Scaling, dimensional analysis, and rough physical reasoning

Matthew Headrick

Abstract: These are notes for the course PHYS 164a: First-year tutorial, at Brandeis University
Scaling relations and dimensional analysis are part of the lore of physics, applied daily by working physicists, yet rarely treated in a systematic way. Similarly, rough physical reasoning often allows us to understand the essential physics of a situation, and to derive scaling relations, without needing to solve a differential equation. Here we attempt to fill this gap in a pedagogical way.

Two excellent resources for this subject, with many more examples than we discuss here, as well as many other aspects of physics lore, are the books [1], [2]. Much of the treatment here is borrowed from or inspired by those sources.
1 Example: Mass vs. size

Scaling, or power-law, relations among physical quantities are ubiquitous in physics, and powerful tools for understanding physical phenomena across a large range of scales. A simple, familiar example is the relation between the mass \( m \), linear size \( l \), and density \( \rho \) of a cube:

\[
m = \rho l^3. \tag{1.1}
\]

Of course, a scaling relation may also include a numerical coefficient. For example, if we replace the cube by a sphere of diameter \( d \), we have

\[
m = \frac{\pi}{6} \rho d^3. \tag{1.2}
\]

In physics, this coefficient is often (but not always) “of order 1”, a loosely defined term that roughly means “somewhere between 0.1 and 10”. Indeed, in (1.2), the coefficient is around \( 1/2 \). More generally, we may not know or care what the numerically coefficient is; in that case we can write the relation using the symbol \( \sim \), which means “goes like” or “is proportional to”:

\[
m \sim \rho d^3, \tag{1.3}
\]

which means that the coefficient depends on some variables that we are not specifying but are holding fixed — in this case, the shape of the object.

A general scaling relation between two variables is of the form

\[
y = C x^\alpha. \tag{1.4}
\]

This becomes a linear relation in terms of the logarithm (in any base you might choose) of the variables:

\[
\log y = \alpha \log x + \log C. \tag{1.5}
\]

(Notice that the slope is independent of the base of the log, although the constant is not.) Therefore, such a scaling relation is easily recognized in data by means of a log-log scatter plot. This is easily generalized to several variables:

\[
y = C x_1^{\alpha_1} x_2^{\alpha_2} \cdots \Rightarrow \log y = \log C + \alpha_1 \log x_1 + \alpha_2 \log x_2 + \cdots. \tag{1.6}
\]

In physics, unless they are true by definition or because of a mathematical identity, scaling relations are almost always approximate, and arise after simplifying the physical situation by neglecting certain effects. Such approximate scaling relations often hold well over impressively large ranges. Consider, for example, the relation (1.1). For cubes, this relation holds exactly. But now suppose we treat the same relation instead as an approximate one, where we neglect the effect of the shape of the object and fix \( \rho \) to be, say, the density of water, \( \rho_w = 10^3 \text{ kg m}^{-3} \); then we have:

\[
m \approx \rho_w l^3. \tag{1.7}
\]

Exercise 1. On a log-log plot, plot the relation (1.7), along with the mass and size of each of the following objects: proton; hydrogen atom; coronavirus; E. coli bacterium; adult human; Empire State building; Halley’s comet; Earth; Sun; solar system; Milky Way galaxy; along with at least ten other objects of your choosing. (I recommend putting mass on the horizontal axis of your plot and size on the vertical one, since the mass varies much more than the length.)
Your plot should show that the relation (1.7) holds to an excellent approximation for common objects ranging from atoms all the way to stars — in all, spanning 19 orders of magnitude in size and an incredible 57 orders of magnitude in mass! This is not a mathematical fact but a fact about the world, and a very non-trivial one. The validity of such a simple law across a huge range of scales is characteristic of physics (and essentially unheard of in any other field of knowledge).

Because they have a simple form, and are often valid over a large range of scales, scaling relations are incredibly powerful. As we will see, many scaling relations in physics can be understood in terms of dimensional analysis (although there exist much more subtle scaling relations, associated with so-called critical phenomena, that are not consequences of dimensional analysis, but have to do with the dynamics in the vicinity of certain kinds of phase transitions). Dimensional analysis fixes the powers, but not the coefficient. The coefficient, even if it cannot be determined theoretically, can be fixed by a single measurement, and the scaling relation can then make a huge number of predictions: an incredible amount of output for a small amount of input!

Understanding when and why a scaling relation holds is an very instructive exercise. Understanding when and why it fails is just as instructive. As a case in point, your plot should show that there are several kinds of objects that fall quite far from the line, both above and below it. Why do these objects not follow the same scaling relation as the others?

2 Example: Throwing a ball

Consider the following problem:

When I throw a ball directly upwards at a certain speed, it reaches a height of 5m. How high does it reach if I throw it upwards at twice the speed? Neglect air resistance and any other forces besides the Earth’s gravitational force.

There are of course many different ways to solve this elementary problem. Here are a few of them.

**Dimensional analysis**: The relevant (or potentially relevant) quantities in this problem are the initial speed \(v_0\), the gravitational acceleration \(g\), the mass of the ball \(m\), and the height \(h\). These quantities have the following dimensions, in terms of the basic dimensions \(M\) (mass), \(L\) (length), \(T\) (time):

\[
[v_0] = LT^{-1}, \quad [m] = M, \quad [g] = LT^{-2}, \quad [h] = L. \tag{2.1}
\]

There is no dimensionless combination of the inputs \(v_0, m, g\), where by combination we mean a product of powers. There is only one combination with the same dimensions as the output \(h\):

\[
[g^{-1} v_0^2] = L = [h]. \tag{2.2}
\]

Hence \(h \sim v_0^2\). Therefore doubling the initial speed quadruples the height.

**Rough physical reasoning**: Since the acceleration is constant, doubling the initial speed doubles the time until the ball stops. The height is proportional to the speed times the time; since both are doubled, the height is quadrupled.

Or, using energy conservation: Kinetic energy is proportional to speed squared, so doubling the initial speed quadruples the initial kinetic energy. Hence the potential energy (relative to the initial position) at the top, where the kinetic energy is zero, is quadrupled. Since the potential energy is proportional to the height difference, the height is also quadrupled.
Solving the equation: Newton’s laws say
\[ \dot{v} = -g, \]  \hspace{2cm} (2.3)
so
\[ v = v_0 - gt \]  \hspace{2cm} (2.4)
and
\[ z = z_0 + v_0 t - \frac{1}{2} gt^2. \]  \hspace{2cm} (2.5)
Solving for the time at which the \( v = 0 \), we have
\[ t_{\text{top}} = \frac{v_0}{g}, \]  \hspace{2cm} (2.6)
hence
\[ z_{\text{top}} = z_0 + \frac{v_0^2}{2g}. \]  \hspace{2cm} (2.7)
Hence doubling \( v_0 \) quadruples the height.

Happily, all the methods give the same answer: 20 m. More importantly, they all give the same scaling relation between the height and the initial speed, \( h \sim v_0^2 \). We also easily derive the facts that the mass is irrelevant, and that the height is inversely proportional to the gravitational force. In the second and third methods, we can see the physical reasons for these facts. Via dimensional analysis, they are inevitable consequences of the nature of the input and output variables. The fact that we ended up with a scaling relation was also a direct consequence of dimensional analysis.

3 Dimensional analysis

All physics students are taught from an early age that, after solving a physics problem, you should check your units: If the answer is supposed to be a length, but your answer has units of kilograms-squared per attosecond, then you have definitely done something wrong. Dimensional analysis (DA) elevates this check to a principle that can be used to predict, or at least constrain, the answer before solving the problem. This is because, in many cases, there is only one way, or only a few ways, to combine the inputs to obtain a quantity with the units that the answer must have.

In order to systematize this method, it’s useful to abstract the notion of a “unit” into the notion of a “dimension”. A dimension is a set of units that are interconvertible. For example, if an answer can be correctly expressed in meters, then it can also be correctly expressed in kilometers or feet (although one of them may be preferable on practical or esthetic grounds). Therefore, we lump all units of length into a single dimension, “length”. On the other hand, kilograms cannot be converted into meters (at least in conventional, everyday physics), so those are different dimensions. Some common dimensions in physics are mass, length, time, energy, speed, charge, current, and temperature. Of course, these are not necessarily independent; for example, speed is length divided by time. We use symbols for dimensions: \( M, L, T, E \) (energy), \( Q \) (charge), \( \Theta \) (temperature). We thus have for example \( E = ML^2T^{-2} \). We use square brackets to denote “dimensions of”, as in (2.1).

In section 10, we will go more deeply into the theory, meaning, and subtleties of DA. Until then, our focus will be on how to use it to help solve and understand physics problems. In this section, we lay it out as an algorithmic method.

The first step in applying DA to a problem is to list the relevant input and output quantities. This includes any relevant constants of nature \( (g, c, G, \hbar, k_B, \text{etc.}) \), so you have to ask yourself: Is
relativity important here? Gravity? Quantum mechanics? The fewer inputs you have the more useful DA will be, so, at least as a first approximation, you will want to strip away as many variables and constants as possible. For the problem of section 2, the inputs were \( v_0, g, \) and \( m \).

The second step is to determine the dimensions of your input and output quantities. (Of course, some may be dimensionless, such as angles and pure numbers. Some constants of nature, such as the fine structure constant, are also dimensionless.) For the above problem, these were given in (2.1).

The third step is to determine whether it’s possible at all to make a combination of the input quantities with the same dimensions as the output quantity.\(^1\) For example, suppose the only input is a length and the output is a speed; there is no way to make a speed out of a length alone. To put it another way, if we change the unit of time but not the unit of length, the numerical value of the speed will change, while the value of the length won’t. The only way out of this paradox is if somehow the numerical value of the speed doesn’t change. Indeed, there are three quantities whose numerical value does not change when we change the units in which they’re measured: 0 and \( \pm \infty \). Without knowing anything else about the problem but the dimensions of the input and output, we can conclude that the output is (1) independent of the input, and (2) either 0 or \( \pm \infty \). (Usually which option is correct is obvious.) This is an amazingly powerful conclusion! We will see an example of this in practice in the next section.

Now suppose, on the other hand, that there is a combination of inputs with the same dimensions as the output:

\[
[I_1^{\alpha_1} I_2^{\alpha_2} \cdots] = [O]. \tag{3.1}
\]

(This was the case for the problem of section 2, as we saw in (2.2).) Then the ratio is a dimensionless function of the inputs:

\[
\frac{O}{I_1^{\alpha_1} I_2^{\alpha_2} \cdots} = f(I_1, I_2, \ldots). \tag{3.2}
\]

The key mathematical fact that gives DA its power is the following:

* A dimensionless quantity can depend on dimensionful ones only through a dimensionless combination.*

The fourth step is thus to determine if there are any dimensionless combinations of the inputs. If not, \( f \) cannot depend on the inputs at all; it must be a constant! Calling that constant \( C \), we have

\[
O = C I_1^{\alpha_1} I_2^{\alpha_2} \cdots. \tag{3.3}
\]

(Again, this was the case for the problem of section 2.) You’ve discovered a scaling relation! Dimensional analysis does not tell you the value of \( C \). Often, but not always, it is of order 1. More dangerously, it may be 0 or \( \pm \infty \). Its sign is also not determined by DA. However, the value of the constant can be determined or estimated by a single honest calculation or experimental measurement.

On the other hand, if there are dimensionless combinations of the inputs, then \( f \) must be expressible in terms of them. Generally speaking, the more different dimensionless combinations exist, the less constrained \( f \) is, and therefore the less powerful DA is. In sections 4–7, we will focus on cases (like the one of the last section) with no dimensionless combinations. Then, in sections 8–9, we will turn to cases with exactly one dimensionless combination, and see that quite a bit of mileage can still

\(^1\)Note that an output may be dimensionless. In that case, there is always a combination of inputs that matches it dimensionally, namely 1. This is indeed a product of powers of the inputs, and is dimensionless — in fact, it is certainly the simplest such combination.
be extracted from DA, particularly when combined with other assumptions, such as the “dull function hypothesis” or knowing how the output depends on one of the inputs.\footnote{2}

In subsection 10.1, we will develop a systematic theory, in the language of linear algebra, for counting the number of dimensionless combinations in a given problem. The short version is this: If there are \( N_{\text{in}} \) inputs made of \( N_{\text{dim}} \) basic dimensions, then as long as \( N_{\text{in}} \leq N_{\text{dim}} \), they usually do not admit a dimensionless combination, whereas if \( N_{\text{in}} > N_{\text{dim}} \) then there is definitely at least one dimensionless combination. For example, in the problem of section 2, there were three inputs \((v_0, m, g)\) made of three dimensions \((M, L, T)\), and indeed they did not admit any dimensionless combinations. However, this is not guaranteed; for example, if the inputs consisted instead of \( v_0, m, \) and \( c \), then we would have had the dimensionless combination \( v_0/c \). Also, if one of the inputs is itself dimensionless, then that already counts as a dimensionless “combination”.

Finally, we note that the basic logic of DA is essentially the same as what we use, for example, when we consider scalar and vector quantities: if a problem involves a vector input \( v \), and no other vectors, a scalar output can only depend on its norm \( v \), and not on its individual components \( v_x, v_y, v_z \). The reason is that the components depend on the choice of frame (or basis), whereas a scalar does not. In dimensional analysis, the role of “change of frame” is played by “change of units”. Thus, a dimensionless quantity cannot depend on our choice of units; the numerical value of a length can depend on our choice of length unit but not our choice of mass unit; and so on.\footnote{3}

4 Example: Quantum harmonic oscillator

Consider the following problem:

What is the ground-state energy of a quantum harmonic oscillator of mass \( m \) and spring constant \( k \)?

We will apply the same three methods as in section 2.

**Dimensional analysis:** The relevant quantities are \( m, k, \hbar \) (inputs), and \( E_0 \) (output), with the following dimensions:

\[
[m] = M, \quad [k] = MT^{-2}, \quad [\hbar] = ML^{2}T^{-1}, \quad [E_0] = ML^2T^{-2}. \tag{4.1}
\]

There are no dimensionless combinations of the inputs, and there is only one combination with the dimensions of the output:

\[
E_0 \sim \hbar k^{1/2}m^{-1/2}. \tag{4.2}
\]
Notice what would have happened if we had asked the same question about the classical harmonic oscillator. We would not have been able to use $\hbar$ as an input, and would not have been able to make a quantity with dimensions of energy from the remaining two inputs, $m$ and $k$. This immediately implies that a classical harmonic oscillator does not possess any intrinsic energy scale, and its ground state energy must be either 0 or $\pm \infty$ — in this case 0 (as we can see by taking the limit $\hbar \to 0$ in (4.2)).

**Rough physical reasoning:** The ground state is the state with the smallest expectation value of the total energy, which is the sum of the kinetic energy $p^2/(2m)$ and potential energy $kx^2/2$:

$$\langle H \rangle = \frac{1}{2m} \langle p^2 \rangle + \frac{k}{2} \langle x^2 \rangle = \frac{1}{2m} \left( \langle p \rangle^2 + \Delta p^2 \right) + \frac{k}{2} \left( \langle x \rangle^2 + \Delta x^2 \right).$$

To minimize this, we first set $\langle x \rangle = \langle p \rangle = 0$. We would like to also set $\Delta x$ and $\Delta p$ to zero, but we can’t do that without running afoul of the uncertainty relation, $\Delta x \Delta p \geq \hbar/2$. In effect, the uncertainty relation creates a tug-of-war between the kinetic and potential energies. The equilibrium occurs where the two terms are equal, that is at

$$\Delta x = \left( \frac{\hbar}{2(km)^{1/2}} \right)^{1/2}, \quad \Delta p = \left( \frac{k \hbar^{1/2}}{2m} \right)^{1/2},$$

yielding

$$E_0 = \frac{1}{2} \hbar \left( \frac{k}{m} \right)^{1/2}.$$

(Note that this method gets the right answer on the nose, including the factor of 1/2. That is a bit of an accident, due to the fact that the quantum harmonic oscillator ground state happens to saturate the uncertainty principle. More generally, this method only gives a lower bound on the energy.)

**Solving the equation:** The relevant equation is the time-independent Schrödinger equation, which can be solved exactly without too much trouble, giving again the answer (4.5). This shows that the numerical coefficient in (4.1) is 1/2, which is indeed “of order 1”.

## 5 Three methods

With the examples of sections 2 and 4 under our belt, we can step back and compare the three methods.

**Dimensional analysis (DA):** This is the most algorithmic. The only physics input is the choice of relevant quantities to include; after that, you turn the crank. Usually, once you’ve done it a few times, it’s very fast. It also explains why you do or do not get a scaling relation. It also immediately tells you that certain variables you might have thought are relevant (such as the mass in section 2) are actually irrelevant. And, if you add another input (e.g. considering the effect of air resistance), it tells you what you should compare the new input variable to in order to get a sense of its relative importance.

**Rough physical reasoning (RPR):** This is the least algorithmic of the three methods; it both requires and provides the most physical insight. The goal here is to obtain a correct general picture of the physics involved, without solving any complicated equations — especially any differential equations — or doing any integrals.
Solving the equation (STE): In most cases, this is a differential equation. It’s the most precise method: It will tell you the value of the coefficient that is undetermined by the other two methods. It is also the most generalizable: for example, if we add air resistance to the example of section 2, then we can get qualitative answers from RPR, but if we want a quantitative result we have to go back and re-solve the equation.

Of course, in practice these are not really separate methods, but rather general tools that can be applied and combined with each other and with other problem-solving tools as the situation requires.

6 Example: Hydrogenic atom

A hydrogenic atom consists of a nucleus and one electron. Such a model also serve as a crude but useful approximation to more complicated situations, such as the innermost electrons in a multi-electron atom, which are not very affected by the other electrons. Suppose we wish to find the binding energy $E_0$ (ignoring fine structure and other corrections). This example will show us that adding a just a little bit of physics knowledge can increase the power of dimensional analysis enormously.

DA: For inputs, we have the mass of the nucleus and electron, $m_n, m_e$; we have their charges $Ze, e$ (where $Z$ is the atomic number); electrostatics is important, so we have the constant $\epsilon_0$; and of course quantum mechanics is important, so we have $\hbar$. We immediately face a problem: too many inputs! Specifically, there are two dimensionless combinations of the inputs: $m_n/m_e$ and $Z$. This will seriously reduce the effectiveness of dimensional analysis. What to do? One perfectly reasonable answer is to fix those constants. First, since we know that the nucleus is much heavier than the electron, to a decent approximation it can be taken as infinitely massive. Second, we can just set $Z = 1$, since anyway hydrogen is the most important (to physicists anyway) hydrogenic atom (hence the name). Both of those are perfectly justifiable choices.

What we want to show, however, is that, with just a little more physics input, we can do much better. The question is how those inputs enter into the physics. First, we know that in 2-body dynamics, only the reduced mass, $\mu := \frac{m_n m_e}{m_n + m_e}$, matters. So we can replace $m_n$ and $m_e$ with $\mu$ — a savings of one input. Second, the only place the atomic number $Z$ enters is in the potential:

$$V = -\frac{Ze^2}{4\pi\epsilon_0 r}.$$  

(6.2)

Hence, in reality, only the combination

$$k := \frac{Ze^2}{4\pi\epsilon_0}$$

(6.3)

matters. So we can replace $Z, e$, and $\epsilon_0$ with $k$ — a savings of two inputs! The lesson is that if, rather than blindly listing the inputs, we carefully consider how they enter into the physical problem, we can often reduce their number, making dimensional analysis more powerful. It is particularly striking that we are able to “smuggle in” a dimensionless input, $Z$, by tying it to dimensionful ones.

Before we proceed, one comment about the factor of $4\pi$ in the denominator: this doesn’t affect the dimensions of $k$, so from the viewpoint of dimensional analysis, strictly speaking it doesn’t matter whether or not we include it. And that is true if we only want to derive the relevant scaling relation: how the binding energy depends on $Z$, for example. But if we are bolder and want to estimate the
binding energy, by guessing that the unknown dimensionless constant $C$ is “of order 1”, we would do better to include the $4\pi$ in the definition of $k$, as we did in (6.3).

From here, the DA is perfectly straightforward. We have

$$[\mu] = M, \quad [k] = ML^3T^{-2}, \quad [\hbar] = ML^2T^{-1}, \quad [E_0] = ML^2T^{-2}. \quad (6.4)$$

We see that $k/\hbar$ is a speed, so squaring and multiplying by $\mu$ gives us an energy:

$$E_0 = C \mu \frac{k^2}{\hbar^2} = C \mu \left( \frac{Ze^2}{4\pi\epsilon_0\hbar} \right)^2. \quad (6.5)$$

We learn that the binding energy is quadratic in the atomic number.

If we are very bold, we can guess that $k/\hbar$ is the average speed of the electron, hence the kinetic energy is $(1/2)\mu(k/\hbar)^2$; applying the virial theorem, which says that the binding energy equals the kinetic energy, then tells us that $C = 1/2$ — spot on! Even if this is a bit dodgy as a derivation, it is handy as a mnemonic: you are now empowered to derive the Bohr energy in two lines, with no calculus.

The above was premised on the motion being non-relativistic; otherwise we would have to include $c$ as an input. But now that we have an expression for the speed, we can check that assumption. The relevant ratio is $v/c$, i.e.

$$\frac{k}{\hbar c} = Z\alpha, \quad (6.6)$$

where $\alpha$ is the fine-structure constant:

$$\alpha := \frac{e^2}{4\pi\epsilon_0\hbar} \approx \frac{1}{137}. \quad (6.7)$$

So the assumption of non-relativistic motion is good for hydrogen and other light elements, but for heavier elements there are significant relativistic corrections. (Here we’re talking about an ion with a single electron, but this also holds for the inner electrons of heavier elements, for which the nucleus is not screened by the other electrons.)

**RPR:** This is similar to the quantum harmonic oscillator. Again, we have a tug-of-war between the kinetic and potential energies, linked by the uncertainty relation: The potential energy is trying to drive the electron into the nucleus, while the kinetic energy is trying to spread the wave function out as much as possible. If the size of the ground state wave function is $a$, then the potential energy expectation value is roughly $-k/a$. The kinetic energy expectation value, $\langle p^2 \rangle / (2m)$ is determined by the uncertainty relation: $\langle p^2 \rangle = \Delta p^2 \approx (\hbar/a)^2$. To find the ground state, we minimize the sum:

$$\langle H \rangle \approx \frac{\hbar^2}{ma^2} - \frac{k}{a} \quad (6.8)$$

(where we’re dropping dimensionless coefficients). The minimum is at

$$a_{\text{min}} = C' \frac{\hbar^2}{\mu k}; \quad (6.9)$$

plugging it back into (6.8) gives agreement with (6.5). As a byproduct, we learn the size of the atom (which we could also have gotten from DA):

$$a_{\text{min}} = a_0 \frac{Z}{Z}, \quad (6.10)$$
where \(a_0\) is the Bohr radius,

\[
a_0 = C' \frac{4 \pi \epsilon_0 \hbar^2}{e^2 m_e} \tag{6.11}
\]

(and where we’ve neglected the small dependence of \(\mu\) on the nuclear mass). Hence, as you might expect, the larger the atomic number, the smaller the atom.

**STE:** This is the standard textbook derivation, giving \(C = 1/2\) in the binding energy (6.5) and \(C' = 1\) in the Bohr radius (6.11). So all that work, setting up and solving a partial differential equation, gives you just two numbers — numbers that you might have guessed anyway! In reality, however, there are many more insights to be gained from the differential equation. For example, the energies of the excited states . . . or could you perhaps have figured those out using DA or RPR as well?

## 7 Exercises

**Exercise 2.** Write down the dimensions of the following kinds of physical quantities in terms of the basic dimensions \(M, L, T, Q, \Theta\): energy; force; power; velocity; acceleration; momentum; angular momentum; torque; mass density; pressure; electric field; electric potential; electric current; capacitance; magnetic field; entropy; heat capacity. Add ten more types of quantities. Similarly for the following constants: \(c, \hbar, G, \epsilon_0, \mu_0, k_B\). Record these in a convenient place for future problem-solving!

**Exercise 3.** Solve each of the following problems using all three of the above methods: DA, RPR, and STE (except where indicated otherwise). These problems are chosen to lie in the sweet spot where (1) there is a combination of the inputs with the dimensions of the output, and (2) there is no dimensionless combination of the input quantities. Hence the answer in every case will be a scaling relation. In each case, \(k\) is a positive constant.

a. Find the stopping distance and time of an object of mass \(m\) with initial speed \(v_0\) subject to a drag force proportional to its speed, \(F = -kv\).

b. Same, but \(F = -kv^2\). You may find a strange answer from dimensional analysis. (In section 9.4, we’ll discuss the regimes in which this drag force and the one of exercise 3a. are applicable.)

c. Find the relation between the amplitude and the period for a particle of mass \(m\) moving in one dimension subject to a potential \(V = \frac{1}{\alpha} k|x|^{\alpha}\), for general real \(\alpha\). For STE, unless you’re brave, only do the important special cases \(\alpha = -1\) (Coulomb potential) and \(\alpha = 2\) (harmonic oscillator).\(^4\)

d. Find the relation between the period and radius for a circular orbit subject to a central potential \(V = -k/r\).

e. Find the position uncertainty and energy of the ground state for a quantum particle of mass \(m\) in a piecewise linear potential, \(V(x) = k|x|\).

\(^4\)For general \(\alpha\), the numerical coefficient can be calculated in STE by a definite integral that gives a fairly complicated answer:

\[
4 \int_0^1 du \sqrt{\frac{\alpha}{1 - u^\alpha}} = \begin{cases} 
2\sqrt{-2\pi \alpha} \frac{\Gamma\left(\frac{\alpha}{2} - \frac{1}{2}\right)}{\Gamma\left(-\frac{1}{2}\right)}, & \alpha < 0 \\
2\sqrt{2}\frac{\pi}{\alpha} \frac{\Gamma\left(\frac{\alpha}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}, & \alpha > 0
\end{cases} \tag{7.1}
\]

This function is actually smooth across \(\alpha = 0\), although I have not found a way to write it that manifests this smoothness.
f. Find the drag force on an object of linear size $a$ moving at speed $v$ through a fluid of density $\rho$, neglecting viscosity, compressibility, and other effects. Only do DA and RPR. (We will return to the drag force in subsections 8.1 and 9.4.)

g. Consider a U-shaped tube partially filled with an incompressible liquid, so that the top of the liquid is in vertical sections of the tube. The liquid can oscillate, going up on one side and down on the other, and then vice-versa. Neglecting viscosity, friction with the side of the tube, and other damping effects, find the frequency of small oscillations. For DA, assume that the only relevant length in the problem is the length of the part of the tube with water in it; for RPR, argue that this is the case. (Hint: for any harmonic oscillator mode, the frequency is the square root of the coefficient of the potential energy to the coefficient of the kinetic energy:

$$K = \frac{1}{2} \mu x^2, \quad V = \frac{1}{2} \kappa x^2 \quad \Rightarrow \quad \omega = \sqrt{\frac{\kappa}{\mu}}.$$  \hspace{1cm} (7.2)

h. Find the speed of a wave on the surface of deep water as a function of its wavelength. Which travels faster, short- or long-wavelength waves? Which is larger, the phase or the group velocity? (We will return to water waves in subsection 9.3. Hints: It is recommended to start by finding the dispersion relation — the frequency as a function of the wavenumber — and to compute the group and phase velocity from that. Deep water means you do not have to include the depth of the water as a variable. You should also exclude surface tension, viscosity, compressibility, and other effects except gravity. You should also work in the linear regime, i.e. assume that the speed is independent of the amplitude. For RPR, the result of exercise 3g. may be useful. For STE, assume the flow is incompressible and irrotational.)

i. Find the energy density of electromagnetic radiation in thermal equilibrium at temperature $T$. (This is closely related to the Stefan-Boltzmann law. Hint for DA: this problem involves quantum mechanics, relativity, and statistical mechanics.)

j. Find the radius of a black hole as a function of its mass. This gives you a scaling relation between size and mass, different from (1.7), that you can add to your plot from exercise 1. (Hint for DA: this is a relativistic gravitational problem. Hint for RPR: the radius of a black hole is the point at which the escape velocity equals the speed of light. Hint for STE: you may look up the answer in a general relativity textbook!)

8 Examples with known dependence on an input

We saw in section 6 about hydrogenic atoms that DA can become much more powerful if we know something about how the output depends on the inputs. There, by considering the reduced mass rather than the separate masses of the electron and nucleus, and the combination $Ze^2/(4\pi\epsilon_0)$ rather than the separate factors, we managed to eliminate all dimensionless combinations from the problem. Here, we consider what happens when we know a priori how the output depends on one of the input. Again, as we will see, this knowledge supercharges DA.

8.1 Drag force I

For our first example, we reconsider the drag force explored in exercise 3f. There, on the basis of dimensional analysis and rough physical reasoning, you found a force that was a function of the linear size $a$ of the object and the speed $v$ of the fluid, and the density $\rho$ of the fluid. We will now find the drag force as a function of the Reynolds number $Re = \frac{va}{\nu}$, where $\nu$ is the kinematic viscosity of the fluid. (We will return to this problem in subsection 9.4.)
size $a$ and speed $v$ of the object and the density $\rho$ of the fluid:

$$F \sim \rho a^2 v^2.$$  \hfill (8.1)

However, this formula does not take into account the fluid’s viscosity. We might expect that, at least in some regimes, viscosity plays an important role in drag. What happens to dimensional analysis when we include viscosity as one of the inputs? Kinematic viscosity has dimensions $L^2T^{-1}$. Unfortunately, with our other inputs, this leads to a dimensionless combination, one that’s important enough to get its own name, the Reynolds number:

$$Re := \frac{av}{\nu}.$$  \hfill (8.2)

From the viewpoint of dimensional analysis alone, the drag force can depend on $Re$ in an arbitrary way.

But let’s bring in a little more physics knowledge. Viscosity is a form of friction, and, at least in some regimes, is ultimately the reason for drag. So it seems reasonable to assume that the drag force is proportional to the viscosity. This immediately fixes the dependence on the rest of the inputs:

$$F \sim \nu \rho a v.$$  \hfill (8.3)

In subsection 9.4 below, we will give a RPR argument for (8.3), which will also clarify the regimes in which it and (8.1) are valid.

### 8.2 Scattering angle

Consider a particle scattering off a central power-law potential $V = \frac{1}{\alpha} kr^\alpha$, where we assume $\alpha < 1$ so the force falls off at large distances. We allow $k$ to be positive (attractive force) or negative (repulsive).

Can we use DA to compute the scattering angle $\theta$?

In all of the problems we’ve studied so far, the output was a dimensionful quantity. There’s a reason for that. Suppose the output is dimensionless. Then either the inputs do not admit a dimensionless combination, in which case the output cannot be expressed in terms of them, or they do, in which case the output is an arbitrary function of this dimensionless combinations, so DA is useless.

But if we have more information — for example, how the output depends on one of the inputs — then this conclusion doesn’t necessarily hold. Here, for example, the scattering is caused by the potential, and therefore it seems reasonable to suppose that, at least for small scattering angles, $\theta$ is proportional to $k$. This guess is backed up by the fact that the sign of the scattering angle correlates with the sign of $k$: if the potential is attractive, the particle scatters towards the center, which we will call $\theta > 0$, while if it is attractive then $\theta < 0$. Given this assumption, and with our other inputs being the particle’s mass $m$, initial speed $v$, and impact parameter $b$, DA now gives a unique answer:

$$\theta \sim \frac{kb^\alpha}{mv^2}.$$  \hfill (8.4)

We can understand this answer with RPR: The particle is subject to a force $F \sim kr^{\alpha-1}$, and therefore an acceleration $a = kr^{\alpha-1}/m$. Most of the acceleration occurs when it is within a distance about $b$ of the center; it spends a time $t \sim b/v$ in that range, hence it acquires a net change in velocity transverse to its original direction of motion of $v_t \sim at \sim kb^\alpha/(mv)$. The scattering angle is $v_t/v$, again giving (8.4).
Exercise 4. Formalize the above RPR argument to compute the numerical coefficient in (8.4) as a function of $\alpha$. This requires doing one integral. You should find $B\left(\frac{1}{2} - \frac{\alpha}{2}, \frac{1}{2}\right)$, where $B$ is the Euler beta function.

The scattering angle can be found exactly by STE only for the case $\alpha = -1$, which is Rutherford scattering. There, the textbook derivation yields

$$\theta = 2\arctan\left(\frac{k}{mv^2b}\right). \quad (8.5)$$

Expanding the arctan to first order shows that in this case the numerical coefficient in (8.4) is 2.

9 Examples with one dimensionless combination

So far we have restricted our considerations to the simplest and most powerful setting for DA, which is when there is no dimensionless combination of the inputs. (Or, in the cases treated in the previous section, no dimensionless combination of the remaining inputs once the dependence on one of the inputs is fixed.) We now turn to the second-simplest and second-most powerful setting, when there is exactly one dimensionless combination. As a result, rather than leaving undetermined a constant undetermined, DA leaves a function undetermined. It is nonetheless enormously powerful, as very different physical situations can be mapped onto each other as long as they share the value of the dimensionless combination. As we’ll see, we can often go even further by making an educated guess about the unknown function. Now that we’ve gained considerable experience with the three methods — DA, RPR, and STE— we will be a little more informal and will not separate them out in the narrative.

9.1 Planetary orbits

We begin with the famous problem of determining the period of a Keplerian orbit. Two bodies orbiting each other gravitationally are equivalent to a single body of mass $m = m_1m_2/(m_1 + m_2)$ subject to the potential $V = -k/r$, where $k = Gm_1m_2$. In exercise 3.d., you computed, in three different ways, the following relation between the period $\tau$ and radius $R$ of a circular orbit subject to this potential:

$$\tau = C\left(\frac{mR^3}{k}\right)^{1/2}. \quad (9.1)$$

You should also have determined the value of the constant: $C = 2\pi$.

Now suppose we consider more general orbits. Solving the full Newtonian equations is a bit complicated but certainly doable, but suppose we want a shortcut. A non-circular orbit is characterized by two distances: the perihelion $R_p$ (closest distance to center) and aphelion $R_a$ (furthest distance to center). Their ratio $R_p/R_a$ gives us a new dimensionless quantity, which by definition ranges between 0 and 1. The relation (9.1) is still valid, with two amendments: (1) The dimensionless coefficient $C$ is now some (dimensionless) function $f(R_p/R_a)$. (2) For $R$ we need to choose $R_a$, $R_p$, or some combination of them (with dimensions of length). All choices are equally valid, as a different choice can be absorbed into a different function. For example, suppose we chose $R_a$; then we will have, instead of (9.1),

$$\tau = f_a(R_p/R_a)\left(\frac{mR^3}{k}\right)^{1/2}. \quad (9.2)$$
On the other hand, if we chose \( R_p \), we would have

\[
\tau = f_p(R_p/R_a) \left( \frac{mR_a^3}{k} \right)^{1/2}.
\]  

(9.3)

These are equivalent, with the functions \( f_a, f_p \) being related as follows:

\[
f_a(x) = x^{3/2} f_p(x).
\]  

(9.4)

Of course, some choices will lead to a simpler formula than others.

To be democratic, let’s use the average \( R := (R_a + R_p)/2 \) (also known as the semimajor axis), and write

\[
\tau = f(R_p/R_a) \left( \frac{mR_a^3}{k} \right)^{1/2}.
\]  

(9.5)

Although it contains a (so far) unknown function, this is already an incredibly powerful equation. It predicts the period of any orbit with some value of the ratio \( R_p/R_a \) in terms of the period of any other orbit with the same ratio — regardless of the size of the orbit or the values of \( k \) and \( m \! \).  

Can we get more information about the function \( f \) without doing too much work? A useful trick is to look for easy cases — which are often limiting cases, lying at the ends of the domain. At one end, where \( R_a = R_p = R \) (circular orbit), we’ve already worked out that the coefficient is \( 2\pi \):

\[
f(1) = 2\pi.
\]  

(9.6)

What about at the other end, \( f(0) \)? In this case, the aphelion is 0, and the motion reduces to a purely radial orbit that shoots out from \( r = 0 \), goes out to \( r = R_a \), and falls back to \( r = 0 \). Luckily, you already solved this problem, in exercise 3c., specializing to the case \( \alpha = -1 \) (and taking only half the period, since a full oscillation in that problem would go from \(-R_a\) to \( R_a \) and back). That answer was:

\[
\tau = \pi \left( \frac{mR_a^3}{2k} \right)^{1/2} = 2\pi \left( \frac{mR_a^3}{k} \right)^{1/2}.
\]  

(9.7)

Well, what do you know? \( f \) is the same at 0 as at 1:

\[
f(0) = 2\pi.
\]  

(9.8)

At this point, we appeal to what Tony Zee calls the “dull function hypothesis” [2]: most dimensionless functions in physics are as simple as they can be, given the constraints they must satisfy. Since there is no particular reason to think that our function \( f \) should increase or decrease between 0 and 1, the simplest guess is just a constant:

\[
f(x) = 2\pi.
\]  

(9.9)

The honest derivation of Kepler’s laws from Newton’s laws is standard textbook material, so we won’t repeat it here. Indeed, the result vindicates our bold guess:

\[
\tau = 2\pi \left( \frac{mR_a^3}{k} \right)^{1/2}.
\]  

(9.10)

This is a good general lesson: look for the simplest curve that fits the data points.
9.2 Artillery range

For our next example, we quote from [2], chapter 1.4:

Here is a freshman physics level problem. A cannonball is launched with speed $v$ at an angle $\theta$. What is its range $R$? While the problem is elementary, it has kept countless physicists in the favor of kings and generals throughout human history. Many eminent physicists have served as artillery officers.

Since $[v] = LT^{-1}$ and $[g] = LT^{-2}$ (the mass of the cannonball cancels out), it follows almost instantly that

$$R = \frac{v^2}{g} f(\theta)$$  \hspace{1cm} (9.11)

with some unknown function $f(\theta)$.

Well, if you fire horizontally, the cannonball would just plop down on the ground. And if you are so idiotic to fire straight up, watch out! You are eliminated from the gene pool. Thus $f(0)$ and $f(\pi/2)$ vanish. A simple guess would be that $f(\theta)$ equals the product of a function that vanishes at $\theta = 0$ and another that vanishes at $\theta = \pi/2$, and if you have any physics sense, you would reject something like $\theta(\pi/2 - \theta)$. Trigonometric functions enter when you decompose a vector into its vertical and horizontal components. Hence you guess $\sin \theta \cos \theta$, or up to some overall factor of 2, $\sin 2\theta$.

As in the planetary orbit example, the unknown function is equal at the two endpoints; but this time we know enough not to guess a constant function.

9.3 Water waves

Surface waves on water are a fascinating and very complicated subject, involving both fluid dynamics and surface physics (surface tension, interaction with the air, etc.). To get a handle on this system, we will simplify it brutally. First, we will treat only the simplest case of small-amplitude waves, which can be treated within the linear regime, on the surface of still water. We will also neglect viscosity, compressibility, and surface physics. This leaves as inputs just the depth $d$ (which we take to be constant), wavenumber $k$, and gravity $g$, and for the output the frequency $\omega$. (The dependence of $\omega$ on $k$, the dispersion relation, will then give us the group and phase velocities.) We could have included also the density $\rho$ of the water, but by now we’re pro enough to know that, as the only quantity involving a mass, it cannot enter. Or, to put it another way, the mass entering into the gravitational restoring force cancels the inertial mass, as with a pendulum. Within this stripped-down system, it is possible to solve the equation; but more importantly, we will be able to understand several important features of real-world water waves using just DA and RPR.

DA immediately gives us the following form for the frequency:

$$\omega = f(kd) \sqrt{gk},$$  \hspace{1cm} (9.12)

with $f$ an undetermined function. As usual, we would do well to start by understanding the two limiting cases: deep water ($kd \gg 1$) and shallow water ($kd \ll 1$).

Deep limit: You studied the deep-water case in exercise 3h., and here is what you (hopefully) found. In this regime the depth is presumably irrelevant: for a wave of wavelength 1 m, it should hardly matter if the ocean is 1 km or 10 km deep. DA thus immediately gives

$$\omega \sim \sqrt{gk}.$$  \hspace{1cm} (9.13)
RPR to back up the result (9.13) can be obtained by thinking of the rising and falling columns of water as being connected by the U-shaped tube studied in exercise 3g. For the length of the tube we take the only length available, the wavelength \( \lambda = \frac{1}{k} \). Here the tube, which is assumed to have a constant cross-section, reaches a depth of \( \frac{1}{k} \); the water below that depth is assumed not to participate in the wave motion.

(9.13) gives the phase and group velocities as follows:

\[
v_g = \frac{1}{2} v_p \sim \sqrt{\frac{g}{k}} \sim \sqrt{g\lambda}
\]  

(9.14)

We learn that the wave disperses, and that longer-wavelength waves travel faster. This prediction is borne out by everyday experience; at sea, for example, short-wavelength waves travel slowly, as they bob up and down due to the passing of long-wavelength waves. On the other hand, the prediction that the (phase or group) velocity goes to infinity in the limit of long wavelength seems unphysical. Presumably the speed gets cut off somehow at long wavelength; indeed we will see shortly that the cutoff is provided by the depth.

Below we will solve the equation for the general case, and find that the numerical coefficient in (9.13) is equal to 1. This accords with experience: in the deep ocean, a wave of wavelength 6 m (\( k = 1 \text{ m}^{-1} \)) has a period around 2 s (\( \omega = 3 \text{ s}^{-1} \)).

**Shallow limit:** In the shallow-water limit, the depth is a relevant input, so we can no longer rely on DA alone.

For RPR, consider a standing wave, with exactly one wavelength, in a tub of length \( \lambda \sim \frac{1}{k} \), width \( w \), and depth \( d \), and let \( h \) be the height of the water, at a given time, above its equilibrium level. As the wave oscillates, the surface goes up in one half of the tub and down in the other. Approximating the movement as uniform in each half, the rate at which water moves from one half to the other is \( \sim \lambda w \dot{h} \) (since the surface area is \( \sim \lambda w \)). We’ll approximate the horizontal flow as being uniform throughout the tub. (This is where the shallowness assumption enters; if the depth is comparable to or larger than the wavelength, then the flow will certainly depend on the vertical position. In particular, as we discussed above for deep water, the flow is essentially zero at depths larger than the wavelength.) This water travels through a tube of cross section \( \sim dw \); therefore, its speed is \( \sim \lambda \dot{h} / d \). The total volume is \( \sim \lambda dw \), hence the total mass is \( \sim \rho \lambda dw \), hence the kinetic energy in this motion is

\[
K \sim (\rho \lambda dw)(\lambda \dot{h} / d)^2 = (\rho \lambda^3 w / d) h^2 .
\]  

(9.15)

The potential energy, meanwhile, is given by the mass lifted, \( \rho \lambda wh \), times \( g \), times the height \( h \):

\[
V \sim \rho g \lambda wh^2 .
\]  

(9.16)

According to (7.2), the frequency is the square root of the ratio of the coefficients in \( V \) and \( K \):

\[
\omega \sim \sqrt{\frac{\rho g \lambda w}{\rho \lambda^3 w / d}} \sim k \sqrt{g d} .
\]  

(9.17)

The exact solution below shows that, in fact, the numerical coefficient in (9.17) is 1. Even without knowing the coefficient, (9.17) makes a striking prediction: the dispersion relation is linear, hence all waves travel at the same speed. That speed (both group and phase velocity) is

\[
v = \sqrt{g d} .
\]  

(9.18)
An excellent example of a shallow-water wave, that is well within the linear regime, is a tsunami. In the open ocean, a tsunami has a wavelength of hundreds of kilometers, much larger than the ocean’s depth of several kilometers, and an amplitude of only a few centimeters. If we take 4 km for the depth, (9.18) gives

\[ v = 200 \text{ m/s} = 700 \text{ km/h}. \]  

This explains how a tsunami can cross an ocean in a matter of hours. Even more importantly, the lack of dispersion explains why it remains coherent. Since it is formed from a sudden and localized event (an undersea landslide or earthquake), a tsunami is a superposition of a wide range of wavelengths. If the dispersion relation were non-linear, the wave would spread out as it travelled, becoming harmless after travelling a long distance. Instead, the whole wave arrives at a distant shore at very close to the same time, to devastating effect.

**General depth:** For intermediate values of \( kd \), we unfortunately have to resort to STE. Luckily, it’s not too hard once we know what equation to solve. We are searching for a normal mode with a wavenumber \( k \). The means that the surface of the water is displaced from its equilibrium height by \( h(t) \sin kx \). The normal mode is a vector function \( \delta r(r) \), where the displacement of the water at position \( r \) and time \( t \) is \( h(t) \delta r(r) \). The normal mode must obey the following conditions:

- Its \( z \)-component equals \( \sin kx \) at \( z = d \) and equals 0 at \( z = 0 \).
- Because the water is incompressible, it is divergenceless.
- Because the flow is irrotational, it is curl-free. (The irrotational condition is due to the fact that we’re neglecting viscosity, so vortical modes decouple.)

These conditions reduce the problem to a fairly textbook boundary-value electrostatics problem, of finding the electric field in a slab with a prescribed normal component of the field on the top and bottom of the slab. The solution is easily found:

\[ \delta r = \nabla \phi, \quad \phi = \frac{\sin kx \cosh kz}{k \sinh kd}. \]  

To find the frequency of this mode, we use (7.2), which means we need the kinetic and potential energies. The velocity is \( \dot{h} \delta r \), so the kinetic energy (per period in the \( x \) direction and per unit length

\[ \dot{h} \delta r = \frac{1}{k^2 d} \left( \frac{\partial^2 g}{\partial x \partial t} \right), \quad \delta r = e^{i k x} \left( \frac{i}{k d} \hat{x} + \frac{z}{d} \hat{z} \right), \]  

where we’ve kept enough terms to see the \( z \)-dependence, and we switched the \( \sin kx \) to avoid fussing with sines and cosines. For an arbitrary superposition of waves \( h(x, t) = \int dk \tilde{h}(k) e^{ikx - i\omega t} \), we then get a very simple displacement field:

\[ \delta r(r, t) = \int dk \tilde{h}(k) \delta r(k, r) e^{-i\omega t} = \frac{1}{d} g(x, t) \hat{x} + \frac{z}{d} h(x, t) \hat{z}, \]  

where \( g \) is the \( x \)-integral of \( h \): \( \partial g/\partial x = h. \)
in the $y$ direction) is

$$K = \int_0^{2\pi/k} dx \int_0^d dz \frac{1}{2} \rho \dot{h}^2 (\delta r)^2$$

(9.23)

$$= \frac{1}{2} \rho \dot{h}^2 \int_0^{2\pi/k} dx \int_0^d (\nabla \phi)^2$$

(9.24)

$$= \frac{1}{2} \rho \dot{h}^2 \int_0^{2\pi/k} dx \ \phi \partial_z \phi_0^d$$

(9.25)

$$= \frac{1}{2} \rho \dot{h}^2 \int_0^{2\pi/k} dx \ \sin^2 kx \cosh kd \over k \sinh kd$$

(9.26)

$$= \rho \dot{h}^2 \frac{\pi \cosh kd}{2k^2 \sinh kd}$$

(9.27)

where in the third equality we integrated by parts and used the fact that $\phi$ is harmonic.

To compute the potential energy, we imagine a column of water of height $h \sin kx$ being lifted from the trough part of the wave ($0 < kx < \pi$) to the crest part ($\pi < kx < 2\pi$). This column is lifted a vertical distance equal to its own height, contributing $\rho gh \sin^2 kx$ to the potential energy. The total potential energy (per period in the $x$ direction and per unit length in the $y$ direction) is therefore

$$V = \int_0^{\pi/k} dx \rho gh \sin^2 kx$$

(9.28)

$$= \rho gh \frac{\pi}{2k}.$$  

(9.29)

According to (7.2), the frequency is then

$$\omega = \sqrt{\frac{g \tanh kd}{x}}.$$

(9.30)

Thus the function $f$ in (9.12), which was undetermined by dimensional analysis, turns out to be $\sqrt{\tanh x}$ — perhaps not the easiest function in the world to guess, but nonetheless one that nicely interpolates between $\sqrt{x}$ for $x \ll 1$ and 1 for $x \gg 1$. The phase and group velocities can easily be derived from (9.30).

### 9.4 Drag force II

Another example with one dimensionless combination of inputs is the drag force on an object (of fixed shape) moving in a fluid. Here, tools like DA and RPR are particularly valuable, because the actual equations are so complicated that solving them analytically is impossible, except for the simplest shapes. We will assume the fluid is incompressible, as this case is already complicated and interesting enough.

You made a first pass on the drag force in exercise 3f., neglecting viscosity. A crude estimate of the drag force on an object of cross-sectional area $A$ moving through a fluid of density $\rho$ at a speed $v$ is that the momentum per unit time exchanged between the fluid and the object is equal to the volume of fluid encountered per unit time, which is $Av$, times the density $\rho$, times the speed $v$:  

$$F \sim \rho Av^2.$$  

(9.31)

This is the only force that can be made from $\rho$, $v$, and $A$. Indeed, the dimensionless drag coefficient $c_d$ is defined by the relation

$$F = \frac{1}{2} c_d \rho Av^2.$$  

(9.32)
This analysis suggests that we should expect $c_d$ to be of order 1 and independent of $v$, $\rho$, and $A$ (since there is no dimensionless combination of these quantities).

However, the result (9.31) is suspect in two ways. First, the viscosity of the fluid does not enter, which is surprising because viscosity is ultimately the reason for drag. Second, the dependence on $v$ is non-analytic. Specifically, taking into account the direction of the force, since the drag force opposes the motion, (9.32) is

$$F = -\frac{1}{2}c_d\rho A v^2 \text{sgn}(v).$$  \hspace{1cm} (9.33)

The sgn function leads to a very suspicious non-analyticity at $v = 0$.

As we discussed in subsection 8.1, we could instead reasonably guess that the drag force is proportional to the viscosity. The kinematic viscosity $\nu$ has units of

$$\nu = L^2 T^{-1}. \hspace{1cm} (9.34)$$

Let $a$ be the length scale of the object, so $A \sim a^2$. The only force that can be made from $\nu$, $a$, $v$, and $\rho$ and that is linear in $\nu$ is

$$F \sim \nu \rho a v.$$  \hspace{1cm} (9.35)

The force is now linear, rather than quadratic, in the speed, which also cures the pathology with the sgn function in (9.33). This analysis predicts

$$c_d \sim \frac{1}{\text{Re}},$$  \hspace{1cm} (9.36)

where the dimensional Reynolds number Re is defined by

$$\text{Re} := \frac{av}{\nu}. \hspace{1cm} (9.37)$$

We can understand the two results (9.31), (9.35) physically — and therefore the regimes in which they apply — as follows. In laminar flow, the momentum transfer per unit area equals the dynamic viscosity $\rho \nu$ times the velocity gradient. The object is surrounded by a boundary layer in which the fluid velocity transitions from that of the object (set by the no-slip boundary condition) to that of the surrounding fluid. Let the thickness of the boundary layer be $\lambda$. Then the velocity gradient is $v/\lambda$, so the force per unit area is $\nu \rho v/\lambda$, and the force is

$$F \sim \frac{\nu \rho v a^2}{\lambda}, \hspace{1cm} (9.38)$$

or

$$c_d \sim \frac{\nu}{\lambda v}. \hspace{1cm} (9.39)$$

Now, what is $\lambda$?

Assume first that $\lambda \ll a$, so that the boundary layer is very thin compared to the object. Then the physics of the boundary layer should be local on the surface, and therefore it should not depend on $a$. The only length we can make with $\rho$, $\nu$, and $v$ is

$$\lambda \sim \frac{\nu}{v}. \hspace{1cm} (9.40)$$

Plugging this into (9.38) gives (9.31). We see why the viscosity dropped out: the thickness of the boundary layer is proportional to $\nu$, cancelling the $\nu$ in (9.38). We can also understand what goes wrong in the limit $v \to 0$: the boundary layer goes to infinity. This doesn’t make any sense: an object...
of size \(a\) is not going to lead to a change in the fluid velocity infinitely far away. A more reasonable assumption is that the object’s size \(a\) sets an upper limit on the thickness of the boundary layer. Thus, as \(v\) decreases, once it is smaller than \(\nu/a\), \(\lambda\) stops increasing, and becomes a constant of order \(a\).

Indeed, \(\lambda \sim a\) reproduces (9.35). In this regime, the velocity field surrounding the object scales in length with \(a\) and in velocity with \(v\), and is independent of \(\rho\) and \(\nu\).

Thus (9.35) applies at low speeds and (9.31) at high speeds. So we should not worry about the non-analyticity of (9.31) at low speeds, since it does not apply there. To summarize the situation in terms of the relation between the dimensionless drag coefficient and the dimensionless Reynolds number, we have

\[
c_d \sim \max(1/\text{Re}, 1) \sim \begin{cases} 
1/\text{Re}, & \text{Re} \ll 1 \\
1, & \text{Re} \gg 1
\end{cases}.
\] (9.41)

In fact, the transition between the two behaviors happens gradually when the Reynolds number is between about 1 and about 10\(^3\) (i.e. when \(\lambda\) ranges from \(a\) to \(10^{-3}a\)).

At very high Reynolds number (around \(10^5\) or \(10^6\)) — high speed, low viscosity, or large object — the above picture breaks down, as the flow ceases being laminar and becomes turbulent. This results in much less efficient momentum transport, leading the drag coefficient to drop by almost an order of magnitude. Unlike the transition between the two laminar regimes discussed above, which is a smooth crossover, this is a sharp transition, called the “drag crisis”. A plot of \(c_d\) versus \(\text{Re}\) for a sphere, showing both the crossover and the drag crisis, can be found at https://kdusling.github.io/teaching/Applied-Fluids/ImageDisplay.html?src=DragSphere.

10 More on dimensional analysis

We conclude in this section with a discussion of various aspects of the theory and philosophy of DA, including the surprisingly fraught question of setting constants to 1.

10.1 Counting dimensionless combinations

As we defined the concept in section 3, a “dimension” is a set of interconvertible units, such as meters, kilometers, feet, miles, etc. Another way to define it is as a set of physical quantities that can be directly compared with each other; we can say that 1 mile is more than 1 kilometer, but it makes no sense (at least in the context of standard undergraduate physics) to say that 1 mile is more or less than 1 kilogram. Dimensions can be combined by taking products of powers, e.g. \(E = ML^2T^{-2}\). It’s useful to fix a set of “basic dimensions”, such that any dimension appearing in a problem (or area of physics) can be written in a unique way as a combination of the basic ones. (The uniqueness requirement means that there is no dimensionless combination of the basic dimensions.) This choice is not unique; for example, instead of \(M, L, T\), we could use \(E, L, T\).

We’ve emphasized that it makes a crucial difference how many dimensionless combinations are admitted by the inputs to a problem. In general, the more dimensionless combinations there are, the less powerful DA is. We’ve seen many examples, first with no such combinations, then with one. Usually the number can be determined by inspection, in other words by just looking for dimensionless combinations. However, it is not too hard to develop a general theory that in principle tells you how many there will be in any given situation.

To begin, suppose for concreteness that our basic dimensions are \(M, L, T\). Then the general dimension is of the form \(M^\alpha L^\beta T^\gamma\), where the exponents are real numbers, which we can assemble
into a vector $\vec{\alpha} = (\alpha_M, \alpha_L, \alpha_T) \in \mathbb{R}^3$. Instead of writing $[x] = M^{\alpha_M} L^{\alpha_L} T^{\alpha_T}$, we will use the shorthand
\[ [x] = \vec{\alpha}. \] (10.1)
Clearly this generalizes to other sets of basic dimensions; if there are $N_{\text{dim}}$ basic dimensions, then a general dimension can be represented as a vector in $\mathbb{R}^{N_{\text{dim}}}$.

Taking a combination of different quantities — in the sense of products of powers — results in a linear combination of their dimension vectors:
\[ [x_1^{\beta_1} x_2^{\beta_2} \cdots] = \beta_1 \vec{\alpha}_1 + \beta_2 \vec{\alpha}_2 + \cdots. \] (10.2)
In this sense, the dimensions of various quantities really should be thought of as living in a vector space. Furthermore, the “basic” dimensions (in the example above, $M, L, T$) are basis vectors in this space. Changing to a different set of basic dimensions (e.g. $L, T, E$) amounts to choosing a different basis in which to write the dimension vectors.

A dimensionless quantity is represented by the zero vector. So a combination of quantities is dimensionless if the corresponding linear combination of their dimension vectors vanishes. Therefore, given a set of $N_{\text{in}}$ input quantities $x_1, \ldots, x_{N_{\text{in}}}$, the number $N_{\text{com}}$ of independent dimensionless combinations equals the number of independent vanishing linear combinations of the vectors $\vec{\alpha}_1, \ldots, \vec{\alpha}_{N_{\text{in}}}$, which in turn equals the difference between $N_{\text{in}}$ and the dimensionality $N_{\text{span}}$ of the space spanned by the vectors:
\[ N_{\text{com}} = N_{\text{in}} - N_{\text{span}}. \] (10.3)
This linear algebra fact can be proven as follows. Since the vectors $\vec{\alpha}_n$ span a space of dimensionality $N_{\text{span}}$, there is a subset of $N_{\text{span}}$ of them that are linearly independent; call them $\vec{\alpha}_1, \ldots, \vec{\alpha}_{N_{\text{span}}}$. Each of the others (of which there are $N_{\text{in}} - N_{\text{span}}$), can be written as a linear combination of those:
\[ \vec{\alpha}_j = \sum_{i=1}^{N_{\text{span}}} \beta_i \vec{\alpha}_i; \] the difference between the two sides is a vanishing linear combination: $\vec{\alpha}_j - \sum_{i=1}^{N_{\text{span}}} \beta_i \vec{\alpha}_i = \vec{0}$.

In practice, the formula (10.3), which is officially called the *Buckingham Pi theorem*, is not that useful, since figuring out what $N_{\text{span}}$ is no easier than just looking for dimensionless combinations directly. However, since $N_{\text{span}} \leq N_{\text{dim}}$, we at least have a lower bound on $N_{\text{com}}$:
\[ N_{\text{com}} \geq N_{\text{in}} - N_{\text{dim}}. \] (10.4)
This is more useful since you definitely know $N_{\text{in}}$ and $N_{\text{dim}}$. Often, $N_{\text{span}} = N_{\text{dim}}$, in which case (10.4) is saturated.

Based on (10.4), we expect that the more basic dimensions — i.e. the larger $N_{\text{dim}}$ — the fewer dimensionless combinations, and therefore the more powerful DA will be. However, this is only true if (1) the dimensions are actually relevant, and (2) the number of inputs is fixed. To show that (1) is necessary, imagine adding electric charge $Q$ as a basic dimension to a problem that has nothing to do with electricity or magnetism. $Q$ will not appear in any of the inputs. Therefore, you will increase $N_{\text{dim}}$ but not $N_{\text{span}}$, which, according to Buckingham Pi, is what really matters. To better understand case (2), in the next subsection we will consider more deeply the question of what determines the allowed dimensions, and what happens when we change them.

### 10.2 Choices of dimensions, and setting constants to 1
Beyond the choice of a set of basic dimensions, the set of allowed dimensions is non-unique in a deeper sense, and can be adapted to the physical context. For example, in non-relativistic physics
we distinguish length from time. In relativistic physics, we may continue to do so, in which case our formulas will contain the speed of light \( c \). But we can also recognize that, in relativity, distances and times are on the same footing, and in fact are mixed with each other by Lorentz transformations, with the speed of light providing a universal way to relate them — which is why light-seconds and light-years are convenient units of distance in relativistic settings. Recognizing this fact, we can combine these two buckets of units into a single one. The speed of light then becomes a mere conversion factor, on the same footing as the factor 1000 m/km that can be used to convert between km and m. Conversion factors between equivalent units are ratios that equal 1:

\[
\frac{1000 \text{ m}}{1 \text{ km}} = 1. \tag{10.5}
\]

Similarly, if we consider the speed of light to be a conversion factor, then it necessarily equals 1:

\[
c = 1, \tag{10.6}
\]

so

\[
1 \text{ second} = 1 \text{ light-second} = 299,792,458 \text{ m}. \tag{10.7}
\]

In general, in order to unify two dimensions, we need to have a universal way to convert between quantities belonging to the two dimensions. Unifying the dimensions then amounts to setting this conversion factor equal to 1. Often, as with relativity, this happens as a result of a deeper understanding of some part of physics. A similar example concerns thermodynamics and statistical mechanics. In thermodynamics, we consider temperature and energy to be different kinds of quantities, but in statistical mechanics a temperature is understood as an energy, with Boltzmann’s constant \( k_B \) relating them. Therefore, we can choose to unify these dimensions, using units of energy to measure temperature, which amounts to setting \( k_B = 1 \).

To emphasize this point further, let us back up, historically speaking. During the Middle Ages in Europe, each town had its own measure for weighing goods. Since most goods were rarely transported from one town to another, it wasn’t particularly useful to have uniform measures.\(^6\) If bags of flour are never carried between Oxford and Cambridge, then the weight of a bag of flour in Oxford and the weight of a bag of flour in Cambridge are different kinds of quantities, and there’s no particular reason to measure them in the same units. If a great scientist travelled between the towns and noticed that the flour in the two towns is fundamentally the same stuff, and it would make sense to directly compare the weight of a bag of one to the weight of a bag of the other, she could do an experiment where she brought together an Oxford standard pound and a Cambridge standard pound and compared them. She could furthermore discover that the factor \( r \) relating them (say, \( r = 1.12 \text{ lb}_{\text{Ox}} / \text{ lb}_{\text{Cam}} \)) is the same no matter where or when she makes the measurement. We should then say that she “discovered and measured the fundamental constant of nature \( r \)”.

\(^6\)An interesting brief history of weights and measures from antiquity through the nineteenth century can be found in the article by Samuel Wesley Stratton, founding director of the U.S. National Bureau of Standards (now NIST), in the 1905 *New International Encyclopedia* [4]. To the point at hand, he says:

> With the fall of the [Roman] Empire and the rise of small principalities, a chaotic condition as to standards developed which extended through the Middle Ages and thereafter, until in Italy alone, as late as the end of the eighteenth century, there were over two hundred lengths called the foot. Every little dukedom and principality had its own standards of weight and measure, and the insignificant intercourse between these small towns did not suffer from these conditions, but the maritime leagues felt the need of common standards.
factor is obvious; it is actually a non-trivial fact about the world, just as non-trivial as the constancy of the speed of light in vacuum. Establishing it requires careful definitions and careful experimental procedures. (For example, if the respective standard pounds are made of different materials, that respond differently to temperature or humidity, then the ratio may seem not to be constant if the conditions are not carefully controlled.) Once it is established experimentally, however, we can shift our thinking to identify the two previously different kinds of quantities, at which point the Oxford and Cambridge pounds just become different units for the same dimension. The constant of nature \( r \) is then denoted to a mere ratio between units, like 1609 m/mi; in other words, we set \( r = 1 \). That shift in thinking is a choice we make for convenience, but it rests upon a non-trivial experimental fact about the world.

A deeper understanding of physics can also point in the other direction, leading us to distinguish dimensions previously considered the same. For example, in everyday life, mass and weight are considered interchangeable concepts, with the ratio 454 g/lb considered a unit conversion (i.e. equal to 1). However, in high school physics we teach that mass and weight are actually not the same kind of quantity, and the ratio 454 g/lb is actually a non-trivial constant \( 1/g \), where \( g \) is the gravitational acceleration at the surface of the Earth. As long as one stays on the surface of the Earth and one is not doing high-precision experiments, whether to distinguish between mass and weight is a choice, to be made based on convenience and conceptual considerations.

Here is another example, related to the direction, rather than the magnitude, of Earth’s gravitational field: In aeronautics and atmospheric physics, horizontal and vertical distance play very different roles and are therefore usually measured in different units (e.g. feet vs. nautical miles); hence it would make sense to maintain distinct dimensions, say \( L_h \) and \( L_v \) — in which case there would be a new “constant of nature” whose value is 6076 feet per nautical mile.

As yet another, perhaps somewhat silly, example, I have been told that in some countries it is taught that “volume” and “capacity” are two different kinds of quantities, with the former to be measured in cubic centimeters and the latter in liters (or maybe it’s the other way around; also, this story may be apocryphal). Students then have to memorize the value of the “constant of nature” 1000 cm\(^3\)/ℓ, which can be used to calculate the volume of a liquid that can be fit in a container given its capacity. It’s not clear to me why this would ever be a useful distinction, but it is a logically consistent one.\(^7\)

What does all this have to do with DA and the Buckingham Pi theorem? Let’s go back to the example of the speed of light. If we’re doing nonrelativistic physics, then length and time are different dimensions. If we’re doing relativistic physics, then we have a choice: we can keep length and time as separate dimensions, in which case we have a new input, \( c \); or we can combine them and set \( c = 1 \). In the first case, we’ve increased \( N_{in} \), while in the second we’ve decreased \( N_{dim} \). Either way the difference \( N_{in} - N_{dim} \) appearing on the right-hand side of (10.4) has decreased by 1. In other words, it makes no difference to DA which choice we make. Similarly, considering vertical and horizontal distance to be different dimensions increases the power of DA only if the conversion factor (6076 feet per nautical mile) doesn’t enter, i.e. only if they are truly distinct kinds of quantities.

Even if technically nothing is gained by considering horizontal and vertical lengths as separate kinds of length, we might nonetheless choose to do so if it fits our mental picture. As another example, \( \hbar \) can be considered as a conversion factor between energy and frequency and between wave number and momentum. A quantum mechanics textbook could therefore set \( \hbar = 1 \) from page 1, and use units

\(^7\)For further discussion about dimensionful versus dimensionless constants, and setting constants to 1, see [5] and references therein.
of $T^{-1}$ for energy and $L^{-1}$ for momentum. But I’ve never seen a textbook that does that. Why not? Because such an approach would break the connection between quantum and classical physics, which is useful not just as a pedagogical bridge, but also to understand the emergence of classicality. In fact, it’s often useful to pretend that $\hbar$ is a variable, allowing one to consider the limit $\hbar \to 0$; obviously that’s hard to do that if you’ve set $\hbar = 1$ at the outset. More on setting constants to zero in the next subsection.

10.3 Setting constants to 0

In the previous subsection, we emphasized that certain constants of nature are fundamentally no different from unit conversion factors in more mundane settings. In practice, however, there is something that distinguishes quantities like $c$, $\hbar$, and $k_B$ that are conventionally glorified by the label “constants of nature”. Namely, on human scales, they have very large or small numerical values; in SI units,

$$1/c \approx 10^{-8}, \quad \hbar \approx 10^{-34}, \quad k_B \approx 1/N_A \approx 10^{-24}.$$  \hspace{1cm} (10.8)

To emphasize the parallels among these quantities, we’ll refer to $1/c$ rather than $c$. ($N_A$ is Avogadro’s number, which, for these purposes, is the same as $1/k_B$; they are related by the “gas constant” $R = N_A k_B$, a dimensionful quantity of order 1 in human units.)

The fact that these quantities are very small on human scales has important implications for the history of physics and for human understanding of physics today. Specifically, humans lived, and to a large extent still live, in a world in which these constants are effectively zero, hence developed a consistent set of laws describing that world. For each one, we therefore have special names for the kind of physics in which it is zero or non-zero:

<table>
<thead>
<tr>
<th>constant</th>
<th>zero</th>
<th>non-zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/c$</td>
<td>nonrelativistic kinematics</td>
<td>relativistic kinematics</td>
</tr>
<tr>
<td>$\hbar$</td>
<td>quantum mechanics</td>
<td>classical mechanics</td>
</tr>
<tr>
<td>$k_B \approx 1/N_A$</td>
<td>thermodynamics</td>
<td>hydrodynamics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>statistical mechanics</td>
</tr>
</tbody>
</table>

Understanding how to incorporate each of these constants into the laws of physics, in other words how to move from the middle to the right column, was a major step forward in the history of physics. It is a remarkable fact that all of three of these steps were taken at almost the same time, namely the end of the nineteenth century and beginning of the twentieth.

10.4 In defense of Boltzmann’s constant

In subsection 10.2, we discussed the fact that, depending on the setting, one may choose to work in units in which certain constants are equal to 1; common choices include $c$ (in relativity), $\hbar$ (in quantum mechanics), or both (in quantum field theory). In my experience it is a fairly common view among physicists that Boltzmann’s constant $k_B$ somehow occupies a lower status than those constants. One occasionally hears opinions along the following lines: “One can choose to set $c$ or $\hbar$ to 1, but $k_B$ really is equal to 1.” To put it another way, given our knowledge of statistical mechanics, to claim that temperature and energy are different sorts of quantities is as silly as claiming that volume and capacity are different sorts of quantities. The fullest expression of this sentiment that I’ve seen is in Tony Zee’s book [2], chapter IV.1.8

8To avoid confusion with the generic numerical coefficient $k$ used in the rest of these notes, I’ve denoted Boltzmann’s constant as $k_B$, including in the quotation from [2].
Setting $\hbar$, $c$, and $G$ to unity, or not

The Planckian system of units amounts to setting $\hbar$, $c$, and $G$ to unity, but often, depending on which are of physics you work in (or which corner of the cube of physics you are on), it would be inconvenient, or even inappropriate, to do so. For instance, in electromagnetism, $\hbar$ and $G$ do not even enter. Setting $c = 1$ would be quite appropriate, but occasionally it would be useful to keep it around, for example, to show the weakness of the magnetic force compared to the electric force.

Particle physicists deal with relativistic quantum phenomena, and so routinely set $\hbar$ and $G$ to 1, but not $c$, which does not enter until one starts discussing quantum gravity. In theories of quantum gravity (for example, string theory), $G$ is routinely set to 1.

So far I pretty much agree. However, unlike Zee, I wouldn’t particularly privilege the constants $G, c, \hbar$. As I discussed in subsection 10.2, generally speaking it’s useful to use the same units to measure quantities that are often directly compared, otherwise it’s not, and this choice will depend on the context. I gave several other examples of constants that are sometimes set to 1 and other times not.

With this larger context in mind, let us return to Zee:

**Did we need to include $k_B$?**

And now I come to my pet peeve. No doubt that Boltzmann’s constant $k_B$ played a pivotal role in physicists’ struggle to understand the discreteness of matter. But now that the reality of the atom has long been established, $k_B$ should be retired. Temperature is an energy, period. Boltzmann’s constant $k_B$ is just a conversion factor between energy units and some quaint markings on a tube of mercury.

Of course I do not object to the fact that temperature is often measured in degrees, but then degree should be considered a unit of energy, just like the erg or the British Thermal Unit, albeit a rather peculiar unit. The constant $k_B$ could then be suppressed. Otherwise, why not introduce a fundamental constant called $\kappa = 2.54$ cm/in, measure in inches, and pepper formulas in physics with expressions like $Gm_1m_2/(\kappa r)^2$? The appearance of $k$ similarly stings my eyes.

We can imagine a world with $\hbar = 0$. Indeed, physicists lived in that world until Planck came along. Similarly, we can imagine worlds with $G = 0$, or $c^{-1} = 0$. But what would it mean to have a world with $k_B = 0$? Instead of filling glass tubes with mercury, we will them with a liquid whose coefficient of thermal expansion is infinite?

I have been surprised that distinguished physicists continue to write $k_B T$, where $T$ would have sufficed. Perhaps they are so unused to it that they think of $k_B T$ as a new letter in some exotic alphabet?

Postscript: Some colleagues who read this chapter in manuscript urge me to strengthen my rant and rave against $k$ even more. Why keep on writing $k_B$ after an entire century has passed and after our long sojourn in the quantum world? Is it merely to confuse some of the weaker students into thinking that $k$ has the same status as the three fundamental constants $G, c, \hbar$?

Apparently I’m one of the weaker students, since to me, $k_B$ has *exactly* “the same status as the three fundamental constants $G, c, \hbar$”. (Actually, as I discuss in subsection 10.5, if I were to give a special status to any of these four constants, it would be $G$.)
First we should be clear that, as illustrated by the volume-vs.-capacity example of subsection 10.2, we can always make artificial distinctions and thereby create new “fundamental constants”. So the question is not whether it’s logically possible to separate temperature and energy, and keep Boltzmann’s constant hanging around, but whether that’s useful to do so, either practically or conceptually. Zee claims it’s not. I think it is, on both counts.

The important physics point here is that thermodynamics is not the same thing as statistical mechanics. They are two different theories, each with a regime of applicability. One is about the macroscopic world, the other the microscopic world. Boltzmann’s constant (or, equivalently, the factor $N_A/\text{mol}$) is the conversion factor that relates them. If you are only doing statistical mechanics, then a temperature is just an energy and you may as well set $k_B = 1$. But in thermodynamics, which governs the macroscopic world, temperatures and energies are different kinds of quantities. Therefore, if you are relating statistical mechanics to thermodynamics, you need Boltzmann’s constant.

This is exactly the same as for $c$: If you are just doing relativistic physics, you might as well set $c = 1$. But to translate your results into a nonrelativistic language, where length and time are different kinds of quantities, you need the speed of light. Ditto for $\hbar$ and quantum vs. classical. The constants $k_B$, $c$, $\hbar$, and many others, are conversion factors needed to go from a context where two kinds of quantities are interchangeable to a context where they are not. Even the Earth’s gravitational acceleration $g$ is such a factor, which takes us from a context where mass and weight are interchangeable concepts (everyday life on the surface of the Earth), to a context where they are distinct (Newtonian physics).

And, unlike Zee, we can easily imagine a world in which $k_B = 0$. That was the world scientists lived in before the atomic hypothesis was proven, when matter was continuous, Avogadro’s number was infinite, there were no statistical fluctuations, and heat was governed by the laws of thermodynamics. In that world, temperature was not an energy, it was a separate quantity that was measured and studied on its own. And yes, in that world, mercury had a finite coefficient of thermal expansion — after all, people were using mercury thermometers long before they knew whether atoms existed or had any idea how big they might be. And that is the world that people still largely live their daily lives in today (and many scientists and engineers work in), just as people live their daily lives (and most scientists and engineers work) in a classical and nonrelativistic world.

There is a modern example that throws into sharp contrast the value of distinguishing between thermodynamics and statistical mechanics. It was understood in the 1960s that black holes obey a set of dynamical laws closely analogous to the laws of thermodynamics, with horizon area playing the role of entropy, surface gravity the role of temperature, etc. Bekenstein then proposed that this was not an analogy, that black holes really do have an entropy proportional to their horizon area, etc. Following up on this proposal, Hawking cleverly used the known statistical mechanics of electromagnetic fields to calibrate the temperature of the black hole, and thereby fix the constant of proportionality; specifically, he showed that

$$S = k_B \frac{c^3}{4G\hbar} \times \text{horizon area}. \quad (10.9)$$

The presence of $k_B$ in this formula has a profound consequence. It shows that this is a statistical entropy, an instance of $S = k_B \ln W$. Formulas for thermodynamic entropies do not have $k_B$. We therefore learn, in a roundabout way, that a black hole has a huge degeneracy of quantum states,

$$W = \exp \left( \frac{c^3}{4G\hbar} \times \text{horizon area} \right). \quad (10.10)$$

Had we chosen to set $k_B = 1$, we would still have been able to figure out that this is a statistical entropy, since it was derived ultimately from the statistical entropy of an electromagnetic field. By
keeping the $k_B$ explicit, however, there’s no need to make such an argument, or even to know how
(10.9) was derived; the mere presence of the $k_B$ immediately establishes its statistical nature.

10.5 What about Newton’s constant?

Newton’s constant $G$ is an interesting case. In general relativity, it relates mass or energy density
to spacetime curvature. When doing GR, it can be useful to set $G = 1$, in which case density has
dimensions $L^{-2}$, and therefore mass has dimensions $L$; essentially we’re using an object’s Schwarzschild
radius to measure its mass. (Here we’re setting $c = 1$, which is obligatory in GR, where space and
time are essentially interchangeable.)

That being said, from a fundamental physics viewpoint it is not obvious that $G$ should be set to 1.
This is quite different from the situation with $k_B$, $c$, and $h$ which (barring some unforeseen revolution
in physics) are definitely equal to 1 in any fundamental theory. Indeed, special relativity and quantum
mechanics are incredibly well tested, with no theoretical or experimental indications of any deviations.
Although our understanding of them (especially of quantum mechanics) may continue to evolve, it is
likely that both theories are exactly true. Statistical mechanics is really a theory of our knowledge,
and is therefore true almost by definition. (This is the one sense in which I am a bit sympathetic to
Zee viewpoint on $k_B$.) The degree to which these theories are well tested is reflected in the fact that,
in the SI unit system, the constants $k_B$, $c$, $h$ have exactly defined values.

On the other hand, not only is general relativity relatively poorly tested experimentally (for
example, the experimental value of $G$ has a relative uncertainty of almost $10^{-4}$), but on theoretical
grounds we already know that it cannot be exactly correct, since its own equations imply the existence
of regions of spacetime, near singularities, where those equations do not hold. Similarly, it cannot be
combined consistently with quantum mechanics without some modification.

In any fundamental theory, $G$ will still, of course, be dimensionless — the fundamental theory
can contain only dimensionless quantities — but it may not equal 1. $G$ may turn out to play a role
more similar to $e$, (minus) the electric charge of the electron. One can work in units where $e = 1,$
and indeed particle physicists conventionally give the charges of elementary particles in those units.
However, when it comes to giving a more fundamental description, in the language of quantum field
theory, it is much more natural and useful to set $h = c = 4\pi\varepsilon_0 = 1$, which leaves $e$ as a dimensionless
quantity, the square root of the fine structure constant, approximately equal to 0.085. (There is
another reason not to set $e = 1$: the strength of the electric force, in other words the effective charge
of the electron, depends on the energy used to measure it, at high energies, so it’s not even really a
constant.) Similarly (and contrary to what Zee says), in quantum gravity theories, $G$ is not always set
equal to 1. For example, in string theory, there is a more fundamental quantity, the string length $\ell_s$,
that can be set to 1; in those units $G$ becomes some dimensionless number, whose value is set by some
complicated and still not well-understood dynamics, involving quantum effects, extra dimensions, etc.

Given the set of fundamental constants $c, h, G$, one is tempted — as Planck was — to define a
“natural” set of units by setting them all to 1. In particular, the Planck length $\ell_P \approx 10^{-35}$ m, which
is the unit of length, seems to be a very special distance (similarly for the Planck mass, time, energy,
etc.). This conclusion from DA can be slightly strengthened by RPR: crudely combining quantum
mechanics and general relativity leads to the conclusion that quantum fluctuations in the spacetime
geometry become so large at the Planck length that geometry ceases to have any meaning. However,
this reasoning really just tells us that $\ell_P$ is a lower bound on the length scale where new physics
must appear that changes the nature of spacetime. For example, as mentioned above, in string theory
the new physics comes in at the string length $\ell_s$, which is necessarily larger than $\ell_P$ (and may be as
large as around $10^{-20}$ m); below the string length, the nature of spacetime changes fundamentally,
as even basic properties like the number of dimensions become probe-dependent or even meaningless. We actually have no direct evidence that anything special happens at the Planck length, since no experiment or observation comes close to probing it.

The above reasoning follows the conventional assumption in physics that equates “fundamental” with “short-distance” (or “ultraviolet”). Intriguingly, however, when it comes to quantum gravity, there is another possibility, supported by recent developments in holographic quantum gravity theories. In fact, $G$ does seem to play a universal role in quantum gravity, aside from its usual role as the strength of gravity, namely as a conversion factor between areas and entropies. Specifically, given an area $A$, the quantity $k_B A/(G \hbar)$ has units of entropy; this holds in any number $d$ of spatial dimensions, as long as we understand an “area” to be a quantity with dimensions $L^{d-1}$. We saw this for black holes in (10.9), but the phenomenon is much more general. Perhaps, in the future, a complete quantum theory of gravity will be understood in which that relation is “baked in”, spacetime is understood as an emergent construct that encodes entanglement and information, and the more familiar role of $G$ as a gravitational coupling will be derived. In fact, the Einstein equation has been derived from the assumption that certain areas represent an entropy $k_B A/(4G \hbar)$ [6–8]: these derivations were done within certain settings and with certain additional assumptions, but they may point the way to a broader truth. If this turns out to be true, then it will indeed be natural to set $G = 1$ (or perhaps $G = 1/4$). In string theory, $\ell_s$ would then be a derived dimensionless number — exactly the opposite of what we said above, that in string theory $\ell_s = 1$ and $G$ is a derived dimensional number.

Which viewpoint is correct, the ultraviolet or the holographic? It is possible that they will both turn out to be correct, just different ways of looking at the same thing. After all, nothing says that there must be only one natural set of units.

References


