there is a mechanism of evolution (physical laws or an optimization algorithm) that causes a time-ordered sequence of different microstates. Accumulated from some start time to some end time, this is referred to as a system.

When we have a system, we can take an average of the macrostates over time. That is from the start time to the end time of the system, we select a certain number of macrostates evenly spaced in time and perform an average. The result is called the *time-average* of the system.

Consider again a particular instance of a problem. Imagine now having many copies of this instance. Each copy is put into a random microstate; many will be different from each other but some may be the same. We shall have something to say about the meaning of the word ‘random’ but will delay it a little. To get a mental picture of this, imagine that the problem consists of a room full of molecules. Now imagine that you have a great many rooms. All the rooms are identical to each other in every aspect except that their microstates – the positions and momenta of the molecules – may be different; as a logical consequence their macrostates may also be different. Each of these copies now evolves over time and thus we have a set of systems. This set of systems is called an *ensemble*. The concept of an ensemble is very important in the treatment of statistical mechanics and thus in our views on the relationship between microstates and macrostates. Please note that we are never going to actually construct an ensemble as this would require too many resources and thus be a practical impossibility. We are just going to consider the existence of an ensemble as a thought-experiment.

At any instant in time, we may record the macrostate of each copy in an ensemble and perform an average over these values. This is called the *ensemble-average*. We can take an ensemble-average at any moment in time including the start time and the end time of the systems in the ensemble. If the value of the ensemble-average does not change with the time at which the average is taken (with the possible exception

of some initial time period), the ensemble is called *stationary*. Physically, this is usually called *equilibrium*. Note that if an ensemble is stationary, the many possible ensemble-averages differing due to their start and end times all take the same value and thus there is in fact only one ensemble-average value. Stationarity is thus a crucial concept for us to speak of *the* ensemble-average as opposed *an* ensemble-average.

Having discussed two averaging procedures, the time and ensemble averages, it is interesting to look at how they differ. In both averages, we list the macrostates of a large number of microstates and perform an average. If the number of microstates is sufficiently large, then the averaging process itself should be stable and the results represent truly underlying differences. In the time case, the microstates are connected by the evolutionary laws of the process (physics or an optimization algorithm). In the ensemble case, the microstates are connected by their initial selection and then their evolution according to the same laws. If an ensemble is stationary and the ensemble-average is equal to the time-average, then the ensemble is called *ergodic*.

To be clear, ergodicity is a good thing. We like ergodic ensembles. Situations where ergodicity is not valid are generally very hairy indeed. The reason for ergodicity being desirable is that if an ensemble is ergodic, we can replace the time-average by the ensemble-average in any mathematics that we will want to do. This is an elemental difference due to the fact that computing a time-average would require the solution of the time-dependent partial differential equations that govern the evolutionary laws of the process. We do not like doing this. Respectively, in many situation we cannot do this. Computing the ensemble-average is relatively easy due to the fact that the individual copies are randomly assigned a microstate and the evolution in time does not play a role (the ensemble is stationary). To perform such an average, we merely need to generate a lot of random microstates, take our average and the deed is done. Computationally speaking, we actually create these many microstates in the computer. Doing this, including the ensuing taking of the average, is a simple presentation of a collection of techniques commonly called *Monte-Carlo computation*. As before, we delay the definition of the word 'random.'

In keeping with the language of statistical mechanics, we are going to use the word *energy* as a synonym for the objective or cost function of the optimization. Physics is effectively one big optimization problem as physics postulates that nature always evolves in order to minimize its energy. Recalling our above definitions, energy is effectively the number representing the macrostate. As every microstate has one corresponding macrostate, we can associate an energy with each microstate.

At this point in the discussion, we are going to create our first basic assumption, namely: The number of possible microstates is finite. Please note that in general the number of microstates is very very large but we demand that it not be infinite. This is important because we want to start counting how many microstates belong to any given macrostate and we want these numbers to be finite so that we can do arithmetic with them. As the number of microstates is finite, we can label them with an integer. The order does not matter for this purpose. We will denote the energy of microstate i by E_i.

The only thing left in the presentation is to be clear about the term 'random.' We will make our second basic assumption: The probability of the system being in any one microstate is equal to that of any other microstate. If there are N microstates in total, then the probability of the system being in microstate i is $P_i' = 1/N$. It is now easy to create an ensemble. We simply select microstates from the set of all microstates each with probability $1/N$. Due to this procedure, we will get an ensemble, we will be able to compute an ensemble average and, if the ensemble is ergodic, this will be equal to the time average and thus give us something interesting.

The probability of the microstate was thus settled by assumption. But what is the probability of the associated macrostate? Well it is simply the number of microstates associated with this macrostate divided by the total number of microstates, $P_i = N_i/N$. While this is an easy formula, it is far from easy to work it out as we, in general, will be hard pressed to compute N_i. Thus, we must find a formulation that is easier to compute.

To discover this, we first talk about temperature. Going back to the physical case of the room full of molecules, we note that this room does not actually exist in isolation but rather it is part of the world and exchanges energy with the world. After some time, so our experience tells us, the temperature in the room will equal the temperature of the world. In statistical mechanics, the world is therefore referred to as a *heat bath*. The concept of *temperature* enters our discussion here as a crucial parameter that is supplied by the external forces that act upon our system; see section 2.4 for this concept. Also, we will assume that we know what the temperature is because we can measure it in the heat bath. We will find that the concept of temperature will play a major role in our later optimization efforts. It should be understood however again that while we are using vocabulary from statistical mechanics, the concepts are much more general and can be applied to non-physical systems. Temperature, for example, is just a macroscopic parameter of the system supplied by the external heat bath forces that govern the system evolution.

Now that we know what temperature is, in statistical mechanics, it is possible to derive what P_i is actually equal to. We will not follow the derivation here as we are concerned only with the interpretation of these results. We have what is called the *Maxwell-Boltzmann distribution*,

$$P_i = \frac{g_i e^{-E_i/kT}}{\sum_j g_j e^{-E_j/kT}} \tag{2.1}$$

where T is the temperature, k a constant known as the *Boltzmann constant* and g_i is the *occupation number* of the energy E_i, i.e. the number of microstates having energy E_i. The denominator of the distribution is referred to as the *partition function* and serves several important uses in statistical mechanics to the extent that complete knowledge of the partition function essentially means complete knowledge about the system – at least with regard to all the things that physics is usually interested in, i.e. the macroscopic description of the system. The partition function cannot practically be evaluated as defined because it is a sum over all microstates and the number of microstates is very large indeed. Supposing that we could write the partition function

in a way to be able to directly evaluate it, we could perform ensemble-averages and thus time-averages.

With the partition function

$$Z = \sum_j g_j e^{-E_j/kT}$$

we may find a number of other crucial thermodynamic concepts such as the energy E, the entropy S or the Helmholtz free energy A,

$$E = -\left[\frac{d\ln Z}{d\beta}\right]_v = kT^2\left[\frac{d\ln Z}{dT}\right]_v,$$

$$S = \frac{d}{dT}\left[kT\ln Z\right]_v,$$

$$A = E - TS = -kT\ln Z$$

where $\beta = 1/(kT)$ and the subscript v indicates that the derivative is to be taken at constant volume. Thus, our knowledge of thermodynamic properties of a particular system is limited by our ability to compute its partition function!

Note that if two microstates have the same energy, their contributions to the sum in the partition function are the same. This is a desirable property as two microstates of the same energy would belong to the same macrostate and should therefore be, macroscopically speaking, indistinguishable. Therefore, it is good that their microscopic contributions are the same.

While the Maxwell-Boltzmann distribution works very well for certain systems in the physical world, it is not necessarily true for all physical systems or indeed for non-physical systems. An abstract optimization problem can be profitably analyzed using the language of statistical mechanics but we must remember which conclusions of statistical mechanics are of a generic nature and which apply particularly to specific elements of physical nature.

2.2 Postulates of the Theory

It is possible to build up statistical mechanics as a formal theory based on axioms. This fact is important beyond making the theory formally clean because it shows the fact that the theory is very generic and applies to many situations that have nothing whatsoever to do with thermodynamics. The postulates are these six [78]:

1. The constituents of the system obey certain laws of motion that themselves do not change. In the thermodynamic context, these are classical or quantum laws of physics. In the industrial context it will usually be classical physics only.
2. An observation is a simultaneous and instantaneous measurement of a set of indicators, which each take the value zero or one only. The instants at which these observations may be made are discrete and equally spaced.

3. Observations cause no visible disturbances in the macrostate of the system under observation[1].
4. The successive observational states as given by the indicator values form a Markov chain.
5. Any microstate may be the initial state.
6. A system with finite energy has finitely many microstates available to it.

All of these postulates are quite clear and simple for a problem that is well defined. The interesting postulate is the fourth concerning the Markov chain. Effectively this means that the system has no memory of its previous states as a Markov chain is defined by a transition probability matrix in which the probability of the future state depends only upon the identity of the current state and not the past states. A Markov process is thus a probabilitistic (stochastic) process without memory.

It is clear that this assumption is not strictly true about every system in nature but it is close enough to being true that the theory leads to interesting results about nature. As we are concerned with optimization in this book, however, we need to ask ourselves whether the optimization method will respect this and the other postulates. If it does, then the following theory will apply to the analysis of its results. In general the methods that are used for optimization do respect the Markov postulate and we may thus proceed.

Through this postulate, we effectively take the probability of a certain observational state to be an intrinsic characteristic and we implicitly assume that this is measurable for example by repeating an experiment several times[2]. The probability of an observational state thus takes on an ontological value similar to that of an object's mass in classical physics.

While the microstates obey physical laws and are thus deterministic[3], the observational state, i.e. the macrostate given by the indicators, does not change deterministically. We want a reliable statistical description of its evolution – that is the purpose of the theory.

Why is that? Note that it is very complex to observe the microstate at all times and to model it deterministically. It is far more expedient to model the macrostate because it has few parameters and we are interested in it. Due to the fact that it does not change deterministically, all we can expect to receive therefore is a statistical description of the macrostate. In order words, while we can say that a certain microstate will *definitely* transit into another specific microstate, we can only say this *with a certain probability* in the context of a macrostate.

From these postulates, it is possible to derive all other statements in statistical mechanics including the famous four laws of thermodynamics, which are the following.

---

[1] From quantum theory we know that this fundamentally wrong but we are dealing with systems far larger than quantum systems and we may thus reasonably assume this. Please do note that there is a large debate about the role of the observer on a physical system and that this point is an assumption and not a statement of fact.

[2] The concept of probability will be further discussed in chapter 5.

[3] We will not concern ourselves with the nature of determinism in quantum mechanics as we are not dealing with the application of this theory to physical situations.

0 If systems A and B are each in equilibrium with system C, then A and B are also in equilibrium with each other.
1 The energy of an isolated system is conserved.
2 The entropy of an isolated system may not decrease.
3 A system cannot be brought to zero absolute temperature.

The concept of energy was already explained. In the following, we will explain entropy and temperature in more detail.

2.3 Entropy

Having covered some basic concepts, we will turn to one of the most central concepts of statistical mechanics, the *entropy* of a system.

> "The first principle of thermodynamics poses the concept of 'energy'; the second principle, the concept 'entropy.' Feeling that we know what energy *is*, we demand to know what entropy *is*. But now, in point of fact, *do* we really know what energy is? The classical dichotomy is matter vs. energy, and energy may then be defined as whatever produces heat. But in the early 20th century this dichotomy was undermined by recognition of the interconvertibility of mass and energy, and to the question 'What is energy?' we can now give only the unsatisfactory reply 'It is *everything*.' Yet however great may be our uncertainty about the intrinsic nature of energy, the thermodynamic significance of that concept remains wholly unimpaired. ... Indeed we don't need to know what energy is, but we do find it satisfying and instructive to use the kinetic-molecular theory to *interpret* internal energy in terms of the kinetic and potential energies of atoms and molecules. Neither need we know what entropy is, but we find it satisfying and instructive to use the kinetic-molecular hypothesis to *interpret* entropy in terms of the 'randomness' with which atoms and molecules are distributed in space and in energy states. A simple illustration of the subtle concept of randomness is found in the ... example of a bullet abruptly stopped by a sheet of armor plate. The bullet's gross kinetic energy disappears, and in its place appears thermal energy that manifests itself in a rise of temperature. *Before* the impact, all the lead atoms comprising the bullet traveled together, as a unit, because all had a single directed component of motion superposed on their uncoordinated thermal motions. *After* the impact, this directed component is randomized: when the bullet's gross motion vanishes, the constituent atoms acquire an increased energy of random thermal motion, which is reflected in the temperature rise. Observe that this molecular picture renders easily intelligible the striking disparity of the following two cases: (i) a moving bullet, when stopped, becomes hotter; and (ii) a stopped bullet, when heated, is *not* thereby set in motion. This otherwise puzzling asymmetry or unidirectionality grows out of a statistical situation amply familiar in everyday experience. Consider for example that a new deck of cards, factory-packed in a regular arrangement of suits and denominations, is soon randomized by shuffling; but we think it highly improbable that, by further shuffling, we will soon return the pack to its original highly-ordered arrangement." [98]

Consider a closed plastic bottle of water. If you squeeze it, the level of the water will rise in the bottle. If you let go, the level will sink back down to its former position. The squeezing is therefore what is called a *reversible process*. Smashing a glass onto the floor and seeing it break into pieces is called a *non-reversible process*. The difference between them lies in the energy that you have to put into reversing

the process. Clearly you can mend a broken glass – but only with effort. Restoring the squeezed bottle to its former state does not require exchange of energy between the system (bottle) and the external world (your hand).

In the real physical world, no action is truly reversible as there is always friction. For instance, when squeezing the bottle, you are actually transferring a small amount of energy to the water which manifests in a rise in temperature. In practical terms, this does not matter because the temperature increase is small but it is there nonetheless and so the process is not quite reversible. However, it is clear that the breaking of the glass is a lot less reversible than the squeezing of the bottle.

Thus, we ask ourselves for a measure of reversibility. This measure will be called entropy.

Suppose that the letter A labels a particular macrostate and $\Omega(A)$ denotes the number of microstates giving rise to that macrostate. Then we define the *Boltzmann entropy* to be $S(A) = k \ln \Omega(A)$ where k is a universal constant known as the *Boltzmann constant* the numerical value of which must be determined by experiments. This definition gives rise to the desirable fact that the entropy is additive. This means that if we make a larger system C by combining two systems A and B, then we have $S(C) = S(A) + S(B)$. While this is not a fundamental requirement of the universe, it is desirable because it makes life easier when computing and also simply makes sense that a system property would add when systems are added. Please note however that this is a definition and not the result of any argument!

The same definition leads to the fact that if macrostate A transits to a different macrostate A', then there could be a change in the entropy of $\Delta S = S(A') - S(A) = k \ln(\Omega(A')/\Omega(A))$. Note that, by the laws of logarithms, this change in the entropy can be both negative, zero or positive.

The second law of thermodynamics states that entropy must not decrease. However, this statement applies to an isolated system and not to a system that has energetic contact with other systems. In the physical world, true isolation is not possible but we can get very close in carefully constructed circumstances.

Let us consider the meaning of changes in entropy for a moment. If the entropy increases, this means that the number of microstates for the observed macrostate has also increased. Macrostates that have a larger number of microstates are more likely to be observed by the fundamental postulate (and please note that this is a postulate) that all microstates are equally likely to occur. Thus entropy increasing means that the system moves to a more likely state. We have a special term for the macrostate that has the highest probability of being observed, we call it the *equilibrium* of the system.

By contrast, if the entropy decreases, the system moves to less likely states. While this is of course allowed, it is by definition unlikely to happen. In colloquial terms, the effort that must be expended to get a system into a less likely state must come from outside the system and will incur an energetic cost in the outside that will lead to an increase in the entropy there. For example, the glass that broke when we dropped it can be fixed by effort expended by us on the system thereby increasing our body's entropy.

Practically, if we could plot the entropy over time of a real system, we would thus see small ups and downs everywhere from natural fluctuations but we should see a global trend upwards towards equilibrium. If this cannot be seen, then something happened: An external action was made that influenced the system towards a less likely state. A state of particularly low likelihood is the *ground state* of the system. This is the state of least energy in the physical world or the least cost function value in the computational world. Generally speaking, the ground state has very few microstates associated with it and often only a single one. Thus this state is very unlikely to be observed naturally. Yet, this is the state we wish to find when doing optimization calculations. In the context of optimization, the algorithm is the external world influencing the system and thus it is the algorithm that constitutes the heat bath.

Therefore, when plotting the entropy over time of a search for the ground state, we should see it decrease. It will generally not do so uniformly but rather have distinct local peaks and even discontinuities associated with it. These discontinuities are very distinctive and interesting features of the evolution because they are the telltale signs of so called *phase transitions*. In the physical world, a phase transition is the change from water to vapor or water to ice and vice versa where the phases are generally solid, liquid, gas and plasma. The ground state will generally be a solid state but we may have to begin searching for it with a system in the gaseous state and thus undergo two phase transitions (to liquid and then to solid). To make things more interesting, there are phase transitions of a more subtle nature, called *second-order phase transitions*, which are not visible in the entropy itself but rather in its first derivative. Formally speaking, a phase transition is a discontinuity in the entropy over time and a second order phase transition is a discontinuity in the derivative of entropy with respect to time.

To be clear, if we want to bring a physical system from a gaseous state to a solid state, we must cross two phase transitions. In cooling the system, we must be very careful at these points because local freezing may set in that will later make it impossible to reach the ground state without re-heating the system. Thus, we must cool very slowly at these points and hence there is a necessity to know where the phase transitions occur.

It is widely documented that phase transitions are observed in computational systems as well as natural ones. These are important events that must be negotiated carefully by the optimization algorithm "cooling" the system because systemic damages can occur during the change of phase. To illustrate this point, consider the freezing of a glass of water. Typically, the boundary will freeze first and slowly the freezing process will permeate to the center of the bit of water. If this occurs in the context of a fragile substance, such as a crystal, the parts that are on the boundary between the frozen and liquid portions experience a stress that may cause local damages in the crystal structure. These damages then freeze and so there is not enough energy anymore to repair the damages by natural fluctuations. This is the so called *freezing in* of local structures. Once this has been done, finding the true ground state is impossible without first heating the system up again and thus allowing the defect to be repaired. The same effect has been observed many times in combinatorial tasks

and so we must beware of phase transitions that will block the path to optimality if they are not negotiated carefully.

So what does negotiating carefully mean? It means two things. First, we must allow the system to cool (i.e. loose energy to the outside world) as slowly as possible so that local stresses are small and damages unlikely. Second, we must apply the cooling as uniformly across the spatial extent of the system as possible so that the the glass of water might freeze as a unit and not from the outside inwards. While being very difficult in physical terms, the second can be achieved more easily in a computational environment.

2.4 Temperature

The concept of temperature is fundamental to statistical mechanics as it is a major macroscopic state variable and closely related to the concept of equilibrium between two systems. The definition of *equilibrium* is[4]: Two systems in the thermodynamic context are in equilibrium with each other if and only if they share the same temperature. It makes little sense to speak of the temperature of a system that is far away from equilibrium as it would then be impractical to determine its temperature. Note that a physical temperature is measured by a thermometer (one system) being put into a substance (another system) and these two systems must be allowed to come to an equilibrium before it is sensible to take a reading of the temperature.

When we speak of something being hot, we mean the subjective impression that the object is transferring a lot of thermal energy to us and that we are experiencing a change in our state of being as a result. When you touch a hot stove plate, you burn your finger – a lot of energy has been transferred causing an increase in your body's entropy. This indicates a high temperature. When you touch an ice cube, some energy is removed from you and lowers your body's entropy as you move to a less probably state. This indicates a low temperature.

In terms of nature, temperature is thus a measure of the molecular agitation of some substance. If the molecules are more agitated, then the temperature is higher. Generally this also means that the pressure increases and thus the body will expand if it is allowed to do so leading to the well observed fact that objects get larger as they get hotter.

The *Temperature*, denoted by T, is defined by the derivative of the energy with respect to the entropy, $T = dE/dS$. Note that temperature is thus a defined concept in terms of the two basic concepts of statistical mechanics: energy and entropy.

We may ask how much heat (or energy) we must supply to a substance in order to increase its temperature by one unit. This is called the heat capacity of the substance. The heat capacity per unit of mass is the *specific heat capacity* or simply specific

[4] This is the definition of equilibrium *between* two systems. What concerns the equilibrium *of* a system, we have already encountered it: Equilibrium is the macrostate corresponding to the most microstates, i.e. the most likely macrostate.

heat of that substance. It is an important concept because it effectively translates between the concepts of energy and temperature.

The specific heat is a characteristic of a particular substance but it is not a constant. In fact, it depends upon temperature, volume and pressure. As the temperature gets large, the specific heat is approximately constant. The specific heat is particularly interesting as we approach absolute zero temperature or the ground state of the system. In nature, it follows *Debye's law* from quantum theory. Here, we have

$$c_v = 9Nk\left(\frac{T}{T_D}\right)^3 \int_0^{T_D/T} \frac{x^4 e^x}{(e^x - 1)^2} dx$$

where N is the volume (effectively a value needed for the physical value of c_v which we may ignore for the purposes of optimization), k is Boltzmann's constant (also a constant that may be ignored for optimization purposes) and T_D is the Debye temperature. The Debye temperature is a material property. For the purposes of optimization theory it may be estimated as that temperature where the specific heat (in the direction of lowering the temperature starting from a high temperature) is first observed to decrease significantly below an initially approximately constant value. In figure 2.1, we display the typical evolution of the specific heat according to the Debye model as compared to the related Einstein model. This is what we would expect to see in the evolution of an optimization problem and we may use this to interpret our progress.

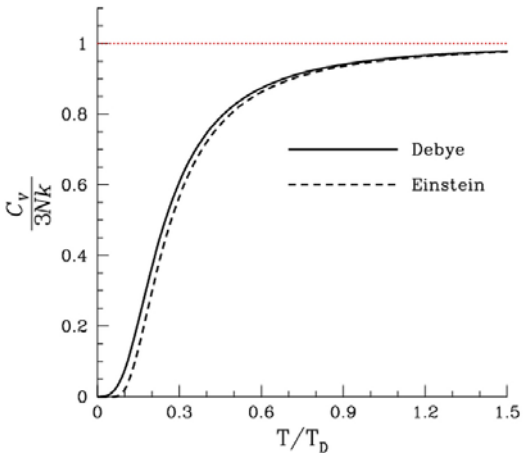

Fig. 2.1 The evolution of specific heat according to the Debye model as compared to the related Einstein model. We use this model to gauge our optimization progress while measuring the specific heat of our currently proposed problem solution.

The reason for studying specific heat in an optimization context is that it allows us to track our progress towards the ground state of the system to be optimized.

We may define two kinds of *specific heat*. The *specific heat at constant volume* c_v and the *specific heat at constant pressure* c_p are defined to be

$$c_v = T\left(\frac{\partial S}{\partial T}\right)_v,$$

$$c_p = T\left(\frac{\partial S}{\partial T}\right)_p$$

where the differential in both cases is made subject to the requirement that either the volume or the pressure respectively must remain constant.

2.5 Ergodicity

One of the most intriguing concepts in statistical mechanics is that of ergodicity. It is related, as discussed earlier, to the relationship between time-averages and ensemble-averages.

In statistical mechanics, we are primarily interested in the equilibrium state for a particular macrostate. We will thus want to know the value of some interesting quantity $G(\alpha)$, a function of the microstate α, in the equilibrium state and denote its value here by G_{eq}. As the system always tends to the equilibrium state, this value is equal to the time average of G for a long time, i.e.

$$G_{eq} = \lim_{t\to\infty} \frac{1}{t} \int_0^t G(\alpha_{t'})\, dt'.$$

The limit in time poses a practical problem. We cannot measure or compute such a limit in practice and so must look for an alternative means by which to obtain this result.

There is a related concept, which takes an average over local phase space (phase space is the set of all microstates), i.e. over microstates with a similar energy. We thus get

$$G_{ps} = \lim_{\Delta E\to 0} \frac{\int_{\omega(E)} G(\alpha) d\alpha}{\int_{\omega(E)} d\alpha}$$

where $\omega(E)$ is that region of phase space, i.e. the set of microstates with energies in $[E-\Delta E, E]$. This phase space average is thus an ensemble average. This is something that we can compute and measure.

We have restricted ourselves to a neighborhood in energy and energy is always an invariant of the motion[5]. If the energy, and functions of the energy, is the only invariant of the motion, the phase space is called *ergodic*. If there are other invariants, then the above phase space integral must be restricted to neighborhoods of these other invariants around the value of the current macrostate also.

The *ergodicity theorem* now states that if phase space is ergodic, then $G_{eq} = G_{ps}$.

This theorem allows us to replace something that we want to know but cannot measure or compute with something else that we can measure and compute. Thus it is very important to know whether phase space is ergodic or not in any particular case.

Let us analyze the situation by focusing on the concept of a Markov chain, which governs the evolution of the system according to our set of postulates. Recall that a Markov chain is a series of microstates where each microstate is arrived at from its predecessor in such a way that the probability to obtain any given microstate depends only upon the present microstate and not upon the history of previous microstates. Such a chain therefore has no memory and is thus particularly easy to model.

An observational state in our Markov chain is called *transient* if there is another state that the system can reach from this one but the system cannot return[6]. A state that is not transient is called *persistent*[7]. The persistent states can now be grouped such that two persistent states will belong to the same group if and only if they can be reached from each other. These groups are called *ergodic sets*. Essentially ergodic sets imply that we may more or less freely move between states within the same ergodic set but are effectively forbidden from going to another ergodic set.

Above, we spoke about the energy being an invariant of the motion and we thus having to restrict attention on states within an energy neighborhood. This is another way of saying that we must focus on one ergodic set of states. If we have more than one ergodic set, we must focus on one of them in order to perform our phase space integral and have it equal the time average of whatever value we are interested in.

So far, things are clean. However there can be a problem. We have differentiated transient and persistent states by the property of being able to return to them. As such the definition is precise. In practice, what also matters is how many transitions (i.e. how much time) are necessary for an eventual return. Sometimes it is quick to return and sometimes a return is possible only after a great many transitions. States

[5] The phrase "invariant of the motion" means that the value does not change over time. In the case of energy, this will not change as we have a law stating that energy is conserved.

[6] An example of a transient state is a hot cup of coffee at room temperature. This will naturally tend to cool down and so it can reach this cooler state. However, it will not be able to return to its hot state – unless it is acted upon by an external system such as a microwave oven.

[7] An example of a persistent state is that of a cup of coffee at the same temperature as the room in which it is located. This state may occasionally transit to other states but can and will return to this state of equilibrium.

that are persistent in principle but only after a time that is longer than our typical observational time period are called *pseudo-persistent*[8].

These will cause the splitting of an ergodic set into several subsets that are each an ergodic set for the realistic time-scale defined by our observational period. The existence of this effect is known as *ergodicity breaking* and represents a major computational problem. The problem has several features: (1) It is hard to know what states are pseudo-persistent in advance, (2) it is hard to diagnose ergodicity breaking when it happens and (3) the inherently long times necessary for a tour around the ergodic set increases the time needed for a reliable computation. During the use of an optimization algorithm, we start somewhere and then move from microstate to microstate until we believe to have found the optimum. This moving process could get stuck in one these pseudo-persistent areas and thus practically prevent our algorithm from exploring other areas. If the true optimum is in that other area, we are unlikely to find it. In the language of optimization, these points are called local minima (of sufficient depth and width to limit our evolution from going away for the observational duration).

Particularly for optimization purposes it is troublesome if we spend a very long time in a restricted section of the ergodic set without exploring the rest of it as it could be that the optimum we are looking for is in that rest. To have reasonable confidence that we will find the optimum, we must therefore increase the observational period. However, the effect of ergodicity breaking can occur on several time-scales and so the period may have to be increased by an impractical amount. Moreover, we cannot know how much we need to increase it by unless we can detect what is going on. In short, ergodicity breaking is a major stumbling block to efficient optimum finding and a response to it needs to be found.

What are appropriate responses? A very general strategy is called *restarting* in which we execute the optimization algorithm several times from different randomly selected starting points in the hope to get into all the pseudo-ergodic sets at least once. In fact, this is the response of choice in the field for a variety of optimization algorithms. Uniformly over the entire space of possible solutions, we select N of them and start a full optimization from these points. To save time, these N optimizations can be run in parallel as they do not interfere at all. We then take the best answer. It is observed that the answer quality improves approximately logarithmically. There is thus a law of diminishing returns as we crank up the effort put into a problem's solution. It is also observed that, for relatively low N, the gain is of the order of a few percentage points and so, in general, substantial enough to be worth the effort. We highly recommend augmenting any optimization algorithm with this simple method.

[8] A pseudo-persistent state is any state that takes so long to change that we are in danger of not seeing the change in our observational period. An example is the heating of water. If we supply heat to a large quantity of water, it takes a long time (respectively a lot of heat) to create even a small rise in temperature. Another example is a lake at high altitude. The water should flow down to reach a lower energy state but it is limited by the mountain. Eventually erosion will bring the mountain and thus the lake down but this takes a very long time.

http://www.springer.com/978-3-642-24973-0

Optimization for Industrial Problems

Bangert, P.

2012, XXII, 246 p. 64 illus., 30 illus. in color., Softcover
ISBN: 978-3-642-24973-0